Jaroslav Fořt · Jiří Fürst,
Jan Halama · Raphaèle Herbin
Florence Hubert *Editors*

Finite Volumes for Complex Applications VI Problems & Perspectives

FVCA 6, International Symposium,
Prague, June 6-10, 2011

Volume 2

Springer

Springer Proceedings in Mathematics

Volume 4

For further volumes:
http://www.springer.com/series/8806

Springer Proceedings in Mathematics

The book series will feature volumes of selected contributions from workshops and conferences in all areas of current research activity in mathematics. Besides an overall evaluation, at the hands of the publisher, of the interest, scientific quality, and timeliness of each proposal, every individual contribution will be refereed to standards comparable to those of leading mathematics journals. It is hoped that this series will thus propose to the research community well-edited and authoritative reports on newest developments in the most interesting and promising areas of mathematical research today.

Jaroslav Fořt • Jiří Fürst • Jan Halama
Raphaèle Herbin • Florence Hubert
Editors

Finite Volumes for Complex Applications VI Problems & Perspectives

FVCA 6, International Symposium,
Prague, June 6-10, 2011, Volume 1

Springer

Editors

Jaroslav Fořt
Czech Technical University
Faculty of Mechanical Engineering
Karlovo náměstí 13
Prague
Czech Republic
Jaroslav.Fort@fs.cvut.cz

Jiří Fürst
Czech Technical University
Faculty of Mechanical Engineering
Karlovo náměstí 13
Prague
Czech Republic
Jiri.Furst@fs.cvut.cz

Jan Halama
Czech Technical University
Faculty of Mechanical Engineering
Karlovo náměstí 13
Prague
Czech Republic
Jan.Halama@fs.cvut.cz

Raphaèle Herbin
Université Aix-Marseille
LATP
Laboratoire d'Analyse
Probabilités et Topologie
rue Joliot Curie 39
13453 Marseille
France
Raphaele.Herbin@latp.univ-mrs.fr

Florence Hubert
Université Aix-Marseille
LATP
Laboratoire d'Analyse
Probabilités et Topologie
rue Joliot Curie 39
13453 Marseille
France
Florence.Hubert@latp.univ-mrs.fr

ISSN 2190-5614 e-ISSN 2190-5622
ISBN 978-3-642-20670-2 e-ISBN 978-3-642-20671-9
DOI 10.1007/978-3-642-20671-9
Springer Heidelberg Dordrecht London New York

Library of Congress Control Number: 2011930263

Cover design: deblik, Berlin

Printed on acid-free paper

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

Editors Preface

The sixth International Symposium on Finite Volumes for Complex Applications, held in Prague (Czech Republic, June 2011) follows the series of symposiums held successively in Rouen (France, 1996), Duisburg (Germany, 1999), Porquerolles (France, 2002), Marrakech (Morocco, 2005), Aussois (France, 2008).

The sixth symposium, similarly to the previous ones, gives the opportunity of a large and critical discussion about the various aspects of finite volumes and related methods: mathematical results, numerical techniques, but also validations via industrial applications and comparisons with experimental test results.

This book tries to assemble the recent advances in both the finite volume method itself (theoretical aspects of the methods, new or improved algorithms, numerical implementation problems, benchmark problems and efficient solvers) as well as its application in complex problems in industry, environmental sciences, medicine and other fields of technology, so as to bring together the academic world and the industrial world. The topics of the proceedings reflect this wide range of perspectives and include: advanced schemes and methods (complex grid topology, higher order methods, efficient implementation), convergence and stability analysis, global error analysis, limits of methods, purely multidimensional difficulties, non homogeneous systems with stiff source terms, complex geometries and adaptivity, complexity, efficiency and large computations, chaotic problems (turbulence, ignition, mixing, . . .), new fields of application, comparisons with experimental results. The proceedings also include the results to a benchmark on three–dimensional anisotropic and heterogeneous diffusion problems, which was designed to test some 16 different schemes, among which finite volume methods, finite element methods, discontinuous Galerkin methods, mimetic methods and discrete gradient schemes. A new feature of this benchmark is the comparison of various iterative solvers on the matrices resulting from the different schemes.

Of course, the success of the symposium crucially depends on the quality of the contributions. Therefore we would like to express many thank all the authors of regular papers, who provided high quality papers on the above mentioned wide range of subjects, or contributed to the 3D anisotropic diffusion benchmark. The level the contributions was ensured by the Scientific Committee members, who organized the reviewing process of each paper. We express our gratitude to members of the Scientific Committee as well as to many other reviewers.

The symposium could not have been organized without the local support of Czech Technical University, Faculty of Mechanical Engineering and financial support of French contributors: CMLA ENS Cachan, IFP Energies nouvelles, IRSN, LATP Université Aix Marseille I, MOMAS group, Université Paris XIII, Université Paris Est Marne la Vallée, Université Pierre et Marie Curie.

Finally we would like to thank Springer Verlag Editor's team for their cooperation in the proceedings preparation, conference secretary T. Němcová and all others, who ensured logistic and communication before and during the conference.

Jaroslav Fořt
Jiří Fürst
Jan Halama
Raphaèle Herbin
Florence Hubert

Organization

Committees

Organizing Committee:

Jaroslav Fořt
Jiří Fürst
Jan Halama
Rémi Abgrall
Fayssal Benkhaldoun
Robert Eymard
Jean-Michel Ghidaglia
Jean-Marc Hérard
Raphaèle Herbin
Martin Vohralík

Scientific Committee:

Rémi Abgrall
Brahim Amaziane
Fayssal Benkhaldoun
Vít Dolejší
François Dubois
Denys Dutykh
Robert Eymard
Jaroslav Fořt
Jürgen Fuhrmann
Jiří Fürst
Thierry Gallouet
Jean-Michel Ghidaglia
Hervé Guillard
Jan Halama
Khaled Hassouni
Jean-Marc Hérard
Raphaèle Herbin
Florence Hubert
Raytcho Lazarov
Karol Mikula
Mario Ohlberger
Frédéric Pascal
Martin Vohralík

Contents

Vol. 2

Part II Invited Papers

Part III Benchmark Papers

Part I
Regular Papers

Volume-Agglomeration Coarse Grid In Schwarz Algorithm

H. Alcin, O. Allain, and A. Dervieux

Abstract The use of volume-agglomeration for introducing one or several levels of coarse grids in an Additive Schwarz multi-domain algorithm is revisited. The purpose is to build an algorithm applicable to elliptic and convective models. The sub-domain solver is ILU. We rely on algebraic coupling between the coarse grid and the Schwarz preconditioner. The Deflation Method and the Balancing Domain Decomposition (BDD) Method are experimented for a coarse grid as well as domain-by-domain coarse gridding. Standard coarse grids are built with the characteristic functions of the sub-domains. We also consider the building of a set of smooth basis functions (analog to smoothed-aggregation methods). The test problem is the Poisson problem with a discontinuous coefficicent. The two options are compared for the standpoint of coarse-grid consistency and for the gain in scability of the global Schwarz iteration.

Keywords domain decomposition, coarse grid
MSC2010: 65F04, 65F05

1 Volume agglomeration in MG and DDM

The idea of Volume Agglomeration is directly inspired by the multi-grid idea, but inside the context of Finite-Volume Method. In this paper the finite-volume partition considered is built as the dual of triangles, Fig. 1, right. In order to

H. Alcin and A. Dervieux
INRIA, B.P. 93, 06902 Sophia-Antipolis, France, e-mail: Hubert.Alcin@inria.fr, Alain.Dervieux@inria.fr

O. Allain
LEMMA, Les Algorithmes (Le Thales A), 2000 route des Lucioles, 06410 BIOT, France, e-mail: olivier.allain@lemma-ing.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_1, © Springer-Verlag Berlin Heidelberg 2011

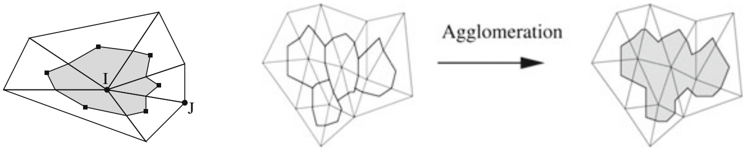


Fig. 1 Left: finite-Volume partition built as dual of a triangulation. Right: Greedy Algorithm for finite-volume cell agglomeration: four fine cells (left) are grouped into a coarse cell

build a coarser grid, it is possible to build coarse cells by sticking together neighboring cells for example with a greedy algorithm, Fig. 1, left. The coarser grid is *a priori* unstructured as is the fine one. By the magic of FVM, a consistent coarse discretisation of a divergence-based first-order PDE is directly available. Indeed, we can consider that the new unknown is constant over the coarse cell and it remains to apply a Godunov quadrature of the fluxes between any couple of two coarse cells. Elliptic PDE can also be addressed in similar although more complicated way.

As a result, consistent linear and non-linear coarse grid approximations are built using the agglomeration principle. Linear and nonlinear MG have been derived, in contrast with AMG algorithms. This method extends to Discontinuous Galerkin approximations [13]. The extension of Agglomeration MG to multi-processor parallel computing, however, are less easily achieved, as compared to Domain Decomposition Methods.

The many works on multi-level methods *à la* Bramble-Pasciak-Xu [2] has drawn attention to the question of basis smoothness. Indeed, the underlying basis function in volume-agglomeration is a characteristic function equal to zero or one. In [10], the agglomeration basis is extended to H^1 consistent ones in an analog way to smoothed-aggregation. In [4], a Bramble-Pasciak-Xu algorithm is built on these bases for an optimal design application.

While MG appeared, at least for a while, as the best CFD solution algorithm, Domain Decomposition methods (DDM) were seen as a new star for computational Structural Dynamics due to matrix stiffness issues. Domain decomposition methods assume the partition of the computational domain into sub-domains and assume that representative sub-problems on sub-domains can be rather easily computed and help convergence towards global problem's solution. An ideal DDM should be weakly scalable, that is, when it produces in some time with p processors a result on a given mesh, the result on a two times larger mesh should be produced in the same time with $2p$ processors. In Schwarz DDM, The set of local problems preconditions the global loop. Boundary conditions for each sub-domain problem are fetched in neighboring domains. The resulting iterative solver generally involves a Krylov iteration and is often refered as Newton-Krylov-Schwarz. It has been shown by S. Brenner [3] that the resulting algorithm is not scalable, unless a extension called coarse grid is added. In [3], the coarse grid correction is computed on a particular coarser mesh, embedded into the main mesh. The advantage of this approach is to produce a convergent coarse mesh solution. However the coarse mesh option is not practical in many cases, in particular for arbitrary unstructured meshes. As a result, it was tried later to build a coarse basis using other principles. An option is to look for a few global eigenvectors of the operator, see for example [15]. For CPU cost reasons, these eigenvectors should not be exactly computed but only approximated. In a recent study [11, 12], it is proposed to compute eigenvectors of the local Dirichlet-to-Neumann operator, which can be computed in parallel on each subdomain. The evaluation of eigenvectors is difficult when the matrix has a dominent Jordan behaviour (as for convection dominent models, the privilegiated domain of finite-volume methods). In the proposed study, we try to build a convergent coarse mesh basis for an arbitrary unstructured fine mesh. It has been observed that coarser meshes for unstructured meshes are elegantly build with volume-agglomeration. In this study, we follow this track, define a convergent basis and examine how it behaves as a coarse grid preconditioner. The test problem we concentrate on is inspired by a pressure-correction phase in Navier-Stokes (see for example [6]), and expresses as a Neumann problem with strongly discontinuous coefficient and writes:

$$-\nabla^* \frac{1}{\rho} \nabla p = RHS \text{ in } \Omega \qquad \frac{\partial p}{\partial n} = 0 \text{ on } \partial\Omega \qquad p(0) = 0.$$

in which the well-posedness is fixed with a Dirichlet condition on one cell.

1.1 Basic Additive exact and ILU Schwarz algorithm

Our discrete model relies on a vertex-centered formulation expressed on a triangulation. Let us assume that the computational domain Ω is split into two sub-domains, Ω_1 and Ω_2 , with an intersection $\overline{\Omega_1} \cap \overline{\Omega_2}$ with a thickness of at least one layer of elements. The *Additive Schwarz* algorithm is written in terms of preconditioning, as $M^{-1} = \sum_{i=1}^{2} A_{|\Omega_i}^{-1}$ where $A_{|\Omega_i}^{-1}$ holds for the Dirichlet problem on sub-domain Ω_i. The preconditioner M^{-1} can be used in a Krylov subspace method. In this paper, in order to keep some generality in our algorithms, we use GMRES, also used in [15]. In the *Additive Schwarz-ILU* version, the exact solution of the Dirichlet on each subdomain is replaced by the less costly Incomplete Lower Upper (ILU) approximate solution.

1.2 Algebraic Coarse grid

As shown by S. Brenner [3], the combination $M^{-1} = A_0^{-1} + \sum_{i=1}^{N} A_{|\Omega_i}^{-1}$ of the Additive Schwarz method with a coarse grid A_0^{-1} reduces the complexity to an essentially scalable one. Two methods have been proposed in the literature for introducing a coarse grid in an *algebraic* manner. Both rely on the following ingredients:

- $A_h u = f_h$ is the linear system to solve in V, fine-grid approximation space.
- $V_0 \subset V$ coarse approximation space. $V_0 = [\Phi_1 \cdots \Phi_N]$.
- Z an extension operator from V_0 in V and Z^T a restriction operator from V in V_0.
- $Z^T A_h Z u_H = Z^T f_h$ is the coarse system.

The Deflation Method (DM) has been introduced by Nicolaides [14] and is used by many authors. Saad *et al.* [15] encapsulates it into a Conjugate Gradient. Aubry *et al.* [1] apply it to a pressure Poisson equation. In DM, the projection operator is defined as:

$$P_D = I_n - A_h Z (Z^T A_h Z)^{-1} Z^T \text{ avec } A_h \in R^{n\times n} \text{ et } Z \in R^{n\times N}$$

The DM algorithm consists in solving first the coarse system $Z^T A_h Z u_H = Z^T f_h$, then the projected system $P_D A_h \check{u} = P_D f_h$ in order to get finally $u = (I_n - P_D^T)u + P_D^T u = Z(Z^T A_h Z)^{-1} Z^T f_h + P_D^T \check{u}$. The Balancing Domain Decomposition has been introduced by J. Mandel [9] and applied to a complex system in [7]. In [16] a formulation close to DM is proposed. It consists in replacing the preconditioner M^{-1} (ex.: global ILU, Schwarz, or Schwarz-ILU) by:

$$P_B = P_D^T M^{-1} P_D + Z(Z^T A_h Z)^{-1} Z^T.$$

1.3 Smooth and non-smooth coarse grid

The coarse grid is then defined by set of basis functions. A central question is the smoothness of these functions. According to Galerkin-MG, smooth enough functions provide consistent coarse-grid solutions. Conversely, DDM methods preferably use the characteristic functions of the sub-domains, $\Phi_i(x_j) = 1 \;\; si \;\; x_j \in \;\; \Omega_i$. In the case of P^1 finite-elements, for example, the typical basis function corresponds to setting to 1 all degrees of freedom in sub-domain. According to [10], the coarse system

$$U^H(x) = \Sigma_i U_i \Phi_i(x) \quad ; \quad \int \nabla U^H \nabla \Phi_i = \int f \Phi_i \quad \forall i$$

produces a solution U^H which does not converge towards the continous solution U when H tends to 0.

In order to build a better basis, we need to introduce a hierarchical coarsening process from the fine grid to a coarse grid which will support the preconditioner. Level j is made of N_j macro-cells C_{jk}, *i.e.* $\mathcal{G}_j = \cup_{k=1}^{N_j} C_{jk}$. Transfer operators are defined between successive levels (from coarse to fine):

$$P_i^j : \mathcal{G}_i \rightarrow \mathcal{G}_j \qquad P_i^j(u)(C_{k'i}) = u(C_{kj}) \text{ with } C_{k'i} \subset C_{kj}$$

Following [10] we introduce the smoothing operator:

$$(L_k u)_i = \sum_{j \in \mathcal{N}(i) \cup \{i\}} \text{meas}(j)\, u_j / \{ \sum_{j \in \mathcal{N}(i) \cup \{i\}} \text{meas}(j) \}$$

where $\mathcal{N}(i)$ holds for the set of cells which are direct neigbors of cell i. The smoothing is applied at each level between the coarse level k defining the characteristic basis and the finest level.

$$\Psi_k = (L_1 P_1^2 L_2 \cdots P_{p-2}^{p-1} L_{p-1} P_{p-1}^{p}) \Phi_k .$$

The resulting smooth basis function is compared with the characteristic one in Fig. 2.

The inconsistency of the characteristic basis and the convergence of this new smooth basis is illustrated by the solution of a Poisson equation with a sin function as exact solution, Fig. 3.

1.4 Three-level Algorithm

Because the local solver is not an exact one but an ILU solver, computing with a larger number of nodes in each sub-domain leads to a degradation of the convergence. It is then interesting to add a coarse grid on each sub-domain. This principle has been investigated in [8], where the authors use a non-smoothed aggregated basis.

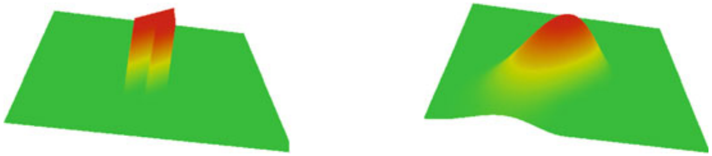


Fig. 2 Left: characteristic coarse grid basis function. Right: smooth coarse grid basis function

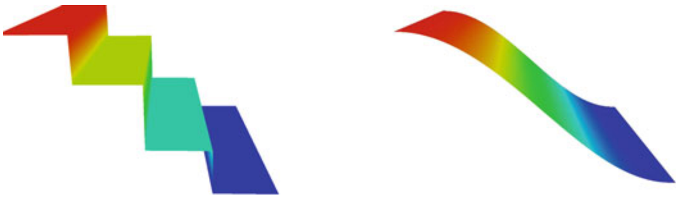


Fig. 3 Accuracy of the coarse grid approximation for a Poisson problem with a sin function (of amplitude 2.) as exact solution. Left: coarse grid solution with the characteristic basis (amplitude is 0.06). Right: coarse grid solution with a smooth basis (amplitude is 1.8)

Our proposition is to build sub-domain bases which are consistent with the Dirichlet condition of the Schwarz interface condition. To satisfy this, the Dirichlet condition is introduced in each smoothing step of the smooth basis function building process.

The *global algorithm* is made of a GMRES iteration preconditioned by the P_B operator combining a global coarse system with sub-domain preconditioners. The latter ones combine the local medium basis and the local ILU solver.

2 Numerical evaluation

We present some performance evaluations for the proposed algorithm. In all cases the conjugate gradient is used as fixed-point. The test case is a Neumann problem with discontinuous coefficient as in Section 2.1. The computational domain is a square. The coefficient takes two values with a ratio 100., on two regions separated by the diagonal of the domain. The right-hand side is a *sin* function. In the sequel, convergence is always measured for a division of the residual by 10^{20}. Convergence at this level were problematic with DM and the results are presented for BDD.

We recall first how behaves the *original Schwarz method* with one layer overlapping when the number of domains is fixed but the number of nodes increased. We compare in Table 1 a 2D calculation with two domains and 400 nodes with the analog computation with two domains and 10,000 nodes, which correspond to a h ratio of 5. We observe (Table 1) that the convergence of a Schwarz-ILU is four times slower on the finer mesh. We also observe that the convergence of the Schwarz algorithm with exact sub-domain solution is also degraded by a factor 2.6, a loss which may be explained by the thinner overlapping.

We continue with the study of the impact of choosing a *smooth basis* for the two-level Additive Schwarz ILU method. We observe that the scalability again does not hold, but it is nearly attained for the smooth basis option. It is rather bad for the characteristic basis. The rest of the paper uses only the smooth basis.

Table 1 Additive Schwarz method

# sub-domains	# cells	Local solver	# Iterations
2	400	ILU	55
2	400	Direct	28
2	10,000	ILU	221
2	10,000	Direct	74

Table 2 Scalability of the two-level AS-ILU method

Cells	10K	20K	47K	94K
Domains	12	28	66	142
Cells/domain	833	714	712	661
Char. basis	480	546	750	810
Smooth basis	400	391	444	491

The impact of the *medium grid* is examined in a third series of experiment is performed on a mesh of 40,000 cells, with 4 sub-domains and a total of 64 medium basis function (8 per sub-domain). In Table 3, we observe that without a coarse grid,

Table 3 Convergence of the different preconditioners (40,000 cells)

Type of preconditioner M^{-1}	# sub-domains	Iterations
Global ILU	1	348
Schwarz-ILU	4	431
Schwarz-ILU+coarse-grid	4	334
Three-level	4	264
Three-level	16	164

the Schwarz-ILU solver is 20% slower than the global (1-sub-domain) ILU solver (in terms of iteration count for 20 decades), the Schwarz-ILU with coarse-grid is slightly faster and the three level is 30% faster.

The *speedup* is measured for a given problem, set on a mesh of 40,000 cells. We compare the iteration count between a 4-sub-domain computation and a 16-sub-domain one. The coarse system solution with 16 unknowns is not parallel, but its cost is very small. Using four times more processors turn into a 6.4 smaller number of iterations before obtaining the solution (Table 3).

For a *scalability* measure, the mesh is taken finer and the number of sub-domain increased accordingly. We compare a 40,000-cell computation on 4 processors with a 160,000-cell on with 16 processors. We would like to mention that the Schwarz method with exact sub-domain resolution is far from being scalable: in Table 4, increase in iteration count is 40%. These bad news were announced by Table 1. We turn the combination of the Schwarz method with our smooth coarse grid. Exact solution is again performed on each sub-domain. Convergence becomes at least twice better. However, passing from 40,000 cells with 4 sub-domains to 160,000 cells with 16 sub-domains increases the iteration count by 60%, Table 4. We have checked that results with characteristic coarse grid are worse. In order to perform the analog comparison for the proposed three-level method (smooth coarse grid, smooth medium grid, ILU), we specify a four times higher number of medium-grid basis functions for the computation with four times higher number of cells (and sub-domains). Scalability in iterations is nearly satisfied, with 7% loss, Table 4.

Table 4 Scalability for the Schwarz, two-level Schwarz and three level Schwarz-ILU

Method	# cells	# sub-domains	# medium basis funct	Iterations
Schwarz	40,000	4		320
Schwarz	160,000	16		451
two-level Schwarz	40,000	4		130
two-level Schwarz	160,000	16		212
Three level	40,000	4	64	164
Three level	160,000	16	256	176

3 Concluding remarks

We have proposed a three-level algorithm for solving a linear system with a Schwarz method. The basis functions are independent of the system to solve and building them is not computationally expensive. The coarse grid solution is obtained after one iteration and yields a good initial solution. A few preliminary results show that the proposed method appears to be suitable for a pressure-projection system. The CPU cost (measured on a 2.6GHz workstation) for the heaviest example is of $0.05nS$ for the coarse factorization, $660nS$ $(20nS$ per processor) for the coarse system assembly while the Schwarz preconditioner cost is $124\mu S$. Further measures and applications to convection-diffusion models are in progress, as well as the introduction into a compressible Navier-Stokes model, [5].

References

1. R. Aubry, F. Mut , R. Lohner , J. R. Cebral, Deflated preconditioned conjugate gradient solvers for the Pressure-Poisson equation, Journal of Computational Physics, 227:24, 10196-10208, 2008.
2. J. Bramble, J. Pasciak, and J. Xu, Parallel multilevel preconditioners, Math. Comput., 55(191):122, 1990.
3. S. Brenner, Two-level additive schwarz preconditioners for plate elements, Numerische Mathematik, 72:4, 1994.
4. F. Courty, A. Dervieux, Multilevel functional Preconditioning for shape optimisation, IJCFD, 20:7, 481-490, 2006.
5. B. Koobus, S. Camarri, M.V. Salvetti, S. Wornom, A. Dervieux, Parallel simulation of three-dimensional flows: application to turbulent wakes and two-phase compressible flows, Advances in Engineering Software, 38, 328-337, 2007.
6. A.-C. Lesage, O. Allain and A. Dervieux, On Level Set modelling of Bi-fluid capillary flow, Int. J. Numer. Methods in Fluids, 53:8, 1297-1314, 2007.
7. P. Le Tallec, J. Mandel, M. Vidrascu, Balancing Domain Decomposition for Plates, Eigth International Symposium on Domain Decomposition Methods for Partial Differential Equations, Penn State, October 1993, Contemporary Mathematics, 180, AMS, Providence, 1994, 515-524.
8. P.T . Lin, M. Sala, J.N. Shadi, R S. Tuminaro, Performance of Fully-Coupled Algebraic Multilevel Domain Decomposition Preconditioners for Incompressible Flow and Transport.
9. J. Mandel, Balancing domain decomposition, Comm. Numer. Methods Engrg., 9, 233-241, 1993.
10. N. Marco, B. Koobus, A. Dervieux, An additive multilevel preconditioning method and its application to unstructured meshes, INRIA research report 2310, 1994 and Journal of Scientific Computing, 12:3, 233-251, 1997.
11. F. Nataf, H. Xiang, V. Dolean, A two level domain decomposition preconditioner based on local Dirichlet-to-Neumann maps, C. R. Mathématiques, 348:21-22, 1163-1167, 2010.
12. F. Nataf, H. Xiang, V. Dolean, N. Spillane, A coarse space construction based on local Dirichlet to Neumann maps, to appear in SIAM J. Sci Comput., 2011.
13. C.R. Nastase, D. J. Mavriplis, High-order discontinuous Galerkin methods using an hp-multigrid approach, Journal of Computational Physics 213:1, 330-357, 2006.
14. R. A. Nicolaides, Deflation of conjugate gradients with applications to boundary value problem, SIAM J.Numer.Anal., 24, 355-365, 1987.

15. Y. Saad, M. Yeung, J. Erhel, and F. Guyomarc'H, A deflated version of the conjugate gradient algorithm, SIAM J. Sci. Comput., 21:5, 1909-1926, 2000.
16. C. Vuik , R. Nabben A comparison of deflation and the balancing preconditionner, SIAM J. Sci. Comput., 27:5, 1742-1759, 2006.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A comparison between the meshless and the finite volume methods for shallow water flows

Yasser Alhuri, Fayssal Benkhaldoun, Imad Elmahi, Driss Ouazar, Mohammed Seaïd, and Ahmed Taik

Abstract A numerical comparison is presented between a meshless method and a finite volume method for solving the shallow water equations. The meshless method uses the multiquadric radial basis functions whereas a modified Roe reconstruction is used in the finite volume method. The obtained results using both methods are compared to experimental measurements.

Keywords Meshless method, shallow water equations, finite volume method, radial basis functions, numerical simulation
MSC2010: 15A09, 65N08, 65F10

1 Introduction

Finite volume method have been widely used to solve shallow water flows due to their conservation properties and the ability to handle complex geometries. Recently, meshless methods using radial basis functions have attracted many researcher in mechanical engineering as well as in computational fluid dynamics. Application of

Yasser Alhuri and Ahmed Taik
Dept. Mathematics, UFR-MASI FST, Hassan II University Mohammedia, Morocco

Fayssal Benkhaldoun
LAGA, Université Paris 13, 99 Av J.B. Clement, 93430 Villetaneuse, France

Imad Elmahi
ENSAO Complex Universitaire, B.P. 669, 60000 Oujda, Morocco

Driss Ouazar
Dept. Genie Civil, LASH EMI, Mohammed V University Rabat, Morocco

Mohammed Seaïd
School of Engineering and Computing Sciences, University of Durham, Durham DH1 3LE, UK

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_2, © Springer-Verlag Berlin Heidelberg 2011

meshless methods for numerical solution of shallow water equations has already been addressed in many references in the literature, see for example [9] and further references are therein. However, to our best knowledge, no comparison between the meshless method and the finite volume method is available in the literature for shallow water flows. The aim of the present work is therefore, to perform a comparative study between the meshless and the finite volume methods for solving the shallow water equations rearranged in a conservative form as

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{W})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{W})}{\partial y} = \mathbf{0}, \tag{1}$$

where **W** is the vector of conserved variables, **F** and **G** are the tensor fluxes

$$\mathbf{W} = \begin{pmatrix} h \\ hu \\ hv \end{pmatrix}, \qquad \mathbf{F}(\mathbf{W}) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \qquad \mathbf{G}(\mathbf{W}) = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix},$$

where g is the gravitational acceleration, $h(t, x, y)$ is the water depth, $u(t, x, y)$ and $v(t, x, y)$ are the depth-averaged velocities in the x- and y-direction, respectively. Note that the equations (1) has to be solved in a bounded spatial domain Ω with smooth boundary Γ, equipped with given boundary and initial conditions. It is well known that the system (1) is strictly hyperbolic with real and distinct eigenvalues.

The basic references for the present finite volume method are [1, 2]. In [1], a description of the overall structure of the finite volume method is presented. In particular, the discretization of the gradient fluxes using the sign matrix of the Jacobian is described in details for both scalar equations and hyperbolic systems of conservation laws with source terms. In [2], the implementation of the finite volume scheme on unstructured grids is analyzed and applied to pollutant transport by shallow water flows. This implementation involves an original treatment of the flux derivatives coupled with the source term in unstructured meshes. The current work aims to compare this finite volume method to a meshless method using the multiquadric radial basis. The numerical results are presented for the shallow water flow in a backward facing step. This test problem has been experimentally investigated in [6]. The obtained results using the meshless and the finite volume methods are compared against the measurements from [6].

2 Solution procedures

In this section we briefly describe the two methods used in solve the shallow water equations (1). Further details on the formulation and implementation of these techniques can be found in the cited references.

2.1 A meshless method

The principal idea of the radial basis interpolation is to interpolate a finite series of an unknown function $f(\mathbf{X})$ at N distinct points $\mathbf{X}_j$ on Ω by the following expansion

$$f(\mathbf{X}) \simeq \sum_{j=1}^{N} \alpha_j \varphi(\|\mathbf{X} - \mathbf{X}_j\|), \tag{2}$$

Here α_j's are the unknown coefficients to be calculated at each time step, and $\varphi(\|\mathbf{X} - \mathbf{X}_j\|)$ is the radial basis function, $X_j \in \mathbb{R}^n$, $j = 1, 2, \ldots, N$, $\|\mathbf{X} - \mathbf{X}_j\| = r_j$, $r_j = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$ is the Euclidean distance and $\mathbf{X} = (x, y)$, $\mathbf{X}_j = (x_j, y_j)$. Since multiquadrics ($MQ$) are infinitely smooth functions, they are often chosen as the trial function for φ, i.e.,

$$\varphi(r_j) = \sqrt{r_j^2 + c^2} = \sqrt{(x - x_j)^2 + (y - y_j)^2 + c^2},$$

where $c \neq 0$ is the shape parameter controlling the fitting of a smooth surface to the data, see for instance [7].

The application of collocation radial basis functions to a system (1) and its boundary conditions start by first selecting a set of $(x_1, y_1), \ldots, (x_b, y_b)$ boundary and $(x_{b+1}, y_{b+1}), \ldots, (x_{d+b}, y_{d+b=N})$ domain nodes. The unknown solution of the problem at each time t can be determined under the form

$$\Phi(X, t) = \sum_{j=1}^{N} \alpha_j(t) \varphi(r_j), \tag{3}$$

where $X = (x, y)^T$ and $\{\alpha_j\}$ are unknown coefficients to be determined. To solve the two-dimensional time-dependent differential equations given by (1), the time explicit forward difference scheme is used, then

$$\Phi_i^{n+1} = \Phi_i^n - \Delta t \Big(\frac{\partial G_i^n}{\partial x} + \frac{\partial F_i^n}{\partial y} \Big), \tag{4}$$

where Δt is the time step, Φ_i^{n+1} is the numerical solution vector at points $X_i = (x_i, y_i)$ in time $n\Delta t$. The values of the interpolate Φ^n are given by the following MQ radial basis function

$$\Phi^n(X_i, t) = \sum_{j=1}^{N} \alpha_j^n(t) \sqrt{r_{ij}^2 + c^2}, \tag{5}$$

where $r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + c^2}$, which are collocating with a set of data $(x_i, y_i)_{i=1}^{N}$ over the domain $\Omega \subset \mathbb{R}^2$, and forms a system of N linear algebraic equations in N unknowns

$$\begin{pmatrix} \Phi_1^n \\ \Phi_2^n \\ \vdots \\ \Phi_N^n \end{pmatrix} = \begin{pmatrix} \varphi(r_{11}) & \varphi(r_{12}) & \dots & \varphi(r_{1N}) \\ \varphi(r_{21}) & \varphi(r_{22}) & \dots & \varphi(r_{2N}) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi(r_{N1}) & \varphi(r_{N2}) & \dots & \varphi(r_{NN}) \end{pmatrix} \begin{pmatrix} \alpha_1^n \\ \alpha_2^n \\ \vdots \\ \alpha_N^n \end{pmatrix}, \tag{6}$$

The numerical scheme (4) gives a system of N linear equations in N unknowns can be expressed in matrix form

$$\overrightarrow{\Phi} = A\overrightarrow{\alpha}, \tag{7}$$

where $A = [\varphi_j(x_i, y_i)]$ is a $N \times N$ matrix, $\overrightarrow{\alpha} = [\alpha_j^n]$ and $\overrightarrow{\Phi} = [\Phi_j^n]$ are N vectors. Note that for $n = 0$ the coefficients $\{\alpha_j^0\}$ can be found using the initial conditions. Hence the solution Φ_1^0 is well-determined and it will be used as initial condition for the scheme (4). The numerical values of the unknown spatial derivatives of $\Phi^n(x_i, y_i)$ is approximated using the multiquadric radial basis functions as

$$\frac{\partial^m \Phi^n}{\partial x^m}(x_i, y_i, t) = \sum_{j=1}^{N} \alpha_j^n(t) \frac{\partial^m \varphi_j}{\partial x^m}(x_i, y_i),$$

$$\frac{\partial^m \Phi^n}{\partial y^m}(x_i, y_i, t) = \sum_{j=1}^{N} \alpha_j^n(t) \frac{\partial^m \varphi_j}{\partial y^m}(x_i, y_i),$$

where $m = 1, 2$. Thus, at each time step n, the numerical solution of the vector $\Phi(x_i, y_i, t)_{i=b+1}^{b+d}$ at the interior points are computed by substituting the Φ and its spatial derivatives into equation (4). The boundary values $\Phi(x_i, y_i, t)_{i=1}^{b}$ are given by boundary conditions.

Finally, the numerical solution is obtained by solving the system of N linear equations

$$\overrightarrow{\Phi}^{n+1} = (A - \Delta t A_L)\overrightarrow{\alpha}^n, \tag{8}$$

where $A_L = [L(\varphi^n(r_{ij}))]$ is an $N \times N$ matrix coefficient of which are defined by

$$L(\varphi^n(r_{ij})) = \Big(\frac{\partial G(\varphi^n(r_{ij}))}{\partial x} + \frac{\partial F(\varphi^n(r_{ij}))}{\partial y}\Big).$$

Hence, the unknown coefficients vector $[\alpha_j^n]$ can be determined using Gaussian elimination or Gmres methods.

2.2 A finite volume method

Discretizing the computational domain in control volumes, the finite volume method applied to (1) results in

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{|\mathscr{T}_i|} \sum_{j \in N(i)} \int_{\Gamma_{ij}} \mathscr{F}(\mathbf{W}^n; \mathbf{n}) \, d\sigma, \tag{9}$$

where $N(i)$ is the set of neighboring triangles of the cell $\mathscr{T}_i$, $\mathbf{W}_i^n$ is an average value of the solution $\mathbf{W}$ in the cell $\mathscr{T}_i$ at time t_n. $|\mathscr{T}_i|$ denotes the area of $\mathscr{T}_i$ and ∂V is the surface surrounding the control volume V. Here, $\mathbf{n} = (n_x, n_y)^T$ denotes the unit outward normal to the surface ∂V, and

$$\mathscr{F}(\mathbf{W}; \mathbf{n}) = \mathbf{F}(\mathbf{W}) n_x + \mathbf{G}(\mathbf{W}) n_y.$$

Following the formulation in [1], the proposed finite volume scheme consists of a predictor stage and corrector stage. It can be formulated as

$$\begin{aligned} \mathbf{W}_{ij}^n &= \frac{1}{2}\left(\mathbf{W}_i^n + \mathbf{W}_j^n\right) - \frac{1}{2}\,\mathrm{sgn}\left[\nabla\mathscr{F}\left(\overline{\mathbf{W}}_{ij}^n; \mathbf{n}_{ij}\right)\right]\left(\mathbf{W}_j^n - \mathbf{W}_i^n\right), \\ \mathbf{W}_i^{n+1} &= \mathbf{W}_i^n - \frac{\Delta t}{|\mathscr{T}_i|} \sum_{j \in N(i)} \mathscr{F}\left(\mathbf{W}_{ij}^n; \mathbf{n}_{ij}\right) |\Gamma_{ij}| + \Delta t \mathbf{S}_i^n, \end{aligned} \tag{10}$$

with sgn $[\mathbf{A}]$ denotes the sign matrix of $\mathbf{A}$, and $\overline{\mathbf{W}}_{ij}^n$ is approximated either by Roe's average state or simply by the mean state

$$\overline{\mathbf{W}}_{ij}^n = \frac{1}{2}\left(\mathbf{W}_i^n + \mathbf{W}_j^n\right). \tag{11}$$

A detailed formulation for the sign matrix in (10) are given in [1,2] and will not be repeated here.

3 Numerical Results

To validate the results obtained using the meshless and finite volume methods we consider the test problem of flow in a backward facing step. Experimental data for this test problem have been provided in [6] and are used here to compare our numerical results. The domain geometry is depicted in Fig. 1 and for the other involved parameters we refer the reader to [6]. On the upstream and downstream boundaries we used the condition as in [6] *i.e.*

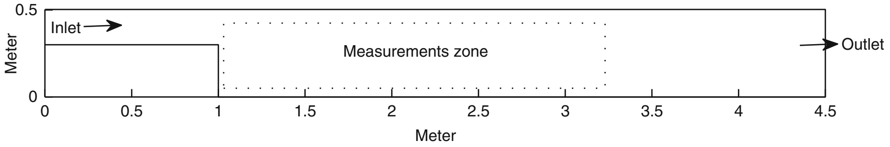


Fig. 1 Schematic domain used in the experimental setup and in our simulations

- On the upstream boundary: The discharge, $Q = 20.2\ l/s$, is imposed.
- On the downstream boundary: The measured depth, $h = 24.2\ cm$, is imposed.

In our finite volume simulations we have used an unstructured mesh shown in the left plot of Fig. 2. This mesh contains 5209 triangles and 2779 nodes. In the computations reported herein, the Courant number C is set to 0.8 and the time stepsize Δt is adjusted at each step according to the stability condition

$$\Delta t = C \min_{\Gamma_{ij}} \left(\frac{|\mathscr{T}_i| + |\mathscr{T}_j|}{2\,|\Gamma_{ij}|\,\max_p \left|(\lambda_p)_{ij}\right|} \right),$$

where λ_p are the eigenvalues of the system (1), Γ_{ij} is the edge between two triangles $\mathscr{T}_i$ and $\mathscr{T}_j$. For the meshless method we used the node distribution shown in the right plot of Fig. 2. Here we used 229 collocation points uniformly distributed in the computational domain. It should be stressed that the stability condition in the meshless method is

$$\Delta t \leq C \frac{d_{min}}{\max\left(\sqrt{U \pm gh}, \sqrt{V \pm gh}\right)}, \tag{12}$$

where d_{min} is the minimum distance between any two adjacent collocation points and C is the courant number set to 0.1 in our simulations. In our computations the shape coefficient in the multiquadric radial basis functions is selected according to [5, 9]. It has been shown in these references that a near-optimal approximation of the model hydrodynamic can be achieved by using the proposed value

$$c = 0.815 d_{min}.$$

Steady-state solutions are presented for both methods.

Figure 3 illustrates the cross sections of the velocity component u at vertical lines located in $x = 2.03$ and in $x = 2.53$. These two location belong to the zone where measurements have been taken. The agreement between the simulations using the finite volume method and measurements is fairly good. The velocity magnitude and recirculation location are well predicted by the both numerical methods. As expected, a reverse flow is formed near the upper and lower walls and propagates

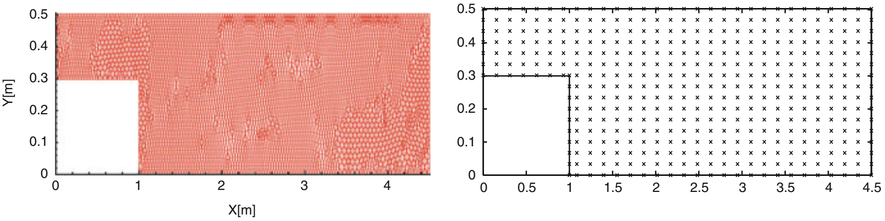


Fig. 2 Finite volume mesh (left plot) and node distribution for meshless method (right plot)

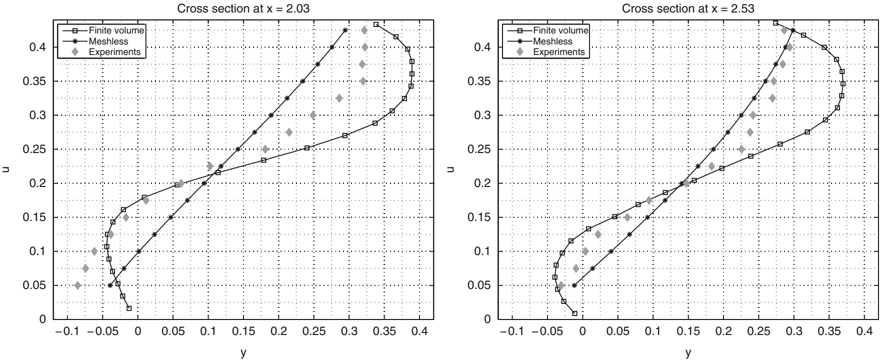


Fig. 3 Comparison results for cross sections of the velocity field u at vertical line $x = 2.03$ (left plot) and at vertical line $x = 2.53$ (right plot)

upstream. However, the location of the recirculation is less accurately predicted by the numerical methods. This may be attributed to the absence of shear stresses from the bed and eddy viscosity in the governing equations (1). It should also be pointed out that the numerical diffusion is more pronounced in the results obtained using the finite volume method than those obtained using the meshless method.

In terms of computational costs for this test problem, the CPU time for the meshless method is about 34 minutes. The considered finite volume method requires more than four times the CPU used for the meshless method. This is a clear indication that the meshless method is more efficient than the finite volume method regardless the number and the distribution of the collocation points in the computational domain.

Acknowledgements The authors would like to thank Prof. Jaime Fe Marqués for providing the experimental data to us.

References

1. F. Benkhaldoun, I. Elmahi, M. Seaïd, “A new finite volume method for flux-gradient and source-term balancing in shallow water equations”, Computer Methods in Applied Mechanics and Engineering. **199** pp:49-52 (2010).
2. F. Benkhaldoun, I. Elmahi, M. Seaïd, “Well-balanced finite volume schemes for pollutant transport by shallow water equations on unstructured meshes”, J. Comp. Physics. **226** pp:180-203 (2007).
3. T. Belytschko, Y. Krongauz, D. Organ, M. Fleming, P. Krysl, “Meshless methods: an overview and recent developments”, Computer methods in applied mechanics and engineering”, special issue on Meshless Methods, 139, pp:3-47, (1996).
4. M. Buhamman, Radial basis function: theory and implementations, Cambridge university press, (2003).
5. R. L. Hardy, “Multiquadric equations of topography and other irregular surfaces”. J. Geophys. Res, 176, pp:1905-1915, (1971).
6. J. Fe, F. Navarrina, J. Puertas, P. Vellando and D. Ruiz, “Experimental validation of two depth-averaged turbulence models”, Int. J. Numer. Meth. Fluids, 60, pp:177-202, (2009).
7. Y. C. Hon, K. F. Cheung, X. Z. Mao and E. J. Kansa, “A Multiquadric solution for the shallow water equations”, ASCE J. of hydrodlic engineering”, vol.125, No.5, pp:524-533, (1999).
8. P. Roe, “Approximate riemann solvers, parameter vectors and difference schemes”, J. Comp. Physics. 43, pp:357-372, (1981)
9. S. M. Wong, Y.C. Hon, M.A. Golberg, “Compactly supported radial basis functions shallow water equations”, J. Appl. Sci. Comput, 127, 79-101, (2002).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Time Compactness Tools for Discretized Evolution Equations and Applications to Degenerate Parabolic PDEs

Boris Andreianov

Abstract We discuss several techniques for proving compactness of sequences of approximate solutions to discretized evolution PDEs. While the well-known Aubin-Simon kind functional-analytic techniques were recently generalized to the discrete setting by Gallouët and Latché [15], here we discuss direct techniques for estimating the time translates of approximate solutions in the space L^1. One important result is the Kruzhkov time compactness lemma. Further, we describe a specific technique that relies upon the order-preservation property. Motivation comes from studying convergence of finite volume discretizations for various classes of nonlinear degenerate parabolic equations. These and other applications are briefly described.

Keywords time translates, Kruzhkov lemma, order-preservation, finite volumes
MSC2010: Primary: 65M12, 46B50, 35K65; Secondary: 46E39, 65M08

1 Introduction

Let us think of evolution equations set on a cylindrical domain $Q := (0,T) \times \Omega \subset \mathbb{R}^+ \times \mathbb{R}^N$. Proving convergence of space-time discretizations of such equations often includes the three following steps: constructing discrete solutions and getting uniform (in appropriate discrete spaces) estimates; extracting a convergent subsequence; writing down a discrete weak formulation (e.g., with discretized test functions) and passing to the limit in the equation in order to infer convergence.

For the first step, obtention of estimates is greatly simplified by preservation, at the discrete level, of the key structure properties of the PDE (such as symmetry, coercivity, monotonicity of the diffusion operators involved; entropy dissipation,

Boris Andreianov
CNRS UMR 6623, 16 route de Gray, Besançon, France, e-mail: boris.andreianov@univ-fcomte.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_3, © Springer-Verlag Berlin Heidelberg 2011

for the nonlinear convection operators in the degenerate parabolic case; etc.). For getting discrete *a priori* estimates test functions are often used, as in the continuous case. Therefore, some analogues of integration-by-parts formulas and chain rules are instrumental for the first step. For the examples we give in this paper, "discrete duality" type schemes (mimetic, co-volume, DDFV; see, e.g., [3] and references therein) can be used to guarantee an exact integration-by-parts feature. In contrast, chain rules for derivation in time or in space must be replaced by approximate analogues, often taking the form of convexity inequalities (see, e.g., [4], [3, Sect. 4]).

In this note, we give some insight into convergence proofs for different subclasses of degenerate elliptic-parabolic-hyperbolic PDEs under the general form[1]

$$u = b(v), w = \varphi(v), \quad u_t - \operatorname{div}\big[\mathbf{G}(v) - \mathbf{a_0}\big(\nabla w\big)\big] + \psi(v) = f \ \text{ in } Q = (0,T)\times\Omega \tag{1}$$

with $b(\cdot), \varphi(\cdot), \psi(\cdot)$ continuous[2] non-decreasing on $\mathbb{R}$, normalized by zero at zero, with a continuous convection flux $\mathbf{G}(\cdot)$ and with $\mathbf{a_0} : \mathbb{R}^N \to \mathbb{R}^N$ of Leray-Lions type (see e.g. [1, 4]; p-laplacian, with $\mathbf{a_0}(\xi) = |\xi|^{p-2}\xi$ is a typical example). For the sake of simplicity, homogeneous Dirichlet boundary condition on $(0,T) \times \partial\Omega$ is taken.

But our main goal is to discuss the second step of the proofs[3], the one of getting compact[4] sequences of discrete solutions. For linear problems, the two latter steps are somewhat trivial; indeed, mere functional-analytic bounds would lead to compactness in a weak topology, which is enough to pass to the limit from the discrete to the continuous weak formulation of the PDE. Thinking of nonlinear problems and passage to the limit in nonlinear terms, bounds in functional spaces can be sufficient when combined with basic compact embeddings; but this requires rather strong bounds involving e.g. some estimates of the derivatives. Regarding evolution PDEs of, say, porous medium type, L^p bounds are available on the space derivatives but not on the time derivatives (those belong to some *negative* Sobolev spaces). In this situation, either compactness in an *ad hoc* strong topology is needed; or the weak compactness coming from uniform boundedness should be combined with some compactification arguments (compensated compactness, Young measures and their reduction, etc.) that exploit in a non-trivial way the particular structure of the PDE in hand (div-curl structure, pseudo-monotonicity, entropy inequalities, etc.).

In this note, we first recall in § 2 the fundamental techniques using only bounds in well-chosen functional spaces (see [2, 9, 11, 17] for the continuous setting; see [12, 15] for the corresponding discrete results). In § 3, we present a collection

[1] See [5] and references therein for well-posedness theory of such "triply nonlinear" equations. These are mathematical models for porous media, sedimentation, Stefan problem, etc..

[2] Actually, we assume that either these functions are uniformly continuous, or v is bounded *a priori*.

[3] When the compactification methods strongly utilize a particular structure of the underlying PDE, this step is in fact combined with the third step of passing to the limit.

[4] Throughout the note, "compact" actually signifies "relavively compact".

of complementary techniques for estimation of time translates of families of functions that already possess some estimate of space translates. In § 4, we describe one indirect method for proving compactness and convergence of families of approximate solutions. The method heavily exploits the order-preservation property, required both for the PDE in hand and for the approximation scheme in use. Throughout the note, the exposition is motivated and illustrated by applications to approximation of several cases of problem (1) (different cases requiring different approaches).

2 Functional-analytic approach of Aubin-Lions-Dubinskii-Simon

In the continuous setting, one celebrated result is the Aubin-Lions or Dubinskii lemma ([9] and [11]) and its generalization by Simon [17] (see also Amann [2]). To give an example relevant for the applications we have in mind, let us simply state here that a sequence $(u_h)_h$ bounded in $L^1(0,T;W^{1,1}(\Omega))$ and such that $((u_h)_t)_h$ is bounded in $L^1(0,T;W^{-1,1}(\Omega))$ is relatively compact in $L^1(Q)$, cf. [15]. More generally, compactness comes from an *a priori bound* on u_h in some space $L^p(0,T,X)$ with X compactly embedded in $L^1(\Omega)$ (e.g., $X = W^{1,1}(\Omega)$) while the PDE brings information on boundedness of the time derivatives $(u_h)_t$ in some space $L^q(0,T;Y)$ where Y can be a subspace of distributions on Ω equipped with a rather weak topology (e.g., $Y = W^{-1,1}(\Omega)$). A discrete version of the Aubin-Simon lemma was recently proposed by Gallouët and Latché in [15]; it is based upon a careful reformulation of estimates in terms of "coherent" families $(X_h)_h$, $(Y_h)_h$ of discrete spaces.

A related result taken from Simon [17] and Amann [2] uses a bound on fractional time derivatives of u_h. As it was demonstrated by Emmrich and Thalhammer in [12], this version is quite appropriate in the time-discretized setting. Indeed, time fractional derivatives of order less than $1/2$ exist even for piecewise constant functions. Technically, this method involves an indirect estimation of weighted time translates, under a form $\displaystyle\int_0^T\int_0^T \frac{|u_h(t)-u_h(s)|^p}{|t-s|^{1+\sigma p}}\,ds\,dt$ with some $p \geq 1$ and $\sigma \in (0,1/2)$.

These results only use bounds in functional spaces and very few of the underlying PDE properties. They offer a very wide spectrum of applications; yet they are difficult to apply on degenerate parabolic problems with non-Lipschitz nonlinearities. The difficulty comes from the fact that non-Lipschitz mappings make bad correspondence between *linear* functional spaces. Yet this difficulty is not a fundamental one; roughly speaking, it is settled by a careful use of translation arguments and of moduli of continuity. This is the object of the next section.

3 Direct estimation of time translates

In this section, the compactness question is studied using the one and only space[5] $L^1(Q)$. By the Fréchet-Kolmogorov compactness criterion in $L^1(Q)$, uniform bounds on space and time translates of u_h are needed. In the setting of the present note, the first ones are readily available. The difficulty lies in estimating the time translates as

$$\forall h \quad \int_0^{T-\delta}\int_\Omega \Big| u_h(t+\delta) - u_h(t) \Big| \leq \omega(\delta) \quad \text{with } \lim_{\delta\to 0} \omega(\delta) = 0, \tag{2}$$

$\omega(\cdot)$ being a modulus of continuity, uniform in h. Here are two ways to obtain (2).

A discrete Kruzhkov lemma[6]

Lemma 1 (Kruzhkov [16]). *Assume that the families of functions $(u_h)_h$, $(F_h^\alpha)_{h,\alpha}$ are bounded in $L^1(Q)$ and satisfy $\frac{\partial}{\partial t} u_h = \sum_{|\alpha|\leq m} D^\alpha F_h^\alpha$ in $\mathscr{D}'(Q)$. Assume that u_h can be extended outside Q, and one has*[7]

$$\iint_Q |u_h(t,x+\delta) - u_h(t,x)|\, dxdt \;\leq\; \omega(\delta), \quad \textit{with } \lim_{\delta\to 0}\omega(\delta) = 0, \tag{3}$$

where $\omega(\cdot)$ does not depend on h. Then $(u_h)_h$ is (relatively) compact in $L^1(Q)$.

Clearly, this is an L^1_{loc} compactness result (one can apply the lemma locally in Q).

For problem (1), the value $m = 1$ is relevant, because an L^1 bound is available for the flux $\big(\mathbf{G}(v) - \mathbf{a_0}\big(\nabla\varphi(v)\big)\big)$; therefore we limit to this case the discussion of discrete analogues of Lemma 1. To give an idea of discrete versions of the Kruzhkov lemma[8], assume we are given a family of meshes of Ω indexed by their size h and satisfying mild proportionality restrictions (e.g., for the case of two-point flux finite volume schemes as described in [13], one needs for all neighbour volumes K,L, diam(K)+diam$(L)\leq const\, d_{K,L}$ uniformly in h). Assume that on these meshes, spaces of

[5] Working in an h-independent space is an advantage for producing discrete versions of compactness arguments; yet the approach of [15] exhibits a simple and efficient use of h-dependent spaces.

[6] There is a strong relation to the method of § 2. The Kruzhkov lemma allows for general moduli of continuity. E.g., for problem (1) with $\varphi = Id$, the Aubin-Lions-Dubinskii-Simon argument can be used if $b(\cdot)$ is Lipschitz continuous (with $X = W_0^{1,p}(\Omega)$) or Hölder continuous (with a fractional Sobolev space chosen for X), and the Kruzhkov lemma can be used for any continuous $b(\cdot)$.

[7] In practice, space translation estimates of the kind (3) can be obtained via an estimate of some discrete gradients; notice that estimates of kind (3) are stable upon composing $(u_h)_h$ by a function $b(\cdot)$ which is uniformly continuous (as in (1), we mean that $u_h = b(v_h)$).

[8] Here we give a rather heuristic presentation; see [7] and [3] for two precise formulations covering, e.g., the two-point flux finite volume schemes ([13]) and DDFV schemes ([3] and ref. therein).

discrete functions $\mathbb{R}_h$ and discrete fields $(\mathbb{R}^N)_h$ are defined (each element $u_h \in \mathbb{R}_h$ or $\mathbf{F}_h \in (\mathbb{R}^N)_h$ is a piecewise constant on Ω function reconstructed from the degrees of freedom of the discretization method). Assume we are given discrete gradient and discrete divergence operators ∇_h and div_h mapping between these spaces. Thus all discrete objects (functions, fields, gradient, divergence) are naturally lifted to $L^1(Q)$.

Let $(\delta_h)_h$ be the associated time steps, let N_h be the entire part of T/δ_h. Assume that we are given an initial condition b_h^0 and discrete evolution equations under the form

$$\text{for } n \in [1, N_h+1], \quad \frac{b(v_h^n) - b(v_h^{n-1})}{\delta_h} = \mathrm{div}_h\,[\mathbf{F}_h^n] + f_h^n \quad \text{in } \mathbb{R}_h, \tag{4}$$

where families $(\big((u_h^n)_n\big)_h$, $\big((f_h^n)_n\big)_h$ (discrete functions) and $\big((\mathbf{F}_h^n)_n\big)_h$ (discrete fields) are bounded in $L^1(Q)$. Assume that the discrete gradients $\big((\nabla_h v_h^n)_n\big)_h$ are bounded in $L^1(Q)$ and that this bound implies a uniform translation bound in space of the family v_h (this is true, e.g., when discrete Poincaré inequalities can be proved). Under these assumptions, reproducing at the discrete level the proof [16] of Lemma 1 as it is done in [3, 7], one concludes that the family $(b(v_h))_h$ is relatively compact in $L^1(Q)$. Note that, the case $m \geq 2$ would require more work.

A classical technique for the "variational" setting

Following [1], by "variational" we mean a setting where the solution w is an admissible test function in the weak formulation of the PDE; e.g., (1) can be tested with $w = \varphi(v)$. It typically comes along with *a priori* estimates that can be reproduced at the discrete level, provided the discretization is somewhat structure-preserving.[9]

The technique of [1] used, in its finite volume version, e.g., in [4, 13, 14], is to integrate[10] the equation in time from t to $t+\delta$, take $w_h(t+\delta)$–$w_h(t)$ for test function, then integrate in (t,x). On problem (1), this leads to a uniform estimate

$$\forall h > 0 \quad \int_0^{T-\delta}\!\!\int_\Omega \Big(b(v_h)(t+\delta) - b(v_h)(t)\Big)\Big(\varphi(v_h)(t+\delta) - \varphi(v_h)(t)\Big) \leq \omega(\delta). \tag{5}$$

Then Lipschitz continuity of $\varphi \circ b^{-1}$ (resp., of $b \circ \varphi^{-1}$) can be used to infer uniform L^2 time translates of $w_h = \varphi(v_h)$ (resp., of $u_h = b(v_h)$). Yet the L^1 time translates can be obtained in the case $\varphi \circ b^{-1}$ (resp., $b \circ \varphi^{-1}$) is a merely continuous function.

[9] Notice that for evolution PDEs governed by accretive in $L^1(\Omega)$ operators, of which (1) is an example, time-implicit discretizations are better suited for structure preservation. Use of numerical schemes in space that possess a kind of discrete duality (mimetic, co-volume, DDFV schemes, etc.) enables getting discrete estimates analogous to the continuous ones. For notions of solution involving some version of chain rule (e.g., entropy, renormalized solutions) orthogonality assumption on the meshes and isotropy assumption on the diffusion operator may be needed, see e.g. [4].

[10] Here, for the sake of simplicity, we stick to the terminology and notation of the continuous case.

• *A technique for L^1 estimates involving non-Lipschitz nonlinearities (see [4])*
Consider the case where $\widetilde{\varphi} := \varphi \circ b^{-1}$ is a uniformly continuous function (moreover, it is non-decreasing). Let π be a concave modulus of continuity for $\varphi \circ b^{-1}$, Π be its inverse, and set $\widetilde{\Pi}(r) = r\,\Pi(r)$. Let $\widetilde{\pi}$ be the inverse of $\widetilde{\Pi}$. Note that $\widetilde{\pi}$ is concave, continuous, and $\widetilde{\pi}(0) = 0$. Set $u^\delta = b(v_h)(t+\delta, x)$ and $u = b(v_h)(t, x)$. We have

$$\int_Q |\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)| = \int_Q \widetilde{\pi} \circ \widetilde{\Pi}\big(|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|\big) \le |Q|\, \widetilde{\pi}\Big(\tfrac{1}{|Q|} \int_Q \widetilde{\Pi}\big(|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|\big)\Big).$$

Since $|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)| \le \pi(|u^\delta - u|)$, we have $\Pi(|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|) \le |u^\delta - u|$ and

$$\widetilde{\Pi}(|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|) = \Pi(|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|)|\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)| \le |u^\delta - u|\, |\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|.$$

Therefore, (5) implies an L^1 estimate of the kind (2) on $w_h = \varphi(v_h)$:

$$\int_Q |w_h(t+\delta) - w_h(t)| \le |Q|\, \widetilde{\pi}\Big(\tfrac{1}{|Q|} \int_Q |u^\delta - u||\widetilde{\varphi}(u^\delta) - \widetilde{\varphi}(u)|\Big) = |Q|\, \widetilde{\pi}\Big(\tfrac{1}{|Q|}\omega(\delta)\Big).$$

• *Use of contraction arguments and absorption terms (see [8])*
Let us mention one more possibility for getting estimates of kind (2) for (1), which takes advantage of the monotonicity of $\psi(\cdot)$. Assume $\varphi = Id$ in (1); to shorten the arguments, assume $f = 0$. Then L^1 translates in time of $u_h = b(v_h)$ can be estimated with every of the two preceding methods, the Kruzhkov lemma and a direct estimation of translates with variational techniques. This makes $(b(v_h))_h$ relatively compact; yet, when $b^{-1}(\cdot)$ is discontinuous, no information on compactness of $(v_h)_h$ is obtained this way. Now, let us use the translation (in time) invariance of the equation and the L^1 contraction property[11] natural for (1). This yields the estimate (see [8])

$$\int_\Omega |b(v_h^\delta) - b(v_h)|(T-\delta) + \int_s^{T-\delta}\!\!\int_\Omega |\psi(v_h^\delta) - \psi(v_h)| \le \int_\Omega |b(v_h^\delta)(s) - b(v_h)(s)| \quad (6)$$

for all $s \in (0, T-\delta)$, where $v_h^\delta(t) = v_h(t+\delta)$. Integrating in $s > \alpha > 0$, using the time translation bound for $(b(v_h))_h$ we get an $L^1((\alpha, T) \times \Omega)$ estimate of time translates of $\psi(v_h)$. If $\psi(\cdot)$ is strictly increasing, this is enough for L^1_{loc} compactness of $(v_h)_h$.

Applications to (1) and some other parabolic PDEs

• *Application to a parabolic-hyperbolic PDE (see [4])*
For problem (1) with $b = Id$, provided $L^p(Q)$ estimates of the discrete gradient of $\varphi(v_h)$ are available, space translates of $\varphi(v_h)$ (and the functions $\varphi(v_h)$ themselves)

[11] Discrete version of (6) (see [8]) assumes the L^1 contraction property (linked to order preservation via the Crandall-Tartar lemma) is preserved at the discrete level. Estimate (6) is exploited in § 4.

can be estimated uniformly, and an estimate of the form (5) can be obtained. Then the above technique for exploiting (5) assesses the $L^1(Q)$ compactness of $(\varphi(v_h))_h$, which is a first step of the convergence proof for this problem (see [4])[12].

• *Application to an elliptic-parabolic PDE with the structure condition (see [3])*
Assume $\varphi = Id$. Estimate (5) controls the L^1 time translates of $b(v_h)$ similarly to what was described above[13]. If the *structure condition* $\mathbf{G}(v) = \mathbf{F}(b(v))$ is satisfied, compactness of $(b(v_h))_h$ is enough to pass to the limit, see [1] (cf. [10] and § 4).

• *Application to a cross-diffusion system (see [7])*
The following kind of models comes from population dynamics:

$$\begin{cases} u_t - D_1 \Delta u - \operatorname{div}\big((u+v)\nabla u + u\nabla v\big) = u(a_1 - b_1 u - c_1 v), \\ v_t - D_2 \Delta v - \operatorname{div}\big(v\nabla u + (u+v)\nabla v\big) = v(a_2 - b_2 u - c_2 v). \end{cases} \tag{7}$$

Natural estimates for approximate solutions of (7) are L^2 bounds on $\sqrt{1+u+v}\ \nabla u$, $\sqrt{1+u+v}\ \nabla v$; this gives only an $L^{4/3}$ bound on the diffusion fluxes in (7), thus we are not in a variational setting[14]. Therefore for a proof of convergence of finite volume approximations of the kind [13], the Kruzhkov lemma was used in [7][15].

• *Application to convergence of some linearized implicit schemes (see [6])*
In [6], discretization of the simplified version of cardioelectrical bidomain model:

$$\begin{cases} v_t - \operatorname{div}\big[\mathbf{M}_i(\cdot)\nabla u_i\big] + H(v) = I_{ap}(\cdot), \\ v_t + \operatorname{div}\big[\mathbf{M}_e(\cdot)\nabla u_e\big] + H(v) = I_{ap}(\cdot), \end{cases} \qquad v = u_i - u_e, \tag{8}$$

was considered; here, the "ionic current" $H(\cdot)$ is a cubic polynomial. This nonlinear reaction term brings an estimate of $vH(v)$ which bounds v in $L^4(Q)$. Time-implicit DDFV discretization of (8) preserves this structure; then the problem falls into the "variational" framework[16] and time translates can be estimated like in [1, 13, 14]. From the practical point of view, it is important to accelerate computations, and to consider a linearized method where the discretization of the reaction term $H(v)$ is not fully implicit. Unfortunately, for theoretical analysis L^4 estimate for v_h is not available any more; only a weaker estimate can be obtained with interpolation arguments. In [6], we applied the Kruzhkov lemma to exploit this weaker estimate[17].

[12] For Lipschitz $\varphi(\cdot)$, also the Aubin-Lions-Dubinskii-Simon and Kruzhkov lemmas could be used. For general $\varphi(\cdot)$, the author thinks that neither of these lemmas can replace the direct use of (5).

[13] Alternatively, the Kruzhkov lemma can be used in a straightforward way, see [3].

[14] From the practical point of view, e.g. the first equation cannot be tested with $u(t+\delta)$.

[15] Alternatively, the discrete Aubin-Lions-Dubinskii-Simon lemma (see [15]) could be used here.

[16] Indeed, we have v_h bounded in $L^4(Q)$ and $H(v_h)$ is bounded in $L^{4/3}(Q) = (L^4(Q))^*$.

[17] A discrete Aubin-Lions-Dubinskii-Simon argument could have been applied as well.

4 Advanced use of the underlying PDE features

Often mere functional-analytic bounds are not enough, but additional constraints coming from the particular structure of the approximated PDE may permit an indirect compactness/convergence proof. E.g., for the parabolic-hyperbolic PDE (1) (case $b = Id$) we proved the compactness of $(\varphi(v_h))_h$ in § 3. The two final steps (see [4]; see also [14]) exploit fine PDE tools. First, the Minty argument (see, e.g., [1]) is used for $(\mathbf{a_0}(\nabla_h \varphi(v_h))_h$; second, the "nonlinear weak-* convergence" ([4, 13, 14]) for $(v_h)_h$ is upgraded to strong convergence using entropy inequalities for (1).

Let us show how one very delicate case of (1), see [10], can be treated indirectly.

Compactness from monotone penalization and order-preservation

For getting (6), we already used the order-preservation structure for (1). Its further use, in conjunction with penalization, may lead to the following convergence proof.

• *The structure needed for compactification*

Assume that one can prove *uniqueness* of a solution to a PDE (Eq^0) under study. Assume that (Eq^0) can be embedded "continuously" into a family (Eq^ε) of perturbed PDEs *having the property that* $v_h^{\varepsilon_1} \le v_h^{\varepsilon_2}$ *when* $\varepsilon_1 \le \varepsilon_2$, where $v_h^{\varepsilon_1}, v_h^{\varepsilon_2}$ are the associated discrete solutions. Continuity in $\varepsilon \in [-1, 1]$ means, we assume that limits as $\varepsilon \to 0$ (if they exist) of exact solutions v^ε of (Eq^ε) solve the limit equation (Eq^0).

Assume that for $\varepsilon \neq 0$, the corresponding sequence $(v_h^\varepsilon)_h$ is well defined and it converges to an exact solution v^ε of (Eq^ε). Then solutions $(v_h^0)_h$ to the discretized equation (Eq^0) converge a.e., as $h \to 0$, to the unique solution of (Eq^0). Indeed, write

$$v_h^{-1} \le v_h^{-1/2} \le \ldots \le v_h^{-1/m} \le \ldots \le v_h^0 \le \ldots \le v_h^{1/m} \le \ldots \le v_h^{1/2} \le v_h^1, \tag{9}$$

and pass to the limit as $h \to 0$ to define $v^{\pm 1/m} := \lim_{h\to 0} v_h^{\pm 1/m}$ (up to extraction of a subsequence) solution to $(Eq^{\pm 1/m})$; then, (9) is inherited at the limit (except that $(v_h^0)_h$ may not have a limit). By monotonicity, we can define $\underline{v} := \lim_{m\to\infty} v^{-1/m}$ and $\overline{v} := \lim_{m\to\infty} v^{1/m}$; furthermore, we have $\underline{v} \le \liminf_{h\to 0} v_h^0 \le \limsup_{h\to 0} v_h^0 \le \overline{v}$. Both $\underline{v}, \overline{v}$ solve (Eq^0). Thus, by uniqueness, $(v_h^0)_h$ converges to $\underline{v} \equiv \overline{v}$ the solution of (Eq^0).

• *Application to an elliptic-parabolic PDE without the structure condition (see [8])*

We assume that $\varphi = Id$, $\psi = 0$ in (1). We have seen that compactness of $(b(v_h))_h$ can be established, e.g., with the Kruzhkov lemma. Under the *structure condition* $\mathbf{G}(v) = \mathbf{F}(b(v))$, this is enough to pass to the limit in the equation. But in general (see [10]) one lacks control of time oscillations of $\mathbf{G}(v_h)$, and the method of [1] fails. Yet it is enough to add penalization term of the form $\psi^\varepsilon(v) = \varepsilon(\arctan v \mp \frac{\pi}{2}\operatorname{sign}\varepsilon)$ to get into the setting where (6) can be exploited to control discrete solutions $(v_h^\varepsilon)_h$ and to pass to the limit, as $h \to 0$, for the ψ^ε-penalized equation (1^ε). The order-preservation assumptions of the above method being fulfilled due to the choice of ψ^ε, we get convergence of $(v_h)_h$ in the cases where uniqueness for (1) can be shown.

Acknowledgements The author thanks E. Emmrich for discussions on the above techniques.

References

1. H.W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Mat. Z.*, (1983), **183**:311–341.
2. H. Amann. Compact embeddings of vector-valued Sobolev and Besov spaces. *Glasnik Matematički*, (2000), **35**:161–177.
3. B. Andreianov, M. Bendahmane and F. Hubert. On 3D DDFV discretization of gradient and divergence operators. II. Discrete functional analysis tools and applications to degenerate parabolic problems. Preprint HAL, (2011), http://hal.archives-ouvertes.fr/hal-00567342
4. B. Andreianov, M. Bendahmane, and K.H. Karlsen. Discrete duality finite volume schemes for doubly nonlinear degenerate hyperbolic-parabolic equations. *J. Hyp. Diff. Eq.*, (2010), **7**:1–67.
5. B. Andreianov, M. Bendahmane, K.H. Karlsen and S. Ouaro. Well-posedness results for triply nonlinear degenerate parabolic equations. *J. Diff. Eq.*, (2009), **247**(1):277–302.
6. B. Andreianov, M. Bendahmane, K.H. Karlsen and Ch. Pierre. Convergence of Discrete Duality Finite Volume schemes for the macroscopic bidomain model of the heart electric activity. *Netw. Het. Media*, (2011), to appear; available at http://hal.archives-ouvertes.fr/hal-00526047
7. B. Andreianov, M. Bendahmane and R. Ruiz Baier. Analysis of a finite volume method to solve a cross-diffusion population system. *Math. Models Meth. Appl. Sci.*, (2011), to appear.
8. B. Andreianov and P. Wittbold. Convergence of approximate solutions to an elliptic-parabolic equation without the structure condition. Preprint, (2011).
9. J.-P. Aubin. Un théorème de compacité. (French) *C.R. Acad. Sc. Paris*, (1963), **256**:5042–5044.
10. Ph. Bénilan and P. Wittbold. Sur un problème parabolique-elliptique. (French) *M2AN Math. Modelling and Num. Anal.*, (1999), **33**(1):121–127.
11. J.A. Dubinskii. Weak convergence for elliptic and parabolic equations. (Russian) *Math. USSR Sbornik*, (1965), **67**:609–642.
12. E. Emmrich and M. Thalhammer. Doubly nonlinear evolution equations of second order: Existence and fully discrete approximation. *J. Diff. Eq.*, (2011), to appear.
13. R. Eymard, T. Gallouët, and R. Herbin. *Finite Volume Methods.* Handbook of Numerical Analysis, Vol. VII (2000). P. Ciarlet, J.-L. Lions, eds., North-Holland.
14. R. Eymard, T. Gallouët, R. Herbin and A. Michel. Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numer. Math.*, (2002), **92**(1):41–82.
15. T. Gallouët and J.-C. Latché. Compactness of discrete approximate solutions to parabolic PDEs - Application to a turbulence model. *Comm. on Pure and Appl. Anal.*, (2011), to appear.
16. S.N. Kruzhkov. Results on the nature of the continuity of solutions of parabolic equations and some of their applications. *Mat. Zametki (Math. Notes)*, (1969), **6**(1):517-523.
17. J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura ed Appl.*, (1987), **146**:65–96.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Penalty Methods for the Hyperbolic System Modelling the Wall-Plasma Interaction in a Tokamak

Philippe Angot, Thomas Auphan, and Olivier Guès

Abstract The penalization method is used to take account of obstacles in a tokamak, such as the limiter. We study a non linear hyperbolic system modelling the plasma transport in the area close to the wall. A penalization which cuts the transport term of the momentum is studied. We show numerically that this penalization creates a Dirac measure at the plasma-limiter interface which prevents us from defining the transport term in the usual sense. Hence, a new penalty method is proposed for this hyperbolic system and numerical tests reveal an optimal convergence rate without any spurious boundary layer.

Keywords hyperbolic problem, penalization method, numerical tests
MSC2010: 00B25, 35L04, 65M85

1 Introduction

A tokamak is a machine to study plasmas and the fusion reaction. The plasma at high temperature ($10^8 K$) is confined in a toroïdal chamber thanks to a magnetic field. One of the main goals is to perform controlled fusion with enough efficiency to be a reliable source of energy. But, since the magnetic confinement is not perfect, the plasma is in contact with the wall. In order to preserve the integrity of the wall and to limit the pollution of the plasma, it is crucial to control these interactions.

We study, using a fluid approximation of the plasma, a simplified system of equations governing the plasma transport in the scrape-off layer, parallel to the magnetic field lines. In this paper, after a numerical study of the penalization

Philippe Angot, Thomas Auphan, and Olivier Guès
Université de Provence, Laboratoire d'Analyse Topologie et Probabilités, Centre de Mathématiques et Informatique, 39 rue Joliot Curie, 13453 Marseille Cedex 13, France,
e-mail: [angot, tauphan, gues]@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_4, © Springer-Verlag Berlin Heidelberg 2011

introduced by Isoardi *et al.* [9], we modify the boundary conditions to ensure the well-posedness of the hyperbolic system and we propose another penalty method which seems to be free of boundary layer.

2 The model hyperbolic problem

In this paper, we consider a very simple model taking only into account the transport in the direction parallel to the magnetic field lines, (see for example [9, 13]). It is a one dimensional 2×2 hyperbolic system of conservation laws for the particle density N and the particle flux Γ, which reads:

$$\begin{cases} \partial_t N + \partial_x \Gamma = S \\ \partial_t \Gamma + \partial_x \left(\dfrac{\Gamma^2}{N} + N \right) = 0 \qquad\qquad (t,x) \in \mathbb{R}_*^+ \times]-L, L[\\ \text{Initial conditions: } N(0,.) = N_0 \text{ and } \Gamma(0,.) = \Gamma_0 \end{cases} \tag{1}$$

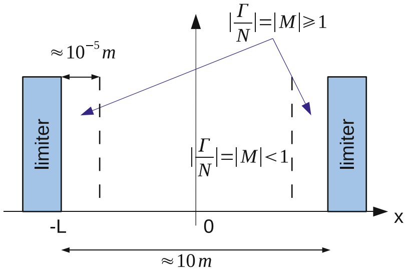
Here, the boundaries of the domain $x = L$ and $x = -L$ correspond to the ”limiters”, which are material obstacles for the fluid (see Fig. 1). In the right-hand side, S is a source term.

There is a difficulty with the choice of the boundary conditions for the system (1). From physical arguments, it follows that the domain (namely the scrape-off layer) is basically divided into two regions [13]:

- One region far from the limiter, the pre-sheath, where the plasma is neutral and the Mach number $M = \Gamma/N$ of the plasma satisfies $|M| \leq 1$.
- One region next to the limiter (in a thin layer called the sheath area, whose typical thickness is of the order of $10^{-5}m$), where the electroneutrality hypothesis does not hold and we have $|M| > 1$. More precisely $M > 1$ close to $x = L$ and $M < -1$ close to the boundary $x = -L$.

It could seem natural to prescribe $M = 1$ (resp. $M = -1$) as a boundary condition at $x = L$ (resp. $x = -L$) for the system, since the physical arguments imply that $M = \pm 1$ very close to the obstacle (Bohm criterion). These are exactly the boundary conditions which are chosen in [9]. However, in that case, as the eigenvalues of the Jacobian of the flux function are $M-1$ and $M+1$, it follows that at the plasma limiter interface one eigenvalue is 0 (the boundary is characteristic) and the other one is outgoing (it is also true at $x = -L$), and clearly the problem does not satisfy the usual sufficient conditions for well posedness, see [3, 8, 11]: the number of boundary conditions ($= 1$) is not equal to the number of incoming eigenvalues ($= 0$).

In order to test our penalty approach with a well-defined hyperbolic boundary value problem, in Sect. 3, we slightly modify the boundary conditions of the paper [9], and impose $M = 1 - \epsilon$ on $x = L$ and $M = -1 + \epsilon$ on $x = -L$ with a fixed $\epsilon > 0$, which leads to a well-posed hyperbolic problem. In our numerical simulations we use $\epsilon = 0.1$.



$|\frac{\Gamma}{N}|=|M|\geqslant 1$
$\approx 10^{-5}m$
limiter
$|\frac{\Gamma}{N}|=|M|<1$
limiter
-L
0
x
$\approx 10m$

Fig. 1 Schematic representation of the scrape-off layer. The x-axis corresponds to the curvilinear coordinate along a magnetic line close to the wall of the tokamak

The numerical tests presented below, use a finite volume scheme with a second order extension: MUSCL reconstruction with the *minmod* slope limiter and the Heun scheme which is a second order Runge–Kutta TVD time discretization. The finite volume scheme is the VFRoe using the non conservative variables for the linearized Riemann solver [7]; here, the non conservatives variables are N and M. To avoid stability issues, the penalized terms are treated implicitly for the time discretization.

3 Study of penalty methods

3.1 A first penalty method

The following penalty approach has been proposed by Isoardi *et al.* [9]. Let's χ be the characteristic function of the limiter, i.e. $\chi(x) = 1$ if x is in the limiter, and $\chi(x) = 0$ elsewhere, and η the penalization parameter. The penalized system is given by:

$$\begin{cases} \partial_t N + \partial_x \Gamma + \dfrac{\chi}{\eta} N = (1-\chi) S_N & \text{in } \mathbb{R}_*^+ \times \mathbb{R} \\ \partial_t \Gamma + (1-\chi)\partial_x \left(\dfrac{\Gamma^2}{N} + N \right) + \dfrac{\chi}{\eta}(\Gamma - M_0 N) = (1-\chi) S_\Gamma & \\ \text{Initial conditions: } N(0,.) = N_0 \text{ and } \Gamma(0,.) = \Gamma_0 & \end{cases} \tag{2}$$

M_0 is a function such that, at the plasma-limiter interface we have $|M_0| = 1$. Here, the two components of the unknown are penalized although there is no incoming wave. At least formally, N is forced to converge to 0 inside the limiter when η tends to 0.

The flux of the second equation is cut inside of the limiter, and this causes some troubles from the mathematical point of view. Indeed, this is an hyperbolic system with discontinuous coefficients and the meaning of the term

$$(1-\chi)\partial_x\left(\frac{\Gamma^2}{N}+N\right)$$

is not clear because it can involve the product of a measure with a discontinuous function which has no distributional sense. As a consequence and as a confirmation of this fact, our numerical tests show the existence of a strong singularity at the interface for the numerical discrete solution. Concerning the interpretation of this numerical singularity, it could happen (but we don't have any rigorous proof and this is just an open question) that this system admits generalized solutions in the spirit of Bouchut–James [4] (see also Poupaud–Rascle [10], or Fornet–Guès [5]) such as measure-valued solutions, which can for example exhibit a Dirac measure at the interface, and this generalized solution could be selected by the numerical approximation process.

For the numerical test, we choose S_N and S_Γ so that the following functions define a solution of the boundary value problem:

$$N(t,x)=\exp\left(\frac{-x^2}{0.16(t+1)}\right)\qquad \Gamma(t,x)=\sin\left(\frac{\pi x}{0.8}\right)\exp\left(\frac{-x^2}{0.16(t+1)}\right)$$

These test solutions are regular (at least inside the plasma area) and has no singularity at the plasma-limiter interface. In the Fig. 2, we observe that a peak appears very quickly, then $|M_i^n|$ become very large (about 10^8) in a few points. The same computations are made for two more refined meshes (respectively for

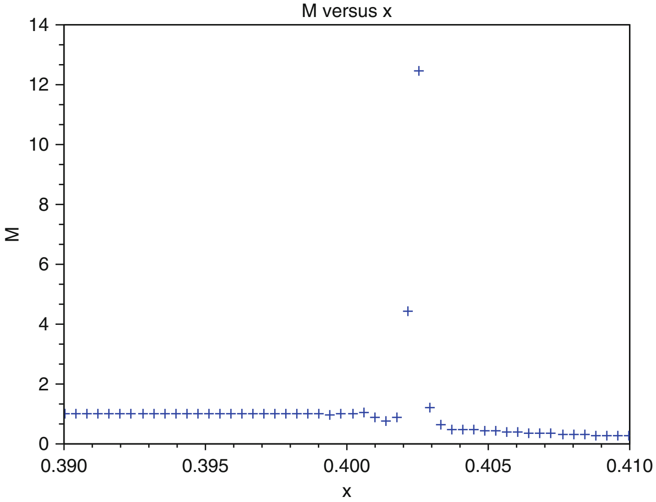


Fig. 2 M versus x with $\eta = 10^{-3}$, a mesh of $J = 1280$ cells using the penalization of Isoardi *et al.* [9]. The computations are stopped when $\max_{i\in\{1,\dots,J\}}(|M_i^n|) > 10$, which corresponds to the time: $t = 0.008822$. The computational domain was $[0, 0.5]$ and $L = 0.4$ (plasma-limiter interface). At $x = 0$, we impose a symmetry condition

2560 and 10240 cells) and we observe that the peak is nearer and nearer to the plasma limiter interface, when the resolution increases. Besides, when the mesh step decreases, the peak appears earlier and earlier. We stop the computations when $\max_{i\in\{1,\dots,J\}}(|M_i^n|) > 10$ but similar results are obtained when the stop criterion is $\max_{i\in\{1,\dots,J\}}(|M_i^n|) > 100$. This leads one to believe that, if the solution converges to a generalized solution of the continuous problem, then this generalized solution must have a singularity supported by the interface (that could be a Dirac measure for example). We notice that the presence of a Dirac measure at the interface is not only a theoretical issue since it has been observed numerically and that the Dirac measure destabilizes numerical schemes. In the following section, we propose a modification of the boundary value problem to obtain a well-posed version.

3.2 A new penalty method for the modified boundary conditions

After the modifications proposed in Sect. 2, the well-posed initial boundary value problem reads:

$$\begin{cases} \partial_t N + \partial_x \Gamma = S \\ \partial_t \Gamma + \partial_x \left(\dfrac{\Gamma^2}{N} + N \right) = 0 \\ M(.,-L) = -1 + \epsilon \text{ and } M(.,L) = 1 - \epsilon \\ N(0,.) = N_0 \text{ and } \Gamma(0,.) = \Gamma_0 \end{cases} \qquad (t,x) \in \mathbb{R}_*^+ \times]-L, L[\qquad (3)$$

For this problem, the boundary is not characteristic, and the boundary conditions are maximally dissipative. Hence, for compatible initial data, the problem has a unique local in time solution, which is regular enough: at least $\mathscr{C}^1$ is sufficient to perform the asymptotic analysis; see e.g. [3, 12].

To penalize (3), we use a method developed in the semi-linear case by Fornet and Guès [6]. In order to have an homogeneous Dirichlet boundary condition for the theoretical study, the system is reformulated with the unknowns $\tilde{u} = \ln(N)$ and $\tilde{v} = \Gamma/N - M_0$. Although our system is quasi-linear (and not semi-linear), the method can be extended to this case. An interesting feature of the method is that it yields to a convergence result without generation of a boundary layer inside the limiter. Up to now, we don't know if this method can be extended to more general quasi-linear first order hyperbolic system with maximally dissipative conditions.

We assume that M_0 is a constant such that $0 < M_0 < 1$. We denote by χ the characteristic function associated to the limiter, i.e. $\chi(x) = 1$ if the point x is in the limiter.

The new penalized problem reads:

$$\begin{cases} \partial_t N + \partial_x \Gamma = S_N \\ \partial_t \Gamma + \partial_x \left(\dfrac{\Gamma^2}{N} + N \right) + \dfrac{\chi}{\eta} \left(\dfrac{\Gamma}{M_0} - N \right) = S_\Gamma \\ N(0,.) = N_0 \text{ and } \Gamma(0,.) = \Gamma_0 \end{cases} \qquad \text{in } \mathbb{R}_*^+ \times \mathbb{R} \qquad (4)$$

The formal asymptotic expansion of a continuous solution to (4) with the BKW (Brillouin–Kramers–Wentzel) method does not contain any boundary layer term [1] and this suggests strongly that there is no boundary layer at all in the solution. Notice that the penalization is incomplete: only one field is penalized, which is natural since there is only one boundary condition.

For the numerical tests, we use a regular solution:

$$N(t,x) = \exp\left(\frac{-x^2}{0.16(t+1)}\right) \qquad \Gamma(t,x) = M_0 \sin\left(\frac{\pi x}{0.8}\right) \exp\left(\frac{-x^2}{0.16(t+1)}\right)$$

and S_N, S_Γ are well chosen. The spatial domain is $[0, 0.5]$ with a symmetry condition at $x = 0$ and the limiter set corresponds to $x \in [0.4, 0.5]$.

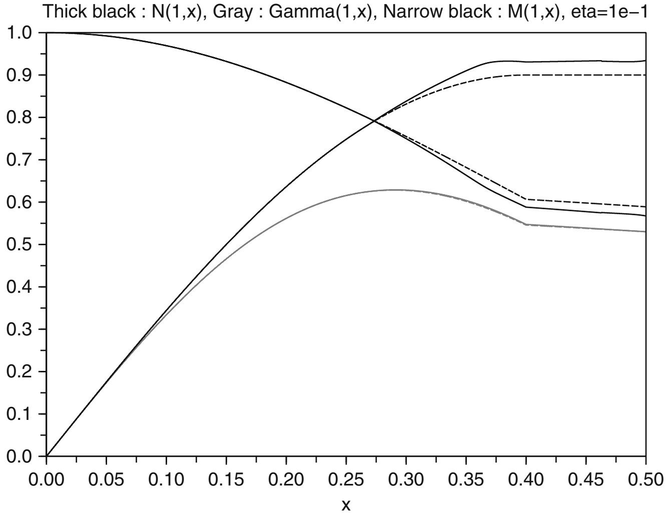


Fig. 3 Plot of N, Γ and M as functions of x (at $t = 1$) with the penalty method free of boundary layer for $\eta = 0.1$. The continuous lines represent the numerical solutions whereas the dashed lines corresponds to the exact solution of the hyperbolic limit problem ($\eta \to 0$). The limiter corresponds to the area $x \in [0.4, 0.5]$. For smaller values of η, for instance for $\eta = 10^{-5}$, the plot is almost the same as the plot of the exact solution (dotted lines)

We analyze the convergence when the penalization parameter η tends to 0 using a uniform spatial mesh of step $\delta x = 10^{-5}$. We calculate the error in L^1 norm for N, $\partial_x N$, Γ and $\partial_x \Gamma$. The goal is to confirm numerically the absence of boundary layer with an optimal rate of convergence as $\mathscr{O}(\eta)$.

One of the main difficulties for the implementation of the penalization, is the choice of a boundary condition at $x = 0.5$ which is necessary for the numerical scheme. As only Γ is penalized, we need a transparent boundary condition for N. For the numerical tests, the boundary condition comes from the asymptotic

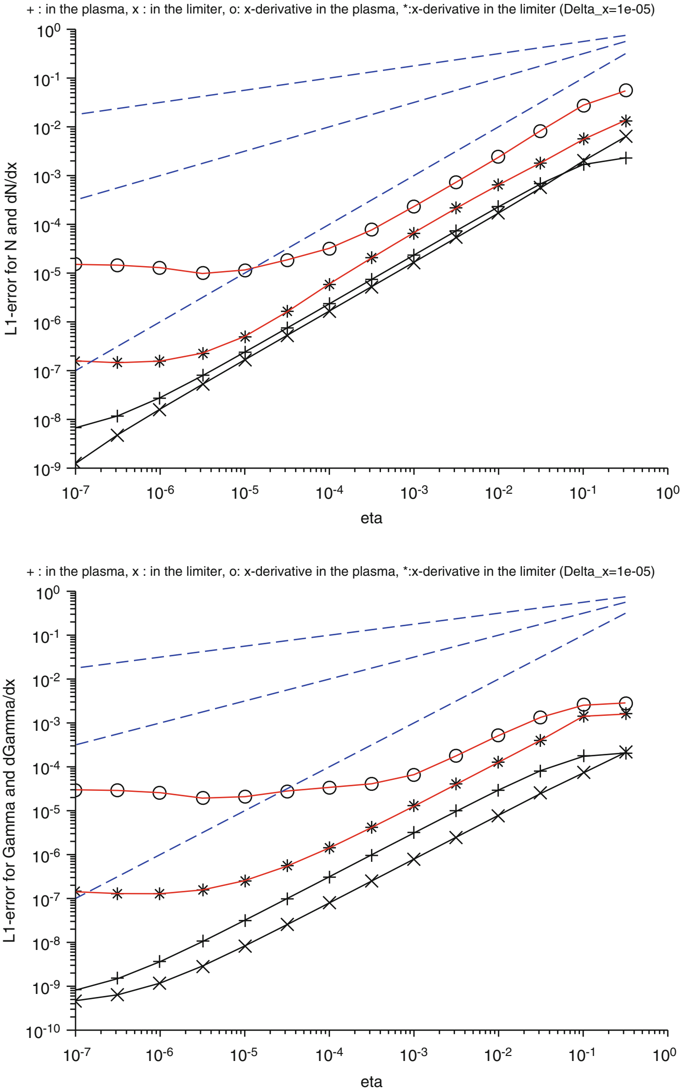


Fig. 4 Errors for N, $\partial_x N$, Γ and $\partial_x \Gamma$ in L^1 norms with the penalization free of boundary layer. The dashed lines represent the curves $\eta^{\frac{1}{4}}$, $\eta^{\frac{1}{2}}$ and η

expansion up to the first order of the BKW analysis. We carry out the computations up to $t = 1$ with an adaptive time step so that the CFL condition is always satisfied. The results are plotted in Fig. 3. In Fig. 4, we observe that the optimal rate of convergence $\mathscr{O}(\eta)$ is reached for the L^1 norms, even for the derivatives. This gives a numerical evidence of the absence of boundary layer. The same numerical results in $\mathscr{O}(\eta)$ are obtained if the penalty term in (4) is replaced by $\frac{\chi}{\eta}\left(\frac{\Gamma}{N} - M_0\right)$, see [2].

When the parameter $\epsilon = 0.01$, i.e. close to a characteristic boundary, the computations show that, for η sufficiently small, $\eta \leq \mathscr{O}(\epsilon)$, the convergence results are similiar; see details in [1].

Acknowledgements This work has been funded by the ANR ESPOIR (Edge Simulation of the Physics Of ITER Relevant turbulent transport)and the *Fédération nationale de Recherche Fusion par Confinement Magnétique* (FR-FCM). We thank Guillaume Chiavassa, Guido Ciraolo and Philippe Ghendrih for fruitful discussions.

References

1. Angot, P., Auphan, P., Guès, O.: An optimal penalty method for the hyperbolic system modelling the edge plasma transport in a tokamak. Preprint in preparation (2011)
2. Auphan, T.: Méthodes de pénalisation pour des systèmes hyperboliques application au transport de plasma en bord de tokamak. Master's thesis, Ecole Centrale Marseille (2010)
3. Benzoni-Gavage, S., Serre, D.: Multidimensional hyperbolic partial differential equations. First-order systems and applications. Oxford Mathematical Monographs. Oxford University Press (2007)
4. Bouchut, F., James, F.: One-dimensional transport equations with discontinuous coefficients. Nonlinear Anal. **32**, 891–933 (1998)
5. Fornet, B.: Small viscosity solution of linear scalar 1-d conservation laws with one discontinuity of the coefficient. Comptes Rendus Mathematique **346**(11-12), 681 – 686 (2008)
6. Fornet, B., Guès, .: Penalization approach of semi-linear symmetric hyperbolic problems with dissipative boundary conditions. Discrete and Continuous Dynamical Systems **23**(3), 827 – 845 (2009)
7. Gallouët, T., Hérard, J.M., Seguin, N.: Some approximate godunov schemes to compute shallow-water equations with topography. Computers and Fluids **32**(4), 479 – 513 (2003)
8. Guès, O.: Problème mixte hyperbolique quasi-linéaire caractéristique. Communications in Partial Differential Equations **15**, 595–654 (1990)
9. Isoardi, L., Chiavassa, G., Ciraolo, G., Haldenwang, P., Serre, E., Ghendrih, P., Sarazin, Y., Schwander, F., Tamain, P.: Penalization modeling of a limiter in the tokamak edge plasma. Journal of Computational Physics **229**(6), 2220 – 2235 (2010)
10. Poupaud, F., Rascle, M.: Measure solutions to the linear multi-dimensional transport equation with non-smooth coefficients. Communications in Partial Differential Equations **22**, 225–267 (1997)
11. Rauch, J.B.: Symmetric positive systems with boundary characteristic of constant multiplicity. Trans. Amer. Math. Soc. **291**(1), 167–187 (1985)
12. Rauch, J.B., Massey, F.J.I.: Differentiability of solutions to hyperbolic initial-boundary value problems. Trans. Amer. Math. Soc. **189**, 303–318 (1974)
13. Tamain, P.: Etude des flux de matière dans le plasma de bord des tokamaks, alimentation, transport et turbulence. Ph.D. thesis, Université de Provence (2007)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Spectacular Vector Penalty-Projection Method for Darcy and Navier-Stokes Problems

Philippe Angot, Jean-Paul Caltagirone, and Pierre Fabrie

Abstract We present a new *fast vector penalty-projection method (VPP$_\varepsilon$)*, issued from noticeable improvements of previous works [3,4,7], to efficiently compute the solution of unsteady Navier-Stokes/Brinkman problems governing incompressible multiphase viscous flows. The method is also efficient to solve anisotropic Darcy problems. The key idea of the method is to compute at each time step an accurate and curl-free approximation of the pressure gradient increment in time. This method performs a *two-step approximate divergence-free vector projection* yielding a velocity divergence vanishing as $\mathscr{O}(\varepsilon\,\delta t)$, δt being the time step, with a penalty parameter ε as small as desired until the machine precision, *e.g.* $\varepsilon = 10^{-14}$, whereas the solution algorithm can be extremely fast and cheap. The method is numerically validated on a benchmark problem for two-phase bubble dynamics where we compare it to the Uzawa augmented Lagrangian (UAL) and scalar incremental projection (SIP) methods. Moreover, a new test case for fluid-structure interaction problems is also investigated. That results in a robust method running faster than usual methods and being able to efficiently compute accurate solutions to sharp test cases whatever the density, viscosity or anisotropic permeability jumps, whereas other methods crash.

Keywords Vector penalty-projection; Penalty method; Splitting method; Multiphase Navier-Stokes/Brinkman; Anisotropic Darcy problem; Incompressible flows
MSC 2010: 35Q30, 35Q35, 65M12, 65M85, 65N12, 65N85, 74F10, 76D05, 76D45, 76M25, 76R10, 76S05, 76T10

Philippe Angot
Aix-Marseille Université, LATP - CMI UMR CNRS 6632, 39 rue F. Joliot Curie, 13453 Marseille Cedex 13 - France, e-mail: angot@cmi.univ-mrs.fr

Jean-Paul Caltagirone
Université de Bordeaux & IPB, IMIB, 16 Av Pey-Berland 33607 Pessac - France,
e-mail: calta@enscbp.fr

Pierre Fabrie
Université de Bordeaux & IPB, IMB UMR CNRS 5251, ENSEIRB-MATMECA,
Talence - France, e-mail: pierre.fabrie@math.u-bordeaux1.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_5, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction to model incompressible multiphase flows

Let $\Omega \subset \mathbb{R}^d$ ($d = 2$ or 3 in practice) be an open bounded and connected domain with a Lipschitz continuous boundary $\Gamma = \partial\Omega$ and $\mathbf{n}$ be the outward unit normal vector on Γ. For $T > 0$, we consider the following unsteady Navier-Stokes/Brinkman problem [9] governing incompressible non-homogeneous or multiphase flows where Dirichlet boundary conditions for the velocity $\mathbf{v}_{|\Gamma} = 0$ on Γ, the volumic force $\mathbf{f}$ and initial data $\mathbf{v}(t = 0) = \mathbf{v}_0$, $\varphi(t = 0) = \varphi_0 \in L^\infty(\Omega)$ with $\varphi_0 \geq 0$ *a.e.* in Ω, are given. For sake of briefness here, we just focus on the model problem (1-3) where $\mathbf{d}(\mathbf{v}) = (\nabla\mathbf{v} + (\nabla\mathbf{v})^T)/2$, as a part of more complex fluid mechanics problems.

$$\rho\,(\partial_t\,\mathbf{v} + (\mathbf{v}\cdot\nabla)\mathbf{v}) - 2\,\nabla\cdot\,(\mu\,\mathbf{d}(\mathbf{v})) + \mu\,\mathbf{K}^{-1}\,\mathbf{v} + \nabla p = \mathbf{f} \qquad \text{in } \Omega \times (0,T) \quad (1)$$

$$\nabla\cdot\mathbf{v} = 0 \qquad \text{in } \Omega \times (0,T) \quad (2)$$

$$\partial_t\,\varphi + \mathbf{v}\cdot\nabla\varphi = 0 \qquad \text{in } \Omega \times (0,T). \quad (3)$$

The permeability tensor $\mathbf{K}$ in the Darcy term is supposed to be symmetric, uniformly positive definite and bounded in Ω. We refer to [1, 9] for the modeling of flows inside complex fluid-porous-solid heterogeneous systems with the Navier-Stokes/Brinkman or Darcy equations. The equation (3) for the positive phase function φ governs the transport by the flow of the interface between two phases, either fluid or solid, respectively in the case of two-phase fluid flows or fluid-structure interaction problems. The force $\mathbf{f}$ may include some volumic forces like the gravity force $\rho\,\mathbf{g}$ as well as the surface tension force to describe the capillarity effects at the phase interfaces Σ. The advection-diffusion equation for the temperature $\mathscr{T}$ is not precised here and we assume some given state laws: $\rho = \rho(\varphi, \mathscr{T})$ and $\mu = \mu(\varphi, \mathscr{T})$ for each phase, where the functions are continuous and positive.

2 The fast vector-penalty projection method (VPP$_\varepsilon$)

2.1 The (VPP$_\varepsilon$) method for multiphase Navier-Stokes/Brinkman

We describe hereafter the two-step vector penalty-projection (VPP$_\varepsilon$) method with a penalty parameter $0 < \varepsilon \ll 1$; see more details in [5]. For φ^0 with $\varphi^0 \geq 0$ *a.e.* in Ω, $\mathbf{v}^0$ and $p^0 \in L^2_0(\Omega)$ given, the method reads as below with usual notations for the semi-discrete setting in time, $\delta t > 0$ being the time step. For all $n \in \mathbb{N}$ such that $(n + 1)\,\delta t \leq T$, find $\tilde{\mathbf{v}}^{n+1}$, $\mathbf{v}^{n+1}$, $p^{n+1} \in L^2_0(\Omega)$, $\varphi^{n+1} \in L^\infty(\Omega)$, such that:

$$\rho^n\left(\frac{\tilde{\mathbf{v}}^{n+1}-\mathbf{v}^n}{\delta t}+(\mathbf{v}^n\cdot\nabla)\tilde{\mathbf{v}}^{n+1}\right)-2\boldsymbol{\nabla}\cdot\left(\mu^n\,\mathbf{d}(\tilde{\mathbf{v}}^{n+1})\right)+\mu^n\,\mathbf{K}^{-1}\,\tilde{\mathbf{v}}^{n+1}+\nabla p^n=\mathbf{f}^n \quad (4)$$

$$\frac{\varepsilon}{\delta t}\,\rho^n\,\hat{\mathbf{v}}^{n+1}-\nabla\left(\boldsymbol{\nabla}\cdot\hat{\mathbf{v}}^{n+1}\right)=\nabla\left(\boldsymbol{\nabla}\cdot\tilde{\mathbf{v}}^{n+1}\right) \quad (5)$$

$$\mathbf{v}^{n+1}=\tilde{\mathbf{v}}^{n+1}+\hat{\mathbf{v}}^{n+1},\qquad\text{and}\qquad\nabla(p^{n+1}-p^n)=-\frac{\rho^n}{\delta t}\,\hat{\mathbf{v}}^{n+1} \quad (6)$$

$$p^{n+1}=p^n+\phi^{n+1}\qquad\text{with }\ \phi^{n+1}\text{ reconstructed from }\ \nabla\phi^{n+1}=-\frac{\rho^n}{\delta t}\,\hat{\mathbf{v}}^{n+1} \quad (7)$$

$$\frac{\varphi^{n+1}-\varphi^n}{\delta t}+\mathbf{v}^{n+1}\cdot\nabla\varphi^n=0 \quad (8)$$

with: $\tilde{\mathbf{v}}^{n+1}_{|\Gamma}=0$, or for non homogeneous Dirichlet conditions: $\tilde{\mathbf{v}}^{n+1}_{|\Gamma}=\mathbf{v}^{n+1}_D$, and $\hat{\mathbf{v}}^{n+1}\cdot\mathbf{n}_{|\Gamma}=0$. Here $\mathbf{v}^n$, p^n are desired to be first-order approximations of the exact velocity and pressure solutions $\mathbf{v}(t_n), p(t_n)$ at time $t_n = n\,\delta t$. Since the end-of-step velocity divergence is not exactly zero, the additional spherical part $\lambda\,\boldsymbol{\nabla}\cdot\mathbf{v}\,\mathbf{I}$ of the Newtonian stress tensor is included within the dynamical pressure gradient ∇p. Once the equations (4-8) have been solved, the advection-diffusion equation of temperature can be solved too for $\mathscr{T}^{n+1}$ and we can find: $\rho^{n+1}=\rho(\varphi^{n+1},\mathscr{T}^{n+1})$ and $\mu^{n+1}=\mu(\varphi^{n+1},\mathscr{T}^{n+1})$.

The key feature of our method is to calculate an accurate and curl-free approximation of the momentum vector correction $\rho^n\,\hat{\mathbf{v}}^{n+1}$ in (5). Indeed (5-6) ensures that $\rho^n\,\hat{\mathbf{v}}^{n+1}$ is exactly a gradient which justifies the choice for $\nabla\phi^{n+1}=\nabla(p^{n+1}-p^n)$ since we have:

$$\rho^n\,\hat{\mathbf{v}}^{n+1}=\frac{\delta t}{\varepsilon}\nabla\left(\boldsymbol{\nabla}\cdot\mathbf{v}^{n+1}\right)\ \Rightarrow\quad\nabla(p^{n+1}-p^n)=-\frac{\rho^n}{\delta t}\,\hat{\mathbf{v}}^{n+1}=-\frac{1}{\varepsilon}\nabla\left(\boldsymbol{\nabla}\cdot\mathbf{v}^{n+1}\right). \quad (9)$$

The (VPP$_\varepsilon$) method effectively takes advantage of the splitting method proposed in [4] for augmented Lagrangian systems or general saddle-point computations to get a very fast solution of (5); see Theorem 1. When we need the pressure field itself, e.g. to compute stress vectors, it is calculated in an incremental way as an auxiliary step. We propose to reconstruct $\phi^{n+1}=p^{n+1}-p^n$ from its gradient $\nabla\phi^{n+1}$ given in (6) with the following method.

Reconstruction of $\phi^{n+1}=p^{n+1}-p^n$ from its gradient.

By circulating on a suitable path starting at a point on the border where $\phi^{n+1}=0$ is fixed and going through all the pressure nodes in the mesh, we get with the gradient formula between two neighbour points A and B using the mid-point quadrature:

$$\phi^{n+1}(B)-\phi^{n+1}(A)=\int_A^B\nabla\phi^{n+1}\cdot d\mathbf{l}=-\int_A^B\frac{\rho^n}{\delta t}\,\hat{\mathbf{v}}^{n+1}\cdot d\mathbf{l}\approx-\frac{\rho^n}{\delta t}\,|\hat{\mathbf{v}}^{n+1}|\,h_{AB} \quad (10)$$

with $h_{AB}=$ distance (A,B). The field ϕ^{n+1} is calculated point by point from the boundary and then passing successively by all the pressure nodes. This fast algorithm is performed at each time step to get the pressure field p^{n+1} from the known field p^n. We refer to [5] for more details and validations on the present method.

2.2 The (VPP$_\varepsilon$) method for anisotropic Darcy problems

We present below the fast solution to incompressible Darcy flow problems in porous media with the (VPP$_\varepsilon$) method. The model problem reads in dimensionless form:

$$s\,\partial_t\,\mathbf{v} + \mu\,\mathbf{K}^{-1}\,\mathbf{v} + \nabla p = \mathbf{f} \qquad \text{in } \Omega \times (0,T) \tag{11}$$

$$\nabla\cdot\mathbf{v} = 0 \qquad \text{in } \Omega \times (0,T) \tag{12}$$

$$\mathbf{v}\cdot\mathbf{n} = 0 \qquad \text{on } \Gamma \times (0,T) \tag{13}$$

where the viscosity $\mu > 0$ is constant and the permeability tensor $\mathbf{K}$ is supposed to be symmetric, bounded in Ω and uniformly positive definite. The dimensionless stationarity parameter $s > 0$ includes the Darcy number: $Da = K_{ref}/L_{ref}^2$ and thus we have $s \ll 1$ for most practical problems or even $s = 0$ for the steady anisotropic Darcy problem. The equations (11-13) also model flows inside heterogeneous porous-solid systems by letting the permeability tend to zero inside the impermeable media; see also [1, 9] for the analysis and validations of the so-called L^2 volume penalty method.

The (VPP$_\varepsilon$) method with $r = \mathscr{O}(\varepsilon) > 0$ and $0 < \varepsilon \ll 1$ to solve (11-13) reads as follows. For all $n \in \mathbb{N}$ such that $(n+1)\,\delta t \le T$, find $\tilde{\mathbf{v}}^{n+1}$, $\mathbf{v}^{n+1}$ and p^{n+1} such that:

$$s\,\frac{\tilde{\mathbf{v}}^{n+1} - \mathbf{v}^n}{\delta t} + \mu\,\mathbf{K}^{-1}\,\tilde{\mathbf{v}}^{n+1} - r\,\nabla\left(\nabla\cdot\tilde{\mathbf{v}}^{n+1}\right) + \nabla p^n = \mathbf{f}^n \tag{14}$$

$$\varepsilon\left(\frac{s}{\delta t} + \mu\,\mathbf{K}^{-1}\right)\hat{\mathbf{v}}^{n+1} - \nabla\left(\nabla\cdot\hat{\mathbf{v}}^{n+1}\right) = \nabla\left(\nabla\cdot\tilde{\mathbf{v}}^{n+1}\right) \tag{15}$$

$$\mathbf{v}^{n+1} = \tilde{\mathbf{v}}^{n+1} + \hat{\mathbf{v}}^{n+1},$$

$$\text{and} \quad \nabla(p^{n+1} - p^n) = -\left(\frac{s}{\delta t} + \mu\,\mathbf{K}^{-1}\right)\hat{\mathbf{v}}^{n+1} - r\,\nabla\left(\nabla\cdot\tilde{\mathbf{v}}^{n+1}\right) \tag{16}$$

$$p^{n+1} = p^n + \phi^{n+1} \quad \text{with } \phi^{n+1} \text{ reconstructed from its gradient } \nabla\phi^{n+1} \tag{17}$$

with the boundary conditions: $\tilde{\mathbf{v}}^{n+1}\cdot\mathbf{n}_{|\Gamma} = 0$ and $\hat{\mathbf{v}}^{n+1}\cdot\mathbf{n}_{|\Gamma} = 0$ on Γ. The space discrete solution to the prediction step (14) is explicit for s and r sufficiently small to invert a perturbation of the Identity matrix with a Neumann asymptotic expansion.

3 On the fast discrete solution to the (VPP$_\varepsilon$) method

The great interest for solving (5) or (15) instead of a usual augmented Lagrangian problem lies in the following result issued from [4] which shows that the method can be ultra-fast and very cheap if $\eta = \varepsilon/\delta t$ is sufficiently small.

Let us now consider any space discretization of our problem. We denote by $B = -div_h$ the $m \times n$ matrix corresponding to the discrete divergence operator, $B^T = grad_h$ the $n \times m$ matrix corresponding to the discrete gradient operator, whereas I denotes the $n \times n$ identity matrix with $n > m$ and D the $n \times n$ diagonal nonsingular matrix containing all the discrete density values of $\rho^n > 0$ *a.e.* in Ω. Here n is the number of velocity unknowns whereas m is the number of pressure unknowns. Then, the discrete vector penalty-projection problem corresponding to (5) with $\varepsilon = \eta\,\delta t$ reads:

$$\left(D + \frac{1}{\eta}\, B^T B\right) \hat{v}_\eta = -\frac{1}{\eta}\, B^T B\, \tilde{v}, \qquad \text{with} \qquad v_\eta = \tilde{v} + \hat{v}_\eta. \tag{18}$$

We proved in [4] the crucial result below due to the *adapted right-hand side* in the correction step (18) which lies in the range of the limit operator $B^T B$. Indeed, (18) can be viewed as a singular perturbation problem with well-suited data in the right-hand side. More precisely, we give in Theorem 1 the zero-order term of the solution $\hat{v}_\eta$ to (18):

$$\hat{v}_\eta = -\frac{1}{\eta}\left(D + \frac{1}{\eta}\, B^T B\right)^{-1} B^T B\, \tilde{v} \tag{19}$$

when the penalty parameter η is chosen sufficiently small; see the asymptotic expansion of $\hat{v}_\eta$ and the proof in [4, Theorem 1.1 and Corollary 1.3].

Theorem 1 (Fast solution of the discrete vector penalty-projection). *Let D be an $n \times n$ positive definite diagonal matrix, I the $n \times n$ identity matrix and B an $m \times n$ matrix. If the rows of B are linearly independent, rank$(B) = m$, then for all η small enough, $0 < \eta < 1/\|S^{-1}\|$ where $S = BD^{-1}B^T$, there exists an $n \times n$ matrix C_1 bounded independently on η such that the solution of the correction step (19) writes for any vector $\tilde{v} \in \mathbb{R}^n$:*

$$\hat{v}_\eta = C_0\, \tilde{v} + \eta\, C_1\, \tilde{v} \quad \textit{with} \quad C_0 = -D^{-1} B^T S^{-1} B = -D^{-1} B^T (BD^{-1}B^T)^{-1} B. \tag{20}$$

If rank$(B) = p < m$, there exists a surjective $p \times n$ matrix T such that $B^T B = T^T T$ and a similar result holds replacing B by T.

Hence, for a constant density $\rho > 0$ and choosing now $\eta = \rho\,\varepsilon/\delta t$, we have: $D = I$, $S = BB^T$ and $C_0 = -B^T S^{-1} B = -B^T(BB^T)^{-1}B$. Moreover, if rank$(B) = p \le m \le n$, the zero-order solution $\hat{v} = C_0\,\tilde{v}$ in (20) is the solution of minimal Euclidean norm in $\mathbb{R}^n$ to the linear system: $B\,\hat{v} = -B\,\tilde{v}$ by the least-squares method, and the matrix $B^\dagger = B^T(BB^T)^{-1}$ is the Moore-Penrose pseudo-inverse of B such that $C_0 = -B^\dagger B$. Indeed, a singular value decomposition (SVD) or a QR factorization of B yields: $C_0 = -I_0$ where I_0 is the $n \times n$ diagonal *matrix having only* 1 *or* 0 *coefficients, the zero entries in the diagonal being the* $n - p$ *null eigenvalues of the operator* $B^T B$.

Hence, for η small enough, the computational effort required to solve (18) amounts to approximate the matrix C_0 which includes both D and D^{-1} inside non commutative products. Thus, we always use the diagonal preconditioning in the case of a variable density which makes the effective condition number quasi-independent on the density or permeability jumps. We also use the Jacobi preconditioner in the prediction step (4) to cope with the viscosity or permeability jumps as performed in [9]. However, for a constant density when $D = I$, we get $C_0 = -I_0$. This explains why the solution can be obtained with only one iteration of a suitable preconditioned Krylov solver whatever the size of the mesh step or the dimension n; see the numerical results in [4].

4 Numerical validations with discrete operator calculus

The (VPP$_\varepsilon$) method has been implemented with discrete exterior calculus (DEC) methods, see the recent review in [6], for the space discretization of the Navier-Stokes equations on unstructured staggered meshes. The (DEC) methods ensure primary and secondary discrete conservation properties. In particular, the space discretization satisfies for the discrete operators: $\nabla_h \times (\nabla_h \phi) = 0$ and $\nabla_h \cdot (\nabla_h \times \psi) = 0$, which is not usually verified by other methods; see [6]. Hence, the (VPP$_\varepsilon$) method is now validated on unstructured meshes both in 2-D or 3-D.

The structure and solver of the computational code are issued from previous works, originally implemented with a Navier-Stokes finite volume solver on the staggered MAC mesh and using the Uzawa augmented Lagrangian (UAL) method to deal with the divergence-free constraint; see [9]. We refer to [1, 2, 9] and the references therein for the analysis and numerical validations of the fictitious domain model using the so-called L^2 or H^1-penalty methods to take account of obstacles in flow problems with the Navier-Stokes/Brinkman equations. Hence, our approach is essentially Eulerian with a Lagrangian front-tracking of the sharp interfaces accurately reconstructed on the fixed Eulerian mesh, see e.g. [10, 11] and the references therein. Thus we use no Arbitrary Lagrangian-Eulerian (ALE) method, no global remeshing nor moving mesh method.

4.1 Multiphase flows: dispersed two-phase bubble dynamics

The (VPP$_\varepsilon$) method is numerically validated for multiphase incompressible flows by performing with the three methods (UAL), (SIP) and (VPP), the benchmark problem studied in [8] for 2-D bubble dynamics. In that problem, we compute the first test case which considers an initial circular bubble of diameter $0.05\,m$ with density and

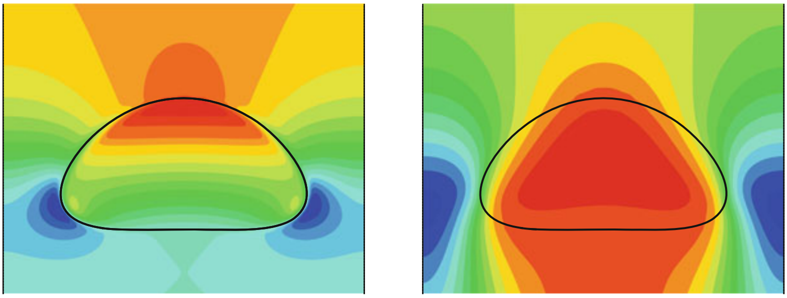


Fig. 1 Benchmark for 2-D bubble dynamics with (VPP$_\varepsilon$) method, $\varepsilon = 10^{-8}$: motion of a circular bubble with surface tension at time $t = 3$ and Re $= 35$ - bubble initial diameter $\oslash = 0.05$, $\rho_1/\rho_2 = 1000/100 = 10$, $\mu_1/\mu_2 = 10/1 = 10$, domain 0.1×0.2, mesh size 128×256, $\delta t = 0.007143$, circular bubble initially with no motion at height $y = 0.05$. LEFT: isobars and isoline $\varphi = 0.5$ of the phase function at interface. RIGHT: superposition of isoline $\varphi = 0.5$ at interface for (UAL), (SIP), (VPP) and vertical velocity field (in absolute referential)

viscosity ratios equal to 10 which undergoes moderate shape deformation. In this case, the bubble is driven up by the external gravity force $\mathbf{f} = \rho\,\mathbf{g}$, whereas the surface tension effect on the interface Σ between the two fluid phases is taken into account through the following force balance at the interface Σ:

$$[\![\mathbf{v}]\!]_\Sigma = 0 \ \text{ and } \ [\![\left(-p\,\mathbf{I} + \mu\,\left(\nabla\mathbf{v} + (\nabla\mathbf{v})^T\right)\right) \cdot \mathbf{n}]\!]_\Sigma = \sigma\,\kappa\,\mathbf{n}_{|\Sigma}, \text{ or } \mathbf{f}_{st} = \sigma\,\kappa\,\mathbf{n}_{|\Sigma}\,\delta_\Sigma$$

where $\sigma = 24.5$ is the surface tension coefficient, κ the local curvature of the interface, $\mathbf{n}_{|\Sigma}$ the outward unit normal to the interface and δ_Σ the Dirac measure supported by the interface Σ. The solution of the phase transport (3) is carried out by the so-called *VOF-PLIC* method, *i.e.* the famous *VOF* method using a piecewise linear interface construction proposed in [12] to precisely reconstruct the sharp interface Σ at the isoline $\varphi = 0.5$, with $\varphi^0 = 0$ in Ω_1 and $\varphi^0 = 1$ in Ω_2; see [10, 11].

The results of the three methods (UAL), (SIP) and (VPP) after 420 time iterations are presented in Fig. 1 by superposing the different fields to get a more precise comparison. We observe an excellent agreement both between the three methods and the reference solution in [8]. However, the (VPP) method runs faster.

4.2 A test case for fluid-structure interaction problems

To evaluate the robustness of the (VPP$_\varepsilon$) method with respect to large density or viscosity ratios, we compute the motion of an heavy solid body which freely falls vertically in air with the gravity force $\mathbf{f} = \rho_s\,\mathbf{g}$. The rigid behaviour of the body is obtained by letting the viscosity μ_s tend to infinity inside the ball in order to penalize the tensor of deformation rate $\mathbf{d}(\mathbf{v})$. This fictitious domain method using a

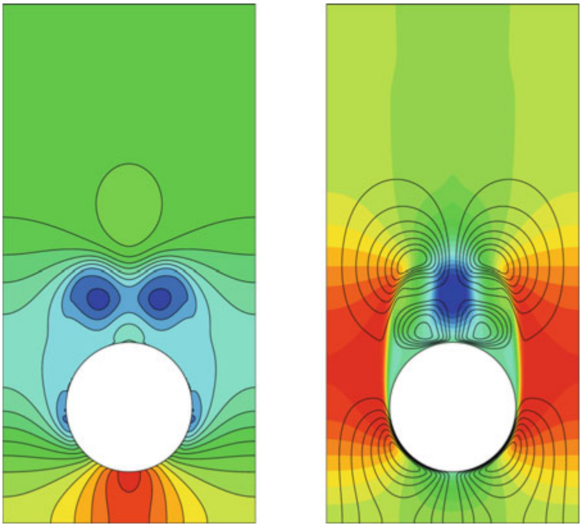


Fig. 2 ACF11-ball with (VPP$_\varepsilon$) method, $\varepsilon = 10^{-6}$: free fall of a heavy solid body in air at time $t = 0.15$ and Re $= 7358$ - Cylinder diameter $\oslash = 0.05$, $\rho_s = 10^6$, $\rho_f = 1$, $\mu_s = 10^{12}$, $\mu_f = 10^{-5}$, domain 0.1×0.2, mesh size 256×512, $\delta t = 0.0002$, cylinder initially with no motion at height $y = 0.15$. LEFT: isobars and isoline $\varphi = 0.5$ of the phase function at interface. RIGHT: vertical velocity field and horizontal velocity isolines

penalty was studied in [1] (see the references therein) to design a numerical wind-tunnel, then numerically validated in several works, e.g. [11], and also analyzed theoretically in [1, 2] where optimal global error estimates are proved for the H^1 penalty method. Moreover, this fictitious domain method allows us to easily compute the forces applied on the obstacle, see [9]; the error estimate being proved in [1] when the nonlinear convection term is neglected inside the solid obstacle.

The results obtained by the (VPP$_\varepsilon$) method are presented in Fig. 2 at time $t = 0.15\,s$ after 750 time iterations when the ball velocity reaches: $V_b = g\,t = 1.4715\,m/s$. The computation shows that the strain rate tensor inside the ball Ω_s vanishes as $\|\mathbf{d}(\mathbf{v})\|_{L^2(\Omega_s)} = \mathcal{O}(\mu_f/\mu_s)$, *i.e.* of the order of the machine precision. Hence, the (VPP$_\varepsilon$) method efficiently ensures both the rigidity of the solid body and a velocity divergence vanishing as $\mathcal{O}(\varepsilon\,\delta t)$ [5], whereas it avoids the blocking effect observed with other methods; see e.g. [11].

The (SIP) method crashes after a few time iterations. The (UAL) method is still able to compute the flow with a larger velocity divergence and the computation is far more expensive than with the (VPP$_\varepsilon$) method.

References

1. PH. ANGOT, Analysis of singular perturbations on the Brinkman problem for fictitious domain models of viscous flows, Math. Meth. in the Appl. Sci. ($M2AS$) **22**(16), 1395-1412, 1999.
2. PH. ANGOT, C.-H. BRUNEAU AND P. FABRIE, A penalization method to take into account obstacles in incompressible viscous flows, Nümerische Mathematik **81**(4), 497-520, 1999.
3. PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, Vector penalty-projection methods for the solution of unsteady incompressible flows, in *Finite Volumes for Complex Applications V*, R. Eymard and J.-M. Hérard (Eds), pp. 169-176, ISTE Ltd and J. Wiley & Sons, 2008.
4. PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, A new fast method to compute saddle-points in constrained optimization and applications, Appl. Math. Letters, 2011 (submitted).
5. PH. ANGOT, J.-P. CALTAGIRONE AND P. FABRIE, A fast vector penalty-projection method for incompressible non-homogeneous or multiphase Navier-Stokes problems, Applied Mathematics Letters, 2011 (submitted).
6. J. BLAIR PEROT, Discrete conservation properties of unstructured mesh schemes, Annu. Rev. Fluid Mech. **43**, 299-318, 2011.
7. J.-P. CALTAGIRONE AND J. BREIL, Sur une méthode de projection vectorielle pour la résolution des équations de Navier-Stokes, C. R. Acad. Sci. Paris, IIb **327**, 1179-1184, 1999.
8. S. HYSING, S. TUREK, D. KUZMIN, N. PAROLINI, E. BURMAN, S. GANESAN AND L. TOBISKA, Quantitative benchmark computations of two-dimensional bubble dynamics, Int. J. Numer. Meth. Fluids, **60**, 1259-1288, 2009.
9. K. KHADRA, PH. ANGOT, S. PARNEIX AND J.-P. CALTAGIRONE, Fictitious domain approach for numerical modelling of Navier-Stokes equations, Int. J. Numer. Meth. in Fluids, **34**(8), 651-684, 2000.
10. A. SARTHOU, S. VINCENT, J.-P. CALTAGIRONE AND PH. ANGOT, Eulerian-Lagrangian grid coupling and penalty methods for the simulation of multiphase flows interacting with complex objects, Int. J. Numer. Meth. in Fluids **56**(8), 1093-1099, 2008.
11. S. VINCENT, A. SARTHOU, J.-P. CALTAGIRONE, F. SONILHAC, P. FÉVRIER, C. MIGNOT AND G. PIANET, Augmented Lagrangian and penalty methods for the simulation of two-phase flows interacting with moving solids. Application to hydroplaning flows interacting with real tire tread patterns, J. Comput. Phys. **230**, 956-983, 2011.
12. D.L. YOUNGS, Time-dependent multimaterial flow with large fluid distortion, in *Numerical Methods for Fluid Dynamics*, K.W. Morton, M.J. Baines (Eds.), Academic Press, 1982.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Numerical Front Propagation Using Kinematical Conservation Laws

K.R. Arun, M. Lukáčová-Medviďová, and P. Prasad

Abstract We use the newly formulated three-dimensional (3-D) kinematical conservation laws (KCL) to study the propagation of a nonlinear wavefront in a polytropic gas in a uniform state at rest. The 3-D KCL forms an under-determined system of six conservation laws with three involutive constraints, to which we add the energy conservation equation of a weakly nonlinear ray theory. The resulting system of seven conservation laws is only weakly hyperbolic and therefore poses a real challenge in the numerical approximation. We implement a central finite volume scheme with a constrained transport technique for the numerical solution of the system of conservation laws. The results of a numerical experiment is presented, which reveals some interesting geometrical features of a nonlinear wavefront.

Keywords kinematical conservation laws, kink, weakly nonlinear ray theory, wavefront, polytropic gas
MSC2010: 35L60, 35L65, 35L67, 35L80

1 Introduction

A curved nonlinear wavefront or a shock front during its evolution develops certain curves of discontinuity, across which the normal to the front and the amplitude

K.R. Arun
Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen, Germany, e-mail: arun@igpm.rwth-aachen.de

M. Lukáčová-Medviďová
Institut für Mathematik, Johannes Gutenberg-Universität Mainz, Staudingerweg 9, D-55099 Mainz, Germany, e-mail: lukacova@mathematik.uni-mainz.de

P. Prasad
Department of Mathematics, Indian Institute of Science, Bangalore - 560012, India, e-mail: prasad@math.iisc.ernet.in

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_6, © Springer-Verlag Berlin Heidelberg 2011

distribution on it are discontinuous. Some of these curves of discontinuity are called kinks, which are shocks in a corresponding ray coordinate system in which a physically realistic system of conservation laws has been formulated. The conservation form of the system of evolution equations of a surface is called kinematical conservation laws (KCL). The KCL is a pure geometrical result and it does not take into consideration any dynamics of the propagating front. This makes the KCL an incomplete system and additional closure equations derived by considering the dynamical conditions of the propagating front are required for applications. Prasad and collaborators have used the KCL in two dimensions along with some closure equations derived on physical considerations to solve several interesting problems, see the review paper [6] and the references therein. The KCL for a surface evolving in three space dimensions, called 3-D KCL, is a system of six conservation laws with three divergence-free type stationary constraints, all three together are termed as 'geometric solenoidal constraint', see [3]. The analysis of the 3-D KCL system, with the closure equation from a weakly nonlinear ray theory (WNLRT), was done in [3] and it has been shown that the resulting system of conservation laws, the so-called conservation laws of 3-D WNLRT give rise to a weakly hyperbolic system; in the sense that the system has zero as a repeated eigenvalue with multiplicity five, but the associated eigenspace is only four-dimensional.

Despite the 3-D WNLRT being a weakly hyperbolic system, in [1, 2] we have been able to develop efficient numerical approximations for it using simple, but robust central schemes. It is well known that the solution to the Cauchy problem for a weakly hyperbolic system (with deficiency in dimension of the eigenspace by one) typically contains a mode, the so-called 'Jordan mode', which grows linearly in time. However, it has been proved in [1] that when the geometric solenoidal constraint is satisfied initially, the solution to the Cauchy problem for linearised 3-D WNLRT at any time does not exhibit the Jordan mode. Motivated by this, a constrained transport technique has been employed to enforce the geometric solenoidal constraint in the numerical solution of 3-D WNLRT, see [1] for more details.

The aim of the present paper is to give a brief overview of the recent results obtained with 3-D WNLRT and to show its efficacy to model propagating wavefronts. The layout of the paper is as follows. In Sect. 2 we introduce the governing equations of 3-D WNLRT. The numerical approximation and the constrained transport strategy are outlined in Sect.3. In Sect.4 we present the results of a numerical experiment, showing the efficiency and robustness of the present method. Finally, we close this article with some concluding remarks in Sect.5.

2 Governing equations

Consider a one parameter family of surfaces in (x_1, x_2, x_3)-space such that it represents the successive positions of a moving surface Ω_t as time varies. Associated with the family, we have a ray velocity $\boldsymbol{\chi}$ at any point (x_1, x_2, x_3) on the surface Ω_t. We consider only the isotropic evolution of Ω_t so that we take $\boldsymbol{\chi}$ to be in the direction of the unit normal $\mathbf{n}$ to Ω_t, i.e. $\boldsymbol{\chi} = m\mathbf{n}$, where m is the normal velocity of propagation of Ω_t. Hence, the evolution of Ω_t is governed by

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = m\mathbf{n}. \tag{1}$$

We introduce a ray coordinate system (ξ_1, ξ_2, t) such that for $t =$ const, we get (ξ_1, ξ_2) as the surface coordinates on Ω_t. Further, $\xi_1 =$ const, $\xi_2 =$ const represent the rays, a two parameter family of curves orthogonal to Ω_t. Let $\mathbf{u}$ and $\mathbf{v}$ be respectively unit tangent vectors to the curves $\xi_2 =$ const and $\xi_1 =$ const on Ω_t. Let $\mathbf{n}$ be a unit normal to Ω_t given by

$$\mathbf{n} = \frac{\mathbf{u} \times \mathbf{v}}{\|\mathbf{u} \times \mathbf{v}\|} \tag{2}$$

so that $(\mathbf{u}, \mathbf{v}, \mathbf{n})$ forms a right handed system. Let an element of distance along a curve $(\xi_2 =$ const$, t =$ const$)$ be $g_1 \mathrm{d}\xi_1$. Analogously, denote by $g_2 \mathrm{d}\xi_2$, the element of distance along a curve $(\xi_1 =$ const$, t =$ const$)$. The element of distance along a ray $(\xi_1 =$ const$, \xi_2 =$ const$)$ is $m\mathrm{d}t$. Based on geometrical considerations we can derive the 3-D KCL [3],

$$(g_1\mathbf{u})_t - (m\mathbf{n})_{\xi_1} = 0, \tag{3}$$

$$(g_2\mathbf{v})_t - (m\mathbf{n})_{\xi_2} = 0 \tag{4}$$

subject to the condition

$$(g_1\mathbf{u})_{\xi_2} - (g_2\mathbf{v})_{\xi_1} = 0. \tag{5}$$

Note that the constraint (5) is an involution, i.e. if it is satisfied at time $t = 0$, then the equations (3)-(4) imply that it is satisfied for every time. Note that each of the scalar equations in (5) can be written as $\mathrm{div}(\mathfrak{B}_k) = 0$, where $\mathfrak{B}_k := (-g_2 v_k, g_1 u_k), k = 1, 2, 3$. Therefore, the vector constraint (5) has been designated as geometric solenoidal constraint. The 3-D KCL (3)-(4), being a system of six evolution equations in seven unknowns $u_1, u_2, v_1, v_2, m, g_1$ and g_2, is underdetermined. We use the closure equation by considering the energy propagation along the rays of a WNLRT, c.f. [6]. The energy transport equation of WNLRT for a polytropic gas initially at rest and in uniform state can be written in a conservation form [3]

$$\left((m-1)^2 e^{2(m-1)} g_1 g_2 \sin\chi\right)_t = 0, \tag{6}$$

where χ is the angle between the vectors $\mathbf{u}$ and $\mathbf{v}$. The system of equations (3)-(4) and (6), hereafter designated as the conservation laws of 3-D WNLRT, is the complete set of equations describing the evolution of the nonlinear wavefront Ω_t.

Remark 1. It has been proved in [3] that the eigenvalues of 3-D WNLRT are $\lambda_1, \lambda_2 (= -\lambda_1), \lambda_3 = \cdots = \lambda_7 = 0$, where λ_1 is given by

$$\lambda_1 = \left\{ \frac{m-1}{2\sin^2\chi} \left(\frac{e_1^2}{g_1^2} - \frac{2e_1e_2}{g_1g_2}\cos\chi + \frac{e_2^2}{g_2^2} \right) \right\}^{1/2}. \tag{7}$$

Here, $(e_1, e_2) \in \mathbb{R}^2$ with $e_1^2 + e_2^2 = 1$. Further, there are only four independent eigenvectors for the eigenvalue zero. Note that λ_1 is real for $m > 1$ and purely imaginary for $m < 1$. Hence, the 3-D WNLRT forms a weakly hyperbolic system when $m > 1$. In this article we consider only the case when $m > 1$.

3 Numerical approximation

In this section we present a numerical approximation of the conservation laws of 3-D WNLRT to study evolution of a weakly nonlinear wavefront Ω_t and formation and propagation of kink curves on it. Note that the system of conservation laws of 3-D WNLRT can be recast in the usual divergence form

$$W_t + F_1(W)_{\xi_1} + F_2(W)_{\xi_2} = 0, \tag{8}$$

where the vector of conserved variables W and the flux-vectors $F_1(W)$ and $F_2(W)$ in the ξ_1- and ξ_2-directions respectively, are given by

$$\begin{aligned} W &= \left(g_1\mathbf{u}, g_2\mathbf{v}, (m-1)^2 e^{2(m-1)} g_1 g_2 \sin\chi\right)^T, \\ F_1(W) &= (m\mathbf{n}, \mathbf{0}, 0)^T, \ F_2(W) = (\mathbf{0}, m\mathbf{n}, 0)^T. \end{aligned} \tag{9}$$

In what follows we briefly summarise the central finite volume scheme for (8), first employed in [1].

1. The cell integral averages $\overline{W}_{i,j}$ of the conservative variable W are used in the discretisation of the system of conservation laws (8).
2. A second order TVD Runge-Kutta method [8] is used for time integration. The time-step is chosen to be inversely proportional to the maximum of the nonzero eigenvalue λ_1, c.f. (7), taken over the entire computational domain.
3. A nonlinear iterative solver is employed to recover the values of $\mathbf{u}, \mathbf{v}, g_1, g_2$ and m from the computed values of W.
4. A second order MUSCL reconstruction with a central weighted essentially non-oscillatory (CWENO) limiter [4] is used to reconstruct the variables at the cell interfaces.
5. The Kurganov-Tadmor high resolution flux [5] is used as the numerical flux at a cell interface, for example at a right hand vertical edge

$$\mathscr{F}_{i+\frac{1}{2},j}\left(W_{i,j}^R, W_{i+1,j}^L\right) = \frac{1}{2}\left(F_1\left(W_{i+1,j}^L\right) + F_1\left(W_{i,j}^R\right)\right) - \frac{a_{i+\frac{1}{2},j}}{2}\left(W_{i+1,j}^L - W_{i,j}^R\right), \tag{10}$$

where $W_{i,j}^{L(R)}$ denote respectively the left and right interpolated states. Here, $a_{i+1/2,j}$ is the maximal wave-speed, which can be computed with the help of the maximum of eigenvalues, c.f. [5]. The numerical flux at a horizontal edge can be computed in an analogous manner.

6. In order that the numerical solution satisfy a discrete version of the geometric solenoidal constraint (5), we use a constrained transport algorithm [7]. We employ three potentials $\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3$, corresponding to the three components of the vectors $g_1\mathbf{u}$ and $g_2\mathbf{v}$. Note that the geometric solenoidal constraint (5) implies the conditions
$$g_1 u_k = \mathbb{A}_{k\xi_1},\ g_2 v_k = \mathbb{A}_{k\xi_2},\ k = 1,2,3. \tag{11}$$
The use of (11) in the 3-D KCL system (3)-(5) immediately yields the evolution equations
$$\mathbb{A}_{kt} - m n_k = 0. \tag{12}$$
We numerically solve (12) to get the updated values of the potentials $\mathbb{A}_k$. The resulting values of $\mathbb{A}_k$ are used to suitably discretise (11) to yield the corrected values of $g_1\mathbf{u}$ and $g_2\mathbf{v}$. It is these updated values which satisfy a discrete version of (5), see [1] for more details.

At any time t, we approximate the wavefront Ω_t by a discrete set of points $\mathbf{x}_{i,j}(t) := \mathbf{x}(\xi_{1_i}, \xi_{2_j}, t)$. To get the successive positions of Ω_t, we numerically solve the system of ODEs (1) in the discretised form $\mathrm{d}\mathbf{x}_{i,j}(t)/\mathrm{d}t = m_{i,j}(t)\mathbf{n}_{i,j}(t)$ where $m_{i,j}(t)$ and $\mathbf{n}_{i,j}(t)$ are the corresponding values of m and $\mathbf{n}$ obtained from $\overline{W}_{i,j}(t)$.

In order to start the algorithm, the conserved variable W has to be initialised at each mesh point. Here, some care has to be taken, so that (11) is satisfied by the initial values. Let us assume that the initial wavefront Ω_0 is given a parametric form $\mathbf{x} = \mathbf{x}_0(\xi_1, \xi_2)$, with some appropriate choice of surface coordinates ξ_1 and ξ_2. The initial values for $g_1\mathbf{u}$ and $g_2\mathbf{v}$ and the potentials $\mathbb{A}_1, \mathbb{A}_2, \mathbb{A}_3$ can be chosen to be

$$g_1\mathbf{u}(\xi_1, \xi_2, 0) = \mathbf{x}_{0\xi_1}(\xi_1, \xi_2),\ g_2\mathbf{v}(\xi_1, \xi_2, 0) = \mathbf{x}_{0\xi_2}(\xi_1, \xi_2), \tag{13}$$
$$\mathbb{A}_k(\xi_1, \xi_2, 0) = x_k(\xi_1, \xi_2),\ k = 1,2,3. \tag{14}$$

Note that (5) and (11) are satisfied by the above choice of initial values. In the numerical test problem considered here, the normal velocity m on Ω_0 has been assigned a constant value $m_0 = 1.2$. For more details of the numerical scheme and its implementation, we refer the reader to [1].

4 Numerical test problem

We choose initial wavefront Ω_0 in such a way that it is not axisymmetric. The front Ω_0 has a single smooth dip. The initial shape of the wavefront is given by

$$\Omega_0\colon x_3 = \frac{-\kappa}{1+\frac{x_1^2}{\alpha^2}+\frac{x_2^2}{\beta^2}}, \tag{15}$$

where the parameter values are set to be $\kappa = 1/2, \alpha = 3/2, \beta = 3$. The ray coordinates (ξ_1, ξ_2) are chosen initially as $\xi_1 = x_1$ and $\xi_2 = x_2$. The computational domain $[-20, 20]\times[-20, 20]$ is divided into 401×401 mesh points. The simulations are done up to $t = 2.0, 6.0, 10.0$. We have set non-reflecting boundary conditions for all the variables.

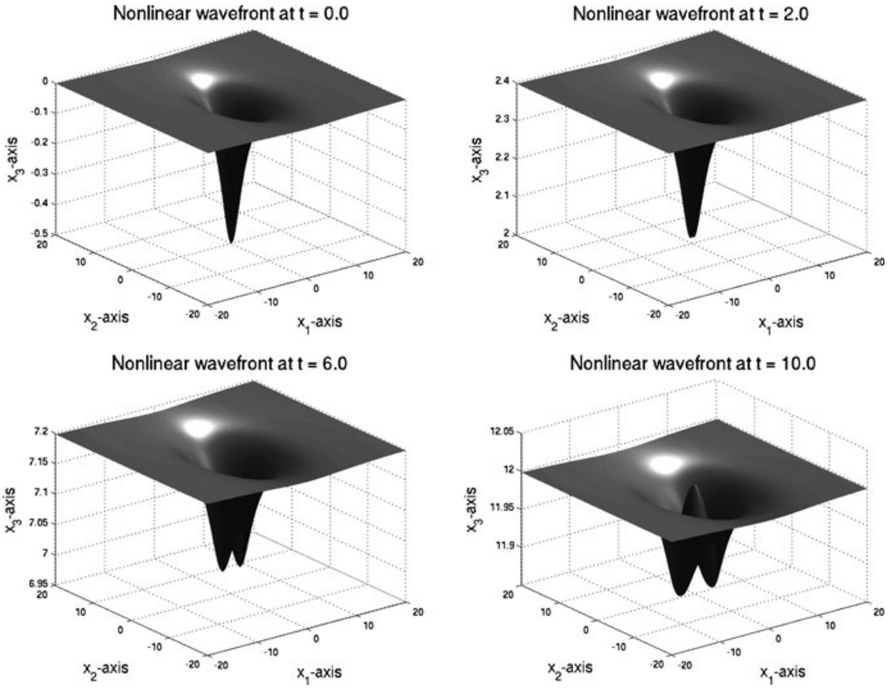


Fig. 1 The successive positions of the nonlinear wavefront Ω_t with an initial smooth dip which is not axisymmetric

In Fig. 1 we plot the initial wavefront Ω_0 and the successive positions of the wavefront Ω_t at times $t = 2.0, 6.0, 10.0$. It can be seen that the wavefront has moved up in the x_3-direction and the dip has spread over a larger area in x_1- and x_2-directions. The lower part of the front moves up leading to a change in shape of the initial front Ω_0. It is very interesting to note that two dips appear in the central part of the wavefront, which are clearly visible at $t = 6.0$ and $t = 10.0$. These two dips are separated by an elevation almost like a wall parallel to the x_2-axis.

To explain the results of convergence of the rays we give in Fig. 2 the slices of the wavefront in $x_2 = 0$ section and $x_1 = 0$ section from time $t = 0.0$ to $t = 10.0$. Due to the particular choice of the parameters α and β in the initial data (15), the section of the front Ω_0 in $x_2 = 0$ plane has a smaller radius of curvature than that of the section in $x_1 = 0$ plane. This results in a stronger convergence of the rays in $x_2 = 0$ plane compared to those in $x_1 = 0$ plane as evident from Fig. 2. In the diagram on the top in Fig. 2, we clearly note a pair of kinks at times $t = 3.0$ onwards in the $x_2 = 0$ section. However, there are no kinks in the bottom diagram in Fig. 2 in $x_1 = 0$ section.

We give now the plots of the normal velocity m in (ξ_1, ξ_2) plane along ξ_1- and ξ_2-directions in Fig. 3. It is observed that m has two shocks in the ξ_1-direction which correspond to the two kinks in the x_1-direction. We have also plotted the numerical values of the divergence of $\mathfrak{B}_1$ at time $t = 10.0$ in Fig. 4. It is evident that the

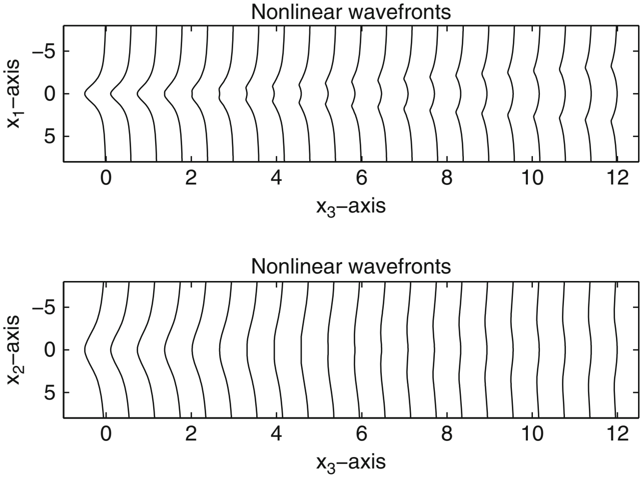


Fig. 2 The sections of the nonlinear wavefront at times $t = 0.0, \ldots, 10.0$ with a time step 0.5. On the top: in $x_2 = 0$ plane. Bottom: in $x_1 = 0$ plane

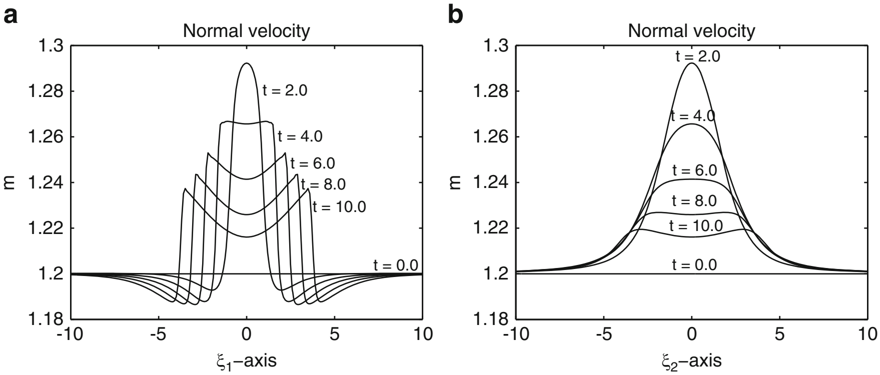


Fig. 3 The time evolution of the normal velocity m. (a): along ξ_1-direction in the section $\xi_2 = 0$. (b): along ξ_2-direction in the section $\xi_1 = 0$

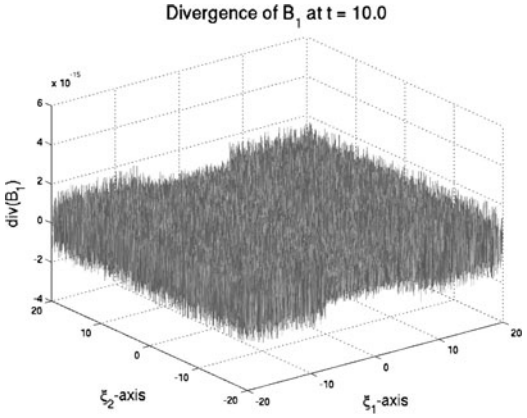


Fig. 4 The divergence of $\mathfrak{B}_1$ at $t = 10.0$. The error is of the order of 10^{-15}

geometric solenoidal condition is satisfied with an error of 10^{-15}. The divergences of $\mathfrak{B}_2$ and $\mathfrak{B}_3$ also show the same trend.

5 Concluding remarks

An efficient central finite volume scheme for the weakly hyperbolic system of conservation laws of 3-D WNLRT has been described and tested. Reconstruction is achieved component-wise and a simple central flux is employed in the numerical flux evaluation. Based on our numerical experiment and the ones reported in [1], it can be concluded that the solenoidal condition is preserved up to machine accuracy if the present finite volume scheme with a constrained transport technique is used. Moreover, none of the solution components exhibits any linearly growing Jordan mode.

Acknowledgements K. R. A. wishes to thank the Alexander von Humboldt Foundation for a postdoctoral fellowship. P. P. is supported by the Department of Atomic Energy, Government of India, under Raja-Ramanna Fellowship Scheme.

References

1. Arun, K.R.: A numerical scheme for three-dimensional front propagation and control of Jordan mode. Tech. rep., Department of Mathematics, Indian Institute of Science, Bangalore (2010)
2. Arun, K.R., Lukáčová-Medviďová, M., Prasad, P., Raghurama Rao, S.V.: An application of 3-D kinematical conservation laws: propagation of a three dimensional wavefront. SIAM. J. Appl. Math. **70**, 2604–2626 (2010)

3. Arun, K.R., Prasad, P.: 3-D kinematical conservation laws (KCL): evolution of a surface in $\mathbb{R}^3$-in particular propagation of a nonlinear wavefront. Wave Motion **46**, 293–311 (2009)
4. Jiang, G.S., Shu, C.W.: Efficient implementation of weighted ENO schemes. J. Comput. Phys. **126**, 202–228 (1996)
5. Kurganov, A., Tadmor, E.: New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations. J. Comput. Phys. **160**, 241–282 (2000)
6. Prasad, P.: Ray theories for hyperbolic waves, kinematical conservation laws (KCL) and applications. Indian J. Pure Appl. Math. **38**, 467–490 (2007)
7. Ryu, D., Miniati, F., Jones, T.W., Frank, A.: A divergence-free upwind code for multidimensional magnetohydrodynamic flow. Astrophys. J. **509**, 244–255 (1998)
8. Shu, C.W.: Total-variation-diminishing time discretizations. SIAM J. Sci. Stat. Comput. **9**, 1073–1084 (1988)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Preservation of the Discrete Geostrophic Equilibrium in Shallow Water Flows

E. Audusse, R. Klein, D.D. Nguyen, and S. Vater

Abstract We are interested in the numerical simulation of large scale phenomena in geophysical flows. In these cases, Coriolis forces play an important role and the circulations are often perturbations of the so-called geostrophic equilibrium. Hence, it is essential to design a numerical strategy that preserves a discrete version of this equilibrium. In this article we work on the shallow water equations in a finite volume framework and we propose a first step in this direction by introducing an auxiliary pressure that is in geostrophic equilibrium with the velocity field and that is computed thanks to the solution of an elliptic problem. Then the complete solution is obtained by working on the deviating part of the pressure. Some numerical examples illustrate the improvement through comparisons with classical discretizations.

Keywords Geostrophic Adjustment, Shallow Water Flows, Finite Volume Method, Well-balanced Scheme
MSC2010: 65M08, 76U05, 86A05

1 Introduction

We are interested in the numerical simulation of large scale phenomena in geophysical flows. At these scales, Coriolis forces play an important role and the atmospheric or oceanic circulations are frequently observed near geostrophic equilibrium situations, see for example [11, 12]. For this reason it is essential to design a numerical strategy that preserves a discrete version of this geostrophic

E. Audusse and D.D. Nguyen
LAGA, Université Paris Nord, 99 av. J.B. Clement, 93430 Villetaneuse, France,
e-mail: audusse@math.univ-paris13.fr, name2@email.adress

R. Klein and S. Vater
Institut für Mathematik, Freie Universität Berlin, Arnimallee 6, D-14195 Berlin, Germany,
e-mail: rupert.klein@math.fu-berlin.de, stefan.vater@math.fu-berlin.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_7, © Springer-Verlag Berlin Heidelberg 2011

equilibrium: if numerical spurious waves are created, they quickly become higher than the physical ones we want to capture. This phenomenon is well known but its solution in the context of finite volume methods is still an open problem. We address this question in this article.

One of the most popular systems that is used to model such quasi-geostrophic flows are the shallow water equations with β-plane approximation

$$h_t + \nabla \cdot (h\overline{u}) = 0, \tag{1}$$

$$(h\overline{u})_t + \nabla \cdot (h\overline{u} \otimes \overline{u}) + \nabla(\frac{gh^2}{2}) = -f\mathbf{e}_z \times (h\overline{u}), \tag{2}$$

The shallow water system is the simplest form of equations of motion that can be used to model Rossby and Kelvin waves in the atmosphere or ocean, and the use of the β-plane approximation allows the model to take into account a non-constant Coriolis parameter f that varies linearly with the latitude without considering a spherical domain. We choose to work in a finite volume framework to discretize the equations because of its ability to deal with complex geometries and its inherent conservation property, see [4, 9]. In this context, the discrete preservation of the geostrophic equilibrium, which is mainly the balance between pressure gradient and Coriolis forces in (2), is a hard touch: the main reason is that the fluxes are upwinded for stability reasons while the source terms are usually discretized in a centered way.

The question of the preservation of non-trivial equilibria in geophysical fluid models has received great attention in the area of numerical modeling in the last decade. Many studies were devoted to the preservation of the so-called hydrostatic and also lake-at-rest equilibria, see [2–4] and references therein. More recently some authors investigated the problem of the geostrophic equilibrium [5, 6, 8, 10]. However, this question is more delicate for two reasons: it is an essentially 2d problem, and it involves a non-zero velocity field. It follows that its solution is still incomplete. In this work we propose a solution to this problem by introducing an auxiliary water depth which is in geostrophic balance with the velocity field and then by working on the deviation between the actual and auxiliary water depths instead of considering the water depth itself. The auxiliary water depth is computed through the solution of a Poisson problem on a dual grid [13].

2 Position of the problem

In this short note we present the method by considering a constant Coriolis parameter. In order to exhibit the importance of the geostrophic equilibrium, we introduce the non-dimensional version of the shallow water equations (1)–(2) written in non-conservative form

$$\begin{aligned} &h_t + \nabla \cdot (h\overline{u}) = 0, \\ &\overline{u}_t + \overline{u} \cdot \nabla\overline{u} + \frac{1}{\mathbf{Fr}^2}\nabla h + \frac{1}{\mathbf{Ro}}2\mathbf{e}_z \times \overline{u} = 0. \end{aligned}$$

Here, h and $\overline{u}$ are the unknown dimensionless depth and velocity fields and

$$\mathbf{Fr} = \frac{\overline{U}}{\sqrt{gH}}, \qquad \mathbf{Ro} = \frac{\overline{U}}{\Omega L}$$

are the Froude and Rossby numbers, respectively, with $\overline{U}$, L and H some characteristic velocity, length and depth for the flow, g the gravity coefficient and Ω the angular velocity of the earth. For large scale phenomena typical values for these numbers are

$$\mathbf{Fr} \approx \mathbf{Ro} \approx \epsilon = 10^{-2},$$

We then expand the unknowns in term of ε

$$h = h_0 + \varepsilon h_1 + \varepsilon^2 h_2 + \ldots, \qquad \overline{u} = \overline{u}_0 + \varepsilon \overline{u}_1 + \varepsilon^2 \overline{u}_2 + \ldots$$

and we keep the leading order terms to exhibit the following stationary state

$$O\left(\epsilon^{-2}\right) : \nabla h_0 = 0 \tag{3}$$

$$O\left(\epsilon^{-1}\right) : \nabla h_1 + 2e_z \times \overline{u}_0 = 0 \tag{4}$$

$$O\left(\epsilon^{0}\right) : \nabla \cdot \overline{u}_0 = 0, \tag{5}$$

This set of equations is called the geostrophic equilibrium. It follows from equation (3) that the water depth is constant at the leading order and from equation (5) that the main part of the velocity field is divergence free. Equation (4) is nothing but the fact that the pressure gradient and the Coriolis term are in balance for leading varying terms h_1 and $\overline{u}_0$. Let's now turn to the numerical point of view. Preservation of the discrete equilibrium (3) is obvious. The divergence free condition (5) is much more delicate to deal with but it has been widely investigated for Stokes or Navier-Stokes equations, mostly in the framework of finite element methods. It is also the subject of a recent work [13], where the authors study the zero Froude number limit of the shallow water equations. In this note we focus on a proper way to preserve the balance in equation (4).

3 The well-balanced finite volume scheme

We choose to discretize the shallow water equations (1)–(2) in a finite volume framework [4, 9]. The reason to consider this particular method is related to its inherent conservation properties that are interesting for geophysical applications and in particular for long time simulations [1]. A second reason is that the finite volume method is also able to deal with sharp fronts that can occur in geophysical applications. We first recall the formulation of the finite volume method and the classical centered discretization of the Coriolis source term. Then, we derive the new well-balanced scheme by introducing an auxiliary pressure that is computed through the solution of a Laplace equation on a dual grid.

System (1)–(2) is a particular case of a 2d conservation law with source term:

$$U_t + (F(U))_x + (G(U))_y = S(U), \tag{6}$$

in which $U = (h, hu, hv)^T$ and

$$F(U) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \end{pmatrix}, \; G(U) = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \end{pmatrix}, \; S(U) = \begin{pmatrix} 0 \\ 2\Omega hv \\ -2\Omega hu \end{pmatrix}.$$

In this note we only consider Cartesian grids. Then, the finite volume discretization of equation (6) leads to the computation of approximated solutions $U_{i,j}^n$ through the discrete formula

$$U_{i,j}^{n+1} = U_{i,j}^n - \frac{\delta t}{\delta x}\left(\mathbf{F}_{i+\frac{1}{2},j}^n - \mathbf{F}_{i-\frac{1}{2},j}^n\right) - \frac{\delta t}{\delta y}\left(\mathbf{G}_{i,j+\frac{1}{2}}^n - \mathbf{G}_{i,j-\frac{1}{2}}^n\right) + \delta t\, S_{i.j}^n,$$

where $\mathbf{F}_{i+\frac{1}{2},j}^n$ is a discrete approximation of the flux $F(U)$ along the interface between cells $C_{i,j}$ and $C_{i+1,j}$ that is constructed through a three points formula

$$\mathbf{F}_{i+\frac{1}{2},j}^n = \mathscr{F}\left((h_{i,j}^n, u_{i,j}^n, v_{i,j}^n), (h_{i+1,j}^n, u_{i+1,j}^n, v_{i+1,j}^n)\right). \tag{7}$$

Here we use the HLL solver [7] to compute these approximations.

The classical discretization of the source term $S_{i.j}^n$ is computed through the centered formula

$$S_{i.j}^n = \begin{pmatrix} 0 \\ -2\Omega_z \times (h_{i,j}^n \overline{u}_{i,j}^n) \end{pmatrix}$$

where $h_{i,j}^n$ denotes the approximated value at time t^n on cell $C_{i,j}$. We will exhibit in the last section that this approach suffers from important drawbacks when we consider applications for small Froude and Rossby numbers.

The main idea of our method to overcome this problem is to introduce an auxiliary water depth h_c that is in balance with Coriolis forces related to the actual velocity field. This idea is an extension of the notion of hydrostatic reconstruction that was introduced in [3] for the Euler equations and in [2] for shallow water flows. Here, h_c will satisfy the equation

$$g\nabla h_c = -2\Omega \times \overline{u}. \tag{8}$$

In our approach, h_c is discretized as a grid function, which is piecewise bilinear on each grid cell and continuous at the interfaces. The second ingredient of the well-balanced scheme is the representation of the Coriolis forces by the gradient of this quantity. Furthermore, the fluxes in the conservative part of the scheme are modified in the following way: For each cell, we introduce a deviation in the water depth by

$$\Delta h^n_{i,j} = h^n_{i,j} - h^n_c(x_i, y_j)$$

Then, the interface water depths are computed by

$$\widehat{h}^{n,k,x}_{i+\frac{1}{2},j} = \frac{1}{2}\left[h^n_c\left(x_{i+\frac{1}{2},j+\frac{1}{2}}\right) + h^n_c\left(x_{i+\frac{1}{2},j-\frac{1}{2}}\right)\right] + \Delta h^n_{k,j}, \quad \text{for} \quad k = i, i+1.$$

and the original three points formula (7) for the flux is replaced by

$$\mathbf{F}^n_{i+\frac{1}{2},j} = \mathscr{F}\left((\widehat{h}^{n,i,x}_{i+\frac{1}{2},j}, u^n_{i,j}, v^n_{i,j}), (\widehat{h}^{n,i+1,x}_{i+\frac{1}{2},j}, u^n_{i+1,j}, v^n_{i+1,j})\right),$$

If the flow satisfies the geostrophic equilibrium, $\hat{h}$ and h_c are equal. The consistency of the flux will then provide some numerical balance between the conservative part and the source term that will directly impact the results. Note also that the time step is now related to the interface water depths. Nevertheless, in the numerical applications, the numerical values remain very close for both methods.

It remains to explain the computation of the auxiliary water depth h_c that is the solution of a discrete equivalent of equation (8). We first take the divergence of this equation and then search for the solution of a Poisson equation

$$-\Delta\phi = \nabla \cdot \left(\overline{k} \times \overline{u}\right).$$

Integration of this equation on the dual cell $C_{i+\frac{1}{2},j+\frac{1}{2}}$ and application of the Gauss theorem leads to

$$\int_{\partial C_{i+\frac{1}{2},j+\frac{1}{2}}} \nabla\phi \cdot \overline{n}\,\mathrm{d}\sigma = -\int_{\partial C_{i+\frac{1}{2},j+\frac{1}{2}}} \overline{k}_z \overline{u} \cdot \overline{t}\,\mathrm{d}\sigma,$$

where $\overline{n}$ (resp. $\overline{t}$) is a normal (resp. tangential) vector to the interface of the dual cell. We solve this equation by using the technique presented in [13] for the solution of a similar problem. We refer the reader to this article for the details of the method that is in particular proved to provide an inf-sup-stable projection. We finally obtain a linear system with a nine point stencil. The boundary conditions for this auxiliary problem are prescribed by using the fact that the computed pressure (or height) field is equivalent to a stream function for the associated balanced geostrophic flow. For example a rigid wall type boundary condition for the flow translates into a Dirichlet type boundary condition for the stream function. Similar types of equivalences can be used to prescribe other types of boundary conidtions.

4 Numerical results

In order to test the new scheme, we consider a stationary vortex in the square domain $[0, 1] \times [0, 1]$. We consider periodic boundary conditions, and as initial conditions we choose a velocity field of the form

$$\overline{u}_0(r, \theta) = v_\theta(r)\overline{e}_\theta, \quad v_\theta(r) = \varepsilon \left[5r \;\chi\left(r < \frac{1}{5}\right) + (2 - 5r)\,\chi\left(\frac{1}{5} \le r < \frac{2}{5}\right)\right],$$

where r is the distance to the center of the domain and χ denotes the characteristic function of a given interval. Some computations show that the vortex is a stationary solution of the shallow water equations (1)-(2), if the initial water depth $h_0(r)$ is a radial solution of the ODE

$$h_0'(r) = \frac{1}{g}\left(2\Omega v_\theta + \frac{v_\theta^2}{r}\right).$$

Note that if we choose a water depth and an angular velocity of order $O(1)$, the Froude and Rossby numbers are of order $O(\epsilon)$. It follows that our interest is for small values of the parameter ϵ.

We first work on a regular grid with 30×30 cells and we consider four Froude resp. Rossby numbers: $0.05, 0.1, 0.5$ and 1. The numerical solution is computed by using both schemes described in the previous section. In order to compare the accuracy of the schemes, we compute the relative L^2 error in the water depth. In Fig. 1 we present the time evolution of this error for the four values of ϵ. It appears that for both schemes the error is increasing with time before reaching a stationary value. More interesting is that for the classical (resp. well-balanced) scheme the error is increasing (resp. decreasing), when the Froude number is decreasing. While the error is of the same order for both schemes when $\epsilon = 1$, for other values of ϵ the well-balanced scheme is always more precise than the classical one.

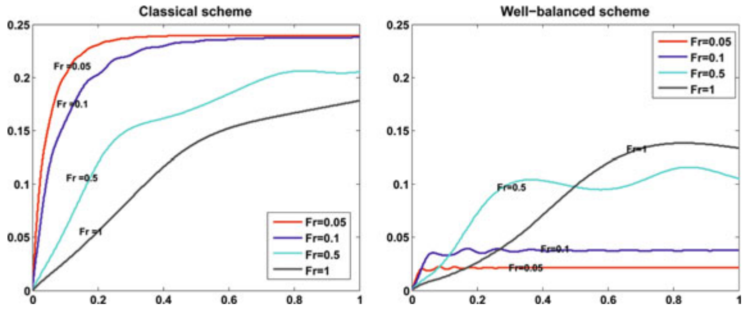


Fig. 1 Error in time for both classical and well-balanced schemes (30×30 grid cells) and for four different Froude numbers

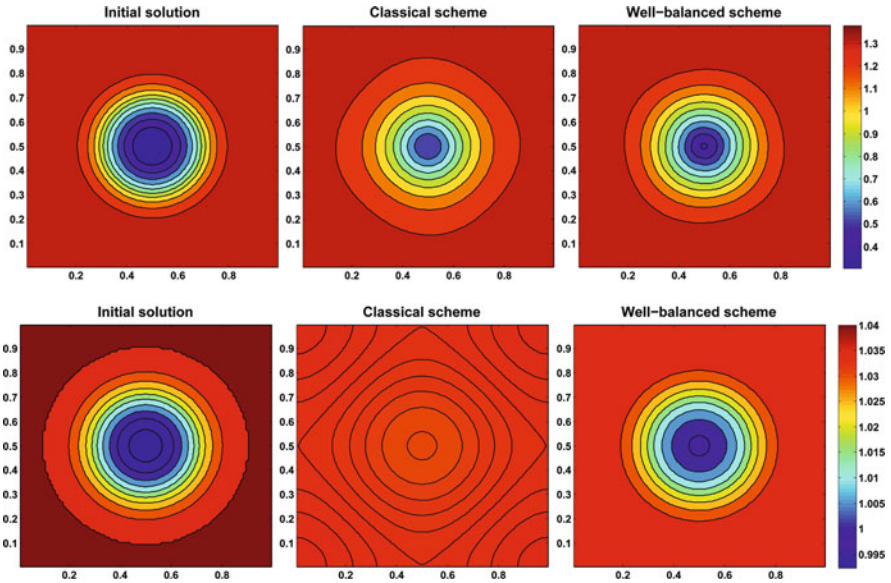


Fig. 2 Contour of the computed fluid depth with 100 × 100 grid cells. **Fr** = 0.95 (Top), **Fr** = 0.1 (bottom)

We then consider a finer grid with 100 × 100 cells and we present the water depth for both schemes and for two values of ϵ: 0.95 (large) and 0.1 (small). In Fig. 2 we present the 2d contour of the water depth. The results look similar and quite close to the initial solution when ϵ is large (top row). But when ϵ is small (bottom row), the classical scheme totally fails to compute the right solution, whereas the water depth computed by the well-balanced scheme stays close to the initial one. In Fig. 3 we give more quantitative results by presenting a cut of the solution along x–axis at $y = 0.5$. These pictures clearly exhibit that the results are very close when ϵ is large, but very different when ϵ is small. In this last case the classical scheme is not able to maintain the vortex, whereas the well-balanced scheme preserves the shape of the free surface. Note that the small diffusion that is observed even for the well-balanced scheme is due to the fact that we consider only first order schemes in this work. We end this short note by some words on the CPU time. We first notice that for the last numerical test case, the time steps are very close for both methods, as it is reported in the table below. We then consider the CPU time for both methods and conclude that it is four times larger for the well-balanced scheme. It is obviously due to the solution of the linear system related to the elliptic problem at each time step. This observation leads to two comments. First, and since the solution of the linear system is only required for the computation of the auxiliary water depth, it is possible to obtain a compromise between accuracy and efficiency of the whole process by considering iterative methods with a small number of iterations. Second we recall that our final objective is to couple the presented process with a numerical

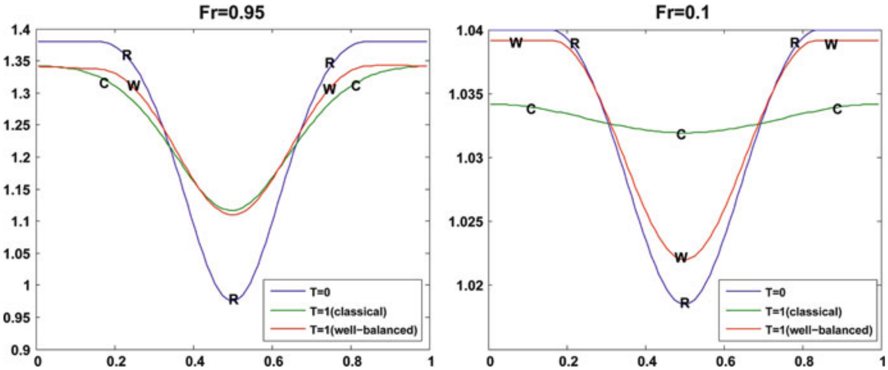


Fig. 3 Fluid depth profiles – cut along the x–axis at $y = 0.5$ with 100×100 grid cells. R line: Initial solution, W line: Well-balanced scheme, C line: Classical scheme

scheme adapted for small Froude number flows and then to generalize the method presented in [13] to rotating flows. Since the technique introduced in [13] already requires for the solution of a related linear system, the additional computational cost of the well-balanced process presented here is very small.

	Classical scheme	Well-balanced scheme
Time Step	9.7702e-005	9.7464e-005
CPU Time	1547 s	5564 s

References

1. E. Audusse, R.Klein, A. Owinoh, Conservative discretization of Coriolis force in a finite volume framework, Journal of Computational Physics, **228**, 2934-2950 (2009).
2. E. Audusse, F. Bouchut, M.O. Bristeau, R. Klein, B. Perthame, A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows, SIAM J. Sci. Comput., **25**, 2050–2065 (2004).
3. N. Botta, R. Klein, S. Langenberg, S. Lützenkirchen, Well-balanced finite volume methods for nearly hydrostatic flows, JCP, **196**, 539–565 (2004).
4. F. Bouchut , Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources, Birkaüser (2004).
5. F. Bouchut, J. Le Sommer, V. Zeitlin, Frontal geostrophic adjustment and nonlinear wave phenomena in one dimensional rotating shallow water. Part 2: high-resolution numerical simulations, J. Fluid Mech., **513**, 35–63 (2004).
6. M.J. Castro, J.A. Lopez, C. Pares, Finite Volume Simulation of the Geostrophic Adjustment in a Rotating Shallow-Water System SIAM J. on Scientific Computing, **31**, 444–477 (2008).
7. A. Harten, P. Lax, B. Van Leer, On upstream differencing and Godunov type schemes for hyperbolic conservation laws, SIAM Review, **25**, 235–261 (1983).

8. A. Kuo, L. Polvani, Time-dependent fully nonlinear geostrophic adjustment, J. Phys. Oceanogr. **27**, 1614–1634 (1997).
9. R.J. LeVeque, Finite Volume Methods for Hyperbolic Problems, Camb. Univ. Press (2002).
10. N. Panktratz, J.R. Natvig, B. Gjevik, S. Noelle, High-order well-balanced finite-volume schemes for barotropic flows. Development and numerical comparisons, Ocean Modelling, **18**, 53–79 (2007).
11. J. Pedlosky, Geophysical Fluid Dynamics, Springer, 2nd edition (1990).
12. G. K. Vallis, Atmospheric and Oceanic Fluid Dynamics: Fundamentals and Large-scale Circulation, Cambridge University Press (2006).
13. S. Vater, R. Klein, Stability of a Cartesian Grid Projection Method for Zero Froude Number Shallow Water Flows, Numerische Mathematik, **113**, 123–161 (2009).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Arbitrary order nodal mimetic discretizations of elliptic problems on polygonal meshes

Lourenço Beirão da Veiga, Konstantin Lipnikov, and Gianmarco Manzini

Abstract We develop and analyze a new family of mimetic methods on unstructured polygonal meshes for the diffusion problem in primal form. The new nodal MFD formulation that we propose in this work extends the original low-order formulation of [3] to arbitrary orders of accuracy by requiring that the consistency condition holds for polynomials of arbitrary degree $m \geq 1$. An error estimate is presented in a mesh-dependent norm that mimics the energy norm and numerical experiments confirm the convergence rate that is expected from the theory.

Keywords mimetic finite difference, diffusion problems, unstructured mesh, polygonal element
MSC2010: 65N06

1 The nodal mimetic finite difference method

We consider a nodal mimetic discretization of the steady diffusion problem for the scalar solution field u given by

$$\operatorname{div}(\mathsf{K}\nabla u) = f \quad \text{in } \Omega, \tag{1}$$

$$u = g \quad \text{on } \Gamma, \tag{2}$$

L. Beirão da Veiga
Dipartimento di Matematica, Università di Milano, Italy, e-mail: lourenco.beirao@unimi.it

K. Lipnikov
Los Alamos National Laboratory, New Mexico (US), e-mail: lipnikov@lanl.gov

G. Manzini
IMATI-CNR and CeSNA-IUSS, Pavia, Italy, e-mail: Marco.Manzini@imati.cnr.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_8, © Springer-Verlag Berlin Heidelberg 2011

where the computational domain Ω is a bounded, open, polygonal subset of $\mathbb{R}^2$ with Lipshitz boundary Γ, the diffusion tensor K is a 2×2 bounded, measurable, strongly elliptic and symmetric tensor describing the material properties, $f \in L^2(\Omega)$ is the forcing term and $g \in H^{1/2}(\Gamma)$ is the boundary function that defines the non-homogeneous Dirichlet boundary condition.

Let us consider the set of functions $H^1_g(\Omega) = \{v \in H^1(\Omega),\ v_{|\Gamma} = g\}$. Problem (1)-(2) can be restated in the variational form:

find $u \in H^1_g(\Omega)$ *such that*

$$\int_\Omega \mathsf{K}\nabla u\cdot\nabla v\, dV = \int_\Omega f\, v\, dV \qquad \forall v \in H^1_0(\Omega). \tag{3}$$

Under the previous assumptions problem (3) is well-posed [5]. The existence and uniqueness of the weak solution follows from continuity and coercivity of the bilinear form in (3).

The mimetic discretizations that are available in the literature for this problem have usually low-order accuracy. The nodal MFD method in [3] uses degrees of freedom associated with the mesh vertices and is derived from a consistency condition, i.e., a discrete integration by parts formula, that is exact for linear polynomials. This method was proven to be first-order accurate in a mesh-dependent H^1-seminorm.

To some extend, the mixed and mixed-hybrid MFD scheme in [4] can be interpreted as a generalization of the $RT_0 - \mathbb{P}_0$ mixed finite element (FE) method to polygonal and polyhedral meshes Similarly, the nodal MFD method in [3] can be viewed as a generalization of the linear Galerkin FE method. The essential difference between FE and MFD methods is that the latter does not use shape functions and operates directly with degrees of freedom. This result in a number of useful consequences such as a family of equivalent MFD methods.

The numerical approximation to (3) is performed on a sequence of polygonal partitions $\{\Omega_h\}_h$ of the domain Ω, which are required to satisfy a suitable set of regularity conditions [4]. For any mesh Ω_h, the subscripted label h is the mesh size and is defined by $h = \max_{\mathsf{P}\in\Omega_h} h_{\mathsf{P}}$ where $h_{\mathsf{P}} = \sup_{\mathbf{x},\mathbf{y}\in\mathsf{P}} |\mathbf{x}-\mathbf{y}|$ is the diameter of the polygonal cell $\mathsf{P} \in \Omega_h$. On a mesh Ω_h, we approximate the scalar fields from $H^1(\Omega)$ through a set of suitable *degrees of freedom* $u_h, v_h \in \mathcal{V}_h$, where $\mathcal{V}_h$ denotes the linear space of the discrete scalar fields. Then, we introduce the bilinear form $\mathscr{A}_h\big(\cdot,\cdot\big) : \mathcal{V}_h \times \mathcal{V}_h \to \mathbb{R}$, which approximates the left-hand side of (3), and the bilinear form $\big(\cdot,\cdot\big)_h : L^2(\Omega) \times \mathcal{V}_h \to \mathbb{R}$, which approximates the right-hand side of (3). In the nodal mimetic finite difference formulation, the Dirichlet boundary conditions are *essential* and are incorporated through the subspace $\mathcal{V}_{h,g}$ of $\mathcal{V}_h$. The set of discrete scalar fields $\mathcal{V}_{h,g}$ is formed by the elements of $\mathcal{V}_h$ whose degrees of freedom associated with the boundary edges approximate the boundary datum g. We also consider the linear space $\mathcal{V}_{h,g}$, which is obtained by setting $g = 0$. Finally, the mimetic finite difference method for (3) reads:

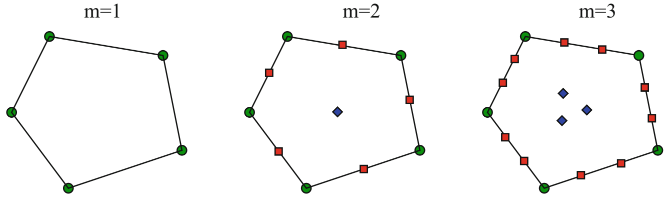


Fig. 1 Degrees of freedom for $m = 1, 2, 3$

Find $u_h \in \mathcal{V}_{h,g}$ such that:

$$\mathscr{A}_h\big(u_h, v_h\big) = \big(f, v_h\big)_h \qquad \forall v_h \in \mathcal{V}_{h,0}. \tag{4}$$

The well-posedness of the numerical approximation (4) follows from the coercivity and the continuity properties of the bilinear form $\mathscr{A}_h\big(\cdot,\cdot\big)$.

Degrees of freedom, norms, and the interpolation operator. Let $\mathscr{V}$ and $\mathscr{F}$ be sets of mesh vertices and edges, respectively. Let m be a positive integer number. A discrete scalar field v_h in $\mathcal{V}_h$ consists of:

(i) one real number v_{v} per mesh vertex $\mathsf{v} \in \mathscr{V}$;
(ii) $(m-1)$ real numbers $v_{\mathsf{f},i}$ per mesh edge $\mathsf{f} \in \mathscr{F}$, where $i = 1, \dots, m-1$;
(iii) $m(m-1)/2$ real numbers $v_{\mathsf{P},k,i}$ per mesh cell $\mathsf{P} \in \Omega_h$, where $k = 0, \dots, m-2$, and $i = 0, \dots, k$.

The first two sets of degrees of freedom represent *nodal degrees of freedom*. They are associated with the vertices of edge f and with the $m-1$ *internal* nodes of the *Gauss-Lobatto* numerical integration rule of order $2m-1$. These nodes are defined uniquely and symmetrically on each edge $\mathsf{f} \in \mathscr{F}$ (see formula 25.4.32 and Table 25.6 of [1] for details). The last set is introduced into the nodal MFD scheme to represent k-th order moments of scalar fields over polygons P. We refer to the latter degrees of freedom as the *internal degrees of freedom*. Examples of the degrees of freedom on a generic cell are shown in Fig. 1 for $m = 1, 2, 3$.

Combining the previous definitions allows us to write

$$v_h = \big\{(v_{\mathsf{v}})_{\mathsf{v}\in\mathscr{V}}, (v_{\mathsf{f},i})_{\mathsf{f}\in\mathscr{F},i=1,\dots,m-1}, (v_{\mathsf{P},k,i})_{\mathsf{P}\in\Omega_h,k=0,\dots,m-2,i=0,\dots,k}\big\} \tag{5}$$

for any $v_h \in \mathcal{V}_h$. Therefore, the global approximation space $\mathcal{V}_h$ has dimension $N^{\mathscr{V}} + N^{\mathscr{F}}(m-1) + N^{\mathscr{P}}m(m-1)/2$, where $N^{\mathscr{V}}$ is the number of mesh vertices, $N^{\mathscr{F}}$ the number of mesh edge, and $N^{\mathscr{P}}$ the number of mesh cells.

The sub-set $\mathcal{V}_{h,g}$ is obtained by approximating the datum g by a (globally continuous) piecewise polynomial function g_m of order m, and then enforcing $v_{h,\mathsf{f}} = g_m|_{\mathsf{f}}$ for every $\mathsf{f} \in \Gamma$. Note that if g is continuous, this can be simply achieved by interpolating g at the Gauss nodes of each edge.

We will find it useful to introduce $\mathscr{V}_{h,\mathsf{P}} := \mathscr{V}_h|_{\mathsf{P}}$, the linear space of discrete scalar fields whose members contain only the degrees of freedom associated with the vertices, the edge and the interior of cell P. The local linear space $\mathscr{V}_{h,\mathsf{P}}$ has dimension $m_{\mathscr{V}_{h,\mathsf{P}}} = N_{\mathsf{P}}^{\mathscr{F}} m + m(m-1)/2$, where $N_{\mathsf{P}}^{\mathscr{F}}$ is the number of the edges forming the boundary $\partial\mathsf{P}$ and the second term is the number of internal degrees of freedom.

Let $\bar{v}_{h,\mathsf{P}} = \sum_{\mathsf{v}\in\partial\mathsf{P}} v_{\mathsf{v}}/N_{\mathsf{P}}^{\mathscr{V}}$ be the arithmetic mean of the vertex values of $v_{h,\mathsf{P}}$. The mesh-dependent norm $||v_h||_{1,h}^2 = \sum_{\mathsf{P}\in\Omega_h} ||v_h||_{1,h,\mathsf{P}}^2$ for the elements of $\mathscr{V}_h$ mimics the $H^1(\Omega)$ seminorm when the summation arguments are given by

$$||v_h||_{1,h,\mathsf{P}}^2 = \sum_{\mathsf{f}\in\partial\mathsf{P}} h_{\mathsf{P}} \left\| \frac{\partial v_{h,\mathsf{f}}}{\partial s} \right\|_{L^2(\mathsf{f})}^2 + \left(v_{\mathsf{P},0,0} - \bar{v}_{h,\mathsf{P}}\right)^2 + \sum_{k=1}^{m-2}\sum_{i=0}^{k} |v_{\mathsf{P},k,i}|^2, \tag{6}$$

where $v_{h,\mathsf{f}}$ is the polynomial on edge f corresponding to the values $v_{\mathsf{f},i}$, $i = 0,\dots,m$ and $\partial v_{h,\mathsf{f}}/\partial s$ is the directional derivative along f.

For every polygon P of Ω_h and every function $v \in H^1(\mathsf{P}) \cap C^0(\overline{\mathsf{P}})$, we define the *interpolation operator* $v_{\mathsf{P}}^{\mathrm{I}}$ to the discrete local space $\mathscr{V}_{h,\mathsf{P}}$. For the nodal degrees of freedom, we set

$$v_{\mathsf{v}}^{\mathrm{I}} = v(\mathbf{x}_{\mathsf{v}}) \qquad \forall \mathsf{v} \in \partial\mathsf{P}; \tag{7}$$

$$v_{\mathsf{f},i}^{\mathrm{I}} = v(\mathbf{x}_{\mathsf{f},i}) \qquad \forall \mathsf{f} \in \partial\mathsf{P},\ i = 1,2,\dots,m-1. \tag{8}$$

Now, for every cell $\mathsf{P} \in \Omega_h$, we consider the set of $m(m-1)/2$ polynomial functions $\varphi_{k,i} : \mathsf{P} \to \mathbb{R}$ for $k = 0,\dots,m-2$ and $i = 0,\dots,k$ such that $\varphi_{0,0} = 1$ and for every k the set $\{\varphi_{k,i}\}_{i=0,\dots,k}$ forms an L^2-orthogonal basis for the linear space of the polynomials of degree exactly equal to k on P and orthogonal to the polynomials of degree up to $(k-1)$. Then, the interior degrees of freedom of the interpolated field v^{I} are given by

$$v_{\mathsf{P},k,i}^{\mathrm{I}} = \frac{1}{|\mathsf{P}|}\int_{\mathsf{P}} v\varphi_{k,i}\, dV, \qquad k = 0,\dots,m-2, \quad i = 0,1,\dots,k. \tag{9}$$

Construction of the mimetic bilinear form $\mathscr{A}_h(\cdot,\cdot)$. The bilinear form $\mathscr{A}_h(\cdot,\cdot)$ is obtained by assemblying the contributions from each polygonal cell

$$\mathscr{A}_h(u_h, v_h) = \sum_{\mathsf{P}\in\Omega_h} \mathscr{A}_{h,\mathsf{P}}(u_{h,\mathsf{P}}, v_{h,\mathsf{P}}) \qquad \forall u_h, v_h \in \mathscr{V}_h,$$

where $\mathscr{A}_{h,\mathsf{P}}(\cdot,\cdot) : \mathscr{V}_{h,\mathsf{P}} \times \mathscr{V}_{h,\mathsf{P}} \to \mathbb{R}$ is a *local* symmetric bilinear form defined on P. To define it, we proceed throughout the following three steps. In the first step, we introduce the linear vector functional $\mathscr{G}^k : (L^2(\mathsf{P}))^2 \to (\mathbb{P}_k(\mathsf{P}))^2$, which is the L^2-orthogonal projection on the linear space of two-dimensional vectors of polynomials of degree k on P. For $p \in H^1(\mathsf{P})$ and $\mathsf{K} \in L^\infty(\mathsf{P})$ it holds that $\mathrm{div}(\mathscr{G}^{m-1}(\mathsf{K}\nabla p))$ belongs to $\mathbb{P}_{m-2}(\mathsf{P})$. Therefore, we can express this divergence as the unique linear combination of the polynomial basis functions $\varphi_{k,i}$ of $\mathbb{P}_{m-2}(\mathsf{P})$:

$$\mathrm{div}(\mathscr{G}^{m-1}(\mathsf{K}\nabla p)) = \sum_{k=0}^{m-2}\sum_{i=0}^{k} \alpha_{k,i}\,\varphi_{k,i}, \tag{10}$$

where the coefficients $\alpha_{k,i}$ depend on p. In the second step, we define the *"numerical integration" formula* $\mathscr{I}_{\mathsf{P}}\big(v_{h,\mathsf{P}}, p\big)$ as:

$$\mathscr{I}_{\mathsf{P}}\big(v_{h,\mathsf{P}},\ p\big) = \sum_{k=0}^{m-2}\sum_{i=0}^{k} |\mathsf{P}|\,\alpha_{k,i}\,v_{\mathsf{P},k,i}, \tag{11}$$

where the real numbers $\alpha_{k,i}$ are the coefficients for the polynomial p in summation (10). In the third step, we assume that the symmetric bilinear form $\mathscr{A}_{h,\mathsf{P}}\big(\cdot,\cdot\big)$ satisfies the following two conditions:

($\mathsf{S1}$) *spectral stability*: there exists two positive constants σ_* and σ^* such that for every $v_{h,\mathsf{P}} \in \mathscr{V}_{h,\mathsf{P}}$ there holds:

$$\sigma_*||v_{h,\mathsf{P}}||^2_{1,h,\mathsf{P}} \le \mathscr{A}_{h,\mathsf{P}}\big(v_{h,\mathsf{P}}, v_{h,\mathsf{P}}\big) \le \sigma^*||v_{h,\mathsf{P}}||^2_{1,h,\mathsf{P}};$$

($\mathsf{S2}$) *local consistency*: for every $v_{h,\mathsf{P}} \in \mathscr{V}_{h,\mathsf{P}}$ and for every $p \in \mathbb{P}_m(\mathsf{P})$ there holds:

$$\mathscr{A}_{h,\mathsf{P}}\big(v_{h,\mathsf{P}}, p^{\mathrm{I}}_{\mathsf{P}}\big) = -\mathscr{I}_{\mathsf{P}}\big(v_{h,\mathsf{P}}, p\big) + \sum_{\mathsf{f}\in\partial\mathsf{P}}\int_{\mathsf{f}} v_{h,\mathsf{f}}(s)\,\mathscr{G}^{m-1}(\mathsf{K}\nabla p)\cdot\mathbf{n}_{\mathsf{P},\mathsf{f}}\,ds. \tag{12}$$

As usual, in the mimetic methods, the bilinear form $\mathscr{A}_{h,\mathsf{P}}$ is not unique. Its representing matrix has two terms, a unique consistency term due to ($\mathsf{S2}$) and a non-unique stabilizing term due to ($\mathsf{S1}$). The possibility to vary the stabilizing term is the unique characteristic of the mimetic approach; see also [2].

Discretization of the load term $\big(\cdot,\cdot\big)_h$. Let $\mathscr{P}^k_{\mathsf{P}} : L^2(\mathsf{P}) \to \mathbb{P}_k(\mathsf{P})$ be the L^2-orthogonal projector of scalar functions onto the space of polynomials of degree at most k. We introduce $\hat{f}_{\mathsf{P}} = \mathscr{P}^{m-2}_{\mathsf{P}}(f)$, where f is the forcing term in (3). Since $\hat{f}_{\mathsf{P}} \in \mathbb{P}_{m-2}(\mathsf{P})$, we can write it as a linear combination of the basis functions $\varphi_{k,i}$:

$$\hat{f}_{\mathsf{P}} = \sum_{k=0}^{m-2}\sum_{i=0}^{k} c_{k,i}\,\varphi_{k,i} \tag{13}$$

using the $(m+1)(m+2)/2$ real coefficients $c_{k,i}$. Then, we define

Table 1 Number of degrees of freedom

	Mesh Family $\mathscr{M}_1$				Mesh Family $\mathscr{M}_2$			
lev	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 2$	$m = 3$	$m = 4$	$m = 5$
0	441	861	1381	2001	881	1521	2261	3101
1	1681	3321	5361	7801	3361	5841	8721	12001
2	6561	13041	21121	30801	13121	22881	34241	47201
3	25921	51681	83841	122401	51841	90561	135681	187201
4	103041	205761	334081	488001	206081	360321	540161	745601

Table 2 Test Case 1: relative errors and convergence rates on mesh family $\mathscr{M}_1$ (randomized quadrilaterals) for different polynomial order $m = 2, 3, 4, 5$ and non-constant diffusion tensor K

	$m = 2$		$m = 3$		$m = 4$		$m = 5$	
lev	Error	Rate	Error	Rate	Error	Rate	Error	Rate
0	$1.529\ 10^{-1}$	——	$9.510\ 10^{-2}$	——	$8.590\ 10^{-3}$	——	$8.237\ 10^{-3}$	——
1	$2.577\ 10^{-2}$	2.60	$1.043\ 10^{-2}$	3.23	$1.496\ 10^{-3}$	2.55	$5.240\ 10^{-4}$	4.03
2	$5.218\ 10^{-3}$	2.29	$1.590\ 10^{-3}$	2.70	$9.476\ 10^{-5}$	3.96	$2.507\ 10^{-5}$	4.37
3	$1.192\ 10^{-3}$	2.19	$1.885\ 10^{-4}$	3.16	$6.620\ 10^{-6}$	3.94	$8.663\ 10^{-7}$	4.98
4	$2.991\ 10^{-4}$	2.06	$2.413\ 10^{-5}$	3.06	$4.199\ 10^{-7}$	4.11	$2.850\ 10^{-8}$	5.09

$$\left(f, v_h\right)_h = \sum_{\mathsf{P}\in\Omega_h} \left(\hat{f}_{\mathsf{P}}, v_{h,\mathsf{P}}\right)_{h,\mathsf{P}}, \qquad \left(\hat{f}_{\mathsf{P}}, v_{h,\mathsf{P}}\right)_{h,\mathsf{P}} = |\mathsf{P}| \sum_{k=0}^{m-2} \sum_{i=0}^{k} c_{k,i}\, v_{\mathsf{P},k,i}\,,$$

where we use the coefficients $c_{k,i}$ from (13).

2 Convergence theorem

For simplicity, we consider the homogeneous boundary value problem (3), i.e. $g = 0$.

Theorem 1. *Let $u \in H^{m+1}(\Omega)$ be the solution of the variational problem (3) under assumptions (**H1**)-(**H3**). Let $u^{\mathcal{I}} \in \mathcal{V}_h$ be its interpolant defined by (7)-(9). Let u_h be solution of the MFD problem (4) under assumption (**HG**) and (**S1**)-(**S2**). Let us assume that $\mathsf{K}|_{\mathsf{P}} \in W^{m,\infty}(\mathsf{P})$ for any polygon P. Finally, let mesh Ω_h be shape regular. Then, there exists a positive constant C, which depends only on the shape regularity constants and is independent of h, such that*

$$||u^{\mathcal{I}} - u_h||_{1,h} \leq C h^m ||u||_{H^{m+1}(\Omega)}. \tag{14}$$

Table 3 Test Case 2: relative errors and convergence rates on mesh family $\mathscr{M}_2$ (non-convex cells) for different polynomial order $m = 2, 3, 4, 5$ and non-constant diffusion tensor K

	$m = 2$		$m = 3$		$m = 4$		$m = 5$	
lev	Error	Rate	Error	Rate	Error	Rate	Error	Rate
0	3.007	−−	$9.873\ 10^{-1}$	−−	$2.059\ 10^{-1}$	−−	$1.988\ 10^{-2}$	−−
1	$8.081\ 10^{-1}$	1.89	$2.760\ 10^{-1}$	1.84	$1.367\ 10^{-2}$	3.92	$1.016\ 10^{-3}$	4.29
2	$2.071\ 10^{-1}$	1.96	$5.621\ 10^{-2}$	2.29	$7.562\ 10^{-4}$	4.18	$3.924\ 10^{-5}$	4.69
3	$5.303\ 10^{-2}$	1.97	$9.083\ 10^{-3}$	2.63	$4.210\ 10^{-5}$	4.17	$1.351\ 10^{-6}$	4.86
4	$1.348\ 10^{-2}$	1.98	$1.292\ 10^{-3}$	2.81	$2.441\ 10^{-6}$	4.11	$4.472\ 10^{-8}$	4.92

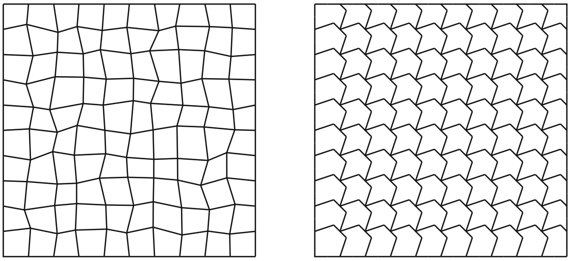


Fig. 2 The first mesh in families $\mathscr{M}_1$ (left) and $\mathscr{M}_2$ (right)

3 Numerical experiments

The numerical experiments presented in this section are aimed to confirm the *a priori* analysis developed in the previous section. To this purpose, we solve the discrete problem (4) on the domain $\Omega =]0,1[\times]0,1[$ with the diffusion tensor given by

$$\mathsf{K}(x,y) = \begin{pmatrix} (x+1)^2 + y^2 & -xy \\ -xy & (x+1)^2 \end{pmatrix}.$$

The forcing term f and the Dirichlet boundary condition g are set in accordance with the following exact solution:

- *test case 1:* $u(x,y) = \big(x - e^{2(x-1)}\big)\big(y^2 - e^{3(y-1)}\big)$;
- *test case 2:* $u(x,y) = e^{-2\pi y}\sin(2\pi x)$.

The performance of the MFD method is investigated by evaluating the rate of convergence on a sequence of refined meshes. Test case 1 is solved using mesh family $\mathscr{M}_1$, where each mesh is formed by randomized quadrilaterals; test case 2 is solved using mesh family $\mathscr{M}_2$, where each mesh is obtained by filling the unit square with a suitably scaled non-convex octagonal reference cell, see the Fig. 2. The meshes are parametrized by the number of partitions in each direction. The

Table 4 Mesh parameters

	Mesh Family $\mathscr{M}_1$				Mesh Family $\mathscr{M}_2$			
lev	$N^{\mathscr{P}}$	$N^{\mathscr{F}}$	$N^{\mathscr{V}}$	h	$N^{\mathscr{P}}$	$N^{\mathscr{F}}$	$N^{\mathscr{V}}$	h
0	100	220	121	$1.922\,10^{-1}$	100	440	341	$1.458\,10^{-1}$
1	400	840	441	$9.705\,10^{-2}$	400	1680	1281	$7.289\,10^{-2}$
2	1600	3280	1681	$4.838\,10^{-2}$	1600	6560	4961	$3.644\,10^{-2}$
3	6400	12960	6561	$2.467\,10^{-2}$	6400	25920	19521	$1.822\,10^{-2}$
4	25600	51520	25921	$1.263\,10^{-2}$	25600	103040	77441	$9.111\,10^{-3}$

starting mesh for both families is built from a 10×10 regular grid, and the refined meshes are obtained by doubling this parameter. The first mesh in each family is shown in Fig. 2. Mesh data for the refinement level lev are reported in Table 4. Here, $N^{\mathscr{P}}$, $N^{\mathscr{F}}$ and $N^{\mathscr{V}}$ are the numbers of mesh elements, edges and vertices, respectively. Table 1 shows the total number of degrees of freedom that are required by the nodal MFD method for $m = 2, \dots, 5$.

The rate of convergence is measured in the mesh-dependent norm (6). Relative errors and convergence rates are reported in Tables 2-3 and are in good accordance with the theoretical prediction of Theorem 1.

4 Conclusions

In this paper, we presented a new family of nodal arbitrary-order accurate MFD methods for unstructured polygonal meshes. The construction of the method is based on a local consistency condition, i.e., a discrete integration by parts formula, that holds for polynomials of degree m. The arbitrary-order accurate MFD methods use nodal degrees of freedom on mesh edges representing solution values at the quadrature nodes of the Gauss-Lobatto formulas and internal degrees of freedom inside polygons representing solution moments.

Acknowledgements The work of the second author was supported by the Department of Energy (DOE) Advanced Scientific Computing Research (ASCR) program in Applied Mathematics. The work of the third author was partially supported by the Italian MIUR through the program PRIN2008.

References

1. M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. Dover, New York, ninth Dover printing, tenth gpo printing edition, 1964.
2. L. Beirão da Veiga, K. Lipnikov, and G. Manzini. High-order nodal mimetic discretizations of elliptic problems on polygonal meshes. Submitted to SIAM J. Numer. Anal. (also IMATI-CNR Technical Report 32PV10/30/0, 2010).

3. F. Brezzi, A. Buffa, and K. Lipnikov. Mimetic finite differences for elliptic problems. *M2AN Math. Model. Numer. Anal.*, 43(2):277–295, 2009.
4. F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 43(5):1872–1896, 2005.
5. P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman, Boston, 1985.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Adaptive cell-centered finite volume method for non-homogeneous diffusion problems: Application to transport in porous media

Fayssal Benkhaldoun, Amadou Mahamane, and Mohammed Seaïd

Abstract We investigate time stepping schemes for the adaptive cell-centered finite volume solution of diffusion equations with heterogeneous diffusion coefficients. The proposed finite volume method uses the cell-centered techniques to discretize the diffusion operators on unstructured grids. Explicit and implicit time integration schemes are used and a comparative study is presented in terms of accuracy and efficiency. Numerical results are presented for a transient diffusion equation with known analytical solution. We also apply these methods to a problem of oil recovery using a two-phase flow problem in porous media.

Keywords Diffusion problems, finite volume, flow in porous media
MSC2010: 76S05, 65N08, 65Y20

1 Introduction

In the last decade, finite volume methods have offered a remarkable level of accuracy and robustness required for solving complex flow problems governed by hyperbolic systems of conservation laws. However, engineering applications often involve coupled hyperbolic and elliptic partial differential equations which have to be solved on complex geometries, thus suggesting the use of the same spatial discretization for both hyperbolic and elliptic equations. As an example of these applications where hyperbolic and elliptic equations coexist we mention

Fayssal Benkhaldoun and Amadou Mahamane
LAGA, Université Paris 13, 99 Av J.B. Clement, 93430 Villetaneuse, France,
e-mail: fayssal@math.univ-paris13.fr, mahamane@math.univ-paris13.fr

Mohammed Seaïd
School of Engineering and Computing Sciences, University of Durham, South Road,
Durham DH1 3LE, UK, e-mail: m.seaid@durham.ac.uk

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_9, © Springer-Verlag Berlin Heidelberg 2011

the multi-phase flows in porous media, see for example [1, 3, 5]. In practice, the focus is on unstructured meshes where a non-trivial reconstruction scheme is required to have a high-order spatial accuracy. Most of upwind finite volume methods for unstructured grids proposed to date employ a cell-vertex discretization, since it allows a natural definition of the flow gradients: using a dual mesh, a gradient-based reconstruction is applied on the two sides of each interface, where an approximate Riemann solver is finally applied to select the proper upwind contributions. However, solving diffusion equations using the finite volume methods is still a considerable task in the case of unstructured meshes; particularly when these equations have to be solved in conjunction with partial differential equations of hyperbolic type. The emphasis in this work is on the time integration of the resultant system of ordinary differential equations induced from cell-centered finite volume discretization in space variable of the transient diffusion problems. The proposed schemes are the explicit Euler and implicit Euler scheme. These two different methods lead to techniques all of which are occurring in time integration framework since years. Theoretical considerations can provide some ideas concerning stability, convergence rates, restriction on time stepsizes, or qualitative behaviour of the solution, but a complete quantitative analysis is not possible today. Therefore, the only way to make a judgment is to perform numerical tests, at least for some problems which seem to be representative. However, looking into the literature, it seems that there have not been many studies of this type which can give satisfactory answers.

2 Adaptive cell-centered finite volume method

Our main concern in the present study is on the finite volume discretization of the two-dimensional gradient operator $\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}\right)^T$ resulting from the weak formulation of the diffusion equations. To this end we discretize the spatial domain $\bar{\Omega} = \Omega \cup \partial\Omega$ in conforming triangular elements K_i as $\bar{\Omega} = \cup_{i=1}^{N} K_i$, with N is the total number of elements. Each triangle represents a control volume and the variables are located at the geometric centers of the cells. To discretize the diffusion operators we adapt the so-called cell-centered finite volume method based on a Green-Gauss diamond reconstruction, see for example [5] and further references are therein. Hence, a co-volume, D_σ, is first constructed by connecting the barycentres of the elements that share the edge σ and its endpoints as shown in Fig. 1. Then, the discrete gradient operator ∇_σ is evaluated at an inner edge σ as

$$\nabla_\sigma u_h = \frac{1}{2\text{meas}(D_\sigma)}\big((u_L - u_K)\text{meas}(\sigma)\mathbf{n}_{K,\sigma} + (u_S - u_N)\text{meas}(s_\sigma)\mathbf{n}'_\sigma\big), \qquad (1)$$

where u_h is the finite volume discretization of a generic function u, meas(D) denotes the area of the element D, $\mathbf{n}_{K,\sigma}$ denotes the unit outward normal to the surface σ, u_K and u_L are the values of the solution u_h in the elements K and L, respectively. In (1),

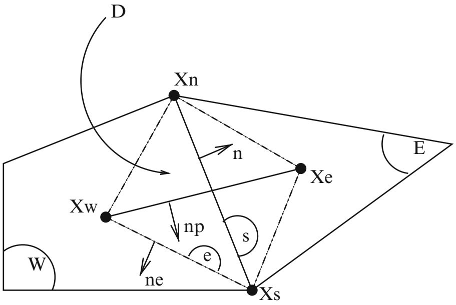


Fig. 1 A generic two-dimensional finite volume and notations

u_S and u_N are the values of the solution u_h at the co-volume nodes approximated by a linear interpolation from the values on the cells sharing the same vertex S and N, respectively. For further details on the formulation and analysis of the considered cell-centered finite volume method we refer to [4,5] among others.

The treatment of boundary conditions in the cell-centered finite volume method is performed using similar techniques as those described in [2,5]. In order to improve the efficiency of the proposed finite volume method, we have performed a mesh adaptation to construct a nearly optimal mesh able to capture the small solution features without relying on extremely fine grid in smooth regions far from steep gradients. In the present work, this goal is achieved by using an error indicator for the gradient of the solution. This indicator requires only information from solution values within a single element at a time and it is easily calculated, see for instance [2].

3 Time stepping schemes

For simplicity in the presentation we consider the transient diffusion problem

$$\frac{\partial u}{\partial t} - \nabla \cdot (\mathbb{K}(\mathbf{x})\nabla u) = f(\mathbf{x},t), \qquad (\mathbf{x},t) \in \Omega \times (0,T], \tag{2}$$

where Ω is a subset of $\mathbb{R}^2$ with smooth boundary $\partial\Omega$, $(0,T]$ is the time interval, $\mathbb{K}(\mathbf{x})$ is a 2×2 matrix with entries k_{ij}, and $f(\mathbf{x},t)$ is an external force. In the current study the spatial discretization of the diffusion equation (2) is carried out using the cell-centered finite volume method and two time stepping schemes are considered for the time integration.

For the time integration of (2) we discretize the time interval into subintervals $[t_n, t_{n+1}]$ with length Δt, $0 = t_0, t_1, \ldots, t_N = T$ and $t_n = n\Delta t$. We use the notation ω^n to denote the value of a generic function ω at time t_n. Hence, using the forward Euler method, the fully discrete version of the diffusion equation (2) reads

$$u_K^{n+1} = u_K^n + \frac{\Delta t}{\text{meas}(K)} \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}\left(u_h^n\right) \text{meas}(\sigma) + \Delta t f_K^n, \qquad \forall\ K \in \mathscr{T}, \quad (3)$$

where $\mathscr{E}_K$ is the set of all edges of the control volume K and $F_{K,\sigma}^n$ are the numerical fluxes reconstructed as

$$F_{K,\sigma}(u) = \mathbb{K}_\sigma \nabla_\sigma u \cdot \mathbf{n}_{K,\sigma}, \quad (4)$$

with $\mathbb{K}_\sigma$ is an averaged values of the diffusion matrix $\mathbb{K}$ on the edge σ and ∇_σ is the cell-centered finite volume discretization of the gradient operator defined in (1).

An implicit time stepping scheme for (2) is formulated using the backward Euler scheme as

$$u_K^{n+1} = u_K^n + \frac{\Delta t}{\text{meas}(K)} \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}\left(u_h^{n+1}\right) \text{meas}(\sigma) + \Delta t f_K^{n+1}, \qquad \forall\ K \in \mathscr{T}. \quad (5)$$

To find the solution u_K^{n+1} from (5) one has to solve, at each time level, a linear system of algebraic equations. In our simulations, the linear system is solved using the preconditioned GMRES solver with a convergence criteria of 10^{-6}, we use the diagonal as a preconditionner.

4 Numerical Results

In this section we examine the accuracy and performance of the proposed time stepping schemes using two test examples. The first example solves a transient diffusion equation with analytical solution that can be used to quantify errors in the time stepping schemes. The second example considers a problem of oil recovery using the equations of two-phase flows in porous media. This last example is used to qualify the considered implicit time stepping scheme for more complicated nonlinear convection-dominated flows. In all the computations reported herein, a two-level refining and fixed CFL numbers are used.

4.1 Accuracy test problem

First we consider the problem of a diffusion problem with manufactured exact solution in a squared domain $\Omega = [-1, 1] \times [-1, 1]$. Here, we solve the transient equation (2) subject to a nonhomogenuous diffusion tensor given by

$$\mathbb{K} = (1 + \alpha x)^2 \begin{pmatrix} 1 & 10^{-2} \\ 10^{-2} & 10^{-6} \end{pmatrix}. \quad (6)$$

The reaction term f is explicitly calculated such that the exact solution of the diffusion problem (2) and (6) is

$$U(x, y, t) = \sin(\pi x)\,\sin(\pi y)\left(1 - \mathrm{e}^{-\lambda t}\right). \tag{7}$$

In our computations $\alpha = \lambda = 0.1$, the initial condition is calculated from the analytical solution (7) and homogeneous Dirichlet boundary conditions are imposed on $\partial\Omega$. To quantify the errors in this test example we consider the L^2-error norm defined as

$$\|e_h\| = \left(\sum_{K\in\mathscr{T}} \mathrm{meas(K)e_K^2}\right)^{\frac{1}{2}},$$

where $e_K(T) = |U_K - u_K|$ with u_K and U_K are respectively, the computed and exact solutions on the control volume K. In the current study, we present numerical results at the transient regime corresponding to the simulation time $T = 1$.

Table 1 Relative error and statistics using the explicit and implicit schemes at the transient time $T = 1$ and different CFL numbers. min Δt is the minimum Δt used in the scheme and GMRES iter refers to the mean number of iterations in the GMRES solver

	Explicit scheme	Implicit scheme			
	CFL = 1	CFL = 5	CFL = 10	CFL = 50	CFL = 100
Coarse mesh					
min Δt	4.70E-04	2.35E-03	4.70E-03	2.35E-02	4.70E-02
Relative error	1.15E-02	9.65E-03	8.07E-03	1.44E-02	3.49E-02
# time steps	2126	426	213	43	22
CPU time	0.76	0.44	0.32	0.20	0.16
GMRES iter	——	5	8	28	44
# elements	896	896	896	896	896
# nodes	481	481	481	481	481
Fine mesh					
min Δt	7.32E-06	3.66E-05	7.32E-05	3.66E-04	7.32E-04
Relative error	1.73E-04	1.41E-004	1.097E-04	1.958E-04	5.569E-04
# time steps	136576	27316	13658	2732	1366
CPU time	4814.781	2187.42	1415.296	1981.735	1297.94
GMRES iter	——	2	4	29	42
# elements	57344	57344	57344	57344	57344
# nodes	28929	28929	28929	28929	28929
Adaptive mesh					
min Δt	8.91E-06	4.46E-05	8.92E-05	4.46E-04	8.91E-04
Relative error	1.83E-03	1.82E-03	1.80E-03	1.918E-03	2.12E-03
# time steps	112120	22383	11166	2192	1070
CPU time	1496.57	595.63	294.846	148.28	124.38
GMRES iter	——	2	3	9	19
# elements	17256	17249	17248	16480	15507
# nodes	8681	8676	8676	8292	7806

To quantify the considered time stepping schemes applied to this example we summarize in Table 1 the results obtained at the transient time $T = 1$. In this table we present the minimum time stepsize, the relative error, the number of time steps required to reach the steady state, the CPU time in seconds, the number of iterations in the GMRES solver, the number of elements and the number of nodes in the considered meshes. A simple inspection of these results shows that the implicit schemes can use larger CFL numbers than those required for a stable explicit scheme. Note that using large CFL numbers in the implicit scheme results in a decrease on the number of time steps needed to reach the steady state and at the same time results in an increase on the number of the iterations in the GMRES solver. As expected the highest accuracy is obtained for both explicit and implicit schemes on the fine mesh but with a large CPU time compared to the results on coarse and adaptive meshes. No remarkable difference is obtained in the relative errors for the implicit and explicit schemes on the coarse and fine meshes. However, using an adaptive mesh the implicit scheme produces smaller errors than the explicit scheme. On the other hand, due to grid adaptation the final mesh at CFL = 100 consists of 15507 cells only compared to the 57344 cells for the fixed fine mesh. This results in a significant reduction of the computational cost. An examination of the CPU times in the tables reveals that, the cell-centered finite volume method on fixed meshes requires more computational work than its adaptive counterpart.

4.2 Transport in porous media

We solve a problem of oil recovery in a two-dimensional porous reservoir. The problem statement is solving a two-phase flow problem in porous media on the computational domain defined by $\bar{\Omega} = \Omega \cup \partial\Omega$ with $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ as illustrated in Fig. 2. Here, the governing equations consist of the pressure equation

$$\begin{aligned} \mathbf{q} &= -d(u)\mathbb{K}(\mathbf{x})\nabla p, && (\mathbf{x},t) \in \Omega \times (0,T], \\ \nabla \cdot \mathbf{q} &= 0, && (\mathbf{x},t) \in \Omega \times (0,T], \\ \mathbf{q}\cdot\mathbf{n}\Big|_{\Gamma_1} = -1.4, \quad \mathbf{q}\cdot\mathbf{n}\Big|_{\Gamma_2} &= 0, \quad p\Big|_{\Gamma_3} = 0, && t \in (0,T], \end{aligned} \tag{8}$$

and the saturation equation

$$\begin{aligned} \phi(\mathbf{x})\frac{\partial u}{\partial t} - \nabla \cdot \big(b(u)\mathbf{q} - \mathbb{K}(\mathbf{x})a(\mathbf{x})\nabla u\big) &= 0, && (\mathbf{x},t) \in \Omega \times (0,T], \\ u\Big|_{\Gamma_1} = 1, \quad \mathbb{K}\cdot\mathbf{n}\Big|_{\Gamma_2} = 0, \quad u\Big|_{\Gamma_3} &= 0, && t \in (0,T], \\ u(\mathbf{x},0) &= u_0(\mathbf{x}), && \mathbf{x} \in \Omega, \end{aligned} \tag{9}$$

where p is the pressure, $\mathbf{q}$ the Darcy velocity, u the saturation, $\mathbb{K}$ is the permeability of the medium, ϕ the porosity and $d(u) = k_w(u) + k_o(u)$ is the total mobility with $k_w(u)$ and $k_o(u)$ are the mobility of water and oil, respectively. In (9),

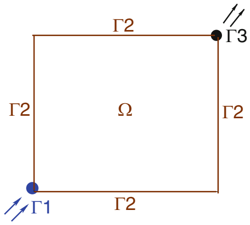


Fig. 2 Computational domain for the example of transport in porous media

$$b(u) = \frac{k_w(u)}{k_w(u) + k_o(u)}, \qquad a(u) = \frac{k_w(u)k_o(u)}{k_w(u) + k_o(u)} p'(u),$$

with $p(u)$ represents the capillary pressure. Note that, in most practical applications, the effects of Darcy velocity dominates the effects of diffusion. As a consequence the saturation equation (9) results in a convection-dominated problem which requires special numerical treatment and many numerical methods from the literature fail to accurately approximate its solution. In addition, since the diffusion in (9) depends on the Darcy velocity, the accuracy on the solution of saturation equation (9) strongly needs an accurate solution of the pressure equation (8). For more details on this model we refer the reader to [1, 3, 5] among others. In our simulations, the permeability $\mathbb{K} = Id$, with Id is the 2×2 identity matrix, the porosity $\phi = 0.2$, the water and oil mobility along with the capillary pressure are given by

$$k_w(u) = \frac{1}{2}u^5, \qquad k_o(u) = \frac{1}{3}(1-u)^3, \qquad p(u) = -\sqrt{\frac{1-u}{u}}.$$

The initial condition u_0 is defined as

$$u_0(\mathbf{x}) = \begin{cases} 1, & \text{if} \quad \mathbf{x} \in \Gamma_1, \\ 0, & \text{elsewhere.} \end{cases}$$

This problem has an interesting structure and will be used to verify the adaptive cell-centered finite volume method namely, to verify that the adaptation methodology is able to compute the right speed of the saturation fronts, and to verify that adaptive refinement is computationally cheaper than fixed mesh for a given level of solution resolution. Based on the conclusions drawn in the previous test example, only results using the implicit time stepping scheme are presented for this example. Figure 3 shows the adapted meshes and plots of the saturation at two different times, namely $t = 0.022$ and $t = 0.048$. The initial mesh contains 3662 cells and a $\Delta t = 2 \times 10^{-5}$ is used in our simulations. At earlier time of the simulation, the front entering the reservoir starts to develop and will be advected later on by the flow at far exit

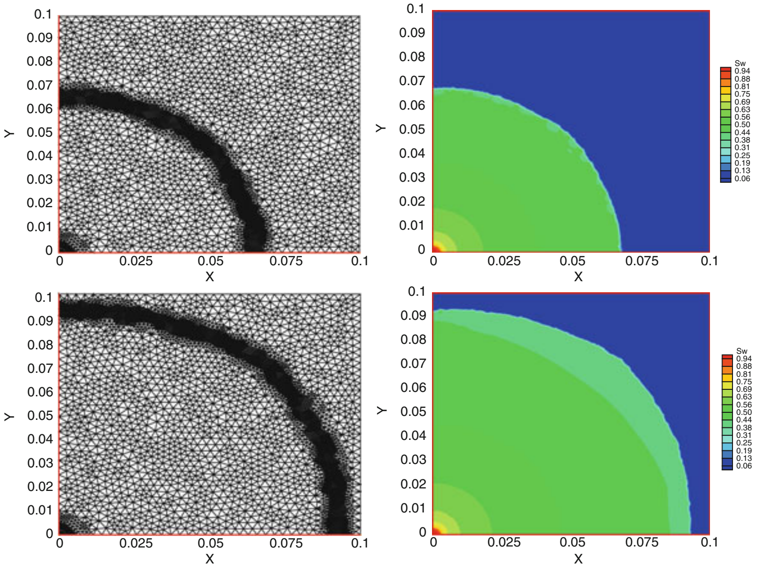


Fig. 3 Adaptive meshes and saturation contours at time $t = 0.022$ (top) and $t = 0.048$ (bottom)

of the reservoir. The interaction between the Darcy pressure and the saturation is detected across the reservoir during the simulation time. It can be clearly seen that the saturation front structures being captured by the cell-centered finite volume method. Another important result is that positions of the saturated waves are not deteriorated by the multiple mesh adaptations. The adaptive cell-centered finite volume method accurately approximates the solution to this problem of two-phase flow in porous media. In addition, the comparison with similar numerical results available in the literature [1] on the same test case is also satisfactory. It should be stressed that, due to grid adaptation the final mesh at times $t = 0.022$ and $t = 0.048$ consists of 4039 and 4947 cells, respectively. This results in a significant reduction of the computational cost compared to a cell-centered finite volume method on fixed meshes.

References

1. M. Afif, B. Amaziane, Convergence of finite volume schemes for degenerate convection-diffusion equation arising in flow in porous media, *Comput. Methods Appl. Mech. Engrg.* (2002) 5265–5286.
2. F. Benkhaldoun, I. Elmahi and M. Seaïd, Well-balanced finite volume schemes for pollutant transport by shallow water equations on unstructured meshes, *J. Comput. Phys.* **226** (2007) 180–203.
3. G. Chavent, J. Jaffré, *Mathematical models and finite element for reservoir simulation*, North-Holland. 1986.

4. Y. Coudière, J.P. Vila, P. Villedieu, Convergence rate of finite volume scheme for a two dimensional convection-diffusion problem, *M2AN.* **33** (1999) 493–516.
5. A. Mahamane, Analyse et estimation d'erreur en volumes finis, Application aux écoulements en milieu poreux et á l'adaptation de maillage, *Dissertation, Université Paris 13*, 2009.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Generalized Rusanov method for Saint-Venant Equations with Variable Horizontal Density

Fayssal Benkhaldoun, Kamel Mohamed, and Mohammed Seaïd

Abstract We present a class of finite volume methods for the numerical solution of Saint-Venant equations with variable horizontal density. The model is based on coupling the Saint-Venant equations for the hydraulic variables with a suspended sediment transport equation for the concentration variable. To approximate the numerical solution of the considered models we propose a generalized Rusanov method. The method is simple, accurate and avoids the solution of Riemann problems during the time integration process. Using flux limiters, a second-order accuracy is achieved in the reconstruction of numerical fluxes. The proposed finite volume method is well-balanced, conservative, non-oscillatory and suitable for Saint-Venant equations for which Riemann problems are difficult to solve. The numerical results are presented for two test examples.

Keywords Shallow water equations, variable density, finite volume method
MSC2010: 35L04, 65N08, 76L05

1 Introduction

In this paper we are interested to develop a robust finite volume method for solving Saint-Venant equations with variable horizontal density. The governing equations

Fayssal Benkhaldoun
LAGA, Université Paris 13, 99 Av J.B. Clement, 93430 Villetaneuse, France,
e-mail: fayssal@math.univ-paris13.fr

Kamel Mohamed
Department of Computer Science, Faculty of Applied Sciences, University of Taibah, Madinah KSA, e-mail: kamel16@yahoo.com

Mohammed Seaïd
School of Engineering and Computing Sciences, University of Durham, South Road, Durham DH1 3LE, UK, e-mail: m.seaid@durham.ac.uk

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_10, © Springer-Verlag Berlin Heidelberg 2011

can be formulated in a conservative form as

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{W})}{\partial x} = \mathbf{Q}(\mathbf{W}), \tag{1}$$

where $\mathbf{W}$, $\mathbf{F}(\mathbf{W})$ and $\mathbf{Q}(\mathbf{W})$ are vector-valued functions in $\mathbb{R}^3$ given by

$$\mathbf{W} = \begin{pmatrix} \rho h \\ \rho h u \\ \rho_s h c \end{pmatrix}, \qquad \mathbf{F}(\mathbf{W}) = \begin{pmatrix} \rho h u \\ \rho h u^2 + \frac{1}{2} g \rho h^2 \\ \rho_s h u c \end{pmatrix}, \qquad \mathbf{Q}(\mathbf{W}) = \begin{pmatrix} 0 \\ -g \rho h \frac{\partial Z}{\partial x} \\ 0 \end{pmatrix},$$

where h is the water height above the bottom, u the water velocity, g the acceleration due to gravity, Z the function characterizing the bottom topography and ρ_s the sediment density. For constant density ρ, the equations (1) reduce to the standard Saint-Venant equations. In the current work, we assume that a sediment transport takes place such that the density depends on space and time variables, *i.e.*, $\rho = \rho(x, t)$. This requires an additional equation for its evolution. Here, the equation used to close the system is given by

$$\rho = \rho_w + (\rho_s - \rho_w)\, c, \tag{2}$$

where ρ_s is the sediment density and c is the depth-averaged concentration of the suspended sediment. Further details on the formulation of the above equations we refer to [3] and further references are therein. It is clear that the system (1) is hyperbolic and the associated eigenvalues λ_k $(k = 1, 2, 3)$ are

$$\lambda_1 = u - \sqrt{gh}, \qquad \lambda_2 = u \qquad \text{and} \qquad \lambda_3 = u + \sqrt{gh}. \tag{3}$$

Note that in the above hydrodynamical model, we have considered only the source terms related to bottom topography while the source terms related to bed friction are neglected. Moreover, the bed-load sediment transport is assumed to be negligible in the considered model compared to the suspended sediment load. It should also be stressed that the transport of suspended sediments involves different physical mechanisms occurring within different time scales according to their time response to the hydrodynamics. In practice, the sediment transport of the bed occurs on a transport time scale much longer than the flow time scale. It is therefore desirable to construct numerical schemes that preserve stability for all time scales. In the current study we propose a modified Rusanov method studied and analyzed in [1] for the numerical solution of conservation laws with source terms. This method is simple, accurate and avoids the solution of Riemann problems during the time integration process. Our main goal is to present a class of numerical methods that are simple, easy to implement, and accurately solves the Saint-Venant equations with variable horizontal density without relying on Riemann solvers or front tracking techniques.

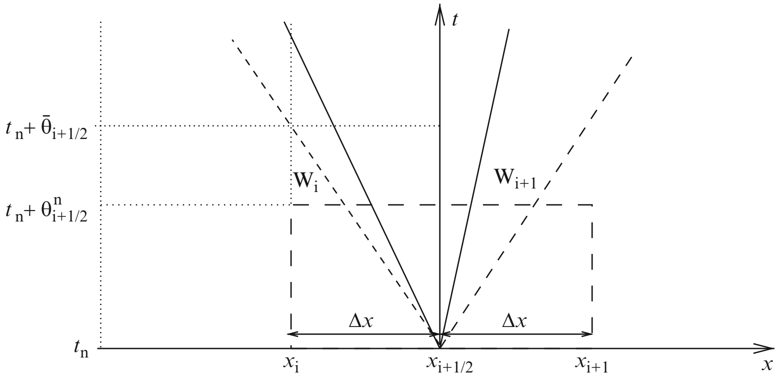


Fig. 1 An illustration of modified Riemann problems in the proposed finite volume method

2 A generalized Rusanov method

To formulate our finite volume method, we discretize the spatial domain into control volumes $[x_{i-1/2}, x_{i+1/2}]$ with uniform size $\Delta x = x_{i+1/2} - x_{i-1/2}$ and we divide the temporal domain into subintervals $[t_n, t_{n+1}]$ with uniform size Δt. Following the standard finite volume formulation, we integrate the equation (1) with respect to time and space over the domain $[t_n, t_{n+1}] \times [x_{i-1/2}, x_{i+1/2}]$ to obtain the following discrete equation

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{\Delta x}\Big(\mathbf{F}(\mathbf{W}_{i+1/2}^n) - \mathbf{F}(\mathbf{W}_{i-1/2}^n)\Big) + \Delta t \mathbf{Q}_i^n, \tag{4}$$

where $\mathbf{W}_i^n$ is the time-space average of the solution $\mathbf{W}$ in the domain $[x_{i-1/2}, x_{i+1/2}]$ at time t_n and $\mathbf{F}(\mathbf{W}_{i\pm1/2}^n)$ is the numerical flux at $x = x_{i\pm1/2}$ and time t_n. The spatial discretization of the equation (4) is complete when a numerical construction of the fluxes $\mathbf{F}(\mathbf{W}_{i\pm1/2}^n)$ is chosen and a discretization of the source term $\mathbf{Q}_i^n$ is performed. In general, the construction of numerical fluxes requires a solution of Riemann problems at the interfaces $x_{i\pm1/2}$.

In order to avoid these difficulties and reconstruct an approximation of $\mathbf{W}_{i+1/2}^n$, we adapt a finite volume method proposed in [1] for numerical solution of conservation laws with source terms. The key idea is to integrate the equation (1) over a control domain $[t_n, t_n + \theta_{i+1/2}^n] \times [x_i, x_{i+1}]$ containing the point $(t_n, x_{i+1/2})$ as depicted in Fig. 1. It should be stressed that, the integration of the equation (1) over the control domain $[t_n, t_n + \theta_{i+1/2}^n] \times [x_i, x_{i+1}]$ is used only at a predictor stage to construct the intermediate states $\mathbf{W}_{i\pm1/2}^n$ which will be used in the corrector stage (4). Here, $\mathbf{W}_{i\pm1/2}^n$ can be viewed as an approximation of the averaged Riemann solution over the control volume $[x_i, x_{i+1}]$ at time $t_n + \theta_{i+1/2}^n$. Thus, the resulting intermediate state is given by

$$\mathbf{W}^n_{i+1/2} = \frac{1}{2}\left(\mathbf{W}^n_i + \mathbf{W}^n_{i+1}\right) - \frac{\theta^n_{i+1/2}}{\Delta x}\Big(F(\mathbf{W}^n_{i+1}) - F(\mathbf{W}^n_i)\Big) + \theta^n_{i+1/2} Q^n_{i+1/2}, \quad (5)$$

where $Q^n_{i+1/2}$ is an approximation of the averaged source term Q *i.e.*

$$Q^n_{i+1/2} = \frac{1}{\theta^n_{i+1/2}\Delta x}\int_{t_n}^{t_n+\theta^n_{i+1/2}}\int_{x_i}^{x_{i+1}} Q(\mathbf{W})\,dt\,dx.$$

In order to complete the implementation of the above finite volume method the parameters $\theta^n_{i+1/2}$ and $Q^n_{i+1/2}$ have to be selected. Based on the stability analysis reported in [1] for conservation laws with source terms, the variable $\theta^n_{i+1/2}$ is selected as

$$\theta^n_{i+1/2} = \alpha^n_{i+1/2}\bar{\theta}_{i+1/2}, \qquad \bar{\theta}_{i+1/2} = \frac{\Delta x}{2S^n_{i+1/2}}, \quad (6)$$

where $\alpha^n_{i+1/2}$ is a positive parameter to be calculated locally and $S^n_{i+1/2}$ is the local Rusanov's velocity defined as

$$S^n_{i+1/2} = \max_{k=1,2,3}\left(\max\Big(\left|\lambda^n_{k,i}\right|, \left|\lambda^n_{k,i+1}\right|\Big)\right), \quad (7)$$

with $\lambda^n_{k,i}$ is the kth eigenvalue in (3) evaluated at the solution state $\mathbf{W}^n_i$. Notice that the introduction of the local time step $\theta^n_{i+1/2}$ in the predictor stage (5) is motivated by the fact that $\theta^n_{i+1/2}$ should not be larger than the value $\bar{\theta}_{i+1/2}$ which corresponds to the time required for the fastest wave generated at the interface $x_{i+1/2}$ to leave the cell $[x_i, x_{i+1}]$, compare Fig. 1.

It is clear that by setting $\alpha^n_{i+1/2} = 1$ the proposed finite volume method reduces to the well-established Rusanov method for linear systems of conservation laws, whereas for $\alpha^n_{i+1/2} = \Delta t/\Delta x S^n_{i+1/2}$ one recovers the well-known Lax-Wendroff scheme. Another choice of the slopes $\alpha^n_{i+1/2}$ leading to a first-order scheme is $\alpha^n_{i+1/2} = \tilde{\alpha}^n_{i+1/2}$ with

$$\tilde{\alpha}^n_{i+1/2} = \frac{S^n_{i+1/2}}{s^n_{i+1/2}}, \quad (8)$$

where

$$s^n_{i+1/2} = \min_{k=1,2,3}\left(\min\Big(\left|\lambda^n_{k,i}\right|, \left|\lambda^n_{k,i+1}\right|\Big)\right). \quad (9)$$

In the current study we incorporate limiters in its reconstruction as

$$\alpha^n_{i+1/2} = \tilde{\alpha}^n_{i+1/2} + \sigma^n_{i+1/2}\Phi\left(r_{i+1/2}\right), \quad (10)$$

where $\tilde{\alpha}^n_{i+1/2}$ is given by (8) and $\Phi_{i+1/2} = \Phi\left(r_{i+1/2}\right)$ is an appropriate limiter which is defined by using a flux limiter function Φ acting on a quantity that measures the ratio $r_{i+1/2}$ of the upwind change to the local change, see for instance [6]. In the present study,

$$\sigma_{i+1/2}^{n} = \frac{\Delta t}{\Delta x} S_{i+1/2}^{n} - \frac{S_{i+1/2}^{n}}{s_{i+1/2}^{n}},$$

and the ratio of the upwind change is calculated locally as

$$r_{i+1/2} = \frac{W_{i+1-q} - W_{i-q}}{W_{i+1} - W_{i}}, \qquad q = \operatorname{sgn}\left[\tilde{\alpha}_{i+1/2}^{n}\right].$$

As a slope limiter function, we consider the Minmod function

$$\Phi(r) = \max\left(0, \min\left(1, r\right)\right). \tag{11}$$

Note that other slope limiter functions functions from [4, 6] can also apply. The reconstructed slopes (10) are inserted in (6) and the numerical fluxes $\mathbf{W}_{i+1/2}^{n}$ are computed from (5). Remark that if we set $\Phi = 0$, the spatial discretization (10) reduces to the first-order scheme.

3 Numerical Results

Two test examples are selected to check the accuracy and the performance of the proposed finite volume scheme. As with all explicit time stepping methods the theoretical maximum stable time step Δt is specified according to the Courant-Friedrichs-Lewy condition

$$\Delta t = Cr \frac{\Delta x}{\max\limits_{i}\left(\left|\alpha_{i+1/2}^{n}\right|\right)}, \tag{12}$$

where Cr is a constant to be chosen less than unity. In all our simulations, the fixed Courant number $Cr = 0.5$ is used and the time step is varied according to (12).

3.1 Example 1

We consider a density dam-break problem with a single initial discontinuity. The problem consists of solving the equations (1) in a flat channel of length 500 m filled with two liquids with density $\rho = 10\ kg/m^3$ in the left section and $\rho = 1\ kg/m^3$ in the right section. Initially, the system is at rest with constant water height $h = 1\ m$ and $g = 1\ m/s^2$. In Fig. 2 we display the time evolution of the density, water height, velocity and concentration variables using a mesh with 500 gridpoints. It is clear from these results that at the initial time, the hydrostatic pressure difference at the interface of the two liquids drives a flow of higher density liquid towards the right, pushing the lower density liquid ahead. To conserve mass, the free surface of the lower density liquid rises and a rightward propagating shock-like bore forms. This flow features have been accurately captured by our generalized Rusanov scheme. It should be stressed that the mechanisms of the density dam-break problems are similar to that of the standard dam-break induced by change in free-surface depth, in that a leftward rarefaction, a rightward shock and a contact wave are formed. Similar wave structures also occur in shock tube gas dynamics.

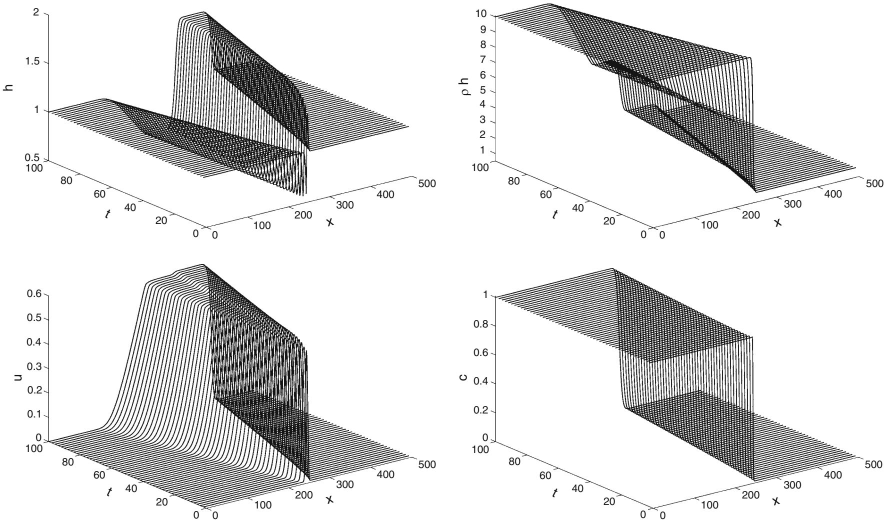


Fig. 2 Numerical results for density dam-break problem with a single initial discontinuity

For the sake of comparison, we present in Fig. 3 the results for the water height at $t = 70\ s$ obtained using the classical Rusanov method and the proposed method using a mesh with 500 and 1000 gridpoints. We have also included a reference solution obtained using a refined mesh with 100000 gridpoints. As can be seen from this figure, the results obtained using the classical Rusanov method are more diffusive than those obtained using our finite volume method. Similar conclusion can be drawn from other results (not reported here) obtained for the velocity field and sediment concentration.

3.2 Example 2

In this example we solve a density dam-break problem with two initial discontinuities. Here, a flat channel of length 100 m is filled at the left-hand side and right-hand side of the channel with a liquid with density $\rho = 1\ kg/m^3$. At the centre of the channel there is a liquid column of density $\rho = 10\ kg/m^3$ and width

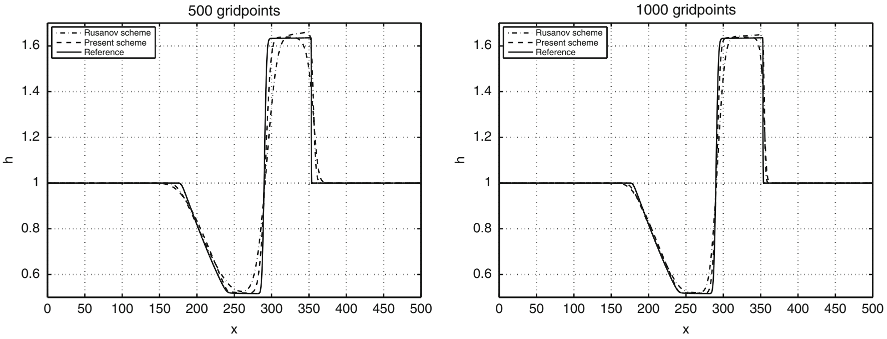


Fig. 3 Comparative results for the water height at $t = 70$ s for density dam-break problem with a single initial discontinuity using a mesh with 500 gridpoints (left) and 1000 gridpoints (right)

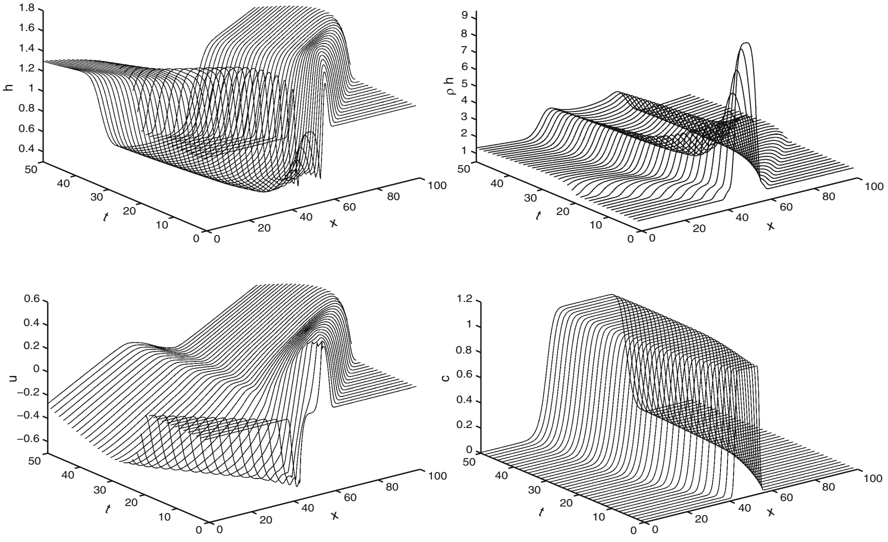


Fig. 4 Numerical results for density dam-break problem with two initial discontinuities

of 1 m. Initially, the system is at rest with constant water height $h = 1$ m and $g = 1$ m/s^2. The computed results are illustrated in Fig. 4 for the t-x phase space. It is evident that the sudden collapse of the denser liquid in the central column causes primary shock waves to be created and propagate as bores in the direction from high to low density. Two outward propagating bores are generated, traveling in opposite directions. Each primary bore decreases in strength with time, which can be seen from the curved shock path. On the other hand, a pair of rarefaction waves travels inward from the interfaces. The rarefaction waves are almost immediately reflected at the center, and then move outward, weakening rapidly. The accuracy of the proposed finite volume is highly achieved in reproducing these physical features.

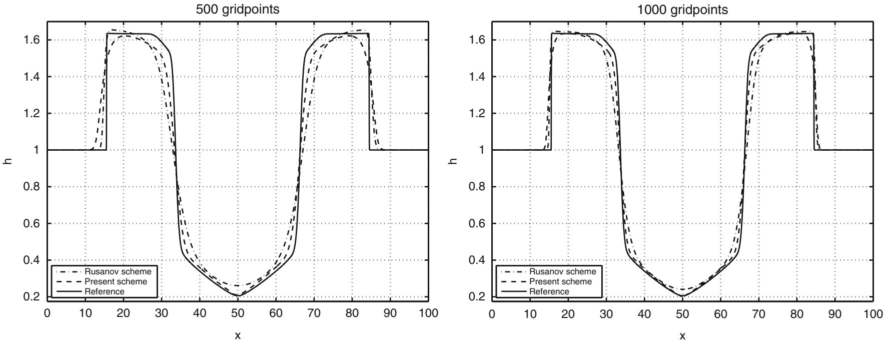


Fig. 5 Comparative results for the water height at $t = 20\ s$ for density dam-break problem with two initial discontinuities using a mesh with 500 gridpoints (left) and 1000 gridpoints (right)

In Fig. 5 we illustrate a comparison between the results for the water height at $t = 20\ s$ obtained using the classical Rusanov method and the proposed method using a mesh with 500 and 1000 gridpoints. Again, a reference solution obtained using a refined mesh with 100000 gridpoints is included in this figure. As in the previous test example, an excessive numerical diffusion is detected in the results obtained using the classical Rusanov method. This numerical diffusion has been noticeably reduced in the results obtained using the proposed finite volume method.

References

1. Mohamed, K.: Simulation numérique en volume finis, de problémes d'écoulements multidimensionnels raides, par un schéma de flux á deux pas. Dissertation, University of Paris 13, (2005)
2. Rusanov, V.: Calculation of interaction of non-steady shock waves with obstacles. Comp. Math. Phys. USSR. **1**, 267–279 (1961)
3. Leighton, F.Z. Borthwick, A.G.L. Taylor, P.H.: 1-D numerical modelling of shallow flows with variable horizontal density. International Journal for Numerical Methods in Fluids. **62**, 1209–1231 (2010)
4. Randall, J.L.: Numerical Methods for Conservation Laws, Lectures in Mathematics. ETH Zürich, (1992)
5. Roe, P.: Approximate riemann solvers, parameter vectors and difference schemes. J. Comp. Physics. **43**, 357–372 (1981)
6. Sweby, P.K.: High resolution schemes using flux limiters for hyperbolic conservation laws. SIAM J. Numer. Anal. **21**, 995–1011 (1984)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Hydrostatic Upwind Schemes for Shallow–Water Equations

Christophe Berthon and Françoise Foucher

Abstract We consider the numerical approximation of the shallow–water equations with non–flat topography. We introduce a new topography discretization that makes all schemes to be well–balanced and robust. At the discrepancy with the well–known hydrostatic reconstruction, the proposed numerical procedure does not involve any cut–off. Moreover, the obtained scheme is able to deal with dry areas. Several numerical benchmarks are performed to assert the interest of the method.

Keywords shallow–water equations, finite volume schemes, source term approximations, well–balanced schemes, positive preserving schemes
MSC2010: 65M12, 76M12, 35L65

1 Introduction

The present work concerns the derivation of finite volume methods to approximate the solutions of the shallow–water equations when involving non–flat topography. The model under interest is given as follows:

$$\begin{cases} \partial_t h + \partial_x hu = 0, \\ \partial_t hu + \partial_x \left(hu^2 + g\dfrac{h^2}{2} \right) = -gh\partial_x z, \end{cases} \tag{1}$$

where $h > 0$ is the local water–depth and $u \in \mathbb{R}$ is the depth–average velocity. Here, $z : \mathbb{R} \to \mathbb{R}^+$ denotes the topography while $g > 0$ is the gravitational constant.

C. Berthon and F. Foucher
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, francoise.foucher@ec-nantes.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_11, © Springer-Verlag Berlin Heidelberg 2011

To shorten the notations, let us set

$$w = \begin{pmatrix} h \\ hu \end{pmatrix}, \qquad f(w) = \begin{pmatrix} hu \\ hu^2 + g\dfrac{h^2}{2} \end{pmatrix},$$

respectively the state vector $\mathbb{R} \times \mathbb{R}^+ \to \Omega$ and the flux function $\Omega \to \mathbb{R}^2$ where Ω stands for the convex space of admissible states:

$$\Omega = \{w \in \mathbb{R}^2;\ h > 0,\ hu \in \mathbb{R}\}.$$

Let us recall that the steady state of the lake at rest, defined by $u = 0$ and $h + z =$ cste, is of primary importance and it must be preserved by the numerical schemes.

The objective is now to derive schemes which preserve positive the water depth and exactly capture the lake at rest. Several approaches have been recently introduced in the literature to approximate the solutions of (1). Most of them propose to consider the discretization of the associated homogeneous system:

$$\partial_t w + \partial_x f(w) = 0, \tag{2}$$

to suggest a relevant approximation of the topography source term able to restore the expected lake at rest. For instance, we refer to [10] where a suitable correction of the VFRoe scheme is proposed (see also [4, 11, 12]). Other approaches consider systematic corrections to enforce the required *well–balanced* property; i.e. to enforce the exact capture of the lake at rest. The hydrostatic reconstruction [1] (for instance, see also [5, 18]) is certainly one of the most celebrate *well–balanced* technique. However, to be water–depth positive preserving, this approach introduces a cut–off procedure which may involve some failures. Indeed, after the work by Delestre [9], let us consider a small water–depth on a topography with constant slope. Next, let us increase the slope to reduce the water–depth. Considering coarse grids, the hydrostatic reconstruction introduces a wrong behavior since the water–depth increases (see Fig. 4).

In this paper, we derive a new strategy to systematically enforce the well–balanced property. In fact, the proposed scheme is able to deal with vanishing water height without any additional correction and it is proved to be robust. Arguing [2, 15–17], we suggest a relevant upwind form of the topography source term but by involving the free surface.

The paper is organized as follows. In the next section, we introduce a new formulation of (1) and we present the suggested scheme. Section 3 is devoted to some essential properties as consistency and robustness. The last section presents numerical experiments to illustrate the interest of the method. Specifically, the adopted numerical scheme is shown to capture the correct behavior for large topography slopes.

2 Upwind source term discretization

In order to derive our scheme, we need considering a reformulation of system (1) by introducing the free surface $H = h + z$ and a fraction of water $X = \frac{h}{H}$. As usual, one can chose arbitrarily the topography origin (see [6, 13] and references therein), and thus we impose a bottom reference so that $H > 0$. As a consequence, X is well–defined. Involving such notations, the weak solutions of (1) satisfy the following system:

$$\begin{cases} \partial_t h + \partial_x XHu = 0, \\ \partial_t hu + \partial_x \left(X \left(Hu^2 + g\dfrac{H^2}{2} \right) - \dfrac{g}{2}hz \right) + gh\partial_x z = 0. \end{cases} \tag{3}$$

Let us remark the following identity:

$$\frac{1}{2}\partial_x(hz) - h\partial_x z = \frac{H^2}{2}\partial_x X,$$

to simplify the discharge equation. Then the smooth solutions of (1) also satisfy:

$$\begin{cases} \partial_t h + \partial_x XHu = 0, \\ \partial_t hu + \partial_x \left(X \left(Hu^2 + g\dfrac{H^2}{2} \right) \right) - g\dfrac{H^2}{2}\partial_x X = 0. \end{cases} \tag{4}$$

For the sake of simplicity in the notations, let us set

$$W = (H, Hu)^T,$$

to rewrite (4) as follows:

$$\partial_t w + \partial_x X f(W) - \begin{pmatrix} 0 \\ g\frac{H^2}{2}\partial_x X \end{pmatrix} = 0.$$

Now, we propose a discretization of (4), but let us note from now on that the discrete form we will obtain for (4) will be consistent with (3). As a consequence, the suggested discretization will be relevant to approximate the weak solutions of (3) or equivalently (1). We suggest to modify any scheme associated with the homogeneous system (2). Hence, let us consider $f^{\Delta x} : \Omega \times \Omega \to \mathbb{R}^2$ a consistent numerical flux function, i.e. $f^{\Delta x}(w, w) = f(w)$. We restrict ourselves to the regular meshes of size Δx such that $\Delta x = x_{i+1/2} - x_{i-1/2}$ for all $i \in \mathbb{Z}$, and we denote the time step by Δt, with $t^{n+1} = t^n + \Delta t$ for all $n \in \mathbb{N}$. The proposed scheme thus reads:

$$w_i^{n+1} = w_i^n - \frac{\Delta t}{\Delta x}\left(X_{i+1/2}^n f^{\Delta x}(W_i^n, W_{i+1}^n) - X_{i-1/2}^n f^{\Delta x}(W_{i-1}^n, W_i^n)\right) + \begin{pmatrix} 0 \\ \frac{\Delta t}{\Delta x}\frac{g}{2} H_{i+1/2}^n H_{i-1/2}^n (X_{i+1/2}^n - X_{i-1/2}^n) \end{pmatrix}, \tag{5}$$

where we have set

$$H_{i+1/2}^n = \begin{cases} H_i^n, & if\ f_h^{\Delta x}(W_i^n, W_{i+1}^n) > 0, \\ H_{i+1}^n, & otherwise, \end{cases} \tag{6}$$

$$X_{i+1/2}^n = \begin{cases} X_i^n, & if\ f_h^{\Delta x}(W_i^n, W_{i+1}^n) > 0, \\ X_{i+1}^n, & otherwise, \end{cases} \tag{7}$$

where $f_h^{\Delta x}$ denotes the first component of the numerical flux function. Here, we have set $W_i^n = X_i^n h_i^n$ where $X_i^n = h_i^n/(h_i^n + Z_i)$.

Before we establish the main properties of the scheme, let us emphasize that the suggested numerical procedure is an extremely easy way to consider the topography from a relevant discretization of the homogeneous shallow–water equation (2). The reader is referred to [15] where similar ideas were introduced.

3 Main properties

First of all, let us remark that the scheme (5) is obviously consistent with (4). Since (4) turns out to be a non conservative formulation of (3), or equivalently (1), after the work by Dal Maso, LeFLoch and Murat [8] (see also [3,7,14]), this may suggest that our approach cannot deal with weak solutions of (1). As a consequence, we first prove the consistency of (5) with (1).

Lemma 1. *Let* $(w_i^{n+1})_{i\in\mathbb{Z}}$ *be given by (5)-(6)-(7). Then* $(w_i^{n+1})_{i\in\mathbb{Z}}$ *satisfies in addition:*

$$h_i^{n+1} = h_i^n - \frac{\Delta t}{\Delta x}\left(X_{i+1/2}^n f_h^{\Delta x}(W_i^n, W_{i+1}^n) - X_{i-1/2}^n f_h^{\Delta x}(W_{i-1}^n, W_i^n)\right), \tag{8}$$

$$\begin{aligned}(hu)_i^{n+1} = (hu)_i^n &- \frac{\Delta t}{\Delta x}\left(X_{i+1/2}^n f_{hu}^{\Delta x}(W_i^n, W_{i+1}^n) - X_{i-1/2}^n f_{hu}^{\Delta x}(W_{i-1}^n, W_i^n)\right) \\ &+ \frac{\Delta t}{\Delta x}\frac{g}{2}\left(h_{i+1/2}^n z_{i+1/2}^n - h_{i-1/2}^n z_{i-1/2}^n\right) \\ &- \frac{\Delta t}{\Delta x}\frac{g}{2}(h_{i+1/2}^n + h_{i-1/2}^n)(z_{i+1/2}^n + z_{i-1/2}^n),\end{aligned} \tag{9}$$

where $h^n_{i+1/2} = H^n_{i-1/2}X^n_{i-1/2}$ and $z^n_{i+1/2} = H^n_{i-1/2}(1-X^n_{i-1/2})$. As a consequence, the scheme (5) is consistent with (3) and thus with (1).

We skip the proof of this result since it is just a reformulation of (5).

At this level, the adopted scheme turns out to be relevant when approximating the weak solutions of (1). Moreover, let us note that the above scheme derivation does not need some smoothness argument concerning the topography function z. Now, let us state our main result.

Theorem 1. *Let w^n_i belongs to Ω for all i in $\mathbb{Z}$. Under a suitable CFL like restriction, let us assume that the numerical flux function $f^{\Delta x}$ is Ω–preserving as follows:*

$$w^n_i - \frac{\Delta t}{\Delta x}\left(f^{\Delta x}(w^n_i, w^n_{i+1}) - f^{\Delta x}(w^n_{i-1}, w^n_i)\right) \in \Omega.$$

1. *Assume an additional CFL restriction given by*

$$\frac{\Delta t}{\Delta x}\left(\max(0, f^{\Delta x}_h(W^n_i, W^n_{i+1})) - \min(0, f^{\Delta x}_h(W^n_{i-1}, W^n_i))\right) < H^n_i,$$

then w^{n+1}_i given by (5) stays in Ω for all i in $\mathbb{Z}$.
2. *The scheme (5) is well–balanced. Assume $u^n_i = 0$ and $h^n_i + z_i = H$ a positive constant, then $u^{n+1}_i = 0$ and $h^{n+1}_i + z_i = H$ for all i in $\mathbb{Z}$.*

We here do not prove the above result but we just establish the preservation of the lake at rest. Let us assume $u^n_i = 0$ and $h^n_i + z_i = H > 0$ for all $i \in \mathbb{Z}$. As a consequence, we have $W^n_i = (H, 0)^T$ for all $i \in \mathbb{Z}$. Arguing the consistency of the numerical flux function $f^{\Delta x}$, we have

$$f^{\Delta x}(W^n_i, W^n_{i+1}) = f(W) = \begin{pmatrix} 0 \\ g\dfrac{H^2}{2} \end{pmatrix}, \qquad \forall i \in \mathbb{Z}.$$

Concerning the water heigh, we immediately deduce $h^{n+1}_i = h^n_i$ for all $i \in \mathbb{Z}$ to obtain $h^{n+1}_i + z_i = h^n_i + z_i = H$. Now, let us rewrite the discretization of the discharge:

$$\begin{aligned}(hu)^{n+1}_i =&(hu)^n_i - \frac{\Delta t}{\Delta x}\left(X^n_{i+1/2}g\frac{H^2}{2} - X^n_{i-1/2}g\frac{H^2}{2}\right)\\ &+ \frac{\Delta t}{\Delta x}\frac{g}{2}H^n_{i+1/2}H^n_{i-1/2}(X^n_{i+1/2} - X_{i-1/2}).\end{aligned}$$

Since $H^n_{i+1/2} = H$ for all $i \in \mathbb{Z}$, we get $(hu)^{n+1}_i = 0$. The scheme is thus well–balanced.

4 Numerical results

The numerical experiments are performed on a grid made of 500 cells. Here, the CFL number is systematically fixed to 0.5. To validate the derived numerical scheme, we first propose to consider transcritical flow with shock over a bump. Figure 1 shows a comparison between the classic hydrostatic reconstruction and our scheme. From now on, let us underline that the scheme formulation (5) depends on the choice of the bottom origin. Hence, the comparison is performed by involving three distinct origins. Since the consistency property is not modified, the approximated solutions stay similar. In Fig. 2, we give the approximation obtained when considering a numerical flux function of HLLC type, VFRoe type and Lax–Friedrichs type. In Fig. 3 we present a comparison between first and second order schemes. Here, a MUSCL second order extension has been performed. The second test concerns the known hydrostatic reconstruction failure. Here, the topography is made of a constant slope. As the slope increases, the expected water height must decrease. Because of the cut–off involved into the hydrostatic reconstruction, a wrong water behavior is noted. The same simulation obtained with the hydrostatic upwind scheme gives the required water behavior, see Fig. 4.

The last experiment, presented in Fig. 5, concerns a drain on a non flat bottom in order to simulate dry areas. Once again we obtain an excellent behavior of the derived numerical scheme. As a perspective of the derived 1D technique, we must now propose an extension for 2D unstructured meshes. To address such an issue, arguing the rotational invariance of the model, we should write the 2D formulation

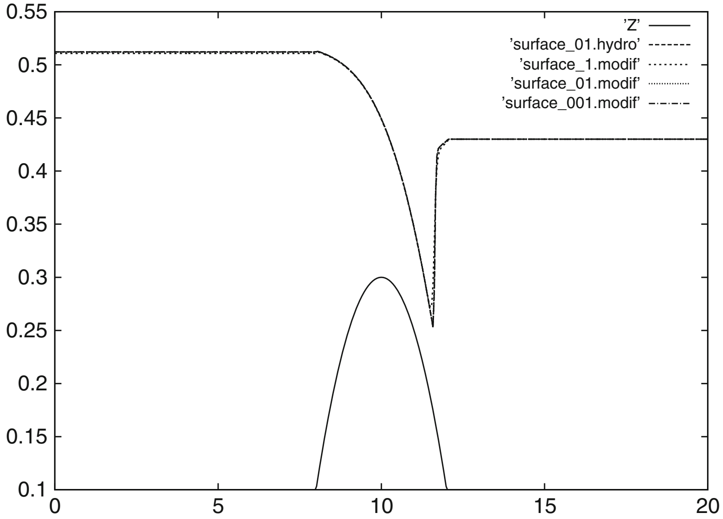


Fig. 1 Comparison hydrostatic reconsturction/hydrostatic upwind involving an arbitrary topography origin so that min(z) = 1, 0.1 and 0.001

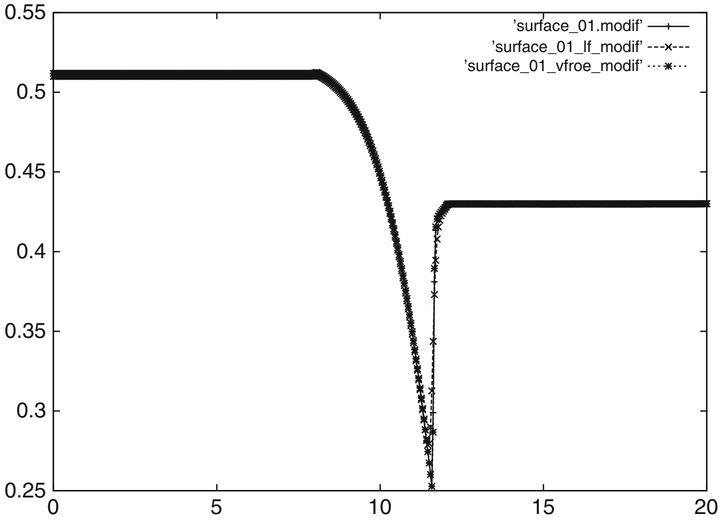


Fig. 2 Comparison between HLLC, VFRoe and Lax–Friedrichs flux functions

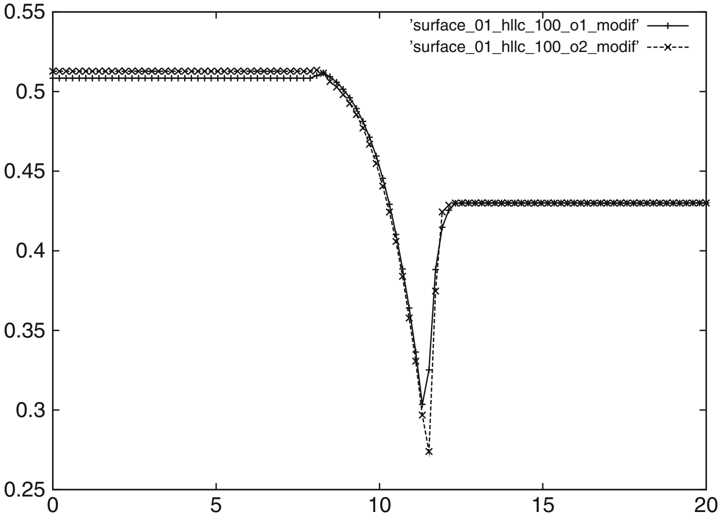


Fig. 3 Comparison first and second order

associated with (4) in the x–direction to obtain the required flux approximation per edge. The interest of such an extension should be an easy topography discretization to obtain 2D well–balanced schemes.

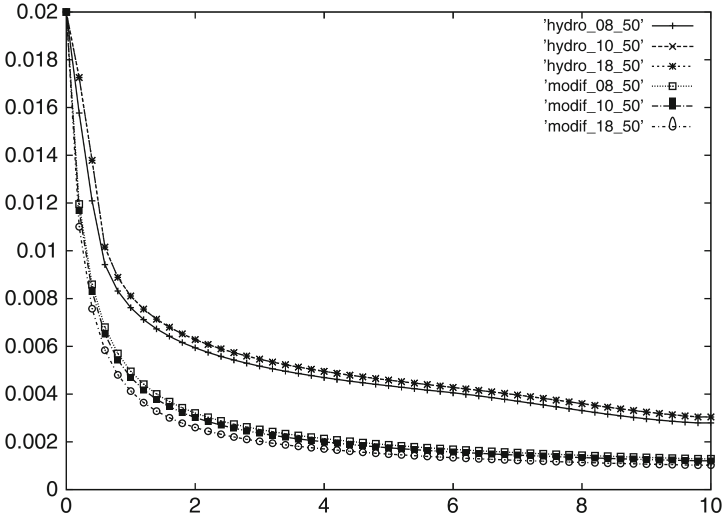


Fig. 4 Failure of the hydrostatic reconstruction. Wrong behavior of h for 8%, 10% and 18% topography slopes. Because of the cut–off the water height increase. Moreover, with 10% and 18% slope, h exactly coincides and thus the obtained approximation is non–physical

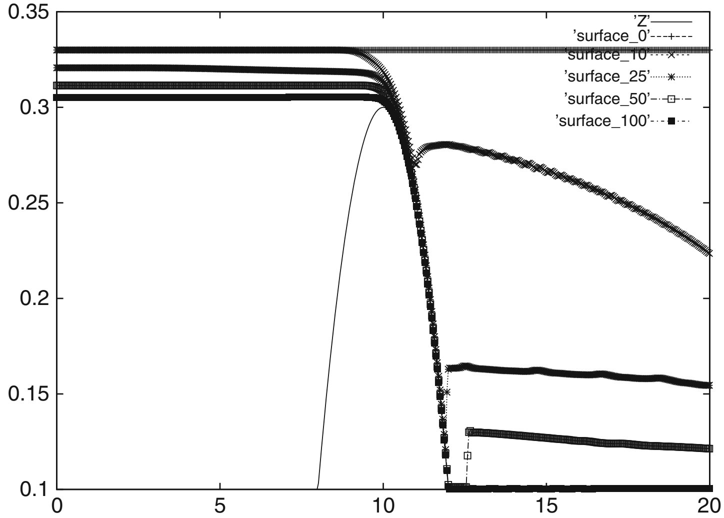


Fig. 5 Drain on a nonflat bottom: Water height at time $t = 0, 1, 10, 25, 50$ and 100

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.Comp., **25**, 2050–2065 (2004).

2. Bermudez A., Vazquez–Cendon M.E., Upwind Methods for Hyperbolic Conservation Laws with Source Terms, Computers and Fluids. **23** 1049–1071 (1994).
3. Berthon C., Coquel F.: Nonlinear projection methods for multi–entropies Navier–Stokes systems, Math. Comput., **76**, 1163–1194 (2007).
4. Berthon C., Dubois J., Dubroca B., Nguyen–Bui T.H., Turpault R.: A Free Streaming Contact Preserving Scheme for the M1 Model, Adv. Appl. Math. Mech., **3**, 259–285 (2010).
5. Berthon C., Marche F.: A positive well–balanced VFRoe–ncv scheme for non–homogeneous shallow–water equations, ISCM–EPMESC proceedings, AIP Conference Proceedings, 1495–1500 (2010).
6. Bouchut F.: Non–linear stability of finite volume methods for hyperbolic conservation laws and well–balanced schemes for sources, Frontiers in Mathematics, Birkhauser, 2004.
7. Castro M., LeFloch P., Munoz–Ruiz M.L., Parés C.: Why many theories of shock waves are necessary: convergence error in formally path–consistent schemes, J. Comput. Phys. **227**, 8107-8129 (2008).
8. Dal Maso G., LeFloch P., Murat F., Definition and weak stability of a non conservative product, J. Math. Pures Appl., **74**, 483–548 (1995).
9. Delestre O., Simulation du ruissellement d'eau de pluie sur des surfaces agricoles, PhD Thesis, University of Orléans, http://tel.archives-ouvertes.fr/tel-00531377/en (2010).
10. Gallouët T., Hérard J.M., Seguin N.: Some recent Finite Volume schemes to compute Euler equations using real gas EOS, Int J. Num. Meth. Fluids, **39**, 1073–1138 (2002).
11. Gallouet T., Hérard J.M., Seguin N.: Some approximate Godunov schemes to compute shallow–water equations with topography, Computers and Fluids, **32**, 479–513 (2003).
12. Gallouet T., Hérard J.M., Seguin N.: On the use of some symetrizing variables to deal with vacuum, Calcolo, **40**, 163–194 (2003).
13. Greenberg J.M., Leroux A.Y.: A well–balanced scheme for the numerical processing of source terms in hyperbolic equations, SIAM J. Numer. Anal., **33**, 1–16 (1996).
14. Hou T.Y., LeFloch P.: Why non–conservative schemes converge to wrong solutions: error analysis, Math. Comput., **206**, pp. 497–530 (1994).
15. Jin S.: A steady–state capturing method for hyperbolic systems with geometrical source terms, M2AN Math. Model. Numer. Anal., **35**, 631–645 (2001).
16. Jin S., Wen X.: Two interface–type numerical methods for computing hyperbolic systems with geometrical source terms having concentrations, SIAM J. Sci. Comput., **26**, 2079–2101 (2005).
17. Jin S., Wen X.: An efficient method for computing hyperbolic systems with geometrical source terms having concentrations, Special issue dedicated to the 70th birthday of Professor Zhong–Ci Shi. J. Comput. Math., **22**, 230–249 (2004).
18. Marche F., Bonneton P., Fabrie P., Seguin N.: Evaluation of well–balanced bore–capturing schemes for 2D wetting and drying processes, Int. J. Numer. Meth. Fluids, **53**, 867–894 (2007).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Finite Volumes Asymptotic Preserving Schemes for Systems of Conservation Laws with Stiff Source Terms

C. Berthon and R. Turpault

Abstract We consider here a numerical technique that allows to build asymptotic-preserving schemes for hyperbolic systems of conservation laws with eventually stiff source terms. The scheme is build in 1D and extended to unstructured 2D meshes. Its behavior is illustrated by numerical experiments on different physical applications.

Keywords systems of conservation laws, stiff, source terms, asymptotic preserving schemes, finite volumes schemes
MSC2010: 65N08, 65Z05

1 Introduction

Our objective is to develop numerical schemes adapted to the resolution of hyperbolic systems of conservation laws with source terms of the form:

$$\partial_t U + \operatorname{div}(\mathbf{F}(U)) = -\gamma R(U), \tag{1}$$

where the state vector $U \in \mathbb{R}^N$ lies in a convex set $\Omega \subset \mathbb{R}^N$. Here, $\gamma \in \mathbb{R}$, which may be a function of U, controls the stiffness of the source term. The function $R(U)$ is supposed to fulfill the compatibility properties required in [2] (see also [10]). In particular, we assume the existence of a constant $n \times N$ matrix Q with rank $n < N$ such that $QR(U) = 0$. It has been showed in [2] that when γ is large, the long-time behavior of such systems degenerates into a nonlinear parabolic system which can be written as:

C. Berthon and R. Turpault
Université de Nantes, Laboratoire de Mathématiques Jean Leray, 2, rue de la Houssinière 44322 Nantes, e-mail: christophe.berthon@univ-nantes.fr, rodolphe.turpault@univ-nantes.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_12, © Springer-Verlag Berlin Heidelberg 2011

$$\partial_t u = -\mathrm{div}\left(\mathscr{M}(u)\nabla u\right), \tag{2}$$

where $u = QU$ and $\mathscr{M}(u)$ is a nonlinear diffusion matrix.

Such systems are involved in numerous physical models found for instance in radiotherapy, radiative transfer or fluid dynamics with friction. Typical applications may involve domains where the source term is neglectable (hyperbolic-dominant zones), very stiff (diffusion-dominant zones) or in-between. Therefore, it is crucial to dispose of a numerical scheme able to handle every regime. The construction of such schemes is generally very difficult. Former works (see for instance [1, 6, 8, 9, 12]) usually concentrate on modifying the HLL scheme [16] to adequately include the source term with respect to the physics of a given problem.

In this article, we will propose a generic numerical technique which extends any approximate Riemann solver into an asymptotic preserving scheme for (1). We will first introduce the construction of a finite volumes scheme adapted to the approximation of the solutions of (1) in 1D. This scheme will then be extended for 2D unstructured meshes. Finally, it will be applied on three numerical simulations that will emphasize the relevance of this numerical technique and underline a possibility to improve it.

2 Description of the Scheme

2.1 Construction in 1D

We first show the construction of the numerical technique as it was introduced in [3] and extended in [2]. It consists in a suitable modification of an approximate Riemann solver designed for the transport part of (1) (ie $\gamma = 0$).

Therefore, we start by selecting such a solver. A Riemann problem is thus approximated at each cell interface:

$$\tilde{U}_{\mathscr{R}}(\frac{x}{t}; U_L, U_R) = \begin{cases} U_L & \text{if } \dfrac{x}{t} < b^-, \\[2ex] \tilde{U}^\star & \text{if } b^- < \dfrac{x}{t} < b^+, \\[2ex] U_R & \text{if } \dfrac{x}{t} > b^+, \end{cases} \tag{3}$$

where $|b^\pm|$ are chosen to be larger than the fastest wave speed of the problem. For the sake of simplicity in the notations, we will consider in the following that $b^+ = -b^- = b > 0$. Furthermore, $\tilde{U}^\star$ represents the value of the intermediate states and hence generally depends on U_L, U_R and x/t.

As soon as the CFL condition $b\frac{\Delta t}{\Delta x} \leq \frac{1}{2}$ holds, we are considering a juxtaposition of non-interacting approximate Riemann solvers denoted $\tilde{U}^n_{\Delta x}(x, t^n + t)$ for $t \in [0, \Delta t)$. The updated approximated solution at time t^{n+1} is then naturally defined as follows:

$$\tilde{U}_i^{n+1} = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} \tilde{U}_{\Delta x}^n(x, t^n + \Delta t)dx. \tag{4}$$

This scheme can be written in the following usual conservation form:

$$\tilde{U}_i^{n+1} = U_i^n - \frac{\Delta t}{\Delta x}(\mathscr{F}_{i+1/2} - \mathscr{F}_{i-1/2}), \tag{5}$$

where $\mathscr{F}_{i+1/2}$ denotes the numerical flux at the interface $x_{i+1/2}$. Any (approximate) Riemann solver enter this framework, including for instance Godunov, HLL, HLLC and relaxation schemes. As an example, in the case of the well-known HLL scheme [16], $\tilde{U}^\star$ and $\mathscr{F}_{i+1/2}$ are given by:

$$\tilde{U}^{\star,HLL} = \frac{1}{2}(U_L + U_R) - \frac{1}{2b}(F(U_R) - F(U_L)), \tag{6}$$

$$\mathscr{F}_{i+1/2} = \frac{1}{2}(F(U_i^n) + F(U_{i+1}^n)) - \frac{b}{2}(U_{i+1}^n - U_i^n). \tag{7}$$

In order to take into account the source term, we now modify the approximate Riemann solver (3) as follows:

$$U_{\mathscr{R}}(\frac{x}{t}; U_L, U_R) = \begin{cases} U_L & \text{if } \frac{x}{t} < -b, \\ U^{\star L} & \text{if } -b < \frac{x}{t} < 0, \\ U^{\star R} & \text{if } 0 < \frac{x}{t} < b, \\ U_R & \text{if } \frac{x}{t} > b, \end{cases} \tag{8}$$

where we have set:

$$\begin{aligned} U^{\star L} &= \underline{\alpha}\tilde{U}^\star + (\mathbb{I}_{\mathrm{d}} - \underline{\alpha})(U_L - \bar{R}(U_L)), \\ U^{\star R} &= \underline{\alpha}\tilde{U}^\star + (\mathbb{I}_{\mathrm{d}} - \underline{\alpha})(U_R - \bar{R}(U_R)). \end{aligned} \tag{9}$$

Here, $\underline{\alpha}$, which denotes a $N \times N$ matrix, and $\bar{R}(U)$ are defined by:

$$\underline{\alpha} = \left(\mathbb{I}_{\mathrm{d}} + \frac{\gamma\Delta x}{2b}(\mathbb{I}_{\mathrm{d}} + \underline{\sigma})\right)^{-1}, \ \bar{R}(U) = (\mathbb{I}_{\mathrm{d}} + \underline{\sigma})^{-1}R(U). \tag{10}$$

The $N \times N$ matrices $\mathbb{I}_{\mathrm{d}}$ and $\underline{\sigma}$ respectively denote the identity matrix and a parameter matrix to be defined. The updated approximated solution at time t^{n+1} is once again naturally defined:

$$U_i^{n+1} = \frac{1}{\Delta x}\int_{x_{i-1/2}}^{x_{i+1/2}} U_{\Delta x}^n(x, t^n + \Delta t)dx. \tag{11}$$

A straightforward computation leads to:

$$\begin{aligned}\frac{1}{\Delta t}(U_i^{n+1} - U_i^n)+&\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2}\mathscr{F}_{i+1/2} - \underline{\alpha}_{i-1/2}\mathscr{F}_{i-1/2})\\ &=\frac{1}{\Delta x}(\underline{\alpha}_{i+1/2} - \underline{\alpha}_{i-1/2})F(U_i^n) - \frac{\gamma}{2}(\underline{\alpha}_{i+1/2} + \underline{\alpha}_{i-1/2})R(U_i^n).\end{aligned} \tag{12}$$

Observe that whenever $\gamma = 0$, then $\underline{\alpha} = \mathbb{I}_{\mathrm{d}}$ and (12) is nothing but (5).

It was proved in [2] that the scheme (12) is consistant with (1) and preserves Ω as soon as the approximate Riemann solver for the transport part does so. These properties hold for any relevant choice of the parameter matrices $\underline{\sigma}$. These matrices may therefore be chosen to enforce the scheme (12) to be consistant with (2) in the asymptotic regimes. Indeed, an asymptotic analysis of the scheme shows that it is asymptotic preserving if $\underline{\sigma}_{i+1/2}$ is chosen so that the following relation holds:

$$Q(\mathbb{I}_{\mathrm{d}} + \underline{\sigma}_{i+1/2})^{-1} = \frac{1}{b^2}\mathscr{M}_{i+1/2}Q, \tag{13}$$

where $\mathscr{M}_{i+1/2}$ is a discretization of the diffusion matrix $\mathscr{M}(u)$ at the interface $x_{i+1/2}$. One of the edges of this scheme is that it allows to consider applications where γ is a nonlinear function of x and U (see examples in [3] and [4]).

2.2 Extension for 2D unstructured grids

In the case of unstructured grids, the 1D scheme (12) can be extended into the following scheme:

$$\begin{aligned}U_K^{n+1} = U_K^n - &\frac{\Delta t}{|C_K|}\sum_{e\in\partial K}|e|\underline{\alpha_e}\Big[\mathscr{F}_e.n_e - F(U_K^n)n_x - G(U_K^n)n_y\Big]\\ &+\frac{c\Delta t}{|C_K|}\sum_{e\in\partial K}|e|\beta_e b_e(\mathbb{I}_{\mathrm{d}} - \underline{\alpha_e})\bar{R}(U_K^n),\end{aligned} \tag{14}$$

where $|K|$ is the measure of the cell K and $|e|$ is the measure of the interface e.

Furthermore, $\underline{\alpha}_e$ is chosen as:

$$\underline{\alpha}_e = |e|\left(|e|\mathbb{I}_{\mathrm{d}} + \frac{\gamma|K|}{2b}(\mathbb{I}_{\mathrm{d}} + \underline{\sigma})\right)^{-1},$$

Finally, β is set to $1/2$. It is to note that the choices of $\underline{\alpha}$ and β are the simplest admissible ones. However, they are not unique and other expressions may even improve the accuracy of the scheme.

This scheme has successfully been used in the case of cartesian grids in 2D (see for example [5]). In the case of unstructured grids however, in order to enforce the asymptotic preserveness of (14), the choice of $\underline{\sigma}_e$ implies the knowledge of a relevant scheme for the diffusion equation (2). Due to the nonlinear nature of the anisotropy of the diffusion matrix $\mathscr{M}(u)$, the classical two-point scheme (aka FV4, see [15]) lacks of consistance. Therefore, efficient compact schemes have to be considered in order to discretize the diffusion operator. In this framework, we are considering Discrete-Duality Finite Volumes schemes (see for instance [7, 11, 13, 17]). The rich structure of the DDFV schemes can obviously also be used to improve the hyperbolic solvers.

3 Numerical Results

In this section, numerical examples illustrate the behavior of the scheme (12) on three different test-cases. For the sake of simplicity, we used the HLL solver for the transport part.

TC1: Euler with friction

We first consider the 1D isentropic Euler equations with friction. The system reads:

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

$$\partial_t q + \partial_x \Big(\frac{q^2}{\rho} + p(\rho)\Big) = -\kappa q,$$

where $\rho > 0$ denotes the density and $q \in \mathbb{R}$ is the fluid momentum. The pressure function $p : \mathbb{R}_+ \to \mathbb{R}_+$ is assumed to be regular enough and to satisfy $p'(\rho) > 0$ in order to ensure the first-order homogeneous associated system to be hyperbolic.

The associated diffusive regime is governed by:

$$\partial_t \rho = \partial_x (p^{'}(\rho)\partial_x \rho). \tag{15}$$

Figure 1 shows the density computed at time $t = 20$. The reference solution is a grid-converged result with a scheme that approximates the diffusion equation (15).

The results of the scheme (12) are in very good agreement with the reference solution even on a coarse grid ($\Delta x = 0.02$). The results of the scheme with $\underline{\sigma} = 0$ are also plotted on Fig. 1. They are representative of what happens with a scheme which is not asymptotic-preserving (although consistant). Indeed, an asymptotic analysis of this scheme shows that it is consistant with a diffusion equation with the wrong diffusion coefficient (see [3]).

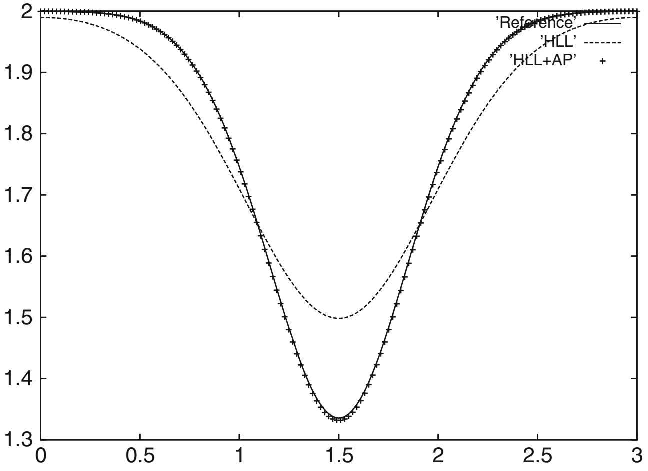


Fig. 1 TC1: computed values of ρ at time $t = 20$. Reference solution (full line) and HLL scheme with (+) or without (dashed line) AP correction

TC2: M1 model for radiative transfer

Now we are interested in the 2D $M1$ model for radiative transfer:

$$\begin{aligned}\partial_t E + \partial_x \mathbf{F}_x + \partial_y \mathbf{F}_y &= c\sigma(aT^4 - E),\\ \partial_t \mathbf{F}_x + c^2\partial_x \mathbf{P}_{xx} + c^2\partial_y \mathbf{P}_{xy} &= -c\sigma\mathbf{F}_x,\\ \partial_t \mathbf{F}_y + c^2\partial_y \mathbf{P}_{xy} + c^2\partial_y \mathbf{P}_{yy} &= -c\sigma\mathbf{F}_y,\\ \rho C_v \partial_t T &= c\sigma(E - aT^4),\end{aligned}$$

where E, $\mathbf{F}$ and $\mathbf{P}$ respectively denote the radiative energy, the radiative flux vector and the radiative pressure tensor. Moreover, T is the material temperature, σ is the opacity, a and c are physical parameters. Finally $\mathbf{P} = \mathbf{P}(\frac{\|\mathbf{F}\|}{cE})$ is a prescribed function (see [14]).

The associated asymptotic regime is described by the so-called equilibrium diffusion equation:

$$\partial_t(\rho C_v T + aT^4) + \operatorname{div}\left(\frac{4acT^3}{3\sigma}\nabla T\right) = 0.$$

In order to obtain a scheme which is consistant with the diffusion operator, unknowns on the triangular mesh were considered at the orthocenter and the classical FV4 scheme (see [15]) was used. Of course, this trick is not valid in general so that other approaches have to be considered as was mentionned in Sect. 2.

Figure 2 shows the results of the scheme (14) on a left-entering Marshak wave inside a square 1m-wide domain with an obstacle. The parameters are $E(t = 0) =$

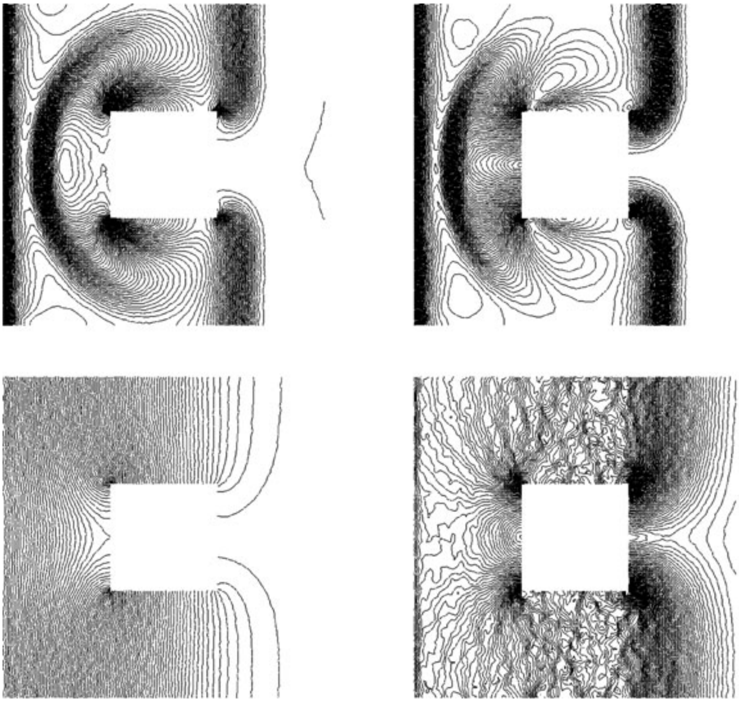


Fig. 2 TC2: Radiative energy (l) and normalized flux (r). Top: $t = 1.e - 8$ and $\sigma = 0$. Bottom: $t = 1.e - 5$ and $\sigma = 10$. Same contours for the energy, same number of contours for the flux (max$\simeq$ 0.8 (T) and 0.1 (B)). Triangular mesh with $h \simeq 6.5e - 3$

$a1000^4$, $F(t = 0) = (0, 0)$, $T(t = 0) = 1000$, $E_L = a2000^4$, $F_L = (0, 0)$ and $T_L = 2000$. Two computations were carried on with $\sigma = 0$ and $\sigma = 10$.

TC3: toy model

For this last application, we consider an interesting toy model that is one of the simplest nontrivial example where the asymptotic regime is described by a system of two equations. It writes:

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

$$\partial_t q + \partial_x \Big(\frac{q^2}{\rho} + p(\rho)\Big) = -\kappa q + \sigma f,$$

$$\partial_t e + \partial_x f = 0,$$

$$\partial_t f + \partial_x \chi \left(\frac{f}{e}\right) e = -\sigma f,$$

where $\chi(\xi) = \frac{3+4\xi^2}{5+2\sqrt{4-3\xi^2}}$.

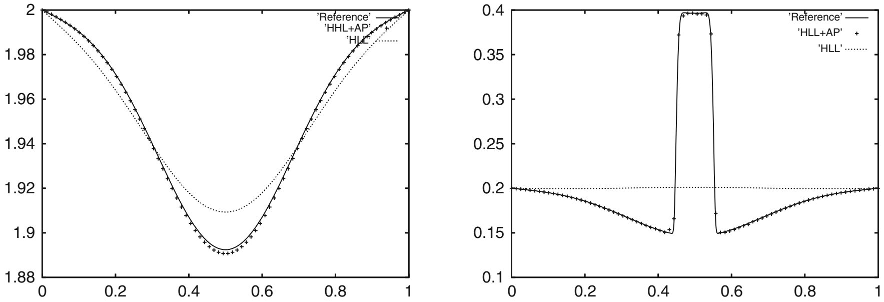


Fig. 3 TC3: computed values of e (l) and ρ (r) at time $t = 50$. Reference solution (full line) and HLL scheme with (+) or without (dashed line) AP correction

The asymptotic regime of this system is given by:

$$\begin{aligned} \partial_t \rho - \frac{1}{\kappa}\partial_x^2 p(\rho) - \frac{1}{3\kappa}\partial_x^2 e &= 0, \\ \partial_t e - \frac{1}{3\sigma}\partial_x^2 e &= 0. \end{aligned} \tag{16}$$

Figure 3 shows the results of scheme (12) at time $t = 50$ compared with a reference solution for the following test-case: the initial values are $\rho(t = 0) = 0.2$, $q(t = 0) = f(t = 0) = 0$ and $e(t = 0) = 2 - 0.5 \times 1_{[0.45;0.55]}$. The other parameters are $\kappa = 2000$, $\sigma = 1000$, $p(\rho) = 10^{-3}\rho^2$.

With these parameters, the solution is governed by the asymptotic system (16). The results given by the AP preserving scheme (12) are in excellent agreement with the reference solution, even on a very coarse grid (only 80 points where used). It is to note that this test-case is very challenging and that a scheme which does not preserve the asymptotics gives poor results here. As a illustration, the results given by the choice $\underline{\sigma} = 0$ are also showed on Fig. 3.

References

1. Berthon C., Charrier P., Dubroca B.: An HLLC Scheme to Solve the M_1 Model of Radiative Transfer in Two Space Dimensions, J. Scie. Comput., J. Sci. Comput., **31** 3, 347-389 (2007).
2. Berthon C., LeFloch P., Turpault, R.: Late-time relaxation limits of nonlinear hyperbolic systems. A general framework. (2010) Available via arXiv. http://arxiv.org/abs/1011.3366
3. Berthon C., Turpault, R.: Asymptotic-preseverving HLL schemes. Numerical Methods for Partial Differential Equations, (2010) doi:10.1002/num.20586
4. Berthon C., Turpault, R.: A numerical correction of the $M1$-model in the diffusive limit. NMCF09 proceedings (2009).
5. Berthon C., Dubois J., Dubroca B., Nguyen Bui T.H., Turpault R.: A Free Streaming Contact Preserving Scheme for the M1 Model, Adv. Appl. Math. Mech., 3 (2010), 259-285.

6. Bouchut F., Ounaissa H., Perthame B.: Upwinding of the source term at interfaces for Euler equations with high friction, J. Comput. Math. Appl. **53**, No. 3-4, 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, 30323070 (2008).
8. Buet C., Cordier S.: An asymptotic preserving scheme for hydrodynamics radiative transfer models: numerics for radiative transfer, Numer. Math. **108**, 199–221 (2007).
9. Buet C., Després B.: Asymptotic preserving and positive schemes for radiation hydrodynamics, J. Comput. Phys. **215**, 717–740 (2006).
10. Chen G.Q., Levermore C.D., Liu T.P.: Hyperbolic Conservation Laws with Stiff Relaxation Terms and Entropy, Comm. Pure Appl. Math. **47**, 787–830 (1995).
11. Coudière Y., Manzini G.: The discrete duality finite volume method for convection-diffusion problems, SIAM J. Numer. Anal. **47**, 6, 41634192 (2010).
12. Degond P., Deluzet F., Sangam A., Vignal M.H.: An Asymptotic Preserving scheme for the Euler equations in a strong magnetic field, J. Comput. Phys. **228**, 3540–3558 (2009).
13. Domelevo K., Omnes P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids., M2AN Math. Model. Numer. Anal. **39**, no. 6, 12031249 (2005).
14. Dubroca B., Feugeas J.L: Entropic Moment Closure Hierarchy for the Radiative Transfer Equation, C. R. Acad. Sci. Paris, Ser. I, **329**, 915–920 (1999).
15. Eymard R., Gallouët T., Herbin R.: Finite Volume Methods, Handbook of Numerical Analysis, Vol. VII, 713-1020 (2000).
16. Harten A., Lax P.D., Van Leer B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws, SIAM Review, **25**, 35–61 (1983).
17. Hermeline F.: Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes, Comput. Methods Appl. Mech. Engrg., **192**, 1939–1959 (2003).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Development of DDFV Methods for the Euler Equations

Christophe Berthon, Yves Coudière, and Vivien Desveaux

Abstract We propose to extend some recent gradient reconstruction, the so–called DDFV approaches, to derive accurate finite volume schemes to approximate the weak solutions of the 2D Euler equations. A particular attention is paid on the limitation procedure to enforce the required robustness property. Some numerical experiments are performed to highlight the relevance of the suggested MUSCL–DDFV technique.

Keywords Finite volume methods for hyperbolic problems, Euler equations, DDFV reconstruction, MUSCL reconstruction, Robustness
MSC2010: 65M08, 65N12, 76N99

1 Introduction

This work is devoted to the numerical approximation of the 2–D Euler equations, given as follows:

$$\partial_t \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ E \end{bmatrix} + \partial_x \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ u(E+p) \end{bmatrix} + \partial_y \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ v(E+p) \end{bmatrix} = 0, \tag{1}$$

where $\rho > 0$ denotes the density, $(u, v) \in \mathbb{R}^2$ the velocity vector and $E > 0$ the total energy. For the sake of the simplicity in the presentation, the pressure is given by

Christophe Berthon, Yves Coudière, and Vivien Desveaux
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, Yves.Coudiere@univ-nantes.fr, Vivien.Desveaux@univ-nantes.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_13, © Springer-Verlag Berlin Heidelberg 2011

the perfect gas law $p = (\gamma - 1)\left[E - \frac{\rho}{2}(u^2 + v^2)\right]$. The forthcoming developments will easily extend to general pressure laws. To shorten the notations, the system can be rewritten as follows:

$$\partial_t W + \partial_x f(W) + \partial_y g(W) = 0, \tag{2}$$

where $W = {}^t(\rho, \rho u, \rho v, E) : \mathbb{R}^2 \times \mathbb{R}^+ \to \Omega$ is the unknown state vector and $f(W) : \Omega \to \mathbb{R}^4$ and $g(W) : \Omega \to \mathbb{R}^4$ are the flux functions which find clear definitions. The convex set of admissible states is defined by:

$$\Omega = \left\{ W \in \mathbb{R}^4; \rho > 0, (u, v) \in \mathbb{R}^2, E - \frac{\rho}{2}\left(u^2 + v^2\right) > 0 \right\}. \tag{3}$$

When approximating (1), several strategies have been proposed to increase the accuracy of the numerical solutions among which the most popular is certainly the MUSCL scheme (for instance see [12, 13, 15, 16]). This scheme extends any first–order scheme into a second–order approximation using a piecewise linear reconstruction. In the 2–D case, the main difficulty is to find a technique to reconstruct gradients that can be extended to unstructured meshes (see [4]).

The DDFV (Discrete Duality Finite Volume) method was introduced in the field of elliptic equations in order to reconstruct gradients on distorted meshes (see [1, 6, 9, 10]). The idea of this method is to combine two distinct finite volume schemes on two overlapping meshes: the primal mesh and the dual mesh whose cells are built around the vertices of the primal mesh. This process adds new numerical unknowns at the vertices of the primal mesh, but it will allow to reconstruct very accurate gradients.

It was first proposed to take advantage of the DDFV gradient in order to built second order schemes for the linear convection–diffusion equation in [5]. In this paper, new values of the unknown are built at the midpoint of the interfaces by mean of some averages of the DDFV gradient. The resulting scheme is proved to be of second order in the diffusive regime.

The aim of this work is to extend DDFV–like methods to the case of the Euler equations. As a first step, we have only developed such a method on structured meshes in order to simplify the computation and to check its efficiency. On unstructured meshes, the extension of the DDFV gradient is straightforward. Our reconstruction and limitation procedures generalize although being more technical. Note that the vertices of the primal cells do not coincide with the center of gravity of the dual cells. It might influence the accuracy of the method and some alternatives will be considered in future work.

The paper is organized as follows. In Sect. 2, we introduce the dual mesh and we describe the reconstruction process and the limitation process of our scheme. Section 3 concerns the robustness of our scheme. Indeed, with most of first–order schemes, if a numerical solution is initially valued in Ω, then it remains in Ω. Such a property must be preserved by the second–order accurate scheme. Section 4 is devoted to numerical experiments to illustrate the relevance of DDFV approach when evaluating second–order reconstructions. We give some conclusions and future developments in Sect. 5.

2 Presentation of the scheme

First let us introduce the main notations. We consider a primal mesh composed of rectangular cells

$$K_{i,j} = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}] \times [y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}], \quad i, j \in \mathbb{Z}. \tag{4}$$

For the sake of simplicity, we will assume that the mesh is uniform, and we enforce $x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}} = y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}} = h$, for all $i, j \in \mathbb{Z}$, where $h > 0$ is fixed.

Let $W_{i,j}^n$ stand for an approximation of the mean value of W on the cell $K_{i,j}$ at time t^n. We denote by $\Delta t > 0$ the time increment. At time $t^{n+1} = t^n + \Delta t$, the updated first–order approximation is given by (see [7, 12, 13]):

$$\begin{aligned} W_{i,j}^{n+1} = W_{i,j}^n - \frac{\Delta t}{h} \Big(& F(W_{i,j}^n, W_{i+1,j}^n) - F(W_{i-1,j}^n, W_{i,j}^n) \\ & + G(W_{i,j}^n, W_{i,j+1}^n) - G(W_{i,j-1}^n, W_{i,j}^n) \Big), \end{aligned} \tag{5}$$

where $F : \Omega \times \Omega \to \mathbb{R}^4$ and $G : \Omega \times \Omega \to \mathbb{R}^4$ are consistent numerical flux functions. In addition, to avoid some instabilities [12, 13], the time step is restricted according to a CFL–like condition given as follows:

$$\frac{\Delta t}{h} \max_{(i,j)\in\mathbb{Z}^2} \left(\left| \lambda_F^{\pm}(W_{i,j}^n, W_{i+1,j}^n) \right|, \left| \lambda_G^{\pm}(W_{i,j}^n, W_{i,j+1}^n \right| \right) \le \frac{1}{4}, \tag{6}$$

where $\lambda_\Phi^{\pm}(W_L, W_R)$ denotes suitable numerical wave velocities associated to the numerical flux function $\Phi(W_L, W_R)$.

2.1 The dual mesh

We denote by $B_{i+\frac{1}{2},j+\frac{1}{2}} = \left(x_{i+\frac{1}{2}}, y_{j+\frac{1}{2}}\right)$ the vertices of the primal mesh and by $B_{i,j} = (x_i, y_j)$ the center of the primal cell $K_{i,j}$. Around each vertex of the primal mesh $B_{i+\frac{1}{2},j+\frac{1}{2}}$, we construct a dual cell $K_{i+\frac{1}{2},j+\frac{1}{2}} = [x_i, x_{i+1}] \times [y_j, y_{j+1}]$. The set of the dual cells $\left(K_{i+\frac{1}{2},j+\frac{1}{2}}\right)_{i,j\in\mathbb{Z}}$ constitutes a second mesh which we call dual mesh. The centers of the dual cells are the vertices of the primal mesh and conversely.

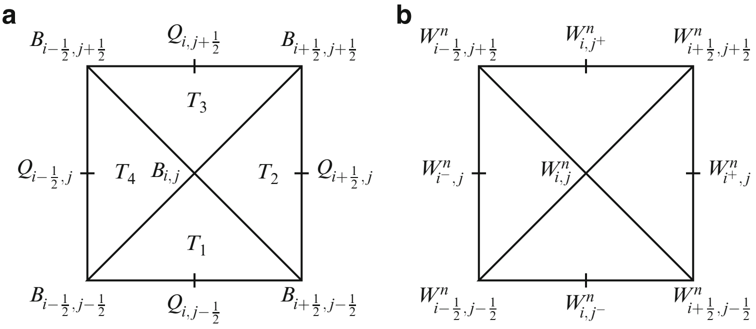

a
$B_{i-\frac{1}{2},j+\frac{1}{2}}$
$Q_{i,j+\frac{1}{2}}$
$B_{i+\frac{1}{2},j+\frac{1}{2}}$
T_3
$Q_{i-\frac{1}{2},j}$
T_4
$B_{i,j}$
T_2
$Q_{i+\frac{1}{2},j}$
T_1
$B_{i-\frac{1}{2},j-\frac{1}{2}}$
$Q_{i,j-\frac{1}{2}}$
$B_{i+\frac{1}{2},j-\frac{1}{2}}$
b
$W^n_{i-\frac{1}{2},j+\frac{1}{2}}$
$W^n_{i,j+}$
$W^n_{i+\frac{1}{2},j+\frac{1}{2}}$
$W^n_{i^-,j}$
$W^n_{i,j}$
$W^n_{i^+,j}$
$W^n_{i-\frac{1}{2},j-\frac{1}{2}}$
$W^n_{i,j-}$
$W^n_{i+\frac{1}{2},j-\frac{1}{2}}$

Fig. 1 (Left) Geometry of the cell $K_{i,j}$. (Right) Location of the known states and of the reconstructed states

At time t^n, we assume known approximations $W^n_{i+\frac{1}{2},j+\frac{1}{2}}$ of the mean values of W on cells $K_{i+\frac{1}{2},j+\frac{1}{2}}$. As a consequence, at time t^n, on each primal or dual cell, we know four approximate values at the vertices and one approximate value at the center (see Fig. 1b).

In the sequel, we will deal simultaneously with primal and dual cells. We thus define the set of the indexes of primal and dual cells $\mathbb{S} = \mathbb{Z}^2 \cup (\mathbb{Z} + \frac{1}{2})^2$. The set of primal and dual cells is then $\{K_{i,j}\}_{(i,j)\in\mathbb{S}}$. For $(i,j) \in \mathbb{S}$, we denote by $Q_{i+\frac{1}{2},j} = (x_{i+\frac{1}{2}}, y_j)$, the middle of the interface between the cells $K_{i,j}$ and $K_{i+1,j}$ and by $Q_{i,j+\frac{1}{2}} = (x_i, y_j + \frac{1}{2})$, the middle of the interface between the cells $K_{i,j}$ and $K_{i,j+1}$ (see Fig. 1a). On each cell $K_{i,j}$ for $(i,j) \in \mathbb{S}$, we reconstruct values $W^n_{i\pm,j}$ and $W^n_{i,j\pm}$ at points $Q_{i\pm\frac{1}{2},j}$ and $Q_{i,j\pm\frac{1}{2}}$ (see Fig. 1b). Arguing these notations, the second order scheme reads as follows:

$$W^{n+1}_{i,j} = W^n_{i,j} - \frac{\Delta t}{h}\Big[F\left(W^n_{i+,j}, W^n_{i+1-,j}\right) - F\left(W^n_{i-1+,j}, W^n_{i-,j}\right) \\ +G\left(W^n_{i,j+}, W^n_{i,j+1-}\right) - G\left(W^n_{i,j-1+}, W^n_{i,j-}\right)\Big]. \quad (7)$$

We now detail the evaluation of $W^n_{i\pm,j}$ and $W^n_{i,j\pm}$. We recall that both the primal and dual unknowns are solutions of a finite volume scheme. The two schemes are coupled through the gradient reconstruction.

2.2 Gradient reconstruction

As a first step, we perform a gradient reconstruction. To address such an issue, we derive a relevant cell splitting. We consider a primal or dual cell $K_{i,j}$, $(i,j) \in \mathbb{S}$. The cell can be decomposed into four triangles using the four vertices and the center.

We denote by T_1 the bottom triangle and the other ones are denoted by T_2, T_3 and T_4, clockwise (see Fig. 1a).

We define a function $\widehat{W} : K_{i,j} \to \mathbb{R}^4$ piecewise linear on the T_l and which coincides with the approximate values at the four vertices and at the center.

Next, we project each coordinate $\widehat{W}_k$ of $\widehat{W}$ on the space of linear function which takes the value ${(W_{i,j})^n}_k$ at the point $B_{i,j}$. This means that for all integers $k \in [1, 4]$, we seek $\mu_k \in \mathbb{R}^2$ which minimizes the functional $E_k(\nu) : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$E_k(\nu) = \int_{K_{i,j}} \left| \widehat{W}_k(X) - \left[\left(W^n_{i,j} \right)_k + \nu \cdot (X - B_{i,j}) \right] \right|^2 dX. \tag{8}$$

Existence and uniqueness of the minimum are immediate since the functional is strictly convex. The numerical computation of the minimum is quite easy since we only need to compute the Jacobian of E_k and to find its zero. For the sake of simplicity in the notations, we denote by $\mu = {}^t(\mu_1, \mu_2, \mu_3, \mu_4)$, the vector of the solutions of these minimization problems. Hence, we define $\widetilde{W}_\mu(X) : K \to \mathbb{R}^4$ the function whose k–th coordinate is $\left(W^n_{i,j} \right)_k + \mu_k \cdot (X - B_{i,j})$.

2.3 Limitation

We assume that the states $W^n_{i,j}$, $(i, j) \in \mathbb{S}$, are in Ω. Let us remark that the reconstructed function $\widetilde{W}_\mu$ does not necessarily remain in Ω. As a consequence, we have to limit the slopes μ_k. To address such an issue, we propose to substitute the slope μ by $\theta\mu$ where $\theta \in [0, 1]$ is a limitation parameter to be fixed according to the required robustness property. To ensure existence and uniqueness of an optimal limited slope, we have to restrict Ω to a close set. We fix a small parameter $\epsilon > 0$ and we define

$$\Omega_\epsilon = \left\{ W \in \mathbb{R}^4; \rho \geq \epsilon, (u, v) \in \mathbb{R}^2, E - \frac{\rho}{2} \left(u^2 + v^2 \right) \geq \epsilon \right\}. \tag{9}$$

Since we need the values of the reconstructed function only at points $B_{i\pm\frac{1}{2},j}$ and $B_{i,j\pm\frac{1}{2}}$, we require $\widetilde{W}_{\theta\mu}(B_{i\pm\frac{1}{2},j}) \in \Omega_\epsilon$ and $\widetilde{W}_{\theta\mu}(B_{i,j\pm\frac{1}{2}}) \in \Omega_\epsilon$. We thus define the optimal slope limiter by

$$\theta = \max \left\{ t \in [0, 1]; \widetilde{W}_{t\mu}(B_{i\pm\frac{1}{2},j}) \in \Omega_\epsilon, \widetilde{W}_{t\mu}(B_{i,j\pm\frac{1}{2}}) \in \Omega_\epsilon \right\}. \tag{10}$$

We emphasize that this set is nonempty since it contains 0. Besides, the maximum is reached because Ω_ϵ is a close set and $t \mapsto \widetilde{W}_{t\mu}(B_{l,m})$ is continuous. Solving for θ requires to find the roots of some quadratic functions (the energy). Finally, the reconstructed states are given by $W^n_{i\pm,j} = \widetilde{W}_{\theta\lambda}(B_{i\pm\frac{1}{2},j})$ and $W^n_{i,j\pm} = \widetilde{W}_{\theta\lambda}(B_{i,j\pm\frac{1}{2}})$.

3 Robustness

We now establish the robustness of the proposed reconstruction. First, let us assume that the directional flux functions F and G are first–order robust on both primal and dual meshes. Indeed, under the CFL condition

$$\frac{\Delta t}{h}\max_{(i,j)\in\mathbb{S}}\left(\left|\lambda_F^{\pm}(W_{i,j}^n, W_{i+1,j}^n)\right|, \left|\lambda_G^{\pm}(W_{i,j}^n, W_{i,j+1}^n)\right|\right) \leq \frac{1}{4}, \tag{11}$$

we assume that the updated states, given by (5) for all pairs (i, j) in $\mathbb{S}$, stay in Ω. Now, let us recall the following statements (for instance see [2,12]) about robustness of the directional numerical flux functions:

Theorem 1. *Let us consider a robust numerical flux Φ. Assume that W_1, W_2 and W_3 are in Ω. Let W_2^- and W_2^+ be two reconstructed states in Ω such that $W_2 = \frac{W_2^- + W_2^+}{2}$. Assume the CFL condition*

$$\frac{\Delta t}{h}\max\left(|\lambda_\Phi^{+}(W_1, W_2^-)|, |\lambda_\Phi^{\pm}(W_2^-, W_2^+)|, |\lambda_\Phi^{-}(W_2^+, W_3)|\right) \leq \frac{1}{4}. \tag{12}$$

Then we have $W_2 - \frac{\Delta t}{h}\left(\Phi(W_2^+, W_3) - \Phi(W_1, W_2^+)\right) \in \Omega$.

We assume that the 1D numerical fluxes F and G are robust. In addition, we assume that the states $W_{i,j}^n$, $(i, j) \in \mathbb{S}$ are in Ω, so that the limitation procedure described in Sect. 2.3 ensures that the reconstructed states $W_{i\pm,j}^n$ and $W_{i,j\pm}^n$, $(i, j) \in \mathbb{S}$, remain in Ω. To shorten the notations, we set

$$\Lambda_F = \max_{(i,j)\in\mathbb{S}}\left(|\lambda_F^{\pm}(W_{i-,j}^n, W_{i+,j}^n)|, |\lambda_F^{\pm}(W_{i+,j}^n, W_{i+1-,j}^n)|\right),$$

$$\Lambda_G = \max_{(i,j)\in\mathbb{S}}\left(|\lambda_G^{\pm}(W_{i,j-}^n, W_{i,j+}^n)|, |\lambda_G^{\pm}(W_{i,j+}^n, W_{i,j+1-}^n)|\right).$$

By applying Theorem 1 we have

$$W_{i,j}^n - \frac{\Delta t}{h}\left[F\left(W_{i+,j}^n, W_{i+1-,j}^n\right) - F\left(W_{i-1+,j}^n, W_{i-,j}^n\right)\right] \in \Omega, \tag{13}$$

as soon as the CFL restriction $\frac{\Delta t}{h}\Lambda_F \leq \frac{1}{4}$ holds, and we get

$$W_{i,j}^n - \frac{\Delta t}{h}\left[G\left(W_{i,j+}^n, W_{i,j+1-}^n\right) - G\left(W_{i,j-1+}^n, W_{i,j-}^n\right)\right] \in \Omega, \tag{14}$$

under the CFL condition $\frac{\Delta t}{h}\Lambda_G \leq \frac{1}{4}$.

Considering half sum of (13) and (14), we finally obtain $W_{i,j}^{n+1} \in \Omega$, for all $(i, j) \in \mathbb{S}$ under the CFL condition [12] $\frac{\Delta t}{h}\max\left(\Lambda_F, \Lambda_G\right) \leq \frac{1}{8}$. The robustness of the proposed numerical method is thus established.

4 Numerical tests

We have chosen two cases from the collection of 2D Riemann problems proposed by [11], namely configuration 3 (p. 594) and 6 (p. 596). They are called case 1 and case 2. These problems are solved on the square $[0, 1] \times [0, 1]$ divided in four quadrants by lines $x = 1/2$ and $y = 1/2$. The Riemann problems are defined by initial constant states on each quadrant. All four 1D Riemann Problems between quadrants have exactly one wave: four shocks for the case 1 and four contact discontinuities for the case 2. Both cases were computed with primal grids of 200 × 200 cells which represent about 80,000 cells counting the dual mesh. In order to complete the scheme (7), the adopted numerical flux functions F and G are given by the well–known HLLC approximate Riemann solver (see [3, 8, 14]). The results are displayed for density in Fig. 2. We also provide a comparison with the classical MUSCL scheme on the line $y = x$ and a comparison of the CPU time between the two methods.

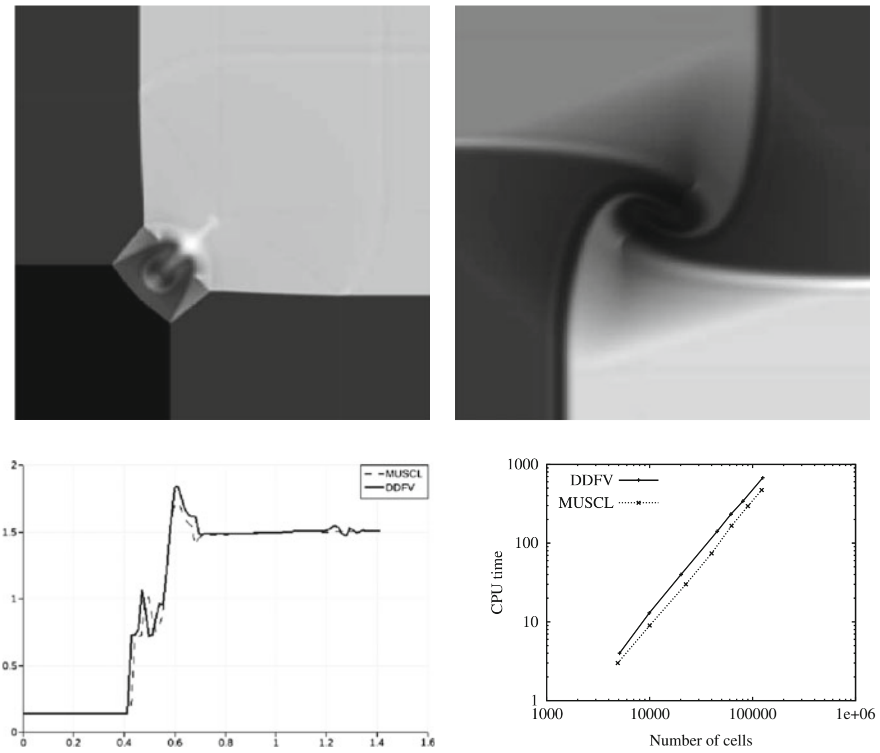


Fig. 2 Results for the 2D Riemann Problem Case 1 (top left) and case 2 (top right) obtained by the derived MUSCL–DDFV scheme. Comparison between the MUSCL–DDFV scheme and the classical MUSCL scheme for case 1: density on the line $y = x$ (bottom left) and CPU time (bottom right)

5 Conclusion

We have presented a second–order robust scheme to approximate the solutions of the 2D Euler equations. The main novelty of this work lies in the gradient reconstruction based on the DDFV methods and the use of two overlapping meshes. We have shown that the method gives good results on structured meshes. Arguing the properties of the DDVF approach, unstructured mesh extensions will be easily obtained.

In order to ensure the robustness, we have enforced that the reconstructed state vectors remain conservative. Another improvement must be performed to propose robust non–conservative reconstructions.

References

1. Andreianov, B., Boyer, F., Hubert, F.: Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numerical Methods for Partial Differential Equations **23**(1), 145–195 (2007)
2. Berthon, C.: Stability of the MUSCL schemes for the Euler equations. Comm. Math. Sci **3**, 133–158 (2005)
3. Bouchut, F.: Nonlinear stability of finite volume methods for hyperbolic conservation laws and well-balanced schemes for sources. Frontiers in Mathematics. Birkhäuser Verlag, Basel (2004)
4. Buffard, T., Clain, S.: Monoslope and multislope MUSCL methods for unstructured meshes. Journal of Computational Physics **229**(10), 3745–3776 (2010)
5. Coudière, Y., Manzini, G.: The Discrete Duality Finite Volume Method for Convection-diffusion Problems. SIAM Journal on Numerical Analysis **47**(6), 4163–4192 (2010)
6. Domelevo, K., Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. Mathematical Modelling and Numerical Analysis **39**(6), 1203–1249 (2005)
7. Godlewski, E., Raviart, P.A.: Numerical approximation of hyperbolic systems of conservation laws, *Applied Mathematical Sciences*, vol. 118. Springer-Verlag, New York (1996)
8. Harten, A., Lax, P., Van Leer, B.: On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. SIAM review pp. 35–61 (1983)
9. Hermeline, F.: A finite volume method for the approximation of diffusion operators on distorted meshes. Journal of computational Physics **160**(2), 481–499 (2000)
10. Hermeline, F.: Approximation of 2-D and 3-D diffusion operators with variable full tensor coefficients on arbitrary meshes. Computer Methods in Applied Mechanics and Engineering **196**(21-24), 2497–2526 (2007)
11. Kurganov, A., Tadmor, E.: Solution of two-dimensional Riemann problems for gas dynamics without Riemann problem solvers. Numerical Methods for Partial Differential Equations **18**(5), 584–608 (2002)
12. LeVeque, R.: Finite volume methods for hyperbolic problems. Cambridge Univ Pr (2002)
13. Toro, E.: Riemann solvers and numerical methods for fluid dynamics: a practical introduction. Springer Verlag (2009)
14. Toro, E., Spruce, M., Speares, W.: Restoration of the contact surface in the HLL-Riemann solver. Shock waves **4**(1), 25–34 (1994)
15. Van Leer, B.: Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method. Journal of Computational Physics **32**(1), 101–136 (1979)
16. Van Leer, B.: A historical oversight: Vladimir P. Kolgan and his high-resolution scheme. Journal of Computational Physics (2010)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Comparison of Explicit and Implicit Time Advancing in the Simulation of a 2D Sediment Transport Problem

M. Bilanceri, F. Beux, I. Elmahi, H. Guillard, and M.V. Salvetti

Abstract The simulation of sediment transport, based on the shallow-water equations coupled with Grass model for the sediment transport equation is considered. The aim of the present paper is to investigate the behavior of implicit linearized schemes in this context. A finite-volume method is considered and second-order accuracy in space is obtained through MUSCL reconstruction. A second-order time accurate explicit version of the scheme is obtained through a two step Runge-Kutta method. Implicit linearized schemes of second-order of accuracy in time are derived thanks to a BDF method associated with a Defect Correction technique. The different time-advancing schemes are compared, using a 2D sediment transport problem, with different types of flow/bed interactions. The implicit one largely outperforms the explicit version for slow flow/bed interactions while in the case of fast flow/bed interactions, the CPU time of both time integration schemes are comparable. Thus, the implicit scheme turns out to be a good candidate to simulate flows with sediment transport in practical applications.

Keywords sediment transport, Grass model, linearized implicit time advancing, automatic differentiation
MSC2010: 65-06

M. Bilanceri and M.V. Salvetti
University of Pisa (Italy), e-mail: marco.bilanceri@gmail.com, mv.salvetti@ing.unipi.it

F. Beux
Alta S.p.A., Pisa (Italy), e-mail: f.beux@alta-space.com

I. Elmahi
EMCS, Ensa, Oujda, Complexe Universitaire, Oujda (Morocco), e-mail: ielmahi@ensa.ump.ma

H. Guillard
INRIA, Sophia Antipolis and Laboratoire Jean-Alexandre Dieudonné, University of Nice Sophia-Antipolis, Parc Valrose 06108 Nice Cedex, (France), e-mail: Herve.Guillard@sophia.inria.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_14, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

A huge amount of work has been done in the last decades to develop numerical methods for the simulation of sediment transport problems (see, e.g., the references in [1, 4]). In this context, the hydrodynamics part is usually modeled through the classical shallow-water equations coupled with an additional equation for the morphodynamical component. The Grass equation [6] is considered herein, which is one of the most popular and simple models. In this context, the treatment of the source terms and of the bed-load fluxes has received the largest attention while time advancing has received much less attention and it is usually carried out by explicit schemes. The focus of the present paper is on the comparison between explicit and implicit schemes in the simulation of a 2D sediment transport problem. We only consider flows over wet areas. The extension to cases in presence of dry areas will the object of further studies. If the interaction of the water flow with the mobile bed is slow, the characteristic time scales of the flow and of the sediment transport can be very different introducing time stiffness in the global problem. Thus, for these cases, it can be advantageous to use implicit schemes. On the other hand, since the considered problems are unsteady, attention must be paid for implicit schemes in the choice of the time step. Another difficulty with implicit schemes is that, in order to avoid the solution of a nonlinear system at each time step, the numerical fluxes must be linearized in time. In order to overcome these difficulties, we use an automatic differentiation tool (Tapenade, [7]). Our starting point was the SRNH numerical scheme, specifically developed and validated for the numerical simulation of sediment transport problems [1]. An implicit version of this scheme is derived herein by computing the Jacobian matrices of the first-order accurate numerical fluxes by the previously mentioned automatic differentiation tool. A defect-correction approach [10] is finally used to obtain second-order accuracy at limited computational costs. The implicit method is compared with the explicit one in a 2D benchmark.

2 Physical model and Numerical Method

The physical model used in this work consists in the well known shallow-water equations coupled with an additional equation to describe the transport of sediment:

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{W})}{\partial x} + \frac{\mathbf{G}(\mathbf{W})}{\partial y} = \mathbf{S}(\mathbf{W}) \tag{1}$$

where x and y are the spatial coordinates, t is the time, and $\mathbf{W}$, $\mathbf{F}(\mathbf{W})$, $\mathbf{G}(\mathbf{W})$ and $\mathbf{S}(\mathbf{W})$ are defined as follows:

$$
\begin{cases}
\mathbf{W} = \left(h, \quad hu, \quad hv, \quad Z \right)^T \\
\mathbf{F}(\mathbf{W}) = \left(hu, \quad hu^2 + \frac{1}{2}gh^2 + ghZ, \quad huv, \quad \frac{1}{1-p}Q_x \right)^T \\
\mathbf{G}(\mathbf{W}) = \left(hv, \quad hvu, \quad hv^2 + \frac{1}{2}gh^2 + ghZ, \quad \frac{1}{1-p}Q_y \right)^T \\
\mathbf{S}(\mathbf{W}) = \left(0, \quad gZ\frac{\partial h}{\partial x}, \quad gZ\frac{\partial h}{\partial y}, \quad 0 \right)^T
\end{cases}
\tag{2}
$$

In (2) h is the height of the flow above the bottom Z, g is acceleration of gravity and u and v are the velocity components in the x and y directions. The first three equations of (1) are the standard 2D Shallow Water equations, recast as in [8] in order to avoid the singularity of the Jacobian of the flux function. The last one is the well-known Exner equation for the evolution of the bed level. We restrict our attention to the case in which the sediment transport porosity p is constant and the bed-load sediment transport fluxes Q_x and Q_y are defined by the Grass model:

$$
Q_x = Au\left(u^2 + v^2\right)^{\frac{m-1}{2}}, \quad Q_y = Av\left(u^2 + v^2\right)^{\frac{m-1}{2}} \tag{3}
$$

where A and $1 \leq m \leq 4$ are experimental constants depending on the particular problem under consideration. The classical case $m = 3$ is considered here.

The numerical method proposed to discretize in space the system of equations (1)-(2) is a finite-volume approach, applicable to unstructured grids. Namely, it is the SRNH scheme introduced in [11]. A brief summary of the main characteristics of the scheme is given herein, for additional details we refer to [1, 11].

The scheme is composed by a predictor and a corrector stage: in the predictor stage an averaged state $\mathbf{U}_{ij}^n$ is computed, then this predicted state is used in the corrector stage to update the solution. The predictor stage is based on primitive variables projected on the normal and tangential directions with respect to the cell interface, $\mathbf{n}$ and τ. Hence, by introducing the normal and tangential components of the velocity, $u_\mathbf{n}$ and u_τ, it is possible to reformulate the system (1) as follows:

$$
\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A_n}(\mathbf{U})\frac{\partial \mathbf{U}}{\partial \mathbf{n}} = 0 \tag{4}
$$

$$
\mathbf{U} = \begin{pmatrix} h \\ u_\mathbf{n} \\ u_\tau \\ Z \end{pmatrix}, \quad \mathbf{A_n}(\mathbf{U}) = \begin{pmatrix} u_\mathbf{n} & h & 0 & 0 \\ g & u_\mathbf{n} & 0 & g \\ 0 & 0 & u_\mathbf{n} & 0 \\ 0 & A(1-p)^{-1}(3u_\mathbf{n}^2 + u_\tau^2) & 2A(1-p)^{-1}u_\mathbf{n}u_\tau & 0 \end{pmatrix} \tag{5}
$$

Starting from (5) it is possible to introduce a Roe average state $\overline{\mathbf{U}}_{ij}$ and a sign matrix $\mathrm{sgn}\left[\mathbf{A_n}(\overline{\mathbf{U}})\right]$ defined as:

$$\overline{\mathbf{U}}_{ij} = \left(\frac{h_i + h_j}{2}, \; \frac{u_{\mathbf{n},i}\sqrt{h_i} + u_{\mathbf{n},j}\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}, \; \frac{u_{\tau,i}\sqrt{h_i} + u_{\tau,j}\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}, \; \frac{Z_i + Z_j}{2} \right)^T \tag{6}$$

$$\operatorname{sgn}\left[\mathbf{A}_{\mathbf{n}}(\overline{\mathbf{U}})\right] = \mathscr{R}(\overline{\mathbf{U}}) \Lambda_{\text{sgn}}(\overline{\mathbf{U}}) \mathscr{R}^{-1}(\overline{\mathbf{U}}) \tag{7}$$

where the elements of the diagonal matrix $\Lambda_{\text{sgn}}(\overline{\mathbf{U}})$ are the sign function of the eigenvalues of $\mathbf{A}_{\mathbf{n}}(\overline{\mathbf{U}})$ and $\mathscr{R}(\overline{\mathbf{U}})$ is the corresponding right-eigenvector matrix.

The explicit SRNH scheme is then formulated as follows:

$$\mathbf{U}^n_{ij} = \frac{1}{2}\left(\mathbf{U}^n_i + \mathbf{U}^n_j\right) - \frac{1}{2}\operatorname{sgn}\left[\mathbf{A}_{\mathbf{n}}(\overline{\mathbf{U}}_{ij})\right]\left(\mathbf{U}^n_j - \mathbf{U}^n_i\right) \tag{8}$$

$$\frac{\mathbf{W}^{n+1}_i - \mathbf{W}^n_i}{\Delta^n t} = -\frac{1}{|V_i|}\sum_{j \in N(i)} \mathscr{F}(\mathbf{W}^n_{ij}, \mathbf{n}_{ij})|\Gamma_{ij}| + \mathbf{S}^n_i \tag{9}$$

where $\mathbf{W}^n_{ij}$ is obtained from $\mathbf{U}^n_{ij}$, $N(i)$ is the set of neighboring cells of the i^{th} cell, $|V_i|$ is the area of the cell, Γ_{ij} is the interface between cell i and j, $\Delta^n t$ is the n^{th} time-step and $\mathscr{F}$ is the analytical flux function. $\mathbf{S}^n_i$ is the discretization of the source term which, in order to satisfy the C-property [2] is defined as follows:

$$\begin{cases} \overline{Z}^n_{x,i} = \dfrac{1}{2}\dfrac{\sum\limits_{j \in N(i)} \left(Z^n_{ij}\right)^2 n_{x,ij}|\Gamma_{ij}|}{\sum\limits_{j \in N(i)} Z^n_{ij} n_{x,ij}|\Gamma_{ij}|}, \quad \overline{Z}^n_{y,i} = \dfrac{1}{2}\dfrac{\sum\limits_{j \in N(i)} \left(Z^n_{ij}\right)^2 n_{y,ij}|\Gamma_{ij}|}{\sum\limits_{j \in N(i)} Z^n_{ij} n_{y,ij}|\Gamma_{ij}|} \\ \mathbf{S}^n_i = \left(0, \quad g\overline{Z}^n_{x,i}\sum\limits_{j \in N(i)} h^n_{ij} n_{x,ij}|\Gamma_{ij}|, \quad g\overline{Z}^n_{y,i}\sum\limits_{j \in N(i)} h^n_{ij} n_{y,ij}|\Gamma_{ij}|, \quad 0\right)^T \end{cases} \tag{10}$$

To switch from an explicit scheme to an implicit one it is sufficient, to compute the quantities $\mathscr{F}^{n+1}_{ij} = \mathscr{F}(\mathbf{W}^{n+1}_{ij}, \mathbf{n}_{ij})$ and $\mathbf{S}^{n+1}_i$ instead of $\mathscr{F}(\mathbf{W}^n_{ij}, \mathbf{n}_{ij})$ and $\mathbf{S}^n_i$. However, from a practical point of view this would require the solution of a large non-linear system of equations at each time step. The computational cost for this operation is in general not affordable in practical applications and generally greatly overcomes any advantage that an implicit scheme could have with respect to its explicit counterpart. A common technique to overcome this difficulty is to linearize the numerical scheme, i.e. to find an approximation of $\mathscr{F}^{n+1}_{ij}$ and $\mathbf{S}^{n+1}_i$ in the form:

$$\Delta^n \mathscr{F}_{ij} \simeq D_{1,ij}\Delta^n \mathbf{W}_i + D_{2,ij}\Delta^n \mathbf{W}_j, \quad \Delta^n \mathbf{S}_i \simeq \sum_{j \in \bar{N}(i)} D_{3,ij}\Delta^n \mathbf{W}_j \tag{11}$$

where $\Delta^n(\cdot) = (\cdot)^{n+1} - (\cdot)^n$ and $\bar{N}(i) = N(i) \cup \{i\}$. Using this approximation, the following linear system must be solved at each time step:

$$\frac{\mathbf{W}_i^{n+1}-\mathbf{W}_i^{n}}{\Delta t}+\frac{1}{|V_i|}\sum_{j\in N(i)}|\Gamma_{ij}|\left(D_{1,ij}\,\Delta^n\mathbf{W}_i+D_{2,ij}\,\Delta^n\mathbf{W}_j\right)-\sum_{j\in\bar{N}(i)}D_{3,ij}\,\Delta^n\mathbf{W}_j$$
$$=-\frac{1}{|V_i|}\sum_{j\in N(i)}\mathscr{F}(\mathbf{W}_{ij}^n,\mathbf{n}_{ij})|\Gamma_{ij}|+\mathbf{S}_i^n \quad (12)$$

The implicit linearized scheme is completely defined once a suitable definition for the matrices $D_{1,ij}, D_{2,ij}, D_{3,ij}$ is given. If the flux function and the source term are differentiable, a common choice is to use the Jacobian matrices. Nevertheless, it is not always possible nor convenient to exactly compute the Jacobian matrices. In fact, it is not unusual to have some lack of differentiability of the numerical flux functions. Furthermore the explicit scheme (9) is composed by a predictor and a corrector stage and this significantly increases the difficulty in linearizing. This problem has been solved herein by computing through the automatic differentiation software Tapenade [7] the flux Jacobians, which are used to approximate$\mathscr{F}_{ij}^{n+1}$ and $\mathbf{S}_i^{n+1}$, as defined in Eq. (11). Given the source code of a routine which computes the explicit numerical fluxes, the differentiation software generates a new source code which computes the flux Jacobians, and, thus, the derivation and the implementation of their analytical expressions can be avoided.

The extension to second-order accuracy in space can be achieved by using a classical MUSCL technique [9], in which (8) is computed by using extrapolated values at the cell interfaces. The extrapolation is done here as in [3] associated with the Minmod slope limiter. For the explicit scheme, second-order accuracy in time is achieved through a two-step Runge-Kutta scheme. Considering the implicit case, it is possible to obtain a space and time second-order accurate formulation by considering the MUSCL technique for space as previously defined and a second-order backward differentiation formula in time. However, the linearization for the second-order accurate fluxes and source terms and the solution of the resulting linear system implies significant computational costs and memory requirements. Thus, a defect-correction technique [10] is used here, which consists in iteratively solving simpler problems obtained, just considering the same linearization as used for the first-order scheme. Thus defining $\mathscr{W}^0=\mathbf{W}^n$, the defect-correction iterations write as follows, the unknown being $\Delta^s\mathscr{W}_i$:

$$\frac{(1+2\tau)}{\Delta^n t\,(1+\tau)}\Delta^s\mathscr{W}_i+\frac{1}{|V_i|}\sum_{j\in N(i)}|\Gamma_{ij}|\left(D_{1,ij}\,\Delta^s\mathscr{W}_i+D_{2,ij}\,\Delta^s\mathscr{W}_j\right)-\sum_{j\in\bar{N}(i)}D_{3,ij}\,\Delta^s\mathscr{W}_j$$
$$=\frac{(1+2\tau)\mathscr{W}_i^s-(1+\tau)^2\mathbf{W}_i^n+\tau^2\mathbf{W}_i^{n-1}}{\Delta^n t\,(1+\tau)}-\frac{1}{|V_i|}\sum_{j\in N(i)}\mathscr{F}\left(\mathscr{W}_{ij}^s,\mathscr{W}_{ji}^s\right)|\Gamma_{ij}|+[\mathbf{S}_2]_i^s \quad (13)$$

for $s=0,\cdots,r-1$. In (13), $\tau=\frac{\Delta^n t}{\Delta^{n-1} t}$, $D_{1,ij}, D_{2,ij}, D_{3,ij}$ are the matrices of the approximation (11) and the update solution is $\mathbf{W}^{n+1}=\mathscr{W}^r$. It can be shown

[10] that only one defect-correction iteration is theoretically needed to reach a second-order accuracy while few additional iterations (one or two) can improve the robustness.

3 Numerical Experiments

The 2D test case considered herein is a well-known benchmark test, proposed in several papers (see, e.g. [1,5]). It is a sediment transport problem in a square domain Ω of dimensions $1000 \times 1000\ m^2$ with a non constant bottom relief. The initial bottom topography is defined as follows:

$$Z(0,x) = \sin^2\left(\frac{(x-300)\pi}{200}\right)\sin^2\left(\frac{(y-400)\pi}{200}\right) \text{ if } (x,y) \in Q_h, \text{ 0 elsewhere} \tag{14}$$

where $Q_h = [300, 500] \times [400, 600]$. Given $Z(0,x)$, the remaining initial conditions are $h(0,x,y) = 10 - Z(0,x,y)$, $u(0,x,y) = \frac{10}{h(0,x,y)}$ and $v(0,x,y) = 0$. Considering the boundaries, Dirichlet boundary conditions are imposed at the inlet, while at the outlet characteristic based conditions are used. Finally, free-slip is imposed on the lateral boundaries. The spatial discretization of the computational domain has been carried out by using two different grids: for the first grid GR1, the number of the nodes and the characteristic length of the elements are, respectively, $l_m = 20\ m$ and $N_c = 2901$. The second grid GR2 is characterized by $l_m = 10\ m$ and $N_c = 11425$.

Two different values of the parameter A are considered, namely a case with slow interaction between the flow and the bed, $A = 0.001$ and a fast one, $A = 1$. Due to the different time scales for the evolution of the bottom topography, different time intervals have been simulated for the considered cases: the total simulation time is 500 seconds for $A = 1$ and 360000 seconds for $A = 0.001$.

For the slow speed of interaction case, Figure 1a shows a comparison of the results obtained by means of the explicit version of the scheme at CFL = 0.8 with those of the implicit one at CFL = 1000 both for 1^{st} and 2^{nd}-order accuracy for grid GR2. For the definition of the CFL number we refer to [1]. There is practically no difference between the solutions obtained with the implicit and explicit version of the schemes, while the results obtained at 1^{st}-order of accuracy significantly differ from the 2^{nd}-order ones. Note that the results shown in Fig. 1a for the 2^{nd}-order implicit scheme are computed using only one DeC iteration. By increasing the number of DeC iterations it is possible to further increment, without loosing in accuracy, the CFL number of the 2^{nd}-order implicit scheme (see Fig. 1b). In particular, when 3 DeC iterations are considered it is possible to use a CFL number equal to 10^4 (see also Table 1). As shown in Fig. 1b, similar results can be obtained by considering the grid GR1 instead of the GR2. The profiles of $h + Z$ are shown if Fig. 1c. Slightly larger oscillations are observed for the second-order implicit scheme, but at the first order both schemes gave practically the same results. As

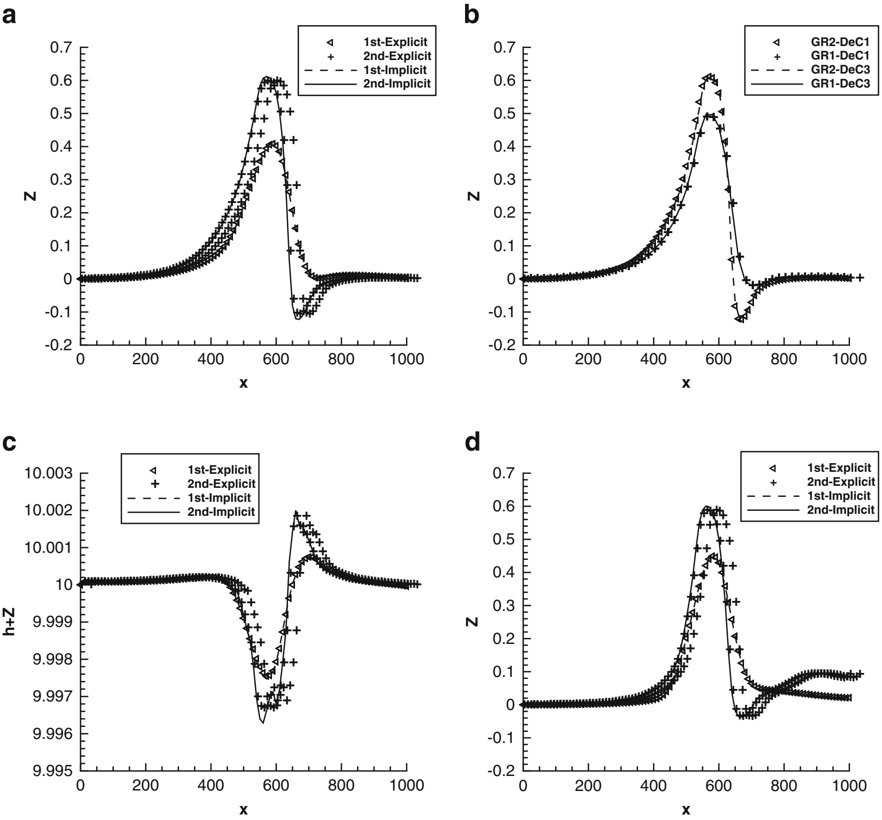
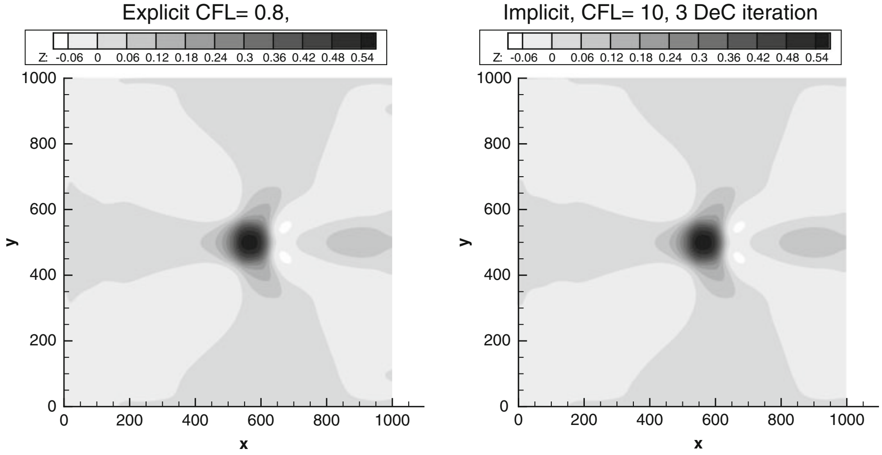


Fig. 1 Comparison of the results given the explicit and implicit schemes; profiles along the line $y = 500$ of: (a) Z for $A = 10^{-3}$ and GR2, (b) Z for implicit schemes and $A = 10^{-3}$, (c) $h + Z$ for $A = 10^{-3}$ GR2, (d) Z for $A = 10^{0}$ and GR2

for the computational costs, Table 1 shows that already at CFL $= 1000$ the gain in CPU time obtained with the implicit scheme is large, both for 1^{st} and 2^{nd}-order of accuracy. The CPU gain obtained with the implicit scheme is significantly larger for 2^{nd}-order accuracy. Indeed, when the implicit formulation is used, there are not significant differences, in terms of CPU time, between the 1^{st} and 2^{nd}-order simulations. Instead in the explicit case an important computational cost increase is observed to reach 2^{nd}-order accuracy: the 2^{nd}-order approach is $\simeq 2.4$ times more expensive than the 1^{st}-order one. As a consequence, already at CFL $= 1000$ using 1 DeC iteration the 2^{nd}-order implicit approach is more than 60 times faster than the explicit one on GR1 and about 30 times faster on GR2. The CPU gain of the 2^{nd}-order implicit approach can be further increased considering 3 DeC iterations and CFL $= 10^4$. For the fast speed of interaction case, to avoid loss of accuracy the CFL number of the implicit scheme must be lowered down to 1. On the other hand, by increasing the number of DeC iterations, it is possible to increase the maximum



Explicit CFL= 0.8,

Implicit, CFL= 10, 3 DeC iteration

Fig. 2 Comparison of the results for the bed profile of the 2^{nd}-order scheme, $A = 1$, Grid GR2

CFL value by a factor 10. As an example Fig. 2 shows a comparison between the explicit and implicit approach at different CFL values for the grid GR2. Due to the reduced CFL number achievable without loss of accuracy for the implicit scheme in this test case the computational cost of the implicit scheme is larger than for the explicit one, both at first and second order of accuracy, as it is shown in Table 1. Summarizing, in order to avoid loss of accuracy, the CFL number of the implicit scheme must be reduced to a value roughly inversely proportional to the velocity of the interaction between the flow and the bed-load. Also, the increase of the number of DeC iterations allows the maximum CFL number achievable without loosing in accuracy to be increased, and therefore the simulation CPU time is reduced. The implicit code has been found to be computationally more efficient than the explicit one for slow rates of the interaction between the bed and the flow.

Table 1 CPU time required for the considered simulations, comparison between explicit and implicit approach, both at first and second-order of accuracy

	$A = 0.001$			$A = 1$		
Method	GR1	GR2	CFL	GR1	GR2	CFL
Explicit 1^{st} order	$12824s$	$103238s$	0.8	$21.0s$	$169.7s$	0.8
Explicit 2^{nd} order	$30996s$	$247215s$	0.8	$52.4s$	$409.9s$	0.8
Implicit 1^{st} order	$323.6s$	$4336s$	10^3	$191.5s$	$1541s$	10^0
Implicit 2^{nd} order 1 DeC	$481.5s$	$8537s$	10^3	$198.7s$	$1582s$	10^0
Implicit 2^{nd} order 3 DeC	$265.9s$	$4866s$	10^4	$74.5s$	$606.8s$	10^1

Acknowledgements This work has been realized in the framework of the EuroMéditerranée 3 + 3 network MhyCoF.

References

1. F. Benkhaldoun, S. Sahmim, M. Seaïd. *A two-dimensional finite volume morphodynamic model on unstructured triangular grids.* Int. J. Numer. Meth. Fluids, 63:1296–1327, 2010.
2. A. Bermudez, M.E. Vazquez. *Upwind methods for hyperbolic conservation laws with source terms.* Computers & Fluids, 23(8):1049–1071, 1994.
3. S. Camarri, M.V. Salvetti, B. Koobus, A. Dervieux. *A low-diffusion MUSCL scheme for LES on unstructured grids.* Computers & Fluids, 33:1101–1129, 2004.
4. M.J. Castro Díaz, E.D. Fernández-Nieto, A.M. Ferreiro. *Sediment transport models in shallow water equations and numerical approach by high order finite volume methods.* Computers & Fluids, 37(3):299–316, 2008.
5. M.J. Castro Díaz, E.D. Fernández-Nieto, A.M. Ferreiro, C. Parés. *Two-dimensional sediment transport models in shallow water equations. A second order finite volume approach on unstructured meshes.* Computer Meth. Appl. Mech. Eng., 198:2520–2538, 2009.
6. A.J. Grass. *Sediments transport by waves and currents.* Tech. Rep., SERC London Cent. Mar. Technol., Report No. FL29, 1981.
7. L. Hascoët, V. Pascual. *TAPENADE 2.1 User's Guide.* Tech. Rep. n 300. INRIA, 2004.
8. J. Hudson, P. K. Sweby. *Formulations for Numerically Approximating Hyperbolic Systems Governing Sediment Transport.* J. Sci. Comput., 19:225-252, 2003.
9. B. van Leer. *Towards the ultimate conservative difference scheme* V*: a second-order sequel to* G*odunov's method.* J. Comput. Phys., 32(1):101–136, 1979.
10. R. Martin, H. Guillard. *A second order defect correction scheme for unsteady problems.* Computers & Fluids, 25(1):9–27, 1996.
11. S. Sahmim, F. Benkhaldoun, F. Alcrudo. *A sign matrix based scheme for non-homogeneous PDE's with an analysis of the convergence stagnation phenomenon.* J. Comput. Phys., 226(2):1753–1783, 2007.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Numerical Simulation of the Flow in a Turbopump Inducer in Non-Cavitating and Cavitating Conditions

M. Bilanceri, F. Beux, and M.V. Salvetti

Abstract A numerical methodology for the simulation of cavitating flows in real complex geometries is presented. A homogeneous-flow cavitation model, accounting for thermal effects and active nuclei concentration, which leads to a barotropic state law is adopted. The continuity and momentum equations are discretized through a mixed finite-element/finite-volume approach, applicable to unstructured grids. A robust preconditioned low-diffusive HLL scheme is used to deal with all speed barotropic flows. Second-order accuracy in space is obtained through MUSCL reconstruction. Time advancing is carried out by a second-order implicit linearized formulation together with the Defect Correction technique. The flow in a real 3D inducer for rockets turbopumps is simulated for a wide range of conditions: different flow rates and rotating speeds as well as non-cavitating and cavitating flows are considered. The results obtained with this numerical approach are compared with experimental data.

Keywords cavitating flows, homogeneous flow model, low diffusive HLL scheme, linearized implicit time advancing
MSC2010: 65-06

1 Introduction

A tool for numerical simulation of 3D compressible flows satisfying a barotropic equation of state is presented in this work. In particular, we are interested in simulating cavitating flows through the barotropic homogeneous flow model proposed in [1]. The numerical method used in this work is based on a mixed

M. Bilanceri and M.V. Salvetti
University of Pisa (Italy), e-mail: marco.bilanceri@gmail.com, mv.salvetti@ing.unipi.it

F. Beux
Alta S.p.A., Pisa (Italy), e-mail: f.beux@alta-space.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_15, © Springer-Verlag Berlin Heidelberg 2011

finite-element/finite-volume spatial discretization on 3D unstructured grids. Viscous fluxes are discretized using P1 finite-elements while for the convective fluxes the LD-HLL scheme [2], a low-diffusive modification of the Rusanov scheme, is adopted. Second-order in space is obtained using a MUSCL reconstruction technique and time-consistent preconditioning is introduced to deal with the low Mach number regime. A linearized implicit time-advancing is associated to a defect-correction technique to obtain a second-order accurate (both in time and space) formulation at a limited computational cost. A non inertial reference frame, rotating at constant angular velocity, is used to account for possible solid-body rotation and the standard $k - \varepsilon$ turbulence model is introduced to capture turbulence effects. The considered numerical tool is used to simulate the flow in a real 3D inducer in both non-cavitating and cavitating conditions.

2 Physical model and numerical method

The physical model considered in this work consists in the standard Navier-Stokes equations for a barotropic flow. Due to the barotropic equation of state (EOS) considered, the energy equation can be discarded since it is decoupled from the mass and momentum balances. Thus, considering a reference frame rotating with constant angular velocity ω, the following system of equations is obtained:

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial}{\partial x_j} F_j(\mathbf{W}) - \frac{\partial}{\partial x_j} \mu V_j(\mathbf{W}, \nabla \mathbf{W}) = \mathbf{S}(\omega, \mathbf{x}, \mathbf{W}) \tag{1}$$

In Eq. (1) the Einstein notation is used, μ is the molecular viscosity of the fluid, $\mathbf{W} = (\rho, \rho u_1, \rho u_2, \rho u_3)^T$ is the unknown vector, where ρ is the density and u_i the velocity component in the i^{th} direction. $\mathbf{F} = (F_1, F_2, F_3)$ and $\mathbf{V} = (V_1, V_2, V_3)$ are, respectively, the convective fluxes and the diffusive ones (not shown here for sake of brevity). Finally, $\mathbf{S}$ is the source term appearing in a frame of reference rotating with constant speed ω:

$$\begin{cases} \overline{\mathbf{S}} & = -\left(2\omega \wedge \rho \mathbf{u} + \rho\omega \wedge (\omega \wedge \mathbf{x})\right) \\ \mathbf{S}(\omega, \mathbf{x}, \mathbf{W}) & = \left(0, \ \overline{\mathbf{S}}^T\right)^T \end{cases} \tag{2}$$

System (1) is completely defined once a suitable constitutive equation $p = p(\rho)$ is introduced. In this work a weakly-compressible liquid at constant temperature T_L is considered as working fluid. The liquid density ρ is allowed to locally fall below the saturation limit $\rho_{Lsat} = \rho_{Lsat}(T_L)$ thus originating cavitation phenomena. A regime-dependent (wet/cavitating) constitutive relation is therefore adopted. As for the wet regime ($\rho \geq \rho_{Lsat}$), a barotropic model of the form

$$p = p_{sat} + \frac{1}{\beta_{sL}} \ln\left(\frac{\rho}{\rho_{Lsat}}\right) \tag{3}$$

is adopted, $p_{sat} = p_{sat}(T_L)$ and $\beta_{sL} = \beta_{sL}(T_L)$ being the saturation pressure and the liquid isentropic compressibility, respectively. As for the cavitating regime ($\rho < \rho_{Lsat}$), a homogeneous-flow model explicitly accounting for thermal cavitation effects and for the concentration of the active cavitation nuclei in the pure liquid has been adopted [1]:

$$\frac{p}{\rho}\frac{d\rho}{dp} = (1-\alpha)\left[(1-\varepsilon_L)\frac{p}{\rho_{Lsat}a^2_{Lsat}} + \varepsilon_L g^{\star}\left(\frac{p_c}{p}\right)^{\eta}\right] + \frac{\alpha}{\gamma_V} \qquad (4)$$

where $g^{\star}$, η, γ_V and p_c are liquid parameters, a_{Lsat} is the liquid sound speed at saturation, $\alpha = 1 - \rho/\rho_{Lsat}$ and $\varepsilon_L = \varepsilon_L(\alpha, \zeta)$ is a given function (see [1] for its physical interpretation and for more details). The resulting unified barotropic state law for the liquid and for the cavitating mixture only depends on the two parameters T_L and ζ. For instance, for water at $T_L = 293.16K$, the other parameters involved in (3) and (4) are: $p_{sat} = 2806.82$ Pa, $\rho_{Lsat} = 997.29$ kg/m^3, $\beta_{sL} = 5\,10^{-10}$ Pa^{-1}, $g^* = 1.67$, $\eta = 0.73$, $\gamma_V = 1.28$, $p_c = 2.21\,10^7$ Pa and $a_{Lsat} = 1415$ m/s [6]. Note that despite the model simplifications leading to a unified barotropic state law, the transition between wet and cavitating regimes is extremely abrupt. Indeed, the sound speed falls from values of order 10^3 m/s in the pure liquid down to values of order 0.1 or 1 m/s in the mixture. The corresponding Mach number variation renders this state law very stiff from a numerical viewpoint. As for the definition of the molecular viscosity, a simple model, which is linear in the cavitating regime, is considered:

$$\mu(\rho) = \begin{cases} \mu_L & \text{if} \quad \rho \geq \rho_{Lsat} \\ \mu_v & \text{if} \quad \rho \leq \rho_v \\ \alpha\mu_v + (1-\alpha)\mu_L & \text{otherwise} \end{cases} \qquad (5)$$

in which μ_v and μ_L are the molecular viscosity of the vapor and of the liquid respectively, which, consistently with the assumptions made in the adopted cavitation model, are considered constant and computed at $T = T_L$.

The spatial discretization of the governing equations is based on a mixed finite-element/finite-volume formulation on unstructured grids. Starting from an unstructured tetrahedral grid, a dual finite-volume tessellation is obtained by the rule of medians. The semi-discrete balance applied to cell C_i reads (not accounting for boundary contributions):

$$V_i \frac{\mathrm{d}\mathbf{W}_i}{\mathrm{d}t} + \sum_{j \in N(i)} \Phi_{ij} + \Upsilon_i = \Omega_i \qquad (6)$$

where $\mathbf{W}_i$ is the semi-discrete unknown associated with C_i, V_i is the cell volume, and $N(i)$ represents the set of neighbors of the i^{th} cell. The numerical discretization of the convective flux crossing the boundary ∂C_{ij} shared by C_i and C_j (positive towards C_j) is denoted Φ_{ij}, while Υ_i and Ω_i are the numerical discretizations for, respectively, the viscous fluxes and the source term. Let us describe, first, the first-order version of the used numerical method. Once defined $\mathbf{n}_{ij} = \left(n_{ij,1}, n_{ij,2}, n_{ij,3}\right)^T$

as the integral over ∂C_{ij} of the outer unit normal to the cell boundary, it is possible to approximate Φ_{ij} by the following preconditioned flux function:

$$\Phi_{ij} = \frac{n_{ij,k}\left(F_k(\mathbf{W}_i) + F_k(\mathbf{W}_j)\right)}{2} -$$
$$\frac{1}{2}\begin{pmatrix} \lambda_1^p & 0 & 0 & 0 \\ 0 & \left(\Delta^{32}\lambda^p\right) n_{ij,1}^2 + \lambda_3^p & \left(\Delta^{32}\lambda^p\right) n_{ij,1}n_{ij,2} & \left(\Delta^{32}\lambda^p\right) n_{ij,1}n_{ij,3} \\ 0 & \left(\Delta^{32}\lambda^p\right) n_{ij,2}n_{ij,1} & \left(\Delta^{32}\lambda^p\right) n_{ij,2}^2 + \lambda_3^p & \left(\Delta^{32}\lambda^p\right) n_{ij,2}n_{ij,3} \\ 0 & \left(\Delta^{32}\lambda^p\right) n_{ij,3}n_{ij,1} & \left(\Delta^{32}\lambda^p\right) n_{ij,3}n_{ij,2} & \left(\Delta^{32}\lambda^p\right) n_{ij,3}^2 + \lambda_3^p \end{pmatrix} (\mathbf{W}_j - \mathbf{W}_i) \tag{7}$$

where $\Delta^{32}\lambda^p = \lambda_2^p - \lambda_3^p, \lambda_1^p = \theta^{-1}\lambda_{ij}, \lambda_2^p = \theta\lambda_{ij}, \lambda_3^p = \lambda_{ij}$ and the parameters θ and λ_{ij} are defined as follows:

$$\theta = \theta(M) = \begin{cases} 10^{-6} & \text{if } M \leq 10^{-6} \\ \min(M, 1) & \text{otherwise} \end{cases}, \quad M = \frac{|\tilde{\mathbf{u}}_{ij}|}{\tilde{a}_{ij}}, \quad \lambda_{ij} = \tilde{\mathbf{u}}_{ij} + \tilde{a}_{ij} \tag{8}$$

$\tilde{\mathbf{u}}_{ij}$ and $\tilde{a}_{ij}$ being the Roe averages for, respectively, the velocity and the sound of speed. The discretization (7) is the 3D extension of LD-HLL scheme defined in [2] as a low diffusive modification of the Rusanov scheme.

The discretization of the viscous fluxes is instead based on P1 finite-elements in which the test functions are linear functions on the tetrahedral element. The source term is discretized as follows:

$$\Omega_i := \begin{pmatrix} 0 \\ -2\,\omega \wedge \rho_i \mathbf{u}_i + \rho_i\,\mathbf{r}_i \end{pmatrix} \quad \mathbf{r}_i := -\,\omega \wedge (\omega \wedge \mathbf{g}_i) \tag{9}$$

$\mathbf{g}_i$ being the centroid of the i^{th} cell.

A first-order implicit Euler method can be used for time-advancing. As a consequence, at each time step it is necessary to compute $\mathscr{F}_i^{n+1} = \mathscr{F}(\mathbf{W}_j^{n+1}, j \in \bar{N}(i))$, where $\bar{N}(i) = N(i) \cup \{i\}$ and $\mathscr{F}_i$ is defined as $\mathscr{F}_i = \sum_{j \in N(i)} \Phi_{ij} + \Upsilon_i - \Omega_i$.

In order to avoid the direct solution of large non-linear system of equations at each time step a linearization can be performed finding an approximation of $\mathscr{F}_i^{n+1}$ in the form:

$$\Delta^n \mathscr{F}_i \simeq \sum_{j \in \bar{N}(i)} D_{ij}\,\Delta^n \mathbf{W}_j \tag{10}$$

where $\Delta^n(\cdot) = (\cdot)^{n+1} - (\cdot)^n$. Using this approximation, the following linear system must be solved at each time step:

$$|V_i| \frac{\mathbf{W}_i^{n+1} - \mathbf{W}_i^n}{\Delta t} + \sum_{j \in \bar{N}(i)} D_{ij}\,\Delta^n \mathbf{W}_j = -\mathscr{F}(\mathbf{W}_j^n, j \in \bar{N}(i)) \tag{11}$$

The implicit linearized scheme is completely defined once a suitable definition for the matrices D_{ij} is given. Since viscous and source terms are easily differentiable, the use of the Jacobian matrices has been considered here to compute their contribution to D_{ij}. However the computation of the Jacobian matrices can be more challenging for the convective fluxes. Thus, in this work the approximate linearization developed in [2] for the numerical flux function (7) has been used. Once matrices D_{ij} are given, the first-order numerical method (11) is completely defined.

Since viscous and source terms are already second-order accurate in space, the extension to second-order accuracy in space can be achieved by simply using a classical MUSCL technique [4], in which the convective fluxes are computed by using extrapolated values at the cell interfaces. The second-order accuracy in time is then achieved through the use of a backward differentiation formula. However, the linearization for the second-order accurate fluxes and the solution of the resulting linear system imply significant computational costs and memory requirements. Thus, a defect-correction technique [5] is used here. It consists in iteratively solving simpler problems obtained just considering the same linearization as used for the first-order scheme. The number of DeC iterations r is typically chosen equal to 2. Indeed, it can be shown [5] that only one defect-correction iteration is theoretically needed to reach a second-order accuracy while few additional iterations (one or two) can improve the robustness.

Finally, in order to account for the turbulence effects the RANS approach together with the standard turbulence model $k - \varepsilon$ have been used. For the sake of brevity the additional terms introduced in the system of equation by this model are not shown. We just mention that the convective and viscous turbulent fluxes are discretized using the same methods considered for their laminar counterparts. Similarly, the turbulent source term appearing in the equations for k and ε is discretized using the same approach considered for the source term associated to the rotating frame of reference.

3 Numerical experiments

In this section the numerical tool described in Sect. 1 is applied to the simulations of the flow in a real three blade axial inducer [6]. It is a three blade inducer with a tip blade radius of 81 mm and 2 mm radial clearance between the blade tip and the external case. Experimental data are available for all the numerical simulations described in the following. In particular the pressure jump between two different stations has been measured for a wide range of working conditions: from small to large mass flow rates, non-cavitating and cavitating conditions and different values of the rotational speed ω_z. The results are presented in terms of the mean adimensionalized pressure jump Ψ as a function of the adimensionalized discharge Φ:

$$\Psi = \frac{\Delta P}{\rho_L \omega_z^2 R_T^2} \qquad \Phi = \frac{Q}{\pi R_T^2 \omega_z R_T} \tag{12}$$

where Q is the discharge, R_T is the radius of the tip of blade, ρ_L the density of the liquid and ω_z is the angular velocity. Note that the numerical pressure jump is averaged over one complete revolution of the inducer. A cylindrical computational domain is used, whose external surface is coincident with the inducer case. The inlet is placed 249 mm ahead of the inducer nose and the outlet is placed 409 mm behind. A second computational domain, characterized by a larger streamwise length (the inlet 1120 mm ahead the inducer nose) has also been considered. Two different grids have been generated to discretize the shorter domain: the basic one G1 (1926773 cells) and G2 (3431721 cells) obtained from G1 by refining the region between the blade tip and the external case. In particular, inside the tip clearance region there are $3-4$ nodes for the grid G1, while there are $9-10$ points for G2. The larger domain has been discretized by grid G1L (2093770 cells), which coincides with G1 in the original domain. The working conditions considered in this work are shown in Table 1, where p_{out} is the outlet pressure of the flow and $\sigma = \frac{p - p_{Lsat}}{0.5\rho\omega_z^2 R_T^2}$ is the cavitating number (only shown for cavitating simulations). Note that, except when differently stated, the simulations do not include turbulence effects.

Table 1 Conditions of the numerical simulations and of the experiments

Benchmark	Ind1	Ind2	Ind3	Ind4	Ind5	Ind6
Φ	0.0584	0.0391	0.0185	0.0531	0.0531	0.0531
ω_z (rpm)	1500	1500	1500	3000	3000	3000
p_{out} (Pa)	125000	125000	125000	60000	85000	82500
T (C°)	25°	25°	25°	16.8°	16.8°	16.8°
σ	-	-	-	0.056	0.084	0.077

As shown in Table 1, all the simulations in non cavitating conditions use the same rotational velocity of 1500 rpm. In the $\Phi - \Psi$ plane the experimental curves of the performances of the inducer are roughly independent from the rotational velocity ω_z [6]. As a consequence, validating the numerical tool for a specific rotational velocity and different flow rates should validate the proposed numerical tool for a generic rotational velocity. Table 2 shows the results for the non-cavitating simulations. It clearly appears that the lower is the discharge Φ, the worse are the results. Already with the coarsest grid G1, rather satisfactory results are obtained for intermediate and high discharge values, Ind2 and Ind1, respectively. Furthermore the quantitative agreement is further improved considering the more refined grid G2 for the case Ind2. Conversely, for the low discharge case, Ind3, the simulations with the grid G1 and G2 greatly overestimate the pressure jump by, respectively, 41% and 30%. The magnitude of this error could be ascribed to the backward flow between the inducer blades and the external case. The correct resolution of this flow is of crucial importance for the determination of the performance of an inducer. Since the smaller is the mass flow rate the greater is the backflow, we investigated

Table 2 Pressure jump in non-cavitating conditions

	Experimental Ψ	Numerical Ψ	Error%
G1-Ind1	0.122	0.114	−6.6%
G1-Ind2	0.186	0.204	+9.7%
G2-Ind2	0.186	0.179	−3.8%
G1-Ind3	0.214	0.302	+41%
G2-Ind3	0.214	0.278	+30%
G1L-Ind3	0.214	0.297	+39%
G1L-Ind3-T	0.214	0.239	+12%

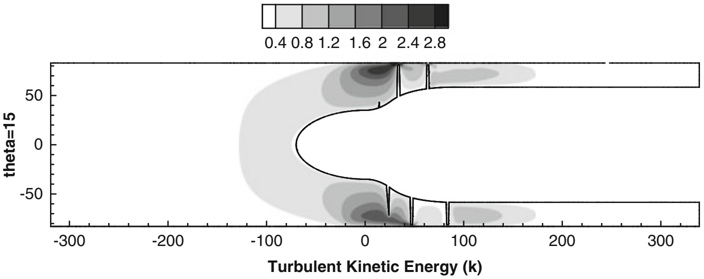


Fig. 1 Cross section of the averaged k field at $\theta = 15°$, simulation G1L-Ind3 (view of the shorter domain)

two possible explanations of this behavior. The first one was that the distance of the inlet from the inducer nose was not large enough to avoid spurious effects on the solution, the second one was that for this case turbulence effects have to be included. The results of the first simulations for the longer computational domain, G1L-Ind3, show that even if there is a small effect, a decrease from 41% to 39%, this is not the source of the error. Instead the results of the simulation G1L-Ind3-T, i.e. the one done considering the RANS model, show that in this case turbulence is a key-issue. Indeed, in this case the error falls down to 12%, less than the error obtained with the refined grid G2 in laminar conditions. As expected the effects of turbulence are particularly important near the gap between the blades and the external case, as it is shown by Fig. 1 by considering the isocontours of k. This strongly affects the backflow and, thus, the pressure jump. This also explains why for larger flow rates, for which the backflow is less important, the effects of turbulence are not so strong and a good agreement with experimental data can be obtained also in laminar simulations.

The mass flow rate for the cavitating cases is large enough to prevent the issues related to the backflow previously described, thus only laminar simulations are considered. The results for the cavitating conditions, reported in Table 3, show that the first grid G1 is not enough refined to correctly describe cavitation for this case. The pressure jump is greatly overestimated. For these conditions the error is related to the underestimation of the cavitating region: the experimental data for $\sigma = 0.056$ show a large cavitating zone and consequently the performance of the

Table 3 Numerical results for the cavitating simulations

	Experimental Ψ	Numerical Ψ	Error%
G1-Ind4	0.105	0.143	+36%
G2-Ind5	0.143	0.130	−8.9%
G2-Ind6	0.137	0.130	−5.0%

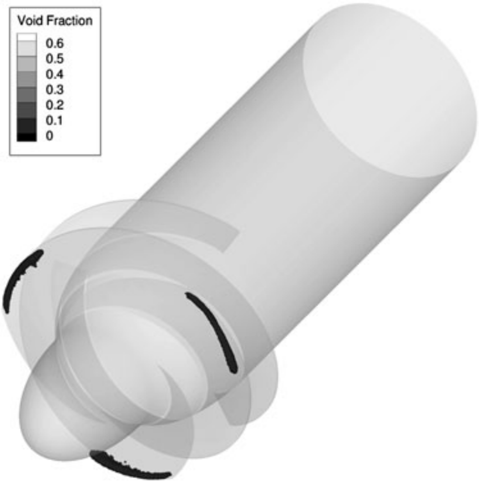


Fig. 2 Isocontours of the cavitating region, $\alpha = 0.005$, for the simulation G2-Ind6

inducer is significantly deteriorated. Instead, in the simulation with grid G1 the extension of the cavitating region is greatly underestimated and, as a consequence, the "numerical" performance of the inducer is similar to the non cavitating case. Grid refinement is particularly effective as shown by the results for the simulations, G2-Ind5 and G2-Ind6. The error in the prediction of the pressure jump is reduced and the extension of the cavitating region, even if it is still underestimated, is closer to the one found in experiments, as it is shown by Fig. 2 which plots the isocontours of the void fraction, corresponding to the cavitating region. Note that when the coarse grid G1 is used the cases Ind5 and Ind6 are not cavitating.

Acknowledgements The support of the European Space Agency under Contract number 20081/06/NL/IA is gratefully acknowledged. The authors also wish to thank the Italian Computer Center CASPUR for having provided computational resources and support.

References

1. L. d'Agostino, E. Rapposelli, C. Pascarella, A. Ciucci *A Modified Bubbly Isenthalpic Model for Numerical Simulation of Cavitating Flows.* 37th AIAA/ASME/SAE/ASEE Joint Propulsion Conference, Salt Lake City, UT, USA, 2001.

2. M. Bilanceri, F. Beux, M.V. Salvetti *An Implicit Low-Diffusive HLL Scheme with Complete Time Linearization: Application to Cavitating Barotropic Flows.* Computer & Fluids, 39(10):1990–2006, 2010.
3. S. Camarri, M.V. Salvetti, B. Koobus, A. Dervieux. *A low-diffusion MUSCL scheme for LES on unstructured grids.* Computers & Fluids, 33:1101–1129, 2004.
4. B. van Leer. *Towards the ultimate conservative difference scheme* V*: a second-order sequel to* G*odunov's method.* J. Comput. Phys., 32(1):101–136, 1979.
5. R. Martin, H. Guillard. *A second order defect correction scheme for unsteady problems.* Computers & Fluids, 25(1):9–27, 1996.
6. L. Torre, G. Pace, P. Miloro, A. Pasini, A. Cervone, L. d'Agostino, *Flow Instabilities on a Three Bladed Axial Inducer at Variable Tip Clearance.* 13^{th} International Symposium on Transport Phenomena and Dynamics of Rotating Machinery, Honolulu, Hawaii, USA.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

On Some High Resolution Schemes for Stably Stratified Fluid Flows

Tomáš Bodnár and Luděk Beneš

Abstract The aim of this paper is to present some high-resolution numerical methods in the context of the solution of stably stratified flow of incompressible fluid. Two different numerical methods are applied to a simple 2D test case of wall bounded flow and results are compared and discussed in detail with emphasize on the specific features of stratified flows. The two numerical methods are the AUSM finite–volume scheme and the high order compact finite-difference scheme.

Keywords finite–volume, finite–difference, stratification, compact, AUSM
MSC2010: 65M08, 65M06, 76D05, 76D50, 76D33

1 Introduction

The numerical solution of stably stratified fluid flows represents a challenging class of problems in modern CFD. This study was motivated by the air flow in the stably stratified Atmospheric Boundary Layer, where the presence of stratification leads to appearance of gravity waves in the proximity of terrain obstacles. These small–amplitude waves are affecting the flow–field at large distances which is in contrast to the typical non-stratified case, where the flow–field is only affected locally in the close proximity of the obstacle. The wavelength of these waves is governed by the Brunt–Väisälä frequency, i.e depends on the product of the gravity acceleration and the background density gradient.

Tomáš Bodnár
Institute of Thermomechanics, Academy of Sciences of Czech Republic, Dolejškova 5, 182 00 Prague 8, Czech Republic, e-mail: bodnar@marian.fsik.cvut.cz

Luděk Beneš
Department of Tech. Mathematics, Faculty of Mech. Engineering, Czech Technical University in Prague, Karlovo Náměstí13, 121 35 Prague 2 Czech Republic, e-mail: Ludek.Benes@fs.cvut.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_16, © Springer-Verlag Berlin Heidelberg 2011

From the numerical point of view, the simulations of stratified fluid flows are in general more demanding than the solution of similar non-stratified flow cases (see our previous work [10], [4], [1], or [6]). First of all the *model of stratified fluid flow* has to be chosen. Such models are based on variable-density incompressible fluid model including gravity force terms. A simple approximation of such model is developed in Sect. 2. The appearance of the gravity waves in the computational field adds some more constrains on the choice of *numerical scheme and grid*. The limiting factor here is the proper resolution of gravity waves in the whole domain with sufficient number of grid points per wavelength and low amount of numerical dumping to preserve the resolved gravity waves rather than excessively dumping them. Last but not least problem comes with *boundary conditions*. Their proper choice and implementation affects the computational field much strongly than in the non–stratified case.

One of the aims of this paper is to demonstrate that the high-order compact finite-difference schemes offer an interesting alternative to the modern finite–volume discretizations. Beside of the high resolving capabilities of both methods, the compact discretizations have well defined dispersion/diffusion properties and thus can safely be applied to the numerical simulations of wave phenomena. These specific properties of compact discretizations have been successfully used in computational aeroacoustics. This paper is one of the first attempts to use these wave resolving capabilities in the numerical solution of stratified fluid flows.

2 Mathematical Model

Full incompressible model The motion equations describing the flow of incompressible Newtonian fluid could be written in the following general form

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{1}$$

$$\frac{\partial \rho}{\partial t} + \frac{\partial (\rho u)}{\partial x} + \frac{\partial (\rho v)}{\partial y} + \frac{\partial (\rho w)}{\partial z} = 0 \tag{2}$$

$$\rho\left(\frac{\partial u}{\partial t} + \frac{\partial (u^2)}{\partial x} + \frac{\partial (uv)}{\partial y} + \frac{\partial (uw)}{\partial z}\right) = -\frac{\partial p}{\partial x} + \mu \Delta u \tag{3}$$

$$\rho\left(\frac{\partial v}{\partial t} + \frac{\partial (uv)}{\partial x} + \frac{\partial (v^2)}{\partial y} + \frac{\partial (vw)}{\partial z}\right) = -\frac{\partial p}{\partial y} + \mu \Delta v \tag{4}$$

$$\rho\left(\frac{\partial w}{\partial t} + \frac{\partial (uw)}{\partial x} + \frac{\partial (vw)}{\partial y} + \frac{\partial (w^2)}{\partial z}\right) = -\frac{\partial p}{\partial z} + \mu \Delta w + \rho g \tag{5}$$

The governing system (1)–(5) for unknowns $\boldsymbol{u}$, p and ρ is sometimes called the Non-homogeneous (incompressible) Navier-Stokes equations.

The small perturbation approximation Now we will assume that the pressure and density fields are perturbation of hydrostatic equilibrium state, i.e.:

$$\rho(x,y,z,t) = \rho_0(z) + \rho'(x,y,z,t) \tag{6}$$

$$p(x,y,z,t) = p_0(z) + p'(x,y,z,t) \tag{7}$$

where the background density and pressure fields are linked by the hydrostatic relation:

$$\frac{\partial p_0}{\partial z} = \rho_0 g. \tag{8}$$

The small perturbation approximation of momentum equations is obtained by introducing the above decomposition of density and pressure into the momentum equations (3), (4) and (5). The density perturbation ρ' is neglected on the left–hand side while on the right–hand side it is retained. On the right–hand side we have removed the hydrostatic pressure using the relation (8) and the fact that according to (7) the horizontal parts of the background pressure gradient are zero.

$$\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} = 0 \tag{9}$$

$$\frac{\partial \rho'}{\partial t} + \frac{\partial(\rho' u)}{\partial x} + \frac{\partial(\rho' v)}{\partial y} + \frac{\partial(\rho' w)}{\partial z} = -w\frac{\partial \rho_0}{\partial z} \tag{10}$$

$$\frac{\partial u}{\partial t} + \frac{\partial(u^2)}{\partial x} + \frac{\partial(uv)}{\partial y} + \frac{\partial(uw)}{\partial z} = \frac{1}{\rho_0}\left(-\frac{\partial p'}{\partial x} + \mu \Delta u\right) \tag{11}$$

$$\frac{\partial v}{\partial t} + \frac{\partial(uv)}{\partial x} + \frac{\partial(v^2)}{\partial y} + \frac{\partial(vw)}{\partial z} = \frac{1}{\rho_0}\left(-\frac{\partial p'}{\partial y} + \mu \Delta v\right) \tag{12}$$

$$\frac{\partial w}{\partial t} + \frac{\partial(uw)}{\partial x} + \frac{\partial(vw)}{\partial y} + \frac{\partial(w^2)}{\partial z} = \frac{1}{\rho_0}\left(-\frac{\partial p'}{\partial z} + \mu \Delta w + \rho' g\right) \tag{13}$$

This model is in 2D (x–z) version used for all the simulations presented in this work.

3 Numerical Methods

Two different numerical methods were chosen to perform a comparative study allowing for cross-comparison of results. The first method is the AUSM finite–volume scheme. For comparison, the compact finite–difference schemes were implemented.

3.1 AUSM Finite–Volume Scheme

This method has been chosen to represent the modern high resolution finite–volume schemes. This particular scheme was previously used for the simulation of stratified flow and compared successfully with other methods in [2], [3].

Space discretizations For numerical solution the artificial compressibility method in dual time was used. Continuity equation is rewritten in the form (in 2D, x–z plane)

$$\frac{\partial p}{\partial \tau} + \beta^2 (\frac{\partial u}{\partial x} + \frac{\partial w}{\partial z}) = 0$$

where τ is the artificial time. The equations (9)–(13) rewritten in the 2D conservative form are

$$PW_t + F(W)_x + G(W)_y = S(W).$$

Here $W = [\rho', u, v, p]^T$, $F = F^i - \nu F^v$ and $G = G^i - \nu G^v$ contain the inviscid fluxes F^i, G^i and viscous fluxes F^v and G^v, S is the source term, and $P = diag(1, 1, 1, 0)$. the fluxes and the source term are

$$F^i(W) = [\rho' u, u^2 + p, uw, \beta^2 u]^T, \qquad G^i(W) = [\rho' w, uw, w^2 + p, \beta^2 w]^T, \quad (14)$$

$$F^v(W) = [0, u_x, w_x, 0]^T, \;\; G^v(W) = [0, u_y, w_y, 0]^T, \;\; S(W) = [-w d\rho_0/dz, 0, \rho' g, 0]^T.$$

The finite volume AUSM scheme was used for spatial discretizations of the inviscid fluxes:

$$\int_{\Omega} (F^i_x + G^i_y) dS = \oint_{\partial\Omega} (F^i n_x + G^i n_y) dl \approx \sum_{k=1}^{4} \left[u_n \begin{pmatrix} \varrho \\ u \\ w \\ \beta^2 \end{pmatrix}_{L/R} + p \begin{pmatrix} 0 \\ n_x \\ n_y \\ 0 \end{pmatrix} \right] \Delta l_k \quad (15)$$

where n is normal vector, u_n is normal velocity vector, and $(q)_{L/R}$ are quantities on left/right hand side of the face respectively. These quantities are computed using MUSCL reconstruction with Hemker–Koren limiter.

$$q_R = q_{i+1} - \frac{1}{2}\delta_R \quad q_L = q_i + \frac{1}{2}\delta_L$$

$$\delta_{L/R} = \frac{a_{L/R}(b^2_{L/R} + 2) + b_{L/R}(2a^2_{L/R} + 1)}{2a^2_{L/R} + 2b^2_{L/R} - a_{L/R} b_{L/R} + 3}$$

$$a_R = q_{i+2} - q_{i+1} \quad a_L = q_{i+1} - q_i \quad b_R = q_{i+1} - q_i \quad b_L = q_i - q_{i-1}$$

The viscous fluxes are discretized in central way on a dual mesh. This scheme is formally of the second order of accuracy in space.

Time integration For the finite–volume AUSM scheme a fully unsteady solver was used. The dual time stepping approach was adopted, so the separate time–discretizations were needed for physical and artificial time. The derivative with respect to the physical time t is discretized by the second order BDF formula,

$$P\frac{3W^{n+1}-4W^{n}+W^{n-1}}{2\Delta t}+F_x^{n+1}(W)+G_y^{n+1}(W)=S^{n+1} \tag{16}$$

$$Rez^{n+1}(W)=P(\frac{3}{2\Delta t}W^{n+1}-\frac{2}{\Delta t}W^{n}+\frac{1}{2\Delta t}W^{n-1})+F_x^{n+1}(W)+G_y^{n+1}(W)-f^{n+1}-S^{n+1}.$$

Arising system of equations is solved by artificial compressibility method in the dual (artificial) time τ by an explicit 3–stage Runge-Kutta method.

3.2 Compact Finite-Difference Schemes

Here again the artificial compressibility method was used. The solver is limited to steady problems solution, employing the time–marching method.

Space discretizations The spatial discretizations used in this work is directly based on the paper [8], where the class of very high order compact finite difference schemes was introduced and analyzed. The main idea used to construct this family of schemes is that instead of approximating the spatial derivatives ϕ' of certain quantity ϕ explicitly from the neighboring values ϕ_i, the (symmetric) linear combination of neighboring derivatives $(\ldots,\phi'_{i-1},\phi'_i,\phi'_{i+1},\ldots)$ is approximated by weighted average of central differences.

The simplest compact finite–difference schemes use the approximation in the form

$$a\,\phi'_{i-1}+\phi'_{i}+a\,\phi'_{i+1}=\alpha_1\frac{\phi_{i+1}-\phi_{i-1}}{2h}+\alpha_2\frac{\phi_{i+2}-\phi_{i-2}}{4h} \tag{17}$$

Here $h=x_i-x_{i-1}$ is the spatial step, while a and α_k are the coefficients determining the specific scheme within the family described by (17). It is easy to see that e.g. for $a=0$, $\alpha_1=1$ and $\alpha_2=0$, the explicit second order central discretizations is recovered. For the simulations presented here, the following coefficients were used:

$$\alpha_1=\frac{2}{3}(a+2) \qquad\qquad \alpha_2=\frac{1}{3}(4a-1). \tag{18}$$

This choice of parameters leads to a one–parametric family of formally fourth order accurate schemes. For $a=0$ the classical explicit fourth order discretizations is recovered, while for $a=0.25$ the well known Padé scheme is obtained.

The above presented schemes are based on central discretizations in space and thus non-physical oscillations can occur in the numerical approximations. A very efficient algorithm for filtering out these high frequency oscillations was proposed in [8]. The low–pass filter (for the filtered values $\widehat{\phi}_i$) can be formulated in a form very similar to (17):

$$b\,\widehat{\phi}_{i-1} + \widehat{\phi}_i + b\,\widehat{\phi}_{i+1} = 2\beta_0\phi_i + \beta_1\frac{\phi_{i+1} + \phi_{i-1}}{2h} + \beta_2\frac{\phi_{i+2} + \phi_{i-2}}{4h} + \dots \quad (19)$$

The filters of different orders could be obtained for various choices of coefficients. Here the sixth order filter with coefficients

$$\beta_0 = \frac{1}{16}(11{+}10b);\ \beta_1 = \frac{1}{32}(15{+}34b);\ \beta_2 = \frac{1}{16}(6b{-}3);\ \beta_3 = \frac{1}{32}(1{-}2b) \quad (20)$$

was used. For other filters see e.g. [12]. The parameter $-0.5 < b < 0.5$ is used to fine–tune the filter.[1] More information on the compact space discretizations can be found in [8], [12], [7].

Temporal discretizations The system of governing Partial Differential Equations was discretized in space using the above described finite–difference technique. This leads to a system of Ordinary Differential Equations for time-evolution of grid values of the vector of unknowns W. Resulting system of ODE's can be solved by a suitable time-integration method. In this study we have used the so called Strong Stability Preserving Runge–Kutta methods [9,11]. The three stage second order SSP Runge–Kutta method was used to obtain the results presented here.

4 Numerical Results

Computational domain The 2D computational domain is selected as a part of wall-bounded half space with low smooth cosine-shaped hill. The hill height is $h = 1m$, while the whole domain has dimensions $90 \times 30\,m$. The numerical simulations were performed on a structured, non-orthogonal wall-fitted grid that has 233×117 points with the minimum cell size in the near-wall region $\Delta z = 0.03m$. The grid is smoothly coarsened from the proximity of the hill towards the far field. The maximum growth of consequent cells is 3%.

Boundary conditions On the *inlet* the velocity profile $\boldsymbol{u} = (u(z), 0, 0)$ is prescribed. The horizontal velocity component u is given by $u(z) = U_0(z/H)^{1/r}$ with $U_0 = 1m/s$ and $r = 40$. Density perturbation ρ' is set to zero, while homogeneous Neumann condition is used for pressure. On the *outlet* the homogeneous Neumann condition is prescribed for all velocity components, as well as for the density perturbations. Pressure is set to a constant. On the *wall* the no-slip conditions are

[1] In order to distinguish between different finite–difference schemes we use the notation $CX_{aaa}FY_{bbb}$ for compact scheme of order X with parameter a = aaa combined with filter of order Y applied with the dumping parameter b = bbb.

used for velocity. Homogeneous Neumann condition is used for pressure and density perturbation. For *free stream* the homogeneous Neumann condition is used for all quantities including pressure and density perturbations.

The background density field is given by $\rho_0(z) = \rho_w + \gamma z$ with $\rho_w = 1.2\,kg\cdot m^{-3}$ and $\gamma = -0.01\,kg\cdot m^{-4}$. The gravity acceleration was set to $g = -50\,m\cdot s^{-2}$ to test the behavior of the model and numerical method for sufficiently high Brunt–Väisälä frequency (i.e. short wavelength). The Reynolds number was in the range 100–500 (i.e. $\mu = 1\cdot 10^{-2} - 2\cdot 10^{-3} kg\cdot m^{-1}\cdot s^{-1}$).

Numerical results The small hill placed at the origin of the coordinate system generates a perturbation in the density field. Due to the buoyancy term in the equation (13), this density perturbation is translated into vertical motion that is superposed to the mean horizontal flow. The gravity waves are best visible in the vertical velocity contours (Fig. 1–4). The same color scale was used in all figures. The results of both schemes are quite close to each other as it is visible from the

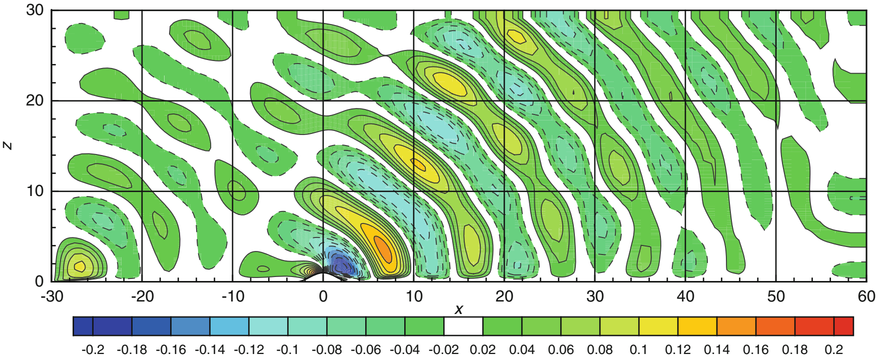


Fig. 1 Vertical velocity contours - $Re = 500$ - Compact scheme $C4_{038}F6_{049}$

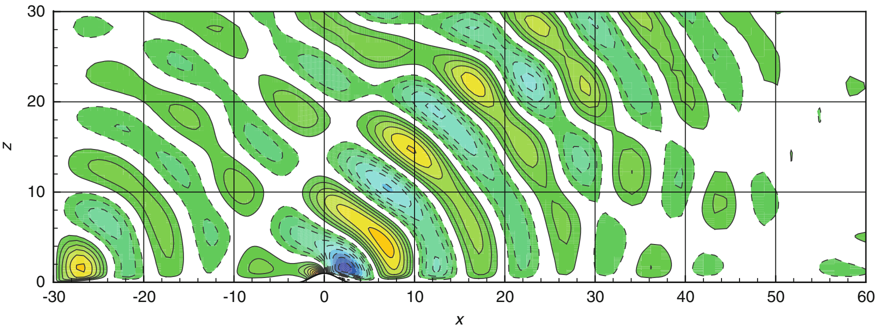


Fig. 2 Vertical velocity contours - $Re = 500$ - AUSM scheme

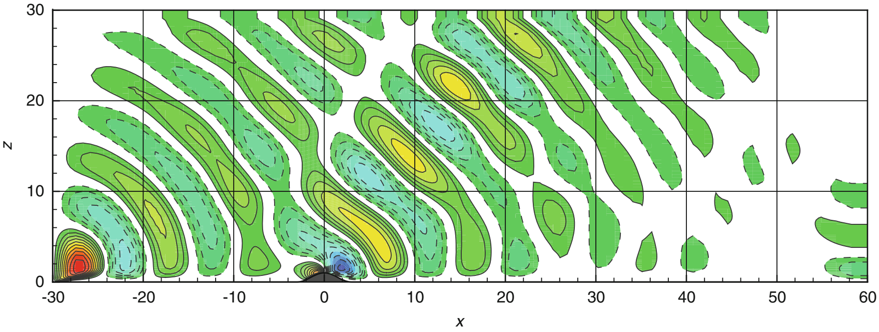


Fig. 3 Vertical velocity contours - $Re = 100$ - Compact scheme $C4_{038}F6_{049}$

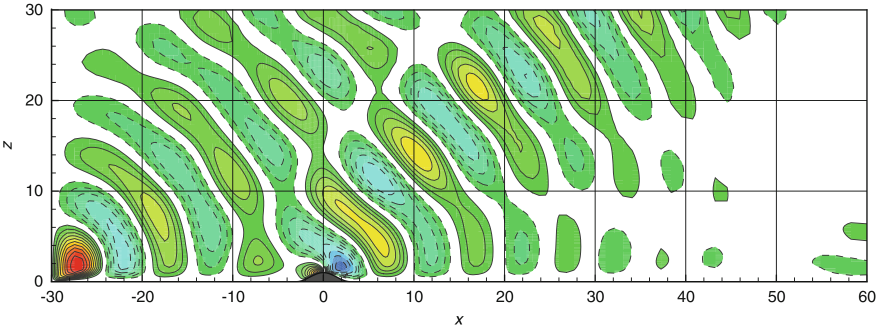


Fig. 4 Vertical velocity contours - $Re = 100$ - AUSM scheme

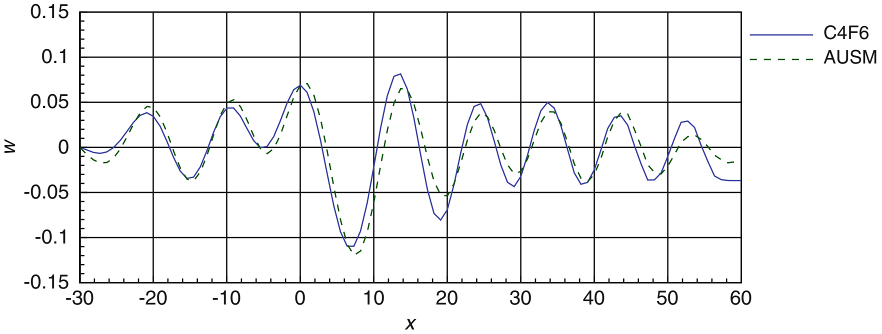


Fig. 5 Vertical velocity profiles comparison for $Re = 500$ in a horizontal cut at the height $z = 10\,m$

vertical velocity profiles shown in the Fig. 5. The basic structure of the results of both numerical methods is very similar. The compact finite difference scheme has clearly an advantage in being able to resolve the gravity waves in the far–field regions where the grid is very coarse.

5 Conclusions and Remarks

The numerical test have demonstrated the ability of both numerical methods to capture the main features of stably stratified flows. The advantage of high order methods was well documented by resolving the gravity waves in the regions of very coarse grid. The numerical simulations have brought some more problems that need to be further explored in detail. From these issues let's mention the the non–physical waves generated and reflected by the artificial boundaries, strong influence of numerical discretization/stabilization techniques on the small amplitude gravity waves, and the grid spacing limits related to Brunt–Väisälä frequency. Some of these problems are discussed in more detail in our recent study [5].

Acknowledgements The financial support for this work was partly provided by the Research Plan *MSM 6840770010* of the Ministry of Education of Czech Republic.

References

1. L. Beneš, T. Bodnár, P. Fraunié, K. Kozel, Numerical modelling of pollution dispersion in 3D atmospheric boundary layer, in: B. Sportisse (Ed.), Air Pollution Modelling and Simulation, Springer Verlag, 2002, pp. 69–78.
2. L. Beneš, J. Fürst, Numerical simulation of stratified flows past a body, in: ENUMATH 2009, Springer, 2009, p. 8p.
3. L. Beneš, J. Fürst, Comparison of the two numerical methods for the stratified flow, in: ICFD 2010 10th Conference on Numerical Methods for Fluid Dynamics, Univ. Reading, 2010, p. 6p.
4. T. Bodnár, L. Beneš, K. Kozel, Numerical simulation of flow over barriers in complex terrain, Il Nuovo Cimento C 31 (5–6) (2008) 619–632.
5. T. Bodnár, L. Beneš, Ph. Fraunié, K. Kozel, Application of Compact Finite-Difference Schemes to Simulations of Stably Stratified Fluid Flows, Preprint submitted to Applied Mathematics and Computation (2011).
6. T. Bodnár, K. Kozel, P. Fraunié, Z. Jaňour, Numerical simulation of flow and pollution dispersion in 3D atmospheric boundary layer, Computing and Visualization in Science 3 (1–2) (2000) 3–8.
7. D. V. Gaitonde, J. S. Shang, J. L. Young, Practical aspects of higher-order numerical schemes for wave propagation phenomena, International Journal for Numerical Methods in Engineering 45 (1999) 1849–1869.
8. S. K. Lele, Compact finite difference schemes with spectral-like resolution, Journal of Computational hysics 103 (1992) 16–42.
9. C. W. Shu, S. Osher, Efficient implementation of essentially non-oscillatory shock-capturing schemes, Journal of Computational Physics 77 (1988) 439–471.
10. I. Sládek, T. Bodnár, K. Kozel, On a numerical study of atmospheric 2D and 3D - flows over a complex topography with forest including pollution dispersion, Journal of Wind Engineering and Industrial Aerodynamics 95 (9–11).
11. R. J. Spiteri, S. J. Ruuth, A new class of optimal high-order strong-stability-preserving time discretization methods, SIAM Journal on Numerical Analysis 40 (2) (2002) 469–491.
12. M. R. Visbal, D. V. Gaitonde, On the use of higher-order finite-difference schemes on curvilinear and deforming meshes, Journal of Computational Physics 181 (2002) 155–185.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Convergence Analysis of the Upwind Finite Volume Scheme for General Transport Problems

Franck Boyer

Abstract This work is devoted to the convergence analysis of the implicit upwind finite volume scheme for the initial and boundary value problem associated to the linear transport equation in any dimension, on general unstructured meshes. We are particularly concerned with the case where the initial and boundary data are in L^∞ and the advection vector field v has low regularity properties, namely $v \in L^1(]0,T[,(W^{1,1}(\Omega))^d)$, with suitable assumptions on its divergence. We prove strong convergence in $L^\infty(]0,T[,L^p(\Omega))$ with $p < +\infty$, of the approximate solution towards the unique weak solution of the problem as well as the strong convergence of its trace.

Keywords Transport equation, upwind finite volume method, renormalized solutions
MSC2010: 35D30, 35L04, 65M08, 65M12

1 Introduction

Let $d \geq 1$, $\Omega \subset \mathbf{R}^d$ a bounded polygonal (or polyhedral) domain, and $T > 0$ given. We are interested here in the following initial and boundary value problem

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho v) + c\rho = 0, \ \ \text{in }]0,T[\times\Omega, \\ \rho(0,\cdot) = \rho_0, \ \ \text{in } \Omega, \\ \rho = \rho^{in}, \ \ \text{on }]0,T[\times\Gamma, \text{ where } (v\cdot\boldsymbol{\nu}) < 0. \end{cases} \tag{1}$$

Franck Boyer
Aix-Marseille Université, LATP, FST Saint-Jérôme, Avenue Escradille Normandie-Niemen,
13397 MARSEILLE Cedex 20, FRANCE, e-mail: fboyer@latp.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_17, © Springer-Verlag Berlin Heidelberg 2011

We consider the following assumptions for the data

$$c \in L^1(]0,T[\times\Omega), \tag{2}$$

$$v \in L^1(]0,T[,(W^{1,1}(\Omega))^d), \text{ and } (v\cdot\boldsymbol{\nu}) \in L^\alpha(]0,T[\times\Gamma), \text{ for some } \alpha > 1, \tag{3}$$

$$(c+\operatorname{div} v)^- \in L^1(]0,T[,L^\infty(\Omega)), \text{ and } (\operatorname{div} v)^+ \in L^1(]0,T[,L^\infty(\Omega)), \tag{4}$$

where x^+ and x^- stands for the positive and negative parts of any real number x. Associated to the vector field v, we introduce the measure $d\mu_v = (v\cdot\boldsymbol{\nu})\,dx\,dt$ on $]0,T[\times\Gamma$ and we denote by $d\mu_v^+$ (resp. $d\mu_v^-$) its positive (resp. negative) part in such a way that $|d\mu_v| = d\mu_v^+ + d\mu_v^-$. The support of $d\mu_v^+$ (resp. $d\mu_v^-$) is the outflow (resp. inflow) part of the boundary. In this framework, Problem (1) is well-posed in the following sense.

Theorem 1 (Existence and uniqueness). *We assume that assumptions* (2), (3), (4) *hold. For any* $\rho^0 \in L^\infty(\Omega)$ *and* $\rho^{in} \in L^\infty(]0,T[\times\Gamma, d\mu_v^-)$, *there exists a unique weak solution* $(\rho,\gamma(\rho)) \in L^\infty(]0,T[\times\Omega)\times L^\infty(]0,T[\times\Gamma,|d\mu_v|)$ *of* (1) *in the sense that*

$$\int_0^T\int_\Omega \rho(\partial_t\phi + v\cdot\nabla\phi - c\phi)\,dx\,dt - \int_0^T\int_\Gamma \gamma(\rho)\phi(v\cdot\boldsymbol{\nu})\,dx\,dt + \int_\Omega \rho^0\phi(0,.)\,dx = 0, \quad \forall\phi\in\mathscr{C}_c^1([0,T[\times\overline{\Omega}), \tag{5}$$

with $\gamma(\rho) = \rho^{in}$, $d\mu_v^-$*-almost everywhere.*

Moreover, $\rho \in \mathscr{C}^0([0,T],L^p(\Omega))$ *for any* $p < +\infty$, *and we have*

$$\|\rho\|_{L^\infty(]0,T[\times\Omega)} \le \max(\|\rho_0\|_{L^\infty}, \|\rho^{in}\|_{L^\infty})e^{\int_0^T \|(c+\operatorname{div} v)^-\|_{L^\infty(\Omega)}\,dt}.$$

Finally, the following **renormalization property** *holds : for any smooth function* $\beta : \mathbf{R} \mapsto \mathbf{R}$, *the following system holds in the weak sense*

$$\begin{cases} \partial_t\beta(\rho) + \operatorname{div}(\beta(\rho)v) + c\beta'(\rho)\rho + (\operatorname{div} v)(\beta'(\rho)\rho - \beta(\rho)) = 0, & \textit{in }]0,T[\times\Omega,\\ \beta(\rho)(0,.) = \beta(\rho^0), \\ \gamma(\beta(\rho)) = \beta(\gamma(\rho)), \quad \textit{on }]0,T[\times\Gamma. \end{cases}$$

This result originates first from DiPerna-Lions theory [10] in the case when $v\cdot\boldsymbol{\nu} = 0$ on the boundary. The initial and boundary value problem is studied in [4] in the case $c = \operatorname{div} v = 0$ and for a smooth domain Ω, whereas the general case is studied in [6]. Note that the assumptions we consider on the vector field v are almost minimal to prove the well-posedness of the transport problem. In fact, for $v \in L^1(]0,T[,(\mathrm{BV}(\Omega))^d)$, it is known that many of the results of the renormalized solutions theory still hold [1], but we will not consider this case here.

The main aim of this work is to prove, in the above weak regularity framework, the convergence of the upwind finite volume method in a strong sense. The detailed proofs of all the results given here can be found in [5]. To our knowledge, the only available result in this framework is a L^∞ weak-$\star$ convergence result, in the case where $v \cdot \boldsymbol{\nu} = 0$ on $\partial\Omega$, which can be found in [12]. When the transport vector field v is more regular (say Lipschitz continuous) and when $v \cdot \boldsymbol{\nu} = 0$ on the boundary, the upwind finite volume scheme was studied in many references (see for instance [3, 7–9, 13, 14]). In summary, it is known that the convergence rate of the scheme is $\frac{1}{2}$ as soon as the initial data is discontinuous, even for regular meshes, whereas this rate is 1 for smooth data and regular meshes.

2 The finite volume setting

2.1 Notation

We introduce here the main notation we need to define and analyse the finite volume method, following quite closely the notation introduced for instance in [11].

A finite volume mesh of the domain Ω is a set $\mathscr{T} = (K)_{K\in\mathscr{T}}$ of closed connected polygonal subsets of $\mathbf{R}^d$, with disjoint interiors and such that $\overline{\Omega} = \bigcup_{K\in\mathscr{T}} K$. The boundary of each control volume $K \in \mathscr{T}$ can be written as the union of a finite number of edges/faces (we will use the word "edge" even for $d > 2$) which are closed connected sets of dimension $d-1$. We denote by $\mathscr{E}_K$ the set of the faces/edges of K. We assume that for any K, L such that $K \neq L$ and $K \cap L$ is of co-dimension 1, then $K \cap L \in \mathscr{E}_K \cap \mathscr{E}_L$, in that case the corresponding face is denoted by $K|L$. The set of all the faces in the mesh is denoted by $\mathscr{E}$ and $\mathscr{E}_{\text{bd}}$ denote the subset of the faces which are included in the boundary $\partial\Omega$, $\mathscr{E}_{\text{int}} = \mathscr{E} \setminus \mathscr{E}_{\text{bd}}$ the set of the interior faces.

- For each $K \in \mathscr{T}$, and $\sigma \in \mathscr{E}_K$, we denote by $\boldsymbol{\nu}_{K\sigma}$ the unit outward normal vector to K on σ. If $\sigma = K|L \in \mathscr{E}_{\text{int}}$, we we observe that $\boldsymbol{\nu}_{K\sigma} = -\boldsymbol{\nu}_{L\sigma}$.
- We will denote by $|K|$ (resp. $|\sigma|$) the d-dimensional Lebesgue measure of the control volume K (resp. the $d-1$ dimensional measure of the face σ).
- The diameter of a control volume K shall be denoted by d_K and the size of the mesh is defined by $h_{\mathscr{T}} = \max_{K\in\mathscr{T}} d_K$.

We will need to measure the regularity of the mesh. To this end, we denote by $\text{reg}(\mathscr{T})$, the smallest positive number such that

$$\|f\|_{L^1(\partial K)} \leq \frac{\text{reg}(\mathscr{T})}{d_K} \|f\|_{W^{1,1}(K)}, \quad \forall K \in \mathscr{T}, \forall f \in W^{1,1}(K). \tag{6}$$

In the convergence results given below we shall assume that $\text{reg}(\mathscr{T})$ remains bounded as $h_{\mathscr{T}} \to 0$, which amounts to assume that the control volumes do not degenerate when one refines the mesh.

2.2 Definition of the scheme

Let us first define the discretization of the data needed to define the scheme.

- For any $K \in \mathscr{T}$, $n \in [\![0, N-1]\!]$, we define

$$c_K^n = \frac{1}{\delta t|K|}\int_{t^n}^{t^{n+1}}\int_K c\,dx\,dt, \text{ and } v_{K\sigma}^n = \frac{1}{\delta t|\sigma|}\int_{t^n}^{t^{n+1}}\int_\sigma (v\cdot \boldsymbol{\nu}_{K\sigma})\,dx\,dt,\ \forall \sigma \in \mathscr{E}_K.$$

  Furthermore, if $\sigma \in \mathscr{E}_{\text{int}}$, with $\sigma = K|L$ we shall use the notation $v_{KL}^n = v_{K\sigma}^n = -v_{L\sigma}^n$, and if $\sigma \in \mathscr{E}_K \cap \mathscr{E}_{\text{bd}}$ we will denote $v_\sigma^n = v_{K\sigma}^n$.
- For any boundary edge $\sigma \in \mathscr{E}_{\text{bd}}$ and any $n \in [\![0, N-1]\!]$, we define

$$\rho_\sigma^{in,n+1} = \frac{1}{\delta t|\sigma|}\int_{t^n}^{t^{n+1}}\int_\sigma \rho^{in}\,dx\,dt. \tag{7}$$

  Notice that ρ^{in} is *a priori* only given $d\mu_v^-$-almost everywhere so that in this formula we need, in fact, to consider an extension of ρ^{in} in $L^\infty(]0,T[\times\Gamma)$.

The implicit upwind finite volume scheme we consider is the following: Find approximate values $\{\rho_K^n, n \in [\![0, N]\!], K \in \mathscr{T}\}$ such that

$$\begin{cases} |K|\dfrac{\rho_K^{n+1}-\rho_K^n}{\delta t} + \displaystyle\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\text{int}}} |\sigma|\left(\left(v_{K\sigma}^n\right)^+\rho_K^{n+1} - \left(v_{K\sigma}^n\right)^-\rho_L^{n+1}\right) \\ \quad + \displaystyle\sum_{\sigma\in\mathscr{E}_K\cap\mathscr{E}_{\text{bd}}} |\sigma| v_{K\sigma}^n\rho_\sigma^{n+1} + |K|c_K^n\rho_K^{n+1} = 0, \quad \forall n \in [\![0, N-1]\!], \forall K \in \mathscr{T}, \\ \rho_K^0 = \dfrac{1}{|K|}\displaystyle\int_K \rho^0\,dx, \quad \forall K \in \mathscr{T}, \\ \rho_\sigma^{n+1} = \rho_\sigma^{in,n+1}, \quad \forall n \in [\![0, N-1]\!], \forall \sigma \in \mathscr{E}_{\text{bd}}, \quad \text{s.t. } v_{K\sigma}^n \le 0, \\ \rho_\sigma^{n+1} = \rho_K^{n+1}, \quad \forall n \in [\![0, N-1]\!], \forall \sigma \in \mathscr{E}_{\text{bd}}, \quad \text{s.t. } v_{K\sigma}^n > 0. \end{cases} \tag{8}$$

Note that the boundary data ρ^{in} is only taken into account on the boundary edges such that $v_{K\sigma}^n \le 0$. Those edges are not necessarily included in the inflow boundary, that is the support of the measure $d\mu_v^-$. That's the reason why we need to extend the definition of ρ^{in} to the whole boundary $]0, T[\times\Gamma$.

2.3 Existence and uniqueness

The first result we can prove is the following existence and uniqueness result.

Theorem 2. *Assume that* (2),(3) *and* (4) *hold. There exists* $\delta t_{\max} > 0$ *(depending only on* $(c + \operatorname{div} v)^-$*) such that for any* $\delta t \leq \delta t_{\max}$*, any mesh* $\mathscr{T}$ *and any data* ρ^{in}, ρ^0*, there exists an unique solution to the scheme* (8).

Moreover, the approximate solution is non-negative as soon as the data is. Finally, we have the following a priori bound

$$\max_{\substack{K\in\mathscr{T}\\ n\in[\![0,N]\!]}} |\rho_K^n| \leq \max(\|\rho^0\|_{L^\infty}, \|\rho^{in}\|_{L^\infty}) \exp\left(2\int_0^T \|(c + \operatorname{div} v)^-\|_{L^\infty}\, dt\right). \quad (9)$$

In the case of the pure transport problem, that is when $c = -\operatorname{div} v$, we find that $\delta t_{\max} = +\infty$. From now on we will denote by $\rho_{\mathscr{T},\delta t}$ (and its trace $(\gamma\rho_{\mathscr{T},\delta t})$) the piecewise constant function build upon the approximate solution as follows

$$\rho_{\mathscr{T},\delta t} = \sum_{n=0}^{N-1}\sum_{K\in\mathscr{T}} \rho_K^{n+1}\mathbf{1}_{]t^n,t^{n+1}[\times K}, \quad \gamma(\rho_{\mathscr{T},\delta t}) = \sum_{n=0}^{N-1}\sum_{\sigma\in\mathscr{E}_{\mathrm{bd}}} \rho_K^{n+1}\mathbf{1}_{]t^n,t^{n+1}[\times\sigma},$$

where, in the last sum, K is the unique control volume in $\mathscr{T}$ such that $\sigma \in \mathscr{E}_K$.

3 Convergence analysis

3.1 Uniform in time strong convergence

The main result of this work is the following

Theorem 3. *Assume that* (2), (3) *and* (4) *hold. Let* $\mathrm{reg}_{\max} > 0$ *be given and consider a family of meshes and time steps, such that* $(\delta t, h_{\mathscr{T}}) \to 0$ *and satisfying the bound*

$$\max\left(\mathrm{reg}(\mathscr{T}), \max_{K\in\mathscr{T}} \frac{\delta t}{d_K}\right) \leq \mathrm{reg}_{\max}.$$

Then, for any bounded data ρ^0 *and* ρ^{in}*, we have the following convergences*

$$\rho_{\mathscr{T},\delta t} \xrightarrow[(\delta t,h_{\mathscr{T}})\to 0]{} \rho, \ \ in\ \ L^\infty(]0,T[, L^p(\Omega)),$$

$$\gamma(\rho_{\mathscr{T},\delta t}) \xrightarrow[(\delta t,h_{\mathscr{T}})\to 0]{} \gamma(\rho), \ \ in\ \ L^p(]0,T[\times\Gamma, |d\mu_v|), \ \ \forall p < +\infty,$$

where $(\rho, \gamma(\rho))$ *is the unique solution to* (5).

We want to emphasize the fact that the convergence of $\rho_{\mathscr{T},\delta t}$ is uniform in time with values in $L^p(\Omega)$. We describe now the main steps of the proof of this result.

- Using the *a priori* L^∞ bound (9), we can extract a subsequence of $\rho_{\mathcal{T},\delta t}$ (resp. $\gamma(\rho_{\mathcal{T},\delta t})$) which weak-$\star$ converges towards a limit denoted by ρ (resp. g) in $L^\infty(]0,T[\times\Omega)$ (resp. in $L^\infty(]0,T[\times\Gamma)$).
  - We prove that $g = \rho^{in}\ d\mu_v^-$-almost-everywhere.
  - We prove that (ρ, g) satisfy the weak formulation of the problem.
  - Since the weak solution $(\rho, \gamma(\rho))$ is unique, we deduce the weak-$\star$ convergence of the whole sequence of approximate solutions.
- For any $\varepsilon > 0$ we construct a smooth function ρ^ε associated to ρ such that
  - $\|\rho^\varepsilon\|_{L^\infty} \leq \|\rho\|_{L^\infty}$;
  - ρ^ε converges to ρ in $\mathscr{C}^0([0,T], L^p(\Omega))$, for any $p < +\infty$;
  - $\gamma(\rho^\varepsilon)$ converges to $\gamma(\rho)$ in $L^p(]0,T[\times\Gamma, |d\mu_v|)$ for any $p < +\infty$;
  - ρ^ε solves
  $$\partial_t\rho^\varepsilon + \operatorname{div}(\rho^\varepsilon v) + c\rho^\varepsilon = R_\varepsilon,$$
  where $R_\varepsilon \in L^1(]0,T[\times\Omega)$ tends to 0 in L^1.

  Following [10], this sequence is built by convolution with a mollifier and the regularity assumptions on v and c let us prove the property of the remainder term R_ε (Friedrichs commutator lemma). Nevertheless, since the boundary of Ω is not characteristic for v, the argument has to be adapted (see [2,4,6]).
- We define the discretization of ρ^ε to be
  $$\rho^\varepsilon_{\mathcal{T},\delta t} = \sum_{n=0}^{N-1}\delta t \sum_{K\in\mathcal{T}} \rho^\varepsilon(t^{n+1}, x_K)\mathbf{1}_{]t^n,t^{n+1}[\times K},$$
  where x_K is an arbitrary point in K.
- We use now the triangle inequality to get
  $$\begin{aligned}\|\rho_{\mathcal{T},\delta t} - \rho\|_{L^\infty(]0,T[,L^2(\Omega))} &\leq \|\rho_{\mathcal{T},\delta t} - \rho^\varepsilon_{\mathcal{T},\delta t}\|_{L^\infty(]0,T[,L^2(\Omega))} \\ &+ \|\rho^\varepsilon_{\mathcal{T},\delta t} - \rho^\varepsilon\|_{L^\infty(]0,T[,L^2(\Omega))} + \|\rho^\varepsilon - \rho\|_{L^\infty(]0,T[,L^2(\Omega))}. \quad (10)\end{aligned}$$
  In this inequality, the last term converges to 0 when $\varepsilon \to 0$ by construction of ρ^ε. The second term can be easily estimated by $C\|\rho^\varepsilon\|_{W^{1,\infty}}(h_{\mathcal{T}} + \delta t)$. Obviously $\|\rho^\varepsilon\|_{W^{1,\infty}}$ blows up when $\varepsilon \to 0$, but if ε is fixed, this term converges to 0 when $(\delta t, h_{\mathcal{T}}) \to 0$.
- Finally, we are led to compare the approximate solution $\rho_{\mathcal{T},\delta t}$ and the projection $\rho^\varepsilon_{\mathcal{T},\delta t}$ of ρ^ε. This can be done by writing the discrete equations satisfied by $\rho^\varepsilon_{\mathcal{T},\delta t}$ in a form which is similar to that of (8) with additional remainder terms in the right-hand side.

  We subtract the two set of equations and we perform an usual $L^\infty(]0,T[, L^2(\Omega))$ estimate of the difference $\rho_{\mathcal{T},\delta t} - \rho^\varepsilon_{\mathcal{T},\delta t}$. It can then be proved that all the contribution of the remainder terms can be controlled by two types of quantities:

- some of them tend to zero when $\varepsilon \to 0$, independently on δt and $h_{\mathcal{T}}$.
- some of them tend to zero when $(\delta t, h_{\mathcal{T}}) \to 0$, as soon as ε is fixed.

The conclusion is then clear. The main tools we use in these estimates are

- The fact that, by the Friedrichs Lemma, we have $\|R_\varepsilon\|_{L^1} \to 0$.
- The weak $L^2(H^1)$ estimate satisfied by the approximate solution which reads

$$\sum_{n=0}^{N-1} \delta t \sum_{\sigma \in \mathcal{E}_{\mathrm{bd}}} |\sigma| |v_{K\sigma}^n| (\rho_K^{n+1} - \rho_\sigma^{n+1})^2 + \sum_{n=0}^{N-1} \delta t \sum_{\substack{\sigma \in \mathcal{E}_{\mathrm{int}} \\ \sigma = K|L}} |\sigma| |v_{KL}^n| (\rho_L^{n+1} - \rho_K^{n+1})^2 \le M, \quad (11)$$

  for some $M > 0$ which only depends on the data. This estimate corresponds in the linear case to the so-called "weak BV estimate" for nonlinear hyperbolic equations (see [11]).
- The density of the set of smooth vector fields in $L^1(]0, T[, (W^{1,1}(\Omega))^d)$.

3.2 Discrete renormalization property and consequences

We now state a result which is a discrete counter-part of the renormalization property given in Theorem 1 for the weak solution of the continuous problem. This kind of result might be important when studying the coupling between problem (1) and some other equations (in the nonhomogeneous incompressible Navier-Stokes system for instance).

Theorem 4. *For any function $\beta : \mathbf{R} \mapsto \mathbf{R}$ which is continuous and piecewise $\mathcal{C}^1$, the approximate solution $\rho_{\mathcal{T},\delta t}$ satisfy the following set of equations*

$$|K| \frac{\beta(\rho_K^{n+1}) - \beta(\rho_K^n)}{\delta t} + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{int}}} |\sigma| \big(v_{K\sigma}^{n+} \beta(\rho_K^{n+1}) - v_{K\sigma}^{n-} \beta(\rho_L^{n+1}) \big) + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{bd}}} |\sigma| v_{K\sigma}^n \beta(\rho_\sigma^{n+1}) + |K| c_K^n \beta'(\rho_K^{n+1}) \rho_K^{n+1}$$
$$+ |K| (\operatorname{div} v)_K^n \big(\beta'(\rho_K^{n+1}) \rho_K^{n+1} - \beta(\rho_K^{n+1}) \big) = |K| R_K^{n+1}, \ \ \forall n \in [\![0, N-1]\!], \forall K \in \mathcal{T}, \quad (12)$$

where the remainder term $(R_K^{n+1})_{K\in\mathscr{T},n\in[\![0,N-1]\!]}$ *strongly converges towards zero in* $L^1(]0,T[\times\Omega)$*, that is*

$$\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathscr{T}}|K||R_K^{n+1}|\xrightarrow[(\delta t,h_{\mathscr{T}})\to 0]{}0.$$

Furthermore, when β *is convex we have* $R_K^{n+1}\leq 0,\quad\forall K\in\mathscr{T},\forall n\in[\![0,N-1]\!]$.

Note that this result holds for any arbitrary choice of the value of β' at singular points. We can deduce from Theorem 4 many properties of the approximate solution. For instance, we can prove:

- For any $\alpha\in\mathbf{R}\setminus\{0\}$, we have

$$\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathscr{T}}|K||c_K^n+(\operatorname{div}v)_K^n|\mathbf{1}_{\{\rho_K^{n+1}=\alpha\}}\xrightarrow[(\delta t,h_{\mathscr{T}})\to 0]{}0.$$

  This is the discrete counter part of the following property of the weak solution of the problem

$$c+\operatorname{div}v=0,\quad\text{for almost every }(t,x)\text{ in the level set }\{\rho=\alpha\}.$$

- The total numerical dissipation term associated with the upwind discretization (that is the left-hand side term in (11)) tends to 0 when $(\delta t,h_{\mathscr{T}})\to 0$. This is an improvement of the weak $L^2(H^1)$ estimate which only says that this quantity is bounded.

Acknowledgements The author wishes to thank T. Gallouët and R. Herbin for many stimulating discussions on the topic of this work.

References

1. Ambrosio, L.: Transport equation and Cauchy problem for BV vector fields. Invent. Math. **158**(2), 227–260 (2004)
2. Blouza, A., Le Dret, H.: An up-to-the-boundary version of Friedrichs's lemma and applications to the linear Koiter shell model. SIAM J. Math. Anal. **33**(4), 877–895 (electronic) (2001)
3. Bouche, D., Ghidaglia, J.M., Pascal, F.: Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation. SIAM J. Numer. Anal. **43**(2), 578–603 (electronic) (2005)
4. Boyer, F.: Trace theorems and spatial continuity properties for the solutions of the transport equation. Differential Integral Equations **18**(8), 891–934 (2005)
5. Boyer, F.: Analysis of the upwind finite volume method for general initial and boundary value transport problems. submitted (2011). http://hal.archives-ouvertes.fr/hal-00559586/fr/
6. Boyer, F., Fabrie, P.: Elements of analysis for the study of some models of incompressible viscous flows. in preparation. Springer-Verlag (2011)

7. Delarue, F., Lagoutière, F.: Probabilistic analysis of the upwind scheme for transport equations. Archive for Rational Mechanics and Analysis **199**, 229–268 (2011)
8. Desprès, B.: Convergence of non-linear finite volume schemes for linear transport. In: Notes from the XIth Jacques-Louis Lions Hispano-French School on Numerical Simulation in Physics and Engineering (Spanish), pp. 219–239. Grupo Anal. Teor. Numer. Modelos Cienc. Exp. Univ. Cádiz, Cádiz (2004)
9. Desprès, B.: Lax theorem and finite volume schemes. Math. Comp. **73**(247), 1203–1234 (electronic) (2004)
10. DiPerna, R., Lions, P.L.: Ordinary differential equations, transport theory and Sobolev spaces. Invent. Math. **98**(3), 511–547 (1989)
11. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of numerical analysis, Vol. VII, Handb. Numer. Anal., VII, pp. 713–1020. North-Holland, Amsterdam (2000)
12. Fettah, A.: Analyse de modèles en mécanique des fluides compressibles. Ph.D. thesis, Université de Provence (2011)
13. Merlet, B.: L^∞- and L^2-error estimates for a finite volume approximation of linear advection. SIAM J. Numer. Anal. **46**(1), 124–150 (2007/08)
14. Merlet, B., Vovelle, J.: Error estimate for finite volume scheme. Numer. Math. **106**(1), 129–155 (2007)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Low Degree Non–Conforming Approximation of the Steady Stokes Problem with an Eddy Viscosity

F. Boyer, F. Dardalhon, C. Lapuerta, and J.-C. Latché

Abstract In the context of Large Eddy Simulation, the use of a turbulence model brings the question of the implementation of the eddy–viscosity. In this communication, we propose to assess the discretization of the diffusive term based on a low–order non–conforming finite element. For this, we build a manufactured solution of the incompressible steady Stokes problem, for which the turbulent viscosity is given either by the Smagorinsky or WALE models. Numerical tests are performed for both models with the finite element approximation and the MAC scheme.

Keywords Large Eddy Simulation, WALE and Smagorinsky models, incompressible steady Stokes equations, low-order finite element approximation
MSC2010:65N30, 76D05

1 Introduction

In the context of turbulence modelling, there is an increasing interest in the Large Eddy Simulation approach (LES), resulting from the augmentation of the computer resources. LES modelling solves large turbulent structures, while small–scale effects are modelled (see [2, 13]). The approach consists in averaging the Navier–Stokes equations in space (by convolution), and then commuting this filtering operation (denoted with the overbar symbol) with space and time derivatives. This

F. Dardalhon, C. Lapuerta, and J.-C. Latché
Institut de Radioprotection et de Sûreté Nucléaire (IRSN), BP3 - 13115 Saint Paul–lez–Durance CEDEX, France, e-mail: [fanny.dardalhon,celine.lapuerta, jean--claude.latche]@irsn.fr

F. Boyer and F. Dardalhon
Université Paul Cézanne, LATP, FST Saint–Jérôme, Case Cour A, Avenue Escadrille Normandie–Niemen, 13397 Marseille Cedex 20, France, e-mail: fboyer@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_18, © Springer-Verlag Berlin Heidelberg 2011

yields balance equations, which keep the same form as the original ones, for the resolved (large scales) velocity $\overline{\mathbf{u}}$ and pressure $\overline{p}$. Due to the presence of the nonlinear convection term, the unclosed quantity $-\mathrm{div}(\overline{\mathsf{T}})$, with $\overline{\mathsf{T}} = \overline{\mathbf{u}\mathbf{u}} - \overline{\mathbf{u}}\,\overline{\mathbf{u}}$ appears at the right–hand side, and must be modelled, *i.e.* recast as a function of the unknowns $\overline{\mathbf{u}}$ and $\overline{p}$.

The key issue in the LES approach is thus to find a suitable expression for the subgrid–scale tensor $\overline{\mathsf{T}}$. A common assumption is to suppose a proportional relation between $\overline{\mathsf{T}}$ and the deformation tensor $\overline{\mathsf{S}}$:

$$\overline{\mathsf{T}} = -2\,\nu_t\,\overline{\mathsf{S}}, \quad \text{with} \quad \overline{\mathsf{S}}_{i,j} = \frac{1}{2}(\partial_j \overline{\mathbf{u}}_i + \partial_i \overline{\mathbf{u}}_j), \quad \forall i,j \in \{1,\cdots,d\},$$

the scalar ν_t being referred to as the turbulent viscosity. We propose to study here two subgrid–scale models often encountered in the literature, namely the Smagorinsky [14] and WALE [10] models.

The Smagorinsky model is the most frequently used because of its quite simple form and reads:

$$\nu_t = (C_s\,\overline{\Delta})^2 \sqrt{2\,\mathrm{Trace}(\overline{\mathsf{S}}\,\overline{\mathsf{S}}^T)} = (C_s\,\overline{\Delta})^2 \left(2 \sum_{i,j\in\{1,\cdots,d\}} \overline{\mathsf{S}}_{i,j}\,\overline{\mathsf{S}}_{i,j}\right)^{\frac{1}{2}}, \tag{1}$$

where C_s is a constant adjusted as a function of the flow and $\overline{\Delta}$ is the cut–off length scale, usually identified to a characteristic size of the cell. However, the viscosity obtained in this way behaves like $\mathcal{O}(1)$ near a wall, contrary to the scaling $\nu_t = \mathcal{O}(y^3)$, where y stands for the distance to the wall, which may be inferred by asymptotic analysis [13]. So this model dissipates the large scales too much near a wall.

The WALE model (Wall Adaptating Local Eddy–viscosity) aims at solving this problem and reads:

$$\nu_t = (C_w\overline{\Delta})^2 \frac{\left(\sum_{i,j} \overline{\varsigma}_{i,j}\,\overline{\varsigma}_{i,j}\right)^{3/2}}{\left(\sum_{i,j} \overline{\mathsf{S}}_{i,j}\,\overline{\mathsf{S}}_{i,j}\right)^{5/2} + \left(\sum_{i,j} \overline{\varsigma}_{i,j}\,\overline{\varsigma}_{i,j}\right)^{5/4}}, \tag{2}$$

C_w being a real constant adjusted as a function of the flow and

$$\overline{\varsigma} = \frac{1}{2}\Big(\boldsymbol{\nabla}\overline{\mathbf{u}}^2 + (\boldsymbol{\nabla}\overline{\mathbf{u}}^2)^T\Big) - \frac{1}{d}\,\mathrm{Trace}(\boldsymbol{\nabla}\overline{\mathbf{u}}^2)\; I_d,$$

I_d being the $d \times d$ identity matrix. Asymptotic analysis of Eq. (2) shows that the proper behaviour $\mathcal{O}(y^3)$ of the viscosity is recovered, without any near–wall modification, which makes this model particularly attractive to deal with complex geometries.

As a first step toward the construction of a scheme for LES equations, we propose in this paper to study the discretization of the nonlinear (due to the presence of the subgrid model) diffusion term of the momentum balance equation. To this purpose, we address a simplified problem, namely the steady incompressible Stokes problem obtained by dropping the time derivative and convection terms in the original Navier-Stokes equations:

$$\begin{cases} -\mathrm{div}(2\nu \mathsf{S}(\overline{\mathbf{u}})) + \nabla\overline{p} = \bar{\mathbf{f}} & \text{in } \Omega, \\ \mathrm{div}\,\overline{\mathbf{u}} = 0 & \text{in } \Omega, \\ \overline{\mathbf{u}} = 0 & \text{on } \partial\Omega, \end{cases} \tag{3}$$

where $\mathsf{S}(\overline{\mathbf{u}}) = \frac{1}{2}(\nabla\overline{\mathbf{u}} + \nabla\overline{\mathbf{u}}^T)$ is the symmetric part of the gradient of $\overline{\mathbf{u}}$ and $\bar{\mathbf{f}}$ is a known forcing term. This problem is posed in Ω, an open, connected, bounded domain of $\mathbb{R}^d$ $(d = 2, 3)$, supposed to be polygonal for the sake of simplicity. The effective viscosity ν is equal to the sum of the laminar and turbulent viscosities denoted by ν_l and ν_t, respectively, the latter one being given as a function of the velocity by Eqs. (1) or (2) with a coefficient $\overline{\Delta}$ supposed here to be fixed (*i.e.* independent of the mesh). Since the velocity is prescribed to zero on the whole boundary, the pressure must be supposed to be mean-valued to obtain a well-posed problem.

Two approaches are considered for the spatial discretization: low order finite element (Rannacher–Turek element) and MAC scheme. We focus the paper on the finite element version, the description of the MAC scheme used for comparison in the numerical experiments being given in [6–8]. The obtained schemes are assessed by numerical experiments, using a manufactured solution technique.

The outline of the article is as follows. After the introduction of the Rannacher–Turek finite element (Sect. 2), we describe the resulting discretization of Problem (3), *i.e.*, essentially, the proposed discrete expression for the Smagorinsky or WALE subgrid viscosity (Sect. 3). Numerical tests are reported in Sect. 4.

To alleviate the notations, we drop in the remainder of this paper the overbar symbol to denote the averaged fields.

2 Mesh and discrete spaces

Let $\mathcal{M}$ be a decomposition of the domain Ω into quadrangles, supposed to be regular in the usual sense of the finite element literature [4]. We denote by $\mathcal{E}$ the set of all faces σ of the mesh; by $\mathcal{E}_{ext}$ the set of faces included in the boundary of Ω, by $\mathcal{E}_{int}$ the set of internal faces (*i.e.* $\mathcal{E} \setminus \mathcal{E}_{\mathrm{ext}}$) and by $\mathcal{E}(K)$ the faces of a particular cell $K \in \mathcal{M}$. By $|K|$ and $|\sigma|$ we denote the measure, respectively, of the control volume K and of the face σ.

The space discretization relies on the Rannacher–Turek mixed finite element. The degrees of freedom for the velocity are located at the mass center of the faces of the mesh, and we use the version of the element where they represent the average of the velocity over the face. The set of degrees of freedom thus reads:

$$\{\mathbf{u}_{\sigma,i},\ \sigma \in \mathcal{E},\ i = 1,\cdots,d\}.$$

The discrete functional space over a cell K is obtained through the usual Q_1 mapping from the space *span* $\left\{1, (x_i)_{i=1,\ldots,d}, (x_i^2 - x_{i+1}^2)_{i=1,\ldots,d-1}\right\}$ over the reference element. The approximation for the velocity is non–conforming: the space X_h is composed of discrete functions which are discontinuous through an edge, but the jump of their integral is imposed to be zero. We denote by $\varphi_\sigma^{(i)}$ the vector shape function associated to $\mathbf{u}_{\sigma,i}$, which, by definition, reads $\varphi_\sigma^{(i)} = \varphi_\sigma\, \mathbf{e}^{(i)}$, where φ_σ is the Rannacher–Turek scalar shape function and $\mathbf{e}^{(i)}$ is the i^{th} vector of the canonical basis of $\mathbb{R}^d$, and we define $\mathbf{u}_\sigma$ by $\mathbf{u}_\sigma = \sum_i \mathbf{u}_{\sigma,i}\, \mathbf{e}^{(i)}$. With these definitions, we have:

$$\mathbf{u} = \sum_{\sigma\in\mathcal{E}} \sum_{i=1,\cdots,d} \mathbf{u}_{\sigma,i}\ \varphi_\sigma^{(i)}(\mathbf{x}) = \sum_{\sigma\in\mathcal{E}} \mathbf{u}_\sigma\ \varphi_\sigma(\mathbf{x}), \qquad \text{for a.e. } \mathbf{x} \in \Omega.$$

Dirichlet boundary conditions are built in the definition of the discrete velocity space X_h by fixing $\mathbf{u}_{\sigma,i} = 0$ for all faces in $\mathcal{E}_{\text{ext}}$ and any component i.

The pressure is piecewise constant, and its degrees of freedom are denoted by p_K for any cell $K \in \mathcal{M}$. We denote by M_h the discrete pressure space.

3 The scheme

In this section, we begin with describing the approximation of the turbulent viscosity, which is chosen piecewise constant by cell, and we then present the discretization of Problem (3).

Expression of the cell viscosity ν_K for the Smagorinsky model – We propose to study two discretizations of the term S in Eq. (1) of the turbulent viscosity. The first one consists in approximating the expression Trace($\mathsf{S}\,\mathsf{S}^T$) by its mean value over a cell K:

$$\overline{\mathsf{S}^2}^K = \frac{1}{|K|} \int_K \mathsf{S}(\mathbf{u}) : \mathsf{S}(\mathbf{u})\,\mathrm{d}x.$$

The second approach is to compute the mean value of the velocity gradient over K and then to use it in the definition of S:

$$\overline{\mathsf{S}_{ij}}^K = \frac{1}{2}\left(\overline{\partial_j \mathbf{u}_i}^K + \overline{\partial_i \mathbf{u}_j}^K\right) = \frac{1}{2}\left(\frac{1}{|K|}\int_K \partial_j \mathbf{u}_i\,\mathrm{d}x + \frac{1}{|K|}\int_K \partial_i \mathbf{u}_j\,\mathrm{d}x\right). \qquad (4)$$

Finally, the expression of the effective viscosity for both approximations is:

– for the method 1:

$$\nu_K = \nu_l + (C_s\,\overline{\varDelta})^2 \left(2\,\overline{\mathsf{S}^2}^K\right)^{\frac{1}{2}}, \tag{5}$$

– for the method 2:

$$\nu_K = \nu_l + (C_s\,\overline{\varDelta})^2 \left(2\sum_{i,j}\overline{\mathsf{S}_{ij}}^K\,\overline{\mathsf{S}_{ij}}^K\right)^{\frac{1}{2}}. \tag{6}$$

These discretizations are different since the discrete velocity field is not piecewise affine. Hovever, as reported hereafter in Sect. 4, they give similar results, so only Method 2 is retained for the discretization of the WALE model, to avoid the computation of integrals needing high order quadrature formulas.

Expression of the cell viscosity ν_K for the WALE model – The discretization of the tensor ς in a cell $K \in \mathcal{M}$ is:

$$\overline{\varsigma_{ij}}^K = \frac{1}{2}\sum_{\ell\in\{1,\cdots,d\}}\left(\overline{\partial_\ell \mathbf{u}_i}^K\,\overline{\partial_j \mathbf{u}_\ell}^K + \overline{\partial_i \mathbf{u}_\ell}^K\,\overline{\partial_\ell \mathbf{u}_j}^K\right) - \left(\frac{1}{d}\sum_{m,n\in\{1,\cdots,d\}}\overline{\partial_m \mathbf{u}_n}^K\,\overline{\partial_n \mathbf{u}_m}^K\right)\delta_{i,j}, \quad \forall i,j\in\{1,\cdots,d\}, \tag{7}$$

where δ is the Kronecker symbol. Using the approximations of S and ς given respectively by (4) and (7), the effective viscosity on K reads:

$$\nu_K = \nu_l + (C_w\,\overline{\varDelta})^2\,\frac{\left(\displaystyle\sum_{i,j}\overline{\varsigma_{ij}}^K\,\overline{\varsigma_{ij}}^K\right)^{3/2}}{\left(\displaystyle\sum_{i,j}\overline{\mathsf{S}_{ij}}^K\,\overline{\mathsf{S}_{ij}}^K\right)^{5/2} + \left(\displaystyle\sum_{i,j}\overline{\varsigma_{ij}}^K\,\overline{\varsigma_{ij}}^K\right)^{5/4}}.$$

Discretization of Problem (3) – The scheme for the solution of Problem (3) consists in searching for $\mathbf{u} \in X_h$ and $p \in M_h$ such that the mean value of p over $\varOmega$ is zero and:

for $1 \le i \le d$ and for any $\sigma \in \mathcal{E}_{\text{int}}$,

$$\sum_{K\in\mathcal{M}} 2\nu_K \int_K \mathsf{S}(\mathbf{u}) : \mathsf{S}(\varphi_\sigma^{(i)})\,\mathrm{d}x - \sum_{K\in\mathcal{M}} \int_K p\,\operatorname{div}(\varphi_\sigma^{(i)})\,\mathrm{d}x = \int_\Omega \mathbf{f}\cdot\varphi_\sigma^{(i)}\mathrm{d}x,$$

$$\forall K \in \mathcal{M}, \quad \int_K \operatorname{div}(\mathbf{u})\,\mathrm{d}x = 0. \tag{8}$$

4 Numerical tests

In this section, we build a manufactured solution to Problem (3) in 2D, and compare the results obtained with the considered discretizations to the analytical solution and to the discrete solutions obtained with the MAC scheme. The simulations are performed with the ISIS software based on the platform PELICANS, both developed at IRSN [9, 11].

Description of the numerical test – The computational domain Ω is the unit square $(0, 1)^2$ and we calculate the forcing term $\mathbf{f}$ such that the exact velocity and pressure fields, $\mathbf{u}_{exact}$ and p_{exact}, are given by:

$$\mathbf{u}_{exact} = \mathbf{curl}(\sin(\pi x)\ \sin(\pi y)), \quad p_{exact} = \cos(\pi x)\ \sin(\pi y).$$

Note that $\mathbf{u}_{exact}$ indeed satisfies homogeneous Dirichlet boundary conditions on $\partial\Omega$, and the mean value over Ω of p_{exact} is zero.

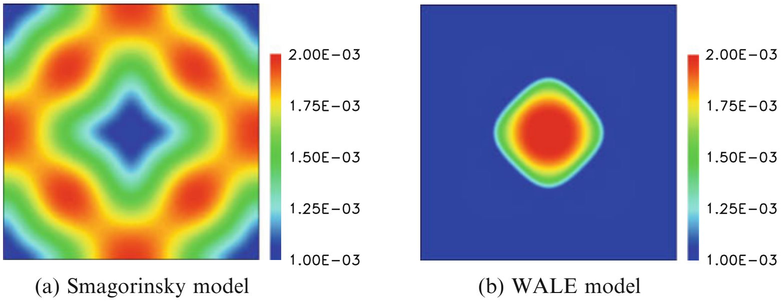
We take $\nu_l = 10^{-3}$ and the coefficient $C_w\overline{\Delta}$ in the expression of ν_t (Eqs. (1) and (2)) is set to $C_s\overline{\Delta} = 0.007$ for the Smagorinsky model and $C_w\overline{\Delta} = 0.009$ for the WALE model, which yields a turbulent and a laminar viscosity of the same range. The Smagorinsky and WALE viscosities obtained for $\mathbf{u}_{exact}$ are plotted on Fig. 1. The profiles are quite different, and one remarks that, as expected, the turbulent viscosity vanishes near the wall with the WALE model while it does not decrease with the Smagorinsky model.

The nonlinear problem (8) is solved using an iterative process analog to a time marching algorithm of pressure correction type [5], computing at each step the value of the turbulent viscosity from the beginning-of-step velocity. The steady state is supposed to be reached when velocity and pressure increments are small enough.

The discrete L^2–norm defined by

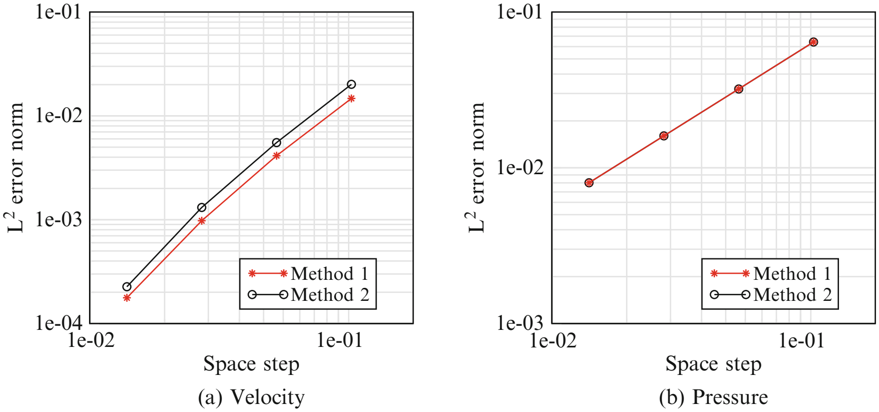
$$||\mathbf{u}||_0^2 = \sum_K \frac{|K|}{4} \sum_\sigma\ |\mathbf{u}_\sigma|^2$$

is used to measure the spatial error for $n\times n$ structured uniform meshes, with $n = 10$, 20, 40 and 80.



(a) Smagorinsky model (b) WALE model

Fig. 1 Repartition of the effective viscosity

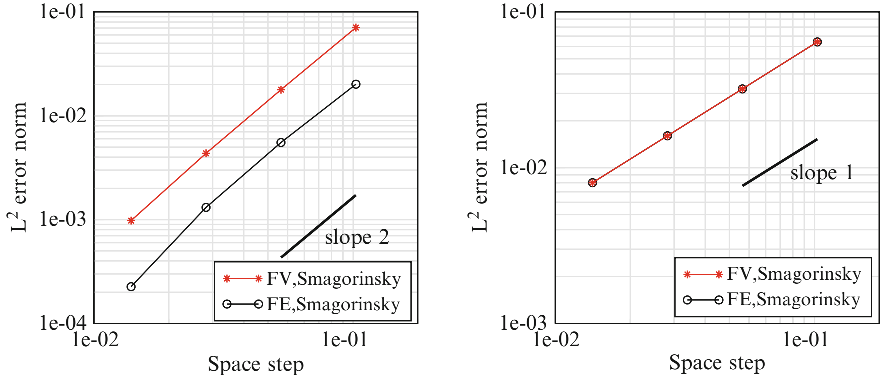


(a) Velocity (b) Pressure

Fig. 2 L^2 error norm for the velocity and the pressure as a function of the space step for both discretizations of the Smagorinsky viscosity: Method 1 corresponds to Eq. (5), Method 2 to Eq. (6)

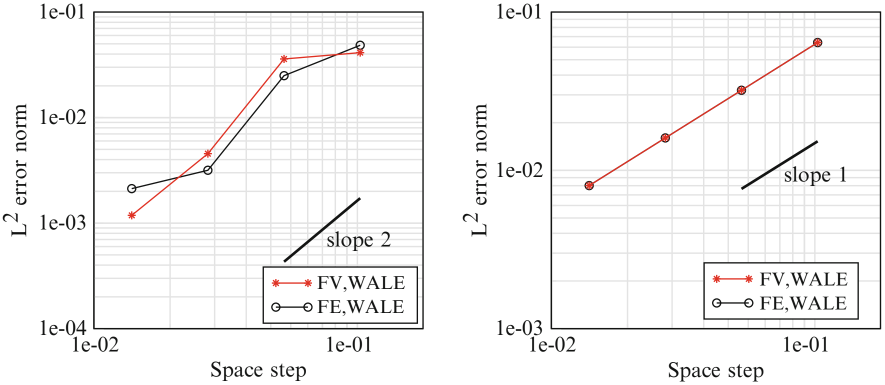
Comparison of both implementations for the Smagorinsky model – On Fig. 2, the spatial error in L^2–norm is plotted for both methods for the computation of the Smagorinsky viscosity. Both implementations give about the same accuracy. Consequently, Method 2 is chosen for further numerical experiments, because its implementation is simpler.

Comparison of the finite element approach and the MAC scheme for both models – On Fig. 3 and Fig. 4, 'FE' and 'FV' represent the discretization chosen, namely the Rannacher–Turek Finite Element and the MAC scheme (Finite Volume) respectively. Both discretizations seem to lead to the same order of convergence in space, that is 2 for the velocity and 1 for the pressure, for the Smagorinsky model. The FE discretization is more accurate than the MAC scheme but, for a given mesh, the number of degrees of freedom for the velocity for the FE discretization is twice (for $d = 2$) greater than for the MAC approximation (the number of degrees of freedom for the pressure being the same in both cases). For the WALE model, results look similar, with a more irregular convergence for the velocity.



(a) Velocity (b) Pressure

Fig. 3 L^2 error norm for the velocity and the pressure as a function of the space step for the Smagorinsky model



(a) Velocity (b) Pressure

Fig. 4 L^2 error norm for the velocity and the pressure as a function of the space step for the WALE model

5 Conclusion

As a conclusion, the space discretizations retained for the Smagorinsky model and the WALE model give satisfactory results for the considered steady nonlinear Stokes problem, both for the finite element method and for the MAC scheme. Next steps will be to extend the scheme to the complete Navier–Stokes equations with the same subgrid models (see [1, 3] for a kinetic energy preserving discretization of the convection term) and assess it on the academic test of the plane channel, before turning to more complex industrial applications.

References

1. G. Ansanay-Alex, F. Babik, J.-C. Latché, D. Vola: An L2-stable Approximation of the Navier-Stokes Convection Operator for Low-Order Non-Conforming Finite Elements. International Journal for Numerical Methods in Fluids, online (2010).
2. L.C. Berselli, T. Illiescu, W.J. Layton: Mathematics of Large Eddy Simulation of Turbulent Flows. Springer (2006).
3. F. Boyer, F. Dardalhon, C. Lapuerta, J.-C. Latché: Stability of a Crank-Nicolson Pressure Correction Scheme based on Staggered Discretizations, in preparation (2011).
4. P.G. Ciarlet: Handbook of Numerical Analysis Volume II: Finite Elements Methods – Basic Error Estimates for Elliptic Problems. North-Holland (1991).
5. F. Dardalhon, J.-C. Latché, S. Minjeaud: Analysis of a Projection Method for Low-Order Non-Conforming Finite Elements. Submitted (2011).
6. F.H. Harlow, J.E. Welsh: Numerical Calculation of Time-Dependent Viscous Incompressible Flow of Fluid with Free Surface. Physics of Fluids, **8**, 2182–2189 (1965).
7. R. Herbin, J.-C. Latché: A Kinetic Energy Control in the MAC Discretization of Compressible Navier-Stokes Equations. International Journal of Finite Volumes, **7**(2) (2010).
8. R. Herbin, W. Kheriji, J.-C. Latché: Discretization of Dissipation Terms with the MAC Scheme. Finite Volumes for Complex Appplications VI, Prague (2011).
9. ISIS: A CFD Computer Code for the Simulation of Reactive Turbulent Flows. https://gforge.irsn.fr/gf/project/isis.
10. F. Nicoud, F. Ducros: Subgrid-Scale Stress Modelling Based on the Square of the Velocity Gradient Tensor. Flow, Turbulence And Combustion, **62**, 183–200 (1999).
11. PELICANS: Collaborative Development Environment. https://gforge.irsn.fr/gf/project/pelicans.
12. R. Rannacher, S. Turek: Simple Nonconforming Quadrilateral Stokes Element. Numerical Methods for Partial Differential Equations, **8**, 97–111 (1992).
13. P. Sagaut: Large Eddy Simulation for Incompressible Flows. Scientific Computation, Springer-Verlag Berlin Heidelberg (2006).
14. J. Smagorinsky: General Circulation Experiments with the Primitive Equations. Monthly Weather Review, **91**(3), 99-164 (1963).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Some Abstract Error Estimates of a Finite Volume Scheme for the Wave Equation on General Nonconforming Multidimensional Spatial Meshes

Abdallah Bradji

Abstract A general class of nonconforming meshes has been recently studied for sationary anisotropic heterogeneous diffusion problems, see [2]. The aim of this contribution is to deal with error estimates, using this new class of meshes, for the wave equation. We present an implicit time scheme to approximate the wave equation. We prove that, when the discrete flux is calculated using a stabilized discrete gradient, the convergence order is $h_{\mathscr{D}} + k$, where $h_{\mathscr{D}}$ (resp. k) is the mesh size of the spatial (resp. time) discretization. This estimate is valid for discrete norms $\mathbb{L}^\infty(0,T;H_0^1(\Omega))$ and $\mathscr{W}^{1,\infty}(0,T;L^2(\Omega))$ under the regularity assumption $u \in \mathscr{C}^3([0,T];\mathscr{C}^2(\overline{\Omega}))$ for the exact solution u. These error estimates are useful because they allow to obtain approximations to the exact solution and its first derivatives of order $h_{\mathscr{D}} + k$.

Keywords second order hyperbolic equation, wave equation, non–conforming grid, SUSHI scheme, implicit scheme, discrete gradient
MSC2010: 65M08, 65M15

1 Motivation and aim of this paper

We consider the wave equation, as a model for second order hyperbolic equations:

$$u_{tt}(x,t) - \Delta u(x,t) = f(x,t),\ \ (x,t) \in \Omega \times (0,T), \tag{1}$$

where Ω is an open polygonal bounded subset in $\mathbb{R}^d$, $T > 0$, and f is a given function.
An initial condition is given by: for given functions u^0 and u^1 defined on Ω

Abdallah Bradji
Department of Mathematics, university of Annaba–Algeria, e-mail: bradji@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_19, © Springer-Verlag Berlin Heidelberg 2011

$$u(x,0) = u^0(x) \text{ and } u_t(x,0) = u^1(x) \; x \in \Omega, \tag{2}$$

Homogeneous Dirichlet boundary conditions are given by

$$u(x,t) = 0, \; (x,t) \in \partial\Omega \times (0,T). \tag{3}$$

2 Definition of the scheme

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, cf. [2, Definition 2.1, Page 1012]) 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.

We define the space $\mathscr{X}_\mathscr{D}$ as the set of all $((v_K)_{K\in\mathscr{M}}, \; (v_\sigma)_{\sigma\in\mathscr{E}})$, and $\mathscr{X}_{\mathscr{D},0} \subset \mathscr{X}_\mathscr{D}$ is the set of all $v \in \mathscr{X}_\mathscr{D}$ such that $v_\sigma = 0$ for all $\sigma \in \mathscr{E}_{\mathrm{ext}}$. Let $H_\mathscr{M}(\Omega) \subset \mathbb{L}^2(\Omega)$ be the space of piecewise constant functions on the control volumes of the mesh $\mathscr{M}$. For all $v \in \mathscr{X}_\mathscr{D}$, we denote by $\Pi_\mathscr{M}\, v \in H_\mathscr{M}(\Omega)$ the function defined by $\Pi_\mathscr{M}\, v(x) = v_K$, for a.e. $x \in K$, for all $K \in \mathscr{M}$.

For all $\varphi \in \mathscr{C}(\Omega)$, we define $\mathscr{P}_{\mathscr{D}}\varphi = ((\varphi(x_K))_{K\in\mathscr{M}}, (\varphi(x_\sigma))_{\sigma\in\mathscr{E}}) \in \mathscr{X}_{\mathscr{D}}$. We denote by $\mathscr{P}_{\mathscr{M}}\varphi \in H_{\mathscr{M}}(\Omega)$ the function defined by $\mathscr{P}_{\mathscr{M}}\,\varphi(x) = \varphi(x_K)$, for a.e. $x \in K$, for all $K \in \mathscr{M}$.

In order to analyze the convergence, we need to consider the size of the discretization $\mathscr{D}$ defined by $h_{\mathscr{D}} = \sup\{\text{diam(K)}, K \in \mathscr{M}\}$ and the regularity of the mesh given by $\theta_{\mathscr{D}} = \max\left(\max_{\sigma\in\mathscr{E}_{\text{int}},K,L\in\mathscr{M}} \frac{d_{K,\sigma}}{d_{L,\sigma}}, \max_{K\in\mathscr{M},\sigma\in\mathscr{E}_K} \frac{h_K}{d_{K,\sigma}}\right)$. The scheme we want to consider in this note (A general framework will be detailed in a future paper.) is based on the use of the discrete gradient given in [2]. For $u \in \mathscr{X}_{\mathscr{D}}$, we define, for all $K \in \mathscr{M}$

$$\nabla_{\mathscr{D}}\, u(x) = \nabla_{K,\sigma}\, u, \ \text{ a. e. } x \in \mathscr{D}_{K,\sigma}, \tag{4}$$

where $\mathscr{D}_{K,\sigma}$ is the cone with vertex x_K and basis σ and

$$\nabla_{K,\sigma} u = \nabla_K\, u + \left(\frac{\sqrt{d}}{d_{K,\sigma}}\left(u_\sigma - u_K - \nabla_K\, u \cdot (x_\sigma - x_K)\right)\right)\mathbf{n}_{K,\sigma}, \tag{5}$$

where $\nabla_K\, u = \dfrac{1}{\text{m}(K)} \displaystyle\sum_{\sigma\in\mathscr{E}_K} \text{m}(\sigma)\,(u_\sigma - u_K)\,\mathbf{n}_{K,\sigma}$ and d is the space dimension.

We define the finite volume approximation for (1)–(3) as $\left(u^n_{\mathscr{D}}\right)_{n=0}^{N+1} \in \mathscr{X}_{\mathscr{D},0}^{N+2}$ with $u^n_{\mathscr{D}} = \left(\left(u^n_K\right)_{K\in\mathscr{M}}, \left(u^n_\sigma\right)_{\sigma\in\mathscr{E}}\right)$, for all $n \in [\![0, N+1]\!]$ and

1. discretization of the initial conditions (2):

$$\langle\, u^0_{\mathscr{D}}, v\rangle_F = -\left(\Delta\, u^0, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)}, \ \ \forall\, v \in \mathscr{X}_{\mathscr{D},0}, \tag{6}$$

   and

$$\langle\, \frac{u^1_{\mathscr{D}} - u^0_{\mathscr{D}}}{k}, v\rangle_F = -\left(\Delta\, u^1, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)}, \ \ \forall\, v \in \mathscr{X}_{\mathscr{D},0}, \tag{7}$$

2. discretization of equation (1): for any $n \in [\![\,1, N]\!]$, $v \in \mathscr{X}_{\mathscr{D},0}$

$$\left(\Pi_{\mathscr{M}}\, \partial^2\, u^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)} + \langle\, u^{n+1}_{\mathscr{D}}, v\rangle_F = \sum_{K\in\mathscr{M}} \text{m}(K) f^n_K v_K, \tag{8}$$

   where

$$\langle\, u, v\rangle_F = \int_\Omega \nabla_{\mathscr{D}}\, u(x)\cdot\nabla_{\mathscr{D}}\, v(x)dx, \ \ \forall\, u, v \in \mathscr{X}_{\mathscr{D}}, \tag{9}$$

$$\partial^2\, v^{n+1} = \frac{v^{n+1} - 2v^n + v^{n-1}}{k^2}, \ \ \forall\, n \in [\![\,1, N]\!], \tag{10}$$

$$f^n_K = \frac{1}{k\text{m}(K)} \int_{t_n}^{t_{n+1}} \int_K f(x,t)d\,x\,dt, \tag{11}$$

   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 1. *(Error estimates for the finite volume scheme (6)–(11)) 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 (weak) of (1)–(3) satisfies $u \in \mathscr{C}^3([0,T];\mathscr{C}^2(\overline{\Omega}))$. Let $k = \frac{T}{N+1}$, with $N \in \mathbb{N}^\star$, and denote by $t_n = nk$, for $n \in [\![0, N+1]\!]$. 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 $\left(u^n_{\mathscr{D}}\right)_{n=0}^{N+1} \in \mathscr{X}^{N+2}_{\mathscr{D},\mathscr{B}}$ for problem (6)–(11).
For each $n \in [\![0, N+1]\!]$, let us define the error $e^n_{\mathscr{M}} \in H_{\mathscr{M}}(\Omega)$ by:

$$e^n_{\mathscr{M}} = \mathscr{P}_{\mathscr{M}}\, u(\cdot, t_n) - \Pi_{\mathscr{M}}\, u^n_{\mathscr{D}}. \tag{12}$$

Then, the following error estimates hold

- *discrete $\mathbb{L}^\infty(0,T;H^1_0(\Omega))$–estimate: for all $n \in [\![0, N+1]\!]$*

$$\|\, e^n_{\mathscr{M}}\|_{1,2,\mathscr{M}} \leq C(k + h_{\mathscr{D}})\|\, u\|_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \tag{13}$$

- *discrete $\mathscr{W}^{1,\infty}(0,T;\mathbb{L}^2(\Omega))$–estimate: for all $n \in [\![1, N+1]\!]$*

$$\|\, \partial^1\, e^n_{\mathscr{M}}\|_{\mathbb{L}^2(\Omega)} \leq C(k + h_{\mathscr{D}})\|\, u\|_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\overline{\Omega}))}, \tag{14}$$

  where $\partial^1 v^n = \dfrac{1}{k}\left(v^n - v^{n-1}\right)$.

- *error estimate in the gradient approximation: for all $n \in [\![0, N+1]\!]$*

$$\|\nabla_{\mathscr{D}}\, u^n_{\mathscr{D}} - \nabla\, u(\cdot, t_n)\|_{\mathbb{L}^2(\Omega)} \leq C(k + h_{\mathscr{D}})\|\, u\|_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\overline{\Omega}))}. \tag{15}$$

The following lemma will help us to prove Theorem 1

Lemma 1. *Let Ω be a polyhedral open bounded subset of $\mathbb{R}^d$, where $d \in \mathbb{N} \setminus \{0\}$, and $\partial\Omega = \overline{\Omega} \setminus \Omega$ its boundary. Let $k = \frac{T}{N+1}$, with $N \in \mathbb{N}^\star$, and denote by $t_n = nk$, for $n \in [\![0, N+1]\!]$. 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 $\theta \geq \theta_{\mathscr{D}}$. Assume in addition that there exists $\left(\eta^n_{\mathscr{D}}\right)_{n=0}^{N+1} \in \mathscr{X}^{N+2}_{\mathscr{D}}$ such that for any $n \in [\![1, N]\!]$, for all $v \in \mathscr{X}_{\mathscr{D}}$*

$$\left(\Pi_{\mathscr{M}}\, \partial^2\, \eta^{n+1}_{\mathscr{D}}, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)} + \langle\, \eta^{n+1}_{\mathscr{D}}, v\rangle_F = \sum_{K\in\mathscr{M}} \mathrm{m}(K)\mathscr{S}^n_K v_K, \tag{16}$$

where $\mathscr{S}^n_K \in \mathbb{R}$, for all $n \in [\![1, N]\!]$ and for all $K \in \mathscr{M}$.
Then the following estimate holds, for all $j \in [\![1, N]\!]$.

$$\begin{aligned} &\|\Pi_{\mathscr{M}}\, \partial^1\, \eta^{j+1}_{\mathscr{D}}\|^2_{\mathbb{L}^2(\Omega)} + C|\,\eta^{j+1}_{\mathscr{D}}|^2_{\mathscr{X}} \\ &\leq C\left(\|\Pi_{\mathscr{M}}\, \partial^1\, \eta^1_{\mathscr{D}}\|^2_{\mathbb{L}^2(\Omega)} + |\eta^1_{\mathscr{D}}|^2_{\mathscr{X}} + (\mathscr{S})^2\right), \end{aligned} \tag{17}$$

where

$$\mathscr{S} = \max\left\{ \left(\sum_{K\in\mathscr{M}} \mathrm{m}(K) \left(\mathscr{S}_K^n\right)^2 \right)^{\frac{1}{2}},\ n \in [\![1, N]\!] \right\}. \tag{18}$$

Proof. Taking $v = \partial^1 \eta_{\mathscr{D}}^{n+1}$ in (16) and summing the result over $n \in [\![1, j]\!]$, where $j \in [\![1, N]\!]$, we get

$$\sum_{n=1}^{j} \left(\Pi_{\mathscr{M}}\, \partial^2 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\, \partial^1 \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)} + \sum_{n=1}^{j} \langle \eta_{\mathscr{D}}^{n+1}, \partial^1 \eta_{\mathscr{D}}^{n+1} \rangle_F$$
$$= \sum_{n=1}^{j} \sum_{K\in\mathscr{M}} \mathrm{m}(K) \mathscr{S}_K^n \partial^1 \eta_K^{n+1}. \tag{19}$$

We need the following two rules

$$\left(\Pi_{\mathscr{M}}\, \partial^2 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\, \partial^1 \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)} = \frac{1}{2k} \| \alpha_{\mathscr{D}}^{n+1} - \alpha_{\mathscr{D}}^{n} \|_{\mathbb{L}^2(\Omega)}^2$$
$$+ \frac{1}{2k} \left(\|\alpha_{\mathscr{D}}^{n+1}\|_{\mathbb{L}^2(\Omega)}^2 - \|\alpha_{\mathscr{D}}^{n}\|_{\mathbb{L}^2(\Omega)}^2 \right), \tag{20}$$

where $\alpha_{\mathscr{D}}^n = \Pi_{\mathscr{M}}\, \partial^1 \eta_{\mathscr{D}}^n$ and

$$\langle \eta_{\mathscr{D}}^{n+1}, \partial^1 \eta_{\mathscr{D}}^{n+1} \rangle_F = \frac{1}{2k} \langle \eta_{\mathscr{D}}^{n+1} - \eta_{\mathscr{D}}^{n}, \eta_{\mathscr{D}}^{n+1} - \eta_{\mathscr{D}}^{n} \rangle_F$$
$$+ \frac{1}{2k} \left\{ \langle \eta_{\mathscr{D}}^{n+1}, \eta_{\mathscr{D}}^{n+1} \rangle_F - \langle \eta_{\mathscr{D}}^{n}, \eta_{\mathscr{D}}^{n} \rangle_F \right\}. \tag{21}$$

Identities (20)–(21) yield

$$\sum_{n=1}^{j} \left(\Pi_{\mathscr{M}}\, \partial^2 \eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\, \partial^1 \eta_{\mathscr{D}}^{n+1} \right)_{\mathbb{L}^2(\Omega)} + \sum_{n=1}^{j} \langle \eta_{\mathscr{D}}^{n+1}, \partial^1 \eta_{\mathscr{D}}^{n+1} \rangle_F$$
$$\geq \frac{1}{2k} \left(\| \alpha_{\mathscr{D}}^{j+1} \|_{\mathbb{L}^2(\Omega)}^2 + \langle \eta_{\mathscr{D}}^{j+1}, \eta_{\mathscr{D}}^{j+1} \rangle_F \right) - \frac{1}{2k} \left(\| \alpha_{\mathscr{D}}^{1} \|_{\mathbb{L}^2(\Omega)}^2 + \langle \eta_{\mathscr{D}}^{1}, \eta_{\mathscr{D}}^{1} \rangle_F \right).$$

This with (19) and [2, Lemma 4.2] implies

$$\frac{1}{2k} \left(\|\alpha^{j+1}\|_{\mathbb{L}^2(\Omega)}^2 + C |\, \eta_{\mathscr{D}}^{j+1} |_{\mathscr{X}}^2 \right) \leq \sum_{n=1}^{j} \sum_{K\in\mathscr{M}} \mathrm{m}(K) \mathscr{S}_K^n \partial^1 \eta_K^{n+1}$$
$$+ \frac{1}{2k} \left(\|\alpha_{\mathscr{D}}^{1}\|_{\mathbb{L}^2(\Omega)}^2 + C |\eta_{\mathscr{D}}^{1}|_{\mathscr{X}}^2 \right). \tag{22}$$

Multiplying both sides of the previous inequality by $2k$ and using the Cauchy Schwarz inequality, we get

$$\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{j+1}\|^2_{\mathbb{L}^2(\Omega)} + C\,|\,\eta_{\mathscr{D}}^{j+1}|^2_{\mathscr{X}} \le 2k\mathscr{S}\sum_{n=1}^{j}\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{n+1}\|_{\mathbb{L}^2(\Omega)}$$
$$+\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{1}\|^2_{\mathbb{L}^2(\Omega)} + C|\eta_{\mathscr{D}}^{1}|^2_{\mathscr{X}}, \quad (23)$$

where $\mathscr{S}$ is given by (18).
This with the inequality $ab \le \frac{T}{k}a^2 + \frac{k}{T}b^2$, (23) implies, for all $j \in [\![1, N]\!]$

$$\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{j+1}\|^2_{\mathbb{L}^2(\Omega)} + C\,|\,\eta_{\mathscr{D}}^{j+1}|^2_{\mathscr{X}} \le \frac{2k}{T}\sum_{n=2}^{j}\left(\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{n}\|^2_{\mathbb{L}^2(\Omega)} + C\,|\,\eta_{\mathscr{D}}^{n}|^2_{\mathscr{X}}\right)$$
$$+2\|\Pi_{\mathscr{M}}\,\partial^1\,\eta_{\mathscr{D}}^{1}\|^2_{\mathbb{L}^2(\Omega)} + C|\eta_{\mathscr{D}}^{1}|^2_{\mathscr{X}} + 8T^2\,(\mathscr{S})^2\,. \quad (24)$$

Using the discrete version of the Gronwall's Lemma, (24) implies estimate (17).

Sketch of the proof of Theorem 1: The uniqueness of $\left(u_{\mathscr{D}}^n\right)_{n\in[\![0,N+1]\!]}$ satisfying (6)–(11) can be deduced from the [2, Lemma 4.2]. As usual, we can use this uniqueness to prove the existence. To prove (13)–(15), we compare the solution $\left(u_{\mathscr{D}}^n\right)_{n\in[\![0,N+1]\!]}$ satisfying (6)–(11) with the solution (it exists and it is unique thanks to [2, Lemma 4.2]): for any $n \in [\![0, N+1]\!]$, find $\bar{u}_{\mathscr{D}}^n \in \mathscr{X}_{\mathscr{D},0}$ such that, see (9)

$$\langle\,\bar{u}_{\mathscr{D}}^n, v\rangle_F = -\sum_{K\in\mathscr{M}} v_K \int_K \Delta\, u(x,t_n)dx,\ \ \forall\, v \in \mathscr{X}_{\mathscr{D},0}. \quad (25)$$

Taking $n = 0$ in (25), using the fact that $u(\cdot,0) = u^0(\cdot)$, and comparing this with (6), we get the following property which will be used below

$$\bar{u}_{\mathscr{D}}^0 = u_{\mathscr{D}}^0. \quad (26)$$

One remarks that the solution of (25) is the same one of [1, (12)], one can use error estimates [1, (13), (15), and (16)] as error estimates for the solution of (25).
Writing (25) in the step $n+1$ and substracting the result from (8) to get

$$\left(\Pi_{\mathscr{M}}\,\partial^2\,\eta_{\mathscr{D}}^{n+1}, \Pi_{\mathscr{M}}\,v\right)_{\mathbb{L}^2(\Omega)} + \langle\,\eta_{\mathscr{D}}^{n+1}, v\rangle_F = \sum_{K\in\mathscr{M}} \mathrm{m}(K)\mathscr{S}_K^n v_K, \quad (27)$$

where $\eta_{\mathscr{D}}^n = u_{\mathscr{D}}^n - \bar{u}_{\mathscr{D}}^n$, for all $n \in [\![0, N+1]\!]$ and

$$\mathscr{S}_K^n = \frac{1}{k\mathrm{m}(K)}\int_{t_n}^{t_{n+1}}\int_K f(x,t)d\,x\,dt + \frac{1}{\mathrm{m}(K)}\int_K \Delta\, u(x,t_{n+1})dx - \partial^2\bar{u}_K^{n+1}. \quad (28)$$

Equation (27) with Lemma 1 implies that, for all $n \in [\![1, N]\!]$

$$\begin{aligned}&\|\Pi_{\mathscr{M}}\, \partial^1\, \eta_{\mathscr{D}}^{n+1}\|^2_{\mathbb{L}^2(\Omega)} + C\,|\,\eta_{\mathscr{D}}^{n+1}|^2_{\mathscr{X}}\\ &\le\; C\left(\|\Pi_{\mathscr{M}}\, \partial^1\, \eta_{\mathscr{D}}^{1}\|^2_{\mathbb{L}^2(\Omega)} + |\eta_{\mathscr{D}}^{1}|^2_{\mathscr{X}} + (\mathscr{S})^2\right). \end{aligned}\tag{29}$$

To estimate the terms on the right hand side of the previous inequality, we consider

$$\xi_{\mathscr{D}}^n = \bar{u}_{\mathscr{D}}^n - \mathscr{P}_{\mathscr{D}}\, u(\cdot, t_n),\;\; \forall\, n \in [\![0, N+1]\!]. \tag{30}$$

It is useful to remark that (recall that $\eta_{\mathscr{D}}^n = u_{\mathscr{D}}^n - \bar{u}_{\mathscr{D}}^n$)

$$u_{\mathscr{D}}^n - \mathscr{P}_{\mathscr{D}}\, u(\cdot, t_n) = \eta_{\mathscr{D}}^n + \xi_{\mathscr{D}}^n. \tag{31}$$

1. *Estimate of* $\|\Pi_{\mathscr{M}}\, \partial^1\, \eta_{\mathscr{D}}^1\|_{\mathbb{L}^2(\Omega)}$: using (31), we get (recall that $u_t(\cdot, 0) = u^1(\cdot)$)

$$\|\Pi_{\mathscr{M}}\, \partial^1\, \eta_{\mathscr{D}}^1\|_{\mathbb{L}^2(\Omega)} \le \sum_{i=1}^{4} \mathbb{T}_i, \tag{32}$$

where

$$\mathbb{T}_1 = \|\Pi_{\mathscr{M}}\partial^1\xi_{\mathscr{D}}^1\|_{\mathbb{L}^2(\Omega)},\;\; \mathbb{T}_2 = \|\Pi_{\mathscr{M}}\, \partial^1\, u_{\mathscr{D}}^1 - u^1\|_{\mathbb{L}^2(\Omega)},$$

$\mathbb{T}_3 = \|\, u_t(\cdot, 0) - \partial^1\, u(\cdot, t_1)\|_{\mathbb{L}^2(\Omega)}$, and $\mathbb{T}_4 = \|\, \partial^1\, u(\cdot, t_1) - \mathscr{P}_{\mathscr{M}}\partial^1\, u(\cdot, t_1)\|_{\mathbb{L}^2(\Omega)}$.

Estimate [1, (15)], when $j = 1$, with (30) leads to

$$\mathbb{T}_1 \le\; C\, h_{\mathscr{D}} \|\, u\|_{\mathscr{C}^1([0,T];\; \mathscr{C}^2(\overline{\Omega}))}. \tag{33}$$

Equation (7) can be written as

$$\langle\, \partial^1\, u_{\mathscr{D}}^1, v\rangle_F = -\left(\Delta\, u^1, \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)},\;\; \forall\, v \in \mathscr{X}_{\mathscr{D},0}. \tag{34}$$

This with [2, (4.25)] and the triangle inequality implies that

$$\mathbb{T}_i \le\; C\, (k + h_{\mathscr{D}}) \|\, u\|_{\mathscr{C}^1([0,T];\; \mathscr{C}^2(\overline{\Omega}))},\;\; \forall\, i \in [\![2, 4]\!]. \tag{35}$$

Thanks to (32), (33), and (35), we have

$$\|\Pi_{\mathscr{M}}\, \partial^1\, \eta_{\mathscr{D}}^1\|_{\mathbb{L}^2(\Omega)} \le\; C\, (k + h_{\mathscr{D}}) \|\, u\|_{\mathscr{C}^1([0,T];\; \mathscr{C}^2(\overline{\Omega}))}. \tag{36}$$

2. *Estimate of* $|\eta_{\mathscr{D}}^1|_{\mathscr{X}}$: let us first remark that thanks to (6) and (7), we have

$$\langle\, u_{\mathscr{D}}^1, v\rangle_F = -\left(\Delta\, (u^0 + k u^1), \Pi_{\mathscr{M}}\, v\right)_{\mathbb{L}^2(\Omega)},\;\; \forall\, v \in \mathscr{X}_{\mathscr{D},0}. \tag{37}$$

In order to bound $|\eta^1_{\mathscr{D}}|_{\mathscr{X}} = |u^1_{\mathscr{D}} - \bar{u}^1_{\mathscr{D}}|_{\mathscr{X}}$, we use the triangle inequality to get

$$|\eta^1_{\mathscr{D}}|_{\mathscr{X}} \leq |u^1_{\mathscr{D}} - \mathscr{P}_{\mathscr{D}}(u^0 + ku^1)|_{\mathscr{X}} + |\mathscr{P}_{\mathscr{D}}(u^0 + ku^1) - \mathscr{P}_{\mathscr{D}}u(\cdot, t_1)|_{\mathscr{X}} + |\mathscr{P}_{\mathscr{D}}u(\cdot, t_1) - \bar{u}^1_{\mathscr{D}}|_{\mathscr{X}}. \tag{38}$$

This with the proof of [2, (4.29)] and suitable Taylor expansions, we get

$$|\eta^1_{\mathscr{D}}|_{\mathscr{X}} \leq C\,(k + h_{\mathscr{D}})\|u\|_{\mathscr{C}^2([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{39}$$

3. *Estimate of $\mathscr{S}$*: substituting f by $u_{tt} - \Delta u$, see (1), in the expansion of $\mathscr{S}^n_K$, we get

$$\mathscr{S}^n_K = \frac{1}{k\mathrm{m}(K)}\int_{t_n}^{t_n+1}\int_K u_{tt}(x,t)d\,x\,dt - \frac{1}{k\mathrm{m}(K)}\int_{t_n}^{t_n+1}\int_K \Delta(x,t)d\,x\,dt + \frac{1}{\mathrm{m}(K)}\int_K \Delta u(x, t_{n+1})dx - \partial^2\bar{u}^{n+1}_K. \tag{40}$$

Thanks to the Taylor expansion and [1, (15)], when $j = 2$, we have

$$\mathscr{S} \leq C\,(k + h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{41}$$

Gathering now (29), (36), (39), and (41) yields, for all $n \in [\![2, N+1]\!]$

$$\|\Pi_{\mathscr{M}}\,\partial^1\,\eta^n_{\mathscr{D}}\|_{\mathbb{L}^2(\Omega)} \leq C\,(k + h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}, \tag{42}$$

and

$$|\eta^n_{\mathscr{D}}|_{\mathscr{X}} \leq C\,(k + h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}. \tag{43}$$

We now combine (42)–(43) with [1, (13), (15), and (16)] to prove the required estimates (13)–(15).

- *Proof of estimate* (13): estimate (43) with [2, (4.6)] implies

$$\|\Pi_{\mathscr{M}}\,\eta^n_{\mathscr{D}}\|_{1,2,\mathscr{M}} \leq C\,(k + h_{\mathscr{D}})\|u\|_{\mathscr{C}^3([0,T];\,\mathscr{C}^2(\overline{\Omega}))}, \quad \forall n \in [\![2, N+1]\!]. \tag{44}$$

  This with (31), the fact that $\Pi_{\mathscr{M}}\,\xi^n_{\mathscr{D}} = \Pi_{\mathscr{M}}\bar{u}^n_{\mathscr{D}} - \mathscr{P}_{\mathscr{M}}\,u(\cdot, t_n)$, estimate [1, (13)], and the triangle inequality implies estimate (13) for all $n \in [\![2, N+1]\!]$. The case when $n = 1$ in (13) can be proved by gathering (39), [2, (4.6)], and the case $n = 1$ of [1, (13)]. Property (26) with the case $n = 0$ of [1, (13)] yields the case $n = 0$ of (13).
- *Proof of estimate* (14): the case when $n \in [\![2, N+1]\!]$ of (14) can be proved by gathering (42), the case when $j = 1$ in [1, (15)], and the triangle inequality. The case $n = 1$ of (14) can be proved by gathering (36), the case when $n = 1$ and $j = 1$ in [1, (15)], and the triangle inequality.

– *Proof of estimate* (15): gathering (39) and (43), and [2, Lemma 4.2] leads to

$$\| \nabla_{\mathscr{D}} \eta_{\mathscr{D}}^n \|_{\mathbb{L}^2(\Omega)} \leq C\,(k + h_{\mathscr{D}}) \| u \|_{\mathscr{C}^3([0,T];\, \mathscr{C}^2(\overline{\Omega}))}, \quad \forall\, n \in [\![1, N+1]\!]. \tag{45}$$

Combining (45), [1, (16)], and the triangle inequality yields (15) for all $n \in [\![1, N+1]\!]$. The case $n = 0$ of (15) can be deduced directly from the case $n = 0$ of [1, (16)] by using (26). □

References

1. Bradji, A., Fuhrmann, J.: Error estimates of the discretization of linear parabolic equations on general nonconforming spatial grids. C. R. Math. Acad. Sci. Paris **348**/19-20, 1119–1122 (2010).
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).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Convergent Finite Volume Scheme for Two-Phase Flows in Porous Media with Discontinuous Capillary Pressure Field

K. Brenner, C. Cancès, and D. Hilhorst

Abstract We consider an immiscible incompressible two-phase flow in a porous medium composed of two different rocks. The flows of oil and water are governed by the Darcy–Muskat law and a capillary pressure law, where the capillary pressure field may be discontinuous at the interface between the rocks. Using the concept of multi-valued phase pressures, we introduce a notion of weak solution for the flow, and prove the convergence of a finite volume approximation towards a weak solution.

Keywords Finite volume, two phase flow, discontinuous capillary pressures
MSC2010: 65M112, 76M12, 35K65, 76S05

1 The Continuous Problem

1.1 Multivalued Phase Pressures

Consider a heterogeneous porous medium, represented by a polygonal domain $\Omega \subset \mathbb{R}^d$, built of two homogeneous and isotropic subdomains, represented by polygonal domains $\Omega_1, \Omega_2 \subset \mathbb{R}^d$. We assume that $\overline{\Omega_1 \cup \Omega_2} = \overline{\Omega}$ and $\Omega_1 \cap \Omega_2 = \emptyset$, and we denote by Γ the interface between the two rocks, i.e. $\overline{\Gamma} = \partial\Omega_1 \cap \partial\Omega_2$. We consider two immiscible incompressible phases (e.g. water and oil), whose flows within Ω_i are described by the conservation of mass equations together with the

Konstantin Brenner and Danielle Hilhorst
CNRS and Université Paris-Sud 11, 91405 Orsay Cedex, e-mail: konstantin.brenner@math.u-psud.fr, danielle.hilhorst@math.u-psud.fr

Clément Cancès
LJLL, UPMC, 75252 Paris cedex 05, e-mail: cances@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_20, © Springer-Verlag Berlin Heidelberg 2011

Darcy–Muskat law:

$$\phi_i \partial_t s - \nabla \cdot (\eta_{o,i}(s)(\nabla p_o - \rho_o \mathbf{g})) = 0, \tag{1}$$

$$-\phi_i \partial_t s - \nabla \cdot (\eta_{w,i}(s)(\nabla p_w - \rho_w \mathbf{g})) = 0, \tag{2}$$

where s denotes the oil saturation of the fluid, $\phi_i > 0$ the porosity of Ω_i, the oil mobility $\eta_{o,i}$ is a Lipschitz continuous increasing function on $[0,1]$ satisfying $\eta_{o,i}(0) = 0$, while the water mobility $\eta_{w,i}$ is Lipschitz continuous, decreasing on $[0,1]$ and such that $\eta_{w,i}(1) = 0$. The density of the phase α ($\alpha \in \{o,w\}$) is denoted by ρ_α, and $\mathbf{g}$ is the gravity vector. Assume first that both phases coexist, i.e. $s \in (0,1)$, then each phase has its own pressure denoted by p_α. Classically, they are supposed to be linked by the capillary pressure relation

$$p_o - p_w = \pi_i(s), \tag{3}$$

where the capillary pressure function π_i is supposed to be increasing and to belong to $\mathscr{C}^1((0,1);\mathbb{R}) \cap L^1(0,1)$. Since the equation (1) degenerates, there is no control on the oil pressure p_o on $\{s = 0\} \cap \Omega_i$, excepted that, because of (3), one has $p_o \le p_w + \pi_i(0)$. Similarly, on $\{s = 1\} \cap \Omega_i$, $p_w \le p_o - \pi_i(1)$. In these cases, the pressure has to be considered as multivalued, i.e.

$$s = 0 \Leftrightarrow p_o = [-\infty, p_w + \pi_i(0)], \qquad s = 1 \Leftrightarrow p_w = [-\infty, p_o - \pi_i(1)]. \tag{4}$$

We deduce from (4) that the capillary pressure function π_i has to be extended into the monotone graph $\tilde{\pi}_i$, already introduced in [3, 5], defined by

$$\tilde{\pi}_i(s) = \begin{cases} [-\infty, \pi_i(0)] & \text{if } s = 0, \\ \pi_i(s) & \text{if } s \in (0,1), \\ [\pi_i(1), +\infty] & \text{if } s = 1. \end{cases} \tag{5}$$

Note that there exists a continuous non-decreasing reciprocal function on $\mathbb{R}$, which we denote by $\tilde{\pi}_i^{-1}$.

At the interface Γ, we prescribe the continuity of the multivalued phase pressures

$$p_{\alpha,1} \cap p_{\alpha,2} \neq \emptyset, \qquad (\alpha \in \{o,w\}) \tag{6}$$

where $p_{\alpha,i}$ denote the trace of the pressure of the phase α on Γ from Ω_i. It is worth noticing that the condition (6) is equivalent to the continuity of the mobile phases prescribed in [8]. The volume conservation of each phase yields

$$\sum_{i=1,2} \eta_{\alpha,i}(s_i)(\nabla p_{\alpha,i} - \rho_\alpha \mathbf{g}) \cdot \mathbf{n}_i = 0, \tag{7}$$

where $\mathbf{n}_i$ denote the outward normal to $\partial\Omega_i$ w.r.t. Ω_i. In order to close the problem, we prescribe the initial condition

$$s_0 \in L^\infty(\Omega), \quad 0 \le s_0 \le 1 \text{ a.e. in } \Omega, \tag{8}$$

and the null-flux boundary condition on $\partial\Omega_i \cap \partial\Omega$:

$$\eta_{\alpha,i}(s_i)(\nabla p_{\alpha,i} - \rho_\alpha \mathbf{g}) \cdot \mathbf{n}_i = 0. \tag{9}$$

1.2 Reformulation of the Problem

We define the fractional flow function $f_i(s) = \frac{\eta_{o,i}(s)}{\eta_{o,i}(s)+\eta_{w,i}(s)}$. We introduce the Kirchhoff transform $\varphi_i(s)$ and the global pressure P defined by

$$\varphi_i(s) = \int_0^s f_i(a)\eta_{w,i}(a)\pi_i'(a)da, \tag{10}$$

$$P = p_w + \lambda_{w,i}(\pi) = p_o + \lambda_{o,i}(\pi) \quad \text{for some } \pi \in \tilde{\pi}_i(s), \tag{11}$$

where $\lambda_{w,i}(\pi) = \displaystyle\int_0^\pi f_i \circ \tilde{\pi}_i^{-1}(p)dp$ and $\lambda_{o,i}(\pi) = \lambda_{w,i}(\pi) - \pi$. Classical computations (see e.g. [7]) allow the rewrite the equations (1) as

$$\phi_i \partial_t s - \nabla \cdot (\eta_{o,i}(s)(\nabla P - \rho_o \mathbf{g}) + \nabla \varphi_i(s)) = 0, \tag{12}$$

while the sum of the equations (1) and (2) yields

$$-\nabla \cdot (M_i(s)\nabla P - \zeta_i(s)\mathbf{g}) = 0, \tag{13}$$

where $M_i(s) = \eta_{o,i}(s) + \eta_{w,i}(s) \ge \alpha_M > 0$ and $\zeta_i(s) = \eta_{o,i}(s)\rho_o + \eta_{w,i}(s)\rho_w$. At the interface, the relations (6) have to be replaced (see [6]) by

$$\exists \pi \in \tilde{\pi}_1(s_1) \cap \tilde{\pi}_2(s_2) \text{ s.t. } P_1 - \lambda_{w,1}(\pi) = P_2 - \lambda_{w,2}(\pi). \tag{14}$$

We solve the problem on the domain $Q = \Omega \times (0, T)$ for some $T > 0$, and we define $Q_i = \Omega_i \times (0, T)$.

Definition 1 (weak solution). A pair (s, P) is said to be a weak solution of the problem if

1. $s \in L^\infty(Q)$ with $0 \le s \le 1$ a.e. in Q, $\varphi_i(s)$ and P belong to $L^2((0, T); H^1(\Omega_i))$;
2. there exists a measurable function π mapping $\Gamma \times (0, T)$ to $\overline{\mathbb{R}}$ such that

$$\pi \in \tilde{\pi}_1(s_1) \cap \tilde{\pi}_2(s_2) \text{ and } P_1 - \lambda_{w,1}(\pi) = P_2 - \lambda_{w,2}(\pi);$$

3. for all $\psi \in C_c^\infty(\overline{\Omega} \times [0, T))$,

$$\iint_Q \phi s \partial_t \psi dxdt + \int_\Omega \phi s_0 \psi(\cdot,0) dxdt$$

$$- \sum_{i=1,2} \iint_{Q_i} (\eta_{o,i}(s)(\nabla P - \rho_o \mathbf{g}) + \nabla \varphi_i(s)) \cdot \nabla \psi dxdt = 0, \qquad (15)$$

$$\iint_Q (M_i(s)\nabla P - \zeta_i(s)\mathbf{g}) \cdot \nabla \psi dxdt = 0. \qquad (16)$$

Because of the choice of the boundary conditions, the global pressure P is only defined up to a constant. In order to eliminate this degree of freedom, we prescribe that

$$\int_\Omega P(x,t)dx = 0, \quad \forall t > 0. \qquad (17)$$

The equation (13) can be reformulated as

$$\nabla \cdot \mathbf{q} = 0, \text{ with } \mathbf{q} = -M_i(s)\nabla P + \zeta_i(s)\mathbf{g}, \qquad (18)$$

while (12) can be rewritten under the form

$$\phi_i \partial_t s + \nabla \cdot (\mathbf{q} f_i(s) + \gamma_i(s)\mathbf{g} - \nabla \varphi_i(s)) = 0, \qquad (19)$$

with $\gamma_i(s) = (\rho_o - \rho_w)\eta_{w,i}(s) f_i(s)$.

2 The Finite Volume Scheme

Since nonlinear test functions are necessary for proving the convergence of the scheme, we must restrict our study to spatial discretizations satisfying an orthogonality condition, as developed in [9].

Definition 2 (admissible discretization of Q).

1. An admissible discretization of Ω is given by $(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}})$ where for all $K \in \mathscr{T}$, K is an open polygonal subset of Ω such that $K \subset \Omega_i$ for some i. We define $\mathscr{T}_i = \{K \subset \Omega_i\}$, and we assume that $\overline{\Omega}_i = \overline{\bigcup_{K\in\mathscr{T}_i} K}$. For $K, L \in \mathscr{T}$ with $K \neq L$, then either the $(d-1)$-Lebesgue measure of $\overline{K} \cap \overline{L}$ is 0, or there exists $\sigma \in \mathscr{E}_K \cap \mathscr{E}_L$ (denoted by $\sigma = K|L$) such that $\overline{\sigma} = \overline{K} \cap \overline{L}$. For all $K \in \mathscr{T}$, there exists $\mathscr{E}_K \subset \mathscr{E}$ such that $\partial K = \bigcup_{\sigma\in\mathscr{E}_K} \overline{\sigma}$. Moreover, $\mathscr{E} = \bigcup_{K\in\mathscr{T}} \mathscr{E}_K$. We define $\mathscr{E}_\Gamma = \{\sigma \in \mathscr{E} : \sigma \subset \Gamma\}$, $\mathscr{E}_i = \{\sigma \in \mathscr{E} : \sigma \subset \Omega_i\}$ and $\mathscr{E}_{\text{ext}} = \{\sigma \in \mathscr{E} : \sigma \subset \partial\Omega\}$, and set $\mathscr{E}_{K,\Gamma} = \mathscr{E}_K \cap \mathscr{E}_\Gamma$, $\mathscr{E}_{K,i} = \mathscr{E}_K \cap \mathscr{E}_i$. The family of points $(x_K)_{K\in\mathscr{T}}$ is such that $x_K \in K$ and if $\sigma = K|L$, the straight line $(x_K x_L)$ is orthogonal to σ. We denote by $d_{K,L}$ the distance between x_K and x_L, and by $d_{K,\sigma}$ the distance between x_K and $\sigma \in \mathscr{E}_K$. For all $K \in \mathscr{T}$ and $\sigma \in \mathscr{E}$ we denote by $m(K)$ and $m(\sigma)$ the corresponding Lebesgue measures.

2. Let N be a positive integer, and $\delta t = T/N$; then a uniform discretization of $(0,T)$ is given by the family $(t^n)_{n\in\{0,\dots,N\}}$, where $t^n = n\delta t$.
3. A discretization $\mathscr{D} = \big(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}}, (t^n)_{n\in\{0,\dots,N\}}\big)$ of Q is said to be admissible if $(\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}})$ is an admissible discretization of Ω and $(t^n)_n$ is a uniform discretization of $(0,T)$.

For a given admissible discretization $\mathscr{D} = (\mathscr{T}, \mathscr{E}, (x_K)_{K\in\mathscr{T}}, (t^n)_{n\in\{0,\dots,N\}})$ of Q, we define the quantities

$$\operatorname{size}(\mathscr{T}) = \max_{K\in\mathscr{T}} \operatorname{diam}(K), \qquad \operatorname{reg}(\mathscr{T}) = \max_{i=1,2}\max_{K\in\mathscr{T}} \left(\sum_{\sigma=K|L\in\mathscr{E}_{K,\mathrm{i}}} \frac{m(\sigma) d_{K,L}}{m(K)} \right),$$

and

$$\operatorname{size}(\mathscr{D}) = \max\left(\operatorname{size}(\mathscr{T}), \delta t\right), \qquad \operatorname{reg}(\mathscr{D}) = \operatorname{reg}(\mathscr{T}).$$

Remark 1. The choice of uniform time steps is not necessary, and all the results presented here can be adapted to the case of nonuniform time steps.

For $K \in \mathscr{T}_i$, we define $g_K(s) = g_i(s)$ for all functions g whose definition depends on the subdomain Ω_i, as for example $\phi_i, \varphi_i, M_i, f_i, \dots$.

We propose a fully implicit cell-centered finite volume scheme for the problem, whose unknowns at each time step are $(s_K^n, P_K^n)_{K\in\mathscr{T}}$ and an interface unknown $\left(\pi_\sigma^n\right)_{\sigma\in\mathscr{E}_\Gamma}$. For all $\sigma \in \mathscr{E}_{K,\Gamma}$, we define $s_{K,\sigma}^n = \tilde{\pi}_K^{-1}(\pi_\sigma^n)$, so that, if $\sigma = K|L$, one directly has that

$$\pi_\sigma^n \in \tilde{\pi}_K(s_{K,\sigma}^n) \cap \tilde{\pi}_L(s_{L,\sigma}^n).$$

The total flux balance equation (18) is discretized by

$$\sum_{\sigma\in\mathscr{E}_K} m(\sigma) Q_{K,\sigma}^n = 0, \qquad \forall n \in \{1,\dots,N\}, \forall K \in \mathscr{T}, \tag{20}$$

with

$$Q_{K,\sigma}^n = \begin{cases} \frac{M_{K,L}(s_K^n, s_L^n)}{d_{K,L}} \left(P_K^n - P_L^n\right) + \mathscr{R}\left(Z_{K,\sigma}; s_K^n, s_L^n\right) & \text{if } \sigma = K|L \in \mathscr{E}_{K,i}, \\ \frac{M_K(s_K^n)}{d_{K,\sigma}} \left(P_K^n - P_{K,\sigma}^n\right) + \mathscr{R}\left(Z_{K,\sigma}; s_K^n, s_{K,\sigma}^n\right) & \text{if } \sigma \in \mathscr{E}_{K,\Gamma}, \\ 0 & \text{if } \sigma \in \mathscr{E}_{K,\mathrm{ext}}, \end{cases} \tag{21}$$

where $M_{K,L}(s_K^n, s_L^n) = M_{L,K}(s_L^n, s_K^n)$ is an average of $M_K(s_K^n)$ and $M_L(s_L^n)$. For example, we can suppose, as in [10] that it is given by the harmonic mean

$$M_{K,L}(s_K^n, s_L^n) = \frac{M_K(s_K^n) M_K(s_L^n) d_{K,L}}{d_{L,\sigma} M_K(s_K^n) + d_{K,\sigma} M_K(s_L^n)}.$$

The function $Z_{K,\sigma}$ is defined by $Z_{K,\sigma}(s) = \zeta_K(s)\mathbf{g}\cdot\mathbf{n}_{K,\sigma}$, where $\mathbf{n}_{K,\sigma}$ denotes the outward normal to σ with respect to K. For a function f, we denote by $\mathscr{R}(f; a, b)$

the Riemann solver

$$\mathscr{R}(f;a,b) = \begin{cases} \min_{c\in[a,b]} f(c) & \text{if } a \le b, \\ \max_{c\in[b,a]} f(c) & \text{if } b \le a. \end{cases}$$

The oil-flux balance equation (19) is discretized in the form

$$\phi_K \frac{s_K^n - s_K^{n-1}}{\delta t} m(K) + \sum_{\sigma\in\mathscr{E}_K} m(\sigma) F_{K,\sigma}^n = 0, \qquad \forall n \in \{1,\dots,N\}, \forall K \in \mathscr{T}, \tag{22}$$

with

$$F_{K,\sigma}^n = \begin{cases} Q_{K,\sigma}^n \; f_K(\overline{s}_{K,\sigma}^n) + \mathscr{R}(G_{K,\sigma}; s_K^n, s_L^n) + \frac{\varphi_K(s_K^n)-\varphi_K(s_L^n)}{d_{K,L}} \text{ if } \sigma = K|L \in \mathscr{E}_{K,i}, \\ Q_{K,\sigma}^n \; f_K(\overline{s}_{K,\sigma}^n) + \mathscr{R}(G_{K,\sigma}; s_K^n, s_{K,\sigma}^n) + \frac{\varphi_K(s_K^n)-\varphi_K(s_{K,\sigma}^n)}{d_{K,\sigma}} \text{ if } \sigma \in \mathscr{E}_{K,\Gamma}, \\ 0 \text{ if } \sigma \in \mathscr{E}_{K,\text{ext}}, \end{cases} \tag{23}$$

where $G_{K,\sigma}(s) = \gamma_K(s)\mathbf{g}\cdot\mathbf{n}_{K,\sigma}$ and $\overline{s}_{K,\sigma}^n$ is the upstream value defined by

$$\overline{s}_{K,\sigma}^n = \begin{cases} s_K^n & \text{if } Q_{K,\sigma}^n \ge 0, \\ s_L^n & \text{if } Q_{K,\sigma}^n < 0 \text{ and } \sigma = K|L \in \mathscr{E}_{K,i}, \\ s_{K,\sigma}^n & \text{if } Q_{K,\sigma}^n < 0 \text{ and } \sigma \in \mathscr{E}_{K,\Gamma}. \end{cases} \tag{24}$$

The interface values $\left(\pi_\sigma^n, P_{K,\sigma}^n, P_{L,\sigma}^n\right)$ for $\sigma = K|L \in \mathscr{E}_\Gamma$ are defined by the following nonlinear system:

$$P_{K,\sigma}^n - \lambda_{w,K}\left(\pi_\sigma^n\right) = P_{L,\sigma}^n - \lambda_{w,L}\left(\pi_\sigma^n\right). \tag{25}$$

$$Q_{K,\sigma}^n + Q_{L,\sigma}^n = 0, \tag{26}$$

$$F_{K,\sigma}^n + F_{L,\sigma}^n = 0. \tag{27}$$

Note that since the equations (25) and (26) are linear with respect to $P_{K,\sigma}^n$ and $P_{L,\sigma}^n$, one can eliminate these interface values, only keeping π_σ^n. We impose the discrete counterpart of (17), that is

$$\sum_{K\in\mathscr{T}} m(K) P_K^n = 0, \qquad \forall n \in \{1,\dots,N\}. \tag{28}$$

The discrete initial data is given by:

$$s_K^0 = \frac{1}{m(K)} \int_K s_0(x)dx, \quad \forall K \in \mathscr{T},$$

so that $0 \le s_K^0 \le 1$.

Proposition 1 (existence of a discrete solution). *For all $n \in \{1, \dots, N\}$, there exists $\left((s_K^n)_{K \in \mathscr{T}}, (P_K^n)_{K \in \mathscr{T}}, (\pi_\sigma^n)_{\sigma \in \mathscr{E}_\Gamma} \right)$ satisfying the relations (20)–(28). Moreover,*

$$0 \le s_K^n \le 1, \quad \forall K \in \mathscr{T}. \tag{29}$$

The proof of Proposition 1 will be given in the forthcoming paper [2].

For an admissible discretization $\mathscr{D}$ of Q, we denote by $s_{\mathscr{D}}$ and $P_{\mathscr{D}}$ the piecewise constant functions defined almost everywhere by

$$s_{\mathscr{D}}(x,t) = s_K^n, \quad P_{\mathscr{D}}(x,t) = P_K^n \quad \text{if } (x,t) \in K \times (t^{n-1}, t^n].$$

We consider now a sequence $(\mathscr{D}_m)_{m \ge 0}$ of admissible discretizations of Q in the sense of Definition 2 such that size$(\mathscr{D}_m)$ tends to 0 and reg$(\mathscr{D}_m)$ remains uniformly bounded as m tends to ∞. We denote by $(s_{\mathscr{D}_m}, P_{\mathscr{D}_m})_m$ a corresponding sequence of discrete solutions, whose existence has been stated in Proposition 1.

Theorem 1 (main result). *There exists a weak solution (s, P) in the sense of Definition 1 such that, up to a subsequence,*

$$s_{\mathscr{D}_m} \to s \text{ and a.e. in } Q \text{ as } m \to \infty,$$

$$P_{\mathscr{D}_m} \to P \text{ weakly in } L^2(Q) \text{ as } m \to \infty.$$

The proof of Theorem 1 that we will present in the forthcoming paper [2] is based on compactness arguments, using the material developed in [9,10]. The proof adapts the steps that are given in [6] for the continuous frame.

3 Numerical Results

We consider a model porous medium $\Omega = (0,1)^2$ composed of two layers $\Omega_1 = \{(x,y) \in \Omega \mid y < \Gamma(x)\}$ and $\Omega_2 = \{(x,y) \in \Omega \mid y > \Gamma(x)\}$, which have different capillary pressure laws. The fluid densities are given by $\rho_o = 0.81$, $\rho_w = 1$, and $\mathbf{g} = -9.81\mathbf{e}_y$. We suppose that the porosity is such that $\phi_i = 1, i \in \{1,2\}$, and we define the oil and water mobilities by

$$\eta_{o,i}(s) = 0.5s^2 \text{ and } \eta_{w,i} = (1-s)^2, \, i \in \{1,2\}.$$

Moreover we suppose that the capillary pressure curves have the form

$$\pi_1(s) = s \text{ and } \pi_2(s) = 0.5 + s.$$

and that the initial saturation is given by

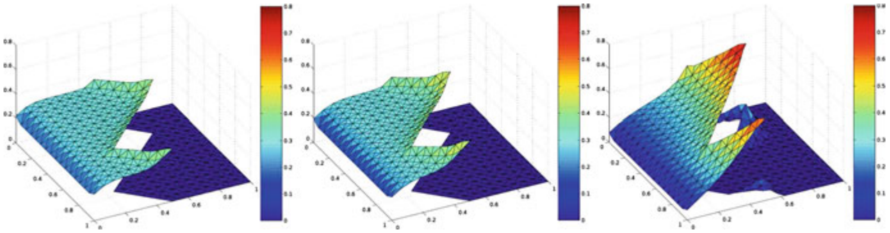


Fig. 1 Saturation for $t = 0.06$, $t = 0.11$ and $t = 0.6$

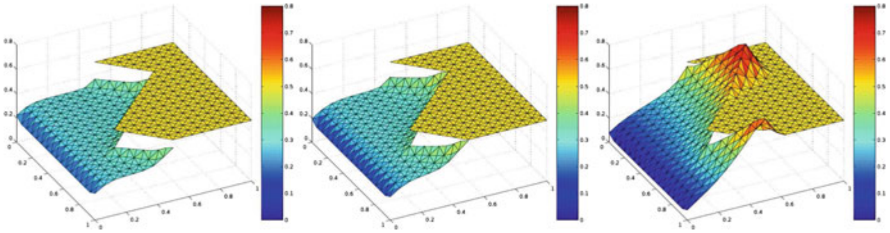


Fig. 2 Capillary pressure for $t = 0.06$, $t = 0.11$ and $t = 0.6$

$$s_0(x) = \begin{cases} 0.3 & \text{if } x \in \Omega_1, \\ 0 & \text{otherwise.} \end{cases}$$

The flow is driven by buoyancy, making the oil move along $\mathbf{e}_y$ until it reaches the interface Γ. For $t \leq 0.11$, oil can not access the domain Ω_2, since the capillary pressure $\pi_1(s_1)$ is lower than the threshold value $\pi_2(0) = 0.5$, which is called *the entry pressure*, see the Fig. 2. Hence the saturation (see the Fig. 1) below the interface s_1 increases, as well as the capillary pressure $\pi_1(s_1)$. As soon as the capillary pressure $\pi_1(s_1)$ reaches the entry pressure $\pi_2(0)$, oil starts to penetrate in the domain Ω_2. Nevertheless, as pointed out in [1,4], a finite quantity of oil remains trapped under the rock discontinuity. This phenomenon is called *oil trapping*.

Acknowledgements This work was supported by the GNR MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN), France.

References

1. M. Bertsch, R. Dal Passo, and C. J. van Duijn. Analysis of oil trapping in porous media flow. *SIAM J. Math. Anal.*, 35:245–267, 2003.
2. K. Brenner, C. Cancès, and D. Hilhorst. Convergence of a finite volume approximation of an immiscible two-phase flow in porous media with discontinuous capillary pressure field. In preparation.

3. F. Buzzi, M. Lenzinger, and B. Schweizer. Interface conditions for degenerate two-phase flow equations in one space dimension. *Analysis*, 29:299–316, 2009.
4. C. Cancès. Finite volume scheme for two-phase flow in heterogeneous porous media involving capillary pressure discontinuities. *M2AN*, 43:973–1001, 2009.
5. C. Cancès, T. Gallouët, and A. Porretta. Two-phase flows involving capillary barriers in heterogeneous porous media. *Interfaces Free Bound.*, 11(2):239–258, 2009.
6. C. Cancès and M. Pierre. An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure fields. HAL : hal-00518219, 2010.
7. G. Chavent and J. Jaffré. *Mathematical Models and Finite Elements for Reservoir Simulation*, volume 17. North-Holland, Amsterdam, stud. math. appl. edition, 1986.
8. G. Enchéry, R. Eymard, and A. Michel. Numerical approximation of a two-phase flow in a porous medium with discontinuous capillary forces. *SIAM J. Numer. Anal.*, 43(6):2402–2422, 2006.
9. R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.
10. A. Michel. A finite volume scheme for two-phase immiscible flow in porous media. *SIAM J. Numer. Anal.*, 41(4):1301–1317 (electronic), 2003.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Uncertainty Quantification for a Clarifier–Thickener Model with Random Feed

Raimund Bürger, Ilja Kröker, and Christian Rohde

Abstract The continuous sedimentation process in a clarifier–thickener can be described by a scalar nonlinear conservation law for the solid volume fraction. The flux is discontinuous with respect to space due to the feed mechanism. Typically the feed flux cannot be given in an exact manner. To quantify uncertainty the unknown solid concentration and the feed bulk flow are expressed by polynomial chaos. A deterministic hyperbolic system for a finite number of stochastic moments is constructed. For the resulting high-dimensional system a simple finite volume scheme is presented. Numerical experiments cover one- and two-dimensional situations.

Keywords Clarifier–Thickener model, Polynomial chaos, Uncertainty quantification, Galerkin projection, Finite–Volume method
MSC2010: 65M08, 68U20, 35R60

1 Introduction

Modelling uncertainty is important in many technical applications. Straightforward Monte-Carlo computations are easy but computationally inefficient or even impossible. The quantification of randomness by stochastic Galerkin or collocation methods seems to be more promising in many situations as this leads to deterministic models for at least a finite number of stochastic moments (cf. [MK05] for an overview).

Raimund Bürger
CI^{2}MA and Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C
Concepcion, Chile, e-mail: rburger@ing-mat.udec.cl

Ilja Kröker and Christian Rohde
IANS, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany,
e-mail: ikroeker|crohde@mathematik.uni-stuttgart.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_21, © Springer-Verlag Berlin Heidelberg 2011

Roughly speaking, there is by now a well-understood theory for models that can be described by linear partial differential equations. What concerns nonlinear problems –we are interested in hyperbolic conservation laws– first steps have been done just recently [Abg07, PDL09, TLMNE10].

As a prototype model in this field we consider a clarifier–thickener (CT) model for the continuous fluid-solid separation of suspensions under gravity. The CT model provides an idealized description of secondary settling tanks in waste water treatment or of thickeners in mineral processing [BCBT99]. Typically, many input parameters can not be described with deterministic accuracy but behave noisily. We take into account two stochastic dimensions: the uncertainty of the rate of inflow of feed suspension and that of the fraction of solid material. This uncertainty produces a hyperbolic equation with a doubly random flux function. To be precise, consider the longitudinal-infinite vessel $D := \mathbb{R} \times S \subset \mathbb{R}^d$ with the cross-sectional domain $S \subset \mathbb{R}^{d-1}$ and coordinates $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T$. The longitudinal direction is aligned with gravity. For a final time $T > 0$ we search then as the unknown the solid volume fraction $u : D_T := D \times (0, T) \to [0, 1]$. According to [BKRT04, BWC00] the sedimentation process can be modelled by the initial value problem

$$\begin{aligned} u_t(\mathbf{x}, t, \omega) + \operatorname{div}\big(\mathbf{h}(\mathbf{x}, t, u(\mathbf{x}, t, \omega))\big) &= \delta(x_1) Q_{\mathrm{F}}(t, \omega_1) u_{\mathrm{F}}(t, \omega_2) && \text{in } D_T \times \Omega, \\ u(., 0) &= 0 && \text{in } D. \end{aligned} \tag{1}$$

The nonlinear flux is given by

$$\mathbf{h}(\mathbf{x}, t, u) = \mathbf{q}(\mathbf{x}, t)u + (\chi_{(-1,1)\times S}(\mathbf{x})b(u), 0, \cdots, 0)^T,$$

where b is the given nonlinear batch flux density function. The vector field $\mathbf{q} = \mathbf{q}(\mathbf{x}, t) \in \mathbb{R}^d$ is the volume average flow velocity which satisfies a coupled Navier–Stokes-like system [BWC00]. For simplicity, we assume $\mathbf{q}$ to be a given deterministic quantity whose transversal components vanish on $\mathbb{R} \times \partial D$. Furthermore, $\chi_{(-1,1)\times S}$ is the characteristic function for the set $(-1, 1) \times S$. This choice describes the upper overflow boundary and the lower discharge boundary of the vessel. The right-hand side in (1) models the stochastic feed process. For probability measures P_1, P_2 let $\Omega = ((\Omega_1, P_1), (\Omega_2, P_2))$ be the vector-valued probability space. By $Q_{\mathrm{F}} = Q_{\mathrm{F}}(t, \omega_1) > 0$, $\omega_1 \in \Omega_1$, we denote the random feed rate and by $u_{\mathrm{F}} = u_{\mathrm{F}}(t, \omega_2) \in [0, 1]$, $\omega_2 \in \Omega_2$, the feed solid volume fraction. For the idealized vessel we assume that the feed source is distributed over the whole cross section $\{0\} \times S$, i.e. δ denotes the Dirac function in (1). As we will show below, the complete feed term in (1) can be rewritten as part of the flux such that (1) gets the form of a nonlinear conservation law with discontinuous flux. To our knowledge such a situation has not yet been treated in the framework of uncertainty quantification.

In Sect. 2 we detail the model and introduce an approximation for the stochastic process u by a polynomial chaos (PC-) ansatz. A numerical scheme for the PC-system on the base of the Lax–Friedrichs approach is presented. Note that the

Engquist–Osher flux, which is usually applied for problems with discontinuous flux, cannot be used for the higher-dimensional PC-system. Finally, in Sect. 3 numerical experiments are displayed.

2 A Polynomial Chaos Approach for Discontinuous Fluxes

2.1 Formulation of the Model

For notational simplicity we choose $d = 1$ (i.e. $S = \emptyset$) in (1) and use $x = x_1$ for the remaining vertical coordinate. The source term is formally rewritten as

$$\delta(x) Q_{\mathrm{F}}(t, \omega_1) u_{\mathrm{F}}(t, \omega_2) = (H(x) Q_{\mathrm{F}}(t, \omega_1) u_{\mathrm{F}}(t, \omega_2))_x, \tag{2}$$

where H denotes the Heaviside function. Following [BKRT04] we obtain then the flux formulation form

$$u_t(x, t, \omega) + \big(g(x, t, u, \omega)\big)_x = 0 \qquad \text{in } \mathbb{R} \times (0, T) \times \Omega. \tag{3}$$

The flux function g is determined for $t \in (0, T)$ and $\omega \in \Omega$ by the flux in (1) (see assumptions below) minus the flux in (2). This leads to

$$g(x, t, u, \omega) := \begin{cases} q_{\mathrm{L}}(t, \omega_1)(u - u_{\mathrm{F}}(t, \omega_2)) & \text{for } x < -1, \\ q_{\mathrm{L}}(t, \omega_1)(u - u_{\mathrm{F}}(t, \omega_2)) + b(u) & \text{for } -1 < x < 0, \\ q_{\mathrm{R}}(t, \omega_1)(u - u_{\mathrm{F}}(t, \omega_2)) + b(u) & \text{for } 0 < x < 1, \\ q_{\mathrm{R}}(u - u_{\mathrm{F}}(t, \omega_2)) & \text{for } x > 1. \end{cases} \tag{4}$$

To obtain this representation, firstly we have made for $q = q(x, t)$ the ansatz

$$q(x, t) = \begin{cases} q_{\mathrm{L}}(t, \omega_1) & \text{for } x < 0, \\ q_{\mathrm{R}} & \text{for } x > 0, \end{cases} \qquad q_{\mathrm{L}}(., \omega_1) \in C^1([0, T)), q_{\mathrm{L}}(., \omega_1) < 0,\ q_{\mathrm{R}} > 0.$$

Stochasticity is solely attached to q_{L}. Secondly, to ensure global conservativity, we have chosen $Q_{\mathrm{F}}(t, \omega_1) = q_{\mathrm{R}} - q_{\mathrm{L}}(t, \omega_1)$.

The flux (4) has discontinuities for $x \in \{-1, 0, 1\}$. We will not directly work with (3) but expand the equation to a system. For $x \in \mathbb{R}$, $t \in [0, T)$, $\omega_1 \in \Omega_1$ we define

$$\gamma^1(x, t, \omega_1) := \begin{cases} q_{\mathrm{L}}(t, \omega_1) & \text{for } x < 0, \\ q_{\mathrm{R}} & \text{for } x > 0, \end{cases} \qquad \gamma^2(x, t) := \begin{cases} 1 & \text{for } x \in (-1, 1), \\ 0 & \text{for } x \notin (-1, 1). \end{cases} \tag{5}$$

With the flux $f(t,u,\gamma^1,\gamma^2,\omega_2) := \gamma^1(\cdot,\omega_1)(u-u_{\mathrm{F}}(t,\omega_2)) + \gamma^2 b(u)$ we can understand (3), (4) as a (only seemingly trivial) system of balance laws

$$\begin{aligned} u(x,t,\omega)_t + \big(f(t,u,\gamma^1,\gamma^2,\omega_2)\big)_x &= 0, \\ \gamma_t^1(x,t,\omega_1) = H(-x)q_{\mathrm{L},t}(t,\omega_1), &\qquad \gamma_t^2(x,t) = 0 \end{aligned} \tag{6}$$

for the unknown vector $(u,\gamma^1,\gamma^2)^T \in [0,1]\times\mathbb{R}^2$.

2.2 Polynomial Chaos Representation

Let $\theta = \theta(\omega) = (\theta_1(\omega_1),\theta_2(\omega_2))^T \in \mathbb{R}^2$ be a vector of i.i.d. (independent identically distributed) random variables. Define

$$\psi_{jk}(\theta) = \phi_j(\theta_1)\phi_k(\theta_2) \quad (j,k\in\mathbb{N}_0),$$

where ϕ_k is the k-th Legendre polynomial. Then $\{\psi_{jk}(\theta)\}_{j,k\in\mathbb{N}_0}$ is a family of $L^2(\Omega_1\times\Omega_2)$-orthonormal polynomials in the sense

$$\big\langle \psi_{jk}(\theta),\psi_{lm}(\theta)\big\rangle_{L^2(\Omega)} := \int_{\Omega_1}\int_{\Omega_2} \psi_{jk}(\theta)\psi_{lm}(\theta)\,dP_1(\omega_1)dP_2(\omega_2) = \delta_{jk}\delta_{lm}. \tag{7}$$

We recall that for some second order random field $w = w(x,t,\omega)$ the polynomial chaos (PC-) representation

$$w(x,t,\omega) = \sum_{j,k\in\mathbb{N}_0} w^{jk}(x,t)\psi_{jk}(\theta(\omega)),\ w^{jk} := \int_{\Omega_1}\int_{\Omega_2} w\psi_{jk}\,dP_1(\omega_1)dP_2(\omega_2) \tag{8}$$

holds [GS91]. For the sake of a more handsome notation let $w^0,\dots,w^P$ for $P = P(M) = (M+1)(M+2)/2-1$ be an arbitrary but fixed re-indexing of the set $\{w^{jk} \mid j,k\in\mathbb{N}_0, j+k\le M\}$. The M-th order approximation of $w(x,t,\omega)$ in (8) is given by

$$(\Pi^P w)(x,t,\omega) := \sum_{p=0}^{P} w^p(x,t)\psi_p(\theta(\omega)).$$

The standard stochastic Galerkin approach (for the first equation in (6)) reads as follows. For $M\in\mathbb{N}_0$ find $u^0,\dots,u^P : D\times(0,T)\to\mathbb{R}$ such that

$$\int_{\Omega_1}\int_{\Omega_2}\!\Big(\Pi^P u + \big(\Pi_2^M\gamma^1\left(\Pi^P u - \Pi_1^M u_{\mathrm{F}}\right) + \gamma^2 b\left(\Pi^P u\right)\big)_x\Big)\,\Psi_q\,dP_1(\omega_1)dP_2(\omega_2) = 0 \tag{9}$$

holds for $q = 0, \dots, P$. We used for the given, stochastically one-dimensional approximation of u_F the notation $\Pi_1^M u_{\mathrm{F}}$. An analogous formulation holds for the unknown (stochastically one-dimensional) approximation $\Pi_2^M \gamma^1$ of γ^1.

Using now the orthogonality from (7) the equations (9) can be written in the form

$$u_t^p + \left(\sum_{m=0}^{M} \sum_{q=0}^{P} {\gamma^1}^m u^q c_{mqp} - \sum_{m,l=0}^{M} {\gamma^1}^m u_{\mathrm{F}}^l(t) d_{mlp} + \gamma^2 \mathbb{E}\left[b\left(\Pi^P u \right) \psi_p \right] \right)_x = 0, \tag{10}$$

with $p = 0, \dots, P$. Here $\mathbb{E}$ denotes the expectation value and

$$\begin{aligned} c_{mqp} &= \int_{\Omega_1} \int_{\Omega_2} \phi_m(\omega_2) \psi_q(\omega) \psi_p(\omega)\, dP(\omega_1)\, dP(\omega_2), \\ d_{mlp} &= \int_{\Omega_1} \int_{\Omega_2} \phi_m(\omega_2) \phi_l(\omega_1) \psi_p(\omega)\, dP(\omega_1)\, dP(\omega_2). \end{aligned} \tag{11}$$

Below we choose b to be a polynomial such that the expectation in (10) can be computed exactly.

We obtain finally from (10) and equations for ${\gamma^1}^0, \dots, {\gamma^1}^M$ the $(P + M + 3)$-dimensional PC-system. Using the definition of the coefficients in (11) and the (weak) hyperbolicity of (6) it can be shown that the PC-system (10) is weakly hyperbolic.

2.3 1D Finite–Volume Method

The PC-system (10) is quite general and it appears hard to construct e.g. a Godunov-type solver. Therefore, at least in this paper, we use the simple Lax–Friedrichs method on a uniform mesh with cells $[x_{i-1/2}, x_{i+1/2})$, $i \in \mathbb{Z}$ and $\Delta x = x_{i+1/2} - x_{i-1/2}$. Restricting to the u-components $u^0, \dots, u^P$ we have for time step $\Delta t^n > 0$ the scheme

$$u_i^{p,n+1} = u_i^{p,n} - \frac{\Delta t^n}{\Delta x} \left(F_{i+1/2}^{p,n} - F_{i-1/2}^{p,n} \right) \qquad (i \in \mathbb{Z},\ n \in \mathbb{N},\ p = 0, \dots, P),$$

$$\begin{aligned} F_{i+1/2}^{p,n} := \frac{1}{2} \Big(& f^p(t^n, u_i^{0,n}, \dots, u_i^{P,n}, {\gamma_i^1}^{0,n}, \dots, {\gamma_i^1}^{M,n}, \gamma_i^{2,n}) \\ & + f^p(t^n, u_{i+1}^{0,n}, \dots, u_{i+1}^{P,n}, {\gamma_{i+1}^1}^{0,n}, \dots, {\gamma_{i+1}^1}^{M,n}, \gamma_{i+1}^{2,n}) \Big) + \frac{\Delta x}{2 \Delta t^n} (u_{i+1}^{p,n} - u_i^{p,n}). \end{aligned}$$

The function f^p is defined by

$$f^p(t,u^0,\dots,u^P,\gamma^{1^0},\dots,\gamma^{1^M},\gamma^2) =$$
$$\sum_{m=0}^{M}\sum_{q=0}^{P}\gamma^{1^m}u^q c_{mqp} - \sum_{m,l=0}^{M}\gamma^{1^m}u_{\mathrm{F}}^l(t)d_{mlp} + \gamma^2\mathbb{E}\left[b\left(\Pi^P u\right)\psi_p\right].$$

Initial values are obtained from $u_i^0 = \dots = u_i^{P,0} = \gamma_i^{1^{1,0}} = \dots = \gamma_i^{1^{M,0}} = 0$ (cf. (1)) and averaging γ^1, γ^2 from (5) for $\gamma_i^{1^{0,0}}, \gamma_i^{2,0}$.

3 Numerical Experiments

Example 1: [1D Computation with one random dimension]
We consider the problem (3) with the batch flux function $b(w) := \frac{27}{4}w((1-w)^2)$ [BKRT04] and $u_0 = 0$. The solid volume feed fraction u_{F} satisfies

$$u_{\mathrm{F}}(t,\omega_1) := 0.6 + 0.2\theta(\omega_2),$$

such that θ is uniformly distributed on [0, 1]. Consequently the random variable u_{F} has the expectation 0.7. No further uncertainty is assumed. We choose $q_{\mathrm{L}} = -1$, $q_{\mathrm{R}} = 0.6$. Figure 1 shows (total view and blow-up close to inflow) the numerical solution with $P = 5$ together with the numerical solution of the deterministic problem using $u_{\mathrm{F}} \equiv 0.7$ and the numerical Monte-Carlo approach with 5000 samples computed with $\Delta x = 0.01$. We use Lax–Friedrichs method for our computation. Almost no differences can be detected.

This is confirmed by the subsequent table which displays the $L^1(\mathbb{R})$-difference between the Monte-Carlo sample solution and the PC-approach for $P = 1,\dots,6$.

P	1	2	3	4	5	6
L^1-Error	1.1372e-02	1.5566e-02	3.2322e-03	1.4975e-03	8.5714e-04	5.0671e-04

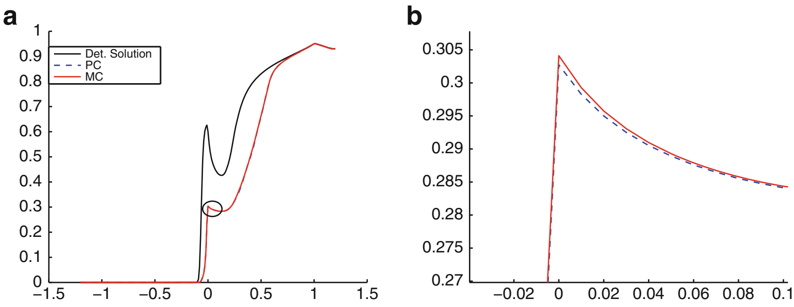


Fig. 1 Solid volume fraction for the deterministic case using the expectation value of the feeding rate, PC-solution, and Monte-Carlo samples. Blow-up in the right figure

We here observe a clear convergence for a reasonable number of stochastic modes.

Example 2: [1D Computation with two random dimensions]
We choose the same setting as in Example 1 but introduce the second random dimension in the suspension feed rate via

$$q_{\mathrm{L}}(t, \omega_1) = -1.2 + 0.4\theta(\omega_1).$$

Again let θ be uniformly distributed on the interval $[0, 1]$. Figure 2 shows the numerical solution with $M = 3$ and $P = 9$. This is compared with the numerical solution of the deterministic problem using the expectation values $q_{\mathrm{L}} = -1$ and $u_{\mathrm{F}} \equiv 0.7$, and the numerical Monte-Carlo approach with 50000 samples at time $T = 1$.

Already for this low random (and spatial) dimension we immediately attain the limits of available computing power. The table below shows the computing time of the PC-approach.

M (P)	1 (2)	2 (5)	3 (9)	4 (14)	5 (20)
cpu-time [s]	1.3721e+03	3.9463e+03	1.2037e+04	3.5001e+04	6.6399e+04

Example 3: [2D Computation with one random dimension]
Let us consider the CT problem (1) for $d = 2$ and $S = (-1.2, 1.2)$, with flux components $h_1(\mathbf{x}, t, u, \omega) = g(x_1, t, u, \omega)$ defined in (4) and $h_2(\mathbf{x}, t, u, \omega) = 0.02{*}\cos(\frac{\pi x_2}{0.6})u$ This corresponds not to a realistic velocity field $\mathbf{q}$ but we understand this example as a test case for the uncertainty quantification. The batch flux function b, solid volume feed fraction $u_F(t, \omega_1)$, and q_L, q_R are as in Example 1. For the numerical approximation we use an adaptive finite-volume method based on

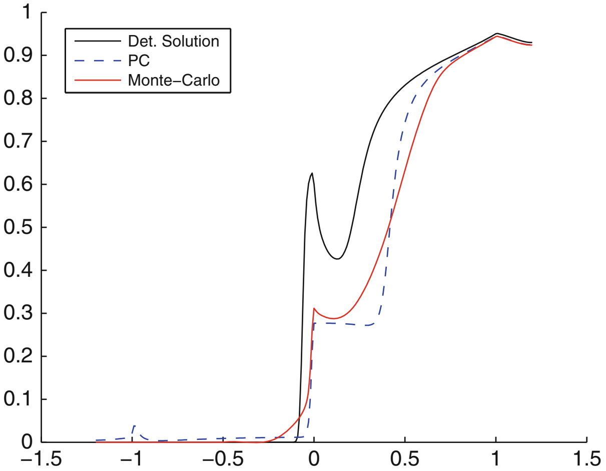


Fig. 2 Solid volume fraction for the deterministic case using the expectation values for solid fraction feeding rate and suspension feeding rate, PC-solution, and Monte-Carlo samples

unstructured triangular meshes with the Lax–Friedrichs flux (cf. [Krö08]). Initially 4608 triangles are used.

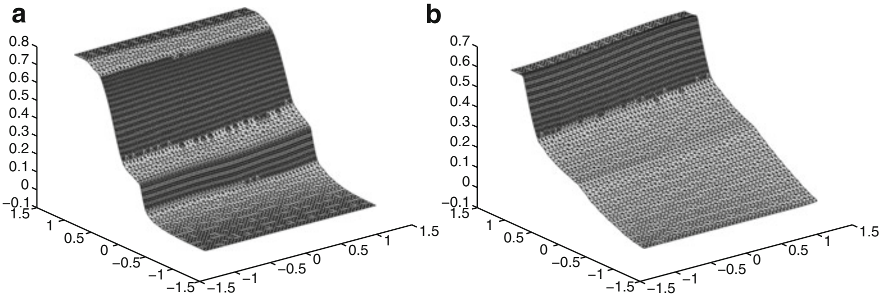


Fig. 3 Solid volume fraction for the deterministic case using the expectation values for solid fraction feeding rate and suspension feeding rate (a), and PC-solution (b) at time $T = 1$

Figure 3(a) shows a deterministic computation with $u_F = 0.7$ and the PC-solution with $P = 7$ (Fig. 3(b)). As in the 1D computations the PC-solution is much smoother and does not develop a peak close to the inlet. As a consequence the adaptive algorithm uses a coarser grid for the PC-solutions. To be specific, at $T = 1$ we had 11826 triangles for the deterministic computation, 8280 for $P = 7$, and 4608 for $P = 1$ (no refinement). Because of the long computation time of each deterministic solution, the computational effort of the Monte-Carlo simulation with a considerable number of samples significantly is higher then the computational effort of the PC approach.

Acknowledgements R. B. acknowledges support by Fondecyt project 1090456, BASAL project CMM, Universidad de Chile and Centro de Investigación en Ingeniería Matemática (CI^{2}MA), Universidad de Concepción. I. K. and C. R. 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

[Abg07] R. Abgrall. A simple, flexible and generic deterministic appoarch to uncertainty quantifications in non linear problems: application to fluid flow problems. 2007.

[BCBT99] M.C. Bustos, F. Concha, R. Bürger, and E. M. Tory. *Sedimentation and thickening*, volume 8 of *Mathematical Modelling: Theory and Applications*. Kluwer Academic Publishers, Dordrecht, 1999. Phenomenological foundation and mathematical theory.

[BKRT04] R. Bürger, K. H. Karlsen, N. H. Risebro, and J. D. Towers. Well-posedness in BV_t and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units. *Numer. Math.*, 97(1):25–65, 2004.

[BWC00] R. Bürger, W. L. Wendland, and F. Concha. Model equations for gravitational sedimentation-consolidation processes. *ZAMM Z. Angew. Math. Mech.*, 80(2): 79–92, 2000.

[GS91] R. G. Ghanem and P. D. Spanos. *Stochastic finite elements: a spectral approach.* Springer-Verlag, New York, 1991.

[Krö08] I. Kröker. Finite volume methods for conservation laws with noise. In *Finite volumes for complex applications V*, pages 527–534. ISTE, London, 2008.

[MK05] H. G. Matthies and A. Keese. Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations. *Comput. Methods Appl. Mech. Engrg.*, 194(12-16):1295–1331, 2005.

[PDL09] G. Poëtte, B. Després, and D. Lucor. Uncertainty quantification for systems of conservation laws. *J. Comput. Phys.*, 228(7):2443–2467, 2009.

[TLMNE10] J. Tryoen, O. Le Maître, M. Ndjinga, and A. Ern. Intrusive Galerkin methods with upwinding for uncertain nonlinear hyperbolic systems. *J. Comput. Phys.*, 229(18):6485–6511, 2010.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Asymptotic preserving schemes in the quasi-neutral limit for the drift-diffusion system

Chainais-Hillairet Claire and Vignal Marie-Hélène

Abstract We are interested in the drift-diffusion system near quasi-neutrality. For this system, classical explicit schemes are decoupled but subject to severe numerical constraints in the quasi-neutral regime. By constrast, the implicit discretizations are unconditionally stable but non linearly coupled. Then, an iterative method must be used yielding a large numerical cost. Here, we propose a new decoupled asymptotic preserving scheme. We perform one and two dimensional numerical experiments which show its good behavior.

Keywords drift-diffusion, asymptotic preserving schemes, quasi-neutral regime
MSC2010: 65M08, 65M12

1 Presentation of the problem

Let $\Omega \subset \mathbb{R}^d$ $(d \geq 1)$ be an open bounded domain describing the geometry of a semiconductor device. The unknowns of the linear drift-diffusion system are the density of electrons and holes, N and P, and the electrostatic potential Ψ. It writes:

$$\partial_t N + \mathrm{div}(-\nabla N + N\nabla\Psi) = 0 \text{ on } \Omega \times [0,T], \tag{1a}$$

$$\partial_t P + \mathrm{div}(-\nabla P - P\nabla\Psi) = 0 \text{ on } \Omega \times [0,T], \tag{1b}$$

$$-\lambda^2 \Delta\Psi = P - N + C \text{ on } \Omega \times [0,T], \tag{1c}$$

Chainais-Hillairet Claire
Laboratoire P. Painlevé, UMR ClNRS 8524, Université Lille 1, 59655 Villeneuve d'Ascq Cédex, e-mail: Claire.Chainais@math.univ-lille1.fr

Vignal Marie-Hélène
Institut de Mathématiques de Toulouse, UMR 5219, Université Paul Sabatier, Toulouse 3, 118 route de Narbonne, 31062 Toulouse Cedex 9, e-mail: mhvignal@math.univ-toulouse.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_22, © Springer-Verlag Berlin Heidelberg 2011

where C is the given doping profile non depending on t. The parameter λ comes from the scaling of the physical model. It is called the rescaled Debye length and is given by the ratio of the Debye length to the size of the domain. The Debye length measures the typical scale of electric interactions in the semiconductor.

The system (1) is supplemented with initial conditions N_0, P_0 and with mixed boundary conditions: Dirichlet boundary conditions on Γ^D (N^D, P^D and Ψ^D) and homogeneous Neumann boundary conditions on Γ^N (with $\partial\Omega = \Gamma^D \cup \Gamma^N$).

We are interested in the so-called quasi-neutral regime. This regime occurs when the parameter λ tends to zero. There has been an intense literature on the rigorous quasi-neutral limit of the drift-diffusion model; we can refer for instance to [9] for a zero doping profile C and to [10] for a regular doping profile.

Many different numerical methods have been already developed for the approximation of (1); see for instance [1] and [12, 13] in the non linear case. The convergence of some finite volume schemes has been proved in [2, 3]. But, up to our knowledge, all the schemes are studied in the case $\lambda = 1$. In this paper, we focus on the behavior of schemes in the quasi-neutral limit, that means when λ tends to zero. In this regime, the local electric charge vanishes everywhere. However, simultaneously, very high frequency oscillations, of order $1/\lambda^2$, are triggered. When a standard explicit scheme is used, the scale of these very high frequency oscillations must be resolved by the time step. Hence, the time step must be smaller than λ^2 otherwise a numerical instability appears. The satisfaction of this constraint requires huge computational resources which makes the explicit methods unusable.

Here, the purpose is to define numerical schemes free of such constraints. For a given time step, we look for schemes which may be used as well as for values of λ of order 1 and for values of λ as small as possible. Furthermore, these schemes must preserve the behavior of the continuous problem in the quasi-neutral limit ($\lambda \to 0$). Such schemes are called asymptotic preserving schemes, this name has been introduced in [11] for relaxation limits of kinetic systems. Asymptotic preserving schemes in the quasi-neutral limit have been developed in [5] for the Euler-Poisson problem and in [6, 7] for the Vlasov-Poisson system. For the drift-diffusion model, implicit strategies have been proposed in [15].

This paper is organized as follows. In Section 2, we present the formal quasi-neutral limit of the drift-diffusion system. Then, in Section 3, we recall two classical schemes and discuss their stability. Section 4 is devoted to the presentation of a new scheme for the drift-diffusion model. Finally, in Section 5, we conclude with numerical simulations.

2 The formal quasi-neutral limit

Formally, passing to the limit $\lambda \to 0$ in system (1) gives the quasi-neutral drift-diffusion system. It is constituted of the mass equations (1a), (1b) and of the quasi-neutrality constraint $P - N + C = 0$. The Poisson equation is lost, and the electrostatic potential becomes the Lagrange multiplier of this constraint. In order to obtain an explicit equation for the potential we subtract the mass equations (1a), (1b) and we remark that thanks to the quasi-neutrality constraint $P - N = -C$. This yields an elliptic equation for the potential: $-\mathrm{div}((P+N)\,\nabla\Psi) = -\Delta C$.

Let us perform the same transformations on the original drift-diffusion system. We begin by subtracting the mass equations. Then, remarking that, thanks to Poisson equation, $\partial_t(P-N) = \partial_t(P-N+C) = \partial_t(-\lambda^2\Delta\Psi)$, we obtain

$$-\lambda^2\,\partial_t\Delta\Psi - \mathrm{div}((P+N)\,\nabla\Psi) = \Delta(P-N). \tag{2}$$

Following [5], we call this equation the reformulated Poisson equation. If P and N are constant, this equation is an order one differential equation on the quantity $-\Delta\Psi$. And, we can note that solutions oscillate in time at the period λ^2.

Thus, an explicit discretization of the electric force terms in (1) will give an explicit discretization of equation (2) and so a stability non uniform in λ. By contrast, an implicit discretization of these terms will give an implicit discretization of (2) and so a stability uniform in λ. This remark will be used in Section 4 for the construction of our decoupled asymptotic preserving scheme.

3 "Classical" schemes

In this section, we present the classical schemes used for the discretization of the drift-diffusion system. The mesh is given by $\mathscr{T}$, a family of control volumes, $\mathscr{E}$, a family of edges and $\mathscr{P} = (x_K)_{K\in\mathscr{T}}$ a family of points. We assume that the mesh is admissible in the sense of [8]. The set of edges will be split into $\mathscr{E} = \mathscr{E}_{int} \cup \mathscr{E}_{ext}$ and for the exterior edges, we distinguish 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 all edge $\sigma \in \mathscr{E}$, 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 $\sigma \in \mathscr{E}_{K,int}$. Then, the transmissibility coefficient is defined by $\tau_\sigma = \mathrm{m}(\sigma)/\mathrm{d}_\sigma$, for all $\sigma \in \mathscr{E}$.

Let Δt be the time step. A finite volume scheme for (1) writes:

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

$$\mathrm{m}(K)\frac{P_K^{n+1} - P_K^n}{\Delta t} + \sum_{\sigma\in\mathscr{E}_K} \mathscr{G}_{K,\sigma}^{n+1} = 0, \forall K \in \mathscr{T}, \forall n \geq 0,$$

$$-\lambda^2 \sum_{\sigma\in\mathscr{E}_K} \tau_\sigma D\Psi_{K,\sigma}^n = \mathrm{m}(K)(P_K^n - N_K^n + C_K), \forall K \in \mathscr{T}, \forall n \geq 0.$$

It remains to define the numerical fluxes $D\Psi_{K,\sigma}^n$, $\mathscr{F}_{K,\sigma}^{n+1}$ and $\mathscr{G}_{K,\sigma}^{n+1}$. As usually, we set $D\Psi_{K,\sigma}^n = \Psi_L^n - \Psi_K^n$ if $\sigma = K|L$, $D\Psi_{K,\sigma}^n = \Psi_\sigma^D - \Psi_K^n$ if $\sigma \in \mathscr{E}_{K,ext}^D$ and $D\Psi_{K,\sigma}^n = 0$

elsewhere. The numerical approximations of the convection-diffusion fluxes in (1a) and (1b), $\mathscr{F}_{K,\sigma}^{n+1}$ and $\mathscr{G}_{K,\sigma}^{n+1}$, are written with the following compact form:

$$\mathscr{F}_{K,\sigma}^{n+1} = \tau_\sigma \left(B(-D\Psi_{K,\sigma}^m) N_K^{n+1} - B(D\Psi_{K,\sigma}^m) N_L^{n+1} \right), \ \forall \sigma \in \mathscr{E}_{int}, \sigma = K|L \quad (3a)$$

$$\mathscr{G}_{K,\sigma}^{n+1} = \tau_\sigma \left(B(D\Psi_{K,\sigma}^m) P_K^{n+1} - B(-D\Psi_{K,\sigma}^m) P_L^{n+1} \right), \ \forall \sigma \in \mathscr{E}_{int}, \sigma = K|L. \quad (3b)$$

If $\sigma \in \mathscr{E}_{K,ext}^D$, we replace N_L^{n+1} by N_σ^D in (3a) and P_L^{n+1} by P_σ^D in (3b). If $\sigma \in \mathscr{E}_{K,ext}^N$, we set $\mathscr{F}_{K,\sigma}^{n+1} = \mathscr{G}_{K,\sigma}^{n+1} = 0$.

The case $m = n$ corresponds to a semi-implicit and decoupled scheme: at each time step $(N_K^{n+1})_{K\in\mathscr{T}}$, $(P_K^{n+1})_{K\in\mathscr{T}}$, and $(\Psi_K^{n+1})_{K\in\mathscr{T}}$, are obtained by solving three linear systems. With $m = n+1$, we write a fully implicit scheme. For the function B, we may choose either $B(x) = 1 - \min(x,0)$ or $B(x) = x/(\exp(x)-1)$ with $B(0) = 1$. The first choice corresponds to a classical two-points discretization of the diffusion with an upwinding for the convection. With the Bernoulli function, we get the Scharfetter-Gummel scheme. One main advantage of this last choice, well-known in semiconductor device simulation, is that the scheme is order 2 in space (see [14]). Moreover, as shown in [4], the Scharfetter-Gummel scheme satisfies some crucial properties like energy and energy dissipation decrease.

The decoupled scheme ($m = n$) has been studied in [2] for $B(x) = 1-\min(x,0)$ and the convergence has been established (for the nonlinear drift-diffusion system). The proof can be extended to the Scharfetter-Gummel scheme (in the linear case). However, in [2], the convergence proof has been done for $\lambda^2 = 1$ and in fact all the a priori estimates (leading to stability, compactness and convergence) depend on λ^2. More precisely, when there is no doping profile or when the doping profile is constant in space, there exists uniform in time L^∞ estimates on the densities N and P (see [10]). In this case, the L^∞ estimates holds at the discrete level, but only under a condition of the form: $\Delta t \leq D\lambda^2$ with $D \in \mathbb{R}$. It means that such a scheme might not be used for small values of λ.

Let us now consider the fully implicit scheme ($m = n+1$). In this case, existence of a solution to the scheme can be proved via a fixed point theorem. Moreover, when the doping profile is constant in space, we can prove that the scheme is unconditionally stable. However, the implementation of the scheme needs the resolution of a nonlinear system of equations at each iteration. This might be done using a Newton's method. It has a numerical cost and the solution is computed up to a precision criterion.

In the next section, we propose a new scheme with the same numerical cost as the decoupled scheme, but remaining stable and consistent when λ tends to 0.

4 Construction of an asymptotic preserving scheme

Following the remark given in Section 2, let us first consider the following semi-discretization of (1) in which the electric force terms are discretized implicitly.

$$\frac{N^{n+1}-N^n}{\Delta t}+\operatorname{div}(-\nabla N^n+N^n\nabla\Psi^{n+1})=0 \text{ on } \Omega\times[0,T], \tag{4a}$$

$$\frac{P^{n+1}-P^n}{\Delta t}+\operatorname{div}(-\nabla P^n-P^n\nabla\Psi^{n+1})=0 \text{ on } \Omega\times[0,T], \tag{4b}$$

$$-\lambda^2\Delta\Psi^{n+1}=P^{n+1}-N^{n+1}+C \text{ on } \Omega\times[0,T]. \tag{4c}$$

We eliminate P^{n+1} and N^{n+1} in (4c) using their expression respectively given in (4b) and (4a). It yields:

$$-\lambda^2\Delta\Psi^{n+1}-\Delta t\operatorname{div}((P^n+N^n)\nabla\Psi^{n+1})=P^n-N^n+C+\Delta t\Delta(P^n-N^n). \tag{5}$$

The semi-discretization given by (4a), (4b) and (5) is uniformally stable in λ but not unconditionaly stable. Then, in order to construct an unconditionally stable semi-discretization we just have to change the discretizations (4a), (4b) into the implicit semi-discretizations of the mass equations.

This corresponds to the following fully discrete scheme:

$$\mathrm{m}(K)\frac{N_K^{n+1}-N_K^n}{\Delta t}+\sum_{\sigma\in\mathscr{E}_K}\mathscr{F}_{K,\sigma}^{n+1}=0, \forall K\in\mathscr{T}, \forall n\geq 0, \tag{6a}$$

$$\mathrm{m}(K)\frac{P_K^{n+1}-P_K^n}{\Delta t}+\sum_{\sigma\in\mathscr{E}_K}\mathscr{G}_{K,\sigma}^{n+1}=0, \forall K\in\mathscr{T}, \forall n\geq 0, \tag{6b}$$

$$\begin{aligned}-\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma(\lambda^2+\Delta t(P_\sigma^n+N_\sigma^n))D\Psi_{K,\sigma}^{n+1}&=\mathrm{m}(K)(P_K^n-N_K^n+C_K)\\&\quad+\Delta t\sum_{\sigma\in\mathscr{E}_K}\tau_\sigma(DP_{K,\sigma}^n-DN_{K,\sigma}^n)\ \forall K\in\mathscr{T}, \forall n\geq 0,\end{aligned} \tag{6c}$$

with the values (3a), (3b) and $m=n+1$ for the numerical fluxes $\mathscr{F}_{K,\sigma}^{n+1}$, $\mathscr{G}_{K,\sigma}^{n+1}$. The interface values, P_σ^n and N_σ^n are defined by taking the mean value between the values of N^n and P^n at two neighboring control volumes. Let us also note that we keep an implicit discretization on N and P in (6a) and (6b) in order to avoid any CFL condition on the time step.

We stress that our scheme is decoupled. It means that, at each time step, if the values $(N_K^n)_{K\in\mathscr{T}}$, $(P_K^n)_{K\in\mathscr{T}}$ are known,

- we first compute $(\Psi_K^{n+1})_{K\in\mathscr{T}}$ by solving the linear system (6c), whose matrix and right-hand-side depend on N^n and P^n,
- then we compute $(N_K^{n+1})_{K\in\mathscr{T}}$ and $(P_K^{n+1})_{K\in\mathscr{T}}$ solutions of the linear systems (6a) and (6b), whose matrices depend on Ψ^{n+1}.

The matrices from (6a) and (6b) are identical to that obtained in the classical decoupled scheme. They are M-matrices, which ensure the positivity at N^n and P^n for all n (starting with positive initial and boundary conditions). However, the numerical analysis of the scheme (6) is not straightforward and is in progress. In the next section, we present the results of numerical simulations in which we compare our new decoupled scheme to the fully implicit scheme. We will focus on the behavior when the rescaled Debye length tends to 0.

5 Numerical experiments

Test case 1. The first test case is a one-dimensional test case (Ω =]0, 1[). The doping profile is a continuous function satisfying $C(x) = -1$ for $0 \leq x \leq 0.4$, $C(x) = +1$ for $0.6 \leq x \leq 1$ and $C(x)$ affine on $[0.4, 0.6]$. The initial and the boundary conditions satisfy the quasi-neutrality condition $P + C - N = 0$, in order to avoid any boundary or initial time layers:

$$N^D = 0, P^D = 1, \Psi^D = 0 \text{ in } x = 0, \quad N^D = 1, \; P^D = 0, \Psi^D = 4 \text{ in } x = 1, \tag{7a}$$

$$N_0(x) = \max(C(x), 0) \quad P_0(x) = -\min(C(x), 0). \tag{7b}$$

With a time step $\Delta t = 10^{-3}$, we run computations with the fully implicit scheme and with the new one for different values of λ^2 on a mesh made of 100 cells. The solution is computed at the final time $T = 1$. For the Newton's method used in the fully implicit scheme the precision criterion is set to 10^{-10} and the maximal number of iterations to 60. In Table 1, we present the CPU times needed by both schemes and also the relative error between the two solutions in a discrete L^2-norm.

We note that the CPU time needed by the new scheme is almost independent of λ. For the fully implicit scheme, we see that for $\lambda^2 \leq 10^{-6}$ the CPU time has a ratio 3 with those of the new scheme. For smaller values of λ^2, it appears some default of convergence of the Newton's method with the given time step for the fully implicit scheme. However, the new scheme still works and we show on Fig. 1(a) the density profiles obtained for $\lambda^2 = 10^{-14}$.

Table 1 Comparison of the fully implicit scheme with the new scheme for the Test Case 1

λ^2	CPU time fully implicit	CPU time new scheme	ratio	relative error on N	relative error on P	relative error on Ψ
1	1.92	0.64	3.00	1.32e-08	1.32e-08	5.94e-09
1e-2	1.82	0.59	3.08	5.73e-06	5.73e-06	2.98e-06
1e-4	2.07	0.59	3.51	2.77e-04	2.77e-04	1.99e-04
1e-6	1.67	0.60	2.78	5.15e-04	5.15e-04	5.70e-04
1e-8	51.46	0.60	85.77	5.24e-04	5.24e-04	5.88e-04

Test Case 2. We change the doping profile for a discontinuous doping profile: $C(x) = -1$ for $x \leq 0.5$ and $C(x) = +1$ for $x \geq 0.5$. We keep (7) as initial

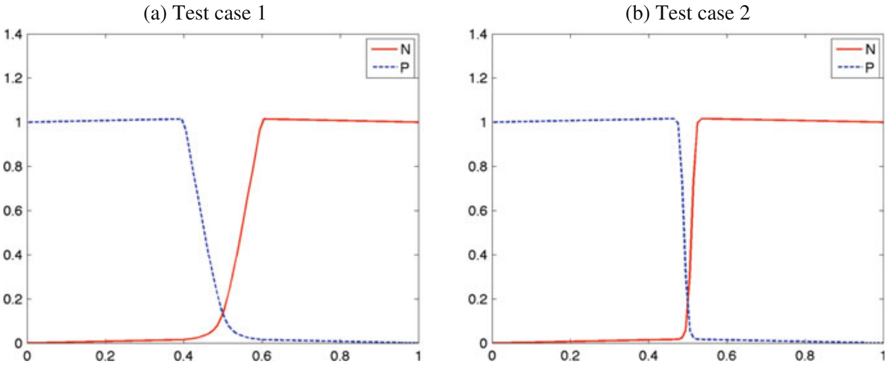


Fig. 1 Density profiles computed by the new scheme for $\lambda^2 = 10^{-14}$ on a mesh made of 100 cells, with $\Delta t = 10^{-3}$

and boundary conditions. The numerical results, presented in Table 2, are similar to those of Test Case 1. We just observe that the relative errors are bigger. This is due to the discontinuity appearing in the density profiles (due to the discontinuity in C): the two schemes do not capture the discontinuity similarly. However, we still note that the new scheme has the same efficiency up to very small values of λ. On Fig. 1(b), we present the density profiles obtained for $\lambda^2 = 10^{-14}$.

Table 2 Comparison of the fully implicit scheme with the new scheme for the Test Case 2

λ^2	CPU time fully implicit	CPU time new scheme	ratio	relative error on N	relative error on P	relative error on Ψ
1	2.09	0.67	3.12	1.31e-08	1.31e-08	5.89e-09
1e-2	1.88	0.60	3.13	7.50e-06	7.50e-06	4.22e-06
1e-4	2.15	0.61	3.52	1.36e-02	1.36e-02	9.51e-03
1e-6	1.73	0.61	2.84	1.03e-01	1.03e-01	6.07e-02
1e-8	51.51	0.60	85.85	1.08e-01	1.08e-01	6.23e-02

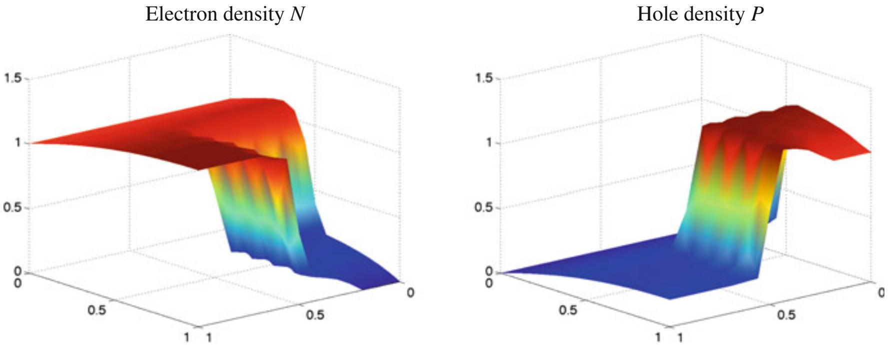
Test Case 3. We consider now the simulation of a two-dimensional forward PN diode. The device is made of two different regions: a P-region with a doping profile equal to -1 and an N-region with a doping profile equal to 1 (see [3]). We use a triangular mesh made of 896 triangles and we set the time step $\Delta t = 5 \cdot 10^{-4}$.

Table 3 shows the efficiency of the new scheme. It really runs faster than the fully implicit scheme. Moreover, the fully implicit scheme did not give results for values of λ^2 less that 10^{-3}, while the new scheme still works. We show on Fig. 2, the density profiles obtained with the new scheme for $\lambda^2 = 10^{-10}$.

As a conclusion, we recall that we have proposed in this paper a new scheme for the drift-diffusion system, whose efficiency is independent of the value of the rescaled Debye length. This scheme can be used at the quasi-neutral limit. Numerical analysis of the scheme is in progress.

Table 3 Comparison of the fully implicit scheme with the new scheme for the Test Case 3

λ^2	CPU time fully implicit	CPU time new scheme	ratio	relative error on N	relative error on P	relative error on Ψ
1	203.28	14.68	13.85	1.13e-01	2.78e-01	2.54e-03
1e-1	219.85	14.52	15.14	8.54e-02	2.19e-01	3.01e-02
1e-2	310.72	14.52	21.40	3.21e-02	1.00e-01	4.50e-02
1e-3	718.09	14.68	48.92	4.84e-02	8.30e-02	7.49e-02



Electron density N

Hole density P

Fig. 2 Test case 3. Density profiles computed by the new scheme for $\lambda^2 = 10^{-10}$ on a mesh made of 896 triangles, with $\Delta t = 5 \cdot 10^{-4}$

References

1. Brezzi F., Marini L.D., Pietra P.: Two-dimensional exponential fitting and applications to drift-diffusion models. SIAM J. Numer. Anal. **26**, 1342–1355 (1989).
2. Chainais-Hillairet, C., Liu, J.-G., Peng, Y.-J.: Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. M2AN **37(2)**,319–338 (2003).
3. Chainais-Hillairet, C., Peng, Y.-J.: Finite volume approximation for degenerate drift-diffusion system in several space dimensions. M3AS **14(3)**, 461–481 (2004).
4. Chatard, M.: Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model. Submitted to this conference.
5. Crispel P., Degond P., Vignal M.-H.: An asymptotic preserving scheme for the two-fluid EulerPoisson model in the quasineutral limit, J. Comput. Phys. **223**, 208–234 (2007).
6. Degond P., Deluzet F., Navoret L.: An asymptotically stable particle-in-cell (PIC) scheme for collisionless plasma simulations near quasineutrality. C.R.Acad. Sci. Paris Ser. I **343**, 613–618 (2006).
7. Degond P., Deluzet F., Navoret L., Sun A-B, Vignal M.-H.: Asymptotic-Preserving Particle-In-Cell method for the Vlasov-Poisson system near quasineutrality. Journal of Computational Physics, **229(16)**, 5630–5652 (2010).
8. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of numerical analysis **VII**, pp. 713–1020. North-Holland, Amsterdam (2000).
9. Gasser I.: The initial time layer problem and the quasineutral limit in a nonlinear drift diffusion model for semiconductors, NoDEA, **8 (3)**, 237–249 (2001).
10. Gasser, I., Levermore, C.D., Markowich, P.A., Schmeiser, C.: The initial time layer problem and the quasineutral limit in the semi-conductor drift-diffusion model. Euro. Jnl of Applied Mathematics **12**, 497–512 (2001).

11. Jin S.: Efficient Asymptotic-Preserving (AP) Schemes for Some Multiscale Kinetic Equations. SIAM J. Sci. Comp. **21(441)** (1999).
12. Jüngel A.: Numerical approximation of a drift-diffusion model for semiconductors with nonlinear diffusion. ZAMM Z. Angew. Math. Mech. **75**, 783–799 (1995).
13. Jüngel A., Pietra P.: A discretization scheme for a quasi-hydrodynamic semiconductor model. Math. Models Methods Appl. Sci. **7**, 935–955 (1997).
14. Lazarov, R.D., Mishev, I.D., Vassilevski, P.S.: Finite volume methods for convection-diffusion problems. SIAM J. Numer. Anal. **33-1**, 31–55 (1996).
15. Ventzek P.L., Hoekstra R., Kushner M.: Two-dimensional modeling of high plasma density inductively coupled sources for materials processing. J. Vac. Sci. Tech. B **12**, 461–477 (1994).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Posteriori Error Estimates for Unsteady Convection–Diffusion–Reaction Problems and the Finite Volume Method

Nancy Chalhoub, Alexandre Ern, Tony Sayah, and Martin Vohralík

Abstract We derive a posteriori error estimates for the discretization of the unsteady linear convection–diffusion–reaction equation approximated with the cell-centered finite volume method in space and the backward Euler scheme in time. The estimates are based on a locally postprocessed approximate solution preserving the conservative fluxes and are established in the energy norm. We propose an adaptive algorithm which ensures the control of the total error with respect to a user-defined relative precision and refines the meshes adaptively while equilibrating the time and space contributions to the error. Numerical experiments illustrate the theory.

Keywords a posteriori estimate, unsteady convection–diffusion–reaction, cell-centered finite volumes, mesh adaptation
MSC2010: 65N15, 76M12, 76S05

1 Introduction

We consider the time-dependent linear convection–diffusion–reaction equation

Nancy Chalhoub and Alexandre Ern
Université Paris-Est, CERMICS, Ecole des Ponts, 77455 Marne-la-Vallée, France,
e-mail: nancy.chalhoub@gmail.com, ern@cermics.enpc.fr

Tony Sayah
Faculté des Sciences, Université Saint-Joseph, B.P. 11-514 Riad El Solh, Beirut 1107 2050, Lebanon, e-mail: tsayah@fs.usj.edu.lb

Martin Vohralík
UPMC Univ. Paris 06, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France &
CNRS, UMR 7598, Laboratoire Jacques-Louis Lions, 75005, Paris, France,
e-mail: vohralik@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_23, © Springer-Verlag Berlin Heidelberg 2011

$$\partial_t u - \nabla\cdot(\mathsf{S}\nabla u) + \nabla\cdot(\boldsymbol{\beta} u) + ru = f \quad \text{a.e. in } Q_T := \Omega \times (0,T), \tag{1a}$$

$$u(\cdot,0) = u_0 \quad \text{a.e. in } \Omega, \tag{1b}$$

$$u = 0 \quad \text{a.e. on } \partial\Omega \times (0,T). \tag{1c}$$

Here S is the diffusion–dispersion tensor, $\boldsymbol{\beta}$ is the velocity field, r is the reaction function, f is the source term, $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is the space domain which we suppose polyhedral, and $(0,T)$ is the time interval. We suppose that $S = (\mathsf{S}_{i,j})$ with $\mathsf{S}_{i,j} \in L^\infty(Q_T)$, $1 \leq i,j \leq d$, is a symmetric, bounded, and uniformly positive definite tensor (we suppose that $\mathsf{S}_{i,j}$ are piecewise constant on space-time meshes defined below), $\boldsymbol{\beta} \in C^0([0,T];[W^{1,\infty}(\Omega)]^d)$, $r \in L^\infty(Q_T)$, $f \in L^2(Q_T)$, and $u_0 \in L^2(\Omega)$.

Several works have studied a posteriori error estimates for the cell-centered finite volume method. Ohlberger derives in [7] estimates in the L^1-norm. Nicaise [6] establishes a posteriori energy-norm estimates using Morley-type interpolants of the original piecewise constant finite volume approximation. Guaranteed flux-based estimates were established in [8] and extended in [3] to the parabolic case. Estimates for vertex-centered unsteady convection–diffusion–reaction problems were derived in [1] and [5].

The purpose of this work is to derive guaranteed a posteriori error estimates for the discretization of (1a)–(1c) by the cell-centered finite volume method in space and the backward Euler scheme in time. We allow for time-varying meshes.

2 Notation and Continuous Problem

2.1 Notation

We consider a strictly increasing sequence of discrete times $\{t^n\}_{0\leq n\leq N}$ such that $t^0 = 0$ and $t^N = T$. For all $1 \leq n \leq N$, we define $\tau^n := t^n - t^{n-1}$ and $I^n := (t^{n-1}, t^n]$. On each time interval I^n, we consider partition $\mathcal{T}^n$ of Ω such that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}^n} K$. For simplicity, we assume that the meshes are simplicial and matching (in the sense that they do not contain hanging nodes). For $1 \leq n \leq N$, $\mathcal{T}^{n-1,n}$ is a common refinement of $\mathcal{T}^{n-1}$ and $\mathcal{T}^n$. For all $0 \leq n \leq N$ and all $K \in \mathcal{T}^n$, h_K denotes the diameter of K. We denote by $c^n_{\mathsf{S},K}$ the smallest eigenvalue of S on K and by $c^n_{\boldsymbol{\beta},r,K}$ the essential minimum of $\frac{1}{2}\nabla\cdot\boldsymbol{\beta} + r$ on $K \times I^n$. We denote by $\mathcal{E}_K$ the set of the sides of $K \in \mathcal{T}^n$, and we fix $\mathbf{n}_{K,\sigma}$ as the unit normal vector to a side σ outward to K.

We denote by $(\cdot,\cdot)_S$ the $L^2(S)$ inner product, by $\|\cdot\|_S$ the associated norm (when $S = \Omega$, the index is dropped), and by $|S|$ the Lebesgue measure of S. Next, we set $\mathbf{H}(\mathrm{div}, S) = \{\mathbf{v} \in \mathbf{L}^2(S); \nabla\cdot\mathbf{v} \in L^2(S)\}$. Moreover, we use the "broken Sobolev space" $H^1(\mathcal{T}^n) := \{\varphi \in L^2(\Omega); \varphi|_K \in H^1(K)\ \forall K \in \mathcal{T}^n\}$. Finally, we use the Raviart–Thomas–Nédélec space $\mathbf{RTN}^0(\mathcal{T}^n) := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div},\Omega); \mathbf{v}_h|_K \in$

$\mathbf{RTN}^0(K)\ \forall K \in \mathscr{T}^n\}$ where $\mathbf{RTN}^0(K) := [\mathbb{P}_0(K)]^d + \mathbf{x}\mathbb{P}_0(K)$. For W, a vector space of functions defined on Ω, we define $\mathscr{P}^1_\tau(W)$ (respectively $\mathscr{P}^0_\tau(W)$) as the vector space of functions v defined on Q_T such that $v(\cdot,t)$ takes values in W and is continuous and piecewise affine (respectively constant) in time.

Because of the nonconformity of the cell-centered finite volume method, we introduce, for all $0 \le n \le N$, the broken gradient operator ∇^n such that for a function $v \in H^1(\mathscr{T}^n)$, $\nabla^n v \in [L^2(\Omega)]^d$ is defined as $(\nabla^n v)|_K := \nabla(v|_K)$ for all $K \in \mathscr{T}^n$. The broken gradient operator $\nabla^{n-1,n}$ on the mesh $\mathscr{T}^{n-1,n}$ is defined similarly.

2.2 Continuous Problem

Let $X := L^2(0,T; H^1_0(\Omega))$, $X' = L^2(0,T; H^{-1}(\Omega))$, and $Y := \{v \in X; \partial_t v \in X'\}$. The weak solution u of the problem (1a)–(1c) is such that $u \in Y$ with $u(\cdot,0) = u_0$. For a.e. $t \in (0,T)$ and for all $\varphi \in H^1_0(\Omega)$, there holds

$$\langle \partial_t u, \varphi\rangle(t) + (\mathsf{S}\nabla u, \nabla\varphi)(t) + (\nabla\cdot(\boldsymbol{\beta} u), \varphi)(t) + (ru, \varphi)(t) = (f, \varphi)(t), \qquad (2)$$

where $\langle\cdot,\cdot\rangle$ stands for the duality pairing between $H^{-1}(\Omega)$ and $H^1_0(\Omega)$.

For $y \in X$, we introduce the space-time energy norm $\|y\|^2_X := \int_0^T |||y|||^2(t)\mathrm{d}t$, where $|||y|||^2 := \|\mathsf{S}^{\frac{1}{2}}\nabla y\|^2 + \|(\frac{1}{2}\nabla\cdot\boldsymbol{\beta} + r)^{\frac{1}{2}} y\|^2$. We extend the energy norm to discrete functions using the broken gradient.

3 The Cell-centered Finite Volume Schemes and Postprocessing

A general cell-centered finite volume scheme for the problem (1a)–(1c) can be written in the following form: for all $1 \le n \le N$, find $\overline{u}^n_h := (u^n_K)_{K\in\mathscr{T}^n}$, such that

$$\frac{1}{\tau^n}(\overline{u}^n_h - u^{n-1}_h, 1)_K + \sum_{\sigma\in\mathscr{E}_K} S^n_{K,\sigma} + \sum_{\sigma\in\mathscr{E}_K} W^n_{K,\sigma} + r^n_K(\overline{u}^n_h, 1)_K = f^n_K|K| \quad \forall K \in \mathscr{T}^n, \quad (3)$$

where $f^n_K = \frac{1}{\tau^n}\int_{I^n}(f(\cdot,t),1)_K/|K|\mathrm{d}t$, $r^n_K = \frac{1}{\tau^n}\int_{I^n}(r(\cdot,t),1)_K/|K|\mathrm{d}t$, $S^n_{K,\sigma}$ and $W^n_{K,\sigma}$ are, respectively, the diffusive and convective fluxes through a side σ of an element K, and u^{n-1}_h is the postprocessed solution that we define below.

For $1 \le n \le N$, we reconstruct a conforming convective flux $\boldsymbol{\psi}^n$ and a conforming diffusive flux $\boldsymbol{\theta}^n$ such that $\boldsymbol{\psi}^n, \boldsymbol{\theta}^n \in \mathbf{RTN}^0(\mathscr{T}^n)$ and verifying

$$\langle \boldsymbol{\psi}^n \cdot \mathbf{n}_{K,\sigma}, 1\rangle_\sigma = W^n_{K,\sigma} \quad \forall K \in \mathscr{T}^n,\ \forall \sigma \in \mathscr{E}_K, \tag{4}$$

$$\langle \boldsymbol{\theta}^n \cdot \mathbf{n}_{K,\sigma}, 1\rangle_\sigma = S^n_{K,\sigma} \quad \forall K \in \mathscr{T}^n,\ \forall \sigma \in \mathscr{E}_K. \tag{5}$$

We refer to [4, 8] for more details on such construction. We define $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ in $\mathscr{P}^0_\tau(\mathbf{H}(\mathrm{div}, \Omega))$ by $\boldsymbol{\theta}|_{I^n} := \boldsymbol{\theta}^n$ and $\boldsymbol{\psi}|_{I^n} := \boldsymbol{\psi}^n$.

Following [8], we introduce a piecewise quadratic approximation u^n_h for all $1 \leq n \leq N$ verifying for all $K \in \mathscr{T}^n$,

$$-\mathsf{S}\nabla u^n_h|_K = \boldsymbol{\theta}^n|_K, \tag{6}$$

$$(u^n_h, 1)_K = |K| u^n_K. \tag{7}$$

When $S = \nu Id$, u^n_h lies in the space $\mathbb{P}_{1,2}(\mathscr{T}^n)$ which is $\mathbb{P}_1(\mathscr{T}^n)$ enriched elementwise with $\sum_{i=1}^d x_i^2$. Finally, we set u^0_h the L^2-projection of u_0 onto $\mathbb{P}_{1,2}(\mathscr{T}^n)$.

Because of the nonconformity of u^n_h, i.e., of the fact that $u^n_h \in H^1(\mathscr{T}^n)$, $u^n_h \notin H^1_0(\Omega)$, we define an averaging interpolate $s^n = I_{\mathrm{av}}(u^n_h) \in H^1_0(\Omega)$ of u^n_h that verifies

$$(s^n, 1)_K = (u^n_h, 1)_K \quad \forall K \in \mathscr{T}^{n,n+1}, \quad \forall 0 \leq n \leq N, \tag{8}$$

with the convention $\mathscr{T}^{N,N+1} := \mathscr{T}^N$. We refer to [3] for the details on such construction. Finally, we consider $u_{h,\tau} \in P^1_\tau(H^1(\mathscr{T}^n))$ and $s \in P^1_\tau(H^1_0(\Omega))$. They are defined by the values u^n_h and s^n for all $0 \leq n \leq N$. We set $\partial^n_t v = \partial_t v|_{I^n}$. An important consequence of this construction is the following, cf. [3],

$$(\partial^n_t s, 1)_K = (\partial^n_t u_{h,\tau}, 1)_K \quad \forall K \in \mathscr{T}^n. \tag{9}$$

4 A Posteriori Error Estimate

Our a posteriori estimate bounds the energy error between the weak solution u and the approximate solution $u_{h,\tau}$. We use the postprocessed solution instead of the original piecewise constant solution since the latter has a zero broken gradient and therefore is not suitable for energy norm estimates.

Let $1 \leq n \leq N$ and $K \in \mathscr{T}^n$. We define the *residual estimator* as

$$\eta^n_{\mathrm{R},K} := m^n_K \| \widetilde{f}^n - \partial^n_t s - \nabla \cdot \boldsymbol{\theta}^n - \nabla \cdot \boldsymbol{\psi}^n - r^n_K s^n \|_K. \tag{10}$$

Here $\widetilde{f}^n = \frac{1}{\tau^n} \int_{I^n} f(\cdot, t)\mathrm{d}t$ and $m^n_K := \min\{C_{\mathrm{P},K} h_K (c^n_{\mathsf{S},K})^{-\frac{1}{2}}, (c^n_{\boldsymbol{\beta},r,K})^{-\frac{1}{2}}\}$ is the constant from the inequality

$$\|\varphi - \varphi_K\|_K \leq m^n_K |||\varphi|||_K \quad \forall K \in \mathscr{T}^n, \quad \forall \varphi \in H^1(K), \tag{11}$$

shown in [8]. Here, $\varphi_K := (\varphi, 1)_K / |K|$ and $C_{\mathrm{P},K} := 1/\pi$ is the constant from the Poincaré inequality (recall that K are convex). We define the *flux estimator* as

$$\eta^n_{\mathrm{F},K}(t) := \|\mathsf{S}^{\frac{1}{2}}\nabla s + \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\theta}^n - \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\beta} s + \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\psi}^n\|_K. \tag{12}$$

Furthermore, we define the following *nonconformity estimator*

$$\eta^n_{\mathrm{NC},K}(t) := |||u_{h,\tau} - s|||_K. \tag{13}$$

Let $\overline{m}^n := \min\{C_{\mathrm{F},\Omega} h_\Omega (c^n_{\mathsf{S},\Omega})^{-\frac{1}{2}}, (c^n_{\boldsymbol{\beta},r,\Omega})^{-\frac{1}{2}}\}$, where $C_{\mathrm{F},\Omega}$ is the Friedrichs inequality constant detailed in [5]. The *quadrature estimator* is given by

$$\eta^n_{\mathrm{Q},K}(t) := \overline{m}^n \|f - \widetilde{f}^n - rs + r^n_K s^n\|_K. \tag{14}$$

Finally, we define the *initial condition estimator* as

$$\eta_{\mathrm{IC}} := 2^{-\frac{1}{2}}\|s^0 - u^0\|. \tag{15}$$

We now state and prove our main result concerning the error upper bound.

Theorem 1 (Energy norm a posteriori estimate). *Let $\eta^n_{\mathrm{R},K}$, $\eta^n_{\mathrm{F},K}$, $\eta^n_{\mathrm{NC},K}$, $\eta^n_{\mathrm{Q},K}$, and η_{IC} be defined by (10) and (12)–(15). Then,*

$$\|u - u_{h,\tau}\|_X \le \eta := \left\{\sum_{n=1}^{N}\int_{I^n}\sum_{K\in\mathscr{T}^n}\left(\eta^n_{\mathrm{R},K} + \eta^n_{\mathrm{F},K}(t)\right)^2 \mathrm{d}t\right\}^{\frac{1}{2}} + \eta_{\mathrm{IC}}$$
$$+\left\{\sum_{n=1}^{N}\int_{I^n}\sum_{K\in\mathscr{T}^n}(\eta^n_{\mathrm{Q},K}(t))^2\mathrm{d}t\right\}^{\frac{1}{2}} + \left\{\sum_{n=1}^{N}\int_{I^n}\sum_{K\in\mathscr{T}^n}(\eta^n_{\mathrm{NC},K}(t))^2\mathrm{d}t\right\}^{\frac{1}{2}}.$$

Proof. For $s \in Y$, we define $\mathscr{R}(s)$ in X' by $\langle \mathscr{R}(s), \varphi\rangle := \int_0^T \{(f - \partial_t s - \nabla\cdot(\boldsymbol{\beta} s) - rs, \varphi) - (\mathsf{S}\nabla s, \nabla\varphi)\}(t)\mathrm{d}t$, for all $\varphi \in X$. We obtain

$$\frac{1}{2}\|u - s\|^2(T) = \frac{1}{2}\|u^0 - s^0\|^2 + \int_0^T \langle \partial_t(u - s), u - s\rangle(t)\mathrm{d}t,$$

which yields

$$\|u - s\|^2_X \le \frac{1}{2}\|u^0 - s^0\|^2 + \langle \mathscr{R}(s), u - s\rangle.$$

Using the definition of the dual norm yields $\|u-s\|^2_X \le \|\mathscr{R}(s)\|_{X'}\|u-s\|_X + \frac{1}{2}\|u^0 - s^0\|^2$. Since $x^2 \le ax + b^2$ implies $x \le a + b$, $(a, b > 0)$, we infer

$$\|u - s\|_X \le \|\mathscr{R}(s)\|_{X'} + 2^{-\frac{1}{2}}\|u^0 - s^0\|. \tag{16}$$

For $1 \le n \le N$, set $\langle \mathscr{R}^n(s), \varphi \rangle := T_{\mathrm{R}}^n(\varphi) + T_{\mathrm{F}}^n(\varphi) + T_{\mathrm{Q}}^n(\varphi)$ with

$$\begin{aligned} T_{\mathrm{R}}^n(\varphi) &:= \sum_{K \in \mathscr{T}^n} (\widetilde{f}^n - \partial_t^n s - \nabla\cdot\boldsymbol{\theta}^n - \nabla\cdot\boldsymbol{\psi}^n - r_K^n s^n, \varphi)_K, \\ T_{\mathrm{F}}^n(\varphi) &:= -(\mathsf{S}\nabla s + \boldsymbol{\theta}^n + \boldsymbol{\psi}^n - \boldsymbol{\beta} s, \nabla\varphi), \\ T_{\mathrm{Q}}^n(\varphi) &:= \sum_{K \in \mathscr{T}^n} (f - \widetilde{f}^n - rs + r_K^n s^n, \varphi)_K. \end{aligned}$$

First, we have $T_{\mathrm{R}}^n(\varphi) = T_{\mathrm{R}}^n(\varphi - \Pi_0\varphi)$, where $\Pi_0\varphi|_K := \varphi_K$ for all K, using $(\widetilde{f}^n - \partial_t^n s - \nabla\cdot\boldsymbol{\theta}^n - \nabla\cdot\boldsymbol{\psi}^n - r_K^n s^n, 1)_K = 0$ from (3), (4), (5), and (7)–(9). Hence, $T_{\mathrm{R}}^n(\varphi) \le \sum_{K \in \mathscr{T}^n} \eta_{\mathrm{R},K}^n |||\varphi|||_K$ using the Cauchy–Schwarz inequality and (11). Moreover, $T_{\mathrm{F}}^n(\varphi)$ is bounded by $\sum_{K \in \mathscr{T}^n} \eta_{\mathrm{F},K}^n |||\varphi|||_K$ using the Cauchy–Schwarz inequality, and $T_{\mathrm{Q}}^n(\varphi)$ is bounded by $\left\{\sum_{K \in \mathscr{T}^n} (\eta_{\mathrm{Q},K}^n)^2\right\}^{1/2} |||\varphi|||$ as in [5]. Using (16), the definition of $\mathscr{R}(s)$, and the Cauchy–Schwarz and triangle inequalities concludes the proof.

In order to make the calculation efficient, it is important to distinguish the space and time errors. To this purpose, the flux estimator $\eta_{\mathrm{F},K}^n(t)$ is split into two contributions using the triangle inequality. We define, for all $1 \le n \le N$,

$$(\eta_{\mathrm{sp}}^n)^2 := 4 \sum_{K \in \mathscr{T}^n} \left\{ \tau^n (\eta_{\mathrm{R},K}^n + \eta_{\mathrm{F},1,K}^n)^2 + \int_{I^n} (\eta_{\mathrm{NC},K}^n)^2(t)\mathrm{d}t \right\},$$

$$(\eta_{\mathrm{tm}}^n)^2 := 4 \sum_{K \in \mathscr{T}^n} \left\{ \int_{I^n} \|\mathsf{S}^{\frac{1}{2}}\nabla(s - s^n) - \mathsf{S}^{-\frac{1}{2}}(\boldsymbol{\beta} s - \boldsymbol{\beta}^n s^n)\|_K^2(t)\mathrm{d}t + \int_{I^n} \left(\eta_{\mathrm{Q},K}^n(t)\right)^2 \mathrm{d}t \right\},$$

where $\boldsymbol{\beta}^n := \frac{1}{\tau^n}\int_{I^n} \boldsymbol{\beta}(\cdot,t)\mathrm{d}t$ and $\eta_{\mathrm{F},1,K}^n := \|\mathsf{S}^{\frac{1}{2}}\nabla s^n + \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\theta}^n - \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\beta}^n s^n + \mathsf{S}^{-\frac{1}{2}}\boldsymbol{\psi}^n\|_K$.

Proceeding as in [3], we obtain

Theorem 2 (A posteriori estimate distinguishing the space and time errors). *There holds*

$$\|u - u_{h,\tau}\|_X \le \left\{ \sum_{n=1}^{N} \left\{ (\eta_{\mathrm{sp}}^n)^2 + (\eta_{\mathrm{tm}}^n)^2 \right\} \right\}^{1/2} + \eta_{\mathrm{IC}}.$$

5 A Space-time Adaptive Time-marching Algorithm

We present here an adaptive algorithm based on our a posteriori error estimates which ensures that the relative energy error between the exact and the approximate solutions is below a prescribed tolerance ε. At the same time, it intends to equilibrate the space and time estimators η_{sp}^n and η_{tm}^n. Recalling Theorem 2 and neglecting η_{IC}

we aim at achieving

$$\frac{\sum_{n=1}^{N}\{(\eta_{\mathrm{sp}}^{n})^2+(\eta_{\mathrm{tm}}^{n})^2\}}{\sum_{n=1}^{N}\|u_{h,\tau}\|^2_{X(t^{n-1},t^{n})}} \leq \varepsilon^2. \tag{17}$$

On a given time level t^{n-1}, we set **Crit** $:= \varepsilon\frac{\|u_{h,\tau}\|_{X(t^{n-1},t^{n})}}{\sqrt{2}}$ and we choose the space mesh $\mathscr{T}^n$ and the time step τ^n such that $\eta_{\mathrm{sp}}^{n} \leq$ **Crit** and $\eta_{\mathrm{tm}}^{n} \leq$ **Crit**. For practical implementation purposes and because of computer limitations, we introduce maximal refinement level parameters N_{sp} and N_{tm}. The actual algorithm is as follows:

```
Choose an initial mesh $\mathscr{T}^0$, an initial time step $\tau^0$, and set $t^0 = 0$
Set $n = 1$ and $t^1 = t^0 + \tau^0$
Loop in time: While $t^n \leq T$
    Set $\mathscr{T}^{n\star} := \mathscr{T}^{n-1}$
    Do
        Solve $u_h^{n\star} = \mathrm{Sol}(u_h^{n-1}, \tau^{n-1}, \mathscr{T}^{n\star})$
        Estimate $\eta_{\mathrm{sp}}^n$ and $\eta_{\mathrm{tm}}^n$
        Refine the elements $K \in \mathscr{T}^{n\star}$ where $\eta_{\mathrm{sp},K}^n \geq \mathbf{Ref}\ \eta_{\mathrm{sp}}^n$ and such
             that their level of refinement is less than $N_{\mathrm{sp}}$
    While $\{\eta_{\mathrm{sp}}^n \geq \mathbf{Crit}$ or $\eta_{\mathrm{sp}}^n$ is much larger than $\eta_{\mathrm{tm}}^n\}$
    If $\{\eta_{\mathrm{tm}}^n \geq \mathbf{Crit}$ or $\eta_{\mathrm{tm}}^n$ is much larger than $\eta_{\mathrm{sp}}^n$ and when
        the level of time refinement is less than $N_{\mathrm{tm}}\}$
        Set $t^n = t^n - \tau^{n-1}$ and $\tau^{n-1} = \tau^{n-1}/2$
    Else
        Save the approximate solution $u_h^n := u_h^{n\star}$, the mesh $\mathscr{T}^n := \mathscr{T}^{n\star}$,
            and the time step $\tau^n$, and set $n = n + 1$
```

In this version we are only refining the elements and time steps where the estimated error is large. In a later version, we will also coarsen elements and time steps where the estimated error is small.

6 Numerical Experiments

We consider (1a)–(1c) on $\Omega = (0,3) \times (0,3)$ with $\mathsf{S} = \nu Id$, $\boldsymbol{\beta} = (\beta_1, \beta_2)$, $r = 0$, and $f = 0$, where $\nu > 0$ determines the amount of diffusion. The initial condition u_0, as well as the Dirichlet boundary condition, are given by the exact solution

$$u(x,y,t) = \frac{1}{200\nu t + 1}\mathrm{e}^{-50\frac{(x-x_0-\beta_1 t)^2+(y-y_0-\beta_2 t)^2}{200\nu t+1}}.$$

Here $x_0 = 0.33$, $y_0 = 1.125$, $\beta_1 = 0.8$, and $\beta_2 = 0.4$. We set $T = 0.6$. We use the DDFV method detailed in [2]. We neglect the additional error from the inhomogeneous Dirichlet boundary condition. We consider two cases $\nu = 0.1$ and $\nu = 0.001$. We start from an initial time step $\tau = 0.05$ and an initial mesh of 336 triangles and we refine uniformly by dividing the time step by 2 and each triangle into 4 subelements. Tables 1 and 2 show the actual and estimated energy error where η is the upper bound from Theorem 1, as well as the contribution of each estimator to the upper bound. Specifically, we define the global-in-time and global-in-space version of the estimators, $(\eta_{\mathrm{R}})^2 := \sum_{n=1}^{N} \tau^n \sum_{K\in\mathscr{T}^n} (\eta_{\mathrm{R},K}^n)^2$, $(\eta_{\mathrm{NC}})^2 := \sum_{n=1}^{N} \int_{I^n} \sum_{K\in\mathscr{T}^n} (\eta_{\mathrm{NC},K}^n(t))^2 \mathrm{d}t$ and $(\eta_{\mathrm{F}})^2 := \sum_{n=1}^{N} \int_{I^n} \sum_{K\in\mathscr{T}^n} (\eta_{\mathrm{F},K}^n(t))^2 \mathrm{d}t$.

Table 1 Convergence results with uniform refinement in the case $\nu = 0.1$

$\Vert u - u_{h,\tau}\Vert_X$	η	η_{R}	η_{F}	η_{NC}	$\frac{\eta}{\Vert u - u_{h,\tau}\Vert_X}$
0.0625	0.2070	0.0420	0.0910	0.0600	3.3102
0.0366	0.1299	0.0242	0.0613	0.0327	3.5464
0.0199	0.0662	0.0065	0.0328	0.0179	3.3182
0.0104	0.0335	0.0017	0.0167	0.0095	3.2104

Table 2 Convergence results with uniform refinement in the case $\nu = 0.001$

$\Vert u - u_{h,\tau}\Vert_X$	η	η_{R}	η_{F}	η_{NC}	$\frac{\eta}{\Vert u - u_{h,\tau}\Vert_X}$
0.0342	1.6490	0.3894	1.0875	0.0101	48.2496
0.0286	1.2341	0.2175	0.8354	0.0091	43.2175
0.0221	0.7992	0.0701	0.5541	0.0083	36.1332
0.0158	0.4773	0.0226	0.3312	0.0076	30.2736

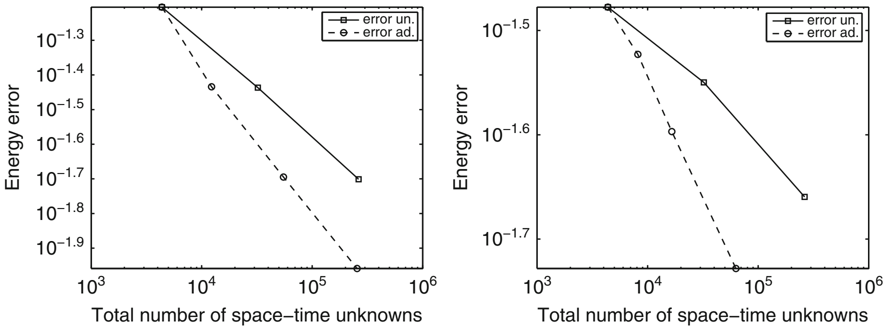


Fig. 1 Energy error in adaptive and uniform refinement for $\nu = 0.1$ (left) and $\nu = 0.001$ (right)

We next compare the uniform and adaptive refinement strategies. We note that the refinement maintains the conformity of the mesh. Figure 1 shows that we obtain a better precision in the adaptive strategy for much fewer space–time unknowns. Figure 2 depicts the approximate solution at the final time for $\nu = 0.001$ obtained

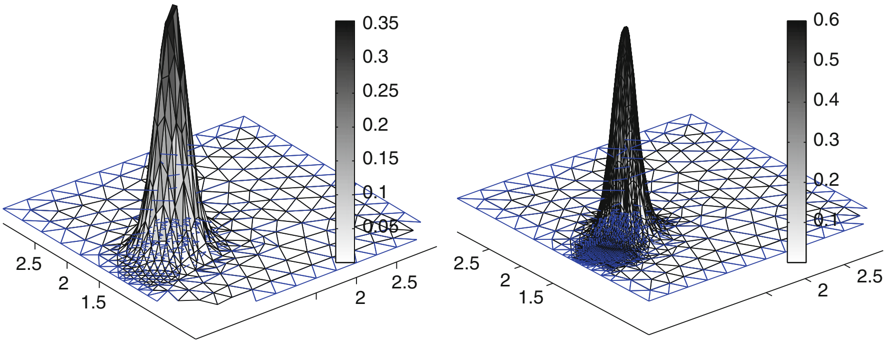


Fig. 2 Approximate solution with adaptive refinement: $N_{\mathrm{sp}} = N_{\mathrm{tm}} = 2$ (left), $N_{\mathrm{sp}} = N_{\mathrm{tm}} = 4$ (right)

with adaptive refinement for $N_{\mathrm{sp}} = N_{\mathrm{tm}} = 2$, and $N_{\mathrm{sp}} = N_{\mathrm{tm}} = 4$. We can see that in the second case the approximate solution better approximates the exact solution.

Acknowledgements Nancy Chalhoub was supported by a joint fellowship from Ecole des Ponts ParisTech and CNRS Lebanon.

References

1. Amaziane, B. and Bergam, A. and El Ossmani, M. and Mghazli, Z.: A posteriori estimators for vertex centred finite volume discretization of a convection-diffusion-reaction equation arising in flow in porous media. Internat. J. Numer. Methods Fluids **59**, 259–284, (2009)
2. Domelevo, K. and Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. M2AN Math. Model. Numer. Anal. **39**, 1203–1249 (2005)
3. Ern, A. and Vohralík, M.: A posteriori error estimation based on potential and flux reconstruction for the heat equation. SIAM J. Numer. Anal. **48**, 198–223 (2010)
4. Eymard, R. and Gallouët, T. and Herbin, R.: Finite volume approximation of elliptic problems and convergence of an approximate gradient. Appl. Numer. Math. **37**, 31–53 (2001)
5. Hilhorst, D. and Vohralík, M.: A posteriori error estimates for combined finite volume–finite element discretizations of reactive transport equations on nonmatching grids. Comput. Methods Appl. Mech. Engrg. **200**, 597–613 (2011)
6. Nicaise, S.: A posteriori error estimations of some cell centered finite volume methods for diffusion-convection-reaction problems. SIAM J. Numer. Anal. **44**, 949–978 (2006)
7. Ohlberger, M.: A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection–diffusion equations. Numer. Math. **87**, 737–761 (2001)
8. Vohralík, M.: Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods. Numer. Math. **11**, 121–158 (2008)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Large Time-Step Numerical Scheme for the Seven-Equation Model of Compressible Two-Phase Flows

Christophe Chalons, Frédéric Coquel, Samuel Kokh, and Nicole Spillane

Abstract We consider the seven-equation model for compressible two-phase flows and propose a large time-step numerical scheme based on a time implicit-explicit Lagrange-Projection strategy introduced in Coquel *et al.* [6] for Euler equations. The main objective is to get a Courant-Friedrichs-Lewy (CFL) condition driven by (slow) contact waves instead of (fast) acoustic waves.

Keywords Compressible two-phase flows, Baer-Nunziato model, seven-equation model, large time-step numerical scheme
MSC2010: 76T99, 74S10

1 Introduction

We are interested in the computation of compressible two-phase flows with the so-called *two-fluid two-pressure* or *seven-equation* model. It was first proposed in Baer & Nunziato [4] and has since aroused more and more interest, see for instance Embid & Baer [7], Stewart & Wendroff [13], Abgrall & Saurel [11], Gallouët, Hérard & Seguin [8], Andrianov & Warnecke [3], Karni *et al.* [9] Schwendeman, Wahle & Kapila [12], Munkejord [10], Tokareva & Toro [14], Ambroso, Chalons,

Christophe Chalons and Samuel Kokh
CEA-Saclay, 91191 Gif-sur-Yvette, France, e-mail: christophe.chalons@cea.fr, samuel.kokh@cea.fr

Frédéric Coquel
CNRS & CMAP, U.M.R. 7641, Ecole Polytechnique, Route de Saclay, 91128 Palaiseau Cedex, France, e-mail: frederic.coquel@cmap.polytechnique.fr

Nicole Spillane
Laboratoire J.-L. Lions, UPMC Univ Paris 06, BC 187, 75252 Paris cedex 05, France, e-mail: spillane@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_24, © Springer-Verlag Berlin Heidelberg 2011

Coquel & Galié [1], Ambroso, Chalons & Raviart [2], and the references therein. One of the main features of this model is that it is hyperbolic, at least in the context of subsonic flows. In particular, an interesting property is that the seven-equation model possesses seven *real* eigenvalues given by $\lambda_k^{\pm}(\mathbf{u}) = u_k \pm c_k$, $\lambda_k^0(\mathbf{u}) = u_k$ and $\lambda_I(\mathbf{u}) = u_I$, where u_k denote the velocities of both phases $k = 1, 2$, c_k the sound speeds, u_I an interfacial velocity and $\mathbf{u}$ the vector of unknowns.

However from a numerical point of view, the seven-equation model raises some issues. The first difficulty is related to the large size of the model and as a consequence to the Riemann problem that is difficult to solve, even approximately. The second difficulty comes from the presence of nonconservative products and more precisely the fact that the model cannot be equivalently recast in full conservative form. However, the nonconservative products naturally vanish when the void fractions α_k are locally constant in space, and the model coincides in that case with two (decoupled) gas dynamics systems. This property will be used in the numerical strategy.

Numerous papers are devoted to the numerical study of two-fluid two-pressure models, see again for instance [8], [3], [9], [12], [10], [14], [1], [2] and the references therein. Many of the proposed methods are based on time-explicit, exact or approximate, Godunov-type methods (Roe or Roe-like scheme, HLL or HLLC scheme...). For stability reasons, the time steps Δt involved in such methods are subject to a usual Courant-Friedrichs-Lewy (CFL) condition that reads

$$\max_{k,\mathbf{u}}(|\lambda_k^{\pm}(\mathbf{u})|, |\lambda_k^0(\mathbf{u})|, |\lambda_I(\mathbf{u})|)\,\Delta t \leq 0.5\Delta x,$$

where Δx represents the space step. It is then clear that the definition of Δt is driven by the fastest eigenvalues $\lambda_k^{\pm}(\mathbf{u})$, associated with the so-called acoustic waves.

In many applications, like for instance in two-phase flows involved in nuclear reactors, the acoustic waves are not predominant physical phenomena. A CFL condition based on the most influent waves, the so-called contact waves associated with eigenvalues $\lambda_k^0(\mathbf{u})$ and $\lambda_I(\mathbf{u})$ would be more adapted. The idea is then to propose a time-implicit treatment of the (fast) acoustic waves, in order to get rid of a too restrictive CFL condition, together with an explicit treatment of the (slow) contact waves in order to preserve accuracy. This was recently proposed in Coquel *et al.* [6] in the context of Euler equations, using a Lagrange-Projection approach. This approach is well-adapted as it naturally decouples the fast and slow waves in the Lagrange and Projection steps respectively.

In this paper, we propose a first attempt to extend this approach to the seven-equation model. The idea is to operate a relevant operator splitting between the conservative and nonconservative parts of the original model, in order to make Euler systems for each phase appear. The latter parts are treated as in [6]. The nonconservative products are then discretized so as to maintain conservativity properties of the model on each partial mass, on the total momentum and total energy. Numerical results are proposed. We underline that this work is still in progress.

2 Governing equations

The model under consideration in this contribution reads as follows:

$$\begin{cases} \partial_t \alpha_k + u_I \partial_x \alpha_k = 0, \qquad t > 0, \quad x \in \mathbb{R}, \\ \partial_t \alpha_k \rho_k + \partial_x \alpha_k \rho_k u_k = 0, \\ \partial_t \alpha_k \rho_k u_k + \partial_x \alpha_k (\rho_k u_k^2 + p_k) - p_I \partial_x \alpha_k = 0, \\ \partial_t \alpha_k \rho_k e_k + \partial_x \alpha_k (\rho_k e_k + p_k) u_k - p_I u_I \partial_x \alpha_k = 0, \end{cases} \tag{1}$$

with $k = 1, 2$. Here, α_k, ρ_k, u_k, e_k and p_k denote the volume fraction, density, velocity, specific total energy and pressure of the phase $k = 1, 2$. The two phases are assumed to be non-miscible that is $\alpha_1 + \alpha_2 = 1$. The structure of (1) is the one of two gas dynamics systems for each phase, coupled with a transport equation on the void fraction α_k at speed u_I. We note that nonconservative products involving the interfacial pressure p_I and velocity u_I (to be precised later on) and the space derivative of the void fractions α_k are present in the momentum and energy equations. These terms act as coupling terms in the evolution of the two phases. Source terms like external forces, pressure and velocity relaxations, dissipation, heat conduction, phase changes and heat exchanges between the two phases are not taken into account.

Each phase is provided with an equation of state $p_k = p_k(\rho_k, \varepsilon_k)$, where $\varepsilon_k = e_k - u_k^2/2$ is the specific internal energy. So far as the definitions of u_I and p_I are concerned, we follow [8] and first observe that the characteristic speeds of (1) are always real and given by u_I, u_k, $u_k \pm c_k, k = 1, 2$, where c_k denotes the speed of sound in phase k. System (1) turns out to be only weakly hyperbolic since there are not enough eigenvectors to span the entire space when $u_I = u_k \pm c_k$ for some index k (resonance occurs). When (1) is hyperbolic, one can easily check that similarly to the classical gas dynamics equations, the characteristic fields associated with the eigenvalues $u_k \pm c_k$ are nonlinear while the one associated with u_k is linearly degenerate. Regarding the characteristic field associated with u_I, it is generally required to be linearly degenerate in practice. This property holds as soon as

$$u_I = \beta u_1 + (1 - \beta) u_2, \quad \beta = \frac{\chi \alpha_1 \rho_1}{\chi \alpha_1 \rho_1 + (1 - \chi) \alpha_2 \rho_2} \tag{2}$$

where $\chi \in [0, 1]$ is a constant (we refer to [8] for the details), which gives a natural definition for the interfacial velocity u_I. The usual choices for χ are $0, 1/2$ and 1. Regarding the interfacial pressure p_I, we set $p_I = \mu p_1 + (1 - \mu) p_2$, $\mu \in [0, 1]$. The choice of the coefficient μ is not detailed here (see again [8]) but is related to the ability to provide the system with an entropy balance equation. Indeed, it can be proved that for a specific choice of μ, smooth solutions of (1) verify the conservation law $\partial_t \eta + \partial_x q = 0$, where (η, q) plays the role of a mathematical entropy pair.

3 A natural operator splitting

The starting point is to propose an equivalent form of (1) where the space derivatives of $\alpha_k p_k$ and $\alpha_k p_k u_k$ are decomposed using a chain rule:

$$\begin{cases} \partial_t \alpha_k + u_I \partial_x \alpha_k = 0, \\ \partial_t \alpha_k \rho_k + \partial_x \alpha_k \rho_k u_k = 0, \\ \partial_t \alpha_k \rho_k u_k + \partial_x \alpha_k \rho_k u_k^2 + \alpha_k \partial_x p_k + (p_k - p_I)\partial_x \alpha_k = 0, \\ \partial_t \alpha_k \rho_k e_k + \partial_x \alpha_k \rho_k e_k u_k + \alpha_k \partial_x p_k u_k + (p_k u_k - p_I u_I)\partial_x \alpha_k = 0. \end{cases} \tag{3}$$

We then suggest to split (3) into two independent and *quasi-classical* gas dynamics equations (their Lagrangian forms will be seen to be *classical*), namely

$$\begin{cases} \partial_t \alpha_k = 0, \\ \partial_t \alpha_k \rho_k + \partial_x \alpha_k \rho_k u_k = 0, \\ \partial_t \alpha_k \rho_k u_k + \partial_x \alpha_k \rho_k u_k^2 + \alpha_k \partial_x p_k = 0, \\ \partial_t \alpha_k \rho_k e_k + \partial_x \alpha_k \rho_k e_k u_k + \alpha_k \partial_x p_k u_k = 0, \end{cases} \tag{4}$$

and into the following genuinely nonconservative system:

$$\begin{cases} \partial_t \alpha_k + u_I \partial_x \alpha_k = 0, \\ \partial_t \alpha_k \rho_k = 0, \\ \partial_t \alpha_k \rho_k u_k + (p_k - p_I)\partial_x \alpha_k = 0, \\ \partial_t \alpha_k \rho_k e_k + (p_k u_k - p_I u_I)\partial_x \alpha_k = 0. \end{cases} \tag{5}$$

This transformation aims at proposing in the next section an implicit-explicit Lagrange-Projection scheme similar to [6], and at treating separately the nonconservative products. Note from now on that the overall algorithm will be conservative on the partial mass $\alpha_k \rho_k$, total momentum $\alpha_1 \rho_1 u_1 + \alpha_2 \rho_2 u_2$ and on the total energy $\alpha_1 \rho_1 e_1 + \alpha_2 \rho_2 e_2$, as it is expected from the original form (1) of the model.

4 Numerical approximation

This section is devoted to the discretization of (1), using (4) and (5). Let us introduce a time step $\Delta t > 0$ and a space step $\Delta x > 0$ that we assume to be constant for simplicity. We set $\lambda = \Delta t / \Delta x$ and define the mesh interfaces $x_{j+1/2} = j\Delta x$ for $j \in \mathbb{Z}$, and the intermediate times $t^n = n\Delta t$ for $n \in \mathbb{N}$. In the sequel, $\mathbf{u}_j^n = (\alpha_1, \mathbf{u}_1, \mathbf{u}_2)_j^n$ where $(\mathbf{u}_k)_j^n = (\alpha_k \rho_k, \alpha_k \rho_k u_k, \alpha_k \rho_k e_k)_j^n$ denotes the approximate value of the unknowns at time t^n and on the cell $\mathscr{C}_j =]x_{j-1/2}, x_{j+1/2}[$.

Implicit-explicit discretization of (4). We first recall that (4) is made of two independent quasi-classical gas dynamics systems, whose eigenvalues are given by

$u_k \pm c_k$, u_k and 0. As already stated, our aim is to propose an implicit treatment of the fast waves $u_k \pm c_k$, and an explicit treatment of u_k. With this in mind, we follow [6] and adopt a Lagrange-Projection scheme, coupled with a pressure relaxation strategy that is well adapted to this purpose. A Lagrange-Projection splitting strategy applied to (4) amounts to introducing the Lagrangian system

$$\begin{cases} \partial_t \alpha_k = 0, \\ \partial_t \alpha_k \rho_k + \alpha_k \rho_k \partial_x u_k = 0, \\ \partial_t \alpha_k \rho_k u_k + \alpha_k \rho_k u_k \partial_x u_k + \alpha_k \partial_x p_k = 0, \\ \partial_t \alpha_k \rho_k e_k + \alpha_k \rho_k e_k \partial_x u_k + \alpha_k \partial_x p_k u_k = 0, \end{cases} \quad \text{or equivalently} \quad \begin{cases} \partial_t \alpha_k = 0, \\ \partial_t \tau_k - \tau_k \partial_x u_k = 0, \\ \partial_t u_k + \tau_k \partial_x p_k = 0, \\ \partial_t e_k + \tau_k \partial_x p_k u_k = 0, \end{cases} \tag{6}$$

with $\tau_k = 1/\rho_k$, and the transport (or projection) system

$$\begin{cases} \partial_t \alpha_k = 0, \\ \partial_t \alpha_k \rho_k + u_k \partial_x \alpha_k \rho_k = 0, \\ \partial_t \alpha_k \rho_k u_k + u_k \partial_x \alpha_k \rho_k u_k = 0, \\ \partial_t \alpha_k \rho_k e_k + u_k \partial_x \alpha_k \rho_k e_k = 0. \end{cases} \tag{7}$$

We note that (6) coincides with two classical gas dynamics systems written in Lagrangian coordinates, the eigenvalues of which are given by $\pm c_k$ and 0. This system is treated using a pressure relaxation approach that consists in introducing a linearized pressure π_k (see for instance [5] and especially the references therein), such that $(\pi_k)_j^n = (p_k)_j^n$, and in solving the partial differential system

$$\begin{cases} \partial_t \alpha_k = 0, \\ \partial_t \tau_k - \tau_k \partial_x u_k = 0, \\ \partial_t u_k + \tau_k \partial_x \pi_k = 0, \\ \partial_t \pi_k + a_k^2 \tau_k \partial_x u_k = 0, \\ \partial_t e_k + \tau_k \partial_x \pi_k u_k = 0, \end{cases} \quad \text{or equivalently} \quad \begin{cases} \partial_t \alpha_k = 0, \\ \partial_t I_k = 0, \\ \partial_t w_k^+ + a_k \tau_k \partial_x w_k^+ = 0, \\ \partial_t w_k^- - a_k \tau_k \partial_x w_k^- = 0, \\ \partial_t e_k + \tau_k \partial_x \pi_k u_k = 0, \end{cases} \tag{8}$$

where $w_k^\pm = \pi_k \pm a_k u_k$, $I_k = \pi_k + a_k^2 \tau_k$, and a_k is a constant satisfying the subcharacteristic condition $a_k > \rho_k c_k$. A natural time-implicit discretization of (8) is

$$\begin{cases} (\alpha_k)_j^{n+1=} = (\alpha_k)_j^n, \\ (I_k)_j^{n+1=} = (I_k)_j^n, \\ (w_k^+)_j^{n+1=} = (w_k^+)_j^n - \lambda (\tau_k)_j^n a_k \big((w_k^+)_j^{n+1=} - (w_k^+)_{j-1}^{n+1=} \big), \\ (w_k^-)_j^{n+1=} = (w_k^-)_j^n + \lambda (\tau_k)_j^n a_k \big((w_k^+)_{j+1}^{n+1=} - (w_k^+)_j^{n+1=} \big), \\ (e_k)_j^{n+1=} = (e_k)_j^n - \lambda (\tau_k)_j^n \big((\pi_k u_k)_{j+1/2}^{n+1=} - (\pi_k u_k)_{j-1/2}^{n+1=} \big), \end{cases} \tag{9}$$

with $(\pi_k u_k)_{j+1/2}^{n+1=} = (\pi_k)_{j+1/2}^{n+1=} (u_k)_{j+1/2}^{n+1=}$ and

$$(\pi_k)_{j+1/2}^{n+1=} = \frac{1}{2} \big((w_k^+)_j^{n+1=} + (w_k^-)_j^{n+1=} \big), \; (u_k)_{j+1/2}^{n+1=} = \frac{1}{2a_k} \big((w_k^+)_j^{n+1=} - (w_k^-)_j^{n+1=} \big).$$

The updated values of u_k, τ_k and ρ_k are recovered from the formulas $u_k = (w_k^+ - w_k^-)/2a_k$, $\pi_k = (w_k^+ + w_k^-)/2$, $\tau_k = (I_k - \pi_k)/a_k^2$ and $\rho_k = 1/\tau_k$. The computation of $(w_k^\pm)_j^{n+1=}$ is cheap and amounts to solving a tridiagonal system of linear equations, while the time-implicit definition of $(e_k)_j^{n+1=}$ explicitly follows. Then, the transport equations involved in (7) are associated with the following classical time-explicit update formula

$$\begin{aligned}(\mathbf{u}_k)_j^{n+1-}=(\mathbf{u}_k)_j^{n+1=}+\lambda\ \Big(&\\ &\max((u_k)_{j-1/2}^{n+1=},0)(\mathbf{u}_k)_{j-1}^{n+1=} - \min((u_k)_{j+1/2}^{n+1=},0)(\mathbf{u}_k)_{j+1}^{n+1=}\\ &+\\ &[\min((u_k)_{j+1/2}^{n+1=},0) - \max((u_k)_{j-1/2}^{n+1=},0)](\mathbf{u}_k)_j^{n+1=}\\ &\Big),\end{aligned} \tag{10}$$

and of course $(\alpha_k)_j^{n+1-} = (\alpha_k)_j^{n+1=}$.

Discretization of (5). Our objective is to propose a consistent approximation of (5) such that the overall algorithm is conservative for each partial mass, for the total momentum and for the total energy, as already motivated. First of all and similarly to (10), the transport equation associated with α_k is treated as follows:

$$\begin{aligned}(\alpha_k)_j^{n+1}=(\alpha_k)_j^{n+1=}+\lambda\ \Big(&\\ &\max((u_I)_{j-1/2}^{n+1=},0)(\alpha_k)_{j-1}^{n+1=} - \min((u_I)_{j+1/2}^{n+1=},0)(\alpha_k)_{j+1}^{n+1=}\\ &+\\ &[\min((u_I)_{j+1/2}^{n+1=},0) - \max((u_I)_{j-1/2}^{n+1=},0)](\alpha_k)_j^{n+1=}\\ &\Big)\end{aligned}$$

where $(u_I)_{j+1/2}^{n+1=} = \beta_{j+1/2}^{n+1=}(u_1)_{j+1/2}^{n+1=} + (1 - \beta_{j+1/2}^{n+1=})(u_2)_{j+1/2}^{n+1=}$ and for instance $\beta_{j+1/2}^{n+1=} = \frac{1}{2}\big(\beta_j^{n+1=} + \beta_{j+1}^{n+1=}\big)$. We set $(\alpha_k\rho_k)_j^{n+1} = (\alpha_k\rho_k)_j^{n+1-}$ for the partial mass, so that only the treatments of the momentum and total energy of each phase are now left. We propose

$$\frac{(\alpha_k\rho_k u_k)_j^{n+1} - (\alpha_k\rho_k u_k)_j^{n+1-}}{\Delta t} + \big((\overline{p_k})_j - (\overline{p_I})_j\big)\frac{(\alpha_k)_{j+1/2}^n - (\alpha_k)_{j-1/2}^n}{\Delta x} = 0,$$

$$\frac{(\alpha_k\rho_k e_k)_j^{n+1} - (\alpha_k\rho_k e_k)_j^{n+1-}}{\Delta t} + \big((\overline{p_k u_k})_j - (\overline{p_I u_I})_j\big)\frac{(\alpha_k)_{j+1/2}^n - (\alpha_k)_{j-1/2}^n}{\Delta x} = 0.$$

In order to get the expected overall conservativity properties, we pay a particular attention to the definitions of $(\overline{p_k})_j$, $(\overline{p_I})_j$, $(\overline{p_k u_k})_j$ and $(\overline{p_I u_I})_j$. For any consistent definition of the flux $(\alpha_k)_{j+1/2}^n$, we set with $\kappa_j \in [0, 1]$

$$\begin{cases} (\alpha_k)^n_j = \kappa_j(\alpha_k)^n_{j+1/2} + (1-\kappa_j)(\alpha_k)^n_{j-1/2}, \\ (\overline{p_k})_j = (1-\kappa_j)(\pi_k)^{n+1=}_{j+1/2} + \kappa_j(\pi_k)^{n+1=}_{j-1/2}, \\ (\overline{p_k u_k})_j = (1-\kappa_j)(\pi_k u_k)^{n+1=}_{j+1/2} + \kappa_j(\pi_k u_k)^{n+1=}_{j-1/2}, \end{cases}$$

and

$$\begin{cases} (\overline{p_I})_j = \mu^{n+1=}_{j+1/2}(\overline{p_1})_j + (1-\mu^{n+1=}_{j+1/2})(\overline{p_2})_j, \text{ with } \mu^{n+1=}_{j+1/2} = \frac{1}{2}\big(\mu^{n+1=}_j + \mu^{n+1=}_{j+1}\big) \\ (\overline{u_I})_j = \beta^{n+1=}_{j+1/2}(\overline{u_1})_j + (1-\beta^{n+1=}_{j+1/2})(\overline{u_2})_j, \text{ with } (\overline{u_k})_j = (\overline{p_k u_k})_j/(\overline{p_k})_j, \\ (\overline{p_I u_I})_j = (\overline{p_I})_j(\overline{u_I})_j. \end{cases}$$

We choosed in practice $(\alpha_k)^n_{j+1/2} = (\alpha_k)^n_j$ or equivalently $\kappa_j = 1$.

With such definitions, it can be proved that under a suitable CFL condition based on the velocities u_k and u_I only, and not on the acoustic waves $u_k \pm c_k$, the void fractions $(\alpha_k)^{n+1}_j$ belong to $(0, 1)$ if $(\alpha_k)^n_j$ do. We can also prove that under the same restriction on the time step $(\rho_k)^{n+1}_j$ is positive, as well as $(\varepsilon_k)^{n+1-}_j$ and $(p_k)^{n+1-}_j$. Unfortunately, the positivity of $(\varepsilon_k)^{n+1}_j$ and $(p_k)^{n+1}_j$ is not proved at the moment.

5 Numerical experiments

For the sake of illustration, we present in this section the results given by our algorithm on three Riemann problems, see the Fig. 1. They are all taken from [2] and are fully described therein. Space and time orders of accuracy are one. The first one (top left) corresponds to an isolated contact discontinuity propagating with a positive velocity, while the second one (top right) and the third one (bottom) involve several distinct waves. The scheme we propose here is denoted LP implicit and is compared with its explicit version (which amounts to replacing (9) by its time-explicit version) and the well-known Rusanov scheme (see [8]). We observe that our approach is clearly less diffusive around the contact discontinuities since the CFL condition is well-adapted to the corresponding speed of propagation, but more diffusive around the acoustic waves since it is implicit. Table 1 gives for each test case the number of iterations needed to perform the computations. As expected, the gain is important when using the proposed implicit-explicit algorithm and the corresponding CFL restriction based on the material waves (instead of the acoustic waves as for the explicit scheme). A careful evaluation of the CPU cost necessitates an additional programming effort that has not been implemented yet.

Table 1 Number of time-iterations for each test case

	Test 1	Test 2	Test 3
Rusanov	4231	550	2630
LP explicit	4297	551	2631
LP implicit	63	41	151

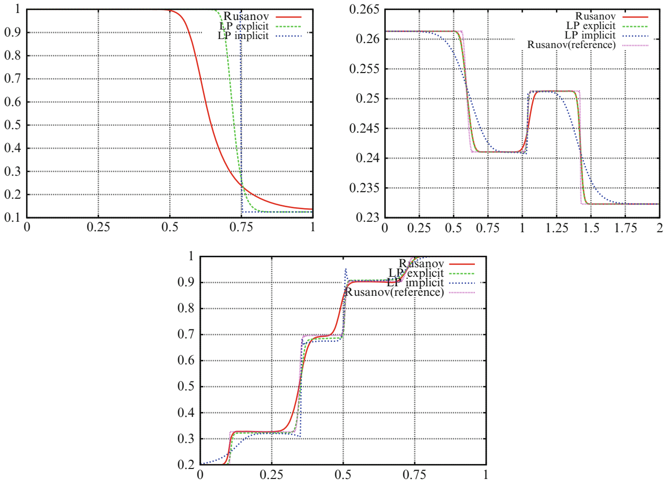


Fig. 1 Comparison of several schemes with a reference solution (density profile)

Acknowledgements Part of this work has been achieved within the framework of the NEPTUNE project, supported by CEA, EDF, IRSN, AREVA-NP.

References

1. A. Ambroso, C. Chalons, F. Coquel, T. Galié. Relaxation and numerical approximation of a two fluid two pressure diphasic model. *M2AN*, vol. 43, pp. 1063-1097, (2009).
2. A. Ambroso, C. Chalons and P.-A. Raviart. A Godunov-type method for the seven-equation model of compressible two-phase flow. *LJLL report number R10020*, http://www.ljll.math.upmc.fr/publications/2010/R10020.php, (2010).
3. N. Andrianov and G. Warnecke. The Riemann problem for the Baer-Nunziato two-phase flow model. *Journal of Computational Physics*, vol. 195, pp. 434-464, (2004).
4. M.R. Baer and J.W. Nunziato, A two phase mixture theory for the deflagration to detonation transition in reactive granular materials. *Int. J. Mult. Flows*, vol. 12(6), pp. 861-889, (1986).
5. C. Chalons and J.-F. Coulombel, Relaxation approximation of the Euler equations. *J. Math. Anal. Appl.*, vol. 348(2), pp. 872-893, (2008).
6. F. Coquel, Q.-L. Nguyen, M. Postel and Q.-H. Tran, Entropy-satisfying relaxation method with large time-steps for Euler IBVPs. *Math. Comp*, vol. 79, pp. 1493-1533, (2010).
7. P. Embid and M. Baer, Mathematical analysis of a two-phase continuum mixture theory, *Contin. Mech. Thermodyn.* vol. 4(4), pp. 279-312, (1992).
8. T. Gallouët, J.-M. Hérard and N. Seguin. Numerical modeling of two-phase flows using the two-fluid two-pressure approach. *M3AS*, vol. 14(5), pp. 663-700, (2004).

9. S. Karni, E. Kirr, A. Kurganov and G. Petrova, Compressible two-phase flows by central and upwind schemes, *M2AN*, vol. 38(3), pp. 477-493, (2004).
10. S.T. Munkejord, Comparison of Roe-type methods for solving the two-fluid model with and without pressure relaxation, *Computers and Fluids*, vol. 36, pp. 1061-1080, (2007).
11. R. Saurel and R. Abgrall, A multiphase Godunov method for compressible multifluid and multiphase flows, *J. Comput. Phys.*, vol. 150, pp. 425-467, (1999).
12. D.-W. Schwendeman, C.-W. Wahle and A.-K. Kapila. The Riemann problem and a high-resolution Godunov method for a model of compressible two-phase flow. *Journal of Computational Physics*, vol. 212, pp. 490-526, (2006).
13. B. Stewart and B. Wendroff, Two-phase flow : models and methods, *J. Comput. Phys.*, vol. 56, pp. 363-409, (1984).
14. S.-A. Tokareva and E.-F. Toro, HLLC-type Riemann solver for the Baer-Nunziato equations of compressible two-phase flow, *J. Comput. Phys.*, to appear, (2010).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Asymptotic Behavior of the Scharfetter–Gummel Scheme for the Drift-Diffusion Model

Marianne CHATARD

Abstract The aim of this work is to study the large-time behavior of the Scharfetter–Gummel scheme for the drift-diffusion model for semiconductors. We prove the convergence of the numerical solutions to an approximation of the thermal equilibrium. We also present numerical experiments which underline the preservation of long-time behavior.

Keywords Drift-diffusion system, finite volume scheme, thermal equilibrium.
MSC2010: 65M08, 76X05, 82D37.

1 Introduction

In the modeling of semiconductor devices, the drift-diffusion system is widely used as it simplifies computations while giving an accurate description of the device physics.
Let $\Omega \subset \mathbb{R}^d$ $(d \geq 1)$ be an open and bounded domain describing the geometry of the semiconductor device. The isothermal drift-diffusion system consists of two continuity equations for the electron density $N(x,t)$ and the hole density $P(x,t)$, and a Poisson equation for the electrostatic potential $V(x,t)$:

$$\begin{cases} \partial_t N - \operatorname{div}(\nabla N - N\nabla V) = 0 & \text{on } \Omega \times (0,T), \\ \partial_t P - \operatorname{div}(\nabla P + P\nabla V) = 0 & \text{on } \Omega \times (0,T), \\ \lambda^2 \Delta V = N - P - C & \text{on } \Omega \times (0,T), \end{cases} \tag{1}$$

Marianne CHATARD
Université Blaise Pascal - Laboratoire de Mathématiques UMR 6620 - CNRS - Campus des Cézeaux, B.P. 80026 63177 Aubière cedex, e-mail: Marianne.Chatard@math.univ-bpclermont.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_25, © Springer-Verlag Berlin Heidelberg 2011

where $C(x)$ is the doping profile, which is assumed to be a given datum, and λ is the Debye length arising from the scaling of the physical model. We supplement these equations with initial conditions $N_0(x)$ and $P_0(x)$ and physically motivated boundary conditions: Dirichlet boundary conditions $\overline{N}$, $\overline{P}$ and $\overline{V}$ on ohmic contacts Γ^D and homogeneous Neumann boundary conditions on insulating boundary segments Γ^N.

There is an extensive literature on numerical schemes for the drift-diffusion equations: finite difference methods, finite elements methods, mixed exponential fitting finite elements methods, finite volume methods (see [1]). The Scharfetter–Gummel scheme is widely used to approximate the drift-diffusion equations in the linear case. It has been proposed and studied in [7] and [10]. It preserves steady-state, and is second order accurate in space (see [9]).

The purpose of this paper is to study the large time behavior of the numerical solution given by the Scharfetter–Gummel scheme for the transient linear drift-diffusion model (1). Indeed, it has been proved by H. Gajewski and K. Gärtner in [5] that the solution to the transient system (1) converges to the thermal equilibrium state as $t \to \infty$ if the boundary conditions are in thermal equilibrium. A. Jüngel extends this result to a degenerate model with nonlinear diffusivities in [8].

The thermal equilibrium is a particular steady-state for which electron and hole currents, namely $\nabla N - N\nabla V$ and $\nabla P + P\nabla V$, vanish.

If the Dirichlet boundary conditions satisfy $\overline{N}, \overline{P} > 0$ and

$$\log(\overline{N}) - \overline{V} = \alpha_N \text{ and } \log(\overline{P}) + \overline{V} = \alpha_P \text{ on } \Gamma^D, \tag{2}$$

the thermal equilibrium is defined by

$$\begin{cases} \Delta V^{eq} = \exp(\alpha_N + V^{eq}) - \exp(\alpha_P - V^{eq}) - C & \text{on } \Omega, \\ N^{eq} = \exp(\alpha_N + V^{eq}), \quad P^{eq} = \exp(\alpha_P - V^{eq}) & \text{on } \Omega, \end{cases} \tag{3}$$

with the same boundary conditions as (1).

Our aim is to prove that the solution of the Scharfetter–Gummel scheme converges to an approximation of the thermal equilibrium as $t \to +\infty$. Long-time behavior of solutions to discretized drift-diffusion systems have been studied in [5], [2] and [6], using estimates of the energy.

In the sequel, we will suppose that the following hypotheses are fulfilled:

(H1) $\overline{N}$, $\overline{P}$ are traces on $\Gamma^D \times (0,T)$ of functions, also denoted $\overline{N}$ and $\overline{P}$, such that $\overline{N},\ \overline{P} \in H^1(\Omega \times (0,T)) \cap L^\infty(\Omega \times (0,T))$ and $\overline{N},\ \overline{P} \geq 0$ *a.e.*,

(H2) $N_0,\ P_0 \in L^\infty(\Omega)$ and $N_0,\ P_0 \geq 0$ *a.e.*,

(H3) there exist $0 < m \leq M$ such that: $m \leq \overline{N}, N_0, \overline{P}, P_0 \leq M$,

(H4) $\overline{N}$, $\overline{P}$ and $\overline{V}$ satisfy the compatibility condition (2).

2 Numerical schemes

In this section, we present the finite volume schemes for the time evolution drift-diffusion system (1) and for the thermal equilibrium (3).

An admissible mesh of Ω is given by a family $\mathcal{T}$ of control volumes (open and convex polygons in 2-D, polyhedra in 3-D), a family $\mathcal{E}$ of edges in 2-D (faces in 3-D) and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 5.1 in [4]. It implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$.

In the set of edges $\mathcal{E}$, we distinguish the interior edges $\sigma \in \mathcal{E}_{int}$ and the boundary edges $\sigma \in \mathcal{E}_{ext}$. We split $\mathcal{E}_{ext}$ into $\mathcal{E}_{ext} = \mathcal{E}_{ext}^D \cup \mathcal{E}_{ext}^N$ where $\mathcal{E}_{ext}^D$ is the set of Dirichlet boundary edges and $\mathcal{E}_{ext}^N$ is the set of Neumann boundary edges. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, $\mathcal{E}_{int,K}$ the set of its interior edges, $\mathcal{E}_{ext,K}^D$ the set of edges of K included in Γ^D and $\mathcal{E}_{ext,K}^N$ the set of edges of K included in Γ^N.

The size of the mesh is defined by $\Delta x = \max\limits_{K\in\mathcal{T}}(\mathrm{diam}(K))$.

We denote by d the distance in $\mathbb{R}^d$ and m the measure in $\mathbb{R}^d$ or $\mathbb{R}^{d-1}$.

We also need some assumption on the mesh:

$$\exists\, \xi > 0 \text{ s. t. } \mathrm{d}(x_K, \sigma) \geq \xi \mathrm{d}(x_K, x_L) \text{ for } K \in \mathcal{T},\ \text{ for } \sigma = K|L \in \mathcal{E}_{int,K}.$$

For all $\sigma \in \mathcal{E}$, we define the transmissibility coefficient $\tau_\sigma = \dfrac{\mathrm{m}(\sigma)}{d_\sigma}$, where $d_\sigma = \mathrm{d}(x_K, x_L)$ for $\sigma = K|L \in \mathcal{E}_{int}$ and $d_\sigma = \mathrm{d}(x_K, \sigma)$ for $\sigma \in \mathcal{E}_{ext}$.

Let $(\mathcal{T}, \mathcal{E}, (x_K)_{K\in\mathcal{T}})$ be an admissible discretization of Ω and let us define the time step Δt, $N_T = E(T/\Delta t)$ and the increasing sequence $(t^n)_{0\leq n\leq N_T}$, where $t^n = n\Delta t$, in order to get a space-time discretization $\mathcal{D}$ of $\Omega \times (0, T)$. The size of the space-time discretization $\mathcal{D}$ is defined by $\delta = \max(\Delta x, \Delta t)$.

First of all, the initial conditions and the doping profile are approximated by $\left(N_K^0, P_K^0, C_K\right)_{K\in\mathcal{T}}$ by taking the mean values of N_0, P_0 and C on each cell K. The numerical boundary conditions $\left(N_\sigma^{n+1}, P_\sigma^{n+1}, V_\sigma^{n+1}\right)_{n\geq 0, \sigma\in\mathcal{E}_{ext}^D}$ are also given by the mean values of $(\overline{N}, \overline{P}, \overline{V})$ on $\sigma \times [t^n, t^{n+1}[$.

2.1 The scheme for the thermal equilibrium

We compute an approximation $(N_K^{eq}, P_K^{eq}, V_K^{eq})_{K\in\mathcal{T}}$ of the thermal equilibrium (N^{eq}, P^{eq}, V^{eq}) defined by (3) with the finite volume scheme proposed by C. Chainais-Hillairet and F. Filbet in [2]:

$$\begin{cases} \lambda^2 \displaystyle\sum_{\sigma\in\mathscr{E}_K} \tau_\sigma DV^{eq}_{K,\sigma} = \mathrm{m}(K)\left(\exp(\alpha_N + V^{eq}_K) - \exp(\alpha_P - V^{eq}_K) - C_K\right) & \forall K \in \mathscr{T}, \\ N^{eq}_K = \exp(\alpha_N + V^{eq}_K), \quad P^{eq}_K = \exp(\alpha_P - V^{eq}_K) & \forall K \in \mathscr{T}, \end{cases} \tag{4}$$

where for a given function f and $(U_K)_{K\in\mathscr{T}}$, $Df(U)_{K,\sigma}$ is defined by:

$$Df(U)_{K,\sigma} = \begin{cases} f(U_L) - f(U_K) & \text{if } \sigma = K|L \in \mathscr{E}_{int,K}, \\ f(U_\sigma) - f(U_K) & \text{if } \sigma \in \mathscr{E}^D_{ext,K}, \\ 0 & \text{if } \sigma \in \mathscr{E}^N_{ext,K}. \end{cases}$$

Assuming that the boundary conditions satisfy hypotheses (H1)–(H4), the scheme (4) admits a unique solution (see [2]).

2.2 The scheme for the transient model

The Scharfetter–Gummel scheme for the system (1) is defined by:

$$\begin{cases} \mathrm{m}(K)\dfrac{N^{n+1}_K - N^n_K}{\Delta t} + \displaystyle\sum_{\sigma\in\mathscr{E}_K} \mathscr{F}^{n+1}_{K,\sigma} = 0, & \forall K \in \mathscr{T}, \forall n \geq 0, \\ \mathrm{m}(K)\dfrac{P^{n+1}_K - P^n_K}{\Delta t} + \displaystyle\sum_{\sigma\in\mathscr{E}_K} \mathscr{G}^{n+1}_{K,\sigma} = 0, & \forall K \in \mathscr{T}, \forall n \geq 0, \\ \lambda^2 \displaystyle\sum_{\sigma\in\mathscr{E}_K} \tau_\sigma DV^n_{K,\sigma} = \mathrm{m}(K)\left(N^n_K - P^n_K - C_K\right), & \forall K \in \mathscr{T}, \forall n \geq 0, \end{cases} \tag{5}$$

with for all $\sigma \in \mathscr{E}_K$

$$\mathscr{F}^{n+1}_{K,\sigma} = \tau_\sigma \left(B\left(-DV^{n+1}_{K,\sigma}\right) N^{n+1}_K - B\left(DV^{n+1}_{K,\sigma}\right) N^{n+1}_\sigma \right), \tag{6}$$

$$\mathscr{G}^{n+1}_{K,\sigma} = \tau_\sigma \left(B\left(DV^{n+1}_{K,\sigma}\right) P^{n+1}_K - B\left(-DV^{n+1}_{K,\sigma}\right) P^{n+1}_\sigma \right), \tag{7}$$

where B is the Bernoulli function defined by:

$$B(x) = \frac{x}{e^x - 1} \text{ for } x \neq 0, \quad B(0) = 1. \tag{8}$$

We consider a fully implicit discretization in time to avoid the restrictive stability condition $\Delta t \leq \lambda^2/M$.
Using a fixed point theorem, we can prove the following result:

Theorem 1. *Let us assume (H1)–(H4) and $C = 0$. Then there exists a solution $\{(N^n_K, P^n_K, V^n_K), K \in \mathscr{T}, 0 \leq n \leq N_T + 1\}$ to the scheme (5)–(6)–(7), and moreover we have*

$$0 < m \leq N^n_K,\ P^n_K \leq M, \quad \forall K \in \mathscr{T}, \quad \forall n \geq 0. \tag{9}$$

3 Asymptotic behavior of the Scharfetter–Gummel scheme

We may now state our main result.

Theorem 2. *Let us assume (H1)–(H4) and* $C = 0$*. Then solution* $(N_\delta, P_\delta, V_\delta)$ *given by the scheme (5)–(6)–(7) satisfies for each* $K \in \mathcal{T}$

$$\left(N_K^n, P_K^n, V_K^n\right) \longrightarrow \left(N_K^{eq}, P_K^{eq}, V_K^{eq}\right) \text{ as } n \to +\infty,$$

where $\left(N_K^{eq}, P_K^{eq}, V_K^{eq}\right)_{K\in\mathcal{T}}$ *is an approximation to the solution of the steady-state equation (3) given by (4).*

The proof is based, as in the continuous case (see [5] and [8]), on an energy estimate and a control of its dissipation, given in Proposition 1 which is valid even if $C \neq 0$. Nevertheless to prove rigorously the convergence to equilibrium, we need the uniform lower bound (9) on N and P which holds under the restrictive assumption $C = 0$.
In the last section, we perform some numerical experiments and observe a convergence to steady-state even when this condition is not satisfied.

3.1 Notations and definitions

For $U = (U_K)_{K\in\mathcal{T}}$, we define the H^1-seminorm as follows:

$$|U|_{1,\Omega}^2 = \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \left|DU_{K,\sigma}\right|^2 + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}} \tau_\sigma \left|DU_{K,\sigma}\right|^2$$

Since the study of the large time behavior of the scheme (5)–(6)–(7) is based on an energy estimate with the control of its dissipation, let us introduce the discrete version of the deviation of the total energy from the thermal equilibrium:

$$\begin{aligned}\mathcal{E}^n = &\sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(N_K^n) - H(N_K^{eq}) - \log(N_K^{eq})\left(N_K^n - N_K^{eq}\right)\right)\\ &+ \sum_{K\in\mathcal{T}} \mathrm{m}(K)\left(H(P_K^n) - H(P_K^{eq}) - \log(P_K^{eq})(P_K^n - P_K^{eq})\right)\\ &+ \frac{\lambda^2}{2}\left|V^n - V^{eq}\right|_{1,\Omega}^2 .\end{aligned}$$

Since $s \mapsto H(s) = \displaystyle\int_1^s \log(\tau)d\tau$ is defined and convex on $\mathbb{R}_+$, we have $\mathcal{E}^n \geq 0$ for all $n \geq 0$. We also introduce the discrete version of the energy dissipation:

$$
\begin{aligned}
\mathscr{I}^n = &\sum_{\substack{\sigma\in\mathscr{E}_{int}\\ \sigma=K|L}} \tau_\sigma \min\left(N_K^n, N_L^n\right)\left[D\left(\log\left(N^n\right)-V^n\right)_{K,\sigma}\right]^2\\
&+\sum_{K\in\mathscr{T}}\sum_{\sigma\in\mathscr{E}_{ext,K}} \tau_\sigma \min\left(N_K^n, N_\sigma^n\right)\left[D\left(\log\left(N^n\right)-V^n\right)_{K,\sigma}\right]^2\\
&+\sum_{\substack{\sigma\in\mathscr{E}_{int}\\ \sigma=K|L}} \tau_\sigma \min\left(P_K^n, P_L^n\right)\left[D\left(\log\left(P^n\right)+V^n\right)_{K,\sigma}\right]^2\\
&+\sum_{K\in\mathscr{T}}\sum_{\sigma\in\mathscr{E}_{ext,K}} \tau_\sigma \min\left(P_K^n, P_\sigma^n\right)\left[D\left(\log\left(P^n\right)+V^n\right)_{K,\sigma}\right]^2 .
\end{aligned}
$$

3.2 Energy estimate

The following Proposition gives the control of energy and dissipation. With this result, Theorem 2 can be proved in the same way as Theorem 2.2 in [2].

Proposition 1. *Under hypotheses (H1)–(H4), we have for all $n \geq 0$:*

$$
0 \leq \mathscr{E}^{n+1} + \Delta t\, \mathscr{I}^{n+1} \leq \mathscr{E}^n .
$$

Proof. Firstly, using the convexity of H and (4), we get

$$
\begin{aligned}
\mathscr{E}^{n+1}-\mathscr{E}^n \leq &\sum_{K\in\mathscr{T}} \mathrm{m}(K)\left(\log\left(N_K^{n+1}\right)-\alpha_N - V_K^{eq}\right)\left(N_K^{n+1}-N_K^n\right)\\
&+\sum_{K\in\mathscr{T}} \mathrm{m}(K)\left(\log\left(P_K^{n+1}\right)-\alpha_P + V_K^{eq}\right)\left(P_K^{n+1}-P_K^n\right)\\
&+\frac{\lambda^2}{2}\left|V^{n+1}-V^{eq}\right|_{1,\Omega}^2-\frac{\lambda^2}{2}\left|V^n-V^{eq}\right|_{1,\Omega}^2 ,
\end{aligned}
$$

and then, by adding $V_K^{n+1}-V_K^{n+1}$ in the two first sums, we have

$$
\mathscr{E}^{n+1}-\mathscr{E}^n \leq T_1 + T_2 + T_3,
$$

where

$$
\begin{aligned}
T_1 &= \sum_{K\in\mathscr{T}} m(K)\left(\log\left(N_K^{n+1}\right)-\alpha_N - V_K^{n+1}\right)\left(N_K^{n+1}-N_K^n\right),\\
T_2 &= \sum_{K\in\mathscr{T}} m(K)\left(\log\left(P_K^{n+1}\right)-\alpha_P + V_K^{n+1}\right)\left(P_K^{n+1}-P_K^n\right),
\end{aligned}
$$

$$T_3 = \sum_{K\in\mathcal{T}} m(K)\left(V_K^{n+1} - V_K^{eq}\right)\left(N_K^{n+1} - N_K^n - P_K^{n+1} + P_K^n\right)$$
$$+\frac{\lambda^2}{2}\left|V^{n+1} - V^{eq}\right|_{1,\Omega}^2 - \frac{\lambda^2}{2}\left|V^n - V^{eq}\right|_{1,\Omega}^2 .$$

Using the scheme (5) and an integration by parts, we get that $T_3 \le 0$ and

$$T_1 = \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \mathcal{R}_{K,\sigma}^{n+1} + \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \tau_\sigma \mathcal{R}_{K,\sigma}^{n+1},$$

where for $\sigma = K|L$,

$$\mathcal{R}_{K,\sigma}^{n+1} = \Big(D\log\left(N^{n+1}\right)_{K,\sigma} - DV_{K,\sigma}^{n+1}\Big)\Big(B\left(-DV_{K,\sigma}^{n+1}\right)N_K^{n+1} - B\left(DV_{K,\sigma}^{n+1}\right)N_L^{n+1}\Big).$$

We now prove that

$$\mathcal{R}_{K,\sigma}^{n+1} \le \mathcal{S}_{K,\sigma}^{n+1} := -\min\left(N_K^{n+1}, N_L^{n+1}\right)\left(D\log\left(N^{n+1}\right)_{K,\sigma} - DV_{K,\sigma}^{n+1}\right)^2 .$$

Indeed, applying the property $B(-x) - B(x) = x$, we obtain

$$\begin{aligned}
&\mathcal{R}_{K,\sigma}^{n+1} - \mathcal{S}_{K,\sigma}^{n+1} = \Big(D\log\left(N^{n+1}\right)_{K,\sigma} - DV_{K,\sigma}^{n+1}\Big)\times\\
&\Big[\Big(B\left(-DV_{K,\sigma}^{n+1}\right) - B\Big(-D\log\left(N^{n+1}\right)_{K,\sigma}\Big)\Big)\left(N_K^{n+1} - \min\left(N_K^{n+1}, N_L^{n+1}\right)\right)\\
&-\Big(B\left(DV_{K,\sigma}^{n+1}\right) - B\Big(D\log\left(N^{n+1}\right)_{K,\sigma}\Big)\Big)\left(N_L^{n+1} - \min\left(N_K^{n+1}, N_L^{n+1}\right)\right)\\
&+ B\Big(-D\log\left(N^{n+1}\right)_{K,\sigma}\Big)N_K^{n+1} - B\Big(D\log\left(N^{n+1}\right)_{K,\sigma}\Big)N_L^{n+1}\Big].
\end{aligned}$$

Now, since B is non-increasing on $\mathbb{R}$, the two first terms are non positive, and by using the definition (8) of B, the third term is equal to zero. Then we can conclude that

$$T_1 \le \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \mathcal{S}_{K,\sigma}^{n+1} + \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{ext,K}^D} \tau_\sigma \mathcal{S}_{K,\sigma}^{n+1},$$

and we obtain in the same way a similar estimate for T_2. To sum up, we have

$$\mathcal{E}^{n+1} - \mathcal{E}^n \le T_1 + T_2 \le -\Delta t\,\mathcal{I}^{n+1},$$

which completes the proof. □

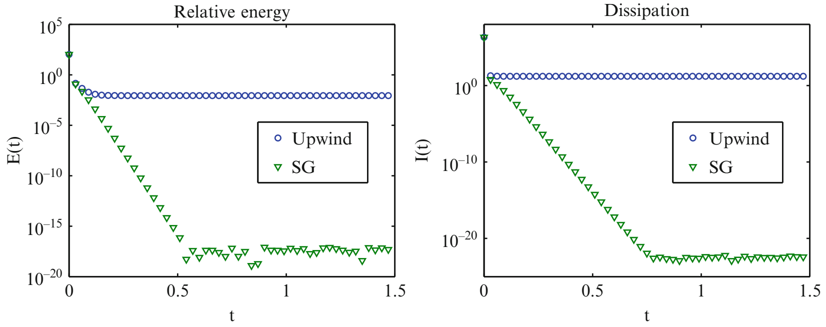


Fig. 1 Evolution of the relative energy $\mathscr{E}^n$ and its dissipation $\mathscr{I}^n$ in log-scale

4 Numerical experiments

We present here a test case for a geometry corresponding to a PN-junction in 1D. The doping profile is piecewise constant, equal to +1 in the N-region $]0.5, 1[$ and -1 in the P-region $]0, 0.5[$. The Debye length is $\lambda = 10^{-2}$.
In Fig. 1 we compare the relative energy $\mathscr{E}^n$ and its dissipation $\mathscr{I}^n$ obtained with the the Scharfetter–Gummel scheme (5) and with the scheme studied by C. Chainais-Hillairet, J. G. Liu and Y. J. Peng in [3], where the diffusion terms are discretized classically and the convection terms are discretized with upwind fluxes. With the Scharfetter–Gummel scheme, we observe that $\mathscr{E}^n$ and $\mathscr{I}^n$ converge to zero when $n \to \infty$, which is in keeping with Theorem 2. On the contrary, the upwind scheme, which does not preserve thermal equilibrium, is not very satisfying to reflect the long time behavior of the solution.

Acknowledgements The author is partially supported by the European Research Council ERC Starting Grant 2009, project 239983-NuSiKiMo.

References

1. F. Brezzi, L.D. Marini, S. Micheletti, P. Pietra, R. Sacco, and S. Wang. Discretization of semiconductor device problems. I. In *Handbook of numerical analysis. Vol. XIII*, pages 317–441. North-Holland, Amsterdam, 2005.
2. C. Chainais-Hillairet and F. Filbet. Asymptotic behavior of a finite volume scheme for the transient drift-diffusion model. *IMA J. Numer. Anal.*, 27(4):689–716, 2007.
3. C. Chainais-Hillairet, J.G. Liu, and Y.J. Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *M2AN Math. Model. Numer. Anal.*, 37(2):319–338, 2003.
4. R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of numerical analysis*, volume VII, pages 713–1020. North-Holland, Amsterdam, 2000.
5. H. Gajewki and K. Gärtner. On the discretization of Van Roosbroeck's equations with magnetic field. *Z. Angew. Math. Mech.*, 76(5):247–264, 1996.

6. A. Glitzky. Exponential decay of the free energy for discretized electro-reaction-diffusion systems. *Nonlinearity*, 21(9):1989–2009, 2008.
7. A.M. Il'in. A difference scheme for a differential equation with a small parameter multiplying the highest derivative. *Math. Zametki*, 6:237–248, 1969.
8. A. Jüngel. Qualitative behavior of solutions of a degenerate nonlinear drift-diffusion model for semiconductors. *Math. Models Methods Appl. Sci.*, 5(5):497–518, 1995.
9. R. D. Lazarov, Ilya D. Mishev, and P. S. Vassilevski. Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.*, 33(1):31–55, 1996.
10. D.L. Scharfetter and H.K. Gummel. Large signal analysis of a silicon Read diode. *IEEE Trans. Elec. Dev.*, 16:64–77, 1969.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Finite Volume Solver for Radiation Hydrodynamics in the Non Equilibrium Diffusion Limit

D. Chauveheid, J.-M.Ghidaglia, and M. Peybernes

Abstract We derive an Implicit Explicit finite volume scheme for the computation of radiation hydrodynamics. The convective part is handled through a classical upwind method while the reactive and diffusive parts are discretized thanks to a centered scheme. These results are compared to semi-analytic solutions obtained by Lowrie and Edwards [10].

Keywords Cell centered finite volume scheme, Radiation hydrodynamics, Implicit/Explicit schemes
MSC2010: 65M08, 76M12, 76N99, 80A99

1 Introduction

Radiation hydrodynamics models are of interest for many applications *e.g.* astrophysics, inertial confinement fusion (ICF) and other flows with very high temperatures. One of the major difficulties for these multi-physics problems is the presence of multiple time scales. From the numerical point of view, this leads to build implicit-explicit schemes with respect to time. The implicit part is here to handle small time scales while the explicit one takes care of larger time scales. In our context, the small time scales result from the radiation transport part (diffusion) while larger time scales come from purely hydrodynamical phenomena (entropy and pressure waves). Our strategy consists in relying on classical cell centered Finite

Daniel Chauveheid and Mathieu Peybernes
CEA, DAM, DIF, F-91297 Arpajon, France, e-mail: daniel.chauveheid@cmla.ens-cachan.fr, mathieu.peybernes@cea.fr

Jean-Michel Ghidaglia
CMLA, ENS Cachan and CNRS UMR 8635, 61 avenue du Président Wilson, F-94235 CACHAN CEDEX, e-mail: jmg@cmla.ens-cachan.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_26, © Springer-Verlag Berlin Heidelberg 2011

Volume schemes based on approximate Riemann solver (namely Flux Schemes see Ghidaglia [5]) for the hydrodynamics part and on an implicit Finite Volume scheme for the radiative one.

This article is a first step towards the derivation of a multi-material solver, that is studying flows with two or more different materials. For example in the ICF applications, we have at least two materials in presence, a metal (Gold) and a highly compressed gas (a mixture of Deuterium and Tritium). The multi material version of the scheme (Chauveheid [3]), relies on a generalization of the method of Braeunig *et al.* [1]. The latter method computes sharp interfaces between non miscible materials whose computation uses directional splitting. Hence in this paper, although we solely address the case of one material, we shall use cartesian meshes.

The governing equations, in non dimensional form (Lowrie and Edwards [10], Lowrie and Morel [9]), read in $3D$ as:

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

$$\frac{\partial (\rho \mathbf{u})}{\partial t} + \nabla \cdot \left(\rho \mathbf{u} \otimes \mathbf{u} + \left(p + \mathscr{P}_0 \frac{\mathscr{E}_r}{3} \right) Id \right) = 0\,, \tag{2}$$

$$\frac{\partial (\rho E)}{\partial t} + \nabla \cdot ((\rho E + p)\,\mathbf{u}) = -\mathscr{P}_0 \left(\sigma (T^4 - \mathscr{E}_r) + \mathbf{u} \cdot \nabla \frac{\mathscr{E}_r}{3} \right)\,, \tag{3}$$

$$\frac{\partial \mathscr{E}_r}{\partial t} + \nabla \cdot (\mathscr{E}_r \mathbf{u}) + \frac{\mathscr{E}_r}{3} \nabla \cdot \mathbf{u} = \nabla \cdot (\kappa \nabla \mathscr{E}_r) + \sigma (T^4 - \mathscr{E}_r)\,, \tag{4}$$

where we denote by ρ the density, $\mathbf{u}$ the velocity field, p the hydrodynamic pressure, related to the density ρ and the internal energy e by an equation of state: $EOS(p, \rho, e) = 0$. The hydrodynamic specific energy $E = e + \frac{1}{2}\|\mathbf{u}\|^2$ is the sum of the specific internal energy e and the kinetic energy, T is the material temperature. The radiative energy is denoted by $\mathscr{E}_r$ and we define the radiation temperature by $T_r^4 = \mathscr{E}_r$. Finally, $\mathscr{P}_0$ is a non dimensional number ([9, 10]). This system is non conservative but adding (3) and $\mathscr{P}_0$ (4) we readily obtain the total energy conservation law:

$$\frac{\partial (\rho E + \mathscr{P}_0 \mathscr{E}_r)}{\partial t} + \nabla \cdot \left(\left(\rho E + p + 4 \mathscr{P}_0 \frac{\mathscr{E}_r}{3} \right) \mathbf{u} \right) = \mathscr{P}_0 \nabla \cdot (\kappa \nabla \mathscr{E}_r)\,. \tag{5}$$

Then, introducing the radiative entropy (as done in [2]) $S_r \equiv T_r^3$, we can rewrite (4) as

$$\frac{\partial S_r}{\partial t} + \nabla \cdot (S_r \mathbf{u}) = \frac{3}{4 T_r} \left[\nabla \cdot (\kappa \nabla \mathscr{E}_r) + \sigma (T^4 - \mathscr{E}_r) \right]\,. \tag{6}$$

The system (1), (2), (5) and (6) is conservative as far as convection terms are concerned. Equation (6) is a non linear heat equation for the radiative temperature T_r. This variable is therefore diffused and the non conservative product appearing in the right hand side of this equation should not induce non uniqueness of solutions.

2 Numerical scheme

We use an operator splitting which consists in solving first the left-hand side of (1), (2), (5) and (6) by means of an upwind explicit finite volume scheme. Then, the diffusion-reaction part is discretized thanks to a centered implicit finite volume scheme. This kind of technique is often referred to as IMEX method (for Implicit/Explicit), see for example [7, 8].

We consider a regular cartesian grid and split also the space differential operators, that is to say we solve successively the x-derivative terms, the y-derivative terms and the z-derivative term.

Therefore, and without loss of generality, we deal only with $1D$ schemes, corresponding to what is done direction by direction. From now on, we call x the generic direction that we are looking at.

2.1 *Cell centered upwind Finite Volume scheme for the convection operator*

We denote by $v = (\rho, \rho\mathbf{u}, \rho E + \mathscr{P}_0\mathscr{E}_r, S_r)$ the conservative variables for the convective part of the system (1), (2), (5) and (6), and $F(v)$ the flux matrix such that:

$$F(v)\cdot\mathbf{n} \equiv (\rho(\mathbf{u}\cdot\mathbf{n}), \rho\mathbf{u}(\mathbf{u}\cdot\mathbf{n}) + (p + \mathscr{P}_0\mathscr{E}_r/3)\,\mathbf{n}, S_r(\mathbf{u}\cdot\mathbf{n}))\,, \tag{7}$$

is the normal flux in the direction $\mathbf{n} \in \mathbb{S}^{d-1}$, d being the physical space dimension.

With these notations, the left-hand side of equations (1)-(2)-(5)-(6) reads:

$$\frac{\partial v}{\partial t} + \nabla\cdot F(v) = 0\,. \tag{8}$$

The integration of (8) over a control volume $K_{i,j,k} = [x_i, x_{i+1}] \times [y_j, y_{j+1}] \times [z_k, z_{k+1}]$, keeping only the terms corresponding to the derivation in the generic x-direction, leads to a system of ordinary differential equations:

$$\frac{dV_{K_{i,j,k}}}{dt} + \frac{1}{|K_{i,j,k}|}\left(A_{i+1/2,j,k}\phi(v_{i+1,j,k}, v_{i,j,k}) - A_{i-1/2,j,k}\phi(v_{i,j,k}, v_{i-1,j,k})\right) = 0\,, \tag{9}$$

where $\phi(v_{i+1,j,k}, v_{i,j,k})$ denotes the numerical flux at the interface between volumes $K_{i,j,k}$ and $K_{i+1,j,k}$. $A_{i+1/2,j,k}$ is the measure of the edge located at $x_{i+1/2} \equiv \frac{x_i + x_{i+1}}{2}$.

The Characteristic Flux Finite Volume (CFFV) scheme. The CFFV scheme [4] consists in choosing, for the numerical flux in (9), the following value:

$$\phi(v, w, \mathbf{n}) = \frac{F(v) + F(w)}{2}\cdot\mathbf{n} - \mathscr{U}(u, v, \mathbf{n})\frac{F(w) - F(v)}{2}\cdot\mathbf{n}\,. \tag{10}$$

Here, $\mathbf{n} = e_x$, for the generic x-direction. $\mathscr{U}(u, v, \mathbf{n})$ is the sign matrix of the jacobian $\frac{\partial F(v)\cdot\mathbf{n}}{\partial v}$, in the sense that it has the same eigenvectors as, and its eigenvalues are the signs of those of $\frac{\partial F(v)\cdot\mathbf{n}}{\partial v}$. Namely, when $\frac{\partial F(v)\cdot\mathbf{n}}{\partial v}$ reads $L(diag(\lambda_i))R$ (which is the case for hyperbolic problems), with λ_i the eigenvalues, R right eigenvectors, and L left eigenvectors such that $LR = Id$, we have $\mathscr{U}(u, v, \mathbf{n}) = L(diag(sign(\lambda_i)))R$.

The boundary conditions use the normal flux method, we refer to [6].

Eigenelements. The jacobian matrix $\frac{\partial F(v)\cdot\mathbf{n}}{\partial v}$ of the normal flux (7) is found to be equal to:

$$\begin{pmatrix} 0 & \mathbf{n} & 0 & 0 \\ K\mathbf{n} - \mathbf{u}(\mathbf{u}\cdot\mathbf{n}) & \mathbf{u}\otimes\mathbf{n} - k\mathbf{n}\otimes\mathbf{u} + (\mathbf{u}\cdot\mathbf{n})Id & k\mathbf{n} & \frac{4}{9}\mathscr{P}_0 T_r(1-3k)\mathbf{n} \\ (K - (H + \frac{4\mathscr{P}_0\mathscr{E}_r}{3\rho}))\mathbf{u}\cdot\mathbf{n} & (H + \frac{4\mathscr{P}_0\mathscr{E}_r}{3\rho})\mathbf{n} - k(\mathbf{u}\cdot\mathbf{n})\mathbf{u} & \mathbf{u}\cdot\mathbf{n}(k+1) & \frac{4}{9}\mathscr{P}_0 T_r(1-3k)\mathbf{u}\cdot\mathbf{n} \\ -\frac{T_r^3}{\rho}\mathbf{u}\cdot\mathbf{n} & \frac{T_r^3}{\rho}\mathbf{n} & 0 & \mathbf{u}\cdot\mathbf{n} \end{pmatrix}.$$

Its eigenvalues are as follows:

$$\begin{cases} \lambda_1(v, \mathbf{n}) = \mathbf{u}\cdot\mathbf{n} - c_s\,, \\ \lambda_2(v, \mathbf{n}) = \cdots = \lambda_{d+2}(v, \mathbf{n}) = \mathbf{u}\cdot\mathbf{n}\,, \\ \lambda_{d+3}(v, \mathbf{n}) = \mathbf{u}\cdot\mathbf{n} + c_s\,. \end{cases} \tag{11}$$

with $k = \frac{1}{\rho T}\left(\frac{\partial p}{\partial s}\right)_\rho$, $c^2 = \left(\frac{\partial p}{\partial \rho}\right)_s$, s being the material entropy, $H = E + \frac{p}{\rho}$, $K = c^2 + k(\|\mathbf{u}\|^2 - H)$ and $c_s^2 = c^2 + \mathscr{P}_0\frac{4\mathscr{E}_r}{9\rho}$.

The right eigenvectors associated to these eigenvalues can be taken equal to:

$$\begin{cases} r_1(v, \mathbf{n}) = (1, \mathbf{u} - c_s\mathbf{n}, H + \frac{4\mathscr{P}_0\mathscr{E}_r}{3\rho} - c_s\mathbf{u}\cdot\mathbf{n}, \frac{T_r^3}{\rho})\,, \\ r_{d+1}(v, \mathbf{n}) = (1, \mathbf{u}, H - \frac{c^2}{k}, 0)\,, \\ r_{d+2}(v, \mathbf{n}) = (\mathscr{P}_0 T_r, \mathscr{P}_0 T_r\mathbf{u}, \mathscr{P}_0 T_r(H - 3c^2), -\frac{9}{4}c^2)\,, \\ r_{d+3}(v, \mathbf{n}) = (1, \mathbf{u} + c_s\mathbf{n}, H + \frac{4\mathscr{P}_0\mathscr{E}_r}{3\rho} + c_s\mathbf{u}\cdot\mathbf{n}, \frac{T_r^3}{\rho})\,, \\ r_2(v, \mathbf{n}) = (0, \mathbf{n}_2^\perp, \mathbf{u}\cdot\mathbf{n}_2^\perp), \cdots, r_d(v, \mathbf{n}) = (0, \mathbf{n}_d^\perp, \mathbf{u}\cdot\mathbf{n}_d^\perp)\,. \end{cases} \tag{12}$$

where $\mathbf{n}_2^\perp \cdots \mathbf{n}_d^\perp$ is an orthonormal basis of the hyperplane orthogonal to $\mathbf{n}$.

The dual basis is then:

$$\begin{cases} \ell_1(v, \mathbf{n}) = \frac{1}{2c_s^2}(K + c_s\mathbf{u}\cdot\mathbf{n}, -k\mathbf{u} - c_s\mathbf{n}, k, \frac{4}{9}\mathscr{P}_0 T_r(1-3k))\,, \\ \ell_{d+1}(v, \mathbf{n}) = \frac{k}{c^2}(H - \|\mathbf{u}\|^2, \mathbf{u}, -1, \frac{4}{9}\mathscr{P}_0 T_r))\,, \\ \ell_{d+2}(v, \mathbf{n}) = \frac{4}{9\rho c^2 c_s^2}(T_r^3 K, -kT_r^3\mathbf{u}, kT_r^3, -\rho c^2 - \frac{4}{9}\mathscr{P}_0 k\mathscr{E}_r)\,, \\ \ell_{d+3}(v, \mathbf{n}) = \frac{1}{2c_s^2}(K - c_s\mathbf{u}\cdot\mathbf{n}, -k\mathbf{u} + c_s\mathbf{n}, k, \frac{4}{9}\mathscr{P}_0 T_r(1-3k))\,, \\ \ell_2(v, \mathbf{n}) = (-\mathbf{u}\cdot\mathbf{n}_2^\perp, \mathbf{n}_2^\perp, 0, 0), \cdots, \ell_d(v, \mathbf{n}) = (-\mathbf{u}\cdot\mathbf{n}_d^\perp, \mathbf{n}_2^\perp, 0, 0)\,. \end{cases} \tag{13}$$

Time discretization and stability condition. We use the explicit Euler's scheme to discretize the time derivative in (9) and then the Courant condition for the linearized scheme reads:

$$\max_{i,j,k} |\lambda^n_{i,j,k}| \frac{A_{i,j,k}}{|K_{i,j,k}|} \Delta t^n \leqslant CFL \leqslant 1, \tag{14}$$

where $A_{i,j,k}$ is either $\Delta x_i \ \Delta y_j$, $\Delta y_j \ \Delta z_k$ or $\Delta z_k \ \Delta x_i$ depending on the direction we solve.

2.2 Implicit centered finite volume scheme for the diffusion equation

The diffusion part consists in the following system:

$$\frac{\partial \rho}{\partial t} = 0\,, \tag{15}$$

$$\frac{\partial (\rho \mathbf{u})}{\partial t} = 0\,, \tag{16}$$

$$\frac{\partial (\rho E)}{\partial t} = -\mathscr{P}_0 \sigma (T^4 - \mathscr{E}_r)\,, \tag{17}$$

$$\frac{\partial \mathscr{E}_r}{\partial t} = \nabla \cdot (\kappa \nabla \mathscr{E}_r) + \sigma (T^4 - \mathscr{E}_r)\,. \tag{18}$$

Since (18) is a heat equation, if we want to use reasonable time step (governed by the Courant Friedrichs Lewy condition (14)), we have to make use of an implicit time discretization.

Writing $E = C_v T + \|\mathbf{u}\|^2/2$, and using (15) and (16), we can show that (17) reduces to an ODE for the temperature T:

$$\rho C_v \frac{\partial T}{\partial t} = -\mathscr{P}_0 \sigma (T^4 - \mathscr{E}_r)\,. \tag{19}$$

The scheme then reads:

$$\rho^n_i C_v \frac{T^{n+1}_i - T^n_i}{\Delta t^n} = -\mathscr{P}_0 \sigma^n_i ((T^{n+1}_i)^4 - \mathscr{E}^{n+1}_{r,i}), \tag{20}$$

$$\frac{\mathscr{E}^{n+1}_{r,i} - \mathscr{E}^n_{r,i}}{\Delta t^n} - 2 \frac{\kappa^n_{i+1/2} \frac{\mathscr{E}^{n+1}_{r,i+1} - \mathscr{E}^{n+1}_{r,i}}{\Delta x_{i+1}} - \kappa^n_{i-1/2} \frac{\mathscr{E}^{n+1}_{r,i} - \mathscr{E}^{n+1}_{r,i-1}}{\Delta x_i}}{(\Delta x_i + \Delta x_{i+1})} = \sigma^n_i ((T^{n+1}_i)^4 - \mathscr{E}^{n+1}_{r,i}), \tag{21}$$

$$\frac{2}{\kappa^n_{i+1/2}} = \frac{1}{\kappa^n_i} + \frac{1}{\kappa^n_{i+1}}\,.$$

It can be shown by a motonicity argument that this system has a unique solution. It is then solved by the Newton method, mainly because of the nonlinear terms, and the GMRES algorithm ([11]) at each Newton iteration to solve the linear system.

3 Numerical results

In this section, we present numerical simulations of radiative shock solutions. These results are compared to semi-analytic solutions obtained following the method described in [10].

We initialize a Riemann problem setting the left-state (subscript 0) to $\rho_0 = 1, Tr_0 = 1, T_0 = 1, u_0 = \mathscr{M}$, for a given $\mathscr{M}$ (some different values are chosen for the tests), and the right-state (subscript 1) is obtained by solving the so-called "overall jump conditions" ([10]), and taking material and radiative temperatures equal to each other:

$$\rho_0 u_0 = \rho_l u_1 \tag{22}$$

$$\rho_0 u_0^2 + p_0 + \mathscr{P}_0 \frac{T_{r,0}^4}{3} = \rho_1 u_1^2 + p_1 + \mathscr{P}_0 \frac{T_{r,1}^4}{3} \tag{23}$$

$$u_0(\rho_0 E_0 + p_0 + \frac{4}{3}\mathscr{P}_0 T_{r,0}^4) = u_1(\rho_1 E_1 + p_1 + \frac{4}{3}\mathscr{P}_0 T_{r,1}^4) \tag{24}$$

Here, $\kappa = 1, \sigma = 10^6$ and $\mathscr{P}_0 = 10^{-4}$.

We take perfect gas equation of state $p = \frac{\rho T}{\gamma}$, with $\gamma = 5/3$.

Figure 1 shows a continuous solution computed over 128 cells. Solutions of Figs. 2 to 4 undergo discontinuities. For these simulations, a finer mesh is used to capture the solutions.

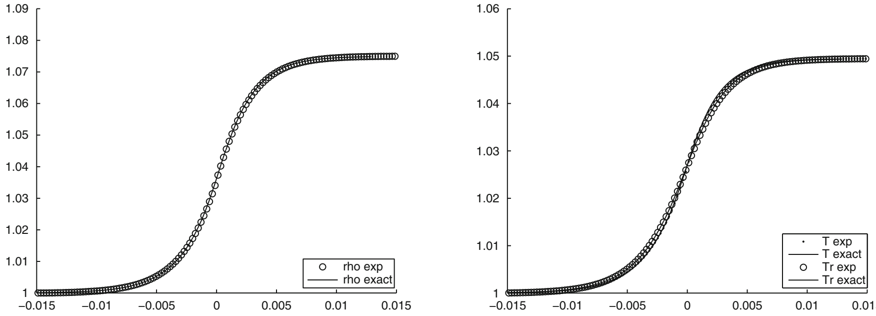


Fig. 1 Solution for density, temperature and radiative temperature for $\mathscr{M} = 1.05$. Comparison with semi-analytic solutions. Number of cells: 128

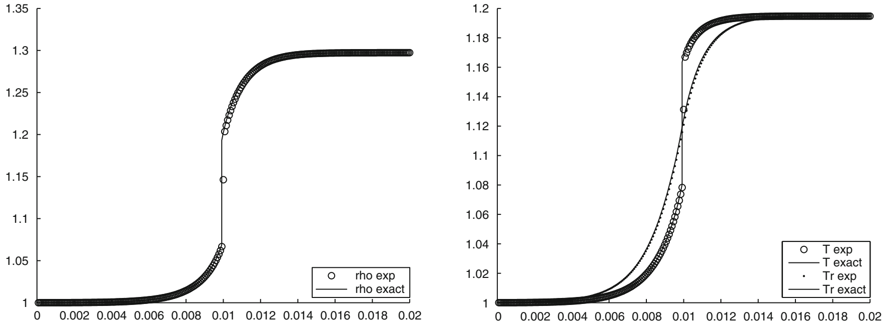


Fig. 2 Solution for density, temperature and radiative temperature for $\mathscr{M} = 1.2$. Comparison with semi-analytic solutions. Number of cells: 256

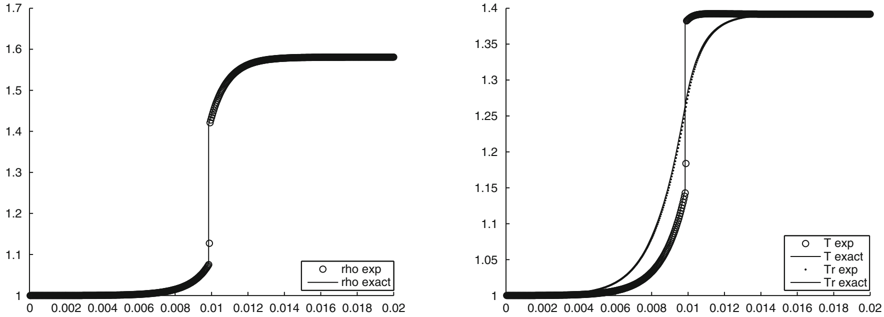


Fig. 3 Solution for density, temperature and radiative temperature for $\mathscr{M} = 1.4$. Comparison with semi-analytic solutions. Number of cells: 512

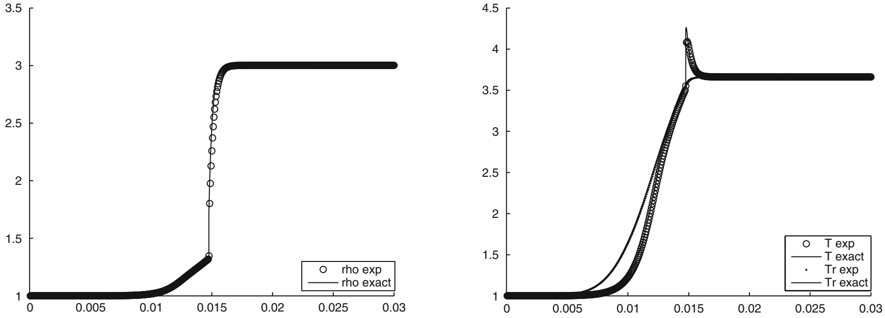


Fig. 4 Solution for density, temperature and radiative temperature for $\mathscr{M} = 3$. Comparison with semi-analytic solutions. Number of cells: 512

Numerical and theoretical results are in good agreement. The conservative formulation chosen in (6) seems to be relevant with regard to these particular physical solutions.

4 Conclusion

As said in the introduction, this work is a first step towards the derivation of a multi material $3D$ solver for multi material radiative hydrodynamics. In this paper we have presented our method for the single material case and shown that on physically relevant non trivial solutions, our solver behaves well. The extension for multi material flows is in progress (Chauveheid [3]). The method presented here was designed in order to make this extension as simple as possible. In fact it only remains to extend the so-called condensate techniques of Braeunig *et. al.* [1] to radiative flows.

References

1. Braeunig J.-P., Desjardins B., Ghidaglia J.-M., A totally Eulerian Finite Volume solver for multi-material fluid flows, Eur. J. Mech. B/Fluids, Vol. 28, pp. 475-485, 2009.
2. Buet, C., Despres B.: Asymptotic preserving and positive schemes for radiation hydrodynamic. J. Comput. Phys. **215**, 717–740, 2006.
3. Chauveheid D., Thesis, École normale supérieure, in preparation.
4. Ghidaglia J.-M., Kumbaro A., Le Coq G.: On the numerical solution to two fluid models *via* a cell centered finite volume method, Eur. J. Mech. B/Fluids, 20, 841-867, 2001.
5. Ghidaglia J.-M., Flux schemes for solving monlinear systems of conservation laws, *in* Innovative Methods for Numerical Solution of Partial Differential Equations, Chattot J.J. and Hafez M. Eds, pp 232-242, WORLD SCIENTIFIC, Singapore, 2001.
6. Ghidaglia J.-M., Pascal F.: The normal flux method at the boundary for multidimensional finite volume approximations in CFD. Eur. J. Mech. B/Fluids, Vol. 24(1), pp. 1-17, 2005.
7. Kadioglu, S.Y. , Knoll, D.A., Lowrie, R.B., Rauenzahn, R.M.: A second order self-consistent IMEX method for radiation hydrodynamics. J. Comput. Phys. **229**, 8313–8332, 2010.
8. Kadioglu, S.Y. , Knoll: A fully second order implicit/explicit time integration technique for hydrodynamics plus nonlinear heat conduction problems. J. Comput. Phys. **229**, 3237–3249 (2010).
9. Lowrie, R.B., Morel, J.E.: The coupling of radiation and hydrodynamics. Astrophys. J. **521**, 423–450, 1999.
10. Lowrie, R.B. , Edwards J.D.: Radiative shock solutions with grey nonequilibrium diffusion. Shock Waves **18**, 129–143, 2008.
11. Saad, Y.: Iterative methods for sparse linear systems 2nd edition, SIAM, 2003.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

An Extension of the MAC Scheme to some Unstructured Meshes

Eric Chénier, Robert Eymard, and Raphaèle Herbin

Abstract We give a variational formulation of the standard MAC scheme for the approximation of the Navier-Stokes problem. This allows an extension of the MAC scheme to locally refined Cartesian meshes. A numerical example is presented, which shows an efficient computation of the solution of the Navier-Stokes problem for a general 2D or 3D domain, using locally refined meshes.

Keywords MAC scheme, incompressible Navier-Stokes, non conforming grid
MSC2010: 65N08,76D05

1 Introduction

Our aim is the approximation on an unstructured mesh, of the weak solution to the steady-state Navier-Stokes equations, defined by

$$\begin{cases} \boldsymbol{u} \in E(\Omega),\ p \in L^2(\Omega) \text{ with } \displaystyle\int_\Omega p(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0, \\ \displaystyle\int_\Omega \nabla \boldsymbol{u}(\boldsymbol{x}) : \nabla \boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \mathrm{R}\int_\Omega (\boldsymbol{u}(\boldsymbol{x})\cdot\nabla)\boldsymbol{u}(x)\cdot\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \\ \qquad\qquad - \displaystyle\int_\Omega p(\boldsymbol{x})\mathrm{div}\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{f}(\boldsymbol{x})\cdot\boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x},\ \forall \boldsymbol{v} \in H_0^1(\Omega)^d, \end{cases} \tag{1}$$

where

E. Chénier and R. Eymard
Université Paris-Est, e-mail: eric.chenier@univ-mlv.fr, robert.eymard@univ-mlv.fr

R. Herbin
Université Aix-Marseille, e-mail: Raphaele.Herbin@latp.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_27, © Springer-Verlag Berlin Heidelberg 2011

$d \in \{2,3\}$ denotes the space dimension,
Ω is an open polygonal bounded and connected subset of $\mathbb{R}^d$,
with Lipschitz-continuous boundary $\partial\Omega$,

$$\mathrm{R} \in [0,+\infty),\ \boldsymbol{f} \in L^2(\Omega)^d,$$
$$E(\Omega) := \{\boldsymbol{v} = (v^{(i)})_{i=1,\dots,d} \in H_0^1(\Omega)^d, \mathrm{div}\boldsymbol{v} = 0 \text{ a.e. in } \Omega\},$$

and, for all $\boldsymbol{u},\boldsymbol{v} \in H_0^1(\Omega)^d$ and for a.e. $\boldsymbol{x} \in \Omega$, $\nabla\boldsymbol{u}(\boldsymbol{x}) : \nabla\boldsymbol{v}(\boldsymbol{x}) = \sum_{i=1}^{d} \nabla u^{(i)}(\boldsymbol{x}) \cdot \nabla v^{(i)}(\boldsymbol{x})$. The approximation of Problem (1) may be performed with several schemes among which the MAC scheme: see e.g. [7] for a presentation of its implementation and [3–6] for its mathematical analysis; the MAC scheme is very popular, in particular because it is simple and needs no stabilisation procedure. Its main drawback is that it only holds on domains which can be gridded by rectangular conforming meshes, in the sense that no hanging node is permitted. This paper is devoted to the presentation of a simple way to extend this scheme to any geometry and to possibly refined meshes, while keeping simplicity and convergence properties. In Sect. 2, we first write a discrete variational formulation of the standard MAC scheme on the Stokes problem, which is (1) with $\mathrm{R} = 0$. Thanks to this formulation, we are able in Sect. 3 to extend this scheme to more complex geometries and to the Navier-Stokes equation (1). Section 3 proposes a numerical example on a non-rectangular domain, using local refinement along the boundary of the domain.

2 The standard MAC scheme for the Stokes equations

Let us consider in this section the standard MAC scheme for the approximation of the Stokes problem, that is (1) with $\mathrm{R} = 0$. We then consider the following case and notations, as depicted in Fig. 1. Let us consider the unit square: $\Omega =]0,1[\times]0,1[$, let N and M be two positive integers. With the notations of Fig.1, we denote by $\mathcal{M}$ the set of pressure grid cells:

$$\mathcal{M} = \left\{]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[\times]y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}[,\ 1 \le i \le N,\ 1 \le j \le M \right\},$$

and by $\mathcal{E} = \mathcal{E}^{(1)} \cup \mathcal{E}^{(2)}$ the set of the edges of the mesh, where $\mathcal{E}^{(1)}$ (resp. $\mathcal{E}^{(2)}$) is the set of vertical (resp. horizontal) edges, associated to the x (resp. y) component of the velocity. In order to define the normal velocity flux from one cell to a neighbouring one, we introduce, for any pair $\sigma, \sigma' \in \mathcal{E}^{(k)}$, $k = 1$ or 2, the transmissivity $\tau_{\sigma,\sigma'}$ between cell $K_\sigma^{(k)}$ and cell $K_{\sigma'}^{(k)}$:

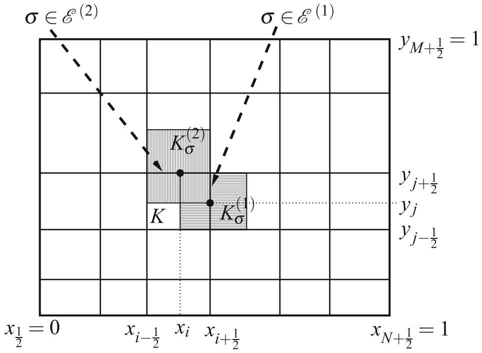


Fig. 1 Notations for the standard MAC scheme

$$\tau_{\sigma,\sigma'} = \frac{|\partial K_\sigma^{(k)} \cap \partial K_{\sigma'}^{(k)}|}{\mathrm{d}(\boldsymbol{x}_\sigma, \boldsymbol{x}_{\sigma'})}, \tag{2}$$

For instance, for a vertical edge $\sigma = \{x_{i-\frac{1}{2}}\}\times]y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}[\in \mathscr{E}^{(1)}$, one has:

$$\tau^{(1)}_{\sigma,\sigma'} = \begin{cases} \dfrac{y_{j+\frac{1}{2}} - y_{j-\frac{1}{2}}}{x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}} & \text{if } \sigma' = \{x_{i+\frac{1}{2}}\}\times]y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}}[, \\ \dfrac{x_i - x_{i-1}}{y_{j+1} - y_j} & \text{if } \sigma' = \{x_{i-\frac{1}{2}}\}\times]y_{j+\frac{1}{2}}, y_{j+\frac{3}{2}}[. \end{cases} \tag{3}$$

Denoting by $(\boldsymbol{e}^{(k)})_{k=1,\ldots,d}$ the canonical orthonormal basis of $\mathbb{R}^d$ and, for $K \in \mathscr{M}$, $\boldsymbol{n}_{K,\sigma}$ the unit normal vector to σ outward to K, the MAC scheme then reads:

$$\text{Find } (u_\sigma)_{\sigma\in\mathscr{E}} \subset \mathbb{R},\ (p_K)_{K\in\mathscr{M}} \subset \mathbb{R}\ ;\ \sum_{K\in\mathscr{M}} |K| p_K = 0,$$

$$\sum_{k=1}^{2} \sum_{\sigma\in\mathscr{E}_K^{(k)}} |\sigma| u_\sigma \boldsymbol{e}^{(k)} \cdot \boldsymbol{n}_{K,\sigma} = 0,\ \forall K \in \mathscr{M}, \tag{4a}$$

$$-\sum_{\sigma'\in\mathscr{E}^{(k)}} \tau^{(k)}_{\sigma,\sigma'}(u_{\sigma'} - u_\sigma) + |\sigma|(p_{L_\sigma} - p_{M_\sigma}) = \int_{K_\sigma^{(k)}} f^{(k)}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \forall\sigma \in \mathscr{E}^{(k)}, k = 1, 2, \tag{4b}$$

where L_σ and $M_\sigma \in \mathscr{M}$ are the two cells which share $\sigma \in \mathscr{E}^{(k)}$ as an edge, and such that $\boldsymbol{e}^{(k)}$ is oriented from L_σ and M_σ, and where the value of u_σ is set to 0 on all exterior edges.

In order to extend the MAC scheme, the idea is to rewrite (4a) and (4b) under a variational formulation. We first define $H_{\mathscr{M}}(\Omega)$ as the set of piecewise functions constant in $K \in \mathscr{M}$, and $H_{\mathscr{E}}^{(k)}(\Omega)$ as the set of piecewise functions which are constant in K_σ, for $\sigma \in \mathscr{E}^{(k)}$, and which are meant to approximate the kth component of the velocity. We finally denote by $H_{\mathscr{E}}(\Omega)$ the set of all $\boldsymbol{v} = (v^{(k)})_{k=1,\ldots,d}$ with $v^{(k)} \in H_{\mathscr{E}}^{(k)}(\Omega)$. We then define the discrete divergence by:

$$\mathrm{div}_K \boldsymbol{v} = \frac{1}{|K|} \sum_{\sigma \in \mathscr{E}_K} |\sigma| v_{K,\sigma}, \ \forall K \in \mathscr{M}, \ \forall \boldsymbol{v} \in H_{\mathscr{E}}(\Omega), \tag{5}$$

where, denoting by $\mathscr{E}_{\mathrm{int}}$ (resp. $\mathscr{E}_{\mathrm{ext}}$) the set of internal (resp. boundary) edges,

$$v_{K,\sigma} = \begin{cases} v_\sigma \boldsymbol{n}_\sigma \cdot \boldsymbol{n}_{K,\sigma} & \forall \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\mathrm{int}}, \\ 0 & \forall \sigma \in \mathscr{E}_K \cap \mathscr{E}_{\mathrm{ext}}, \end{cases} \quad \forall K \in \mathscr{M}, \ \forall \boldsymbol{v} \in H_{\mathscr{E}}(\Omega), \tag{6}$$

where $\boldsymbol{n}_\sigma$ denotes the basis vector $\boldsymbol{e}$ to which σ is orthogonal. Using (5), we may define the following operator:

$$\mathrm{div}_{\mathscr{D}} \boldsymbol{v}(\boldsymbol{x}) = \mathrm{div}_K \boldsymbol{v}, \ \text{ for a.e. } \boldsymbol{x} \in K, \ \forall K \in \mathscr{M}, \ \forall \boldsymbol{v} \in H_{\mathscr{E}}(\Omega), \tag{7}$$

and remark that (4a) can be written

$$\mathrm{div}_{\mathscr{D}} \boldsymbol{u}(\boldsymbol{x}) = 0, \ \text{for a.e. } \boldsymbol{x} \in \Omega. \tag{8}$$

Next, for $k = 1, \ldots, d$, we define an inner product on the space $H_{\mathscr{E}}^{(k)}$:

$$\langle u, v \rangle_k = \sum_{\{\sigma,\sigma'\} \subset \mathscr{E}^{(k)}} \tau_{\sigma,\sigma'}^{(k)} (u_\sigma - u_{\sigma'})(v_\sigma - v_{\sigma'}), \ \forall u, v \in H_{\mathscr{E}}^{(k)}(\Omega); \tag{9}$$

this allows the definition of the following inner product on $H_{\mathscr{E}}(\Omega)$ which is expected to approximate $\int_\Omega \nabla \boldsymbol{u}(\boldsymbol{x}) : \nabla \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$:

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathscr{E}} = \sum_{k=1}^{d} \langle u^{(k)}, v^{(k)} \rangle_k, \ \forall \boldsymbol{u}, \boldsymbol{v} \in H_{\mathscr{E}}(\Omega). \tag{10}$$

We then obtain, multiplying (4b) by v_σ and summing on $k = 1, 2$ and $\sigma \in \mathscr{E}^{(k)}$,

$$\langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathscr{E}} - \int_\Omega p(\boldsymbol{x}) \mathrm{div}_{\mathscr{D}} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \int_\Omega \boldsymbol{f}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \ \ \forall \boldsymbol{v} \in H_{\mathscr{E}}(\Omega), \tag{11}$$

A discrete variational formulation of the MAC scheme (4) is therefore:

$$\text{Find } \boldsymbol{u} \in H_{\mathscr{E}}(\Omega) \text{ and } p \in H_{\mathscr{M}}(\Omega) \text{ s. t. } \sum_{K \in \mathscr{M}} |K| p_K = 0 \text{ and (8) and (11) hold.} \tag{12}$$

3 The extended MAC scheme for the Navier-Stokes equations

We extend the standard MAC scheme to cases where all internal edges (2D) or faces (3D) whose normal is parallel to a basis vector $\boldsymbol{e}^{(k)}$, such as the pressure grid depicted in Fig. 2 (left). Because of possibly hanging nodes, we may no longer define the velocity meshes by dual rectangles, but use instead the Voronoi cells associated with the barycentres of the edges $(\boldsymbol{x}_\sigma)_{\sigma\in\mathscr{E}}$; they are defined as follows:

$$K_\sigma^{(k)} = \{\boldsymbol{x} \in \Omega, \mathrm{d}(\boldsymbol{x}, \boldsymbol{x}_\sigma) < \mathrm{d}(\boldsymbol{x}, \boldsymbol{x}_{\sigma'}), \sigma' \in \mathscr{E}^{(k)} \setminus \{\sigma\}\}, \ \forall \sigma \in \mathscr{E}^{(k)},$$

Note that in the case of a uniform rectangular mesh, the Voronoi cells thus defined are equal to the velocity cells defined in the previous section. However, this is no longer true if a non uniform mesh is used, even in the conforming case; indeed, in this latter case, the Voronoi cells $K_\sigma^{(k)}$ are again rectangles, but they are not equal to the rectangular cells $K_\sigma^{(k)}$ defined previously. In the case of hanging nodes, they are no longer rectangular, as can be seen in Fig. 2, where we depict the pressure mesh, the horizontal and vertical velocity grids.

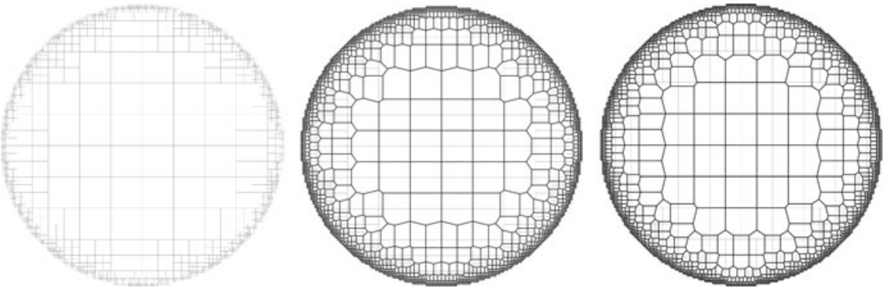


Fig. 2 The pressure and velocity grids

The diffusion term is again approximated by the discrete inner product defined by (9)-(10)-(2), but the expression of $\tau_{\sigma,\sigma'}$ given by (2) can no longer be written as in (3) for non rectangular cells. For Voronoï cells $K_\sigma^{(k)}$ and $K_{\sigma'}^{(k)}$ separated by a (dual) edge ε, such as those depicted in Fig. 3, one has

$$\tau_{\sigma,\sigma'} = \frac{|\varepsilon|}{d_\varepsilon} \tag{13}$$

where $|\varepsilon|$ denotes the length of the edge ε shared by $K_\sigma^{(k)}$ and $K_{\sigma'}^{(k)}$, and $d_\varepsilon = d(\boldsymbol{x}_\sigma, \boldsymbol{x}_{\sigma'})$ the distance between the two cell centres $\boldsymbol{x}_\sigma$ and $\boldsymbol{x}_{\sigma'}$, which are also the barycentres of the edges σ and σ'. We can again define $H_{\mathscr{M}}(\Omega)$

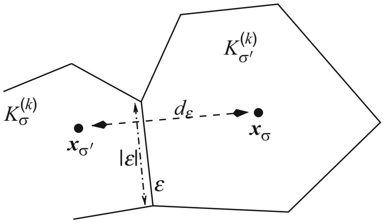


Fig. 3 Notations for a velocity cell

as the set of piecewise functions constant on the pressure cells $K \in \mathscr{M}$, the set $H^{(k)}_{\mathscr{E}}(\Omega)$ of piecewise constant functions on the grid cells K_σ, for $\sigma \in \mathscr{E}^{(k)}_{\text{int}} \cup \mathscr{E}_{\text{ext}}$ which vanish on any grid cell K_σ for a boundary edge $\sigma \subset \partial\Omega$; this discrete set is the space of functions meant to approximate the k-th component of the velocity. We finally denote by $H_{\mathscr{E}}(\Omega)$ the set of all $\boldsymbol{v} = (v^{(k)})_{k=1,\ldots,d}$ with $v^{(k)} \in H^k_{\mathscr{E}}(\Omega)$. The extended MAC scheme for the Stokes equations (R $=0$) is again (5)-(12), with the new definition (13) for $\tau_{\sigma,\sigma'}$.

In order to write this generalized scheme for the Navier-Stokes equations, we need to add the discretization of the nonlinear term $\int_\Omega (\boldsymbol{u}(\boldsymbol{x}) \cdot \nabla)\boldsymbol{u}(x) \cdot \boldsymbol{v}(\boldsymbol{x})\mathrm{d}\boldsymbol{x}$. For $\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in H_{\mathscr{E}}(\Omega)$, we define the discrete nonlinear convection term

$$b_{\mathscr{E}}(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = \sum_{K \in \mathscr{M}} \sum_{\substack{\sigma \in \mathscr{E}_K \\ \mathscr{M}_\sigma = \{K,L\}}} |\sigma| u_{K,\sigma} \frac{\boldsymbol{\Pi}_K \boldsymbol{v} + \boldsymbol{\Pi}_L \boldsymbol{v}}{2} \cdot \boldsymbol{\Pi}_K \boldsymbol{w},$$

where $u_{K,\sigma}$ is defined by (6) $\boldsymbol{\Pi}_K \boldsymbol{v}$ is a reconstruction of the full velocity on each pressure cell K defined by its components $(\boldsymbol{\Pi}_K \boldsymbol{v})_k, k = 1, \ldots, d$:

$$(\boldsymbol{\Pi}_K \boldsymbol{v})_k = \frac{1}{\sum_{\sigma \in \mathscr{E}_K \cap \mathscr{E}^{(k)}} |K^{(k)}_\sigma|} \sum_{\sigma \in \mathscr{E}_K \cap \mathscr{E}^{(k)}} |K^{(k)}_\sigma| v_\sigma.$$

The extended MAC scheme for the Navier-Stokes equation then reads:

$$\begin{cases} \text{Find } \boldsymbol{u} \in H_{\mathscr{E}}(\Omega) \text{ and } p \in H_{\mathscr{M}}(\Omega) \text{ s. t. } \displaystyle\sum_{K \in \mathscr{M}} |K| p_K = 0, \\ \operatorname{div}_{\mathscr{D}} \boldsymbol{u}(\boldsymbol{x}) = 0, \text{ for a.e. } \boldsymbol{x} \in \Omega. \\ \langle \boldsymbol{u}, \boldsymbol{v} \rangle_{\mathscr{E}} - \displaystyle\int_\Omega p(\boldsymbol{x}) \operatorname{div}_{\mathscr{D}} \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \mathrm{R}\, b_{\mathscr{E}}(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) = \displaystyle\int_\Omega \boldsymbol{f}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \forall \boldsymbol{v} \in H_{\mathscr{E}}(\Omega). \end{cases}$$

With this scheme, a control over the discrete kinetic energy can be obtained, which allows to prove some discrete H^1 estimates on the velocity. Then an L^2 estimate is proved for the discrete pressure, using the standard Necas lifting, which is particularly easy thanks to the staggered grids. The proof of convergence is then completed, considering the interpolation of regular test functions. Details may be found in [2].

4 Numerical example

We consider a problem where the continuous solution of the Navier–Stokes equations (1) with R $=1$ is given by:

$$\bar{u}_1(x_1, x_2) = 2\pi \sin^2(\pi x_1)\cos(\pi x_2)\sin(\pi x_2)$$
$$\bar{u}_2(x_1, x_2) = -2\pi \cos(\pi x_1)\sin(\pi x_1)\sin^2(\pi x_2)$$
$$\bar{p}(x_1, x_2) = \sin^2(\pi x_1)\sin^2(\pi x_2)$$

in a circle with centre $(0, 0)$ and radius 0.45. We consider four meshes for the mass conservation $\mathscr{M}_j$, $j = 0, \ldots, 3$, defined in the following way:

1. a structured square 10×10 is given on the square $[0, 1] \times [0, 1]$,
2. for $i = 0, \ldots, 3$, let us split in 4 control volumes each grid block whose centre (x_1, x_2) satisfies
$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} \geq 0.45 - 0.25/2^i,$$
3. for $i = 0, \ldots, j$, let us split in 4 control volumes each grid block K,
4. get rid of all the control volumes K with centre (x_1, x_2) such that
$$\sqrt{(x_1 - 0.5)^2 + (x_2 - 0.5)^2} > 0.45.$$

Let us denote card$(\mathscr{M}_j)$ the number of control volumes of the mesh $\mathscr{M}_j$. We get that card$(\mathscr{M}_0) = 1604$, card$(\mathscr{M}_1) = 6416$, card$(\mathscr{M}_2) = 25592$ and card$(\mathscr{M}_3) = 102324$. The L^2 errors of unknowns u_1, u_2, p, respectively denoted by $e_2(u_1), e_2(u_2), e_2(p)$, are respectively computed in the Voronoi grids associated to the velocity components and in $\mathscr{M}_j$.

Left part of Fig. (4) shows the errors $\log 10(e_2(u_1))$ and $\log 10(e_2(p))$ with respect to $\log 10(1/\sqrt{\text{card}(\mathscr{M}_j)})$ for $j = 0, \ldots, 3$. On right part of Fig. (4) are

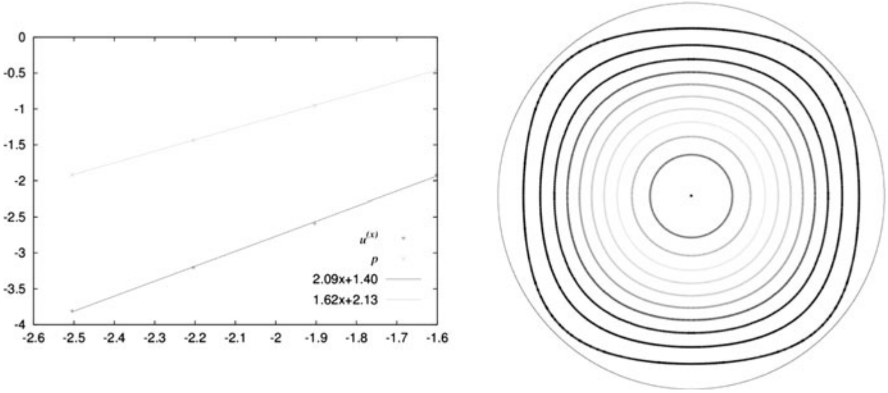


Fig. 4 Left: The L^2 error with respect to the number of control volumes. Right: Stream lines

plotted the stream lines for the finest mesh. The velocity components and the pressure are respectively shown in Figs. (5), (6) and (7). Although the velocity

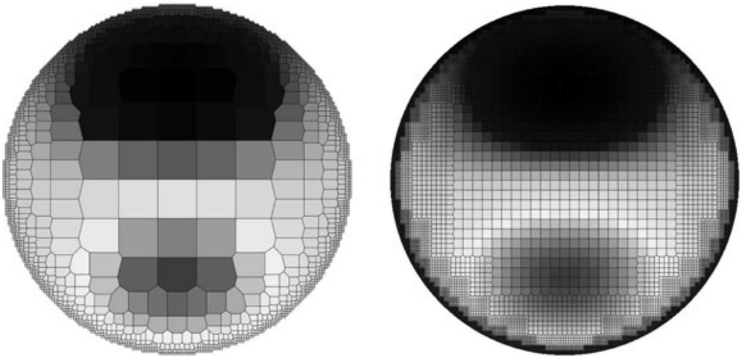


Fig. 5 Horizontal component of the velocity for $j = 0$ and $j = 2$

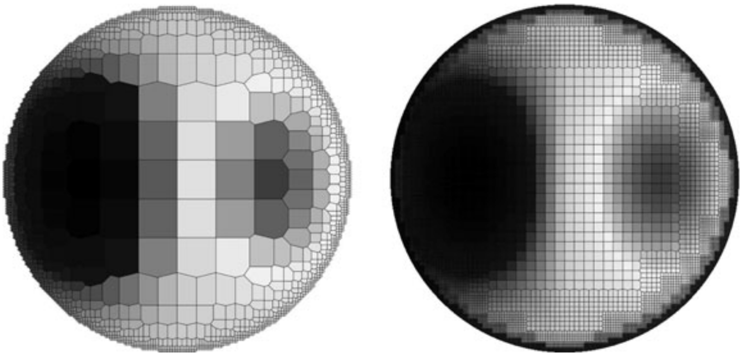


Fig. 6 Vertical component of the velocity for $j = 0$ for $j = 2$

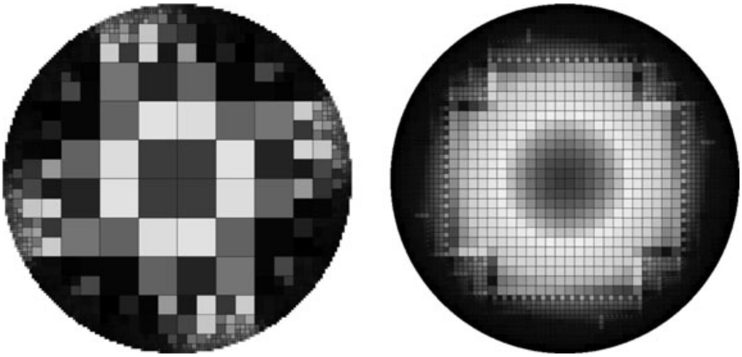


Fig. 7 Pressure for $j = 0$ and $j = 2$

fields are accurately computed on the coarsest mesh, the pressure fields show oscillations where neighbouring control volumes have contrasted sizes. However, these oscillations disappear while refining the mesh.

5 Conclusion

The generalised MAC scheme seems very efficient on meshes which are parallel to the axes. In particular, the scheme keeps a five-point stencil on all non-refined regions. It can also be extended to more general non-structured grids. However for these latter grids, the stencil may become large, which can be a problem when solving the linear systems in the Newton iteration.

References

1. P. Blanc. Convergence of a finite volume scheme on a MAC mesh for the Stokes problem with right-hand side in H^{-1}. In *Finite volumes for complex applications IV*, pages 133–142. ISTE, London, 2005.
2. E. Chénier, R. Eymard and R. Herbin. The MAC scheme on general meshes. in preparation, 2011.
3. V. Girault and J. Lopez. Finite-element error estimates for the MAC scheme., *IMA J. Numer. Anal., 16, 3, 247–379, 1996.*
4. R. Nicolaïdes. Analysis and convergence of the mac scheme i: The linear problem. *SIAM J. Numer. Anal.*, 29:1579–1591, 1992.
5. R. Nicolaïdes and X. Wu. Analysis and convergence of the mac scheme ii, Navier-Stokes equations. *Math. Comp.*, 65:29–44, 1996.
6. D. Shin and J.C. Strikwerda Inf-sup conditions for finite-difference approximations of the Stokes equations. *J. Austral. Math. Soc. Ser. B, 39, 1 121–134, 1997.*
7. S.V. Patankar. *Numerical heat transfer and fluid flow. Series in Computational Methods in Mechanics and Thermal Sciences*, volume XIII. Washington - New York - London: Hemisphere Publishing Corporation; New York. McGraw-Hill Book Company, 1980.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Multi-dimensional Optimal Order Detection (MOOD) — a Very High-Order Finite Volume Scheme for Conservation Laws on Unstructured Meshes

S. Clain, S. Diot, and R. Loubère

Abstract The Multi-dimensional Optimal Order Detection (MOOD) method is an original Very High-Order Finite Volume (FV) method for conservation laws on unstructured meshes. The method is based on an *a posteriori* degree reduction of local polynomial reconstructions on cells where prescribed stability conditions are not fulfilled. Numerical experiments on advection and Euler equations problems are drawn to prove the efficiency and competitiveness of the MOOD method.

Keywords MOOD, high-order, finite volume, unstructured meshes, conservation laws
MSC2010: 65M08, 65Z05, 76M12

1 Introduction

The Multi-dimensional Optimal Order Detection has been introduced in [6] as an original High-Order Finite Volume method for conservation laws on unstructured meshes. As multi-dimensional MUSCL [2–4, 8] or ENO/WENO methods [1, 7, 10], the MOOD method is based on a high-order space discretization with local polynomial reconstructions coupled with a high-order TVD Runge–Kutta method for time discretization.

S. Clain
Departamento de Matemática e Aplicações, Campus de Gualtar - 4710-057 Braga Campus de Azurm - 4800-058 Guimares, Portugal, e-mail: clain@math.uminho.pt

S. Diot, and R. Loubère
Institut de Mathématiques de Toulouse, Université de Toulouse, France,
e-mail: steven.diot@math.univ-toulouse.fr,raphael.loubere@math.univ-toulouse.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_28, © Springer-Verlag Berlin Heidelberg 2011

The main difference between classical high-order methods and the MOOD one is that the limitation procedure is done *a posteriori*. Inside a time step, a first solution is computed with numerical fluxes evaluated from unlimited high-order polynomial reconstructions. Then polynomial degrees are reduced on cells where prescribed stability conditions are not fulfilled and the solution is re-evaluated. That iterative procedure provides a solution which respects the stability constraints.

The present article is devoted to an extension of the MOOD method to a sixth-order space discretization on triangular meshes. Numerical tests for the advection problem and Euler equations with gravity are given in last section.

2 Framework

We consider the scalar hyperbolic equation defined on a bounded polygonal domain $\Omega \subset \mathbb{R}^2$ written in its conservative form

$$\begin{aligned} \partial_t u + \nabla \cdot F(u) &= 0, \\ u(\cdot, 0) &= u_0, \end{aligned} \tag{1}$$

where $u = u(\mathbf{x}, t)$ is the unknown function with $t > 0$, $\mathbf{x} \in \Omega$, F is the physical flux and u_0 stands for the initial condition. We consider a triangular tessellation of Ω where K_i is a generic triangle with centroid $\mathbf{c}_i$. Moreover $\mathbf{n}_{ij}$ is the unit normal vector of edge e_{ij} from K_i to K_j and $q_{ij}^r, r = 1, 2, 3$, are the Gaussian quadrature points of e_{ij}. Finally $\underline{\nu}(i)$ (resp. $\overline{\nu}(i)$) is the index set of cells which share an edge (resp. an edge or a node with K_i). This notation is summarized in Fig. 1.
We recall the generic first-order Finite Volume discretization of (1)

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|e_{ij}|}{|K_i|} G(u_i^n, u_j^n, \mathbf{n}_{ij}), \tag{2}$$

where u_i^n is an approximation of the mean value of u on cell K_i at time t^n and $|e_{ij}|$, $|K_i|$ stand for the edge length and the cell surface respectively. We assume that the numerical flux $G(u_i^n, u_j^n, \mathbf{n}_{ij})$ satisfies the consistency and monotonicity properties such that, under an adequate CFL condition, the following Discrete Maximum Principle (DMP) is fulfilled

$$\min_{j \in \overline{\nu}(i)} (u_i^n, u_j^n) \le u_i^{n+1} \le \max_{j \in \overline{\nu}(i)} (u_i^n, u_j^n). \tag{3}$$

Only few modifications of (2) are needed to get the following High-Order Finite Volume scheme

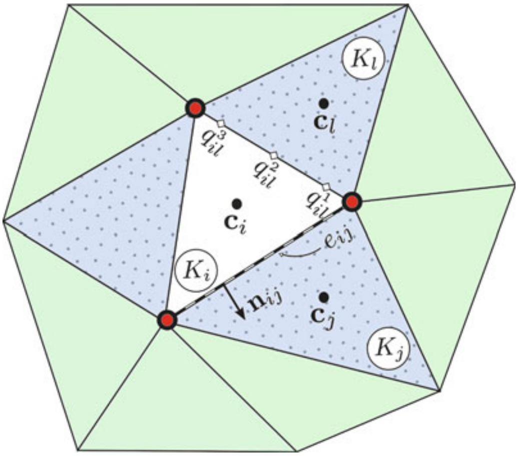
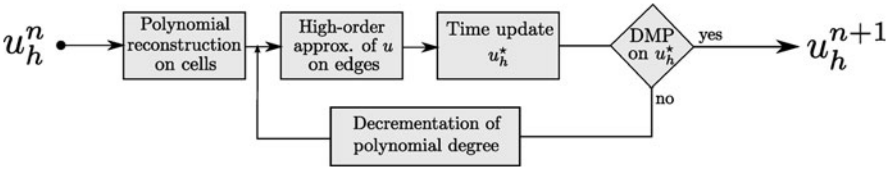


Fig. 1 Mesh notation. Index set $\underline{\nu}(i)$ corresponds to blue cells with dots and $\overline{\nu}(i)$ corresponds to every non-white cells

$$u_i^{n+1} = u_i^n - \Delta t \sum_{j \in \underline{\nu}(i)} \frac{|e_{ij}|}{|K_i|} \sum_{r=1}^{3} \xi_r G(u_{ij,r}^n, u_{ji,r}^n, \mathbf{n}_{ij}), \tag{4}$$

namely the use of a sixth-order Gaussian quadrature rule with weights ξ_r $(r = 1, 2, 3)$ and the replacement of u_i^n (resp. u_j^n) by $u_{ij,r}^n$ (resp. $u_{ji,r}^n$) which is an approximation of $u(q_{ij}^r, t^n)$ from the high-order polynomial reconstruction on K_i (resp. K_j). Notice that the high-order scheme (4) corresponds to a convex combination of the first-order one (2), that is important from a practical point of view for an easy and effective implementation.

It is well known that methods based on high-order reconstructions without limiting procedure produce spurious oscillations in the vicinity of discontinuities. In order to prevent such oscillations, the today's effective high-order methods (MUSCL, WENO...) use *a priori* limitation procedures.
The Multi-dimensional Optimal Order Detection (MOOD) method breaks away from this approach through an original effective iterative procedure based on an *a posteriori* detection of such unphysical oscillations (see Fig. 2). The details of MOOD method are recalled in next section



u_h^n
Polynomial reconstruction on cells
High-order approx. of u on edges
Time update $u_h^\star$
DMP on $u_h^\star$
yes
u_h^{n+1}
no
Decrementation of polynomial degree

Fig. 2 A simplistic view of the Multi-dimensional Optimal Order Detection concept

3 MOOD method

For the sake of clarity, we only consider a forward Euler method and one quadrature point per edge. Consequently we denote by u_{ij} (resp. u_{ji}) the high-order approximation of u on edge e_{ij} from cell K_i (resp. K_j).

3.1 Basics

Polynomial reconstruction.

High-order approximations of the solution at quadrature points are mandatory. To this end, multi-dimensional polynomial reconstructions from mean values are carried out. There exist several techniques [1, 5] to obtain such reconstructions, but we choose to use the one from [7] where a over-determined linear system is solved using a QR decomposition. The reconstructed polynomial of arbitrary high-order $d_{max} + 1$ has the form

$$\widetilde{u}(x, y) = \bar{u} + \sum_{1 \leq \alpha+\beta \leq d_{max}} \mathcal{R}_{\alpha\beta} \left((x - c_x)^\alpha (y - c_y)^\beta - \frac{1}{|K|} \int_K (x - c_x)^\alpha (y - c_y)^\beta \; dxdy \right),$$

where (c_x, c_y) is the centroid of a generic cell K and $\mathcal{R}_{\alpha\beta}$ are the unknowns polynomial coefficients. In this way mean value on K is conserved and the truncation of all terms of degree $\alpha + \beta > \bar{d}$ produces a relevant approximation of u as a polynomial of degree $\bar{d} \leq d_{max}$.
At least $\mathcal{N}(d) = (d + 1)(d + 2)/2 - 1$ neighbors are needed to perform reconstructions. However for the sake of robustness at least $1.5 \times \mathcal{N}(d)$ elements are involved. We first take the neighbors by nodes of K and then the neighbors by faces of already picked elements. Lastly, since the condition number of the generated system is dependent of spatial characteristic length, we use the technique proposed in [5] to overcome this problem.

CellPD and EdgePD.

We recall the fundamental notions introduced in [6].

- d_i is the Cell Polynomial Degree (CellPD) which represents the degree of the polynomial reconstruction on cell K_i.
- d_{ij} and d_{ji} are the Edge Polynomial Degrees (EdgePD) which correspond to the effective degrees used to respectively build u_{ij} and u_{ji} on both sides of edge e_{ij}.

We now detail the MOOD method using both notions in the case of the scalar problem (1).

3.2 Algorithm for the scalar case.

The MOOD method consists of the following iterative procedure which details the concept depicted in Fig. 2.

1. **CellPD initialization.** Each CellPD is initialized with d_{max}.
2. **EdgePD evaluation.** Each EdgePD is set up as the minimum of the two neighboring CellPD.
3. **Quadrature points evaluation.** Each u_{ij} is evaluated with the polynomial reconstruction of degree d_{ij}.
4. **Mean values update.** The updated values $u_h^\star$ are computed using the finite volume scheme (4).
5. **DMP test.** The DMP criterion is checked on each cell K_i

$$\min_{j\in\overline{\nu}(i)} (u_i^n, u_j^n) \le u_i^\star \le \max_{j\in\overline{\nu}(i)} (u_i^n, u_j^n). \tag{5}$$

   If $u_i^\star$ does not satisfy (5) the CellPD is decremented, $\mathsf{d}_i := \max(0, \mathsf{d}_i - 1)$.
6. **Stopping criterion.** If all cells satisfy the DMP property, the iterative procedure stops with $u_h^{n+1} = u_h^\star$ else go to Step 2.

Since only problematic cells and their neighbors in the compact stencil $\underline{\nu}(i)$ have to be checked and re-updated during the iterative MOOD procedure, the computational cost is dramatically reduced.

3.3 Algorithm for the Euler equations case.

We now extend the MOOD method to the Euler system, namely

$$\partial_t \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix} + \partial_x \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho u v \\ u(E+p) \end{pmatrix} + \partial_y \begin{pmatrix} \rho v \\ \rho u v \\ \rho v^2 + p \\ v(E+p) \end{pmatrix} = 0, \tag{6}$$

where ρ, $\mathbf{V} = (u, v)$ and p are the density, velocity and pressure respectively while the total energy per unit volume E is given by

$$E = \rho \left(\frac{1}{2}\mathbf{V}^2 + e \right), \quad \mathbf{V}^2 = u_1^2 + u_2^2, \quad e = \frac{p}{\rho(\gamma - 1)},$$

where e is the specific internal energy and γ the ratio of specific heats.

The reconstruction is classically done on the primitive variables ρ, u, v, p while $U = (\rho, \rho u, \rho v, E)$ and we use the same CellPD and EdgePD for all variables in a cell. In other words, the two notions are linked to cells and edges and not affected by the number of variables. Furthermore steps 5 and 6 of the previous MOOD algorithm are substituted with the following stages.

5. **Density DMP test.** The DMP criterion is checked on the density

$$\min_{j\in\overline{\nu}(i)} (\rho_i^n, \rho_j^n) \leq \rho_i^\star \leq \max_{j\in\overline{\nu}(i)} (\rho_i^n, \rho_j^n). \tag{7}$$

   If $\rho_i^\star$ does not satisfy (7) the CellPD is decremented, $\mathsf{d}_i := \max(0, \mathsf{d}_i - 1)$.
6. **Pressure positivity test.** The pressure positivity is checked and if $p_i^\star \leq 0$ and d_i has not been altered by step 5 then the CellPD is decremented, $\mathsf{d}_i := \max(0, \mathsf{d}_i - 1)$.
7. **Stopping criterion.** If for all $i \in \mathcal{E}_{el}$, d_i has not been altered by steps 5 and 6 then the iterative procedure stops with $U_h^{n+1} = U_h^\star$ else go to step 2.

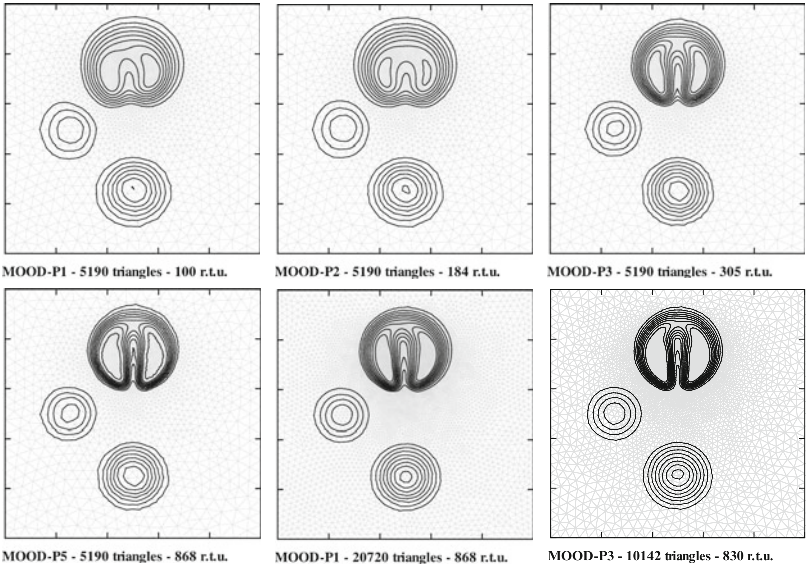
4 Numerical results

The reader should refer to [6] for a study on the effective convergence rate and for more hydrodynamics test cases. In this paper, we restrict the presentation to two representative tests.

Scalar case

We first deal with the classical Solid Body Rotation (see [6] for details) test case for the advection problem. We plot in Fig. 3 isolines top views of the solution obtained with the MOOD method applied to different polynomial degrees and meshes. Method name, triangles number and computational times are embedded in each figure. Time is given in relative time units (r.t.u) where MOOD-P1 is taken as reference with 100 r.t.u.

First solutions obtained on the 5190 cells mesh (3 top) clearly show that the MOOD method is able to handle high-order polynomials with a great improvement of solutions while enforcing a strict DMP. Then for the sake of comparison, results with lower degrees on finer meshes are given in the bottom line of Fig. 3. Finally notice that the computational cost increase is mainly due to the reconstruction step. However since profiles are not smooth the DMP is often violated and the iterative procedure cost more than in a smooth case. For example a sixth-order unlimited version of the scheme costs 586 r.t.u., thus the iterative procedure costs about a third of the total time of the MOOD-P5 computation.


MOOD-P1 - 5190 triangles - 100 r.t.u.

MOOD-P2 - 5190 triangles - 184 r.t.u.

MOOD-P3 - 5190 triangles - 305 r.t.u.

MOOD-P5 - 5190 triangles - 868 r.t.u.

MOOD-P1 - 20720 triangles - 868 r.t.u.

MOOD-P3 - 10142 triangles - 830 r.t.u.

Fig. 3 Solid Body Rotation. 10 isolines (0 to 1). Time in relative time units (r.t.u.)

Euler equations case

For the system case, a Rayleigh–Taylor Instability for the Euler equations with gravity is considered. The reader should refer to [9] for complete description of the test case. A zoom on the pattern of the unstructured symmetric triangular mesh of 28800 cells and the density solutions for MOOD-P1, MOOD-P3 and MOOD-P5 are plotted in Fig. 4.

As for the scalar case the MOOD method is plainly able to improve the solution through the use of high-order polynomial reconstructions. From a computational cost point of view, computational times given in Fig. 4 prove that the MOOD iterative procedure is effective since the time raise from a degree to a bigger one is mainly due to the reconstruction cost itself.

Decrementation procedure

In Table 1, we give the mean percentage over all the calculation of polynomial degrees actually used to compute the solution, *i.e.* the CellPD at the end of the iterative procedure. Three test cases are taken as examples (see [6] for details), the Solid Body Rotation of Fig. 3 with MOOD-P3, the classical Double Mach Reflection on a 57600 cells uniform mesh with MOOD-P2 and the Mach 3 Wind Tunnel on a 4978 cells Delaunay mesh with MOOD-P3. Results show that only few cells are affected by the *a posteriori* limitation.

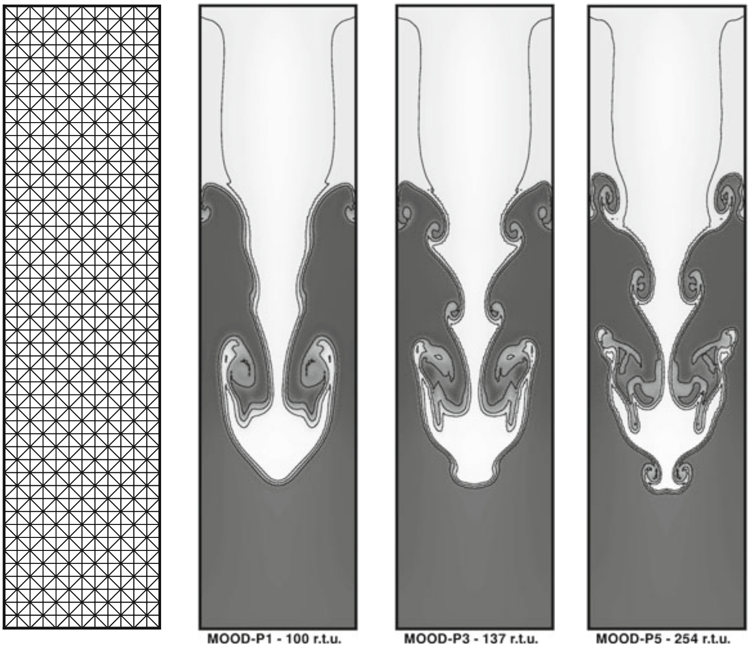


Fig. 4 Rayleigh–Taylor Instability. Density. 5 isolines from 0.8 (dark) to 2.3 (light)

Test case	P0	P1	P2	P3
Solid Body Rotation	7.16%	0.78%	0.64%	91.42%
Double Mach Reflection	5.69%	0.72%	93.69%	—
Mach 3 Wind Tunnel	3.02%	0.36%	0.16%	96.46%

Fig. 5 Mean percentage of polynomial degrees actually used with MOOD method

References

1. R. Abgrall, On Essentially Non-oscillatory Schemes on Unstructured Meshes: Analysis and Implementation, J. Comput. Phys. **114** 45–58 (1994)
2. T. J. Barth, Numerical methods for conservation laws on structured and unstructured meshes, VKI March 2003 Lectures Series
3. T. J. Barth, D. C. Jespersen, The design and application of upwind schemes on unstructured meshes, AIAA Report 89-0366 (1989)
4. T. Buffard, S. Clain, Monoslope and Multislope MUSCL Methods for unstructured meshes, J. Comput. Phys. **229** 3745-3776 (2010)
5. O. Friedrich, Weighted Essentially Non-Oscillatory Schemes for the Interpolation of Mean Values on Unstructured Grids, J. Comput. Phys. 144 (1998) 194–212.
6. S. Clain, S. Diot, R. Loubère A high-order finite volume method for systems of conservation laws — Multi-dimensional Optimal Order Detection (MOOD), accepted in J. Comput. Phys. (2011)

7. C. F. Ollivier-Gooch, Quasi-ENO Schemes for Unstructured Meshes Based on Unlimited Data-Dependent Least-Squares Reconstruction, J. Comput. Phys. **133** 6–17 (1997)
8. J. S. Park, S.-H. Yoon, C. Kim, Multi-dimensional limiting process for hyperbolic conservation laws on unstructured grids, J. Comput. Phys. **229** 788–812 (2010)
9. J. Shi, Y-T Zhang, C-W Shu, Resolution of high order WENO schemes for complicated flow structures, J. Comput. Phys. **186** 690–696 (2003)
10. W. R. Wolf , J. L. F. Azevedo, High-order ENO and WENO schemes for unstructured grids, International Journal for Numerical Methods in Fluids, **55** Issue 10 917—943 (2007)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Relaxation Approach for Simulating Fluid Flows in a Nozzle

Frédéric Coquel, Khaled Saleh, and Nicolas Seguin

Abstract We present here a Godunov-type scheme to simulate one-dimensional flows in a nozzle with variable cross-section. The method relies on the construction of a relaxation Riemann solver designed to handle all types of flow regimes, from subsonic to supersonic flows, as well as resonant transonic flows. Some computational results are also provided, in which this relaxation method is compared with the classical Rusanov scheme and a modified Rusanov scheme.

Keywords Relaxation scheme, Godunov-type scheme, resonant transonic flows.
MSC2010: 76M12, 76H05, 76S05, 65M12

1 Introduction

In this paper, we are interested in the numerical approximation of the solutions of a model describing one-dimensional barotropic flows in a nozzle. In this model, ρ and w are respectively the density and the velocity of the fluid while α stands for the cross-section of the nozzle, which is assumed to be constant in time. Under the classical assumption that α is small with respect to a characteristic length in the

Frédéric Coquel
CMAP, UMR 7641, Ecole Polytechnique, route de Saclay, F-91128 Palaiseau,
e-mail: frederic.coquel@cmap.polytechnique.fr

Khaled Saleh
EDF R&D, MFEE, 6 Quai Watier, F-78400 Chatou
and
UPMC & CNRS, UMR 7598, LJLL, F-75005 Paris, e-mail: khaled.saleh@edf.fr,
saleh@ann.jussieu.fr

Nicolas Seguin
UPMC & CNRS, UMR 7598, LJLL, F-75005 Paris, e-mail: seguin@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_29, © Springer-Verlag Berlin Heidelberg 2011

mainstream direction, the flow can be supposed to be one-dimensional and described by the following set of partial differential equations:

$$\begin{aligned} &\partial_t \alpha = 0, \\ &\partial_t(\alpha\rho) + \partial_x(\alpha\rho w) = 0, \qquad\qquad t > 0,\ x \in \mathbb{R}, \\ &\partial_t(\alpha\rho w) + \partial_x(\alpha\rho w^2 + \alpha p(\tau)) - p(\tau)\partial_x \alpha = 0, \end{aligned} \tag{1}$$

where $\tau = \rho^{-1}$ is the specific volume and $\tau \mapsto p(\tau)$ is a barotropic pressure law (satisfying $p'(\tau) < 0$ and $p''(\tau) > 0$). System (1) takes the condensed form:

$$\partial_t \mathbb{U} + \partial_x \mathbf{f}(\mathbb{U}) + \mathbf{c}(\mathbb{U})\partial_x \mathbb{U} = 0, \tag{2}$$

where the state vector is $\mathbb{U} = (\alpha, \alpha\rho, \alpha\rho w)^T$. The solutions are sought in the phase space of positive solutions defined as

$$\Omega = \{\mathbb{U} = (\alpha, \alpha\rho, \alpha\rho w)^T \in \mathbb{R}^3, \alpha > 0, \alpha\rho > 0\}. \tag{3}$$

We recall the properties of this model:

- **Property 1.1 (Hyperbolicity)** *System (1) admits, for $\mathbb{U}$ in Ω, the following eigenvalues*

$$\lambda_0(\mathbb{U}) = 0, \qquad \lambda_1(\mathbb{U}) = w - c(\tau), \qquad \lambda_2(\mathbb{U}) = w + c(\tau), \tag{4}$$

  where $c(\tau) = \tau\sqrt{-p'(\tau)}$. The system is hyperbolic (i.e. the corresponding eigenvectors span $\mathbb{R}^3$) if and only if $|w| \neq c(\tau)$. Besides, the fields associated with the λ_1 and λ_2 eigenvalues are genuinely non-linear while the field associated with λ_0 is linearly degenerate.
- **Property 1.2 (Entropy)** *The entropy solutions of system (1) satisfy the following inequality in the weak sense*

$$\partial_t\left(\alpha\rho\mathscr{E}\right) + \partial_x\left(\alpha\rho\mathscr{E} w + \alpha p(\tau) w\right) \leq 0 \tag{5}$$

  where $\mathscr{E} = \frac{w^2}{2} + e(\tau)$ is the total energy and where the function $\tau \mapsto e(\tau)$ is given by $e'(\tau) = -p(\tau)$.

The Godunov scheme for this model is difficult to implement because the Riemann problem for system (1) is hard to solve due to the non linearities of the pressure law (giving rise to the genuinely non-linear acoustic fields), to the absence of a satisfactory definition of the non-conservative product $p(\tau)\partial_x\alpha$ and to the resonance phenomenon that appears for transonic flows causing the model to lose hyperbolicity [5]. For these reasons, we rather follow the classical approach of [7] and design an approximate Riemann solver, relying on a relaxation method. With this end in view, the solutions of system (1) are approximated by the solutions of the following enlarged relaxation system in the limit of a vanishing positive parameter ε:

$$\begin{aligned}
&\partial_t \alpha^\varepsilon = 0, \\
&\partial_t (\alpha\rho)^\varepsilon + \partial_x (\alpha\rho w)^\varepsilon = 0, \qquad\qquad t > 0,\ x \in \mathbb{R}, \\
&\partial_t (\alpha\rho w)^\varepsilon + \partial_x (\alpha\rho w^2 + \alpha\pi(\tau, \mathscr{T}))^\varepsilon - \pi(\tau, \mathscr{T})^\varepsilon \partial_x \alpha^\varepsilon = 0, \\
&\partial_t (\alpha\rho\mathscr{T})^\varepsilon + \partial_x (\alpha\rho\mathscr{T} w)^\varepsilon = \frac{1}{\varepsilon}(\alpha\rho)^\varepsilon(\tau - \mathscr{T})^\varepsilon,
\end{aligned} \tag{6}$$

with a linearization of the pressure law given by $\pi(\tau, \mathscr{T}) = p(\mathscr{T}) + a^2(\mathscr{T} - \tau)$. The variable $\mathscr{T}$ is an additionnal unknown relaxing towards the specific volume τ in the limit $\varepsilon \searrow 0$, and the constant a is a numerical parameter that must be taken large enough so as to guarantee the non-linear stability of the numerical approximation. The state vector for the relaxation system is $\mathbb{W} = (\alpha, \alpha\rho, \alpha\rho w, \alpha\rho\mathscr{T})^T$ and the solutions are sought in the phase space

$$\Omega^r = \{\mathbb{W} = (\alpha, \alpha\rho, \alpha\rho w, \alpha\rho\mathscr{T})^T \in \mathbb{R}^4, \alpha > 0, \alpha\rho > 0, \alpha\rho\mathscr{T} > 0\}. \tag{7}$$

The following property motivates the introduction of this relaxation system

Property 1.3 (Hyperbolicity) *The convective part of (6) admits, for $\mathbb{W}$ in Ω^r, the following eigenvalues*

$$\sigma_0(\mathbb{W}) = 0, \qquad \sigma_1(\mathbb{W}) = w - a\tau, \qquad \sigma_2(\mathbb{W}) = w, \qquad \sigma_3(\mathbb{W}) = w + a\tau. \tag{8}$$

The system is hyperbolic (i.e. the corresponding eigenvectors span $\mathbb{R}^4$) if and only if $|w| \neq a\tau$, and all the fields are linearly degenerate.

2 The Riemann problem for the relaxation system

In this section, we give the main ideas leading to the construction of solutions to the Riemann problem for the convective part of the relaxation system (6). Being given $\mathbb{W}_L$ and $\mathbb{W}_R$ two states in Ω^r, we look for solutions of

$$\begin{cases} \partial_t \mathbb{W} + \partial_x \mathbf{g}(\mathbb{W}) + \mathbf{d}(\mathbb{W})\partial_x \mathbb{W} = 0, \\ \mathbb{W}(x, 0) = \mathbb{W}_L \quad \text{if} \quad x < 0 \quad \text{and} \quad \mathbb{W}_R \quad \text{if} \quad x > 0. \end{cases} \tag{9}$$

As all the characteristic fields are linearly degenerate, the solution turns out to be simpler to construct than a solution of the Riemann problem for the equilibrium system (1). Indeed, the solution is sought in the form of a self-similar function consisting in constant intermediate states separated by contact discontinuities. The linear degeneracy of the fields provides natural jump relations across each discontinuity and yields a set of equations eventually leading to the expessions of the wave speeds and intermediate states. However, some issues related to the resonance phenomenon still need to be handled with care (see [2] for details).

We show that the solutions can be expressed in terms of the physical data $\mathbb{V}_L = (\rho_L, w_L, \mathscr{T}_L)$ and $\mathbb{V}_R = (\rho_R, w_R, \mathscr{T}_R)$ (*i.e.* all the initial data excluding the cross-section α) and of the ratio of left and right initial sections $\nu := \frac{\alpha_L}{\alpha_R}$. More precisely, we introduce the following quantities depending only on $(\mathbb{V}_L, \mathbb{V}_R)$

$$w^\sharp := \frac{1}{2}(w_L + w_R) - \frac{1}{2a}(\pi_R - \pi_L), \tag{10}$$

$$\tau_L^\sharp := \tau_L + \frac{1}{a}(w^\sharp - w_L) = \tau_L + \frac{1}{2a}(w_R - w_L) - \frac{1}{2a^2}(\pi_R - \pi_L), \tag{11}$$

$$\tau_R^\sharp := \tau_R - \frac{1}{a}(w^\sharp - w_R) = \tau_R + \frac{1}{2a}(w_R - w_L) + \frac{1}{2a^2}(\pi_R - \pi_L), \tag{12}$$

where $w^\sharp$ has the dimension of a speed and $\tau_L^\sharp$, $\tau_R^\sharp$ the dimension of specific volumes. These quantities appear in the explicit expressions of the solutions and it can be proved that these specific volumes need to be positive in order to guarantee the positivity of the solutions. In the numerical applications however, a will be chosen large for stability matters (see Sect. 4) and it will always be possible to impose the positivity of $\tau_L^\sharp$ and $\tau_R^\sharp$ by taking a large enough.

The main result of this section is the existence theorem for the Riemann problem.

Theorem 2.1 *Let $\mathbb{W}_L$ and $\mathbb{W}_R$ be two positive states in Ω^r. Assume that a is such that $\tau_L^\sharp > 0$ and $\tau_R^\sharp > 0$. Then the Riemann problem (9) admits a positive self-similar solution whatever the ratio $\nu = \frac{\alpha_L}{\alpha_R}$ is.*

Sketch of the proof *(see [2] for details).* The proof consists in the effective construction of a solution. For the relaxation system, the eigenvalues are not naturally ordered because of the existence of a standing wave, and a resonance phenomenon does appear for transonic flows. Therefore, in order to construct solutions, we investigate all admissible wave configurations (including sonic and supersonic ones) and for each admissible ordering of the eigenvalues, we determine sufficient conditions on the initial states $\mathbb{W}_L$ and $\mathbb{W}_R$ for the solution to have this particular ordering. Eventually, we check *a posteriori* that the determined conditions totally cover the whole space of initial conditions $\Omega^r \times \Omega^r$. □

Figure 1 represents the map of the admissible solutions given by Theorem 2.1 with respect to the initial states $\mathbb{W}_L$ and $\mathbb{W}_R$. The right part of the chart corresponds to the solutions with positive material speed, while the left part depicts the symmetric configurations with negative material speed.

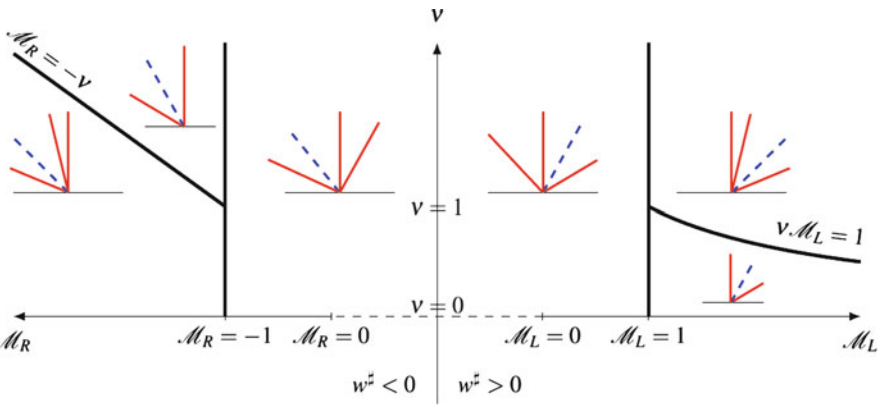


Fig. 1 Wave configuration of the solution of the Riemann problem (9) with respect to $\mathbb{W}_L$ and $\mathbb{W}_R$. $\mathcal{M}_L = \frac{w_L}{a\tau_L}$ and $\mathcal{M}_R = \frac{w_R}{a\tau_R}$ are the Mach numbers of the initial left and right states $\mathbb{W}_L$ and $\mathbb{W}_R$. The material wave is represented by a dashed line

3 Numerical approximation

In this section, we derive a numerical scheme from the relaxation approximation introduced in Sect. 1, the aim being to approximate the weak solutions of a Cauchy problem associated with system (1):

$$\begin{cases} \partial_t \mathbb{U} + \partial_x \mathbf{f}(\mathbb{U}) + \mathbf{c}(\mathbb{U})\partial_x \mathbb{U} = 0, \\ \mathbb{U}(x,0) = \mathbb{U}_0(x). \end{cases} \tag{13}$$

Let Δx be a space step and Δt a time step. The space is partitioned into cells $\mathbb{R} = \bigcup_{j\in\mathbb{Z}} C_j$ with $C_j = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}[$, where $x_{j+\frac{1}{2}} = (j + \frac{1}{2})\Delta x$ are the cell interfaces. At the discrete times $t^n = n\Delta t$, the solution of (13) is approximated on each cell C_j by a constant value denoted by $\mathbb{U}_j^n = \left(\alpha_j^n, (\alpha\rho)_j^n, (\alpha\rho w)_j^n\right)^T$. We now describe the two-step splitting method associated with the relaxation system (6) in order to calculate the values of the approximate solution at time t^{n+1} $(\mathbb{U}_j^{n+1})_{j\in\mathbb{Z}}$ from those at time t^n.

Step 1: Time evolution $(t^n \to t^{n+1,-})$

We first introduce the piecewise constant approximate solution of the relaxation system at time t^n: $x \mapsto \mathbb{W}(x,t^n) = \mathbb{W}_j^n$ in C_j with $\mathbb{W}_j^n = \left(\alpha_j^n, (\alpha\rho)_j^n, (\alpha\rho w)_j^n, (\alpha\rho\mathscr{T})_j^n\right)$, where $\mathscr{T}_j^n := \tau_j^n$, *i.e.* $\mathbb{W}_j^n$ is at equilibrium. Then, the following Cauchy problem is **exactly solved** for $t \in [0, \Delta t]$ with Δt small enough (see condition (15) below)

$$\begin{cases} \partial_t \widetilde{\mathbb{W}} + \partial_x \mathbf{g}(\widetilde{\mathbb{W}}) + \mathbf{d}(\widetilde{\mathbb{W}})\partial_x \widetilde{\mathbb{W}} = 0, \\ \widetilde{\mathbb{W}}(x,0) = \mathbb{W}(x,t^n). \end{cases} \tag{14}$$

Since the initial condition $x \mapsto \mathbb{W}(x,t^n)$ is piecewise constant, the exact solution of (14) is obtained by gluing together the solutions of the Riemann problems set at each cell interface $x_{j+\frac{1}{2}}$, provided that these solutions do not interact during the period Δt, *i.e.* provided the following classical CFL condition

$$\frac{\Delta t}{\Delta x} \max_{\mathbb{W}} |\sigma_i(\mathbb{W})| < \frac{1}{2},\ i \in \{0, ..., 3\}, \tag{15}$$

for all $\mathbb{W}$ under consideration. More precisely, if (x,t) is in $[x_j, x_{j+1}] \times [0, \Delta t]$, then

$$\widetilde{\mathbb{W}}(x,t) = \mathbb{W}_r \left(\frac{x - x_{j+1/2}}{t}; a_{j+1/2}, \mathbb{W}_j^n, \mathbb{W}_{j+1}^n \right), \tag{16}$$

where $(x,t) \mapsto \mathbb{W}_r \left(\frac{x}{t}; a, \mathbb{W}_L, \mathbb{W}_R\right)$ is the self-similar solution of the Riemann problem constructed in Sect. 1, which clearly depends on the local choice of the parameter a. Then, in order to define a piecewise constant approximate solution at time $t^{n+1,-}$, the solution $\widetilde{\mathbb{W}}(x,t)$ is averaged on each cell C_j at time Δt:

$$\mathbb{W}(x, t^{n+1,-}) = \mathbb{W}_j^{n+1,-} := \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \widetilde{\mathbb{W}}(x, \Delta t) dx, \quad \forall x \in C_j, \quad \forall j \in \mathbb{Z}. \tag{17}$$

Step 2: Instantaneous relaxation ($t^{n+1,-} \to t^{n+1}$)

The second step consists in sending ε to zero instantaneously in the piecewise constant function $\mathbb{W}(x, t^{n+1,-})$ obtained at the end of the first step. This amounts to imposing $\mathscr{T}_j^{n+1} := \tau_j^{n+1}$, thus we have

$$\mathbb{W}_j^{n+1} = \left(\alpha_j^{n+1,-}, (\alpha\rho)_j^{n+1,-}, (\alpha\rho w)_j^{n+1,-}, \alpha_j^{n+1,-} \right)^T. \tag{18}$$

Finally, the new cell value at time t^{n+1} of the approximate solution reads

$$\mathbb{U}_j^{n+1} = \left(\alpha_j^{n+1,-}, (\alpha\rho)_j^{n+1,-}, (\alpha\rho w)_j^{n+1,-} \right)^T. \tag{19}$$

We can prove that this two-step relaxation method can be equivalently rewritten in the form of a Godunov-type finite volume scheme [7].

4 Non-linear stability of the scheme

Non-linear stability issues are usually dealt with through a so-called *discrete entropy inequality*, which is the discrete counterpart of the entropy inequality (5) satisfied by the weak solutions of the model. We have the following definition:

Definition 4.1 *We say that a numerical scheme satisfies a discrete entropy inequality if there exists a numerical entropy flux $G(\mathbb{U}_L, \mathbb{U}_R)$ which is consistent with the exact entropy flux $\mathscr{G} = \alpha\rho\mathscr{E}w + \alpha p(\tau)w$ (in the sense that $G(\mathbb{U}, \mathbb{U}) = \mathscr{G}(\mathbb{U})$ for all $\mathbb{U}$) such that, under some CFL condition, the discrete values $(\mathbb{U}_j^n)_{j\in\mathbb{Z},n\in\mathbb{N}}$ computed by the scheme automatically satisfy*

$$(\alpha\rho\mathscr{E})(\mathbb{U}_j^{n+1}) - (\alpha\rho\mathscr{E})(\mathbb{U}_j^n) + \frac{\Delta t}{\Delta x}(G(\mathbb{U}_j^n, \mathbb{U}_{j+1}^n) - G(\mathbb{U}_{j-1}^n, \mathbb{U}_j^n)) \leq 0. \qquad (20)$$

As seen in Sect. 3, under the CFL condition (15), the different Riemann problems at each interface do not interact and the parameter $a = a_{j+\frac{1}{2}}$ can be chosen locally interface by interface. Usually, if $a_{j+\frac{1}{2}}$ is large enough, so as to satisfy a so-called Whitham condition (see [1]), then a discrete entropy inequality (20) is guaranteed. In order to define $a_{j+\frac{1}{2}}$, we propose a weak Whitham-like condition that handles the resonance phenomenon and still guarantees a discrete entropy inequality under the CFL condition (15) (see [2] for details).

5 Numerical tests

In this section, we run the relaxation scheme described in Sect. 3 on a Riemann problem that contains the standing wave associated with the constant cross-section α, a left-going λ_1-rarefaction wave, a sonic right-going λ_1-rarefaction wave and a right-going λ_2-shock. The chosen pressure law is an ideal gas barotropic pressure law $p(\tau) = \tau^{-\gamma}$, with $\gamma = 3$. The left and right initial conditions are given by $\alpha_L = 3.0$, $\rho_L = 1.0$, $w_L = 0$, $\alpha_R = 1.0$, $\rho_R = 0.1$, and $w_R = 0$. The outcome of the relaxation method is compared with two other numerical schemes. The first one is the classical Rusanov scheme where the cross-section α is preserved throughout time:

$$\alpha_j^{n+1} := \alpha_j^n. \qquad (21)$$

The second one is a modification of the Rusanov scheme that consists in applying the scheme to the whole state vector $\mathbb{U}$ (including the cross-section α) causing α to be dissipated:

$$\alpha_j^{n+1} := \alpha_j^n - \frac{\Delta t}{\Delta x}\left(q_{j+\frac{1}{2}}^n - q_{j-\frac{1}{2}}^n\right), \qquad (22)$$

with $q_{j+\frac{1}{2}}^n = -r(\mathbb{U}_j^n)(\alpha_{j+1}^n - \alpha_j^n)$ where the scalar $r(\mathbb{U}_j^n)$ is the maximal value of the spectral radius of the Jacobian matrices $(\nabla\mathbf{f} + \mathbf{c})(\mathbb{U}_k^n)$ for $k = j, j+1$.

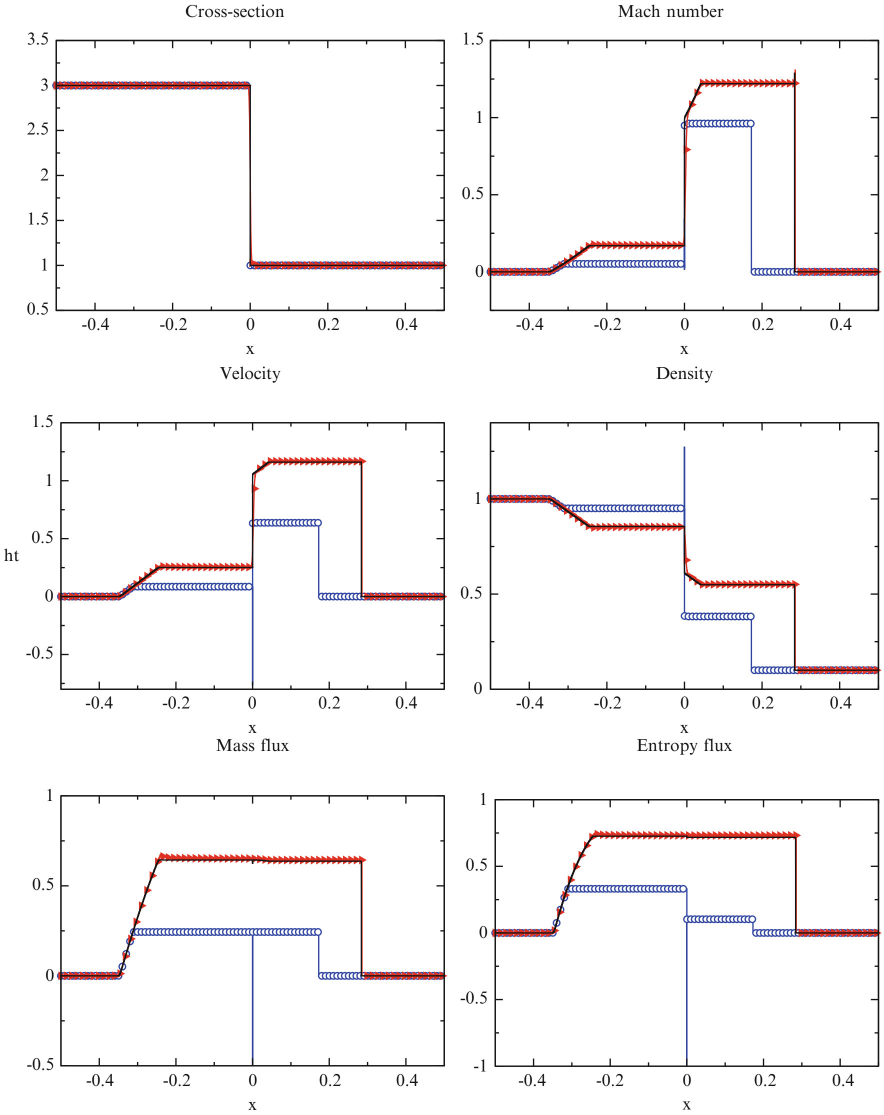


Fig. 2 Solution of the Riemann problem at time $T = 0.2$. Space step $\Delta x = 10^{-5}$. Straight line: relaxation scheme, circles: classical Rusanov scheme, triangles: Rusanov scheme with dissipation of the cross-section

In Fig. 2, we can see that, due to a smoothing effect, the dissipation of the cross-section α provides a notable improvement for the Rusanov scheme (see [4] and [8] for different approaches to improve the Rusanov scheme). The L^1-norm of the error on α, at the final time T, vanishes as the space step Δx goes to zero (with $\Delta t/\Delta x$ constant) with the order $\mathcal{O}(\Delta x^{1/2})$.

Acknowledgements The second author receives a financial support by ANRT through an EDF-CIFRE contract 529/2009. Computational facilities were provided by EDF. The third author is partially supported by the LRC Manon (Modélisation et Approximation Numérique Orientées pour l'énergie Nucléaire – CEA/DM2S-LJLL).

References

1. F. Bouchut. *Nonlinear Stability of Finite Volume Methods for Hyperbolic Conservation Laws*. Birkhauser. Frontiers in Mathematics. 2004.
2. F. Coquel, K. Saleh, N. Seguin. Relaxation and numerical approximation for fluid flows in a nozzle. *Preprint to be published.*
3. C.M. Dafermos. *Hyperbolic Conservation Laws in Continuum Physics*. Springer-Verlag. Grundlehren der mathematischen Wissenschaften. **Vol 325**. 2000.
4. L. Girault, J-M. Hérard. A two-fluid hyperbolic model in a porous medium. *M2AN*, **Vol 44(6)**, pp 1319-1348, 2010.
5. P. Goatin, P.G. LeFloch. The Riemann problem for a class of resonant hyperbolic systems of balance laws. *Ann. Inst. H. Poincaré Anal. Non Linéaire 21*, **no 6**, pp 881-902, 2004.
6. E. Godlewski, P-A. Raviart. *Numerical Approximation of Hyperbolic Systems ofConservation Laws*. Springer-Verlag. Applied Mathematical Sciences. **Vol 118**. 1996.
7. A. Harten, P.D. Lax & B. Van Leer. On upstream differencing and Godunov-type schemes for hyperbolic conservation laws. *Comm. Math. Sci.* **Vol 1**. pp 763-796. 2003.
8. D. Kröner, M.D. Thanh. Numerical solution to compressible flows in a nozzle with variable cross-section. *SIAM J. Numer. Anal.*, **Vol 43(2)**, pp 796-824, 2006.
9. P.G. LeFloch, M.D. Thanh. The Riemann problem for fluid flows in a nozzle with discontinuous cross-section. *Comm. Math. Sci.* **Vol 1**, pp 763-796, 2003.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A CeVeFE DDFV scheme for discontinuous anisotropic permeability tensors

Yves Coudière, Florence Hubert, and Gianmarco Manzini

Abstract In this work we derive a formulation for discontinuous diffusion tensor for the Discrete Duality Finite Volume (DDFV) framework that is exact for affine solutions. In fact, DDFV methods can naturally handle anisotropic or non-linear problems on general distorded meshes. Nonetheless, a special treatment is required when the diffusion tensor is discontinuous across an internal interfaces shared by two control volumes of the mesh. In such a case, two different gradients are considered in the two subdiamonds centered at that interface and the flux conservation is imposed through an auxiliary variable at the interface.

Keywords Finite volume schemes, Darcy flow
MSC2010: 65N08, 76S05

1 Introduction

In this proceeding we propose a Discrete Duality Finite Volume (DDFV) method that can handle *discontinuous* permeability coefficients. This method is a variant of the DDFV formulation proposed by Y. Coudière and F. Hubert in [6] to extend to three-dimensional (3D) problems the original two-dimensional finite volume schemes by F. Hermeline [11] and K. Domelevo and P. Omnès [9]. In the DDFV approach the diffusive flux is approximated using a piecewise constant

Yves Coudière
Laboratoire Jean Leray, Nantes, FRANCE, e-mail: Yves.Coudiere@univ-nantes.fr

Florence Hubert
LATP, Université de Provence, Marseille, FRANCE, e-mail: fhubert@cmi.univ-mrs.fr

Gianmarco Manzini
IMATI and CESNA-IUSS, Pavia, ITALY, e-mail: gm.manzini@gmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_30, © Springer-Verlag Berlin Heidelberg 2011

approximation of the solution gradient over a set of edge-based cells called *diamond cells*. In the two dimensional formulation, the gradient is approximated by a formula that requires the vertex values of the scalar solution. Following the DDFV approach, such vertex values are the solution of another finite volume method whose control volumes are built around the vertices. Therefore, the resulting scheme combines two distinct finite volume methods for the cell unknowns and the vertex unknowns on two overlapping meshes. Effectiveness and efficiency of such coupled finite volume formulation are documented in [5, 10].

Several generalizations of the two-dimensional DDFV formulation have been proposed in the literature; it is worth mentioning the works by F. Hermeline in [12], C. Pierre in [8, 13], and B. Andreianov and collaborators in [1–4]. Here, we consider the alternative construction proposed in [6], which uses two families of additional unknowns. In the first family, the unknowns are located at the vertices of the mesh and are the solution of a finite volume method whose control volumes are built around the vertices. In the second family, the unknowns are located at the centers of mesh edges and faces and are the solution of a finite volume method whose control volumes are built around such geometric objects. Therefore, the resulting scheme couples three distinct finite volume methods through a 3D gradient formula that generalizes the 2D one on a set of special cells, the so called *diamond cells*, built around edges and faces as will be discussed in the next sub-section.

The outline of the paper is as follows. In Sect. 2 we present a short review of the DDFV method. In Sect. 3 we present the numerical treatment that we propose for the case of discontinuous permeabilities. In Sect. 6 we offer final remarks and conclusions.

2 The Discrete Duality Finite volume formulation

Meshes

Given a general finite volume mesh $\mathscr{M}$ of the computational domain Ω, composed of polyhedra, three additional polyhedral partitions of Ω are built, denoted by $\mathscr{N}$, $\mathscr{F}\mathscr{E}$ and $\mathscr{D}$, hereafter described.

We denote the control volumes of the initial mesh $\mathscr{M}$ by K or L. The set $\partial\mathscr{M}$ gathers the boundary faces, which we consider as degenerated control volumes, and we complete the initial mesh as $\overline{\mathscr{M}} = \mathscr{M} \cup \partial\mathscr{M}$. We associate a set of points $x_{\mathrm{K}} \in \mathrm{K}$ with the control volumes in $\overline{\mathscr{M}}$; specifically, in the current applications we use the arithmetic average of the vertex position vectors for each polyhedral cell. We denote the vertices, the edges, and the faces of mesh $\mathscr{M}$ by x_{A}, E and F, respectively, and we define some additional points: the center of gravity x_{F} of each face F and the midpoint x_{E} of each edge E. These points are ordered following the relation

$$x_{\mathrm{A}} \prec x_{\mathrm{E}} \prec x_{\mathrm{F}} \prec x_{\mathrm{K}} \quad \text{which means that} \quad x_{\mathrm{A}} \subset \partial\mathrm{E}, \quad \mathrm{E} \subset \partial\mathrm{F}, \quad \mathrm{F} \subset \partial\mathrm{K}.$$

The 3D gradient formula that we will introduce in the next subsection provides a piecewise constant approximation of the solution gradient on the mesh $\mathscr{D}$, which is the set of *diamond cells* D. To each one of the pairs "(edge, face)" (E,F) related by $x_{\mathrm{E}} \prec x_{\mathrm{F}}$ there corresponds a different diamond cell D that we define as follows. Cell D is the convex polyhedra with vertices $x_{\mathrm{A}}, x_{\mathrm{B}}, x_{\mathrm{E}}, x_{\mathrm{F}}, x_{\mathrm{K}}, x_{\mathrm{L}}$, where x_{A} and x_{B} denote the endpoints of E, K and L the two cells sharing the common face F. Specifically, it holds that $\mathrm{D} = \mathrm{hull}(x_{\mathrm{A}}, x_{\mathrm{F}}, x_{\mathrm{B}}, x_{\mathrm{K}}) \cup \mathrm{hull}(x_{\mathrm{A}}, x_{\mathrm{F}}, x_{\mathrm{B}}, x_{\mathrm{L}})$. We associate with each diamond cell D the point $x_{\mathrm{D}} = \frac{1}{2}(x_{\mathrm{E}} + x_{\mathrm{F}}) \in \mathrm{D}$.

We partition each diamond cell into eight tetrahedra sharing x_{D} as common vertex and having the remaining three vertices chosen within the pairs $(x_{\mathrm{A}}, x_{\mathrm{B}})$, $(x_{\mathrm{E}}, x_{\mathrm{F}})$ and $(x_{\mathrm{K}}, x_{\mathrm{L}})$, respectively. Formally, we denote the eight possible combinations by

$$\mathrm{D} = \mathrm{hull}\left(x_{\mathrm{D}}, \begin{pmatrix} x_{\mathrm{A}} \\ x_{\mathrm{B}} \end{pmatrix}, \begin{pmatrix} x_{\mathrm{E}} \\ x_{\mathrm{F}} \end{pmatrix}, \begin{pmatrix} x_{\mathrm{K}} \\ x_{\mathrm{L}} \end{pmatrix} \right), \quad \text{with} \quad \begin{pmatrix} x_{\mathrm{A}} \\ x_{\mathrm{B}} \end{pmatrix} \prec x_{\mathrm{E}} \prec x_{\mathrm{F}} \prec \begin{pmatrix} x_{\mathrm{K}} \\ x_{\mathrm{L}} \end{pmatrix}.$$

We assume the six vertices $x_{\mathrm{K}}, x_{\mathrm{L}}$, $x_{\mathrm{A}}, x_{\mathrm{B}}$ and $x_{\mathrm{E}}, x_{\mathrm{F}}$ of the diamond cell $\mathrm{D}(\mathrm{E},\mathrm{F})$ to be ordered in such a way that $\varDelta_{\mathrm{EF}} := \det(x_{\mathrm{B}} - x_{\mathrm{A}}, x_{\mathrm{F}} - x_{\mathrm{E}}, x_{\mathrm{L}} - x_{\mathrm{K}}) > 0$. Thus, the measure of D is $|\mathrm{D}| = \frac{1}{6}\varDelta_{\mathrm{EF}}$.

We denote the control volume associated with a vertex x_{A} of the mesh by A. This control volume is built by gathering the contributions (i.e., sub-tetraedra) of the diamond cells that share vertex x_{A} as:

$$\mathrm{A} = \bigcup_{\mathrm{D}\in\mathrm{D}_{\mathrm{A}}} \mathrm{hull}\left(x_{\mathrm{D}}, x_{\mathrm{A}}, \begin{pmatrix} x_{\mathrm{E}} \\ x_{\mathrm{F}} \end{pmatrix}, \begin{pmatrix} x_{\mathrm{K}} \\ x_{\mathrm{L}} \end{pmatrix} \right),$$

where $\mathrm{D}_{\mathrm{A}} = \{\mathrm{D} \in \mathscr{D}, \text{ such that } x_{\mathrm{A}} \prec x_{\mathrm{E}} \prec x_{\mathrm{F}}\}$ for x_{A} fixed. The resulting finite volume partition of $\varOmega$, denoted by $\overline{\mathscr{N}}$, forms the *vertex mesh*. The vertex mesh is split into interior and boundary controls volumes, respectively denoted by $\mathscr{N}$ and $\partial\mathscr{N}$; formally, it holds that $\overline{\mathscr{N}} = \mathscr{N} \cup \partial\mathscr{N}$.

Similarly, we associate a control volume denoted either by F or by E, with the point x_{F} (face center) or the point x_{E} (edge midpoint) in accordance with the following formula:

$$\mathrm{E} = \bigcup_{\mathrm{D}\in\mathrm{D}_{\mathrm{E}}} \mathrm{hull}\left(x_{\mathrm{D}}, \begin{pmatrix} x_{\mathrm{A}} \\ x_{\mathrm{B}} \end{pmatrix}, x_{\mathrm{E}}, \begin{pmatrix} x_{\mathrm{K}} \\ x_{\mathrm{L}} \end{pmatrix} \right), \quad \mathrm{F} = \bigcup_{\mathrm{D}\in\mathrm{D}_{\mathrm{F}}} \mathrm{hull}\left(x_{\mathrm{D}}, \begin{pmatrix} x_{\mathrm{A}} \\ x_{\mathrm{B}} \end{pmatrix}, x_{\mathrm{F}}, \begin{pmatrix} x_{\mathrm{K}} \\ x_{\mathrm{L}} \end{pmatrix} \right),$$

where $\mathrm{D}_{\mathrm{E}} = \{\mathrm{D} \in \mathscr{D}, \text{ with } x_{\mathrm{E}} \prec x_{\mathrm{F}}\}$ with x_{E} fixed and $\mathrm{D}_{\mathrm{F}} = \{\mathrm{D} \in \mathscr{D}, \text{ with } x_{\mathrm{E}} \prec x_{\mathrm{F}}\}$ with x_{F} fixed. The resulting finite volume partition of $\varOmega$, denoted by $\overline{\mathscr{F}\mathscr{E}}$, is the *face-edge mesh*. This partition contains both control volumes associated with the faces and the edges of the initial mesh and is split into the interior and boundary controls volumes, respectively denoted by $\mathscr{F}\mathscr{E}$ and $\partial\mathscr{F}\mathscr{E}$; formally, it holds that $\overline{\mathscr{F}\mathscr{E}} = \mathscr{F}\mathscr{E} \cup \partial\mathscr{F}\mathscr{E}$.

The 3D "Cell-Vertex-Face/Edge" DDFV Scheme

We say that $u^{\mathscr{T}} = (u^{\mathscr{M}}, u^{\mathscr{N}}, u^{\mathscr{F}\mathscr{E}})$ is a *discrete function* on Ω whenever its three components are piecewise constant functions on the meshes $\mathscr{M}$, $\mathscr{N}$ and $\mathscr{F}\mathscr{E}$, respectively, and take the form

$$u^{\mathscr{M}} = \sum_{\mathrm{K}\in\mathscr{M}} u_{\mathrm{K}}\chi_{\mathrm{K}}, \quad u^{\mathscr{N}} = \sum_{A\in\mathscr{N}} u_{\mathrm{A}}\chi_{A}, \quad u^{\mathscr{F}\mathscr{E}} = \sum_{\mathrm{F}\in\mathscr{F}} u_{\mathrm{F}}\chi_{\mathrm{F}} + \sum_{\mathrm{E}\in\mathscr{E}} u_{\mathrm{E}}\chi_{\mathrm{E}}.$$

Let X denote the set of the degrees of freedom of the form

$$u^{\mathscr{T}} = \big((u_{\mathrm{K}})_{K\in\mathscr{M}}, (u_{\mathrm{A}})_{A\in\mathscr{N}}, (u_{\mathrm{E}})_{E\in\mathscr{E}}, (u_{\mathrm{F}})_{F\in\mathscr{F}}\big).$$

In order to take into account the Dirichlet boundary conditions, this set is supplemented by the boundary data

$$\delta u^{\mathscr{T}} = \left((u_{\mathrm{K}})_{x_{\mathrm{K}}\in\partial\mathscr{M}},\ (u_{\mathrm{A}})_{x_{\mathrm{A}}\in\partial\mathscr{N}},\ (u_{\mathrm{E}})_{x_{\mathrm{E}}\in\partial\mathscr{F}\mathscr{E}},\ (u_{\mathrm{F}})_{x_{\mathrm{F}}\in\partial\mathscr{F}\mathscr{E}}\right),$$

which form the set ∂X. We will search the numerical approximation to the scalar solution field u in the product set $(u^{\mathscr{T}}, \delta u^{\mathscr{T}}) \in X \times \partial X$. Note that X, ∂X, and $X \times \partial X$ can be given the algebraic structure of a linear space after introducing (in the obvious way) the addition of two elements of the set and the multiplication of an element of the set by a real number.

The gradient of the discrete unknown $u^{\mathscr{T}}$, denoted by $\nabla^{\mathscr{T}} u^{\mathscr{T}}$, is a constant vector field on each diamond cell and is identified with a piecewise constant vector field on mesh $\mathscr{D}$. It depends on the boundary data $\delta u^{\mathscr{T}}$ and can be written as $\nabla^{\mathscr{T}}_{\delta u} u^{\mathscr{T}} = \sum_{\mathrm{D}\in\mathscr{D}} \nabla^{\mathrm{D}}_{\delta u} u^{\mathscr{T}} \chi_{\mathrm{D}}$ where

$$\nabla^{\mathrm{D}}_{\delta u} u^{\mathscr{T}} = \frac{1}{3|\mathrm{D}|}\big((u_{\mathrm{L}} - u_{\mathrm{K}})N_{\mathrm{KL}} + (u_{\mathrm{B}} - u_{\mathrm{A}})N_{\mathrm{AB}} + (u_{\mathrm{F}} - u_{\mathrm{E}})N_{\mathrm{EF}}\big). \tag{1}$$

for any $\mathrm{D} \in \mathscr{D}$ and with the vectors $N_{\mathrm{KL}} = \frac{1}{2}(x_{\mathrm{B}} - x_{\mathrm{A}}) \times (x_{\mathrm{F}} - x_{\mathrm{E}})$, $N_{\mathrm{AB}} = \frac{1}{2}(x_{\mathrm{F}} - x_{\mathrm{E}}) \times (x_{\mathrm{L}} - x_{\mathrm{K}})$ and $N_{\mathrm{EF}} = \frac{1}{2}(x_{\mathrm{L}} - x_{\mathrm{K}}) \times (x_{\mathrm{B}} - x_{\mathrm{A}})$. This procedure defines a gradient operator, denoted by $\nabla^{\mathscr{T}}_{\delta u}$, mapping the discrete space X onto the space of the discrete vector fields $\nabla^{\mathscr{T}} u^{\mathscr{T}}$, which we conveniently denote by $\mathbf{Q}$.

Using the gradient formula we define the flux through each interface of the control volumes of the three meshes $\mathscr{M}$, $\mathscr{N}$ and $\mathscr{F}\mathscr{E}$. The three finite volume schemes are written by using a discrete divergence operator that maps each vector field in $\mathbf{Q}$ to a triple of scalar functions in X. Formally, we introduce the operator

$$\mathrm{div}^{\mathscr{T}} : \xi = (\xi_{\mathrm{D}})_{\mathrm{D}\in\mathscr{D}} \in \mathbf{Q} \mapsto (\mathrm{div}^{\mathscr{M}}\xi, \mathrm{div}^{\mathscr{N}}\xi, \mathrm{div}^{\mathscr{F}\mathscr{E}}\xi) \in X$$

where $\mathrm{div}^{\mathscr{M}}\xi = (\mathrm{div}_{\mathrm{K}}\xi)_{\mathrm{K}}$, $\mathrm{div}^{\mathscr{N}}\xi = (\mathrm{div}_{\mathrm{A}}\xi)_{\mathrm{A}}$ and $\mathrm{div}^{\mathscr{F}\mathscr{E}}\xi = \{(\mathrm{div}_{\mathrm{E}}\xi)_{\mathrm{E}}, (\mathrm{div}_{\mathrm{F}}\xi)_{\mathrm{F}}\}$ are given by

$$|\mathrm{K}|\mathrm{div}_{\mathrm{K}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{K}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{KL}}, \quad |\mathrm{A}|\mathrm{div}_{\mathrm{A}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{A}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{AB}}, \tag{2}$$

$$|\mathrm{E}|\mathrm{div}_{\mathrm{E}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{E}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{EF}}, \quad |\mathrm{F}|\mathrm{div}_{\mathrm{F}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{F}}} \xi_{\mathrm{D}} \cdot (-N_{\mathrm{EF}})\,. \tag{3}$$

In the previous statements, the symbols D_{K}, D_{A}, D_{E}, D_{F} refer to the diamond cells which overlap the cells labeled by the corresponding subscripted indices K, A, E, and L.

Since each of the $\mathrm{div}_{\mathrm{C}}\xi$ approximates $\frac{1}{|\mathrm{C}|}\int_{\mathrm{C}}\mathrm{div}\xi$ (for $\mathrm{C} = \mathrm{K}, \mathrm{A}, \mathrm{E}, \mathrm{F}$), the right hand side of the discrete problem is given by the piecewise constant projection of the function f onto the space X, $\pi^{\mathscr{T}} f = \{(f_{\mathrm{K}})_{\mathrm{K}\in\mathscr{M}}, (f_{\mathrm{A}})_{\mathrm{A}\in\mathscr{N}}, (f_{\mathrm{E}}, f_{\mathrm{F}})_{\mathrm{E}\in\mathscr{E},\mathrm{F}\in\mathscr{F}}\}$ with $f_{\mathrm{C}} = \frac{1}{|\mathrm{C}|}\int_{\mathrm{C}} f(x)dx$ for any cell $\mathrm{C} = \mathrm{K} \in \mathscr{M}$ or $\mathrm{A} \in \mathscr{N}$ or F or $\mathrm{E} \in \mathscr{F}\mathscr{E}$.

Finally, the DDFV scheme reads as

$$-\,\mathrm{div}^{\mathscr{T}}(\mathbf{K}_{\mathrm{D}}\nabla^{\mathrm{D}}_{\delta u}u^{\mathscr{T}}) = \pi^{\mathscr{T}} f \tag{4}$$

where $\mathbf{K}_{\mathrm{D}} = \frac{1}{|\mathrm{D}|}\int_{\mathrm{D}}\mathbf{K}(x)dx$ is defined piecewise on the diamond cells. The scheme in (4) originates a symmetric and positive-definite linear system of equations (see [6] for a thourough discussion of the other properties). Assembling the matrix of the system amounts to gathering the local contributions of the discrete gradient associated to each diamond cell. These contributions are explicitly taken into account by the local Gram matrix

$$\mathbb{K}_{\mathrm{D}} = \begin{pmatrix} \mathbf{K}_{\mathrm{D}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{KL}} \cdot N_{\mathrm{AB}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{KL}} \cdot N_{\mathrm{EF}} \\ \mathbf{K}_{\mathrm{D}}N_{\mathrm{AB}} \cdot N_{\mathrm{KL}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{AB}} \cdot N_{\mathrm{AB}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{AB}} \cdot N_{\mathrm{EF}} \\ \mathbf{K}_{\mathrm{D}}N_{\mathrm{EF}} \cdot N_{\mathrm{KL}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{EF}} \cdot N_{\mathrm{AB}} & \mathbf{K}_{\mathrm{D}}N_{\mathrm{EF}} \cdot N_{\mathrm{EF}} \end{pmatrix}$$

The right hand side in (4) is split similarly in elementary contributions on the eight tetrahedra that compose the diamond cells D.

3 Treatment of discontinuous permeability tensors

The case of a discontinuous permeability tensor in the DDFV framework deserves a special treatment that we discuss in this subsection. Let us suppose that the permeability tensor is discontinuous across the interfaces of the control volumes of mesh $\mathscr{M}$. We decompose each diamond cell into two sub-diamonds D_{K} and D_{L}, i.e., $\mathrm{D} = \mathrm{D}_{\mathrm{K}} \cup \mathrm{D}_{\mathrm{L}}$, where D_{K} is the union of the four tetrahedra with vertices x_{D}, x_{K}, the third vertex being x_{A} or x_{B}, and the fourth vertex being x_{E} or x_{F}.

Then, we introduce an additional degree of freedom at x_{D}, the center of the diamond cell, and we write a gradient formula that is exact for affine functions on the two sub-diamonds. We obtain the two following formulas

$$\nabla^{\mathscr{T}}_{\mathrm{K}} u^{\mathscr{T}} = \frac{1}{3|\mathrm{D_K}|}\big((u_{\mathrm{D}} - u_{\mathrm{K}})N_{\mathrm{KL}} + (u_{\mathrm{B}} - u_{\mathrm{A}})N^{\mathrm{K}}_{\mathrm{AB}} + (u_{\mathrm{F}} - u_{\mathrm{E}})N^{\mathrm{K}}_{\mathrm{EF}}\big)$$

$$\nabla^{\mathscr{T}}_{\mathrm{L}} u^{\mathscr{T}} = \frac{1}{3|\mathrm{D_L}|}\big((u_{\mathrm{L}} - u_{\mathrm{D}})N_{\mathrm{KL}} + (u_{\mathrm{B}} - u_{\mathrm{A}})N^{\mathrm{L}}_{\mathrm{AB}} + (u_{\mathrm{F}} - u_{\mathrm{E}})N^{\mathrm{L}}_{\mathrm{EF}}\big)$$

using the geometric vectors $N^{\mathrm{K}}_{\mathrm{AB}} = \frac{1}{2}(x_{\mathrm{F}} - x_{\mathrm{E}}) \times (x_{\mathrm{D}} - x_{\mathrm{K}})$, $N^{\mathrm{L}}_{\mathrm{AB}} = \frac{1}{2}(x_{\mathrm{F}} - x_{\mathrm{E}}) \times (x_{\mathrm{L}} - x_{\mathrm{D}})$, $N^{\mathrm{K}}_{\mathrm{EF}} = \frac{1}{2}(x_{\mathrm{D}} - x_{\mathrm{K}}) \times (x_{\mathrm{B}} - x_{\mathrm{A}})$, $N^{\mathrm{L}}_{\mathrm{EF}} = \frac{1}{2}(x_{\mathrm{L}} - x_{\mathrm{D}}) \times (x_{\mathrm{B}} - x_{\mathrm{A}})$, and introducing the two volume factors $|\mathrm{D_K}| = \frac{1}{6}\det(x_{\mathrm{B}} - x_{\mathrm{A}}, x_{\mathrm{F}} - x_{\mathrm{E}}, x_{\mathrm{D}} - x_{\mathrm{K}})$ and $|\mathrm{D_L}| = \frac{1}{6}\det(x_{\mathrm{B}} - x_{\mathrm{A}}, x_{\mathrm{F}} - x_{\mathrm{E}}, x_{\mathrm{L}} - x_{\mathrm{D}})$. Also, we remark that $|\mathrm{D}| = |\mathrm{D_K}| + |\mathrm{D_L}|$, $N_{\mathrm{AB}} = N^{\mathrm{K}}_{\mathrm{AB}} + N^{\mathrm{L}}_{\mathrm{AB}}$ $N_{\mathrm{EF}} = N^{\mathrm{K}}_{\mathrm{EF}} + N^{\mathrm{L}}_{\mathrm{EF}}$ and it holds that $|\mathrm{D_K}|N^{\mathrm{L}}_{\mathrm{AB}} - |\mathrm{D_L}|N^{\mathrm{K}}_{\mathrm{AB}} = |\mathrm{D_K}|N_{\mathrm{AB}} - |\mathrm{D}|N^{\mathrm{L}}_{\mathrm{AB}}$.

Let $\mathbf{K}_{\mathrm{D_K}} = \frac{1}{|\mathrm{D_K}|}\int_{\mathrm{D_K}} \mathbf{K}(x)dx$ and $\mathbf{K}_{\mathrm{D_L}} = \frac{1}{|\mathrm{D_L}|}\int_{\mathrm{D_L}} \mathbf{K}(x)dx$ be the constant approximation of the diffusion tensor on the two sub-diamonds D_K and D_L. We determine the additional unknown u_{D} in terms of the other local degrees of freedom $u_{\mathrm{K}}, u_{\mathrm{L}}, u_{\mathrm{A}}, u_{\mathrm{B}}, u_{\mathrm{E}}$ and u_{F} by imposing that

$$\mathbf{K}_{\mathrm{D_K}}\nabla^{\mathscr{T}}_{\mathrm{K}} u^{\mathscr{T}} \cdot N_{\mathrm{KL}} = \mathbf{K}_{\mathrm{D_L}}\nabla^{\mathscr{T}}_{\mathrm{L}} u^{\mathscr{T}} \cdot N_{\mathrm{KL}},$$

which is the flux conservation through the common face $\mathrm{D}_K|\mathrm{D}_L$. Moreover, let us introduce the following geometric factors that also depend on the permeability coefficients:

$$\beta_{KL} = +|\mathrm{D_L}|\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}} + |\mathrm{D_K}|\mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}$$
$$\beta_{AB} = -|\mathrm{D_L}|\mathbf{K}_{\mathrm{D_K}}N^{\mathrm{K}}_{\mathrm{AB}} \cdot N_{\mathrm{KL}} + |\mathrm{D_K}|\mathbf{K}_{\mathrm{D_L}}N^{\mathrm{L}}_{\mathrm{AB}} \cdot N_{\mathrm{KL}}$$
$$\beta_{EF} = -|\mathrm{D_L}|\mathbf{K}_{\mathrm{D_K}}N^{\mathrm{K}}_{\mathrm{EF}} \cdot N_{\mathrm{KL}} + |\mathrm{D_K}|\mathbf{K}_{\mathrm{D_L}}N^{\mathrm{L}}_{\mathrm{EF}} \cdot N_{\mathrm{KL}}$$

A straightforward calculation yields the formula for u_{D}

$$u_{\mathrm{D}} = \frac{|\mathrm{D_L}|\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}u_{\mathrm{K}} + \frac{|\mathrm{D_K}|\mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}u_{\mathrm{L}} + \frac{\beta_{AB}}{\beta_{KL}}\big(u_{\mathrm{B}} - u_{\mathrm{A}}\big) + \frac{\beta_{EF}}{\beta_{KL}}\big(u_{\mathrm{F}} - u_{\mathrm{E}}\big),$$

and the formulas for the numerical gradients:

$$3\nabla^{\mathscr{T}}_{\mathrm{K}} u^{\mathscr{T}} = \frac{\mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}N_{\mathrm{KL}}\big(u_{\mathrm{L}} - u_{\mathrm{K}}\big) + \left(\frac{\beta_{AB}}{|\mathrm{D_K}|\beta_{KL}}N_{\mathrm{KL}} + \frac{1}{|\mathrm{D_K}|}N^{\mathrm{K}}_{\mathrm{AB}}\right)(u_{\mathrm{B}} - u_{\mathrm{A}})$$
$$+ \left(\frac{\beta_{EF}}{|\mathrm{D_K}|\beta_{KL}}N_{\mathrm{KL}} + \frac{1}{|\mathrm{D_K}|}N^{\mathrm{K}}_{\mathrm{EF}}\right)(u_{\mathrm{F}} - u_{\mathrm{E}}),$$

$$3\nabla^{\mathscr{T}}_{\mathrm{L}} u^{\mathscr{T}} = \frac{\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}N_{\mathrm{KL}}\big(u_{\mathrm{L}} - u_{\mathrm{K}}\big) + \left(\frac{\beta_{AB}}{|\mathrm{D_L}|\beta_{KL}}N_{\mathrm{KL}} + \frac{1}{|\mathrm{D_L}|}N^{\mathrm{L}}_{\mathrm{AB}}\right)(u_{\mathrm{B}} - u_{\mathrm{A}})$$
$$+ \left(\frac{\beta_{EF}}{|\mathrm{D_L}|\beta_{KL}}N_{\mathrm{KL}} + \frac{1}{|\mathrm{D_L}|}N^{\mathrm{L}}_{\mathrm{EF}}\right)(u_{\mathrm{F}} - u_{\mathrm{E}}).$$

Finally, we define the divergence operator for a discrete vector field which is piecewise constant on $\mathrm{D}_K \cap \mathrm{D}_L$ and may be discontinuous across $\mathrm{D}_K|\mathrm{D}_L$ as

$$|\mathrm{K}|\mathrm{div}_{\mathrm{K}}\xi^{\mathscr{D}} = \sum_{\mathrm{D}|\mathrm{K}} \xi_{\mathrm{D_K}} \cdot N_{\mathrm{KL}} = \sum_{\mathrm{D}|\mathrm{K}} \xi_{\mathrm{D_L}} \cdot N_{\mathrm{KL}} = \sum_{\mathrm{D}|\mathrm{K}} \left(\frac{|\mathrm{D_K}|}{|\mathrm{D}|}\xi_{\mathrm{D_K}} + \frac{|\mathrm{D_L}|}{|\mathrm{D}|}\xi_{\mathrm{D_L}} \right) \cdot N_{\mathrm{KL}}, \tag{5}$$

$$|\mathrm{A}|\mathrm{div}_{\mathrm{A}}\xi^{\mathscr{D}} = \sum_{\mathrm{D}|\mathrm{A}} (\xi_{\mathrm{D_K}} \cdot N^{\mathrm{K}}_{\mathrm{AB}} + \xi_{\mathrm{D_L}} \cdot N^{\mathrm{L}}_{\mathrm{AB}}), \tag{6}$$

$$|\mathrm{E}|\mathrm{div}_{\mathrm{E}}\xi^{\mathscr{D}} = \sum_{\mathrm{D}|\mathrm{E}} (\xi_{\mathrm{D_K}} \cdot N^{\mathrm{K}}_{\mathrm{EF}} + \xi_{\mathrm{D_L}} \cdot N^{\mathrm{L}}_{\mathrm{EF}}), \tag{7}$$

$$|\mathrm{F}|\mathrm{div}_{\mathrm{F}}\xi^{\mathscr{D}} = \sum_{\mathrm{D}|\mathrm{D_F}} (\xi_{\mathrm{D_K}} \cdot (-N^{\mathrm{K}}_{\mathrm{EF}}) - \xi_{\mathrm{D_L}} \cdot (-N^{\mathrm{L}}_{\mathrm{EF}})). \tag{8}$$

The DDFV method for the discontinuous case follows by using (5)-(8) with the approximate permeability tensors $\mathbf{K}_{\mathrm{D_K}}$ and $\mathbf{K}_{\mathrm{D_L}}$ instead of (2)-(3) in the scheme formulation (4). Let $\xi^{\mathscr{D}} = \mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}}u^{\mathscr{T}}$ and evaluate the quantities:

$$\begin{aligned}
\left(\frac{|\mathrm{D_K}|}{|\mathrm{D}|}\xi_{\mathrm{D_K}} + \frac{|\mathrm{D_L}|}{|\mathrm{D}|}\xi_{\mathrm{D_L}} \right) \cdot N_{\mathrm{KL}} &= \alpha_{\mathrm{KL\cdot KL}}(u_{\mathrm{L}} - u_{\mathrm{K}}) + \alpha_{\mathrm{KL\cdot AB}}(u_{\mathrm{B}} - u_{\mathrm{A}}) + \alpha_{\mathrm{KL\cdot EF}}(u_{\mathrm{F}} - u_{\mathrm{E}}) \\
\xi_{\mathrm{D_K}} \cdot N^{\mathrm{K}}_{\mathrm{AB}} + \xi_{\mathrm{D_L}} \cdot N^{\mathrm{L}}_{\mathrm{AB}} &= \alpha_{\mathrm{AB\cdot KL}}(u_{\mathrm{L}} - u_{\mathrm{K}}) + \alpha_{\mathrm{AB\cdot AB}}(u_{\mathrm{B}} - u_{\mathrm{A}}) + \alpha_{\mathrm{AB\cdot EF}}(u_{\mathrm{F}} - u_{\mathrm{E}}) \\
\xi_{\mathrm{D_K}} \cdot N^{\mathrm{K}}_{\mathrm{EF}} + \xi_{\mathrm{D_L}} \cdot N^{\mathrm{L}}_{\mathrm{EF}} &= \alpha_{\mathrm{EF\cdot KL}}(u_{\mathrm{L}} - u_{\mathrm{K}}) + \alpha_{\mathrm{EF\cdot AB}}(u_{\mathrm{B}} - u_{\mathrm{A}}) + \alpha_{\mathrm{EF\cdot EF}}(u_{\mathrm{F}} - u_{\mathrm{E}})
\end{aligned}$$

using the entries of the coefficient matrix

$$\mathbf{K}^{new}_{\mathrm{D}} = \begin{pmatrix} \alpha_{\mathrm{KL\cdot KL}} & \alpha_{\mathrm{KL\cdot AB}} & \alpha_{\mathrm{KL\cdot EF}} \\ \alpha_{\mathrm{AB\cdot KL}} & \alpha_{\mathrm{AB\cdot AB}} & \alpha_{\mathrm{AB\cdot EF}} \\ \alpha_{\mathrm{EF\cdot KL}} & \alpha_{\mathrm{EF\cdot AB}} & \alpha_{\mathrm{EF\cdot EF}} \end{pmatrix}.$$

Since $\mathbf{K}^{new}_{\mathrm{D}}$ is a 3×3 symmetric elements we have only six independent entries, which after a straightforward calculations are given by:

$$\begin{aligned}
\alpha_{\mathrm{KL\cdot KL}} &= \frac{1}{3}\frac{\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}\ \mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}, \\
\alpha_{\mathrm{KL\cdot AB}} &= \frac{\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N^{\mathrm{K}}_{\mathrm{AB}}\ \mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}} + \mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N^{\mathrm{L}}_{\mathrm{AB}}\ \mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}, \\
\alpha_{\mathrm{KL\cdot EF}} &= \frac{1}{3}\frac{\mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N^{\mathrm{K}}_{\mathrm{EF}}\ \mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}} + \mathbf{K}_{\mathrm{D_L}}N_{\mathrm{KL}} \cdot N^{\mathrm{L}}_{\mathrm{EF}}\ \mathbf{K}_{\mathrm{D_K}}N_{\mathrm{KL}} \cdot N_{\mathrm{KL}}}{\beta_{KL}}, \\
\alpha_{\mathrm{AB\cdot AB}} &= \frac{1}{3}\left(-\frac{\beta^2_{AB}}{|\mathrm{D_K}||\mathrm{D_L}|\beta_{KL}} + \frac{1}{|\mathrm{D_K}|}\mathbf{K}_{\mathrm{D_K}}N^{\mathrm{K}}_{\mathrm{AB}} \cdot N^{\mathrm{K}}_{\mathrm{AB}} + \frac{1}{|\mathrm{D_L}|}\mathbf{K}_{\mathrm{D_L}}N^{\mathrm{L}}_{\mathrm{AB}} \cdot N^{\mathrm{L}}_{\mathrm{AB}} \right),
\end{aligned}$$

$$\alpha_{\mathrm{AB}\cdot\mathrm{EF}} = \frac{1}{3}\Big(-\frac{\beta_{AB}\beta_{EF}}{|\mathrm{D_K}||\mathrm{D_L}|\beta_{KL}} + \frac{1}{|\mathrm{D_K}|}\mathbf{K}_{\mathrm{D_K}}N^{\mathrm{K}}_{\mathrm{EF}}\cdot N^{\mathrm{K}}_{\mathrm{AB}} + \frac{1}{|\mathrm{D_L}|}\mathbf{K}_{\mathrm{D_L}}N^{\mathrm{L}}_{\mathrm{EF}}\cdot N^{\mathrm{L}}_{\mathrm{AB}}\Big),$$

$$\alpha_{\mathrm{EF}\cdot\mathrm{EF}} = \frac{1}{3}\Big(-\frac{\beta^2_{EF}}{|\mathrm{D_K}||\mathrm{D_L}|\beta_{KL}} + \frac{1}{|\mathrm{D_K}|}\mathbf{K}_{\mathrm{D_K}}N^{\mathrm{K}}_{\mathrm{EF}}\cdot N^{\mathrm{K}}_{\mathrm{EF}} + \frac{1}{|\mathrm{D_L}|}\mathbf{K}_{\mathrm{D_L}}N^{\mathrm{L}}_{\mathrm{EF}}\cdot N^{\mathrm{L}}_{\mathrm{EF}}\Big).$$

and the remaining coefficients are determined by symmetry, i.e., $\alpha_{\mathrm{AB}\cdot\mathrm{KL}} = \alpha_{\mathrm{KL}\cdot\mathrm{AB}}$, $\alpha_{\mathrm{EF}\cdot\mathrm{KL}} = \alpha_{\mathrm{KL}\cdot\mathrm{EF}}$, and $\alpha_{\mathrm{EF}\cdot\mathrm{AB}} = \alpha_{\mathrm{AB}\cdot\mathrm{EF}}$.

4 Conclusions

In this work, we discussed how a discontinuous permeability can be treated in the numerical framework offered by the DDFV method. Whenever the discontinuity is across an internal interfaces shared by two control volumes of the primal mesh, two different gradients are considered on the two subdiamonds centered at that interface. Introducing an auxiliary variable at the interface and imposing flux conservation makes it possible to derive a formula for both gradients that is exact for affine functions. Then, a DDFV method can be formulated using a discrete divergence operator to express the flux balance on the overlapping meshes for primal control volumes, vertex control volumes and face-edge control volumes. The numerical experiments in [7] show the effectiveness of the method.

Acknowledgements The work of the first two authors was supported by Groupement de Recherche MOMAS. The work of the third author was partially supported by the Italian MIUR through the program PRIN2008.

References

1. Andreianov, B., Bendahmane, M., Hubert, F.: On 3D DDFV discretization of gradient and divergence operators. Part II. (2011). HAL, http://hal.archives-ouvertes.fr/hal-00567342
2. Andreianov, B., Bendahmane, M., Hubert, F., Krell, S.: On 3D DDFV discretization of gradient and divergence operators. Part I. (2011) HAL, http://hal.archives-ouvertes.fr/hal-00355212.
3. Andreianov, B., Bendahmane, M., Karlsen, K.: A gradient reconstruction formula for finite-volume schemes and discrete duality. FVCA5, Wiley, (2008).
4. Andreianov, B., Hubert, F., Krell, S.: Benchmark 3D: a version of the DDFV scheme with cell/vertex unknowns on general meshes, this volume (2011).
5. Boyer, F., Hubert, F.: Benchmark on anisotropic problems, the DDFV discrete duality finite volumes and m-DDFV schemes. In: R. Eymard, J.M. Hérard (eds.) FVCA5, Wiley, (2008).
6. Coudière, Y., Hubert, F.: A 3D discrete duality finite volume method for nonlinear elliptic equation (2010) HAL, URL: http://hal.archives-ouvertes.fr/hal-00456837/fr.
7. Coudière, Y., Hubert, F., Manzini, G.: Benchmark 3D: CeVeFE-DDFV, a discrete duality scheme with cell/vertex/face+edge unknowns. (this volume) (2011).
8. Coudière, Y., Pierre, C., Rousseau, O., Turpault, R.: A 2D/3D discrete duality finite volume scheme. Application to ECG simulation. International Journal on Finite Volumes (2009) **6**(1).

9. Domelevo, K., Omnès, P.: A finite volume method for the laplace equation on almost arbitrary two-dimensional grids. M2AN, Math. Model. Numer. Anal. (2005) **39**(6), 1203–1249.
10. Herbin, R., Hubert, F.: Benchmark on discretization schemes for anisotropic diffusion problems on general grids. FVCA5, Wiley, (2008).
11. Hermeline, F.: Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. Comp. Meth. Appl. Mech. Eng. (2003) **192**(16), 1939–1959.
12. Hermeline, F.: A finite volume method for approximating 3D diffusion operators on general meshes. Journal of computational Physics (2009) **228**(16), 5763–5786.
13. Pierre, C.: Modélisation et simulation de l'activité électrique du coeur dans le thorax, analyse numérique et méthodes de volumes finis. Ph.D. thesis, Université de Nantes (2005).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Multi-Water-Bag Model And Method Of Moments For The Vlasov Equation

Anaïs Crestetto and Philippe Helluy

Abstract The kinetic Vlasov-Poisson model is very expensive to solve numerically. It can be approximated by a multi-water-bag model in order to reduce the complexity. This model amounts to solve a set of Burgers equations, which can be done easily by finite volume methods. However, the solution is naturally multivalued (filamentation). The multivalued solution can be computed by a moment method. We present here several numerical experiments.

Keywords Vlasov-Poisson, water-bag approximation, Burgers equation, multivalued solution
MSC2010: 35Q83, 44A60, 65M08

1 Introduction

Kinetic equations are used in several domains, such as plasma physics or bubble flows in gases or liquids. The distribution function depends on space and time but also on an additional velocity variable. Computations are thus very expensive. It is of great interest to reduce the complexity of the resolution by using fluid methods.

We consider a plasma containing ions of positive charge and electrons of negative charge. Ions are much heavier than electrons so that we can neglect their displacement and assume that their density n_0 is constant. The electrons move following the system of Vlasov-Poisson in a periodic domain in $\mathbf{x}$:

Anaïs Crestetto
IRMA, University of Strasbourg & INRIA Nancy-Grand Est, e-mail: crestetto@math.unistra.fr

Philippe Helluy
IRMA, University of Strasbourg, e-mail: helluy@math.unistra.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_31, © Springer-Verlag Berlin Heidelberg 2011

$$\partial_t f(\mathbf{x},\mathbf{v},t) + \mathbf{v}\cdot\nabla_{\mathbf{x}} f(\mathbf{x},\mathbf{v},t) + \frac{q}{m}\mathbf{E}(\mathbf{x},t)\cdot\nabla_{\mathbf{v}} f(\mathbf{x},\mathbf{v},t) = 0, \tag{1}$$

$$\operatorname{div}\mathbf{E}(\mathbf{x},t) = \rho(\mathbf{x},t) = \frac{q}{m}\left(\int f(\mathbf{x},\mathbf{v},t)\; d\mathbf{v} - n_0\right), \tag{2}$$

$$\int \mathbf{E}(\mathbf{x},t)\; d\mathbf{x} = 0, \tag{3}$$

$$f(\mathbf{x},\mathbf{v},0) = f_0(\mathbf{x},\mathbf{v}) \approx n_0. \tag{4}$$

The unknowns are the distribution function of electrons f, and the electric field $\mathbf{E}$. The electric field depends on space and time $\mathbf{x}, t$, while the distribution function depends also on an additional velocity variable $\mathbf{v}$. The charge and the mass of one electron are noted $q < 0$ and $m > 0$ respectively. Without loss of generality, we can take $\frac{q}{m} = -1$ and $n_0 = 1$. In one dimension of space, this system becomes:

$$\partial_t f(x,v,t) + v\partial_x f(x,v,t) - E(x,t)\,\partial_v f(x,v,t) = 0, \tag{5}$$

$$\partial_x E(x,t) = \rho(x,t) = 1 - \int f(x,v,t)\; dv, \tag{6}$$

$$\int E(x,t)\; dx = 0, \tag{7}$$

$$f(x,v,0) = f_0(x,v) \approx 1. \tag{8}$$

The distribution function f is initially a perturbation of the equilibrium n_0. After simple calculations, Equation (6) can also be written

$$\partial_t E(x,t) = \int v f(x,v,t)\; dv - \int v f(0,v,t)\; dv. \tag{9}$$

In higher dimensions, Equation (6) would be replaced by a Poisson equation, assuming that $E = -\nabla\Phi$, where Φ is the electric potential.

The solution can be stable (for example in the case of Landau damping). But since there is no dissipation, the solution can become unstable and filamentation can appear.

In order to reduce the complexity of our model, we propose to approximate the kinetic equation by fluid models. We consider two possibilities:

- the multi-water-bag model,
- the method of moments.

Then we propose numerical approximations of these models by simple finite volume schemes. The numerical results will be compared to those obtained with a full resolution of the kinetic model by the popular Particle-In-Cell (PIC) method (described for example in [3]).

2 Multi-water-bag model

The multi-water-bag model, detailed in [2], generalizes the water-bag model (see for example [1]). It consists of replacing f by a piecewise constant approximation in the velocity variable. Each piece is called a "water-bag". It is possible to compute only the boundaries of the water-bags.

2.1 Presentation of the model

Let N be an integer, $v_j^+(x,t)$ and $v_j^-(x,t)$ velocities, A_j constants, $j = 1,\ldots,N$, such that we can write:

$$f(x,v,t) = \sum_{j=1}^{N} A_j \left(H\left(v_j^+(x,t) - v\right) - H\left(v_j^-(x,t) - v\right)\right), \tag{10}$$

where H is the Heaviside function: $H(u) = \begin{cases} 0 \text{ if } u < 0, \\ 1 \text{ if } u > 0. \end{cases}$

Injecting this expression in the Vlasov equation, and assuming that there is no two $v_j^\pm(x,t)$ equal, we obtain the following system:

$$\partial_t v_j^\pm + v_j^\pm \partial_x v_j^\pm + E = 0, \quad \forall j = 1,\ldots,N, \tag{11}$$

$$\partial_t E = -\sum_{j=1}^{N} A_j \left(\frac{v_j^{+^2}(x,0) - v_j^{-^2}(x,0)}{2} - \frac{v_j^{+^2}(x,t) - v_j^{-^2}(x,t)}{2} \right). \tag{12}$$

The velocities follow a Burgers equation with a source term. Instead of evolving the distribution function f, we evolve these velocities. The natural solution can become multivalued (filamentation). The weak entropy solution is only an approximation in this context, when shocks appear.

2.2 Numerical scheme

We discretize our domain: $x_i = x_0 + i\Delta x$, with Δx being the spacial step, and consider a time step Δt such that $t^n = n\Delta t$. We evolve each velocity independently by using, for example, the Godunov scheme:

$$v_{j,i}^{\pm,n+1} = v_{j,i}^{\pm,n} - \frac{\Delta t}{\Delta x} \left(\frac{\left(v_{j,i+\frac{1}{2}}^{\pm,\star}\right)^2}{2} - \frac{\left(v_{j,i-\frac{1}{2}}^{\pm,\star}\right)^2}{2} \right) - \Delta t E_i^n, \tag{13}$$

where $v_{j,i}^{\pm,n} \simeq v_j^{\pm}(x_i, t^n)$ and $E_i^n \simeq E(x_i, t^n)$. We compute the $v_j{}^{\pm,\star}$ with an exact Riemann solver.

We compute the electric field with the following scheme:

$$E_i^{n+1} = E_i^n - \Delta t \sum_{j=1}^{N} A_j \left(\frac{\left(v_{j,i}^{+,0}\right)^2 - \left(v_{j,i}^{-,0}\right)^2}{2} - \frac{\left(v_{j,i}^{+,n}\right)^2 - \left(v_{j,i}^{-,n}\right)^2}{2} \right). \tag{14}$$

In higher dimension, this step should be replaced by the numerical resolution of a Poisson equation.

2.3 Remarks

Before the shock, the weak solution and the multivalued solution coincide. After the shock, the natural solution is multivalued, filaments or branches appear, and the weak solution becomes discontinuous. In Sect. 4, we compare numerically the two kinds of solutions. It is also possible to compute several branches by the method of moments [4].

3 Method of moments

The method of moments, presented in [5, 6], consists in taking the first moments of the equation that we have to solve in order to reduce the number of variables. The system is closed by assuming that the distribution function is made of water-bags.

3.1 Presentation of the method

Definition 1. The moment M_k of order k of f is defined by:

$$M_k(x,t) = \int_{-\infty}^{+\infty} v^k f(x,v,t)\ dv. \tag{15}$$

Taking the $2N$ first moments of the Vlasov equation:

$$\int v^k \partial_t f(x,v,t)\ dv + \int v^{k+1} \partial_x f(x,v,t)\ dv - E(x,t) \int v^k \partial_v f(x,v,t)\ dv = 0, \tag{16}$$

we obtain the following system of moments, coupled with the equation for the electric field:

$$\partial_t M_0 + \partial_x M_1 = 0 \tag{17}$$

$$\partial_t M_k + \partial_x M_{k+1} + kEM_{k-1} = 0, \text{ for } k = 1, \ldots, 2N-1, \tag{18}$$

$$\partial_t E(x,t) = M_1(x,t) - M_1(0,t). \tag{19}$$

We now have a system of $2N+1$ equations, in which the velocity no longer appears. There are $2N+2$ unknowns: the moments M_k for $k = 0, \ldots, 2N$ and E. We have to close this system by finding an expression for M_{2N}.

3.2 Closure relation

We represent f by water-bags for closing the system:

$$f(x,v,t) = \sum_{j=1}^{N} A_j \left(H\left(v_j^+(x,t) - v\right) - H\left(v_j^-(x,t) - v\right)\right). \tag{20}$$

We obtain:

$$M_k(x,t) = \sum_{j=1}^{N} A_j \frac{{v_j^+}^{k+1}(x,t) - {v_j^-}^{k+1}(x,t)}{k+1}, \ \forall\, k = 0, \ldots, 2N, \tag{21}$$

and thus have an expression of M_{2N}, assuming that we know the $v_j^\pm$.

3.3 Numerical scheme

At time t^n, we know the $2N$ first moments of f and the electric field. We solve the system:

$$M_k(x,t) = \sum_{j=1}^{N} A_j \frac{{v_j^+}^{k+1}(x,t) - {v_j^-}^{k+1}(x,t)}{k+1}, \ \forall\, k = 0, \ldots, 2N-1, \tag{22}$$

to obtain $v_j^\pm$, at time t^n. In general, the system may have several solutions. Uniqueness can be recovered through an entropy argument [4].

To solve this system, we can use the Newton method. But we have no rigorous result of convergence.

When the problem is such that f can be written:

$$f(x,v,t) = \sum_{j=1}^{2N} (-1)^{j-1} H\left(a_j^+(x,t) - v\right), \tag{23}$$

where $-\infty < a_{2N} \leq \cdots \leq a_1 < +\infty$, we can use the algorithm described in [4,7], which solves rigorously such a system. It is then possible to catch numerically the multivalued solutions.

For approximating the system of moments, we use a natural kinetic scheme. Formally, we write :

$$\partial_t f(x,v,t) + v^+ \partial_x f(x,v,t) + v^- \partial_x f(x,v,t) - E(x,t)\,\partial_v f(x,v,t) = 0, \tag{24}$$

where $v^+ = \max(0,v)$ and $v^- = \min(0,v)$. Denoting Δx as the space step, Δt as the time step, and f_i^n as the approximation of $f(x_i, v, t^n)$, we use the following upwind discretization for the Vlasov equation:

$$\frac{f_i^{n+1} - f_i^n}{\Delta t} + \frac{1}{\Delta x}\left(v^+\left(f_i^n - f_{i-1}^n\right) + v^-\left(f_{i+1}^n - f_i^n\right)\right) - E_i^n \partial_v f_i^n = 0. \tag{25}$$

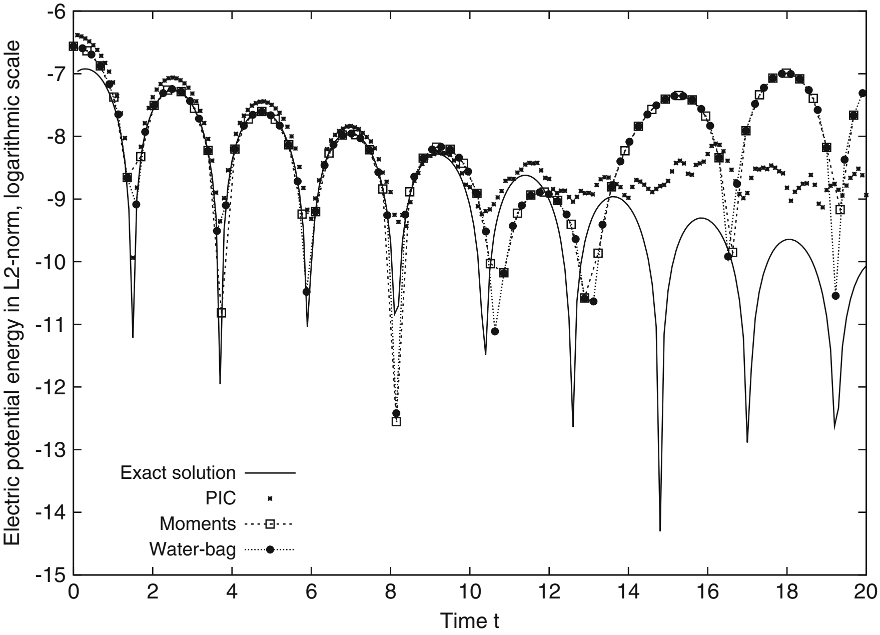


Fig. 1 Landau damping. Multi-water-bag model, method of moments and PIC method compared to the exact solution

We multiply it by v^k and integrate it in v:

$$\begin{aligned}&\frac{1}{\Delta t}\int v^k\left(f_i^{n+1}-f_i^n\right)\,dv\\&+\frac{1}{\Delta x}\int\left(v^{+^{k+1}}\left(f_i^n-f_{i-1}^n\right)+v^{-k+1}\left(f_{i+1}^n-f_i^n\right)\right)\,dv\\&-E_i^n\int v^k\partial_v f_i^n\,dv=0.\end{aligned}\tag{26}$$

We obtain a finite volume scheme:

$$\frac{M_{k,i}^{n+1}-M_{k,i}^n}{\Delta t}+\frac{1}{\Delta x}\left(F\left(M_{k,i}^n,M_{k,i+1}^n\right)-F\left(M_{k,i-1}^n,M_{k,i}^n\right)\right)-kE_i^nM_{k-1,i}^n=0.\tag{27}$$

The scheme for the electric field is:

$$\frac{E_i^{n+1}-E_i^n}{\Delta t}=M_{1,i}^n-M_{1,0}^n.\tag{28}$$

4 Numerical results

We validate our models on classical test cases: Landau damping and two stream instability, and obtain good decrease rates for the electric potential energy, when N is big enough. An example for a Landau damping test case is given in Fig. 1, with $N=5$.

We are now interested in solutions that can initially exactly be depicted by the multi-water-bag model, and that are unstable. We compare the three methods in Fig. 2, for $N=1$:

- the method of moments with an approximation by water-bags,
- the multi-water-bag model with the scheme of Godunov,
- the Particle-In-Cell (PIC) method, considered as the reference.

Before the shock, the two fluid methods describe precisely the solution. After the shock, they describe the main part of the solution, but cannot catch the filaments. More test cases will be presented at the conference, with higher N.

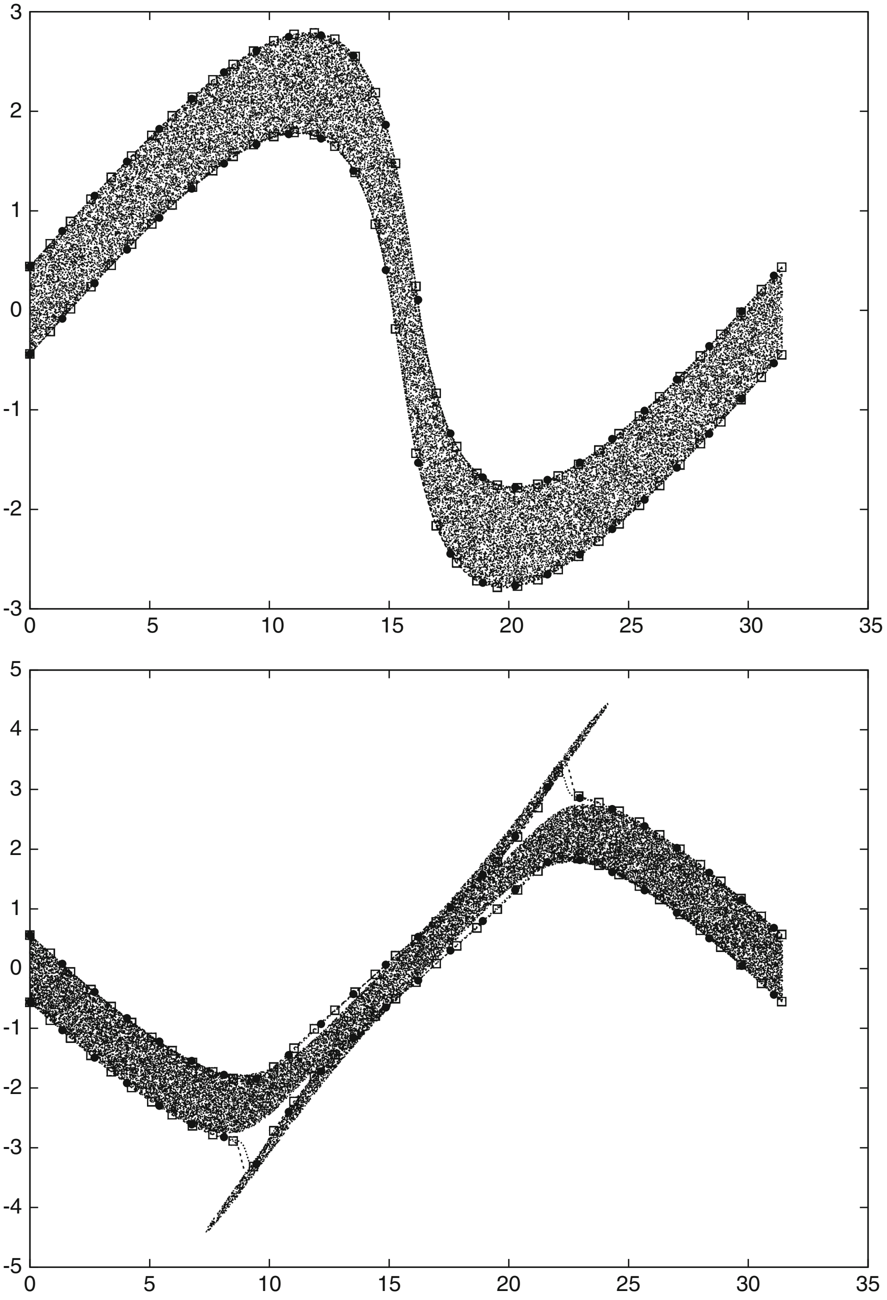


Fig. 2 Test case for $N = 1$ in the phase space at times $T = 2$ and $T = 5$. Multi-water-bag model (black circles), method of moments (empty squares) and PIC method (dots)

References

1. Bertrand, P., Feix, M. R.: Non linear electron plasma oscillation: the water bag model. Phys. Lett. **28A**, 68–69 (1968)
2. Besse, N., Berthelin, F., Brenier, Y., Bertrand, P.: The multi-water-bag equations for collisionless kinetic modeling. Kinetic and Related Models **2**, 39–80 (2009)
3. Birdsall, C. K. and Langdon, A. B.: Plasma Physics via Computer Simulation. Institute of Physics (1991)
4. Brenier, Y. and Corrias, L.: A kinetic formulation for multi-branch entropy solutions of scalar conservation laws. Annales de l'Institut Henri Poincaré. Analyse Non Linéaire **15**, 169–190 (1998)
5. Desjardins, O., Fox, R. O., Villedieu, P.: A quadrature-based moment method for dilute fluid-particle flows. Journal of Computational Physics **227**, 2514–2539 (2008)
6. Fox, R. O., Laurent, F., Massot, M.: Numerical simulation of spray coalescence in an Eulerian framework: direct quadrature method of moments and multi-fluid method. Journal of Computational Physics **227**, 3058–3088 (2008)
7. Gosse, L. and Runborg, O.: Resolution of the finite Markov moment problem. Comptes Rendus Mathématique. Académie des Sciences. Paris **12**, 775–780 (2005)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Comparison of Upwind and Centered Schemes for Low Mach Number Flows

Thu–Huyen DAO, Michael NDJINGA, and Frédéric MAGOULES

Abstract In this paper, fully implicit schemes are used for the numerical simulation of compressible flows at low Mach number. The compressible Navier–Stokes equations are discretized classically using the finite volume framework and a Roe type scheme for the convection flux. Though explicit Godunov type schemes are inaccurate for low Mach number flows on Cartesian meshes, we claim that their implicit counterpart can be more precise for that type of flow. Numerical evidence from the lid driven cavity benchmark shows that the centered implicit scheme can capture low Mach vortices, unlike the upwind scheme. We also propose a Scaling strategy based on the convection spectrum to reduce the computational cost and accelerate the convergence of both linear system and Newton scheme iterations.

Keywords Low Mach number, centered scheme, upwind scheme, compressible flows, scaling preconditioner
MSC2010: 35L65, 65F35, 76E19, 76M12

1 Introduction

Accurate numerical simulation of compressible flows at low Mach number is of great practical importance in the design and safety analysis of nuclear reactors and core thermal-hydraulic studies (see [6] and [7]). The numerical solutions of the

Thu–Huyen DAO
CEA–Saclay, DEN, DM2S, SFME, LGLS, F–91191 Gif-sur–Yvette, France and MAS, Ecole Centrale Paris, 92295 Châtenay–Malabry, France, e-mail: thu-huyen.dao@cea.fr

Michael NDJINGA
CEA–Saclay, DEN, DM2S, SFME, LGLS, F–91191 Gif-sur–Yvette, France, e-mail: michael.ndjinga@cea.fr

Frédéric MAGOULES
MAS, Ecole Centrale Paris, 92295 Châtenay–Malabry, France,
e-mail: frederic.magoules@hotmail.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_32, © Springer-Verlag Berlin Heidelberg 2011

corresponding two-phase flow models are based on Riemann approximate solvers which are robust and can efficiently capture shock wave solutions using an upwind strategy. However, when the flow is at low Mach number, especially on Cartesian meshes, these schemes are inaccurate, and corrections have to be made to capture the correct dynamics (see for example [8]). In [5], a detailed analysis of the behavior of Godunov type schemes applied to the compressible Euler system at low Mach number is proposed. The upwind part of the Roe scheme is identified as bringing excessive numerical diffusion and several corrections are proposed. These corrections aim at reducing the numerical diffusion of the explicit schemes, as well as maintaining their stability.

In this paper we present a more general strategy that could be easily applied to simulate various multiphase models at low Mach number. Such a strategy is inspired by single phase analysis and is first tested on the compressible Navier–Stokes equations in the present paper. In order to reduce the numerical diffusion, we consider a scheme that is order two in space such as the implicit centered scheme, already studied for example in [9].

In Sect. 2, we briefly recall the mathematical model and the considered numerical schemes. In Sect. 4 we give numerical evidence that the centered implicit scheme is much less diffusive than the upwind scheme (whether explicit or implicit) and can capture low Mach vortices. In order to reduce the computational cost involved by the resolution of many linear systems, Sect. 3 presents preconditioning strategy based on the scaling of the linear system matrix. This strategy is based on the underlying hyperbolic operator and could be applied to other set of equations.

2 Mathematical model and Numerical method

2.1 Mathematical model

The model consists of the following three balance laws for the mass, the momentum and the energy:

$$\left\{\begin{array}{lllll} \frac{\partial\rho}{\partial t} & +\nabla.\mathbf{q} & & & =0\\ \frac{\partial\mathbf{q}}{\partial t} & +\nabla.\left(\mathbf{q}\otimes\frac{\mathbf{q}}{\rho}+p\mathbb{I}_d\right) & -\nu\Delta(\frac{\mathbf{q}}{\rho}) & & =0\\ \frac{\partial(\rho E)}{\partial t} & +\nabla.\left[(\rho E+p)\frac{\mathbf{q}}{\rho}\right] & -\lambda\Delta T & & =0 \end{array}\right. \tag{1}$$

where ρ is the density, $\mathbf{v}$ the velocity, $\mathbf{q}=\rho\mathbf{v}$ the momentum, p the pressure, ρe the internal energy, $\rho E=\rho e+\frac{||\mathbf{q}||^2}{2\rho}$ the total energy, T the absolute temperature, ν the viscosity and λ the thermal conductivity. We close the system (1) by the ideal gas law $p=(\gamma-1)\rho e$. For the sake of simplicity, we consider constant viscosity and conductivity, and neglect the contribution of viscous forces in the energy equation. By denoting $U=(\rho,\mathbf{q},\rho E)^t$ the vector of conserved variables, the Navier–Stokes system (1) can be written as a nonlinear system of conservation laws:

$$\frac{\partial U}{\partial t} + \nabla \cdot (\mathscr{F}^{conv}(U)) + \nabla \cdot \left(\mathscr{F}^{diff}(U)\right) = 0, \tag{2}$$

where $\mathscr{F}^{conv}(U) = \begin{pmatrix} \mathbf{q} \\ \mathbf{q} \otimes \frac{\mathbf{q}}{\rho} + p\mathbb{I}_d \\ (\rho E + p)\frac{\mathbf{q}}{\rho} \end{pmatrix}$, $\mathscr{F}^{diff}(U) = \begin{pmatrix} 0 \\ -\nu\nabla(\frac{\mathbf{q}}{\rho}) \\ -\lambda\nabla T \end{pmatrix}$.

2.2 Numerical method

The conservation form (2) enables to define the concept of weak solutions, which can be discontinuous ones. Discontinuous solutions such as shock waves are of great importance in transient calculations. In order to correctly capture shock waves, one needs a robust, low diffusive conservative scheme. The finite volume framework is the best appropriate setup to build such schemes as it enables to write discrete equations that express the conservation laws at each cell (see for example [1]).

We decompose the computational domain into N disjoint cells C_i with volume v_i. Two neighboring cells C_i and C_j have a common boundary ∂C_{ij} with area s_{ij}. We denote $N(i)$ the set of neighbors of a given cell C_i and $\mathbf{n}_{ij}$ the exterior unit normal vector of ∂C_{ij}. Integrating the system (2) over C_i and setting $U_i(t) = \frac{1}{v_i}\int_{C_i} U(x,t)dx$ and $U_i^n = U_i(n\Delta t)$, the discretized equations can be written:

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \sum_{j \in N(i)} \frac{s_{ij}}{v_i}\left(\overrightarrow{\Phi}_{ij}^{conv} + \overrightarrow{\Phi}_{ij}^{diff}\right) = 0. \tag{3}$$

with: $\overrightarrow{\Phi}_{ij}^{conv} = \frac{1}{s_{ij}}\int_{\partial C_{ij}} \mathscr{F}^{conv}(U).\mathbf{n}_{ij}\,ds$, $\overrightarrow{\Phi}_{ij}^{diff} = \frac{1}{s_{ij}}\int_{\partial C_{ij}} \mathscr{F}^{diff}(U).\mathbf{n}_{ij}\,ds$.

To approximate the convection numerical flux $\overrightarrow{\Phi}_{ij}^{conv}$ we solve an approximate Riemann problem at the interface ∂C_{ij}. Using the Roe local linearisation of the fluxes [2], we obtain the following formula:

$$\overrightarrow{\Phi}_{ij}^{conv} = \frac{\mathscr{F}^{conv}(U_i) + \mathscr{F}^{conv}(U_j)}{2}.\mathbf{n}_{ij} - \mathscr{D}(U_i, U_j)\frac{U_j - U_i}{2} \tag{4}$$

$$= \mathscr{F}^{conv}(U_i)\mathbf{n}_{ij} + A^-(U_i, U_j)(U_j - U_i), \tag{5}$$

where $\mathscr{D}$ is an upwinding matrix, $A(U_i, U_j)$ the Roe matrix and $A^- = \frac{A-\mathscr{D}}{2}$. The choice $\mathscr{D} = 0$ gives the centered scheme, whereas $\mathscr{D} = |A|$ gives the upwind scheme. For the Euler equations, we can build $A(U_i, U_j)$ explicitly using the Roe averaged state (see [1]).

The diffusion numerical flux $\overrightarrow{\Phi}_{ij}^{diff}$ is approximated on structured meshes using the formula:

$$\overrightarrow{\Phi}_{ij}^{diff} = D(\frac{U_i + U_j}{2})(U_j - U_i) \tag{6}$$

with the matrix $D(U) = \begin{pmatrix} 0 & \mathbf{0} & 0 \\ \frac{\nu \mathbf{q}}{\rho^2} & \frac{-\nu}{\rho}\mathbb{I}_d & 0 \\ \frac{\lambda}{c_v}\left(\frac{c_v T}{\rho} - \frac{||\mathbf{q}||^2}{2\rho^3}\right) & \frac{\mathbf{q}^t \lambda}{\rho^2 c_v} & -\frac{\lambda}{c_v \rho} \end{pmatrix}$, where c_v is the heat capacity at constant volume.

2.3 Newton scheme

Finally, since $\sum_{j \in N(i)} \mathscr{F}^{conv}(U_i).\mathbf{n}_{ij} = 0$, using (5) and (6) the equation (3) of the numerical scheme becomes:

$$\frac{U_i^{n+1} - U_i^n}{\Delta t} + \sum_{j \in N(i)} \frac{s_{ij}}{v_i} \{(A^- + D)(U_i^{n+1}, U_j^{n+1})\}(U_j^{n+1} - U_i^{n+1}) = 0. \tag{7}$$

The system (7) is nonlinear. We use the following Newton iterative method to obtain the required solutions:

$$\begin{aligned} &\frac{\delta U_i^{k+1}}{\Delta t} + \sum_{j \in N(i)} \frac{s_{ij}}{v_i} \left[(A^- + D)(U_i^k, U_j^k)\right] \left(\delta U_j^{k+1} - \delta U_i^{k+1}\right) \\ &\qquad = -\frac{U_i^k - U_i^n}{\Delta t} - \sum_{j \in N(i)} \frac{s_{ij}}{v_i} \left[(A^- + D)(U_i^k, U_j^k)\right] (U_j^k - U_i^k), \end{aligned}$$

where $\delta U_i^{k+1} = U_i^{k+1} - U_i^k$ is the variation of the k-th iterate that approximate the solution at time $n + 1$. Defining the unknown vector $\mathscr{U} = (U_1, \dots, U_N)^t$, each Newton iteration for the computation of $\mathscr{U}$ at time step $n+1$ requires the numerical solution of the following linear system:

$$\mathscr{A}(\mathscr{U}^k)\delta\mathscr{U}^{k+1} = b(\mathscr{U}^n, \mathscr{U}^k). \tag{8}$$

2.4 The low Mach problem

When the flow is smooth and the Mach number $\frac{||\mathbf{v}||}{c}$ (where $c = \sqrt{\frac{\gamma p}{\rho}}$ is the sound speed) is small, the solutions of the system (1) should behave as those of an incompressible Navier–Stokes model (see [10]). However, in general, Godunov type schemes do not preserve the asymptotic behavior and generate spurious solutions when applied to low Mach number flows (see [5]). The analysis presented in [5] suggests that the inaccuracies originate from the anisotropy of the upwind matrix $\mathscr{D}$, and various " Low Mach Schemes " are proposed in the explicit context. In order to avoid the stability issue, we propose to use implicit schemes and to consider the simpler case $\mathscr{D} = 0$ (no upwinding). The resulting centered scheme can be applied to any system of conservation law, and we present in Sect. 4 our first numerical experiments.

3 Description of the Scaling strategy

The larger the time step, the worse the condition number of the matrix $\mathscr{A}$ in (8). As a consequence, it is important to apply a preconditioner before solving the linear system. The most popular choice is the Incomplete LU factorisation (later named ILU, see [3] for more details). The error made by the approximate factorisation using an ILU preconditioner depends on the size of the off diagonal coefficients of the matrix. For a better performance of the preconditioner, it is desirable that off diagonal entries of the matrix have small magnitudes.

As we are interested in convection dominated flows, the main contributions to the matrix $\mathscr{A}$ come from the convective part discretisation of the equations through the matrix A^-. Unfortunately, the coefficients of the Roe matrix have very different magnitudes for low Mach number flows. Consequently, A^- and hence $\mathscr{A}$ have coefficients with very different magnitudes.

We are now going to detail a procedure that scales the matrix coefficients so that they have the same magnitude. The matrix A^- can be expressed using a complete eigenstructure decomposition of the Roe matrix: $A = \sum_k \lambda_k L^k \otimes R^k$. The three eigenvalues of the Roe matrix are $v_n + c$, v_n (multiplicity d), and $v_n - c$. As we are interested in flows at low Mach number, we can assume $\mathbf{v} \approx 0$ and in that case the eigenvalues of A become $\lambda^- = -c, \lambda_v = 0$, and $\lambda^+ = +c$. The right and left eigenvectors $R^{\pm}$ and $L^{\pm}$ associated to the sound waves are:

$$R^{\pm} = (1,\ \pm c\mathbf{n},\ \frac{c^2}{\gamma - 1})^t, \qquad L^{\pm} = \frac{1}{2}(0,\ \pm\frac{1}{c}\mathbf{n},\ \frac{\gamma-1}{c^2})^t. \tag{9}$$

We have:

$$\begin{aligned} A^- &= -cL^- \otimes R^- \qquad \text{for the upwind scheme,} \\ A^- &= \frac{1}{2}(cL^+ \otimes R^+ - cL^- \otimes R^-) \text{ for the centered scheme.} \end{aligned}$$

One sees from (9) that the disequilibrium in A^- coefficients comes from the difference in the magnitude of the components of the left and right eigenvectors of A.

If we multiply A^- to the left (respectively to the right) by a diagonal matrix with the coefficients $d_{sca} = diag(1, c\mathbf{n}, \frac{c^2}{\gamma-1})$, respectively $d_{sca}^{-1} = diag(1, \frac{1}{c}\mathbf{n}, \frac{\gamma-1}{c^2})$ ($\mathbf{n}$ is the unit normal vector), we obtain vectors and matrices with better balanced coefficients:

$$d_{sca}^{-1} R^{\pm} = (1, \pm\mathbf{n}, 1)^t, \qquad d_{sca} L^{\pm} = (0, \pm\mathbf{n}, 1)^t,$$

$$L^{\pm} \otimes R^{\pm} = \frac{1}{2}\begin{pmatrix} 0 & \mathbf{0} & 0 \\ \pm\frac{1}{c}\mathbf{n} & \mathbf{n}\otimes\mathbf{n} & \pm\frac{c}{\gamma-1}\mathbf{n} \\ \frac{\gamma-1}{c^2} & \pm\frac{\gamma-1}{c}\mathbf{n}^t & 1 \end{pmatrix}, \quad d_{sca} L^{\pm} \otimes R^{\pm} d_{sca}^{-1} = \frac{1}{2}\begin{pmatrix} 0 & \mathbf{0} & 0 \\ \pm\mathbf{n} & \mathbf{n}\otimes\mathbf{n} & \pm\mathbf{n} \\ 1 & \pm\mathbf{n}^t & 1 \end{pmatrix}$$

Any mesh can be associated with two diagonal matrices D_{sca} and D_{sca}^{-1} having the size of the mesh and containing the successive coefficients of the local matrices d_{sca} and d_{sca}^{-1}. Instead of solving system (8), one can rather solve:

$$\tilde{\mathscr{A}}\mathscr{V} = \tilde{b}, \tag{10}$$

where $\tilde{\mathscr{A}} = D_{sca}\mathscr{A}D_{sca}^{-1}$, $\mathscr{V} = D_{sca}\mathscr{U}$ and $\tilde{b} = D_{sca}b$. System (10) can be resolved more easily using an ILU preconditioner. Once the solution $\mathscr{V}$ is obtained we compute $D_{sca}^{-1}\mathscr{V}$ to obtain the original unknown vector $\mathscr{U}$.

4 Numerical results

4.1 Upwind scheme vs Centered one

Figures 1 and 2 present the streamlines of the steady state result obtained using either the upwind or the centered schemes to discretize the convective part of the Navier–Stokes equations with the fully implicit scheme presented in Sect. 2.2. Our test case is a lid driven cavity flow at Reynolds number 400 solved on a cartesian 50 × 50 cell mesh. This case is described in [4], with the correct solution given by an incompressible solver. The lid speed is $1\,m/s$, the maximum Mach number of the flow is 0.008. The Roe approximate Riemann solver [2] employed for the convection fluxes is known to have problem solving such low Mach number flows when the scheme is explicit, especially on multidimensional cartesian meshes (see [5]). It can be seen on Fig. 1 that the upwind scheme does not capture the correct streamlines. However, on Fig. 2, it can be seen that the implicit centered scheme is much less diffusive and captures the correct solution with its expected three vortices.

4.2 Assessment of the Scaling strategy

We now study the performance of our numerical methods on the same lid driven cavity test case presented in Sect. 4.1. In this section, we vary the time step

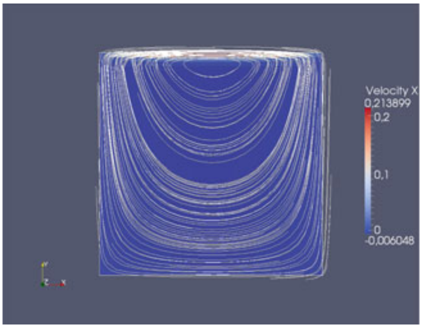


Fig. 1 Steady state, upwind scheme

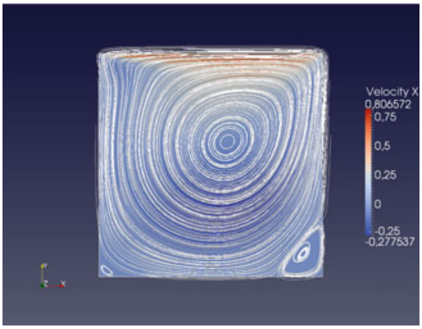


Fig. 2 Steady state, centered scheme

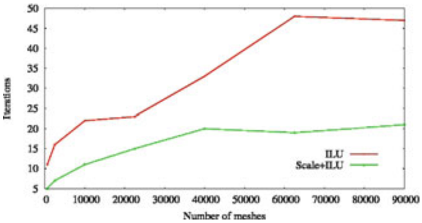


Fig. 3 Number of GMRES iterations for the upwind scheme, CFL 1000

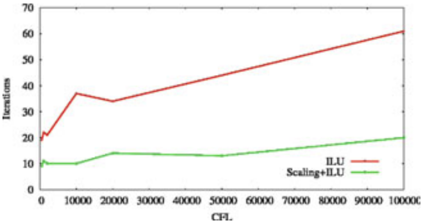


Fig. 4 Number of GMRES iterations for the upwind scheme, mesh 100 × 100

(CFL number) and the mesh size. We also compare the direct solver with the iterative one and the effect of different preconditioners on the resolution of the linear systems.

Considering first the upwind scheme, we remark that the ILU preconditioner with no level of fill-in performs well. Figs. 3 and 4, show the average number of GMRES iterations at each Newton iteration. We observe that the use of our Scaling strategy presented in Sect. 3 reduces more than twice the iteration number.

When we use the centered scheme, the system matrix has a poor diagonal, and ILU preconditioner with no fill-in is not efficient in preconditioning the linear system. One needs to use an incomplete factorisation with two levels of fill-in to solve linear system up to the CFL 100, and the Scaling strategy enables to save a considerable number of iterations (Fig. 5). Beyond that value, only a direct solver is able to solve the system. However, one can remark that the Scaling strategy enables a reduction of the number of Newton iterations using a direct solver (Fig. 6). We also stress that the steady state solution obtained with very large CFL numbers is still accurate and displays the expected vortices.

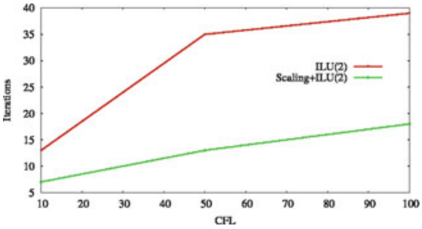


Fig. 5 Number of GMRES iterations for the centered scheme, mesh 50 × 50

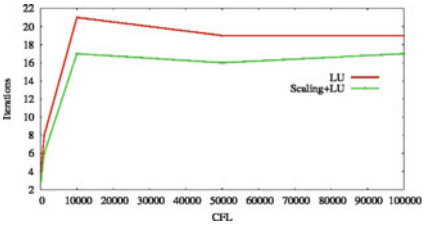


Fig. 6 Number of Newton iterations for the centered scheme, mesh 50 × 50s

5 Conclusion and Perspectives

In this paper, two simple and general fully implicit schemes have been presented for the simulation of compressible Navier–Stokes equations at low Mach number. We have shown that the centered scheme is able to capture low Mach vortices unlike the upwind scheme. However, ILU preconditioning performs better with the upwind scheme than with the centered scheme. Thanks to the particular features of Roe matrix for compressible Navier–Stokes equations, we have proposed a preconditioning strategy Scaling+ILU that considerably reduces the computation time. The centered scheme and the scaling strategy can be applied to other sets of equations than Navier–Stokes. Study of these techniques applied to two-phase flow models will follow.

References

1. E. Godlewski, P.A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws* Springer Verlag, 1996.
2. P.L Roe, Approximate Riemann solvers, parameter vectors and difference schemes *J. Comput. Phys.*, 43 (1981), 537-372.
3. Michele Benzi, Preconditioning Techniques for Large Linear Systems: A Survey *J. Comput. Phys.*, 182 (2002), 418-477.
4. U. Ghia, K.N. Ghia, C.T. Shin, High-Re Solutions for Incompressible Flow Using the Navier-Stokes Equations and a Multigrid Method *J. Comput. Phys.*, 48 (1982), p 387-411.
5. S. Dellacherie, Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number *J. Comput. Phys.*, 229(2010), 701-727.
6. I. Toumi, A. Bergeron, D. Gallo, and D. Caruge, FLICA-4: a three-dimensional two-phase flow computer code with advanced numerical methods for nuclear applications *Nucl. Eng. Design*, 200 (2000), p 139-155.
7. P. Fillion, A. Chanoine, S. Dellacherie, A. Kumbaro, FLICA-OVAP: a New Platform for Core Thermal-hydraulic Studies *NURETH-13* Japon, Sep 27-Oct 2, (2009).
8. H. Guillard, C. Viozat, On the behavior of upwind schemes in the low Mach number limit *Comput. Fluids* 28 (1999).

9. J.-A. Désidéri, P. W. Hemker, Convergence Analysis of the Defect-Correction Iteration for Hyperbolic Problems *SIAM J. Sci. Comput.*, Vol. 16 (1995), No.1, pp. 88-118.
10. S. Schochet, Fast Singular Limits of Hyperbolic PDEs *Journal of Differential Equations* 114 (1994).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

On the Godunov Scheme Applied to the Variable Cross-Section Linear Wave Equation

Stéphane Dellacherie and Pascal Omnes

Abstract We investigate the accuracy of the Godunov scheme applied to the variable cross-section acoustic equations. Contrarily to the constant cross-section case, the accuracy issue of this scheme in the low Mach number regime appears even in the one-dimensional case; on the other hand, we show that it is possible to construct another Godunov type scheme which is accurate in the low Mach number regime.

Keywords Variable cross-section, wave equation, low Mach, Godunov scheme
MSC2010: 35L05, 35L45, 65M08

1 Introduction

It is well-known that Godunov type schemes suffer from an accuracy problem at low Mach number. The analysis of this scheme applied to the linear wave equation has shown that this problem already occurs for such a simple submodel, except in the one-dimensional case [1]. However, it has also been proved that in higher dimensions, simplicial meshes perform much better than rectangular meshes [2]. These results are obtained by the analysis of the invariant space of the discrete wave operator: when this invariant space is rich enough to approach well the invariant space of the continuous wave operator (that is to say the incompressible fields), then the Godunov scheme is accurate at low Mach number. With the same type of analysis, we show in the present work that accuracy problems may already occur in the one-dimensional case for the variable cross-section linear wave equation, if one

Stéphane Dellacherie
CEA, DEN, DM2S, SFME F-91191 Gif-sur-Yvette, France, e-mail: stephane.dellacherie@cea.fr

Pascal Omnes
CEA, DEN, DM2S, SFME F-91191 Gif-sur-Yvette, France and LAGA, Université Paris 13, 99 Av. J.-B. Clément, F-93430 Villetaneuse, France, e-mail: pascal.omnes@cea.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_33, © Springer-Verlag Berlin Heidelberg 2011

is not careful about the expression of the diffusion terms inherent to the Godunov scheme. This equation may be seen as a simple model for diphasic flows in which the volumic fraction plays the role of the variable cross-section.

2 The variable cross-section wave equation

For regular solutions, the dimensionless barotropic Euler system with variable cross-section may be written as

$$\partial_t(A\rho) + \nabla\cdot(A\rho u) = 0 \qquad \text{and} \qquad \rho(\partial_t u + u\cdot\nabla u) + \frac{\nabla p}{M^2} = 0, \tag{1}$$

where the Mach number M is supposed to be small and where $p = p(\rho)$ with $p'(\rho) > 0$. Denoting by a_* a reference sound velocity, and setting

$$\rho(t,x) := \rho_*\left[1 + \frac{M}{Aa_*}s(t,x)\right], \tag{2}$$

system (1) may be written, after some simplifications

$$\partial_t q + \mathscr{H}(q) + \frac{\mathscr{L}_{A,M}}{M}(q) = 0 \tag{3}$$

with

$$\begin{cases} q = \big(s\ ,\ u\big)^T, & \text{(a)} \\ \mathscr{H}(q) = \big(\nabla\cdot(us)\ ,\ (u\cdot\nabla)u\big)^T, & \text{(b)} \\ \mathscr{L}_{A,M}(q) = \left(a_*\nabla\cdot(Au)\ ,\ \dfrac{p'[\rho_*(1+\frac{M}{Aa_*}s)]}{a_* + \frac{M}{A}s}\nabla\left(\dfrac{s}{A}\right)\right)^T. & \text{(c)} \end{cases} \tag{4}$$

2.1 The linear wave equation with variable cross-section

When A is bounded by below and by above independently of M, we formally have that $\frac{M}{Aa_*}s(t,x) \ll 1$ in (2) and $\mathscr{O}(||\mathscr{L}_{A,M}(q)||) = 1$ in (3) when $\mathscr{O}(||q||) = 1$. In that case, (3) contains a transport contribution whose characteristic time scale is a $\mathscr{O}(1)$ and a non-linear acoustic contribution whose characteristic time scale is a $\mathscr{O}(M)$, like in the usual barotropic low Mach number Euler system. In that case, at least in a first approach, we may drop the transport contribution and study the linearized cross-section acoustic equation which reads

$$\partial_t q + \frac{L_A}{M} q = 0 \qquad \text{with} \qquad L_A q = a_* \left(\nabla \cdot (Au) \, , \, \nabla \left(\frac{s}{A} \right) \right)^T . \tag{5}$$

3 Basic properties of the variable cross-section linear wave equation

3.1 General properties

In this section, we are interested in basic properties of (5) solved on a periodic torus $\mathbb{T}^{d \in \{1,2,3\}}$. For this, we define the energy space

$$(L^2_A(\mathbb{T}^d))^{1+d} := \left\{ q := \left(s \, , \, u \right)^T \text{ such that } \int_{\mathbb{T}^d} s^2 \frac{dx}{A(x)} + \int_{\mathbb{T}^d} |u|^2 A(x) dx < +\infty \right\}$$

endowed with the scalar product

$$\langle q_1, q_2 \rangle_A = \int_{\mathbb{T}^d} s_1 s_2 \frac{dx}{A(x)} + \int_{\mathbb{T}^d} u_1 \cdot u_2 \, A(x) dx. \tag{6}$$

On the other hand, we set

$$\begin{cases} \mathscr{E}_A = \left\{ q := \left(s \, , \, u \right)^T \in (L^2_A(\mathbb{T}^d))^{1+d} \text{ such that } s = aA, \, a \in \mathbb{R} \text{ and } \nabla \cdot (Au) = 0 \right\}, \\ \mathscr{E}^{\perp} = \left\{ q := \left(s \, , \, u \right)^T \in (L^2_A(\mathbb{T}^d))^{1+d} \right. \\ \qquad\qquad \left. \text{such that } \int_{\mathbb{T}^d} s dx = 0 \text{ and } \exists \phi \in H^1(\mathbb{T}^d), u = \nabla \phi \right\}. \end{cases}$$

We remark that $\mathscr{E}_A \perp \mathscr{E}^{\perp}$ for the scalar product (6). We shall admit the following extension of the Hodge decomposition $(L^2_A(\mathbb{T}^d))^{1+d} = \mathscr{E}_A \oplus \mathscr{E}^{\perp}$. Moreover, we have

$$\mathscr{E}_A = KerL_A. \tag{7}$$

Finally, for all $q \in (L^2_A(\mathbb{T}^d))^{1+d}$, we define the energy $E_A := \langle q, q \rangle_A$. The following lemma is an easy extension of the energy conservation property to the variable cross-section case:

Lemma 1. *Let $q(t,x)$ be the solution of (5) on $\mathbb{T}^{d \in \{1,2,3\}}$. Then:*

$$E_A(t \geq 0) = E_A(t = 0).$$

We also have the following result:

Lemma 2. *Let $q(t,x)$ be the solution of (5) on $\mathbb{T}^{d\in\{1,2,3\}}$ with initial condition q^0. Then:*
1) $\forall q^0 \in \mathscr{E}_A : q(t \geq 0) \in \mathscr{E}_A$;
2) $\forall q^0 \in \mathscr{E}^{\perp} : q(t \geq 0) \in \mathscr{E}^{\perp}$.

Proof of Lemma 2: The first point is a direct consequence of (7). The second point is inferred from the first item, from Lemma 1, and from the following Lemma, a proof of which may be found in the appendix A of [1].

Lemma 3. *Let $\mathscr{A}$ be a linear isometry in a Hilbert space $\mathbb{H}$ and let $\mathscr{E}$ be a linear subspace of $\mathbb{H}$. Then:* $\qquad \mathscr{A}\mathscr{E} = \mathscr{E} \qquad \Longrightarrow \qquad \mathscr{A}\mathscr{E}^{\perp} \subset \mathscr{E}^{\perp}$.

3.2 The one-dimensional case

In the particular case of the one-dimensional geometry, equation (5) is now set in $\mathbb{T}^{d=1}$ and writes

$$\partial_t q + \frac{L_A}{M} q = 0 \tag{8}$$

with

$$L_A q = a_* \left(\partial_x (Au) \ , \ \partial_x \left(\frac{s}{A} \right) \right)^T . \tag{9}$$

The subspaces $\mathscr{E}_A$ and $\mathscr{E}^{\perp}$ are now characterized by

$$\begin{cases} \mathscr{E}_A = \left\{ q := \left(s \ , \ u \right)^T \in (L^2_A(\mathbb{T}^1))^2 \text{ such that } s = aA \text{ and } u = \dfrac{b}{A}, \ (a,b) \in \mathbb{R}^2 \right\}, \\ \mathscr{E}^{\perp} = \left\{ q := \left(s \ , \ u \right)^T \in (L^2_A(\mathbb{T}^1))^2 \text{ such that } \displaystyle\int_{\mathbb{T}^d} s \, dx = \int_{\mathbb{T}^d} u \, dx = 0 \right\} . \end{cases}$$

In the one-dimensional case, we remark that, as soon as $A'(x) \neq 0$, the variables s and u do not play the same role, while when $A = 1$, they do play symmetrical roles.

4 Numerical approximation in the one-dimensional geometry

We now consider the numerical approximation of (8)–(9) on a mesh with N cells $[x_{i-1/2}, x_{i+1/2}]$ of constant size Δx. We denote by x_i the midpoints of the cells and by $u_i(t)$ and $s_i(t)$ the numerical approximation of u and s in the cell $[x_{i-1/2}, x_{i+1/2}]$.

4.1 A first numerical scheme

Integrating (8) over the cell $[x_{i-1/2}, x_{i+1/2}]$, we obtain

$$\begin{cases} \dfrac{d}{dt}s_i + \dfrac{a_*}{M} \cdot \dfrac{A_{i+1/2}u_{i+1/2}(t) - A_{i-1/2}u_{i-1/2}(t)}{\Delta x} = 0, \\ \dfrac{d}{dt}u_i + \dfrac{a_*}{M} \cdot \dfrac{\dfrac{s_{i+1/2}(t)}{A_{i+1/2}} - \dfrac{s_{i-1/2}(t)}{A_{i-1/2}}}{\Delta x} = 0, \end{cases} \tag{10}$$

where $A_{i+1/2} := A(x_{i+1/2})$ and where the interface values $(s_{i+1/2}(t), u_{i+1/2}(t))$ are determined by the solution of a Riemann problem (R.P.) based on the equation

$$\partial_t q + \frac{a_*}{M}\left(A_{i+1/2}\partial_x u, \frac{1}{A_{i+1/2}}\partial_x s\right)^T = 0,$$

which amounts to locally neglect the variations of A in (8). The left and right initial states of the R. P. being $(s_i(t), u_i(t))$ and $(s_{i+1}(t), u_{i+1}(t))$ respectively, its solution is

$$\begin{cases} s_{i+1/2} = \frac{1}{2}(s_i + s_{i+1}) + \frac{A_{i+1/2}}{2}(u_i - u_{i+1}), \\ u_{i+1/2} = \frac{1}{2A_{i+1/2}}(s_i - s_{i+1}) + \frac{1}{2}(u_i + u_{i+1}). \end{cases} \tag{11}$$

Plugging (11) into (10), we obtain the following scheme

$$\begin{cases} \dfrac{d}{dt}s_i + \dfrac{a_*}{M} \cdot \dfrac{A_{i+1/2}(u_i + u_{i+1}) - A_{i-1/2}(u_{i-1} + u_i)}{2\Delta x} = \dfrac{a_*}{2M\Delta x}(s_{i+1} - 2s_i + s_{i-1}), \\ \dfrac{d}{dt}u_i + \dfrac{a_*}{M} \cdot \dfrac{\dfrac{(s_i + s_{i+1})}{A_{i+1/2}} - \dfrac{(s_{i-1} + s_i)}{A_{i-1/2}}}{2\Delta x} = \dfrac{a_*}{2M\Delta x}(u_{i+1} - 2u_i + u_{i-1}), \end{cases} \tag{12}$$

whose first-order modified equation is given by

$$\partial_t q + \frac{L_A}{M} q = \left(\nu_s \partial^2_{xx} s \ , \ \nu_u \partial^2_{xx} u\right)^T \tag{13}$$

with $(\nu_s, \nu_u) = \frac{a_* \Delta x}{2M}(1, 1)$. This shows that for all non trivial $(\nu_s, \nu_u) \in \mathbb{R}^2$, the space $\mathscr{E}_A$ is no more invariant as soon as $A' \neq 0$. In particular, even when $\nu_u = 0$, the space $\mathscr{E}_A$ is not invariant as soon as $A' \neq 0$: this property stresses the fact that the Godunov scheme, as well as its low Mach modification obtained by simply removing the dissipative term in the right-hand side of the second equation of (12) like in [1, 2], may not be accurate at low Mach number, including in the 1D case, contrarily to the case $A' = 0$.

4.2 Study of a second numerical scheme

In order to propose a numerical scheme which will be accurate at low Mach number, we proceed like in [1]. That is to say:

- First, we try to modify the diffusion term in (13) such that the new equation preserves the invariance of $\mathscr{E}_A$.
- Then, we identify a numerical scheme whose modified equation corresponds to the equation with the new diffusion term.

To these two points, we add something new with respect to what is done in [1]: we shall show that it is possible to define a Godunov type scheme which corresponds to the numerical scheme proposed in the second point above. This stresses the fact that it is possible to build a particular Godunov scheme that is accurate at low Mach number for the linear wave equation with variable cross-section, if one discretizes equation (8) in a adequate set of variables. Another interest of this scheme is that it doesn't suffer from any checkerboard mode (see [3] when $A = 1$).

4.2.1 Modification of the diffusion term

Let us replace the diffusion term

$$\left(\nu_s \partial^2_{xx} s \ , \ \nu_u \partial^2_{xx} u \right)^T \tag{14}$$

in equation (13) by the diffusion term

$$\left(\nu_s \partial_x \left[A \partial_x \left(\frac{s}{A} \right) \right] \ , \ \nu_u \partial_x \left[\frac{1}{A} \partial_x \left(A u \right) \right] \right)^T \tag{15}$$

with $(\nu_s, \nu_u) \in \mathbb{R}^2$. Then, by construction, the space $\mathscr{E}_A$ is invariant for equation

$$\partial_t q + \frac{L_A}{M} q = \left(\nu_s \partial_x \left[A \partial_x \left(\frac{s}{A} \right) \right] \ , \ \nu_u \partial_x \left[\frac{1}{A} \partial_x \left(A u \right) \right] \right)^T . \tag{16}$$

Moreover, we have the following result:

Lemma 4. *Let $q(t,x)$ be solution of (16) over $\mathbb{T}^1$. Then:*

$$E_A(t \geq 0) \leq E_A(t = 0).$$

A numerical scheme associated to (16) is then likely to be stable.

Proof of Lemma 4: Let $q(t,x)$ be solution of (16). There holds

$$\frac{1}{2}\frac{d}{dt}E_A = \nu_s \int_{\mathbb{T}^d} s\partial_x\left[A\partial_x\left(\frac{s}{A}\right)\right]\frac{dx}{A(x)} + \nu_u\int_{\mathbb{T}^d} u\partial_x\left[\frac{1}{A}\partial_x\left(Au\right)\right]A(x)dx$$

$$= -\nu_s\int_{\mathbb{T}^d}\left[\partial_x\left(\frac{s}{A}\right)\right]^2 A(x)dx - \nu_u\int_{\mathbb{T}^d}\left[\partial_x\left(Au\right)\right]^2\frac{dx}{A(x)} \quad \leq 0.$$

This proves that $E_A(t \geq 0) \leq E_A(t=0)$.□

4.2.2 Identifying the numerical scheme

A numerical scheme whose modified equation corresponds to (16) is given by

$$\begin{cases} \dfrac{d}{dt}s_i + \dfrac{a_*}{M}\cdot\dfrac{A_{i+1}u_{i+1} - A_{i-1}u_{i-1}}{2\Delta x} = \\ \qquad \dfrac{a_*}{2M\Delta x}\left[\dfrac{A_{i+1/2}}{A_{i+1}}s_{i+1} - \left(\dfrac{A_{i+1/2}+A_{i-1/2}}{A_i}\right)s_i + \dfrac{A_{i-1/2}}{A_{i-1}}s_{i-1}\right] & \text{(a)} \\ \dfrac{d}{dt}u_i + \dfrac{a_*}{M}\cdot\dfrac{\dfrac{s_{i+1}}{A_{i+1}} - \dfrac{s_{i-1}}{A_{i-1}}}{2\Delta x} = \\ \qquad \dfrac{a_*}{2M\Delta x}\left[\dfrac{A_{i+1}}{A_{i+1/2}}u_{i+1} - A_i\left(\dfrac{1}{A_{i+1/2}} + \dfrac{1}{A_{i-1/2}}\right)u_i + \dfrac{A_{i-1}}{A_{i-1/2}}u_{i-1}\right] & \text{(b)} \end{cases}, \tag{17}$$

where $A_i := A(x_i)$. A discrete analogue of Lemma 4 may be proved through "discrete integration by parts" and shows that the scheme is stable and that the discrete invariant space is the set

$$\mathscr{E}_A^h = \left\{ q := \left(s\ ,\ u\right)^T \in (\mathbb{R}^N)^2 \text{ such that } s_i = aA_i \text{ and } u_i = \frac{b}{A_i}\ ,\ (a,b)\in\mathbb{R}^2\right\},$$

which admits the orthogonal set

$$(\mathscr{E}^h)^\perp = \left\{ q := \left(s\ ,\ u\right)^T \in (\mathbb{R}^N)^2 \text{ such that } \sum_{i=1}^N s_i = \sum_{i=1}^N u_i = 0\right\}$$

for the discrete scalar product $\langle q_1, q_2\rangle_A^h := \sum_{i=1}^N \Delta x\left(\frac{(s_1)_i(s_2)_i}{A_i} + (u_1)_i(u_2)_iA_i\right).$

4.2.3 The associated Godunov scheme

It is possible to obtain scheme (17) from (10) by the following process: we set

$$\widetilde{q} := \left(r\ ,\ j\right)^T\ ,\ r := \frac{s}{A}\ ,\ j := Au$$

and solve the Riemann Problem based on the equation

$$\partial_t \widetilde{q} + \frac{a_*}{M}\left(\partial_x\left(\frac{j}{A_{i+1/2}}\right)\ ,\ \partial_x(A_{i+1/2}r)\right)^T = 0$$

with initial left and right states given by $\left(\frac{s_i}{A_i}, A_i u_i\right)^T$ and $\left(\frac{s_{i+1}}{A_{i+1}}, A_{i+1}u_{i+1}\right)^T$ respectively. This provides an expression for $\left(r_{i+1/2}, j_{i+1/2}\right)^T$ which is plugged into (10) for the evaluation of $\frac{s_{i+1/2}}{A_{i+1/2}}$ and $A_{i+1/2}u_{i+1/2}$, and we obtain (17).

5 Numerical results in 1D

In this section, we chose $A(x) = \frac{1}{4}\sin(2\pi x) + \frac{1}{2}$. As an initial condition, we choose $s^0(x) = A(x)$ and $u^0(x) = 1/A(x)$. At the discrete level, we choose $s_i^0 = A(x_i)$ and $u_i^0 = 1/A(x_i)$, so that the initial condition belongs to $\mathscr{E}_A^h$. Then, with (17), this initial condition is left unchanged for all times, as is the case with the continuous solution. On the other hand, with (12), the solution $(s_i(t), u_i(t))_{i\in[1,N]}^T$ has a non zero component in the space $(\mathscr{E}^h)^\perp$ as soon as $t > 0$, which may be computed by an orthogonal projection. Figure 1 shows the discrete weighted L^2

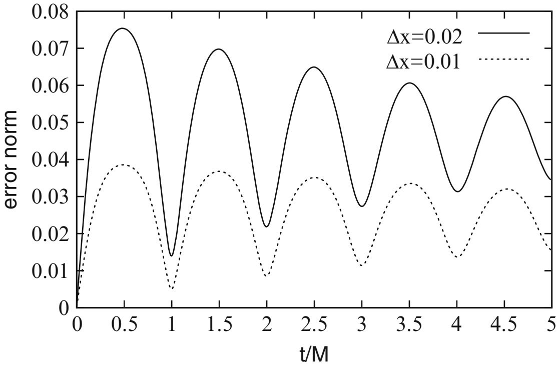


Fig. 1 norm of the spurious component for $M = 10^{-4}$ as a function of t/M

norm of this spurious component in $(\mathscr{E}^h)^\perp$ as a function of time scaled by M, with $M = 10^{-4}$ and for two different values of Δx. The size of the spurious wave grows up from 0 at $t = 0$ to $\mathscr{O}(\Delta x)$ at $t = \mathscr{O}(M)$, which is much greater than $\mathscr{O}(M)$, since $M \ll \Delta x$.

References

1. Dellacherie, S.: Analysis of Godunov type schemes applied to the compressible Euler system at low Mach number. J. Comp. Phys. **229**(4), 978–1016 (2010)
2. Dellacherie, S., Omnes, P., Rieper, F.: The influence of cell geometry on the Godunov scheme applied to the linear wave equation. J. Comp. Phys. **229**(14), 5315–5338 (2010)
3. Dellacherie, S.: Checkerboard modes and wave equation. In: Proc. of the Algoritmy 2009 Conference on Scientic Computing (March 15-20, 2009, Vysoke Tatry, Podbanske, Slovakia), pp. 71-80 (2009)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Towards stabilization of cell-centered Lagrangian methods for compressible gas dynamics

Bruno Després and Emmanuel Labourasse

Abstract We propose a sub-cell procedure for the stabilization of cell-centered Lagrangian numerical schemes for the computation of compressible gas dynamics. This procedure is intended to stabilize the mesh, indeed cell-centered schemes are already stable for shocks since they are based on a Riemann solver technology. In this work we focus on the basic principles and on the compatibility with the entropy. We show that a sub-cell decomposition into four triangles is always mesh-stable provided the scheme is entropy increasing. Numerical examples serve as illustration. We also discuss the consistency issue.

Keywords Finite Volume, Lagrangian scheme, stabilization
MSC2010: 65M08, 65M12, 65Z99, 76M12

1 Introduction

Cell centered Finite Volume numerical methods for the calculation of the equations of Lagrangian gas dynamics [1]

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \mathbf{u}) = 0, \\ \partial_t (\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p = 0, \\ \partial_t (\rho e) + \nabla \cdot (\rho \mathbf{u} e + p \mathbf{u}) = 0, \\ \partial_t (\rho S) + \nabla \cdot (\rho \mathbf{u} S) \geq 0. \end{cases} \tag{1}$$

Bruno Després
Lab. LJLL-UPMC, France, e-mail: despres@ann.jussieu.fr

Emmanuel Labourasse
CEA, DAM, DIF, 91297 Arpajon, France

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_34, © Springer-Verlag Berlin Heidelberg 2011

receive increasing interest nowadays due to three facts: 1) these are cell centered methods, like any Finite Volume scheme this is easy to handle in a multidimensional code [4, 9]; 2) in the context of compressible gas dynamics new corner based cell centered Riemann solvers have been developed which make these methods appropriate for shock calculations; 3) remeshing and projection techniques are easy to developed for such algorithms (but not that easy to optimize). However these methods suffer for the clue of all numerical methods on moving grids [2, 8, 10]. The mesh may become pathological (tangling) due to physical features of the flow (vortex, shear). Spurious modes, like checkerboard modes, may indeed be responsible of local negative volumes. This problem does not show up in dimension one but is the rule in dimension two and three.

In this paper, we focus on the underlying subcell finite volume structure in the context of finite Volume discretization and on possible ideas which can be used to develop an unconditionally stable lagrangian algorithm for compressible gas dynamics. Notice that other stabilization processes have been developed for Lagrangian hydrodynamics discretization [2], using subcell modeling [3] and also for ALE techniques which are another way to stabilize Lagrangian calculations [7].

2 An example

In order to establish a guideline for further developments, we want first to consider the example of a Sod shock tube problem computed with two different meshes and with the GLACE scheme [4]. We display a zoom in the shocked zone is in Fig. 1.

One sees a difference between triangles and quadrangles. Meshes made with triangles are stable in the sense that the volume of all cells always remain positive. Numerical observation with quadrangles show that the total volume is also positive, but the local volumes may become negative. This phenomenon is generated by the numerical scheme used to move the mesh. In our case the scheme is the cell-centered Glace scheme. But this behavior is common to any Lagrangian scheme.

In what follows we propose to introduce some aspects of the computation with triangles into the computation with quadrangles in order to improve the robustness.

3 Subcell modeling

To overcome the difficulties encountered with quadrangles we propose to consider a subcell modeling. The idea is to consider one cell at time step n. The volume of the cell is referred to as V^n. The total mass in the cell is

$$M = V^n \rho^n.$$

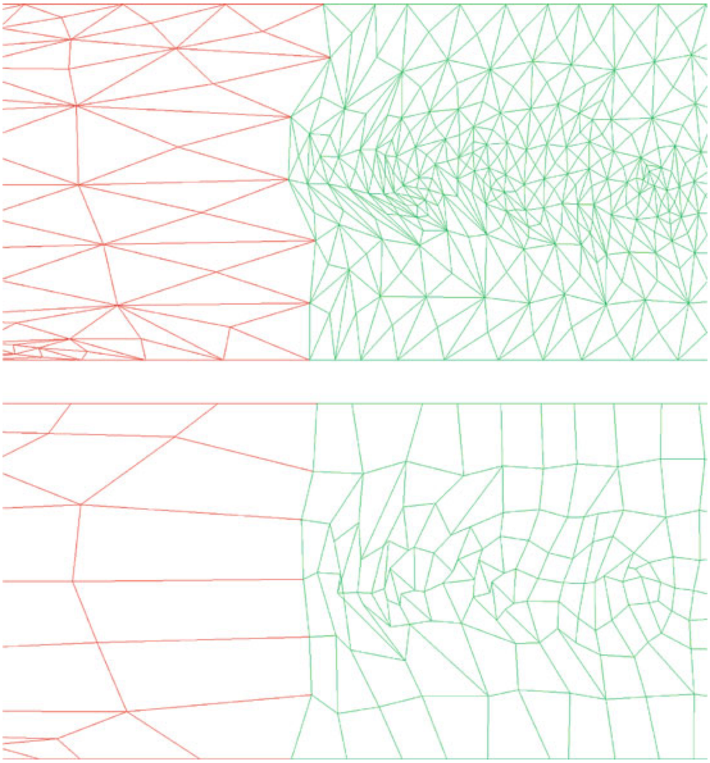


Fig. 1 The first mesh on the top is made with triangles: all cells have a positive volume. The second mesh on the bottom is made with quadrangles: pathological cells are visible to rows on the right of the interface between the coarse mesh and the fine mesh. The calculation stops

Notice that the total mass does not depend upon n since the scheme is Lagrangian. The momentum in the cell is the vector

$$\mathbf{I}^n = V^n \rho^n \mathbf{u}^n$$

and the total energy is

$$E^n = V^n \rho^n e^n.$$

The internal energy is

$$\varepsilon^n = E^n - \frac{1}{2} |\mathbf{I}^n|^2.$$

Next we split the cell into triangles. For example the quadrangle of Fig. 2 is split into four triangles $\overline{V} = \overline{T_1 \cup T_2 \cup T_3 \cup T_4}$ where the center O is simply the average of the corners

$$0 = \frac{1}{4} (A + B + C + D).$$

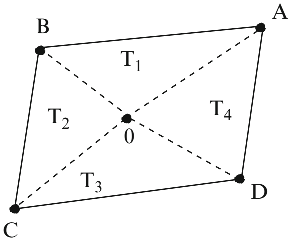


Fig. 2 Decomposition of a quad into four triangles

Then we decide arbitrarily that the total mass is split into four equal parts and affected in each triangle

$$m_i = \frac{1}{4}M$$

and the same for the internal energy

$$\varepsilon_i^n = \frac{1}{4}\varepsilon^n.$$

In our mind it is absolutely essential to characterize this very simple operation at the thermodynamical level, that is for the entropy variable. We present here a somewhat naive procedure for doing this.

Assume for example a perfect gas pressure law $p = (\gamma - 1)\rho\varepsilon$ and $S = \log(\varepsilon\rho^{-\gamma})$. The entropy in T_i is

$$S_i = \log\left(\varepsilon_i\rho_i^{-\gamma}\right).$$

The density in triangle T_i is

$$\rho_i = \frac{m_i}{|T_i|} = \frac{1}{4}\frac{m}{|T_i|} = \frac{1}{4}f_i\rho$$

where $f_i = \frac{|T_i|}{V}$ is the volume fraction. So the entropy in T_i is

$$S_i = \log(\varepsilon\rho^{-\gamma}) - \gamma\log f_i + (\gamma - 1)\log 4.$$

The new total entropy in the cell is

$$\overline{S} = \frac{S_1 + S_2 + S_3 + S_4}{4} = S - \frac{\gamma}{4}\sum_i \log f_i + (\gamma - 1)\log 4.$$

This very basic example shows that subcell modeling is somehow equivalent to modifying the entropy S into a new entropy variable

$$\overline{S} = S + \varphi(f_1, \ldots, f_4). \tag{2}$$

Definition 1. For any pressure law, we say that the subcell model (2) is entropy consistent if the function φ satisfies two conditions: it is a concave function and

$$f_1 f_2 f_3 f_4 = 0 \text{ implies that } \varphi = -\infty.$$

We say that φ is a subzonal entropy.

In order to stabilize a given Lagrangian computation, the general principle is to introduce a subzonal entropy in the numerical method which can be either staggered [2] or cell-centered as in our case [4]. The next proposition explains a fundamental advantage of subzonal entropies.

Proposition 1. *Consider a Lagrangian scheme which has the property to be entropy consistent, that the entropy in the cell increases from time step n to time step $n+1$ $S^{n+1} \geq S^n$.*

Assume that we are able to modify this scheme in order to introduce a subzonal entropy in a way such that

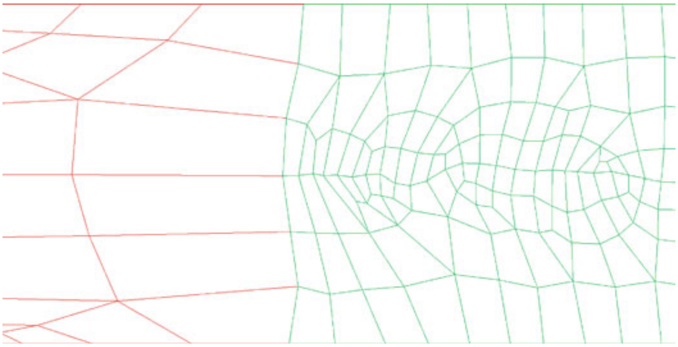
$$\overline{S}^{n+1} \geq \overline{S}^{n}. \tag{3}$$

Then the mesh is never pathological.

Indeed by continuity a pathological mesh is such that one volume fraction tends to zero, becomes zero and after become negative. In this case $\varphi^{n+1} \approx -\infty$ and it is in contradiction with the assumption (3).

Once the general framework of a thermodynamically consistent subcell model has been identified, it remains to use these new subzonal entropies in the scheme. In our case we focus on the GLACE scheme [4] which is cell-centered and of Finite Volume type. It is quite technical and there are many options, this is why we prefer to skip this issue in this presentation. However we managed to stabilize some computations which were unstable before. We present a numerical result that has been computed with the function $\varphi = \sum_i \log f_i$ in order to illustrate the numerical performances of the method proposed in this work. One sees on Fig. 3 that the robustness of the simulation is achieved with a result which is still physically correct.

Another major point is the convergence (as the mesh size tends to zero) of the new scheme with the new entropy. Indeed such a modification could be a source of major errors and of some fundamental inconsistency with the real problem (1). However since the proposed procedure is compatible with the idea of a geometric sub-cell modeling which is a kind of interpolation between a computation with quadrangles and a computation between triangles, it is reasonable to think the numerical solution is indeed consistent. At least the numerical test displayed in Fig. 3 shows the correctness of the result. In some situations it is also possible to show that this procedure is weakly consistent as proposed in [5]: the key property is to show consistency in the mimetic sense of the scheme [6].



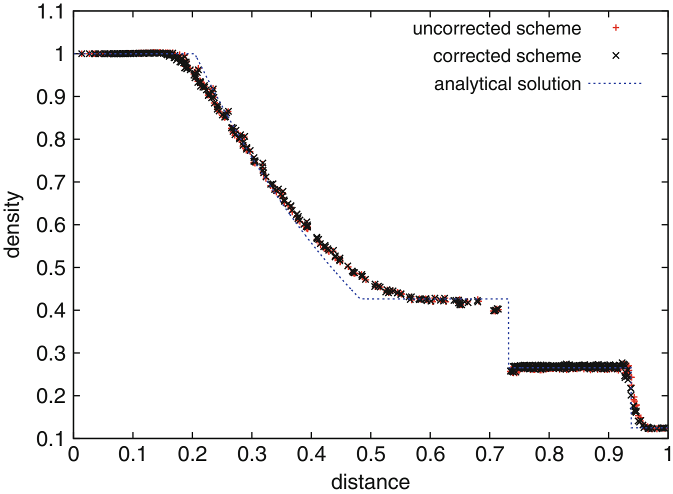


Fig. 3 Top: the quad mesh of Fig. 1 with the subzonal entropy; the mesh is no more pathological. Bottom: we plot the reference density of the Sod shock tube problem, the density calculated with the classical scheme and the density calculated with the subzonal entropy: there all agree which means that the subzonal entropy does not perturb the accuracy for this particular problem

Acknowledgements he first author kindly acknowledges the support of CEA.

References

1. D.J. Benson. Computational methods in Lagrangian and Eulerian hydrocodes. *Comp. Meth. Appl. Mech. Eng.*, 99:235–394, 1992.
2. E.J. Caramana, D.E. Burton, M.J. Shashkov, and P.P. Whalen. The construction of compatible hydrodynamics algorithms utilizing conservation of total energy. *J. Comput. Phys.*, 146:227–262, 1998.

3. E.J. Caramana and M.J. Shashkov. Elimination of Artificial Grid Distortion and Hourglass-Type Motions by Means of Lagrangian Subzonal Masses and Pressures. *J. Comput. Phys.*, 142:521–561, 1998.
4. G. Carré, S. Delpino, B. Després, and E. Labourasse. A cell-centered Lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. *J. Comput. Phys.*, 228:5160–5183, 2009.
5. B. Després. Weak consistency of the cell-centered lagrangian glace scheme on general meshes in any dimension. *CMAME*, 199:2669–2679, 2010.
6. B. Després and E. Labourasse. Subzonal entropy stabilization of cell-centered lagrangian methods. in preparation.
7. Liska, Shashkov, Vchal, and Wendroff. Optimization-based synchronized flux-corrected conservative interpolation (remapping) of mass and momentum for arbitrary lagrangian-eulerian methods. *J. Comput. Phys.*, 229(5):1467–1497, 2010.
8. R. Loubere, J. Ovadia, and R. Abgrall. A Lagrangian discontinuous Galerkin type method on unstructured meshes to solve hydrodynamics problems. *Int. J. Num. Meth. in Fluids*, 2000.
9. P.H. Maire, R. Abgrall, J. Breil, and J. Ovadia. A cell-centered lagrangian scheme for 2D compressible flow problems. *Siam J. Sci. Comp.*, 29, 2007.
10. G. Scovazzi, E. Love, and M.J. Shashkov. Multi-scale Lagrangian shock hydrodynamics on Q1/P0 finite elements: Theoretical framework and two-dimensional computations. *Comp. Meth. in Applied Mech. and Eng.*, 197:1056–1079, 2008.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Hybrid Finite Volume Discretization of Linear Elasticity Models on General Meshes

Daniele A. Di Pietro, Robert Eymard, Simon Lemaire, and Roland Masson

Abstract This paper presents a new discretization scheme for linear elasticity models using only one degree of freedom per face corresponding to the normal component of the displacement. The scheme is based on a piecewise constant gradient construction and a discrete variational formulation for the displacement field. The tangential components of the displacement field are eliminated using a second order linear interpolation. Our main motivation is the coupling of geomechanical models and porous media flows arising in reservoir or CO2 storage simulations. Our scheme guarantees by construction the compatibility condition between the displacement discretization and the usual cell centered finite volume discretization of the Darcy flow model. In addition it applies on general meshes possibly non conforming such as Corner Point Geometries commonly used in reservoir and CO2 storage simulations.

Keywords Hybrid finite volumes, linear elasticity, general meshes
MSC2010: 74S10, 74B05

1 Introduction

The oil production in unconsolidated, highly compacting porous media (such as Ekofisk or Bachaquero) induces a deformation of the pore volume which (i) modifies significantly the production, and (ii) may have severe consequences such as surface subsidence or damage of well equipments. This explains the growing interest in reservoir modeling for simulations coupling the reservoir Darcy multiphase flow with the geomechanical deformation of the porous media [3]. Similarly,

D.A. Di Pietro, S. Lemaire, and R. Masson
IFP Énergies nouvelles, FRANCE, e-mail: dipietrd, simon.lemaire, roland.masson@ifpen.fr

R. Eymard
Université Paris-Est Marne-la-Vallée, FRANCE, e-mail: robert.eymard@univ-mlv.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_35, © Springer-Verlag Berlin Heidelberg 2011

poromechanical models are also used in CO2 storage simulations to predict the over pressure induced by the injection of CO2 in order to assess the mechanical integrity of the storage in the injection phase.

The most commonly used geometry in reservoir and CO2 storage models is the so called Corner Point Geometry or CPG [7]. Although the CPG discretization is initially a structured hexahedral grid, vertical edges of the cells may typically collapse to account for the erosion of the geological layers and vertices may be dedoubled and slide along the coordlines (*i.e.* straight lines orthogonal to the geological layers) to model faults. In addition non conforming local grid refinement is used in near well regions. The resulting mesh is unstructured, non conforming, it includes all the degenerate cells obtained from hexahedra by collapsing edges, and hence it is not adapted to conforming finite element discretizations.

The objective of this paper is to introduce a new discretization scheme for linear elasticity equations which should

- apply on general meshes possibly non conforming;
- guarantee the stability of the coupling with Darcy flow models using cell centered finite volume discretization for the Darcy equation [1].

Our discretization is based on the family of Hybrid Finite Volume schemes introduced for diffusion problems on general meshes in [5] and also closely related to Mimetic Finite Difference schemes [6] as shown in [4]. The degrees of freedom are defined by the displacement vector $\mathbf{u}_\sigma$ at the center of gravity of each face σ of the mesh as well as the displacement vector $\mathbf{u}_K$ at a given point $\mathbf{x}_K$ of each cell K of the mesh. Following [5], a piecewise constant gradient is built and can be readily used to define the discrete variational formulation which mimics the continuous variational formulation for the displacement vector field. In order to stabilize this formulation and to reduce the degrees of freedom, the tangential components of the displacement are interpolated in terms of the neighbouring normal components. Numerical experiments on two dimensional and three dimensional meshes show that the resulting discretization is stable and convergent. In addition, this discretization satisfies by construction the compatibility condition or LBB (see [2], [1]) condition for poroelastic models when coupled with a cell centered finite volume scheme for the Darcy flow equation.

2 Discretization of Linear Elasticity Models

Let Ω be a d-dimensional polygonal or polyhedral domain ($d = 2$ or 3) and let us consider the following linear elasticity problem in infinitesimal strain theory:

$$\begin{cases} \mathbf{div}\,(\sigma(\mathbf{u})) + \mathbf{f} = \mathbf{0} & \text{on } \Omega, \\ \mathbf{u} = \mathbf{u}^D & \text{on } \partial\Omega^D, \\ \sigma(\mathbf{u}) \cdot \mathbf{n} = \mathbf{g} & \text{on } \partial\Omega^N, \end{cases} \tag{1}$$

where $\mathbf{u} \in \mathbb{R}^d$ is the unknown displacement field and D and N the two exponents standing respectively for Dirichlet and Neumann boundary conditions. $\sigma(\mathbf{u})$ is the Cauchy stress tensor and is given by Hooke's law $\sigma(\mathbf{u}) = 2\mu\varepsilon(\mathbf{u}) + \lambda \mathrm{tr}\,(\varepsilon(\mathbf{u}))\,\mathbf{1}$, where μ, λ are the Lamé parameters and $\varepsilon(\mathbf{u}) = \frac{1}{2}\left(\nabla\mathbf{u} + \nabla\mathbf{u}^T\right)$ is the infinitesimal strain tensor.

2.1 Hybrid Finite Volume Discretization

The simulation domain Ω is discretized by a set of polygonal or polyhedral control volumes $K \in \mathscr{K}$, such that $\overline{\Omega} = \bigcup_{K\in\mathscr{K}} \overline{K}$. The set of faces of the mesh is denoted by $\mathscr{E}$ and splits into boundary interfaces $\mathscr{E}_{ext}$ and inner interfaces $\mathscr{E}_{int}$. Among the boundary interfaces, we denote by $\mathscr{E}_{ext}^D$ and $\mathscr{E}_{ext}^N$ the subsets of boundary faces verifying Dirichlet or Neumann conditions. The center of gravity of the face σ is denoted by $\mathbf{x}_\sigma$ and its $d-1$ dimensional measure by $|\sigma|$. A point $\mathbf{x}_K$ is defined inside each cell K of the mesh. The set of faces of each cell K is denoted by $\mathscr{E}_K$, and the distance between $\mathbf{x}_K$ and σ by $d_{K,\sigma}$. The cone of base $\sigma \in \mathscr{E}_K$ and top $\mathbf{x}_K$ is denoted by K_σ.

Brief reminder of the hybrid finite volume discretization of a scalar diffusion problem (see [5]):

We first define the discrete hybrid spaces $V = \{(v_K \in \mathbb{R})_{K\in\mathscr{K}}, (v_\sigma \in \mathbb{R})_{\sigma\in\mathscr{E}}\}$ and $V^0 = \left\{v \in V \mid v_\sigma = 0\ \forall\sigma \in \mathscr{E}_{ext}^D\right\}$. V^0 is endowed with the discrete $H^1_{0,D}(\Omega)$ norm

$$\|v\|_{V^0} = \left(\sum_{K\in\mathscr{K}}\sum_{\sigma\in\mathscr{E}_K} \frac{|\sigma|}{d_{K,\sigma}} |v_\sigma - v_K|^2\right)^{\frac{1}{2}}. \tag{2}$$

Then, following [5], a discrete gradient is defined on each cone K_σ which only depends on v_K and $v_{\sigma'}$ for $\sigma' \in \mathscr{E}_K$. This gradient is exact on linear functions and satisfies a weak convergence property. According to [5], it can be written

$$\nabla_{K_\sigma} v = \sum_{\sigma'\in\mathscr{E}_K} (v_{\sigma'} - v_K)\mathbf{y}_K^{\sigma\sigma'} \quad \forall\, v \in V, \tag{3}$$

where $\mathbf{y}_K^{\sigma\sigma'} \in \mathbb{R}^d$ only depends on the geometry.

Hybrid finite volume discretization of the linear elasticity model:

As we did above, we introduce $\mathbf{V} = \left\{(\mathbf{v}_K \in \mathbb{R}^d)_{K\in\mathscr{K}}, (\mathbf{v}_\sigma \in \mathbb{R}^d)_{\sigma\in\mathscr{E}}\right\}$ as the discrete hybrid space. With an equivalent definition for $\mathbf{V}^0$, the discrete norm is now defined as $\|\mathbf{v}\|^2_{\mathbf{V}^0} = \sum_{i=1}^{d} \|v_i\|^2_{V^0}$. Let us also introduce the space $\mathbf{W} = \left\{(\mathbf{w}_K \in \mathbb{R}^d)_{K\in\mathscr{K}},\ (\mathbf{w}_\sigma \in \mathbb{R}^d)_{\sigma\in\mathscr{E}^D_{ext}}, (w^n_\sigma \in \mathbb{R})_{\sigma\in\mathscr{E}_{int}\cup\mathscr{E}^N_{ext}}\right\}$ and the following projection operator $P_\mathbf{W} : \mathbf{V} \to \mathbf{W},\ \mathbf{v} \mapsto \left((\mathbf{v}_K)_{K\in\mathscr{K}}, (\mathbf{v}_\sigma)_{\sigma\in\mathscr{E}^D_{ext}}, (\mathbf{v}_\sigma \cdot \mathbf{n}_\sigma)_{\sigma\in\mathscr{E}_{int}\cup\mathscr{E}^N_{ext}}\right)$, where $\mathbf{n}_\sigma$ is a unit vector normal to σ which orientation is fixed once and for all. Let us finally define the space $\mathbf{W}^0 = P_\mathbf{W}(\mathbf{V}^0)$ endowed with the discrete norm $\|\mathbf{w}\|_{\mathbf{W}^0} = \inf_{\mathbf{v}\in\mathbf{V}^0\,|\,P_\mathbf{W}(\mathbf{v})=\mathbf{w}} \|\mathbf{v}\|_{\mathbf{V}^0}$.

The main novelty of our discretization lies in the definition of a linear interpolation operator $\mathbf{I} : \mathbf{W} \to \mathbf{V}$. This linear interpolation operator must be second order accurate to preserve the order of approximation of the scheme and interpolant in the sense that $P_\mathbf{W}(\mathbf{I}(\mathbf{w})) = \mathbf{w}$ for all $\mathbf{w} \in \mathbf{W}$. It must also be local in the sense that it computes the displacement vector $\mathbf{v}_\sigma$ at a given face $\sigma \in \mathscr{E}_{int} \cup \mathscr{E}^N_{ext}$ in terms of a given number of normal components $\mathbf{v}_{\sigma'} \cdot \mathbf{n}_{\sigma'}$ taken on a stencil $\mathscr{S}_\sigma \subset \mathscr{E}$ of neighbouring faces σ' of σ (with $\sigma \in \mathscr{S}_\sigma$). An example of construction of such an interpolator is given in subsection 2.3.

The use of the interpolation operator $\mathbf{I}$ will bring two crucial improvements to the discretization: first a drastic reduction of the degrees of freedom and secondly a stabilization of the discretization.

Finally, generalizing the scalar framework to the vectorial case of the linear elasticity model, we introduce a piecewise constant discrete gradient for each cone K_σ:

$$\nabla_{K_\sigma}\mathbf{v} = \sum_{\sigma'\in\mathscr{E}_K} (\mathbf{v}_{\sigma'} - \mathbf{v}_K) \otimes \mathbf{y}^{\sigma\sigma'}_K \quad \forall\, \mathbf{v} \in \mathbf{V}. \tag{4}$$

2.2 Discrete Variational Formulation

Starting from (1), we deduce a discrete weak formulation of the problem in $\mathbf{W}^0$.

Setting $\varepsilon_{K_\sigma}(\mathbf{v}) = \frac{1}{2}\left(\nabla_{K_\sigma}\mathbf{v} + \nabla_{K_\sigma}\mathbf{v}^T\right)$ and $\sigma_{K_\sigma}(\mathbf{v}) = 2\mu\varepsilon_{K_\sigma}(\mathbf{v}) + \lambda\mathrm{tr}\left(\varepsilon_{K_\sigma}(\mathbf{v})\right)\mathbf{1}$ for all $\mathbf{v} \in \mathbf{V}$, we introduce the discrete bilinear form on $\mathbf{W} \times \mathbf{W}$

$$a_\mathscr{D}(\mathbf{u}, \mathbf{v}) = \sum_{K\in\mathscr{K}} \sum_{\sigma\in\mathscr{E}_K} |K_\sigma|\, \sigma_{K_\sigma}(\mathbf{I}(\mathbf{u})) : \varepsilon_{K_\sigma}(\mathbf{I}(\mathbf{v}))\,. \tag{5}$$

Then, the discrete variational formulation reads: find $\mathbf{u} \in \mathbf{W}$ such that $\mathbf{u}_\sigma = \mathbf{u}^D_\sigma$ for all $\sigma \in \mathscr{E}^D_{ext}$ and such that, for all $\mathbf{v} \in \mathbf{W}^0$,

$$a_\mathscr{D}(\mathbf{u}, \mathbf{v}) = \sum_{K\in\mathscr{K}} |K|\,\mathbf{f}_K \cdot \mathbf{v}_K + \sum_{\sigma\in\mathscr{E}^N_{ext}} |\sigma|\,\mathbf{g}_\sigma \cdot \mathbf{I}(\mathbf{v})_\sigma, \tag{6}$$

where $\mathbf{u}^D_\sigma = \frac{1}{|\sigma|}\int_\sigma \mathbf{u}^D \mathrm{d}\sigma$, $\mathbf{f}_K = \frac{1}{|K|}\int_K \mathbf{f}\,\mathrm{d}\mathbf{x}$ and $\mathbf{g}_\sigma = \frac{1}{|\sigma|}\int_\sigma \mathbf{g}\,\mathrm{d}\sigma$ are average values.

It is important to keep in mind that numerical experiments show that without interpolation, the bilinear form $a_\mathscr{D}$ on the space $\mathbf{V} \times \mathbf{V}$ leads to an unstable scheme with vanishing eigenvalues on triangular or tetrahedral meshes with mixed Neumann Dirichlet boundary conditions.

Note also that for the solution of the linear system (6), the unknowns $\mathbf{u}_K$ can easily be eliminated without any fill in, reducing the degrees of freedom to the face normal components of the displacement.

2.3 Interpolation of the tangential components of the displacement

Given a face $\sigma \in \mathscr{E}_{int} \cup \mathscr{E}_{ext}^N$, for each component $i \in [\![1,d]\!]$ of the displacement field $\mathbf{u}_\sigma$, we look for a linear interpolation of the form

$$\bar{u}_\sigma^{\,i}(\mathbf{x}) = \sum_{j=1}^{d} \alpha_\sigma^{ij} x_j + \beta_\sigma^i. \tag{7}$$

In order to determine the $d(d+1)$ coefficients $(\alpha_\sigma^{ij})_{i,j\in[\![1,d]\!]}$, $(\beta_\sigma^i)_{i\in[\![1,d]\!]}$ as linear combinations of normal components $\mathbf{u}_{\sigma'} \cdot \mathbf{n}_{\sigma'}$, $\sigma' \in \mathscr{S}_\sigma$, we look for a set $\mathscr{S}_\sigma$ of $d(d+1)$ neighbouring faces σ' of the face σ, with $\sigma \in \mathscr{S}_\sigma$ and such that the system of equations $\bar{\mathbf{u}}_\sigma(\mathbf{x}_{\sigma'}) \cdot \mathbf{n}_{\sigma'} = \mathbf{u}_{\sigma'} \cdot \mathbf{n}_{\sigma'}$ is non singular. The set $\mathscr{S}_\sigma$ is built using the following greedy algorithm:

1. Initialization: for a given number $k > d(d+1)$, we select the k closest neighbouring faces of the face σ which are sorted from the closest to the farthest using the distance between the face center and $\mathbf{x}_\sigma$: $\sigma_0 = \sigma, \sigma_1, \cdots, \sigma_{k-1}$. We set $\mathscr{S}_\sigma = \{\sigma\}$ and $q = 1, l = 0$;
2. while $q < d(d+1)$ and $l < k-1$:
    - $l = l + 1$;
    - if the equation $\bar{\mathbf{u}}_\sigma(\mathbf{x}_{\sigma_l}) \cdot \mathbf{n}_{\sigma_l} = \mathbf{u}_{\sigma_l} \cdot \mathbf{n}_{\sigma_l}$ is linearly independent with the set of equations $\bar{\mathbf{u}}_\sigma(\mathbf{x}_{\sigma'}) \cdot \mathbf{n}_{\sigma'} = \mathbf{u}_{\sigma'} \cdot \mathbf{n}_{\sigma'}$ for all $\sigma' \in \mathscr{S}_\sigma$, then $\mathscr{S}_\sigma = \{\sigma_l\} \cup \mathscr{S}_\sigma$, $q = q + 1$;
3. if $q < d(d+1)$, the algorithm is rerun with a larger value of k.

Note that since $\sigma \in \mathscr{S}_\sigma$, the interpolation operator satisfies as required the property $P_{\mathbf{W}}(\mathbf{I}(\mathbf{u})) = \mathbf{u}$ for all $\mathbf{u} \in \mathbf{W}$.

2.4 Compatibility condition with cellwise constant pressure for poroelastic models

Let $L_0^2(\Omega)$ be the subspace of functions of $L^2(\Omega)$ with vanishing mean values. For the sake of simplicity but without any loss of generality, we consider here

$\partial\Omega^D = \partial\Omega$ and $\mathbf{u}^D = \mathbf{0}$. It is well known (see [1]) that the well-posedness of linear poroelasticity models relies on the well-posedness of the following saddle point problem: find $(\mathbf{u}, p) \in H_0^1(\Omega)^d \times L_0^2(\Omega)$ such that

$$\begin{cases} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v})_{L^2(\Omega)^d} & \text{for all } \mathbf{v} \in H_0^1(\Omega)^d, \\ b(\mathbf{u}, q) = (h, q)_{L^2(\Omega)} & \text{for all } q \in L_0^2(\Omega), \end{cases} \tag{8}$$

where a is the bilinear form of the linear elasticity model and $b(\mathbf{v}, p) = -(\operatorname{div} \mathbf{v}, p)_{L^2(\Omega)}$. The stability of this saddle point problem results from the coercivity of a and the following LBB condition (see [2]) which guarantees the existence and uniqueness of the solution: $\inf_{p \in L_0^2(\Omega)} \sup_{\mathbf{v} \in H_0^1(\Omega)^d} \frac{b(\mathbf{v},p)}{\|\mathbf{v}\|_{H_0^1(\Omega)^d} \|p\|_{L^2(\Omega)}} \geq \gamma > 0$.

The following theorem states that the LBB condition holds on the discrete spaces $\mathbf{W}^0 \times M_0$, where M_0 is the space of cellwise constant functions on the mesh $\mathscr{K}$ with vanishing mean values endowed with the $L^2(\Omega)$ norm, and for the discrete divergence operator defined by:

$$b_{\mathscr{D}}(\mathbf{w}, p) = -(\operatorname{div}_{\mathscr{D}} \mathbf{w}, p)_{L^2(\Omega)} = -\sum_{K \in \mathscr{K}} p_K \sum_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} |\sigma| \, w_\sigma^n \, (\mathbf{n}_\sigma \cdot \mathbf{n}_{K,\sigma}), \tag{9}$$

for all $(\mathbf{w}, p) \in \mathbf{W}^0 \times M_0$, and where $\mathbf{n}_{K,\sigma}$ is the normal to the face σ outward K. It implies that, assuming the coercivity of $a_{\mathscr{D}}$, the coupling of our discretization for the elasticity model with a cell centered finite volume scheme for the Darcy pressure equation will lead to a convergent and stable scheme for the poroelastic model.

Theorem 1. *The bilinear form $b_{\mathscr{D}}$ defined on $\mathbf{W}^0 \times M_0$ satisfies the LBB condition*

$$\inf_{p \in M_0} \sup_{\mathbf{w} \in \mathbf{W}^0} \frac{b_{\mathscr{D}}(\mathbf{w}, p)}{\|\mathbf{w}\|_{\mathbf{W}^0} \|p\|_{L^2(\Omega)}} \geq \gamma_{\mathscr{D}} > 0, \tag{10}$$

with a constant $\gamma_{\mathscr{D}}$ depending only on usual regularity parameters of the mesh.

Proof. From the continuous LBB condition, for all $p \in M_0$, there exists a displacement field $\mathbf{v} \in H_0^1(\Omega)^d$ such that $b(\mathbf{v}, p) \geq \gamma \, \|\mathbf{v}\|_{H_0^1(\Omega)^d} \|p\|_{L^2(\Omega)}$. Let $\mathbf{u}$ be the element of $\mathbf{V}^0$ such that $\mathbf{u}_K = \frac{1}{|K|} \int_K \mathbf{v} \, d\mathbf{x}$ for $K \in \mathscr{K}$ and $\mathbf{u}_\sigma = \frac{1}{|\sigma|} \int_\sigma \mathbf{v} \, d\sigma$ for $\sigma \in \mathscr{E}$. Then, $\mathbf{w} = P_{\mathbf{W}}(\mathbf{u}) \in \mathbf{W}^0$ satisfies $b_{\mathscr{D}}(\mathbf{w}, p) = b(\mathbf{v}, p)$.

Since it is shown in [9] that $\|\mathbf{u}\|_{\mathbf{V}^0} \leq \kappa \, \|\mathbf{v}\|_{H_0^1(\Omega)^d}$, with a constant κ depending on usual regularity parameters of the mesh, and since we have by definition the inequality $\|\mathbf{w}\|_{\mathbf{W}^0} \leq \|\mathbf{u}\|_{\mathbf{V}^0}$, the discrete LBB condition holds with $\gamma_{\mathscr{D}} = \frac{\gamma}{\kappa}$. □

3 Numerical experiments

In this section, the convergence of the scheme is tested on the linear elasticity model with exact solution

$$u_i = e^{\cos\left(\sum_{j=1}^{d} \chi^{ij}\, x_j\right)}, \qquad i = 1, \cdots, d. \tag{11}$$

The right hand side and the Dirichlet boundary conditions are defined by the exact solution. The Lamé parameters λ and μ are set to 1.

The tests have been held using an object oriented C++ implementation which original approach is described in [8]. The relative l^2 error on the displacement and on the gradient of the displacement are plotted function of the number of inner faces. In the computation of these errors, the cellwise constant discrete solution and discrete gradient are used. We first consider the triangular meshes (mesh family 1), the Cartesian grids (mesh family 2), the local grid refinement (mesh family 3) and the Kershaw meshes (mesh family 4) from the FVCA5 2D benchmark. The exact solution is defined by

$$\chi = \begin{pmatrix} 1 & 1 \\ 2 & -1 \end{pmatrix}.$$

The results presented on Fig. 1 show the good convergence behaviour of the scheme. The expected order of convergence is reached for all the meshes (for the solution and its gradient) and is even exceeded for the gradient on Cartesian grids. Next, let us consider the Cartesian grids (mesh family A), the randomly distorted grids (mesh family AA), and the tetrahedral meshes (mesh family B) from the FVCA6 3D benchmark. The exact solution is defined by

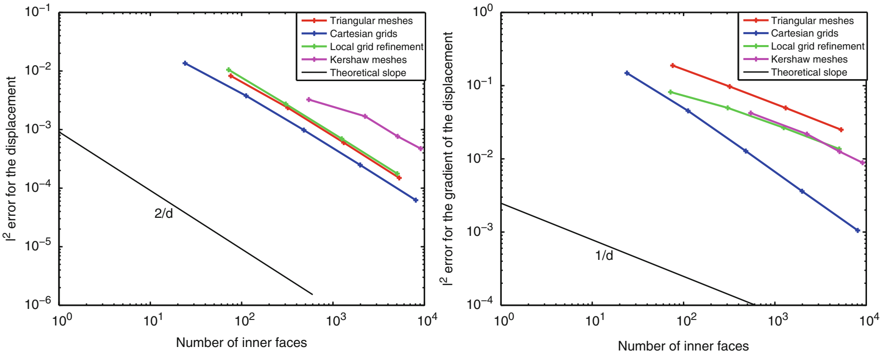


Fig. 1 l^2 error for the displacement and for the gradient of the displacement function of the number of inner faces for the triangular meshes, the Cartesian grids, the local grid refinement and the Kershaw meshes

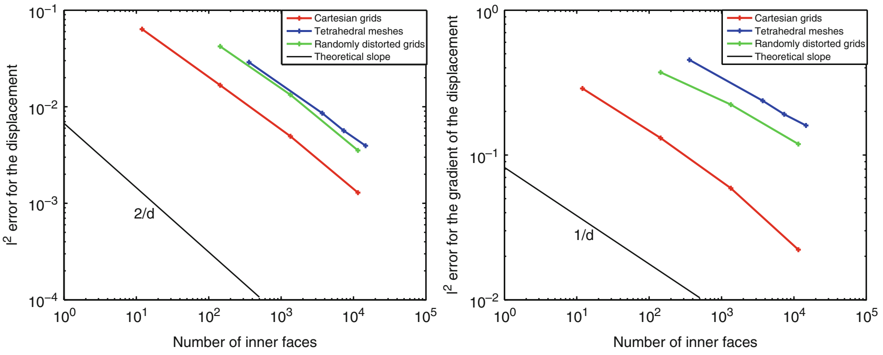


Fig. 2 l^2 error for the displacement and for the gradient of the displacement function of the number of inner faces for the Cartesian grids, the randomly distorted grids and the tetrahedral meshes

$$\chi = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -1 & 1 & 2 \end{pmatrix}.$$

The results exhibited on Fig. 2 show again the good convergence behaviour of the scheme with the expected order on the three meshes for both the discrete solution and its gradient.

4 Conclusion

In this paper, a new discretization method has been introduced for linear elasticity using only one degree of freedom per face. It applies to general polygonal and polyhedral meshes possibly non conforming. In addition this discretization satisfies the compatibility condition when coupled with cell centered finite volume schemes for the Darcy equation in poroelastic models.

First numerical experiments in 2D and 3D exhibit the stability and convergence of the scheme. In the near future, further testings will be performed on CPG grids with erosions, local grid refinement and faults, to assess the potential of this scheme for reservoir and CO2 storage simulations.

References

1. M.A. Murad and A.F.D. Loula. On Stability and Convergence of Finite Element Approximations of Biot's Consolidation Problem. *Int. Jour. Numer. Eng.*, 37:645–667, 1994.
2. F. Brezzi and M. Fortin. Mixed and Hybrid Finite Element Methods. *Springer-Verlag, New-York*, 1991.

3. A. Settari and F.M. Mourits. Coupling of Geomechanics and Reservoir simulation models. *Comp. Meth. Adv. Geomech., Siriwardane and Zaman and Balkema, Rotterdam*, 2151–2158, 1994.
4. J. Droniou, R. Eymard, T. Gallouët and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.
5. R. Eymard, T. Gallouët and R. Herbin. 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(4):1009–1043, 2010. See also http://hal.archives-ouvertes.fr/.
6. F. Brezzi, K. Lipnikov and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15:1533–1553, 2005.
7. D.K. Ponting. Corner Point Geometry in reservoir simulation. *In Clarendon Press, editor*, Proc. ECMOR I, 45–65, Cambridge, 1989.
8. D.A. Di Pietro and J.M. Gratien. Lowest order methods for diffusive problems on general meshes: A unified approach to definition and implementation. *These proceedings*, 2011.
9. R. Eymard, T. Gallouët and R. Herbin. Finite Volume Methods. *Handbook of Numerical Analysis*, 7:713–1020, 2000.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

An A Posteriori Error Estimator for a Finite Volume Discretization of the Two-phase Flow

Daniele A. Di Pietro, Martin Vohralík, and Carole Widmer

Abstract We derive a posteriori error estimates for a multi-point finite volume discretization of the two-phase Darcy problem. The proposed estimators yield a fully computable upper bound for the selected error measure. The estimate also allows to distinguish, estimate separately, and compare the linearization and algebraic errors and the time and space discretization errors. This enables, in particular, to design a discretization algorithm so that all the sources of error are properly balanced. Namely, the linear and nonlinear solvers can be stopped as soon as the algebraic and linearization errors drop to the level at which they do not affect to the overall error. This can lead to significant computational savings, since performing an excessive number of unnecessary iterations can be avoided. Similarly, the errors in space and in time can be equilibrated by time step and local mesh adaptivity.

Keywords Finite volumes, a posteriori error estimates, darcy model, fully computable upperbound, twophase flow.
MSC2010:65M08,65M50,76S05.

1 The two-phase flow model

Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, denote a bounded connected polygonal domain and let $t_\mathrm{F} > 0$. Let w denote the wetting phase (e.g., water) and o the non-wetting phase (e.g., oil), and let there be given sources $f_\mathrm{o},\ f_\mathrm{w} \in L^2((0, t_\mathrm{F}); L^2(\Omega))$ and a (constant) porosity

Daniele A. Di Pietro and Carole Widmer
IFP Energies nouvelles, 1&4, avenue du Bois-Préau, Rueil-Malmaison, France,
e-mail: dipietrd@ifpen.fr, carole.widmer@ifpen.fr

Martin Vohralík
UPMC Univ. Paris 06, UMR 7598, Laboratoire J.-L. Lions, 75005, Paris, France & CNRS, UMR 7598, Laboratoire J.-L. Lions, 75005, Paris, France, e-mail: vohralik@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_36, © Springer-Verlag Berlin Heidelberg 2011

$\phi \in (0,1]$. We consider the two-phase flow (see, e.g., [3]): Find $\mathbf{U} := \{P, S_{\text{o}}, S_{\text{w}}\}$, with P the pressure and S_p, $p \in \{\text{o}, \text{w}\}$, the saturations, such that

$$\begin{aligned}
\partial_t(\phi S_{\text{o}}) + \nabla \cdot (\nu_{\text{o}}(P, S_{\text{o}})\mathbf{u}_{\text{o}}(P, S_{\text{o}})) &= f_{\text{o}} && \text{in } \Omega \times (0, t_{\text{F}}),\\
\partial_t(\phi S_{\text{w}}) + \nabla \cdot (\nu_{\text{w}}(P, S_{\text{w}})\mathbf{u}_{\text{w}}(P, S_{\text{w}})) &= f_{\text{w}} && \text{in } \Omega \times (0, t_{\text{F}}),\\
S_{\text{o}} + S_{\text{w}} &= 1 && \text{in } \Omega \times (0, t_{\text{F}}).
\end{aligned} \tag{1}$$

For $p \in \{\text{o}, \text{w}\}$, ν_p denotes here the mobility of the phase p defined as the ratio of the relative permeability to the viscosity. In (1), $\mathbf{u}_{\text{o}}$ and $\mathbf{u}_{\text{w}}$ are such that

$$\mathbf{u}_p(P, S_p) := -K\nabla\left(P + P_{\text{c},p}(S_p)\right), \quad \text{for } p \in \{\text{o}, \text{w}\}, \text{ in } \Omega \times (0, t_{\text{F}}), \tag{2}$$

where $P_{\text{c},p}(S_p)$ is the capillary pressure and K denotes a piecewise constant, uniformly elliptic tensor-valued field corresponding to the absolute permeability. To find some example of the physics laws (capillarity pressure, phase mobility) or of the absolute permeability see [7].

Problem (1) is complemented by the initial conditions:

$$S_{\text{o}}(\cdot, 0) = S_{\text{o}}^0 \text{ and } P(\cdot, 0) = P^0, \quad \text{in } \Omega, \tag{3}$$

as well as by no-flow boundary conditions:

$$\mathbf{u}_p(P, S_p) \cdot \mathbf{n}_\Omega = 0, \qquad \text{in } \partial\Omega \times (0, t_{\text{F}}). \tag{4}$$

The purpose of this paper is to propose fully computable a posteriori error estimates for the discretization of (1)–(4) by cell-centered finite volume methods in space and the backward Euler scheme in time. In particular, we consider the multi-point finite volume method proposed in [1]. Using a dual error norm is motivated by, e.g., [8]. Developing the ideas of [4–6, 9], we in particular separate the estimate into contributions representing the *space discretization error*, *time discretization error*, *linearization error*, and *algebraic error*. Then, at each time step, the linearization algorithm and the iterative algebraic solver can be stopped as soon as the corresponding errors no longer affect the total error, and space and the time errors can be equilibrated.

2 Discretization by the finite volume method

2.1 Notations

Let $\mathscr{T} = \{T\}$ denotes a partition of Ω into simplices or rectangular parallelepipeds (the extension to general polygonal meshes is possible via the introduction of simplicial submeshes). For rectangular parallelepipeds, we further assume that K

is diagonal to perform $H(\mathrm{div};\Omega)$-conforming reconstructions. For every element $T \in \mathscr{T}$, we denote by $|T|$ its measure and by h_T its diameter. Let $\mathscr{F} = \{\sigma\}$ be the set of faces of the mesh and, for all $T \in \mathscr{T}$, set $\mathscr{F}_T := \{\sigma \in \mathscr{F} \mid \sigma \subset \partial T\}$. The time discretization is defined by a strictly increasing sequence of discrete times $\{t^n\}_{0\le n\le N}$ such that $t^0 = 0$ and $t^N = t_{\mathrm{F}}$. For $1 \le n \le N$, we define the time interval $I_n := (t^{n-1}, t^n]$ and the time step $\tau^n := t^n - t^{n-1}$.

2.2 The finite volume scheme

The discrete problem reads: For all $1 \le n \le N$, all $T \in \mathscr{T}$, and all $p \in \{\mathrm{o}, \mathrm{w}\}$, find $\mathbf{U}^n_T := \{P^n_T, S^n_{\mathrm{o},T}, S^n_{\mathrm{w},T}\}$ such that

$$\phi\frac{|T|}{\tau^n}\left(S^n_{p,T} - S^{n-1}_{p,T}\right) + \sum_{\sigma\in\mathscr{F}_T} \nu_p(P^{n-1}_{T^\star_p(\sigma)}, S^{n-1}_{p,T^\star_p(\sigma)})F^n_{p,T,\sigma} - f^n_{p,T} = 0, \tag{5}$$

where $f^n_{p,T} = (f^n_p, 1)_T$ and $f^n_p = \frac{1}{\tau^n}\int_{t^{n-1}}^{t^n} f_p(t)\,\mathrm{d}t$. We set $P^0_T := (P^0, 1)_T\,/|T|$, $S^0_{\mathrm{o},T} := (S^0_{\mathrm{o}}, 1)_T/|T|$, and impose $S^n_{\mathrm{o},T} + S^n_{\mathrm{w},T} = 1$ for all $0 \le n \le N$. Furthermore, $F^n_{p,T,\sigma} = F_{p,T,\sigma}(\{\mathbf{U}^n_{T'}\}_{\mathscr{S}_\sigma})$ is a multi-point approximation of the flux of the phase p leaving $T \in \mathscr{T}$ through the face $\sigma \in \mathscr{F}_T$ that depends on the unknowns associated to the elements of the face stencil $\mathscr{S}_\sigma \subset \mathscr{T}$. The numerical flux is assumed to be conservative, i.e., for all internal faces $\sigma \subset \partial T_1 \cap \partial T_2$, there holds $F^n_{p,T_1,\sigma} = -F^n_{p,T_2,\sigma}$. The upwind cell $T^\star_p(\sigma)$ is equal to T_1 if $F^n_{p,T_1,\sigma} \ge 0$, to T_2 otherwise. For boundary faces $\sigma \subset \partial T \cap \partial\Omega$, $F^n_{p,T,\sigma} = 0$ to honor the no-flow boundary condition (4), and we can leave $T^\star_p(\sigma)$ undefined.

For all $0 \le n \le N$ and $T \in \mathscr{T}$, the unknown $S^n_{\mathrm{w},T}$ is eliminated using the local volume conservation equation $S^n_{\mathrm{o},T} + S^n_{\mathrm{w},T} = 1$. We introduce the reduced set of unknowns $\overline{\mathbf{U}}^n := \{\mathbf{P}^n, \mathbf{S}^n_{\mathrm{o}}\}$, where $\mathbf{P}^n = \{P^n_T\}_{T\in\mathscr{T}}$ and $\mathbf{S}^n_{\mathrm{o}} = \{S^n_{\mathrm{o},T}\}_{T\in\mathscr{T}}$. With a little abuse of notation, for a function $\xi(S_{\mathrm{w}})$, we write $\xi(S_{\mathrm{o}})$ to mean $\xi(1 - S_{\mathrm{o}})$. As a consequence, $\nu_{\mathrm{w}}(S_{\mathrm{o}})$ and $P_{\mathrm{c,w}}(S_{\mathrm{o}})$ are equal to $\nu_{\mathrm{w}}(1 - S_{\mathrm{o}})$, $P_{\mathrm{c,w}}(1 - S_{\mathrm{o}})$ and $\mathbf{u}_{\mathrm{w}}(P, 1 - S_{\mathrm{o}})$ respectively. Equation (5) becomes, for all $1 \le n \le N$, all $T \in \mathscr{T}$, and all $p \in \{\mathrm{o}, \mathrm{w}\}$

$$\mathbf{D}^n_{p,T}(\overline{\mathbf{U}}^n) = 0, \text{ with,} \tag{6}$$

$$\mathbf{D}^n_{p,T}(\overline{\mathbf{U}}^n) := \phi\frac{|T|}{\tau^n}(-1)^j(S^n_{\mathrm{o},T} - S^{n-1}_{\mathrm{o},T}) + \sum_{\sigma\in\mathscr{F}_T} \nu_p(P^{n-1}_{T^\star_p(\sigma)}, S^{n-1}_{\mathrm{o},T^\star_p(\sigma)})F^n_{p,T,\sigma} - f^n_{p,T}, \tag{7}$$

where $j = 1$ if $p = \mathrm{w}$ and 0 otherwise.

2.3 Linearization

Problem (6) is a system of nonlinear algebraic equations that can be solved using the Newton algorithm. For a fixed $1 \leq n \leq N$, let $\overline{\mathbf{U}}^{n,0}$ be given (typically, $\overline{\mathbf{U}}^{n,0} = \overline{\mathbf{U}}^{n-1}$). For $1 \leq k$, a new estimate $\overline{\mathbf{U}}^{n,k}$ is computed from the previous $\overline{\mathbf{U}}^{n,k-1}$ by solving the following system of linear algebraic equations: For all $T \in \mathscr{T}$ and all $p \in \{\mathrm{o}, \mathrm{w}\}$,

$$\sum_{T' \in \mathscr{T}} \frac{\partial \mathbf{D}^n_{p,T}}{\partial \overline{\mathbf{U}}_{T'}} \left(\overline{\mathbf{U}}^{n,k}_{T'} - \overline{\mathbf{U}}^{n,k-1}_{T'} \right) = -\mathbf{D}^n_{p,T}(\overline{\mathbf{U}}^{n,k-1}), \tag{8}$$

where $\overline{\mathbf{U}}^{n,k}_T = \{P^{n,k}_T, S^{n,k}_{\mathrm{o},T}\}$ denotes the approximate solutions in T at the n-th time step and k-th Newton iteration. We suppose that (8) is solved using an iterative linear solver. For a fixed $1 \leq n \leq N$ and $k \geq 1$, let $\overline{\mathbf{U}}^{n,k,0}$ be given (typically, $\overline{\mathbf{U}}^{n,k,0} = \overline{\mathbf{U}}^{n,k-1}$). Then, at a given step $i \geq 1$, we have, for all $T \in \mathscr{T}$ and $p \in \{\mathrm{o}, \mathrm{w}\}$,

$$\sum_{T' \in \mathscr{T}} \frac{\partial \mathbf{D}^n_{p,T}}{\partial \overline{\mathbf{U}}_{T'}} \left(\overline{\mathbf{U}}^{n,k,i}_{T'} - \overline{\mathbf{U}}^{n,k-1}_{T'} \right) + \mathbf{D}^n_{p,T}(\overline{\mathbf{U}}^{n,k-1}) = \mathbf{R}^{n,k,i}_{p,T}, \tag{9}$$

where $\mathbf{R}^{n,k,i}_{p,T}$ is the algebraic residual, while $\overline{\mathbf{U}}^{n,k,i}_T = \{P^{n,k,i}_T, S^{n,k,i}_{\mathrm{o},T}\}$ denotes the approximate solution at the n-th time step, k-th Newton iteration, and i-th linear solver iteration.

3 A posteriori error estimate

3.1 Space-time approximate solutions

Let, for $0 \leq n \leq N$ and $p \in \{\mathrm{o}, \mathrm{w}\}$, $S^n_{p,h}$ be the piecewise constant function such that $S_{p,h|T} = S_{p,T}$ for all $T \in \mathscr{T}$. We introduce the space-time function $S_{p,h\tau}$ continuous and piecewise affine in time, and such that $S_{p,h\tau}(t^n) = S^n_{p,h}$ for $0 \leq n \leq N$. In order to give a meaning to the gradient operator appearing in (2), we need to postprocess the approximate cell pressures $\{P^n_T\}_{T \in \mathscr{T}}$ and capillary pressures $\{P^n_{\mathrm{c},p,T}\}_{T \in \mathscr{T}}$, $P^n_{\mathrm{c},p,T} := P_{\mathrm{c},p}(S^n_{p,T})$, $p \in \{\mathrm{o}, \mathrm{w}\}$. As in [5, 6, 9], we introduce an elementwise postprocessing of $\{P^n_T\}_{T \in \mathscr{T}}$ and $\{P^n_{\mathrm{c},p,T}\}_{T \in \mathscr{T}}$, $1 \leq n \leq N$, yielding piecewise quadratic functions $\tilde{P}^n_h$ and $\tilde{P}^n_{\mathrm{c},p,h}$ ($\tilde{P}^0_h$ is given by a projection of the initial pressure P^0). As for the saturations, $\tilde{P}_{h\tau}$ and $\tilde{P}_{\mathrm{c},p,h\tau}$ are the space-time functions, continuous and piecewise affine in time, and such that $\tilde{P}_{p,h\tau}(t^n) := \tilde{P}^n_h$ and $\tilde{P}_{\mathrm{c},p,h\tau}(t^n) := \tilde{P}^n_{\mathrm{c},p,h}$, respectively.

3.2 Error measure

Set $X := L^2((0,t_{\mathrm{F}}); H^1(\Omega))$. For $\varphi \in X$, let $\|\varphi\|_X^2 := \int_0^{t_{\mathrm{F}}} \|\nabla\varphi\|^2 \mathrm{d}t$ and $\|\cdot\|$ denotes the L^2-norm on Ω. We suppose that the solution $(P, S_{\mathrm{o}}, S_{\mathrm{w}})$ of the problem (1)–(4) has the necessary regularity to permit the following weak formulation characterization: For all $\varphi \in X$, and all $p \in \{\mathrm{o}, \mathrm{w}\}$,

$$\int_0^{t_{\mathrm{F}}} \left\{ \langle \partial_t(\phi S_p), \varphi \rangle - \left(\nu_p(P,S_p)\mathbf{u}_p(P,S_p), \nabla\varphi\right)_\Omega - (f_p, \varphi)_\Omega \right\} \mathrm{d}t = 0. \tag{10}$$

The aim of the following measure is to evaluate the residual of the approximate solution and the nonconformity of the approximate pressure (i.e., the facts that $(\tilde{P}_{h\tau}, S_{\mathrm{o},h\tau}, S_{\mathrm{w},h\tau})$ do not satisfy (10) and that $\tilde{P}_{h\tau} \notin X$ in general). Note that if $S_{p,h\tau}$ coincide with S_p, $p \in \{\mathrm{o},\mathrm{w}\}$, and $\tilde{P}_{h\tau}$ with P, the error measure equals zero:

$$\begin{aligned} & |||(S_p - S_{p,h\tau}, P - \tilde{P}_{h\tau})||| \\ := & \sup_{\varphi\in X,\, \|\varphi\|_X = 1} \int_0^{t_{\mathrm{F}}} \Bigg\{ \langle \partial_t(\phi S_p) - \partial_t(\phi S_{p,h\tau}), \varphi \rangle \\ & - \left(\nu_p(P,S_p)\mathbf{u}_p(P,S_p) - \nu_p(\tilde{P}_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\tilde{P}_{h\tau}, S_{p,h\tau}), \nabla\varphi\right) \Bigg\} \mathrm{d}t \\ & + \inf_{\delta\in X} \left\{ \int_0^{t_{\mathrm{F}}} \left\| \nu_p(\tilde{P}_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\tilde{P}_{h\tau}, S_{p,h\tau}) - \nu_p(\delta, S_{p,h\tau})\mathbf{u}_p(\delta, S_{p,h\tau}) \right\|^2 \mathrm{d}t \right\}^{\frac{1}{2}}. \end{aligned} \tag{11}$$

3.3 A posteriori error estimate

We let $\mathbf{RTN}(T) := [\mathbb{P}_0(T)]^d + \mathbb{P}_0(T)\mathbf{x}$ and $\mathbf{RTN}(T) := [\mathbb{P}_0(T)]^d + [\mathbb{P}_0(T)]^d\mathbf{x}$, on simplices and on rectangular parallelepipeds respectively, and we introduce the Raviart–Thomas–Nédélec space

$$\mathbf{RTN}(\mathscr{T}) := \{\mathbf{v}_h \in \mathbf{H}(\mathrm{div},\Omega) \,|\, \mathbf{v}_{h|T} \in \mathbf{RTN}(T),\ \forall T \in \mathscr{T}\}.$$

Following [2, 4, 5, 9], in order to obtain an estimate on (11), we introduce for $1 \le n \le N$ and $p \in \{\mathrm{o},\mathrm{w}\}$ the flux reconstructions $\boldsymbol{\theta}_{p,h}^n \in \mathbf{RTN}(\mathscr{T})$ such that for $1 \le n \le N$, $T \in \mathscr{T}$, $T' \in \mathscr{T}_T$, $(T \cap T' = \sigma_{T,T'})$, and $p \in \{\mathrm{o},\mathrm{w}\}$,

$$\langle \boldsymbol{\theta}_{p,h}^n \cdot \mathbf{n}_T\, |_{\sigma_{T,T'}}, 1\rangle_{\sigma_{T,T'}} := \nu_p(P^{n-1}_{T_p^\star(\sigma)}, S^{n-1}_{\mathrm{o},T_p^\star(\sigma)}) F^n_{p,T,\sigma}. \tag{12}$$

The following local conservation property is obtained by the Green theorem from (6) and (12):

$$(f_p^n - \partial_t(\phi S_{p,h\tau}) - \nabla\cdot\boldsymbol{\theta}_{p,h}^n, 1)_T = 0. \tag{13}$$

Let us now define the *residual estimators* $\eta^n_{\mathrm{R},T,p}$, the *diffusive flux estimators* $\eta^n_{\mathrm{DF},T,p}$, and the *nonconformity estimators* $\eta^n_{\mathrm{NC},T,p}$ as

$$\begin{aligned}
\eta^n_{\mathrm{R},T,p} &:= \frac{h_T}{\pi}\|f_p - \partial_t(\phi S_{p,h\tau}) - \nabla\cdot\boldsymbol{\theta}^n_{p,h}\|_T,\\
\eta^n_{\mathrm{DF},T,p}(t) &:= \left\|\boldsymbol{\theta}^n_{p,h} - \nu_p(\tilde{P}_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\tilde{P}_{h\tau}, S_{p,h\tau})(t)\right\|_T, \qquad (14)\\
\eta^n_{\mathrm{NC},T,p}(t) &:= \left\|\nu_p(\tilde{P}_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\tilde{P}_{h\tau}, S_{p,h\tau})(t) - \nu_p(\delta_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\delta_{h\tau}, S_{p,h\tau})(t)\right\|_T.
\end{aligned}$$

Here $\delta_{h\tau} \in X$ is continuous and piecewise affine in time and such that $\delta_{h\tau}(t^n) = \delta^n_h$, with $\delta^n_h := \mathscr{I}_{av}(\tilde{P}^n_h)$ for all $0 \le n \le N$; $\mathscr{I}_{av}$ is an averaging operator as in [5, 6, 9].

Theorem 1 (Guaranteed a posteriori error estimate). *Let* $p \in \{\mathrm{o}, \mathrm{w}\}$. *Then*

$$\begin{aligned}
|||(S_p - S_{p,h\tau}, P - \tilde{P}_{h\tau})||| \le &\left\{\sum_{n=1}^{N}\int_{I_n}\sum_{T\in\mathscr{T}}(\eta^n_{\mathrm{R},T,p} + \eta^n_{\mathrm{DF},T,p}(t))^2\,\mathrm{d}t\right\}^{\frac{1}{2}}\\
&+\left\{\sum_{n=1}^{N}\int_{I_n}\sum_{T\in\mathscr{T}}(\eta^n_{\mathrm{NC},T,p}(t))^2\,\mathrm{d}t\right\}^{\frac{1}{2}}. \qquad (15)
\end{aligned}$$

Proof. The proof is straightforward using the definition of the error measure (11) and following the techniques of [5]. The second term in (15) clearly issues from the second term in the right hand-side of (11). We thus only have to prove that the first term is an upper bound on the first term in the right hand-side of (11). Let $\varphi \in X$, $\|\varphi\|_X = 1$, and $p \in \{\mathrm{o}, \mathrm{w}\}$. Set $\mathbf{w}_p := \nu_p(P, S_p)\mathbf{u}_p(P, S_p)$ and $\mathbf{w}_{p,h\tau} := \nu(\tilde{P}_{h\tau}, S_{p,h\tau})\mathbf{u}_p(\tilde{P}_{h\tau}, S_{p,h\tau})$. Then using the characterization of the weak solution (10),

$$\begin{aligned}
&\int_0^{t_{\mathrm{F}}}\{\langle\partial_t(\phi S_p) - \partial_t(\phi S_{p,h\tau}), \varphi\rangle - (\mathbf{w}_p - \mathbf{w}_{p,h\tau}, \nabla\varphi)\}\mathrm{d}t\\
&= \int_0^{t_{\mathrm{F}}}\{(f_p - \partial_t(\phi S_{p,h\tau}), \varphi) + (\mathbf{w}_{p,h\tau}, \nabla\varphi)\}\mathrm{d}t.
\end{aligned}$$

Let now $1 \le n \le N$ be given. Adding and subtracting $(\boldsymbol{\theta}^n_{p,h}, \nabla\varphi)$, using the Green theorem, the local conservativity property (13), the Poincaré inequality, and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned}
&(f_p, \varphi) - (\partial_t(\phi S_{p,h\tau}), \varphi) + (\mathbf{w}_{p,h\tau}, \nabla\varphi)\\
&= (f_p - \partial_t(\phi S_{p,h\tau}) - \nabla\cdot\boldsymbol{\theta}^n_{p,h}, \varphi) + (\mathbf{w}_{p,h\tau} - \boldsymbol{\theta}^n_{p,h}, \nabla\varphi)\\
&= (f_p - \partial_t(\phi S_{p,h\tau}) - \nabla\cdot\boldsymbol{\theta}^n_{p,h}, \varphi - \Pi_0\varphi) + (\mathbf{w}_{p,h\tau} - \boldsymbol{\theta}^n_{p,h}, \nabla\varphi)\\
&\le \sum_{T\in\mathscr{T}}(\eta^n_{\mathrm{R},T,p} + \eta^n_{\mathrm{DF},T,p}(t))\|\nabla\varphi\|_T,
\end{aligned}$$

where Π_0 denotes the L^2-orthogonal projection onto piecewise constants on $\mathscr{T}$. The assertion follows by the Cauchy–Schwarz inequality and by $\|\varphi\|_X = 1$. □

3.4 Identification of different components of the error

Let $1 \le n \le N$, $T \in \mathscr{T}$, and $p \in \{\mathrm{o}, \mathrm{w}\}$. In Section 2.2, we define the nonlinear system (6) and we solve it in Section 2.3 using an iterative solver for the Newton algorithm. Let assume we are at the n-th time step, k-th Newton step and i-th linearization step. We introduce the following notations:

$$A^{n,k,i}_{p,T} := \phi \frac{|T|}{\tau^n}\left[(S^{n,k,i}_{p,T} - S^{n,k-1}_{p,T}) - S^{n-1}_{p,T}\right],\quad B^{n,k,i}_{p,T,\sigma} := \nu_p(P^{n,k-1}_{T_p^\star(\sigma)}, S^{n,k-1}_{p,T_p^\star(\sigma)})F^{n,k,i}_{p,T,\sigma}.$$

The linear system (9) is then equivalent to the following sum of diagonal terms and face fluxes:

$$\frac{\partial A^{n,k,i}_{p,T}}{\partial \overline{\mathbf{U}}_T} + \sum_{\sigma\in\mathscr{F}_T}\sum_{T'\in\mathscr{S}_\sigma}\frac{\partial B^{n,k,i}_{p,T,\sigma}}{\partial \overline{\mathbf{U}}_{T'}} + \mathbf{D}^n_{p,T}(\overline{\mathbf{U}}^{n,k-1}) = \mathbf{R}^{n,k,i}_{p,T}. \tag{16}$$

Let us now define a linearization flux $\overline{\boldsymbol{\theta}}^{n,k,i}_{p,h} \in \mathbf{RTN}(\mathscr{T})$ and algebraic solver flux $\mathbf{r}^{n,k,i}_{p,h} \in \mathbf{RTN}(\mathscr{T})$ such that $\boldsymbol{\theta}^{n,k,i}_{p,h} := \overline{\boldsymbol{\theta}}^{n,k,i}_{p,h} + \mathbf{r}^{n,k,i}_{p,h}$ and such that

$$\langle \overline{\boldsymbol{\theta}}^{n,k,i}_{p,h}\cdot \mathbf{n}_T\,|_{\sigma_{T,T'}}, 1\rangle_{\sigma_{T,T'}} := \sum_{T'\in\mathscr{T}_T}\frac{\partial B^{n,k,i}_{p,T,\sigma}}{\partial \overline{\mathbf{U}}_{T'}} \text{ and } (\nabla\cdot \mathbf{r}^{n,k,i}_{p,h}, 1)_T = -\mathbf{R}^{n,k,i}_{p,T}. \tag{17}$$

Note that $\overline{\boldsymbol{\theta}}^{n,k,i}_{p,h}$ is fully specified; $\mathbf{r}^{n,k,i}_{p,h}$ can be constructed as in [6]. This gives

$$(f^n_p - \partial_t(\phi S^{k,i}_{p,h\tau}) - \nabla\cdot \overline{\boldsymbol{\theta}}^{n,k,i}_{p,h}, 1)_T = (\nabla\cdot \mathbf{r}^{n,k,i}_{p,h}, 1)_T, \qquad p\in\{\mathrm{o},\mathrm{w}\}. \tag{18}$$

We can now define the same estimators as in (14) and we have:

$$\eta^{n,k,i}_{\mathrm{R},T,p} + \eta^{n,k,i}_{\mathrm{DF},T,p}(t) + \eta^{n,k,i}_{\mathrm{NC},T,p}(t) \le \eta^{n,k,i}_{\mathrm{tm},T,p}(t) + \eta^{n,k,i}_{\mathrm{sp},T,p}(t) + \eta^{n,k,i}_{\mathrm{lin},T,p}(t) + \eta^{n,k,i}_{\mathrm{alg},T,p},$$

with

$$\begin{aligned}
\eta^{n,k,i}_{\mathrm{tm},T,p}(t) &:= \left\|\nu_p(\tilde P^{k,i}_{h\tau}, S^{k,i}_{p,h\tau})\mathbf{u}_p(\tilde P^{k,i}_{h\tau}, S^{k,i}_{p,h\tau})(t) - \nu_p(\tilde P^{n,k,i}_{h}, S^{n,k,i}_{p,h})\mathbf{u}_p(\tilde P^{n,k,i}_{h}, S^{n,k,i}_{p,h})\right\|_T,\\
\eta^{n,k,i}_{\mathrm{sp},T,p}(t) &:= \eta^{n,k,i}_{\mathrm{R},T,p} + \eta^{n,k,i}_{\mathrm{NC},T,p}(t),\\
\eta^{n,k,i}_{\mathrm{lin},T,p}(t) &:= \left\|\nu_p(\tilde P^{n,k,i}_{h}, S^{n,k,i}_{p,h})\mathbf{u}_p(\tilde P^{n,k,i}_{h}, S^{n,k,i}_{p,h}) - \overline{\boldsymbol{\theta}}^{n,k,i}_{p,h}\right\|_T,\\
\eta^{n,k,i}_{\mathrm{alg},T,p} &:= \|\mathbf{r}^{n,k,i}_{p}\|_T.
\end{aligned} \tag{19}$$

3.5 Adaptive algorithm

To solve the nonlinear system (6), let us introduce the following algorithm, for $1 \leq n \leq N$.

1. Choose initial saturations $\mathbf{S}_{\mathrm{o}}^{n,0}$ and pressures $\mathbf{P}^{n,0}$ according to (3). Typically, we put $\mathbf{S}_{\mathrm{o}}^{n,0} = \mathbf{S}_{\mathrm{o}}^{n-1}$ and $\mathbf{P}^{n,0} = P^{n-1}$. Set $k = 1$.
2. Set up the linear system (8).

   a. Choose some initial saturation $\mathbf{S}_{\mathrm{o}}^{n,k,0}$ and pressure $\mathbf{P}^{n,k,0}$. Typically, we let $\mathbf{S}_{\mathrm{o}}^{n,k,0} = \mathbf{S}_{\mathrm{o}}^{n,k-1}$ and $\mathbf{P}^{n,k,0} = \mathbf{P}^{n,k-1}$. Set $i = 1$.
   b. Perform a step of a chosen iterative method for the solution of (8), starting from $\mathbf{S}_{\mathrm{o}}^{n,k,i-1}$ and $\mathbf{P}^{n,k,i-1}$. This gives approximations $\mathbf{S}_{\mathrm{o}}^{n,k,i}$ and $\mathbf{P}^{n,k,i}$.
   c. Postprocess locally the pressures $\mathbf{P}^{n,k,i}$.
   d. Construct the fluxes $\overline{\boldsymbol{\theta}}_{p,h}^{n,k,i} \in \mathbf{RTN}(\mathscr{T})$, $p \in \{\mathrm{o}, \mathrm{w}\}$, according to Section 3.4.
   e. For $p \in \{\mathrm{o}, \mathrm{w}\}$, from the algebraic residual vectors $\mathbf{R}_p^{n,k,i}$ construct the fluxes $\mathbf{r}_{p,h}^{n,k,i} \in \mathbf{RTN}(\mathscr{T})$, as described in Section 3.4.
   f. We evaluate all the indicators (19)and define their global versions by their Hilbertian sums. The convergence criterion for the linear solver is:

   $$\eta_{\mathrm{alg},p}^{n,k,i} \leq \gamma_{\mathrm{alg}}(\eta_{\mathrm{sp},p}^{n,k,i} + \eta_{\mathrm{tm},p}^{n,k,i} + \eta_{\mathrm{lin},p}^{n,k,i}), \qquad p \in \{\mathrm{o}, \mathrm{w}\}. \tag{20}$$

   Here, $0 < \gamma_{\mathrm{alg}} \leq 1$ is a user-given weight, typically close to 1. Criterion (20) expresses that there is no need to continue with the algebraic solver iterations if the overall error is dominated by the other components. If (20) is reached, set $\mathbf{S}_{\mathrm{o}}^{n,k} := \mathbf{S}_{\mathrm{o}}^{n,k,i}$ and $\mathbf{P}^{n,k} := \mathbf{P}^{n,k,i}$. If not, $i := i + 1$ and go back to step 2(b).

3. The convergence criterion for the nonlinear solver is:

   $$\eta_{\mathrm{lin},p}^{n,k,i} \leq \gamma_{\mathrm{lin}}(\eta_{\mathrm{sp},p}^{n,k,i} + \eta_{\mathrm{tm},p}^{n,k,i}), \qquad p \in \{\mathrm{o}, \mathrm{w}\}. \tag{21}$$

   Here $0 < \gamma_{\mathrm{lin}} \leq 1$ is a user-given weight, typically close to 1. Criterion (21) expresses that there is no need to continue with the linearization iterations if the overall error is dominated by the other components. If criterion (21) is reached, finish. If not, $k := k + 1$ and go back to step 1.

Additionally, for all $1 \leq n \leq N$, the space and time estimators $\eta_{\mathrm{sp},p}^{n}$ and $\eta_{\mathrm{tm},p}^{n}$ should be made of similar size.

References

1. L. Agélas, D.A. Di Pietro, and R. Masson, *A symmetric and coercive finite volume scheme for multiphase porous media flow with applications in the oil industry*, (2008), pp. 35–52.
2. C. Cancès and M. Vohralík, *A posteriori error estimtate for immiscible incompressible two-phase flows*. In preparation, 2011.
3. Z. Chen, G. Huan, and Y. Ma, *Computational methods for multiphase flows in porous media*, Computational Science & Engineering, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.
4. L. El Alaoui, A. Ern, and M. Vohralík, *Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems*, Comput. Methods Appl. Mech. Engrg., (2010). DOI 10.1016/j.cma.2010.03.024.
5. A. Ern and M. Vohralík, *A posteriori error estimation based on potential and flux reconstruction for the heat equation*, SIAM J. Numer. Anal., 48 (2010), pp. 198–223.
6. P. Jiránek, Z. Strakoš, and M. Vohralík, *A posteriori error estimates including algebraic error and stopping criteria for iterative solvers*, SIAM J. Sci. Comput., 32 (2010), pp. 1567–1590.
7. C. Marles, *Cours de production, Tome 4*, Technip, 1984.
8. R. Verfürth, *Robust a posteriori error estimates for nonstationary convection-diffusion equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1783–1802.
9. M. Vohralík, *A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows*. In preparation, 2011.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Two-Dimensional Relaxation Scheme for the Hybrid Modelling of Two-Phase Flows

Kateryna Dorogan, Jean-Marc Hérard, and Jean-Pierre Minier

Abstract Recently, a new relaxation scheme for hybrid modelling of two-phase flows has been proposed. This one allows to obtain stable unsteady approximations for a system of partial differential equations containing non-smooth data. This paper is concerned with a two-dimensional extension of the present method, in which two alternative relaxation schemes are compared. A short stability analysis is given.

Keywords Finite Volumes, Relaxation schemes, Hybrid methods
MSC2010: 76M12, 65M12

1 Introduction

This paper deals with the modelling and the numerical simulation of polydispersed turbulent two-phase flows, where one phase is a turbulent fluid (considered to be a continuum) and the other appears as separate inclusions carried by the fluid (solid particles, droplets or bubbles). Such a kind of flows can be encountered in many industrial situations (combustion, water sprays, smokes) and in some environmental problems. Despite the need of their accurate prediction, the physical complexity of these processes is so broad that existing methods are either too expensive (in calculation cost) or not sufficiently accurate. A hybrid approach recently proposed in [2] enables to reach an acceptable compromise between the

Kateryna Dorogan
EDF R&D, MFEE, 6 quai Watier, F-78400 Chatou
and
LATP, CMI, 39 rue Joliot Curie, F-13453 Marseille, e-mail: kateryna.dorogan@edf.fr

Jean-Marc Hérard and Jean-Pierre Minier
EDF R&D, MFEE, 6 quai Watier, F-78400 Chatou, e-mail: jean-marc.herard@edf.fr, jean-pierre.minier@edf.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_37, © Springer-Verlag Berlin Heidelberg 2011

physical realism and a cheap numerical treatment. For two-phase flows, it consists in coupling two classic approaches (Eulerian and Lagrangian) in the particle phase description. This method allows to gather the advantages of classic approaches: high level of physical description, lower calculation costs, correct treatment of non-linearities and polydispersity, expected values free from statistical error. From now on, *"L"* and *"E"* superscripts respectively refer to all quantities calculated with the Lagrangian and Eulerian descriptions, and subscript *"p"* is used for the particle phase. The Lagrangian part of the particle phase description is given by the stochastic differential equations:

$$dZ_i(t) = A_i(t, Z, f(t;z), Y)dt + \sum_j B_{ij}(t, Z, f(t;z), Y)dW_j(t), \tag{1}$$

where $f(t;z)$ stands for the probability density function (pdf) of the particle state vector $Z = (x_p, U_p, U_s)$ with $x_p(t)$ the particle position, $U_p(x_p(t), t)$ the particle velocity, $U_s(x_p(t), t)$ the fluid velocity seen at the particle position and the local relative velocity $U_r = U_s - U_p$, whereas external mean fields, i.e. the fluid mean fields defined at particle locations [9, 10] are denoted by Y. A_i and B_{ij} represent the drift vector and the diffusion matrix, and $W(t)$ the vector of independent Wiener processes. Here we assume that the particles are only influenced by the drag and the gravity forces. Then, using corresponding Fokker-Planck equation we deduce from (1) a system of equations for the mean particle concentration α_p^E and flow rate $\alpha_p^E \langle U_{p,i}^E \rangle$, which represents an Eulerian description of the particle phase:

$$\begin{aligned} &\partial_t \alpha_p^E + \partial_{x_i}\left(\alpha_p^E \left\langle U_{p,i}^E \right\rangle\right) = 0 \\ &\partial_t \left(\alpha_p^E \langle U_{p,i}^E \rangle\right) + \partial_{x_j}\left(\alpha_p^E \left(\langle U_{p,i}^E \rangle \langle U_{p,j}^E \rangle + \langle u_{p,i} u_{p,j} \rangle^L\right)\right) = \alpha_p^E (g_i + \left\langle \frac{U_{r,i}^L}{\tau_p^L} \right\rangle) \end{aligned} \tag{2}$$

Usually, only one among the two systems (1), (2) is solved. However, in this case we are faced with shortcomings of the standard methods. In fact, system (1) contains a bias-error and thus needs calculations with a larger number of particles, whereas the Reynolds stress term $\langle u_{p,i} u_{p,j} \rangle^L$ in system (2) is not closed. The new hybrid approach consists in solving both of these systems at the same time. Thus, the terms with superscript *"L"*, calculated with a better accuracy in the Lagrangian part of the model, are provided to the Eulerian part (2). The latter, in turn, gives the values of $\langle U_{p,i}^E \rangle$ free from statistical error, that enable computations with a smaller number of particles in (1). Hence, for the same accuracy, the total calculation cost is reduced with reference to the pure Lagrangian approach. However, such a coupling introduces noisy quantities (computed by the stochastic equations) in the Eulerian part of the model, which presents an important convective part and thus requires a stabilization. A specific relaxation approach was proposed in [3,4] in order to tackle this problem in a one-dimensional case. It relies both on upwinding techniques and relaxation tools [8] and it allows to obtain stable unsteady approximations of solutions of system (2), even with noisy data $\langle u_{p,i}u_{p,j}\rangle^L$. Actually, two slightly distinct relaxation systems were examined and compared in references [3,4]. The present paper is concerned with a two-dimensional extension of these relaxation approaches. In section 2, we propose two forms of the relaxation system that are very similar and give the motivation for such a choice. Some stability results are presented in section 3 and we briefly describe the numerical treatment and results in section 4. We recall that the density of particles is constant.

2 Relaxation approach in a two-dimensional framework

From now on, we omit the superscripts *"E"* and subscripts *"p"*, and introduce the -constant- density of particles ρ_p. Thus we denote by $\rho = \alpha_p^E \rho_p$ the mean density distribution of the particles in the domain, by $U_i = \left\langle U_{p,i}^E \right\rangle, i = 1,2$ the mean particle velocity. Hence, for given non-smooth values of the Lagrangian Reynolds stress tensor $\underline{\underline{R}}_{ij}^L = \langle u_{p,i}u_{p,j}\rangle^L$, we want to compute stable approximations of solutions of:

$$\begin{aligned} &\partial_t \rho + \partial_{x_j}(\rho U_j) = 0 \\ &\partial_t(\rho U_i) + \partial_{x_j}(\rho U_i U_j) + \partial_{x_j}(\rho R_{ij}^L) = \rho g_i + \rho \left\langle U_{r,i}/\tau_p \right\rangle^L \end{aligned} \tag{3}$$

By construction, $\underline{\underline{R}}_{ij}^L$ complies with the *realisability condition*: $\underline{x}^t \underline{\underline{R}}^L \underline{x} \geq 0$ for all $\underline{x} \in \mathscr{R}^2$. Since non-smooth external data $\underline{\underline{R}}_{ij}^L$ are introduced in the system (3), we are formally interested in finding discontinuous solutions. In order to overcome this difficulty, a relaxation technique was proposed in [5,6], which is in fact grounded on ideas developed in [1]. It consists in introducing new variables $\underline{\underline{R}}_{ij}$ (that are expected to relax towards $\underline{\underline{R}}_{ij}^L$ when a given relaxation time scale τ_p^R tends to 0), and supplementary partial differential equations that govern the time evolution of the Reynolds stresses $\underline{\underline{R}}_{ij}$, in such a way that the new relaxation system is hyperbolic and preserves the realizability of solutions ($\underline{x}^t \underline{\underline{Rx}} \geq 0$ for $\underline{x} \in \mathscr{R}^2$). On the basis of [1,7], the following relaxation system naturally arises:

$$\begin{aligned} &\partial_t \rho + \partial_{x_j}(\rho U_j) = 0 \\ &\partial_t(\rho U_i) + \partial_{x_j}(\rho U_i U_j) + \partial_{x_j}(\rho R_{ij}) = 0 \\ &\partial_t(\rho R_{ij}) + \partial_{x_k}(\rho U_k R_{ij}) + \rho(R_{ik}\partial_{x_k}U_j + R_{jk}\partial_{x_k}U_i) = \rho(R_{ij}^L - R_{ij})/\tau_p^R \end{aligned} \tag{4}$$

Since this system is invariant under frame rotation, we consider the reference frame $(\underline{n}, \underline{\tau})$: $\underline{n} = (n_x, n_y)$, $\underline{\tau} = (-n_y, n_x)$, such that $n_x^2 + n_y^2 = 1$, for a given interface whose normal is $\underline{n}$. We also introduce: $U_n = \underline{U}.\underline{n}$, $U_\tau = \underline{U}.\underline{\tau}$, $R_{nn} = \underline{n}^t.\underline{\underline{R}}.\underline{n}$, $R_{n\tau} = \underline{n}^t.\underline{\underline{R}}.\underline{\tau} = \underline{\tau}^t\underline{\underline{R}}.\underline{n} = R_{\tau n}$, $R_{\tau\tau} = \underline{\tau}^t.\underline{\underline{R}}.\underline{\tau}$. When neglecting transverse variations (i.e. $\forall\phi : \partial\phi/\partial\tau = 0$), the relaxation system corresponding to system

(4) written in terms of variable $Z^t = (\rho, U_n, U_\tau, \rho R_{nn}, \rho R_{n\tau}, S)$ takes the following form for smooth solutions:

$$\partial_t Z + A_n(Z)\partial_n Z = \mathscr{S}(Z), \tag{5}$$

with: $Z = Z(t, x_n)$, $S = \left((\rho R_{nn})(\rho R_{\tau\tau}) - (\rho R_{n\tau})^2\right)/\rho^4$ and noting $\vartheta(x,t) = 1/\rho(x,t)$:

$$A_n(Z) = \begin{pmatrix} U_n & \rho & 0 & 0 & 0 & 0 \\ 0 & U_n & 0 & \vartheta & 0 & 0 \\ 0 & 0 & U_n & 0 & \vartheta & 0 \\ 0 & \Psi_{nn} & 0 & U_n & 0 & 0 \\ 0 & 2\rho R_{n\tau} & \Phi_{n\tau} & 0 & U_n & 0 \\ 0 & 0 & 0 & 0 & 0 & U_n \end{pmatrix}, \quad \mathscr{S}(Z) = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \rho(R^L_{nn} - R_{nn})/\tau^R_p \\ \rho(R^L_{n\tau} - R_{n\tau})/\tau^R_p \\ (S^L - S)/\tau^R_p \end{pmatrix}$$

$$\textit{where:} \qquad \Psi_{nn} = 3\rho R_{nn}, \quad \Phi_{n\tau} = \rho R_{nn}. \tag{6}$$

Eigenvalues of the homogeneous part of system (5) are:

$$\lambda_{1,6} = U_n \pm c_1, \quad \lambda_{2,5} = U_n \pm c_2, \quad \lambda_3 = \lambda_4 = U_n, \tag{7}$$

with $c_1^2 = \Psi_{nn}/\rho = 3R_{nn}$ and $c_2^2 = \Phi_{n\tau}/\rho = c_1^2/3$. Thus, system (5) is hyperbolic (unless vacuum occurs in the solution) if $\Psi_{nn} > 0$ and $\Phi_{n\tau} > 0$, thus if $R_{nn} > 0$. This first method associated with the choice (6), and refered to as **(A1)**, takes advantage of the hyperbolic structure of the set of PDE that governs Eulerian Reynolds stress components, while assuming classical closure laws [1]. Actually, we note that system (5) is characterized by four linearly-degenerate (LD) fields associated with $\lambda_{2,3,4,5}$ and by two genuinely non-linear (GNL) fields associated with $\lambda_{1,6}$. Details can be found in [6, 7]. A nice feature is that the whole set of partial differential equations in the evolution step preserves the realisability of the Reynolds stress tensor R_{ij}, both at the continuous and the discrete levels. This is in fact mandatory since eigenvalues remain real if and only if the quadratic form $n_i R_{ij} n_j$ remains positive (see the form of c_1, c_2 above). *However, a drawback in this approach is due to the true non-conservative form of the governing equations for the Reynolds stress components in (5).* Thus, non-conservative products that are active in genuinely non-linear fields are not uniquely defined.

This has motivated the introduction of a second form for (Ψ_{nn}, $\Phi_{n\tau}$) - corresponding to **(A2)**. The aim is to comply with specifications (i,ii): (i) the relaxation system should be hyperbolic, (ii) jump conditions in the relaxation system should be uniquely defined, field by field. The idea is to introduce functions which are close to (6), but such that non-conservative products are only effective through linearly degenerate fields. Introducing $(R_{nn})_0 > 0$ and choosing functions (Ψ_{nn}, $\Phi_{n\tau}$) as:

$$\Psi_{nn} = 3\rho_0^2(R_{nn})_0\vartheta, \quad \Phi_{n\tau} = \Psi_{nn}/3 = \rho_0^2(R_{nn})_0\vartheta, \tag{8}$$

we note that the relaxation system corresponding to *system (5) with the choice (8) is hyperbolic and it is characterized by 6 LD fields*; thus the jump relations are uniquely defined. Eigenvalues of the homogeneous part of the modified system (5) are now:

$$\lambda'_{1,6} = U_n \pm c'_1, \quad \lambda'_{2,5} = U_n \pm c'_2, \quad \lambda'_3 = \lambda'_4 = U_n, \tag{9}$$

with $(c'_1)^2 = 3(c'_2)^2 = 3\rho_0^2(R_{nn})_0\vartheta^2$. This method associated with the choice (8) will be refered to as **(A2)**. We provide below some properties of approaches **(A1, A2)**, assuming that the initial conditions are physically relevant:

$$\rho > 0, \quad \underline{x}^t.\underline{\underline{R}}.\underline{x} > 0. \tag{10}$$

Property 1 (Existence and Uniqueness of the solution of the Riemann problem for A1). *The Riemann problem associated with (5), (6), approximate jump relations given in [6, 7], and initial conditions for left and right states Z_L, Z_R in agreement with condition (10), admits a unique solution if:*

$$(U_n)_R - (U_n)_L < \sqrt{3}\left(\sqrt{(R_{nn})_L} + \sqrt{(R_{nn})_R}\right). \tag{11}$$

The solution is composed of six constant states $Z_L, Z_1, Z_2, Z_3, Z_4, Z_R$ separated by 2 GNL waves associated with $\lambda_{1,6}$ and 4 LD waves associated with $\lambda_{2,3,4,5}$.

Property 2 (Existence and Uniqueness of the solution of the Riemann problem for A2). *Assume that $\rho_0^2(R_{nn})_0 \geq 0$ is such that it satisfies the Wave Ordering Condition (WOC): $\lambda'_1 < \lambda'_2 < \lambda'_3 = \lambda'_4 < \lambda'_5 < \lambda'_6$. Then the Riemann problem associated with (5), (8) and initial conditions Z_L, Z_R satisfying (10), admits a unique solution composed of six constant states $Z_L, Z'_1, Z'_2, Z'_3, Z'_4, Z_R$ separated by 6 LD waves. The WOC is the same as in the pure one-dimensional framework (see [3, 4]).*

Property 3 (Positivity of interface values of the density).

• *The realisability in approach* **(A1)** *is ensured by condition (11) and density intermediate states are such that: $\rho_1 = \rho_2 > 0$, $\rho_3 = \rho_4 > 0$;*

• *For* **(A2)***, the latter condition (11) is replaced by the WOC, that guarantees the positvity of the densities in intermediate states: $\rho'_1 = \rho'_2 > 0$, $\rho'_3 = \rho'_4 > 0$.*

Remark 1 (Positivity of interface values of Reynolds stresses). *In approach (***A1***), the realisability of the Reynolds stress tensor is required to ensure the hyperbolicity property for the corresponding relaxation system and, at the same time, is preserved by the very construction of this system. In approach (***A2***) the realisability of Reynolds stresses in the intermediate states is not preserved for any initial condition; however, the hyperbolicity of the relaxation system in (***A2***) holds since $\vartheta^2 > 0$ and the realisability is recovered through the instantaneous relaxation step.*

3 Stability properties of approaches A1, A2

We focus now on the evolution step in the relaxation procedure, thus on the homogeneous system corresponding to the left hand side of (4). In order to give an estimation of the mean kinetic energy, which characterises the initial system of equations (3), we focus only on smooth solutions (we assume that: $\rho(\underline{x},t)$, $U_i(\underline{x},t)$, $R_{ij}(\underline{x},t) \in \mathscr{C}^1$, $i,j=1,2$), and we study the evolution of the "total" energy in the relaxation system (4). Let us denote by

$$\mathscr{E}_1(t) = \frac{1}{2}\int_\Omega \rho U_i^2(\underline{x},t)d\Omega \quad \text{and} \quad \mathscr{E}_2(t) = \frac{1}{2}\int_\Omega \rho\, tr(\underline{\underline{R}})(\underline{x},t)d\Omega, \quad i=1,2. \tag{12}$$

the kinetic energy of the drift (the mean motion) and the energy of the fluctuating particle motion. The total particle energy is given by $\mathscr{E}(t) = \mathscr{E}_1(t) + \mathscr{E}_2(t)$. We also assume that: $\forall \underline{x} \in \partial\Omega \quad \underline{U}.\underline{n} = 0$.

Property 4 (Energy estimation for A1). *We define: $\delta = R_{11}R_{22} - R_{12}^2$ and we assume that $\delta(\underline{x} \in \partial\Omega, t > t_0) > 0$, $\delta(\underline{x} \in \Omega, t_0) > 0$. Then smooth solutions of the homogeneous relaxation system corresponding to approach (**A1**) (left-hand side of system (4)) satisfy the following energy estimate:*

$$0 \le \mathscr{E}_1(t) = \mathscr{E}(t_0) - \mathscr{E}_2(t) \le \mathscr{E}(t_0), \quad since \quad \mathscr{E}_2(t) \ge 0. \tag{13}$$

An important ingredient in the proof is linked with the fact that the governing equation of $X = \delta/\rho^2$ reads:

$$\partial_t X + (\underline{U}\cdot\nabla)X = 0. \tag{14}$$

However we can only give a partial estimation for approach (**A2**). Actually, for the system corresponding to (5), (8), we must introduce a modified definition of the total energy in a *pure 1D framework* in order to get some estimation (see [4]):

$$\tilde{\mathscr{E}} = \mathscr{E}_1(t) + \mathscr{E}_2(t) + \int_\Omega \frac{\rho(a_0^2\vartheta - 3\rho R_{nn})^2}{16a_0^2}d\Omega, \quad \text{with } a_0^2 = 3\rho_0^2(R_{nn})_0, \tag{15}$$

Remark 2 (Energy estimation for A2). *In a one-dimensional framework, smooth solutions of the homogeneous relaxation system corresponding to (5), (8) satisfy:*

$$0 \le \mathscr{E}_1(t) \quad \text{and} \quad \mathscr{E}_1(t) + \mathscr{E}_2(t) \le \tilde{\mathscr{E}}(t_0). \tag{16}$$

4 Numerical algorithm and results

In order to compute the approximations of solutions of system (3) at each time step, the Finite Volume method relies on a classical fractional step method, which proceeds in three distinct steps (Evolution/Instantaneous Relaxation/Sources):

- **Step 1 (Evolution):** Starting from ρ^n, $(\rho U_i)^n$, $(\rho R_{ij})^n$, compute approximate solutions $\rho^{n+1,-}$, $(\rho U_i)^{n+1,-}$, $(\rho R_{ij})^{n+1,-}$ of the homogeneous system corresponding to the left hand side of (4) at time t^{n+1}, using an approximate Godunov solver for (**A1**) [7] or an exact Godunov solver for (**A2**) (using property 2, see [3,4]).

- **Step 2 (Relaxation):** restore local values of the Reynolds stresses $R_{ij} = R^L_{ij}$:

$$\rho^{n+1} = \rho^{n+1,-}, \quad (\rho U_i)^{n+1} = (\rho U_i)^{n+1,-}, \quad (\rho R_{ij})^{n+1} = \rho^{n+1}(R^L_{ij})^{n+1}.$$

- **Step 3 (Sources):** account for physical source terms (right hand side of (3)).

An extensive validation of both methods **(A1, A2)** has been achieved in [3, 4] in the one-dimensional framework, by computing the L^1 norm of the error for analytical solutions of Riemann problems associated with the homogeneous part of (3), assuming specific forms for $R^L_{ij} = r_{ij}(\rho, U)$. We only show here a few computations and we put emphasis on the main conclusions.

Analytical test cases: In order to validate the two approaches (**A1, A2**), we consider some test cases where analytical solutions are known and we focus especially on the most difficult configurations. Assuming the following closure relation:

$$R^L_{ij} = S_0 \rho^{\gamma-1} \delta_{ij}$$

with $S_0 = 10^5$ and $\gamma = 3$ (this value of γ corresponds to the isentropic case arising in [1, 7]), we focus on two 1D Riemann problems. The computational domain is a square $[-1, 1]^2$, the time step is in agreement with the CFL condition (CFL = 0.49), and the regular meshes contain from 2×10^2 up to 2×10^5 cells. The figures below (Fig. 1) represent the L^1-norm of the errors w.r.t. the mesh size. *On the whole, both methods (A1, A2) guarantee the correct convergence of approximations, even when the solution contains strong shocks*. This is very encouraging and not obvious since **(A1)** involves non conservative products in GNL fields, which means that we might expect to retrieve convergence of approximations towards *wrong* shock solutions. Though we have no proof at all for that, the fact that the scheme preserves the conservative form of the first two equations of (4) might explain this good behaviour. Moreover, we note (see Figs. 1) that *we retrieve the classical h^1 convergence* since no LD wave is involved here. Whereas (**A1**) and (**A2**) schemes exhibit almost the same accuracy, *(**A2**) seems to be a little bit more stable than (**A1**). Eventually, both schemes can handle vacuum occurence and strong shock waves.*

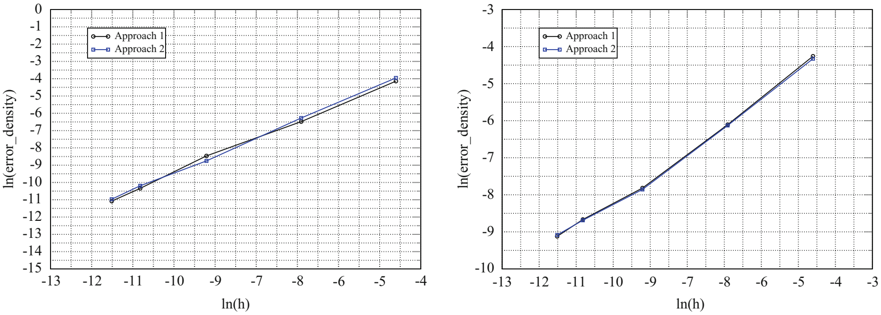


Fig. 1 L1 convergence curves for symmetric double shock (left) and symmetric double rarefaction waves with vacuum occurence (right). Coarser mesh: 200 cells; finer mesh: 200000 cells

Numerical results with noisy Reynolds stresses: We choose the initial conditions of a subsonic shock tube problem and we plug noisy Reynolds stresses in the system of equations (3) at each time step in the cells that belong to the region $(x, y) \in [-0.25, 0.25] \times [-1, 1]$ with: $R_{ij}^L = S_0\rho^{\gamma-1}(1 + rms(0.5 - rand(0, 1)))\delta_{ij}$, where rms stands for the noise intensity and $rand$ allow to manage the noise amplitude. The noisy region is not developping in time (Fig. 2). The same remark holds for other values of the noise intensity. Other test cases with noisy data [3] show that the noise is independent of the mesh refinement. Eventually, the difference between noisy approximations and those without a noise is increasing with rms in a linear manner. Both methods are stable.

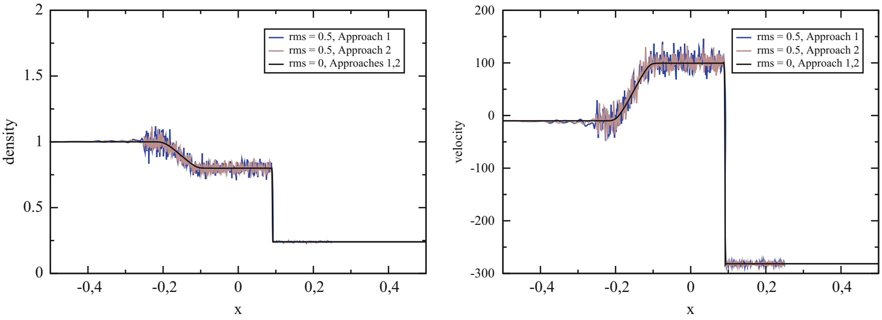


Fig. 2 Approximations of the density (left) and the velocity (right) with rms = 0.5 and rms = 0 in time. Mesh size: 1000 cells in the x-direction

Acknowledgements The first author receives a financial support through an EDF-CIFRE contract 203/2009. Computational facilities were provided by EDF.

References

1. Berthon, C., Coquel, F., Hérard, J.M., Uhlmann, M.: An approximate solution of the Riemann problem for a realisable second-moment turbulent closure. Shock Waves, **11**, 245–269 (2002)
2. Chibbaro, S., Hérard, J.M., Minier, J.P.: A novel Hybrid Moments/Moments-PDF method for turbulent two-phase flows. Final Technical Report Activity Marie Curie Project. TOK project LANGE Contract MTKD-CT-2004 509849 (2006)
3. Dorogan, K., Hérard, J.M., Minier, J.P.: Development of a new scheme for hybrid modelling of gas-particle two-phase flows. EDF report H-I81-2010-2352-EN, unpublished, 1–50 (2010)
4. Dorogan, K., Hérard, J.M., Minier, J.P.: A relaxation scheme for hybrid modelling of gas-particle flows. Submitted (2011)
5. Hérard, J.M.: A relaxation tool to compute hybrid Euler-Lagrange compressible models. AIAA paper 2006-2872 (2006) http://aiaa.org
6. Hérard, J.M., Minier, J.P., Chibbaro, S.: A Finite Volume scheme for hybrid turbulent two-phase flow models. AIAA paper 2007-4587, http://aiaa.org
7. Hérard, J.M., Uhlmann, M., Van der Velden, D.: Numerical techniques for solving hybrid Eulerian Lagrangian models for particulate flows. EDF report H-I81-2009-3961-EN (2009)
8. Jin, S., Xin, Z.: The relaxation schemes for systems of conservation laws in arbitrary space dimensions. Comm. Pure Appl. Math., **48**, 235–276 (1995)
9. Minier, J.P., Peirano, E.: The pdf approach to polydispersed turbulent two-phase flows. Physics reports, **352**, 1–214 (2001)
10. Peirano, E., Chibbaro, S., Pozorski, J., Minier, J.P.: Mean-field/PDF numerical approach for polydispersed turbulent two-phase flows. Prog. Ene. Comb. Sci., **32**, 315–371 (2006)

The paper is in final form and no similar paper has been submitted elsewhere.

Finite Volume Method for Well-Driven Groundwater Flow

Milan Dotlić, Dragan Vidović, Milan Dimkić, Milenko Pušić, and Jovana Radanović

Abstract Finite volume method for well-driven porous media flow which uses a computational mesh tailored for finite elements is presented. It replaces one-dimensional elements used to model well drains in the original mesh with one-dimensional cells. It does not modify the original mesh by adding or moving nodes. It can handle the discontinuous anisotropic hydraulic conductivity. Special discretization of the flux between the porous medium and the drain is proposed. Numerical results are compared to an analytical solution.

Keywords Finite volume method, well-driven flow, porous media, privileged routes, 1d elements
MSC2010: 76S05, 65N08

1 Introduction

A significant number of finite element codes for well-driven groundwater flow simulation is available (FEFLOW [1], HydroGeoSphere [6], PAKP–Lizza [2], etc). Well drains are represented in these codes as arrays of one-dimensional elements, i.e. mesh edges. Triangulators, such as Triangle [5], allow the user to specify the exact location of these drains prior to triangulation, and place the mesh nodes and the edges at the specified locations.

Milan Dotlić, Dragan Vidović, Milan Dimkić, and Jovana Radanović
Jaroslav Černi Institute, Jaroslava Černog 80, 11226 Pinosava, Belgrade, Serbia,
e-mail: milandotlic@gmail.com, draganvid@gmail.com jdjcerni@jcerni.co.rs,
jovanaradanovic@gmail.com

Milenko Pušić
University of Belgrade, Faculty of Mining and Geology, Đušina 7, 11000 Belgrade, Serbia,
e-mail: mpusic@ptt.rs

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_38, © Springer-Verlag Berlin Heidelberg 2011

This arrangement is appropriate for finite element method because it associates the discrete variables with the mesh nodes. Cell-centered finite volume methods associate the discrete variables with mesh cells. Thus, matching the cell center with the exact well location and representing a drain as an array of cells would be more appropriate.

Since numerous tools exist to construct finite element meshes, our goal is to find a suitable way to use these meshes with finite volumes.

One possibility to obtain a suitable finite volume mesh is to construct a dual of a finite element mesh. However, the conductivity, which is associated with the finite element mesh cells and may be discontinuous between the cells, will now be associated with the nodes of the dual mesh. In finite volumes the conductivity is also associated with cells. Therefore, some kind of interpolation must be performed in order to compute the dual cell conductivity, which may introduce significant error because the conductivity may vary by several orders of magnitude between geological layers.

Another possibility is to associate a fictive one-dimensional cell with each 1d element. These new cells are connected with the surrounding three-dimensional cells by one-dimensional faces, and with each other by zero-dimensional faces. In order to compute non-zero fluxes and finite hydraulic heads, physical surfaces and volumes of the real drain portions must be associated with these new entities.

In this paper we present details of such discretization. Groundwater flow equation is given in Section 2, together with boundary conditions and a well clogging model. Interpretation of the mesh is explained in Section 3, and the flux and the boundary conditions discretization is specified. Obtained numerical results are compared to an analytical solution in Section 4, and a correction to the flux discretization between the porous medium and a drain is proposed.

2 Problem formulation

Correlation between the hydraulic head gradient ∇h and the flux density $\mathbf{q}$ is known as Darcy's law

$$\mathbf{q} = -K\nabla h, \tag{1}$$

where the hydraulic conductivity K is in general a symmetric positive piecewise-continuous anisotropic tensor.

Mass conservation is expressed trough the groundwater flow equation

$$S\frac{\partial h}{\partial t} = -\nabla \cdot \mathbf{q}, \tag{2}$$

where S is the specific storage. Substituting (1) into (2) results in a form suitable for solving

$$S\frac{\partial h}{\partial t} = \nabla \cdot (K\nabla h). \tag{3}$$

Domain boundary is divided in two parts $\partial\Omega = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, and Dirichlet and Neumann boundary conditions are specified:

$$h = g_D \qquad \text{on } \Gamma_D, \tag{4}$$

$$(-K\nabla h)\mathbf{n} = g_N \qquad \text{on } \Gamma_N, \tag{5}$$

where $\mathbf{n}$ is the outward unit normal to $\partial\Omega$. In addition to these common boundary conditions, either the hydraulic head or the total flux per unit of time Q is specified in each well.

Initial condition is

$$h|_{t=0} = h_0, \tag{6}$$

and the final time is $t = T$.

Drain clogging, which happens due to complex mechanical, chemical, and biological processes [3, 4], results in a colmated layer along the drain wall, which causes an additional hydraulic resistance. This can be expressed as

$$\mathbf{q} \cdot \mathbf{n} = \Psi(h_f - h_w), \tag{7}$$

where h_w is the hydraulic head inside the well, h_f is the hydraulic head just outside the colmated layer, $\mathbf{n}$ is the unit normal to the drain wall pointing inside, $\Psi = K_c/d_c$ is the transfer coefficient, K_c is the unknown conductivity of the colmated layer, and d_c is it's unknown thickness.

3 Discretization

Integrating (3) over polyhedral control volume T, and using the divergence theorem and implicit Euler time integration results in

$$|T|S\frac{h^{n+1} - h^n}{\Delta t} = \sum_{f\in\partial T} \chi_{T,f} Q_f^{n+1}, \qquad Q_f = \int_f \mathbf{q} \cdot \mathbf{n}_f ds, \tag{8}$$

where Q_f is the flux through face f, $\mathbf{n}_f$ is a unit vector normal to face f fixed once and for all, and $\chi_{T,f} = 1$ if $\mathbf{n}_f$ points outside of T, or -1 otherwise. At boundary faces, fixed normal vectors point outside.

3.1 Drains

It is assumed that well drains coincide with the mesh edges (see Fig. 1). Each well may have one or more connected drains.

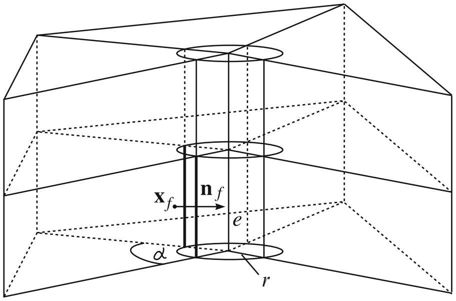


Fig. 1 Drain discretization

A cylindrical cell called 1d cell is associated with each edge e belonging to a well drain. This cell is logically plugged into the grid by defining interfaces between it and the surrounding cells, but nodes are not added or moved. A volume $r^2\pi|e|$, where r is the drain radius and $|e|$ is the edge length, is associated with this cell. Each volume T sharing the edge e is reduced by $r^2\frac{\alpha}{2}|e|$, where α is the angle between the faces of the cell T sharing the edge e.

Interface f between the 3d cell T and a 1d cell is called 1d face. This is a portion of the cylinder with surface $\alpha r|e|$. Unit normal $\mathbf{n}_f$ belongs to the bisector plain of the angle α and it points into the cylinder. Center $\mathbf{x}_f$ of the face f is obtained by shifting the centroid of the 1d cell by vector $-r\mathbf{n}_f$.

Interfaces between the 1d cells are so-called 0d faces. These are circles with radius r associated with nodes where the drain edges meet.

If more drains meet in a node, a 0d cell is introduced. This cell has zero volume and it is connected with each of the drains by a 0d face.

Hydraulic head or the total well flux is specified at a boundary 0d face, which may be at a drain end, or it may be an extra face introduced in a 0d cell in order to impose a boundary condition.

Hagen–Poiseuille law is used for a flow through a pipe, which means that within the drains we take $k = r^2/8$, $K = k\rho g/\mu$, where $\rho = 1000\text{kg/m}^3$ is the water density, $\mu = 0.001307\text{kg/(ms)}$ is the dynamic viscosity of water, and $g = 9.81\text{m/s}^2$.

3.2 Flux

It is assumed that K is continuous within cells. Possible discontinuities happen along faces. At the interface between a 3d cell and a drain, K is discontinuous.

If f is an internal face and if K is continuous in f, then

$$Q_f = -|f|\|K\mathbf{n}\|\frac{h_{out} - h_{in}}{\|\mathbf{x}_{out} - \mathbf{x}_{in}\|}, \tag{9}$$

where $|f|$ is the face f area, and $\mathbf{x}_{in}$ and $\mathbf{x}_{out}$ are the centroids of the cells sharing the face f such that n points from cell in to cell out. This is the most basic finite volume flux discretization used here for simplicity, and it is not very accurate. A more accurate non-linear flux discretization [7] is planned.

If K is discontinuous in f, then two one-sided flux approximations

$$Q_f = -|f|\|K_{out}\mathbf{n}\|\frac{h_{out} - h_f}{\|\mathbf{x}_{out} - \mathbf{x}_f\|}, \qquad Q_f = -|f|\|K_{in}\mathbf{n}\|\frac{h_f - h_{in}}{\|\mathbf{x}_f - \mathbf{x}_{in}\|} \tag{10}$$

are combined to compute the face hydraulic head and eliminate it from (10):

$$Q_f = -|f|\frac{\Psi_{in}\Psi_{out}}{\Psi_{in} + \Psi_{out}}(h_{out} - h_{in}), \tag{11}$$

$$\Psi_{in} = \frac{\|K_{in}\mathbf{n}\|}{\|\mathbf{x}_{in} - \mathbf{x}_f\|}, \qquad \Psi_{out} = \frac{\|K_{out}\mathbf{n}\|}{\|\mathbf{x}_f - \mathbf{x}_{out}\|}. \tag{12}$$

At the interface between a drain and a 3d cell, transfer coefficient Ψ defined in (7) is substituted in (11) instead of Ψ_{out}.

Second formula in (10) is inadequate when the drain radius is much smaller than the mesh size. This is demonstrated in Section 4, and a correction is proposed.

3.3 Boundary conditions

If f is a boundary face, one-sided flux approximation is used:

$$Q_f = -|f|\|K\mathbf{n}\|\frac{h_f - h_{in}}{\|\mathbf{x}_f - \mathbf{x}_{in}\|}. \tag{13}$$

Flux Q_f or hydraulic head h_f imposed at f is used trough this relation. If f is a 0d face, then $K = K_{drain}$.

4 Flux correction

Example 1. If K is constant, then

$$h(\rho) = A + B\ln\rho \tag{14}$$

is a stationary solution of (3), where ρ is the distance from the well central axis. If domain Ω is a cylinder with radius R and a well of radius r at the center (see Fig. 2), then A and B can be found from the requirement that $h(r) = h_r$ and $h(R) = h_R$,

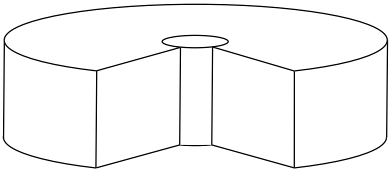


Fig. 2 Example domain

for some specified h_r and h_R. The resulting solution is

$$h(\rho) = \frac{h_r \ln \frac{R}{\rho} + h_R \ln \frac{\rho}{r}}{\ln \frac{R}{r}}. \tag{15}$$

Total well flux is

$$Q = 2\pi K H \frac{h_R - h_r}{\ln \frac{R}{r}}, \tag{16}$$

where H is the cylinder height. Axial well resistance has been neglected here.

In order to incorporate the resistance of the colmated layer, we compute Ψ from (7) using the exact flux obtained in (16), so that we obtain a desired hydraulic head decrease in the well $h_r - h_w$ due to colmation.

We choose $K = 10^{-4}$, $R = 20$, $H = 10$, $h_R = 100$, $h_r = 60$, $h_w = 55$, and compute fluxes for two different well radii, $r = 0.5$ and $= 0.01$, to test the scheme in cases when the well radius is close to the mesh size, and when it is much smaller than the mesh size. The computational mesh with the maximal base triangle area of 0.5 is shown in Fig. 3.

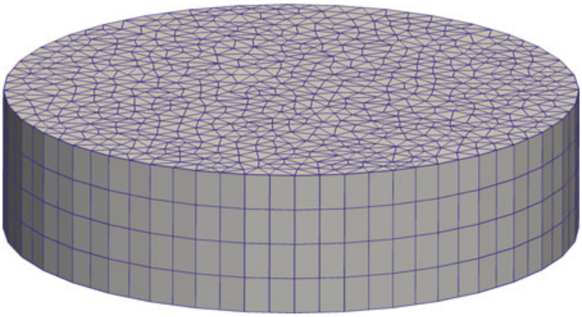


Fig. 3 Computational grid

Exact hydraulic head is specified at the inner and the outer cylinder. Zero flux is specified at the flat boundaries.

The exact and the numerical fluxes are given in Table 1. If the well radius is much smaller than the mesh size, the flux is about ten times smaller than what it should

Table 1 Exact and numerical fluxes for $r = 0.01$ and $r = 0.5$, with and without the proposed correction

	Exact Q	Numerical Q	Q with correction
$r = 0.5$	0.06813	0.06634	0.06894
$r = 0.01$	0.03306	0.00322	0.03312

be. The reason is that $h'(\rho)$ is very sharp at $\rho = r$, and it is not well approximated with a finite difference.

To allow the computation on coarse meshes, we replace R in (15) with the distance between $\mathbf{x}_{in}$ and the central well axis $\rho(\mathbf{x}_{in})$, and use the derivative of this formula to derive a replacement for the second formula in (10) for the case of a 3d-1d cell interface:

$$Q_f = -|f|\|K_{in}\mathbf{n}\| \frac{h_f - h_{in}}{r \ln \frac{\rho(\mathbf{x}_{in})}{r}}. \tag{17}$$

Results presented in the last column of Table 1 show that the total well flux computed with this correction is much more accurate.

Example 2. Another stationary analytical solution of (3) can be obtained by superposing a linear solution $h = cy$ to (15), where c is an arbitrary constant:

$$h = \frac{h_r \ln \frac{R}{\rho} + h_R \ln \frac{\rho}{r}}{\ln \frac{R}{r}} + cy. \tag{18}$$

The fluxes in the well which are due to the linear term cancel out, and the total well flux is the same as in Example 1. This solution is not constant at $\rho = r$ and we cannot set such a boundary condition in our method. However, presuming that r is small, h varies little around the well, thus specifying a constant h_r should give a close approximation. We take $c = 1$ and specify the exact hydraulic head at $\rho = R$. The obtained total well flux given in Table 2 shows that correction (17) improves the accuracy even if formula (15) on which the correction is based is not the exact solution, because the influence of the well is still dominant near the well.

Table 2 Numerical fluxes in Example 2 for $r = 0.01$ with and without the correction

Q without correction	Q with correction
0.00322	0.03309

5 Conclusion

We have presented a finite volume method for well-driven groundwater flow that uses a computational grid tailored for a finite element method, in which well drains are represented by one-dimensional elements. In its interpretation of the grid, our method adds faces and cells that correspond to well drains with geometry that is not fully resolved in the original grid. However, the original grid is not modified in the sense that nodes are not added or moved.

We compared the obtained numerical fluxes with the analytical solution in cases when the well radius is close to the grid resolution, and also when it is much smaller than the grid resolution, and we found that the match is poor for the small radius case. We proposed a correction to the discretization of the flux between the 3d porous medium and a drain. Total well flux obtained with this correction is very accurate in all cases, bearing in mind that the inaccurate linear two-point flux discretization was used.

References

1. H.-J.G. Diersch. *FEFLOW 5.3 user's manual.* WASY GmbH, Berlin, Germany, 2006.
2. M. Dimkić, M. Pušić, D. Vidović, N. Filipović, V. Isailović, and B. Majkić. Numerical model assessment of radial-well ageing. *ASCE's Journal of computing in civil engineering*, 25(1):43–49, 2010.
3. M. Dimkić and M. Pušić. Preporuke za projektovanje bunara uzevši u obzir kolmiranje gvoždjem na osnovu iskustva beogradskog izvorišta. *Gradjevinski kalendar*, 40:430–496, 2008.
4. M. Dimkić, M. Pušić, V. Obradović, and D. Djurić. Several natural indicators of radial well ageing at the belgrade groundwater source, part 2. *Water Science and Technology*, 2011. submitted WST-S-10-02140[1].
5. J. R. Shewchuk. Triangle: Engineering a 2d quality mesh generator and delaunay triangulator. *Computational Geometry: Theory and Applications*, 22(1):21–74, 2002.
6. R. Therrien, R.G. McLaren, E.A. Sudicky, and S.M. Panday. *HydroGeoSphere - A three-dimensional Numerical Model Describing Fully-integrated Subsurface and Surface Flow and Solute Transport.* Groundwater Simulations Group, 2010. Available on Internet:www.hydrogeosphere.org/hydrosphere.pdf.
7. D. Vidović, M. Dimkić, and M. Pušić. Accelerated non-linear finite volume method for diffusion. *Journal of Computational Physics*, 2011. doi: 10.1016/j.jcp.2011.01.016.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Adaptive Reduced Basis Methods for Nonlinear Convection–Diffusion Equations

Martin Drohmann, Bernard Haasdonk, and Mario Ohlberger

Abstract Many applications from science and engineering are based on parametrized evolution equations and depend on time-consuming parameter studies or need to ensure critical constraints on the simulation time. For both settings, model order reduction by the reduced basis methods is a suitable means to reduce computational time. In this proceedings, we show the applicability of the reduced basis framework to a finite volume scheme of a parametrized and highly nonlinear convection-diffusion problem with discontinuous solutions. The complexity of the problem setting requires the use of several new techniques like parametrized empirical operator interpolation, efficient a posteriori error estimation and adaptive generation of reduced data. The latter is usually realized by an adaptive search for base functions in the parameter space. Common methods and effects are shortly revised in this presentation and supplemented by the analysis of a new strategy to adaptively search in the time domain for empirical interpolation data.

Keywords Finite volume methods, model reduction, reduced basis methods, empirical interpolation
MSC2010: 65M08, 65J15, 65Y20

1 Introduction

Reduced basis (RB) methods are popular methods for model order reduction of problems with parametrized partial differential equations that need to be solved for many parameters. Such scenarios might occur in parameter studies, optimization,

M. Drohmann and M. Ohlberger
Institute of Computational and Applied Mathematics, University of Muenster, Einsteinstr. 62, 48149 Muenster, e-mail: mdrohmann,ohlberger@uni-muenster.de

B. Haasdonk
Institute of Applied Analysis and Numerical Simulation, University of Stuttgart, 70569 Stuttgart, e-mail: haasdonk@mathematik.uni-stuttgart.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_39, © Springer-Verlag Berlin Heidelberg 2011

control, inverse problems or statistical analysis for a given parametrized problem. Such problems deal with different solutions $u_h(\boldsymbol{\mu}) \in \mathscr{W}_h$ from a high dimensional discrete function space $\mathscr{W}_h \subset L^2(\Omega)$ which are characterized by a parameter $\boldsymbol{\mu} \in \mathscr{P} \subset \mathbb{R}^p$. For evolution problems, a discrete solution forms a series of what we call "snapshot solutions" $u_h^k(\boldsymbol{\mu})$ indexed by a time-step number $k = 0, \dots, K$.

By applying the reduced basis method, these solution trajectories need to be computed for a few parameters only and can then be used to span a problem-specific subspace $\mathscr{W}_{\text{red}} \subset \mathscr{W}_h$. If this subspace captures a broad solution variety, a numerical scheme based on this reduced basis space $\mathscr{W}_{\text{red}}$ can produce reduced solutions $u_{\text{red}}(\boldsymbol{\mu}) \in \mathscr{W}_{\text{red}}$ very inexpensively for every parameter $\boldsymbol{\mu} \in \mathscr{P}$. In case of nonlinear discretizations or complex dependencies of the equations on the parameter, the reduced scheme requires an empirical interpolation method [1] to efficiently interpolate operator evaluations in a low-dimensional discrete function space.

The applicability of the reduced scheme has been successfully demonstrated for stationary, instationary, linear and nonlinear problems mainly based on finite element schemes (cf. [7] and the references therein). In this presentation, we focus on a scalar, but highly nonlinear convection–diffusion problem: For a given parameter $\boldsymbol{\mu} \in \mathscr{P}$ determine solutions $u = u(x, t; \boldsymbol{\mu})$ fulfilling

$$\partial_t u + \nabla \cdot (\mathbf{v}(u; \boldsymbol{\mu}) u) - \nabla \cdot (d(u; \boldsymbol{\mu}) \nabla u) = 0 \qquad \text{in } \Omega \times [0, T_{\max}] \tag{1}$$

$$u(0; \boldsymbol{\mu}) = u_0(\boldsymbol{\mu}) \qquad \text{in } \Omega \times \{0\} \tag{2}$$

plus Dirichlet boundary conditions $u(\boldsymbol{\mu}) = u_{\text{dir}}(\boldsymbol{\mu})$ on $\Gamma_{\text{dir}} \times [0, T_{\max}]$, Neumann boundary conditions $(\mathbf{v}(u; \boldsymbol{\mu}) u - d(u; \boldsymbol{\mu}) \nabla u) \cdot \mathbf{n} = u_{\text{neu}}(\boldsymbol{\mu})$ on $\Gamma_{\text{neu}} \times [0, T_{\max}]$ with suitable parametrized functions $\mathbf{v}(\cdot; \boldsymbol{\mu}) \in C(\mathbb{R}, \mathbb{R}^d)$ and $d(\cdot; \boldsymbol{\mu}) \in C(\mathbb{R}, \mathbb{R}^+)$.

For complex data functions, solutions of this problem can depend on the parameter in a highly nonlinear way, and the convection term can lead to a variety of solution snapshots which is difficult to capture by a linear subspace $\mathscr{W}_{\text{red}}$. This makes the construction of the reduced basis space $\mathscr{W}_{\text{red}}$ difficult and therefore requires sophisticated construction algorithms for the reduced data. After elaborating on the empirical operator interpolation and the reduced basis scheme for problem (1)-(2) in Section 2, we provide an overview of such algorithms in Section 3 with a focus on the time-adaptive construction of interpolation for the empirical interpolation. In Section 4, we numerically discuss the effects and costs of the introduced algorithms based on a finite volume discretization of a Buckley–Leverett type problem.

2 Reduced basis method

In this section, we present a reduced basis method for general operator based discretizations of equations (1), (2). We show that the reduced scheme depends both in memory and computational complexity on the low dimensions of suitable reduced spaces only and can therefore be efficiently evaluated. We first introduce the basic approach, and discuss the main ingredients to efficiently compute the reduced solutions at the end of this section. For a more detailed presentation, we refer to [2].

As a starting point for the reduced basis scheme, we assume a high dimensional discretization scheme producing for each parameter $\boldsymbol{\mu} \in \mathscr{P}$ a sequence of solution snapshots $u_h^k(\boldsymbol{\mu})$ stemming from an H-dimensional discrete function space $\mathscr{W}_h$. The sequence indices $k = 0, \dots, K$ correspond to strictly increasing time steps $t^k := k\Delta t$ from the interval $[0, T_{\max}]$, where $\Delta t > 0$ is a global time step size. For the high-dimensional scheme, first, the initial data is projected on the discrete function space yielding a discrete solution $u_h^0(\boldsymbol{\mu}) = \mathscr{P}_h [u_0(\boldsymbol{\mu})]$, where $\mathscr{P}_h : L^2(\Omega) \to \mathscr{W}_h$ is a projection operator. Subsequently, equations of the form

$$(\mathrm{Id} + \Delta t \mathscr{L}_I(\boldsymbol{\mu}))\left[u_h^{k+1}(\boldsymbol{\mu})\right] - (\mathrm{Id} + \Delta t \mathscr{L}_E(\boldsymbol{\mu}))\left[u_h^k(\boldsymbol{\mu})\right] = 0 \tag{3}$$

are solved with the Newton–Raphson method. The operators $\mathscr{L}_I(\boldsymbol{\mu}), \mathscr{L}_E(\boldsymbol{\mu}) : \mathscr{W}_h \to \mathscr{W}_h$ describe the explicit and implicit discretization terms of a first order Runge–Kutta scheme. For our numerical experiments presented in Section 4, the operators implement finite volume fluxes for the diffusive respectively convective terms.

For the reduced basis scheme, we first assume a given reduced basis space $\mathscr{W}_{\mathrm{red}} \subset \mathscr{W}_h$ of dimension $N \ll H$. This space is spanned by selected solution snapshots and its construction implies a computationally expensive preprocessing step. This allows to solve for reduced solutions $u_{\mathrm{red}}^k(\boldsymbol{\mu}) \in \mathscr{W}_{\mathrm{red}}$. These are computed by projection of the initial data on the reduced basis space and with the same evolution scheme as in (3), but with the operators $\mathscr{L}_I(\boldsymbol{\mu}), \mathscr{L}_E(\boldsymbol{\mu})$ substituted by reduced counterparts

$$\mathscr{L}_{\mathrm{red},I}^{k+1}(\boldsymbol{\mu}) := \mathscr{P}_{\mathrm{red}} \circ \mathscr{I}_{M^{k+1}}^{k+1} \circ \mathscr{L}_I(\boldsymbol{\mu}) \quad \text{and} \quad \mathscr{L}_{\mathrm{red},E}^{k}(\boldsymbol{\mu}) := \mathscr{P}_{\mathrm{red}} \circ \mathscr{I}_{M^k}^{k} \circ \mathscr{L}_E(\boldsymbol{\mu}) \tag{4}$$

at each time instance $k = 0, \dots, K-1$. Here, $\mathscr{P}_{\mathrm{red}} : \mathscr{W}_h \to \mathscr{W}_{\mathrm{red}}$ is a further projection operator and the actual operator evaluations are substituted by approximations in a further low dimensional function space $\mathscr{W}_M \subset \mathscr{W}_h$. This approximation, the so-called *empirical operator interpolation*, is denoted by $\mathscr{I}_M \circ \mathscr{L}$ and shortly summarized in the next subsection. Note that in this scheme the empirical operator interpolation and therefore also the reduced function spaces can vary over time.

Empirical operator interpolation and offline/online splitting The idea of empirical interpolation was first introduced in [1]. The empirical operator interpolation presented here is extracted from [2].

The principal idea is to interpolate functions $v_h \in \mathscr{W}_h$ in a *collateral reduced basis space* $\mathscr{W}_M$ spanned by basis functions $q_m, m = 1, \dots, M$ with exact evaluations at interpolation points $x_m \in T_M$, i.e.

$$\mathscr{I}_M[v_h](x_m) = \sum_{m=1}^{M} \sigma_m q_m(x_m) = v_h(x_m), \tag{5}$$

where the coefficients can be determined easily because the construction process for the basis functions ensures for each $m = 0, \dots, M$ that the condition $q_m(x_{m'}) = 0$ is fulfilled for all $m' < m$. By optimizing the collateral reduced basis space such that it well approximates operator evaluations $\mathscr{L}_h(\boldsymbol{\mu})\,[v_h] \in \mathscr{W}_h$ of a parameterized discrete operator $\mathscr{L}_h(\boldsymbol{\mu})$ on solution snapshots v_h, we obtain an approximation $\mathscr{I}_M\,[\mathscr{L}_h(\boldsymbol{\mu})\,[v_h]] \approx \mathscr{L}_h(\boldsymbol{\mu})\,[v_h]$ which can be computed by evaluating the operator locally at M given interpolation points. If such an evaluation depends only on a few degrees of freedom of the argument function (H independent Dof-dependence) and $M \ll H$, the interpolation can be computed very efficiently. The interpolant is therefore suitable for the reduced basis method. Furthermore, it can be verified that the same argumentation also applies to Fréchet derivatives of discrete operators fulfilling the H independent Dof-dependence. This result is needed for the efficient implementation of the Newton–Raphson method. In Section 3.2, we summarize how the discrete function space $\mathscr{W}_M$ can be constructed by a greedy search algorithms in a finite set of operator evaluations.

In order to evaluate the reduced numerical scheme efficiently, the high dimensional data needs to be precomputed in an expensive offline phase and to be reduced to low-dimensional matrices and vectors. Afterwards, every Newton step of a reduced simulation can be computed with complexity $\mathscr{O}(NM^2 + N^3)$ including the costs of the linear equation solver. In [2], the computations leading to these results are presented in detail. The same article also introduces an efficiently computable a posteriori error estimator $\eta(\boldsymbol{\mu})$ estimating the error

$$\max_{k=0,\dots,K} \left\| u_h^k(\boldsymbol{\mu}) - u_{\mathrm{red}}^k(\boldsymbol{\mu}) \right\| \leq \eta(\boldsymbol{\mu}) \tag{6}$$

for a suitable problem-specific norm $\|\cdot\|$.

3 Adaptive basis generation strategies

In this section, we give an introduction on how reduced basis functions and empirical interpolation data are constructed by algorithms that greedily search in a finite subset of the parameter space for new basis functions. For complex parameter sets or complex dependencies of the solution on the parameter, these algorithms, however, can result in very large reduced basis spaces and therefore make the speed advantages of the reduced simulations obsolete. For this reason, we also discuss variations of the algorithms adapting the parameter search set during the basis construction which lead to smaller and better basis spaces.

3.1 POD-greedy algorithm

The "POD-greedy" algorithm introduced in [4] is used to generate the reduced basis space $\mathscr{W}_{\text{red}}$. Its purpose is to minimize the error $\|u_h(\boldsymbol{\mu}) - u_{\text{red}}(\boldsymbol{\mu})\|$ for all $\boldsymbol{\mu} \in \mathscr{P}$ in a suitable problem-specific norm. We assume the existence of an estimator $\eta(\boldsymbol{\mu})$ as introduced in (6), a finite training set $M_{\text{train}} \subset \mathscr{P}$ and an initial choice for the reduced basis $\Phi_{N_0} := \{\varphi_n\}_{n=1}^{N_0}$. For evolution problems, the span of this initial reduced basis usually comprises all initial data functions. Then, the reduced basis can be iteratively extended by searching for the parameter $\boldsymbol{\mu}_{\max} := \arg\max_{\boldsymbol{\mu} \in M_{\text{train}}} \eta(\boldsymbol{\mu})$ of the worst approximated trajectory, and adding the first and most significant mode gained from a proper orthogonal decomposition of this trajectory's projection errors $\left\{u_h^k(\boldsymbol{\mu}_{\max}) - \mathscr{P}_{\text{red}}\left[u_h^k(\boldsymbol{\mu}_{\max})\right]\right\}_{k=0}^K$ as a new basis function. This algorithm is repeated, until $\eta(\boldsymbol{\mu}_{\max})$ falls beneath a given tolerance.

Adaptation techniques: The basic algorithm described above depends on a fixed initial choice for the training subset M_{train}. In case of complex dependencies of the solution trajectories on the parameter, the reduced basis approximation can therefore turn out to be very bad for parameters not in the training set. In [6] this problem is addressed by adaptively refining the parameter space if indicated by bad approximations from a further validation training set.

Other variations of the POD-Greedy algorithm adaptively partition the parameter space and construct different reduced bases for each of these partitions [3,5] leading to faster reduced simulations at the cost of a more expensive offline phase.

3.2 Time-adaptive empirical operator interpolation

The construction of the collateral reduced basis space and corresponding interpolation points follows a similar idea like the "POD-greedy" algorithms. For the empirical interpolation of an operator $\mathscr{I}_M \circ \mathscr{L}_h$, the interpolation error $\left\|v_h - \sum_{j=1}^M \mathscr{I}_M\left[v_h\right]\right\|$ is minimized over all $v_h \in \mathbf{L} := \left\{\mathscr{L}_h(\boldsymbol{\mu})\left[u_h^k(\boldsymbol{\mu})\right] \mid \boldsymbol{\mu} \in \mathscr{P}, k = 0, \cdots, K-1\right\}$. Analogously to the reduced basis generation, we define a finite subset $L_{\text{train}} \subset \mathbf{L}$ and pick one of this set's snapshots as an initial collateral reduced basis function. The extension step for the empirical interpolation then looks as follows:

1. Find the approximation with the worst error $v_M \leftarrow \arg\sup_{v_h \in L_{\text{train}}} \|u_h - \mathscr{I}_M\left[v_h\right]\|$.
2. Compute the residual between v_M and its interpolant $r_M \leftarrow v_M - \mathscr{I}_M\left[v_h\right]$.
3. Find the interpolation point maximizing the residual $x_M \leftarrow \arg\sup_{x \in X_h} |r_M(x)|$.
4. Normalize to construct new reduced basis space function $q_M \leftarrow \frac{r_M}{r_M(x_M)}$.

These steps are repeated until the maximum interpolation error falls beneath a given tolerance. We call this algorithm EIDETAILED in the sequel.

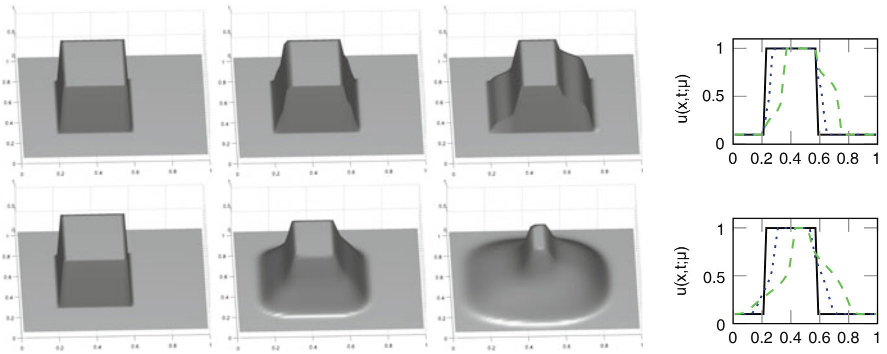


Fig. 1 Detailed simulation solution snapshots at time instants $t = 0.0$ (first column), $t = 0.1$ (second column), $t = 0.3$ (third column) and for different parameters $\boldsymbol{\mu} = (0, 0.1, 0.4)$ (first row) and $\boldsymbol{\mu} = (2, 0.1, 0.4)$ (second row). The last column shows the reduced solution on cross-sections at $y = 0.5$ for the time instants $t = 0$ (solid line), $t = 0.1$ (dotted line), $t = 0.3$ (dashed line)

Adaptation techniques: The adaptation techniques mentioned in Section 3.1 can also be applied to the empirical interpolation algorithm EIDETAILED, but so far no actual implementation for this is known to us. Supplementary to the adaptive search in the parameter space, we propose to build different collateral reduced basis spaces for different time instant sets $\mathscr{K} \subset \{0, \ldots, K-1\}$. As this time-adaptation strategy is the main focus of this article, we want to give a detailed description of the algorithm:

```
procedure TIMESLICEDEI($\mathscr{W}_{\text{init}}, \mathscr{K}, L_{\text{train}}^{\mathscr{K}}$)
    $\mathscr{W}_M \leftarrow$ EIDETAILED($\mathscr{W}_{\text{init}}, L_{\text{train}}^{\mathscr{K}}, M_{\max}, \varepsilon_{tol}$)
    if $\varepsilon_{tol}$ reached then
        $M^k \leftarrow M$ and $\mathscr{W}_{M^k}^k \leftarrow \mathscr{W}_M$ for all $k \in \mathscr{K}$.
    else if card($\mathscr{K}$) $\leq 2c_{\min}$ then
        $\mathscr{W}_{M^k}^k \leftarrow$ EIDETAILED($\mathscr{W}_M, L_{\text{train}}^{\mathscr{K}}, \infty, \varepsilon_{tol}$) for all $k \in \mathscr{K}$.
    else % maximum number of extensions $M_{\max}$ reached
        $\mathscr{K}_1, \mathscr{K}_2 \leftarrow$SPLITTIMEINTERVAL($\mathscr{K}, \mathscr{W}_M$)
        TIMESLICEDEI($\mathscr{W}_M^{\mathscr{K}_1}, L_{\text{train}}^{\mathscr{K}_1}$)
        TIMESLICEDEI($\mathscr{W}_M^{\mathscr{K}_2}, L_{\text{train}}^{\mathscr{K}_2}$)
    end if
end procedure
```

The training sets $L_{\text{train}}^{\mathscr{K}}$ are restrictions of the full training set L_{train} to operator evaluations on solutions snapshots at time steps t^k for $k \in \mathscr{K}$. Likewise $\mathscr{W}_M^{\mathscr{K}}$ is a restriction of the discrete space $\mathscr{W}_M$ build only out of solution snapshots with time indices stemming from $\mathscr{K}$. This strategy reduces the computation time, as no computed reduced basis function needs to be thrown away. The method SPLITTIMEINTERVAL splits the interval $\mathscr{K}$ such that afterwards the spaces $\mathscr{W}_M^{\mathscr{K}_1}$ and $\mathscr{W}_M^{\mathscr{K}_2}$ are of equal dimension. The threshold $c_{\min}$ asserts a lower bound on the size of the time intervals.

Table 1 Comparison of the number of bases, the reduced basis sizes averaged over sub-intervals, offline time, averaged online reduced simulation times and maximum errors for non-adaptive and adaptive runs with threshold $c_{\min} = 5$, and $= 1$. The average online run-times and maximum errors are obtained from 20 simulations with randomly selected parameters $\boldsymbol{\mu}$

adaptation	no. of bases	ø-dim(CRB)	offline time[h]	ø-runtime[s]	max. error
no	1	350	1.47	6.79	$5.88 \cdot 10^{-4}$
yes, $c_{\min} = 5$	11	223.09	2.08	4.06	$5.80 \cdot 10^{-4}$
yes, $c_{\min} = 1$	26	198.42	8.40	3.38	$5.75 \cdot 10^{-4}$

4 Example: Buckley–Leverett equation

We consider a Buckley–Leverett type problem in two space dimensions fulfilling the equations (1)-(2) on a rectangular domain $\Omega := [0,1]^2$ with initial data function $u_0(\boldsymbol{\mu}) = c_{low} + (1 - c_{low})\chi_{[0.2,0.6]\times[0.25,0.75]}$, velocity vector $\mathbf{v}(u;\boldsymbol{\mu}) = (0,1)^t f(u;\boldsymbol{\mu})$ and diffusion $d(u;\boldsymbol{\mu}) = KD(s;\boldsymbol{\mu})$. Here $f(u;\boldsymbol{\mu}) = \frac{u^3}{\mu_1} \cdot \left(\frac{u^3}{\mu_1} + \frac{(1-u)^3}{\mu_2}\right)^{-1}$ denotes the fractional flow rate, $D(u;\boldsymbol{\mu}) = \frac{(1-u)^3}{\mu_2} f(u;\boldsymbol{\mu}) p_c'(u;\boldsymbol{\mu})$ the capillary diffusion for a capillary pressure $p_c(u;\boldsymbol{\mu}) = u^{-\lambda}$. The variable parameters are chosen as $\boldsymbol{\mu} := (K, c_{\text{low}}, \lambda)$ and the parameter space is given by $\mathscr{P} := [1,2] \times [0,0.1] \times [0.1,0.4]$. The scalar viscosities are fixed at $\mu_1 = \mu_2 = 5$. At the boundary of the domain a Dirichlet condition applies with $u_{\mathcal{N}_{\text{dir}}}(\boldsymbol{\mu}) = c_{\text{low}}$.

Discretization The problem is discretized with a standard finite volume scheme comprising an explicitly computed Engquist–Osher flux for the convective terms and an implicit discretization of the diffusive terms. The underlying grid has a dimension of $H = 25 \times 25$ grid cells and the time interval $[0, T_{\max}]$ is discretized by 60 uniformly distributed time steps. Fig. 1 illustrates solution snapshots for two different parameters with different diffusion levels $K = 0$ respectively $K = 2$. The cross-section plots in the last column show the expected behaviour of combinations of rarefaction waves and smoothed shocks.

Offline phase In order to assess the effects of the adaptation algorithms, the reduced basis algorithms are run three times, once without the time adaptive empirical operator interpolation and two times with adaptation, but different thresholds $c_{\min}$ to bound the time interval size from below. The results concerning reduced basis sizes, offline and reduced simulation time, are summarized in Table 1.

In order to assure that the generated reduced basis leads to equally small reduction errors for all parameters of the parameter space, the parameter training set for the "POD-greedy" algorithm has been adapted with a validation set of randomly chosen parameters $\boldsymbol{\mu}$ in both runs. In the test runs, after three refinement steps the training parameter set comprises 255 elements, and the chosen validation ratio of 1.4 is assured after the maximum error for the training parameters has fallen beneath the targeted level of $5 \cdot 10^{-4}$. The target interpolation error for the empirical interpolation was set to 10^{-6} in all runs. This error is reached with an average number of 198

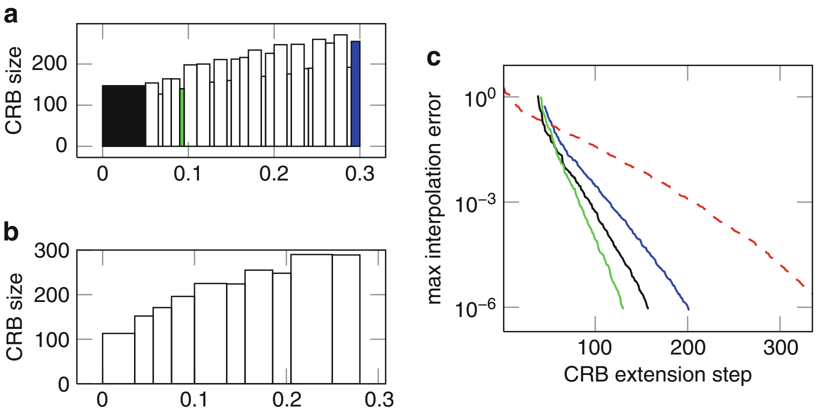


Fig. 2 Illustration of basis sizes on time intervals after adaptation with (a) $c_{\min} = 1$ and (b) $c_{\min} = 5$. Plot (c) illustrates the error decrease during generation of bases on three intervals marked with the same color in plot (a). The dashed line graph shows the slower decrease for a single basis without adaptation

respectively 223 basis functions in the adaptive cases, and 350 basis functions without adaptation. In the adaptive runs, the time interval has been decomposed into 11 respectively 26 sub-intervals (cf. Fig 2(a)&(b)). Fig. 2(c) illustrates the error decrease during the generation of the reduced spaces for selected time intervals (dashed lines) for the run with $c\text{min} = 1$. It can be observed that the slopes for the error graphs are much steeper than in the non-adaptive case illustrated with a dashed line. Because of the larger variation of the solutions for larger time steps, however, the basis on the last interval $[0.29, 0.30]$ still shows the slowest error decrease. Fig. 2(a+b) show that for both adaptive runs the bases dimensions for all intervals stay significantly below the non-adaptive basis size of 350.

Conclusion We observed that the adaptive search in the time domain can lead to faster reduced simulations. However, the costs of 26 generated basis spaces for an average dimension reduction by a factor of approximately 0.56 turned out to be very expensive. We therefore advice to combine the time domain search with a parameter domain search to obtain a further improvement of the method.

Acknowledgements This work was supported by the German Science Foundation (DFG) under the contract number OH 98/2-1. The second author was supported by the Baden Württemberg Stiftung gGmbH.

References

1. Barrault, M., Maday, Y., Nguyen, N., Patera, A.: An 'empirical interpolation' method: application to efficient reduced-basis discretization of partial differential equations. C. R. Math. Acad. Sci. Paris Series I **339**, 667–672 (2004)

2. Drohmann, M., Haasdonk, B., Ohlberger, M.: Reduced Basis Approximation for Nonlinear Parametrized Evolution Equations based on Empirical Operator Interpolation. Tech. rep., FB10, University of Münster (2010)
3. Eftang, J.L., Patera, A.T., Rønquist, E.M.: An hp Certified Reduced Basis Method for Parametrized Parabolic Partial Differential Equations. Technical report, MIT, Cambridge, 2009. Submitted to SISC
4. Haasdonk, B. and Ohlberger, M.: Reduced basis method for finite volume approximations of parametrized evolution equations. M2AN Math. Model. Numer. Anal., **4**2(2):277-302 (2008)
5. Haasdonk, B., Dihlmann, M., Ohlberger, M.: A training set and multiple bases generation approach for parametrized model reduction based on adaptive grids in parameter space. Tech. rep., University of Stuttgart (submitted) (2010)
6. Haasdonk, B., Ohlberger, M.: Adaptive basis enrichment for the reduced basis method applied to finite volume schemes. In: Proc. 5th International Symposium on Finite Volumes for Complex Applications, pp. 471–478 (2008)
7. Patera, A., Rozza, G.: Reduced Basis Approximation and a Posteriori Error Estimation for Parametrized Partial Differential Equations. MIT (2007). http://augustine.mit.edu/methodology/methodology_bookPartI.htm. Version 1.0, Copyright MIT 2006-2007, to appear in (tentative rubric) MIT Pappalardo Graduate Monographs in Mechanical Engineering

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Adaptive Time-Space Algorithms for the Simulation of Multi-scale Reaction Waves

Max Duarte, Marc Massot, Stéphane Descombes, and Thierry Dumont

Abstract We present a new resolution strategy for multi-scale reaction waves based on adaptive time operator splitting and space adaptive multiresolution, in the context of localized and stiff reaction fronts. The main goal is to perform computationally efficient simulations of the dynamics of multi-scale phenomena under study, considering large simulation domains with conventional computing resources. We aim at time-space accuracy control of the solution and splitting time steps purely dictated by the physics of the phenomenon and not by stability constraints associated with mesh size or source time scales. Numerical illustrations are provided for 2D and 3D combustion applications modeled by reaction-convection-diffusion equations.

Keywords time adaptive integration, space adaptive multiresolution, combustion
MSC2010: 65M08, 65M50, 65Z05, 65G20

1 Introduction

Numerical simulations of multi-scale phenomena are commonly used for modeling purposes in many applications such as combustion, chemical vapor deposition, or air pollution modeling. In general, all these models raise several difficulties created by the high number of unknowns, the wide range of temporal scales due to large and

M. Duarte and M. Massot
Laboratoire EM2C - UPR CNRS 288, Ecole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry Cedex, France, email: {max.duarte,marc.massot}@em2c.ecp.fr

S. Descombes
Laboratoire J. A. Dieudonné - UMR CNRS 6621, Université de Nice - Sophia Antipolis, Parc Valrose, 06108 Nice Cedex 02, France, e-mail: sdescomb@unice.fr

T. Dumont
Institut Camille Jordan - UMR CNRS 5208, Université de Lyon, 43 Boulevard du 11 novembre 1918, 69622 Villeurbanne Cedex, France, e-mail: tdumont@math.univ-lyon1.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_40, © Springer-Verlag Berlin Heidelberg 2011

detailed chemical kinetic mechanisms, as well as steep spatial gradients associated with localized fronts of high chemical activity. In this context, faced with the induced stiffness of these time dependent problems, a natural stumbling block to perform 3D simulations with all scales resolution is either the unreasonably small time step due to stability requirements or the unreasonable memory and computing time required by implicit methods. Furthermore, an accurate description of such spatial multi-scale phenomena would also lead to large and sometimes unfeasible computation domains, if no adaptive meshing technique is used.

To overcome these difficulties, we present a new numerical strategy with a time operator splitting that considers dedicated high order time integration methods for reaction, diffusion and convection problems, in order to build a time operator splitting scheme that exploits efficiently the special features of each problem. Based on recent theoretical studies of numerical analysis, such a strategy leads to a splitting time step which is not restricted neither by the fastest scales in the source term nor by restrictive stability limits of diffusive or convective steps, but only by the physics of the phenomenon. Moreover, this splitting time step is dynamically adapted taking into account local error estimates [4]. The time integration is performed over a dynamic adapted grid obtained by multiresolution techniques in a finite volumes framework [2, 9, 11], which on the one hand, yield important savings in computing resources and on the other hand, allow to somehow control the spatial accuracy of the compressed representation based on a solid mathematical background.

Even though, the strategy was developed for the resolution of general multi-scale phenomena in various domains as biomedical applications [7] or nonlinear chemical dynamics [6], we will focus here on multidimensional combustion problems at large Reynolds numbers in order to assess the capability of the method. The paper is organized as follows: section 2 describes briefly the numerical strategy, based on spatial adaptive multiresolution and second order adaptive time integration. Physical configuration and modeling equations are presented in section 3 for laminar premixed flames interacting with vortices, along with 2D and 3D numerical illustrations. We end in the last part with some concluding remarks.

2 Construction of the Numerical Strategy

We detail briefly the developed operator splitting strategy with splitting time step adaptation, and some fundamental aspects of the adaptive multiresolution method.

2.1 Adaptive Time Operator Splitting

Given a general convection-reaction-diffusion system of equations

$$\partial_t \mathbf{u} - \partial_{\mathbf{x}} \left(\mathbf{F}\left(\mathbf{u}\right) + \mathbf{D}(\mathbf{u}) \partial_{\mathbf{x}} \mathbf{u} \right) = \mathbf{f}\left(\mathbf{u}\right), \quad \mathbf{x} \in \mathbb{R}^d, \ t > 0, \tag{1}$$

with $\mathbf{u}(0,\mathbf{x}) = \mathbf{u}_0(\mathbf{x})$, where $\mathbf{F}, \mathbf{f} : \mathbb{R}^m \to \mathbb{R}^m$ and $\mathbf{u} : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^m$, with diffusion matrix $\mathbf{D}(\mathbf{u})$: a tensor of order $d \times d \times m$; an operator splitting procedure allows to consider dedicated solvers for the reaction part which is decoupled from the other physical phenomena like convection, diffusion or both, for which there also exist dedicated numerical methods. These dedicated methods chosen for each subsystem are then responsible for dealing with the fast scales associated with each one of them, in a separate manner, while the reconstruction of the global solution by the splitting scheme should guarantee an accurate description with error control of the global physical coupling, without being related to the stability constraints of the numerical resolution of each subsystem.

A second order Strang scheme is then implemented [12]

$$\mathscr{S}^{\Delta t}(\mathbf{u}_0) = \mathscr{R}^{\Delta t/2}\mathscr{D}^{\Delta t/2}\mathscr{C}^{\Delta t}\mathscr{D}^{\Delta t/2}\mathscr{R}^{\Delta t/2}(\mathbf{u}_0), \tag{2}$$

where operators $\mathscr{R}$, $\mathscr{D}$, $\mathscr{C}$ indicate respectively the independent resolution of the reaction, diffusion and convection problems with Δt defined as the splitting time step. Usually, for propagating reaction waves where for instance, the speed of propagation is much slower than some of the chemical scales, the fastest scales are not directly related to the global physics of the phenomenon, and thus, larger splitting time steps might be considered. Nevertheless, order reductions may then appear due to short-life transients associated to fast variables and in these cases, it has been proved in [5] that better performances are expected while ending the splitting scheme by operator $\mathscr{R}$ or in a more general case, the part involving the fastest time scales of the phenomenon.

An adaptive splitting time step strategy, based on a local error estimate at the end of each Δt, is implemented in order to control the accuracy of computations. A second, embedded and lower order Strang splitting method $\widetilde{\mathscr{S}}^{\Delta t}$ was developed [4] in order to dynamically calculate a local error estimate that should verify

$$\left\| \mathscr{S}^{\Delta t}(\mathbf{u}_0) - \widetilde{\mathscr{S}}^{\Delta t}(\mathbf{u}_0) \right\| \approx \mathscr{O}(\Delta t^2) < \eta_{\text{split}}, \tag{3}$$

in order to accept current computation with Δt, and thus, the new splitting time step is given by

$$\Delta t_{\text{new}} = \Delta t \sqrt{\frac{\eta_{\text{split}}}{\left\| \mathscr{S}^{\Delta t}(\mathbf{u}_0) - \widetilde{\mathscr{S}}^{\Delta t}(\mathbf{u}_0) \right\|}}. \tag{4}$$

The choice of suitable time integration methods to approximate numerically $\mathscr{R}$, $\mathscr{D}$ and $\mathscr{C}$ during each Δt is mandatory not only to guarantee the theoretical framework of the numerical analysis but also to take advantage of the particular features of each independent subproblem. A new operator splitting for reaction-diffusion systems was recently introduced [6], which considers a high fifth order, A-stable, L-stable method like Radau5 [8], based on implicit Runge-Kutta schemes for stiff ODEs, that solves with a local cell by cell approach the reaction term: a system of stiff ODEs without spatial coupling. On the other hand, a high fourth order method was chosen, like ROCK4 [1], based on explicit stabilized Runge-Kutta schemes which features extended stability domains along the negative real axis, very appropriate for diffusion problems because of the usual predominance of negative real eigenvalues. Both methods incorporate adaptive time integration tools, similar to (4), in order to control accuracy for given η_{Radau5} and η_{ROCK4}.

An explicit high order in time and space one step monotonicity preserving scheme OSMP [3] is used as convective scheme. It combines monotonicity preserving constraints for non-monotone data to avoid extrema clipping, with TVD features to prevent spurious oscillations around discontinuities or sharp spatial gradients. Classical CFL stability restrictions are though imposed during each splitting time step Δt. Notice that the overall combination of explicit treatment of spatial phenomena as convection and diffusion, with local implicit integration of stiff reaction implies important savings in computing time and memory resources. For the reaction, local treatment plus adaptive time stepping allow to discriminate cells of high reactive activity in the neighborhood of the localized wavefront, saving as a consequence a large quantity of integration time.

2.2 Mesh Refinement Technique

We are concerned with the propagation of reacting wavefronts, hence important reactive activity as well as steep spatial gradients are localized phenomena. This implies that if we consider the resolution of reactive problem, a considerable amount of computing time is spent on nodes that are practically at (partial) equilibrium. Moreover, there is no need to represent these quasi-stationary regions with the same spatial discretization needed to describe the reaction front, so that convection and diffusion problems might also be solved over a smaller number of nodes. An adapted mesh obtained by a multiresolution process which discriminates the various space scales of the phenomenon, turns out to be a very convenient solution to overcome these difficulties [6, 7].

In practice, if one considers a set of nested spatial grids from the coarsest to the finest one, a multiresolution transformation allows to represent a discretized function as values on the coarsest grid plus a series of local estimates at all other levels of such nested grids. These estimates correspond to the wavelet coefficients of a wavelet decomposition obtained by inter-level transformations, and retain the information on local regularity when going from a coarse to a finer grid. Hence, the main idea is to use the decay of the wavelet coefficients to obtain information on local regularity of the solution: lower wavelet coefficients are associated to local regular spatial configurations and vice-versa. This representation yields to a thresholding process that builds dynamically the corresponding adapted grid on which the solutions are represented; then the error committed by the multiresolution transformation is proportional to η_{MR}, where η_{MR} is a threshold parameter [2, 9].

3 Numerical Illustration

In these illustrating examples, we are concerned with the numerical simulation of premixed flames interacting with vortex structures: a pair of counter rotating vortices in a 2D configuration and a 3D toroidal vortex. This is usually a difficult problem to solve because of the localized and stiff reactive fronts, even more with large Reynolds numbers. Nevertheless, in order to properly evaluate the proposed strategy we consider only time evolution problems for which the hydrodynamics is not solved but a large Reynolds number velocity field is imposed. Based on a model presented in [10], we consider that the chemistry may be modeled by a global, single step, irreversible reaction characterized by an Arrhenius law; and a thermodiffusive approach of laminar flame theory is adopted in order to decouple velocity field computation from determination of species mass fractions and temperature. Known solutions of incompressible Navier-Stokes equations may then be imposed, and the problem is reduced to solving the standard species and energy balance equations.

Following [10], a progress variable $c(x, y, t)$ is introduced:

$$c = \frac{T - T_{\mathrm{o}}}{T_{\mathrm{b}} - T_{\mathrm{o}}}, \tag{5}$$

where subscripts $(\ )_{\mathrm{o}}$ and $(\ )_{\mathrm{b}}$ indicate respectively, fresh mixture zone and burnt product zone; and we finally obtain for a 2D configuration

$$\frac{\partial c}{\partial t_\star} + u_\star \frac{\partial c}{\partial x_\star} + v_\star \frac{\partial c}{\partial y_\star} - \left(\frac{\partial^2 c}{\partial x_\star^2} + \frac{\partial^2 c}{\partial y_\star^2} \right) = \mathrm{Da}(1-c)\exp\left(-\frac{T_{\mathrm{a}}}{T_{\mathrm{o}}(1+\tau c)} \right), \tag{6}$$

where Da is a Damköhler number, T_{a} the activation energy, $\tau = T_{\mathrm{b}}/T_{\mathrm{o}} - 1$, and $(\ )_\star$ indicates dimensionless variables. The velocity field $(u_\star(t), v_\star(t))$ is deduced analytically and imposed into (6), considering a 2D viscous core vortex with a dimensionless azimuthal velocity of the form:

$$v_{\theta\star}(r_\star, t_\star) = \frac{\mathrm{Re}\,\mathrm{Sc}}{r_\star}\left(1 - \exp\left(-\frac{r_\star^2}{4\,\mathrm{Sc}\, t_\star}\right)\right), \tag{7}$$

with $r_\star(x_\star, y_\star)$, the distance to the vortex center, Reynolds and Schmidt numbers.

Figure 1 shows the interaction of the premixed flame with two counter rotating vortices modeled each one of them by (7), centered at $(-0.25, -0.5)$ and $(0.25, -0.5)$ for a 2D spatial domain of $[-1, 1]^2$. The upper (red) and lower (blue) regions correspond respectively to burnt product ($c = 1$) and fresh mixture ($c = 0$) zones. The corresponding adapted mesh tightens around the stiff regions and propagates along the wavefronts.

The following modeling values were considered into (6) and (7): $\mathrm{Da} = 2.5 \times 10^9$, $T_{\mathrm{a}} = 20000$ K, $T_{\mathrm{o}} = 300$ K, $T_{\mathrm{b}} = 2315.4$ K, $\tau \approx 6.72$, $\mathrm{Sc} = 1$ and $\mathrm{Re} = 1000$. The initial condition corresponds to a planar premixed flame at $y = -0.5$ and the

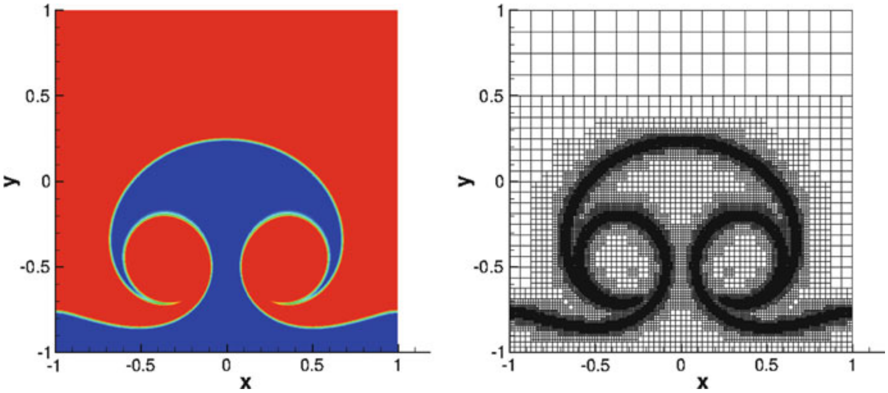


Fig. 1 2D premixed flame interacting with two counter rotating vortices. Solution of variable c at $t_\star = 4 \times 10^{-4}$ (left) and corresponding adapted mesh (right). Finest grid: 1024^2

phenomenon is studied over a time domain of $[0, 4 \times 10^{-3}]$. The MR procedure considers a set of 10 grids, equivalent to $1024^2 = 1048576$ cells on the finest grid. MR and adaptive splitting time step tolerances were set to $\eta_{\mathrm{MR}} = 10^{-2}$ and $\eta_{\mathrm{split}} = 10^{-3}$, with $\eta_{\mathrm{Radau5}} = \eta_{\mathrm{ROCK4}} = 10^{-5}$.

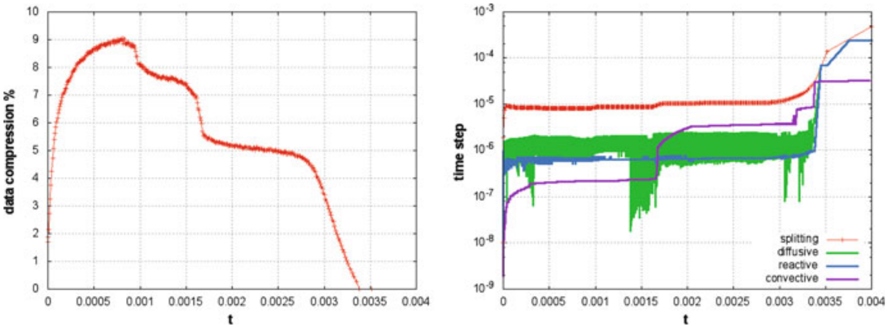


Fig. 2 2D premixed flame interacting with two counter rotating vortices. Time evolution of data compression in the solution representation (left) and splitting, diffusive, reactive and convective time steps (right). Finest grid: 1024^2

Figure 2 shows data compression obtained by MR representation of the solution, measured as the percentage of active cells with respect to the finest grid representation; in this case, lower than 9% of 1024^2. On the other hand, splitting time step starts from an initial value set to 10^{-8} in order to handle correctly the initial sudden apparition of the vortices, that evolves rapidly to a final quasi stable value of 10^{-5}, which indicates the decoupling degree achieved within the accuracy prescribed to describe the global propagating phenomenon. The corresponding convective time step with CFL $= 1$ illustrates the time scale decoupling obtained by a splitting technique and highlights the eventual inconveniences of solving (6) considering all phenomena at once. The same conclusion is valid concerning reactive and diffusive time steps. By the way, larger convective time steps are used thanks to the adapted grid representation which allows to discriminate locally large velocity values (in this case $|u_\star|, |v_\star| \approx 40000$) from the refined regions around the wavefront, as we can see in the "jumps" of convective time steps in Fig. 2. Reactive time steps correspond to cells at the wavefront (for furthest cells, reactive time steps are equal to splitting ones), while lower diffusive time steps are needed in order to fulfill each splitting time step, which explains the "oscillations". Diffusive time steps might take values beyond classical stability constraints (of the order of 10^{-6} for explicit RK4 [8] and eigenvalues of -2.2×10^6), and it is finally set by the accuracy criterion.

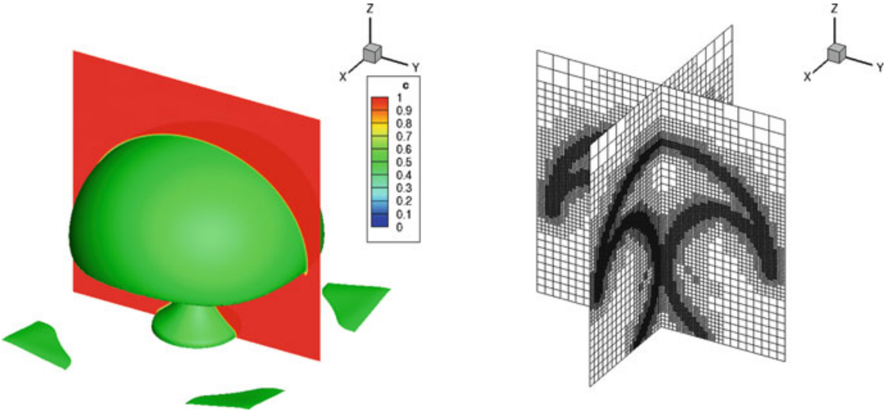


Fig. 3 3D premixed flame interacting with a toroidal vortex. Solution of variable c at $t_\star = 1.1 \times 10^{-3}$ showing isosurface $c = 0.5$ (left) and corresponding adapted mesh (right). Finest grid: 256^3

This resolution technique has a straightforward extension to 3D configurations. Figure 3 shows the interaction of the premixed flame with a toroidal vortex modeled by (7) centered at $\sqrt{x_\star^2 + y_\star^2} = 0.25$, $z_\star = -0.5$ for a 3D spatial domain of $[-1, 1]^3$. The modeling and tolerance parameters are taken equal to the 2D case and the MR procedure considers a set of 8 grids, equivalent to $256^3 = 16777216$ cells on the finest grid. The splitting time step shows the same behavior as for the previous case with same order of values, while the data compression is lower than 17%, taking into account that a lower scale discrimination is available with 8 different grids. All the computations have been performed on a AMD Shanghai processor of 2.7 GHz with memory capacity of 4 GB. Computing times for the 2D and 3D configurations were about of 0h57m and 14h40m, respectively.

4 Concluding Remarks

The present work proposes a new numerical approach which is shown to be computationally efficient. It couples adaptive multiresolution techniques with a new operator splitting strategy with high order time integration methods to properly solve the entire spectrum of scales of each phenomenon. The splitting time step is chosen on the sole basis of the structure of the continuous system and its decoupling capabilities, but not related to stability requirements of the numerical methods involved in order to integrate each subsystem, even if stiffness is present. The global accuracy of the simulation is controlled and dynamically evaluated based on theoretical and numerical results. As a consequence, the resulting highly compressed data representations as well as the accurate and feasible resolution of these stiff phenomena prove that large computational domains previously out of reach can be successfully simulated with conventional computing resources. At this stage of development, the same numerical strategy can be coupled to a hydrodynamics solver, considering though that an important amount of work is still in progress concerning programming features such as data structures and parallelization strategies.

Acknowledgements This research was supported by fundamental project grants from ANR *Séchelles* (project leader S. Descombes), and DIGITEO *MUSE* (project leader M. Massot). M. Duarte has a Ph.D. grant from Mathematics (INSMI) and Engineering (INSIS) Institutes of CNRS and supported by INCA project. Authors express special thanks to Christian Tenaud (LIMSI-CNRS) for providing the basis of the multiresolution kernel of MR CHORUS (DI 03760-01).

References

1. Abdulle, A.: Fourth order Chebyshev methods with recurrence relation. SIAM J. Sci. Comput. **23**, 2041–2054 (2002)
2. Cohen, A., Kaber, S., Müller, S., Postel, M.: Fully adaptive multiresolution finite volume schemes for conservation laws. Math. of Comp. **72**, 183–225 (2003)
3. Daru, V., Tenaud, C.: High order one-step monotonicity-preserving schemes for unsteady compressible flow calculations. Journal of Computational Physics **193**(2), 563–594 (2004)
4. Descombes, S., Duarte, M., Dumont, T., Louvet, V., Massot, M.: Adaptive time splitting method for multi-scale evolutionary PDEs. Confluentes Mathematici (to app.) (2011)
5. Descombes, S., Massot, M.: Operator splitting for nonlinear reaction-diffusion systems with an entropic structure: Singular perturbation and order reduction. Numer. Math. **97**(4), 667–698 (2004)
6. Duarte, M., Massot, M., Descombes, S., Tenaud, C., Dumont, T., Louvet, V., Laurent, F.: New resolution strategy for multi-scale reaction waves using time operator splitting, space adaptive multiresolution and dedicated high order implicit/explicit time integrators. Submitted to SIAM J. Sci. Comput., available on HAL (http://hal.archives-ouvertes.fr/hal-00457731) (2010)
7. Duarte, M., Massot, M., Descombes, S., Tenaud, C., Dumont, T., Louvet, V., Laurent, F.: New resolution strategy for multi-scale reaction waves using time operator splitting and space adaptive multiresolution: Application to human ischemic stroke. ESAIM Proc. (to app.) (2011)

8. Hairer, E., Wanner, G.: Solving ordinary differential equations II, second edn. Springer-Verlag, Berlin (1996). Stiff and differential-algebraic problems
9. Harten, A.: Multiresolution algorithms for the numerical solution of hyperbolic conservation laws. Comm. Pure and Applied Math. **48**, 1305–1342 (1995)
10. Laverdant, A., Candel, S.: Computation of diffusion and premixed flames rolled up in vortex structures. Journal of Propulsion and Power **5**, 134–143 (1989)
11. Müller, S.: Adaptive multiscale schemes for conservation laws, vol. 27. Springer-Verlag, Heidelberg (2003)
12. Strang, G.: On the construction and comparison of difference schemes. SIAM J. Numer. Anal. **5**, 506–517 (1968)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Dispersive wave runup on non-uniform shores

Denys Dutykh, Theodoros Katsaounis, and Dimitrios Mitsotakis

Abstract Historically the finite volume methods have been developed for the numerical integration of conservation laws. In this study we present some recent results on the application of such schemes to dispersive PDEs. Namely, we solve numerically a representative of Boussinesq type equations in view of important applications to the coastal hydrodynamics. Numerical results of the runup of a moderate wave onto a non-uniform beach are presented along with great lines of the employed numerical method (see D. Dutykh *et al.* (2011) [6] for more details).

Keywords dispersive wave, runup, Boussinesq equations, shallow water
MSC2010: 65M08, 76B15

1 Introduction

The simulation of water waves in realistic and complex environments is a very challenging problem. Most of the applications arise from the areas of coastal and naval engineering, but also from natural hazards assessment. These applications may require the computation of the wave generation [5, 12], propagation [17], interaction with solid bodies, the computation of long wave runup [16, 18] and even the extraction of the wave energy [15]. Issues like wave breaking, robustness

Denys Dutykh
LAMA, UMR 5127 CNRS, Université de Savoie, Campus Scientifique, 73376 Le Bourget-du-Lac Cedex, France, e-mail: Denys.Dutykh@univ-savoie.fr

Theodoros Katsaounis
Department of Applied Mathematics, University of Crete, Heraklion, 71409 Greece Inst. of App. and Comp. Math. (IACM), FORTH, Heraklion, 71110, Greece, e-mail: thodoros@tem.uoc.gr

Dimitrios Mitsotakis
IMA, University of Minnesota, Minneapolis MN 55455, USA, e-mail: dmitsot@gmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_41, © Springer-Verlag Berlin Heidelberg 2011

of the numerical algorithm in wet-dry processes along with the validity of the mathematical models in the near-shore zone are some basic problems in this direction [11]. During past several decades the classical Nonlinear Shallow Water Equations (NSWE) have been essentially employed to face these problems [7]. Mathematically, these equations represent a system of conservation laws describing the propagation of infinitely long waves with a hydrostatic pressure assumption. The wave breaking phenomenon is commonly assimilated to the formation of shock waves (or hydraulic jumps) which is a common feature of hyperbolic PDEs. Consequently, the finite volume (FV) method has become the method of choice for these problems due to its excellent intrinsic conservative and shock-capturing properties [3, 7].

In the present article we report on recent results concerning the extension of the finite volume method to dispersive wave equations steming essentially from water wave modeling [4, 6, 14].

2 Mathematical model and numerical methods

Consider a cartesian coordinate system in two space dimensions (x, z) to simplify notations. The z-axis is taken vertically upwards and the x-axis is horizontal and coincides traditionally with the still water level. The fluid domain is bounded below by the bottom $z = -h(x)$ and above by the free surface $z = \eta(x,t)$. Below we will also need the total water depth $H(x,t) := h(x) + \eta(x,t)$. The flow is supposed to be incompressible and the fluid is inviscid. An additional assumption of the flow irrotationality is made as well.

In the pioneering work of D.H. Peregrine (1967) [14] the following system of Boussinesq type equations has been derived:

$$\eta_t + \big((h+\eta)u\big)_x = 0, \tag{1}$$

$$u_t + uu_x + g\eta_x - \frac{h}{2}(hu)_{xxt} + \frac{h^2}{6}u_{xxt} = 0, \tag{2}$$

where $u(x,t)$ is the depth averaged fluid velocity, g is the gravity acceleration and underscripts (u_x, η_t) denote partial derivatives.

In our recent study [6] we proposed an improved version of this system which contains higher order nonlinear terms which should be neglected from asymptotic point of view and can be written in conservative variables $(H, Q) = (H, Hu)$ as:

$$H_t + Q_x = 0, \tag{3}$$

$$\Big(\big(1+\frac{1}{3}H_x^2-\frac{1}{6}HH_{xx}\big)Q_t-\frac{1}{3}H^2Q_{xxt}-\frac{1}{3}HH_xQ_{xt}\Big)+\Big(\frac{Q^2}{H}+\frac{g}{2}H^2\Big)_x = gHh_x. \tag{4}$$

Obviously the linear characteristics of both systems (1), (2) and (3), (4) coincide since they differ only by nonlinear terms.

However, this modification has several important implications onto structural properties of the obtained system. First of all, the magnitude of the dispersive terms tends to zero when we approach the shoreline $H \to 0$. This property corresponds to our physical representation of the wave shoaling and runup process. On the other hand, the resulting system becomes invariant under vertical translations (subgroup G_5 in Theorem 4.2, T. Benjamin & P. Olver (1982) [2]):

$$z \leftarrow z + d, \quad \eta \leftarrow \eta - d, \quad h \leftarrow h + d, \quad u \leftarrow u, \tag{5}$$

where d is some constant. This property is straightforward to check since we use only the total water depth variable $H = h + \eta$ which remains invariant under transformation (5).

Remark 1. In this paper we will consider the initial-boundary value problem posed in a bounded domain $I = [b_1, b_2]$ with reflective boundary conditions. In this case one needs to impose boundary conditions only in one of the two dependent variables, cf. [8]. In the case of reflective boundary conditions it is sufficient to take $u(b_1, t) = u(b_2, t) = 0$.

2.1 Finite volume discretization

Let $\mathscr{T} = \{x_i\},\ i \in \mathbb{Z}$ denotes a partition of $\mathbb{R}$ into cells $C_i = (x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}})$ where $x_i = (x_{i+\frac{1}{2}} + x_{i-\frac{1}{2}})/2$ denotes the midpoint of C_i. Let $\Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ be the length of the cell C_i, $\Delta x_{i+\frac{1}{2}} = x_{i+1} - x_i$. (Here, we consider only uniform grids with $\Delta x_i = \Delta x_{i+\frac{1}{2}} = \Delta x$.)

The governing equations (3), (4) can be recast in the following vector form:

$$[\mathbf{D}(\mathbf{v_t})] + [\mathbf{F}(\mathbf{v})]_x = \mathbf{S}(\mathbf{v}),$$

where

$$\mathbf{D}(\mathbf{v_t}) = \begin{pmatrix} H_t \\ (1 + \frac{1}{3}H_x^2 - \frac{1}{6}HH_{xx})Q_t - \frac{1}{3}H^2Q_{xxt} - \frac{1}{3}HH_xQ_{xt} \end{pmatrix}, \tag{6}$$

$$\mathbf{F}(\mathbf{v}) = \begin{pmatrix} Q \\ \frac{Q^2}{H} + \frac{g}{2}H^2 \end{pmatrix}, \qquad \mathbf{S}(\mathbf{v}) = \begin{pmatrix} 0 \\ gHh_x \end{pmatrix}. \tag{7}$$

We denote by H_i and U_i the corresponding cell averages. To discretize the dispersive terms in (6) we consider the following approximations:

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\left[1+\frac{1}{3}(H_x)^2-\frac{1}{6}HH_{xx}\right]\,Q\,dx \approx$$

$$\left(1+\frac{1}{3}\left(\frac{H_{i+1}-H_{i-1}}{2\Delta x}\right)^2-\frac{1}{6}H_i\,\frac{H_{i+1}-2H_i+H_{i-1}}{\Delta x^2}\right)Q_i,$$

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\frac{1}{3}HH_xQ_x\,dx \approx \frac{1}{3}H_i\,\frac{H_{i+1}-H_{i-1}}{2\Delta x}\frac{Q_{i+1}-Q_{i-1}}{2\Delta x},$$

$$\frac{1}{\Delta x}\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}}\frac{1}{3}H^2Q_{xx}\,dx \approx \frac{1}{3}H_i^2\,\frac{Q_{i+1}-2Q_i+Q_{i-1}}{\Delta x^2}.$$

We note that we approximate the reflective boundary conditions by taking the cell averages of u on the first and the last cell to be $u_0 = u_{N+1} = 0$. We do not impose explicitly boundary conditions on H. The reconstructed values on the first and the last cell are computed using neighboring ghost cells and taking odd and even extrapolation for u and H respectively. These specific boundary conditions appeared to reflect incident waves on the boundaries while conserving the mass.

This discretization leads to a linear system with tridiagonal matrix denoted by $\mathbf{L}$ that can be inverted efficiently by a variation of Gauss elimination for tridiagonal systems with computational complexity $O(n)$, n-being the dimension of the system. We note that on the dry cells the matrix becomes diagonal since H_i is zero on dry cells. For the time integration the explicit third-order TVD-RK method is used. In the numerical experiments we observed that the fully discrete scheme is stable and preserves the positivity of H during the runup under a mild restriction on the time step Δt.

Therefore, the semidiscrete problem of (6) - (7) is written as a system of ODEs in the form:

$$\mathbf{L}_i\mathbf{v}_{i\,t}+\frac{1}{\Delta x}(\mathscr{F}_{i+\frac{1}{2}}-\mathscr{F}_{i-\frac{1}{2}})=\frac{1}{\Delta x}\mathbf{S_i},$$

where $\mathbf{L}_i$ is the $i-$th row of matrix $\mathbf{L}$ and $\mathscr{F}_{i+\frac{1}{2}}$ can be chosen as one of the numerical flux functions [6] (in computations presented below we choose the FVCF flux [9]). In the sequel we will use the KT and the CF numerical fluxes. In this case the Jacobian of $\mathbf{F}$ is given by the matrix

$$A=\begin{pmatrix}0 & 1\\ gH-(Q/H)^2 & 2Q/H\end{pmatrix},$$

and the eigenvalues are $\lambda_{1,2} = Q/H \pm \sqrt{gH}$. Therefore, the characteristic numerical flux [9] takes the form

$$\mathscr{F}_{i+\frac{1}{2}}=\frac{\mathbf{F}(\mathbf{V}^L_{i+\frac{1}{2}})+\mathbf{F}(\mathbf{V}^R_{i+\frac{1}{2}})}{2}-\mathbf{U}(\boldsymbol{\mu})\frac{\mathbf{F}(\mathbf{V}^R_{i+\frac{1}{2}})-\mathbf{F}(\mathbf{V}^L_{i+\frac{1}{2}})}{2},$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ are the Roe average values,

$$\mu_1 = \frac{H^L_{i+\frac{1}{2}} + H^R_{i+\frac{1}{2}}}{2}, \quad \mu_2 = \frac{\sqrt{H^L_{i+\frac{1}{2}}}U^L_{i+\frac{1}{2}} + \sqrt{H^R_{i+\frac{1}{2}}}U^R_{i+\frac{1}{2}}}{\sqrt{H^L_{i+\frac{1}{2}}} + \sqrt{H^R_{i+\frac{1}{2}}}}$$

and

$$\mathbf{U}(\boldsymbol{\mu}) = \begin{pmatrix} \frac{s_2(\mu_2+c)-s_1(\mu_2-c)}{2c} & \frac{s_1-s_2}{2c} \\ \frac{(s_2-s_1)(\mu_2^2-c^2)}{2c} & \frac{s_1(\mu_2+c)-s_2(\mu_2-c)}{2c} \end{pmatrix}, \; c = \sqrt{g\mu_1}, \; s_i = \operatorname{sign}(\lambda_i).$$

For more details on the discretization and reconstruction procedures, (that are based on the hydrostatic reconstraction, [1]), we refer to our complete work on this subject [6].

3 Numerical results

In the present section we show a numerical simulation of a solitary wave runup onto a non-uniform sloping beach. More precisely, we add a small pond along the slope. As our results indicate, this small complication is already sufficient to develop some instabilities which remain controlled in our simulations.

As an initial condition we used an approximate solitary wave solution of the following form:

$$\eta_0(x) = A_s \operatorname{sech}^2\big(\lambda(x - x_0)\big), \quad u_0(x) = -c_s \frac{\eta_0(x)}{1 + \eta_0(x)},$$

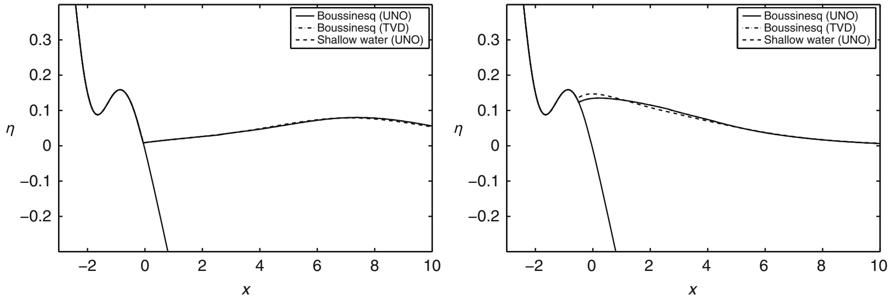
where A_s is the amplitude relative to the constant water depth taken to be unity in our study. The solitary wave speed c_s along with the wavelength λ are given here:

$$\lambda = \sqrt{\frac{3A_s}{4(1 + A_s)}}, \quad c_s = \sqrt{g}\frac{\sqrt{6}(1 + A_s)}{\sqrt{3 + 2A_s}} \cdot \frac{\sqrt{(1 + A_s)\log(1 + A_s) - A_s}}{A_s}.$$

The solitary wave is centered initially at $x_0 = 10.62$ and has amplitude $A_s = 0.08$. The constant slope β is equal to 2.88°. The sketch of the computational domain can be found in [6].

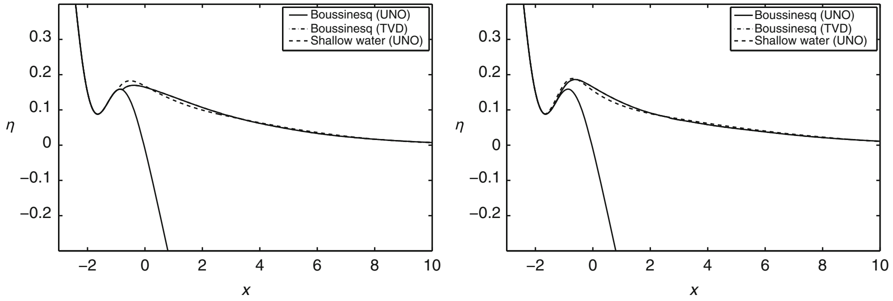
In numerical simulations presented below we used a uniform space discretization with $\Delta x = 0.025$ and very fine time step $\Delta t = \Delta x/100$ to guarantee the accuracy and stability during the whole simulation.

Snapshots of numerical results are presented on Figs. 1 – 6. We present simultaneously three different computational results:



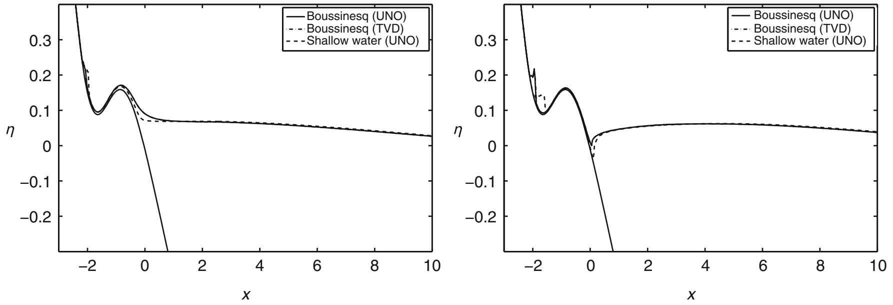
(a) t = 1 s (b) t = 3 s

Fig. 1 Solitary wave aproaching a sloping beach with a pond



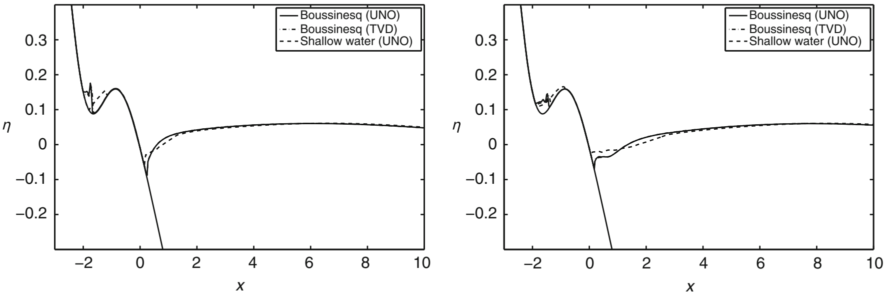
(a) t = 3.5 s (b) t = 4 s

Fig. 2 Beginning of the pond inundation



(a) t = 5 s (b) t = 5.5 s

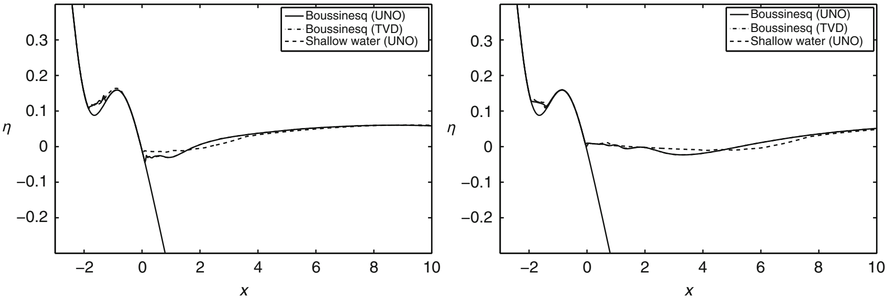
Fig. 3 A part of the wave mass is trapped in the pond volume



(a) t = 6 s

(b) t = 6.5 s

Fig. 4 Wave oscillations in the pond



(a) t = 7 s

(b) t = 8 s

Fig. 5 Stabilization of wave oscillations

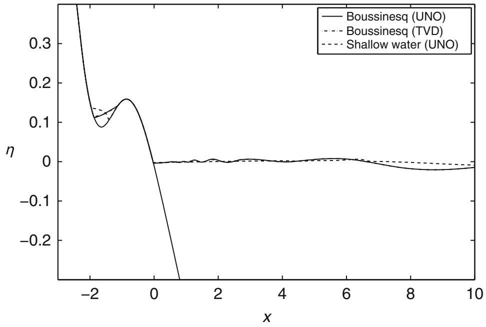


Fig. 6 The whole system is tending to the rest position ($t = 10$ s)

- Modified Peregrine system solved with UNO2 reconstruction [10]
- The same system with classical MUSCL TVD2 scheme [13]
- Nonlinear Shallow Water Equations (NSWE) with UNO2 scheme [10]

Surprisingly good agreement was obtained among all three numerical models. Presumably, the complex runup process under consideration is governed essentially by nonlinearity. However, on Figs. 1(b) and 2(a) the amplitude predicted by NSWE is slightly overestimated.

On Figs. 3(b) – 4(b) some oscillations (due to the small-dispersion effect characterizing dispersive wave breaking procedures) can be observed. However, their amplitude remains small for all times and does not produce any blow up phenomena. Later these oscillations decay tending gradually to the "lake at the rest" state (see Figs. 5, 6).

In the specific experiment a friction term could be beneficial to reduce the amplitude of oscillations (or damp them out completely). However, we prefer to present the computational results of our model without adding any ad-hoc term to show its original performance.

4 Conclusions

In this study we presented an improved version of the Peregrine system which is particularly suited for the simulation of dispersive waves runup. This system allows for the description of higher amplitude waves due to improved nonlinear characteristics. Better numerical stability properties have been obtained since most of the dispersive terms tend to zero when we approach the shoreline. Consequently, our model naturally degenerates to classical Nonlinear Shallow Water Equations (NSWE) for which the runup simulation technology is completely mastered nowadays. However we underline that there is no artificial parameter to turn off dispersive terms. Their importance is naturally governed by the underlying physical process.

Acknowledgements D. Dutykh acknowledges the support from French Agence Nationale de la Recherche, project MathOcean (Grant ANR-08-BLAN-0301-01) and Ulysses Program of the French Ministry of Foreign Affairs under the project 23725ZA. The work of Th. Katsaounis was partially supported by European Union FP7 program Capacities(Regpot 2009-1), through ACMAC (http://acmac.tem.uoc.gr).

References

1. Audusse, E., Bouchut, F., Bristeau, O., Klein, R., Perthame, B.: A fast and stable well-balanced scheme with hydrostatic reconstruction for shallow water flows. SIAM J. of Sc. Comp. **25**, 2050–2065 (2004)
2. Benjamin, T., Olver, P.: Hamiltonian structure, symmetries and conservation laws for water waves. J. Fluid Mech **125**, 137–185 (1982)

3. Delis, A.I., Katsaounis, T.: Relaxation schemes for the shallow water equations. Int. J. Numer. Meth. Fluids **41**, 695–719 (2003)
4. Dutykh, D., Dias, F.: Dissipative Boussinesq equations. C. R. Mecanique **335**, 559–583 (2007)
5. Dutykh, D., Dias, F.: Water waves generated by a moving bottom. In: A. Kundu (ed.) Tsunami and Nonlinear waves. Springer Verlag (Geo Sc.) (2007)
6. Dutykh, D., Katsaounis, T., Mitsotakis, D.: Finite volume schemes for dispersive wave propagation and runup. Accepted to Journal of Computational Physics **http://hal.archives-ouvertes.fr/hal-00472431/** (2011)
7. Dutykh, D., Poncet, R., Dias, F.: Complete numerical modelling of tsunami waves: generation, propagation and inundation. Submitted **http://arxiv.org/abs/1002.4553** (2010)
8. Fokas, A.S., Pelloni, B.: Boundary value problems for Boussinesq type systems. Math. Phys. Anal. Geom. **8**, 59–96 (2005)
9. Ghidaglia, J.M., Kumbaro, A., Coq, G.L.: On the numerical solution to two fluid models via cell centered finite volume method. Eur. J. Mech. B/Fluids **20**, 841–867 (2001)
10. Harten, A., Osher, S.: Uniformly high-order accurate nonscillatory schemes, I. SIAM J. Numer. Anal. **24**, 279–309 (1987)
11. Hibberd, S., Peregrine, D.: Surf and run-up on a beach: a uniform bore. J. Fluid Mech. **95**, 323–345 (1979)
12. Kervella, Y., Dutykh, D., Dias, F.: Comparison between three-dimensional linear and nonlinear tsunami generation models. Theor. Comput. Fluid Dyn. **21**, 245–269 (2007)
13. van Leer, B.: Towards the ultimate conservative difference scheme V: a second order sequel to Godunov' method. J. Comput. Phys. **32**, 101–136 (1979)
14. Peregrine, D.H.: Long waves on a beach. J. Fluid Mech. **27**, 815–827 (1967)
15. Simon, M.: Wave-energy extraction by a submerged cylindrical resonant duct. Journal of Fluid Mechanics **104**, 159–187 (1981)
16. Tadepalli, S., Synolakis, C.E.: The run-up of N-waves on sloping beaches. Proc. R. Soc. Lond. A **445**, 99–112 (1994)
17. Titov, V., González, F.: Implementation and testing of the method of splitting tsunami (MOST) model. Tech. Rep. ERL PMEL-112, Pacific Marine Environmental Laboratory, NOAA (1997)
18. Titov, V.V., Synolakis, C.E.: Numerical modeling of tidal wave runup. J. Waterway, Port, Coastal, and Ocean Engineering **124**, 157–171 (1998)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

MAC Schemes on Triangular Meshes

Robert Eymard, Jürgen Fuhrmann, and Alexander Linke

Abstract We present numerical results for two generalized MAC schemes on triangular meshes, which are based on staggered meshes using the Delaunay–Voronoi duality. In the first one, the pressures are defined at the vertices of the mesh, and the discrete velocities are tangential to the edges of the triangles. In the second one, the pressures are defined in the triangles, and the discrete velocities are normal to the edges of the triangles. In both cases, convergence results are obtained.

Keywords Navier–Stokes, MAC scheme, Delaunay mesh
MSC2010: 76D07, 65N08

1 Introduction

We consider in this paper two different generalizations of the classical MAC scheme [1] for the incompressible Stokes problem

$$\begin{aligned} -\Delta \boldsymbol{u} + \nabla p &= \boldsymbol{f} && \boldsymbol{x} \in \Omega, \\ \nabla \cdot \boldsymbol{u} &= 0 && \boldsymbol{x} \in \Omega, \\ \int_\Omega p \,\mathrm{d}\boldsymbol{x} &= 0 && \\ \boldsymbol{u} &= 0 && \boldsymbol{x} \in \partial\Omega. \end{aligned} \tag{1}$$

R. Eymard
Université Paris–Est, Paris, France, e-mail: robert.eymard@univ-mlv.fr

J. Fuhrmann and A. Linke
Weierstrass Institute, Berlin, Germany, e-mail: juergen.fuhrmann@wias-berlin.de, alexander.linke@wias-berlin.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_42, © Springer-Verlag Berlin Heidelberg 2011

We assume that $\boldsymbol{f} \in L^2(\Omega)^2$ holds, where Ω is an open polygonal bounded and connected subset of $\mathbb{R}^2$ without holes and with boundary $\partial\Omega$.

The MAC scheme [1] is based on a staggered approach on structured grids, where the velocity and the pressure control volumes are dual to each other and have square or rectangular shape. Since the scheme is staggered, the pressure is not prone to instabilities. In this situation, convergence proofs for the linear Stokes and the nonlinear Navier–Stokes problems (with small data assumption) have been presented by Nicolaides [2, 3]. But in spite of its success, this scheme has the main drawback that complex geometries cannot be well approximated by structured grids. Therefore, several attempts have been made to generalize it for unstructured grids, see e.g., [4], where the unstructured simplex grid possesses the Delaunay property. Then the dual Voronoi grid can be defined in a sensible way, and two different staggered approaches are possible, where the pressure is discretized either in the triangles or at the vertices of the mesh:

1. in the first scheme, in the sequel called *tangential velocity scheme*, the velocity is approximated by its tangential values along the edges of the triangles, whereas the pressures are approximated at the vertices of the triangles;
2. in the second scheme, in the sequel called *normal velocity scheme*, the velocity is approximated by its normal values to the edges of the triangles, whereas the pressures are approximated at the center of the triangles.

For these generalized MAC schemes on unstructured grids, no convergence proofs have been found up to now. Therefore, we will present in this paper an appropriate discrete weak formulation of the problem, which allows to give a convergence proof [5]. Moreover, we show experimental orders of convergence in appropriate norms for a test problem with known analytical solution. It is worth noticing that for both schemes, the discrete rotation operator is consistent, but not the discrete divergence operator. In order to obtain a consistent discrete rotation operator for the tangential velocity scheme, the locations of the discrete velocity degrees of freedom are imagined as the midpoints of the triangle edges. On the other hand, in order to obtain a consistent discrete rotation operator for the normal velocity scheme, the locations of the discrete velocity degrees of freedom are imagined as the midpoints of the Voronoi faces. We note that for the tangential velocity scheme, the proposed discretization of $\nabla \cdot \boldsymbol{u} = 0$ exactly coincides with the discrete solenoidal condition allowing to prove a discrete maximum principle for the Voronoi finite volume method for convective transport of a dissolved species in the velocity field $\boldsymbol{u}$ [6].

The structure of the paper is as follows: In the second section, the notions of a Delaunay mesh and its dual, the Voronoi mesh, are introduced, and related quantities are defined. With these notions, discrete divergence and rotation operators for the tangential and the normal scheme are introduced, and both discretization schemes for the incompressible Stokes equations are presented. In the third section, a numerical example exhibits the convergence properties of both schemes on structured and unstructured grids. Experimental convergence rates for the tangential and normal scheme are given for the corresponding discrete L^2 norms for the velocities, the corresponding discrete L^2 norms for the pressure, and the corresponding discrete norms for the discrete rotation of the velocities.

2 Definition of the schemes

We define primal and dual meshes of the domain Ω as follows:

1. The set $\mathscr{T}$ is the finite set of disjoint triangles (considered as open subsets of $\mathbb{R}^2$) such that $\bigcup_{T\in\mathscr{T}} \overline{T} = \overline{\Omega}$. It is considered as the primal mesh. We denote by h_{mesh} the greatest diameter of all triangles. For all $T \in \mathscr{T}$, the point $\boldsymbol{x}_T$, defined as the center of the circumcircle of T, is such that $\boldsymbol{x}_T \in T$.
2. The set $\mathscr{V}$ contains the vertices of all the triangles (and therefore of the edges of the triangles). For all $\boldsymbol{y} \in \mathscr{V}$, we denote by $V_{\boldsymbol{y}}$ the Voronoi box around the vertex $\boldsymbol{y} \in \mathscr{V}$, defined as $V_{\boldsymbol{y}} = \{\boldsymbol{x} \in \Omega, |\boldsymbol{x} - \boldsymbol{y}| < |\boldsymbol{x} - \boldsymbol{y}'| \text{ for all } \boldsymbol{y}' \in \mathscr{V}, \boldsymbol{y}' \neq \boldsymbol{y}\}$. The set of Voronoi boxes is considered as the dual mesh.
3. The set $\mathscr{E}$ contains all the edges of the triangles, and is such that, for all $\sigma \in \mathscr{E}$, either σ is located on the boundary of Ω (we denote by $\mathscr{E}_{\text{bnd}}$ the set of these boundary edges), either σ is common to two neighboring triangles (we denote by $\mathscr{E}_{\text{int}}$ the set of these interior edges). We then denote by $\boldsymbol{x}_\sigma$ the middle of σ and by θ_{mesh} the infimum of all quantities $|\boldsymbol{x}_\sigma - \boldsymbol{x}_T|/h_T$, for all triangles T, and $h_T/h_{T'}$, for any pair of neighboring triangles T and T'.

We note that the circumcenter condition $\boldsymbol{x}_T \in T\ \forall T \in \mathscr{T}$ implies that the triangulation is acute, and, therefore, Delaunay. In agreement with the numerical results, we believe that it is possible to weaken the conditions on the triangulation to boundary conforming Delaunay meshes, i.e. Delaunay meshes with the additional property that $\boldsymbol{x}_T \in \Omega\ \forall T \in \mathscr{T}$ [7].

For every edge σ, we define a fixed orientation, which is given by a unit vector $\boldsymbol{t}_\sigma$ parallel to σ, and we define $\boldsymbol{n}_\sigma$ the normal vector to σ, obtained from $\boldsymbol{t}_\sigma$ by a rotation with angle $\pi/2$ in the counterclockwise sense (this rotation operator will be denoted as $\rho_{\frac{\pi}{2}}$, see Fig. 1). We further assume that the edges $\sigma \in \mathscr{E}_{\text{bnd}}$ at the border of Ω build a counterclockwise path around Ω. Then, for any edge $\sigma \in \mathscr{E}_{\text{bnd}}$ the exterior of Ω is located to the right of σ. For every $T \in \mathscr{T}$ we denote by $\mathscr{E}_T$ the set of edges of the triangle T, and we denote, for any $\sigma \in \mathscr{E}_T$, by $\boldsymbol{t}_{T,\sigma}$ the unit vector parallel to σ oriented in the counterclockwise sense around T, by $\boldsymbol{n}_{T,\sigma}$ the unit vector normal to σ and outward to T, and by $D_{T,\sigma}$ the cone with basis σ and vertex $\boldsymbol{x}_T$. For any $\sigma \in \mathscr{E}_{\text{int}}$, let T and T' be the two neighboring triangles such that σ is an edge of T and T'. We denote by $\sigma^\perp$ the segment $[\boldsymbol{x}_T, \boldsymbol{x}_{T'}]$ and by $D_\sigma = D_{T,\sigma} \cup D_{T',\sigma}$. For any $\sigma \in \mathscr{E}_{\text{bnd}}$, let T be the triangle such that σ is an edge of T. We then denote by $\sigma^\perp$ the segment $[\boldsymbol{x}_T, \boldsymbol{x}_\sigma]$ and by $D_\sigma = D_{T,\sigma}$.

For any $\boldsymbol{y} \in \mathscr{V}$, we denote by $\mathscr{E}_{\boldsymbol{y}}$ the set of all the edges where $\boldsymbol{y}$ is a vertex of, and we denote, for any $\sigma \in \mathscr{E}_{\boldsymbol{y}}$, by $\boldsymbol{t}_{\boldsymbol{y},\sigma}$ the unit vector parallel to σ oriented

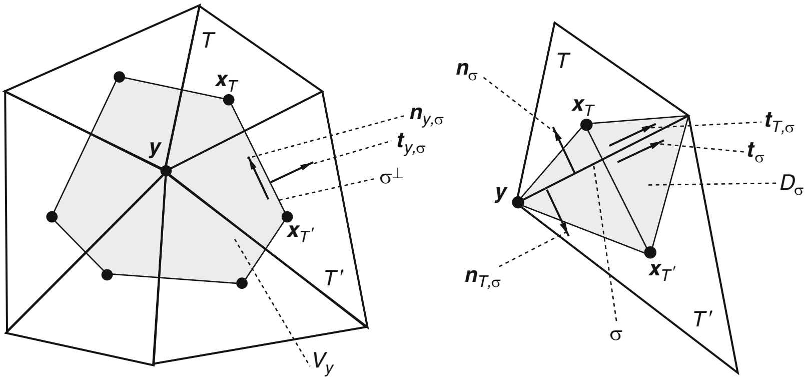


Fig. 1 Notations for the mesh: Left: the Voronoi box associated to a vertex. Right: Zoom on a diamond

from $\boldsymbol{y}$ to the other vertex of σ and by $\boldsymbol{n}_{\boldsymbol{y},\sigma}$ the unit vector normal to σ and in the counterclockwise sense around $\boldsymbol{y}$.

The space of degrees of freedom at edges, vertices and triangles are respectively defined by $X_{\mathscr{E}} = \mathbb{R}^{\mathscr{E}}$, $X_{\mathscr{V}} = \mathbb{R}^{\mathscr{V}}$ and $X_{\mathscr{T}} = \mathbb{R}^{\mathscr{T}}$.

For the tangential velocity scheme, the degrees of freedom for the velocity represent the tangential velocity components $\boldsymbol{v} \cdot \boldsymbol{t}_\sigma$ at the midpoint of the edges $\sigma \in \mathscr{E}$, which are oriented in the direction $\boldsymbol{t}_\sigma$. The degrees of freedom for the pressure represent the pressure at the vertices of the triangulation. The space

$$\dot{X}_{\mathscr{E}} = \{v \in X_{\mathscr{E}}, v_\sigma = 0, \forall \sigma \in \mathscr{E}_{\text{ext}}\} \tag{2}$$

represents the degrees of freedom for the velocity, when homogeneous Dirichlet boundary conditions are prescribed at the boundary edges. We introduce the following discrete differential operators:

$$\mathrm{rot}_T v = \frac{1}{|T|} \sum_{\sigma \in \mathscr{E}_T} |\sigma| v_\sigma \boldsymbol{t}_\sigma \cdot \boldsymbol{t}_{T,\sigma} \qquad \forall v \in X_{\mathscr{E}},\ \forall T \in \mathscr{T},$$

$$\mathrm{div}_{\boldsymbol{y}} v = \frac{1}{|V_{\boldsymbol{y}}|} \sum_{\sigma \in \mathscr{E}_{\boldsymbol{y}}} |\sigma^\perp| v_\sigma \boldsymbol{t}_\sigma \cdot \boldsymbol{t}_{\boldsymbol{y},\sigma} \qquad \forall v \in X_{\mathscr{E}},\ \forall \boldsymbol{y} \in \mathscr{V}.$$

Then the tangential velocity scheme reads:

$$\text{find } (v, p) \in \dot{X}_{\mathscr{E}} \times X_{\mathscr{V}} \text{ such that}$$

$$\sum_{T \in \mathscr{T}} |T| \mathrm{rot}_T v \mathrm{rot}_T w - \sum_{\boldsymbol{y} \in \mathscr{V}} |V_{\boldsymbol{y}}| p_{\boldsymbol{y}} \mathrm{div}_{\boldsymbol{y}} w = \sum_{\sigma \in \mathscr{E}} 2 w_\sigma \int_{D_\sigma} \boldsymbol{f} \cdot \boldsymbol{t}_\sigma \mathrm{d}\boldsymbol{x}, \ \forall w \in \dot{X}_{\mathscr{E}}$$

$$\sum_{\boldsymbol{y} \in \mathscr{V}} |V_{\boldsymbol{y}}| p_{\boldsymbol{y}} = 0,$$

$$\mathrm{Div}_{\boldsymbol{y}} v = 0, \ \forall \boldsymbol{y} \in \mathscr{V}.$$

For the normal velocity scheme, the degrees of freedom for the velocity represent the normal velocity components $\boldsymbol{v} \cdot \boldsymbol{n}_\sigma$ at the midpoints of the Voronoi faces $\sigma^\perp$ for all $\sigma \in \mathscr{E}$, which are oriented in the direction $\boldsymbol{n}_\sigma$, and the degrees of freedom for the pressure represent the pressure at the center of the triangles. Using the discrete differential operators

$$\mathrm{rot}_{\boldsymbol{y}} v = \frac{1}{|V_{\boldsymbol{y}}|} \sum_{\sigma \in \mathscr{E}_{\boldsymbol{y}}} |\sigma^\perp| v_\sigma \boldsymbol{n}_\sigma \cdot \boldsymbol{n}_{\boldsymbol{y}\sigma} \qquad \forall v \in X_{\mathscr{E}}, \ \forall \boldsymbol{y} \in \mathscr{V},$$

$$\mathrm{div}_T v = \frac{1}{|T|} \sum_{\sigma \in \mathscr{E}_T} |\sigma| v_\sigma \boldsymbol{n}_\sigma \cdot \boldsymbol{n}_{T\sigma} \qquad \forall v \in X_{\mathscr{E}}, \ \forall T \in \mathscr{T},$$

the normal velocity scheme writes:

$$\text{find } (v, p) \in \dot{X}_{\mathscr{E}} \times X_{\mathscr{T}} \text{ such that}$$

$$\sum_{\boldsymbol{y} \in \mathscr{V}} |V_{\boldsymbol{y}}| \mathrm{rot}_{\boldsymbol{y}} v \mathrm{rot}_{\boldsymbol{y}} w - \sum_{T \in \mathscr{T}} |T| p_T \mathrm{div}_T w = \sum_{\sigma \in \mathscr{E}} 2 w_\sigma \int_{D_\sigma} \boldsymbol{f} \cdot \boldsymbol{n}_\sigma \mathrm{d}\boldsymbol{x}, \ \forall w \in \dot{X}_{\mathscr{E}}$$

$$\sum_{T \in \mathscr{T}} |T| p_T = 0,$$

$$\mathrm{div}_T v = 0, \ \forall T \in \mathscr{T}.$$

3 Numerical results

In order to investigate numerically the convergence rate that can be achieved with the extended MAC schemes introduced above, we define an academic Stokes problem on two sequences of meshes. We remark that we achieved the same experimental convergence rates for the full nonlinear Navier–Stokes equations [5], where the nonlinear term was discretized in rotational form. The problem is posed on $\Omega = [0, 1]^2$, has homogeneous Dirichlet boundary conditions and reads

$$\boldsymbol{v} = \begin{pmatrix} 2(x-1)^2x^2(y-1)y(2y-1) \\ -2(2x-1)(x-1)x(y-1)^2y^2 \end{pmatrix},$$

$$p = x^3 + y^3 - 0.5.$$

The vector $\boldsymbol{f}$ is computed such that $\boldsymbol{v}$ and p fulfill the Stokes equations (1).

In the first sequence of meshes, every mesh is build up from small squares, where the side length of such a square defines the mesh size. Every square in the mesh is split into two triangles. This mesh is not admissible in the strict sense of the above definition, since the circumcenters of these two triangles coincide. But this does not pose any problem, since in this degenerated case, the discrete method is equivalent to a method where the squares take over the role of triangles, and the diagonals of the squares can be removed from the above considerations, as the measure of their corresponding Voronoi faces are zero. At the same time, on these meshes, triangle edge midpoints and Voronoi face midpoints coincide. This fact will result in superior convergence behavior on these meshes in comparison to "purely" triangular meshes.

In Table 1 we show some information about the degrees of freedom in these square meshes. The last two columns of this Table show some quite interesting information. The penultimate column reveals that the tangential velocity scheme is quite efficient in terms of degrees of freedom, since the ratio between the number of degrees of freedom corresponding to discretely divergence-free velocities and the total number of degrees of freedom is about 0.5. For the normal velocity scheme, the corresponding ratio is only 0.20.

Table 1 Number of edges, vertices and triangles in different square meshes. The penultimate column shows the ratio between discretely divergence-free degrees of freedom and the total number of degrees of freedom for the tangential velocity scheme. The last column shows the ratio between discretely divergence-free degrees of freedom and the total number of degrees of freedom for the normal velocity scheme

mesh size	$\lvert E\rvert$	$\lvert V\rvert$	$\lvert T\rvert$	$\frac{\lvert E\rvert-\lvert V\rvert}{\lvert E\rvert+\lvert V\rvert}$	$\frac{\lvert E\rvert-\lvert T\rvert}{\lvert E\rvert+\lvert T\rvert}$
$\frac{1}{32}$	2945	1024	1922	0.484	0.210
$\frac{1}{64}$	12033	4096	7938	0.492	0.205
$\frac{1}{128}$	48641	16384	32258	0.496	0.203
$\frac{1}{256}$	195585	65536	130050	0.498	0.201
$\frac{1}{512}$	784385	262144	522242	0.499	0.201
$\frac{1}{1024}$	3141633	1048576	2093058	0.500	0.200

The second sequence of meshes are made up of isotropic, unstructured boundary conforming Delaunay meshes. They have been generated by the mesh generator TRIANGLE [8]. We remark, that this approach does not guarantee that the triangulation is acute. In Table 2 we show some information about the degrees of freedom in these triangle meshes. An approximate mesh size was defined according to the largest triangle area that the mesh generator was allowed to generate within a mesh.

From Tables 1 and 2 we recognize that the degrees of freedom of corresponding meshes in the two mesh families are quite similar, such that the definition of the mesh size for unstructured meshes seems to be reasonable. The two schemes are implemented within the framework of the software package PDELIB[9]. All the discrete linear systems are solved with the direct solver PARDISO[10, 11].

Table 2 Number of edges, vertices and triangles in different Delaunay meshes generated by the mesh generator TRIANGLE[8]. The penultimate column shows the ratio between discretely divergence-free degrees of freedom and the total number of degrees of freedom for the tangential velocity scheme. The last column shows the ratio between discretely divergence-free degrees of freedom and the total number of degrees of freedom for the normal velocity scheme

mesh size	$\|E\|$	$\|V\|$	$\|T\|$	$\frac{\|E\|-\|V\|}{\|E\|+\|V\|}$	$\frac{\|E\|-\|T\|}{\|E\|+\|T\|}$
$\frac{1}{32}$	3121	1084	2038	0.484	0.210
$\frac{1}{64}$	12326	4195	8132	0.492	0.205
$\frac{1}{128}$	48664	16393	32272	0.496	0.203
$\frac{1}{256}$	194879	65302	129578	0.498	0.201
$\frac{1}{512}$	779506	260519	518988	0.499	0.201
$\frac{1}{1024}$	3114404	1039501	2074904	0.500	0.200

In Figs. 2 and 3, for both schemes and series of meshes, we plot various measures of the error between the discrete solution and a projection of the exact solution onto the grid. We used two different projections for both schemes. For the tangential velocity scheme, we evaluate the tangential velocities at the edge midpoints and assign them to the corresponding velocity degrees of freedom. For the normal velocity scheme, we evaluate the normal velocities at the Voronoi face midpoints and assign them to the corresponding velocity degrees of freedom, likewise.

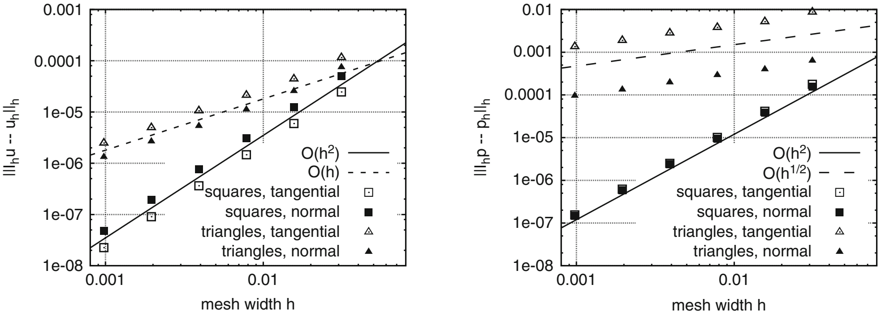


Fig. 2 Discrete L^2-norm of the error between the projected exact solution and the discrete solution. Left: velocity, right: pressure

We start the discussion with the approximation of the velocity, see Fig. 2, left. We observe similar behavior for the two discretization schemes proposed. On triangular

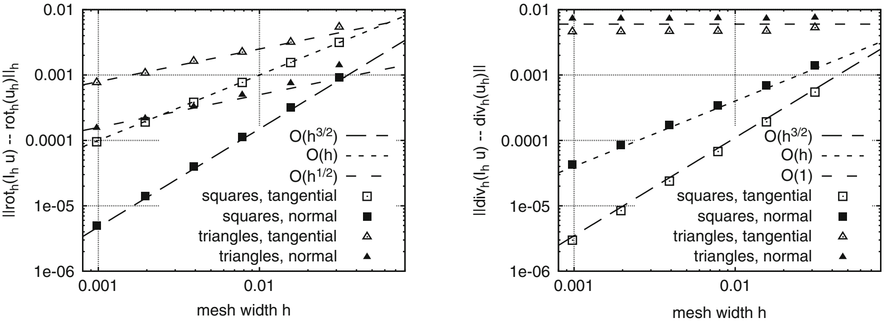


Fig. 3 Discrete L^2-norm of the discrete vector calculus operators applied to the difference between the projected exact velocity and the velocity component of the discrete solution. Left: rotational, right: divergence

meshes the convergence order is approximately $O(h)$. On square meshes, we gain an order of magnitude in the convergence rate in comparison to the triangular meshes.

Also, concerning the approximation orders of the pressure, both schemes behave in a similar way, including second order convergence on square meshes, see Fig. 2, right. We observe that on triangular meshes, the convergence order drops to $O(h^{\frac{1}{2}})$. At the same time, the accuracy of the normal velocity scheme on triangular meshes is better by a factor of $\approx$ 10 in comparison to the tangential velocity scheme.

As shown by the mathematical analysis [5], the discrete rotation is convergent for both schemes. This is confirmed by Fig. 3 (left), where we observe the convergence of the difference between the discrete rotation of the discrete solution and the discrete rotation of the projected exact solution. On square meshes, for the normal velocity scheme, the L^2 norm of this difference exhibits $O(h^{\frac{3}{2}})$-convergence, while the convergence order of the tangential velocity scheme is only $O(h)$. On triangular meshes, both schemes exhibit $O(h^{\frac{1}{2}})$ convergence with an advantage for the normal velocity scheme concerning the constants.

By construction, for both schemes, the discrete divergence of the velocity component of the discrete solution is zero. Therefore, the error shown in Fig. 3 (right) coincides with the discrete divergence of the projected exact velocity. On square meshes, for both schemes, the discrete divergence operator is consistent, since mid points of an edge coincide with mid points of the orthogonal Voronoi faces. Therefore, the discrete divergence converges on square meshes to zero with order $O(h^{1.5})$ for the tangential velocity scheme and $O(h)$ for the normal velocity scheme. On the triangular meshes, edge mid points and Voronoi face mid points do not coincide and the discrete divergence operator is not consistent resulting in no convergence at all if it is applied to the projection of the velocity component of the exact solution.

We note that the convergence behavior on the boundary conforming Delaunay meshes, which are not acute, is consistent with the theoretical considerations which for technical reasons had been constrained to acute triangulations.

References

1. F. H. Harlow and J. E. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *Physics of fluids*, 8(12):2182–2189, 1965.
2. R. A. Nicolaides. Analysis and convergence of the MAC scheme. I. The linear problem. *SIAM J. Numer. Anal.*, 29(6):1579–1591, 1992.
3. R. A. Nicolaides and X. Wu. Analysis and convergence of the MAC scheme. II. Navier-Stokes equations. *Math. Comp.*, 65(213):29–44, 1996.
4. J. Nicolaides, T. A. Porsching, and C. A. Hall. Covolume methods in computational fluid dynamics. In M. Hafez and K. Oshma, editors, *Computation Fluid Dynamics Review*, pages 279–299. John Wiley and Sons, New York, 1995.
5. J. Fuhrmann R. Eymard and A. Linke. Extended MAC schemes on Delaunay meshes for the incompressible Navier-Stokes equations, 2011. In preparation.
6. J. Fuhrmann, A. Linke, and H. Langmach. Mass conservative coupling between fluid flow and solute transport. In *Finite Volumes for Complex Application VI*. Springer, 2011.
7. H. Si, K. Gärtner, and J. Fuhrmann. Boundary conforming Delaunay mesh generation. *Comput. Math. Math. Phys.*, 50:38–53, 2010.
8. J. Shewchuk. Triangle: A two-dimensional quality mesh generator and Delaunay triangulator. http://www.cs.cmu.edu/ quake/triangle.html, University of California at Berkeley.
9. J. Fuhrmann et al. Pdelib. www.wias-berlin.de/software/pdelib/.
10. O. Schenk, K. Gärtner, and W. Fichtner. Efficient sparse LU factorization with left-right looking strategy on shared memory multiprocessors. *BIT*, 40(1):158–176, 1999.
11. O. Schenk, K. Gärtner, G. Karypis, S. Röllin, and M. Hagemann. PARDISO Solver Project. URL: http://www.pardiso-project.org, 2010. Retrieved 2010-02-15.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Multiphase Flow in Porous Media Using the VAG Scheme

Robert Eymard, Cindy Guichard, Raphaèle Herbin, and Roland Masson

Abstract We present the use of the Vertex Approximate Gradient scheme for the simulation of multiphase flow in porous media. The porous volume is distributed to the natural grid blocks and to the vertices, hence leading to a new finite volume mesh. Then the unknowns in the control volumes may be eliminated, and a 27-point scheme results on the vertices unknowns for a hexahedral structured mesh. Numerical results show the efficiency of the scheme in various situations, including miscible gas injection.

Keywords two-phase flow in porous media, vertex approximate gradient scheme, reservoir simulation.
MSC2010: 65M08,76S05

1 Introduction

Simulation of multiphase flow in porous media is a complex task, which has been the object of several works over a long period of time, see the reference books [12] and [3]. Several types of numerical schemes have been proposed in the past decades. Those which are implemented in industrial codes are mainly built upon cell centred approximations and discrete fluxes, in a framework which is also that of the method

R. Eymard
Université Paris-Est, France, e-mail: robert.eymard@univ-mlv.fr

C. Guichard
Université Paris-Est and IFP Energies nouvelles, France, e-mail: cindy.guichard@ifpen.fr

R. Herbin
Université Aix-Marseille, France, e-mail: raphaele.herbin@latp.univ-mrs.fr

R. Masson
IFP Energies nouvelles, France, e-mail: roland.masson@ifpen.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_43, © Springer-Verlag Berlin Heidelberg 2011

we propose here. Let us briefly sketch this framework. The 3D simulation domain Ω is meshed by control volumes $X \in \mathscr{M}$. Let us denote by Λ the diffusion matrix (which is a possibly full matrix depending on the point of the domain).

For each control volume $X \in \mathscr{M}$, the set of neighbours $Y \in \mathscr{N}_X$ is the set of all control volumes involved in the mass balance in X, which means that the following approximation formula is used: $-\int_X \nabla \cdot \Lambda\, grad\ p\, \mathrm{d}x \simeq \sum_{Y \in \mathscr{N}_X} F_{X,Y}(P)$, where $P = (p_Z)_{Z \in \mathscr{M}}$ is the family of all pressure unknowns in the control volumes, and where the flux $F_{X,Y}(P)$, between control volumes X and Y, is a linear function of the components of P which ensures the following conservativity property:

$$F_{X,Y}(P) = -F_{Y,X}(P). \tag{1}$$

Such a linear function, which is expected to vanish on constant families, may be defined by

$$F_{X,Y}(P) = \sum_{Z \in \mathscr{M}_{X,Y}} a_{X,Y}^{Z} p_Z, \tag{2}$$

where the family $(a_{X,Y}^{Z})_{Z \in \mathscr{M}_{X,Y}}$ and $\mathscr{M}_{X,Y} \subset \mathscr{M}$ are such that $\sum_{Z \in \mathscr{M}_{X,Y}} a_{X,Y}^{Z} = 0$.

Assuming N_c constituents and N_α phases, the discrete balance laws then read

$$\begin{aligned} &\frac{\Phi_X}{\delta t}(A_{X,i}^{(n+1)} - A_{X,i}^{(n)}) + \sum_{\alpha=1}^{N_\alpha} \sum_{Y \in \mathscr{N}_X} M_{X,Y,i}^{(n+1),\alpha} F_{X,Y}^{(n+1),\alpha} = 0,\ \forall i = 1, \ldots, N_c, \\ &F_{X,Y}^{(n+1),\alpha} = F_{X,Y}(P^{(n+1),\alpha}) - \rho_{X,Y}^{(n+1),\alpha} g \cdot (x_Y - x_X),\ \forall \alpha = 1, \ldots, N_\alpha, \end{aligned} \tag{3}$$

where n is the time index, δt is the time step, Φ_X is the porous volume of the control volume $X \in \mathscr{M}$, $A_{X,i}$ represents the accumulation of constituent i in the control volume X per unit pore volume (assumed to take into account the dependence of the porosity with respect to the pressure), $M_{X,Y,i}^{\alpha}$ is the amount of constituent i transported by phase α from the control volume X to the control volume Y (generally computed by taking the upstream value with respect to the sign of $F_{X,Y}$), P^α is the family of the pressure unknowns of phase α, g is the gravity acceleration, $\rho_{X,Y}^{\alpha}$ is the bulk density of phase α between control volumes X and Y and x_X is the centre of control volume X. In addition to these relations, the differences between the phase pressures are ruled by capillary pressure laws. Thermodynamical equilibrium and standard closure relations are used.

When applying scheme (3), one should be very wary of the use of conformal finite elements in the case of highly heterogeneous media. Indeed, assuming that the control volumes are vertex centred with vertices located at the interfaces between different media, then the porous volume concerned by the flow of very permeable medium includes that of non permeable medium. This may lead to surprisingly wrong results on the component velocities. A possible interpretation of these poor results is that, when seen as a set of discrete balance laws, the finite element method provides the same amount of impermeable and permeable porous volume for the accumulation term for a node located at a heterogeneous interface.

We present in this paper the use of a new scheme, called Vertex Approximate Gradient (VAG) scheme [8, 9], which can be implemented in (3) so that the components velocities are correctly approximated, thanks to a special choice of the control volumes and of the discrete fluxes, which respect to the form (2). The purpose of respecting the form (3)-(2) is to be able to plug it easily into an existing reservoir code, say Multi-Point Flux Approximation (MPFA), by simply redefining the control volumes and the coefficients $a^{Z}_{X,Y}$ of the discrete flux.

Although part of this scheme is vertex centred, we show that the solution obtained on a very heterogeneous medium with a coarse mesh remains accurate. This is a great advantage of this scheme, which is also always coercive, symmetric, and leads to a 27-stencil on hexahedral structured meshes. In addition the VAG scheme is very efficient on meshes with tetrahedra since the scheme can then be written with the nodal unknowns only, thus inducing a reduction of the number of degrees of freedom by a factor 5 compared with cell centred finite volume schemes such as MPFA schemes [1, 2, 4, 5].

2 Presentation of the scheme

The VAG scheme is described in [8,9], and its gradient scheme properties are related to those presented in [7]; therefore we focus here on the use of this scheme for a multiphase flow simulation of the form (3). Let $\mathscr{M}$ be a general mesh of Ω, defined by a set $\mathscr{G}$ of grid blocks and the set $\mathscr{V}$ of their vertices; this is a mesh of control volumes in the sense of the preceding section: a control volume is either a grid block $K \in \mathscr{G}$ or a vertex $v \in \mathscr{V}$. In particular, a porous volume must be associated to each control volume, *i.e.* to each grid block and to each vertex. Finally a flux $F_{X,Y}$ from the control volume X to the control volume Y must be specified.

Any given grid block $K \in \mathscr{G}$ has, say, N_K vertices; let us denote by $\mathscr{V}_K \subset \mathscr{V}$ the set of these vertices. We wish to define a flux between neighbouring control volumes $X = K$ and $Y = v \in \mathscr{V}_K$, and between neighbouring control volumes $X = v \in \mathscr{V}_K$ and $Y = K \in \mathscr{G}_v = \{Y = K \in \mathscr{G}$ such that $v \in \mathscr{V}_K\}$; for this purpose, we introduce a local discrete gradient $\nabla_{K,v}(P_K) \in \mathbb{R}^3$ (see [8,9] for the precise definitions), which only depends on the values $P_K = (P_{K,v})_{v\in\mathscr{V}_K} = (p_v - p_K)_{v\in\mathscr{V}_K}$. We then introduce the matrices $(A_K^{v,v'})_{v,v'\in\mathscr{V}_K}$, which are defined by the following relation

$$\frac{|K|}{N_K}\sum_{v\in\mathscr{V}_K} \Lambda_K \nabla_{K,v} P_K \cdot \nabla_{K,v} Q_K = \sum_{v\in\mathscr{V}_K}\sum_{v'\in\mathscr{V}_K} A_K^{v,v'} P_{K,v'} Q_{K,v}, \ \forall P_K, Q_K \in \mathbb{R}^{\mathscr{V}_K}.$$

The flux from control volume $X = K$ to control volume $Y = v$ is then given by

$$F_{X,Y}(P) = F_{K,v}(P) = -\sum_{v'\in\mathscr{V}_K} A_K^{v,v'}(p_{v'} - p_K),$$

which is of the same form as (2) ; using (1), we get $F_{Y,X}(P) = -F_{X,Y}(P)$. Let us now turn to the definition of porous volumes for all $X \in \mathscr{M}$. The question is to associate to each vertex a porous volume in such a way that the component velocities are well approximated. Let us denote by $\widetilde{\Phi}_K = \int_K \Phi(x)\,\mathrm{d}\,x$ the total porous volume of each grid block $K \in \mathscr{G}$. We shall then take out a little bit of this porous volume of each grid block to associate it with the control volumes of the vertices. In order to obtain a systematic way to redistribute the porous volume between the grid blocks and the vertices, we define a first indicator of the transmissivity between K and v by $B_{K,v} = \sum_{v' \in \mathscr{V}_K} A_K^{v,v'} > 0, \forall K \in \mathscr{G}_v$, and then, for a global small value $\mu \in]0,1[$ (for example, $\mu = 0.05$), a weighted relative transmissivity (which is larger for permeable regions than for impermeable ones):

$$\widetilde{B}_{K,v} = \mu \frac{B_{K,v}}{\sum\limits_{L \in \mathscr{G}_v} B_{L,v}},\ \forall v \in \mathscr{V},\ \forall K \in \mathscr{G}_v, \tag{4}$$

Note that it might be expected that a too small value for μ lead to some numerical problems; nevertheless, such consequences have not been observed within the range $\mu \in [0.01, 0.05]$. The total porous volume can then be redistributed between all control volumes $X \in \mathscr{M}$, that is between the grid blocks and the vertices, by the following relations:

$$\Phi_X = \begin{cases} \sum_{K \in \mathscr{G}_v} \widetilde{B}_{K,v} \widetilde{\Phi}_K & \text{if } X = v \in \mathscr{V}, \\ \widetilde{\Phi}_K (1 - \sum_{v \in \mathscr{V}_K} \widetilde{B}_{K,v}) & \text{if } X = K \in \mathscr{G}. \end{cases}$$

Hence, we distribute a small amount of the porous volume of K to its vertices, in a conservative way; indeed, by construction, we get that

$$\sum_{X \in \mathscr{M}} \Phi_X = \sum_{K \in \mathscr{G}} \widetilde{\Phi}_K,$$

with all $\Phi_X > 0$, provided that the value μ be chosen sufficiently small. We can remark that:

1. the porous volume of a vertex $v \in \mathscr{V}$ located at the interface between high and low permeability regions is mainly extracted from the higher permeability region,
2. the part of the lower permeability region distributed to the vertices is reduced by the factor μ.

We recall that we keep the property ensured in the monophasic case on the full system: indeed, the linear systems issued from Newton's method may be solved by first eliminating all unknowns $K \in \mathscr{G}$, and then solve a 27-point system on $v \in \mathscr{V}$.

3 Numerical applications

3.1 *Heterogeneous case*

The first example is the injection of CO_2, considered as immiscible with the liquid phase, at the middle point of an isotropic and heterogeneous reservoir, with size $[-100, 100] \times [0, 50] \times [0, 45]\ m^3$. The reservoir includes three 15 m-thick layers. The top and bottom layers are assumed to be weakly permeable ($|\Lambda| = 10^{-16}\ m^2$) and the medium layer is much more permeable ($|\Lambda| = 10^{-12}\ m^2$). A regular coarse $100 \times 10 \times 15$ mesh is used for the simulation (depicted in Fig. 1). The values $\mu = 0.01$ and $\mu = 0.05$ have been tested in (4), without significant influence on the results both in terms of accuracy and CPU time. The results of the VAG scheme are compared to those obtained using the two-point flux approximation (TPFA) scheme, which is available on such a regular mesh. We observe in Fig. 2 that the numerical diffusion along the axes of the mesh leads, after a short injection time, to a distorted profile of the gas saturation in the case of the TPFA scheme, known as Grid Orientation Effect (GOE), see also [10]. This phenomenon is clearer in the profile of the saturation at the end of the gas injection. We see in Fig. 3 the important GOE due to the TPFA scheme, whereas this effect is nearly invisible in the results obtained with the VAG scheme. Moreover, this distortion, also visible in the vertical section (Fig. 4), is again corrected using the VAG scheme. These results

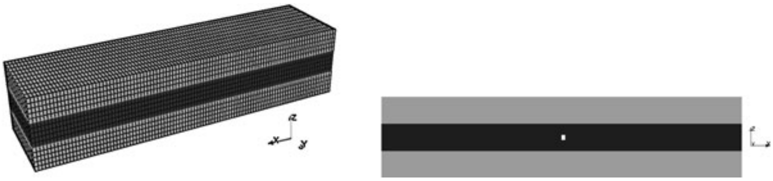


Fig. 1 First example. Left: mesh and layers. Right: the well is depicted at the centre of the section $y = 25\ m$, illustrated by the white block

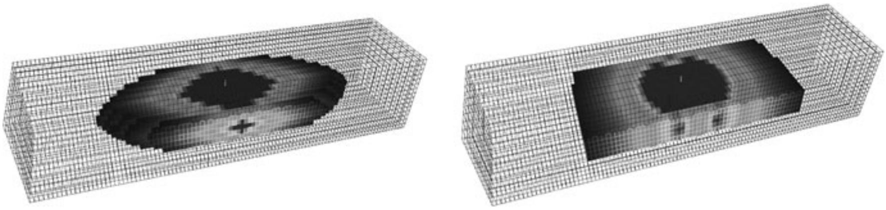


Fig. 2 View of the gas saturation in the reservoir, after a short injection time. Farthest to the well: $S = 0.001$. Closest to the well : $S = 0.042$. Left: TPFA scheme. Right: VAG scheme

can be explained by the construction of the fluxes. In fact, after elimination of the cell centred unknowns, the resulting scheme on the vertex unknowns has a 27-point stencil, whereas it remains a 7-point scheme on the control volumes unknowns using the TPFA scheme.

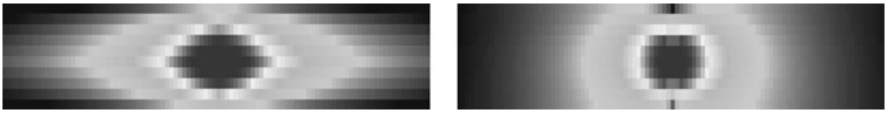


Fig. 3 Gas saturation at the end of the gas injection. Section $z = 22.5\ m$. Farthest to the well : $S = 0$. Closest to the well : $S = 1$. Left: TPFA scheme. Right: VAG scheme

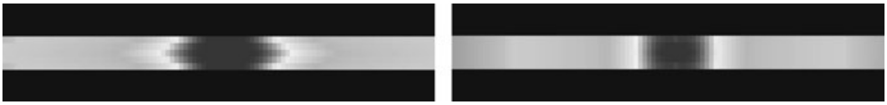


Fig. 4 Gas saturation at the end of the gas injection. Section $y = 25\ m$. Farthest to the well : $S = 0$. Closest to the well : $S = 1$. Left: TPFA scheme. Right: VAG scheme

3.2 Near-Well case

In the second example, we consider the numerical simulation of the injection of CO_2 in near-well regions for a deviated well. A hexahedral radial part is connected to the outside boundary either by a hexahedral mesh or a hybrid mesh (using both pyramids and tetrahedra) as illustrated in Fig. 5. The number of cells is roughly the same for both types of grids. This family of meshes is also used in the 3D benchmark on monophasic diffusion [11]. The medium is homogeneous, but anisotropic. We consider that the CO_2 can be dissolved in the aqueous phase.

We consider in Figs. 7 and 6 the mass outflow rate of CO_2 in the two phases at the outer boundary using the VAG scheme and the MPFA O-scheme on both types of grids. The values 0.01 and 0.05 have been tested for the parameter μ used in (4) and the results are almost the same. In order to keep the output clearer, the curves are only plotted for $\mu = 0.05$. We observe that the VAG scheme produces results which are not very sensitive to the type of the grid. On the contrary, the MPFA O-scheme shows a significant sensitivity to the type the grid, since the production of CO_2 is slowed down by the use of the tetrahedral mesh.

We finally remark that there are 74 679 cell unknowns and 74 800 nodal unknowns for the hexahedral mesh, to be compared with 77 599 cell unknowns (including 28 704 tetrahedra) and only 37 883 nodal unknowns for the hybrid mesh.

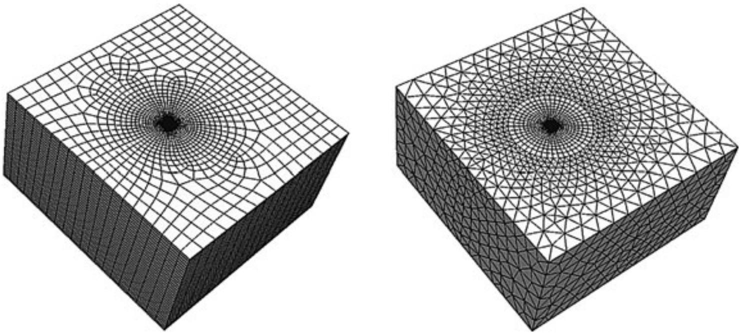


Fig. 5 Near-well grid : the hexahedral mesh (left) and the hybrid mesh (right)

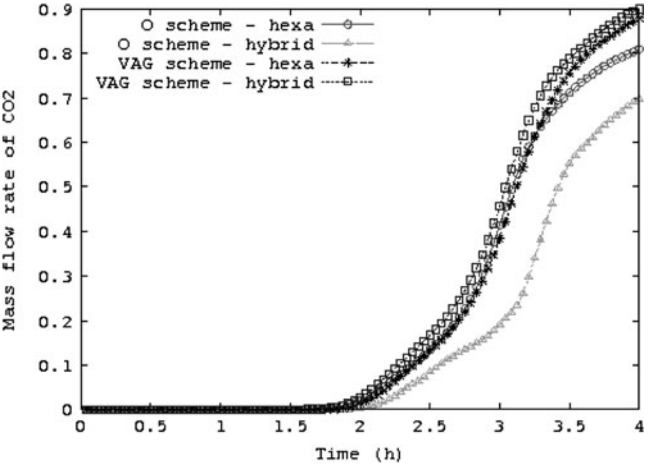


Fig. 6 Outflowing mass flow rate of CO_2 in the water phase at the outer boundary

As stated in the introduction, we see on this example that computing costs of the VAG scheme may be reduced in the case of meshes with tetrahedra.

4 Conclusion

The above numerical results show that the VAG scheme seems to be an efficient scheme for multiphase flow simulation of a heterogeneous anisotropic reservoir; it features the following properties:

1. it may be implemented, without any additional cost, into an MPFA industrial code;

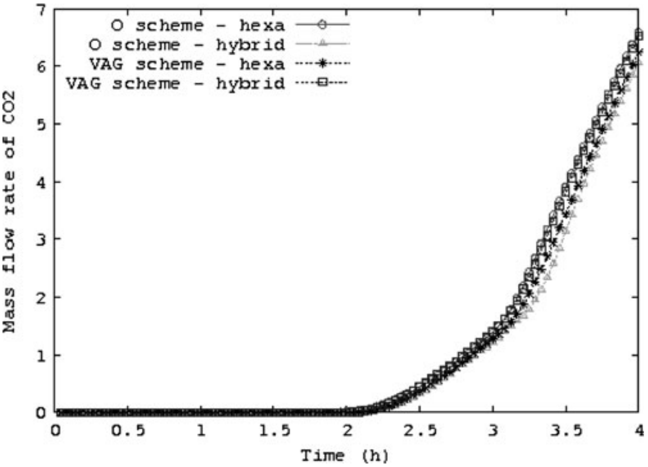


Fig. 7 Outflowing mass flow rate of CO_2 in the gas phase at the outer boundary

2. it leads to a 27-point compact stencil and a symmetric and coercive operator for the treatment of the diffusion terms, even in the case of distorted meshes and heterogeneous and anisotropic diffusion,
3. its cost is considerably reduced in the case of meshes with tetrahedra compared with cell centred MPFA schemes;
4. it remains accurate on coarse meshes thanks to a well-chosen distribution of the porous volume between the centre of the control volumes and the vertices;
5. since a pore volume is assigned to the Neumann boundary nodes, the Neumann conditions are obtained by writing the conservation and closure equations as in the inner control volumes.

Full scale reservoir simulations will be performed in order to confirm the efficiency of the method.

Acknowledgements Work sponsored by GNR MOMAS and ANR VFSitCom.

References

1. I. Aavatsmark, T. Barkve, O. Bøe and T. Mannseth. Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.*, 127 (1):2-14, 1996.
2. I. Aavatsmark, GT. Eigestad, BT. Mallison and JM. Nordbotten. A compact multipoint flux approximation method with improved robustness. *Numer. Methods Partial Diff. Eqns.*, 27 (5):1329-1360, 2008.
3. K. Aziz and A. Settari. Petroleum reservoir simulation. *Chapman & Hall*, 1979.
4. MG. Edwards. Unstructured, control-volume distributed, full-tensor finite-volume schemes with flow based grids. *Computational Geosciences*, 6 (3):433-452, 2002.

5. MG. Edwards and CF. Rogers. Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Computational Geosciences*, 2 (4):259-290, 1998.
6. R. Eymard, T. Gallouët, and R. Herbin. 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(4):1009–1043, 2010. see also http://hal.archives-ouvertes.fr/.
7. R. Eymard and R. Herbin. Gradient schemes for diffusion problem. *in FVCA VI proc.*, Prague, June 6-10, 2011.
8. R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3D schemes for diffusive flows in porous media. *submitted*, 2010. see also http://hal.archives-ouvertes.fr/.
9. R. Eymard, C. Guichard and R. Herbin. Benchmark 3D: the VAG scheme. *in FVCA VI proc.*, Prague, June 6-10, 2011
10. R. Eymard, C. Guichard and R. Masson. Grid orientation effect and multipoint flux approximation. *in FVCA VI proc.*, Prague, June 6-10, 2011.
11. R. Herbin and F. Hubert. Benchmark 3D on discretization schemes for anisotropic diffusion problem on general grids. *in FVCA VI proc.*, Prague, June 6-10, 2011.
12. D. Peaceman. Fundamentals of numerical reservoir simulation. *Elsevier*, 1977.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Grid Orientation Effect and MultiPoint Flux Approximation

Robert Eymard, Cindy Guichard, and Roland Masson

Abstract Some cases of nonlinear coupling between a diffusion equation, related to the computation of a pressure field within a porous medium, and a convection equation, related to the conservation of a species, lead to the apparition of the so-called grid orientation effect. We propose in this paper a new procedure to eliminate this Grid Orientation Effect, only based on the modification of the stencil of the discrete version of the convection equation. Numerical results show the efficiency and the accuracy of the method.

Keywords Grid Orientation Effect, Two Phase Porous Media Flow, Finite Volume Methods
MSC2010: 76S05, 65M08, 76E06

1 Introduction

In the 1980's, numerous papers have been concerned with the so-called grid orientation effect, in the framework of oil reservoir simulation. This effect is due to the anisotropy of the numerical diffusion induced by the upstream weighting scheme, and the computation of a pressure field, solution to an elliptic equation in which the diffusion coefficient depends on the value of the convected unknown. This problem has been partly solved in the framework of industrial codes, in which the meshes are structured and regular (mainly based on squares and cubes). The literature on

R. Eymard
Université Paris-Est, France, e-mail: robert.eymard@univ-mlv.fr

C. Guichard (work supported by ANR VFSitCom)
Université Paris-Est and IFP Energies nouvelles, France, e-mail: cindy.guichard@ifpen.fr

R. Masson
IFP Energies nouvelles, France, e-mail: roland.masson@ifpen.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_44, © Springer-Verlag Berlin Heidelberg 2011

this problem is huge, and is impossible to exhaustively quote; let us only cite [3, 4, 6, 10, 11] and references therein. In the 2000's, a series of new schemes have been introduced in order to compute these coupled problems on general grids [1, 2, 5, 8]. But, in most of the cases, the non regular meshes conserve structured directions, although the shape of the control volumes is no longer that of a regular cube. This is the case for the Corner Point Geometries [9] widely used in industrial reservoir simulations. The control volumes which are commonly used in 3D reservoir simulations are generalised "hexahedra", in the sense that each of them is neighboured by 6 other control volumes. In this case, the stencil for the pressure resolution may have a 27-point stencil (using for instance a MPFA scheme). Nevertheless, selecting a 27-point stencil instead of a 7-point stencil for the pressure resolution has no influence on the Grid Orientation Effect, which results from the stencil used in upstream weighted mass exchanges coupled with the pressure resolution.

In order to overcome this problem, we study here a generalisation of methods consisting in increasing the stencil of the convection equation, without modifying the pressure equation. The method will be presented on a simplified problem, modelling immiscible two-phase flow within a porous medium. Let $\Omega \subset \mathbb{R}^d$ (with $d = 2$ or 3) be the considered space domain. We consider the following two-phase flow problem in Ω:

$$\begin{cases} u_t - \mathrm{div}(k_1(u)\Lambda\nabla p) &= 0 \\ (1-u)_t - \mathrm{div}(k_2(u)\Lambda\nabla p) &= 0, \end{cases} \tag{1}$$

where $u(x,t) \in [0,1]$ is the saturation of phase 1 (for example water), and therefore $1 - u(x,t)$ is the saturation of phase 2, k_1 is the mobility of phase 1 (increasing function such that $k_1(0) = 0$), k_2 is the mobility of phase 2 (decreasing function such that $k_2(1) = 0$), and p is the common pressure of both phases (the capillary pressure is assumed to be negligible in front of the pressure gradients due to injection and production wells) and we consider a horizontal medium with permeability tensor Λ. It is therefore possible to see System (1) as the coupling of an elliptic problem with unknown p and a nonlinear scalar hyperbolic problem with unknown u:

$$\begin{cases} m(u) = k_1(u) + k_2(u),\, f(u) = \dfrac{k_1(u)}{m(u)} \\ \mathrm{div}F = 0 \text{ with } F = -m(u)\Lambda\nabla p \\ u_t + \mathrm{div}(f(u)F) = 0 \end{cases} \tag{2}$$

We then consider a MultiPoint Flux Approximation finite volume scheme for the approximation of Problem (1), coupled with an upstream weighting scheme for the mass exchanges. Such a scheme may be written:

$$F_{K,L}^{(n)} = m_{KL}^{(n)} \sum_{M\in\mathcal{M}} a_{KL}^M p_M^{(n+1)} \text{ with } \sum_{M\in\mathcal{M}} a_{KL}^M = 0 \tag{3}$$

$$\sum_{L\in\mathcal{N}_K} F_{K,L}^{(n)} = 0 \tag{4}$$

$$F_{K,L}^{(n)} + F_{L,K}^{(n)} = 0 \quad (5)$$

$$|K| \left(u_K^{(n+1)} - u_K^{(n)}\right) + \delta t^n \sum_{L\in\mathscr{N}_K} \left(f(u_K^{(n)})(F_{K,L}^{(n)})^+ - f(u_L^{(n)})(F_{K,L}^{(n)})^-\right) = 0. \quad (6)$$

In the above system, we denote by $\mathscr{M}$ the finite volume mesh of Ω, K, L are control volumes, $\mathscr{N}_K$ is the set of the neighbours of K (*i.e.* control volumes exchanging fluid mass with K), n is the time index and δt^n is the time step ($\delta t^n = t^{(n+1)} - t^{(n)}$), $p_M^{(n)}$ and $u_M^{(n)}$ are respectively the pressure and the saturation in control volume M at time $t^{(n)}$. The coefficients a_{KL}^M are computed with respect to the geometry of the mesh and to Λ. The value $m_{KL}^{(n)}$ is any average value (arithmetic or harmonic) of the values $m(u_K^{(n)})$ and $m(u_L^{(n)})$. Then $F_{K,L}^{(n)}$ is the approximation of $F \cdot n$ at the interface $K|L$ between control volumes K and L at time step n, and, for all real a, the values a^+ and a^- are respectively defined by $\max(a, 0)$ and $\max(-a, 0)$.

The set $\mathscr{N}_K$ of the neighbours of K is classically defined as all the control volumes which have a common face with K. But, as we show in this paper, this notion may be relaxed. Defining the notion of "stencil" $S \subset \mathscr{M}^2$ by $S = \{(K, L) \in \mathscr{M}^2, L \in \mathscr{N}_K\}$, this stencil is then equal to the set of all $(K, L) \in \mathscr{M}^2$ such that $F_{K,L}^{(n)}$ may be different from 0. In view of (5), S must verify the symmetry property

$$S \subset \mathscr{M}^2 \text{ and } \forall (K, L) \in S, (L, K) \in S. \quad (7)$$

As we stated in the introduction, the drawback of the use of this stencil for practical problems, where $F_{K,L}^{(n)}$ is computed from the resolution of a pressure equation, is that it leads to the Grid Orientation Effect. Therefore, we want to replace (6) by

$$|K| \left(u_K^{(n+1)} - u_K^{(n)}\right) + \delta t^n \sum_{L\in\widehat{\mathscr{N}}_K} \left(f(u_K^{(n)})(\widehat{F}_{K,L}^{(n)})^+ - f(u_L^{(n)})(\widehat{F}_{K,L}^{(n)})^-\right) = 0, \quad (8)$$

where the new stencil $\widehat{S}$, defined by $\widehat{S} = \{(K, L) \in \mathscr{M}^2, L \in \widehat{\mathscr{N}}_K\}$, is such that the Grid Orientation Effect is suppressed. In (8), the values of the fluxes $(\widehat{F}_{K,L}^{(n)})_{(K,L)\in\widehat{S}}$ will be set such that the two following properties hold: the flux continuity holds

$$\widehat{F}_{K,L}^{(n)} + \widehat{F}_{L,K}^{(n)} = 0,\ \forall (K, L) \in \widehat{S}, \quad (9)$$

and the balance in the control volumes is the same as that satisfied by the fluxes $(F_{K,L}^{(n)})_{(K,L)\in S}$:

$$\sum_{L,(K,L)\in\widehat{S}} \widehat{F}_{K,L}^{(n)} = \sum_{L,(K,L)\in S} F_{K,L}^{(n)},\ \forall K \in \mathscr{M}. \quad (10)$$

In view of (15), we again prescribe the symmetry property

$$\widehat{S} \subset \mathscr{M}^2 \text{ and } \forall (K,L) \in \widehat{S},\ (L,K) \in \widehat{S}. \tag{11}$$

The section 2 of this paper is devoted to the description of a method for constructing $\widehat{F}^{(n)}_{K,L}$ for a given stencil $\widehat{S}$, which ensures properties (9) and (10) (corresponding, for a given n, to (15) and (16) below). The application of this method to the case of an initial five-point pattern stencil S and of a nine-point stencil $\widehat{S}$ is detailed in Section 3. Then numerical tests show the efficiency of the method to fight the Grid Orientation Effect (section 4).

2 Construction of $\widehat{F}_{K,L}$ in the new stencil $\widehat{S}$

The method presented in this section concerns the reconstruction of the fluxes, which has to be applied to each time step. Hence, for the simplicity of notation, we drop the index n in this section. For a stencil $\widehat{S} \subset \mathscr{M}^2$ such that (11) holds and for given $(K,L) \in \mathscr{M}^2$, the set $\widehat{\mathscr{P}}_{K,L}$ of the paths from K to L following $\widehat{S}$ is defined by

$$\widehat{\mathscr{P}}_{K,L} := \left\{ P = \left\{ \begin{array}{l} (K_i, K_{i+1}), i = 1, \ldots, N-1 \text{ with } K_1 = K,\ K_N = L \\ \text{and } K_i \neq K_j \text{ for } i \neq j = 1, \cdots, N \end{array} \right\} \subset \widehat{S} \right\}. \tag{12}$$

We denote by $\sharp\widehat{\mathscr{P}}_{K,L}$ the cardinality of $\widehat{\mathscr{P}}_{K,L}$, i.e. the number of paths P from K to L following $\widehat{S}$. For any $P = \{(K_i, K_{i+1}), i = 1, \ldots, N-1\} \in \widehat{\mathscr{P}}_{K,L}$, we denote by $P^{\leftarrow}$ the inverse path from L to K following $\widehat{S}$, defined by $P^{\leftarrow} = \{(K_{i+1}, K_i), i = 1, \ldots, N-1\}$.

We may now state the following result.

Lemma 1 (New stencil and fluxes). *Let $\mathscr{M}$ be a finite set, let $S \subset \mathscr{M}^2$ be given such that (7) holds. Let $(F_{K,L})_{(K,L)\in S}$ be a family such that the property*

$$F_{K,L} + F_{L,K} = 0,\ \ \forall (K,L) \in \mathscr{M}^2$$

holds. Let $\widehat{S} \subset \mathscr{M}^2$ be given such that (11) holds and such that

$$\forall (K,L) \in S,\ \ \sharp\widehat{\mathscr{P}}_{K,L} > 0.$$

For all $(K,L) \in S$, let $(F^P_{K,L})_{P\in\widehat{\mathscr{P}}_{K,L}}$ be a family such that

$$\forall (K,L) \in S,\ \ \textstyle\sum_{P\in\widehat{\mathscr{P}}_{K,L}} F^P_{K,L} = F_{K,L},$$

satisfying the property

$$\forall (K,L) \in S,\ \forall P \in \widehat{\mathscr{P}}_{K,L},\ F^P_{K,L} + F^{P^{\leftarrow}}_{L,K} = 0. \tag{13}$$

Then the family $(\widehat{F}_{K,L})_{(K,L)\in\widehat{S}}$, defined by

$$\forall (I,J) \in \widehat{S},\ \widehat{F}_{I,J} = \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathscr{P}}_{K,L}} \xi_{I,J,P} F^P_{K,L}, \tag{14}$$

where $\xi_{I,J,P}$ *is such that* $\xi_{I,J,P} = 1$ *if* $(I,J) \in P$ *and* $\xi_{I,J,P} = 0$ *otherwise, satisfies*

$$\widehat{F}_{K,L} + \widehat{F}_{L,K} = 0,\ \forall (K,L) \in \widehat{S}, \tag{15}$$

and

$$\sum_{L,(K,L)\in\widehat{S}} \widehat{F}_{K,L} = \sum_{L,(K,L)\in S} F_{K,L},\ \forall K \in \mathscr{M}. \tag{16}$$

Proof. Firstly, using definitions, for a given $(I,J) \in \widehat{S}$, we have $(J,I) \in \widehat{S}$ and $\widehat{F}_{J,I} = \sum_{(L,K)\in S} \sum_{P\in\widehat{\mathscr{S}}_{L,K}} \xi_{J,I,P} F^P_{L,K}$. Then, thanks to the following equivalences

$$\begin{cases} (L,K) \in S \iff (K,L) \in S \\ P \in \widehat{\mathscr{S}}_{L,K} \iff P^{\leftarrow} \in \widehat{\mathscr{S}}_{K,L} \\ (J,I) \in P \iff (I,J) \in P^{\leftarrow}, \end{cases}$$

and using (13), we can rewrite $\widehat{F}_{J,I}$ as follows

$$\widehat{F}_{J,I} = - \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathscr{S}}_{K,L}} \xi_{I,J,P} F^P_{K,L} = -\widehat{F}_{I,J}\ ,$$

which proves (15).

Secondly, for a given $I \in \mathscr{M}$, by reordering the sums, we can write that

$$\sum_{J,(I,J)\in\widehat{S}} \widehat{F}_{I,J} = \sum_{J,(I,J)\in\widehat{S}} \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathscr{S}}_{K,L}} \xi_{I,J,P} F^P_{K,L} = \sum_{(K,L)\in S} \sum_{P\in\widehat{\mathscr{S}}_{K,L}} \chi_{I,P} F^P_{K,L}$$

where $\chi_{I,P} = \sum_{J,(I,J)\in\widehat{S}} \xi_{I,J,P}$ is equal to 1 if there exists $J \in \mathscr{M}$ such that $(I,J) \in P$ (therefore $I \neq L$), and to 0 otherwise. Note that, for $(K,L) \in S$ with $K \neq I$ and for $P \in \widehat{\mathscr{S}}_{K,L}$ with $\chi_{I,P} = 1$, we have $I \neq L$, $(L,K) \in S$, $P^{\leftarrow} \in \widehat{\mathscr{S}}_{L,K}$ and $\chi_{I,P^{\leftarrow}} = 1$. So, using (13), we obtain

$$\sum_{(K,L)\in S \text{ s.t. } K\neq I} \sum_{P\in\widehat{\mathscr{S}}_{K,L}} \chi_{I,P} F^P_{K,L} = 0.$$

Therefore we can write

$$\sum_{J,(I,J)\in\widehat{S}} \widehat{F}_{I,J} = \sum_{L,(I,L)\in S} \sum_{P\in\widehat{\mathscr{S}}_{I,L}} \chi_{I,P} F^P_{I,L} = \sum_{L,(I,L)\in S} \sum_{P\in\widehat{\mathscr{S}}_{I,L}} F^P_{I,L} = \sum_{L,(I,L)\in S} F_{I,L},$$

which proves (16).

3 Application to an initial five-point stencil on a structured quadrilateral mesh

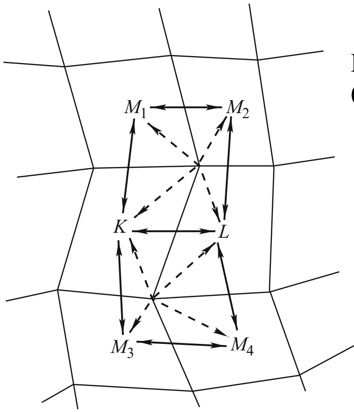
Let us assume, taking the example of a 2D situation, that the initial stencil S is a five-point stencil, defined on a regular quadrilateral mesh

$$S = \{(K, L) \in \mathscr{M}^2,\ \overline{K} \textit{ and } \overline{L} \textit{ have a common edge}\}, \tag{17}$$

and that the new stencil $\widehat{S}$ is the nine-point stencil (see the figure below), defined by

$$\widehat{S} = S \cup \{(K, L) \in \mathscr{M}^2,\ \overline{K} \textit{ and } \overline{L} \textit{ have a common point}\}. \tag{18}$$

Then we define $(F^P_{K,L})_{P \in \widehat{\mathscr{S}}_{K,L}}$, for all $P \in \widehat{\mathscr{S}}_{K,L}$ and all $(K, L) \in S$ (remark that in this case, $S \subset \widehat{S}$):



For a given $\omega > 0$ (we take the value $\omega = 0.1$ in the numerical examples), we define

$$\begin{cases} F^{P_0}_{K,L} = (1 - 4\omega) F_{K,L} \text{ for } P_0 = \{(K, L)\}, \\ F^{P_i}_{K,L} = \omega F_{K,L} \\ \text{for } P_i = \{(K, M_i), (M_i, L)\},\ \forall i = 1, \dots, 4, \\ F^{P}_{K,L} = 0 \text{ otherwise.} \end{cases}$$

Assuming that this procedure has been applied to all initial five-point connection, let us give the resulting values of $\widehat{F}_{K,L}$ deduced from (14) in two cases:

$$\begin{cases} \widehat{F}_{K,L} = (1 - 4\omega) F_{K,L} \\ \widehat{F}_{K,M_2} = \omega(F_{K,L} + F_{L,M_2} + F_{K,M_1} + F_{M_1,M_2}). \end{cases}$$

4 Numerical results

The numerical tests presented here are inspired by [7]. The domain is defined by $\Omega = [-0.5, 0.5]\text{x}[-0.5, 0.5]\text{x}[-0.15, 0.15]$. The permeability $\Lambda(x), x \in \Omega$ is equal to 1 if the distance from x to the vertical axis $0z$ is lower than 0.48, and to 10^{-3} otherwise (see Fig. 1), which ensures the confinement of the flow in the cylinder with axis $0z$ and radius 0.48. We use two Cartesian grids, the second one deduced from the first one by a rotation of angle $\theta = \frac{\pi}{6}$ with axis Oz. The number of cells in each

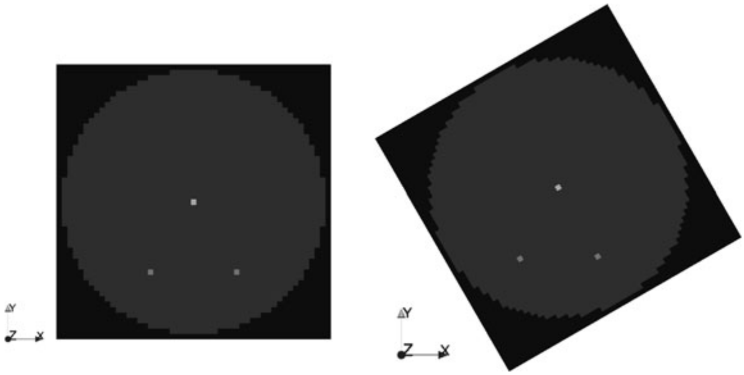
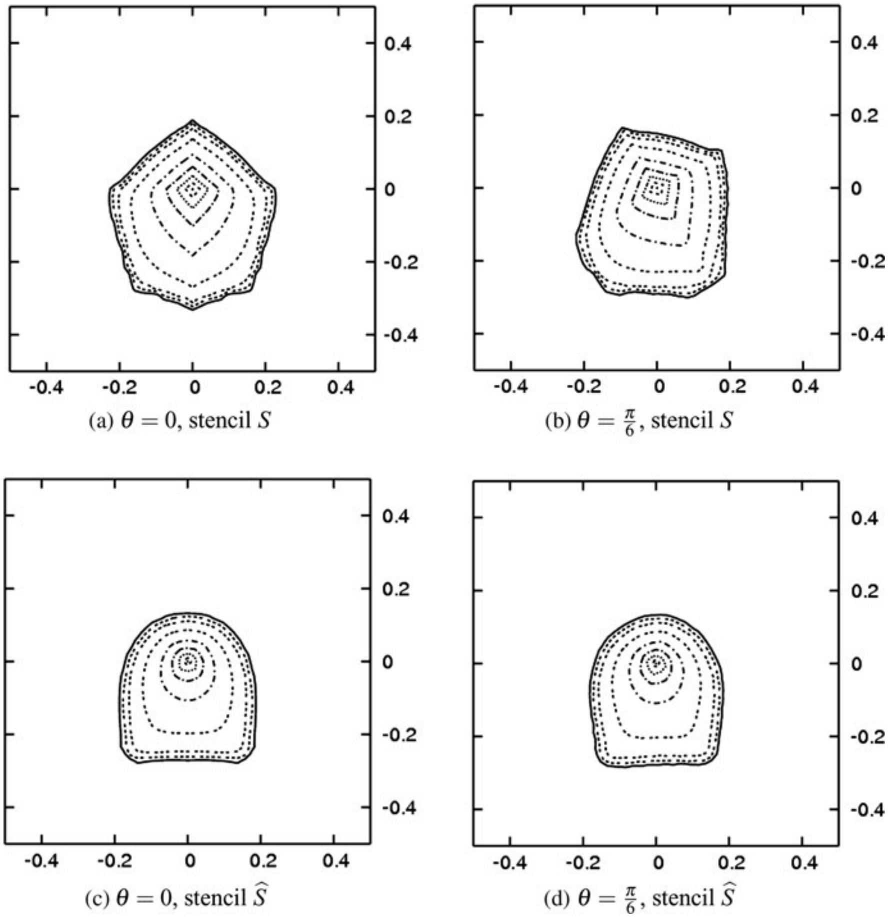


Fig. 1 The two meshes used. In grey scale, the highest permeability zone, in black the lower permeability zone. Squares indicate wells

direction (x, y, z) are $N_x = N_y = 51$ and $N_z = 3$. At the initial state, the reservoir is assumed to be saturated by the oil phase. Water is injected at the origin by an injection well. Two production wells, denoted by P_1 and P_2, are respectively located at the points $(-0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$ and $(0.3\cos\frac{\pi}{3}, -0.3\sin\frac{\pi}{3}, 0)$ (that means that the three wells are numerically taken into account as source terms in the middle layer of the mesh). The oil and water properties are respectively denoted by the index o and w. The viscosity ratio between the two phases is given by $\mu_o/\mu_w = 100$ and, the density ratio is given by $\rho_o/\rho_w = 0.8$. We use Corey-type relative permeability, $k_{r_w} = S_w^4$ and $k_{r_o} = S_o^2$. We use the method described in Sections 2 and 3, with $\omega = 0.1$ for all grid blocks which are inscribed in the cylinder (this value, also used in [6], provides the less sensitive numerical results with respect to the grid orientation). The same value for the time step is used for all the computations, which are stopped once a given quantity of water has been injected. Note that, in the mesh depicted on the right part of Fig. 1, the line (P_2, O) is the axis $0y$ of the mesh. We then see on Fig. 2 the resulting contours of the saturation. We observe that the results obtained using the method described in Sections 2 and 3 look very similar in the two grids, whereas the ones obtained using the five-point stencil are strongly distorted by the Grid Orientation Effect.

5 Conclusion

The method presented in this paper is a natural extension of the nine-point schemes defined some decades ago on regular grids. Its advantage is that it applies on the structured but not regular grids used in reservoir simulation, in association with MultiPoint Flux Approximation finite volume schemes. It demands no further modification to the standard industrial codes, since the modification are only the definition of new coefficients a_{KL}^M used in (3).



(a) $\theta = 0$, stencil S

(b) $\theta = \frac{\pi}{6}$, stencil S

(c) $\theta = 0$, stencil $\widehat{S}$

(d) $\theta = \frac{\pi}{6}$, stencil $\widehat{S}$

Fig. 2 Water saturation contours $S_w = 0.1,\ 0.2, \ldots, 1$ at the same time

References

1. Aavatsmark, I., Eigestad, G.T.: Numerical convergence of the MPFA O-method and U-method for general quadrilateral grids. Int. J. Numer. Meth. Fluids **51**, 939–961 (2006)
2. Agelas, L., Masson, R.: Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes. C. R. Math. **346**, 1007–1012 (2008)
3. Aziz, K., Ramesh, A.B., Woo, P.T.: Fourth SPE comparative solution project: comparison of steam injection simulators. J. Pet. Tech. **39**, 1576–1584 (1987)
4. Corre, B., Eymard, R., Quettier, L.: Applications of a thermal simulator to field cases, SPE ATCE (1984)
5. Dawson, C., Sun, S., Wheeler, M.F.: Compatible algorithms for coupled flow and transport. Comput. Meth. Appl. Mech. Eng. **193**, 2565–2580 (2004)
6. Eymard, R., Sonier, F.: Mathematical and Numerical Properties of Control-Volumel Finite-Element Scheme for Reservoir Simulation. SPE Reservoir Eng. **9**, 283–289 (1994)

7. Keilegavlen, E., Kozdon, J., Mallison, B.T.: Monotone Multi-dimensional Upstream Weighting on General Grids. Proceeding of ECMOR XII (2010)
8. Lipnikov, K., Moulton, J.D., Svyatskiy, D.: A multilevel multiscale mimetic (M3) method for two-phase flows in porous media. J. Comput. Phys. **14**, 6727–6753 (2008)
9. D.K. Ponting. Corner Point Geometry in reservoir simulation. *In Clarendon Press, editor*, Proc. ECMOR I, 45–65, Cambridge, 1989
10. Vinsome, P., Au, A.: One approach to the grid orientation problem in reservoir simulation. Old SPE J. **21**, 160–161 (1981)
11. Yanosik, J.L., McCracken, T.A.: A nine-point, finite-difference reservoir simulator for realistic prediction of adverse mobility ratio displacements. Old SPE J. **19**, 253–262 (1979)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Gradient Schemes for Image Processing

Robert Eymard, Angela Handlovičová, Raphaèle Herbin, Karol Mikula, and Olga Stašová

Abstract We present a gradient scheme (which happens to be similar to the MPFA finite volume O-scheme) for the approximation to the solution of the Perona-Malik model regularized by a time delay and to the solution of the nonlinear tensor anisotropic diffusion equation. Numerical examples showing properties of the method and applications in image filtering are discussed.

Keywords advection equation, semi-implicit scheme, finite volume method
MSC2010: 35L04, 65M08, 65M12

1 Introduction

A series of methods for image processing are based on the use of approximate solutions to equations of the type

$$u_t - \text{div}\ (G(u,x,t)\nabla u) = r(x,t),\ \text{for a.e. } (x,t) \in \Omega\times]0,T[\tag{1}$$

with the initial condition

$$u(x,0) = u_{\text{ini}}(x),\ \text{for a.e. } x \in \Omega, \tag{2}$$

Robert Eymard
Université Paris-Est, 5 boulevard Descartes Champs-sur-Marne F-77454 Marne la Vallée, France, e-mail: robert.eymard@univ-mlv.fr

Raphaèle Herbin
Centre de Mathmatiques et Informatique, Université de Provence, 39 rue Joliot Curie, 13453 Marseille 13, France, e-mail: raphaele.herbin@cmi.univ-mrs.fr

Angela Handlovičová, Karol Mikula, and Olga Stašová
Department of Mathematics, Slovak University of Technology, Radlinského 11, 81368 Bratislava, Slovakia, e-mail: angela@math.sk, mikula@math.sk, stasova@math.sk

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_45, © Springer-Verlag Berlin Heidelberg 2011

and the homogeneous Neumann boundary condition

$$G(u,x,t)\nabla u(x,t)\cdot \mathbf{n}_{\partial\Omega}(x) = 0, \text{ for a.e. } (x,t)\in \partial\Omega\times\mathbb{R}_+, \tag{3}$$

where Ω is an open bounded polyhedron in $\mathbb{R}^d$, $d\in\mathbb{N}^\star$, with boundary $\partial\Omega$, $T>0$, $u_{\text{ini}}\in L^2(\Omega)$, $r\in L^2(\Omega\times]0,T[)$, and G is such that, for all $v\in L^2(\Omega)$ and a.e. $(x,t)\in\Omega\times]0,T[$, $G(v,x,t)$ is a self-adjoint linear operator with eigenvalues in $(\underline{\lambda},\overline{\lambda})$ with $0<\underline{\lambda}\le\overline{\lambda}$, and $G(v,x,t)$ is continuous with respect to v and measurable with respect to x,t. In image processing applications, u_{ini} represents an original noisy image, the solution $u(x,t)$ represents its filtering which depends on scale parameter t and $d=2$ for $2D$ image filtering, $d=3$ for $3D$ image or $2D$+time movie filtering and $d=4$ for $3D$+time filtering of spatio-temporal image sequences.

The image processing methods based on approximations of equation (1) differ by definition of the function G. The first such model was proposed by Perona-Malik in 1987 [9], and nowadays, its regularization (by spatial convolution) due to Catte, Lions, Morel and Coll [2] is usually used. The regularized equation has the following form

$$\partial_t u - \nabla.(g(|\nabla G_\sigma * u|)\nabla u) = 0 \tag{4}$$

where $g(s)$ is a Lipschitz continuous decreasing function, $g(0)=1,\ 0<g(s)\to 0$ for $s\to\infty$, $G_\sigma\in C^\infty(\mathbb{R}^d)$ is a smoothing kernel, e.g. the Gauss function or mollifier with a compact support, for which $\int_{\mathbb{R}^d} G_\sigma(x)dx = 1$. Thanks to convolution, the nonlinearity in difusion term depends on the unknown function u, opposite to the original Perona-Malik equation (without convolution) where it depends on the gradient of solution. For the regularized model, the finite volume scheme were suggested and convergence and error estimates were proved in [3, 8].

Next interesting image processing model with the structure of equation (1) is the so-called nonlinear tensor anisotropic diffusion introduced by Weickert [11]. In that case, the matrix $G(u,x,t)$ represents the so-called diffusion tensor depending on the eigenvalues and eigenvectors of the (regularized) structure tensor

$$J_\rho(\nabla u_{\tilde t}) = G_\rho * (\nabla u_{\tilde t}\nabla u_{\tilde t}^{\,T}), \tag{5}$$

where

$$u_{\tilde t}(x,t) = (G_{\tilde t} * u(\cdot,t))(x) \tag{6}$$

and $G_{\tilde t}$ and G_ρ are Gaussian kernels. In computer vision, the matrix $J_\rho = \begin{pmatrix} a & b\\ b & c\end{pmatrix}$, which is symmetric and positive semidefinite, is also known as the interest operator or second moment matrix. If we denote $x=(x_2,x_2)$ we can write $a = G_\rho * \left(\frac{\partial G_{\tilde t}}{\partial x_1} * u\right)^2$, $b = G_\rho * \left(\left(\frac{\partial G_{\tilde t}}{\partial x_1} * u\right)\left(\frac{\partial G_{\tilde t}}{\partial x_2} * u\right)\right)$ and $c = G_\rho * \left(\frac{\partial G_{\tilde t}}{\partial x_2} * u\right)^2$. The orthogonal set of eigenvectors (v,w) of J_ρ corresponding to its eigenvalues

(μ_1, μ_2), $\mu_1 \ge \mu_2$, is such that the orientation of the eigenvector w, which corresponds to the smaller eigenvalue μ_2, gives the so-called coherence orientation. This orientation has the lowest fluctuations in image intensity. The diffusion tensor G in equation (1) is then designed to steer a smoothing process such that the filtering is strong along the coherence direction w and increasing with the coherence defined by difference of eigenvalues $(\mu_1 - \mu_2)^2$. To that goal, G must possess the same eigenvectors $v = (v_1, v_2)$ and $w = (-v_2, v_1)$ as the structure tensor $J_\rho(\nabla u_{\tilde{t}})$ and the eigenvalues of G can be chosen as follows

$$\kappa_1 = \alpha, \quad \alpha \in (0,1),\ \alpha \ll 1, \tag{7}$$

$$\kappa_2 = \begin{cases} \alpha, & \text{if } \mu_1 = \mu_2, \\ \alpha + (1-\alpha)\exp\left(\frac{-C}{(\mu_1-\mu_2)^2}\right), & C > 0 \quad \text{else.} \end{cases}$$

So, the matrix G is finally defined by

$$G = ABA^{-1}, \quad \text{where} \quad A = \begin{pmatrix} v_1 & -v_2 \\ v_2 & v_1 \end{pmatrix} \quad \text{and} \quad B = \begin{pmatrix} \kappa_1 & 0 \\ 0 & \kappa_2 \end{pmatrix}. \tag{8}$$

By the construction, again thanks to convolutions, we see that diffusion matrix depends nonlinearly on the solution u and it satisfies smoothness, symmetry and uniform positive definitness properties. The so-called diamond-cell finite volume schemes for the nonlinear tensor anisotropic diffusion were suggested and analyzed in [6, 7].

In this paper, we use a new class of finite volume schemes, the so-called gradient schemes [5], for solving image processing models based on equation (1). Moreover, we suggest and study numerically new type of regularization of the classical Perona-Malik approach by considering the gradient information from delayed time $t - \bar{t}$. We called this model time-delayed Perona-Malik equation, and consider (1) with $u_{\text{ini}} \in H^1(\Omega)$, and we define $u(x,t) = u_{\text{ini}}(x)$ for $x \in \Omega$ and $t < 0$ and function G is defined by

$$G(u,x,t) = \max\left(\frac{1}{1 + |\nabla u(x, t-\bar{t})|^2}, \alpha\right) \tag{9}$$

where $\bar{t}$ is a time delay and $\alpha > 0$ is a parameter. It turns out that for any $k \in \mathbb{N}$ in the time interval $]k\bar{t}, (k+1)\bar{t}[$, G is a given function of (x,t) only, which leads to a construction of efficient linear numerical scheme for this type of problems.

2 Gradient scheme approximation

In order to describe the scheme, we now introduce some notations for the space discretisation, see the Fig. 1.

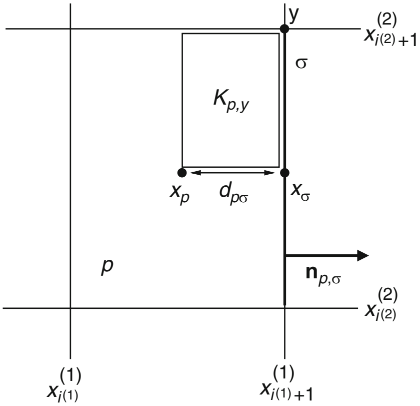


Fig. 1 Notations for the meshes

1. A rectangular discretisation of Ω is defined by the increasing sequences $a_i = x_0^{(i)} < x_1^{(i)} < \ldots < x_{n^{(i)}}^{(i)} = b_i, i = 1, \ldots, d$.
2. We denote by

$$\mathscr{M} = \left\{]x^{(1)}_{i^{(1)}}, x^{(1)}_{i^{(1)}+1}[\times \ldots \times]x^{(d)}_{i^{(d)}}, x^{(d)}_{i^{(d)}+1}[,\ 0 \le i^{(1)} < n^{(1)},\ \ldots,\ 0 \le i^{(d)} < n^{(d)} \right\}$$

   the set of the control volumes. The elements of $\mathscr{M}$ are denoted $p, q, \ldots$. We denote by $\boldsymbol{x}_p$ the centre of p. For any $p \in \mathscr{M}$, let $\partial p = \overline{p} \setminus p$ be the boundary of p; let $|p| > 0$ denote the measure of p and let h_p denote the diameter of p and $h_{\mathscr{D}}$ denote the maximum value of $(h_p)_{p \in \mathscr{M}}$.
3. We denote by $\mathscr{E}_p$ the set of all the faces of $p \in \mathscr{M}$, by $\mathscr{E}$ the union of all $\mathscr{E}_p$, and for all $\sigma \in \mathscr{E}$, we denote by $|\sigma|$ its $(d-1)$-dimensional measure. For any $\sigma \in \mathscr{E}$, we define the set $\mathscr{M}_\sigma = \{p \in \mathscr{M}, \sigma \in \mathscr{E}_p\}$ (which has therefore one or two elements), we denote by $\mathscr{E}_p$ the set of the faces of $p \in \mathscr{M}$ (it has $2\,d$ elements) and by $\boldsymbol{x}_\sigma$ the centre of σ. We then denote by $d_{p\sigma} = |\boldsymbol{x}_\sigma - \boldsymbol{x}_p|$ the orthogonal distance between $\boldsymbol{x}_p$ and $\sigma \in \mathscr{E}_p$ and by $\mathbf{n}_{p,\sigma}$ the normal vector to σ, outward to p.
4. We denote by $\mathscr{V}_p$ the set of all the vertices of $p \in \mathscr{M}$ (it has 2^d elements), by $\mathscr{V}$ the union of all $\mathscr{V}_p$, $p \in \mathscr{M}$. For $y \in \mathscr{V}_p$, we denote by $K_{p,y}$ the rectangle whose faces are parallel to those of p, and whose the set of vertices contains $\boldsymbol{x}_p$ and y. We denote by $\mathscr{V}_\sigma$ the set of all vertices of $\sigma \in \mathscr{E}$ (it has 2^{d-1} elements), and by $\mathscr{E}_{p,y}$ the set of all $\sigma \in \mathscr{E}_p$ such that $y \in \mathscr{V}_\sigma$ (it has d elements).
5. We define the set $X_{\mathscr{D}}$ of all $u = ((u_p)_{p \in \mathscr{M}}, (u_{\sigma,y})_{\sigma \in \mathscr{E}, y \in \mathscr{V}_\sigma})$, where all u_p and $u_{\sigma,y}$ are real numbers.
6. We denote, for all $u \in X_{\mathscr{D}}$, by $\Pi_{\mathscr{D}} u \in L^2(\Omega)$ the function defined by the constant value u_p a.e. in $p \in \mathscr{M}$.
7. For $u \in X_{\mathscr{D}}$, $p \in \mathscr{M}$ and $y \in \mathscr{V}_p$, we denote by

$$\nabla_{p,y} u = \frac{2}{|p|} \sum_{\sigma \in \mathscr{E}_{p,y}} |\sigma|(u_{\sigma,y} - u_p)\mathbf{n}_{p,\sigma} = \sum_{\sigma \in \mathscr{E}_{p,y}} \frac{u_{\sigma,y} - u_p}{d_{p\sigma}} \mathbf{n}_{p,\sigma}, \tag{10}$$

and by $\nabla_{\mathscr{D}} u$ the function defined a.e. on Ω by $\nabla_{p,y} u$ on $K_{p,y}$.

Let $T > 0$ be given, and $\tau > 0$ such that there exists $N_T \in \mathbb{N}$ with $T = N_T \tau$, We then define $X_{\mathscr{D},\tau} = X_{\mathscr{D}}^{N_T} = \{(u^n)_{n=1,\ldots,N_T}, u^n \in X_{\mathscr{D}}\}$, and we define the mappings $\Pi_{\mathscr{D},\tau} \ : \ X_{\mathscr{D},\tau} \to L^2(\Omega)$ and $\nabla_{\mathscr{D},\tau} \ : \ X_{\mathscr{D},\tau} \to L^2(\Omega)^d$ by

$$\Pi_{\mathscr{D},\tau} u(x,t) = \Pi_{\mathscr{D}} u^n(x), \text{ for a.e. } x \in \Omega, \ \forall t \in](n-1)\tau, n\tau], \ \forall n = 1, \ldots, N_T, \tag{11}$$

$$\nabla_{\mathscr{D},\tau} u(x,t) = \nabla_{\mathscr{D}} u^n(x), \text{ for a.e. } x \in \Omega, \ \forall t \in](n-1)\tau, n\tau], \ \forall n = 1, \ldots, N_T. \tag{12}$$

We then define the following gradient scheme approximation [5] for the discretization of Problem (1):

$$\begin{aligned} &u \in X_{\mathscr{D},\tau}, \ D_\tau u(x,t) := \frac{1}{\tau}(\Pi_{\mathscr{D}} u^1(x) - u_{\text{ini}}(x)), \text{ for a.e. } x \in \Omega, \ \forall t \in]0, \tau], \\ &D_\tau u(x,t) = \frac{1}{\tau}(\Pi_{\mathscr{D}} u^n(x) - \Pi_{\mathscr{D}} u^{n-1}(x)), \\ &\text{ for a.e. } x \in \Omega, \ \forall t \in](n-1)\tau, n\tau], \forall n = 2, \ldots, N_T, \end{aligned} \tag{13}$$

and

$$\begin{aligned} &\int_0^T \int_\Omega (D_\tau u \, \Pi_{\mathscr{D},\tau} v + G_{\mathscr{D},\tau}(\Pi_{\mathscr{D},\tau} u, x, t) \nabla_{\mathscr{D},\tau} u \cdot \nabla_{\mathscr{D},\tau} v) \, \mathrm{d}x\mathrm{d}t \\ &= \int_0^T \int_\Omega r \Pi_{\mathscr{D},\tau} v \mathrm{d}x\mathrm{d}t, \ \forall v \in X_{\mathscr{D},\tau}, \end{aligned} \tag{14}$$

where $G_{\mathscr{D},\tau}(v,x,t)$ is a suitable approximation of $G(v,x,t)$. The mathematical properties of this scheme are studied in [4].

Remark 1. The equations obtained, for a given $y \in \mathscr{V}$, defining $v \in X_{\mathscr{D}}$ for a given $\sigma \in \mathscr{E}_y$ by $v_{\sigma,y} = 1$ and all other degrees of freedom null, constitute a local invertible linear system, allowing for expressing all $(u_{\sigma,y})_{\sigma \in \mathscr{E}_y}$ with respect to all $(u_p)_{p \in \mathscr{M}}$. This leads to a nine-point stencil on rectangular meshes in 2D, 27-point stencil in 3D (this property is the basis of the MPFA O-scheme [1]).

3 Numerical experiments

3.1 Numerical study of the error for the time-delayed Perona-Malik model

We consider equation (1) in case of G defined by (9) and with a right hand side computed such that the function $u(x,y,t) = ((x^2 + y^2)/2 - (x^3 + y^3)/3)t$ is its exact solution. The domain Ω is square $[0,1] \times [0,1]$. We consider two cases, first, the time delay $\bar{t} = 0.0625$ and the overal time $T = 0.625$, and then $\bar{t} = 0.625$ and $T = 1.25$. In both cases we used coupling between space and time step $\tau \approx h^2$, where $h = \frac{1}{n}$ is length of the side of finite volume in uniform squared partition of Ω. We observe the second order convergence in L^2 and L^∞ norms of solution (denoted by E_2 and E_∞) and its gradient (denoted by EG_2 and EG_∞) in this special example, see Tables 1 and 2.

Table 1 The errors and EOC for the time-delayed Perona-Malik model, $\bar{t} = 0.0625$, $T = 0.625$

n	τ	E_2	EOC	E_∞	EOC	EG_2	EOC	EG_∞	EOC
4	0.0625	4.771e-4	-	1.022e-3	-	7.184e-3	-	1.450e-2	-
8	0.015625	1.172e-4	1.429	2.692e-4	1.925	1.707e-3	2.073	3.615e-3	2.004
16	0.00390625	2.913e-5	2.604	6.812e-5	1.982	4.213e-4	2.019	9.031e-4	2.001
32	0.0009765625	7.270e-6	2.002	1.708e-5	1.996	1.050e-4	2.004	2.257e-4	2.000
64	0.000244140625	1.815e-6	2.001	4.273e-6	1.999	2.624e-5	2.000	5.643e-5	1.999

Table 2 The errors and EOC for the time-delayed Perona-Malik model, $\bar{t} = 0.625$, $T = 1.25$

n	τ	E_2	EOC	E_∞	EOC	EG_2	EOC	EG_∞	EOC
4	0.0625	1.482e-3	-	2.237e-3	-	1.913e-2	-	2.848e-2	-
8	0.015625	3.745e-4	1.985	5.889e-4	1.925	4.651e-3	2.040	7.083e-3	2.007
16	0.00390625	9.379e-5	1.998	1.450e-4	2.022	1.155e-3	2.009	1.768e-3	2.002
32	0.0009765625	2.346e-5	1.999	3.735e-5	1.957	2.881e-4	2.003	4.419e-4	2.000
64	0.000244140625	5.865e-6	2.000	9.343e-6	1.999	7.201e-5	2.003	1.105e-4	2.000

3.2 *Image filtering by the time-delayed Perona-Malik model*

The example of image filtering by the gradient scheme applied to the time-delayed Perona-Malik equation is presented in Fig. 2. The original clean image can be seen in Fig. 2 left top. It is damaged by 40% additive noise, see Fig. 2 right top. In the bottom raws of Fig. 2 we present 5th, 10th and 20th denoising step which show the reconstruction of the original. In the last step we see the correct shape reconstruction with the keeping of the edge, with only slighly changed intensity values inside and outside quatrefoil due to diffusion. The following parameters were used in computations: $n^{(1)} = n^{(2)} = 200,\ h = 0.0125, \tau = 0.01, \bar{t} = 0.1$.

3.3 *Image filtering by the nonlinear anisotropic tensor diffusion*

In this example we present the image denoising by the nonlinear tensor diffusion and show improvement of the coherence of the line structures, which is the basic

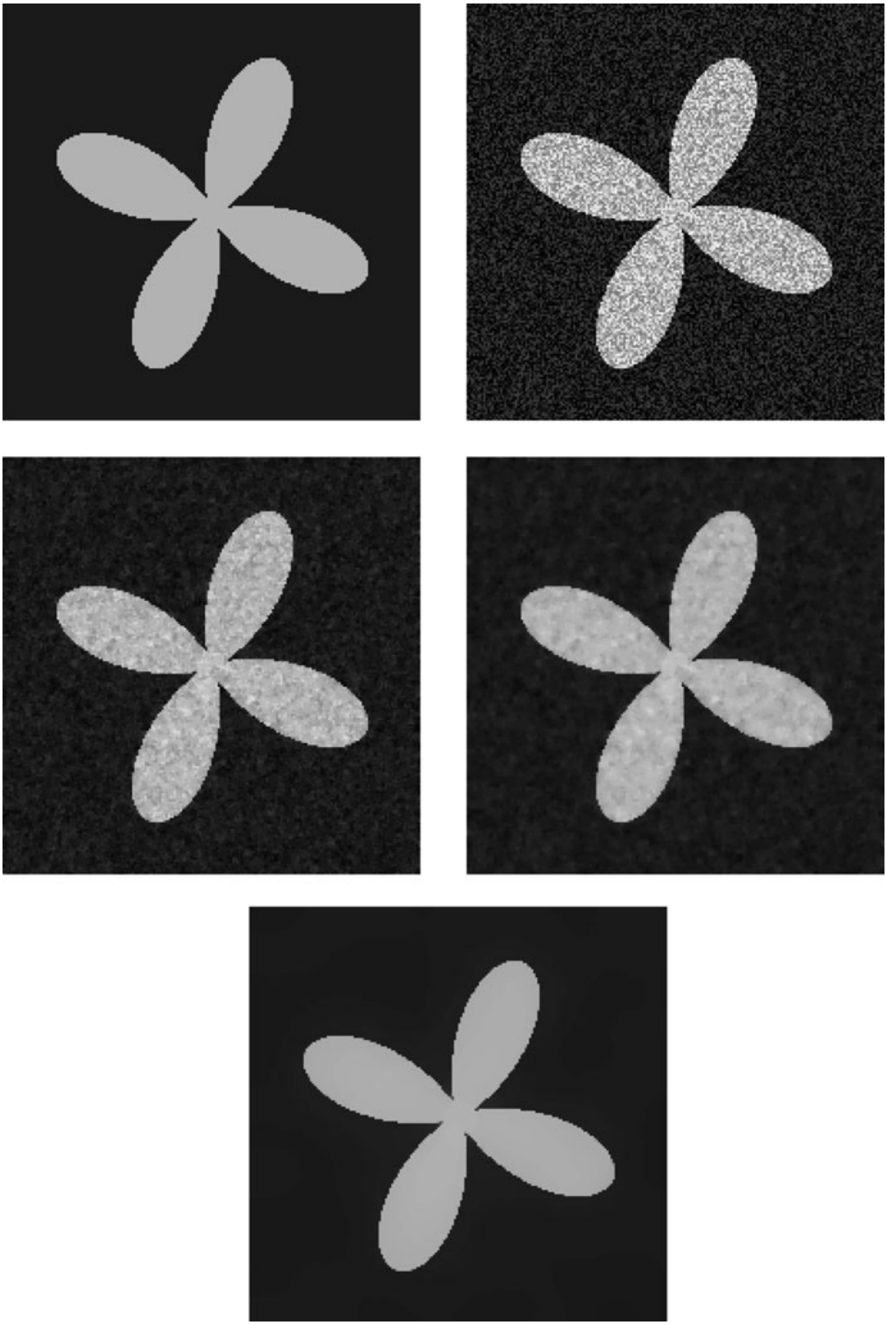


Fig. 2 Image filtering by the time-delayed Perona-Malik model: the original image (left top), the noisy image (right top) and the results after 5, 10 and 20 filtering steps

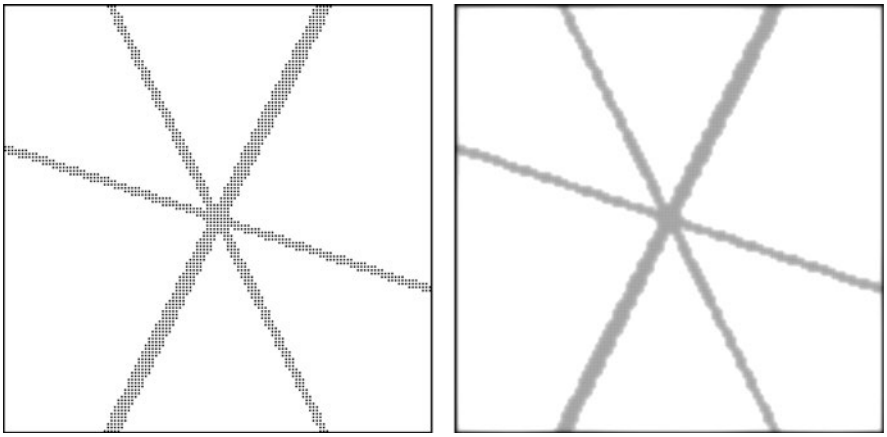


Fig. 3 The enhancement of the coherence by the nonlinear anisotropic tensor diffusion, original image (left) and the result of filtering after 100 time steps (right)

property of such models. Here, in the evaluation of diffusion matrix we use the semi-implicit approach, which means that in (6) we use the solution shifted by one time step backward, $u_{\tilde{t}}(x,t) = (G_{\tilde{t}} * u(\cdot, t-\tau))(x)$, cf. also [6]. The original image with three crackling lines can be seen in Fig. 3 left. On the right, one can see its filtering after 100 time steps which indeed enhance the coherence of those line structures. In this experiment we used the following parameters: $n^{(1)} = n^{(2)} = 250$, $h = 0.01$, $\tau = 0.0001$, $\tilde{t} = 0.0001$, $\rho = 0.01$, $\alpha = 0.001$, $C = 1$.

References

1. I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.*, 127(1):2–14, 1996.
2. F. Catté, P.L. Lions, J.M. Morel and T. Coll. Image selective smoothing and edge detection by nonlinear diffusion. *SIAM J. Numer. Anal.* 29:182–193,1992.
3. A. Handlovičová and Z. Krivá. Error estimates for finite volume scheme for Perona - Malik equation. *Acta Math. Univ. Comenianae*,74,(1):79–94, 2005.
4. R. Eymard, A. Handlovičová, R. Herbin, K. Mikula and O. Stašová. Applications of approximate gradient schemes for nonlinear parabolic equations. *in preparation*, 2011.
5. R. Eymard, R. Herbin. Gradient Scheme Approximations for Diffusion Problems. *these proceedings*, 2011.
6. O. Drblíková and K. Mikula. Convergence Analysis of Finite Volume Scheme for Nonlinear Tensor Anisotropic Diffusion in Image Processing. *SIAM Journal on Numerical Analysis*, 46 (1): 37–60,2007.
7. O. Drblíková A. Handlovičová and K. Mikula. Error estimates of the Finite Volume Scheme for the Nonlinear Tensor -Driven Anisotropic Diffusion. *Applied Numerical Mathemtaics*, 59: 2548–2570,2009.

8. K. Mikula and N. Ramarosy. Semi-implicit finite volume scheme for solving nonlinear diffusion equations in image processing. *Numerische Mathematik*, 89, (3):561–590,2001.
9. P. Perona, J. Malik. Scale space and edge detection using anisotropic diffusion. In: Proc. IEEE Computer Society Workshop on Computer Vision (1987).
10. N. J. Walkington. Algorithms for computing motion by mean curvature. *SIAM J. Numer. Anal.*, 33(6):2215–2238, 1996.
11. J. Weickert. Coherence-enhancing diffusion filtering. *Int. J. Comput. Vision*, 31: 111–127, 1999.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Gradient Scheme Approximations for Diffusion Problems

Robert Eymard and Raphaèle Herbin

Abstract We propose in this paper the definition and main properties of a family of nonconforming methods, dedicated to the approximation of diffusion problems on general meshes. We give an example of theoretical convergence result in the case of a nonlinear diffusion problem. We then review a few schemes that are part of this family, such as standard conforming and nonconforming finite element schemes, mixed finite element schemes, the SUSHI scheme, the vertex gradient approximation and particular DDFV schemes in 3D.

Keywords Gradient Scheme Approximation, diffusion problems
MSC2010: 65N08

1 Introduction

The 2D [13] and 3D [12] benchmarks for the approximation of heterogeneous and anisotropic diffusion show the large range of schemes which can be used in this setting. The aim of this paper is to propose a simple framework for nonconforming approximation methods, which can include a number of schemes such as some finite volume schemes or nonconforming finite element methods. The interest of this framework is that it provides a simple assessment of the approximation error with respect to some consistency errors. We consider the following problem, posed on an open bounded subset $\Omega \subset \mathbb{R}^d$ (where d is the space dimension), with boundary $\partial\Omega = \overline{\Omega} \setminus \Omega$:

$$\begin{cases} -\mathrm{div}(\Lambda(\overline{u})\nabla\overline{u}) = f \text{ in } \Omega, \\ \overline{u} = 0 \text{ on } \partial\Omega, \end{cases}$$

R. Eymard
Université Paris-Est, France, e-mail: robert.eymard@univ-mlv.fr

R. Herbin
Université Aix-Marseille, France, e-mail: Raphaele.Herbin@latp.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_46, © Springer-Verlag Berlin Heidelberg 2011

where $\overline{u}$ is an unknown field (temperature, pressure, ...), $f \in L^2(\Omega)$ is a volumetric source term, and $\Lambda : L^2(\Omega) \to (L^\infty(\Omega))^{d\times d}$ is a continuous operator with respect to the L^2 norm on both $L^2(\Omega)$ and $(L^\infty(\Omega))^{d\times d}$. Furthermore, we assume that for any $u \in L^2(\Omega)$ and a.e. $x \in \Omega$, the matrix $\Lambda(u)(x)$ is symmetric and the eigenvalues of $\Lambda(u)(x)$ belong to $[\underline{\lambda}, \overline{\lambda}]$, $0 < \underline{\lambda} \leq \overline{\lambda}$. Note that, if $\Lambda(u)(x)$ only depends on u through the value $u(x)$ for a.e. $x \in \Omega$, it may be defined by $\Lambda(u)(x) = \tilde{\Lambda}(u(x), x)$ where $\tilde{\Lambda}$ is a Caratheodory function.

We wish to approximate a function $\overline{u}$ solution of the weak form of the problem, that is:

$$\overline{u} \in H_0^1(\Omega) \text{ and } \forall \overline{v} \in H_0^1(\Omega), \int_\Omega \Lambda(\overline{u})\nabla\overline{u}(x)\cdot\nabla\overline{v}(x)\mathrm{d}x = \int_\Omega f(x)\overline{v}(x)\mathrm{d}x. \quad (1)$$

In order to obtain a consistent approximation of this problem, we define the following nonconforming method, called in this paper a Gradient Scheme Approximation. Defining the set $X_{\mathcal{D},0}$ of all families of discrete unknowns (which may take into account the homogeneous Dirichlet boundary condition if the discrete unknowns include approximate values at the boundary of the domain), we denote for a family of discrete unknowns $u \in X_{\mathcal{D},0}$ by $\Pi_{\mathcal{D}}u \in L^2(\Omega)$ a reconstruction of a measurable function and by $\nabla_{\mathcal{D}}u \in L^2(\Omega)^d$ a discrete approximation of its gradient.

Then Problem (1) is naturally approximated by the discrete weak formulation

$$u \in X_{\mathcal{D},0},\ \forall v \in X_{\mathcal{D},0},\ \int_\Omega \Lambda(\Pi_{\mathcal{D}}u)\nabla_{\mathcal{D}}u(x)\cdot\nabla_{\mathcal{D}}v(x)\mathrm{d}x = \int_\Omega f(x)\Pi_{\mathcal{D}}v(x)\mathrm{d}x, \quad (2)$$

which yields a numerical scheme once the set $X_{\mathcal{D},0}$ and the operators $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$ are defined. In Section 2, we provide the characterisation of the coercivity, compactness, strong and dual approximation properties for given $X_{\mathcal{D},0}$, $\Pi_{\mathcal{D}}$ and $\nabla_{\mathcal{D}}$. In the case where Λ does not depend on u and where the Gradient Scheme Approximation checks suitable properties in terms of coercivity, strong and dual approximation, then Scheme (2) may be shown to converge to (1). In the general case of an operator $\Lambda(u)$, a requirement on the compactness property is then needed for proving the convergence of the scheme (2). We then review in Section 3 a few known schemes which can be seen as Gradient Scheme Approximations.

2 Gradient Scheme Approximation

2.1 Definition and properties

Definition 1 (Gradient scheme discretization). Let Ω be an open bounded domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$. A gradient scheme discretization $\mathcal{D}$ is defined by $\mathcal{D} = (X_{\mathcal{D},0}, h_{\mathcal{D}}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$, where:

1. the set of discrete unknowns $X_{\mathscr{D},0}$ is a finite dimensional vector space on $\mathbb{R}$,
2. the space step $h_{\mathscr{D}} \in (0, +\infty)$ is a positive real number,
3. the mapping $\Pi_{\mathscr{D}} : X_{\mathscr{D},0} \to L^2(\Omega)$ is the reconstruction of the approximate function (for any $u \in X_{\mathscr{D},0}$, $\Pi_{\mathscr{D}} u$ is prolonged by 0 outside Ω),
4. the mapping $\nabla_{\mathscr{D}} : X_{\mathscr{D},0} \to L^2(\Omega)^d$ is the reconstruction of the gradient of the function (for any $u \in X_{\mathscr{D},0}$, $\nabla_{\mathscr{D}} u$ is prolonged by 0 outside Ω); accounting for the homogeneous Dirichlet boundary condition, it must be chosen such that $\| \cdot \|_{\mathscr{D}} = \|\nabla_{\mathscr{D}} \cdot\|_{L^2(\Omega)^d}$ is a norm on $X_{\mathscr{D},0}$.

Remark 1. In the case of the homogeneous Neumann boundary condition, one requires that
$\| \cdot \|_{\mathscr{D}} = ((\int_\Omega \Pi_{\mathscr{D}} \cdot \mathrm{d}x)^2 + \|\nabla_{\mathscr{D}} \cdot\|^2_{L^2(\Omega)^d})^{1/2}$ be a norm on $X_{\mathscr{D},0}$.

Then the **coercivity** of the discretization is measured through the norm $C_{\mathscr{D}}$ of the linear mapping $\Pi_{\mathscr{D}}$, defined by

$$C_{\mathscr{D}} = \max_{v \in X_{\mathscr{D},0} \setminus \{0\}} \frac{\|\Pi_{\mathscr{D}} v\|_{L^2(\Omega)}}{\|v\|_{\mathscr{D}}}. \tag{3}$$

Note that, in the homogeneous Dirichlet boundary condition framework, (3) yields the following "discrete Poincaré" inequality:

$$\|\Pi_{\mathscr{D}} v\|_{L^2(\Omega)} \le C_{\mathscr{D}} \|\nabla_{\mathscr{D}} v\|_{L^2(\Omega)^d}, \ \forall v \in X_{\mathscr{D},0}.$$

The **consistency** of the discretization is measured through the interpolation error function $S_{\mathscr{D}} : H_0^1(\Omega) \to [0, +\infty)$, defined by

$$S_{\mathscr{D}}(\varphi) = \min_{v \in X_{\mathscr{D},0}} \left(\|\Pi_{\mathscr{D}} v - \varphi\|^2_{L^2(\Omega)} + \|\nabla_{\mathscr{D}} v - \nabla\varphi\|^2_{L^2(\Omega)^d} \right)^{\frac{1}{2}}, \ \forall \varphi \in H_0^1(\Omega), \tag{4}$$

The **dual consistency** of the discretization is measured through the conformity error function $W_{\mathscr{D}}$: $H_{\mathrm{div}}(\Omega) \to [0, +\infty)$, defined by

$$\begin{aligned} W_{\mathscr{D}}(\boldsymbol{\varphi}) = & \max_{v \in X_{\mathscr{D},0} \setminus \{0\}} \frac{\int_\Omega (\nabla_{\mathscr{D}} v(x) \cdot \boldsymbol{\varphi}(x) + \Pi_{\mathscr{D}} v(x) \mathrm{div} \boldsymbol{\varphi}(x))\,\mathrm{d}x}{\|v\|_{\mathscr{D}}}, \\ & \forall \boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega). \end{aligned} \tag{5}$$

The **compactness** of the discretization is measured through the function $T_{\mathscr{D}} : \mathbb{R}^d \to \mathbb{R}^+$, defined by

$$T_{\mathscr{D}}(\xi) = \max_{v \in X_{\mathscr{D}} \setminus \{0\}} \frac{\|\Pi_{\mathscr{D}} v(\cdot + \xi) - \Pi_{\mathscr{D}} v\|_{L^2(\mathbb{R}^d)}}{\|v\|_{\mathscr{D}}}, \ \forall \xi \in \mathbb{R}^d. \tag{6}$$

We may remark that the function $\Pi_{\mathscr{D}} u$ lies in a finite dimensional subspace of $L^2(\Omega)$ and therefore $T_{\mathscr{D}}$ is such that $\lim_{|\xi|\to 0} T_{\mathscr{D}}(\xi) = 0$ and that, for any $|\xi| \in \mathbb{R}^d$, $T_{\mathscr{D}}(\xi) \le 2C_{\mathscr{D}}$, showing the link between compactness and coercivity.

If $\mathscr{D} = (X_{\mathscr{D},0}, h_{\mathscr{D}}, \Pi_{\mathscr{D}}, \nabla_{\mathscr{D}})$ is an approximate gradient discretization, we shall say that (2) is a Gradient Scheme Approximation.

In [11] the following results are proved:

Lemma 1 (Control of the approximation error, linear case). *Let Ω be a bounded open domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$, let $f \in L^2(\Omega)$ and let $\Lambda \in (L^\infty(\Omega))^{d\times d}$ be such that, for a.e. $x \in \Omega$, the matrix $\Lambda(x)$ is symmetric and the eigenvalues of $\Lambda(x)$ belong to $[\underline{\lambda}, \overline{\lambda}]$, $0 < \underline{\lambda} \le \overline{\lambda}$. Let $\overline{u} \in H_0^1(\Omega)$ be the solution of (1) (remark that since $f \in L^2(\Omega)$, one has $\Lambda\nabla\overline{u} \in H_{\mathrm{div}}(\Omega)$).*

Let $\mathscr{D}$ be an approximate gradient discretization in the sense of Definition 1. Then there exists one and only one $u_{\mathscr{D}} \in X_{\mathscr{D},0}$, solution to the Gradient Scheme Approximation (2), which moreover satisfies the following inequalities:

$$\|\nabla\overline{u} - \nabla_{\mathscr{D}} u_{\mathscr{D}}\|_{L^2(\Omega)^d} \le \frac{1}{\underline{\lambda}}(W_{\mathscr{D}}(\Lambda\nabla\overline{u}) + (\overline{\lambda} + \underline{\lambda})S_{\mathscr{D}}(\overline{u})), \tag{7}$$

and

$$\|\overline{u} - \Pi_{\mathscr{D}} u_{\mathscr{D}}\|_{L^2(\Omega)} \le \frac{1}{\underline{\lambda}}(C_{\mathscr{D}} W_{\mathscr{D}}(\Lambda\nabla\overline{u}) + (C_{\mathscr{D}}\overline{\lambda} + \underline{\lambda})S_{\mathscr{D}}(\overline{u})). \tag{8}$$

Corollary 1 (Convergence, linear case). *Under the assumptions of Lemma 1, let $\mathscr{F}$ be a family of gradient discretizations in the sense of Definition 1, which satisfies the following assumptions:*

(P1) there exists $C_P \in \mathbb{R}$ such that $C_{\mathscr{D}} \le C_P$ for any $\mathscr{D} \in \mathscr{F}$,
(P2) for all $\varphi \in H_0^1(\Omega)$ and $\mathscr{D} \in \mathscr{F}$, $S_{\mathscr{D}}(\varphi)$ tends to 0 as $h_{\mathscr{D}} \to 0$,
(P3) for all $\boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega)$ and $\mathscr{D} \in \mathscr{F}$, $W_{\mathscr{D}}(\boldsymbol{\varphi})$ tends to 0 as $h_{\mathscr{D}} \to 0$.

For $\mathscr{D} \in \mathscr{F}$, let $u_{\mathscr{D}} \in X_{\mathscr{D},0}$ be the solution to the Gradient Scheme Approximation (2), then $\Pi_{\mathscr{D}} u_{\mathscr{D}}$ converges to $\overline{u}$ in $L^2(\Omega)$ and $\nabla_{\mathscr{D}} u_{\mathscr{D}}$ converges to $\nabla\overline{u}$ in $L^2(\Omega)^d$ as $h_{\mathscr{D}} \to 0$.

Lemma 2. *Let Ω be a bounded open domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$. Let $\mathscr{F}$ be a family of approximate gradient discretizations in the sense of Definition 1. Then, for any dense subspace $\mathscr{R}$ of $H_0^1(\Omega)$, the two properties:*

$$\lim_{h_{\mathscr{D}}\to 0} S_{\mathscr{D}}(\varphi) = 0, \ \forall \varphi \in \mathscr{R}, \tag{9}$$

and

$$\lim_{h_{\mathscr{D}}\to 0} S_{\mathscr{D}}(u) = 0, \ \forall u \in H_0^1(\Omega), \tag{10}$$

are equivalent. Furthermore, if there exists $C_P > 0$ such that the following uniform discrete Poincaré inequality holds:

$$C_{\mathscr{D}} \leq C_P, \ \forall \mathscr{D} \in \mathscr{F}, \tag{11}$$

then for any dense subspace $\mathscr{S}$ of $H_{\mathrm{div}}(\Omega)$, the two properties:

$$\lim_{h_{\mathscr{D}} \to 0} W_{\mathscr{D}}(\boldsymbol{\varphi}) = 0, \ \forall \boldsymbol{\varphi} \in \mathscr{S}, \tag{12}$$

and

$$\lim_{h_{\mathscr{D}} \to 0} W_{\mathscr{D}}(\boldsymbol{U}) = 0, \ \forall \boldsymbol{U} \in H_{\mathrm{div}}(\Omega), \tag{13}$$

are equivalent.

Let us now prove a convergence result in the nonlinear case.

Lemma 3 (Convergence of the scheme, nonlinear case). *Let Ω be a bounded open domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$, let $f \in L^2(\Omega)$ and let $\Lambda : L^2(\Omega) \to (L^\infty(\Omega))^{d \times d}$ be a continuous operator with respect to the L^2 norm on both $L^2(\Omega)$ and $(L^\infty(\Omega))^{d \times d}$; furthermore, we assume that for any $u \in L^2(\Omega)$ and a.e. $x \in \Omega$, the matrix $\Lambda(u)(x)$ is symmetric and the eigenvalues of $\Lambda(u)(x)$ belong to $[\underline{\lambda}, \overline{\lambda}]$, $0 < \underline{\lambda} \leq \overline{\lambda}$. Let $\mathscr{F}$ be a family of gradient discretizations in the sense of Definition 1, which satisfies properties (P1), (P2) and (P3) of Corollary 1, and moreover satisfies*

(P4) the family of functions $(T_{\mathscr{D}})_{\mathscr{D} \in \mathscr{F}}$ (which is bounded by $2C_P$ thanks to (P1)) is such that

$$\lim_{|\xi| \to 0} \sup_{\mathscr{D} \in \mathscr{F}} T_{\mathscr{D}}(\xi) = 0. \tag{14}$$

Then, for any $\mathscr{D} \in \mathscr{F}$, there exists at least one $u_{\mathscr{D}} \in X_{\mathscr{D},0}$, solution to the Gradient Scheme Approximation (2). Moreover, for a sequence $(\mathscr{D}_n)_{n \in \mathbb{N}}$ of elements of $\mathscr{F}$ such that $h_{\mathscr{D}_n} \to 0$ as $n \to \infty$, there exists $\overline{u} \in H_0^1(\Omega)$, solution to (1) and a subsequence of $(\mathscr{D}_n)_{n \in \mathbb{N}}$, again denoted $(\mathscr{D}_n)_{n \in \mathbb{N}}$, such that then $\Pi_{\mathscr{D}_n} u_{\mathscr{D}_n}$ converges to $\overline{u}$ in $L^2(\Omega)$ and $\nabla_{\mathscr{D}_n} u_{\mathscr{D}_n}$ converges to $\nabla \overline{u}$ in $L^2(\Omega)^d$ as $n \to \infty$.

Remark 2. It is possible to find a family of gradient discretizations in the sense of Definition 1, which only satisfies (P1), (P2), (P3) and not (P4).

Proof. The existence of a solution to (2) is an immediate consequence of the topological degree argument and of the estimate

$$\underline{\lambda} \|\nabla_{\mathscr{D}} u\|_{L^2(\Omega)^d} \leq \|f\|_{L^2(\Omega)} C_{\mathscr{D}}. \tag{15}$$

We then define, for all $n \in \mathbb{N}$, a solution u_n to (2) for $\mathscr{D} = \mathscr{D}_n$. Thanks to properties (P1) and (P4), to (15) and to the Kolmogorov theorem, the family $(\Pi_{\mathscr{D}_n} u_n)_{n \in \mathbb{N}}$ is relatively compact in $L^2(\Omega)$. Then there exists $\overline{u} \in L^2(\Omega)$ and a subsequence of $(\mathscr{D}_n)_{n \in \mathbb{N}}$, again denoted $(\mathscr{D}_n)_{n \in \mathbb{N}}$, such that $\Pi_{\mathscr{D}_n} u_n$ converges to $\overline{u}$ in $L^2(\Omega)$. Extracting again a subsequence, we get that $\nabla_{\mathscr{D}_n} u_n$ converges weakly in $L^2(\mathbb{R}^d)$ to some function $G \in L^2(\mathbb{R}^d)$. Using (P3), we get that $G = \nabla \overline{u}$, hence showing that $\overline{u} \in H_0^1(\Omega)$. Then, for all $v \in H_0^1(\Omega)$, denoting by $v_n \in X_{\mathscr{D}_n,0}$ the element minimising $S_{\mathscr{D}_n}(v)$, we get from (P2) that $\nabla_{\mathscr{D}_n} v_n$ converges in $L^2(\Omega)^d$ to ∇v. It is then possible, by weak/strong limit, to pass to the limit as $n \to \infty$ in (2); thus, $\overline{u}$ is solution to (1). Letting $v = u_n$ in (2) and $\overline{v} = \overline{u}$ in (1) shows that

$$\lim_{n\to\infty} \int_\Omega \Lambda(\Pi_{\mathscr{D}_n} u_n) \nabla_{\mathscr{D}_n} u_n(x) \cdot \nabla_{\mathscr{D}_n} u_n(x) \mathrm{d}x = \int_\Omega \Lambda(\overline{u}) \nabla \overline{u}(x) \cdot \nabla \overline{u}(x) \mathrm{d}x,$$

and therefore:

$$\lim_{n\to\infty} \int_\Omega \Lambda(\Pi_{\mathscr{D}_n} u_n)(\nabla_{\mathscr{D}_n} u_n(x) - \nabla \overline{u}(x)) \cdot (\nabla_{\mathscr{D}_n} u_n(x) - \nabla \overline{u}(x)) \mathrm{d}x = 0,$$

hence proving the convergence of $\nabla_{\mathscr{D}_n} u_{\mathscr{D}_n}$ to $\nabla \overline{u}$ in $L^2(\Omega)^d$ as $n \to \infty$.

3 Application to some schemes

Let us notice that **standard conforming finite element** discretizations may be seen as Gradient Scheme Approximations. If $V_h \subset H_0^1(\Omega)$ is the usual conforming finite element space spanned by the basis functions $\varphi_1, \ldots \varphi_N$, the space $X_{\mathscr{D},0}$ is then $\mathbb{R}^N$ and for $u = (u_1, \ldots, u_N) \in X_{\mathscr{D},0}$, $\Pi_{\mathscr{D}} u = \sum_{i=1}^N u_i \varphi_i$, and $\nabla_{\mathscr{D}} u = \sum_{i=1}^N u_i \nabla \varphi_i = \nabla \Pi_{\mathscr{D}} u$. Hence

$$W_{\mathscr{D}}(\boldsymbol{\varphi}) = 0 \text{ for all } \boldsymbol{\varphi} \in H_{\mathrm{div}}(\Omega). \tag{16}$$

Note that in fact, an approximate gradient discretization is conforming if and only if (16) holds. The compactness property (P4) is satisfied since in this conforming case, we have $T_{\mathscr{D}}(\xi) = |\xi|$.

Let us now turn to the case of the **non conforming P1 finite element discretization** on conforming simplicial meshes. In this case, the basis functions of the finite element space V_h are associated with the N internal faces of the mesh, and V_h is spanned by the basis functions $\varphi_1, \ldots \varphi_N$ which are piecewise affine and continuous at the barycentre of the faces. In this case, the space $X_{\mathscr{D},0}$ is then again $\mathbb{R}^N$ and for $u = (u_1, \ldots, u_N) \in X_{\mathscr{D},0}$, $\Pi_{\mathscr{D}} u = \sum_{i=1}^N u_i \varphi_i$, but $\nabla_{\mathscr{D}} u$ cannot be defined as in the conformal case; it is only piecewise defined as the gradient of $\Pi_{\mathscr{D}} u$. It is possible, under some geometrical conditions on the mesh (see e.g. [9]) to get from classical results that for all $\boldsymbol{\varphi} \in (C^1(\mathbb{R}^d))^d$, $W_{\mathscr{D}}(\boldsymbol{\varphi}) \le h_{\mathscr{D}} C_{\boldsymbol{\varphi}}$. Property (P4) is also classically shown.

In fact, the **mixed finite element discretization** may also be seen as a Gradient Scheme Approximation. We denote by $(\varphi_i)_{i=1,\ldots,N} \subset L^2(\Omega)$ the basis functions for the approximation of $\overline{u}$, and $(\boldsymbol{\varphi}_i)_{i=1,\ldots,M} \subset H_{\mathrm{div}}(\Omega)$ the basis functions for the approximation of $\Lambda \nabla \overline{u}$. We then define $X_{\mathscr{D},0} \subset \mathbb{R}^{N+M}$ as the set of all families $u = \big((u_1, \ldots, u_N), (q_1, \ldots, q_M)\big)$ such that, denoting $\Pi_{\mathscr{D}} u = \sum_{i=1}^N u_i \varphi_i$ and $\nabla_{\mathscr{D}} u = \sum_{i=1}^M q_i \Lambda^{-}1 \boldsymbol{\varphi}_i$, the relation $\int_\Omega \big(\boldsymbol{\varphi}_j(x) \cdot \nabla_{\mathscr{D}} u(x) + \Pi_{\mathscr{D}} u(x) \mathrm{div} \boldsymbol{\varphi}_j(x)\big) \mathrm{d}x = 0$ holds for all $j = 1, \dots, M$. Then the mixed finite element scheme may be written as (2). The property (P3) is a direct consequence of the imposed relation between $\Pi_{\mathscr{D}} u$ and $\nabla_{\mathscr{D}} u$. The property (P1) is the consequence of the so-called "infsup" condition, and the properties (P2) and (P4) may be shown to be satisfied.

The **SUSHI scheme** [10], as well as the **vertex gradient scheme** [11] are explicitly defined through the space $X_{\mathscr{D},0}$, the reconstruction operator $\Pi_{\mathscr{D}}$ and the discrete gradient $\nabla_{\mathscr{D}}$. The compactness property (P4) is detailed in the appendix of [10]. The study of $S_{\mathscr{D}}$ is also detailed in [10], and that of $W_{\mathscr{D}}$ in [11]. Note that the SUSHI scheme is part of the Mimetic Mixed Hybrid family [8]; however, it does not seem easy to write a general mimetic scheme as a Gradient Scheme Approximation, because the stabilisation term which is needed for the coercivity of the scheme (except in its SUSHI implementation) be included in the gradient term, and therefore the scheme cannot be written under the form (2).

The **DDFV scheme**, see [3, 7, 14] for the two dimensional case and [1, 2, 4–6, 15, 16] for the three dimensional case may also be seen, in some cases, as a Gradient Scheme Approximation. Consider the case where the domain Ω is the union of octahedra which are the so-called diamond cells (such a cell is depicted in Figure 1). Octahedral meshes may be obtained from general hexahedral meshes by introducing an internal point to each hexahedron. We show in Fig. 1 a locally refined face of hexahedral cell where we depict a octahedron constructed with an internal point of the cell and the barycentre of the four points of a face. With such a construction, we can easily take into account boundary conditions and heterogeneous media (each octahedron is homogeneous). The unknown at the centre of the internal faces (point B on the right side of Figure 1), may be easily eliminated. Let us define the space $X_{\mathscr{D},0}$ as $X_{\mathscr{D},0} = \{(u_s)_{s \in \mathscr{V}}, u_s = 0, \ \forall s \in \mathscr{V}_{\text{ext}}\}$, where $\mathscr{V}$ denotes the set of vertices of the octahedral mesh $\mathscr{M}$ and $\mathscr{V}_{\text{ext}}$ denotes the set of the elements of $\mathscr{V}$ located on the boundary of Ω. Referring to Fig. 1, we define a discrete piecewise constant

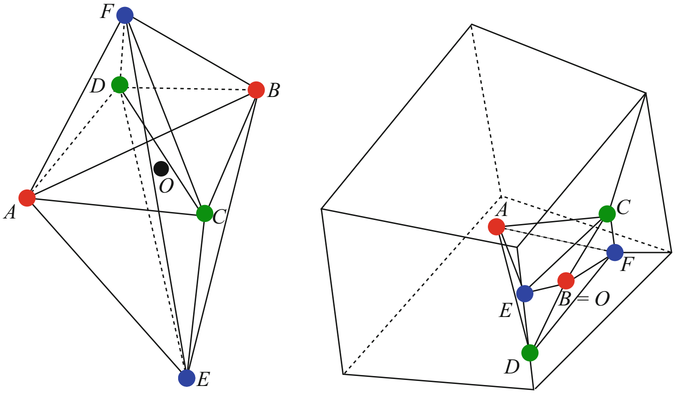


Fig. 1 Left: A generic octahedral cell for the DDFV scheme - Right: An example of construction of an octahedron from a locally refined face of a hexahedron

gradient by its value on the octahedron $K \in \mathscr{M}$:

$$\nabla_{\mathscr{D}} u(x) = \frac{1}{\Delta_K}\Big((u_B - u_A)\overrightarrow{CD} \wedge \overrightarrow{EF} + (u_D - u_C)\overrightarrow{EF} \wedge \overrightarrow{AB} \\ +(u_F - u_E)\overrightarrow{AB} \wedge \overrightarrow{CD}\Big), \qquad \forall x \in K, \tag{17}$$

where $\Delta_K = \mathrm{Det}(\overrightarrow{AB}, \overrightarrow{CD}, \overrightarrow{EF})$. Let O be a well chosen point in $\overline{K}$, for instance the barycentre of the six vertices A, B, C, D, E and F. Taking the example of the vertex F, we denote by σ_{EF} the union of the four triangles OAC, OCB, OBD and ODA, and we denote by $V_{K,F}$ the subset of K of all points which are on the same side of σ_{EF} as F. We proceed similarly for the five other vertices. The reconstruction operator is then defined for $x \in K$ by:

$$\Pi_{\mathscr{D}} u(x) = \tfrac{1}{3}\Big(u_A 1_{V_{K,A}}(x) + u_B 1_{V_{K,B}}(x) + u_C 1_{V_{K,C}}(x) \\ +u_D 1_{V_{K,D}}(x) + u_E 1_{V_{K,E}}(x) + u_F 1_{V_{K,F}}(x)\Big).$$

With these definitions, (2) is identical to a DDFV scheme [4] formulated on three grids.

Acknowledgements Work supported by Groupement MOMAS and ANR VFSitCom.

References

1. B. Andreianov, M. Bendahmane, and K. Karlsen. A gradient reconstruction formula for finite-volume schemes and discrete duality. In *Finite volumes for complex applications V*, pages 161–168. ISTE, London, 2008.
2. B. Andreianov, M. Bendahmane, K. H. Karlsen, and C. Pierre. Convergence of discrete duality finite volume schemes for the cardiac bidomain model. see http://hal.archives-ouvertes.fr/hal-00526047/PDF/ABKP-submitted.pdf.
3. F. Boyer and F. Hubert. Finite volume method for 2d linear and nonlinear elliptic problems with discontinuities. *SIAM Journal on Numerical Analysis*, 46(6):3032–3070, 2008.
4. Y. Coudière and F. Hubert. A 3D discrete duality finite volume method for nonlinear elliptic equations. 35J65, 65N15, 74S10.
5. Y. Coudière, C. Pierre, O. Rousseau, and R. Turpault. 2D/3D discrete duality finite volume scheme (DDFV) applied to ECG simulation. A DDFV scheme for anisotropic and heterogeneous elliptic equations, application to a bio-mathematics problem: electrocardiogram simulation. In *Finite volumes for complex applications V*, pages 313–320. ISTE, London, 2008.
6. Y. Coudière, C. Pierre, O. Rousseau, and R. Turpault. A 2D/3D discrete duality finite volume scheme. Application to ECG simulation. *Int. J. Finite Vol.*, 6(1):24, 2009.
7. K. Domelevo and P. Omnes. A finite volume method for the laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, 39(6):1203–1249, 2005.
8. J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.
9. A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

10. R. Eymard, T. Gallouët, and R. Herbin. 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.
11. R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3d schemes for diffusive flows in porous media. *submitted*.
12. R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klöfkorn, and G. Manzini. 3d benchmark on discretization schemes for anisotropic diffusion problem on general grids. In *Finite volumes for complex applications VI*. SPringer, 2011.
13. R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids for anisotropic heterogeneous diffusion problems. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 659–692. Wiley, 2008.
14. F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, 192(16-18):1939–1959, 2003.
15. F. Hermeline. Approximation of 2-D and 3-D diffusion operators with variable full tensor coefficients on arbitrary meshes. *Comput. Methods Appl. Mech. Engrg.*, 196(21-24):2497–2526, 2007.
16. F. Hermeline. A finite volume method for approximating 3D diffusion operators on general meshes. *J. Comput. Phys.*, 228(16):5763–5786, 2009.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Cartesian Grid Method for the Compressible Euler Equations

M. Asif Farooq and B. Müller

Abstract The accuracy of the Cartesian grid method has been investigated for the 2D compressible Euler equations. We impose wall boundary conditions at ghost points by interpolation or extrapolation at the corresponding mirror points either linearly or quadratically. We find that linear or quadratic interpolation does not affect the accuracy of our node-centered finite volume method. Two different ghost point treatments have been compared.

Keywords Cartesian Grid Method, Ghost Point Treatment, Compressible Euler Equations, Conservation Laws, Oblique Shock Wave
MSC2010: 76J20, 76L05, 35L03, 35L65, 76N15

1 Introduction

The Cartesian grid method has recently become one of the widely used methods in CFD [1–7]. This is due to its simplicity, faster grid generation, simpler programming, lower storage requirements, lower operation count, and easier post processing compared to body fitted structured and unstructured grid methods. The Cartesian grid method is also advantageous in constructing higher order methods. Problems occur at the boundary, when this method is applied to complex domains [8]. When the Cartesian grid method is applied at curved boundaries the cells at the boundaries are not rectangular and these cut-cells create problems for the scheme to be implemented. The time step restriction problem caused by small cut-cells can be solved by merging those cut-cells with neighboring cells [7].

M. Asif Farooq and B. Müller
Department of Energy and Process Engineering, Norwegian University of Science and Technology (NTNU), 7491, Trondheim, Norway, e-mail: asif.m.farooq@ntnu.no, bernhard.muller@ntnu.no

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_47, © Springer-Verlag Berlin Heidelberg 2011

Cut cells are avoided altogether by ghost point treatment at the boundary. In this method symmetry conditions with respect to the boundary are imposed at ghost points in the solid adjacent to the boundary [9]. However, conservativity is lost in this process. Nevertheless, the simplicity of the ghost point treatment has motivated us to use that approach instead of the more complicated cut-cells.

The Cartesian grid method is also called immersed boundary method, in particular when it is applied to the incompressible Navier-Stokes equations. Often the effect of solid boundaries cutting a Cartesian grid has been modelled by a force term in the incompressible momentum equations [10]. Since this approach is not so practical for compressible flow due to the sensitive coupling of all flow variables, it has not been used for compressible flow simulation except for [2, 11]. Instead, the effect of the tangency or slip condition at solid boundaries for inviscid compressible flow is used in the Cartesian grid method to determine the flow variables in ghost cells or at ghost points near solid boundaries [9, 12–16]. In the ghost point treatment we divide our domain into three types of points: fluid, ghost and solid points. For first and second order schemes the methods require one and two ghost points, respectively. Solid and ghost points are flagged inactive.

In this paper we employ the Sjögreen and Petersson ghost point treatment [16], while in [17] we applied a simplified ghost point treatment for the 2D compressible Euler equations. A comparison of these two ghost point treatments at the boundary is also presented in this paper. Sjögreen and Petersson [16] used linear interpolation at the boundary, while we use linear and quadratic interpolation at the boundary. We impose the wall boundary conditions at the ghost points by interpolating the numerical solution at their mirror points with respect to the wall in the fluid domain and mirroring the interpolated values to ensure reflective boundary conditions. If the numerical solution at a mirror point cannot be approximated by interpolation, we employ extrapolation. We employ the local Lax-Friedrichs (lLF) method for the spatial discretization. To increase the accuracy we apply the MUSCL approach with the minmod limiter. For time integration we use the first order explicit Euler and the third order TVD Runge-Kutta (RK3) methods. As a test case, we consider supersonic flow over a wedge and solve the 2D compressible Euler equations by time stepping for the steady state.

The paper is organized as follows. In Section 2 we present the governing equations, i.e. the 2D compressible Euler equations. In Section 3 we outline the boundary conditions. In Section 4 we explain the ghost point treatment at the embedded boundary. In section 5 we present results and discussions. Conclusions are given in section 6.

2 Compressible Euler Equations

The 2D compressible Euler equations serve as a model for a 2D nonlinear hyperbolic system. In conservative form the 2D compressible Euler equations read

$$\frac{\partial U}{\partial t} + \frac{\partial F}{\partial x} + \frac{\partial G}{\partial y} = 0, \tag{1}$$

where

$$U = \begin{bmatrix} \rho \\ \rho u \\ \rho v \\ \rho E \end{bmatrix}, F = \begin{bmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (\rho E + p)u \end{bmatrix}, G = \begin{bmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (\rho E + p)v \end{bmatrix}, \tag{2}$$

with ρ , u, v, E, and p are density, velocity components in x and y-directions, total energy per unit mass and pressure, respectively.

For perfect gas we have the following relation

$$p = (\gamma - 1)(\rho E - \frac{1}{2}\rho(u^2 + v^2)), \tag{3}$$

where γ is the ratio of specific heats. We consider γ = 1.4 for air.

3 Approximation of Boundary Conditions

The inflow boundary conditions for supersonic flow at $x = 0$ are imposed as $U_{0,j}(t) = g(y_j, t)$. The flow variables at the outlet $x = L_1$ are approximated by $U_{I,j}(t) = U_{I-1,j}(t)$, i.e. by constant extrapolation. This approximation implies that the upwind finite volume method is used to determine the numerical fluxes $F_{I-\frac{1}{2},j}$. The symmetry boundary conditions are implemented by considering an extra line below $y = 0$. There we use $U_{i,1}(t) = diag(1, 1, -1, 1)U_{i,3}(t)$. The boundary conditions at $y = L_2$ are treated as $U_{i,J}(t) = U_{i,J-1}(t)$.

4 Ghost Point Treatment at Embedded Boundary

4.1 Sjögreen and Petersson [16] Ghost Point Treatment for Two Dimensional Embedded Boundary

In Fig. 1 we show the flagging strategy. We flag the ghost and solid points by assigning them 0 and -1 values. The fluid points are assigned values equal to 1. In Fig. 2(a) we show a 2D graphical description of the treatment at the boundary [16]. The distance of ghost point g from the wedge is denoted by b_1. The straight line through g normal to the wedge is intersecting the horizontal lines at three points denoted by vertical lines. At the first intersection point I we obtain the primitive variables V_I by linear interpolation of the values at the neighboring horizontal grid points. And similarly we get V_{II} and V_{III}. We introduce a coordinate s on the line in the direction of the outer unit normal **n** of the boundary. Now we proceed as follows. Subtract the distance b_1 from the boundary coordinate s_{wedge} to obtain the mirror point s_m. Then we reach between intersection points I and II on the straight line normal to the boundary. Here we apply either linear interpolation between

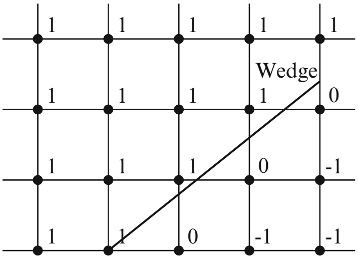


Fig. 1 Flagging strategy for fluid (1), ghost (0) and solid points (-1)

V_I and V_{II} or quadratic interpolation among V_I, V_{II} and V_{III} for normal and tangential components of velocity, pressure p and density ρ. The mathematical description of this strategy is explained as follows.

$$b_1 = s_g - s_{wedge}, \tag{4}$$

$$s_m = s_{wedge} - b_1, \tag{5}$$

$$V_m = V_I + \frac{V_{II} - V_I}{\Delta s}(s_I - s_m). \tag{6}$$

$$V_m = V_{III} + \frac{V_{II} - V_{III}}{s_{II} - s_{III}}(s_m - s_{III}) + \frac{\frac{V_I - V_{II}}{s_I - s_{II}} - \frac{V_{II} - V_{III}}{s_{II} - s_{III}}}{s_I - s_{III}}(s_m - s_{III})(s_m - s_{II}) \tag{7}$$

where $V = (\rho, u, v, p)$ and $\Delta s = s_I - s_{II}$. Then we use reflection boundary conditions

$$u_{t_g} = u_{t_m}, u_{n_g} = -u_{n_m}, p_g = p_m, \rho_g = \rho_m, \tag{8}$$

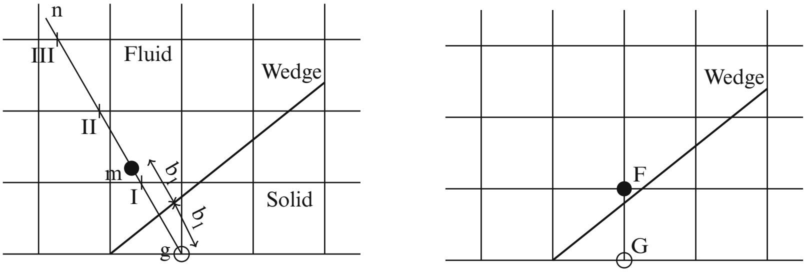
where u_t and u_n denote the tangential and normal components of the velocity vector, respectively.

4.2 Simplified Ghost Point Treatment for Two Dimensional Embedded Boundary

In Fig. 2(b) we show a simplified ghost point treatment at the solid boundary [17]. A ghost point is denoted by G. In the simplified ghost point treatment we consider the fluid point F on the vertical grid line through G adjacent to the boundary as the mirror point. Then, we assume the wedge is in the middle between ghost and fluid points. The mathematical description of this strategy is given as

$$\rho_G = \rho_F, p_G = p_F, u_G = u_F - 2(n_1 u_F + n_2 v_F)n_1, v_G = v_F - 2(n_1 u_F + n_2 v_F)n_2, \tag{9}$$

where n_1 and n_2 are the x-and y-components of the outer unit normal **n** of the boundary.



(a) Ghost point treatment at the boundary [16].

(b) Simplified ghost point treatment [17].

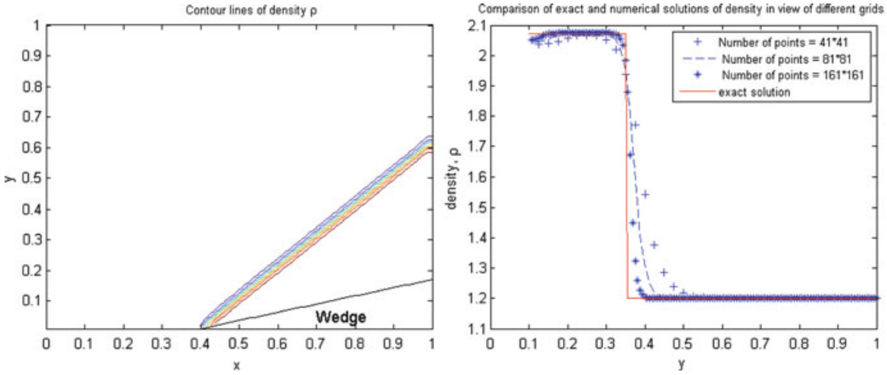
Fig. 2 Ghost point treatment

5 Results

5.1 Two Dimensional Compressible Euler Equations

We verify our 2D code of the Cartesian grid method for an oblique shock wave. For the spatial discretization we use the local Lax-Friedrichs (lLF) method, and to increase the order of our method we employ the MUSCL scheme with the minmod limiter. For time integration we use the first order explicit Euler and third order TVD Runge-Kutta (RK3) methods. A supersonic flow moves from left to right and hits a wedge with the wedge angle $\Theta = 15$. The supersonic upstream flow conditions are given as

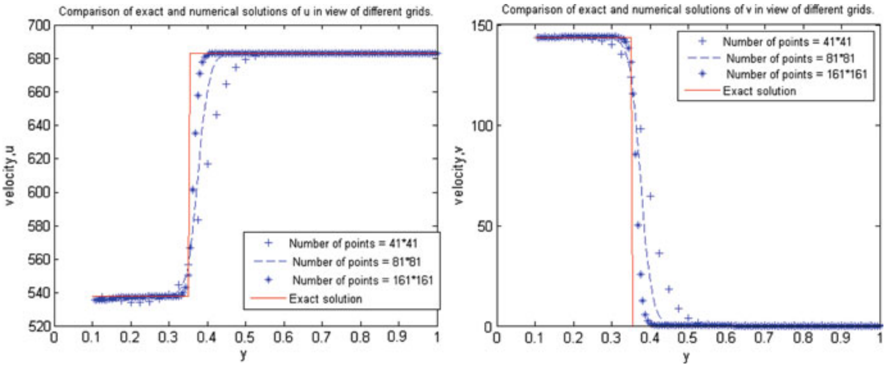
$$M = 2, p_\infty = 10^5 Pa, \rho_\infty = 1.2kg/m^3 \tag{10}$$



(a) Density contours for supersonic wedge flow ($M_\infty = 2, \Theta = 15$ degrees).

(b) Comparison of exact and numerical solutions for density at different grid levels.

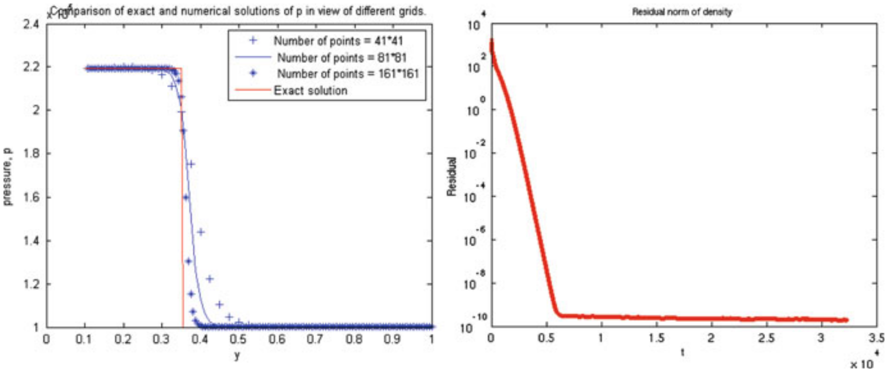
Fig. 3 Left: Density contours. Right: Comparison of exact and numerical solutions for density



(a) Comparison of exact and numerical solutions for velocity component u at different grid levels.

(b) Comparison of exact and numerical solutions for velocity component v at different grid levels.

Fig. 4 Comparison of exact and numerical solutions for velocity components



(a) Comparison of exact and numerical solutions for pressure p at different grid levels.

(b) Residual of density.

Fig. 5 Left: Comparison of exact and numerical solutions for pressure. Right: Residual of density

In Fig. 3(a) we present density contours obtained with the TVD RK3 method in time and the local Lax-Friedrichs (lLF) method in space with MUSCL and minmod limiter using the Sjögreen and Petersson ghost point treatment [16]. The apex of the wedge is placed at $x = 0.4$. When the supersonic flow hits the wedge an oblique shock wave is produced which begins at the apex of the wedge.

In Fig. 3(b) and Fig. 4(a) we compare the exact and numerical solutions for density ρ and velocity component u at $x = 0.75m$. We observe that ρ and u are getting closer to the exact solution as we refine the grid. However, there is some discrepancy between the exact and computed solutions near the wall of the wedge. This might be due to the ghost point method not guaranteeing conservativity and

Table 1 Mass flow error for simplified ghost point treatment [17]

	2D Compressible Euler Equations			
	First Order Method		MUSCL with minmod limiter	
Number of points	$\Delta\dot{m}[\frac{kg}{s}]$	$\frac{\Delta\dot{m}}{\dot{m}}$ %	$\Delta\dot{m}[\frac{kg}{s}]$	$\frac{\Delta\dot{m}}{\dot{m}}$ %
41×41	23.7115	2.9797	17.0569	2.1275
81×81	11.5532	1.43	8.3530	1.0298
161×161	5.6317	0.6920	4.0201	0.4930

Table 2 Mass flow error for Sjögreen and Petersson [16] method

	2D Compressible Euler Equations			
	Linear Interpolation		Quadratic Interpolation	
Number of points	$\Delta\dot{m}[\frac{kg}{s}]$	$\frac{\Delta\dot{m}}{\dot{m}}$ %	$\Delta\dot{m}[\frac{kg}{s}]$	$\frac{\Delta\dot{m}}{\dot{m}}$ %
41×41	23.3362	2.9311	23.3219	2.9293
81×81	11.3940	1.41	11.3653	1.4064
161×161	5.5727	0.6847	5.5465	0.6814

due to numerical problems near the apex of the wedge. This apex is acting like a singular point where the flow variables are multivalued.

In Figs. 4(b) and 5(a) we compare the exact and numerical solutions for velocity component v and pressure p. The computed results for v and p are in good agreement with the exact solutions.

In Fig. 5(b) we show the l_2-norm of the residual $(\frac{\rho_{i,j}^{n+1}-\rho_{i,j}^{n}}{\Delta t})$ of the density for the first order method in time and space. We see that the residual has dropped to machine accuracy after 5500 time levels.

In Tables 1 and 2 we present the mass flow error for the simplified ghost point treatment and the Sjögreen and Petersson [16] method. In Table 1 we present results for the first order node-centered finite volume method and the corresponding method with MUSCL and minmod limiter. From Table 1 we observe that by doubling the number of grid points in each direction the percentage of mass flow error is almost halved. In Table 2 we present results for the first order method by using linear and quadratic interpolation using the Sjögreen and Petersson [16] ghost point treatment. We see that linear and quadratic interpolation is not affecting the accuracy of our first order method. The mass flow error in Table 2 obtained with [16] is only slightly lower than the mass flow error of the first order method in Table 1 obtained with the simplified ghost point treatment [17].

6 Conclusions

The Cartesian grid method has been applied to the compressible Euler equations. Local symmetry boundary conditions have been employed at ghost points. The ghost point treatments at the solid boundary are not conservative, and the mass flow error is calculated. We find that linear or quadratic interpolation does not affect the results for the Sjögreen and Petersson method. For supersonic wedge flow, the simplified ghost point treatment on vertical grid lines yields similar results as the ghost point treatment on lines normal to the boundary.

Acknowledgements The current research has been funded by Higher Education Commission (HEC) of Pakistan.

References

1. Almgren, A. S., Bell, J. B., Colella, P., Marthaler, T.: A Cartesian grid projection method for the incompressible Euler equations in complex geometries, SIAM J. Sci. Comput **18**, 1289–1309 (1997).
2. Palma, P. D., de Tullio, M. D., Pascazio, G., Napolitano, M.: An immersed-boundary method for compressible viscous flows, Comput. Fluids **35**, 693–702 (2006).
3. Marshall, D. D., Ruffin, S. M: A new inviscid wall boundary condition treatment for boundary Cartesian grid method, AIAA 2004-0583 42nd AIAA Aerospace Sciences Meeting and Exhibit, Reno, Nevada (2004).
4. Udaykumar, H. S., Krishnann, S. and Marella, S. V. : Adaptively refined parallelisded sharp interface Cartesian grid method for three dimensional moving boundary problem, Int. J. Comput. Fluid Dyn. **23**, 1–24 (2009).
5. Uzgoren, E., Sim, J., Shyy, W.: Marker based 3-D adaptive Cartesian grid method for multiphase flow around irregular geometries, Commun. Comput. Phys. **5**, 1–41 (2009).
6. Wang, Z., Fan, J., Cen, K.: Immersed boundary method for the simulations of 2D viscous flow based on vorticity-velocity formulations., J. Comput. Phys **228**, 1504–1520 (2009).
7. Mittal, R., Iaccarino, G: Immersed boundary method, Annu. Rev. Fluid Mech. **37**, 239–261 (2005).
8. Quirk, J. J.: An alternative to unstructred grids for computing gas dynamic flows around arbitrarily complex two dimensional bodies, Comput. Fluids **23**, 125–142 (1994).
9. Forrer, H., Jeltsch, R: A higher-order boundary treatment for Cartesian grid methods, J. Comput. Phys. **140**, 259–277 (1998).
10. Peskin, C. S. : Flow pattern around heart valves: A numerical method, J. Comput. Phys. **10**, 252–271 (1972).
11. de Tullio, M. D., Palma, P. D. D., Iaccarino, G., Pascazio, G., Napolitano, M.: An immersed boundary method for compressible flows using local grid refinement, J. Comput. Phys. **225**, 2098–2117 (2007).
12. Berger, M. J., Leveque, R. J.: A rotated difference scheme for Cartesian grids in complex geometries, AIAA Paper **CP-91-1602**, 1–9 (1991).
13. Pember, R. B., Bell, J. B., Colella, P., Curtchfield, W. Y., Welcome, M. L.: An adaptive Cartesian grid method for unsteady compressible flow in irregular regions, J. Comput. Phys. **117**, 121–131 (1995).
14. Coirier, W. J., Powell, K. G.: An accuracy assessment of Cartesian-mesh approaches for the Euler equations, J. Comput. Phys. **117**, 121–131 (1995).
15. 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).
16. Sjögreen, B., Petersson, N. A.: A Cartesian embedded boundary method for hyperbolic conservation laws, Commun. Comput. Phys. **2**, 1199–1219 (2007).
17. Farooq, M. A., Müller, B.: Investigation of the accuracy of the Cartesian grid method, Proceedings of International Bhurban Conference on Applied Sciences and Technology Islamabad, Pakistan, January 10-13 (2011).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Compressible Stokes Problem with General EOS

A. Fettah and T. Gallouët

Abstract In this paper, we propose a discretization for the compressible Stokes problem with an equation of state of the form $p = \varphi(\rho)$ (where p stands for the pressure, ρ for the density and φ is a nondecreasing function belonging to $C^1(\mathbb{R}_+, \mathbb{R})$). 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, and two additional terms. We prove existence of a discrete solution and convergence of this approximate solution to a solution of the continuous problem.

Keywords Compressible Stokes, finite element, finite volume
MSC2010: 35Q30,65N12,65N08,65N30

1 Introduction

Let Ω be a bounded open set of $\mathbb{R}^d$, polygonal if $d = 2$ and polyhedral if $d = 3$, and $\mu > 0$. For $M > 0$, $f \in \mathrm{L}^2(\Omega)^d$ and $\varphi \in C^1(\mathbb{R}_+, \mathbb{R})$ a nondecreasing function satisfying:

$$\forall s \in \mathbb{R}_+, as^\gamma - b \le \varphi(s) \le \tilde{a}s^{2\gamma-1} + \tilde{b}, \tag{1}$$

with $a, \tilde{a}, b, \tilde{b} > 0$ and $\gamma > 1$, we consider the following problem:

A. Fettah and T. Gallouët
Université Aix-Marseille, e-mail: afettah@cmi.univ-mrs.fr, gallouet@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_48, © Springer-Verlag Berlin Heidelberg 2011

$$-\mu\Delta u-\frac{\mu}{3}\nabla(divu)+\nabla p=f \text{ in } \Omega,\ \ u=0 \text{ on } \partial\Omega, \tag{2a}$$

$$\operatorname{div}(\rho u)=0 \ \text{ in } \Omega,\ \rho\geq 0 \ \text{ in } \Omega,\ \int_{\Omega}\rho(x)\,\mathrm{d}x=M, \tag{2b}$$

$$p=\varphi(\rho) \text{ in } \Omega. \tag{2c}$$

Remark 1. The second inequality in (1) is used only in Section 3, for the passage to the limit in the EOS. It can be replaced by an hypothesis of convexity of φ.

Definition 1. Let $f\in \mathrm{L}^2(\Omega)^d$ and $M>0$. A weak solution of Problem (2) is a function (u,p,ρ) satisfying:

$$(u,p,\rho)\in \mathrm{H}_0^1(\Omega)^d\times \mathrm{L}^2(\Omega)\times \mathrm{L}^{2\gamma}(\Omega), \tag{3a}$$

$$\mu\int_{\Omega}\nabla u:\nabla v\,\mathrm{d}x+\frac{\mu}{3}\int_{\Omega}\operatorname{div}(u)\operatorname{div}(v)\,\mathrm{d}x-\int_{\Omega}p\,\operatorname{div}(v)\,\mathrm{d}x=\int_{\Omega}f\cdot v\,\mathrm{d}x$$
$$\text{for all } v\in(\mathrm{H}_0^1(\Omega))^d, \tag{3b}$$

$$\int_{\Omega}\rho u\cdot\nabla\psi\,\mathrm{d}x=0 \text{ for all } \psi\in \mathrm{W}^{1,\infty}(\Omega), \tag{3c}$$

$$\rho\geq 0 \text{ a.e. in } \Omega,\ \int_{\Omega}\rho\,\mathrm{d}x=M,\ p=\varphi(\rho) \text{ a.e. in } \Omega. \tag{3d}$$

In Section 2, we give a possible discretization of this problem and we prove the existence of a solution of the discrete problem. In Section 3 we prove the convergence (up to a subsequence, since no uniqueness result of a solution of (3) is avalaible), as the mesh size goes to zero, of this approximate solution to a solution of (3). In particular, we then obtain existence of a solution of (3). The present paper generalizes the results of [3] where convergence was proven if $\varphi(\rho)=\rho^{\gamma}$. The main additional difficulties of the present paper with respect to [3] are in the proof of the crucial lemma 1 (which yields the estimate on the approximate velocity), in the proof of the estimate of the approximate pressure (Inequality (13)) and in the last proof of the paper, which consists in proving $p=\varphi(\rho)$. The proof of $p=\varphi(\rho)$ cannot be done using a strict monotony argument as in [3] because φ is not necessarily an increasing function. We overcome this difficulty by using the so called "Minty trick" (the drawback of this method is that we do not obtain the a.e. convergence, up to a subsequence, of pressure and density).

2 Discrete spaces and scheme

Let $\mathscr{T}$ be a decomposition of the domain Ω in simplices, which we call hereafter a triangulation of Ω, regardless of the space dimension. By $\mathscr{E}(K)$, we denote the set of the edges ($d=2$) or faces ($d=3$) of the element $K\in\mathscr{T}$; for short, each edge or face will be called an edge hereafter. The set of all edges of the mesh is denoted by $\mathscr{E}$; the set of edges included in the boundary of Ω is denoted by $\mathscr{E}_{\rm ext}$ and the set of internal edges (*i.e.* $\mathscr{E} \setminus \mathscr{E}_{\rm ext}$) is denoted by $\mathscr{E}_{\rm int}$. The decomposition $\mathscr{T}$ is assumed to be regular in the usual sense of the finite element literature. For each internal edge of the mesh $\sigma = K|L$, n_{KL} stands for the normal vector of σ, oriented from K to L (so that $n_{KL} = -n_{LK}$). 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:

$$\theta = \inf \{\frac{\xi_K}{h_K},\ K \in \mathscr{T}\} \tag{4}$$

where ξ_K stands for the diameter of the largest ball included in K. Finally, as usual, we denote by h the quantity $\max_{K\in\mathscr{T}} h_K$. The space discretization relies on the Crouzeix-Raviart element (see [1] for the seminal paper and, for instance, [2, pp. 199–201] for a synthetic presentation). The space of approximation for the velocity is the space W_h of vector-valued functions each component of which belongs to V_h: $W_h = (V_h)^d$, where V_h is the discrete space defined as follows:

$$\begin{aligned} V_h = \{\ v \in L^2(\Omega)\ :\ \forall K \in \mathscr{T},\ v|_K \in P_1(K)\,; \\ \forall \sigma \in \mathscr{E}_{\rm int},\ \sigma = K|L,\ \ F_\sigma(v|_K) = F_\sigma(v|_L);\ \forall \sigma \in \mathscr{E}_{\rm ext},\ F_\sigma(v) = 0\}, \end{aligned} \tag{5}$$

where $F_\sigma(v)$ is the mean value of v on σ, denoted hereafter by v_σ. The pressure and the density are approximated in the space L_h of piecewise constant functions, namely $L_h = \big\{q \in L^2(\Omega)\ :\ q|_K = \text{ constant},\ \forall K \in \mathscr{T}\big\}$. For $u \in W_h$, the discrete gradient and discrete divergence of u are defined by $\nabla_h u = \nabla u$ and $\mathrm{div}_h(u) = \mathrm{div}(u)$ on K, for $K \in \mathscr{T}$. The Crouzeix-Raviart pair of approximation spaces for the velocity and the pressure is *inf-sup* stable, in the sense that there exists $c_{\rm i} > 0$ only depending on Ω and, in a monotone way, on θ, such that:

$$\forall p \in L_h, \qquad \sup_{v\in W_h} \frac{\displaystyle\int_\Omega p\ \mathrm{div}_h(v)\,\mathrm{d}x}{||v||_{1,b}} \geq c_{\rm i}\ ||p - m(p)||_{\mathrm{L}^2(\Omega)}\ ,$$

where $m(p)$ is the mean value of p over Ω and $||\cdot||_{1,b}$ stands for the broken Sobolev H^1 semi-norm, which is defined for scalar as well as for vector-valued functions by:

$$||v||^2_{1,b} = \sum_{K\in\mathscr{T}} \int_K |\nabla v|^2\,\mathrm{d}x = \int_\Omega |\nabla_h v|^2\,\mathrm{d}x.$$

This norm is known to control the L^2 norm by a Poincaré inequality (*e.g.* [2, lemma 3.31]). We also define a discrete semi-norm on L_h, similar to the usual H^1 semi-norm used in the finite volume context:

$$\forall p \in L_h, \qquad |p|^2_{\mathscr{T}} = \sum_{\substack{\sigma\in\mathscr{E}_{\rm int},\\ \sigma=K|L}} \frac{|\sigma|}{h_\sigma}\,(p_K - p_L)^2.$$

We refer to [1] for the usual properties of the interpolation operator from the Sobolev spaces to the Crouzeix-Raviart spaces. We now describe the numerical scheme. Let ρ^* be the mean density, *i.e.* $\rho^* = M/|\Omega|$ where $|\Omega|$ stands for the measure of Ω. We consider the following numerical scheme for the discretization of (2):

$$u \in W_h,\ p \in L_h,\ \rho \in L_h, \tag{6a}$$

$$\forall v \in W_h,\ \ \mu \int_\Omega \nabla_h u : \nabla_h v \,{\rm d}x + \frac{\mu}{3}\int_\Omega {\rm div}_h(u){\rm div}_h(v)\,{\rm d}x - \int_\Omega p\ {\rm div}_h(v)\,{\rm d}x$$

$$= \int_\Omega f\cdot v \,{\rm d}x, \tag{6b}$$

$$\forall K \in \mathscr{T},\ \ \sum_{\sigma=K|L} {\rm v}^+_{\sigma,K}\,\rho_K - {\rm v}^-_{\sigma,K}\,\rho_L + M_K + T_K = 0, \tag{6c}$$

$$\forall K \in \mathscr{T},\ \ p_K = \varphi(\rho_K), \tag{6d}$$

where:

- ${\rm v}_{\sigma,K} = |\sigma| u_\sigma \cdot n_{KL}$, ${\rm v}^+_{\sigma,K} = \max({\rm v}_{\sigma,K}, 0)$, ${\rm v}^-_{\sigma,K} = -\min({\rm v}_{\sigma,K}, 0)$,
- the terms M_K and T_K read, with $\zeta = \max(0, 2-\gamma)$, $\alpha \geq 1$ and $0 < \xi < 2$,

$$M_K = h^\alpha\,|K|\,\left(\rho_K - \rho^*\right), \tag{7a}$$

$$T_K = \sum_{\sigma=K|L} (h_K + h_L)^\xi\,\frac{|\sigma|}{h_\sigma}\,(|\rho_K| + |\rho_L|)^\zeta\,(\rho_K - \rho_L)\,. \tag{7b}$$

As it is proven in [3], if $(u,\rho) \in W_h \times L_h$ is solution of (6c), one has necessarily $\rho_K > 0$ for all $K \in \mathscr{T}$, so that (6d) makes sense, and $\sum_{K\in\mathscr{T}} |K|\rho_K = M$. The existence of a solution to the numerical scheme (6) can be proven with the Brouwer fixed point Theorem, using a simple adaptation of the proof of [3], which we therefore omit.

3 Convergence of approximate solutions

We first have to obtain some estimates on the approximate solution. In order to obtain an estimate on the velocity, we will use the following crucial lemma 1.

Lemma 1. *Let $\mathscr{T}$ be a triangulation of the computational domain Ω and $(u,\rho) \in W_h \times L_h$ satisfy Equation (6c). (As above mentioned, this gives $\rho > 0$.) Then:*

$$\int_{\Omega} \varphi(\rho)\mathrm{div}_h(u)\,\mathrm{d}x \le 0.$$

Proof. Let $\psi \in C^1(\mathbb{R}_+^\star)$ be a function satisfying $\psi'(s) = \frac{\varphi'(s)}{s}$ (so that ψ is nondecreasing). Multiplying (6c) by $\psi_K = \psi(\rho_K)$ and summing over $K \in \mathscr{T}$ yields $T_1 + T_2 + T_3 = 0$ with:

$$T_1 = \sum_{K\in\mathscr{T}} \psi_K \sum_{\sigma=K|L} |\sigma|\rho_\sigma\, u_\sigma \cdot n_{KL},\ T_2 = \sum_{K\in\mathscr{T}} h^\alpha\,|K|\,\psi(\rho_K)\left(\rho_K - \rho^*\right),$$
$$T_3 = \sum_{K\in\mathscr{T}} \psi(\rho_K) \sum_{\sigma=K|L} (h_K + h_L)^\xi\,\frac{|\sigma|}{h_\sigma}\,(\rho_K+\rho_L)^\zeta\,(\rho_K - \rho_L)\,.$$

Let $T_4 = \sum_{K\in\mathscr{T}} \int_K \varphi(\rho_K)\mathrm{div}(u) = \sum_{\sigma=K|L} |\sigma| u_\sigma \cdot n_{KL}(\varphi(\rho_K) - \varphi(\rho_L))$. We have $T_4 = T_4 - T_1 - T_2 - T_3$ and then

$$T_4 = \sum_{\sigma=K|L} |\sigma| u_\sigma \cdot n_{KL}[\varphi(\rho_K) - \varphi(\rho_L) - \rho_\sigma(\psi(\rho_K) - \psi(\rho_L))] - T_2 - T_3. \quad (8)$$

The fact that ψ is nondecreasing (and $\sum_{K\in\mathscr{T}} |K|\rho_K = M$) yields:

- $T_2 \ge \sum_{K\in\mathscr{T}} h^\alpha |K| \psi(\rho^\star)\,(\rho_K - \rho^*) = 0$
- $T_3 = \sum_{\sigma=K|L} (h_K + h_L)^\xi\,\frac{|\sigma|}{h_\sigma}\,(\rho_K+\rho_L)^\zeta\,(\rho_K-\rho_L)\,(\psi(\rho_K) - \psi(\rho_L)) \ge 0.$

In order to conclude that $T_4 \le 0$, we now introduce, for $\alpha > 0$, the function Φ defined on $\mathbb{R}_+^\star$ by

$$\Phi(s) = \varphi(\alpha) - \varphi(s) - \alpha(\psi(\alpha) - \psi(s)).$$

Since $s\psi'(s) = \varphi'(s)$ (for $s > 0$), one has $\Phi(s) \le 0$ for all $s > 0$ and then:

$$\sum_{\sigma=K|L} |\sigma| u_\sigma \cdot n_{KL}[\varphi(\rho_K) - \varphi(\rho_L) - \rho_\sigma(\psi(\rho_K) - \psi(\rho_L))] \le 0.$$

We then conclude, with (8), that $\int_\Omega \varphi(\rho)\mathrm{div}_h(u)\,\mathrm{d}x = T_4 \le 0$.

Theorem 1. *Let $\theta_0 > 0$ and let $\mathscr{T}$ be a triangulation of the computational domain Ω such that $\theta \ge \theta_0$, where θ is defined by (4). Let (u, p, ρ) be a solution of (6). Then, there exist C, only depending on the data of the problem (Ω, μ, f, M and φ) and on θ_0 such that:*

$$\|u\|_{1,b} \le C,\ \|p\|_{\mathrm{L}^2(\Omega)} \le C,\ \|\rho\|_{\mathrm{L}^{2\gamma}(\Omega)} \le C \text{ and } h^{\xi/2}\,|\rho|_{\mathscr{T}} \le C. \quad (9)$$

Proof. Let (u, p, ρ) be a solution of (6) . Taking u as test function in (6b) yields:

$$\mu\,\|u\|_{1,b}^2 + \frac{\mu}{3}\int_\Omega \mathrm{div}_h^2(u)\,\mathrm{d}x - \int_\Omega p\,\mathrm{div}(u)\,\mathrm{d}x = \int_\Omega f\cdot u\,\mathrm{d}x. \quad (10)$$

Using Lemma 1, a discrete Poincaré Inequality (as in [3]) and the Hölder inequality, yields the existence of C_1 only depending on Ω, f, μ and θ_0 such that $\|u\|_{1,b} \leq C_1$. Using the *inf-sup* stability of the discretization, we hence get from (10) a control of $\|p - m(p)\|_{\mathrm{L}^2(\Omega)}$ (where $m(p)$ stands for the mean value of p over Ω).

In order to obtain an estimate on p, we set (for simplicity) $\varphi(s) = s + \varphi(0)$ for $s < 0$ and we define the function Φ from $\mathbb{R}$ to $\mathbb{R}$ by $\Phi(s) = \inf\{t \in \mathbb{R}_+; s = 3\varphi(t)\}$. The function Φ satisfies the following properties:

$$s = 3\varphi(t) \Rightarrow \Phi(s) \leq t, \tag{11a}$$

$$s = 3\varphi(\Phi(s)), \tag{11b}$$

$$\Phi(s) \to +\infty, \text{ as } s \to +\infty, \tag{11c}$$

$$\Phi \text{ is nondecreasing.} \tag{11d}$$

For all $x \in \Omega$ one has $m(p) \leq |m(p) - p(x)| + |p(x)| \leq |m(p) - p(x)| + 2|\varphi(0)| + p(x)$. Then, using (11d),

$$\Phi(m(p)) \leq \Phi(3|m(p) - p(x)|) + \Phi(6|\varphi(0)|) + \Phi(3p(x)).$$

Since $3p(x) = 3\varphi(\rho(x))$, (11a) gives $\Phi(m(p)) \leq \Phi(3|m(p) - p(x)|) + \Phi(6|\varphi(0)|) + \rho(x)$. By summing equation (6c) for $K \in \mathscr{T}$, we obtain that the integral of ρ over Ω is M, which yields:

$$\int_\Omega \Phi(m(p))dx \leq \int_\Omega \Phi(3|m(p) - p(x)|)dx + M + \Phi(6|\varphi(0)|)|\Omega|. \tag{12}$$

On the other hand, if $\Phi(s) \geq 0$, one has, with (11b) and the first inequality of (1),

$$\frac{s}{3} = \varphi(\Phi(s)) \geq a(\Phi(s))^\gamma - b,$$

and then $\Phi(s) \leq (\frac{|s|}{3a} + \frac{b}{a})^{\frac{1}{\gamma}} \leq (\frac{|s|}{3a} + \frac{b}{a} + 1)^2$. This inequality gives an estimate on $\int_\Omega \Phi(3|m(p) - p(x)|)dx$ from the L^2-estimate on $(p - m(p))$. We hence get, with (12), an estimate on $\Phi(m(p))$. Using (11c) yields an estimate on $m(p)$. Finally, the estimate on $[m(p)]$ and $[p - m(p)]$ gives the existence of C_2 (depending on the data and θ_0) such that

$$\|p\|_{\mathrm{L}^2(\Omega)} \leq C_2. \tag{13}$$

Finally, thanks to $p = \varphi(\rho)$ and the first inequality of (1), the estimate on ρ follows. For the estimate on $|\rho|_{\mathscr{T}}$, which comes form the T_K term in (6c), we refer to [3] where the proof is the same.

Let us now state the final convergence result:

Theorem 2. *Let a sequence of triangulations $(\mathscr{T}^{(n)})_{n\in\mathbb{N}}$ of Ω be given. We assume that h_n tends to zero when $n \to \infty$. In addition, we assume that the sequence of* *discretizations is regular, in the sense that there exists* $\theta_0 > 0$ *such that* $\theta_n \geq \theta_0$ *for all* $n \in \mathbb{N}$. *For* $n \in \mathbb{N}$, *we denote by* $W_h^{(n)}$ *and* $L_h^{(n)}$ *the discrete spaces associated to* $\mathscr{T}^{(n)}$ *and b* (u_n, p_n, ρ_n) *a corresponding solution to the discrete problem* (6)*, with* $\alpha \geq 1$ *and* $0 < \xi < 2$. *Then, up to the extraction of a subsequence, when* $n \to \infty$*:*

1. *the sequence* $(u_n)_{n\in\mathbb{N}}$ *(strongly) converges in* $\mathrm{L}^2(\Omega)^d$ *to a limit* $u \in \mathrm{H}_0^1(\Omega)^d$*,*
2. *the sequence* $(p_n)_{n\in\mathbb{N}}$ *weakly converges in* $\mathrm{L}^2(\Omega)$*,*
3. *the sequence* $(\rho_n)_{n\in\mathbb{N}}$ *weakly converges in* $\mathrm{L}^{2\gamma}(\Omega)$*,*
4. (u, p, ρ) *is a solution to Problem* (3)*.*

Proof. The first item of Theorem 2 (namely the convergence, up to the extraction of a subsequence, of the sequence $(u_n)_{n\in\mathbb{N}}$ and the fact that the limit belongs to $\mathrm{H}_0^1(\Omega)^d$) is a consequence of the uniform (with respect to n) estimate of Theorem 1, applying a compactness result wich is proven for instance in [4, Theorem 3.3]. The second and third item of Theorem 2 are trivial consequences of the uniform (with respect to n) estimate of Theorem 1. It remains to prove that (u, p, ρ) is solution to (3b)–(3d).

The proof that the limit satisfies $\rho \geq 0$ a.e. in Ω, $\int_\Omega \rho \,\mathrm{d}x = M$ and Equation (3b) is strictly the same as the proof of the same result for a linear equation of state, *i.e.* Theorem 6.1 in [4]. The fact that (u, ρ) satisfies Equation (3c) follows the proof of [3]. Then, we only need here to prove that the equation of state is satisfied, that is $p = \varphi(\rho)$ a.e. in Ω.

The fact that $\rho \in \mathrm{L}^{2\gamma}(\Omega)$, $\rho \geq 0$ a.e. in Ω, $u \in (\mathrm{H}_0^1(\Omega))^d$ and that (ρ, u) satisfies (3c) yields, see Lemma 2.1 in [3]:

$$\int_\Omega \rho \,\mathrm{div}(u) \,\mathrm{d}x = 0. \tag{14}$$

Then, using (14), we have, following the proof given in [3]:

$$\lim_{n\to\infty} \int_\Omega \big(p_n - \mathrm{div}_h(u_n)\big)\, \rho_n \,\mathrm{d}x - \int_\Omega p\,\rho \,\mathrm{d}x = 0.$$

As in [3], we also have $\limsup_{n\to\infty} \int_\Omega \mathrm{div}_h(u_n)\, \rho_n \,\mathrm{d}x \leq 0$. Hence:

$$\limsup_{n\to\infty} \int_\Omega p_n\, \rho_n \,\mathrm{d}x \leq \int_\Omega p\,\rho \,\mathrm{d}x. \tag{15}$$

We want to deduce from (15) that $p = \varphi(\rho)$. But, since φ in only nondecreasing (and not necessarily increasing), we cannot use the proof given in [3]. We use here the so called Minty trick.

For simplicity, we define φ on $\mathbb{R}^-$ setting $\varphi(s) = \varphi(0)$ if $s < 0$. Let $\bar\rho \in L^{2\gamma}$ and, for $n \in \mathbb{N}$, $G_n = (\varphi(\rho_n) - \varphi(\bar\rho))(\rho_n - \bar\rho)$. One has $G_n \geq 0$ a.e. in Ω (since φ is nondecreasing). Thanks to the second inequality of (1) (which is used only in this proof) one has $\varphi(\bar\rho) \in L^{2\gamma/(2\gamma-1)}(\Omega)$ and then $\varphi(\bar\rho)\bar\rho \in L^1(\Omega)$. Then, one has

$G_n \in L^1(\Omega)$ and

$$0 \leq \int_\Omega G_n \, dx = \int_\Omega (p_n \rho_n - p_n \bar{\rho} - \varphi(\bar{\rho})\rho_n + \varphi(\bar{\rho})\bar{\rho}) \, \mathrm{d}x.$$

Using (15) and the weak convergences of p_n to p and ρ_n to ρ in $L^2(\Omega)$ and $L^{2\gamma}(\Omega)$ respectively, we obtain:

$$0 \leq \limsup_{n\to\infty} \int_\Omega G_n \, dx \leq \int_\Omega (p - \varphi(\bar{\rho}))(\rho - \bar{\rho}) \, \mathrm{d}x.$$

We have thus proven that

$$\int_\Omega (p - \varphi(\bar{\rho}))(\rho - \bar{\rho}) \, \mathrm{d}x \geq 0 \text{ for all } \bar{\rho} \in L^{2\gamma}(\Omega). \tag{16}$$

We now have to choose $\bar{\rho}$ conveniently to deduce $p = \varphi(\rho)$ from (16). Let $\psi \in C_c^\infty(\Omega, \mathbb{R})$. For $n \in \mathbb{N}^\star$, we set $\rho_n = \rho + \frac{1}{n}\psi$. Since $\rho_n \in L^{2\gamma}$, we can choose $\bar{\rho} = \rho_n$ in (16). We obtain

$$\int_\Omega (p - \varphi(\rho + \frac{1}{n}\psi))\psi \leq 0.$$

We now use the Dominated Convergence Theorem on the sequence $(g_n)_{n\in\mathbb{N}^\star}$ with $g_n = (p - \varphi(\rho + \frac{1}{n}\psi))\psi$. The continuity of φ gives $g_n \to (p - \varphi(\rho))\psi$ a.e. in Ω. Since φ is nondecreasing, one has, for all $n \in \mathbb{N}^\star$,

$$|g_n| \leq G = |p\psi| + |\varphi(\rho + \|\psi\|_\infty)\psi| + |\varphi(0)\psi| \text{ a.e. in } \Omega.$$

The second inequality of (1) gives $\varphi(\rho + \|\psi\|_\infty) \in L^1(\Omega)$. Then one has $G \in L^1(\Omega)$ and the Dominated Convergence Theorem yields $\int_\Omega (p - \varphi(\rho))\psi \leq 0$. Changing ψ in $-\psi$, we conclude that $\int_\Omega (p - \varphi(\rho))\psi = 0$ for all $\psi \in C_c^\infty(\Omega, \mathbb{R})$. This gives $p = \varphi(\rho)$ a.e. in Ω. The proof of Theorem 2 is now complete.

If φ is increasing, we can prove, as in [3], the a.e. convergence, up to a subsequence, of p and ρ.

Conclusion We gave a scheme for the discretization of the compressible Stokes problem with a quite general EOS and we proved the convergence of the approximate solution to an exact solution (up to a subsequence) as the mesh size goes to zero. The main difficulty of the paper is in the passage to the limit in EOS. This difficulty is due to the nonlinearity of the EOS and the fact that the estimates on pressure and density only lead to weak convergences. It will be now interesting to consider the Navier-Stokes problem along with the evolution problem.

References

1. M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations I. *Revue Française d'Automatique, Informatique et Recherche Opérationnelle (R.A.I.R.O.)*, R-3:33–75, 1973.
2. A. Ern and J.-L. Guermond. Theory and practice of finite elements. Number 159 in Applied Mathematical Sciences. Springer, New York, 2004.
3. R. Eymard, T. Gallouët, R. Herbin, and J.-C. Latché. A convergent finite element-finite volume scheme for the compressible Stokes problem. Part II: the isentropic case. *to appear in Mathematics of Computation*, 2009.
4. T. Gallouët, R. Herbin, and J.-C. Latché. A convergent finite element-finite volume scheme for the compressible Stokes problem. Part I: the isothermal case. *Mathematics of Computation*, 267:1333–1352, 2009.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Asymptotic Preserving Finite Volumes Discretization For Non-Linear Moment Model On Unstructured Meshes

Emmanuel Franck, Christophe Buet, and Bruno Després

Abstract In this work we present a new finite volume discretization of the nonlinear model M_1 [2]. This new method is based on nodal solver for hyperbolic systems [3, 6] and overcomes, on 2-D unstructured meshes, the problem of the inconsistent diffusion limit for schemes based on classical edge formulation. We provide numerical examples to illustrate the properties of the method.

Keywords asymptotic preserving, M_1 model, unstructured, diffusion limit, GLACE scheme
MSC2010: 65M08, 65M08, 65M22, 65Z05, 65M08

1 Introduction

Our physical motivation stems from the discretization of the linear transport equation $\partial_t f(t,\mathbf{x},\omega) + \frac{1}{\epsilon}\omega\nabla f(t,\mathbf{x},\omega) = \frac{\sigma}{\epsilon^2}Q(f)$ in diffusive regime. $f(t,\mathbf{x},\omega) \geq 0$ is the distribution function associated to particles located in $\mathbf{x}$ and having a direction ω. $Q(f)$ is a Lorentz operator for scattering of lights particles. It is well known that on coarse grids, numerical schemes for such hyperbolic systems does not capture the diffusion limit ($\varepsilon << 1$) correctly. Since many years, many Asymptotic Preserving (AP) schemes have been proposed to correct this problem. But extended in 2-D unstructured meshes these methods are not consistent with a diffusion operator in diffusive regimes and coarse grids.

Emmanuel Franck and Christophe Buet
CEA, DAM, DIF, F-91297 Arpajon, France, e-mail: efranck21@gmail.com, christophe.buet@cea.fr

Bruno Després
Laboratoire Jacques Louis Lions, Université Pierre et Marie Curie, 75252 Paris, Cedex 5, France

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_49, © Springer-Verlag Berlin Heidelberg 2011

In this work we present an attempt to overcome this difficulty on a simplified non linear model which is the M_1 model: this model is the first element of a family of angular discretization based on minimization of entropy procedure [2]. It writes

$$\left\{\begin{array}{c} \partial_t E + \dfrac{1}{\epsilon}\nabla.\mathbf{F} = 0 \\ \partial_t \mathbf{F} + \dfrac{1}{\epsilon}\nabla(\hat{P}) = -\dfrac{\sigma}{\epsilon^2}\mathbf{F} \end{array}\right. \tag{1}$$

E is the energy, $\mathbf{F}$ is the flux and $\hat{P}$ the pressure tensor. The pressure tensor is defined by

$$\hat{P} = \frac{1}{2}((1-\chi(\mathbf{f}))Id + (3\chi(\mathbf{f})-1)\frac{\mathbf{f}\otimes\mathbf{f}}{\parallel \mathbf{f} \parallel})E$$

with $\mathbf{f} = \mathbf{F}/E$, $\chi(\mathbf{f}) = \dfrac{3+4\mathbf{f}^2}{5+2\sqrt{4-3\mathbf{f}^2}}$ for M_1 model. When ϵ tends to zero these models tends to the linear diffusion equation $\partial_t E(t,\mathbf{x}) - \frac{1}{3\sigma}\triangle E(t,\mathbf{x}) = 0$.

The original contribution of this work concerns news results for the construction of an AP scheme for the non-linear M_1 models on 2-D unstructured meshes. For this we first rewrite the M_1 model as a compressible gas dynamics like equation as in [4], and second we adapt the linear scheme developed in [5] to the non linear M_1 model. For the P_1 model [5]: one shows that in diffusive regime and on coarse grids the asymptotic limit of the finit volume based scheme is consistent with the right diffusion equation. Numerical results that it is also the case for the discretization of the M_1 model. To finish we present news numerical results for the equilibrium radiative model, with a non-linear coupling between a moment model and a non linear temperature relaxation.

2 Notations on 2-D unstructured meshes

Jin, Levermore in [9] or Gosse, Toscani in [7] proposed methods based on the incorporation to the source term in the fluxes in order to obtain AP schemes. We introduce some notations which are used to define a particular AP scheme on unstructured mesh.

Our idea is to use a nodal scheme like “GLACE” [6] or “CHIC” [3] for the linearized Euler equations, since the hyperbolic heat equation is a special case of them and incorporate the source term in the Riemann solver by Jin-Levermore procedure [9] to obtain an AP scheme. Let us consider the 2D unstructured mesh of Fig. 2. The mesh is defined by the vertices $\mathbf{x}_r$ and the cells Ω_j. We denote by $\mathbf{x}_j$ the gravity center of Ω_j. In each cell j, we define the length and the normal associated to the node of local index r

$$l_{jr} = \frac{1}{2} \parallel \mathbf{x}_{r+1} - \mathbf{x}_{r-1} \parallel \text{ and } \mathbf{n}_{jr} = \frac{1}{2l_{jr}}\begin{pmatrix} -y_{r-1}+y_{r+1} \\ x_{r-1}-x_{r+1} \end{pmatrix}. \tag{2}$$

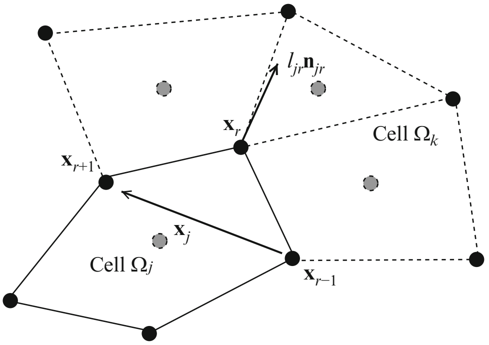


Fig. 1 Notation for the nodal formulation

where (x_r, y_r) are the coordinates of $\mathbf{x}_r$. We use a tensor definition of nodal schemes introduced in [8]. We define the scheme by

$$\begin{cases} \left|\Omega_j\right| \partial_t E_j(t) + \dfrac{1}{\varepsilon}\displaystyle\sum_r l_{jr}(\mathbf{F}_r.\mathbf{n}_{jr}) = 0 \\ \left|\Omega_j\right| \partial_t \mathbf{F}_j(t) + \dfrac{1}{\varepsilon}\displaystyle\sum_r \mathbf{G}_{jr} = -\dfrac{\sigma}{\varepsilon^2}\displaystyle\sum_r \widehat{\beta}_{jr}\mathbf{F}_r \end{cases} \tag{3}$$

The fluxes associated to these schemes are

$$\begin{cases} \mathbf{G}_{jr} = l_{jr}E_j\mathbf{n}_{jr} + \widehat{\alpha}_{jr}(\mathbf{F}_j - \mathbf{F}_r) - \dfrac{\sigma}{\varepsilon}\widehat{\beta}_{jr}\mathbf{F}_r \\ \left(\displaystyle\sum_j \widehat{\alpha}_{jr} + \dfrac{\sigma}{\varepsilon}\widehat{\beta}_{jr}\right)\mathbf{F}_r = \displaystyle\sum_j l_{jr}E_j\mathbf{n}_{jr} + \widehat{\alpha}_{jr}\mathbf{F}_j \end{cases} \tag{4}$$

$\widehat{\alpha}_{jr}$ and $\widehat{\beta}_{jr}$ are defined by $\widehat{\alpha}_{jr} = l_{jr}\mathbf{n}_{jr} \otimes \mathbf{n}_{jr}$, $\widehat{\beta}_{jr} = l_{jr}\mathbf{n}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$ For the $\widehat{\alpha}_{jr}$ tensor we can use also the CHIC tensor [5].

The matrix $\sum_j l_{jr}\widehat{\alpha}_{jr}$ is always invertible on non-degenerate meshes. For the matrix $A_r = \sum_j l_{jr}\mathbf{n}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$ we do not have a complete result. However we give a sufficient condition for positivity of the matrix. We prove that $A_r = V_r\hat{I}_d + P$, where P is the matrix with $Tr(P) = 0$ and V_r is the control volume around the node r. Studying P we obtain the sufficient condition. For example, on triangular meshes the matrix is positive definite if the angles are superior to 11 degrees. In practice, these matrix are always non singular. Under the condition that the matrix is invertible we prove that the scheme is L^2 stable for the different tensor defined in [5].

The previous scheme tends to a new diffusion scheme on coarse grids

$$\begin{cases} E_j'(t) + \dfrac{1}{|\Omega_j|}\displaystyle\sum_r l_{jr}\left(\mathbf{n}_{jr}, \mathbf{F}_r\right) = 0, \\ A_r\sigma\mathbf{F}_r = \displaystyle\sum_j l_{jr}\mathbf{n}_{jr}E_j, \qquad\qquad \text{with } A_r = \left(\sum_j l_{jr}\mathbf{n}_{jr}\otimes\left(\mathbf{x}_r - \mathbf{x}_j\right)\right). \end{cases} \tag{5}$$

In [5] we prove the first order convergence to the solution of the linear P_1 model. The numerical results show a convergence at the second order on different unstructured meshes. This scheme may exhibit spurious modes but this problem can be solve by a modification of the normal and length associated to the node.

3 An AP scheme for the M_1 model on 2-D unstructured meshes

We reformulate the M_1 model as gas dynamics equations, see [4] in 1D, and we use a nodal solver as in [3, 6]. The new formulation writes

$$\begin{cases} \partial_t\rho + \dfrac{1}{\epsilon}div(\rho\mathbf{u}) = 0 \\ \partial_t\rho\mathbf{v} + \dfrac{1}{\epsilon}div(\rho\mathbf{u}\otimes\mathbf{v}) + \dfrac{1}{\epsilon}\nabla q = -\dfrac{\sigma}{\epsilon^2}\rho\mathbf{v} \\ \partial_t\rho e + \dfrac{1}{\epsilon}div(\rho\mathbf{u}e + q\mathbf{u}) = 0 \\ \partial_t\rho s + \dfrac{1}{\epsilon}div(\rho\mathbf{u}s) = 0 \end{cases} \tag{6}$$

with $S = \rho s$ (S the radiation entropy), $\mathbf{F} = \rho\mathbf{v}$ and $E = \rho e$. We define also the hydrodynamics variables

- $q = \dfrac{1-\chi}{2}E,$
- $\mathbf{u} = \dfrac{3\chi - 1}{2}\dfrac{\mathbf{f}}{|\,\mathbf{f}\,|^2}$

with $\mathbf{f} = \dfrac{|\,\mathbf{F}\,|}{E}$.

To discretize this model we use a Lagrange+remap scheme. The lagrangian step is solved by a nodal scheme which is a non-linear generalization of (11) coupled with the Jin-Levermore method

$$\begin{cases} M_j \dfrac{\tau_j^{n+1} - \tau_j^n}{\Delta t} - \dfrac{1}{\epsilon} \displaystyle\sum_r l_{jr} (\mathbf{u}_r, \mathbf{n}_{jr}) = 0 \\ M_j \dfrac{\mathbf{v}_j^{n+1} - \mathbf{v}_j^n}{\Delta t} + \dfrac{1}{\epsilon} \displaystyle\sum_r \mathbf{G}_{jr} = -\dfrac{\sigma}{\epsilon^2} \displaystyle\sum_r \widehat{\beta}_{jr} k_r \mathbf{u}_r \\ M_j \dfrac{e_j^{n+1} - e_j^n}{\Delta t} + \dfrac{1}{\epsilon} \displaystyle\sum_r (\mathbf{u}_r, \mathbf{G}_{jr}) = 0 \end{cases} \tag{7}$$

with the fluxes

$$\begin{cases} \mathbf{G}_{jr} = l_{jr} q_j \mathbf{n_{jr}} + r_{jr} \hat{\alpha}_{jr} (\mathbf{u_j} - \mathbf{u_r}) - \dfrac{\sigma}{\epsilon} k_r \hat{\beta}_{jr} \mathbf{u_r} \\ (\displaystyle\sum_j r_{jr} \hat{\alpha_{jr}} + \dfrac{\sigma}{\epsilon} k_r \hat{\beta}_{jr}) \mathbf{u_r} = \displaystyle\sum_j l_{jr} q_j \mathbf{n_{jr}} + r_{jr} \hat{\alpha}_{jr} \mathbf{u_j} \end{cases} \tag{8}$$

where $M_j =\mid \Omega_j \mid^{n+1} \rho^{n+1} =\mid \Omega_j \mid^n \rho^n$, and $k_r = \frac{2E_r |\mathbf{f_r}|^2}{(3\chi-1)}$, $\quad r_{jr} = \frac{4}{\sqrt{3}} \frac{E_j}{3+|\mathbf{u}|^2}$, r_{jr} is the wave-speed calculated for the one dimensional Riemann solver. In diffusive regime, the previous lagrangian scheme gives the following non-linear positive diffusion scheme.

$$\begin{cases} \mid \Omega_j \mid \dfrac{E_j^{n+1} - E_j^n}{\Delta t} + \displaystyle\sum_r \dfrac{1}{12\sigma} ((l_{jr} E_j \mathbf{n}_{jr} - \sigma \widehat{\beta}_{jr} \mathbf{u}_r), \dfrac{\mathbf{u}_r}{E_r}) = 0 \\ \sigma \left(\displaystyle\sum_j \hat{\beta}_{jr} \right) \mathbf{u}_r = \displaystyle\sum_j l_{jr} E_j \mathbf{n_{jr}} \end{cases} \tag{9}$$

This scheme can be seen as the non linear extension to the previous limit scheme. If $\widehat{\beta}_{jr} \simeq l_{jr} \mathbf{n}_{jr} \otimes (\mathbf{x}_r - \mathbf{x}_j)$ and $\mathbf{u}_r$ discretize correctly the gradient then

$$(l_{jr} E_j \mathbf{n}_{jr} - \sigma \widehat{\beta}_{jr} \mathbf{u}_r) \simeq l_{jr} E_r \mathbf{n}_{jr}$$

We obtain a result very close to the linear limit diffusion scheme. For the remap step, we can use any advection scheme. In asymptotic regime the lagrangian step gives a diffusion coefficient $\frac{1}{12\sigma}$. Studying the asymptotic limit we remark that $\mathbf{u}_r$ is homogeneous to $\frac{\nabla E}{4E\sigma}$ when ε tends to zero. Therefore the remap step gives a first order diffusion scheme with the coefficient $\frac{1}{4\sigma}$. The two steps are necessary to obtain the good diffusion coefficient. To obtain a second order scheme, we must use a MUSCL procedure with slop limiter in the remap step.

4 Numerical results

4.1 Numerical results for the limit diffusion scheme

We begin by study he convergence to the limit diffusion scheme. The studied scheme is the sum to (9) and the limit scheme of advection step. The initial condition is the fundamental solution of the heat equation at $t = 0.001$. the final time is $t = 0.01$. We obtain the following order of convergence K the coefficient of deformation for the Kershaw mesh. This results show that the scheme is a valid second order scheme on unstructured meshes.

Mesh	order	negative coef
Cartesian	1.92	0
Rand. quad. mesh	1.9	0
Cartesian trig. mesh	2.23	0
Rand. trig. mesh	2.16	0
Kershaw K=1	1.93	0
Kershaw K=1.5	2.02	0

Remark: Other test cases show that the scheme is convergent with the second order for th free-streaming regime ($\sigma = 0$) to the M_1 model.

4.2 Numerical results for radiation equilibrium models

The convergence results show that the limit scheme and the hyperbolic for all ϵ scheme are convergent. We solve

$$\begin{cases} \partial_t E + \dfrac{1}{\epsilon}\nabla.\mathbf{F} = \frac{\sigma}{\varepsilon^2}(aT^4 - E) \\ \partial_t \mathbf{F} + \dfrac{1}{\epsilon}\nabla(\hat{P}) = -\dfrac{\sigma}{\epsilon^2}\mathbf{F} \\ \rho C_v \partial_t T_j = -\frac{\sigma}{\varepsilon^2}(aT^4 - E) \end{cases} \tag{10}$$

Where T is the material temperature. We define $T_r = (E/a)^{\frac{1}{4}}$ the radiation temperature. To treat this model, we use a splitting strategy. The moment model part is solve with the previous scheme and the temperature relaxation part part is solve with a implicit fixed point procedure.

To obtain this implicit procedure we linearize in time the equation of T. We define $\Theta = aT^4$, consequently we obtain for the relaxation part the scheme

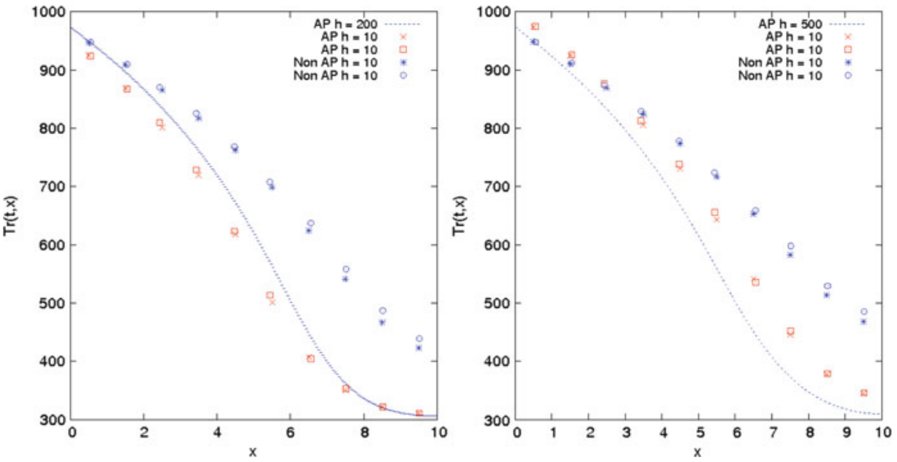


Fig. 2 The curve represent the solution T_r at the final time. The square and cross correspond to the solution with the AP correction on Cartesian and random mesh with 10 cells by direction. The point and circle correspond to the solution without the AP correction on Cartesian and random mesh with 10 cells by direction. At left this is the result for the P_1 model and at right for the M_1 model

$$\begin{cases} \frac{E_j^{q+1}-E_j}{\Delta t} = \frac{\sigma}{\varepsilon^2}(\Theta_j^{q+1} - E_j) \\ \rho C_v \mu_j \frac{E_j^{q+1}-E_j}{\Delta t} = -\frac{\sigma}{\varepsilon^2}(\Theta_j^{q+1} - E_j) \end{cases} \tag{11}$$

with $\mu_j = \frac{T_j^q - T_j}{\Theta_j^q - \Theta^q}$. This method is convergent and preserve the positivity of E and T.

We use the Marshak test describe in [1]. For the test case we consider a material initially cold and at radiative equilibrium. A heat wave enters the domain and we observe this evolution. The calculation is realized on a 2D mesh. We present the results for one line of cells. This test show the AP scheme capture the correct solution on coarse grid contrary to the classical scheme. The Fig. 2 show also that the scheme give the good result on random grid.

5 Conclusion

Starting from an asymptotic preserving scheme on 2-D unstructured meshes obtained in [5] for the linear P_1 model, we propose in this work its extension for the non-linear M_1 model. The scheme is valid on unstructured meshes, and is asymptotic preserving for the non-equilibrium regime (without coupling with matter) and for the equilibrium regime (with the coupling). Future works will be devoted to higher order angular discretization of the linear transport equations such as the discrete ordinates method or P_N equations.

References

1. C. Berthon, P. Charrier and B. Dubroca: An HLLC scheme to solve M_1 model of radiative transfer in two space dimensions, J. Scie. Comput., 31 (2007), pp. 347–389
2. J.L. Feugeas and B. Dubroca, Entropy moment closure hierarchy for the radiative transfer equation, C. R. Acad. Sci., Paris, Sér. I, Math. 329, No.10, 915-920 (1999).
3. P-H. Maire, R. Abgrall, J. Breil, J. Ovadia *A cell-centered lagragian scheme for two-dimensional compressible flow problems.* SIAM J. Sci. Comput. Vol 29, No. 4, pp. 1781-1824. 2007.
4. C. Buet, B. Després: A gas dynamics scheme for a two moments model of radiative transfer, Mathematical models and numerical methods for radiative transfer, Panorama er synthse 2009.
5. C. Buet, B. Després, E. Franck: Design of asymptotic preserving schemes for hyperbolic heat equation on unstructured meshes. Preprint LJLL UPMC, 2010.
6. G. Carré, S. Del Pino, B. Desprès, E. Labourasse: A Cell-centered lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension, JCP vol. 228 (2009) no14, pp. 5160-518.
7. L. Gosse, G. Toscani: An asymptotic-preserving well-balanced scheme for the hyperbolic heat equations, C. R. Acad. Sci Paris,Ser, I 334 (2002) 337-342.
8. G. Kluth, B. Després: Discretization of hyperelasticity on unstructured mesh with a cell-centered Lagrangian scheme. Journal of Computational Physics, Volume 229, December 2010
9. S. Jin, D. Levermore: Numerical schemes for hyperbolic conservation laws with stiff relaxation terms. JCP 126, 449-467, 1996.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Mass Conservative Coupling Between Fluid Flow and Solute Transport

Jürgen Fuhrmann, Alexander Linke, and Hartmut Langmach

Abstract We present a coupled discretization approach for species transport in an incompressible fluid. The Navier-Stokes equations for the flow are discretized by the divergence-free Scott-Vogelius element. The convection-diffusion equation for species transport is discretized by the Voronoi finite volume method. The species concentration fulfills discrete global and local maximum principles. We report convergence results for the coupled scheme and an application of the scheme to the interpretation of limiting current measurements in an electrochemical flow cell.

Keywords Incompressible Navier-Stokes Equations, Convection-Diffusion Equation, Finite Element Method, Finite Volume Method, Limiting Current
MSC2010: 76V05 76D05 65N08 65N30

1 Introduction

For the transport of a substance dissolved in a dilute solution in an incompressible fluid characterized by a velocity field $\mathbf{v}$, local mass conservation and maximum principle for the substance concentration c are directly connected to the solenoidal condition $\nabla \cdot \mathbf{v} = 0$ on the velocity field.

The Scott-Vogelius mixed finite element P_k-P_{k-1}^{disc} with order $k \geq 1$ for the Navier-Stokes equations guarantees a point-wise divergence-free discrete velocity field.

Upwinded Voronoi finite volume methods guarantee the desired qualitative properties for the discrete transport problem if the discrete velocity field fulfills a discrete counterpart of the solenoidal condition and if the underlying simplicial mesh fulfills

Jürgen Fuhrmann, Alexander Linke, and Hartmut Langmach
Weierstrass Institute, Mohrenstr. 39, 10117 Berlin, Germany,
e-mail: juergen.fuhrmann|alexander.linke|hartmut.langmach@wias-berlin.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_50, © Springer-Verlag Berlin Heidelberg 2011

the boundary conforming Delaunay property [1,2]. Recent developments in mesh generation [3,4] allow to consider this approach as a realistic option.

Using exact integration of the normal component of the discrete flow through the faces of the Voronoi volumes, we couple both schemes [5]. We discuss the application to the limiting current problem in a thin layer flow cell [6].

Let $\Omega \subset \mathbb{R}^d$ be a simply connected Lipschitz domain with $d \in \{2,3\}$. We regard the stationary, incompressible Navier-Stokes equations coupled to the equation of stationary transport of a dissolved species The flow is described using the steady, incompressible Navier-Stokes equations:

$$(\mathbf{v}\cdot\nabla)\mathbf{v} + \nabla p - \eta\Delta\mathbf{v} = \mathbf{f}, \qquad \nabla\cdot\mathbf{v} = 0. \tag{1}$$

Here, **v** is the fluid velocity, p is the pressure, η is the fluid viscosity, and **f** is a force vector. The steady transport of a species dissolved in the fluid is described by

$$\nabla\cdot\mathbf{q} = s, \quad \mathbf{q} = -(D\nabla c - c\mathbf{v}) \tag{2}$$

Here, **q** is the species molar flux, c is the species concentration, D is the diffusion coefficient, and s is a given source term.

The boundary conditions correspond to the limiting current problem in a flow cell [6]. Let $\mathscr{I}_\Gamma = \{A, I, O, S, W\}$ be a set of labels for boundary segments. We assume that the boundary $\Gamma = \partial\Omega = \bigcup_{i\in\mathscr{I}_\Gamma}\Gamma_i$ is subdivided into an inlet Γ_I, an outlet Γ_O, an anode Γ_A, and a symmetry boundary on Γ_S. For an illustration, see Fig. 1. The remaining part of Γ is assumed to consist of inert, impermeable walls Γ_W. We further assume that $\Gamma_A, \Gamma_I, \Gamma_O$ are separated from each other by sections belonging either to Γ_W or Γ_S. We impose the following boundary conditions:

Section	c	$(\mathbf{v}, p)$
Inlet Γ_I	$c = c_I(\mathbf{x})$	$\mathbf{v} = \mathbf{v}_I(\mathbf{x})$
Anode Γ_A	$c = 0$	$\mathbf{v} = \mathbf{0}$
Outlet Γ_O	$\frac{\partial c}{\partial\mathbf{n}} = 0$	$\eta\frac{\partial\mathbf{v}}{\partial\mathbf{n}} = p\mathbf{n}$
Symmetry Γ_S	$\frac{\partial c}{\partial\mathbf{n}} = 0$	$\mathbf{v}\cdot\mathbf{n} = 0,\ \frac{\partial(\mathbf{v}\cdot\mathbf{t})}{\partial\mathbf{n}} = 0$
Wall Γ_W	$\frac{\partial c}{\partial\mathbf{n}} = 0$	$\mathbf{v} = \mathbf{0}.$

(3)

The flow boundary condition at the outlet Γ_O states that the stress $\eta\nabla\mathbf{v} - p\cdot\mathrm{Id}$ projected onto the outward normal direction **n** is zero. For the concentration, it states that all solute transported to Γ_O by convection leaves the domain there [2].

Let $\Gamma_D^{\mathrm{NS}} = \Gamma_I \cup \Gamma_A \cup \Gamma_W$ denote the Dirichlet boundary for the Navier-Stokes equations. Let $\mathbf{v}_D$ be a vector function on Γ_D^{NS} which is defined by the corresponding boundary values in (3). By applying the differential operator to the extension of $\mathbf{v}_D$ into Ω, and adding the result to the right hand **f** side we derive a new right hand side, also denoted by **f** which allows to assume that the solution **v** is in the space $V = \{\mathbf{v} \in [H^1(\Omega)]^d | \mathbf{v} = 0 \text{ on } \Gamma_D^{\mathrm{NS}}, \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \Gamma_S\}$. The weak formulation of (1) arises as follows: Find $(\mathbf{v}, p) \in V \times L^2(\Omega)$ such that for all $(\mathbf{w}, q) \in V \times L^2(\Omega)$

$$\begin{aligned}\int_\Omega \eta\nabla\mathbf{v} : \nabla\mathbf{w}\,dx + \int_\Omega ((\mathbf{v}\cdot\nabla)\mathbf{v})\cdot\mathbf{w}\,dx + \int_\Omega p\nabla\cdot\mathbf{w}\,dx &= \int_\Omega \mathbf{f}\cdot\mathbf{w}\,dx \\ \int_\Omega q\nabla\cdot\mathbf{v}\,dx &= 0.\end{aligned} \tag{4}$$

The weak formulation of (2) relies on the particular choice of boundary conditions for $\mathbf{v}$. Let $\Gamma_D^T = \Gamma_A \cup \Gamma_I$ be the Dirichlet boundary for the transport equation, and let s be the right hand side containing the Dirichlet boundary conditions.

Let $W = \{c \in H^1(\Omega)|\ c|_{\Gamma_D^T} = 0\}$. Then we look for $c \in W$ such that for all $\phi \in W$,

$$\int_\Omega (D\nabla c - c\mathbf{v})\cdot\nabla\phi\,dx + \int_{\Gamma_O} \mathbf{v}\cdot\mathbf{n}c\phi\,ds = \int_\Omega s\phi dx. \tag{5}$$

2 Scott Vogelius mixed finite elements for fluid flow

Let $\bar{\mathscr{T}}_h$ denote a regular finite element triangulation of the domain Ω in the sense of [7], called macro triangulation. For each simplex $\bar{T} \in \bar{\mathscr{T}}_h$ we connect its barycenter with its vertices, and we thereby get $d + 1$ new simplices from each macro simplex.This new triangulation $\mathscr{T}_h$ is called an SV-admissible mesh. We define $V_h := \left\{\mathbf{v}_h \in [C(\Omega)]^d \cap V : \mathbf{v}_{h|T} \in [P_d(T)]^d\ \forall T \in \mathscr{T}_h\right\}$ as the space of continuous element-wise polynomial vector functions of order d on the triangulation $\mathscr{T}_h$. The pressure space $P_h := \left\{q \in L^2(\Omega) : q_{|T} \in P_{d-1}\ \forall T \in \mathscr{T}_h\right\}$ is defined as the space of element-wise polynomial functions of degree $d - 1$ without the constraint of continuity between elements. The derivation of the triangulation $\mathscr{T}_h$ from a macro-triangulation $\bar{\mathscr{T}}_h$ assures that the discrete saddle point problem derived from the Scott-Vogelius element has a unique solution by fulfilling in a stable manner the necessary and sufficient $\inf-\sup$ condition $0 < \beta \leq \beta_h = \inf_{p_h\in P_h, p_h\neq 0} \sup_{\mathbf{v}_h\in V_h} \frac{(\nabla\cdot\mathbf{v}_h, p_h)}{||p_h||\,||\mathbf{v}_h||}$ [8–10].

The discretization of of the Navier-Stokes equations is derived in a standard manner from (4): find $(\mathbf{v}_h, p_h) \in V_h \times P_h$ such that $\forall\ (\mathbf{w}_h, q_h) \in V_h \times P_h$,

$$\begin{aligned}\int_\Omega (\mathbf{v}_h\cdot\nabla)\mathbf{v}_h\cdot\mathbf{w}_h\,dx - \int_\Omega p_h\nabla\cdot\mathbf{w}_h\,dx + \int_\Omega \eta\nabla\mathbf{v}_h : \nabla\mathbf{w}_h\,dx &= \int_\Omega \mathbf{f}\cdot\mathbf{w}_h\,dx \\ -\int_\Omega q_h\nabla\cdot\mathbf{v}_h\,dx &= 0.\end{aligned} \tag{6}$$

3 Voronoi Finite Volumes for solute transport

Let $\partial\Omega$ be the union of straight lines resp. planar polygons. Let $\mathscr{P} = \{\mathbf{x}_K\} \subset \bar{\Omega}$ be a set of points which includes all the vertices of the polygons constituting $\partial\Omega$.

A simplicialization of this point set is Delaunay if no circumball of any simplex contains a point $\mathbf{x}_K$ of $\mathscr{P}$. Besides the fact that it is related to the same domain Ω, this simplicialization may be completely independent of the triangulations introduced in section 2. For a point $\mathbf{x}_K \in \mathscr{P}$, the Voronoi cell $V_K^0 \subset \mathbb{R}^d$ around $\mathbf{x}_K$ is defined as the set of points $\mathbf{x} \in \mathbb{R}^d$ which are closer to $\mathbf{x}_K$ than to any other point $\mathbf{x}_L$ of $\mathscr{P}$. We define as the control volume around $\mathbf{x}_K$ the Voronoi box V_K associated with $\mathbf{x}_K$ as $V_K = V_K^0 \cap \Omega$. The Delaunay simplicialization is boundary conforming [1] if

1. Ω is the union of all simplices;
2. no simplex circumball contains any other discretization vertex;
3. all simplex circumcenters are contained in $\bar{\Omega}$;
4. the boundary sections Γ_i ($i \in \mathscr{I}_\Gamma$) are the unions of simplex faces, and all circumcenters of boundary simplices from Γ_i are contained in $\bar{\Gamma}_i$.

We will use K in order to denote the Voronoi boxes V_K. Let $\mathscr{K}$ denote the set of control volumes K, and $\mathscr{K}_i$ denote the set of control volumes K which share facets with Γ_i. Let $\mathscr{K}_D = (\mathscr{K}_I \cup \mathscr{K}_A)$ denote the set of Dirichlet control volumes and $\mathscr{K}^0 = \mathscr{K} \setminus \mathscr{K}_D$ denote the set of non-Dirichlet control volumes. For two neighboring control volumes K, L, $\mathbf{x}_K\mathbf{x}_L$ is an edge of the boundary conforming Delaunay simplicialization which is known to be orthogonal to the Voronoi box face $\partial K \cap \partial L$. Let $\mathscr{N}_K$ denote the set of neighbors of K. For $i \in \mathscr{I}_\Gamma$, let $\mathscr{G}_K^i$ be the set of facets of K with nonempty intersection with boundary section Γ_i. Then $\partial K \cap \Gamma_i = \bigcup_{\sigma\in\mathscr{G}_K^i} \sigma$ and

$$\partial K = \left(\bigcup_{L\in\mathscr{N}_K} \partial K \cap \partial L \right) \cup \left(\bigcup_{i\in\mathscr{I}_\Gamma} (\cup_{\sigma\in\mathscr{G}_K^i} \sigma) \right).$$

Let $\mathbf{v} \in [H^1(\Omega)]^d$ fulfill the boundary conditions (3) and $\nabla \cdot \mathbf{v} = 0$. These conditions are fulfilled by every solution $\mathbf{v}$ of (4) and every solution $\mathbf{v}_h$ of (6). For any $K \in \mathscr{K}$, the H^1-regularity of $\mathbf{v}$ allows to define the scaled velocity projections

$$v_{KL} = \frac{1}{|\partial K \cap \partial L|} \int_{\partial K\cap\partial L} \mathbf{v} \cdot (\mathbf{x}_K - \mathbf{x}_L) ds, \qquad L \in \mathscr{N}_K \tag{7}$$

$$v_\sigma = \frac{1}{|\sigma|} \int_\sigma \mathbf{v} \cdot \mathbf{n}_\sigma ds, \qquad \sigma \in \mathscr{G}_K^i \tag{8}$$

They are discretely divergence-free in the sense that for all $K \in \mathscr{K}$ holds

$$\sum_{L\in\mathscr{N}_K} \frac{|\partial K \cap \partial L|}{|\mathbf{x}_K - \mathbf{x}_L|} v_{KL} + \sum_{i\in\mathscr{I}} \sum_{\sigma\in\mathscr{G}_K^i} |\sigma| v_\sigma = 0. \tag{9}$$

We introduce the space of functions $W_h = \{c_h \in L^2(\Omega) : c_h|_K = c_K\}$, consisting of scalar functions which are piecewise constant on each control volume.

For a given upwind function $U(z)$, the average normal flux of $\mathbf{q} = -D\nabla c + c\mathbf{v}$ between two neighboring control volumes K, L is approximated by a flux function $g(c_K, c_L, v_{KL}) = D\left(U\left(\frac{v_{KL}}{D}\right)c_K - U\left(-\frac{v_{KL}}{D}\right)c_L\right)$, depending on the values of the discrete solution in the adjacent control volumes and the velocity projection [2].

Further, we define the discrete right-hand side of the discrete convection-diffusion equation by the average value of the continuous right-hand side over the control volume K $s_K = \frac{1}{|K|}\int_L s(\mathbf{x})\,dx$. Then, the finite volume scheme for the transport equation (5) reads as: we look for $c_h \in W_h$ such that

$$\begin{cases} \sum_{L\in\mathscr{N}_K} \frac{|\partial K\cap\partial L|}{|\mathbf{x}_K-\mathbf{x}_L|} g(c_K, c_L, v_{KL}) + \sum_{\sigma\in\mathscr{G}_K^O} |\sigma| g(c_K, c_K, v_\sigma) = s_K & K \in \mathscr{K}^0 \\ c_K = c_D(\mathbf{x}_K), & K \in \mathscr{K}_D, \end{cases} \tag{10}$$

where the treatment of the outflow boundary conditions is taken from [2]. For $U(z) = U_{\text{dcd}}(z) = 1 + \frac{z}{2}$ we yield the central difference scheme. The simple upwind discretization is given by the upwind function $U_{\text{dsu}}(z) = 1 + \max\{0, z\}$. Our preferable choice is the Bernoulli function $U_{\exp}(z) = B(z) = \frac{z}{1-e^{-z}}$, leading to the the so-called exponential fitting scheme [11, 12].

4 Convergence of the coupled FVM-FEM scheme

The convergence results are given for homogeneous Dirichlet boundary conditions on $\Gamma = \partial\Omega$: $c|_\Gamma = 0 \quad \mathbf{v}|_\Gamma = \mathbf{0}$. As a consequence, in the weak formulations (4) and (5), we assume that $\Gamma = \Gamma_I$ and $\mathbf{v} \in V = [H_0^1(\Omega)]^d$ and $c \in H_0^1(\Omega)$.

We investigate a sequence of mesh pairs $(\mathscr{T}_h, \mathscr{V}_h)$ indexed by the mesh parameter h, and where $\mathscr{T}_h, \mathscr{V}_h$ are SV-admissible and boundary conforming Delaunay, respectively and possess uniform bounds for their respective mesh regularities. The only geometrical assumption relating both sequences is that there are h-independent constants C_1 and C_2 such that $C_1 h_{\text{FEM}}(h) \leq h_{\text{FVM}}(h) \leq C_2 h_{\text{FEM}}(h)$.

Theorem 1 (Finite Element Convergence).

1. *Equation (6) has at least one solution $(\mathbf{v}_h, p_h)$ on every SV-admissible grid.*
2. *For a sequence of Scott-Vogelius solutions $(\mathbf{v}_h)$ in (6) we can extract a subsequence which converges weakly in $V = [H_0^1(\Omega)]^d$ to some $\mathbf{v} \in V$. Moreover, this convergence is strong in $[L^2(\Omega)]^d$ and the limit $\mathbf{v}$ is divergence-free.*
3. *The limit $\mathbf{v}$ of said subsequence $(\mathbf{v}_h)_h$ is a solution of (4), and $\mathbf{v}_h \overset{H_0^1}{\to} \mathbf{v}$.*

Theorem 2 (Convergence of the coupled scheme). *We assume that $(\mathbf{v}_h, c_h)$ is a sequence of pairs of discrete solutions of (6) and (10) such that the sequence $(\mathbf{v}_h)$ converges strongly in $[H_0^1(\Omega)]^d$ to a solution $\mathbf{v}$ of (4).*

1. *From the sequence (c_h) we can extract a subsequence which converges strongly in $L^2(\Omega)$ to some $c \in H_0^1(\Omega)$.*
2. *The accumulation point $c \in H_0^1(\Omega)$ of said subsequence $(c_h)_h$ is the unique solution of the continuous problem (5), where the solution $\mathbf{v}$ of (4) drives the convection. Therefore, also the entire sequence (c_h) converges strongly in L^2 to the unique c, and not only a subsequence.*

The discretization matrix of (10) has the M property. Furthermore, (9) leads to

Lemma 1. *For any solution $(c_K)_{K\in\mathscr{K}}$ of (10) with $(s_K) = 0$, we have:*

1. *Global minimax principle:* $0 \leq c_K \leq c_I \quad \forall K \in \mathscr{K}$
2. *Local minimax principle:* $\min_{L\in\mathscr{N}_K} c_L \leq c_K \leq \max_{L\in\mathscr{N}_K} c_L \quad \forall K \in \mathscr{K}^0$

The convergence of the finite volume scheme for an analytically given velocity in the discrete L^2, discrete maximum, and discrete H^1 norms has been investigated [5]. On a mesh obtained from a rectangular mesh by subdividing each rectangle into two triangles, the exponential fitting and the central schemes exhibit second order convergence in all three norms, while the simple upwind scheme is first order. On a genuinely triangular mesh, first order convergence in the discrete H^1-norm for all three schemes has been indicated. Replacing the analytical velocities by velocities obtained using the Scott-Vogelius element resulted in the same asymptotic behavior.

5 Interpretation of a limiting current experiment

We report results from [6]. At the inlet Γ_I, a sulphuric acid (H_2SO_4) based electrolyte with given velocity profile $\mathbf{v}_I(\mathbf{x})$ derived from Poiseuille flow is injected with a concentration c_I of dissolved hydrogen H_2. At a certain potential applied between the anode Γ_A covered with a platinum catalyst, and a counter electrode placed in the electrolyte outside the domain of consideration, the part of the H_2 reaching Γ_A reacts immediately according to $H_2 \rightarrow 2H^+ + 2e^-$. The flow containing the unreacted H_2 leaves the cell at the outlet Γ_O. All H_2 reaching the anode Γ_A is consumed by the reaction, so homogeneous Dirichlet boundaries for the concentration are assumed. The source terms $\mathbf{f}, s$ in (1), (2) are zero. The geometry is depicted in Fig. 1. The symmetry of the cell allows to reduce the computational domain to one twelfth of the original problem by applying symmetry boundary conditions at the corresponding cut planes Γ_S. The anode current $I_E = 2F\int_{\Gamma_A} D\frac{\partial c}{\partial \mathbf{n}}ds$ is called the *limiting current*.

Figure 2 compares the concentration isosurfaces obtained with the Scott-Vogelius and Taylor-Hood Elements, respectively. We clearly see a striking difference concerning the preservation of the maximum principle.

Figure 3 (left) shows the maximum concentration vs. flow rate for the two finite element discretizations. For the Taylor-Hood element, we are unable to control the violation of the maximum principle. For the Scott-Vogelius element, we see that the a-priori bound for the concentration given by the inlet velocity is observed.

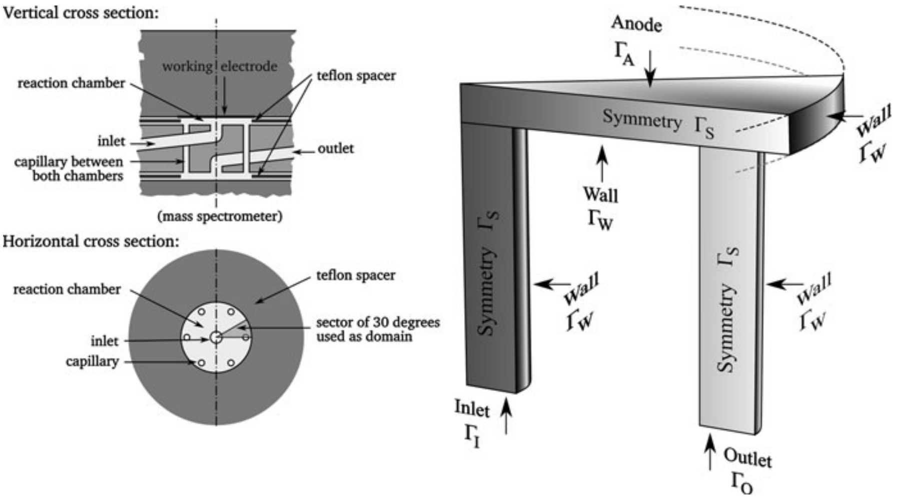


Fig. 1 Left: Schematic of a thin layer flow cell [13]. By symmetry, the problem is reduced to the 30 degrees (gray) circular arc shown. Right: computational domain with boundary segments. Reprinted with permission from [5]

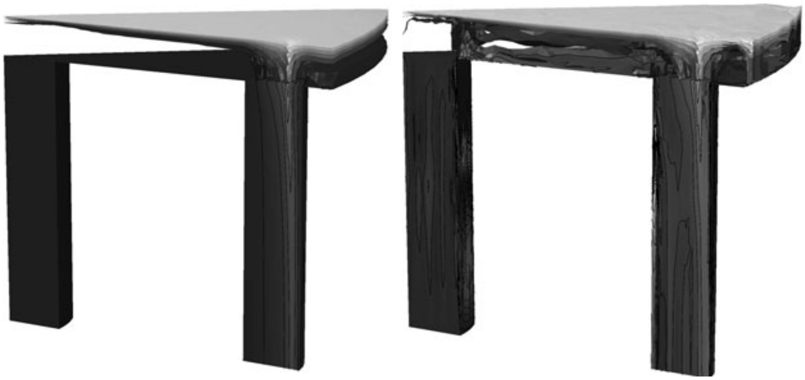


Fig. 2 Concentration profiles for flow rate $80mm^3/s$ on a coarse grid: Flow calculated using Scott-Vogelius element (left) and Taylor-Hood element (right). Isosurfaces ($c = 1.0, 2.0 \ldots 6.0$) are shown in the interior of the working chamber. Isolines and grayscale color code at surfaces are shown at the inlet, the outlet, and the bottom of the working chamber. The graphical representation has been stretched by a factor around 10 in z direction. Reprinted with permission from [5]

The right plot in Fig. 3 compares the values of the limiting current for different grids and discretizations with those measured in [14]. The grid dependency of this value is well below the accuracy of the experimental data [6]. At the same time one observes that the violation of the maximum principle does not significantly influence the value of the limiting current.

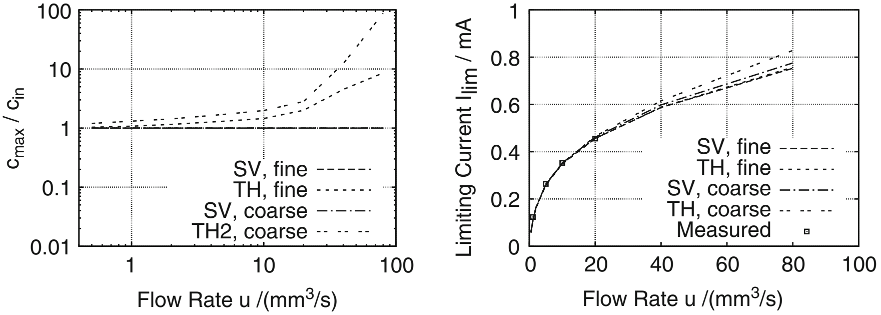


Fig. 3 Maximum concentration vs. flow rate (left). Measured [14] and calculated limiting current for different grids and discretizations (right). Reprinted with permission from [5]

6 Conclusions and Outlook

We presented a new scheme allowing for mass conservative coupling of solute transport and Navier-Stokes flow. It shows the expected convergence properties, and can be used in relevant applications. The approach has some drawbacks. Whereas in the implementation of the scheme (10), only the areas $|\partial K \cap \partial L|$ are used, which can be assembled from simplicial contributions, for the velocity projections (7), the entities $\partial K \cap \partial L$ need to be constructed [5]. The Scott Vogelius element is expensive. Static condensation may allow for more efficient assembly. There may be other routes to the discrete solenoidal condition (9) – for the tangential velocity MAC scheme [15], it exactly corresponds to the discretization of the mass balance for fluid flow.

References

1. H. Si, K. Gärtner, and J. Fuhrmann. Boundary conforming Delaunay mesh generation. *Comput. Math. Math. Phys.*, 50:38–53, 2010.
2. J. Fuhrmann and H. Langmach. Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. *Appl. Numer. Math.*, 37(1–2):201–230, 2001.
3. J. R. Shewchuk. triangle version 1.6. URL: http://www.cs.cmu.edu/~quake/triangle.html, 2007. Retrieved 2007-09-26.
4. H. Si. TetGen version 1.4.2. URL: http://tetgen.berlios.de/, 2010. Retrieved 2010-09-21.
5. J. Fuhrmann, A. Linke, and H. Langmach. A numerical method for mass conservative coupling between fluid flow and solute transport. *Appl. Numer. Math.*, 61(4):530 – 553, 2011.
6. J. Fuhrmann, A. Linke, H. Langmach, and H. Baltruschat. Numerical calculation of the limiting current for a cylindrical thin layer flow cell. *Electrochimica Acta*, 55(2):430–438, 2009.
7. P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*, volume 4 of *Studies in Mathematics and its Applications*. North-Holland, 1978.
8. J. Qin. *On the convergence of some low order mixed finite elements for incompressible fluids*. PhD thesis, Penn. State Univ., 1994.

9. D. N. Arnold and J. Qin. Quadratic velocity/linear pressure Stokes elements. In R. Vichnevetsky, D. Knight, and G. Richter, editors, *Advances in Computer Methods for Partial Differential Equations VII*, pages 28–34. IMACS, 1992.
10. S. Zhang. A new family of stable mixed finite elements for the 3D Stokes equations. *Math. Comp.*, 74(250):543–554, 2005.
11. D. N. Allen and R. V. Southwell. Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. and Appl. Math.*, 8:129–145, 1955.
12. A. M. Il'in. A difference scheme for a differential equation with a small parameter multiplying the second derivative. *Mat. zametki*, 6:237–248, 1969.
13. Z. Jusys, H. Massong, and H. Baltruschat. A new approach for simultaneous DEMS and EQCM: Electrooxidation of adsorbed CO on Pt and Pt-Ru. *J. Electrochem. Soc.*, pages 1093–1098, 1999.
14. H. Wang. *Electrocatalytic oxidation of adsorbed CO and methanol on Mo, Ru and Sn modified poly- and mono-crystalline platinum electrodes: A quantitative DEMS study (chinese)*. PhD thesis, Beijing Normal Univ., 2001.
15. R. Eymard, A. Linke, and J. Fuhrmann. MAC schemes on triangular meshes. In *Finite Volumes for Complex Application VI*. Springer, 2011. submitted.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Large Eddy Simulation of the Stable Boundary Layer

Vladimír Fuka and Josef Brechler

Abstract The model CLMM (Charles University Large-eddy Microscale Model) is a large-eddy simulation model for atmospheric flows. It solves Navier-Stokes equations for incompressible flow using the projection method and the 3rd order Runge-Kutta method in time. The spatial discretization is performed using the finite volume method on a uniform staggered grid.

The capability of the model to compute flows influenced by buoyancy is evaluated in this study in the case of stable stratification of the planetary boundary layer. The results are compared to the results of the project GABLS [2] with a good agreement.

Keywords large eddy simulation, planetary boundary layer, stable stratification, atmosphere
MSC2010: 76F40, 76F45, 76F65

1 Introduction

The large eddy simulation (LES) has been an important tool in boundary layer meteorology for several decades [4]. The presented model CLMM (Charles University Large-eddy Microscale Model) is a nonhydrostatic model for flows in the planetary boundary layer (PBL) and uses LES as it's main framework. It has been extended for the effects of buoyancy (or temperature stratification). The aim of this study is to present the numerical methods used in the dynamical core of the model and evaluate

Vladimír Fuka and Josef Brechler
Dep. of Meteorology and Environment Protection, Fac. of Mathematics and Physics, Charles University, V Holešovičkách 2, 18000, Prague 8, Czech Republic, e-mail: vladimir.fuka@mff.cuni.cz, josef.brechler@mff.cuni.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_51, © Springer-Verlag Berlin Heidelberg 2011

it's results in the situations influenced by buoyancy. All computations presented here are performed above a flat homogeneous terrain for simplicity.

2 Numerical methods

The model CLMM solves the Navier-Stokes equations for incompressible flow in the Boussinesq approximation. These equations in the filtered form for the use in LES are as follows

$$\frac{\partial \overline{u_i}}{\partial t} + \frac{\partial \overline{u}_i \overline{u}_j}{\partial x_j} = -\frac{\partial \overline{p}}{\partial x_i} + \frac{\partial \tau_{ij}}{\partial x_j} + \delta_{i3} \frac{g}{\theta_{\mathrm{ref}}} (\overline{\theta} - \theta_{\mathrm{ref}}) + f \epsilon_{ij3} \overline{u}_j \tag{1}$$

$$\frac{\partial \overline{\theta}}{\partial t} + \frac{\partial \overline{\theta} \overline{u}_j}{\partial x_j} = \frac{\partial q_j}{\partial x_j} \tag{2}$$

$$\tau_{ij} = \overline{u}_i \overline{u}_j - \overline{u_i u_j} \tag{3}$$

$$q_i = \overline{u}_i \overline{\theta} - \overline{u_i \theta}, \tag{4}$$

$$\frac{\partial \overline{u}_i}{\partial x_i} = 0 \tag{5}$$

where θ_{ref} is the reference potential temperature, f the Coriolis parameter and τ_{ij} and q_i are the subgrid stress tensor and the subgrid temperature fluxes respectively. The molecular viscosity and diffusivity is neglected. CLMM uses implicit filtering, i.e. the use of finite grid is considered as a sort of filtering. The overlines will be omitted hereinafter.

The solution of equations (1-5) is based on the method of lines (MOL), i.e. on discretization of time and space separately. The time discretization is based on a projection method [3] and the 3rd order low storage Runge-Kutta method combined with the Crank-Nicolson method [7]. The semi-discretized system can be written as

$$\begin{aligned}\frac{\hat{u}_i^k - u_i^{k-1}}{\Delta t} &= -\gamma_k \left[\frac{\partial u_i u_j}{\partial x_j}\right]^{k-1} - \rho_k \left[\frac{\partial u_i u_j}{\partial x_j}\right]^{k-2} - \\ &- \alpha_k \frac{\partial p}{\partial x_i} + \frac{\alpha_k}{2}\left(\frac{\partial \tau_{ij}^k}{\partial x_j} + \frac{\partial \tau_{ij}^{k-1}}{\partial x_j}\right) + \alpha_k f \epsilon_{ij3} u_j + \\ &+ \gamma_k \delta_{i3} \frac{g}{\theta_{\mathrm{ref}}} (\theta^{k-1} - \theta_{\mathrm{ref}}) + \rho_k \delta_{i3} \frac{g}{\theta_{\mathrm{ref}}} (\theta^{k-2} - \theta_{\mathrm{ref}})\end{aligned} \tag{6}$$

$$\frac{\partial^2 \varphi}{\partial x_i^2} = \frac{1}{\alpha_k \Delta t} \frac{\partial \hat{u}_i}{\partial x_i} \tag{7}$$

$$u_i^k = \hat{u}_i^k - \alpha_k \Delta t \frac{\partial \varphi}{\partial x_i}, \tag{8}$$

$$p^k = p^{k-1} + \varphi - \frac{\alpha_k \Delta t \nu_t}{2} \frac{\partial^2 \varphi}{\partial x_i^2}, \tag{9}$$

$$\frac{\theta^k - \theta^{k-1}}{\Delta t} = -\gamma_k \left[\frac{\partial \theta u_j}{\partial x_j}\right]^{k-1} - \rho_k \left[\frac{\partial \theta u_j}{\partial x_j}\right]^{k-2} + \\ + \alpha_k \left(\frac{1}{2}\frac{\partial q_j^k}{\partial x_j} + \frac{1}{2}\frac{\partial q_j^{k-1}}{\partial x_j}\right), \tag{10}$$

where

$$k = (1, 2, 3), \tag{11}$$

$$\gamma_k = (8/15, 5/12, 3/4), \tag{12}$$

$$\rho_k = (0, -17/60, -5/12), \tag{13}$$

$$\alpha_k = (8/15, 2/15, 1/3) \tag{14}$$

$$\tag{15}$$

and $\hat{u}_i$ and φ are auxiliary intermediate variables. The $\hat{u}_i$ does not fulfill the continuity equation (4) and is corrected in the latter steps.

The spatial discretization is carried out using the finite volume method on a uniform staggered grid. The standard second order central differences are used for most of the terms. In the case of scalar advection this method is not adequate because negative values and spurious oscillations have to be avoided at the cost of slightly increased numerical diffusion. For this reason CLMM employs a third order non-split semi-discrete advection method [5] which employs a flux limiter. This method is conservative, positive, but is not TVD. It still prevents the spurious oscillations to emerge and for 1D problems TVD and positive schemes may be equivalent [8].

The Poisson equation (7) is solved using a multigrid method with the Gauss-Seidel smoother and the Gaussian elimination solver on the smallest grid.

A crucial part in LES is the evaluation of the subgrid stresses. Many approaches are possible, but the eddy viscosity models are the basic and still widely used ones [6]. In these models it is assumed, that the subgrid stresses and fluxes are correlated to the strain rates and gradients in the same way, as in the case of molecular diffusion. CLMM uses the Vreman [9] algebraic model. It is simple to use, but it's results are claimed to be close to that of a dynamic model. The eddy viscosity is computed using the equation

$$\nu_t = c\sqrt{\frac{B_\beta}{\alpha_{ij}\alpha_{ij}}}, \tag{16}$$

where

$$\alpha_{ij} = \frac{\partial \overline{u}_j}{\partial x_i}, \tag{17}$$

$$\beta_{ij} = \sum_m \Delta_m^2 \alpha_{mi}\alpha_{mj} \tag{18}$$

$$B_\beta = \beta_{11}\beta_{22} - \beta_{12}^2 + \beta_{22}\beta_{33} - \beta_{23}^2 + \beta_{33}\beta_{11} - \beta_{31}^2. \tag{19}$$

The constant c was set to 0.05 in present computations. The temperature diffusivity is computed using a constant subgrid Prandtl number $\mathrm{Pr}_{\mathrm{sgs}} = 0.5$ in all presented cases.

At the surface the flow cannot be accurately resolved and the subgrid terms have to be computed using a wall model. because of the buoyancy effects the Monin-Obukhov similarity theory is employed [1]. The surface stress and the surface temperature flux are computed using the following expressions

$$\frac{U}{u_\star} = \frac{\ln(z/z_0) - \Psi_{\mathrm{M}}(z/L)}{\kappa} \tag{20}$$

$$\frac{\theta - \theta_0}{\theta_\star} = \frac{\ln(z/z_0) - \Psi_{\mathrm{H}}(z/L)}{\kappa} \tag{21}$$

where $\kappa = 0.4$ is the von Kármán constant, z_0 is the roughness parameter, θ_0 is the surface potential temperature, $u_\star = \sqrt{\tau_0}$ is the friction velocity and $\theta_\star = -(\overline{w'\theta'})_0/u_*$ is the friction temperature, τ_0 is the surface stress and

$$L = -\frac{u_\star^3 \theta_0}{\kappa g (\overline{w'\theta'})_0} \tag{22}$$

is the Obukhov length. The empirical similarity functions are set according to the GABLS [2] recommendation

$$\Psi_{\mathrm{M}} = -4.8\frac{z}{L}, \tag{23}$$

$$\Psi_{\mathrm{H}} = -7.8\frac{z}{L}. \tag{24}$$

3 Boundary and initial conditions

The boundary and initial conditions in the present study follow the GABLS intercomparison project. It is based on the results of the Beaufort Sea Arctic Stratus Experiment (BASE) and should approximate a quasi-stationary stable boundary layer over sea ice. The Coriolis parameter was set a value at the latitude of 73° north and the roughness parameter was set to a value of $z_0 = 0.1$ m.

The computational domain measured 400 m in all three dimensions. The computations were been carried out using resolutions $16 \times 16 \times 17$, $32 \times 32 \times 33$, $64 \times 64 \times 65$ and $128 \times 128 \times 129$. The vertical resolution is different from the horizontal one due to a limitation of the multigrid solver in various boundary conditions.

The boundary conditions at the limits of the domain in x and y directions were periodic. At the upper boundary the free-slip condition was used with a sponge layer damping the oscillations in the upper 100 meters.

The initial conditions consists of a neutrally stratified layer in the lowest 100 m and a stable layer with the lapse rate of 0.01 K/m. In the lowest 50 m random fluctuations with amplitude 0.1 K are applied to start-up turbulence. The initial surface temperature is 265 K and drops with a cooling rate of 0.25 K per hour.

4 Results

The model was run with described conditions for 9 hours. The last one hour was used for computing statistics. In the next paragraphs the results of CLMM are combined with the results of the groups participating in the project GABLS. All presented profiles are averaged temporally on the last one hour and spatially on horizontal planes.

4.1 Mean quantities

In Fig. 1 are the profiles of potential temperature at different resolutions. It is obvious, that a proper grid convergence has not been achieved.The GABLS paper [2] suggest the importance of even larger resolution. For the finest computed grid the comparison with GABLS results in Fig. 2 shows noticeable difference in the temperature gradient in the boundary layer. This value is sensitive to the choice of the subgrid model [2] and should be investigated more in the future.

For the wind velocity magnitude (Fig. 3 and the wind direction (Ekman spiral, Fig. 4) the agreement with GABLS results is better. The super-geostrophic jet and the wind turning in the boundary layer are well pronounced.

4.2 Turbulent fluxes

The vertical buoyancy and momentum fluxes are depicted in Figs. 5 and 6 respectively. In both cases the profiles follow the shape and fall within the range

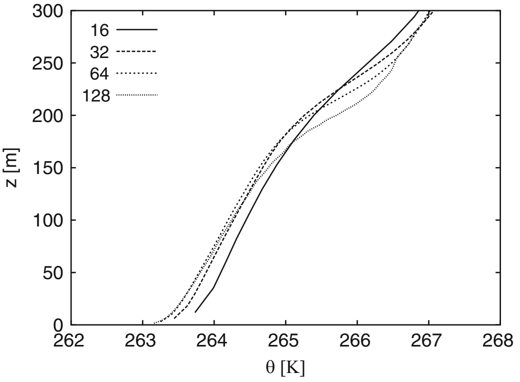


Fig. 1 Grid convergence study for the vertical temperature profile. Other variables yield similar results

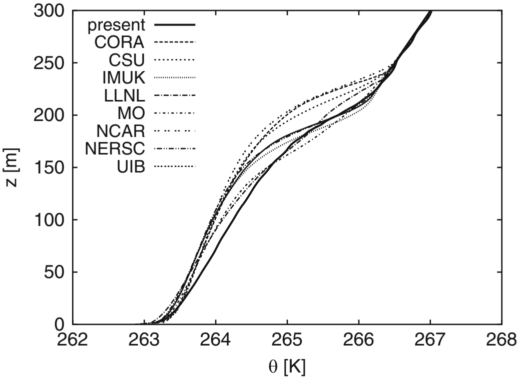


Fig. 2 The vertical temperature profile in comparison with the GABLS results

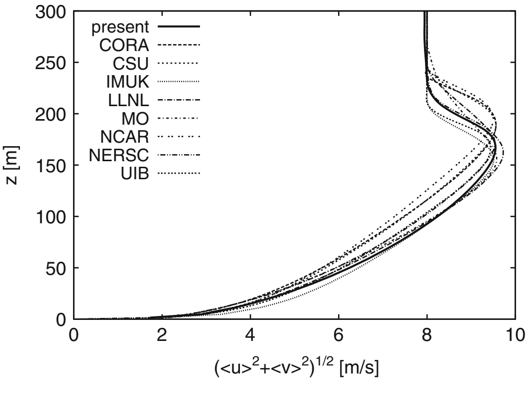


Fig. 3 The vertical profile of the wind velocity in comparison with the GABLS results

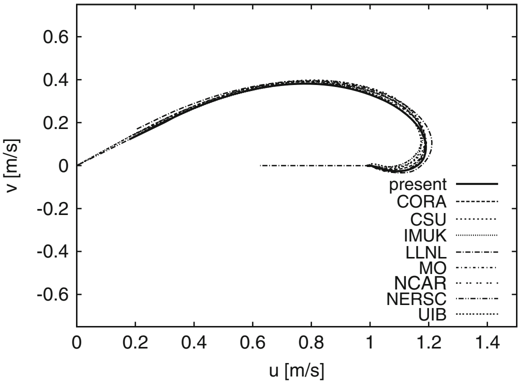


Fig. 4 The Ekman spiral (graph of horizontal velocity components) in comparison with the GABLS results

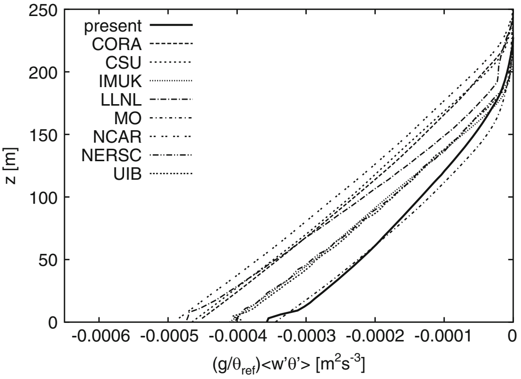


Fig. 5 The vertical buoyancy flux profile in comparison with the GABLS results

of the referenced GABLS simulations. The value of the fluxes is at the lower side of the range.

The gradient Richardson number and the flux Richardson number profiles are in Fig. 7. Their values are almost equal throughout the boundary layer reaching approximately the critical value 0.25 at it’s top.

5 Conclusion

The ability of the model CLMM to simulate turbulent flow in the stable boundary layer has been tested. The results agree to those of model intercomparison initiative GABLS. Some inconsistency has been found in the temperature gradient in the boundary layer and will be investigated further. In next development the model will

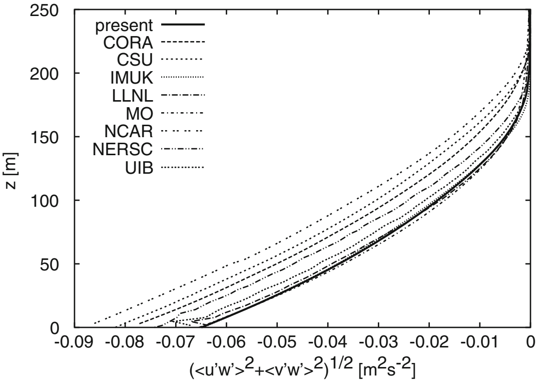


Fig. 6 The vertical momentum flux profile in comparison with the GABLS results

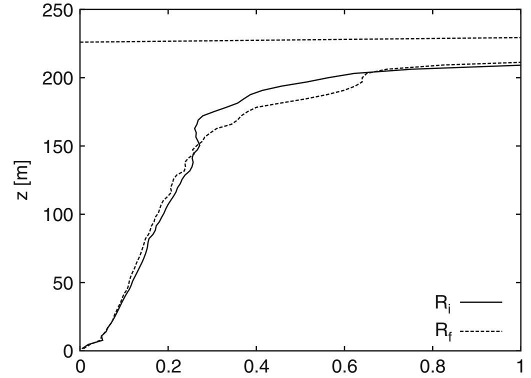


Fig. 7 The vertical profile of the gradient and flux Richardson number

be extended for inhomogeneous or spatially developing flows and for flows over a complex terrain. The model is also aimed to atmospheric dispersion studies.

Acknowledgements The work was supported by grant SVV-2010-261308 and by the Czech Ministry of Education, Youth and Sports under the framework of research plan MSM0021620860. The authors would like to thank Ing. Petr Bauer for valuable help with the implementation of the multigrid method.

References

1. Basu, S., Holtslag, A., van de Wiel, B., Moene, A., Steeneveld, G.J.: An inconvenient truth about using sensible heat flux as a surface boundary condition in models under stably stratified regimes. Acta Geophys. **56**(1), 88–99 (2008)

2. Beare, R.J., Macvean, M.K., Holtslag, A.A.M., Cuxart, J., Esau, I., Golaz, J.C., Jimenez, M.A., Khairoutdinov, M., Kosovic, B., Lewellen, D., Lund, T.S., Lundquist, J.K., Mccabe, A., Moene, A.F., Noh, Y., Raasch, S., Sullivan, P.: An intercomparison of large-eddy simulations of the stable boundary layer. Boundary-Layer Meteorology **118**(2), 247–272 (2006)
3. Brown, D.L., Cortez, R., Minion, M.L.: Accurate projection methods for the incompressible Navier-Stokes equations. J. Comput. Phys. **168**, 464–499 (2001)
4. Deardorff, J.W.: Numerical investigation of neutral and unstable planetary boundary layers. J. Atmos. Sci. **29**, 91–115 (1972)
5. Hundsdorfer, W., Koren, B., van Loon, M., Verwer, J.G.: A Positive Finite-Difference Advection Scheme. J. Comput. Phys. **117**(1), 35–46 (1995)
6. Lilly, D.K.: The representation of small-scale turbulence in numerical simulation experiments. Proc. IBM Sci. Comput. Symp. on Environ. Sci. **29**, 91–115 (1967)
7. Spalart, P.R., Moser, R.D., Rogers, M.M.: Spectral methods for the Navier-Stokes equations with one infinite and two periodic directions. J. Comput. Phys. **96**(2), 297–324 (1991). DOI 10.1016/0021-9991(91)90238-G
8. Thuburn, J.: TVD schemes, positive schemes, and the universal limiter. Monthly Weather Review **125**(8), 1990–1993 (1997)
9. Vreman, A.W.: An eddy-viscosity subgrid-scale model for turbulent shear flow: Algebraic theory and applications. Phys. Fluids **16**(10), 3670–3681 (2004)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

3D Unsteady Flow Simulation with the Use of the ALE Method

Petr Furmánek, Jiří Fürst, and Karel Kozel

Abstract This works deals with three-dimensional numerical simulation of transonic and subsonic inviscid compressible steady and unsteady flow. The problem is solved using finite volume method, namely the so–called Modified Causon's scheme [3] in combination with Arbitrary Lagrangian–Eulerian method [5]. This scheme is based on TVD form of classical MacCormack scheme. Although it is not TVD it retains almost the same precision as the original TVD scheme, but demands approximately 30% less computational memory and power. Both subsonic and transonic regimes of flow over oscillating wings are simulated. The subsonic case (flow over the AS28 wing) is compared with experimental data with a very good agreement. Comparison for the transonic unsteady case (flow over the ONERA M6 wing) is unfortunately not possible, but numerical results show very good properties.

Keywords ALE, FVM, TVD, unsteady flow
MSC2010: 65M08, 65Y20

1 Introduction

Unsteady effects appear in many physical phenomena including flows in external aerodynamics. Their appearance usually entails very unpleasant problems, sometimes even with fatal consequences (e.g. flutter). It is therefore necessary to research unsteady behaviour of the flow - both forced and induced. The authors made a series of numerical experiments featuring subsonic and transonic flow over an oscillating wing using the finite volume method (FVM) [4] in combination with the Arbitrary

Petr Furmánek
VZLÚ a.s., Beranovch 130, 199 05 Praha - Letany, e-mail: petr.furmanek@fs.cvut.cz

Jiří Fürst and Karel Kozel
ÈVUT v Praze, Fakulta strojní, Ústav technick matematiky, Karlovo námstí 13, 12135 Praha,
e-mail: jiri.furst@fs.cvut.cz, kozelk@fsik.cvut.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_52, © Springer-Verlag Berlin Heidelberg 2011

Lagrangian–Eulerian method (ALE) in order to study behaviour of the flow field and its development towards unsteady state.

2 Mathematical Model

The flow was considered inviscid and compressible and hence system of the Euler equations was employed as a mathematical model. It can be written down in the following conservative vector form:

$$W_t + F(W)_x + G(W)_y + H(W)_z = 0, \tag{1}$$

where subscripts denote partial derivatives and

$$\begin{aligned} W &= \big(\rho, \rho u, \rho v, \rho w, e\big)^T, \\ F(W) &= \big(\rho u, \rho u^2 + p, \rho uv, \rho uw, (e+p)u\big)^T, \\ G(W) &= \big(\rho v, \rho uv, \rho v^2 + p, \rho vw, (e+p)v\big)^T, \\ H(W) &= \big(\rho w, \rho uw, \rho vw, \rho w^2 + p, (e+p)w\big)^T. \end{aligned} \tag{2}$$

W is vector of conservative variables with components: ρ - density, $\mathbf{w} = (u, v, w)$ - velocity vector, e - total energy per unit volume and p - static pressure. F, G, H are inviscid fluxes. System (1) is enclosed by the Equation of State:

$$p = (\gamma - 1)\Big[e - \frac{1}{2}\rho(u^2 + v^2 + w^2)\Big], \quad \gamma = \frac{c_p}{c_v}. \tag{3}$$

where c_p and c_v are specific heat capacities under constant pressure (at constant volume).

3 Numerical Method

When solving (1) by the finite volume method for the case of steady flow the computational domain Ω is divided into a number of quadrilateral cells D_i such that $\Omega = \bigcup_i D_i$ and $i \in \langle 1, N_i \rangle$ where N_i is total number of cells. For each i the following relation must be fulfilled

$$\frac{d}{dt}\int_{D_i} W \, d\Omega_X + \int_{\partial D_i} \big(F(W)_x, G(W)_y, H(W)_z\big) \cdot \mathbf{n} \, d\Omega_S = 0, \tag{4}$$

where $\mathbf{n}$ is unit normal outer vector of D_i. In the unsteady case system (4) is altered in order to meet the needs of the ALE method. Computational cells D_i are now time-dependent and

$$\frac{d}{dt}\int_{D_i(t)} W\, d\Omega_X + \int_{\partial D_i(t)} \big(\tilde{F}(W,w_1)_x, \tilde{G}(W,w_2)_y, \tilde{H}(W,w_3)_z\big)\cdot \mathbf{n}\, d\Omega_S = 0 \quad (5)$$

where

$$\begin{aligned} \tilde{F}(W,w_1)_x &= F(W)_x - w_1 W, \\ \tilde{G}(W,w_2)_y &= G(W)_y - w_2 W, \\ \tilde{H}(W,w_3)_z &= H(W)_z - w_3 W. \end{aligned} \quad (6)$$

(w_1, w_2, w_3) is velocity of mesh vertices during motion [5]. System (5) is now solved by the Modified Causon's scheme in ALE formulation [3]. This cell-centred scheme is derived from TVD form of the MacCormack scheme. It is not TVD but saves approximately 30% of computational time and memory with almost no loss in accuracy. The ALE method uses computation on moving meshes and hence an algorithm for mesh modification is needed. The actual position of mesh vertices $\mathbf{x}_i$ is in our case given by the following prescription

$$\mathbf{x}_i(t) = \mathbb{Q}\big[\phi(t, ||\mathbf{x}_i(0) - \mathbf{x}_{ref}||)\big](\mathbf{x}_i(0) - \mathbf{x}_{ref}) + \mathbf{x}_{ref}, \quad (7)$$

where

$$\mathbb{Q}(\phi) = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \quad (8)$$

and

$$\phi(t,r) = \begin{cases} -\alpha_1(t) & \text{for } r < r_1, \\ -\alpha_1(t) f_D(r) & \text{for } r_1 \le r < r_2, \\ 0 & \text{for } r_2 < r. \end{cases} \quad (9)$$

where

$$f_D(r) = \left[2\left(\frac{r-r_1}{r_2-r_1}\right)^3 - 3\left(\frac{r-r_1}{r_2-r_1}\right)^2 + 1\right] \quad (10)$$

The computational area is divided into three regions by spheres (or hemispheres) with different radius. The hemisphere with centre in $\mathbf{x}_{ref}$ and radius r_1 is rotating according to the prescribed change of pitching angle as a solid body. Outer area of the second hemisphere with radius $r_2 > r_1$ is motion-less and in space between these two hemispheres motion of the mesh is damped by damping function $f_D(\cdot)$. Wing moves according to the following prescription for pitching angle:

$$\alpha_1(t) = \alpha_{init} + A\sin(\omega t) \quad (11)$$

with angular velocity

$$\omega = \frac{2\pi k U_\infty}{c}, \tag{12}$$

$U_\infty = M_\infty$ is the free-stream velocity, c is chord length (in the wing-root) and k is reduced frequency (or $\omega = 2\pi f$ with f being real /dimensional/ frequency). In both simulated cases structured C-mesh was used for discretization of the computational domain. In the case of the AS28 wing the mesh consisted from 396000 cells, in the case of the ONERA M6 wing it was made from 493000 cells.

4 Numerical Results

Forced oscillations of the wing were both in subsonic and transonic regimes given by formally the same relation (11) but with various values of α_{init}, A and ω.

4.1 Unsteady Subsonic Flow over the AS28 wing

The initial conditions for unsteady subsonic flow over the AS28 wing were as follows: inlet Mach number $M_\infty = 0.51$, $\alpha_{init} = -0.5°$, $f = 45\text{Hz}$, $A = 3°$. The wing oscillated around reference axis parallel with wing span and going through point $\mathbf{x}_{ref} = [0.25, 0, 0]$. Numerical and experimental results were compared on behaviour of lift coefficient in cuts along the wing (Fig. 2). Development of the periodic state can be observed on Fig. 1.

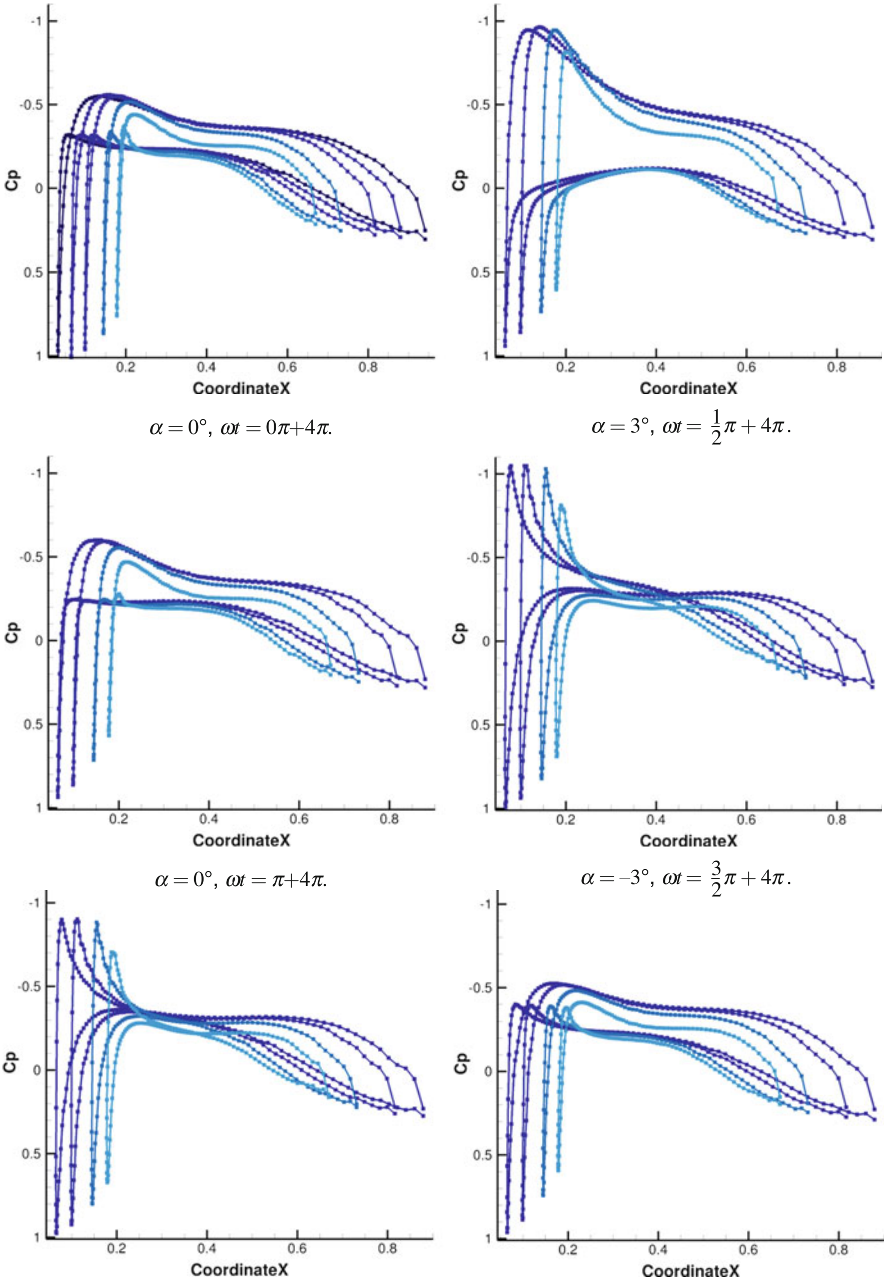
4.2 Steady Transonic Flow over the ONERA M6 wing

A well-known test case published in AGARD report no. 138 [1] and characterised by inlet Mach number $M_\infty = 0.8395$ and angle of attack $\alpha_{init} = 3.06°$ was chosen as initial condition for unsteady transonic flow computation. Numerical results obtained by the MCS scheme are compared to results of WLSQR scheme [6] with HLLC numerical flux [7] and also to the experimental data (Fig. 3). Agreement between numerical and experimental results is more than satisfactory.

4.3 Unsteady Transonic Flow over the ONERA M6 wing

Simulation of transonic flow over the ONERA M6 wing was based on a test case mentioned Sect. 4.2 with initial conditions: $M_\infty = 0.8395$, $\alpha_{init} = 3.06°$, $f = 10\text{Hz}$, $A = 1.5°$. The wing oscillated around reference axis parallel with its span, this time going through point $\mathbf{x}_{ref} = [0.35, 0, 0]$.

As can be seen from Figs. 1 to 4 the scheme delivers very good results for both steady and unsteady flow. Considering subsonic regime, behaviour of c_l coefficient



$\alpha = 0°$, $\omega t = 0\pi{+}4\pi.$

$\alpha = 3°$, $\omega t = \frac{1}{2}\pi + 4\pi\,.$

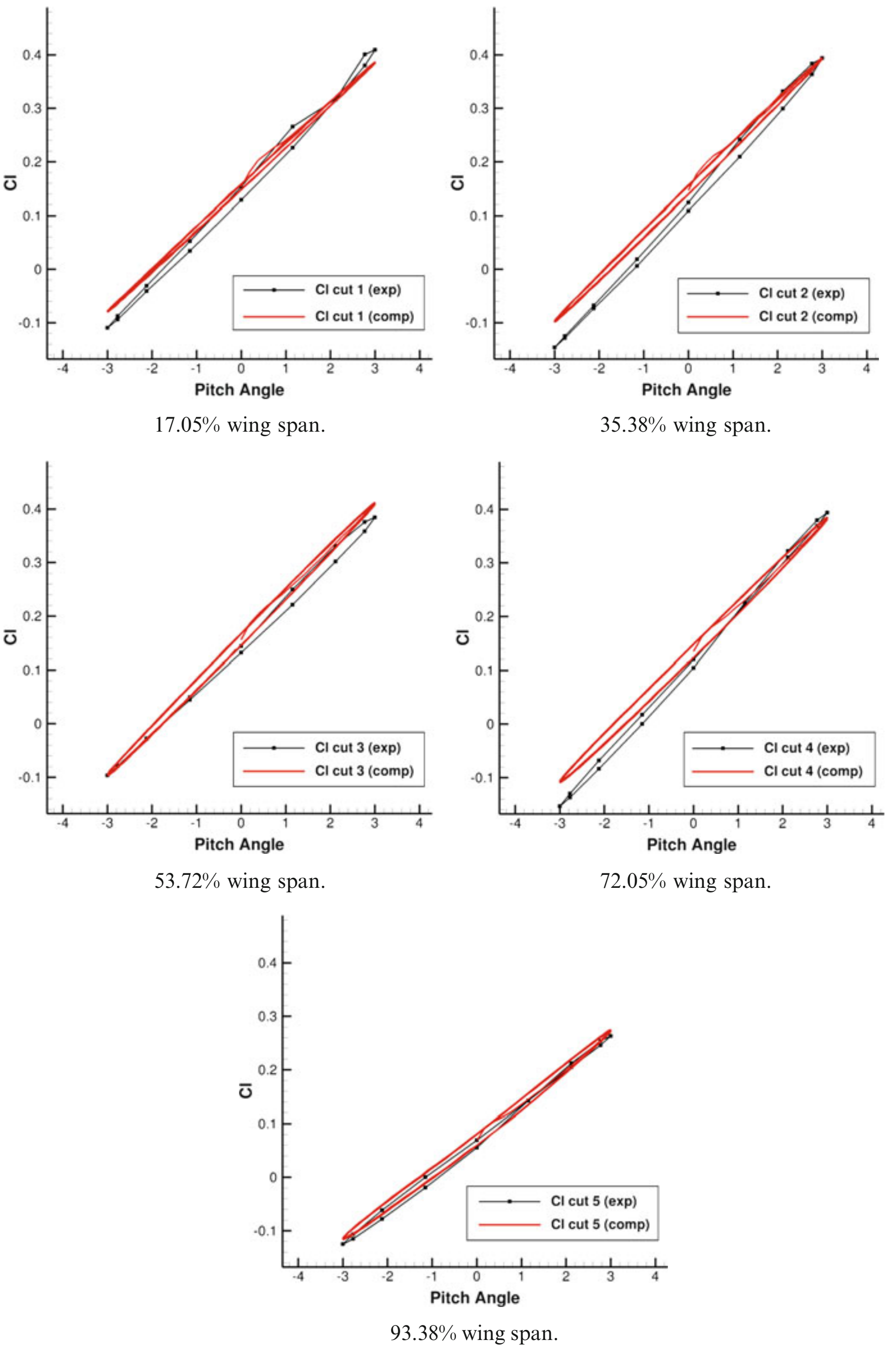
$\alpha = 0°$, $\omega t = \pi{+}4\pi.$

$\alpha = -3°$, $\omega t = \frac{3}{2}\pi + 4\pi\,.$

$\alpha = -2.12°$

$\alpha = 0°$, $\omega t\ = 6\pi$

Fig. 1 c_p coefficient in cuts alongside the AS28 wing during 3^{rd} period of forced oscillatory motion. Cuts are placed in 17.05%, 35.38%, 53.72%, 72.05% and 93.38% of the wing span



17.05% wing span.

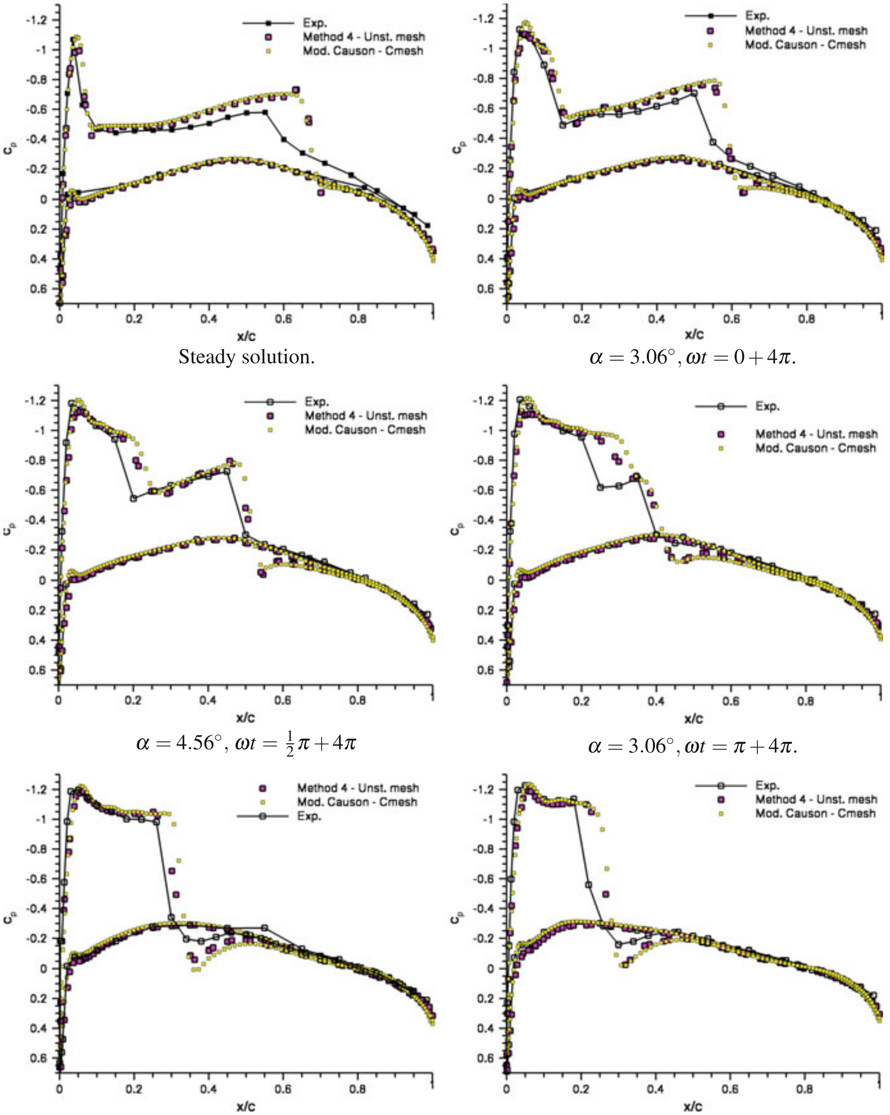
35.38% wing span.

53.72% wing span.

72.05% wing span.

93.38% wing span.

Fig. 2 c_l coefficient in cuts alongside the AS28 wing, forced oscillatory motion. Red line - numerical results, black line - experimental results



Steady solution.

$\alpha = 3.06^\circ, \omega t = 0 + 4\pi.$

$\alpha = 4.56^\circ$, $\omega t = \frac{1}{2}\pi + 4\pi$

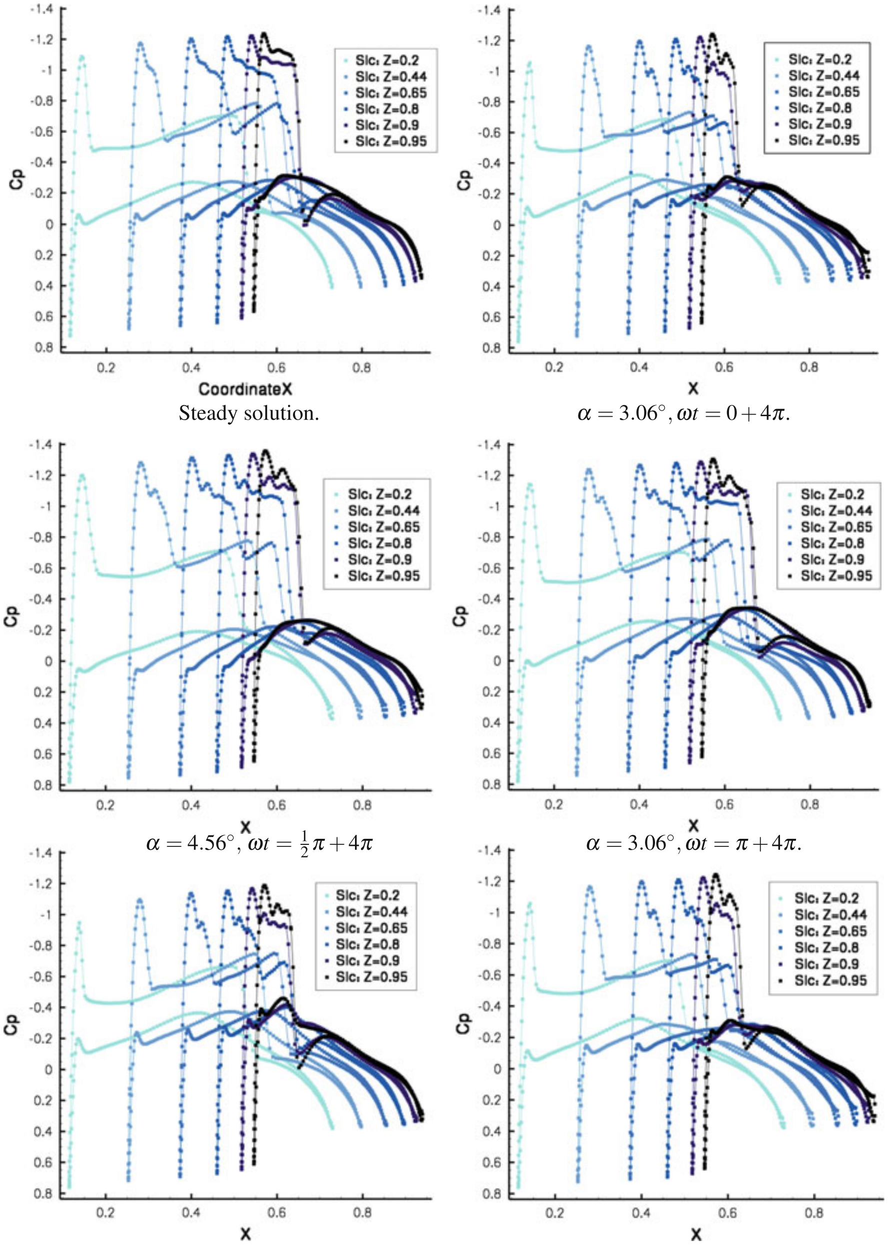
$\alpha = 3.06^\circ, \omega t = \pi + 4\pi.$

$\alpha = 1.56^\circ$, $\omega t = \frac{3}{2}\pi + 4\pi$

$\alpha = 0^\circ, \omega t = 6\pi.$

Fig. 3 c_p coefficient in cuts alongside the ONERA M6 wing during 3^{rd} period of forced oscillatory motion

obtained by numerical computation corresponds very well to the experimental observations. Moreover, the results show that fully periodic state has been achieved at least during the 3^{rd} period of oscillatory motion. Pressure coefficient decreases with increasing angle of attack (and vice versa) and the scheme does not produce



Steady solution.

$\alpha = 3.06°, \omega t = 0 + 4\pi.$

$\alpha = 4.56°, \omega t = \frac{1}{2}\pi + 4\pi$

$\alpha = 3.06°, \omega t = \pi + 4\pi.$

$\alpha = 1.56°, \omega t = \frac{3}{2}\pi + 4\pi$

$\alpha = 0°, \omega t = 6\pi.$

Fig. 4 c_p coefficient in cuts alongside the ONERA M6 wing during 3^{rd} period of forced oscillatory motion spurious oscillations. Comparison between experimental and numerical data is unfortunately not available in the transonic case, but the numerical results have all the mentioned characteristics as in subsonic flow.

5 Conclusion

The scheme is able to capture important flow characteristics even in the case of inviscid flow and can be used as a reliable numerical simulation of mentioned problems. From Figs. 1 to 4 can be seen that fully periodic state was achieved during at least 3^{rd} period of oscillatory motion. The future steps intended are implementation of implicit version of the scheme and its extension to aero-elastic problems.

Acknowledgements This work was supported by Research Plans MSM 6840770010, MSM 0001066902 and grant GACR No. 201/08/0012 and P101/10/1329.

References

1. Schmitt, V., Charpin, F.: Pressure Distributions on the ONERA-M6-Wing at Transonic Mach Numbers. Experimental Data Base for Computer Program Assessment. Report of the Fluid Dynamics Panel Working Group 04, AGARD AR 138, May 1979.
2. Fürst J.,: A weighted least square scheme for compressible flows. *Flow, Turbulence and Combustion*, 76(4):331–342, June 2006.
3. Furmánek, P., Fürst, J., Kozel, K.: High Order Finite Volume Schemes for Numerical Solution of 2D and 3D Transonic Flows in Kybernetika, Volume 45 no. 4, 567-579, 2009.
4. LeVeque, R., J.: Numerical Methods for Conservation Laws, Basel, 1990, ISBN 3-7643-2464-3.
5. Donea, J.: An arbitrary Lagrangian-Eulerian finite element method for transient fluid- structur interactions. Comput. Methods Apll. Mech. Eng., (1982), 33:689-723.
6. Fürst, J.: A weighted least square scheme for compressible fows. Submitted to "Flow, Turbulence and Combustion", (2005).
7. Batten, P., Leschziner, M. A., Goldberg, U. C.: Average-State Jacobians and Implicit Methods for Compressible Viscous and Turbulent Flows, Journal of computational physics 137, 1997.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

FVM-FEM Coupling and its Application to Turbomachinery

J. Fořt, J. Fürst, J. Halama, K. Kozel, P. Louda, P. Sváček, Z. Šimka, P. Pánek, and M. Hajsman

Abstract The paper deals with the numerical solution of turbulent flows through a 2D turbine cascade considering heat exchange between the gas and the solid blade. The flow field is described by the Favre averaged Navier-Stokes equations, and the temperature field inside the solid blade is given by the Laplace equation. Both parts are coupled in order to achieve continuity of the temperature as well as of the heat flux along the fluid-solid boundary. The analysis of simplified model case is presented and the results obtained with two in-house codes with several two-equation turbulence models are compared to results of commercial software (Fluent).

Keywords finite volumes, finite elements, heat transfer
MSC2010: 65M08, 65M12, 65N30, 76M10, 76M12

1 Introduction

The objective of this paper is to describe the coupled solution of turbulent flows through turbine cascade with heat transfer inside the blade. Due to the geometry of blades and expansion of the compressible fluid there is a temperature jump between suction and pressure side of blade profile, which is overestimated for commonly used adiabatic case compared to case with blade-fluid heat exchange. The correct modeling of heat exchange between blade and fluid is important for blades with high thermal conductivity and of course it is essential when considering some heat source inside the blade.

J. Fürst, J. Fořt, J. Halama, K. Kozel, P. Louda, and P. Sváček
Dept. of Tech. Math., CTU FME Prague, Karlovo nám. 13, CZ-12135 Praha 2, Czech Republic,
e-mail: Jiri.Furst@fs.cvut.cz

Z. Šimka, P. Pánek, and M. Hajsman
ŠKODA POWER a.s., A Doosan company, Tylova 1/57, CZ-30128 Plzeň, Czech Republic,
e-mail: Pavel.Panek@doosanskoda.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_53, © Springer-Verlag Berlin Heidelberg 2011

The solution of fluid part is obtained with the help of our in-house code using finite volumes whereas the heat equation is solved with finite elements. In order to achieve the continuity of temperature filed between the fluid and solid parts, the Dirichlet-Neumann coupling is used.

It is well known that the Dirichlet-Neumann coupling is under some conditions divergent in FEM-FEM case (see e.g. [8]). Therefore we do an analysis of simplified 2D problem in our FVM-FEM case in section 2 and we show that the method is under certain conditions stable.

The section 3 describes an application of the coupled algorithm to quite complex case of heat transfer between the turbulent flow field in turbine cascade and the solid blade.

2 Model problem

Our goal is to solve the heat transfer problem in turbomachinery (see section 3). The analysis of full model involving the solution of Navier–Stokes equations with a turbulence model is rather difficult, therefore we will analyze simplified model of Dirichlet–Neumann coupling of temperature field with different heat conductivities. We assume that the domain Ω is divided onto two parts: Ω_f covered by fluid with heat conductivity k_f and thermal capacity C and Ω_s corresponding to solid part with heat conductivity k_s. The solid part is in the interior of Ω (see Fig. 1 a).

We assume that the temperature field is described by parabolic heat equation in fluid part and by elliptic equation in solid part:

$$C\frac{\partial T(\mathbf{x},t)}{\partial t} = k_f \Delta T(\mathbf{x},t), \text{ for } \mathbf{x} \in \Omega_f, \tag{1}$$

$$\Delta\theta(\mathbf{x},t) = 0 \text{ for } \mathbf{x} \in \Omega_s. \tag{2}$$

Here T and θ denote the temperature in fluid and solid parts. The initial-boundary value problem for fluid part is equipped with the initial and boundary condition

$$T(\mathbf{x},0) = T_0(\mathbf{x}) \text{ for } \mathbf{x} \in \Omega_f, \tag{3}$$

$$T(\mathbf{x},t) = g(\mathbf{x},t) \text{ for } \mathbf{x} \in \partial\Omega \text{ and } t > 0. \tag{4}$$

We assume, that the temperature is continuous at the interface $\Gamma = \overline{\Omega_f} \cap \overline{\Omega_s}$. Moreover, the conservation of energy dictates also the continuity of heat fluxes across Γ. Hence

$$\theta(\mathbf{x},t) = T(\mathbf{x},t) \text{ for } \mathbf{x} \in \Gamma, \tag{5}$$

$$k_f\frac{\partial T}{\partial \mathbf{n}} = k_s\frac{\partial \theta}{\partial \mathbf{n}} \text{ for } \mathbf{x} \in \Gamma. \tag{6}$$

Here $\mathbf{n}$ is the outer normal with respect to Ω_f.

The solution is calculated using following semi-discrete time marching algorithm

1. Set $n = 0$ and $T^0 = T_0$.
2. Solve the Laplace equation (2) for θ^n using Dirichlet boundary condition $\theta^n|_\Gamma = T^n|_\Gamma$.
3. Calculate T^{n+1} using the parabolic equation (1) with Dirichlet boundary condition (4) at the $\partial\Omega$ and Neumann boundary condition (6) at the interface Γ calculating the heat flux using θ^n.
4. Increment n and repeat steps 2-4.

2.1 FE solution in $\boldsymbol{\Omega}_s$

We are solving the Laplace equation with Dirichlet boundary condition using standard FEM method with piece-wise linear base functions with triangular mesh in 2D. Let θ_i denotes the solution at FEM mesh node i (we omit the superscript n in this section). The standard discretization then leads to the algebraic system of equations

$$A\theta = \mathbf{b}, \tag{7}$$

where $b_i = 0$ for internal nodes and $b_j = T_j^n$ for boundary nodes.

It is known, that the Delaunay triangulation in 2D implies for piece-wise linear elements discrete maximum principle, i.e.

$$\min_{j\in\Gamma} T_j^n \le \theta_i \le \max_{j\in\Gamma} T_j^n, \tag{8}$$

where the shorthand notation $j \in \Gamma$ denotes boundary nodes (see e.g. [9] or [3]). This discrete maximum principle is equivalent in our case to the fact, that the solution in internal points i is a convex combination of boundary values T_j^n, i.e.

$$\theta_i = \sum_{j\in\Gamma} \alpha_{ij} T_j^n, \tag{9}$$

with $\alpha_{ij} \ge 0$ and $\sum_j \alpha_{ij} = 1$.

2.2 FV solution in $\boldsymbol{\Omega}_f$

Assume that the parabolic equation is solved with the explicit cell-centered FV scheme using an unstructured mesh. Assume that the mesh is orthogonal in the sense, that the face between two adjacent cell is orthogonal to the line connecting the cell centers. In that case the flux through the interface between cells i and k is proportional to $T_k - T_i$ and the explicit scheme for internal points reads

$$T_i^{n+1} = T_i^n + \Delta t \sum_k \beta_{ik}(T_k^n - T_i^n), \tag{10}$$

with $\beta_{ik} \geq 0$ (k goes over cells adjacent to cell i).

Including the Neumann boundary condition $k_f \partial T/\partial \mathbf{n} = \dot{q}$, the scheme for cells adjacent to Γ is

$$T_i^{n+1} = T_i^n + \Delta t \sum_k \beta_{ik}(T_k^n - T_i^n) + \Delta t \beta_i^b \dot{q}_i, \tag{11}$$

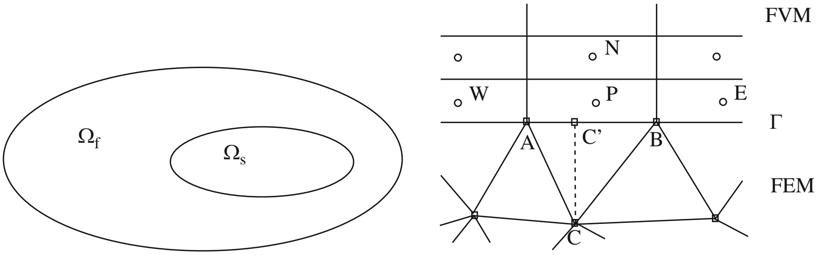
with $\beta_i^b > 0$.

The scheme for internal cells is positive and hence stable for small enough time steps. On the other hand the positivity is not obvious for boundary cells.

2.3 Coupling method

Here we assume that the nodes for FEM correspond to the cell vertices of FVM method at the boundary (see Fig. 1b, quadrilateral FV mesh in the upper part is connected to triangular FE mesh in the lower part).

Before solving the FE problem, we have to obtain a value at the boundary (points A and B at Fig. 1b). We calculate the boundary values using a weighted average of the cell-centered values adjacent to the boundary node with non-negative weights, e.g. $T_A^n = 0.5T_W^n + 0.5T_P^n$. Let us note, that this interpolation implies low order of accuracy at the interface. On the other hand we have to use very fine mesh spacing and high aspect ration cells in the fluid part near the interface due to thin boundary layers. Therefore we hope that the low order interpolation doesn't impair the overall accuracy.



(a) Domain (b) FVM-FEM coupling

Fig. 1 Domain $\Omega = \Omega_f \cup \Omega_s$, meshes for FVM-FEM coupling

Combining this positive interpolation with the equation (9) we get the solution in the solid part as

$$\theta_i = \sum_{k \in \Gamma} \gamma_{ik} T_k^n \text{ with } \gamma_{ik} \geq 0, \tag{12}$$

where $k \in \Gamma$ means that k goes over FV cells adjacent to the boundary Γ, i.e. $k = W, P, E, ...$ at Fig. 1b.

The "natural" evaluation of the normal derivative of θ in the triangle ABC yields

$$\frac{\partial \theta}{\partial \mathbf{n}}|_{ABC} = f(\theta_A, \theta_B, \theta_C) = \frac{\theta_C - \theta_{C'}}{|CC'|}, \tag{13}$$

where C' is the orthogonal projection of C onto line AB and $\theta_{C'}$ is the obtained with linear interpolation of θ_A and θ_B. Unfortunately it is very difficult to analyze the scheme with formula. Therefore we propose to use an approximation (see Fig. 1b)

$$\theta_{C'} = T_P^n. \tag{14}$$

Then the gradient of θ with respect to $\mathbf{n}$ is

$$\frac{\partial \theta}{\partial \mathbf{n}}|_{ABC} = \frac{\theta_C - T_P^n}{|CC'|}, \tag{15}$$

and taking into account the relation (12) we get

$$\frac{\partial \theta}{\partial \mathbf{n}}|_{ABC} = \frac{\sum_{k \in \Gamma} \gamma_{Ck} T_k^n - T_P^n}{|CC'|}, \tag{16}$$

Then the final scheme for cell P is

$$T_P^{n+1} = T_P^n + \Delta t \sum_{k \in \{N,W,E\}} \beta_{Pk}(T_k^n - T_P^n) + \Delta t \beta_P^b k_s \frac{\sum_{j \in \Gamma} \gamma_{Cj} T_j^n - T_P^n}{|CC'|} \tag{17}$$

with β, and γ being non-negative. Therefore we can make the scheme positive using small enough time step.

Note 1: the final scheme for T is under appropriate limit for time step Δt positive

$$T_i^{n+1} = \sum_j b_{ij} T_j^n,\ b_{ij} \geq 0. \tag{18}$$

Moreover taking $T_j^n = \tau$ a constant, we can easily show that the boundary values T_A^n, T_B^n,... are equal to τ too. Then $\theta_C = \tau$ and finally $T_i^{n+1} = \tau$. Canceling τ yields $\sum_j b_{ij} = 1$. Therefore the T_i^{n+1} is convex combination of values T_j^n and therefore it satisfies the discrete maximum principle (as far as the FE mesh is Delaunay and the FV mesh is orthogonal).

Note 2: the approximation (14) introduces an error in the heat flux. Nevertheless as we stated before, we are using fine near-wall scaling dictated by the turbulence model in the fluid part. In order to eliminate the error caused by "side shift" (i.e. different x coordinate of P and C' at the Fig. 1b) we can use isosceles triangles near the boundary in the FEM part.

3 Fluid-solid heat exchange in turbine cascade

Here we describe the application of this method to the solution of turbulent flows including heat transfer in the solid blade.

The fluid field is described by the set of Favre-averaged Navier-Stokes equations for compressible flows using several two-equations turbulence models. The temperature field inside the blade Ω_s is described by the Laplace equation (2).

The problem was numerically solved using commercial software Fluent at Škoda Plzeň and two different versions of in-house software developed at Czech Technical University.

3.1 Commercial software

The calculation by Škoda was performed in Fluent version 6.2 with the two-dimensional, double-precision, pressure-based solver. The turbulence models tested were the RNG $k-\epsilon$ model and the $k-\omega$ SST model. A second-order discretization scheme was employed. All calculations started on a coarse initial grid generated in Gambit, which was gradually refined at walls using the hanging-node adaption method to achieve adequate mesh resolution for low Reynolds turbulence models. The software does the computation of temperature field in both parts with the same FVM, hence it does not use the above described coupling procedure.

3.2 In-house codes

The coupling algorithm from previous section was used in the combination of two in-house codes developed at the Czech Technical University in Prague. The first one, denoted as solver 1 in later text, uses structured multiblock mesh, AUSM flux by [6] and quasi one-dimensional reconstructions with Van Leer limiter. The time discretization is achieved with backward Euler method for details see [1]. The second solver (solver 2) uses AUSMPW+ flux of [4], unstructured meshes, and multidimensional weighted least squares reconstruction described in [2].

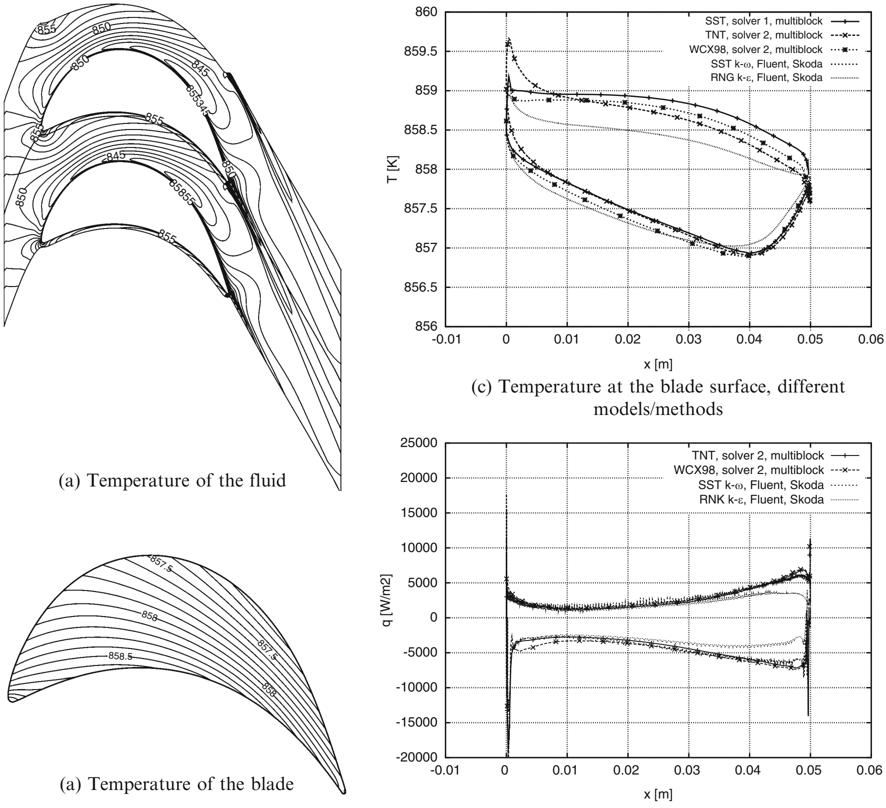
The turbulence is modeled using Low-Reynolds $k-\omega$ model by [10], TNT $k-\omega$ model by [5], and SST model by [7].

The temperature field in solid is solved with FEM and is coupled to fluid part using the algorithm described above with "natural" approximation of heat fluxes, see eq. (13). Moreover, the FV mesh is not orthogonal in the above mentioned sense, therefore the sufficient conditions for positivity of the scheme were not satisfied. Nevertheless we didn't met serious problems with stability in this case.

3.3 Results

Calculations were performed using structured and unstructured hybrid meshes with 10-20 000 cells in fluid part with in-house codes and using an extremely fine mesh with 650 000 cells using Fluent. The flow regime is characterized by the outlet isentropic Mach number $M_{2i} = 0.34$ and with the Reynolds number $Re = 820\,000$ related to the parameters at the outlet and to the blade pitch.

Figure 2 shows the iso-lines of the temperature obtained with solver 2 in the fluid part of the domain (a), the iso-lines of the temperature in the blade (b) calculated



(a) Temperature of the fluid

(a) Temperature of the blade

(c) Temperature at the blade surface, different models/methods

(d) Heat flux through the blade surface, different models/methods

Fig. 2 Temperature and heat flux for the conjugated heat transfer problem

with FEM. Moreover, it compares the distribution of temperature (c) and heat flux (d) along the blade surface obtained with both in-house methods including several turbulence models and the results of calculation made by commercial software. However there are some differences in the temperature and consequently in the heat flux, we can say that the agreement of all methods is satisfactory.

4 Conclusions

The article shows some results concerning the solution of heat transfer between turbulent flow and solid blades. The analysis shows that it is possible to couple FVM with FEM for this kind of problems. The second part shows an application of the procedure to conjugated heat transfer problem. Due to missing experimental data for our case, we were able to compare only our solution to the results obtained with commercial software which uses different approach. However the comparison was quite satisfactory, we have to do a comparison with experimental data in the future.

Acknowledgements The work was supported by the grant No. P101/10/1329 of the Grant Agency of Czech Republic and by the Research Plan MSM No. 6840770010.

References

1. J. Dobeš, J. Fořt, J. Fürst, P. Louda, K. Kozel, and L. Tajč. Numerical methods for transonic flows, application for design of axial and radial stator turbine cascades. In *8th ISAIF Conference Proceedings*, volume 2, pages 569–578. Ecole Centrale de Lyon, 2007.
2. Jiří Fürst. The third order WLSQR scheme on unstructured meshes with curvilinear boundaries. In *Proceedings of the ENUMATH conference*, Graz, 2007. submitted.
3. Antti Hannukainen, Sergey Korotov, and Tom Vejchodsk. On weakening conditions for discrete maximum principles for linear finite element schemes. *Numerical Analysis and Its Applications*, 5434:297–304, 2009.
4. Kyu Hong Kim, Chongam Kim, and Oh-Hyun Rho. Methods for the accurate computations of hypersonic flows I. AUSMPW+ scheme. *Journal of Computational Physics*, (174):38–80, 2001.
5. J. C. Kok. Resolving the dependence on free stream values for the k-omega turbulence model. Technical Report NLR-TP-99295, NLR, 1999.
6. M. S. Liou. A sequel to AUSM: AUSM+. *Journal of Computational Physics*, (129):364–82, 1996.
7. F. R. Menter. Two-equation eddy-viscosity turbulence models for engeenering applications. *AIAA J.*, 8(32):1598–1605, 1994.
8. A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford University Press, Oxford, 1999.
9. Reiner Vanselow. About delaunay triangulations and discrete maximum principles for the linear conforming FEM applied to the Poisson equation. *Applications of Mathematics*, 46(1):13–28, 2001.
10. David C. Wilcox. *Turbulence Modeling for CFD*. DCW Industries, Inc., second edition edition, 1998.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Charge Transport in Semiconductors and a Finite Volume Scheme

Klaus Gärtner

Abstract The van Roosbroeck system describes the transport of holes and electrons in semiconductors in a drift-diffusion approximation (a special type of Nernst-Planck-Poisson systems). The classic finite volume scheme used in the field allows to prove the existence of bounded steady state solutions and the uniqueness of the thermodynamic equilibrium solution by using the duality of the boundary conforming Delaunay grid and the Voronoi diagram. The article gives an overview over properties proven for this discrete version. The time dependent problem is dissipative in case of the implicit Euler scheme. The free energy decays exponentially in case of boundary conditions compatible with the thermodynamic equilibrium. The interesting qualitative properties of the analytic problem can be carried over to the discrete case for any h and τ (spatial, time step size respectively).
A weak interpretation of the scheme is helpful: using test functions one to gets estimates, and the weak discrete maximum principle allows to prove the bounds.
An implementation following the theory strictly (Oskar3) is used to solve 3d silicon detector problems, characterized by large volumes, multiple floating regions per detector pixel and extreme charge conservation requirements. An example is discussed to illustrate the problem.

Keywords reaction-diffusion systems, discrete bounded solutions, Delaunay grids, discrete weak maximum principle
MSC2010: 65N08, 65N12

Klaus Gärtner
WIAS, Mohrenstr.39, 10117 Berlin, Germany, e-mail: gaertner@wias-berlin.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_54, © Springer-Verlag Berlin Heidelberg 2011

1 The van Roosbroeck system

The continuity equations for the particle densities of electrons and holes are given on a bounded polyhedral domain $\Omega = \cup_i \Omega_i$, Ω_i a subdomain containing one material.

$$\frac{\partial n}{\partial t} - \nabla(D_n n_i \cdot \nabla \frac{n}{n_i} - \mu_n n \cdot \nabla w) + R(n,p) = 0, \tag{1}$$

$$\frac{\partial p}{\partial t} - \nabla(D_p n_i \cdot \nabla \frac{p}{n_i} + \mu_p p \cdot \nabla w) + R(n,p) = 0. \tag{2}$$

The main interaction of electrons and holes is described by the Poisson equation

$$-\nabla \cdot (\varepsilon_r \varepsilon_s \nabla w) = C - n + p. \tag{3}$$

The meaning of the quantities is:

- w - electrostatic potential,
- $n = n_i e^{w-\phi_n}$ - electron density, ϕ_n - quasi-Fermi potential of electrons,
- $p = n_i e^{\phi_p - w}$ - hole density, ϕ_p - quasi-Fermi potential of holes,
- C - density of impurities, n_i intrinsic carrier density,
- $\varepsilon = \varepsilon_r \varepsilon_s$ - dielectric permittivity, ε_r relative permittivity, ε_s scaled permittivity
- R - recombination-generation rate $R = r(x,n,p)(np-1)$, $r(x,n,p) \geq 0$,
- $\mu_{n,p}$ - carrier mobilities $\mu_{n,p} > 0$.

The Einstein relation is supposed to hold (diffusion constant $D_i = \mu_i k_B T/q_e$, k_B Boltzmann constant, T temperature, q_e elementary charge). Hence a natural scaling is to 'measure' all potentials in thermal voltages U_T and densities in a $n_{ref} \approx n_i$ resulting in $\varepsilon_s \approx 1.4 \cdot 10^{-13}\mathrm{m}^2$, $1\mathrm{V} \approx 40 U_T$ at room temperature, and $n + p$ can easily be of the order 10^{10} or 10^{-10} in different parts or states of a device.

The free energy of van Roosbroeck system is (compare [8, 16, 17])

$$F(w,n,p) = \int [n(\ln \frac{n}{n^*} - 1) + n^* + p(\ln \frac{p}{p^*} - 1) + p^*]\, d\Omega + \frac{1}{2} |||w - w^*|||^2, \tag{4}$$

with $|||h|||^2 = \int \varepsilon |\nabla h|^2 \, d\Omega + \int \alpha h^2 \, d\Gamma$, $\nu \cdot \nabla w^* + \alpha(w^* - w_\Gamma) = 0$, ν outer normal, and

$$-\nabla \cdot (\varepsilon \nabla w^*) = C + n_i e^{-w^*} - n_i e^{w^*} \text{ in } \Omega \tag{5}$$

the (w^*, n^*, p^*) (unique weak) thermal equilibrium solution. The dissipation rate is given by

$$d(w,n,p) = \int [n\mu_n |\nabla \phi_n|^2 + p\mu_p |\nabla \phi_p|^2 + r(x,n,p)(np-1)\ln(np)]\, d\Omega \geq 0. \tag{6}$$

The problem is supplemented by boundary conditions on $\Gamma = \Gamma_D \cup \Gamma_N$ describing contacts (Dirichlet boundary conditions for w, n, p on $\Gamma_D = \bar{\Gamma}_D$), gate contacts (third kind boundary condition for w, homogeneous Neumann boundary condition for n, p), and homogeneous Neumann boundary conditions on different parts of the boundary of the domain. The analytic results obtained by different techniques and for different assumptions on data [6, 9, 19–21] can be summarized for the purpose in mind here by: existence of steady state solutions (in $H^1 \cap L^\infty$).

2 Spatial discretization

The aim from this application point of view is to have discretizations, that carry over the analytic properties for classes of grids, hence parameters like step sizes are a question of precision and not of existence of the solution. For boundary conforming Delaunay meshes and the Scharfetter- Gummel scheme together with an implicit Euler time discretization the following table summarizes the proven properties: At a first glance that may look like a pretty comfortable position, but the headroom for improvement by better understanding is large.

A short summary with respect to Delaunay meshes [3] introduces notations and the following part reviews results establishing the lower right part of Table 1, see [10]. Let a vertex $v_k \in I\!R^N$ be denoted by $\mathbf{v}_k = (x_1, \ldots, x_N)^T$, $\mathbf{E}_l^N$ is simplex l in the Delaunay grid, $B(E_l^M)$ its circumscribed ball (if $M < N$ the smallest circumscribed ball). Vertex numbers are chosen such that the local coordinate system for each simplex defined by the matrix $P_{l,k} = (\mathbf{v}_{k+1} - \mathbf{v}_k, \ldots, \mathbf{v}_{k+N} - \mathbf{v}_k)$ results in the volume $|P_{l,k}|/N > 0$. Interfaces and Γ are given by $N-1$ dimensional simplices in the grid. The Delaunay property requires that for all l $\mathbf{v}_j \in\!| \; B(E_l^N)$, $\forall \mathbf{v}_j \neq P_l$, and $B(E_l^N)$ is the circumscribed ball of E_l^N. The Voronoi volume V_i is the set of all points closer to $\mathbf{v}_i$ than to $\mathbf{v}_j, j \neq i$. $\partial V_i = \bar{V_i} \setminus V_i$ denotes the surface of the Voronoi volume and the intersections with the simplex $\mathbf{E}_j^N$ are $V_{ij} = V_i \cap \mathbf{E}_j^N$ and $\partial V_{ij} :\, = \partial(V_i \cap \mathbf{E}_j^N)$. ∂V_{ij} is the union of planar pieces of ∂V_i and those $E_l^{N-1} \in E_j^N$ sharing the vertex i. The Delaunay property guarantees a non negative surface measure per edge in the interior of each subdomain Ω_i, the boundary conformity [5] (per subdomain) requires that all lower dimensional simplices on the boundary have empty smallest circumscribed balls, too. Together with the fact that all interfaces (boundaries) coincide with a set of E_l^{N-1} both types of surface measures (in or orthogonal to each interface) per edge and subdomain are non negative.

Starting with the equation $-\nabla \cdot \varepsilon \nabla w = f$, using Gauss's theorem on V_{ij}, and assuming $\varepsilon = const$ per simplex yields:

Table 1 Proven properties, compare [7, 8, 10–15]

property	analytic	discrete
dissipativity	yes	yes
exponential decay free energy	yes	yes
existence of bounded steady state sol.	yes	yes
uniqueness for small applied voltages	yes	yes

$$\int_{V_{ij}} -\nabla \cdot \varepsilon_l \nabla w \, dV = -\varepsilon_l \int_{\partial V_{ij}} \nabla w \cdot d\mathbf{S_k} = -\varepsilon_l \sum_{k(j)} \int_{\partial V_{i,k(j)}} \nabla w \cdot d\mathbf{S_k} + BI_{V_{ij}}$$

$$\approx -\varepsilon_l \sum_{k} \frac{\partial V_{i,k(j)}}{|\mathbf{e}_{ik(j)}|}(w_k - w_i) + BI_{V_{ij}} = \varepsilon_l [\gamma_{k(i)}] \tilde{G}_N \mathbf{w}|_{E_j^N} + BI_{V_{ij}},$$

where $BI_{V_{ij}}$ denotes boundary integrals in case of boundary faces. It is compensated in the interior by the neighboring $BI_{V_{ij'}}$, $\mathbf{e}_{i,k(j)}$ is the edge from node i to k in simplex j, and $\tilde{G}_N$ is a difference matrix, mapping from nodes to edges.

$$(\tilde{G}^T \tilde{G})_{ii} > 0, \ \ (\tilde{G}^T \tilde{G})_{i>j} < 0, \text{ and } \mathbf{1}^T \tilde{G}^T = \mathbf{0}^T. \tag{7}$$

$$\gamma_{k(i)} = \frac{\partial V_{i,k(i)}}{|\mathbf{e}_{ik(i)}|} \tag{8}$$

denotes the elements of a diagonal matrix of geometric weights per simplex. Functions are approximated by

$$\int_{V_{ij}} f dV \approx V_{ij} f(x_i), \ \ [V]_i = \sum_j V_{ij}, \tag{9}$$

where $[\cdot]$ denotes a diagonal matrix. Summing over all vertices of the simplex j yields

$$\sum_{V_{ij} \in \mathbf{E}_j^N} \int_{V_{ij}} -\nabla \cdot \varepsilon \nabla w \, dV \approx \varepsilon \tilde{G}^T [\gamma] \tilde{G} \mathbf{w}|_{E_j^N} + BI. \tag{10}$$

The explicit form of the boundary integrals (in the generic situation $\xi_1 w + \xi_2 \partial w / \partial \nu + \xi_3 = 0$, with ξ_i defined on Γ, $\xi_1(x, w, \dots) \geq 0$, $\xi_2(x, w, \dots) > 0$) is given by

$$BI_{V_{ij}} = \sum_{i' \neq i, i' \in \mathbf{E}_j^N} \int_{E_{i'}^{N-1} \cap \partial V_{ij}} -\varepsilon \nabla w \cdot d\mathbf{S} \approx \sum_{i' \neq i, i' \in \mathbf{E}_j^N} |E_{i'}^{N-1} \cap \partial V_{ij}| \frac{\varepsilon}{\xi_{2_{i'}}} (\xi_{1_{i'}} w_i + \xi_{3_{i'}}),$$

where $E_{i'}^{N-1}$ denotes the $N-1$ dimensional simplex opposite to $i' \in \mathbf{E}_j^N$, $E_{i'}^{N-1} \in \Gamma$, and $BI = \sum_{i \in \mathbf{E}_j^N} BI_{V_{ij}}$.

Remark 1. A 'discrete weak maximum principle' holds ($\mathbf{w}^+$ pos. part)

$$(\mathbf{w} - w_0)^{+T} \tilde{G}^T [\gamma] \tilde{G} \mathbf{w} > 0,$$

if $\mathbf{w} > w_0 = const$ at least for one $\mathbf{x}_i \in \Omega$, as long as the Voronoi faces related to each edge and subdomain fulfill

$$\sum_{E_j^N \ni \mathbf{e}_{ik}, E_j^N \in \Omega_l} \partial V_{ik} \geq 0. \tag{11}$$

This is exactly the requirement fulfilled by a 'boundary conforming Delaunay mesh' and has to be preserved for acceptable averages $\bar{\varepsilon}_{ij}$ in case of $\varepsilon = \varepsilon(x, n, p, |\nabla w|, \ldots)$. In the sequel the notation

$$G := [\sqrt{\gamma}]\tilde{G},$$

is used to denote the discrete gradient and summation (elements, edges, nodes) is not indicated any more—the context should indicate a local or global use.

The continuity equations can be transformed by changing variables $n = n_i e^w u$, $p = n_i e^{-w} v$, hence the steady state case reads:

$$-\nabla \cdot n_i D_n e^w \nabla u + R = 0, \tag{12}$$

$$-\nabla \cdot n_i D_p e^{-w} \nabla v + R = 0, \tag{13}$$

with an elliptic main part for bounded electrostatic potentials. Application of the discretization scheme and integration along each edge (the term $w_k - w_i$ in the discrete current expression (7) is just a special case of integrating the equation $w(s)'' = 0$ along the edge from s_i to s_k) using

- $(\bar{\mu} e^{w(x)}(e^{-\phi})')' = 0$, $\bar{\mu}$ edge average,
- $w(x)$ piecewise linear,
- $\mathrm{sh}(s) := \sinh(s)/s$, $\mathrm{b}(2s) = e^{-s}/\mathrm{sh}(s) = 2s/(e^{-2s}-1)$, b Bernoulli function:

$$G^T[\varepsilon]G\mathbf{w} = [V]\mathbf{g}(\mathbf{C}, \mathbf{n}, \mathbf{p}),\ \mathbf{g} = \mathbf{C} - \mathbf{n} + \mathbf{p},\ \mathbf{n} = n_i[e^w]\mathbf{u},\ \mathbf{p} = n_i[e^{-w}]\mathbf{v}, \tag{14}$$

$$A_{S_n}(D_n, \mathbf{w})e^{-\phi_n} = G^T[\bar{D}_n e^{\bar{w}}/\mathrm{sh}(\tilde{G}\mathbf{w}/2)]G\mathbf{u} = [V][r(\mathbf{x}, \mathbf{n}, \mathbf{p})](\mathbf{1} - [v]\mathbf{u}), \tag{15}$$

$$A_{S_p}(D_p, -\mathbf{w})e^{\phi_p} = G^T[\bar{D}_p e^{-\bar{w}}/\mathrm{sh}(\tilde{G}\mathbf{w}/2)]G\mathbf{v} = [V][r(\mathbf{x}, \mathbf{n}, \mathbf{p})](\mathbf{1} - [u]\mathbf{v}). \tag{16}$$

The diagonal transformations $\mathbf{n} = [n_i e^w]\mathbf{u}$, $\mathbf{p} = [n_i e^{-w}]\mathbf{v}$ yield the well known Scharfetter-Gummel (Il'in) scheme, dating back to Allen and Southwell ([1, 18, 22], see [4], too), generalized to boundary conforming Delaunay grids and used since the early eighties in semiconductor device simulations (compare [2, 23]).

The thermodynamic equilibrium solution is given by ($\mathbf{w}^*$, $\mathbf{u}^* = \mathbf{1}$, $\mathbf{v}^* = \mathbf{1}$) and $\mathbf{w}^*$ solution of (14) with $\mathbf{u} = \mathbf{u}^*$, $\mathbf{v} = \mathbf{v}^*$.

The proof of bounded steady state solutions for the system (14, 15, 16) proceeds in the following steps (for details see [10]):
a) for $-\infty < \check{w}_0 \leq w \leq \hat{w}_0 < \infty$ matrices A_{S_n}, A_{S_p} are weakly diagonally dominant (positive Dirichlet boundary measure).
b) for some $\check{u} \leq u_i^0 \leq \hat{u}$, $\check{v} \leq v_i^0 \leq \hat{v}$ $\forall$ $x_i \in \bar{\Omega}$, $\check{C} = \min(C(x))$ and $\hat{C} = \max(C(x))$, the extreme boundary values, and the monotone mapping with respect to w_j, $g_j(C_j, n_j, p_j) = V_j(C_j - n_j + p_j)$ (right hand side of the discrete Poisson equation) allow to bound the solution of (14) by construction contradictions using the weak discrete maximum principle.

c) supposing properly chosen bounds for $\tilde{w} = \max |w_i|$ one proves $e^{-\tilde{w}} \leq \mathbf{u}^0, \quad \mathbf{v}^0 \leq e^{\tilde{w}}$ using the maximum principle and the properties of the recombination-generation term, hence $e^{-\tilde{w}}$, $e^{\tilde{w}}$ is a lower, upper solution respectively for equations (15, 16) with properly frozen $\mathbf{u}^0$, $\mathbf{v}^0$ in $r(x, \mathbf{u}^0, \mathbf{v}^0)$ and $(1 - [v^0]\mathbf{u}^0)$ $(1 - [u^0]\mathbf{v}^0)$ in (15, 16) respectively.
d) Brouwer's fixed point theorem guarantees the existence of at least one fixed point. The bounds are identical with the analytic ones.
The discrete dissipation expression is obtained from testing the discrete equations by the discrete quasi-Fermi potentials.
The uniqueness in the neighborhood of the discrete thermodynamic equilibrium follows by linearization at $(\mathbf{w}^*, \mathbf{u}^*, \mathbf{v}^*)$, resulting in a decoupling of the continuity equations and the Poisson equation, the Schur complement due to elimination of $\mathbf{u}$ is a weakly diagonally dominant M-matrix (compare [10]).

3 Example

Silicon detectors for high energy and astrophysics are nice examples to stress the algorithms used: each new detector is in some sense an extreme design for one special purpose. The new X-ray lasers for instance require high speed, low power, high spatial and energy resolution and the best possible signal to noise ratio. Extreme charge conservation requirements (see Fig. 4) in the interesting parts of the detector, hence in the computations, are typical.
The example shown is a pnCCD for SLAC's Linac Coherent Light Source designed at the MPI HLL, Munich. The Fig. 1 shows the relative simple geometry of two quarter CCD registers, two times two half CCD registers, and again two quarter CCD registers. This is the minimal configuration for testing the charge shift properties of the CCD (Fig. 3). Questions of interest are:

- The maximum number of electrons to be stored in one register (see Fig. 2)?
- Appeare losses to the surface, hence recombination with holes from contacts?
- How fast is shifting by changing boundary conditions on top of the registers?
- Do electrons stay in the start register, reach all the aim register, see Fig. 4?

The computations predicted the possibility to store 5 to 10 times more electrons in an optimized pnCCD. That was verified by experiments just now.

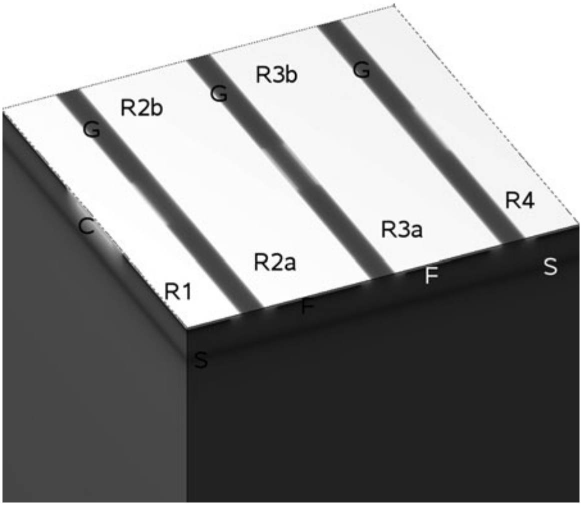


Fig. 1 Doping as equivalent equilibrium potential, white negative, dark positive, R1 quarter of a register, R2a one half register, side a is separated by the 'channel stop' C from side b, R3b one half register, F floating region to create a potential minimum beneath each register in the shift layer S, G gate contact to tune register separation in shift direction; dimensions $x = 75$¯m, $y = 75$¯m, $z = 150$¯m, 958 399 nodes, BACK: -50V, REGISTER 1, 3, 4: -18V, REGISTER 2: -10V, GATE 1, 2, 3: 5V

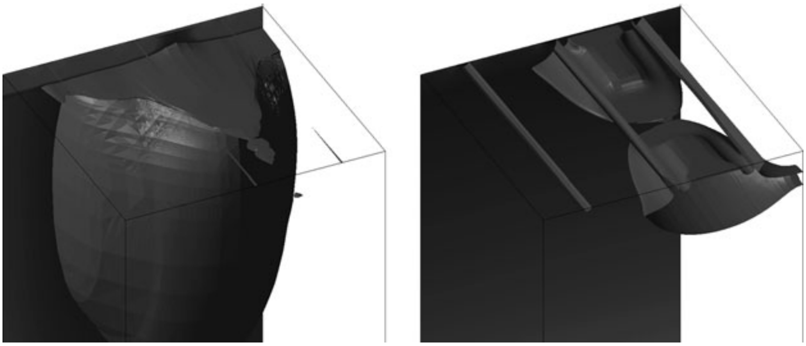


Fig. 2 Overflow of electrons at the arrival of the charge cloud created close to the bottom in the center of register 2 at y_{max}, $10n_i$ iso-surface of $\log(n)$ (left), weakest point in the potential barrier (properly selected electrostatic potential iso-surface, right)

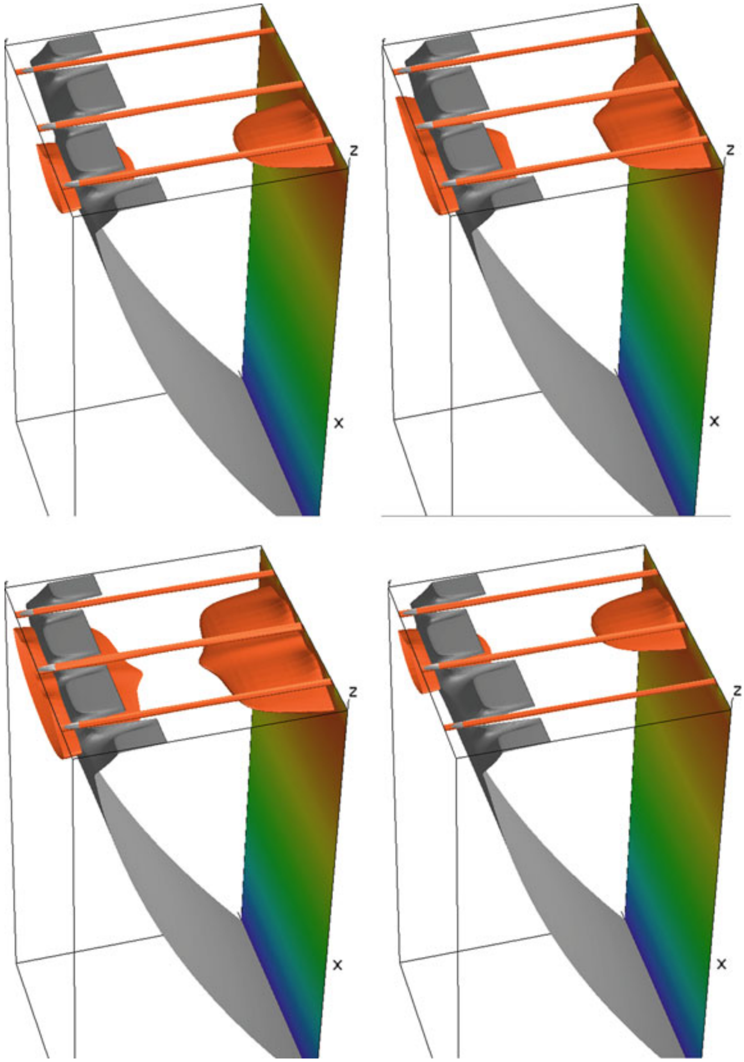


Fig. 3 In computations one can not shift the electrons 1000 times in one direction, hence a minimal configuration is used, to shift them back and forth. Shifting of electrons (not shown) takes place inside the moving potential barriers (iso-surfaces) due to time dependent boundary conditions at register 2 and 3 (compare the plotted electrostatic potential elevation over the $y = y_{max}$ surface). Initial state, the electrons are inside the iso-surface centered at register 2 (top left), the boundary value at register 3 reduces the potential barrier between register 2 and 3 (top right), both registers are 'open' (bottom left), register 2 has pushed out the electrons to register 3 (bottom right). Graphics by `gltools`

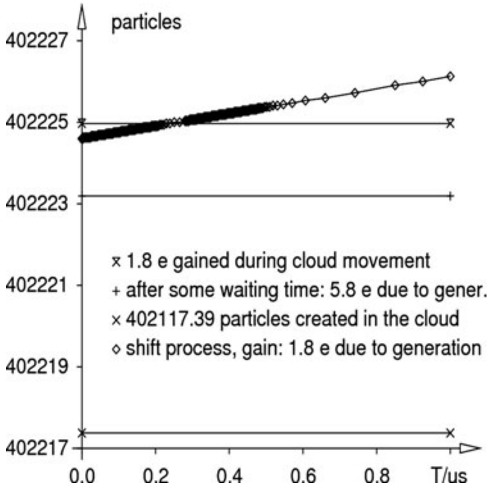


Fig. 4 The electron balance in the volume of interest is one crucial point: after order 1000 shift operations total losses of 0.1% are acceptable in the detector. The charge balance in the computations can be explained up to one third missing electron (out of 402225) after two integrations over 7 orders in time or 668 time steps

Acknowledgements The author has to thank the European XFEL for collaboration and partially funding the work. MPI HLL at Munich is a partner with interesting problems since many years.

References

1. Allen, D.N., Southwell, R.V.: Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. Quart. J. Mech. and Appl. Math. **8**, 129–145 (1955)
2. Bank, R., Rose, D., Fichtner, W.: Numerical methods for semiconductor device simulation. SIAM Journal on Scientific and Statistical Computing **4**(3), 416–435 (1983)
3. Delaunay, B.: Sur La Sphére Vide. Izvestia Akademii Nauk SSSR. Otd. Matem. i Estestv. Nauk **7**, 793 – 800 (1934)
4. Eymard, R., Fuhrmann, J., Gärtner, K.: A finite volume scheme for nonlinear parabolic equations derived from one-dimensionsl local Dirichlet problems. Numer. Math. **102**, 463–495 (2006)
5. Gabriel, K., Sokal, R.: A new statistical approach to geographic analysis. Systematic Zoology **18**, 259–278 (1969)
6. Gajewski, H.: On existence, uniqueness and asymptotic behavior of solutions of the basic equations for carrier transport in semiconductors. Z. Angew. Math. Mech. **65**, 101–108 (1985)
7. Gajewski, H., Gärtner, K.: On the discretization of van Roosbroeck's equations with magnetic field. Z. Angew. Math. Mech. **76**, 247–264 (1996)
8. Gajewski, H., Gröger, K.: On the basic equations for carrier transport in semiconductors. J. Math. Anal. Appl. **113**, 12–35 (1986)

9. Gajewski, H., Gröger, K.: Initial-boundary value problems modelling heterogeneous semiconductor devices. In: Surveys on Analysis, Geometry and Mathematical Physics, *Teubner-Texte Math.*, vol. 117, pp. 4–53. Teubner, Leipzig (1990)
10. Gärtner, K.: Existence of bounded discrete steady state solutions of the van roosbroeck system on boundary conforming delaunay grids. SIAM J. Sci. Comput. **31**, 1347–1362 (2009)
11. Glitzky, A.: Exponential decay of the free energy for discretized electro-reaction-diffusion systems. Nonlinearity **21**, 1989–2009 (2008)
12. Glitzky, A.: Uniform exponential decay of the free energy for Voronoi finite volume discretized reaction-diffusion systems. Preprint 1443, Weierstraß-Institut für Angewandte Analysis und Stochastik, Berlin (2009, to appear in Math. Nachr.)
13. Glitzky, A., Gärtner, K.: Energy estimates for continuous and discretized electro-reaction-diffusion systems. Nonlinear Analysis **70**, 788–805 (2009)
14. Glitzky, A., Gärtner, K.: Existence of bounded steady state solutions to spin-polarized drift-diffusion systems. SIAM J. Math. Anal. **41**, 2489–2513 (2010)
15. Glitzky, A., Hünlich, R.: Energetic estimates and asymptotics for electro–reaction–diffusion systems. Z. Angew. Math. Mech. **77**, 823–832 (1997)
16. Gokhale, B.: Numerical solutions for a one-dimensional sislicon p-n-p transistor. IEEE Trans. Electron Devices **ED-17**, 594–602 (1970)
17. Horn, F., Jackson, R.: General mass action kinetics. Arch. Rat. Mech. Anal. **47**, 81–116 (1972)
18. Il'in, A.M.: A difference scheme for a differential equation with a small parameter multiplying the second derivative. Matematičeskije zametki **6**, 237–248 (1969)
19. Jerome, J.W.: Consistency of Semiconductor Modeling: An Existence/Stability Analysis for the Stationary Van Roosbrook System. SIAM J. Appl. Math. **45**, 565–590 (1985)
20. Markowich, P.A.: The Stationary Semiconductor Device Equations. Springer, Wien (1986)
21. Mock, M.S.: Analysis of Mathematical Models of Semiconductor Devices. Boole Press, Dublin (1983)
22. Scharfetter, D.L., Gummel, H.K.: Large–signal analysis of a silicon read diode oscillator. IEEE Trans. Electr. Dev. **16**, 64 – 77 (1969)
23. Selberherr, S.: Analysis and Simulation of Semiconductor Devices. Springer, Wien-New York (1984)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Playing with Burgers's Equation

T. Gallouët, R. Herbin, J.-C. Latché, and T.T. Nguyen

Abstract The 1D Burgers equation is used as a toy model to mimick the resulting behaviour of numerical schemes when replacing a conservation law by a form which is equivalent for smooth solutions, such as the total energy by the internal energy balance in the Euler equations. If the initial Burgers equation is replaced by a balance equation for one of its entropies (the square of the unknown) and discretized by a standard scheme, the numerical solution converges, as expected, to a function which is not a weak solution to the initial problem. However, if we first add to Burgers' equation a diffusion term scaled by a small positive parameter ϵ before deriving the entropy balance (this yields a non conservative diffusion term in the resulting equation), and then choose ϵ and the discretization parameters adequately and let them tend to zero, we observe that we recover a convergence to the correct solution.

Keywords Burgers equation, compressible flows, Euler equations, finite volumes
MSC2010: 65M08,76N99

1 Introduction

Computer codes developed for the simulation of inviscid and non heat-conducting compressible flows are in general based on the conservative form of the Euler equations, which read in the one-dimensional case:

T. Gallouët and R. Herbin
Université Aix-Marseille, e-mail: [gallouet,herbin]@cmi.univ-mrs.fr

J.-C. Latché and T.T. Nguyen
Institut de Radioprotection et de Sûreté Nucléaire (IRSN), e-mail: [jean-claude.latche,tan-trung.nguyen]@irsn.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_55, © Springer-Verlag Berlin Heidelberg 2011

$$\partial_t \rho + \partial_x(\rho u) = 0, \tag{1a}$$

$$\partial_t(\rho u) + \partial_x(\rho u^2) + \partial_x p = 0, \tag{1b}$$

$$\partial_t E + \partial_x\big((E + p)u\big) = 0, \tag{1c}$$

where t stands for the time, ρ, u and p are the density, velocity and pressure in the flow, and E stands for the total energy, $E = \rho u^2/2 + \rho e$, with e the internal energy. This system must be complemented by an equation of state, giving for instance the pressure as a function of the density and the internal energy $p = \wp(\rho, e)$.

For physical reasons, the density and internal energy must be non-negative (in usual applications, positive). In addition, for the continuous problem as well as, at the discrete level, for a wide range of schemes (the so-called conservative schemes), the non-negativity of these variables allows a (weak) control on the solution; assuming that ρ and E are known on the parts of the boundary where the flow is entering the computational domain, Equations (1a) and (1c) indeed yield an $\mathrm{L}^\infty(0,T;\mathrm{L}^1(\Omega))$-estimate (with $\Omega \times (0,T)$ the space-time domain of computation) for the density and the total energy respectively. The positivity of the density at the discrete level is easily obtained from a convenient discretization of (1a). The positivity of the internal energy does not seem easily obtained other than by replacing Equation (1c) by a balance equation for the internal energy in the discrete problem; this balance equation is formally derived (*i.e.* supposing that the solution is regular) from (1b) and (1c) and reads:

$$\partial_t(\rho e) + \partial_x(\rho e u) + p\partial_x u = 0. \tag{2}$$

In this relation, the discrete convection operator may be built so as to respect the positivity of e: provided that the equation of state is such that for any value of ρ, p vanishes for $e = 0$, testing the discrete counterpart of (2) by the negative part of e proves $e \geq 0$ (see [5] for the initial paper, [2, Appendix B] for another proof suitable in this context, and [4] in the framework of the compressible Navier-Stokes equations).

Instead of Equation (1c), one may also prefer to use a conservation equation for the physical entropy s, because this equation (derived for regular solutions) is a simple transport equation:

$$\partial_t(\rho s) + \partial_x(\rho s u) = 0. \tag{3}$$

Let us then consider that, for computational efficiency or robustess reasons, (2) or (3) are prefered to (1c). Since both (2) and (3) are derived from (1c) assuming a regular solution, there is no reason for their discretization to yield the correct weak solutions in the presence of shocks. Nevertheless, we may reasonably expect to recover the correct shock solutions if we use the following strategy:

(i) regularize the problem by adding a small diffusion term,

(ii) derive the counterpart of (2) or (3) taking into account the diffusion terms,

(iii) solve these equations,
(iv) let ϵ tend to zero.

Of course, step (*iii*) is performed numerically, and convergence is monitored by the space and time discretization steps h and k; the question which arises is then to find a convenient way to let ϵ and the numerical parameters h and k tend to zero. The aim of this paper is to perform numerical experiments in order to investigate this issue on a toy problem, namely the inviscid Burgers equation. Note that we only consider explicit schemes in this study.

2 The equations and the numerical schemes

The inviscid Burgers equation reads:

$$\partial_t u + \partial_x(u^2) = 0, \qquad \text{for } x \in \mathbb{R},\ t \in (0,T), \tag{4}$$

which we complement with the initial condition:

$$u(x,0) = u_0(x), \qquad \text{for } x \in \mathbb{R}. \tag{5}$$

Following the above mentioned strategy (items (i)-(iv)), we first add to (4) a viscous term, to obtain: $\partial_t u + \partial_x(u^2) - \epsilon \partial_{xx} u = 0$. Now, multiplying this relation by $2u$ yields the following perturbed equation:

$$\partial_t u^2 + \frac{4}{3}\,\partial_x u^3 - 2u\epsilon\partial_{xx} u = 0. \tag{6}$$

For $\varepsilon = 0$, we get the following "Burgers square entropy" equation:

$$\partial_t u^2 + \frac{4}{3}\,\partial_x u^3 = 0. \tag{7}$$

which also reads, setting $v = u^2$:

$$\partial_t v + \frac{4}{3}\partial_x(v^{\frac{3}{2}}) = 0. \tag{8}$$

We consider the following initial data, chosen such that the entropy solution of (4)-(5) contains a discontinuity:

$$u_0(x) = \begin{cases} 10, \ x \leq -0.25 \\ 1, \quad x > -0.25 \end{cases}. \tag{9}$$

It is well known that for such an initial condition, the entropy weak solutions of equations (4) and (7) differ. Let us then turn to their numerical approximations.

Since the chosen initial data (9) is positive, the celebrated Godunov scheme reduces for both equations to the classical upwind scheme, thanks to the fact that the upwind scheme preserves (for these equations) the sign of the solution; it is well known that it leads to an approximate solution which converges, under a so called CFL condition, to the exact solution as the discretization parameters go to zero [1] (note that this is not the case for the centred finite volume scheme, although it is conservative). For the sake of simplicity, we consider constant time and space steps h and k. For $i \in \mathbb{Z}$, we set $x_i = ih$ and for $n \in \{0, \dots, M\}$, with $(M-1)k < T \leq Mk$, we set $t_n = nk$. The discrete unknowns are the real numbers $u_i^{(n)}$, with $i \in \mathbb{Z}$ and $n \in \{0, \dots, M\}$. The values $u_i^{(0)}$ are obtained with the initial condition:

$$u_i^{(0)} = \frac{1}{h} \int_{x_i - \frac{h}{2}}^{x_i + \frac{h}{2}} u_0(x) dx. \tag{10}$$

Since the discrete solution is positive, the upwind scheme for Equation (4) reads:

$$u_i^{(n)} = u_i^{(n-1)} + \frac{k}{h} \Big[\big(u_{i-1}^{(n-1)}\big)^2 - \big(u_i^{(n-1)}\big)^2 \Big]. \tag{11}$$

For this particular problem and scheme, the maximum value for the solution is reached at the initial time step so that the CFL number is the number G such that:

$$k = G \frac{h}{\max\{2s,\ s \in [1, 10]\}} = G \frac{h}{20}. \tag{12}$$

Similarly, the upwind scheme for Equation (8) reads:

$$v_i^{(n)} = v_i^{(n-1)} + \frac{4k}{3h} \Big[\big(v_{i-1}^{(n-1)}\big)^{\frac{3}{2}} - \big(v_i^{(n-1)}\big)^{\frac{3}{2}} \Big], \tag{13}$$

and the CFL number is the same number G. The numerical solutions obtained with (11) for the Burgers equation (4) and with (13) for the Burgers square entropy equation (7) are depicted in Fig. 1. Both are obtained with CFL equal to 1, for $T = 1/20$ and with various values of N, starting from $N = 200$ and multiplying successively by two the number of cells up to $N = 1600$. As expected, the upwind scheme (13) yields a numerical solution which converges (as the discretization parameters go to zero and under a CFL condition) to a weak solution of (7) (and even to its entropy solution), which is not a weak solution of (4), since the Rankine-Hugoniot conditions differ. At time $T = 1/20$, the shock for the solution of (4) is located at $x = 0.3$, while the shock of the solution of (7) is located at $x > 0.4$.

Remark 1 (Link with a non-conservative diffusion term). For the Burgers equation (4), upwinding may be seen as adding a diffusion, namely discretizing (since $u > 0$):

$$\partial_t u + \partial_x(u^2) - \partial_x((hu - 2ku^2)\partial_x u) = 0.$$

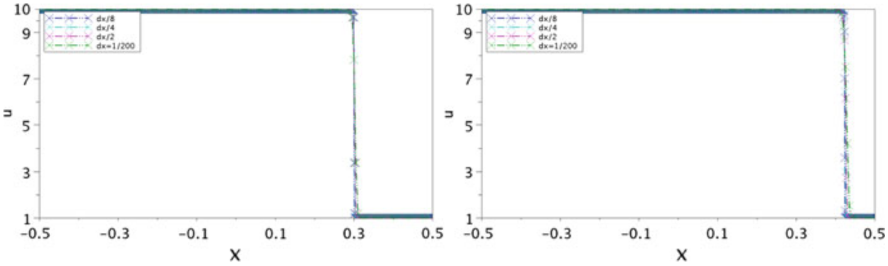


Fig. 1 Upwind Scheme for (4)-(9) (left) and (7)-(9) (right) with different mesh sizes, $CFL = 1$

Note that one has $hu - 2ku^2 \geq 0$ thanks to the CFL condition. For the Burgers square entropy equation (7), upwinding may be seen, formally, as solving the following parabolic equation (since $u > 0$): $\partial_t u^2 + (4/3)\partial_x(u^3) - \partial_x((2hu^2 - 4ku^3)\partial_x u) = 0$. This equation is equivalent to the following parabolic perturbation of the Burgers equation:

$$\partial_t u + \partial_x(u^2) - \frac{1}{u}\partial_x((hu^2 - 2ku^3)\partial_x u) = 0.$$

The third term at the left-hand side may be seen as a numerical diffusion (thanks to the CFL condition) which is not in a conservative form, because of the factor $1/u$. The above numerical results show that such a non conservative diffusion may lead to wrong discontinuities.

3 Numerical solution of the perturbed equation

We then discretize the perturbed equation (6) with $\epsilon = \epsilon_0 h^\alpha$, where $\epsilon_0 > 0$ and $\alpha > 0$ are fixed. Note that, setting $v = u^2$, (6) can also be recast as:

$$\partial_t v + \frac{4}{3}\partial_x(v^{\frac{3}{2}}) - v^{\frac{1}{2}}\epsilon_0 h^\alpha \partial_x(v^{-\frac{1}{2}}\partial_x v) = 0,$$

that is a nonlinear hyperbolic equation augmented with a nonlinear nonconservative diffusion term. The upwind finite volume discretization of this equation reads (in the u variable), with $u_i^{(0)}$ given by (10),

$$\left(u_i^{(n)}\right)^2 = \left(u_i^{(n-1)}\right)^2 + \frac{4k}{3h}\left[\left(u_{i-1}^{(n-1)}\right)^3 - \left(u_i^{(n-1)}\right)^3\right] + \frac{k}{h^2}\epsilon_0 h^\alpha\, u_i^{(n-1)}\left[u_{i-1}^{(n-1)} - 2u_i^{(n-1)} + u_{i+1}^{(n-1)}\right]. \quad (14)$$

We present in Figs. 2, 3 and 4 the numerical solutions obtained with (14) for $\alpha = 0.5$, $\alpha = 1$ and $\alpha = 2$ respectively, and for the same time $T = 1/20$,

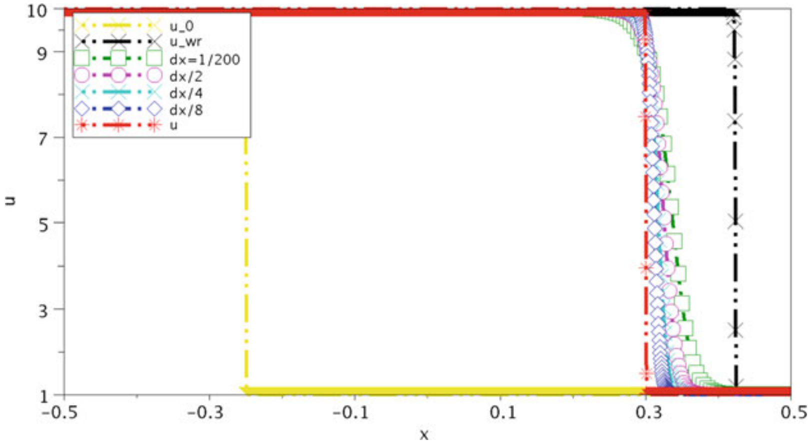


Fig. 2 Upwind Scheme for (6) with non conservative diffusion term, $\alpha = 0.5$

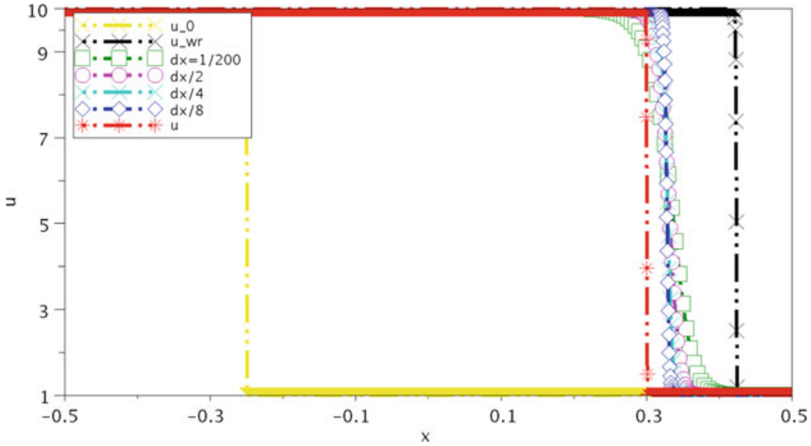


Fig. 3 Upwind Scheme for (6) with non conservative diffusion term, $\alpha = 1$

CFL $= 0.1$ and meshes as in Sect. 2. The parameter ϵ_0 is such that $\epsilon_0 h^\alpha = 0.2$ for $N = 200$ (whatever α may be). Figure 2 shows that for $0 < \alpha < 1$, the sequence of approximate solutions given by (14) converges to a weak solution of the initial Burgers equation (4), as h and k tend to 0, under a stability condition, which, since $\alpha < 1$, becomes more stringent than a CFL condition when h tends to zero. Figure 3 shows that for $\alpha > 1$, we obtain the convergence to the solution of (7); Fig. 4 shows that for $\alpha = 1$, the location of the discontinuity lies in between the discontinuities of the solution to (6) and (7). These results seem to indicate that the convergence to the solution of (7) (resp. (6)) occurs when the added diffusion dominates (resp. is dominated by) the numerical one.

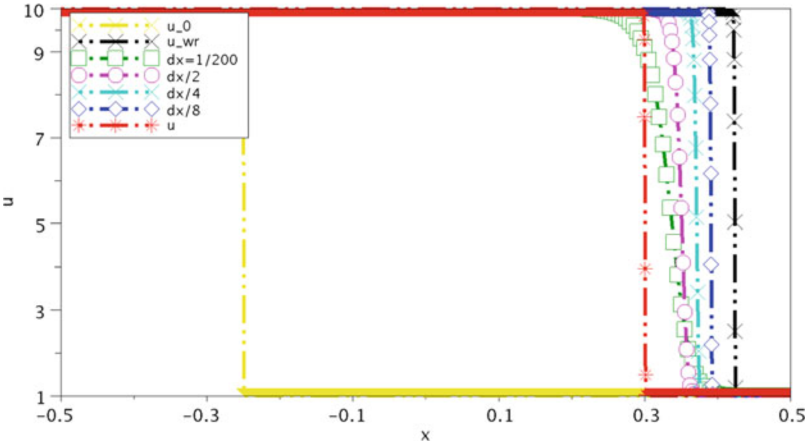


Fig. 4 Upwind Scheme for (6) with non conservative diffusion term, $\alpha = 2$

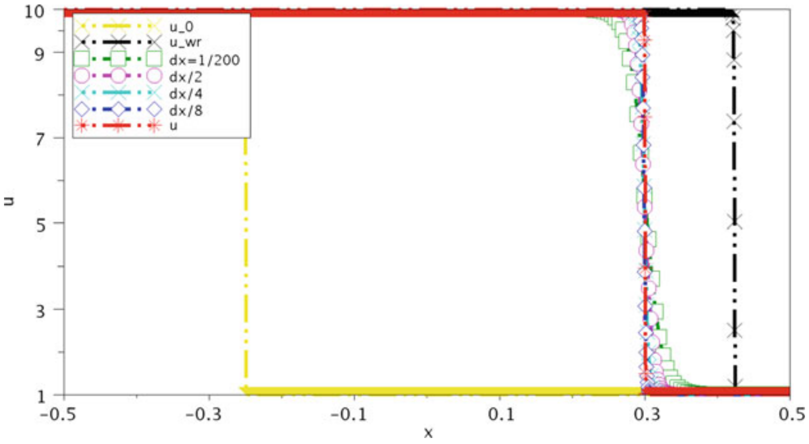


Fig. 5 Centered Scheme for (6) with non conservative diffusion term, $\alpha = 1$

Let us finally study the following finite volume centred scheme for Equation (7), which reads:

$$\big(u_i^{(n)}\big)^2 = \big(u_i^{(n-1)}\big)^2 + \frac{4k}{3h}\Big[\big(\frac{u_{i-1}^{(n-1)} + u_i^{(n-1)}}{2}\big)^3 - \big(\frac{u_i^{(n-1)} + u_{i+1}^{(n-1)}}{2}\big)^3\Big] + \frac{k}{h^2}\,\epsilon_0 h^\alpha\, u_i^{(n-1)}\Big[u_{i-1}^{(n-1)} - 2u_i^{(n-1)} + u_{i+1}^{(n-1)}\Big]. \quad (15)$$

Results for $\alpha = 1$, $\alpha = 1.5$ and $\alpha = 2$ (and ϵ_0 such that $\epsilon_0 h^\alpha = 0.2$ for $N = 200$, whatever α may be) are reported on Figs. 5, 6 and 7, respectively. The numerical

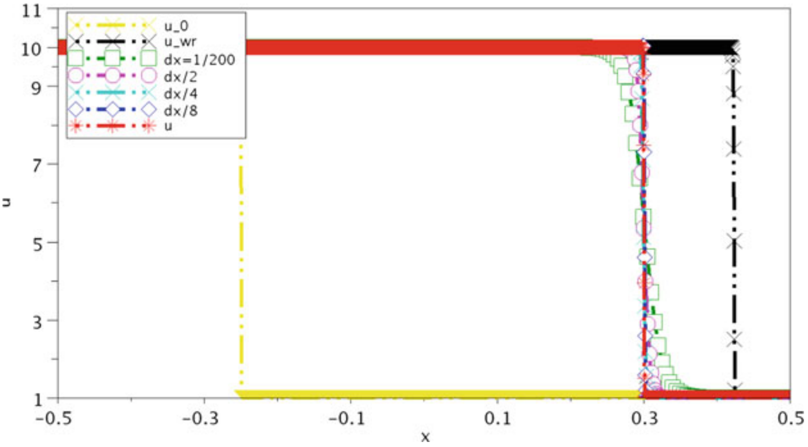


Fig. 6 Centered Scheme for (6) with non conservative diffusion term, $\alpha = 1.5$

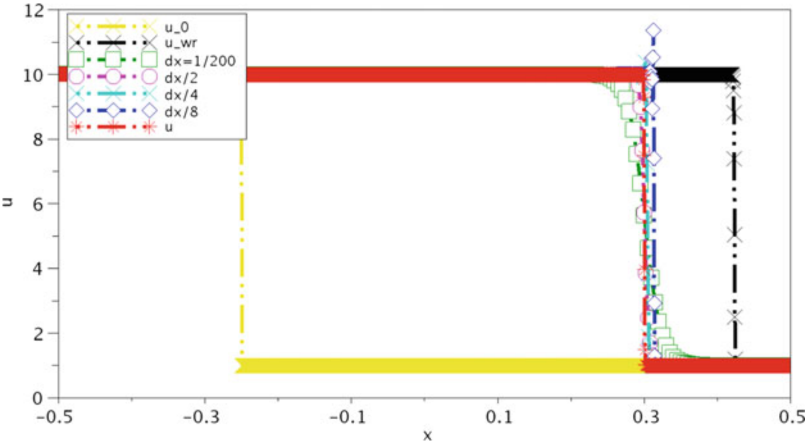


Fig. 7 Centered Scheme for (6) with non conservative diffusion term, $\alpha = 2$

solution now seems to converge to the solution of (7), at least for $\alpha \in (0, 2)$. For the finest mesh and $\alpha = 2$, the diffusion is no longer sufficient to prevent some spurious oscillations near the shock. Last but not least, the additional diffusion which is necessary to recover the right shock location is considerably reduced with respect to the upwind scheme (even if the scheme still appears more diffusive than the standard upwind scheme applied to (4)), which is encouraging in view of practical extensions to Euler equations.

Conclusion We tested two discretizations for the modified equation (6):

- an upwind scheme for which the solution converges to the weak solution of (4) if the viscous term is predominant with respect to the numerical diffusion, that is if $\epsilon = \epsilon_0 h^\alpha$, with $\epsilon_0 > 0$ and $\alpha \in (0, 1)$.
- a centred scheme which yields correct solutions for all values $\alpha \in (0, 2)$.

The extension of this work to Euler equations is under way, and results are encouraging. Indeed, it seems that we are able to build convergent schemes, even in the presence of shocks, using either the entropy or internal energy balance. A next step might be to use a nonlinear viscosity to avoid an excessive smearing of the solutions, following the ideas developed in [3].

References

1. R. Eymard, T. Gallouët, R. Herbin: Finite Volume Methods. Handbook of Numerical Analysis, Volume VII, North Holland (2000).
2. T. Gallouët, A. Larcher, J.-C. Latché: Convergence of a finite volume scheme for the convection-diffusion equation with L^1 data. To appear in Mathematics of Computation (2011).
3. J.-L. Guermond, R. Pasquetti, B. Popov: Entropy viscosity methods for nonlinear conservation laws. To appear in Journal of Computational Physics (2011).
4. R. Herbin, W. Kheriji, J.-C. Latché: An unconditionally stable Finite Element-Finite Volume pressure correction scheme for compressible Navier-Stokes equations. In preparation (2011).
5. B. Larrouturou: How to preserve the mass fractions positivity when computing compressible multi-component flows. Journal of Computational Physics, **95**, 59–84 (1991).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

On Discrete Sobolev–Poincaré Inequalities for Voronoi Finite Volume Approximations

Annegret Glitzky and Jens A. Griepentrog

Abstract We prove a discrete Sobolev–Poincaré inequality for functions with arbitrary boundary values on Voronoi finite volume meshes. We use Sobolev's integral representation and estimate weakly singular integrals in the context of finite volumes. We establish the result for star shaped polyhedral domains and generalize it to the finite union of overlapping star shaped domains.

Keywords Discrete Sobolev inequality, Sobolev integral representation, Voronoi finite volume mesh
MSC2010: 46E35, 46E39, 31B10

1 Introduction and notation

In this paper we study discrete Sobolev inequalities. In the continuous situation the Sobolev embedding estimates

$$\|u\|_{L^q(\Omega)} \leq C_q \|u\|_{H^1(\Omega)} \quad \forall u \in H^1(\Omega) \tag{1}$$

for $q \in [1, \infty)$ in two space dimensions and for $q \in [1, 2n/(n-2)]$ in $n \geq 3$ space dimensions are well known [1, 10, 15].

For the finite volume discretized situation some results can be found in [3, 6]. But these estimates concern only the case of zero boundary values. The two-dimensional case for admissible finite volume meshes (see [6, Definition 9.1]) is treated in [6, Lemma 9.5]. The corresponding three-dimensional result is proved in [3, Lemma 1]. For $p \in [1, 2]$, a discrete Sobolev inequality estimating the L^{p^*}-

Annegret Glitzky and Jens A. Griepentrog
WIAS, Mohrenstr. 39, D-10117 Berlin, Germany, e-mail: glitzky@wias-berlin.de, griepent@wias-berlin.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_56, © Springer-Verlag Berlin Heidelberg 2011

norm (where $p^* = np/(n-p)$ if $p < n$ and $p^* < \infty$ if $n = p = 2$) by the discrete $W^{1,p}$-norm is presented in [5, Proposition 2.2]. Moreover, for the zero boundary value case and $1 \leq p < \infty$, the discrete embedding of $W_0^{1,p}$ into L^q for some $q > p$, $1 \leq p < \infty$ is established in [7, sect. 5]. A corresponding result for discontinuous Galerkin methods working in the spaces of piecewise polynomial functions on general meshes is obtained in [4, Theorem 6.1]. The idea there is to follow Nirenberg's proof of Sobolev embeddings. Recently in [2], in the context of discontinuous Galerkin finite element methods, broken Sobolev–Poincaré inequalities were proved. There, known classical results in $BV(\Omega)$ and in Sobolev spaces $W^{1,p}(\Omega)$, together with local norm equivalence and global estimates for the reconstruction operator, lead to the desired estimates.

According to our knowledge and to the information of authors of the cited papers concerning finite volume schemes, finding discrete versions of the Sobolev inequality (1) for functions with arbitrary boundary values has been an open question up to now. Only a discrete Poincaré inequality ($q = 2$) is available in [6, Lemmas 10.2, 10.3] and [9, Lemma 4.2]. But in both papers the second step of the proof is done only for two space dimensions.

The aim of the present paper is to establish a discrete Sobolev–Poincaré inequality for functions with nonzero boundary values on Voronoi finite volume meshes. Such results can be applied to more general boundary value problems, for instance, to problems with inhomogeneous Dirichlet, Neumann, or mixed boundary conditions. The technique used here is an adaptation of Sobolev's integral representation and of the treatment of weakly singular integrals in the context of Voronoi finite volume meshes. The Voronoi property of the mesh essentially comes into play in the proofs of the potential theoretical results, Lemmas 1–3.

The plan of the paper is as follows. In the remainder of this section we introduce our notation. In Sect. 2 we formulate our assumptions and our main result, the discrete Sobolev–Poincaré inequality for star shaped domains (see Theorems 1 and 2 for a uniform estimate for a class of Voronoi finite volume meshes having comparable mesh quality). In Sect. 3 we collect three potential theoretical lemmas needed for the proof of our main result. In Sect. 4 we generalize the discrete Sobolev inequality to domains which are a finite union of overlapping star shaped domains (see Theorem 3). The last section contains some remarks concerning applications of discrete Sobolev inequalities.

Let $\Omega \subset B(0, R_0) \subset \mathbb{R}^n$, $n \in \mathbb{N}$, $n \geq 2$, be a bounded, open, polyhedral domain, and let $\partial\Omega$ be its boundary. We work with Voronoi finite volume meshes of Ω, and our notation is basically taken from [3, 6]. Moreover, for set valued arguments we write diam$(\cdot)$ for the diameter of the corresponding set. By mes$(\cdot)$ and mes$_d(\cdot)$ we denote the n- and d-dimensional Lebesgue measures, respectively.

A Voronoi finite volume mesh of Ω denoted by $\mathscr{M} = (\mathscr{P}, \mathscr{T}, \mathscr{E})$ is formed by a family of grid points $\mathscr{P}$ in $\overline{\Omega}$, a family $\mathscr{T}$ of Voronoi control volumes, and a family of relatively open parts of hyperplanes in $\mathbb{R}^n$ denoted by $\mathscr{E}$ (which represent the faces of the Voronoi boxes). For a Voronoi mesh we use the following notation:

For each grid point x_K of the set $\mathscr{P}$ the control volume K of the Voronoi mesh belonging to the point x_K is defined by

$$K = \{x \in \Omega : |x - x_K| < |x - x_L| \quad \forall x_L \in \mathscr{P},\ x_L \neq x_K\}, \quad K \in \mathscr{T}.$$

For $K, L \in \mathscr{T}$ with $K \neq L$ either the $(n-1)$-dimensional Lebesgue measure of $\overline{K} \cap \overline{L}$ is zero or $\overline{K} \cap \overline{L} = \overline{\sigma}$ for some $\sigma \in \mathscr{E}$. In the latter case the symbol $\sigma = K|L$ denotes the Voronoi surface between K and L. We introduce the following subsets of $\mathscr{E}$: The sets of interior and external Voronoi surfaces are denoted by $\mathscr{E}_{int}$ and $\mathscr{E}_{ext}$, respectively. Additionally, for every $K \in \mathscr{T}$ we call $\mathscr{E}_K$ the subset of $\mathscr{E}$ such that $\partial K = \overline{K} \setminus K = \cup_{\sigma \in \mathscr{E}_K} \overline{\sigma}$. Then $\mathscr{E} = \cup_{K \in \mathscr{T}} \mathscr{E}_K$. Moreover, for $\sigma \in \mathscr{E}$ we use the following notation: m_σ represents the $(n-1)$-dimensional measure of the Voronoi surface σ, and x_σ corresponds to the coordinates of the center of gravity of σ.

For $\sigma = K|L \in \mathscr{E}_{int}$ let d_σ be the Euclidean distance between x_K and x_L. For $K \in \mathscr{T}$, $\sigma \in \mathscr{E}_K$ we define $d_{K,\sigma}$ to be the Euclidean distance between x_K and the hyperplane containing σ. Then, in the case of (isotropic) Voronoi meshes we have $d_{K,\sigma} = \frac{d_\sigma}{2}$ for $\sigma \in \mathscr{E}_{int}$.

We work with half-diamonds $D_{K\sigma} = \{t x_K + (1-t)y : t \in (0,1),\ y \in \sigma\}$, where $n\,\mathrm{mes}(D_{K\sigma}) = m_\sigma d_{K,\sigma}$. Then due to our definitions,

$$n\,\mathrm{mes}(K) = \sum_{\sigma \in \mathscr{E}_K} m_\sigma d_{K,\sigma} \quad \forall K \in \mathscr{T}.$$

The mesh size is defined by $\mathrm{size}(\mathscr{M}) = \sup_{K \in \mathscr{T}} \mathrm{diam}(K)$. We denote by $X(\mathscr{M})$ the set of functions from Ω to $\mathbb{R}$ which are constant on each Voronoi box of the mesh. For $u \in X(\mathscr{M})$ the value in the Voronoi box $K \in \mathscr{T}$ is denoted by u_K. Finally, for $u \in X(\mathscr{M})$ the discrete H^1-seminorm $|u|_{1,\mathscr{M}}$ of u is defined by

$$|u|_{1,\mathscr{M}}^2 = \sum_{\sigma \in \mathscr{E}_{int}} \frac{m_\sigma}{d_\sigma} (D_\sigma u)^2,$$

where $D_\sigma u = |u_K - u_L|$, u_K is the value of u in the Voronoi box K, and $\sigma = K|L$.

2 Main result

First we formulate our *assumptions on the geometry and the meshes* as follows:

Assumption 1. We assume that the open, polyhedral domain $\Omega \subset B(0, R_0) \subset \mathbb{R}^n$ is star shaped with respect to some ball $B(0, R)$.

Let the function $\rho : \mathbb{R}^n \to [0, 1]$ be given by

$$\rho(y) = \begin{cases} \exp\left\{-\frac{R^2}{R^2 - |y|^2}\right\} & \text{if } |y| < R, \\ 0 & \text{if } |y| \geq R. \end{cases}$$

We introduce the piecewise constant approximations $\rho^{\mathscr{M}} \in X(\mathscr{M})$ as

$$\rho_K^{\mathscr{M}}(x) = \min_{y \in \overline{K}} \rho(y) \quad \text{for } x \in K. \tag{2}$$

Assumption 2. Let $\mathscr{M} = (\mathscr{P}, \mathscr{T}, \mathscr{E})$ be a Voronoi finite volume mesh of Ω with the property that $\mathscr{E}_K \cap \mathscr{E}_{ext} \neq \emptyset$ implies $x_K \in \partial\Omega$. Moreover, the local mesh size near $B(0, R)$ is assumed to be so small that there exists a constant $\rho_0 > 0$ such that $\int_\Omega \rho^{\mathscr{M}}(x)\,dx \geq \rho_0$.

Let us remark that the constant ρ_0 in Assumption 2 can be fixed, for instance, if we demand that for some $r \leq R/4$ we have $\mathrm{diam}(K) < r$ for all $x_K \in \mathscr{P}$ with $x_K \in B(0, R)$. Then, for almost all $x \in B(0, r)$ we find

$$\rho^{\mathscr{M}}(x) \geq \exp\left\{-\frac{R^2}{R^2 - (2r)^2}\right\} \geq \exp\left\{-\frac{4}{3}\right\}$$

and

$$\int_\Omega \rho^{\mathscr{M}}(x)\,dx \geq \mathrm{mes}(B(0,r)) \exp\left\{-\frac{R^2}{R^2 - (2r)^2}\right\},$$

which can be taken as ρ_0 in Assumption 2.

Under Assumption 2 there exist minimal constants $\kappa_1(\mathscr{M}) > 0$, $\kappa_2(\mathscr{M}) \geq 1$ such that the geometric weights fulfill

$$0 < \mathrm{diam}(\sigma) \leq \kappa_1(\mathscr{M})\, d_\sigma \quad \forall \sigma \in \mathscr{E}_{int} \tag{3}$$

and

$$\max_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} \max_{x \in \overline{\sigma}} |x_K - x| \leq \kappa_2(\mathscr{M}) \min_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} d_{K,\sigma} \quad \forall x_K \in \mathscr{P}. \tag{4}$$

Having in mind that

$$R_{K,out} := \max_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} \max_{x \in \overline{\sigma}} |x_K - x|, \quad R_{K,inn} := \min_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} d_{K,\sigma}$$

are the smallest radius of a circumscribed ball of K centered at x_K and the greatest radius of a ball fully contained in K and centered at x_K, respectively, inequality (4) implies that

$$R_{K,out} \leq \kappa_2(\mathscr{M}) R_{K,inn}.$$

Moreover, inequality (4) implies that

$$\max_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} |x_K - x_\sigma| \leq \kappa_2(\mathscr{M}) \min_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} d_{K,\sigma} \quad \forall x_K \in \mathscr{P}. \tag{5}$$

In this prescribed setting of a Voronoi finite volume mesh we establish the discrete Sobolev–Poincaré inequality:

Theorem 1. *Let Assumptions 1 and 2 be satisfied, and let $q \in [1, \infty)$ for $n = 2$ and $q \in [1, \frac{2n}{n-2})$ for $n \geq 3$, respectively. Then there exists a constant $c_q(\mathscr{M}) > 0$ depending only on n, q, Ω and the constants ρ_0, $\kappa_1(\mathscr{M})$, and $\kappa_2(\mathscr{M})$ such that*

$$\|u - m_\Omega(u)\|_{L^q(\Omega)} \leq c_q(\mathscr{M})\, |u|_{1,\mathscr{M}} \quad \forall u \in X(\mathscr{M}),$$

where $m_\Omega(u) = \mathrm{mes}(\Omega)^{-1} \int_\Omega u(x)\, dx$.

For the proof this theorem (see [14, sect. 4]) we adapt techniques used in [16, 17] to the discretized situation using Voronoi diagrams. To do so, we establish some discrete analogue for Sobolev's integral representation (see [16, sect. 116]) and of the treatment of weakly singular integral operators (see [16, sect. 115]).

Note that for $n \geq 3$, the discrete version of the Sobolev embedding $H^1(\Omega) \hookrightarrow L^{2n/(n-2)}(\Omega)$ for the critical Sobolev exponent can not be obtained by using only Sobolev's integral representation. This is exactly the same situation as for the continuous case (see [10, Chap. 7.8], [16, sect. 114–116], [17, sect. 8]).

We generalize our result to a class of Voronoi finite volume meshes having a unified mesh quality. Namely, we additionally assume the following for the meshes:

Assumption 3. There exist constants $\kappa_1 > 0$ and $\kappa_2 \geq 1$ such that the geometric weights fulfill $0 < \mathrm{diam}(\sigma) \leq \kappa_1\, d_\sigma$ for all $\sigma \in \mathscr{E}_{int}$ and $\max_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} |x_K - x_\sigma| \leq \kappa_2 \min_{\sigma \in \mathscr{E}_K \cap \mathscr{E}_{int}} d_{K,\sigma}$ for all $x_K \in \mathscr{P}$.

Now we can formulate the main theorem of our paper, the discrete Sobolev inequality uniformly on a class of Voronoi finite volume meshes $\mathscr{M}$ characterized by Assumptions 2 and 3:

Theorem 2. *Let Ω be an open bounded polyhedral subset of $\mathbb{R}^n$, and let $\mathscr{M}$ be a Voronoi finite volume mesh such that additionally Assumptions 1–3 are fulfilled. Let $q \in [1, \infty)$ for $n = 2$ and $q \in [1, \frac{2n}{n-2})$ for $n \geq 3$, respectively. Then there exists a constant $c_q > 0$ depending only on n, q, Ω and the constants in Assumptions 1–3 such that*

$$\|u - m_\Omega(u)\|_{L^q(\Omega)} \leq c_q\, |u|_{1,\mathscr{M}} \quad \forall u \in X(\mathscr{M}).$$

Note that the constant c_q in Theorem 2 depends on the fixed constants κ_1, κ_2 from Assumption 3 instead of $\kappa_1(\mathscr{M})$, $\kappa_2(\mathscr{M})$. The dependency on ρ_0 is of the same quality as in Theorem 1.

3 Potential theoretical lemmas

In this section we introduce three potential theoretical lemmas which are essential for the proof of the discrete Sobolev–Poincaré inequality, Theorem 1. The proofs of these lemmas can be found in [14, sect. 5].

Lemma 1. *Let $n \in \mathbb{N}$, $n \geq 2$ and Assumptions 1 and 2 be satisfied. Let $x_{K_0} \in \mathscr{P}$ be a fixed grid point and let $\sigma \in \mathscr{E}_{int}$ be an internal Voronoi surface with gravitational* *center x_σ. Then*

$$\operatorname{mes}\bigl(\{x \in B(0,R) : [x_{K_0}, x] \cap \sigma \neq \emptyset\}\bigr) \leq A_n \frac{m_\sigma}{|x_{K_0} - x_\sigma|^{n-1}},$$

where $A_n := \frac{1}{n} \max\{2,\ 4\kappa_1(\mathscr{M})\}^{n-1} \operatorname{diam}(\Omega)^n$.

Lemma 2. *Let $n \in \mathbb{N}$, $n \geq 2$ and Assumptions 1 and 2 be satisfied. Let $q \in (2,\infty)$ for $n = 2$ and $q \in (2, \frac{2n}{n-2})$ for $n \geq 3$. Moreover, let $\beta > 0$ be given by $2\beta = \frac{n}{q} - \frac{n}{2} + 1$, and let $x_{K_0} \in \mathscr{P}$ be a fixed grid point. Then*

$$\sum_{K \in \mathscr{T}} \sum_{\sigma \in \mathscr{E}_K} |x_{K_0} - x_\sigma|^{-n+2\beta}\, m_\sigma d_{K,\sigma} \leq B_n,$$

where $B_n := \frac{n}{2\beta} \max\{1 + 2\kappa_1(\mathscr{M}), 2\}^{n-2\beta} (2R_0)^{2\beta} \operatorname{mes}_{n-1}(\partial B(0,1))$.

Lemma 3. *Let $n \in \mathbb{N}$, $n \geq 2$ and Assumptions 1 and 2 be satisfied. Let $q \in (2,\infty)$ for $n = 2$ and $q \in (2, \frac{2n}{n-2})$ for $n \geq 3$. Moreover, let $\beta > 0$ be given by $2\beta = \frac{n}{q} - \frac{n}{2} + 1$, let $\sigma \in \mathscr{E}_{int}$ be a fixed inner Voronoi surface, and let x_σ denote its center of gravity. Then*

$$\sum_{K_0 \in \mathscr{T}} \sum_{\sigma_0 \in \mathscr{E}_{K_0}} |x_{K_0} - x_\sigma|^{-n+q\beta}\, m_{\sigma_0} d_{K_0,\sigma_0} \leq D_n,$$

where $D_n := \frac{n}{q\beta}\bigl(1 + \kappa_2(\mathscr{M})(1 + 2\kappa_1(\mathscr{M}))\bigr)^{n-q\beta} (2R_0)^{q\beta} \operatorname{mes}_{n-1}(\partial B(0,1))$.

4 Sobolev–Poincaré inequalities for more general domains

In this section we discuss how the results of Theorems 1 and 2, which hold true for star shaped domains Ω, can be used to obtain assertions for a more general situation. In the nondiscretized situation the result can be carried over to domains Ω, which are a finite union of star shaped domains Ω_i (see [16, sect. 118], [17, pp. 69–70]). In our discretized situation we assume the following:

Assumption 4. The open, connected, polyhedral domain $\Omega \subset B(0,R_0)$ is a finite union of open, polyhedral sets Ω_i, $i = 1,\dots,N$, and there are $\delta > 0$, $R > 0$, and points $z^i \in \Omega$ such that Ω_i, as well as the set $\Omega_{i\delta} := \Omega_i \cup \cup_{j \neq i}\{x \in \Omega_j : \operatorname{dist}(x,\Omega_i) < \delta\}$, is star shaped with respect to the ball $B(z^i, R)$, $i = 1,\dots,N$.

We introduce the functions

$$\rho_i : \mathbb{R}^n \to [0,1], \quad \rho_i(y) = \begin{cases} \exp\left\{-\frac{R^2}{R^2 - |y - z^i|^2}\right\} & \text{if } |y - z^i| < R, \\ 0 & \text{if } |y - z^i| \geq R \end{cases}$$

and their piecewise constant approximations $\rho_i^{\mathscr{M}} \in X(\mathscr{M})$. Concerning the mesh, we assume the following:

Assumption 5. Let $\mathscr{M} = (\mathscr{P}, \mathscr{T}, \mathscr{E})$ be a Voronoi finite volume mesh of Ω with the property that $\mathscr{E}_K \cap \mathscr{E}_{ext} \neq \emptyset$ implies $x_K \in \partial\Omega$. Moreover, the local mesh size near $B(z^i, R)$, $i = 1, \dots, N$, is assumed to be so small that there exists a constant $\rho_0 > 0$ such that $\int_\Omega \rho_i^{\mathscr{M}}(x)\, dx \geq \rho_0$, $i = 1, \dots, N$.

Then the discrete Sobolev–Poincaré inequalities remain true also for finite unions of δ-overlapping star shaped domains:

Theorem 3. *Let Assumptions 3–5 be satisfied, and $q \in [1, \infty)$ for $n = 2$ and $q \in [1, \frac{2n}{n-2})$ for $n \geq 3$, respectively. Then there exists a constant $C_q > 0$ depending only on n, q, Ω, and the constants in Assumptions 3–5 such that*

$$\|u - m_\Omega(u)\|_{L^q(\Omega)} \leq C_q\, |u|_{1,\mathscr{M}} \quad \forall u \in X(\mathscr{M}).$$

For a proof we refer to [14, sect. 6]. Since Theorem 2 is a direct consequence of Theorem 1 (with fixed κ_1, κ_2), this statement also remains true for more general domains characterized by Assumption 4.

5 Applications of discrete Sobolev inequalities

A functional analytic tool like a discrete Sobolev–Poincaré inequality enables us to treat discretized boundary value problems similarly to the corresponding continuous boundary value problems. Especially, if the embedding constants hold true for a class of meshes, uniform results with respect to the mesh can be obtained which can be used, for instance, for convergence results, too.

We were forced to prove the discrete Sobolev–Poincaré inequality by the analytical and numerical treatment of (nonlinear) reaction-diffusion systems. For the nondiscretized systems under consideration the free energy decays exponentially to its equilibrium value. We introduced a discretization scheme (Voronoi finite volume in space and fully implicit in time) which has the special property that it preserves the main features of the continuous problem, namely, positivity, dissipativity, and flux conservation (see [13]). For each fixed mesh we proved the exponential decay of the discretized free energy, too (see [11]). This proof works with the finite dimensional quantities.

To obtain uniform decay rates for a class of Voronoi finite volume meshes we had to translate the quantities from the finite dimensional discretized problems into expressions of functions from $X(\mathscr{M})$ being defined on Ω and being constant on Voronoi boxes of the corresponding meshes, and we had to consider limits of such functions belonging to sequences of Voronoi finite volume meshes to find a contradiction in the indirect proof of an estimate of the free energy by the dissipation rate (see [12, Theorem 3.2]). The essential ingredient in that proof is the discrete Sobolev–Poincaré inequality, Theorem 2.

For the application of discrete versions of Sobolev's inequality in the case of homogeneous Dirichlet boundary conditions we refer to [7, sect. 5]. Moreover, this inequality in the discrete $W_0^{1,p}$-setting (where $p \in (1,\infty)$) comes into play in the discretization of nonlinear elliptic problems of the form

$$-\operatorname{div} a(x,\nabla u) = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega$$

on general polyhedral meshes in n space dimensions. In [8, sect. 3] it is used to obtain an estimate of the approximate solution. In this setting the Caratheodory function $a : \Omega \times \mathbb{R}^n$ fulfills, with suitable positive constants c_1, c_2, and $d \in L^{p'}(\Omega)$,

$$\begin{aligned}
&a(x,\zeta)\cdot\zeta \geq c_1|\zeta|^p \quad \text{for almost all } x \in \Omega, \forall \zeta \in \mathbb{R}^n,\\
&(a(x,\zeta) - a(x,\chi))\cdot(\zeta - \chi) > 0 \quad \text{for almost all } x \in \Omega, \forall \zeta \neq \chi \in \mathbb{R}^n,\\
&|a(x,\zeta)| \leq d(x) + c_2|\zeta|^{p-1} \quad \text{for almost all } x \in \Omega, \forall \zeta \in \mathbb{R}^n.
\end{aligned}$$

Acknowledgements Jens A. Griepentrog gratefully acknowledges financial support by the DFG Research Center Matheon *Mathematics for key technologies* in the framework of the Matheon project C32: *Modeling of phase separation and damage processes in alloys*. Research about the topic of the text has also been carried out under financial support by the European XFEL GmbH in the framework of the XFEL project *Charge cloud simulations*.

References

1. Adams, R.A.: Sobolev Spaces. Academic Press, New York (1975)
2. Buffa, A., Ortner, C.: Compact embeddings of broken Sobolev spaces and applications. IMA J. Numer. Anal. **29**, 827–855 (2009)
3. Coudière, Y., Gallouët, T., Herbin, R.: Discrete Sobolev Inequalities and L^p error estimates for finite volume solutions of convection diffusion equations. M2AN Math. Model. Numer. Anal. **35**, 767–778 (2001)
4. Di Pietro, D., Ern, A.: Discrete functional analysis tools for discontinuous Galerkin methods with applications to the incompressible Navier–Stokes equations. Math. Comp. **79**, 1303–1330 (2010)
5. Droniou, J., Gallouët, T., Herbin, R.: A finite volume scheme for a noncoercive elliptic equation with measure data. SIAM J. Numer. Anal. **41**, 1997–2031 (2003)
6. 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. North-Holland, Amsterdam (2000)
7. 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**, 1009–1043 (2010)
8. Eymard, R., Gallouët, T., Herbin, R.: Cell centered discretisation of non linear elliptic problems on general multidimensional polyhedral grids. J. Numer. Math. **17**, 173–193 (2009)
9. Gallouët, T., Herbin, R., Vignal, M.H.: Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. SIAM J. Numer. Anal. **37**, 1935–1972 (2000)
10. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order. Springer, Berlin, Heidelberg, New York (1977)

11. Glitzky, A.: Exponential decay of the free energy for discretized electro-reaction-diffusion systems. Nonlinearity **21**, 1989–2009 (2008)
12. Glitzky, A.: Uniform Exponential Decay of the Free Energy for Voronoi Finite Volume Discretized Reaction-Diffusion Systems. WIAS Preprint **1443**, Berlin (2009)
13. Glitzky, A., Gärtner, K.: Energy estimates for continuous and discretized electro-reaction-diffusion systems. Nonlinear Anal. **70**, 788–805 (2009)
14. Glitzky, A., Griepentrog, J.A.: Discrete Sobolev–Poincaré inequalities for Voronoi finite volume approximations. SIAM J. Numer. Anal. **48**, 372–391 (2010)
15. Kufner, A., John, O., Fučik, S.: Function Spaces. Academia, Prague (1977)
16. Smirnow, W.I.: Lehrgang der höheren Mathematik V. Deutscher Verlag der Wissenschaften, Berlin (1962)
17. Sobolew, S.L.: Einige Anwendungen der Funktionalanalysis auf Gleichungen der mathematischen Physik. Akademie-Verlag, Berlin (1964)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Simple Second Order Cartesian Scheme for Compressible Flows

Y. Gorsse, A. Iollo, and L. Weynans

Abstract A simple second order scheme for compressible inviscid flows on cartesian meshes is presented. An appropriate Rieman solver is used to impose the impermeability condition. The level set function defines the immersed body and provides some useful geometrical data to increase the scheme accuracy. A modification of the convective fluxes computation for the cells near the solid ensures the boundary condition at second order accuracy. The same procedure is performed in each direction independently. An application to the simulation of a Ringleb flow is presented to demonstrate the accuracy of the method.

Keywords compressible flow, second order scheme, level set method, Riemann solver, cartesian meshes
MSC2010: 65M08, 65M12, 76N15

1 Introduction

The computation of flows in complex unsteady geometries is a crucial issue to perform realistic simulations of physical or biological applications like for instance biolocomotion (fish swimming or insect flight), turbomachines, windmills... To this end several class of methods exist. Here we are concerned with immersed boundary methods, i.e., integration schemes where the grid does not fit the geometry. These methods have been widely developed in the last 15 years, though the first methods were designed earlier (see for example [2,3,10]). The general idea behind immersed boundary methods is to take into account the boundary conditions by a modification of the equations to solve, either at the continuous level or at the discrete one,

Y. Gorsse, A. Iollo, and L. Weynans
Institut de Mathematiques de Bordeaux and INRIA Bordeaux Sud-Ouest, Université Bordeaux 1,
e-mail: yannick.gorsse@math.u-bordeaux1.fr, angelo.iollo@math.u-bordeaux1.fr,
lisl.weynans@math.u-bordeaux1.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_57, © Springer-Verlag Berlin Heidelberg 2011

rather than by the use of an adapted mesh. The main advantages of using these approaches, compared to methods using body-conforming grids, are that they are easily parallelizable and allow the use of powerful line-iterative techniques. They also avoid to deal with grid generation and grid adaptation, a prohibitive task when the boundaries are moving. A recent through review of immersed boundary methods is provided by Mittal and Iaccarino [6].

In this paper we present a simple globally second order scheme inspired by ghost cell approaches to solve compressible inviscid flows. In the fluid domain, away from the boundary, we use a classical finite-volume method based on an approximate Riemann solver for the convective fluxes and a centered scheme for the diffusive term. At the cells located on the boundary, we solve an *ad hoc* Riemann problem taking into account the relevant boundary condition for the convective fluxes by an appropriate definition of the contact discontinuity speed. These ideas can be adapted to reach higher order accuracy. However, here our objective is to device a method that can easily be implemented in existing codes and that is suitable for massive parallelization.

In section 2 we describe the finite volume scheme used to solve the flow equations in the fluid domain, away from the interface. In section 3 we introduce our method to impose impermeability condition. Finally, in section 4 we present a numerical test in two dimensions to validate the expected order of convergence and to discuss performance compared to others immersed boundary or body fitted methods.

2 Resolution in the fluid domain

We briefly describe how we solve the Euler equations in the fluid domain.

2.1 Governing equations

The compressible Euler equations are:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \mathbf{u} = 0 \tag{1}$$

$$\frac{\partial \rho \mathbf{u}}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u} + p\mathbf{n}) = 0 \tag{2}$$

$$\frac{\partial E}{\partial t} + \nabla \cdot ((E + p)\mathbf{u}) = 0 \tag{3}$$

where E denotes the total energy per unit volume. For a perfect gas

$$E = \frac{p}{\gamma - 1} + \frac{1}{2}\rho\mathbf{u}^2 \text{ and } p = \rho R T \tag{4}$$

2.2 Discretization

We focus on a two-dimensional setting. Let i and j be integers and consider the rectangular lattice generated by i and j, with spacing h_x and h_y in the x and y direction, respectively.

Let W be the conservative variables, $\mathscr{F}^x(W)$, $\mathscr{F}^y(W))$ the convective flux vectors in the x and y directions, respectively. By averaging the governing equations over any cell of the rectangular lattice we have

$$\frac{dW_{i\,j}}{dt} + \frac{1}{h_x}(\mathscr{F}^x{}_{i+1/2\,j} - \mathscr{F}^x{}_{i-1/2\,j}) + \frac{1}{h_y}(\mathscr{F}^y{}_{i\,j+1/2} - \mathscr{F}^y{}_{i\,j-1/2}) = 0 \tag{5}$$

where $W_{i\,j}$ is the average value of the conservative variables on the cell considered, $\mathscr{F}^x_{i+1/2\,j}$ the average flux in the x direction taken on the right cell side, and similarly for the other sides.

The average convective fluxes at cell interfaces are approximated using the Osher numerical flux function [9].

A second order Runge-Kutta scheme is used for the time integration.

3 A second order impermeability condition

For Euler equations, the boundary condition on the interface is the impermeability assumption, i.e., given normal velocity to the boundary (zero for a static wall, but non-zero for a moving solid). We are concerned with recovering second order accuracy on the impermeability condition.

3.1 Level set method

In order to improve accuracy at the solid walls crossing the grid cells we need additional geometric information. This information, mainly the distance from the wall and the wall normal, is provided by the distance function. The level set method, introduced by Osher and Sethian [8], is used to implicitly represent the interface of solid in the computational domain. We refer the interested reader to [11, 12] and [7] for recent reviews of this method. The zero isoline of the level set function represents the boundary Σ of the immersed body. The level set function is defined by:

$$\varphi(x) = \begin{cases} dist_\Sigma(x) & \text{outside of the solid} \\ -dist_\Sigma(x) & \text{inside of the solid} \end{cases} \tag{6}$$

A useful property of the level set function is:

$$\mathbf{n}(x) = \nabla\varphi(x) \tag{7}$$

where $\mathbf{n}(x)$ is the outward normal vector of the isoline of φ passing on x. In particular, this allows to compute the values of the normal to the interface, represented by the isoline $\varphi = 0$.

3.2 The impermeability condition in one dimension

To make the ideas clear, let us start from a simple case. The typical situation for a grid that does not fit the body is shown in Fig. 1. The plan is to modify the flux at the cell interface nearest to the boundary of the solid, in order to impose the boundary condition at the actual fluid-solid interface location. For a fixed body, we want to impose $u_b = u(x_b) = 0$ at the boundary point x_b where $\varphi(x_b) = 0$.

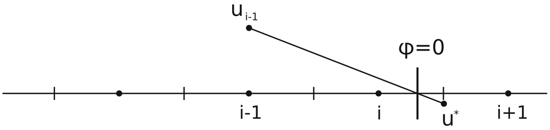


Fig. 1 Mesh near the solid. The interface lies between the center of cell i (fluid) and the center of cell $i+1$ (solid). The flux in $i+1/2$ has to be modified in order to account for the boundary conditions

Let u^* be the contact discontinuity speed resulting from the solution of the Riemann problem defined at the interface between cell i and cell $i+1$. The plan is to define a fictitious fluid state in $i+1$ such that the resulting velocity at the interface u^* takes into account, at the desired degree of accuracy, the boundary condition $u_b = 0$ in x_b. In particular, taking a second order Taylor expansion of the velocity at x_b, we have

$$u^* = u_b + \left(\frac{h_x}{2} - \varphi_i\right) \left.\frac{\partial u}{\partial x}\right|_{x_b} + O(h_x^2) \tag{8}$$

The boundary can be located anywhere between x_i and x_{i+1}, so to ensure a well defined derivative (if $x_b \to x_i$, $\frac{u_b - u_i}{x_b - x_i}$ is not numerically well defined), we use x_{i-1} instead of x_i to compute the first order derivative at the interface:

$$\left.\frac{\partial u}{\partial x}\right|_{x_b} = \frac{u_b - u_{i-1}}{h_x + \varphi_i} \tag{9}$$

To obtain u^* as the contact discontinuity speed of the Riemann problem, having computed the left state of the Riemann problem with the MUSCL reconstruction and slope limiters: $U_- = (u_-, p_-, c_-)$, we create the right state $U_+ = (-u_- + 2u^*, p_-, c_-)$, where c is the speed of sound. The left and right state of the variables p and c are chosen identical to express the continuity of these variables on the interface.

3.3 The impermeability condition in two dimensions

In two dimensions the flow equations are solved by computing independently the flux in each direction, so we want to apply in each direction the same kind of ideas as in one dimension in order to accurately enforce the boundary conditions. When the level set function changes sign between two cells, we need to modify the fluxes at the interface between these cells.

The interface point is the intersection between the interface ($\varphi = 0$) and the segment connecting the two cell centers concerned by the sign change (for example the points A, B and C on Fig. 2). For the flux computation, a fictitious state is created for instance between the cells (i, j) and $(i + 1, j)$ on Fig. 2. The boundary condition that we have to impose now is $\mathbf{u}_b.\mathbf{n}_b = 0$, where $\mathbf{u}_b$ is the speed of the fluid at the boundary, and $\mathbf{n}_b$ the outward normal vector of the body.

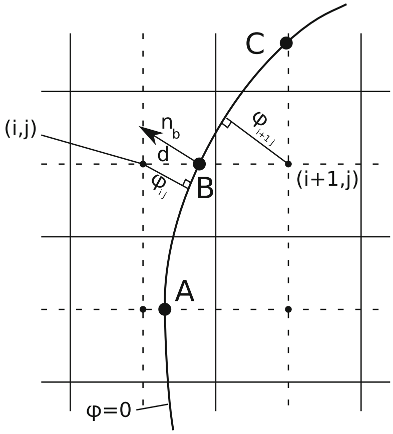


Fig. 2 Example of geometric configuration at the interface. B is the interface point located between (i, j) and $(i + 1, j)$. The flux on cell interface $(i + 1/2, j)$ is modified to enforce the boundary condition on B

With reference to Fig. 2, the level set function changes sign between $x_{i,j}$ and $x_{i+1,j}$ at point B. Let the normal vector point to the fluid side. If we assume that the boundary $\varphi = 0$ is locally rectilinear, using the side splitter theorem, the distance between $x_{i,j}$ and B is

$$d = \frac{h_x |\varphi_{i,j}|}{|\varphi_{i,j}| + |\varphi_{i+1,j}|} \tag{10}$$

and the normal vector $\mathbf{n}_b$ is computed by

$$\mathbf{n}_b = \mathbf{n}_{i,j} + \frac{d}{h_x} \left(\mathbf{n}_{i+1,j} - \mathbf{n}_{i,j} \right) \tag{11}$$

where $\mathbf{n}_{i,j}$ is a second order centered finite-difference approximation of $\nabla\varphi$ at point (i, j). To impose the boundary condition at the interface point B, we determine a value of the contact discontinuity speed $\mathbf{u}^*$, relative to a Riemann problem defined in the direction normal to the cell side through $x_{i+1/2,j}$, consistent at second order accuracy with $\mathbf{u}_b \cdot \mathbf{n}_b = 0$ in B. Figure 3 illustrates graphically the following steps. Let the normal component of $\mathbf{u}^*$ be $u^*_n = \mathbf{u}^* \cdot \mathbf{n}_b$.
u^*_n is computed with a second order Taylor expansion of the normal velocity at the interface point, that is:

$$u^*_n = \mathbf{u}_b \cdot \mathbf{n} + \left(\frac{h_x}{2} - d \right) \frac{\mathbf{u}_b \cdot \mathbf{n} - \mathbf{u}_{i-1} \cdot \mathbf{n}}{h_x + d} + O(h_x^2). \tag{12}$$

Then, let determine u^*_τ the tangential component of $\mathbf{u}^*$ by $u^*_\tau = \mathbf{u}^* \cdot \tau_b$, τ_b being the vector tangential to the interface at point B. We use the continuity property of the tangential component of the velocity to define u^*_τ according to

$$u^*_\tau = \mathbf{u}_- \cdot \tau \tag{13}$$

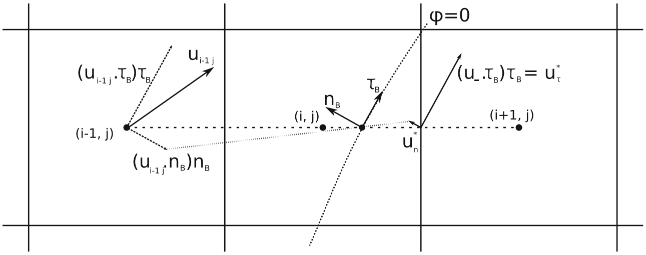


Fig. 3 Graphical illustration of the construction of the $\mathbf{u}^*$ vector

Finally we decompose $\mathbf{u}^*$ in the canonical basis by its horizontal and vertical components u^* and v^*, that is

$$u^* = u_n^* n_x + u_\tau^* \tau_x \tag{14}$$

$$v^* = u_n^* n_y + u_\tau^* \tau_y \tag{15}$$

To obtain $\mathbf{u}^*$ as the contact discontinuity speed of the Riemann problem, the left state resulting from the MUSCL reconstruction with slope limiters being $U_- = (u_-, v^*, p_-, c_-)$, we choose the right state to be $U_+ = (-u_- + 2u^*, v^*, p_-, c_-)$.

4 Numerical illustration: The Ringleb flow

The objective is to ascertain the actual accuracy obtained at the solid interface.

The Ringleb flow refers to an exact solution of Euler equations. The solution is obtained with the hodograph method, see [13]. The streamlines and iso-Mach lines are shown on Fig. 4.

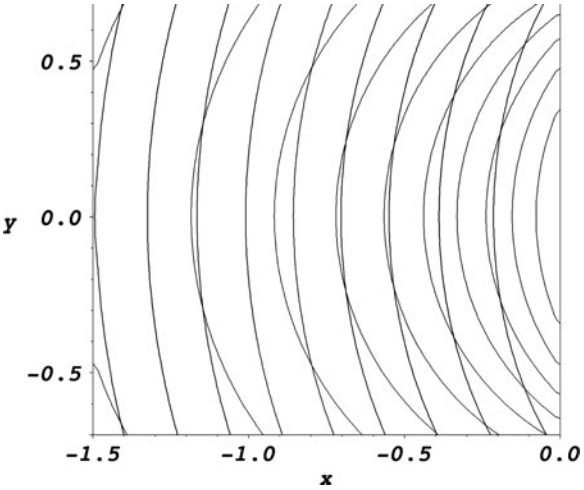


Fig. 4 Streamlines (black) and iso-Mach lines (grey) of the Ringleb flow

The exact solution is formulated in (θ, V) variables with $u = V\cos\theta$, $v = V\sin\theta$ and $V = \sqrt{u^2 + v^2}$.
The stream function is given by $\Psi = \frac{\sin\theta}{V}$.
The streamlines equations are:

$$x = \frac{1}{2\rho}\left(\frac{1}{V^2} - 2\Psi^2\right) + \frac{L}{2}, \qquad y = \frac{\sin\theta\cos\theta}{\rho V^2} \tag{16}$$

with:

$$L = -\left(\frac{1}{2}\ln\frac{1+c}{1-c} - \frac{1}{c} - \frac{1}{3c^3} - \frac{1}{5c^5}\right), \quad c^2 = 1 - \frac{\gamma-1}{2}V^2, \quad \rho = c^5 \tag{17}$$

In our test case, the computational domain is [–0.5; –0.1] × [–0.6; 0] and we numerically solve the flow between the streamlines $\Psi_1 = 0.8$ and $\Psi_2 = 0.9$. The inlet and outlet boundary condition are supersonic for $y = -0.6$ and $y = 0$ respectively. The convergence orders are calculated for each variable in L_1, L_2, L_∞ norms on four different grids 32×48, 64×96, 128×192 and 256×384 and presented on Table 1.

Table 1 Global orders of convergence for each variable

Variables	L_1 norm	L_2 norm	L_∞ norm
x-velocity	2.04	1.68	1.28
y-velocity	1.97	1.6	1.13
pressure	2.0	2.02	1.97
sound velocity	1.95	1.58	1.03
entropy	1.9	1.49	1.08

The error for several variables is order 1 for the L_{infty} norm. Colella et al. [5] obtain the same kind of results on other test cases. One argument developed in [5] to explain this order degradation is that the solid wall is characteristic for entropy, and hence the error on this variable accumulates from inlet to outlet. For the same test case, Coirier and Powell [4] observed also a convergence order between one and two in the case of their own cartesian method. In [1], Abgrall et al. obtain a L^2 numerical order of convergence for the density equal to 1.5 with their second order residual distribution scheme.

5 Conclusions

In this paper we have presented a new second order cartesian method to solve compressible flows in complex domains. This method is based on a classical finite volume approach, but the values used to compute the fluxes at the cell interfaces near the solid boundary are determined so to satisfy the boundary conditions with a second order accuracy. A test case for inviscid flows was presented. The order of convergence of the method is similar to those observed in the literature. This method is particularly simple to implement, as it doesn't require any special cell reconstruction at the solid-wall interface. The extension to three-dimensional cases is natural as the same procedure at the boundary is repeated in each direction. Forthcoming work will concern the extension of the present approach to multi-physics problems.

References

1. ABGRALL, R., LARAT, A., AND RICCHIUTO, M. Construction of very high order residual distribution schemes for steady inviscid flow problems on hybrid unstructured meshes, in press, 2010.
2. BERGER, M., AND LEVEQUE, R. An adaptive cartesian mesh algorithm for the euler equations in arbitrary geometries, 1989.
3. BERGER, M., AND LEVEQUE, R. Stable boundary conditions for cartesian grid calculations. *Computing systems in Engineering 1*, 2-4 (1990), 305–311.
4. COIRIER, W., AND POWELL, K. An accuracy assessment of cartesian mesh approaches for the euler equations. *J. Comput. Phys. 117* (1995), 121–131.
5. COLELLA, P., GRAVES, D., KEEN, B., AND MODIANO, D. A cartesian grid embedded boundary method for hyperbolic conservation laws. *J. Comput. Phys. 211*, 1 (2006), 347–366.
6. MITTAL, R., AND IACCARINO, G. Immersed boundary methods, 2005.
7. OSHER, S., AND FEDKIW, R. *Level Set Methods and Dynamic Implicit Surfaces*. Springer, 2003.
8. OSHER, S., AND SETHIAN, J. A. Fronts propagating with curvature-dependent speed: Algorithms based on hamiltonjacobi formulations. *J. Comput. Phys. 79*, 12 (1988).
9. OSHER, S., AND SOLOMAN, F. Upwind difference schemes for hyperbolic systems of conservation laws. *Math. Comp. 38*, 158 (April 1982), 339–374.
10. PESKIN, C. The fluid dynamics of heart valves: experimental, theoretical and computational methods. *Annu. Rev. Fluid Mech. 14* (1981), 235–259.
11. SETHIAN, J. A. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, Cambridge, UK, 1999.
12. SETHIAN, J. A. Evolution, implementation, and application of level set and fast marching methods for advancing fronts. *J. Comput. Phys. 169* (2001), 503–555.
13. SHAPIRO, A. *The Dynamics and Thermodynamics of Compressible Fluid Flow*. Ronald Press, 1953.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Efficient Implementation of High Order Reconstruction in Finite Volume Methods

Florian Haider, Pierre Brenner, Bernard Courbet, and Jean-Pierre Croisille

Abstract The paper presents a new algorithm for high order piecewise polynomial reconstruction. This algorithm computes a high order approximant in a given cell using data from adjacent cells in several steps, eliminating the need to handle directly large reconstruction stencils. The resulting high order finite volume method is well suited for modern parallel and vector (array) computers.

Keywords High Order Scheme, Unstructured Grid, Finite Volume Method, MUSCL
MSC2010: 65M08,65D15

1 Introduction

The finite volume MUSCL method to solve hyperbolic conservation laws was introduced by B. Van Leer in [7] thirty years ago. The main idea is to increase the accuracy of the first order finite volume scheme by a piecewise linear reconstruction used to evaluate upwinded fluxes at the cell interfaces.

Practical applications for convection dominated flows in complex geometries have motivated many extensions of the MUSCL approach to unstructured grids.

Florian Haider and Bernard Courbet
ONERA 29 avenue de la Division Leclerc 92322 Châtillon France,
e-mail: florian.haider@onera.fr, bernard.courbet@onera.fr

Pierre Brenner
ASTRIUM Space Transportation - Aerodynamics BP3002 - 78133 Les Mureaux France,
e-mail: pierre.brenner@astrium.eads.net

Jean-Pierre Croisille
Université Paul Verlaine-Metz - UFR MIM Département de Mathématiques
Ile du Saulcy 57045 Metz France, e-mail: croisil@poncelet.univ-metz.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_58, © Springer-Verlag Berlin Heidelberg 2011

A typical example is the flow solver CEDRE developed at ONERA. It uses a cell centered finite volume scheme with piecewise linear reconstruction on general polyhedral grids to solve the compressible Navier Stokes equations. A large choice of physical models is available in CEDRE: turbulence, combustion, diphasic flow, radiation etc.

Our experience has shown that second order accuracy becomes insufficient for LES and to capture contact discontinuities. The easiest way to increase the spatial accuracy is to replace the linear interpolation with quadratic or cubic ones. Indeed, the MUSCL scheme with quadratic reconstruction (3rd order) was already discussed by B. Van Leer [7]. The quadratic approach was extended to unstructured grids [1, 2]. The need for large (non compact) stencils seems to have limited the use of cubic reconstructions (4th order), although some practical applications exist [6].

For reasons of performance, the computation of a polynomial reconstruction on a grid cell must be local, using data in neighboring cells only. On the other hand, high order approximation requires sufficiently many data samples, which means that data from cells beyond adjacent cells in the grid must be accessed.

This paper shows how to compute a high order approximant in a given cell using data from adjacent cells in several steps, eliminating the need to handle directly large reconstruction stencils. No additional degrees of freedom are added: the independent variables are the cell averages of the conserved quantity. The resulting high order finite volume method is well suited for modern parallel and vector (array) computers. This aspect is of primary importance for large scale industrial software.

2 Semi-discrete High Order Finite Volume Scheme

- *Geometric notation*: an unstructured grid is a triangulation of a domain $\Omega \subset \mathbb{R}^d$ consisting of N general polyhedra. The cell with number α is denoted $\mathcal{T}_\alpha$, with barycenter $\boldsymbol{x}_\alpha$ and d-volume $|\mathcal{T}_\alpha|$. The face $\mathcal{A}_{\alpha\beta}$, with barycenter $\boldsymbol{x}_{\alpha\beta}$, has a normal vector $\boldsymbol{a}_{\alpha\beta}$ oriented from cell $\mathcal{T}_\alpha$ to $\mathcal{T}_\beta$ and of length $\left\|\boldsymbol{a}_{\alpha\beta}\right\|$ equal to the surface $\left|\mathcal{A}_{\alpha\beta}\right|$. The oriented normal unit vector of the face $\mathcal{A}_{\alpha\beta}$ is $\boldsymbol{\nu}_{\alpha\beta}$. Furthermore, define $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha$ and

$$z_{\alpha\beta}^{(k)} \triangleq \frac{1}{\left|\mathcal{T}_\beta\right|}\int_{\mathcal{T}_\beta} (\boldsymbol{x}-\boldsymbol{x}_\alpha)^{\otimes k}\, dx = \frac{1}{\left|\mathcal{T}_\beta\right|}\int_{\mathcal{T}_\beta} \underbrace{(\boldsymbol{x}-\boldsymbol{x}_\alpha)\otimes\cdots\otimes(\boldsymbol{x}-\boldsymbol{x}_\alpha)}_{m \text{ factors}}\, dx \tag{1}$$

The k^{th} moment of cell $\mathcal{T}_\alpha$ is then defined as $\boldsymbol{x}_\alpha^{(k)} \triangleq z_{\alpha\alpha}^{(k)}$. Note that $\boldsymbol{x}_\alpha^{(1)} = 0$. For a locally integrable function u, define its average over cell $\mathcal{T}_\alpha$ as $\overline{u}_\alpha$.

- *Semi-discrete* MUSCL *scheme*: consider a hyperbolic conservation law with flux $\boldsymbol{f}$. Its balance equation over a cell $\mathcal{T}_\alpha$ can be written as

$$\frac{\mathrm{d}\overline{u}_\alpha(t)}{\mathrm{d}t} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}\,(u(\boldsymbol{x},t))\; d\sigma\,. \tag{2}$$

The semi-discrete MUSCL discretization of such a conservation law gives the finite volume scheme

$$\frac{\mathrm{d}\overline{u}_\alpha(t)}{\mathrm{d}t} = -\frac{1}{|\mathcal{T}_\alpha|}\sum_\beta \int_{\mathcal{A}_{\alpha\beta}} \widetilde{f}_{\alpha\beta}\left(w_\alpha\left[\overline{\mathrm{u}}(t)\right](\boldsymbol{x}), w_\beta\left[\overline{\mathrm{u}}(t)\right](\boldsymbol{x})\right) d\sigma. \tag{3}$$

In (3), $\widetilde{f}_{\alpha\beta} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ is a *numerical flux* that is *consistent* with $\boldsymbol{f}$: $\widetilde{f}_{\alpha\beta}(u,u) = \boldsymbol{\nu}_{\alpha\beta}\cdot\boldsymbol{f}(u)$. The functions w_α and w_β are reconstructed from the cell averages $\overline{\mathrm{u}}(t) = (\overline{u}_1(t), \dots, \overline{u}_N(t))$. The dependence of w_α on the cell averages is denoted by square brackets $w_\alpha[\overline{\mathrm{u}}(t)]$ and the dependence on $\boldsymbol{x}$ by $w_\alpha[\overline{\mathrm{u}}(t)](\boldsymbol{x})$.

- *Accuracy*: the piecewise reconstruction operates on each cell so that only the cell averages in a certain neighborhood – *the reconstruction stencil* – of the cell $\mathcal{T}_\alpha$ determine the approximant w_α. Assume that the reconstruction satisfies for all smooth functions u and uniformly in $\boldsymbol{x} \in \mathcal{T}_\alpha$ for all cells $\mathcal{T}_\alpha$

$$\left| w_\alpha\left[\overline{\mathrm{u}}(t)\right](\boldsymbol{x}) - u(\boldsymbol{x},t)\right| \le O\left(h^{k+1}\right). \tag{4}$$

  Assuming $\boldsymbol{f}$ is Lipschitz continuous, one verifies easily that (3) is k^{th} order accurate.
- *Conservation*: the reconstruction is required to be *conservative*, i.e. the mean value of the function $w_\alpha[\overline{\mathrm{u}}(t)]$ over the cell $\mathcal{T}_\alpha$ must always be $\overline{u}_\alpha(t)$.

3 High Order Polynomial Reconstruction

This section gives a short overview of *k-exact reconstruction* along the line of [1,4]. The goal is to reconstruct the functions w_α used in (3) in such a way that they satisfy (4). The time dependency is dropped to simplify the notation.

Let $\mathbb{P}_k\left(\mathbb{R}^d\right)$ be the space of polynomials of degree k in $\mathbb{R}^d$. In each cell $\mathcal{T}_\alpha$, the reconstruction procedure is represented by the linear operator

$$\mathfrak{R}_\alpha : \mathbb{R}^N \to \mathbb{P}_k\left(\mathbb{R}^d\right) \quad ; \quad \overline{\mathrm{u}} \longmapsto w_\alpha[\overline{\mathrm{u}}]\,. \tag{5}$$

Define a *neighborhood* of cell $\mathcal{T}_\alpha$ as a set of cells $\mathbb{W}_\alpha \subset \{1\dots,N\}$ such that $\alpha \in \mathbb{W}_\alpha$ and associate with $\mathbb{W}_\alpha$ a local cell average operator

$$\mathfrak{P}_{k;\mathbb{W}_\alpha} : \mathbb{P}_k\left(\mathbb{R}^d\right) \to \mathbb{R}^N \tag{6}$$

given by $\left(\mathfrak{P}_{k;\mathbb{W}}(p)\right)_\beta = \overline{p}_\beta$ if $\beta \in \mathbb{W}_\alpha$ and $\left(\mathfrak{P}_{k;\mathbb{W}}(p)\right)_\beta = 0$ if $\beta \notin \mathbb{W}_\alpha$.

A reconstruction operator $\mathfrak{R}_\alpha : \mathbb{R}^N \to \mathbb{P}_k\left(\mathbb{R}^d\right)$ is called *k-exact* if it is a left inverse of (6)

$$\mathfrak{R}_\alpha \mathfrak{P}_{k;\mathbb{W}_\alpha} = \mathrm{Id}_{\mathbb{P}_k\left(\mathbb{R}^d\right)} \,. \tag{7}$$

It can be shown that, under certain conditions, (7) provides an approximation error (4) that is $O\left(h^{k+1}\right)$ [3].

The space of symmetric tensors of rank m in $\mathbb{R}^d$ is denoted $\mathcal{S}^m\left(\mathbb{R}^d\right)$. For all $\boldsymbol{a}, \boldsymbol{b} \in \mathcal{S}^m\left(\mathbb{R}^d\right)$ and $\boldsymbol{c} \in \mathbb{R}^d$ define

$$\boldsymbol{a} \bullet \boldsymbol{b} \triangleq \sum_{i_1=1}^{d} \cdots \sum_{i_m=1}^{d} a_{i_1\cdots i_m} b_{i_1\cdots i_m} \tag{8}$$

$$(\boldsymbol{a} \cdot \boldsymbol{c})_{j_1\cdots j_{m-1}} \triangleq \sum_{j_m=1}^{d} a_{j_1\cdots j_{m-1} j_m} c_{j_m} \,. \tag{9}$$

A function u is called *k-exact* on $\mathbb{W}_\alpha$ if the restriction of u to the cells in $\mathbb{W}_\alpha$ is a polynomial of degree k. Note that the m^{th} derivative of u can be considered as an element of $\mathcal{S}^m\left(\mathbb{R}^d\right)$.

A *k-exact m^{th} derivative on the neighborhood* $\mathbb{W}_\alpha$ *at cell* $\mathcal{T}_\alpha$ is defined to be a *linear map* $\boldsymbol{w}_\alpha^{(m|k)} : \mathbb{R}^N \longrightarrow \mathcal{S}^m\left(\mathbb{R}^d\right)$ such that for all polynomials p of degree k

$$\boldsymbol{w}_\alpha^{(m|k)} \left[\mathfrak{P}_{k;\mathbb{W}_\alpha}(p)\right] = D^{(m)} p\Big|_{\boldsymbol{x}_\alpha} \,. \tag{10}$$

Since a polynomial is determined by its cell average and its m^{th} derivatives at a point $\boldsymbol{x}_\alpha$, a k-exact reconstruction operator is equivalent to a set of *k-exact m^{th}* derivatives $\boldsymbol{w}_\alpha^{(m|k)}$ for $1 \le m \le k$. By linearity, (10) can be expressed as

$$\boldsymbol{w}_\alpha^{(m|k)} [\overline{\mathrm{u}}] = \sum_{\beta \in \mathbb{W}_\alpha} \boldsymbol{w}_{\alpha\beta}^{(m|k)} \overline{u}_\beta \tag{11}$$

In (11), the symmetric tensors $\boldsymbol{w}_{\alpha\beta}^{(m|k)}$, called the *reconstruction coefficients* of $\boldsymbol{w}_\alpha^{(m|k)}$, depend only on the local cell geometry. In principle, a complete set of $\boldsymbol{w}_{\alpha\beta}^{(m|k)}$ can be computed by applying (10) to a basis of the space $\mathbb{P}_k\left(\mathbb{R}^d\right)$ and solving the resulting linear system in the least squares sense. Since this algorithm computes the $\boldsymbol{w}_\alpha^{(m|k)}$ directly from the cell averages, we will refer to this method in Sect. 5 as *direct least squares reconstruction* (DLS). However, an obvious drawback of this method is that its implementation requires the computation of (11) over large stencils $\mathbb{W}_\alpha$.

Taking into account the constraint of conservation and using k-exact m^{th} derivatives (10), one can write the reconstructed polynomial at cell $\mathcal{T}_\alpha$ in the general form

$$w[\overline{\mathrm{u}}](\boldsymbol{x}) = \overline{u}_\alpha + \sum_{m=1}^{k} \frac{1}{m!} \boldsymbol{w}_\alpha^{(m|k)} [\overline{\mathrm{u}}] \bullet \left[(\boldsymbol{x} - \boldsymbol{x}_\alpha)^{\otimes m} - \boldsymbol{x}_\alpha^{(m)}\right] \,. \tag{12}$$

In (12), $(\boldsymbol{x}-\boldsymbol{x}_\alpha)^{\otimes m}$ is defined as in (1) and $\boldsymbol{x}_\alpha^{(m)} \triangleq \boldsymbol{z}_{\alpha\alpha}^{(m)}$.

When a k-exact m^{th} derivative (11) is applied to a polynomial p of degree $(k+1)$, the reconstruction error can be expressed as

$$\boldsymbol{w}_\alpha^{(m|k)}\left[\mathbb{p}\right] - D^{(m)}p\Big|_{\boldsymbol{x}_\alpha} = \frac{1}{(k+1)!}\sum_{\beta\in\mathbb{W}_\alpha} \boldsymbol{w}_{\alpha\beta}^{(m|k)}\left(\boldsymbol{z}_{\alpha\beta}^{(k+1)} \bullet D^{(k+1)}p\Big|_{\boldsymbol{x}_\alpha}\right). \qquad (13)$$

The interest of (13) is that a $(k+1)$-exact $(k+1)^{\text{th}}$ derivative can be used to compute the right hand side of (13) and to subtract it from $\boldsymbol{w}_\alpha^{(m|k)}$, making $\boldsymbol{w}_\alpha^{(m|k)}$ $(k+1)$-exact.

Finally, we introduce the following smoothing technique: let $\mathbb{V}_\alpha$ be the set of direct neighbors of cell $\mathcal{T}_\alpha$, including $\mathcal{T}_\alpha$ itself. Let $0 \le \xi_\beta \le 1$ be such that $\sum_{\beta\in\mathbb{V}_\alpha}\xi_\beta = 1$. If a set of *$k$-exact k^{th} derivatives* $\boldsymbol{w}_\beta^{(k|k)}$ is known, one can define a new *k-exact k^{th} derivative* $\widetilde{\boldsymbol{w}}_\alpha^{(k|k)}$ as a convex combination

$$\widetilde{\boldsymbol{w}}_\alpha^{(k|k)}\left[\overline{\mathbb{u}}\right] = \sum_{\beta\in\mathbb{V}_\alpha}\xi_\beta \boldsymbol{w}_\beta^{(k|k)}\left[\overline{\mathbb{u}}\right]. \qquad (14)$$

It is natural to choose the weights ξ_β in (14) proportional to the cell volumes $\left|\mathcal{T}_\beta\right|$.The stencil of (14) is larger which increases stability, see [3,5].

4 Efficient Algorithms for High Order Reconstruction

The computation of (11) involves large (non compact) stencils. To avoid this undesirable feature, we compute a $(k+1)$-exact $(k+1)^{\text{th}}$ derivative not directly from the cell averages, but from a family of k-exact k^{th} derivatives $\boldsymbol{w}_\beta^{(k|k)}$ at cells $\mathcal{T}_\beta$ for β in a small neighborhood $\mathbb{W}_\alpha$ of cell $\mathcal{T}_\alpha$. This is done as follows:

Let $\mathbb{W}_\alpha$ be a neighborhood of cell $\mathcal{T}_\alpha$. Let $\boldsymbol{w}_\beta^{(k|k)}$ be a family of k-exact k^{th} derivatives with stencils $\mathbb{W}_\beta^{(k)}$ at cells $\mathcal{T}_\beta$ for $\beta\in\mathbb{W}_\alpha$. Assume that $\bigcup_{\beta\in\mathbb{W}_\alpha}\mathcal{T}_\beta$ is path connected where the paths are piecewise C^1. Let $m_\alpha \triangleq |\mathbb{W}_\alpha| - 1$ and define the linear operator

$$\mathfrak{J}_{\mathbb{W}_\alpha}^{(k+1)} : \mathcal{S}^{(k+1)}\left(\mathbb{R}^d\right) \longrightarrow \left(\mathcal{S}^{(k)}\left(\mathbb{R}^d\right)\right)^{m_\alpha}. \qquad (15)$$

The i^{th} component of (15) is defined using $\boldsymbol{h}_{\alpha\beta} = \boldsymbol{x}_\beta - \boldsymbol{x}_\alpha$, (1), (8) and (9) as

$$\left(\mathfrak{J}_{\mathbb{W}_\alpha}^{(k+1)}(\boldsymbol{b})\right)_i \triangleq \boldsymbol{b}\cdot\boldsymbol{h}_{\alpha\beta_i} + \frac{1}{(k+1)!}\sum_\gamma \boldsymbol{w}_{\beta_i\gamma}^{(k|k)}\left(\boldsymbol{z}_{\beta\gamma}^{(k+1)}\bullet\boldsymbol{b}\right) - $$
$$ - \frac{1}{(k+1)!}\sum_\gamma \boldsymbol{w}_{\alpha\gamma}^{(k|k)}\left(\boldsymbol{z}_{\alpha\gamma}^{(k+1)}\bullet\boldsymbol{b}\right). \qquad (16)$$

Proposition 1 **(Functional Identity for Reconstruction).** *Let u be a function that is $(k+1)$-exact on $\bigcup_{\beta \in \mathbb{W}_\alpha} \mathbb{W}_\beta^{(k)}$. Then the following identity holds*

$$\mathfrak{J}_{\mathbb{W}_\alpha}^{(k+1)}\left(D^{(k+1)} u \Big|_{\boldsymbol{x}_\alpha} \right) = \\ = \left(\boldsymbol{w}_{\beta_1}^{(k|k)}\left[\overline{\mathrm{u}}\right] - \boldsymbol{w}_{\alpha}^{(k|k)}\left[\overline{\mathrm{u}}\right], \ldots, \boldsymbol{w}_{\beta_{m_\alpha}}^{(k|k)}\left[\overline{\mathrm{u}}\right] - \boldsymbol{w}_{\alpha}^{(k|k)}\left[\overline{\mathrm{u}}\right] \right) . \quad (17)$$

The *main result* of this section is

Proposition 2 **($(k+1)$-exact $(k+1)^{\text{th}}$ derivative).** *Assume that the operator $\mathfrak{J}_{\mathbb{W}_\alpha}^{(k+1)}$ defined in (16) has a left inverse $\mathfrak{D}_{\mathbb{W}_\alpha}^{(k+1)}$. Then the following expression defines a $(k+1)$-exact $(k+1)^{th}$ derivative on the neighborhood $\bigcup_{\beta \in \mathbb{W}_\alpha} \mathbb{W}_\beta^{(k)}$:*

$$\widetilde{\boldsymbol{w}}_{\alpha}^{(k+1|k+1)}\left[\overline{\mathrm{u}}\right] \triangleq \\ \triangleq \mathfrak{D}_{\mathbb{W}_\alpha}^{(k+1)}\left(\boldsymbol{w}_{\beta_1}^{(k|k)}\left[\overline{\mathrm{u}}\right] - \boldsymbol{w}_{\alpha}^{(k|k)}\left[\overline{\mathrm{u}}\right], \ldots, \boldsymbol{w}_{\beta_m}^{(k|k)}\left[\overline{\mathrm{u}}\right] - \boldsymbol{w}_{\alpha}^{(k|k)}\left[\overline{\mathrm{u}}\right] \right) \quad (18)$$

Prop. 2 gives the following algorithm.

Definition 1 **(k-exact Coupled Least Squares Algorithm (CLS)).**

1. Compute a 1-exact 1st derivative directly from the cell averages on a small stencil.
2. Iterate the following step from $m = 1$ to $m = k - 1$ at each cell:
   a. Compute a $(m+1)$-exact $(m+1)^{\text{th}}$ derivative from a m-exact m^{th} derivative, using the pseudo inverse of (15) in (18).
   b. On tetrahedral grids, apply (14) to the $(m+1)$-exact $(m+1)^{\text{th}}$ derivative.
3. Use (13) to obtain k-exact m^{th} derivatives for $1 \leq m \leq k-1$.

Remark 1. The smoothing step 2b is important on tetrahedral meshes due to stability considerations, see [5].

5 Numerical Results

As a test case, we apply the cell centered finite volume scheme (3) to the linear advection equation with constant velocity $\boldsymbol{c} = \left(\frac{1}{10}, \frac{1}{5}, 1\right)$ on the unit cube with periodic boundaries. The numerical flux is the classical upwinded flux

$$\widetilde{f}_{\alpha\beta}\left(u_\alpha, u_\beta\right) \triangleq \left(\mathbf{c} \cdot \boldsymbol{\nu}_{\alpha\beta}\right)_+ u_\alpha + \left(\mathbf{c} \cdot \boldsymbol{\nu}_{\alpha\beta}\right)_- u_\beta \, .$$

Table 1 Grid convergence: series of tetrahedral grids (CLS *with* smoothing)

h_{avg}	N	CLS D2(4)	CLS D3(6)	DLS D1(2)	DLS D2(3)	DLS D3(4)
0.042316	5928					
0.037870	8406	2.2661	4.4984	1.9190	1.5232	4.8690
0.032909	12817	2.3654	4.4386	1.9881	1.8080	4.8354
0.027354	22493	2.7521	4.5814	2.2780	2.3162	5.0287
0.022707	39518	2.7832	4.3773	2.3307	2.5409	4.8292
0.018133	77770	2.8736	4.2354	2.3583	2.7250	4.7772
0.013422	192972	2.9989	4.3158	2.3962	2.9245	4.8433

Table 2 Grid convergence: series of cartesian grids (CLS *with* smoothing)

h_{avg}	N	CLS D2(4)	CLS D3(6)	DLS D1(2)	DLS D2(3)	DLS D3(4)
0.045455	10648					
0.035714	21952	2.4243	4.4868	1.9355	1.9313	4.4325
0.029412	39304	2.6872	4.5062	2.1330	2.3892	4.4078
0.025000	64000	2.8119	4.4633	2.1923	2.6240	4.3497
0.021739	97336	2.8791	4.4349	2.2001	2.7541	4.3267
0.019231	140608	2.9175	4.3359	2.1894	2.8310	4.2191
0.017241	195112	2.9406	4.3225	2.1720	2.8779	4.2215

Table 3 Grid convergence: series of polyhedral grids (CLS *without* smoothing)

h_{avg}	N	CLS D2(2)	CLS D3(3)	DLS D1(2)	DLS D2(3)	DLS D3(4)
0.044784	13819					
0.041544	17933	3.1771	5.4782	1.3943	1.0037	4.9159
0.038507	22983	2.9794	4.9878	1.4959	1.2227	4.5057
0.033027	35595	2.4432	3.3753	1.4035	1.3847	3.9068
0.029400	52487	3.3681	4.8044	2.1503	2.2422	5.1493
0.025212	80995	2.5399	2.3076	1.6894	1.9848	4.0666
0.021547	135609	3.5112	5.4666	2.4056	2.8601	5.3057

The scheme has been tested with the CLS reconstruction of Def. 1 for $k = 2$ – named CLS D2 – and $k = 3$ – named CLS D3. The direct least squares reconstruction mentioned in Sect. 3 serves as comparison, named DLS Dk for $k = 1, 2, 3$. Tables 1, 2 and 3 display the convergence rate for the ℓ_2 error at $t = 10$ as a function of the average cell diameter h_{avg} on three different shapes of grids for the initial value $u_0(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$. The column N displays the number of cells. The number (n) indicates that the effective stencil is the n^{th} neighborhood: The first neighborhood of the cell $\mathcal{T}_\alpha$ consists of the cell $\mathcal{T}_\alpha$ itself and its adjacent cells $\mathcal{T}_\beta$. The second neighborhood of the cell $\mathcal{T}_\alpha$ is the union of the first neighbors of the first neighbors, etc.

Observe that the algorithm CLS in Def. 1 gives the desired convergence rates for quadratic (3$^{\text{rd}}$ order) and cubic reconstruction (4$^{\text{th}}$ order). The rates are comparable to those of the direct method DLS.

6 Conclusion

The algorithm of Def. 1 avoids large stencils in implementing high order finite volume schemes (3), without introducing additional degrees of freedom. The integration of the CLS algorithm in the CEDRE software is an ongoing work. This requires appropriate limiting techniques to deal with monotonicity.

References

1. Barth, T.J., Frederickson, P.O.: Higher order solution of the Euler equation on unstructured grids using quadratic reconstruction. In: AIAA 90, AIAA-90-0013, pp. 1–12. AIAA, Reno Nevada (1990)
2. Delanaye, M., Essers, J.A.: Quadratic-reconstruction finite volume scheme for compressible flows on unstructured adaptive grids. AIAA Journal **35**(4), 631 – 639 (1997)
3. Haider, F.: Discrétisation en maillage non structuré et applications les. Ph.D. thesis, Université Pierre et Marie Curie Paris VI (2009)
4. Haider, F., Brenner, P., Courbet, B., Croisille, J.P.: High order approximation on unstructured grids: Theory and implementation. Preprint (2011)
5. Haider, F., Croisille, J.P., Courbet, B.: Stability analysis of the cell centered finite-volume MUSCL method on unstructured grids. Numer. Math. **113**, 555 – 600 (2009). DOI 10.1007/s00211-009-0242-6
6. Khosla, S., Dionne, P., Lee, M., Smith, C.: Using fourth order spatial integration on unstructured meshes to reduce LES run time. AIAA 2008-782. 46th AIAA Aerospace Sciences Meeting and Exhibit, AIAA (2008)
7. van Leer, B.: Towards the ultimate conservative difference scheme. IV. A new approach to numerical convection. Journal of Computational Physics **23**(3), 276 – 299 (1977). DOI 10.1016/0021-9991(77)90095-X. URL http://www.sciencedirect.com/science/article/B6WHY-4DD1MM2-4J/2/61bfce9111ba17f514bbf0fbdb2f2ee4

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Well-Balanced Scheme For Two-Fluid Flows In Variable Cross-Section ducts

Philippe Helluy and Jonathan Jung

Abstract We propose a finite volume scheme for computing two-fluid flows in variable cross-section ducts. Our scheme satisfies a well-balanced property. It is based on the VFRoe approach. The VFRoe variables are the Riemann invariants of the stationnary wave and the cross-section. In order to avoid spurious pressure oscillations, the well-balanced approach is coupled with an ALE (Arbitrary Lagrangian Eulerian) technique at the interface and a random sampling remap.

Keywords Well-balanced scheme, Glimm scheme, Lagrange projection, two-fluid flows.
MSC2010: 65M08, 76M12, 76T10, 35Q31

Introduction

Classical finite volume solvers generally have a bad precision for solving two-fluid interfaces or flows in varying cross-section ducts. Several cures have been developed for improving the precision.

- For cross-section ducts, the well-balanced approach of Greenberg and Leroux [4] (see also [7] and [5]) is an efficient tool to improve the precision.
- For two-fluid flows the pressure oscillations phenomenon (see [6] and [2] for instance) can be cured by a recent tool developed in [3] and [1]. It is based on an ALE (Arbitrary Lagrangian Eulerian) scheme followed by a random sampling projection step.

In this paper, we show that is is possible to mix the two approaches in order to design an efficient scheme for computing two-fluid flows in variable cross-section ducts.

Philippe Helluy, Jonathan Jung
IRMA, Université de Strasbourg, 7 rue Descartes 67084 Strasbourg,
e-mail: jonathan.jung@math.unistra.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_59, © Springer-Verlag Berlin Heidelberg 2011

1 A well-balanced two-fluid ALE solver

1.1 Model

We consider the flow of a mixture of two compressible fluids (a gas (1) and a liquid (2), for instance) in a cross-section duct. The time variable is noted t and the space variable along the duct is x. We denote by $A(x)$ the cross-section at position x. The unknowns are the density $\rho(x,t)$, the velocity $u(x,t)$, the internal energy $e(x,t)$ and the fraction of gas $\varphi(x,t)$. Following Greenberg and Leroux [4] it is now classical to consider the cross-section A as an artificial unknown. The equations are the Euler equations in a duct, which read

$$\partial_t(A\rho) + \partial_x(A\rho u) = 0, \tag{1}$$

$$\partial_t(A\rho u) + \partial_x(A(\rho u^2 + p)) = p\partial_x A, \tag{2}$$

$$\partial_t(A\rho E) + \partial_x(A(\rho E + p)u) = 0, \tag{3}$$

$$\partial_t(A\rho\varphi) + \partial_x(A\rho\varphi u) = 0, \tag{4}$$

$$\partial_t A = 0, \tag{5}$$

with

$$p = p(\rho, e, \varphi), \tag{6}$$

$$E = e + \frac{u^2}{2}. \tag{7}$$

Without loss of generality, in this paper we consider a stiffened gas pressure law (see [8] and included references)

$$p(\rho, e, \varphi) = (\gamma(\varphi) - 1)\rho e - \gamma(\varphi)\pi(\varphi). \tag{8}$$

The mixture pressure law parameters $\gamma(\varphi)$ and $\pi(\varphi)$ are obtained from the pure fluid parameters $\gamma_i > 1, \pi_i, i = 1, 2$ thanks to the following interpolation, which is justified in [2]

$$\frac{1}{\gamma(\varphi) - 1} = \varphi\frac{1}{\gamma_1 - 1} + (1-\varphi)\frac{1}{\gamma_2 - 1}, \tag{9}$$

$$\frac{\gamma(\varphi)\pi(\varphi)}{\gamma(\varphi) - 1} = \varphi\,\frac{\gamma_1\pi_1}{\gamma_1 - 1} + (1-\varphi)\frac{\gamma_2\pi_2}{\gamma_2 - 1}. \tag{10}$$

We define the vector of conservative variables

$$W = (A\rho, A\rho u, A\rho E, A\rho\varphi, A)^T. \tag{11}$$

The conservative flux is

$$F(W) = (A\rho u, A(\rho u^2 + p), A(\rho E + p)u, A\rho\varphi u, 0)^T, \tag{12}$$

and the non-conservative source term is

$$S = (0, p\partial_x A, 0, 0, 0), \tag{13}$$

such that the system (1)-(5) becomes

$$\partial_t W + \partial_x F(W) = S(W). \tag{14}$$

We define the vector of primitive variables

$$Y = (\rho, u, p, \varphi, A)^T. \tag{15}$$

We define also the following quantities

$$Q = \text{mass flow rate} = \rho A u, \tag{16}$$

$$s = \text{entropy} = (p + \pi(\varphi))\rho^{-\gamma(\varphi)}, \tag{17}$$

$$h = \text{enthalpy} = e + \frac{p}{\rho}, \tag{18}$$

$$H = \text{total enthalpy} = h + \frac{u^2}{2}. \tag{19}$$

The entropy is solution of the partial differential equation

$$Tds = de - \frac{p}{\rho^2}d\rho + \lambda d\varphi. \tag{20}$$

It is useful to express also the pressure p and the enthalpy h as functions of (ρ, s, φ)

$$p = p(\rho, s, \varphi), \quad h = h(\rho, s, \varphi). \tag{21}$$

Then in these variables the sound speed c satisfies

$$c^2 = p_\rho = \rho h_\rho. \tag{22}$$

The jacobian matrix $F'(W)$ in system (14) admits real eigenvalues

$$\lambda_0 = 0, \quad \lambda_1 = u - c, \quad \lambda_2 = \lambda_3 = u, \quad \lambda_4 = u + c. \tag{23}$$

However, the system may be resonant (when $\lambda_0 = \lambda_1$ or $\lambda_0 = \lambda_4$.) The quantities φ, s, Q and H are independant Riemann invariants of the stationnary wave λ_0. In the sequel, the vector of "stationary" variables Z will play a particular role

$$Z = (A, \varphi, s, Q, H)^T. \tag{24}$$

1.2 VFRoe ALE numerical flux

We recall now the principles of the VFRoe solver. We first consider a arbitrary change of variables $U = U(W)$. In practice, we will take the set of primitive variables $U = Y$ (15) or the set of stationnary variables $U = Z$ (24). The vector U satisfies a non-conservative set of equations

$$\partial_t U + C(U)\partial_x U = 0. \tag{25}$$

The system (1)-(5) is approximated by a finite volume scheme with cells $]x_{i-1/2}, x_{i+1/2}[$, $i \in \mathbb{Z}$. We denote by τ the time step and by $\Delta x_i = x_{i+1/2} - x_{i-1/2}$ the size of cell i. We denote by W_i^n the conservative variables in cell i at time step n. The cross-section A is approximated by a piecewise constant function, $A = A_i$ in cell i.

We consider first a very general scheme where the boundary of the cell $x_{i+1/2}$ moves at the velocity $v_{i+1/2}^n$ between time steps n and $n+1$, thus we have

$$x_{i+1/2}^{n+1} = x_{i+1/2}^n + \tau v_{i+1/2}^n. \tag{26}$$

In a VFRoe-type scheme, we have to define linearized Riemann problems at interface $i + 1/2$ between the state $W_L = W_i^n$ and $W_R = W_{i+1}^n$, we introduce

$$\overline{U} = \frac{1}{2}(U_L + U_R). \tag{27}$$

In this way, it is possibe to define

$$\overline{W} = W(\overline{U}), \quad \overline{C} = C(\overline{U}). \tag{28}$$

We then consider the linearized Riemann problem

$$\partial_t U + \overline{C}\partial_x U = 0, \tag{29}$$

$$U(x,0) = \begin{cases} U_L \text{ if } x < 0, \\ U_R \text{ if } x > 0. \end{cases} \tag{30}$$

We denote its solution by

$$U(U_L, U_R, \frac{x}{t}) = U(x,t). \tag{31}$$

Because of the stationary wave, $U(U_L, U_R, \frac{x}{t})$ is generally discontinuous at $x/t = 0$. We are then able to define a discontinuous Arbitrary Lagrangian Eulerian (ALE) numerical flux

$$F(W_L, W_R, v^{\pm}) := F(W(U(U_L, U_R, v^{\pm}))) - vW(U(U_L, U_R, v^{\pm})). \tag{32}$$

The sizes of the cells evolve as

$$\Delta x_i^{n+1} = \Delta x_i^n + \tau(v_{i+1/2}^n - v_{i-1/2}^n). \tag{33}$$

If $v_{i+1/2}^n \leq 0$ and $v_{i-1/2}^n \geq 0$, the ALE scheme is

$$\begin{aligned} \Delta x_i^{n+1} W_i^{n+1,-} - \Delta x_i^n W_i^n + \\ \tau \left(F(W_i^n, W_{i+1}^n, v_{i+1/2}^{n,-}) - F(W_{i-1}^n, W_i^n, v_{i-1/2}^{n,+}) \right) = 0. \end{aligned} \tag{34}$$

If $v_{i+1/2}^n > 0$ then we have to add the following term to the left of the previous equation

$$\tau \left(F(W_i^n, W_{i+1}^n, 0^-) - F(W_i^n, W_{i+1}^n, 0^+) \right). \tag{35}$$

If $v_{i-1/2}^n < 0$ then we have to add also the following term

$$\tau \left(F(W_{i-1}^n, W_i^n, 0^-) - F(W_{i-1}^n, W_i^n, 0^+) \right). \tag{36}$$

1.3 ALE velocity

We have now to detail the choice of the variable U and the velocity v according to the data W_L and W_R. The idea is to use the classical well-balanced scheme everywhere but at the interface between the two fluids, where we use the Lagrange flux. When our initial data satisfy $\varphi \in \{0, 1\}$, the algorithm reads

- If we are not at the interface, i.e. if $\varphi_L = \varphi_R$, we take $U = Z$ and $v = 0$. This choice corresponds to the VFRoe well-balanced scheme described in [5].
- If we are at the interface, i.e. if $\varphi_L \neq \varphi_R$ then we choose $U = Y$. This choice ensures that the linearized Riemann solver presents no jump of pressure and velocity at the contact discontinuity. We thus denote by $u^*(W_L, W_R)$ and $p^*(W_L, W_R)$ the velocity and the pressure at the contact. We take $v = u^*(W_L, W_R)$, $A^* = A_L$ if $v < 0$ and $A^* = A_R$ if $v > 0$. The lagrangian numerical flux then takes the form

$$F(W_L, W_R, v^{\pm}) = (0, A^* p^*, A^* u^* p^*, 0, -A^* u^*)^T. \tag{37}$$

1.4 Glimm remap

We go back to the original Euler grid by the Glimm procedure.

We construct a sequence of pseudo-random numbers $\omega_n \in [0,1[$. In practice, we consider the $(5,3)$ van der Corput sequence [1]. According to this number we take

$$W_i^{n+1} = W_{i-1}^{n+1,-} \text{ if } \omega_n < \frac{\tau_n}{\Delta x_i} \max(v_{i-1/2}^n, 0), \tag{38}$$

$$W_i^{n+1} = W_{i+1}^{n+1,-} \text{ if } \omega_n > 1 + \frac{\tau_n}{\Delta x_i} \min(v_{i+1/2}^n, 0), \tag{39}$$

$$W_i^{n} = W_{i}^{n+1,-} \text{ if } \frac{\tau_n}{\Delta x_i} \max(v_{i-1/2}^n, 0) \le \omega_n \le 1 + \frac{\tau_n}{\Delta x_i} \min(v_{i+1/2}^n, 0). \tag{40}$$

1.5 Properties of the scheme

The constructed scheme has many interesting properties:

- it is well-balanced in the sense that it preserves exactly all stationary states (i.e. initial data for which the quantities φ, s, Q, H are constant);
- for constant cross-section ducts, it computes exactly the contact discontinuities, with no smearing of the density and the mass fraction;
- if at the initial time the mass fraction is in $\{0, 1\}$, then this property is exactly preserved at any time.

For detailed proofs, we refer to [5] and [1]. Some other subtleties are given in the same references. For instance, the change of variables $Z = Z(W)$ is not always invertible. This implies to define a special procedure for constructing completely rigorously the well-balanced VFRoe solver.

2 Numerical results

In order to test our algorithm, we consider a Riemann problem for which we know the exact solution. The initial data are discontinuous at $x = 1$. The data of the problem are given in Table 1.

The pressure law parameters are $\gamma_1 = 1.4$, $\pi_1 = 0$, $\gamma_2 = 1.6$ and $\pi_2 = 2$. We compute the solution on the domain [0.4; 1.6] with approximately 2000 cells. The final time is $T = 0.2$ and the CFL number is 0.6. The density, the velocity and the pressure are represented on Figs. 1, 2 and 3. We observe an excellent agreement between the exact and the approximate solution. The mass fraction is not represented: it is not smeared at all and perfectly matches the exact solution.

quantity	Left	Right
ρ	2	3.230672602
u	0.5	-0.4442565900
p	1	12
φ	1	0
A	1.5	1

Table 1 Numerical results. Data of the Riemann problem

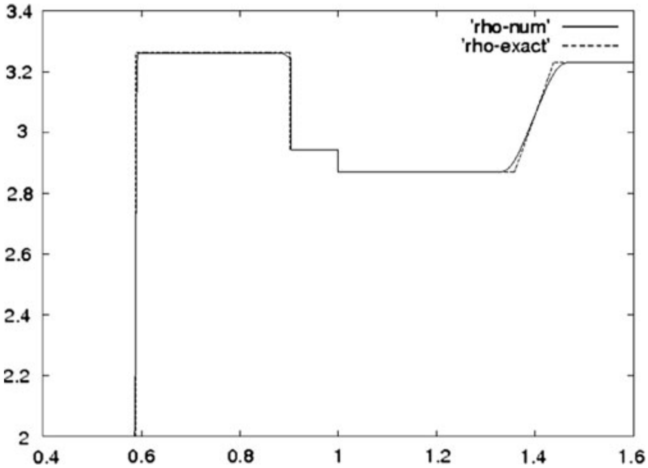


Fig. 1 Two-fluid, discontinuous cross-section Riemann problem. Density plot. Comparison of the exact solution (dotted line) and the approximate one (continuous line)

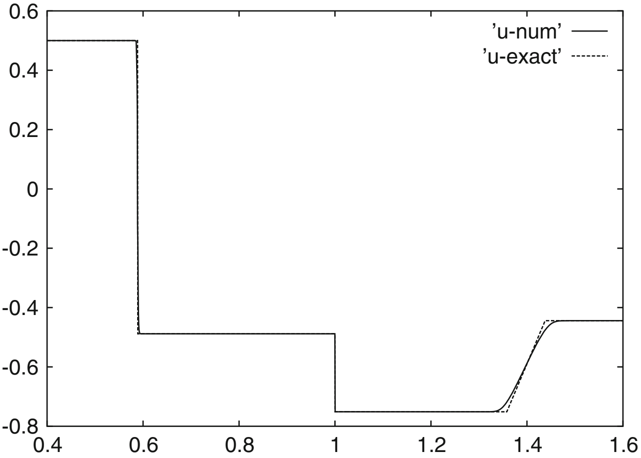


Fig. 2 Two-fluid, discontinuous cross-section Riemann problem. Pressure plot. Comparison of the exact solution (dotted line) and the approximate one (continuous line)

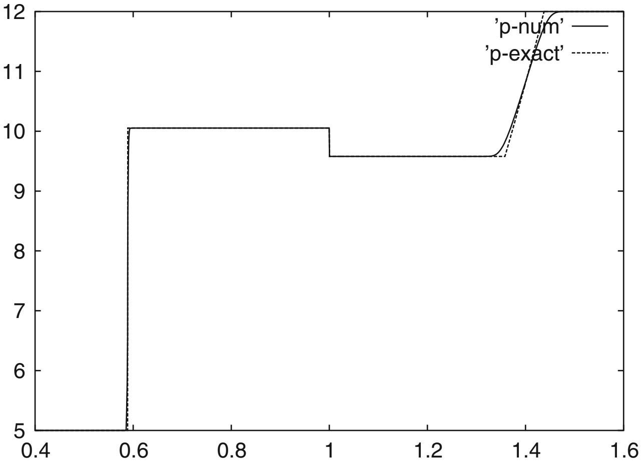


Fig. 3 Two-fluid, discontinuous cross-section Riemann problem. Velocity plot. Comparison of the exact solution (dotted line) and the approximate one (continuous line)

3 Conclusion

We have constructed and validated a new scheme for computing two-fluid flows in variable cross-section ducts. Our scheme relies on two ingredients:

- a well-balanced approach for dealing with the varying cross-section;
- a Lagrange plus remap technique in order to avoid pressure oscillations at the interface. The random sampling remap ensures that the interface is not diffused at all.

On preliminary test cases, our approach gives very satisfactory results. We intend to apply it to the computation of the oscillations of cavitation bubbles. More results will be presented at the conference.

The authors wish to thank Jean-Marc Hérard for many fruitful discussions.

References

1. M. Bachmann, P. Helluy, H. Mathis, S. Mueller. Random sampling remap for compressible two-phase flows. Preprint HAL http://hal.archives-ouvertes.fr/hal-00546919/fr/
2. T. Barberon, P. Helluy, S. Rouy. Practical computation of axisymmetrical multifluid flows. Int. J. Finite Vol. 1 (2004), no. 1, 34 pp. http://ijfv.org
3. C. Chalons, F. Coquel. Computing material fronts with a Lagrange-Projection approach. HYP2010 Proc. http://hal.archives-ouvertes.fr/hal-00548938/fr/
4. J.-M. Greenberg, A.Y., Leroux. A well balanced scheme for the numerical processing of source terms in hyperbolic equations", SIAM J. Num. Anal., vol. 33 (1), pp. 1–16, 1996.

5. P. Helluy, J.-M. Hérard, H. Mathis. A Well- Balanced Approximate Riemann Solver for Variable Cross- Section Compressible Flows. AIAA-2009-3540. 19th AIAA Computational Fluid Dynamics. June 2009.
6. S. Karni. Multicomponent flow calculations by a consistent primitive algorithm. J. Comput. Phys. 112 (1994), no. 1, 31–43.
7. D. Kroner, M.-D. Thanh. Numerical solution to compressible flows in a nozzle with variable cross-section", SIAM J. Numer. Anal., vol. 43(2), pp. 796–824, 2006.
8. R. Saurel, R. Abgrall. A simple method for compressible multifluid flows. SIAM J. Sci. Comput. 21 (1999), no. 3, 11151145.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Discretization of the viscous dissipation term with the MAC scheme

F. Babik, R. Herbin, W. Kheriji, and J.-C. Latché

Abstract We propose a discretization for the MAC scheme of the viscous dissipation term $\tau(u) : \nabla u$ (where $\tau(u)$ stands for the shear stress tensor associated to the velocity field u), which is suitable to obtain an unconditionally stable scheme for the compressible Navier-Stokes equations. It is also shown, in some model cases, to ensure the strong convergence in L^1 of the dissipation term.

Keywords compressible Navier-Stokes, MAC scheme
MSC2010: 65M12

1 Introduction

Let us consider the compressible Navier-Stokes equations, which may be written as:

$$\partial_t \rho + \mathrm{div}(\rho u) = 0, \tag{1a}$$

$$\partial_t(\rho u) + \mathrm{div}(\rho u \otimes u) + \nabla p - \mathrm{div}(\tau(u)) = 0, \tag{1b}$$

$$\partial_t(\rho e) + \mathrm{div}(\rho e u) + p\mathrm{div} u + \mathrm{div}(q) = \tau(u) : \nabla u, \tag{1c}$$

$$\rho = \wp(p, e), \tag{1d}$$

where t stands for the time, ρ, u, p and e are the density, velocity, pressure and internal energy in the flow, $\tau(u)$ stands for the shear stress tensor, q for the energy

F. Babik, W. Kheriji, and J.-C. Latché
Institut de Radioprotection et Sûreté Nucléaire (IRSN),
e-mail: [fabrice.babik,walid.kheriji,jean-claude.latche]@irsn.fr

R. Herbin
Université de Provence, e-mail: herbin@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_60, © Springer-Verlag Berlin Heidelberg 2011

diffusion flux, and the function $\wp$ is the equation of state. This system of equations is posed over $\Omega \times (0, T)$, where Ω is a domain of $\mathbb{R}^d$, $d \leq 3$. It must be supplemented by a closure relation for $\tau(u)$ and for q, assumed to be:

$$\tau(u) = \mu(\nabla u + \nabla^t u) - \frac{2\mu}{3} \operatorname{div} u\, I, \quad q = -\lambda \nabla e, \tag{2}$$

where μ and λ stand for two (possibly depending on x) positive parameters.

Let us suppose, for the sake of simplicity, that u is prescribed to zero on the whole boundary, and that the system is adiabatic, *i.e.* $q \cdot n = 0$ on $\partial\Omega$. Then, formally, taking the inner product of (1b) with u and integrating over Ω, integrating (1c) over Ω, and, finally, summing both relations yields the stability estimate:

$$\frac{d}{dt} \int_\Omega \big[\frac{1}{2} \rho\, |u|^2 + \rho e\big] \,\mathrm{d}x \leq 0. \tag{3}$$

If we suppose that the equation of state may be set under the form $p = f(\rho, e)$ with $f(\cdot, 0) = 0$ and $f(0, \cdot) = 0$, Equation (1c) implies that e remains positive (still at least formally), and so (3) yields a control on the unknown. Mimicking this computation at the discrete level necessitates to check some arguments, among them:

(i) to have available a discrete counterpart to the relation:

$$\int_\Omega \big[\partial_t(\rho u) + \operatorname{div}(\rho u \otimes u)\big] \cdot u \,\mathrm{d}x = \frac{d}{dt} \int_\Omega \frac{1}{2} \rho\, |u|^2 \,\mathrm{d}x.$$

(ii) to identify the integral of the dissipation term at the right-hand side of the discrete counterpart of (1c) with the (discrete) L^2 inner product between the velocity and the diffusion term in the discrete momentum balance equation (1b).

(iii) to be able to prove that the right-hand side of (1c) is non-negative, in order to preserve the positivity of the internal energy.

The point (i) is extensively discussed in [5] (see also [6]), and is not treated here. Indeed, we focus here on a discretization technique which allows to obtain *(ii)* and *(iii)* with the usual Marker and Cell (MAC) discretization [3, 4], and which is implemented in the ISIS free software developed at IRSN [8] on the basis of the software component library PELICANS [10]. We complete the presentation by showing how (ii) may also be used, in some model problems, to prove the convergence in L^1 of the dissipation term.

2 Discretization of the dissipation term

2.1 The two-dimensional case

Let us begin with a two-dimensional case. The first step is to propose a discretization for the diffusion term in the momentum equation. We begin with the x-component of the velocity, for which we write a balance equation on $K^x_{i-\frac{1}{2},j} = (x_{i-1}, x_i) \times (y_{j-\frac{1}{2}}, y_{j+\frac{1}{2}})$ (see Figs. 1 and 2 for the notations). Integrating the x component of the momentum balance equation over $K^x_{i-\frac{1}{2},j}$, we get for the diffusion term:

$$\bar{T}^{\rm dif}_{i-\frac{1}{2},j} = -\Big[\int_{K^x_{i-\frac{1}{2},j}} {\rm div}\big[\tau(u)\big]\,{\rm d}x\Big] \cdot \boldsymbol{e}^{(x)} = -\Big[\int_{\partial K^x_{i-\frac{1}{2},j}} \tau(u)\,n\,{\rm d}\gamma\Big] \cdot \boldsymbol{e}^{(x)}, \qquad (4)$$

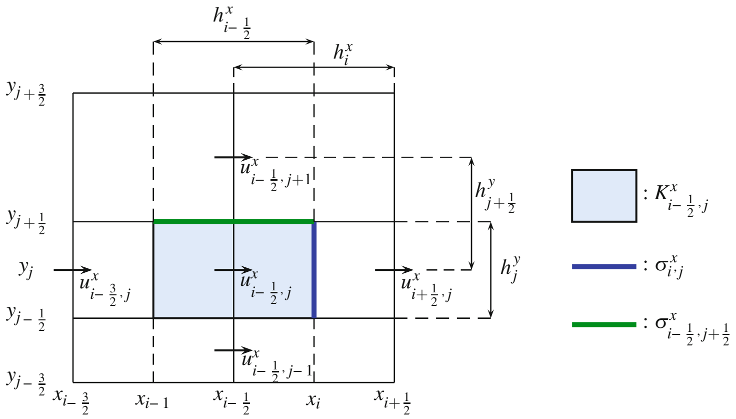


Fig. 1 Dual cell for the x-component of the velocity

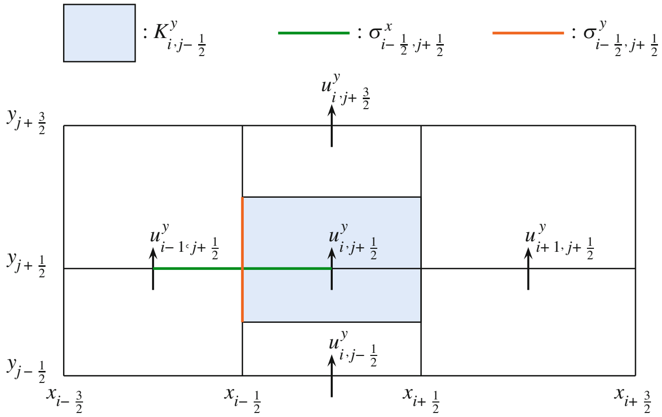


Fig. 2 Dual cell for the y-component of the velocity

where $\boldsymbol{e}^{(x)}$ stands for the first vector of the canonical basis of $\mathbb{R}^2$. We denote by $\sigma^x_{i,j}$ the right face of $K^x_{i-\frac{1}{2},j}$, *i.e.* $\sigma^x_{i,j} = \{x_i\} \times (y_{j-\frac{1}{2}},\, y_{j+\frac{1}{2}})$. Splitting the boundary integral in (4), the part of $\bar{T}^{\text{dif}}_{i-\frac{1}{2},j}$ associated to $\sigma^x_{i,j}$, also referred to as the viscous flux through $\sigma^x_{i,j}$, reads:

$$-\Big[\int_{\sigma^x_{i,j}} \tau(u)\, n \,\mathrm{d}\gamma\Big] \cdot \boldsymbol{e}^{(x)} = -2\int_{\sigma^x_{i,j}} \mu\ \partial_x u^x \,\mathrm{d}\gamma + \frac{2}{3}\int_{\sigma^x_{i,j}} \mu\ (\partial_x u^x + \partial_y u^y)\,\mathrm{d}\gamma,$$

and the usual finite difference technique yields the following approximation for this term:

$$\begin{aligned} &-\frac{4}{3}\int_{\sigma^x_{i,j}} \mu\ \partial_x u^x \,\mathrm{d}\gamma + \frac{2}{3}\int_{\sigma^x_{i,j}} \mu\ \partial_y u^y \,\mathrm{d}\gamma \\ &\qquad \approx -\frac{4}{3}\,\mu_{i,j}\ \frac{h^y_j}{h^x_i}\ (u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}) + \frac{2}{3}\,\mu_{i,j}\ \frac{h^y_j}{h^y_j}\ (u^y_{i,j+\frac{1}{2}} - u^y_{i,j-\frac{1}{2}}), \end{aligned} \tag{5}$$

where $\mu_{i,j}$ is an approximation of the viscosity at the face $\sigma^x_{i,j}$. Similarly, let $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}} = (x_{i-1},\, x_i) \times \{y_{j+\frac{1}{2}}\}$ be the top edge of the cell. Then:

$$\begin{aligned} &-\Big[\int_{\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}} \tau(u)\, n\,\mathrm{d}\gamma\Big] \cdot \boldsymbol{e}^{(x)} = -\int_{\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}} \mu\ (\partial_y u^x + \partial_x u^y)\,\mathrm{d}\gamma \\ &\quad \approx -\mu_{i-\frac{1}{2},j+\frac{1}{2}} \Big[\frac{h^x_{i-\frac{1}{2}}}{h^y_{j+\frac{1}{2}}}\ (u^x_{i-\frac{1}{2},j+1} - u^x_{i-\frac{1}{2},j}) + \frac{h^x_{i-\frac{1}{2}}}{h^x_{i-\frac{1}{2}}}\ (u^y_{i,j+\frac{1}{2}} - u^y_{i-1,j+\frac{1}{2}})\Big], \end{aligned}$$

where $\mu_{i-\frac{1}{2},j+\frac{1}{2}}$ stands for an approximation of the viscosity at the edge $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$.

Let us now multiply each discrete equation for u^x by the corresponding degree of freedom of a velocity field v (*i.e.* the balance over $K^x_{i-\frac{1}{2},j}$ by $v^x_{i-\frac{1}{2},j}$) and sum over i and j. The viscous flux at the face $\sigma^x_{i,j}$ appears twice in the sum, once multiplied by $v^x_{i-\frac{1}{2},j}$ and the second one by $-v^x_{i+\frac{1}{2},j}$, and the corresponding term reads:

$$\begin{aligned} &T^{\text{dis}}_{i,j}(u,v) = \\ &\mu_{i,j}\Big[-\frac{4}{3}\,\frac{h^y_j}{h^x_i}\ (u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}) + \frac{2}{3}\,\frac{h^y_j}{h^y_j}\ (u^y_{i,j+\frac{1}{2}} - u^y_{i,j-\frac{1}{2}})\Big]\ (v^x_{i-\frac{1}{2},j} - v^x_{i+\frac{1}{2},j}) \\ &\quad = \mu_{i,j}\ h^y_j h^x_i\ \Big[\frac{4}{3}\ \frac{u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}}{h^x_i} - \frac{2}{3}\ \frac{u^y_{i,j+\frac{1}{2}} - u^y_{i,j-\frac{1}{2}}}{h^y_j}\Big]\ \frac{v^x_{i+\frac{1}{2},j} - v^x_{i-\frac{1}{2},j}}{h^x_i}. \end{aligned} \tag{6}$$

Similarly, the term associated to $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$ appears multiplied by $v^x_{i-\frac{1}{2},j}$ and by $-v^x_{i-\frac{1}{2},j+1}$, and we get:

$$T^{\mathrm{dis}}_{i-\frac{1}{2},j+\frac{1}{2}}(u,v) = \mu_{i-\frac{1}{2},j+\frac{1}{2}}\, h^x_{i-\frac{1}{2}} h^y_{j+\frac{1}{2}} \\ \Bigl[\frac{u^x_{i-\frac{1}{2},j+1} - u^x_{i-\frac{1}{2},j}}{h^y_{j+\frac{1}{2}}} + \frac{u^y_{i,j+\frac{1}{2}} - u^y_{i-1,j+\frac{1}{2}}}{h^x_{i-\frac{1}{2}}}\Bigr] \frac{v^x_{i-\frac{1}{2},j+1} - v^x_{i-\frac{1}{2},j}}{h^y_{j+\frac{1}{2}}}. \quad (7)$$

Let us now define the discrete gradient of the velocity as follows:

- The derivatives involved in the divergence, $\eth^{\mathcal{M}}_x u^x$ and $\eth^{\mathcal{M}}_y u^y$, are defined over the primal cells by:

$$\eth^{\mathcal{M}}_x u^x(x) = \frac{u^x_{i+\frac{1}{2},j} - u^x_{i-\frac{1}{2},j}}{h^x_i}, \quad \eth^{\mathcal{M}}_y u^y(x) = \frac{u^y_{i,j+\frac{1}{2}} - u^y_{i,j-\frac{1}{2}}}{h^y_j}, \quad \forall x \in K_{i,j}. \quad (8)$$

- For the other derivatives, we introduce another mesh which is vertex-centred, and we denote by K^{xy} the generic cell of this new mesh, with $K^{xy}_{i+\frac{1}{2},j+\frac{1}{2}} = (x_i, x_{i+1}) \times (y_j, y_{j+1})$. Then, $\forall x \in K^{xy}_{i+\frac{1}{2},j+\frac{1}{2}}$:

$$\eth^{\mathcal{M}}_y u^x(x) = \frac{u^x_{i+\frac{1}{2},j+1} - u^x_{i+\frac{1}{2},j}}{h^y_{j+\frac{1}{2}}}, \quad \eth^{\mathcal{M}}_x u^y(x) = \frac{u^y_{i+1,j+\frac{1}{2}} - u^y_{i,j+\frac{1}{2}}}{h^x_{i+\frac{1}{2}}}. \quad (9)$$

With this definition, we get:

$$T^{\mathrm{dis}}_{i,j}(u,v) = \mu_{i,j} \int_{K_{i,j}} \Bigl[\frac{4}{3}\eth^{\mathcal{M}}_x u^x - \frac{2}{3}\eth^{\mathcal{M}}_y u^y\Bigr]\eth^{\mathcal{M}}_x v^x \,\mathrm{d}x,$$

and:

$$T^{\mathrm{dis}}_{i-\frac{1}{2},j+\frac{1}{2}}(u,v) = \mu_{i-\frac{1}{2},j+\frac{1}{2}} \int_{K^{xy}_{i-\frac{1}{2},j+\frac{1}{2}}} (\eth^{\mathcal{M}}_y u^x + \eth^{\mathcal{M}}_x u^y)\, \eth^{\mathcal{M}}_y v^x \,\mathrm{d}x.$$

Let us now perform the same operations for the y-component of the velocity. Doing so, we are lead to introduce an approximation of the viscosity at the edge $\sigma^y_{i-\frac{1}{2},j+\frac{1}{2}} = \{x_{i-\frac{1}{2}}\} \times (y_j,\ y_{j+1})$ (see Fig. 2). Let us suppose that we take the same approximation as on $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$. Then, the same argument yields that multiplying each discrete equation for u^x and for u^y by the corresponding degree of freedom of a velocity field v, we obtain a dissipation term which reads:

$$T^{\mathrm{dis}}(u,v) = \int_{\Omega} \tau^{\mathcal{M}}(u) : \nabla^{\mathcal{M}} v \,\mathrm{d}x, \quad (10)$$

where $\nabla^{\mathcal{M}}$ is the discrete gradient defined by (8)-(9) and $\tau^{\mathcal{M}}$ the discrete tensor:

$$\tau^{\mathcal{M}}(u) = \\ \begin{bmatrix} 2\mu\, \partial_x^{\mathcal{M}} u_x & \mu^{xy}\, (\partial_y^{\mathcal{M}} u_x + \partial_x^{\mathcal{M}} u_y) \\ \mu^{xy}\, (\partial_y^{\mathcal{M}} u_x + \partial_x^{\mathcal{M}} u_y) & 2\mu\, \partial_y^{\mathcal{M}} u_y \end{bmatrix} - \frac{2}{3}\, \mu\, (\partial_x^{\mathcal{M}} u_x + \partial_y^{\mathcal{M}} u_y)\, I, \quad (11)$$

where μ is the viscosity defined on the primal mesh by $\mu(x) = \mu_{i,j}$, $\forall x \in K_{i,j}$ and μ^{xy} is the viscosity defined on the vertex-centred mesh, by $\mu(x) = \mu_{i+\frac{1}{2},j+\frac{1}{2}}$, $\forall x \in K^{xy}_{i+\frac{1}{2},j+\frac{1}{2}}$.

Now the form (10) suggests a natural to discretize the viscous dissipation term in the internal energy balance in order for the consistency property (ii) to hold. Indeed, if we simply set on each primal cell $K_{i,j}$:

$$(\tau(u) : \nabla u)_{i,j} = \frac{1}{|K_{i,j}|} \int_{K_{i,j}} \tau^{\mathcal{M}}(u) : \nabla^{\mathcal{M}} u \, \mathrm{d}x, \quad (12)$$

then, thanks to (10), the property *(ii)* which reads:

$$T^{\mathrm{dis}}(u,u) = \sum_{i,j} |K_{i,j}|\, (\tau(u) : \nabla u)_{i,j}.$$

holds. Furthermore, we get from Definition (11) that $\tau^{\mathcal{M}}(u)(x)$ is a symmetrical tensor, for any i, j and $x \in K_{i,j}$, and therefore an elementary algebraic argument yields:

$$\begin{aligned}(\tau(u) : \nabla u)_{i,j} &= \frac{1}{|K_{i,j}|} \int_{K_{i,j}} \tau^{\mathcal{M}}(u) : \nabla^{\mathcal{M}} u \, \mathrm{d}x \\ &= \frac{1}{2\,|K_{i,j}|} \int_{K_{i,j}} \tau^{\mathcal{M}}(u) : \left[\nabla^{\mathcal{M}} u + (\nabla^{\mathcal{M}} u)^t\right] \mathrm{d}x \geq 0.\end{aligned}$$

Remark 1 (Approximation of the viscosity). Note that, for the symmetry of $\tau^{\mathcal{M}}(u)$ to hold, the choice of the same viscosity at the edges $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$ and $\sigma^y_{i-\frac{1}{2},j+\frac{1}{2}}$ is crucial even though other choices may appear natural. Assuming for instance the viscosity to be a function of an additional variable defined on the primal mesh, the following construction seems reasonable:

1. define a constant value for μ on each primal cell,
2. associate a value of μ to the primal edges, by taking the average between the value at the adjacent cells,

3. finally, split the integral of the shear stress over $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$ in two parts, one for the part included in the (top) boundary of $K_{i-1,j}$ and the second one in the boundary of $K_{i,j}$.

Then the viscosities on $\sigma^x_{i-\frac{1}{2},j+\frac{1}{2}}$ and $\sigma^y_{i-\frac{1}{2},j+\frac{1}{2}}$ coincide only for uniform meshes, and, in the general case, the symmetry of $\tau^{\mathcal{M}}(u)$ is lost.

2.2 Extension to the three-dimensional case

Extending the computations of the preceding section to three space dimensions yields the following construction.

– First, define three new meshes, which are "edge-centred": $K^{xy}_{i+\frac{1}{2},j+\frac{1}{2},k} = (x_i, x_{i+1}) \times (y_i, y_{j+1}) \times (z_{k-\frac{1}{2}}, z_{k+\frac{1}{2}})$ is staggered from the primal mesh $K_{i,j,k}$ in the x and y direction, $K^{xz}_{i+\frac{1}{2},j,k+\frac{1}{2}}$ in the x and z direction, and $K^{yz}_{i,j+\frac{1}{2},k+\frac{1}{2}}$ in the y and z direction.
– The partial derivatives of the velocity components are then defined as piecewise constant functions, the value of which is obtained by natural finite differences:
  - for $\partial^{\mathcal{M}}_x u^x$, $\partial^{\mathcal{M}}_y u^y$ and $\partial^{\mathcal{M}}_z u^z$, on the primal mesh,
  - for $\partial^{\mathcal{M}}_y u^x$ and $\partial^{\mathcal{M}}_x u^y$ on the cells $(K^{xy}_{i+\frac{1}{2},j+\frac{1}{2},k})$,
  - for $\partial^{\mathcal{M}}_z u^x$ and $\partial^{\mathcal{M}}_x u^z$ on the cells $(K^{xz}_{i+\frac{1}{2},j,k+\frac{1}{2}})$,
  - for $\partial^{\mathcal{M}}_y u^z$ and $\partial^{\mathcal{M}}_z u^y$ on the cells $(K^{yz}_{i,j+\frac{1}{2},k+\frac{1}{2}})$.
– Then, define four families of values for the viscosity field, μ, μ^{xy}, μ^{xz} and μ^{yz}, associated to the primal and the three edge-centred meshes respectively.
– The shear stress tensor is obtained by the extension of (11) to $d = 3$.
– And, finally, the dissipation term is given by (12).

3 A strong convergence result

We conclude this paper by showing how the consistency property *(ii)* may be used, in some particular cases, to obtain the strong convergence of the dissipation term, and then pass to the limit in a coupled equation having the dissipation term as right-hand side. To this purpose, let us just address the model problem:

$$-\Delta \underline{u} = \underline{f} \text{ in } \Omega = (0,1) \times (0,1), \qquad \underline{u} = 0 \text{ on } \partial\Omega, \tag{13}$$

with $\underline{u}$ and $\underline{f}$ two scalar functions, $\underline{f} \in \mathrm{L}^2(\Omega)$. Let us suppose that this problem is discretized by the usual finite volume technique, with the uniform MAC mesh associated to the x-component of the velocity. We define a discrete function as a piecewise constant function, vanishing on the left and right sides of the domain (so on the left and right stripes of staggered (half-)meshes adjacent to these boundaries), and we define the discrete H^1-norm of a discrete function v by:

$$\|v\|_1^2 = \int_\Omega (\partial_x^{\mathcal{M}} v)^2 + (\partial_y^{\mathcal{M}} v)^2 \,\mathrm{d}x.$$

Let $(\mathcal{M}^{(n)})_{n\in\mathbb{N}}$ be a sequence of such meshes, with a step h^n tending to zero, and let $(u^{(n)})_{n\in\mathbb{N}}$ be the corresponding sequence of discrete solutions. Then, with the variational technique employed in the preceding section, we get, with the usual discretization of the right-hand side:

$$\|u^{(n)}\|_1^2 = \int_\Omega (\partial_x^{\mathcal{M}} u^{(n)})^2 + (\partial_y^{\mathcal{M}} u^{(n)})^2 \,\mathrm{d}x = \int_\Omega \underline{f}\, u^{(n)} \,\mathrm{d}x. \tag{14}$$

Since the discrete H^1-norm controls the L^2-norm (*i.e.* a discrete Poincaré inequality holds [2]), this yields a uniform bound for the sequence $(u^{(n)})_{n\in\mathbb{N}}$ in discrete H^1-norm. Hence the sequence $(u^{(n)})_{n\in\mathbb{N}}$ converges in $\mathrm{L}^2(\Omega)$ to a function $\bar{u} \in \mathrm{H}_0^1(\Omega)$, possibly up to the extraction of a subsequence [2], and he discrete derivatives $(\partial_x^{\mathcal{M}} u^{(n)})_{n\in\mathbb{N}}$ and $(\partial_y^{\mathcal{M}} u^{(n)})_{n\in\mathbb{N}}$ weakly converge in $\mathrm{L}^2(\Omega)$ to $\partial_x \bar{u}$ and $\partial_y \bar{u}$ respectively. This allows to pass to the limit in the scheme, and we obtain that $\bar{u}$ satisfies the continuous equation (13). Thus, taking $\bar{u}$ as a test function in the variational form of (13):

$$\int_\Omega (\partial_x \bar{u})^2 + (\partial_y \bar{u})^2 \,\mathrm{d}x = \int_\Omega \underline{f}\, \bar{u} \,\mathrm{d}x.$$

But, passing to the limit in (14), we get:

$$\lim_{n\mapsto\infty} \int_\Omega (\partial_x^{\mathcal{M}} u^{(n)})^2 + (\partial_y^{\mathcal{M}} u^{(n)})^2 \,\mathrm{d}x = \lim_{n\mapsto\infty} \int_\Omega \underline{f}\, u^{(n)} \,\mathrm{d}x = \int_\Omega \underline{f}\, \bar{u} \,\mathrm{d}x,$$

which, comparing to the preceding relation, yields:

$$\lim_{n\to\infty} \int_\Omega (\partial_x^{\mathcal{M}} u^{(n)})^2 + (\partial_y^{\mathcal{M}} u^{(n)})^2 \,\mathrm{d}x = \int_\Omega (\partial_x \bar{u})^2 + (\partial_y \bar{u})^2 \,\mathrm{d}x.$$

Since the discrete gradient weakly converges and its norm converges to the norm of the limit, the discrete gradient strongly converges in $\mathrm{L}^2(\Omega)^2$ to the gradient of the solution. Let us now imagine that Equation (13) is coupled to a balance equation for another variable, the right-hand side of which is $|\nabla \underline{u}|^2$; this situation occurs for instance in models involving ohmic losses [1], or RANS turbulence models [9]. The discretization (12) of the dissipation term in the cell K, which reads here:

$$\big(|\nabla u^{(n)}|^2\big)_K = \frac{1}{|K|} \int_K (\partial_x^{\mathcal{M}} u^{(n)})^2 + (\partial_y^{\mathcal{M}} u^{(n)})^2 \,\mathrm{d}x,$$

yields a convergent right-hand side, in the sense that, for any regular function $\varphi \in C_c^\infty(\Omega)$, we have:

$$\lim_{n\to\infty} \sum_K \int_K \big(|\nabla u^{(n)}|^2\big)_K \varphi \,\mathrm{d}x = \int_\Omega |\nabla \underline{u}|^2 \varphi \,\mathrm{d}x.$$

(A declination of) this argument has been used to prove the convergence of numerical schemes in [1, 9].

References

1. A. Bradji, R. Herbin: Discretization of the coupled heat and electrical diffusion problems by the finite element and the finite volume methods. IMA Journal of Numerical Analysis, **28**, 469–495 (2008).
2. R. Eymard, T. Gallouët, R. Herbin, Finite Volume Methods. Handbook of Numerical Analysis, Volume VII, 713–1020, North Holland (2000).
3. F.H. Harlow, J.E. Welsh: Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. Physics of Fluids, **8**, 2182–2189 (1965).
4. F.H. Harlow, A.A. Amsden: A numerical fluid dynamics calculation method for all flow speeds. Journal of Computational Physics, **8**, 197–213 (1971).
5. L. Gastaldo, R. Herbin, W. Kheriji, C. Lapuerta, J.-C. Latché: Staggered discretizations, pressure correction schemes and all speed barotropic flows. Finite Volumes for Complex Applications VI (FVCA VI), Prague, Czech Republic, June 2011.
6. R. Herbin, J.-C. Latché: A kinetic energy control in the MAC discretization of compressible Navier-Stokes equations. International Journal of Finite Volumes **2** (2010).
7. R. Herbin, W. Kheriji, J.-C. Latché: An unconditionally stable Finite Element-Finite Volume pressure correction scheme for compressible Navier-Stokes equations. In preparation (2011).
8. ISIS: a CFD computer code for the simulation of reactive turbulent flows, `https://gforge.irsn.fr/gf/project/isis`.
9. A. Larcher, J.-C. Latché: Convergence analysis of a finite element-finite volume scheme for a RANS turbulence model. Submitted (2011).
10. PELICANS: Collaborative Development Environment. `https://gforge.irsn.fr/gf/project/pelicans`.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Sharp Contact Discontinuity Scheme for Multimaterial Models

Angelo Iollo, Thomas Milcent, and Haysam Telib

Abstract We present a method to capture the evolution of a contact discontinuity separating two different materials. This method builds on the ghost-fluid idea: a locally non-conservative scheme allows an accurate and stable simulation of problems involving non-miscible media that have significantly different physical properties. Compared to the ghost-fluid approach, the main difference is that with the present method no ghost fluid is introduced. Numerical illustrations involving one-dimensional interfaces show that with this scheme the contact discontinuity stays sharp and oscillation free.

Keywords sharp contact discontinuity, multimaterial, compressible elasticity
MSC2010: 65N08, 65Z05,74S10,76M12

1 Introduction

Physical and engineering problems that involve several materials are ubiquitous in nature and in applications: multi-phase flows, fluid-structure interaction, particle flows, to cite just a few examples. The main contributions in the direction of simulating these phenomena go back to [7] and [8] for the model and to [1] for a consistent and stable discretization. The idea is to model the eulerian stress tensor through a constitutive law reproducing the mechanical characteristics of the

Angelo Iollo, Thomas Milcent
Institut de Mathématiques de Bordeaux UMR 5251 Université Bordeaux 1 and INRIA Bordeaux-Sud Ouest, équipe-projet MC2, 351, Cours de la Libération, 33405 Talence, France, e-mail: angelo.iollo@math.u-bordeaux1.fr, thomas.milcent@math.u-bordeaux1.fr

Haysam Telib
Dipartimento di Ingegneria Aeronautica e Spaziale, Politecnico di Torino, C.so Duca degli Abruzzi 24, 10129 Torino, Italy, e-mail: haysam.telib@polito.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_61, © Springer-Verlag Berlin Heidelberg 2011

medium under consideration. Hence, for example, an elastic material or a gas will be modeled by the same set of equations except for the constitutive law relating the deformation and the stress tensor. The system of conservation laws thus obtained can be cast in the framework of quasi-linear hyperbolic partial differential equations (PDEs). From the numerical view point this is convenient since classical integration schemes can be employed in each material. However, it turns out that the evolution of the interface, which is represented in this model by a contact discontinuity, is particularly delicate because standard Godunov schemes fail. In [1] it was shown that a simple and effective remedy to this problem is the definition of a ghost fluid across the interface. A remarkable application based on this approach is presented in [6]. This method, however, has the disadvantage that the interface is diffused over a certain number of grid points. From a practical view point this can be a serious drawback if one is interested in the geometric properties of the evolving interface, as, for example, in the case of surface tension or when the interface itself is elastic. In this paper we propose a simple first-order accurate method to recover a sharp interface description keeping the solution stable and non-oscillating.

2 The model

This approach was discussed in [4, 6, 7], and [8]. We develop here the principal elements of the formulation. The starting point is classical continuum mechanics. Let Ω_0 be the reference or initial configuration of a single material and Ω_t the deformed configuration at time t. We define $X(\xi,t)$ as the image at time t of a material point ξ belonging to the initial configuration, in the deformed configuration, i.e., $X : \Omega_0 \times [0,T] \longrightarrow \Omega_t$, $(\xi,t) \mapsto X(\xi,t)$, and the corresponding velocity field u as $u : \Omega_t \times [0,T] \longrightarrow \mathbb{R}^3$, $(x,t) \mapsto u(x,t)$ where $X_t(\xi,t) = u(X(\xi,t),t)$ completed by the initial condition $X(\xi,0) = \xi$. Also we introduce the backward characteristics $Y(x,t)$ that for a time t and a point x in the deformed configuration, gives the corresponding initial point ξ in the initial configuration, i.e., $Y : \Omega_t \times [0,T] \longrightarrow \Omega_0$, $(x,t) \mapsto Y(x,t)$ with the initial condition $Y(x,0) = x$. Of course, we have $[\nabla_\xi X(\xi,t)] = [\nabla_x Y(x,t)]^{-1}$ and $Y_t + (u \cdot \nabla)Y = 0$.

In elasticity, the internal energy is a function of the strain tensor $\nabla_\xi X$ and the entropy s: $W = W(\nabla_\xi X, s)$. The potential W has to be Galilean invariant and, eventually, isotropic. It can be proven that (Rivlin-Eriksen theorem [3]) this is the case if, and only if, $\mathscr{E}$, the energy, is expressed as a function of $s(\xi,t)$, the entropy, and of the invariants of $C(\xi,t) = [\nabla_\xi X]^T[\nabla_\xi X]$, the right Cauchy-Green tensor. The invariants often considered in the literature are $J(\xi,t) = \det([\nabla_\xi X])$, $\mathrm{Tr}(C(\xi,t))$ and $\mathrm{Tr}(\mathrm{Cof}(C(\xi,t)))$.We assume that the internal energy per unit volume is the sum of a term depending on volume variation and entropy, and a term accounting for isochoric deformation. In general the term relative to an isochoric transformation will also depend on entropy. Here, we will limit the discussion to materials where shear forces are conservative. The governing equations derived from the above formulation in the deformed configuration are:

$$\begin{cases} \rho_t + \operatorname{div}_x(\rho u) = 0 \\ \rho(u_t + (u \cdot \nabla_x)u) - \operatorname{div}_x \sigma = 0 \\ \rho(\varepsilon_t + (u \cdot \nabla_x)\varepsilon) - \sigma : \nabla_x u = 0 \\ Y_t + (u \cdot \nabla_x)Y = 0 \end{cases} \tag{1}$$

where $\sigma(x,t)$ is the Cauchy stress tensor in the physical domain. The unknowns are the backward characteristics of the problem $Y(x,t)$, the velocity $u(x,t)$, the internal energy per unit mass $\varepsilon(x,t) = W/\rho_0$ and the density $\rho(x,t)$. The initial velocity $u(x,0)$, the initial internal energy $\varepsilon(x,0)$ and $Y(x,0) = x$ are given. If the initial density $\rho_0(\xi)$ is known, the equation of mass conservation is actually redundant because $\rho(x,t) = \det(\nabla_x Y(x,t))\rho_0(Y(x,t))$.

To close the system, a constitutive law which connects σ to Y is necessary. In the deformed domain energy can be written

$$\mathscr{E} = \int_{\Omega_t} \Big(W_{\text{vol}}(J,s) + W_{\text{iso}}(\operatorname{Tr}(\overline{B}), \operatorname{Tr}(\operatorname{Cof}(\overline{B}))\Big) J^{-1} dx \tag{2}$$

where

$$\overline{B}(x,t) = \frac{B(x,t)}{\det(B(x,t))^{\frac{1}{3}}} \qquad\qquad B(x,t) = [\nabla_x Y(x,t)]^{-1}[\nabla_x Y(x,t)]^{-T} \tag{3}$$

with $B(x,t)$ the left Cauchy-Green tensor and

$$J(x,t) = \det([\nabla_x Y(x,t)])^{-1} = \det(B(x,t))^{\frac{1}{2}}. \tag{4}$$

It can be shown that

$$\sigma(x,t) = W'_{\text{vol}}(J,s)I + 2J^{-1}\left(\overline{\sigma}_{\text{iso}} - \frac{1}{3}I(\overline{\sigma}_{\text{iso}} : I)\right) \tag{5}$$

with

$$\overline{\sigma}_{\text{iso}} = \frac{\partial W_{\text{iso}}}{\partial a}\overline{B} - \frac{\partial W_{\text{iso}}}{\partial b}\overline{B}^{-1}. \tag{6}$$

By definition pressure is given by $p = -\dfrac{1}{3}\operatorname{Tr}(\sigma) = -W'_{\text{vol}}(J,s)$.

For the objectives of this paper, we restrict our investigation to an elastic one-dimensional isoentropic case with non-zero transverse velocity. Let x_i, $i = 1\ldots 2$ be the coordinates in the canonical basis of $\mathbb{R}^2$, u_i the velocity components, Y^i the components of Y and σ^{ij} the components of the stress tensor. Also, let us denote by $,i$ differentiation with respect to x_i. We consider the governing equations in two space dimensions and we assume that ∇Y is a function only of one direction (x_1), as well as u_1 and u_2. In this case we have that $(Y^1_{,2})_t = (Y^2_{,2})_t = 0$. Since $Y(x,0) = x$, we have also that $Y^1_{,2} = 0$, $Y^2_{,2} = 1$ and hence

$$[\nabla Y] = \begin{pmatrix} Y^1_{,1} & 0 \\ Y^2_{,1} & 1 \end{pmatrix}. \tag{7}$$

The governing equations in conservative form become

$$\Psi_t + (F(\Psi))_{,1} = 0$$

with

$$\Psi = \begin{pmatrix} \rho \\ \phi_1 \\ \phi_2 \\ Y^1_{,1} \\ Y^2_{,1} \end{pmatrix} \qquad F(\Psi) = \begin{pmatrix} \phi_1 \\ \frac{(\phi_1)^2}{\rho} - \sigma^{11} \\ \frac{\phi_1\phi_2}{\rho} - \sigma^{21} \\ \frac{\phi_1 Y^1_{,1}}{\rho} \\ \frac{\phi_1 Y^2_{,1} + \phi_2}{\rho} \end{pmatrix}$$

In two dimensions we define $\overline{B} = \dfrac{B}{\det(B)^{\frac{1}{2}}}$, so that $\det(\overline{B}) = 1$. Let now

$$\varepsilon = \frac{W_{vol} + W_{iso}}{\rho_0} = \frac{\exp\left(\dfrac{s - s_0}{c_v}\right)\rho^{\gamma-1}}{\gamma - 1} + \frac{p_\infty}{\rho} + \chi(\mathrm{Tr}\,(\overline{B}) - 2) \tag{8}$$

where s_0 is the reference entropy, $\phi_i = \rho u_i$ and γ, $p_\infty, \chi \in \mathbb{R}^+$ are constants that characterize a given material. This model accounts for elastic deformations in the transverse direction, i.e., $\sigma^{21} \neq 0$. Here the term W_{vol} represents a stiff gas, the term W_{iso} a Mooney-Rivlin solid.

3 Multimaterial solver

We assume that the initial condition at time t_n, the n-th time step, is known. Let $\Psi^n_k = \Psi(x_k, t_n)$, with x_k the spatial coordinate x of grid point k. The discretization points are $N + 1$ and let $I = \{1, \cdots, N\}$. Consider two non-miscible materials separated by a physical interface located, at time t_n, in x^n_f and let $\iota = i$ such that $x_i \leq x_f < x_{i+1}$, $i \in I$. The space and time discretization $\forall k \in I$ and $k \neq \iota, \iota + 1$ is as follows

$$\frac{\Psi^{n+1}_k - \Psi^n_k}{\Delta t} = -\frac{\mathscr{F}^n_{k+1/2}(\Psi^n_k, \Psi^n_{k+1}) - \mathscr{F}^n_{k-1/2}(\Psi^n_{k-1}, \Psi^n_k)}{\Delta x} \tag{9}$$

where $\Delta t = t^{n+1} - t^n$, $\Delta x = x_{k+1/2} - x_{k-1/2}$ and $\mathscr{F}^n_{k\pm 1/2}$ are the numerical fluxes evaluated at the cell interface located at $x_{k\pm 1/2}$. For consistency $\mathscr{F}$ is a regular enough function of both arguments and $\mathscr{F}(\Psi, \Psi) = F(\Psi)$. Numerical conservation requires that $\mathscr{F}(\Psi', \Psi) = \mathscr{F}(\Psi, \Psi')$. The numerical flux function $\mathscr{F}_{k+1/2}(\Psi^n_k, \Psi^n_{k+1})$ is computed by an approximate Riemann solver. In the following we use the HLLC [9] approximate solvers.

In any case, we assume that Riemann solver employed defines at least two intermediate states Ψ^n_- and Ψ^n_+, in addition to Ψ^n_k and Ψ^{n+1}_{k+1} and a contact discontinuity of speed u^n_*. The fluid speed is continuous across the states Ψ^n_- and Ψ^n_+. These states are defined so that mechanical equilibrium is ensured at the contact discontinuity.

Let us assume that Ψ^n_- is the state to the left of the contact discontinuity and Ψ^n_+ to the right. The main idea is to use a standard numerical flux function $\mathscr{F}(\Psi_k, \Psi_{k+1})$, $\forall k \in I, k \neq \iota, \iota + 1$ and from (9) to deduce Ψ^{n+1}_k. In contrast, for Ψ^{n+1}_ι and $\Psi^{n+1}_{\iota+1}$ we have

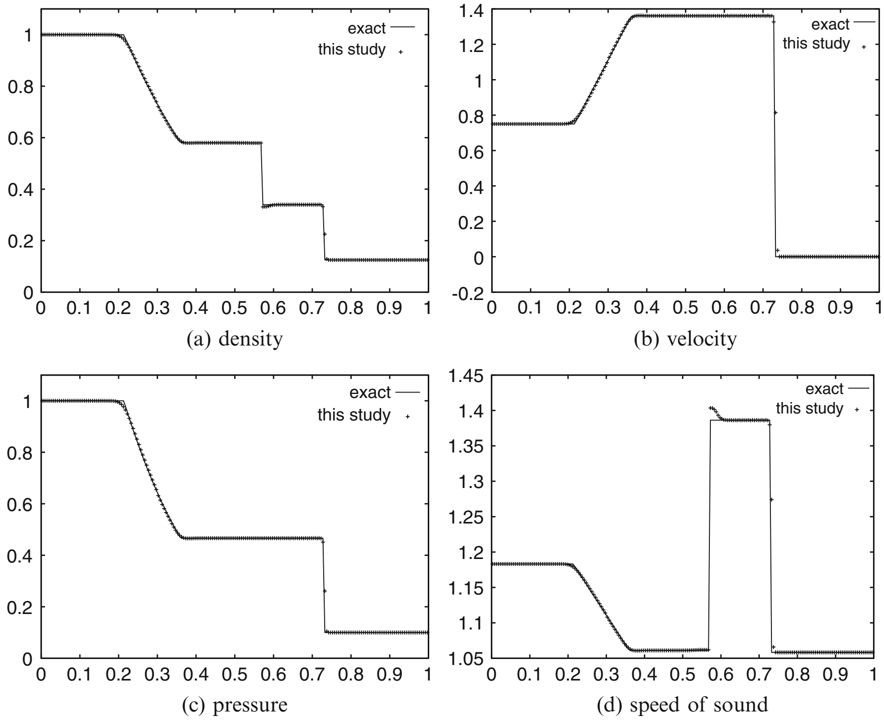
$$\begin{cases} \dfrac{\Psi^{n+1}_\iota - \Psi^n_\iota}{\Delta t} = -\dfrac{\mathscr{F}^n_-(\Psi^n_-) - \mathscr{F}^n_{\iota-1/2}(\Psi^n_{\iota-1}, \Psi^n_\iota)}{\Delta x} \\ \dfrac{\Psi^{n+1}_{\iota+1} - \Psi^n_{\iota+1}}{\Delta t} = -\dfrac{\mathscr{F}^n_{\iota+3/2}(\Psi^n_{\iota+1}, \Psi^n_{\iota+2}) - \mathscr{F}^n_+(\Psi^n_+)}{\Delta x} \end{cases} \tag{10}$$

where $\mathscr{F}^n_\pm = F(\Psi^n_\pm)$. The scheme is locally non conservative since $\mathscr{F}^n_+ \neq \mathscr{F}^n_-$. However, the effect on the approximation of shocks is negligible. The interface position is updated in time using u^n_*, i.e., $x^{n+1}_f = x^n_f + u^n_* \, \Delta t$. For numerical stability, the integration step is limited by the fastest of the characteristics over the grid points. Hence, the interface position will belong to the same interval between two grid points for more than one time step. When the physical interface overcomes a grid point, i.e., $x^{n+1}_f \geq x_{i+1}$ or $x^{n+1}_f < x_i$ then $\iota = i \pm 1$ accordingly. In other words, the above integration scheme is simply shifted of one point to the right or to the left.

When the interface crosses a grid point, however, the corresponding conservative variables Ψ^{n+1}_ι do not correspond anymore to the material present at that grid point before the integration step. When $\iota = i + 1$, i.e., the physical interface moves to the right of $i + 1$, then we take $\Psi^{n+1}_\iota = \Psi^n_-$, whereas if $\iota = i - 1$, $\Psi^{n+1}_\iota = \Psi^n_+$. The scheme proposed in [2] can be recast in a form similar to what we presented here.

4 Results

As a first test case we show the results (Fig. 1) of a computation on 200 grid points of the classical perfect gas ($\gamma = 1.4$) shock tube with conditions $\rho = 1$, $u_1 = 1.0$ and $p = 0.75$ to the left and $\rho = 0.125$, $u_1 = 0$ and $p = 0.1$ to the right. The governing equations are in this case the usual one-dimensional compressible Euler equations. The multimaterial solver is applied across the contact discontinuity, that stays sharp during the transient.



(a) density

(b) velocity

(c) pressure

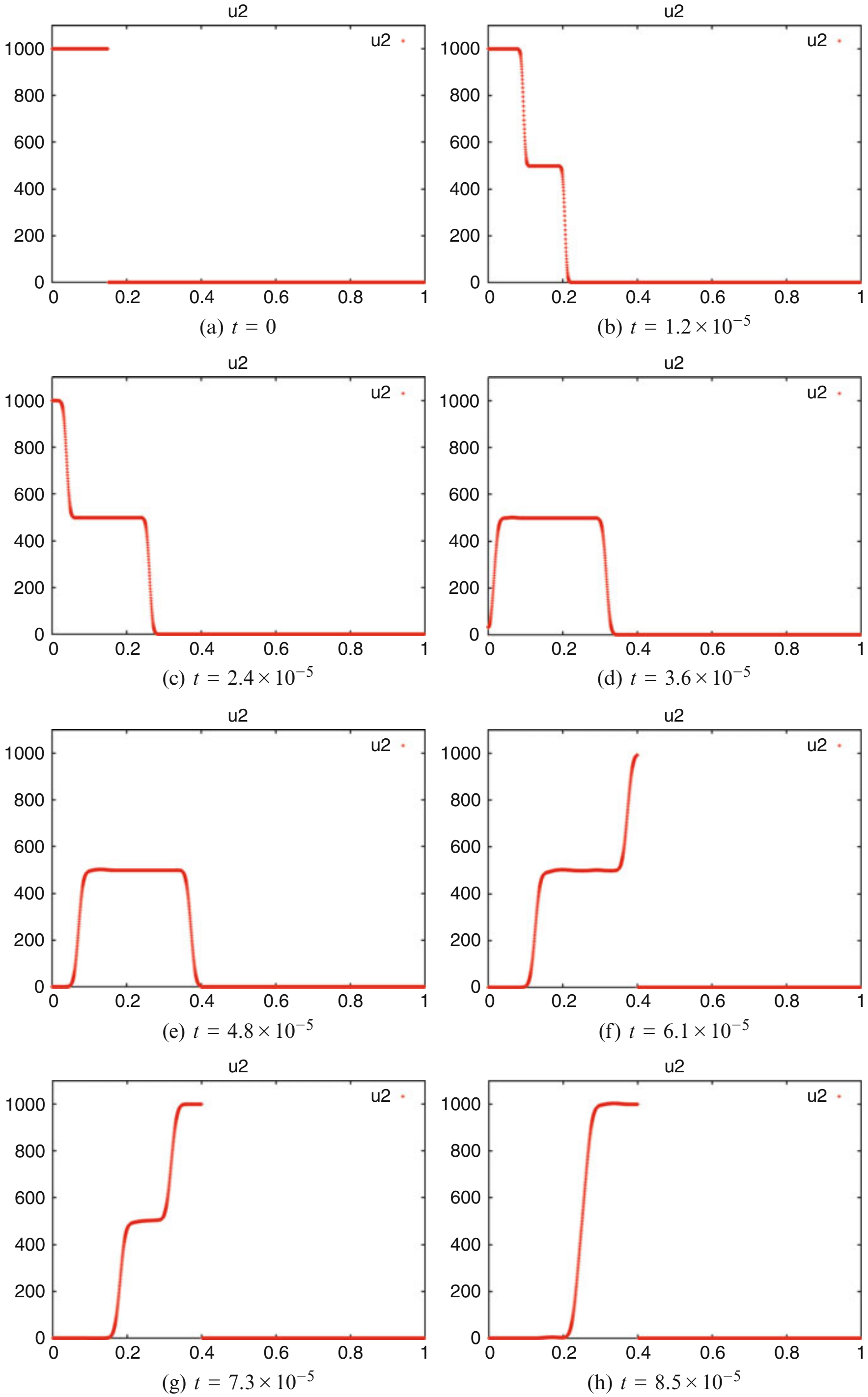
(d) speed of sound

Fig. 1 Shock tube

The subsequent test case is relevant to elastic materials with different physical properties separated by an interface. It represents a one-dimensional configuration with non-zero transverse velocity. This velocity is constant in the transverse direction but may be variable in the longitudinal direction. This configuration is similar to that presented in [5].

The test case concerns a copper-air interface with discontinuous initial conditions in the copper. The copper-air interface is at $x_1 = 0.4$. To left of the interface there is copper with $p_\infty = 342 \cdot 10^8$, $\gamma = 4.22$, $\chi = 9.2 \cdot 10^{10}$and $\rho_0 = 8.9 \cdot 10^3$. To the right there is air with $p_\infty = 0$, $\gamma = 1.4$, $\chi = 0$ and $\rho_0 = 1$. The initial conditions are uniform static pressure (10^5) and uniform horizontal velocity (0) across the materials. Inside copper the vertical velocity is 10^3 between $x_1 = 0$ and $x_1 = 0.15$ and 0 elsewhere. The vertical velocity in air is 0. The left boundary conditions are such that the solution is symmetric.

The results obtained on 2000 grid points are presented in Fig. 2 and Fig. 3. The initial discontinuity in vertical velocity at time 0 breaks down in two waves travelling in opposite directions. These waves are reflected on the left border and on the copper-air interface, giving rise to subsequent wave interactions. The sharp contact discontinuity is between copper and air. When the transverse wave hits this interface, since $\sigma^{21} = 0$, the transverse speed is discontinuous. The results are in good accordance with those presented in [5].



(a) $t = 0$

(b) $t = 1.2 \times 10^{-5}$

(c) $t = 2.4 \times 10^{-5}$

(d) $t = 3.6 \times 10^{-5}$

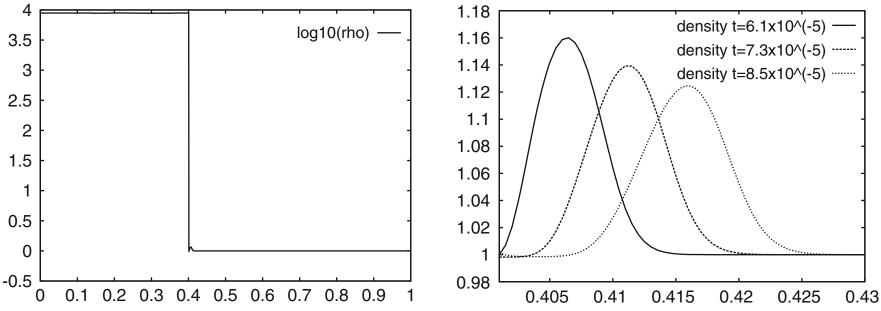
(e) $t = 4.8 \times 10^{-5}$

(f) $t = 6.1 \times 10^{-5}$

(g) $t = 7.3 \times 10^{-5}$

(h) $t = 8.5 \times 10^{-5}$

Fig. 2 Copper-air.Vertical velocity u_2



(a) $t = 6.1 \times 10^{-5}$

(b) Acoustic wave in air

Fig. 3 Copper-air. (a) Density in logarithmic scale: copper to the left of the interface, air to the right. The time snapshot shown correspond to the first shear wave interaction with the copper-air interface ((f) in Fig. 2). Copper density is nearly constant during the transient. In (b) a zoom of the acoustic wave in air for several time steps ((f), (g) and (h) in Fig. 2)

References

1. R. Abgrall and S. Karni. Computations of compressible multifluids. *Journal of computational physics*, 169(2):594–623, 2001.
2. A. Chertock, S. Karni, and A. Kurganov. Interface tracking method for compressible multifluids. *ESAIM: Mathematical Modelling and Numerical Analysis*, 42:991–1019, 2008.
3. P.G. Ciarlet. *Mathematical elasticity Vol I, Three dimensional elasticity*. Volume 20 of Studies in Mathematics and its Applications, 1994.
4. G.-H. Cottet, E. Maitre, and T. Milcent. Eulerian formulation and level set models for incompressible fluid-structure interaction. *M2AN*, 42:471–492, 2008.
5. N. Favrie, S.L. Gavrilyuk, and R. Saurel. Solidfluid diffuse interface model in cases of extreme deformations. *Journal of computational physics*, 228(16):6037–6077, 2009.
6. S.L. Gavrilyuk, N. Favrie, and R. Saurel. Modelling wave dynamics of compressible elastic materials. *Journal of computational physics*, 227(5):2941–2969, 2008.
7. S.K. Godunov. Elements of continuum mechanics. *Nauka Moscow*, 1978.
8. G.H. Miller and P. Colella. A high-order eulerian godunov method for elasticplastic flow in solids. *Journal of computational physics*, 167(1):131–176, 2001.
9. E.F. Toro, M. Spruce, and W. Speares. Restoration of the contact surface in the hll-riemann solver. *Shock Waves*, 4:25–34, 1994.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Numerical Simulation of Viscous and Viscoelastic Fluids Flow by Finite Volume Method

Radka Keslerová and Karel Kozel

Abstract This paper deals with the numerical modeling of steady incompressible laminar flows of viscous and viscoelastic fluids. The governing system of the equations is based on the system of balance laws for mass and momentum for incompressible fluid. Two models for the stress tensor are tested. The models used in this study are generalized Newtonian model with power-law viscosity model and Oldroyd-B model with constant viscosity. The numerical results for these models are presented.

Keywords viscous and viscoelastic stress tensor, generalized Newtonian and Oldroyd-B model, Runge–Kutta method, finite volume method
MSC2010: 65N08, 76D05, 65L06, 76A05, 76A10

1 Introduction

Generalized Newtonian fluids can be subdivided according to the viscosity behavior. For Newtonian fluids the viscosity is constant and is independent of the applied shear stress. Shear thinning fluids are characterized by decreasing viscosity with increasing shear rate. Shear thickening fluids are characterized by increasing viscosity with increasing shear rate. The Fig. 1 shows the dependence of viscosity on the shear rate, see e.g. [2].

R. Keslerová and K. Kozel
CTU in Prague, Karlovo nám. 13, Praha, Czech Republic, e-mail: keslerov@marian.fsik.cvut.cz, kozelk@fsik.cvut.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_62, © Springer-Verlag Berlin Heidelberg 2011

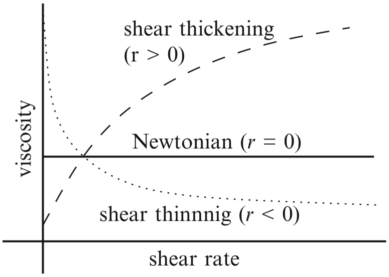


Fig. 1 Viscosity generalized Newtonian fluid as a function of shear rate for power-law fluid

2 Mathematical Model

The governing system of equations is the system of balance laws of mass and momentum for incompressible fluids [1], [7]:

$$\operatorname{div} \boldsymbol{u} = 0 \tag{1}$$

$$\rho\frac{\partial \boldsymbol{u}}{\partial t} + \rho(\boldsymbol{u}.\nabla)\boldsymbol{u} = -\nabla P + \operatorname{div} \mathsf{T} \tag{2}$$

where P is the pressure, ρ is the constant density, $\boldsymbol{u}$ is the velocity vector. The symbol T represents the stress tensor.

2.1 Stress tensor

In this work the different choices of the definition of the stress tensor are used.

a) Viscous fluids

The simple viscous model is *Newtonian model*:

$$\mathsf{T} = 2\mu\mathsf{D} \tag{3}$$

where μ is the dynamic viscosity and tensor D is the symmetric part of the velocity gradient.

This model could be generalized by extending Newtonian model for shear thinning and thickening fluids flow. For this case the viscosity μ is no more constant, but is defined as the viscosity function by the power-law model [4]

$$\mu = \mu(\dot{\gamma}) = \mu_{\epsilon}\left(\sqrt{\mathrm{tr}\mathsf{D}^2}\right)^r, \tag{4}$$

where μ_ϵ is a constant, e.g. the dynamic viscosity for Newtonian fluid. The symbol tr D^2 denotes the trace of the tensor D^2. The exponent r is the power-law index. This model includes Newtonian fluids as a special case ($r = 0$). For $r > 0$ the power-law fluid is shear thickening, while for $r < 0$ it is shear thinning, (see Fig. 1).

b) Viscoelastic fluids

The behavior of the mixture of viscous and viscoelastic fluids can be described by *Oldroyd-B model* and it has the form

$$\mathsf{T} + \lambda_1 \frac{\delta \mathsf{T}}{\delta t} = 2\mu \left(\mathsf{D} + \lambda_2 \frac{\delta \mathsf{D}}{\delta t} \right). \tag{5}$$

The parameters λ_1, λ_2 are *relaxation* and *retardation time*.

The stress tensor T is decomposed to the Newtonian part T_s and viscoelastic part T_e ($\mathsf{T} = \mathsf{T}_s + \mathsf{T}_e$) and

$$\mathsf{T}_s = 2\mu_s \mathsf{D}, \qquad \mathsf{T}_e + \lambda_1 \frac{\delta \mathsf{T}_e}{\delta t} = 2\mu_e \mathsf{D}, \tag{6}$$

where

$$\frac{\lambda_2}{\lambda_1} = \frac{\mu_s}{\mu_s + \mu_e}, \qquad \mu = \mu_s + \mu_e. \tag{7}$$

The *upper convected derivative* $\frac{\delta}{\delta t}$ is defined (for general tensor) by the relation (see [7])

$$\frac{\delta \mathsf{M}}{\delta t} = \frac{\partial \mathsf{M}}{\partial t} + (\boldsymbol{u}.\nabla)\mathsf{M} - (\mathsf{WM} - \mathsf{MW}) - (\mathsf{DM} + \mathsf{MD}) \tag{8}$$

where D is the symmetric part of the velocity gradient $\mathsf{D} = \frac{1}{2}(\nabla \boldsymbol{u} + \nabla \boldsymbol{u}^T)$ and W is the antisymmetric part of the velocity gradient $\mathsf{W} = \frac{1}{2}(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}^T)$.

The governing system (1), (2) of equations is completed by the equation for the viscoelastic part of the stress tensor

$$\frac{\partial \mathsf{T}_e}{\partial t} + (\boldsymbol{u}.\nabla)\mathsf{T}_e = \frac{2\mu_e}{\lambda_1}\mathsf{D} - \frac{1}{\lambda_1}\mathsf{T}_e + (\mathsf{W}\mathsf{T}_e - \mathsf{T}_e\mathsf{W}) + (\mathsf{D}\mathsf{T}_e + \mathsf{T}_e\mathsf{D}). \tag{9}$$

3 Numerical Solution

Numerical solution of the described models is based on cell-centered finite volume method using explicit Runge–Kutta time integration. The unsteady system of equations with steady boundary conditions is solved by finite volume method. Steady state solution is achieved for $t \to \infty$. In this case the artificial compressibility method can be applied. It means that the continuity equation is completed by the time derivative of the pressure in the form (for more details see e.g. [8]):

$$\frac{1}{\beta^2}\frac{\partial p}{\partial t} + \operatorname{div} \boldsymbol{u} = 0, \quad \beta \in \mathbb{R}^+. \tag{10}$$

The system of equations (including the modified continuity equation) could be rewritten in the vector form.

$$\tilde{R}_\beta W_t + F^c_x + G^c_y = F^v_x + G^v_y + S, \qquad \tilde{R}_\beta = \operatorname{diag}(\frac{1}{\beta^2}, 1, 1, 1, 1, 1). \tag{11}$$

where W is the vector of unknowns, F^c, G^c are inviscid fluxes, F^v, G^v are viscous fluxes defined as

$$W = \begin{pmatrix} p \\ u \\ v \\ t_{11} \\ t_{12} \\ t_{22} \end{pmatrix}, \quad F^c = \begin{pmatrix} u \\ u^2 + p \\ uv \\ ut_{11} \\ ut_{12} \\ ut_{22} \end{pmatrix}, \quad G^c = \begin{pmatrix} v \\ uv \\ v^2 + p \\ vt_{11} \\ vt_{12} \\ vt_{22} \end{pmatrix}, \tag{12}$$

$$F^v = \begin{pmatrix} 0 \\ 2\mu(\dot\gamma)u_x \\ \mu(\dot\gamma)(u_y + v_x) \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad G^v = \begin{pmatrix} 0 \\ \mu(\dot\gamma)(u_y + v_x) \\ 2\mu(\dot\gamma)v_y \\ 0 \\ 0 \\ 0 \end{pmatrix} \tag{13}$$

and the source term S is defined as where t_{ij} are components of the symmetric tensor T_e

$$S = \begin{pmatrix} 0 \\ t_{11x} + t_{12y} \\ t_{12x} + t_{22y} \\ 2\frac{\mu_e}{\lambda_1}u_x - \frac{t_{11}}{\lambda_1} + 2(u_x t_{11} + u_y t_{12}) \\ \frac{\mu_e}{\lambda_1}(u_y + v_x) - \frac{t_{12}}{\lambda_1} + (u_x t_{12} + u_y t_{22} + v_x t_{11} + v_y t_{12}) \\ 2\frac{\mu_e}{\lambda_1}v_y - \frac{t_{22}}{\lambda_1} + 2(v_x t_{12} + v_y t_{22}) \end{pmatrix} \tag{14}$$

The following special parameters are used:

Newtonian	$\mu(\dot\gamma) = \mu_s = const.$	$\mathsf{T}_e \equiv 0$
Generalized Newtonian	$\mu(\dot\gamma)$	$\mathsf{T}_e \equiv 0$
Oldroyd-B	$\mu(\dot\gamma) = \mu_s = const.$	T_e

The eq. (11) is discretized in space by the cell-centered finite volume method (see [3]) and the arising system of ODEs is integrated in time by the explicit multistage Runge–Kutta scheme (see [4], [6], [9]):

$$\begin{aligned} W_i^n &= W_i^{(0)} \\ W_i^{(s)} &= W_i^{(0)} - \alpha_{s-1} \Delta t \mathscr{R}(W)_i^{(s-1)} \\ W_i^{n+1} &= W_i^{(M)} \qquad\qquad s = 1, \dots, M, \end{aligned} \tag{15}$$

where $M = 3, \alpha_0 = \alpha_1 = 0.5, \alpha_2 = 1.0$, the steady residual $\mathscr{R}(W)_i$ is defined by finite volume method as

$$\mathscr{R}(W)_i = \frac{1}{\sigma_i} \sum_{k=1}^{4} \left[\left(\overline{F}_k^c - \overline{F}_k^v \right) \Delta y_k - \left(\overline{G}_k^c - \overline{G}_k^v \right) \Delta x_k \right] + \overline{S}, \tag{16}$$

where σ_i is the volume of the cell, $\sigma_i = \int \int_{C_i} dx\ dy$. The symbols $\overline{F}_k^c, \overline{G}_k^c$ and $\overline{F}_k^v, \overline{G}_k^v$ denote the numerical approximation of the inviscid and viscous fluxes, for more details see [5], symbol $\overline{S}$ represents the numerical approximation of the source term with central approximation of derivatives.

4 Numerical results

The steady numerical results in the branching channel for two dimensional generalized Newtonian fluids are shown in the Sect. 4.1. In the Sect. 4.2 the comparison of Newtonian and Oldroyd-B fluids is presented for simple 2D channel.

4.1 Two Dimensional Case

In this section the steady numerical results are presented. The comparison of Newtonian and non-Newtonian shear thickening and shear thinning fluids for Re = 400 in the geometry of the branching channel in the form of the velocity isolines is shown in the Fig. 2.

The following choices of the power-law index were used: for Newtonian fluid $r = 0$, for shear thickening and shear thinning fluid values $r = 0.5$ and $r = -0.5$. In the inlet the velocity is prescribed by the parabolic profile. The histories of the convergence are also presented in the Fig. 2. One can observe some differences between tested fluids in the size of the separation region.

The nondimensional axial velocity profile for steady fully developed flow of considered fluids is shown in the Fig. 3. In these figure the small channel is sketched. The line (inside the domain) marks the position where the cuts for the velocity profile were done.

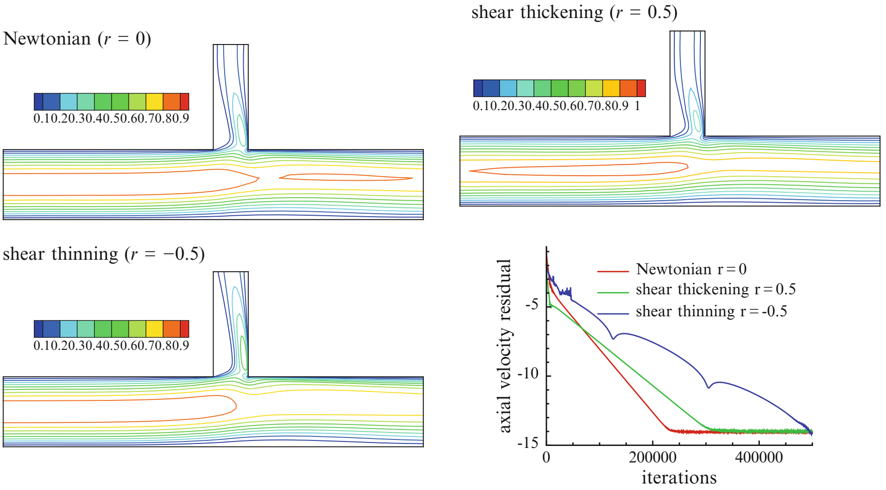


Fig. 2 Velocity isolines and history of the convergence of steady flows for generalized Newtonian fluids

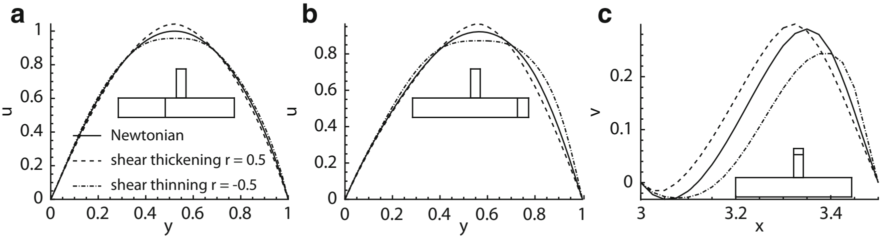


Fig. 3 Nondimensional velocity profile for steady fully developed flow of generalized Newtonian fluids in the branching channel (the line legend in the a) is the same for all figures)

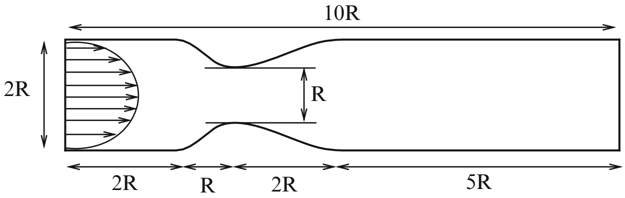


Fig. 4 Structure of the computational domain

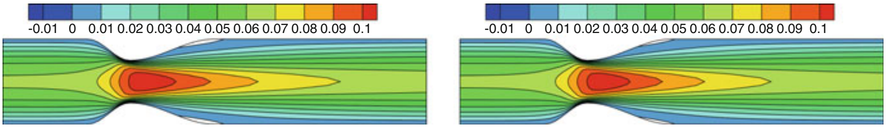
4.2 Viscous and Visoelastic Model

This section deals with the comparison of the numerical results of Newtonian and Oldroyd-B fluids. Fig. 4 shows the shape of the tested domain.

The following model parameters are:

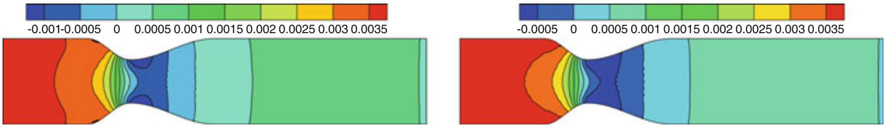
$$\begin{array}{cc}
\mu_e = 4.0 \cdot 10^{-4} Pa \cdot s & \mu_s = 3.6 \cdot 10^{-3} Pa \cdot s \\
\lambda_1 = 0.06s & \lambda_2 = 0.054s \\
U_0 = 0.0615m \cdot s^{-1} & L_0 = 2R = 0.0062m \\
\mu_0 = \mu = \mu_s + \mu_e & \rho = 1050kg \cdot m^{-3}
\end{array}$$

In the Figs. 5 and 6 the comparison of the axial velocity isolines and the pressure distributions is presented.



(a) Newtonian (b) Oldroyd-B

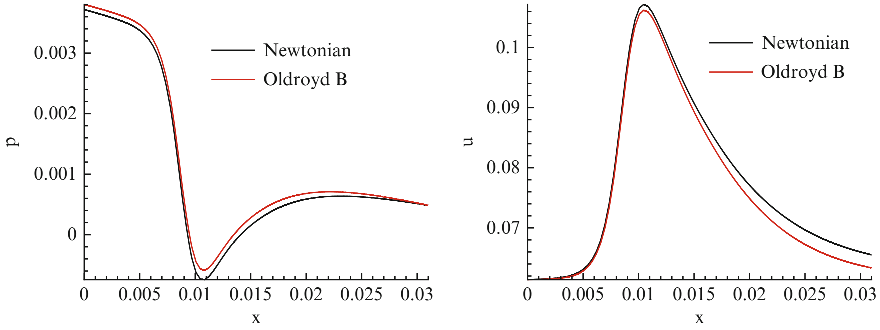
Fig. 5 Axial velocity isolines for Newtonian and Oldroyd-B fluids



(a) Newtonian (b) Oldroyd-B

Fig. 6 Pressure distribution for Newtonian and Oldroyd-B fluids

Pressure and velocity distribution along the axis for both tested fluids models is shown in the Fig. 7. By simple observation one can conclude that the main effect of the Oldroyd-B fluids behavior is visible mainly in the recirculation zone.



(a) pressure (b) axial velocity

Fig. 7 Pressure and axial velocity distribution along the central axis of the channel

5 Conclusions

Newtonian model with its generalized modification and Oldroyd-B model have been considered for numerical simulation of fluids flow in the branching channel and in the idealized axisymmetric stenosis. The cell-centered finite volume solver for incompressible laminar viscous and viscoelastic fluids flow has been described. Generalized Newtonian model was used for testing of different choices of the power-law index r in the branching channel. For time integration the explicit Runge–Kutta method was considered. The numerical results obtained by this method are presented. We can conclude that the numerical results of the tested fluids agrees with well-known non-Newtonian behavior. The differences between these three fluids are given mainly in the separation region.

In the idealized stenosis we tested the Newtonian and Oldroyd-B fluids models. Here the two definitions of the stress tensor were used. Based on the above numerical results we can conclude that the difference between the viscous and viscoelastic fluids is visible in the recirculation zone.

Acknowledgements This work was partly supported by the grant GACR 101/09/1539 and GACR 201/09/0917, Research Plan MSM 684 077 0003 and Research Plan MSM 684 077 0010.

References

1. Dvořák, R., Kozel, K.: Mathematical Modelling in Aerodynamics (in Czech). CTU, Prague, Czech Republic (1996).
2. Robertson, A.M., Sequeira, A., Kameneva, M.V.: Hemorheology. Birkhäuser Verlag Basel, Switzerland (2008).
3. LeVeque, R.: Finite-Volume Methods for Hyperbolic Problems. Cambridge University Press, (2004).
4. Keslerová, R., Kozel, K.: Numerical modelling of incompressible flows for Newtonian and non-Newtonian fluids, Mathematics and Computers in Simulation, **80**, 1783–1794 (2010).
5. Keslerová, R., Kozel, K.: Numerical solution of laminar incompressible generalized Newtonian fluids flow, Applied Mathematics and Computation, **217**, 5125–5133 (2011).
6. Jameson, A., Schmidt, W., Turkel, E.: Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes, AIAA 14th Fluid and Plasma Dynamic Conference California (1981).
7. Bodnar, T., Sequeira, A.: Numerical study of the significance of the non-Newtonian nature of blood in steady flow through stenosed vessel (Editor: R. Ranacher, A. Sequeira), Advances in Mathematical Fluid Mechanics, 83–104 (2010).
8. Chorin, A.J.: A numerical method for solving incompressible viscous flow problem, Journal of Computational Physics, **135**, 118–125 (1967).
9. Vimmr, J., Jonášová, A.: Non-Newtonian effects of blood flow in complete coronary and femoral bypasses, Mathematics and Computers in Simulation, **80**, 1324–1336 (2010).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

An Aggregation Based Algebraic Multigrid Method Applied to Convection-Diffusion Operators

Sana Khelifi, Namane Méchitoua, Frank Hülsemann, and Frédéric Magoulès

Abstract The paper focuses on an aggregation-based algebraic multigrid method applied to convection/diffusion problems. We show that for an unstructured finite volume approach on arbitrary shaped cells, the separation of the two operators associated with suitable smoothers improves the aggregation-based multigrid. While the convection is treated by a piecewise constant prolongation, the off-diagonals entries of the diffusion P_0 Galerkin operator are scaled by a parameter representative of the mesh spacing ratio between the fine and coarse mesh in the vicinity of the coarse mesh cell boundaries. Some numerical examples are shown to assess the rate of convergence and the robustness of the proposed approach.

Keywords finite volumes, algebraic multigrid, convection-diffusion
MSC2010: 76M12

Introduction

The ongoing increase in computing power renders the solving of ever larger linear systems possible, so that the use of algorithms with optimal algebraic complexity becomes necessary. From this point of view, multigrid methods (MG) represent a

Sana Khelifi
EDF R&D, MFEE, 6 Quai Wattier F-78401 Chatou Cedex and Ecole Centrale de Paris, Grande voie des Vignes, F-92295 Chatenay-Malabry, e-mail: sana.khelifi@edf.fr

Namane Méchitoua
EDF R&D, MFEE, 6 Quai Wattier F-78401 Chatou, Cedex, e-mail: namane.mechitoua@edf.fr

Frank Hülsemann
EDF R&D, SINETICS, 1, avenue du Général de Gaulle, F-92140 Clamart Cedex, e-mail: frank.hulsemann@edf.fr

Frédéric Magoulès
Ecole Centrale de Paris, Grande voie des Vignes, F-92295 Chatenay-Malabry, e-mail: frederic.magoules@hotmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_63, © Springer-Verlag Berlin Heidelberg 2011

viable alternative to other solution strategies since their theoretical computational complexities scale quasi-linearly with the problem size, especially for elliptic dominated problems [11]. The key ingredient of the multigrid technique relies on the use of a hierarchy of grids for solving a linear set of equations. The fast convergence of the multigrid scheme is based on the fact that, for each component of the error, there exists a grid level on which the error component in question is efficiently reduced. Compared to the standard geometric procedure, the algebraic multigrid has become very competitive. The hierarchy of grids is created automatically, taking into account the matrix entries of the discretized operator. This procedure allows the effective solution of a large class of linear systems arising from highly non homogeneous PDE discretized with unstructured meshes. Among the different interpolation schemes used for the restriction and prolongation operators involved in the coarse level matrix calculation, the piecewise constant interpolation is a limiting case of the "Ruge-Stüben" multigrid procedure [11]. Consisting of agglomerating the fine mesh points for creating coarse mesh points, this approach is widely used in finite volume based CFD solvers [8]. In order to recover the theoretical convergence for elliptic second order operators [6], the rescaling of the coarse level P_0 Galerkin operator is a judicious and quite simple mean. The trivial method, consisting of rescaling by a global number is very simple to implement but it is limited to problems that can be solved effectively with a geometric multigrid procedure [9]. A second approach called "smoothed aggregation" [12] improves the MG convergence, but it increases the number of off-diagonal entries of the coarse level matrix and therefore destroys the simplicity of the original approach. A third approach, although only studied with a finite volume scheme on fully unstructured meshes, greatly improves the scaled P_0 Galerkin multigrid procedure, thanks to an original face based rescaling [10]. It maintains the simplicity of P_0 interpolation. However, for equations mixing convection and diffusion operators, acting preferentially at different grid scales, the optimal use of multilevel techniques is less evident, with the presence of several different strategies for overcoming these difficulties [2], [4], [5], [7]. The aim of the paper is to present a strategy for solving such systems, in the framework of the face based rescaling algebraic multigrid procedure.

1 Finite Volume procedure

The scalar convection/diffusion equation $div(\overrightarrow{Q}C - \kappa\overrightarrow{\nabla}C) = b$, where $\overrightarrow{Q}$, κ and C represent respectively a divergence free velocity field, a diffusion coefficient and the unknown scalar to be solved, is representative of the transport/diffusion terms of the momentum, energy or stationary mass fraction equations used in CFD solvers [1]. The integration over a discrete cell Ω_I is written as:

$$\int_{\Omega_I} div(\overrightarrow{Q}C - \kappa\overrightarrow{\nabla}C) = \sum_{J\in V_I}(\overrightarrow{Q}_{IJ}.\overrightarrow{N}_{IJ})C_{IJ} + \sum_{J\in V_I}(-\kappa\overrightarrow{\nabla}C)_{IJ}.\overrightarrow{N}_{IJ} = b_I|\Omega_I|, \quad (1)$$

where V_I represents the neighbourhood of the cell I, i.e the set of cells J sharing a non-zero area surface IJ with the cell I and $\overrightarrow{N}_{IJ}$ designs the normal vector to the face IJ pointing from Ω_I to Ω_J. Numerical consistency and precision for diffusive and convective fluxes for non-orthogonal cells are taken into account using a gradient reconstruction technique. This technique is useful for increasing the order of some numerical schemes, when applied to complex situations as unstructured meshes for instance. It concerns both first order (convection) and second order (diffusion) differential equations, discretized with finite volume methods. Among the various numerical fluxes, assuming regular coefficient κ, the following one is used for the diffusion:

$$(-\kappa\overrightarrow{\nabla}C)_{IJ}.\overrightarrow{N}_{IJ} = \frac{\kappa_{IJ}(C_I - C_J) + (\overrightarrow{II'} - \overrightarrow{JJ'}).(\kappa\overrightarrow{\nabla}C)^c_{IJ}}{I'J'}.|\overrightarrow{N}_{IJ}|, \tag{2}$$

where $(\kappa\overrightarrow{\nabla}C)^c_{IJ} \approx 0.5(\kappa_I\overrightarrow{\nabla}C_I + \kappa_J\overrightarrow{\nabla}C_J)$ represents the cell gradient projected at the face centre F, either evaluated with a finite volume or a least square formulation and κ_{IJ} represents the face interpolation of the diffusion coefficient (with arithmetic, harmonic or geometric interpolation). For the convective part of the fluxes, the simplest upwind scheme remains consistent for non orthogonal smooth meshes, but the order can be less than one. Among the various numerical higher order convective fluxes, the following one is used for a centred interpolation of the convected variable C at the face centre F (also named IJ) of the interface separating two cells I and J (see Fig. 1):

$$\begin{aligned} C_{IJ} &= C_F = C_O + \overrightarrow{OF}.\overrightarrow{\nabla}C^c_{IJ} \\ C_O &= \tfrac{OJ}{IJ}C_I + \tfrac{OI}{IJ}C_J \end{aligned} \tag{3}$$

In order to avoid instabilities, the convective schemes for all variables, except the pressure, are non-linear centred or second order upwind schemes. The switch between first order upwind and higher order interpolation is triggered if the non-monotony of the variable in the neighbourhood of the interpolation point is detected. The linear set of equations arising from discrete formulations (1) with the higher order diffusive flux (2) and convective flux (3), is not solved directly with GMRES or BICG-STAB method, because the resolution can be too expensive

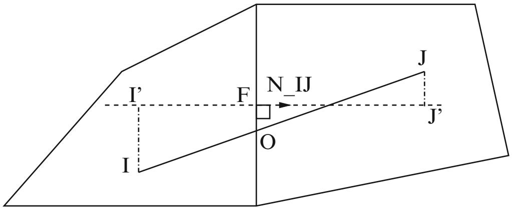


Fig. 1 Geometrical parameters at the face separating cells I and J. I' (res. J') is the orthogonal projection of I (res. J) on the normal through the face centre F

or even impossible without robust pre-conditioners. The computation of such systems is made through a defect correction technique. The explicit (or initial) convective/diffusive flux (named old) computed with the scheme (2) and (3) is taken into account in the right hand side. The iteration matrix for solving the correction, making use only of the original finite volume neighbourhood, is a positive M matrix, which can be solved by suitable iterative methods, such as conjugate gradient or multigrid solvers.

$$\begin{aligned} \int_{\Omega_I} div(\overrightarrow{Q}\phi - \kappa\overrightarrow{\nabla}\phi) &\approx \sum_{J\in V_I} (\overrightarrow{Q}_{IJ}.\overrightarrow{N}_{IJ})\phi_{IJ}^{upwind} + \frac{\kappa_{IJ}(\phi_I-\phi_J)}{I'J'}|\overrightarrow{N}_{IJ}| \\ &= b_I\Omega_I - \sum_{J\in V_I}(\overrightarrow{Q}_{IJ}.\overrightarrow{N}_{IJ}\phi_{IJ} - \kappa\overrightarrow{\nabla}C_{IJ}^{old}.\overrightarrow{N}_{IJ}) \qquad (4) \\ C^{new} &= C^{old}+\phi \end{aligned}$$

In (4) , $\phi_{IJ}^{upwind} = \phi_I$ if $(\overrightarrow{Q}_{IJ}.\overrightarrow{N}_{IJ}) > 0$ and $\phi_{IJ}^{upwind} = \phi_J$ otherwise. The defect correction procedure (4) is repeated, by replacing C^{old} by C^{new}, until the residual tends towards a user defined value.

2 Multigrid procedure

Many of the multigrid approaches for CFD solvers are based on the idea of separating the elliptic part from the non-elliptic one in the PDE [3]. System (4) can be written as $(C + D)\phi = f$, where C is a non-symmetric M-matrix obtained by upwind-biased discretization of the convection and D is a symmetric M-matrix corresponding to a low order discrete scheme (on non orthogonal meshes) of diffusion. The piecewise constant interpolation for restriction and prolongation operators involved in the coarse level matrix construction preserves the M-matrix properties of the 2 operators, and hence the smoothing properties of simple Gauss-Seidel or Jacobi-type relaxation schemes. The coarse convection operator is constructed based on the Galerkin product with a piecewise constant interpolation operator [11]. The off-diagonal entries XC^0 of the coarse convection operator read as:

$$\begin{aligned} XC^0_{I_C J_C} &= \sum_{(I_k,J_k)} Min((\overrightarrow{Q}_{I_kJ_k}\overrightarrow{N}_{I_kJ_k}),0) \\ XC^0_{J_C I_C} &= -\sum_{(I_k,J_k)} Max((\overrightarrow{Q}_{I_kJ_k}\overrightarrow{N}_{I_kJ_k}),0) \end{aligned}$$

The upwind character of the finest level discretization is propagated to all coarse levels. This property, associated with an algebraic multigrid procedure allowing the aggregation along the streamlines and associated with an appropriate smoother ensures a physical coarse grid correction for the convective part. The piecewise constant interpolation for the elliptic part is not optimal from a theoretical viewpoint

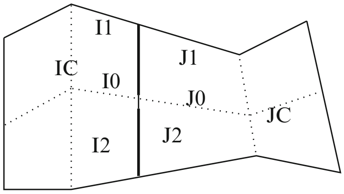


Fig. 2 Sketch of an aggregate and coarse mesh boundary where I_C and J_C are the gravity centres of the coarse cells, (I_k,J_k) are the fine mesh cells situated on both sides of the coarse mesh interface and I_0 (resp. J_0) is an average value of I_1, I_2 (resp. J_1, J_2)

[6]. The optimal multigrid convergence of the elliptic part is obtained in a quite simple manner, considering a finite volume discretization on fully unstructured meshes [10]. A geometrical face-based rescaling of the off-diagonal entries XD^0 of the P$_0$ coarse mesh matrix takes into account the mesh spacing ratio between the fine and coarse level in the vicinity of the coarse mesh cell boundaries, as represented in Fig.2. In our notation, the rescaling reads as follows:

$$XD_{I_C J_C} = \frac{I'_0 J'_0}{I'_C J'_C} XD^0_{I_C J_C}, \quad \text{with } XD^0_{I_C J_C} = -\sum_{(I_k,J_k)} \frac{\kappa_{IJ}}{I'_k J'_k} |\overrightarrow{N}_{I_k J_k}|$$

The detailed derivation of the rescaling method for the elliptic part is given in [10]. The singularly perturbed character of convection complicates the use of multigrid techniques, because of the possible poor coarse grid approximation of the convective part. The typical approach is to use a smoother which eliminates the singular perturbation errors on the finest level, so that the coarse grid correction can handle efficiently the remaining elliptic part of the errors. Gauss-Seidel like methods are quite fast iterative solvers for convection operators discretized with upwind biased techniques. Downwind numbering w.r.t constant characteristics [2] or, more generally, curved characteristics following the vortices in the flow [5] renders the Gauss-Seidel iteration sufficiently robust. Nevertheless, these formulations cannot be easily applied to complex flows inside complex geometries, in terms of implementation and set up phase computing time. The symmetric Gauss-Seidel (SSOR) relaxation scheme, for which upwind and downwind directions are swept, represents a viable alternative for complex situations. Previous numerical assessments have shown that it is nearly as robust as circular ordering [7].

The simplest cycle in the multigrid resolution, the V-cycle, is used, in combination with SSOR methods acting as smoothers. The algebraic multigrid procedure is based upon the strength of the matrix connectivity, defined in a symmetric way for an aggregation-based procedure. Two cells numbered i and j are merged if $max(A_{ij}^2, A_{ji}^2)/A_{ii}A_{jj}$ is greater than a threshold value, progressively relaxed until the targeted coarsening is reached. Each coarse grid has approximately one third of the number of cells of the previous fine grid, representing a good trade off between the efficiency of the smoothing with few iterations (3 for a coarsening ratio of about 3) and grid complexity.

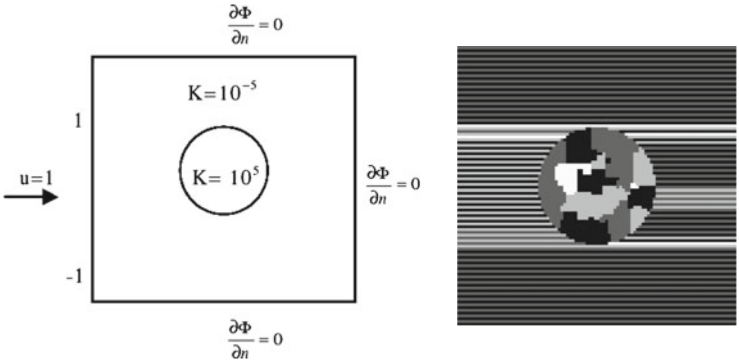


Fig. 3 The square test case: initial parameters and the coarsest level with a finest mesh of 100×100

3 Numerical examples

In the following examples, we tested the based piecewise constant multigrid scheme (referred as P_0) and the new scheme with the rescaled diffusion operator (referred as P_1). The multigrid scheme is used as a stand alone solver. Our unit of measure is the equivalent of the number of matrix-vector products performed on the finest level. It is a representative measure of the arithmetic and memory access operations performed during the resolution that does not depend on the computer.

3.1 The square test

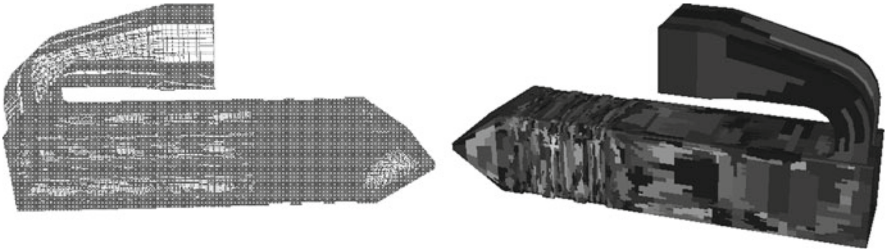
In this example, the diffusion coefficient is piecewise constant with a strong jump and the convection velocity is horizontal and equal to 1, as shown on Fig. 3. The right hand side is homogeneous and equal to 0. The initial solution is zero on the whole domain with the boundary conditions shown on Fig. 3. The stopping criterion is based on a threshold on a normalised residual. The number of iterations presented in Table 1 stands for the equivalent of the number of matrix-vector products performed on the initial mesh (the finest grid). The number between brackets represents the number of cycles performed. For Gauss-Seidel as a stand alone solver, one iteration counts for 3 matrix-vector products (2 sweeps and the computing of the residual). Observing the different results, we notice a stability of the P_1 multigrid solver while the P_0 one exhibits a mesh dependency. The mesh dependency of the aggregation P_0 solver is a well known fact. The rescaling of this operator yields much better results. For the SSOR stand alone solver, it is clear that it is not competitive. It is not efficient because of diffusive part.

Table 1 The number of matrix-vector products (MG cycles), for different solvers on a sequence of 2D meshes for the square problem

Number of cells in one direction	10	100	500	1,000
Symmetric Gauss-Seidel	63	3,372	74,409	284,301
P_0 V-cycle	71 (5)	547 (31)	1,588 (94)	5,353 (298)
P_1 V-cycle	55 (4)	126 (8)	135 (9)	135 (9)

3.2 *The 600 MW corner fired boiler*

The second example concerns the steady transport/diffusion source term of NO_x polluant inside a 600 MW corner fired boiler, see the Fig. 4. The boiler is fitted with 24 burners displayed at 3 levels. The velocity field and the diffusivity are relatively complex [13]. The hierarchy of grids obtained is summarised in Table 2. A reduction by a factor around 3 is noticed between all the levels. The agglomeration is stopped when the coarse level size drops below 1% of the cells number in the finest level.


(a) The geometry

(b) The coarsest level

Fig. 4 The geometry, the initial mesh and the coarsest level of the boiler test case

Table 2 Hierarchy of grids obtained for the boiler test case

Grid level	0	1	2	3	4	5
Ncel_k	462,784	158,169	52,631	17,403	5,739	1,898
Nfac_k	1,363,784	629,626	278,518	113,296	42,760	15,106

The different results obtained with the multigrid scheme, the stand alone symmetric Gauss-Seidel solver and the stand alone BICG-Stab solver are summarised in Table 3. As we recompute the residual at the end of each BICG-Stab iteration, each iteration counts for 3 matrix-vector products. The number between brackets represents the number of MG cycles performed. It is clear that symmetric Gauss-Seidel is not competitive because of the presence of diffusive dominated regions and recirculation zones. The results obtained with BICG-Stab are reasonable but far away from those obtained with the proposed scheme.

Conclusion: Based on a finite volume approach on arbitrary cell shapes, the rescaling of the piecewise constant elliptic part of the convection/diffusion equation

Table 3 Number of matrix-vector products (and multigrid cycles) for the boiler test case

solver	P_0 V-cycle	P_1 V-cycle	SSOR	BICG-Stab
number of iterations	919 (49)	387(20)	115,806	4227

was successfully accomplished. The separation of the convection and the diffusion operators in order to enable the use of the basic piecewise constant interpolation operator for the convection while rescaling the Galerkin coarse grid operator of the diffusion, associated with suitable smoothers, yields better results than using the same interpolation for both operators. The numerical examples show the robustness and the convergence rate of the proposed technique.

References

1. Archambeau F., Mechitoua N., Sakiz M.: Code_Saturne: a finite volume code for the computation of turbulent incompressible flows- Industrial Applications. Int. J. on Finite Volumes, **1** (2004). http://www.latp.univ-mrs.fr/IJFV/.
2. Bey J., Wittum G.: Downwind numbering: robust multigrid for convection-diffusion problems. Applied Numerical Mathematics, **23**,177–192 (1997).
3. Brandt A., Yavneh I.: On multigrid solution of high Reynolds incompressible entering flow. J. Comp. Phys., **101**, 151–164 (1992).
4. Guillard H., Vanek P.: An Aggregation Multigrid Solver for Convection-diffusion Problems on Unstructured Meshes. University of Colorado at Denver, technical report(1998).
5. Hackbusch W., Probst T.: Downwind Gauss-Seidel Smoothing for Convection Dominated Problems. Numerical Linear Algebra with Applications, **4**, 85–102 (1997).
6. Hemker P.W.: On the order of prolongations and restrictions in multigrid procedures. J. Comp. Applied Math., 423–429 (1990).
7. Kanschat G.: Robust smoothers for high-order discontinuous Galerkin discretizations of advection-diffusion problems. J. of Comp. and App. Math., **218**, 53–60 (2008).
8. Lonsdale R.D.: An algebraic multigrid solver for the Navier Stokes equations on unstructured meshes. Int. J. Num. Meth. Heat and FluidFlow, **3**, 3–14 (1993).
9. Mavriplis D.J., Venkatakrishnan V.: Agglomeration Multigrid for 2 Dimensional Viscous Flows. Computers and Fluids, **24**, 553–570, (1995).
10. Mechitoua N., Hülsemann F., Fournier Y.: Improvement of a Finite Volume Based Multigrid Method Applied to Elliptic Problems. Int. Conf. on Math., Comp. Methods & Reactor Physics (M&C09), Saratoga Springs, New York, May 3-7, (2009).
11. Trottenberg U., Oosterlee C., Schuller A.: Multigrid. Elsevier Academic Press. ISBN 0-12-701070-X, (2001).
12. Vanek P., Mandel J., Brezina M.: Algebraic Multigrid by smoothed aggregation for 2nd order and 4th order elliptic problems. Computing, **56**, 179–196, (1996).
13. Dal Secco S., Schuck Y.: Using three dimensional simulation of pulverized coal combustion in a 600 MW corner fired boiler to identify low NO_x operating conditions. VGB conference in Potsdam, “Power plants in competition, operation, technology and environment”, (2005).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Stabilized DDFV Schemes For The Incompressible Navier-Stokes Equations

Stella Krell

Abstract "Discrete Duality Finite Volume" schemes (DDFV for short) on general meshes are studied here for the Navier-Stokes problem with Dirichlet boundary conditions. The DDFV method falls in the class of the so-called staggered scheme: the discrete unknowns, the components of the velocity and the pressure, are located on different nodes. The scheme is stabilized using a finite volume analogue to Brezzi-Pitkäranta techniques. We prove the wellposedness of the scheme for general meshes and we derive the first energy estimates. Finally, we illustrate the convergence properties with numerical experimentations.

Keywords Finite-volume methods, Navier-Stokes problem, DDFV scheme.
MSC2010: 65N08, 76D03, 76D05.

1 Introduction

We restrict here the presentation to the Navier-Stokes problem with homogeneous Dirichlet boundary conditions and a smooth viscosity which depends on the spatial variable. The system reads as follows:

$$\begin{cases} \partial_t \mathbf{u} + \mathrm{div}\,(-2\eta(x)\mathrm{D}\mathbf{u} + p\mathrm{Id}) + (\mathbf{u}\cdot\nabla)\mathbf{u} = \mathbf{f}, \text{ in }]0,T[\times\Omega, \\ \qquad\qquad\qquad\qquad\qquad\qquad \mathrm{div}(\mathbf{u}) = 0, \text{ in }]0,T[\times\Omega, \end{cases} \tag{1}$$

where the unknowns are the velocity $\mathbf{u}\ :]0,T[\times\Omega \to \mathbb{R}^2$ and the pressure $p\ :\ ]0,T[\times\Omega \to \mathbb{R}$ such that $\displaystyle\int_\Omega p(t,x)\mathrm{d}x = 0$, for all $t\ \in]0,T[$, Ω is a polygonal open bounded connected subset of $\mathbb{R}^2$, $T > 0$. We recall that $\mathrm{D}\mathbf{u} = \frac{1}{2}(\nabla\mathbf{u} + {}^t\nabla\mathbf{u})$

Stella Krell
INRIA, Lille, France, e-mail: stella.krell@inria.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_64, © Springer-Verlag Berlin Heidelberg 2011

and $(\mathbf{u}\cdot\nabla)\mathbf{u} = \sum_{i=1}^{2} \mathbf{u}_i\,\partial_i \mathbf{u}$ for $\mathbf{u} = (\mathbf{u}_1, \mathbf{u}_2)$. We supplement the system (1) with the following boundary and initial conditions:

$$\begin{cases} \mathbf{u} = 0, \text{ on }]0,T[\times\partial\Omega, \\ \mathbf{u}(0,.) = \mathbf{u}_{\text{ini}}, \text{ in } \Omega. \end{cases}$$

We assume that $\mathbf{f}$ is a function in $(L^2(]0,T[\times\Omega))^2$, $\mathbf{u}_{\text{ini}}$ is a function in $(L^\infty(\Omega))^2$ and the viscosity η is a function in $W^{1,\infty}(\Omega)$ with $\operatorname*{Inf}_{\Omega} \eta > 0$.

Finite volume approximation of Navier-Stokes problem is a current research topic, we refer to [5, 6, 9–11, 17] for the description and the analysis of the main available schemes up to now. We consider here the class of finite volume schemes called DDFV, which have been first introduced and studied in [7, 12] to approximate the solution of the Laplace equation on a large class of 2D meshes including non-conformal and distorted meshes and without "orthogonality" assumptions as for classical finite volume methods. This strategy has been extended to a wide class of PDE problems [1–4, 13, 15] and gives a staggered method for the Navier-Stokes equations: the approximate velocity is located at the centers and at the vertices of the mesh and the approximate pressure at the edges of the mesh. In a previous work [15], we proposed a stabilized DDFV scheme for the Stokes problem with variable viscosity, which is equivalent (except on the boundary) to two uncoupled MAC schemes [14, 16] when the grid is cartesian and the viscosity is constant. We will use here the same discretization for the viscous part of momentum conservation and the mass conservation equations of (1).

This paper is organized as follows. In Sect. 2, we construct the approximation of the non-linear convective term. In Sect. 3, we introduce the DDFV stabilized scheme for the Navier-Stokes problem (1), we begin with existence and uniqueness of the approximate solution, then we present the first energy estimates. Finally, in Sect. 4, we illustrate the convergence with numerical results.

2 The DDFV framework

We use the same notation as in [15] for the Stokes problem. We do not recall here the complete description of meshes and operators:

- the DDFV meshes $(\mathfrak{T},\mathfrak{D})$: $\mathfrak{T}$ is constituted by the primal mesh $\mathfrak{M}\cup\partial\mathfrak{M}$, which is the initial mesh, and the dual mesh $\mathfrak{M}^*\cup\partial\mathfrak{M}^*$, whose cells κ^* are built around the vertices of the primal mesh (Fig. 1), and $\mathfrak{D}$ is the diamond mesh, whose cells D are built around the edges of the primal mesh.
- a discrete gradient $\nabla^{\mathfrak{D}} : \left(\mathbb{R}^2\right)^{\mathfrak{T}} \to (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}}$, its discrete dual operator $\mathbf{div}^{\mathfrak{T}} : (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}} \to \left(\mathbb{R}^2\right)^{\mathfrak{T}}$, its trace $\mathrm{div}^{\mathfrak{D}} : \left(\mathbb{R}^2\right)^{\mathfrak{T}} \to \mathbb{R}^{\mathfrak{D}}$, a discrete strain rate tensor $\mathrm{D}^{\mathfrak{D}} : \left(\mathbb{R}^2\right)^{\mathfrak{T}} \to (\mathrm{M}_2(\mathbb{R}))^{\mathfrak{D}}$ and a stabilization term $\Delta^{\mathfrak{D}} : \mathbb{R}^{\mathfrak{D}} \to \mathbb{R}^{\mathfrak{D}}$.

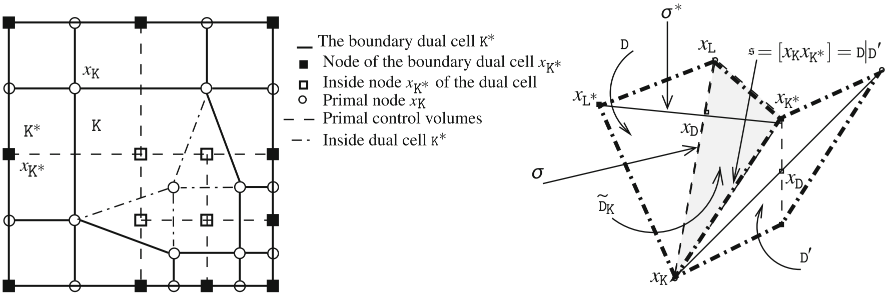

x_{K}
K*
K
x_{K^*}
The boundary dual cell K*
Node of the boundary dual cell x_{K^*}
Inside node x_{K^*} of the dual cell
Primal node x_{K}
Primal control volumes
Inside dual cell K*
σ^*
D
x_{L}
$\mathfrak{s} = [x_{\mathrm{K}}x_{\mathrm{K}^*}] = \mathrm{D}|\mathrm{D}'$
x_{L^*}
x_{K^*}
x_{D}
σ
$\widetilde{\mathrm{D}}_{\mathrm{K}}$
$x_{\mathrm{D}'}$
D'
x_{K}

Fig. 1 The mesh $\mathfrak{T}$ (left). A diamond D with a neighbour diamond D′(right)

Concerning the discretization of the nonlinear term $(\mathbf{u}^n \cdot \nabla)\mathbf{u}^{n+1}$, a conflict appears when writing the DDFV scheme because $\mathbf{u}^n$ is defined on centers and vertices of the mesh whereas $\nabla\mathbf{u}^{n+1}$ is defined on the diamond mesh. We have to approach $\displaystyle\int_V (\mathbf{u}^n \cdot \nabla)\mathbf{u}^{n+1}\mathrm{d}x$ on both the primal and dual cells. Using a Stokes formula, we get $\displaystyle\sum_{\sigma\in\partial V}\int_\sigma (\mathbf{u}^n \cdot \mathbf{n}_{\sigma,V})\mathbf{u}^{n+1}\mathrm{d}s$, this quantity will be approached using a scheme of the form: $\displaystyle\sum_{\sigma\in\partial V} F_{\sigma,V}\mathbf{u}_{\oe}^{n+1}$. In the following section, we explain how to define $F_{\sigma,V}$ and in Sect. 2.2, how to define $\mathbf{u}_{\oe}^{n+1}$.

2.1 Approximation of the normal flux

The major difficulty in the construction of the scheme lies in the approximation of

$$\int_\sigma (\mathbf{u}^n \cdot \mathbf{n}_{\sigma,V})\mathrm{d}s. \tag{2}$$

We use the idea already presented in [8, 11] that is to define discrete mass fluxes taking into account the stabilization term. This allows to ensure the convenient property given below in Proposition 2.

Expression of discrete mass fluxes through a diamond edge. At the continuous level, the Stokes formula gives:

$$\int_{\mathrm{D}} \mathrm{div}(\mathbf{u}^n)\mathrm{d}x = \sum_{\mathfrak{s}\in\partial\mathrm{D}}\int_{\mathfrak{s}} \mathbf{u}^n \cdot \mathbf{n}_{\mathfrak{s}\mathrm{D}}\mathrm{d}s.$$

The discrete counterpart of this equality is:

$$|\mathrm{D}|\mathrm{div}^{\mathrm{D}}(\mathbf{u}^n) - \lambda d_{\mathrm{D}}^2|\mathrm{D}|\Delta^{\mathrm{D}}p^n = \sum_{\mathfrak{s}\in\partial\mathrm{D}} G_{\mathfrak{s},\mathrm{D}}(\mathbf{u}^n, p^n),$$

where d_{D} is the diameter of D and for $\mathfrak{s} = [x_{\mathrm{K}}, x_{\mathrm{K}^*}] = \mathrm{D}|\mathrm{D}'$ (see Fig. 1), we have:

$$G_{\mathfrak{s},\mathrm{D}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) = m_{\mathfrak{s}} \frac{\mathbf{u}_{\mathrm{K}} + \mathbf{u}_{\mathrm{K}^*}}{2} \cdot \mathbf{n}_{\mathfrak{s}\mathrm{D}} - \lambda(d_{\mathrm{D}}^2 + d_{\mathrm{D}'}^2)(p^{\mathrm{D}'} - p^{\mathrm{D}}).$$

We can approach the mass fluxes $\displaystyle\int_{\mathfrak{s}} \mathbf{u}^n \cdot \mathbf{n}_{\mathfrak{s}\mathrm{D}} \mathrm{d}s$ by using $G_{\mathfrak{s},\mathrm{D}}(\mathbf{u}^n, p^n)$.

Link between the integral (2) and the mass conservation equation. Noting $\widetilde{\mathrm{D}_{\mathrm{K}}}$ the triangle whose vertices are x_{K}, x_{K^*} and x_{L^*} (see Fig. 1), we remark that:

$$0 = \int_{\widetilde{\mathrm{D}_{\mathrm{K}}}} \mathrm{div}(\mathbf{u}^n)\mathrm{d}x = \int_{\sigma} \mathbf{u}^n \cdot \mathbf{n}_{\sigma\mathrm{K}} \mathrm{d}s + \sum_{\mathfrak{s}\in\mathfrak{S}_{\mathrm{K}}\cap\partial\mathrm{D}} \int_{\mathfrak{s}} \mathbf{u}^n \cdot \mathbf{n}_{\mathfrak{s}\mathrm{D}} \mathrm{d}s.$$

where $\mathfrak{S}_{\mathrm{K}} = \{\mathfrak{s} \in \mathfrak{S}, \text{s. t. } \mathfrak{s} \subset \mathrm{K}\}$ for all $\mathrm{K} \in \mathfrak{M}$ and $\mathfrak{S}_{\mathrm{K}^*} = \{\mathfrak{s} \in \mathfrak{S}, \text{s. t. } \mathfrak{s} \subset \mathrm{K}^*\}$ for all $\mathrm{K}^* \in \mathfrak{M}^*$, recalling that $\mathfrak{S}$ is the set of interior diamond sides. Thus, we define $F_{\sigma,\mathrm{K}}$ and $F_{\sigma^*,\mathrm{K}^*}$, the approximation of mass fluxes (2), as follows:

$$\begin{aligned} F_{\sigma,\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) &= - \sum_{\mathfrak{s}\in\mathfrak{S}_{\mathrm{K}}\cap\partial\mathrm{D}} G_{\mathfrak{s},\mathrm{D}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}), \quad \text{where } \sigma \subset \mathrm{D},\ \forall \mathrm{K} \in \mathfrak{M}, \\ F_{\sigma^*,\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) &= - \sum_{\mathfrak{s}\in\mathfrak{S}_{\mathrm{K}^*}\cap\partial\mathrm{D}} G_{\mathfrak{s},\mathrm{D}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}), \quad \text{where } \sigma^* \subset \mathrm{D},\ \forall \mathrm{K}^* \in \mathfrak{M}^*. \end{aligned} \tag{3}$$

We remark that if $(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}})$ satisfies $\mathrm{div}^{\mathfrak{D}}(\mathbf{u}^{\mathfrak{T}}) - \lambda d_{\mathfrak{D}}^2 \Delta^{\mathfrak{D}} p^{\mathfrak{D}} = 0$, we have the conservativity of the fluxes:

$$\begin{aligned} F_{\sigma,\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) &= - F_{\sigma,\mathrm{L}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}), \quad \forall \sigma = \mathrm{K}|\mathrm{L}, \\ F_{\sigma^*,\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) &= - F_{\sigma^*,\mathrm{L}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}), \quad \forall \sigma^* = \mathrm{K}^*|\mathrm{L}^*. \end{aligned}$$

With our choice, we obtain that the approximation of the integral of the velocity divergence on the primal and dual cells vanishes:

Proposition 1. *Let $\mathfrak{T}$ be a DDFV mesh. For all $(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) \in \left(\mathbb{R}^2\right)^{\mathfrak{T}} \times \mathbb{R}^{\mathfrak{D}}$, we have*

$$\forall \mathrm{K} \in \mathfrak{M}, \sum_{\sigma\in\partial\mathrm{K}} F_{\sigma,\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) = 0 \ \text{ and } \ \forall \mathrm{K}^* \in \mathfrak{M}^*, \sum_{\sigma^*\in\partial\mathrm{K}^*} F_{\sigma^*,\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) = 0.$$

2.2 Discretization of the non-linear term

Using the definition of the mass fluxes given by (3), we can define the discretization of the non-linear term with an upwind method.

Definition 1. We define $\mathbf{b}^{\mathfrak{T}} : \left(\mathbb{R}^2\right)^{\mathfrak{T}} \times \mathbb{R}^{\mathfrak{D}} \times \left(\mathbb{R}^2\right)^{\mathfrak{T}} \to \left(\mathbb{R}^2\right)^{\mathfrak{T}}$, as follows:

$$\mathbf{b}_{\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}, \mathbf{v}^{\mathfrak{T}}) = \frac{1}{|\mathrm{K}|} \sum_{\sigma\in\partial\mathrm{K}} F_{\sigma,\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) \mathbf{v}_{\sigma^+},\ \forall \mathrm{K} \in \mathfrak{M},$$

$$\mathbf{b}_{\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}, \mathbf{v}^{\mathfrak{T}}) = \frac{1}{|\mathrm{K}^*|} \sum_{\sigma^* \in \partial \mathrm{K}^*} F_{\sigma^*,\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}})\mathbf{v}_{\sigma^*+}, \ \forall \mathrm{K}^* \in \mathfrak{M}^*,$$

where

$$\mathbf{v}_{\sigma+} = \begin{cases} \mathbf{v}_{\mathbf{K}} & \text{if } F_{\sigma,\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) \geq 0, \\ \mathbf{v}_{\mathbf{L}} & \text{elsewhere.} \end{cases} \qquad \mathbf{v}_{\sigma^*+} = \begin{cases} \mathbf{v}_{\mathbf{K}^*} & \text{if } F_{\sigma^*,\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}) \geq 0, \\ \mathbf{v}_{\mathbf{L}^*} & \text{elsewhere.} \end{cases}$$

The unconditional stability of the scheme is ensured by the crucial result:

Proposition 2. *Let $\mathfrak{T}$ be a DDFV mesh. For all $(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}, \mathbf{v}^{\mathfrak{T}}) \in \mathbb{E}_0 \times \mathbb{R}^{\mathfrak{D}} \times \mathbb{E}_0$ such that $\mathrm{div}^{\mathfrak{D}}(\mathbf{u}^{\mathfrak{T}}) - \lambda d_{\mathfrak{D}}^2 \Delta^{\mathfrak{D}} p^{\mathfrak{D}} = 0$, we have*

$$\sum_{\mathrm{K} \in \mathfrak{M}} |\mathrm{K}| \mathbf{b}_{\mathrm{K}}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}, \mathbf{v}^{\mathfrak{T}}) \cdot \mathbf{v}_{\mathbf{K}} + \sum_{\mathrm{K}^* \in (\mathfrak{M}^* \cup \partial\mathfrak{M}^*)} |\mathrm{K}^*| \mathbf{b}_{\mathrm{K}^*}(\mathbf{u}^{\mathfrak{T}}, p^{\mathfrak{D}}, \mathbf{v}^{\mathfrak{T}}) \cdot \mathbf{v}_{\mathbf{K}^*} \geq 0.$$

3 DDFV schemes for the Navier-Stokes equation

Let $N \in \mathbb{N}^*$. We note $\delta t = \frac{T}{N}$ and $t_n = n\delta t$ for $n \in \{0, \cdots, N\}$. We use an implicit Euler time discretization except for the non-linear term which is linearized thanks to a standard semi-implicit approximation $(\mathbf{u}^n \cdot \nabla)\mathbf{u}^{n+1}$. The DDFV scheme for the problem (1) reads as follows:

- **Initialization:** we define $\mathbf{u}^0 \in \mathbb{E}_0$ and $p^0 \in \mathbb{R}^{\mathfrak{D}}$ as follows:

$$\begin{cases} \mathbf{u}^0 = \mathbb{P}_m^{\mathfrak{T}} \mathbf{u}_{\text{ini}} \in \mathbb{E}_0, \\ p^0 \in \mathbb{R}^{\mathfrak{D}}, \text{ s. t. } \Delta^{\mathfrak{D}} p^0 = \dfrac{1}{\lambda d_{\mathfrak{D}}^2} \mathrm{div}^{\mathfrak{D}}(\mathbb{P}_m^{\mathfrak{T}} \mathbf{u}_{\text{ini}}) \text{ with } \displaystyle\sum_{\mathrm{D} \in \mathfrak{D}} |\mathrm{D}| p_{\mathrm{D}}^0 = 0. \end{cases} \tag{4}$$

  Note that with this choice of $(\mathbf{u}^0, p^0)$, we have $\mathrm{div}^{\mathfrak{D}}(\mathbf{u}^0) - \lambda d_{\mathfrak{D}}^2 \Delta^{\mathfrak{D}} p^0 = 0$.
- **Time stepping:** assume that $(\mathbf{u}^n, p^n) \in \mathbb{E}_0 \times \mathbb{R}^{\mathfrak{D}}$ are given ($n \in \{0, \cdots, N-1\}$). We have to find $\mathbf{u}^{n+1} \in \mathbb{E}_0$ and $p^{n+1} \in \mathbb{R}^{\mathfrak{D}}$ such that:

$$\begin{cases} \forall \mathrm{K} \in \mathfrak{M}, \dfrac{\mathbf{u}_{\mathrm{K}}^{n+1} - \mathbf{u}_{\mathrm{K}}^n}{\delta t} + \mathbf{div}^{\mathbf{K}}(-2\eta^{\mathfrak{D}} \mathrm{D}^{\mathfrak{D}} \mathbf{u}^{n+1} + p^{n+1}\mathrm{Id}) + \mathbf{b}_{\mathrm{K}}(\mathbf{u}^n, p^n, \mathbf{u}^{n+1}) = \mathbf{f}_{\mathrm{K}}^{n+1}, \\ \forall \mathrm{K}^* \in \mathfrak{M}^*, \dfrac{\mathbf{u}_{\mathrm{K}^*}^{n+1} - \mathbf{u}_{\mathrm{K}^*}^n}{\delta t} + \mathbf{div}^{\mathbf{K}^*}(-2\eta^{\mathfrak{D}} \mathrm{D}^{\mathfrak{D}} \mathbf{u}^{n+1} + p^{n+1}\mathrm{Id}) + \mathbf{b}_{\mathrm{K}^*}(\mathbf{u}^n, p^n, \mathbf{u}^{n+1}) = \mathbf{f}_{\mathrm{K}^*}^{n+1}, \\ \mathrm{div}^{\mathfrak{D}}(\mathbf{u}^{n+1}) - \lambda d_{\mathfrak{D}}^2 \Delta^{\mathfrak{D}} p^{n+1} = 0, \\ \displaystyle\sum_{\mathrm{D} \in \mathfrak{D}} |\mathrm{D}| p_{\mathrm{D}}^{n+1} = 0, \end{cases} \tag{5}$$

where $\eta^{\mathfrak{D}} = (\eta(x_{\text{D}}))_{\text{D}\in\mathfrak{D}}$, $\lambda > 0$ given, $\mathbf{f}_{\text{K}}^{n+1} = \dfrac{1}{\delta t\ |\text{K}|}\displaystyle\int_{t_n}^{t_{n+1}}\int_{\text{K}} \mathbf{f}(t,x)\mathrm{d}x\mathrm{d}t$ for all $\text{K} \in \mathfrak{M}$ and $\mathbf{f}_{\text{K}^*}^{n+1} = \dfrac{1}{\delta t\ |\text{K}^*|}\displaystyle\int_{t_n}^{t_{n+1}}\int_{\text{K}^*} \mathbf{f}(t,x)\mathrm{d}x\mathrm{d}t$ for all $\text{K}^* \in \mathfrak{M}^*$.

Note that, in order to be able to apply Proposition 2, we have to ensure the property $\mathrm{div}^{\mathfrak{D}}(\mathbf{u}^n) - \lambda d_{\mathfrak{D}}^2 \Delta^{\mathfrak{D}} p^n = 0$ even for the initial time (*i.e.* for $n \in \{0,\cdots,N\}$). This permits to prove the following stability proposition.

Proposition 3 (Discrete energy estimates). *Let $\mathfrak{T}$ be a DDFV mesh. The finite volume scheme (4)-(5) with $\lambda > 0$ admits a unique solution $(\mathbf{u}^n, p^n)_{n\in\{0,\cdots,N\}}$. For $N > 1$, there exists a constant $C > 0$, depending only on Ω, λ, η, $\mathbf{u}_{ini}$ and $\mathbf{f}$, such that:*

$$\sum_{n=1}^{N} \delta t \|\!|\nabla^{\mathfrak{D}}\mathbf{u}^n\|\!|_2^2 \le C,\ \sum_{n=1}^{N} \delta t |p^n|_h^2 \le C,\ \sum_{n=0}^{N-1} \|\mathbf{u}^{n+1} - \mathbf{u}^n\|_2^2 \le C \ \ \textit{and}\ \ \|\mathbf{u}^N\|_2^2 \le C.$$

4 Numerical results

We show here some numerical results obtained on a domain $\Omega =]0,1[^2$ with $T = 1$ and $\delta t = 10^{-2}$. Error estimates are given on a test with a stabilization coefficient chosen to be $\lambda = 10^{-3}$. In order to illustrate error estimates, the family of meshes (see Fig. 2) are obtained by successive global refinement of the original mesh.

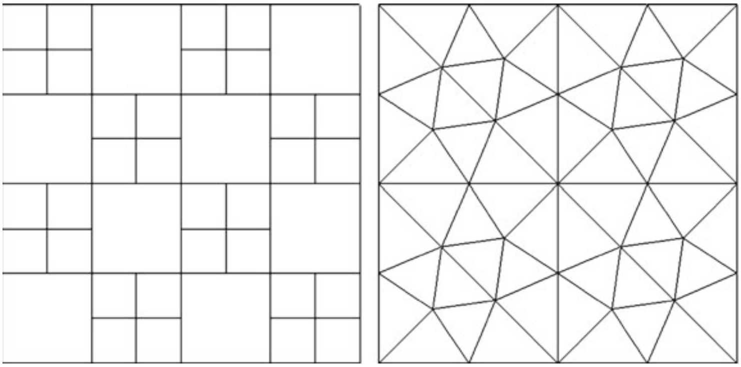

(a) Non conformal square mesh.
(b) Triangle mesh.

Fig. 2 Family of meshes

The exact solution is the Green-Taylor vortex:

$$\mathbf{u} = \begin{pmatrix} -\cos(2\pi x)\sin(2\pi y)e^{-2t\eta} \\ \sin(2\pi x)\cos(2\pi y)e^{-2t\eta} \end{pmatrix}, \quad p = -\frac{1}{4}(\cos(4\pi x) + \cos(4\pi y))e^{-4t\eta}.$$

The viscosity η being chosen, we define the source term **f** and the boundary data **g** in such a way that (1) is satisfied.

We compare the relative $L^2(\Omega\times]0,T[)$-norm of the error obtained with the DDFV scheme, for the pressure (denoted Erpre), for the velocity gradient (denoted Ergradvel) and for the velocity (denoted Ervel) respectively. On the two tables, we give the number of primal cells (denoted NbCell) and the convergence rates (denoted Ratio).

Table 1 $\eta = 1$ on the non conformal square mesh Fig. 2(a)

NbCell	Ervel	Ratio	Ergradvel	Ratio	Erpre	Ratio
208	2.804E-02	-	8.508E-02	-	1.526E+00	-
736	6.761E-03	2.052	4.309E-02	0.9815	6.574E-01	1.215
2752	1.803E-03	1.907	2.158E-02	0.9973	3.237E-01	1.022
10624	6.045E-04	1.577	1.079E-02	1.001	1.633E-01	0.9874

When the viscosity is equal to 1 (Table 1), we observe a first order convergence for the L^2-norm of the velocity gradient and of the pressure, which seems to be optimal. We obtain a super-convergence for the L^2-norm of the velocity. Furthermore, let us emphasize that the convergence rate is not sensitive to the presence of non conformal control volumes.

Table 2 $\eta = 10^{-3}$ on the triangle mesh Fig. 2(b)

NbCell	Ervel	Ratio	Ergradvel	Ratio	Erpre	Ratio
256	2.952E-01	-	4.403E-01	-	5.181E-01	-
960	2.080E-01	0.5049	3.551E-01	0.3105	3.718E-01	0.4788
3712	1.292E-01	0.6871	2.465E-01	0.5262	2.420E-01	0.6195
14592	7.432E-02	0.7975	1.643E-01	0.5858	1.432E-01	0.7573

When the viscosity is equal to 10^{-3} (Table 2), the convective term is dominant and we observe that the scheme is still convergent even if the convergence of the velocity gradient deteriorates.

5 Conclusion

In this paper, we proposed a stabilized DDFV scheme for the Navier-Stokes problem. This scheme is well-posed on 2D general meshes. Its convergence properties were illustrated with numerical experimentations. In a work in progress, we provide a proof of this result.

References

1. B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general 2D-meshes. *Numer. Methods PDE*, (2007), **23**(1):145–195.
2. F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Num. Anal.*, (2008), **46**(6):3032–3070.
3. Y. Coudière and G. Manzini. The Discrete Duality Finite Volume Method for Convection-diffusion Problems. *SIAM J. Numer. Anal. Volume*, (2010), **47**(6):4163–4192.
4. S. Delcourte, K. Domelevo, and P. Omnes. A discrete duality finite volume approach to Hodge decomposition and div-curl problems on almost arbitrary two-dimensional meshes. *SIAM J. Numer. Anal.*, (2007), **45**(3):1142–1174.
5. J. Droniou and R. Eymard. Study of the mixed finite volume method for Stokes and Navier-Stokes equations. *Num. Meth. PDEs*, (2009), **25**(1):137–171.
6. S. Delcourte. *Développement de méthodes de volumes finis pour la mécanique de fluides.* Ph.D. thesis, Univ. Paul Sabatier, Toulouse, 2007.
7. K. Domelevo and P. Omnès. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *Math. Model. Numer. Anal.*, (2005), **39**(6):1203–1249.
8. R. Eymard, T. Gallouët, R. Herbin, and J.-C. Latché. Analysis tools for finite volume schemes. *Acta Math. Univ. Comenian. (N.S.)*, (2007) **76**(1):111–136.
9. R. Eymard and R. Herbin. A new colocated finite volume scheme for the incompressible Navier-Stokes equations on general non matching grids. *C. R. Math. Acad. Sci. Paris*, (2007) **344**(10):659–662.
10. R. Eymard, R. Herbin, and J.-C. Latché. Convergence analysis of a colocated finite volume scheme for the incompressible Navier-Stokes equations on general 2D or 3D meshes. *SIAM J. Numer. Anal.*, (2007) **45**(1):1–36.
11. R. Eymard, R. Herbin, J.-C. Latché, and B. Piar. Convergence analysis of a locally stabilized collocated finite volume scheme for incompressible flows. *Math. Model. Numer. Anal.*, (2009), **43**(5):889–927.
12. F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, (2000), **160**(2):481–499.
13. F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, (2003), **192**(16-18):1939–1959.
14. F. Harlow and J. Welch. Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. *The physics of fluids*, (1965), **8**(12):2182–2189.
15. S. Krell. Stabilized DDFV schemes for Stokes problem with variable viscosity on general 2D meshes. *Num. Meth. PDEs*, (2011), available on-line: http://dx.doi.org/10.1002/num.20603.
16. R. A. Nicolaides. Analysis and convergence of the MAC scheme. I. The linear problem. *SIAM J. Numer. Anal.*, (1992), **29**(6):1579–1591.
17. S. Perron, S. Boivin, and J.-M. Hérard. A finite volume method to solve the 3D Navier-Stokes equations on unstructured collocated meshes. *Comput. & Fluids*, (2004), **33**(10):1305–1333.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Higher-Order Reconstruction: From Finite Volumes to Discontinuous Galerkin

Václav Kučera

Abstract This work is concerned with the introduction of a new numerical scheme based on the discontinuous Galerkin (DG) method. We follow the methodology of higher order finite volume (FV) and spectral volume (SV) schemes and introduce a reconstruction operator into the discontinuous Galerkin (DG) method. This operator constructs higher order piecewise polynomial reconstructions from the lower order DG scheme. We present two variants, the generalization of standard FV schemes, already proposed by Dumbser et al. (2008) and the generalization of the SV method. Theoretical aspects are discussed and numerical experiments are carried out.

Keywords Discontinuous Galerkin, finite volumes, reconstruction
MSC2010: 65M15, 65M60, 65M12

1 Problem formulation and notation

For simplicity, we shall be concerned with a scalar hyperbolic equation, although the same arguments basically hold for any time-dependent PDE. We treat a nonlinear nonstationary scalar hyperbolic equation in a bounded domain $\Omega \subset I\!R^d$ with a Lipschitz-continuous boundary $\partial\Omega$. We seek $u : \Omega \times [0,T] \to I\!R$ such that

$$\frac{\partial u}{\partial t} + \mathrm{div}\mathbf{f}(u) = 0 \quad \text{in} \;\; \Omega \times (0,T) \tag{1}$$

along with an appropriate initial and boundary condition. Here $\mathbf{f} = (f_1, \cdots, f_d)$ and $f_s, s = 1, \ldots, d$ are Lipschitz continuous fluxes in the direction $x_s, s = 1, \ldots, d$.

Václav Kučera
Charles University in Prague, Faculty of Mathematics and Physics, Sokolovská 83, Praha 8, 186 75, Czech Republic, e-mail: vaclav.kucera@email.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_65, © Springer-Verlag Berlin Heidelberg 2011

Let $\mathscr{T}_h$ be a partition (triangulation) of the closure $\overline{\Omega}$ into a finite number of closed simplices $K \in \mathscr{T}_h$. In general we do not require the standard conforming properties of $\mathscr{T}_h$ used in the finite element method (i.e. we admit the so-called hanging nodes). We shall use the following notation. By ∂K we denote the boundary of an element $K \in \mathscr{T}_h$ and set $h_K = \mathrm{diam}(K),\ h = \max_{K\in\mathscr{T}_h} h_K$.

Let $K, K' \in \mathscr{T}_h$. We say that K and K' are *neighbours*, if they share a common *face* $\Gamma \subset \partial K$. By $\mathscr{F}_h$ we denote the system of all faces of all elements $K \in \mathscr{T}_h$.

For each $\Gamma \in \mathscr{F}_h$ we define a unit normal vector $\mathbf{n}_\Gamma$, such that for $\Gamma \in \mathscr{F}_h^B$ the normal $\mathbf{n}_\Gamma$ has the same orientation as the outer normal to $\partial\Omega$.

Over a triangulation $\mathscr{T}_h$ we define the *broken Sobolev spaces*

$$H^k(\Omega, \mathscr{T}_h) = \{v;\ v|_K \in H^k(K),\ \forall K \in \mathscr{T}_h\}.$$

For each face $\Gamma \in \mathscr{F}_h^I$ there exist two neighbours $K_\Gamma^{(L)}, K_\Gamma^{(R)} \in \mathscr{T}_h$ such that $\Gamma \subset K_\Gamma^{(L)} \cap K_\Gamma^{(R)}$. We use the convention that $\mathbf{n}_\Gamma$ is the outer normal to $K_\Gamma^{(L)}$. For $v \in H^1(\Omega, \mathscr{T}_h)$ and $\Gamma \in \mathscr{F}_h^I$ we introduce the following notation:

$$v|_\Gamma^{(L)} = \text{ trace of } v|_{K_\Gamma^{(L)}} \text{ on } \Gamma, \quad v|_\Gamma^{(R)} = \text{ trace of } v|_{K_\Gamma^{(R)}} \text{ on } \Gamma, \quad [v]_\Gamma = v|_\Gamma^{(L)} - v|_\Gamma^{(R)}.$$

On boundary edges we define $v|_\Gamma^{(R)} = [v]_\Gamma := v|_\Gamma^{(L)}$.

Let $n \geq 0$ be an integer. We define the space of discontinuous piecewise polynomial functions

$$S_h^n = \{v;\ v|_K \in P^n(K), \forall K \in \mathscr{T}_h\},$$

where $P^n(K)$ is the space of all polynomials on K of degree $\leq n$. Specifically,

- S_h^0: is the space of piecewise constant functions as known from the FV method,
- $S_h^n,\ n \geq 0$: the DG solution lies in this space of piecewise nth degree polynomials,
- $S_h^N,\ N > n$: the higher order reconstructed DG solution will lie in this space.

2 Discontinuous Galerkin (DG) formulation

We multiply (1) by an arbitrary $\varphi_h^n \in S_h^n$, integrate over an element $K \in \mathscr{T}_h$ and apply Green's theorem. By summing over all $K \in \mathscr{T}_h$ and rearranging, we get

$$\frac{d}{dt}\int_\Omega u(t)\,\varphi_h^n\,dx + \sum_{\Gamma\in\mathscr{F}_h}\int_\Gamma \mathbf{f}(u)\cdot\mathbf{n}\,[\varphi_h^n]\,dS - \sum_{K\in\mathscr{T}_h}\int_K \mathbf{f}(u)\cdot\nabla\varphi_h^n\,dx = 0. \quad (2)$$

The boundary convective terms will be treated similarly as in the finite volume method, i.e. with the aid of a numerical flux $H(u, v, \mathbf{n})$:

$$\int_\Gamma \mathbf{f}(u)\cdot\mathbf{n}\,[\varphi_h^n]\,dS \approx \int_\Gamma H(u^{(L)},u^{(R)},\mathbf{n})[\varphi_h^n]\,dS. \tag{3}$$

We assume that H is *Lipschitz continuous, consistent* and *conservative*, cf. [4].

Finally, we define the *convective form* $b_h(\cdot,\cdot)$ defined for $v,\varphi\in H^1(\Omega,\mathscr{T}_h)$:

$$b_h(v,\varphi)=\int_{\mathscr{F}_h} H(v^{(L)},v^{(R)},\mathbf{n})[\varphi]\,dS-\sum_{K\in\mathscr{T}_h}\int_K \mathbf{f}(v)\cdot\nabla\varphi\,dx.$$

Definition 1 (Standard DG scheme). We seek $u:[0,T]\to S_h^n$ such that

$$\frac{d}{dt}\big(u_h(t),\varphi_h^n\big)+b_h\big(u_h(t),\varphi_h^n\big)=0,\quad \forall\varphi_h^n\in S_h^n,\ \forall t\in(0,T). \tag{4}$$

We note that if we take $n=0$, i.e. $u_h:(0,T)\to S_h^0$, then from the definition of b_h, we see that the DG scheme (4) is equivalent to the standard FV method.

3 Reconstructed discontinuous Galerkin (RDG) formulation

For $v\in L^2(\Omega)$, we denote by $\Pi_h^n v$ the $L^2(\Omega)$-projection of v on S_h^n:

$$\Pi_h^n v\in S_h^n,\quad \big(\Pi_h^n v-v,\,\varphi_h^n\big)=0,\qquad \forall\,\varphi_h^n\in S_h^n. \tag{5}$$

The basis of the proposed method lies in the observation that (2) can be viewed as an equation for the evolution of $\Pi_h^n u(t)$, where u is the exact solution of (1). In other words, due to (5), $\Pi_h^n u(t)\in S_h^n$ satisfies the following equation for all $\varphi_h^n\in S_h^n$:

$$\frac{d}{dt}\int_\Omega \Pi_h^n u(t)\,\varphi_h^n\,dx+\int_{\mathscr{F}_h}\mathbf{f}(u)\cdot\mathbf{n}\,[\varphi_h^n]\,dS-\sum_{K\in\mathscr{T}_h}\int_K\mathbf{f}(u)\cdot\nabla\varphi_h^n\,dx=0. \tag{6}$$

Now, let $N>n$ be an integer. We assume that there exists a piecewise polynomial function $U_h^N(t)\in S_h^N$, which is an approximation of $u(t)$ of order $N+1$, i.e.

$$U_h^N(x,t)=u(x,t)+O(h^{N+1}),\quad \forall x\in\Omega,\ \forall t\in[0,T]. \tag{7}$$

This is possible, if u is sufficiently regular in space, e.g. $u(t)\in W^{N+1,\infty}(\Omega)$, cf.[1]. Now we incorporate the approximation $U_h^N(t)$ into (6): the exact solution u satisfies

$$\frac{d}{dt}\big(\Pi_h^n u(t),\varphi_h^n\big)+b_h\big(U_h^N(t),\varphi_h^n\big)=E(\varphi_h^n,t),\quad \forall\varphi_h^n\in S_h^n,\ \forall t\in(0,T), \tag{8}$$

where $E(\varphi_h^n)$ is an error term defined as

$$E(\varphi_h^n,t)=b_h\big(U_h^N(t),\varphi_h^n\big)-b_h\big(u(t),\varphi_h^n\big). \tag{9}$$

Lemma 1. *The following estimate holds for all $t \in [0,T]$:*

$$E(\varphi_h^n, t) = O(h^N)\|\varphi_h^n\|_{L^2(\Omega)}. \tag{10}$$

Proof: Due to the consistency and Lipschitz continuity of H, we have on $\Gamma \in \mathscr{F}_h$

$$\mathbf{f}(u)\cdot\mathbf{n} - H(U_h^{N,(L)}, U_h^{N,(R)}, \mathbf{n}) = H(u,u,\mathbf{n}) - H(U_h^{N,(L)}, U_h^{N,(R)}, \mathbf{n}) = O(h^{N+1}).$$

Furthermore, due to the Lipschitz-continuity of $\mathbf{f}$, we have on element $K \in \mathscr{T}_h$

$$\mathbf{f}(u) - \mathbf{f}(U_h^N) = O(h^{N+1}).$$

Estimate (10) follows from these results and the application of the *inverse* and *multiplicative trace inequalities*, cf [4]. □

It remains to construct a sufficiently accurate approximation $U_h^N(t) \in S_h^N$ to $u(t)$, such that (7) is satisfied. This leads to the following problem.

Definition 2 (Reconstruction problem). Let $v : \Omega \to I\!R$ be sufficiently regular. Given $\Pi_h^n v \in S_h^n$, find $v_h^N \in S_h^N$ such that $v - v_h^N = O(h^{N+1})$ in Ω. We define the corresponding reconstruction operator $R : S_h^n \to S_h^N$ by $R\,\Pi_h^n v := v_h^N$.

By setting $U_h^N(t) := R\,\Pi_h^n u(t)$ in (8), we obtain the following semidiscrete, formally Nth order scheme for the $L^2(\Omega)$-projections of the exact solution u onto S_h^n:

$$\frac{d}{dt}\big(\Pi_h^n u(t), \varphi_h^n\big) + b_h\big(R\,\Pi_h^n u(t), \varphi_h^n\big) = O(h^N)\|\varphi_h^n\|_{L^2(\Omega)}, \quad \forall \varphi_h^n \in S_h^n. \tag{11}$$

By neglecting the right-hand side and approximating $u_h^n(t) \approx \Pi_h^n u(t)$, we arrive at the following definition of the *reconstructed discontinuous Galerkin* (RDG) scheme.

Definition 3 (Reconstructed DG scheme). We seek $u_h^n : [0,T] \to S_h^n$ such that

$$\frac{d}{dt}\big(u_h^n(t), \varphi_h^n\big) + b_h\big(Ru_h^n(t), \varphi_h^n\big) = 0, \quad \forall \varphi_h^n \in S_h^n,\ \forall t \in (0,T). \tag{12}$$

There are several points worth mentioning.

- The derivation of the RDG scheme follows the methodology of higher order FV and SV schemes, cf. [7]. The basis of these schemes is an equation for the evolution of averages of the exact solution on individual elements (i.e. an equation for $\Pi_h^0 u(t)$). Equation (11) is a direct generalization for the case of higher order $L^2(\Omega)$-projections $\Pi_h^n u(t)$, $n \geq 0$.
- Both $u_h^n(t)$ and φ_h^n lie in S_h^n. Only $Ru_h^n(t)$, lies in the higher dimensional space S_h^N. Despite this fact, equation (11) indicates that we may expect $u - Ru_h^n = O(h^{N+1})$, although $u - u_h^n = O(h^{n+1})$.

- Numerical quadrature must be employed to evaluate surface and volume integrals in (12). Since test functions are in S_h^n, as compared to S_h^N in the corresponding Nth order standard DG scheme, we may use lower order (i.e. more efficient) quadrature formulae as compared to standard DG.

As in the case of higher order FVM, we use an explicit time stepping method. For simplicity, we formulate the forward Euler method, which is only first order accurate, however in Section 5, higher order Adams-Bashforth methods are used.

Let us construct a partition $0 = t_0 < t_1 < t_2 \dots$ of the time interval $[0, T]$ and define the time step $\tau_k = t_{k+1} - t_k$. We use the approximation $u_h^n(t_k) \approx u_h^{n,k} \in S_h^n$. The forward Euler scheme is given by:

Definition 4 (Explicit RDG scheme). We seek $u_h^{n,k} \in S_h^n,\ k = 0, 1, \dots$ such that

$$\left(\frac{u_h^{n,k+1} - u_h^{n,k}}{\tau_k}, \varphi_h^n\right) + b_h\big(Ru_h^{n,k}, \varphi_h^n\big) = 0, \quad \forall \varphi_h^n \in S_h^n,\ k = 0, 1, \dots, \tag{13}$$

where $u_h^{n,0} = u_{h,0}$ is an S_h^n approximation of the initial condition u^0.

The upper limit on stable time steps, given by a CFL-like condition, is more restrictive with growing N. However, in the RDG scheme, stability properties are inherited from the lower order scheme, therefore a larger time step is possible as compared to the corresponding Nth order standard DG scheme.

3.1 Construction of the reconstruction operator

3.1.1 'Standard' approach

In the *standard approach*, a stencil (a group of neighboring elements and the element under consideration) is used to build an Nth-degree polynomial approximation to u on the element under consideration ([5] [6]). In the FV method, the von Neumann neighborhood of an element is used as a stencil to obtain a piecewise linear reconstruction, cf. Fig. 1, 1). However, for higher order reconstructions, the size of the stencil increases dramatically, cf. Fig. 1, 2), rendering higher degrees than quadratic very time consuming. In the case of the RDG scheme, we need not increase the stencil size to obtain higher order accuracy, it suffices to take the von Neumann neighborhood and increase the order of the underlying DG scheme.

In analogy to the FV method, the reconstruction operator R is constructed on each stencil independently and satisfies that $R\Pi_h^n$ is in some sense *polynomial preserving*. Specifically, for each element K and its corresponding stencil S, we require that for all $p \in P^N(S)$

$$\Big(\big(R\Pi_h^n\big)\big|_S\, p\Big)\Big|_K = p\big|_K. \tag{14}$$

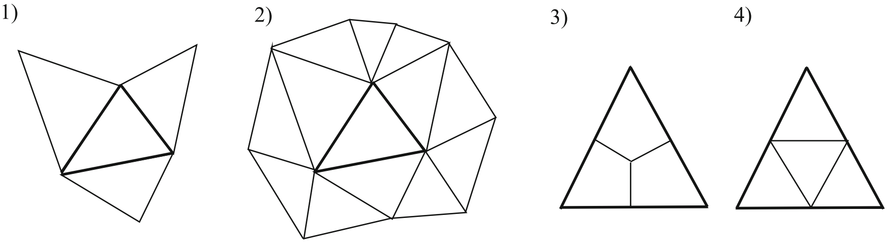


Fig. 1 1) FV stencil for linear reconstruction, 2) FV stencil for quadratic reconstruction, 3) Control volumes in a spectral volume for linear reconstruction, 4) Analogy to the SV approach for DG - partition of triangle into control volumes, e.g. cubic reconstruction from linear data

This requirement allows us to study approximation properties of R using the Bramble–Hilbert technique as in the standard finite element method, [1]. The disadvantage of this approach is that for unstructured meshes, the coefficients of the reconstruction operator must be stored for each individual stencil.

In the FV method, different conditions on R than (14) are often used, e.g. continuous or discrete least squares. Special care must be taken in the vicinity of steep gradients and discontinuities, where the Gibbs phenomenon may occur. In this case different strategies are employed, e.g. limiting, ENO and WENO schemes, TVD etc. The generalization of these concepts to the RDG method is left for future work.

3.1.2 Spectral volume approach

In the *spectral volume approach*, we start with a partition of Ω into so-called *spectral volumes* S, for example triangles in 2D. The triangulation $\mathscr{T}_h$ is formed by subdividing each spectral volume S into sub-cells K, called *control volumes*, cf. [7]. In the FV method, the order of accuracy of the reconstruction determines the number of control volumes to be generated in each spectral volume. For example, for a linear reconstruction on a triangle, the triangle is divided into three control volumes, Fig. 1, 3). Again, in the RDG scheme, we may use only the smallest available partition into control volumes, and increase the accuracy by increasing the order of the underlying scheme, cf. Fig. 1, 4).

The reconstruction operator is constructed on each spectral volume independently such that it is in some sense polynomial preserving, i.e. for each stencil S, we require that for all $p \in P^N(S)$

$$\left(R\Pi_h^n\right)\big|_S\, p = p. \tag{15}$$

The advantage of this approach is that all spectral volumes are affine equivalent, we construct the reconstruction operator R only on one reference spectral volume.

4 Relation between RDG and standard DG

The only difference between the DG scheme (4) and RDG scheme (12) is the presence of the reconstruction operator R in the first variable of $b_h(\cdot,\cdot)$. While the error analysis of (4) is well understood (at least for convection-diffusion problems [4]), the analysis of (12) or (13) poses a new challenge. The problem lies in the fact that we cannot test (12) with $\varphi_h^n := Ru_h^{n,k}$ or something similar, since $Ru_h^{n,k} \notin S_h^n$. Therefore, we need to establish a relation between (12) and Nth order DG, instead of only nth order DG.

Definition 5 (Auxiliary problem). We seek $\tilde{u}_h^{N,k} \in S_h^N$ such that

$$\left(\frac{\tilde{u}_h^{N,k+1} - \tilde{u}_h^{N,k}}{\tau_k}, \varphi_h^N\right) + b_h\big(R\Pi_h^n \tilde{u}_h^{N,k}, \varphi_h^N\big) = 0, \quad \forall \varphi_h^N \in S_h^N, \;\; k = 0, 1, \ldots, \tag{16}$$

where $\tilde{u}_h^{N,0}$ is an S_h^N approximation of the initial condition u^0.

Lemma 2. *Let $u_h^{n,0} = \Pi_h^n \tilde{u}_h^{N,0}$. Then $u_h^{n,k} \in S_h^n$, the solution of (13) and the solution $\tilde{u}_h^{N,k} \in S_h^N$ of (16) satisfy*

$$u_h^{n,k} = \Pi_h^n \tilde{u}_h^{N,k}, \quad \forall k = 0, 1, \cdots. \tag{17}$$

Proof: We prove (17) by induction:

$k = 1$: Since $u_h^{n,0} = \Pi_h^n \tilde{u}_h^{N,0}$, we have for all $\varphi_h^n \in S_h^n$

$$\begin{aligned}(\Pi_h^n \tilde{u}_h^{N,1}, \varphi_h^n) &= (\tilde{u}_h^{N,1}, \varphi_h^n) = (\tilde{u}_h^{N,0}, \varphi_h^n) - \tau_k b_h\big(R\Pi_h^n \tilde{u}_h^{N,0}, \varphi_h^n\big) \\ &= (u_h^{n,0}, \varphi_h^n) - \tau_k b_h\big(Ru_h^{n,0}, \varphi_h^n\big) = (u_h^{n,1}, \varphi_h^n),\end{aligned}$$

hence $(\Pi_h^n \tilde{u}_h^{N,1} - u_h^{n,1}, \varphi_h^n) = 0$ for all $\varphi_h^n \in S_h^n$. Therefore $\Pi_h^n \tilde{u}_h^{N,1} = u_h^{n,1}$.

$k > 1$: Assume (17) holds for some $k > 1$. Then for all $\varphi_h^n \in S_h^n$

$$\begin{aligned}\big(\Pi_h^n \tilde{u}_h^{N,k+1}, \varphi_h^n\big) &= \big(\tilde{u}_h^{N,k+1}, \varphi_h^n\big) = \big(\tilde{u}_h^{N,k}, \varphi_h^n\big) - \tau_k b_h\big(R\Pi_h^n \tilde{u}_h^{N,k}, \varphi_h^n\big) \\ &= \big(u_h^{n,k}, \varphi_h^n\big) - \tau_k b_h\big(Ru_h^{n,k}, \varphi_h^n\big) = \big(u_h^{n,k+1}, \varphi_h^n\big),\end{aligned}$$

therefore $\Pi_h^n \tilde{u}_h^{N,k+1} = u_h^{n,k+1}$. This completes the induction step $k \to k+1$. □

As a corollary, error estimates for the auxiliary problem imply error estimates for the RDG scheme (12). Problem (16) is basically the standard Nth order DG scheme with the operator $R\Pi_h^n$ in the first variable of $b_h(\cdot,\cdot)$. Therefore, sufficient knowledge of the properties of $R\Pi_h^n$ (which is polynomial preserving) and standard DG error estimates would imply the estimates for the RDG scheme.

5 Numerical experiments

We present numerical experiments for the periodic advection of a 1D sine wave on uniform meshes. Experimental orders of accuracy α in various norms on meshes with N elements are given in Tables 1 and 2. Here $e_h = u - Ru_h^n$ at t corresponding to ten periods. The increase in accuracy due to reconstruction is clearly visible.

Table 1 1D advection of sine wave, P^2 RDG scheme with P^8 reconstruction

N	$\|\|e_h\|\|_{L^\infty(\Omega)}$	α	$\|\|e_h\|\|_{L^2(\Omega)}$	α	$\|e_h\|_{H^1(\Omega,\mathscr{T}_h)}$	α
4	5.82E-03	–	3.49E-03	–	3.65E-02	–
8	7.53E-05	6.27	4.43E-05	6,30	1.06E-03	5,11
16	9.07E-07	6.38	5.95E-07	6,22	3.58E-05	4,89
32	1.82E-08	5.64	8.70E-09	6,10	1.16E-06	4,95
64	3.41E-10	5.74	1.33E-10	6,03	3.67E-08	4,98

Table 2 1D advection of sine wave, P^2 RDG scheme with P^8 reconstruction

N	$\|\|e_h\|\|_{L^\infty(\Omega)}$	α	$\|\|e_h\|\|_{L^2(\Omega)}$	α	$\|e_h\|_{H^1(\Omega,\mathscr{T}_h)}$	α
4	2.90E-03	–	1.85E-03	–	1.63E-02	–
8	7.75E-06	8.55	3.56E-06	9.02	1.03E-04	7.30
16	2.10E-08	8.53	6.64E-09	9.07	4.34E-07	7.89
32	7.21E-11	8.18	4.02E-11	7.37	1.76E-09	7.94

6 Conclusions

We have presented a possible generalization of higher-order reconstruction operators as used in the FV method to the DG method. Two constructions of the reconstruction operator R are presented, the first analogous to the standard FV case (already treated in [2]) and the construction analogous to the SV method. The resulting scheme has many advantages over standard DG, FV and SV schemes:

- To increase the order of the scheme, the reconstruction stencil need not be enlarged, we may simply increase the order of the underlying DG scheme.
- Test functions are from the lower order space, hence more efficient quadratures may be used than in the corresponding higher order DG scheme.
- Since the RDG scheme is basically a lower order DG scheme with higher order reconstruction, the CFL condition is less restrictive than for the corresponding higher order DG scheme.

Acknowledgements This work is a part of the research project MSM 0021620839 of the Ministry of Education of the Czech Republic. The research was also supported by the project P201/11/P414 of the Czech Science Foundation.

References

1. P.G. Ciarlet: *The Finite Elements Method for Elliptic Problems*, North-Holland, Amsterdam, New York, Oxford, 1979.
2. M. Dumbser, D. Balsara, E.F. Toro, C.D. Munz: *A unified framework for the construction of one-step finite-volume and discontinuous Galerkin schemes*, J. Comput. Phys. 227 (2008), pp. 8209–8253.
3. M. Feistauer, J. Felcman, I. Straškraba: *Mathematical and Computational Methods for Compressible Flow*, Oxford University Press, Oxford, 2003.
4. M. Feistauer, V. Kučera: *Analysis of the DGFEM for nonlinear convection-diffusion problems*, Electronic Transactions on Numerical Analysis, Vol. 32, No.1, (2008), pp. 33–48.
5. D. Kröner: *Numerical Schemes for Conservation Laws*, Wiley und Teubner, 1996.
6. R.J. LeVeque: *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, 2002.
7. Z. J. Wang: *Spectral (Finite) Volume Method for Conservation Laws on Unstructured Grids. Basic Formulation*, J. Comput. Phys. 178 (2002), pp. 210 – 251.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Flux-Based Approach for Conservative Remap of Multi-Material Quantities in 2D Arbitrary Lagrangian-Eulerian Simulations

Milan Kucharik and Mikhail Shashkov

Abstract Remapping is one of the essential parts of most Arbitrary Lagrangian-Eulerian (ALE) methods. It conservatively interpolates all fluid quantities from the original (Lagrangian) computational mesh to the new (rezoned) one. This paper focuses on the situation when more materials are present in the computational domain – the multi-material remap. We present a new remapping method based on the computation of the material exchange integrals (using intersections), and construction of the inter-cell fluxes of all quantities from them. As we are interested in the staggered ALE, we also briefly discuss the remap of nodal mass and velocity. Properties of the method are demonstrated on a selected numerical example.

Keywords Multi-material remap, conservative interpolation, staggered arbitrary Lagrangian-Eulerian methods
MSC2010: 35L65, 41A45, 65D05, 76T99

1 Introduction

Traditionally, there have been two families of numerical method for computational fluid dynamics, utilizing the Lagrangian or the Eulerian framework, each with its own advantages and disadvantages. In the pioneering paper [1], Hirt et al. developed the formalism combining both frameworks, and showed that this general framework could be used to combine the best properties of the Lagrangian and Eulerian methods. This class of methods has been termed Arbitrary Lagrangian-Eulerian or

Milan Kucharik
Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, Brehova 7, Praha 1, 115 19, Czech Republic, e-mail: kucharik@newton.fjfi.cvut.cz

Mikhail Shashkov
XCP-4 Group, MS-B284, Los Alamos National Laboratory, Los Alamos, NM, 87545, USA, e-mail: shashkov@lanl.gov

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_66, © Springer-Verlag Berlin Heidelberg 2011

ALE. This methodology has become very popular in recent years and many authors contributed to this topic, see for example [2–6].

For multi-material flows, the initial mesh is usually aligned with the material interfaces – each cell of the mesh contains only one material. For simple flows, it is possible to rezone the mesh in each material separately and keep the interfaces aligned with the mesh that is, do not move nodes on the interface at all or move them along the interface. Unfortunately for realistic simulations, the material interfaces get often distorted and their rezoning leads to the appearance of mixed cells containing two or more materials. We focus on the explicit material representation in form of pure material sub-polygons in each mixed cell, constructed by the modern moment-of-fluid (MOF) [7] method, which appears to be most optimal for this kind of application [8].

The ALE algorithm is usually separated into three distinct stages: 1) a Lagrangian stage in which the solution and the computational mesh are updated; 2) a rezoning stage in which the nodes of the computational mesh are moved to more optimal positions; and 3) a remapping stage in which the Lagrangian solution is interpolated onto the rezoned mesh. Here, we focus on the last stage of the ALE algorithm – the multi-material remapping. In the multi-material case, the fast and simple swept region method [9] cannot be used and one must switch to an intersection-based method.

In this paper, we present a new remapping method for multi-material quantities in the staggered discretization. The remapping algorithm is based on the computation of the exchange integrals between the Lagrangian and rezoned meshes, which represent fluxes of the basic geometry integrals through the computational cell boundaries, and are computed using intersections (overlays) of the original and rezoned meshes. These exchange integrals can be pre-computed at the beginning of the remapping step, and fluxes of all quantities are composed from these integrals. Due to the flux form of the remapper, this method is best suitable for continuous remap, where the original and rezoned meshes are similar.

This paper is organized as follows. In Section 2, we describe the construction of the exchange integrals. In the following two Sections 3 and 4, we describe the construction of material/average quantity fluxes and remapping all cell-centered and nodal fluid quantities. In Section 5, we demonstrate the properties of the method on a selected numerical example. The whole paper is concluded in Section 6.

2 Construction of Exchange Integrals

Our approach is based on expressing the standard overlay formula in the equivalent flux form [10],

$$\tilde{c} = c \cup \left(\bigcup_{c' \in C'(c)} c' \cap \tilde{c} \right) \setminus \left(\bigcup_{\tilde{c}' \in C'(\tilde{c})} c \cap \tilde{c}' \right) , \tag{1}$$

where ˜ denotes a particular cell in the new mesh, and $C'(c)$ is the set of all cells neighboring with c (including the corner neighbors). For the construction of the intersection polygons, we intersect the original cell with the halfplanes defined by the edges of the rezoned cell [11]. This robust approach works well for intersection of the generally non-convex polygons (Lagrangian cells) with the convex polygons (rezoned cells). The situation is demonstrated in Fig. 1. The first term in parentheses represents the outward part of the flux, while the second term is the inward part of the flux. Both inward and outward parts can be seen as two light triangles in image (c) of the Figure. The same expression can be written for each pure material polygon of cell c. An example of original material polygons is shown in images (a) and (b) of the Figure, the fluxes of different materials are shown in different shades in images (d-f) of the Figure. As we can see, the flux between c and a particular neighbor can have non-zero values of both inward and outward components of the flux, and each component can include fluxes of several materials (including the corner fluxes).

Now, suppose that we want to remap volume of a particular materials k of cell c to the new cell $\tilde{c}$. The new material volume can be written as

$$V_{\tilde{c},k} = \int\limits_{\tilde{c}_k} 1\, dV\,, \tag{2}$$

and after employing formula (1), we can rewrite it as

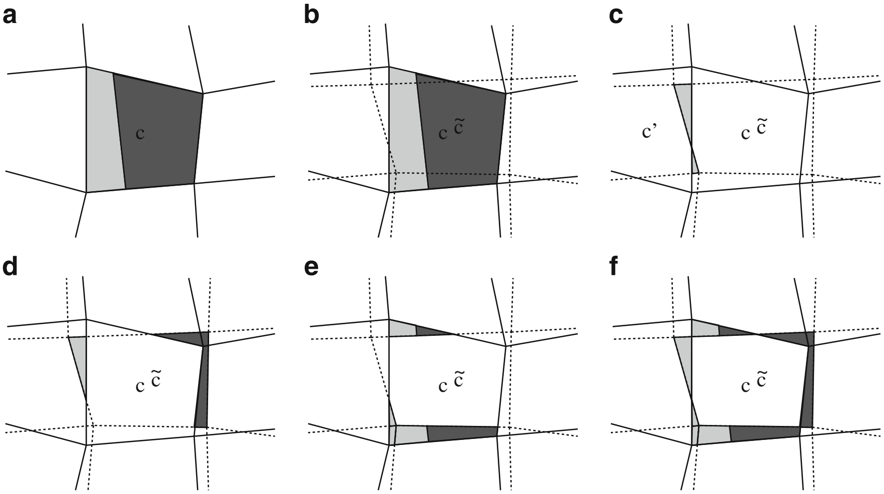


Fig. 1 (a) Original cell c (solid line) containing light and dark materials. (b) New cell $\tilde{c}$ (dashed line). (c) Fluxes between cell c and its left neighbor c'. (d) All outward fluxes around c. (e) All inward fluxes around c. (f) All fluxes around c

$$V_{\tilde{c},k} = V_{c,k} + \sum_{c' \in C'(c)} F^V_{c,c',k} \,, \tag{3}$$

where the material volume fluxes are defined as

$$F^V_{c,c',k} = I^1_{c'_k \cap \tilde{c}} - I^1_{c_k \cap \tilde{c}'} \,, \qquad I^f_P = \int_P f \, dV \,. \tag{4}$$

Here c_k is the polygon of pure material k in cell c, and the total material volume flux has its outward and inward components. The exchange integral of function f over polygon P is denoted by the I^f_P symbol. The exchange integrals can be pre-computed at the beginning of the remapping step from the mesh geometry, and can be used for the construction of fluxes of all fluid quantities. Later, we will need the exchange integrals for polynomials up to the second order, i.e. $f = 1, x, y, x^2, x\,y, y^2$. Let us note that these are all integrals of polynomials over polygons ($c'_k \cap \tilde{c}$ or $c_k \cap \tilde{c}'$), which can be evaluated analytically.

3 Remap of Cell-Centered Quantities

In this Section, we demonstrate the remap of the cell-centered and material-centered quantities. All material quantities will be remapped in the same form as we have shown in equation (3) for material volumes. Material centroids $x_{c,k}$ (needed as reference centroids for MOF) are remapped as

$$x_{\tilde{c},k} \, V_{\tilde{c},k} = x_{c,k} \, V_{c,k} + \sum_{c' \in C'(c)} F^x_{c,c',k} \,, \qquad F^x_{c,c',k} = I^x_{c'_k \cap \tilde{c}} - I^x_{c_k \cap \tilde{c}'} \tag{5}$$

and similarly for $y_{c,k}$.

Material density is reconstructed in a piece-wise linear way

$$\rho_{c,k}(x, y) = \rho_{c,k} + S^x_{c,k} \, (x - x_{c,k}) + S^y_{c,k} \, (y - y_{c,k}) \,, \tag{6}$$

where the material density mean values $\rho_{c,k}$ are known, and the slopes $S^{x,y}_{c,k}$ are determined by minimization of the discrepancy between the reconstructed values in the centroids of the same material polygons in the neighboring cells from the mean values there [12]. Limiting by the Barth-Jespersen limiter [13] guarantees preservation of the local density extrema, while in the mixed cells the 0 and $+\infty$ limits are used for limiting to avoid excessive slope degradation in case of thin material filaments with only few neighbors containing the same material. This approach implies second order of accuracy of the remapper. The material mass remap is then performed in a similar form,

$$m_{\tilde{c},k} = m_{c,k} + \sum_{c' \in C'(c)} F^m_{c,c',k} \,, \qquad F^m_{c,c',k} = F^m_{c'_k \cap \tilde{c}} - F^m_{c_k \cap \tilde{c}'} \,, \tag{7}$$

where the inward and outward mass fluxes are obtained by the integration of the reconstructed density over the intersection, which can be composed from the precomputed exchange integrals, for example

$$F^m_{c_k \cap \tilde{c}'} = \int\limits_{c_k \cap \tilde{c}'} \rho_{c,k}(x,y)\, dV = \left(\rho_{c,k} - S^x_{c,k}\, x_{c,k} - S^y_{c,k}\, y_{c,k}\right) I^1_{c_k \cap \tilde{c}'} + S^x_{c,k}\, I^x_{c_k \cap \tilde{c}'} + S^y_{c,k}\, I^y_{c_k \cap \tilde{c}'} \,. \tag{8}$$

For the material internal energy, same approach

$$\varepsilon_{\tilde{c},k}\, m_{\tilde{c},k} = \varepsilon_{c,k}\, m_{c,k} + \sum_{c' \in C'(c)} F^\varepsilon_{c,c',k} \,, \qquad F^\varepsilon_{c,c',k} = F^\varepsilon_{c'_k \cap \tilde{c}} - F^\varepsilon_{c_k \cap \tilde{c}'} \tag{9}$$

and (8) with the material specific internal energy ε instead of the density ρ can be used. However, this approach does not guarantee satisfaction of the local-bound conservation condition, so a more advanced approach described in [14] must be used, which constructs the energy fluxes by integration of the reconstructed density multiplied by the reconstructed specific internal energy, for example

$$F^\varepsilon_{c_k \cap \tilde{c}'} = \int\limits_{c_k \cap \tilde{c}'} \rho_{c,k}(x,y)\, \varepsilon_{c,k}(x,y)\, dV \,. \tag{10}$$

The reconstruction of the specific internal energy cannot be done the same way as we did for density in (6), it must be centered in material centers of mass $x^m_{c,k} = (\int_{c_k} \rho_{c,k}(x,y)\, x\, dV)/m_{c,k}$ instead of material centroids,

$$\varepsilon_{c,k}(x,y) = \varepsilon_{c,k} + S^{x,\varepsilon}_{c,k}\left(x - x^m_{c,k}\right) + S^{y,\varepsilon}_{c,k}\left(y - y^m_{c,k}\right) \,. \tag{11}$$

Both centers of mass and energy fluxes (10) can be composed from the precomputed exchange integrals as we did for mass (8), however, integrals of the second order polynomials are needed now.

The last cell-centered quantity we need to remap is the average cell pressure needed for the next Lagrangian step (the material pressures are updated from the remapped material energies using the equation of state). We suggest to remap the average pressure in the following form

$$p_{\tilde{c}}\, V_{\tilde{c}} = p_c\, V_c + \sum_{c' \in C'(c)} F^p_{c,c'} \,, \qquad F^p_{c,c'} = F^p_{c' \cap \tilde{c}} - F^p_{c \cap \tilde{c}'} \,, \tag{12}$$

where the pressure fluxes are obtained as the exchange volumes multiplied by the pressure reconstructed by same formula as (6) in the centroid of the intersection polygon, for example

$$F^{p}_{c\cap\tilde{c}'} = I^{q}_{c\cap\tilde{c}'}\, p_c(x_{c\cap\tilde{c}'}, y_{c\cap\tilde{c}'})\,, \qquad \{x,y\}_{c\cap\tilde{c}'} = \frac{I^{\{x,y\}}_{c\cap\tilde{c}'}}{I^{1}_{c\cap\tilde{c}'}}\,. \tag{13}$$

All terms here can be composed from the pre-computed exchange integrals again.

4 Remap of Nodal Quantities

Nodal mass is tied with the total cell mass through the sub-zonal masses [15]. Our approach is to remap the nodal mass in a similar flux form,

$$m_{\tilde{n}} = m_n + \sum_{n'\in N'(n)} F^{m}_{n,n'} \tag{14}$$

where $N'(n)$ is the set of nodes neighboring with n and $F^{m}_{n,n'}$ are the inter-nodal mass fluxes, which can be defined either by interpolation from the inter-cell fluxes [16], or by minimizing of their difference from given reference fluxes [17]. All remaining nodal quantities are remapped in the same form as (14) by attaching the particular nodal quantity to the inter-nodal mass flux, for example

$$u_{\tilde{n}}\, m_{\tilde{n}} = u_n\, m_n + \sum_{n'\in N'(n)} u_{n,n'}\, F^{m}_{n,n'} \tag{15}$$

for nodal velocity, where the value of $u_{n,n'}$ is the reconstructed velocity inside the inter-nodal swept region, for example a simple bilinear interpolation or the kinetic-energy conservative approach [18]. Similarly, the kinetic energy can be remapped in order to perform the standard energy fix [2] ensuring total energy conservation.

5 Numerical Example

To demonstrate the properties of the described remapping method, we present simulation of the triple point problem described in [19]. The initial data are shown in image (a) of Fig. 2. This problem contains three materials, the interfaces are initially aligned with the mesh edges. The simulation was performed in the context of our 2D staggered research multi-material ALE (RMALE) code on the orthogonal 140 × 60 mesh. To stress the influence of the remapper, we run the simulation in the Eulerian manner – remapping to the initial mesh is done after every single Lagrangian step. The light material generates a shock wave propagating in different speeds into the gray and dark materials, causing development of a vertex.

The material distribution in the final time $t = 5$ can be seen in image (b) of Fig. 2. We can see the thin filament of the dark material which stays compact and does not break apart even though the width of its tail is smaller than 1 cell size. The profiles of material density, specific internal energy, and pressure can be seen in images (c-e) of Fig. 2. As we can see, all fields are smooth without any numerical problems. Finally, in image (f) of Fig. 2, the material velocity field is shown displaying the vortex around the triple point. Again, the velocity field is smooth and does not contain any numerical artifacts.

6 Conclusion

We have briefly described a new method for remapping of all fluid quantities between similar meshes in the context of staggered multi-material 2D ALE. This method is flux based, and fluxes of all fluid quantities are constructed from the pre-computed exchange integrals. As these integrals are computed just once, at the beginning of the remapping step, computational cost of this method is not excessive

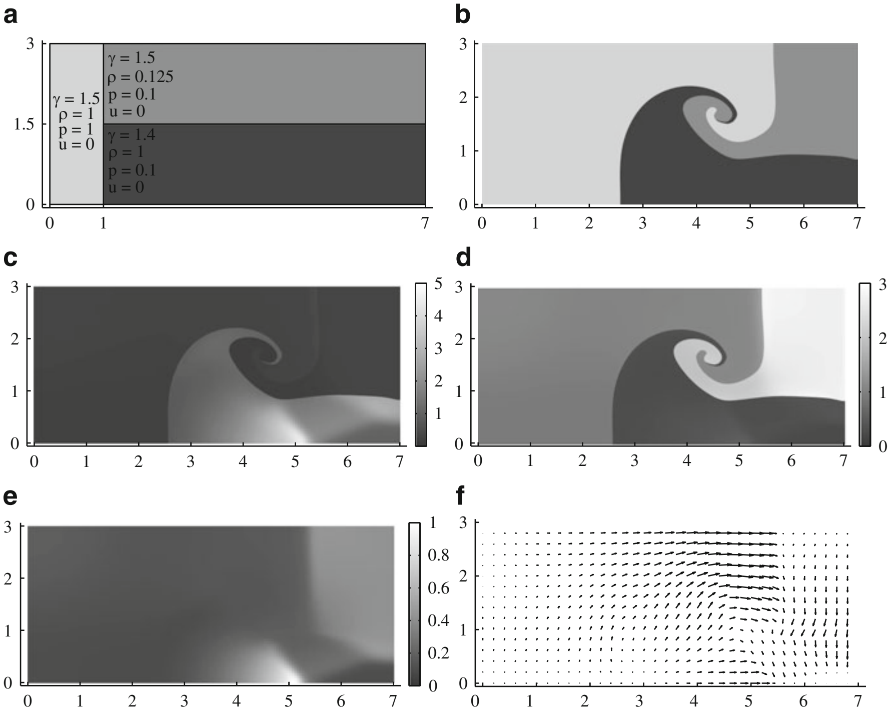


Fig. 2 Triple point problem: (a) materials in $t = 0$; (b) materials in $t = 5$; (c) material density in $t = 5$; (d) material energy in $t = 5$; (e) material pressure in $t = 5$; (f) velocity field in $t = 5$

although it involves intersections. Due to the flux nature of the method, this method is conservative for all quantities (total energy conservation is assured by the energy fix). If high order reconstructions are used for fluxes of all quantities, this method is second order accurate. We have demonstrated that this method can be used as a remapper in the framework of a full hydrodynamic code.

Acknowledgements 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 and supported by the DOE Advanced Simulation and Computing (ASC) program. The authors acknowledge the partial support of the DOE Office of Science ASCR Program. Milan Kucharik was supported by the Czech Ministry of Education grants MSM 6840770022, MSM 6840770010, and LC528, and the Czech Science Foundation project P201/10/P086. The authors thank H. Ahn, D. Bailey, A. Barlow, K. Lipnikov, R. Loubere, P.-H. Maire, and M. Owen for fruitful and stimulating discussions over many years.

References

1. C. W. Hirt, A. A. Amsden, and J. L. Cook. An arbitrary Lagrangian-Eulerian computing method for all flow speeds. *Journal of Computational Physics*, 14(3):227–253, 1974.
2. D. J. Benson. Computational methods in Lagrangian and Eulerian hydrocodes. *Computer Methods in Applied Mechanics and Engineering*, 99(2-3):235–394, 1992.
3. L. G. Margolin. Introduction to ”An arbitrary Lagrangian-Eulerian computing method for all flow speeds”. *Journal of Computational Physics*, 135(2):198–202, 1997.
4. J. S. Peery and D. E. Carroll. Multi-material ALE methods in unstructured grids. *Computer Methods in Applied Mechanics and Engineering*, 187(3-4):591–619, 2000.
5. R. W. Anderson, N. S. Elliott, and R. B. Pember. An arbitrary Lagrangian-Eulerian method with adaptive mesh refinement for the solution of the Euler equations. *Journal of Computational Physics*, 199(2):598–617, 2004.
6. R. Loubere, P.-H. Maire, M. Shashkov, J. Breil, and S. Galera. ReALE: A reconnection-based arbitrary-LagrangianEulerian method. *Journal of Computational Physics*, 229(12):4724–4761, 2010.
7. V. Dyadechko and M. Shashkov. Reconstruction of multi-material interfaces from moment data. *Journal of Computational Physics*, 227(11):5361–5384, 2008.
8. M. Kucharik, R.V. Garimella, S.P. Schofield, and M.J. Shashkov. A comparative study of interface reconstruction methods for multi-material ALE simulations. *Journal of Computational Physics*, 229(7):2432–2452, 2010.
9. M. Kucharik, M. Shashkov, and B. Wendroff. An efficient linearity-and-bound-preserving remapping method. *Journal of Computational Physics*, 188(2):462–471, 2003.
10. L. G. Margolin and M. Shashkov. Second-order sign-preserving conservative interpolation (remapping) on general grids. *Journal of Computational Physics*, 184(1):266–298, 2003.
11. M. Kucharik and M. Shashkov. Conservative multi-material remap for staggered discretization. 2011. In prep.
12. D. J. Mavriplis. Revisiting the least-squares procedure for gradient reconstruction on unstructured meshes. In *AIAA 2003-3986*, 2003. 16th AIAA Computational Fluid Dynamics Conference, June 23-26, Orlando, Florida.
13. T. J. Barth. Numerical methods for gasdynamic systems on unstructured meshes. In C. Rohde D. Kroner, M. Ohlberger, editor, *An introduction to Recent Developments in Theory and Numerics for Conservation Laws, Proceedings of the International School on Theory and Numerics for Conservation Laws*, Berlin, 1997. Lecture Notes in Computational Science and Engineering, Springer. ISBN 3-540-65081-4.

14. J. K. Dukowicz and J. R. Baumgardner. Incremental remapping as a transport/advection algorithm. *Journal of Computational Physics*, 160(1):318–335, 2000.
15. R. Loubere and M. Shashkov. A subcell remapping method on staggered polygonal grids for arbitrary-Lagrangian-Eulerian methods. *Journal of Computational Physics*, 209(1):105–138, 2005.
16. R. B. Pember and R. W. Anderson. A comparison of staggered-mesh Lagrange plus remap and cell-centered direct Eulerian Godunov schemes for Eulerian shock hydrodynamics. Technical report, LLNL, 2000. UCRL-JC-139820.
17. J. M. Owen and M. J. Shashkov. Arbitrary Lagrangian Eulerian remap treatments consistent with corner based compatible total energy conserving Lagrangian methods. In prep., 2010.
18. D. Bailey, M. Berndt, M. Kucharik, and M. Shashkov. Reduced-dissipation remapping of velocity in staggered arbitrary Lagrangian-Eulerian methods. *Journal of Computational and Applied Mathematics*, 233(12):3148–3156, 2010.
19. S. Galera, J. Breil, and P.-H. Maire. A 2D unstructured multi-material cell-centered arbitrary lagrangianeulerian (CCALE) scheme using MOF interface reconstruction. *Computers & Fluids*, 2010. In press. doi:10.1016/j.compfluid.2010.09.038.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Optimized Riemann Solver to Compute the Drift-Flux Model

Anela Kumbaro and Michaël Ndjinga

Abstract This paper discusses the development of an approximated optimized Riemann solver applied to the two-phase flow drift-flux model. The solver makes use of a partial eigenstructure information while maintaining the Roe solver accuracy. Moreover, it allows to take into account the contribution of the dynamic and thermal non-equilibrium in the upwinding matrix. A further optimization of the solver is realized by scaling the global matrix which results in better preconditioning. Both the partial eigenstructure decomposition and the scaling of the matrix are inspired from the eigenstructure of the two-phase flow model. A number of physical benchmarks are presented to illustrate this method. Comparison between the computational results obtained with the optimized solver and the conventional Roe-type solver demonstrates the efficiency of the new methodology.

Keywords Riemann solver, eigenstructure decomposition, drift-flux model
MSC2010: 76T10, 76M12, 35L65, 76E19, 65F35

1 Introduction

The drift-flux is commonly used to simulate water-vapor flows in nuclear power plants. Various industrial codes within the nuclear community, for example FLICA4 code of CEA, or THYC of EDF, both dedicated to design and safety studies of nuclear reactors, rely on this model. When compared with codes that use more advanced two-phase models, such as the two-fluid or the multifield model, their strong point is the code-efficiency. Reducing furthermore the CPU time cost is crucial for the survival of these type of codes. Our work is done within the FLICA-OVAP code [3], which is a new platform dedicated to core thermal-hydraulic studies,

Anela Kumbaro and Michaël Ndjinga
CEA-Saclay, DEN/DM2S/SFME/LETR, F-91191 Gif-sur-Yvette, France, e-mail: anela.kumbaro@cea.fr, michael.ndjinga@cea.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_67, © Springer-Verlag Berlin Heidelberg 2011

funded by the Thermal-hydraulics Simulation project of CEA. To provide a relevant response to different core concepts and multiple industrial applications, several models coexist in FLICA-OVAP platform: the Homogeneous Equilibrium model, the drift flux model which is directly derived from the previous CEA core code FLICA-4 [1]-[2], the two-fluid model, and finally, a general multifield model [4], with a variable number of fields for both vapor and liquid phases. We present in this paper two techniques to reduce the execution time and improve the code's performance while using the drift-flux model. Our starting point solver is the weak formulation of the Roe's approximate Riemann solver, adapted to low Mach number [6]. Based on the eigenstructure of the drift-flux model we propose to rewrite the solver in a more optimized form.

On the other hand, to go forward in time, a fully implicit integrating step is used that provides fast running steady state calculations. We introduce a scaling of the implication matrix so that the matrix coefficients have the same order of magnitude. This allows for a much better preconditioning of the matrix and significantly reduces the global CPU time.

This paper is organized as follows: to begin with, Sect. 2 briefly describes the standard two-phase flow drift-flux model we deal with. Next, we introduce an evaluation of its eigenstructure. In Sect. 4 we present the numerical solver based on the specific eigenstructure of the drift-flux model and discuss its accuracy. Section 5 introduces the scaling of the matrix. We show that the coefficients of the upwinding matrix which have different orders of magnitude, have the same magnitude after the scaling. Some numerical results are presented in Sect. 6 to illustrate the behavior of the numerical solver. Finally, some conclusions are presented in the last section.

2 Drift-flux two-phase flow model

We introduce here the FLICA-OVAP drift-flux model. For the sake of simplicity this model is represented without taking into account the porosity variable and the viscous term. The balance equations for the drift-flux model read:

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

$$\frac{\partial}{\partial t}(\rho\mathbf{u}) + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u} + \rho c(1-c)\mathbf{u}_r\otimes\mathbf{u}_r) + \nabla p = \mathbf{F}_{ext} + \mathbf{F}^w \tag{2}$$

$$\frac{\partial}{\partial t}(\rho E) + \nabla\cdot(\rho H\mathbf{u} + \rho c(1-c)\mathbf{u}_r(L + \frac{\mathbf{u}_v^2 - \mathbf{u}_l^2}{2})) = Q_{tot} + \mathbf{F}_{ext}\cdot\mathbf{u} \tag{3}$$

$$\frac{\partial}{\partial t}(\rho c) + \nabla\cdot(\rho c\mathbf{u} + \rho c(1-c)\mathbf{u}_r) = \Gamma, \tag{4}$$

where c is the vapor concentration, $\mathbf{u}$, ρ, p, E, H are the mixture velocity, mixture density, pressure, total energy and enthalpy, respectively, $\mathbf{u}_r$ is the relative velocity between vapor and liquid phases given by a drift-flux model, Γ is the mass transfer term, $\mathbf{F}_{ext}$ is the external forces term, F^w is the wall friction term, and L is the latent heat.

The model is closed by a general equation of state $\rho = \rho(p, h, c)$, and by the assumption that the vapor is saturated in presence of liquid: $h_v = h_v^{sat}(P)$, where h_g is the vapor enthalpy. Closure laws (wall transfer, mass exchange, diffusion, ...) for this model come from FLICA-4 code and have been described in [1].

3 Eigenstructure of the drift-flux model

If we introduce an orthonormal basis ($\mathbf{n}$, τ_1, τ_2) of the three dimensional space R^3, the one-dimensional formulation of the above drift-flux system (1-4) is:

$$\frac{\partial \mathbf{V}}{\partial t} + \frac{\partial \mathbf{F}_n}{\partial n} = \mathbf{S} \tag{5}$$

with the conservative vector $\mathbf{V} = (\rho, \rho\mathbf{u}, \rho E, \rho c) \in \mathbb{R}^m$, where $m = d + 3$ and d is the space dimension, and S the source term vector. The expression of the flux is separated into two part, the zero order relative velocity part, $\mathbf{F}_{n0}$, and the relative velocity dependent part, $\mathbf{F}_{nr}$:

$$\mathbf{F}_{n0} = \begin{pmatrix} \rho\mathbf{u}\cdot\mathbf{n} \\ \rho(\mathbf{u}\cdot\mathbf{n})\mathbf{u} + p\mathbf{n} \\ \rho H\mathbf{u}\cdot\mathbf{n} \\ \rho c\mathbf{u}\cdot\mathbf{n} \end{pmatrix}, \qquad \mathbf{F}_{nr} = \begin{pmatrix} 0 \\ \rho c(1-c)\mathbf{u}_r\cdot\mathbf{n}\mathbf{u}_r \\ \rho c(1-c)(L + \frac{u_v^2 - u_l^2}{2})(\mathbf{u}_r\cdot\mathbf{n}) \\ \rho c(1-c)\mathbf{u}_r\cdot\mathbf{n} \end{pmatrix} \tag{6}$$

Let first consider only the part without relative velocity contribution. The eigenvalues are $\lambda^- = \mathbf{u}\cdot\mathbf{n} - a$, $\lambda_u = \mathbf{u}\cdot\mathbf{n}$ (multiplicity $d+1$), and $\lambda^+ = \mathbf{u}\cdot\mathbf{n} + a$, where a is the mixture sound velocity. The right and left eigenvectors associated to the sound waves are

$$\mathbf{r}^\pm = \begin{bmatrix} 1 \\ \mathbf{u} - (\mathbf{u}\cdot\mathbf{n} - \lambda^\pm)\mathbf{n} \\ H - (\mathbf{u}\cdot\mathbf{n} - \lambda^\pm)\mathbf{u}\cdot\mathbf{n} \\ c \end{bmatrix} \qquad \mathbf{l}^\pm = \frac{1}{2a^2}\begin{bmatrix} \chi \mp a\mathbf{u}\cdot\mathbf{n} \\ -\kappa\mathbf{u} \pm a\mathbf{n} \\ \kappa \\ \xi \end{bmatrix} \tag{7}$$

where $\chi = \frac{\partial P}{\partial \rho}$, $\kappa = \frac{\partial P}{\partial \rho E}$, and $\xi = \frac{\partial P}{\partial \rho c}$. If we consider the relative velocity dependent part, the problem becomes very complex. Indeed, the relative velocity is drift flux model dependent, and the drift flux model depends on the flow configuration. We will not represent here any analytical expression about the eigenvalues but we will only assume that the drift-flux model, like the general multifield model [4], has two fast eigenvalues of $\mathbf{u} \pm a$ order of magnitude, while the other eigenvalues are between $\mathbf{u}_v \cdot \mathbf{n}$ and $\mathbf{u}_l \cdot \mathbf{n}$. Hence, these so called intermediate eigenvalues have the same order of magnitude as the mixture velocity.

4 Simplified eigenstructure decomposition solver (SEDES)

Let τ be a meshing of Ω defined as the union of control volumes K. The discrete unknowns are denoted by $\mathbf{V}_K^n$ and represent the approximation of a mean value of $\mathbf{V}$ on the control volume K at time t^n.

The Roe-type approximate Riemann solver is the current solver in the CEA industrial code FLICA-4 [2] and it will be used as the reference solver for this study. This solver requires the solution of a one-dimensional Riemann problem at cell interfaces on a non-staggered grid, to define backward and forward differences used to approximate the spatial derivatives. The numerical flux is the following:

$$\Phi_{K,L}^{n+1} = \frac{\mathbf{F}_n(\mathbf{V}_K^{n+1}) + \mathbf{F}_n(\mathbf{V}_L^{n+1})}{2} - \frac{|\mathbf{A}_n(\mathbf{V}_K^n, \mathbf{V}_L^n)|}{2}(\mathbf{V}_L^{n+1} - \mathbf{V}_K^{n+1}) \tag{8}$$

So, at the heart of such scheme is the so-called Roe matrix, first introduced in [5] for the single-phase flow equations, taken at the Roe average state on the interface between K and L.

To construct this matrix the FLICA-4 standard Roe-type solver makes use of a complete eigenstructure decomposition:

$$|\mathbf{A}_n(\mathbf{V}_K^n, \mathbf{V}_L^n)| = \sum_{k=1}^{m} |\lambda_k| l_k \otimes r_k \tag{9}$$

with λ_k, l_k and r_k the eigenvalues, the left and right eigenvectors of the system matrix.

First case: If the relative velocity contribution is not considered into the system matrix, the eigenvalues are $\lambda^- = \mathbf{u} \cdot \mathbf{n} - a$, $\lambda_u = \mathbf{u} \cdot \mathbf{n}$ (multiplicity d), and $\lambda^+ = \mathbf{u} \cdot \mathbf{n} + a$, and the eigenvectors are easily obtained. Therefore, we propose to rewrite Equation (9)

$$|\mathbf{A}_n| = (|\lambda^-| - |u_n|) l^- \otimes r^- + (|\lambda^+| - |u_n|) l^+ \otimes r^+ + |u_n| \mathbb{I}. \tag{10}$$

with $l^\pm$ and $r^\pm$ given by (7). Eq. (10), while corresponding exactly to the Roe upwinding matrix, provides a more efficient way to calculate this matrix.

Second case: The relative velocity contribution is considered into the system matrix. In this case Eq. (9) requires the computation of all the eigenstructure of the system matrix. This computation has to be done using a numerical algorithm and this means a consistent increase in CPU time. For this reason the FLICA-4 standard Roe-type solver does not take into account the relative velocity in the upwind part of the numerical fluxes.

On the other hand, based on the structure of the complete matrix eigenvalues, we remark that to compute the absolute value of the upwinding matrix it is essential to take into account the contributions of the fastest eigenvalues, while the remaining eigenvalues which have more or less the same order of magnitude, can be represented by an unique candidate, for instance, the fastest one of this group, that we will denote simply by $\tilde{\lambda}$.

We can rewrite Eq. (10) using $\tilde{\lambda}$ instead of $|u_n|$:

$$|\mathbf{A}_n^{SEDES}| = (|\lambda^-| - \tilde{\lambda})l^- \otimes r^- + (|\lambda^+| - \tilde{\lambda})l^+ \otimes r^+ + \tilde{\lambda}\mathbb{I}. \tag{11}$$

Eq. (11) corresponds to a simplified eigenstructure decomposition, hence the name SEDES of the solver, as it uses only the fastest waves contributions. To construct the SEDES flux we need the eigenvalues and eigenvectors associated to the sound waves, which are determined using a shifted power method, and $\tilde{\lambda}$ calculated as $\tilde{\lambda} = \mathbf{u} \cdot \mathbf{n} + |\mathbf{u}_r \cdot \mathbf{n}|$.

The spectrum of the approximated upwind matrix $|\mathbf{A}_n^{SEDES}|$ is very close to the spectrum of the standard complete Roe decomposition (9), since the spectrum of the following matrix is close to zero

$$|\mathbf{A}_n^{SEDES}| - |\mathbf{A}_n^{Roe}| = \sum_{k=2}^{m-1} l_k \otimes r_k (\tilde{\lambda} - |\lambda_k|). \tag{12}$$

The general structure of the drift-flux system eigenvalues ensures that $\tilde{\lambda} - |\lambda_k| \leq |\mathbf{u}_r \cdot \mathbf{n}|$, so the truncation error in upwinding matrix remains small, especially when compared with the two first terms on the right hand side of Eq. (11), as the velocities are small compared to the sound speeds.

5 Matrix scaling for better preconditioning

We are interested in using a fully implicit method for transient calculations. To this end we rewrite the numerical flux (8) that gives its contribution on the right hand side of the disretized system, in either of the two equivalent forms:

$$\Phi_{K,L}^{n+1} = \mathbf{F}_n(\mathbf{V}_K^{n+1}) + \mathbf{A}_n^-(\mathbf{V}_K^n, \mathbf{V}_L^n)(\mathbf{V}_L^{n+1} - \mathbf{V}_K^{n+1}) \tag{13}$$

or

$$\Phi_{K,L}^{n+1} = \mathbf{F}_n(\mathbf{V}_L^{n+1}) - \mathbf{A}_n^+(\mathbf{V}_K^n, \mathbf{V}_L^n)(\mathbf{V}_L^{n+1} - \mathbf{V}_K^{n+1}) \tag{14}$$

where $\mathbf{A}_n^{\pm}(\mathbf{V}_K^n, \mathbf{V}_L^n)$ are the negative/positive part of the upwind matrix. The derivatives of these fluxes give contribution on the implicitation matrix that depends only on the matrices $\mathbf{A}_n^{\pm}(\mathbf{V}_K^n, \mathbf{V}_L^n)$. We remark that both vapor and liquid velocity projections on the normal at the cells interface have the same sign in most kind of two-phase flow configurations, as the relative velocity is smaller compared with the mixture velocity. Hence, we expect the eigenvalues of the two-phase flow system to have rather the same sign, except one. We choose in the code to compute first whichever matrix $\mathbf{A}_n^{+}(\mathbf{V}_K^n, \mathbf{V}_L^n)$ or $\mathbf{A}_n^{-}(\mathbf{V}_K^n, \mathbf{V}_L^n)$ corresponding to a minimal number of eigenvalues of the same sign. Then, we obtain the other one using the relation $\mathbf{A}_n^{+}(\mathbf{V}_K^n, \mathbf{V}_L^n) - \mathbf{A}_n^{-}(\mathbf{V}_K^n, \mathbf{V}_L^n) = |\mathbf{A}_n(\mathbf{V}_K^n, \mathbf{V}_L^n)|$, with the absolute value matrix obtained using the partial eigenstructure decomposition method explained in Sect. 5.

The implicit numerical method finally leads to the solving of the system

$$MX = b. \tag{15}$$

n In order to solve efficiently (15) using an iterative solver [7], one needs to find an approximation of M^{-1}. This is usually done through an approximate factorization $M \approx LU$ where U is upper triangular and L is lower triangular. The error made in the approximate factorisation using an incomplete Gauss factorisation depends on the size of off-diagonal coefficients of the matrix. Hence one may benefit from working with matrix having off-diagonal coefficient of smallest possible magnitude.

When looking at the coefficients of the system eigenvectors (7), one sees that they have very different magnitudes. Indeed in the particular case where $\mathbf{u} = 0$, the Roe matrix has only two non zero eigenvalues, $\pm a$, with the respective eigenvectors

$$\mathbf{r}^{\pm} = \left[\, 1, \pm a\mathbf{n}, h, c \,\right] \qquad \mathbf{l}^{\pm} = \tfrac{1}{2a^2}\left[\, \chi, \pm a\mathbf{n}, \kappa, \xi \,\right] \tag{16}$$

For better readability, the rest of the analysis is presented in the 1D case ($\mathbf{n} = 1$) and derivatives χ and ξ having the same order as κh will be replaced by κh. One has $A^{\pm} = \pm a\,(\mathbf{l}_{\pm} \otimes \mathbf{r}_{\pm})$:

$$A^{\pm} = \frac{1}{2a^2}\begin{pmatrix} h\kappa & \pm ah\kappa & h^2\kappa & ch\kappa \\ \pm a & a^2 & \pm ah & \pm ac \\ \kappa & \pm a\kappa & h\kappa & c\kappa \\ h\kappa & \pm ah\kappa & h\kappa & ch\kappa \end{pmatrix}. \tag{17}$$

We remark that $h\kappa$ has the same order of magnitude as a^2. One can see that the disequilibrium in $A^{\pm}$ coefficients comes from the difference in magnitude of the left and right eigenvectors of A. Multiplying $A^{\pm}$ to the left (respectively to the right) by a diagonal matrix with coefficients $d_{scale} = diag(1, a, \frac{a^2}{\kappa}, 1))$ (respectively $d_{scale}^{-1} = diag(1, \frac{1}{a}, \frac{\kappa}{a^2}, 1))$ one obtains a new matrix with better balanced coefficients

$$\tilde{A} = d_{scale} A^{\pm} d_{scale}^{-1} = \frac{1}{2a^2} \begin{pmatrix} h\kappa & \pm h\kappa & \frac{h^2\kappa^2}{a^2} & ch\kappa \\ \pm a^2 & a^2 & \pm h\kappa & \pm a^2 c \\ a^2 & \pm a^2 & h\kappa & a^2 c \\ h\kappa & \pm h\kappa & \frac{h^2\kappa^2}{a^2} & ch\kappa \end{pmatrix} \sim \frac{1}{2} \begin{pmatrix} 1 & \pm 1 & 1 & c \\ \pm 1 & 1 & \pm 1 & \pm c \\ 1 & \pm 1 & 1 & c \\ 1 & \pm 1 & 1 & c \end{pmatrix} \tag{18}$$

We propose to build two diagonal matrices D_{scale} and D_{scale}^{-1} having the size of the mesh and containing the successive coefficients of the local matrices d_{scale} and d_{scale}^{-1}. Instead of solving system (15) it is equivalent to solve

$$\tilde{M} Y = \tilde{b} \tag{19}$$

where $\tilde{M} = D_{scale} M D_{scale}^{-1}$, $Y = D_{scale} X$ and $\tilde{b} = D_{scale} b$. System (19) can be more easily resolved using an ILU preconditioner. Once the solution Y is obtained we compute $D_{scale}^{-1} Y$ to obtain the original unknown vector X.

6 Numerical investigation

We present here three applications to demonstrate the overall efficiency of the new solver. All the simulations are realized without taking into account the relative velocity into the upwinding matrix. In this case the SEDES solver gives the identical results with Roe solver and we can concentrate our attention only to the solver efficiency. The first two test cases correspond to 1D configurations. The first test case is a boiling flow in a 1D heated channel and the second one corresponds to a flow in a PWR reactor core. We have used a 50-cells mesh and a CFL number of 30 and 833, respectively. Table 1 represents the dimensionless CPU time for the simplified eigenstructure decomposition solver (SEDES) and the old solver (Roe solver). The last test corresponds to a steady state computation of a full charge 3D PWR reactor core configuration. The simulation is run using a 157 assemblies and 32 cells in the axial direction. The CFL number is equal to 2000. We have realized two runs with and without the scaling using the SEDES solver and compared the CPU time with the standard Roe solver time results. For this simulation we use the standard steady state algorithm of FLICA-OVAP code such as described in [2] which saves the matrix of the linear system at the first time step and uses it for the whole steady-state calculation. Nevertheless, the scaling decreases both the number

Table 1 Dimensionless CPU time: 1D test cases

CPU Time	Boiling	PWR
SEDES	0.63	0.55
Roe	1	1

Table 2 Dimensionless CPU time: 3D PWR

Solver	CPU Time
SEDES without scaling	0.89
SEDES with scaling	0.70
Roe	1

of iterations and the cost of an iteration during the resolution of the non linear system and the new solver is still more efficient as the old one as shown by the results represented in Table 2. In this last case, the result is nearly the same when relative velocity is taken into account in the upwind matrix, as the relative velocity is smaller than mixture velocity which is about 4m/s and both, are much smaller than the sound velocity.

7 Conclusions

This paper has presented how a simplified account for the system's eigenvalues can be considered in order to build a more efficient Riemann solver for the resolution of two-phase flow drift-flux model considered in industrial codes to assess the safety of nuclear plants or to support research on two phase thermal-hydraulics which conserves the accuracy of a Roe solver. Moreover, a procedure is presented to scale the coefficient of the upwinding matrix in order to obtain a better preconditioning, which is particularly efficient in complex geometry. Various test cases have shown that the methodology greatly improves the code efficiency during the simulation of two-phase flows with realistic state equations in mono-dimensional and multi-dimensional settings.

References

1. E. Royer, S. Aniel, A. Bergeron, P. Fillion, D. Gallo, F. Gaudier, O. Grgoire, M. Martin, E. Richebois, P. Salvadore, S. Zimmer, T. Chataing, P. Clment, and F. Franois, FLICA4: Status of numerical and physical models and overview of applications, in Proceedings of NURETH-11, (Avignon, France), October 2-6 (2005)
2. I. Toumi, A. Bergeron, D. Gallo, E. Royer, and D. Caruge, FLICA-4: a three-dimensional two-phase flow computer code with advanced numerical methods for nuclear application, Nucl. Eng. Design, vol. 200, pp. 139-155 (2000)
3. Fillion P, Chanoine A, Dellacherie S, Kumbaro A, FLICA-OVAP: a New Platform for Core Thermal-hydraulic Studies, NURETH-13, Japan, September 27-October 2, (2009)
4. Kumbaro A., Application of the Simplified Eigenstructure Decomposition Solver to the Simulation of General Multifield Models, 7th International Topical Meeting on Nuclear Reactor Thermal Hydraulics, Operation and Safety, Seoul, Korea, October 5-9, (2008)

5. Roe P.L., Approximate Riemann solvers, parameter vectors, and difference schemes, J. Comp. Phys. 43(2) (1981)357-372.
6. Toumi I, A weak formulation of Roe's approximate Riemann solver , *J. Comput. Phys.*, 102 (1992) 360-373.
7. Michele Benzi, Preconditioning Techniques for Large Linear Systems: A Survey *J. Comput. Phys.*, 182 (2002), 418-477.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Finite Volume Schemes for Solving Nonlinear Partial Differential Equations in Financial Mathematics

Pavol Kútik and Karol Mikula

Abstract In order to estimate a fair value of financial derivatives, various generalizations of the classical linear Black–Scholes parabolic equation have been made by adjusting the constant volatility to be a function of the option price itself. We present a second order numerical scheme, based on the finite volume method discretization, for solving the so–called Gamma equation of the Risk Adjusted Pricing Methodology (RAPM) model. Our new approach is based on combination of the fully implicit and explicit schemes where we solve the system of nonlinear equations by iterative application of the semi–implicit approach. Presented numerical experiments show its second order accuracy for the RAPM model as well as for the test with exact Barenblatt solution of the porous–medium equation which has a similar character as the Gamma equation.

Keywords finite volume method, second-order scheme, financial mathematics
MSC2010: 65A05, 35K20, 35K55, 65M08

1 Motivation from Financial Mathematics

Black–Scholes linear model In 1973 Black and Scholes in [4] and independently Merton in [9] derived a simple model for pricing financial derivatives based on the solution of a linear PDE. To obtain the governing equation, they assumed that the underlying asset S follows a geometric Brownian motion $dS = (\mu - q)Sdt + \hat{\sigma}SdW$, where $\mu > 0$ is a constant drift, $\hat{\sigma} > 0$ is a constant volatility, $q > 0$ is a dividend yield rate and W is a standard Wiener process. Denoting the price of an option as $V(S,t)$ and applying It$\overline{o}$'s lemma to obtain the stochastic differential dV,

Pavol Kútik and Karol Mikula
Department of Mathematics, Radlinského 11, 813 68, Bratislava, Slovak University of Technology, e-mail: kutik@math.sk,mikula@math.sk

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_68, © Springer-Verlag Berlin Heidelberg 2011

the equation takes the following form [7]:

$$\frac{\partial V}{\partial t} + \frac{\hat{\sigma}^2}{2} S^2 \frac{\partial^2 V}{\partial S^2} + (r-q) S \frac{\partial V}{\partial S} - rV = 0, \tag{1}$$

where r represents the riskless interest rate. In the case of an European call option the terminal pay–off condition in time $t = T$ for the strike price E looks as follows:

$$V(S,T) = \max(S-E,0). \tag{2}$$

For plain vanilla options, an exact solution to (1)–(2) is known (see [7]).

Nonlinear extensions If we assume the volatility parameter to be non-constant, it can be defined by a function $\sigma = \sigma(\partial_S^2 V, S, T-t)$, where $\partial_S^2 V$ is the so–called Γ of an option. In financial theory and practice various nonlinear generalizations of Black–Scholes linear model exist with such defined volatility function. For instance, Leland in [8] proposed a model which takes transaction costs into account. Avellaneda et al. in [1] described option pricing in incomplete markets. Barles and Soner in [3] adjusted the volatility depending on investor's preferences. Illiquid market effects were studied by Schönbucher and Wilmott in [11]. Another model which we deal with in this paper is the so–called **Risk Adjusted Pricing Methodology (RAPM) model** derived by Kratka in [6] and further generalized by Jandačka and Ševčovič in [5]. Notice that the numerical scheme presented in the next section can be applied to all the above mentioned models since they can be represented by a PDE in the general form (5). Interestingly, the nonlinear porous–medium equation (13) which we deal with in the last section is also a special case of the Gamma equation (5).

The RAPM model assumes that the portfolio is rehedged only at discrete times, since continuous rehedging would lead to infinite costs. The more often the portfolio is being rehedged, the higher the risk associated with transaction costs becomes. On the other hand, seldom rehedging implies higher risk arising from its weak protection against the movement of the assets's price. Hence, there exists an optimal time step, representing the hedge interval, for which the sum of both risks is minimal. Using such ideas, the governing PDE in the following form is obtained [5]:

$$\frac{\partial V}{\partial t} + \frac{1}{2}\hat{\sigma}^2 S^2 \left[1 + \mu \left(S \frac{\partial^2 V}{\partial S^2} \right)^{\frac{1}{3}} \right] \frac{\partial^2 V}{\partial S^2} + (r-q) S \frac{\partial V}{\partial S} - rV = 0, \tag{3}$$

where $\mu = 3\left(\frac{C^2 R}{2\pi}\right)^{\frac{1}{3}}$, $C \geq 0$ represents the relative transaction costs for buying or selling one stock and $R \geq 0$ is the marginal value of investor's exposure to risk. Since $1 + \mu\,(S\Gamma)^{\frac{1}{3}} \geq 1$, the option price computed by this equation is slightly above that from the linear Black–Scholes model, i.e. we obtain a so–called Ask price. On the contrary, if $1 - \mu\,(S\Gamma)^{\frac{1}{3}} \leq 1$, then we get the lower Bid price of an option.

Gamma equation Let us define a function $\beta(H) := \frac{1}{2}\hat{\sigma}^2\left(1+\mu H^{\frac{1}{3}}\right)H$. Since the equation (3) contains the term $S\Gamma$ we introduce the following transformation:

$$H(x,\tau) = S\Gamma = S\partial_S^2 V(S,t), \tag{4}$$

where $x = \ln\left(\frac{S}{E}\right)$, $x \in R$ and $\tau = T - t$, $\tau \in (0,T)$. Moreover, if we take the second derivative of equation (3) with respect to x it turns out that the function $H(x,\tau)$ is a solution to the following nonlinear PDE, the so–called **Gamma equation** [12]:

$$\frac{\partial H(x,\tau)}{\partial \tau} = \frac{\partial^2 \beta(H)}{\partial x^2} + \frac{\partial \beta(H)}{\partial x} + (r-q)\frac{\partial H(x,\tau)}{\partial x} - qH(x,\tau). \tag{5}$$

Notice that unlike in equation (3), all terms containing spatial derivatives in the Gamma equation (5) are in divergent form, thus it is suitable to use finite volume method discretization which follows. Furthermore, since $\partial_S^2 V$ tends asymptotically to zero as $S \to 0$, respectively $S \to \infty$, from (4) it follows that the transformed Dirichlet boundary conditions are $H(-\infty,\tau) = H(\infty,\tau) = 0$.

2 Finite Volume Approximation Schemes

The most general form of the Gamma equation is as follows:

$$\frac{\partial H(x,\tau)}{\partial \tau} = \frac{\partial^2 \beta(H,x,\tau)}{\partial x^2} + \frac{\partial \beta(H,x,\tau)}{\partial x} + f(x)\frac{\partial H(x,\tau)}{\partial x} + g(x)H(x,\tau), \tag{6}$$

Notice that

$$\frac{\partial^2 \beta(H(x,\tau),x,\tau)}{\partial x^2} = \frac{\partial}{\partial x}\left(\beta'_H(H,x,\tau)\frac{\partial H(x,\tau)}{\partial x} + \beta'_x(H,x,\tau)\right), \tag{7}$$

where $\beta'_H(H,x,\tau)$ and $\beta'_x(H,x,\tau)$ are partial derivatives of the function $\beta(H(x,\tau),x,\tau)$ by H and x, respectively. Moreover,

$$f(x)\frac{\partial H(x,\tau)}{\partial x} = \frac{\partial}{\partial x}(f(x)H(x,\tau)) - H(x,\tau)f'_x(x). \tag{8}$$

Inserting (7) and (8) into (6) and integrating over the finite volume $\left(x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}\right)$, with center point denoted by x_i, we get

$$\int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial H}{\partial \tau}\, dx = \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \frac{\partial}{\partial x} \left(\beta'_H \frac{\partial H}{\partial x} + \beta'_x + \beta + f(x)H \right) dx$$
$$+ \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \left(g(x) - f'_x(x)\right) H\, dx. \tag{9}$$

Using central spatial differences, Newton–Leibniz formula and notations $\beta^\star_{i+\frac{1}{2}} = \beta(H^\star_{i+\frac{1}{2}}, x_{i+\frac{1}{2}}, \tau_\star)$, $\beta'^{\star}_{x\,i+\frac{1}{2}} = \beta'_x(H^\star_{i+\frac{1}{2}}, x_{i+\frac{1}{2}}, \tau_\star)$, $\beta'^{\star}_{H\,i+\frac{1}{2}} = \beta'_H(H^\star_{i+\frac{1}{2}}, x_{i+\frac{1}{2}}, \tau_\star)$,

we obtain the following *general numerical scheme* for solving (6):

$$h\, \frac{H_i^{j+1} - H_i^j}{k} = \beta'^{\star}_{H\,i+\frac{1}{2}} \frac{H^\star_{i+1} - H^\star_i}{h} - \beta'^{\star}_{H\,i-\frac{1}{2}} \frac{H^\star_i - H^\star_{i-1}}{h} + \beta'^{\star}_{x\,i+\frac{1}{2}} - \beta'^{\star}_{x\,i-\frac{1}{2}} + \beta^\star_{i+\frac{1}{2}}$$
$$-\beta^\star_{i-\frac{1}{2}} + f\left(x_{i+\frac{1}{2}}\right) \frac{H^\star_{i+1} + H^\star_i}{2} - f\left(x_{i-\frac{1}{2}}\right) \frac{H^\star_i + H^\star_{i-1}}{2} + hH^\star_i\Big(g(x_i) - f'_x(x_i)\Big), \tag{10}$$

where H_i^j represents the approximate value of the solution in point x_i at time τ_j and $\star \in \{j, j+1\}$ represents the chosen time layer. Depending on in which time we evaluate the terms on the right–hand side in (10) we obtain three distinct first–order schemes.

Explicit scheme is obtained by taking all terms from the old time layer, i.e. $\star = j$.

Semi–implicit scheme is obtained by taking all linear terms from the old time layer, i.e. $\star = j$, and all nonlinear terms from the new time layer, i.e. $\star = j + 1$. The solution is found by solving a tridiagonal system of linear equations by the Thomas algorithm.

Fully–implicit scheme is obtained if all terms are taken from the new time layer, i.e. $\star = j + 1$. We get a system of nonlinear equations. The algorithm for solving such a system is based on iterative solution of the semi–implicit scheme. We start the iterative process by assigning the old time step solution vector to the starting iteration solution vector for the new time step. Then, in each iteration, we insert the solution vector into the nonlinear terms, to get their actual iteration. If we collect all unknowns from the solution vector, i.e. the linear terms from the new layer, on the left–hand side and all remaining terms, i.e. the nonlinear terms and the linear term from the old layer, on the right–hand side we obtain a linear tridiagonal system for determining next iteration of the solution vector. The whole process is terminated when the successive solution vectors are close enough [2].

New second–order scheme is of the **Crank–Nicolson type** and is obtained by the arithmetic average of the explicit and the fully–implicit scheme. The system of nonlinear equations has a similar structure to that from the fully–implicit scheme, thus we solve it using the same principles.

Stability As noticed above, the linear systems arising in our schemes are solved by the Thomas algorithm. Its numerical stability is guaranteed by the strict diagonal dominance of the system's matrix which can be always achieved by a suitable choice of time step k in (10). Another important issue is the study of stability which is usually related to the approximation of diffusion and advection terms. Inspecting the Gamma equation (5), one can see that the diffusion coefficient is given by β'_H while the speed of the advection is proportional to $\beta'_H + r - q$ and thus they are comparable ($\hat{\sigma}^2 \approx r - q$). The fully explicit scheme gives oscillations for the coupling $k \approx h$ due to violating the CFL condition in approximation of the diffusion term. On the other hand, all other schemes are implicit and we did not observe any oscillations, mainly due to the fact that the advection does not dominate the diffusion.

3 Numerical Experiments

Three different numerical experiments were made. The first two are concerned with the approximate solution to the RAPM Gamma equation and the last one deals with the numerical solution to a nonlinear porous–medium PDE.

RAPM Gamma equation experiments As no comparative exact solution to such an equation is known, a natural choice is to take the exact solution of the linear Black–Scholes model. Clearly, to maintain the equality in the Gamma equation we have to add a residual term $Res(x, \tau)$ into (5) which balances the difference between the Black–Scholes solution and the higher Ask price of the RAPM model:

$$\frac{\partial H}{\partial \tau} = \frac{\partial^2 \beta}{\partial x} + \frac{\partial \beta}{\partial x} + (r - q)\frac{\partial H}{\partial x} - qH + Res, \tag{11}$$

where $\beta(H) = \frac{\hat{\sigma}^2}{2}(1 + \mu H^{\frac{1}{3}})H$. The first two experiments differ from each other in two main aspects: the coefficient μ and the initial condition. Following parameters were set for both cases the same: $\hat{\sigma} = 0.30,\ r = 0.03,\ q = 0.01,\ E = 25$. In all numerical experiments we impose boundary conditions $H(x_L, \tau) = H(x_R, \tau) = 0$, where x_L and x_R are boundaries of the space interval.

The intention of *the first experiment* is to show how well the proposed numerical schemes can handle the nonlinearity in the Gamma equation (11). We put the coefficient $\mu = 0.2$, hence the function $\beta(H)$ is nonlinear. As the initial condition $H(x, \tau_0)$ we consider Black–Scholes solution $V(S, T - \tau_0)$ transformed by (4), in time $\tau_0 = 1$. Measurements of the estimated error $||e_n^m||_{L_2}$ are done by comparison with the exact solution $H(x, \tau)$ to (11) for $\tau > \tau_0$. Since all first–order schemes exhibited very similar features, we show here outputs just for the semi–implicit scheme. The reason for exclusion of the explicit scheme was its instability using coupling $k = h$. Regarding the fully–implicit scheme, experiments show that the accuracy of the semi–implicit scheme is very close to the fully–implicit scheme, thus it is sufficient to use just the former one which is less time consuming. The

Table 1 Outputs obtained by solving the RAPM Gamma equation (11) ($\tau_0 = 1, k = h$) using the semi–implicit scheme: estimated error $||e_n^m||_{L_2}$, CPU–time and EOC with respect to $||e_n^m||_{L_2}$

n	h	$\|\|e_n^m\|\|_{L_2}$	CPU	$EOC_{k\sim h}$
20	0.1	0.00777657	2.231	–
40	0.05	0.00333385	9.126	1.22194
80	0.025	0.00153036	36.614	1.12332
160	0.0125	0.00073141	147.078	1.06512
320	0.00625	0.00035733	582.929	1.03343

Table 2 Outputs obtained by solving the RAPM Gamma equation (11) ($\tau_0 = 1, k = h$) using the Crank–Nicolson type scheme: estimated error $||e_n^m||_{L_2}$, CPU–time and EOC with respect to $||e_n^m||_{L_2}$

n	h	$\|\|e_n^m\|\|_{L_2}$	CPU	$EOC_{k\sim h}$
20	0.1	0.00272286	4.383	–
40	0.05	0.000666762	17.785	2.02988
80	0.025	0.000165182	71.136	2.01311
160	0.0125 5	0.0000412598	294.062	2.00125
320	0.00625	0.0000108204	1206.53	1.93099

experiment was done on the time–space domain $(x, \tau) = [-2, 2] \times [1, 2]$. Tables 1 and 2 indicate that for this type of problem the semi–implicit scheme is first order accurate while the Crank–Nicolson type scheme is second order accurate.

In the *the second experiment* we set $\mu = 0$ and we show how the regularization of the transformed initial condition and the backward transformation of the Gamma equation solution affects the total accuracy of the method. In this case the solution of the Gamma equation coincides with the transformed solution $H(x, \tau)$ of the linear Black–Scholes equation (1) which implies that the residual term in (11) is zero. The initial condition $H(x, \tau_0)$ is considered for $\tau_0 = 0$. Hence the transformed payoff function, see (2) and (4), is the Dirac delta function, $H(x, 0) = \delta(x),\ \ x \in R$. In order to get a suitable initial condition for our computation, we consider its regularization given by the function $H(x, 0) = \frac{N'(d)}{\hat{\sigma}\sqrt{\tau^*}}$, where $\tau^* > 0$ is sufficiently small, $N(d)$ is the cumulative distribution function of the normal distribution and $d = \frac{x+(r-q-\hat{\sigma}^2/2)\tau^*}{\hat{\sigma}\sqrt{\tau^*}}$ [12]. The backward transformation of numerical solution is done by using formula

$$V(S_k, T-\tau_j) = h \sum_{i=-n}^{n} \max(S_k - Ee^{x_i}, 0)\, H_i^j = h \sum_{i=-n}^{k} (S_k - Ee^{x_i})\, H_i^j$$

$$= h\, S_k \sum_{i=-n}^{k} H_i^j - h\, E \sum_{i=-n}^{k} e^{x_i}\, H_i^j = h\, S_k\, F_k - h\, E\, G_k, \quad (12)$$

Table 3 Outputs obtained by solving numerically Gamma equation (11) ($\tau_0 = 0, k = h/4$) using the Crank–Nicolson type scheme and using formula (12) for backward transformation

n	h	τ^*	$\|\|e_n^m\|\|_{L_2}$	$EOC_{k\sim h}$	CPU Gamma	CPU Transform	CPU Total
5	0.4	0.46765	4.0644	–	0.047	0.011	0.058
10	0.2	0.14602	1.4586	1.4784	0.141	0.016	0.157
20	0.1	0.04371	0.4617	1.6595	0.624	0.047	0.671
40	0.05	0.01269	0.1379	1.7432	2.372	0.187	2.559
80	0.025	0.00361	0.0399	1.787	9.173	0.843	10.016
160	0.0125	0.00101	0.0113	1.816	41.091	3.323	44.414
320	0.00625	0.00028	0.0031	1.8270	150.525	12.87	163.396

where $F_k = F_{k-1} + H_k^j$, $G_k = G_{k-1} + e^{x_k} H_k^j$ and $S_k = Ee^{x_k}$. Formula (12) is obtained by integration of (4). Measurements of the estimated error $||e_n^m||_{L_2}$ are done by comparison with the Black–Scholes solution $V(S,t)$. However, in practice, doing computations with such an initial condition is not as straightforward task as in the first experiment. The problem is that we do not know a priori the optimal value of τ^* for a given time–space mesh. We consider the optimal value of τ^* as a value for which the estimated error of the numerical solution is minimized. Numerical outputs for the discretized time–space domain $(x,\tau) = [-2,2] \times [0,1]$ are summarized in the Table 3. Since the total error is influenced not only by the discretization error, but also by the error related to the regularization and backward transformation, the Crank–Nicolson method exhibits EOC slightly below the second order. Finally, in Fig. 1 we present the numerical solution of the RAPM model for a call option using parameter τ^* obtained by the above described strategy but considering nonzero μ. Such an experiment is of particular interest also for practical applications.

Experiment with an exact (Barenblatt) solution The goal of the *third experiment* was to investigate the accuracy of the proposed Crank–Nicolson type scheme using exact solution of the following (porous–medium type) equation [10]:

$$\partial_t v = \partial_x^2(v^\omega), \quad x \in R, \quad t > 0, \quad \omega > 1 \tag{13}$$

which is a special case of the Gamma equation (5). The exact solution has the form $v(x,t) = \frac{1}{\omega(t)} \max\left[0, 1 - \left(\frac{x}{\omega(t)}\right)^2\right]^{\frac{1}{\omega-1}}$, where $\lambda(t) = \left[\frac{2\omega(\omega+1)}{\omega-1}(t+1)\right]^{\frac{1}{\omega+1}}$ represents a sharp interface of the solution's finite support. EOC of the Crank–Nicolson type scheme in L_1–norm which is used due to the singularity in the exact solution, is equal to 2, see Table 4.

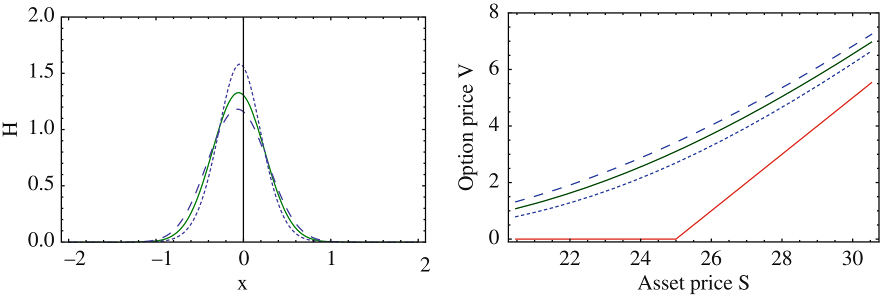


Fig. 1 A comparison of Bid and Ask option prices computed by means of the RAPM model for a call option in time $T - t = 1$. Left (right) figure presents the results before (after) the backward transformation. The dashed (fine–dashed) curve indicates the Ask (Bid) price of a call option. The solid curve represents the option prices computed by the linear Black–Scholes model and the solid broken line is the payoff function. Parameters: $n = 80$, $h = 0.025$, $m = 160$, $k = 0.00625$, $\tau^* = 0.00391$, $\hat{\sigma} = 0.30$, $\mu = \pm 0.2$, $r = 0.011$, $q = 0.0$, $X = 25$

Table 4 Numerical approximation of the Barenblatt exact solution using Crank–Nicolson type scheme

n	h	$\|\|e_n^m\|\|_{L_1}$	CPU	$EOC_{k\sim h}$
25	0.1	0.000629	0.312	–
50	0.05	0.000173	1.139	1.8584
100	0.025	0.000048	4.258	1.8543
200	0.0125	0.000012	17.036	1.9161
400	0.00625	$3.31 \cdot 10^{-}6$	67.798	1.9399
800	0.003125	$8.52 \cdot 10^{-}7$	250.475	1.9597
1600	0.0015625	$2.16 \cdot 10^{-}7$	881.905	1.97824

4 Conclusions

In this paper we proposed a new nonlinear second order Crank–Nicolson type numerical scheme based on the finite volume method. Our main goal was to provide an efficient and precise numerical solution to nonlinear PDEs arising in financial mathematics. Various experiments have shown such properties of the new scheme.

Acknowledgements This work was supported by grants APVV–0351–07 and VEGA 1/0733/10.

References

1. Avellaneda, M., Levy, A., and Paras, A.: Pricing and hedging derivative securities in markets with uncertain volatilities. Applied Mathematical Finance 2, (1995) 73-88
2. Balažovjech M., Mikula K.: A Higher Order Scheme for the Curve Shortening Flow of Plane Curves. Proceedings of ALGORITMY, STU Bratislava, (2009) 165–175

3. Barles, G., and Soner, H. M.: Option pricing with transaction costs and a nonlinear Black-Scholes equation. Finance Stoch. 2, 4 (1998) 369-397
4. Black, F., and Scholes, M.: The pricing of options and corporate liabilities. The Journal of Political Economy 81, (1973) 637-654
5. Jandačka, M., Ševčovič, D.: On the risk-adjusted pricing-methodology-based valuation of vanilla options and explanation of the volatility smile. J. Appl. Math. 3, (2005) 235-258
6. Kratka, M.: No mystery behind the smile. Risk 9, (1998) 67-71
7. Kwok, Y. K.: Mathematical models of financial derivatives. Springer-Verlag, Singapore (1998)
8. Leland, H. E.: Option pricing and replication with transaction costs. Journal of Finance 40, (1985) 1283-1301
9. Merton, R.: Theory of rational option pricing. The Bell Journal of Economics and Management Science, (1973) 141-183
10. Mimura, M., Tomoeda, K.: Numerical approximations to interface curves for a porous media equation. Hiroshima Math. J. 13, Hiroshima University, (1983) 273–294
11. Schönbucher, P. J., and Wilmott, P.: The feedback effect of hedging in illiquid markets. SIAM J. Appl. Math. 61, 1 (2000) 232-272
12. Ševčovič, D., Stehlíkova, B., Mikula, M.: Analytical and numerical methods for pricing financial derivatives. Nova Science Publishers, Hauppauge NY (2011)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Monotonicity Conditions in the Mimetic Finite Difference Method

Konstantin Lipnikov, Gianmarco Manzini, and Daniil Svyatskiy

Abstract The maximum principle is a major property of solutions of partial differential equations. In this work, we analyze a few constructive algorithms that allow one to embed this property into a mimetic finite difference (MFD) method. The algorithms search in the parametric family of MFD methods for a member that guarantees the discrete maximum principle (DMP). A set of sufficient conditions for the DMP is derived for a few types of meshes. For general meshes, a numerical optimization procedure is proposed and studied numerically.

Keywords Mimetic method, discrete maximum principle, M-matrix
MSC2010: 35B50, 65N06, 65N50

1 Mimetic finite difference method with parameters

The maximum principle is one of the most important properties of solutions of partial differential equations [3, 4]. Its numerical analog, the discrete maximum principle (DMP), is one of the most difficult properties to achieve in numerical methods, especially when the computational mesh is distorted to adapt and conform to the physical domain or the problem coefficients are highly heterogeneous and anisotropic. In this work, we investigate sufficient conditions to ensure the DMP in the mimetic finite difference (MFD) method [2]. We extend the analysis proposed in [5] by considering an optimization procedure as a way to achieve the DMP.

K. Lipnikov · D. Svyatskiy
Los Alamos National Laboratory, USA, e-mail: lipnikov@lanl.gov,dasvyat@lanl.gov

G. Manzini
IMATI-CNR and CESNA-IUSS, Pavia, Italy, e-mail: marco.manzini@imati.cnr.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_69, © Springer-Verlag Berlin Heidelberg 2011

We consider the mimetic discretization of the steady diffusion problem for the scalar and vector solution fields, p and $\mathbf{u}$, given by

$$\mathbf{u} + \mathsf{K}\nabla p = 0 \quad \text{in } \Omega, \tag{1}$$

$$\text{div}(\mathbf{u}) = q \quad \text{in } \Omega, \tag{2}$$

$$p = g^D \quad \text{on } \Gamma. \tag{3}$$

Here, Ω is an open bounded polygonal subset of R^d with Lipshitz boundary Γ; K is a $d \times d$ bounded, strongly elliptic and symmetric diffusion tensor; $q \in L^2(\Omega)$ is the forcing term and $g^D \in H^{1/2}(\Gamma)$ is the given boundary function.

Hopf's lemmas in weak and strong form can be summarized as follows [4]. Let $-\text{div}(\mathsf{K}\nabla p) \le 0$ in Ω. Then, the weak maximum principle holds:

$$\max_{\mathbf{x}\in\overline{\Omega}} p(\mathbf{x}) \le \max\big(0, \max_{\mathbf{x}\in\Gamma} p(\mathbf{x})\big).$$

This implies immediately that $p \ge 0$ if f and g^D are non-negative functions. Finally, the strong maximum principle says that if p attains a nonnegative maximum $\widehat{p}$ at an interior point of Ω, then $p = \widehat{p}$ in $\overline{\Omega}$.

Let Ω_h denote a conforming and *face-connected* partition of Ω into control volumes P, which are general polyhedra in 3-D and polygons in 2-D. The degrees of freedom for the scalar variable p are p_{P} and p_{f}. They approximate the average of p over elements P and faces f, respectively. The degrees of freedom of the vector variable $\mathbf{u}$ are $U_{\mathsf{P},\mathsf{f}}$. They approximate the normal component of $\mathbf{u}$ over mesh faces f. Any internal face f shared by two elements P' and P'' is characterized by two flux unknowns $U_{\mathsf{P}',\mathsf{f}}$ and $U_{\mathsf{P}'',\mathsf{f}}$ that must satisfy the flux conservation condition:

$$U_{\mathsf{P}',\mathsf{f}} + U_{\mathsf{P}'',\mathsf{f}} = 0. \tag{4}$$

Let the boundary of P be formed by the m faces f_i, $i = 1, \dots, m$, with measure $|\mathsf{f}_i|$ (length in 2-D, surface area in 3-D). We consider the numerical discretization of (1) that reads

$$\begin{pmatrix} U_{\mathsf{P},\mathsf{f}_1} \\ \vdots \\ U_{\mathsf{P},\mathsf{f}_m} \end{pmatrix} = \mathbb{W}_{\mathsf{P}} \begin{pmatrix} |\mathsf{f}_1|(p_{\mathsf{f}_1} - p_{\mathsf{P}}) \\ \vdots \\ |\mathsf{f}_m|(p_{\mathsf{f}_m} - p_{\mathsf{P}}) \end{pmatrix}, \tag{5}$$

where $\mathbb{W}_{\mathsf{P}}$ is a symmetric and positive definite (SPD) matrix.

Let $\mathbf{U}_{\mathsf{P}} = (U_{\mathsf{P},\mathsf{f}_1}, U_{\mathsf{P},\mathsf{f}_2}, \dots, U_{\mathsf{P},\mathsf{f}_m})^T$ be the m-sized vector of numerical fluxes across faces f_i of P. We write the numerical approximation of (2) as

$$\text{div}_{\mathsf{P}}\mathbf{U}_{\mathsf{P}} = \overline{q}_{\mathsf{P}}, \qquad \text{and} \qquad \text{div}_{\mathsf{P}}\mathbf{U}_{\mathsf{P}} = \frac{1}{|\mathsf{P}|}\sum_{i=1}^{m} |\mathsf{f}_i| U_{\mathsf{P},\mathsf{f}_i}. \tag{6}$$

where $\overline{q}_{\mathsf{P}}$ is the average of q over P, and $\mathrm{div}_{\mathsf{P}}$ is the primary mimetic divergence operator. The MFD method is given by equations (4), (5), (6). The Dirichlet boundary conditions are imposed by assigning average values of g^D on boundary faces f to corresponding unknowns p_{f}.

2 Construction of monotone mimetic methods

In the MFD method, the SPD matrix $\mathbb{W}_{\mathsf{P}}$ is is built in accordance with a *stability* and a *consistency* conditions [2]. A rich family of matrices satisfies these conditions. To achieve the DMP, we will impose additional constraints on this family.

The *stability condition* states that

$$\frac{\sigma_*}{|\mathsf{P}|}\,\mathbf{V}_{\mathsf{P}}^T\mathbf{V}_{\mathsf{P}} \le \mathbf{V}_{\mathsf{P}}^T\mathbb{W}_{\mathsf{P}}\mathbf{V}_{\mathsf{P}} \le \frac{\sigma^*}{|\mathsf{P}|}\,\mathbf{V}_{\mathsf{P}}^T\mathbf{V}_{\mathsf{P}} \qquad \forall\mathbf{V}_{\mathsf{P}}, \tag{7}$$

where σ_* and σ^* are two constants independent of P and of the mesh Ω_h. This condition states that matrix $\mathbb{W}_{\mathsf{P}}$ is spectrally equivalent to the scalar matrix $|\mathsf{P}|^{-1}\,\mathbb{I}_m$.

Let $\mathbf{x}_{\mathsf{P}}$ and $\mathbf{x}_{\mathsf{f}}$ be centers of gravity of element P and face f, respectively. Let $\mathbf{n}_{\mathsf{f}}$ be the external unit normal vector to f. We introduce the following two matrices:

$$\mathbb{R}_{\mathsf{P}} = \begin{pmatrix} |\mathsf{f}_1|(\mathbf{x}_{\mathsf{f}_1} - \mathbf{x}_{\mathsf{P}})^T \\ \vdots \\ |\mathsf{f}_m|(\mathbf{x}_{\mathsf{f}_m} - \mathbf{x}_{\mathsf{P}})^T \end{pmatrix} \quad \text{and} \quad \mathbb{N}_{\mathsf{P}} = \begin{pmatrix} \mathbf{n}_{\mathsf{f}_1}^T \\ \vdots \\ \mathbf{n}_{\mathsf{f}_m}^T \end{pmatrix}\mathsf{K}. \tag{8}$$

The *consistency condition* takes the form $\mathbb{W}_{\mathsf{P}}\mathbb{R}_{\mathsf{P}} = \mathbb{N}_{\mathsf{P}}$.

A straightforward calculation shows that $\mathbb{N}_{\mathsf{P}}^T\,\mathbb{R}_{\mathsf{P}} = |\mathsf{P}|\,\mathsf{K}$. It is proved in [2] that matrix $\mathbb{W}_{\mathsf{P}}$ is given by

$$\mathbb{W}_{\mathsf{P}} = \mathbb{N}_{\mathsf{P}}(\mathbb{N}_{\mathsf{P}}^T\mathbb{R}_{\mathsf{P}})^{-1}\mathbb{N}_{\mathsf{P}}^T + \mathbb{D}_{\mathsf{P}}\mathbb{U}_{\mathsf{P}}\mathbb{D}_{\mathsf{P}}^T, \tag{9}$$

where $\mathbb{D}_{\mathsf{P}}$ is a maximum rank $d \times (m-d)$-sized matrix such that $\mathbb{R}_{\mathsf{P}}^T\mathbb{D}_{\mathsf{P}} = 0$, and $\mathbb{U}_{\mathsf{P}}$ is a $(m-d)\times(m-d)$-sized SPD matrix of parameters.

An effective way to ensure that the monotonicity property holds is to construct a numerical method such that the final discretization matrix is an M-matrix [1]. In the MFD method, this occurs when $\mathbb{W}_{\mathsf{P}}$ satisfies two geometric conditions formulated below. Let $\mathbb{W}_{\mathsf{P}} = \left\{w_{ij}\right\}_{i,j=1}^{m}$. Since this is an SPD matrix, we obtain that $w_{ii} > 0$. We assume that

($\mathsf{A1}$) The matrix $\mathbb{W}_{\mathsf{P}}$ satisfies the geometric constraint:

$$w_{ii}|\mathsf{f}_i| + \sum_{j \neq i} w_{ij}|\mathsf{f}_j| \geq 0 \quad \forall i,$$

and the inequality is strict for at least one matrix row.
($\mathsf{A2}$) The matrix $\mathbb{W}_{\mathsf{P}}$ is a Z-matrix, i.e., $w_{ij} \leq 0$ for $i \neq j$.

Sufficient conditions ($\mathsf{A1}$) and ($\mathsf{A2}$) together with positive definiteness of matrix $\mathbb{U}_{\mathsf{P}}$ result in a set of inequalities for every element P. These local optimization problems can be solved analytically on special meshes or numerically to provide an MFD method for which the following theorem holds.

Theorem 1 (Discrete Maximum Principle). *Let p_{P}, p_{f} and $U_{\mathsf{P},\mathsf{f}}$ be the solutions of the MFD method under assumptions ($\mathsf{A1}$) and ($\mathsf{A2}$). Let q and g^D be nonnegative functions. Then, $p_{\mathsf{P}} \geq 0$ for any $\mathsf{P} \in \Omega_h$. Furthermore, if $q = 0$, then the values of p_{P} are bounded by the maximum and minimum values of p_{f} on boundary faces* f.

3 Oblique parallelepipeds

Let us consider a mesh Ω_h consisting of regular oblique parallelepipeds. We assume that parallelepiped faces are planar. To construct matrices $\mathbb{N}_{\mathsf{P}}$ and $\mathbb{R}_{\mathsf{P}}$, we refer to the numbering order shown in Fig. 1. Let $\mathbf{n}_1 = \mathbf{n}_{BCGF}, \mathbf{n}_2 = \mathbf{n}_{DCGH}, \mathbf{n}_3 = \mathbf{n}_{EFGH}$ and $\alpha := |\mathsf{f}_{BCGF}| = |\mathsf{f}_{ADHE}|$, $\beta := |\mathsf{f}_{DCGH}| = |\mathsf{f}_{ABFE}|$, $\gamma := |\mathsf{f}_{EFGH}| = |\mathsf{f}_{ABCD}|$. We define the *rotated diffusion tensor*:

$$\mathsf{K}^{\theta} = (\mathsf{K}^{\theta}_{ij})^3_{i,j=1} \qquad \mathsf{K}^{\theta}_{ij} = \mathbf{n}_i^T \mathsf{K} \mathbf{n}_j.$$

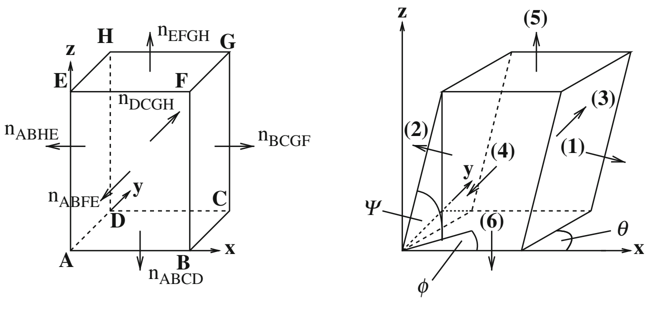


Fig. 1 Geometry of an orthogonal (left) and oblique (right) parallelepiped

Let us choose matrix $\mathbb{D}_{\mathsf{P}}^{T}$ as follows:

$$\mathbb{D}_{\mathsf{P}}^{T} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}. \tag{10}$$

This choice of $\mathbb{D}_{\mathsf{P}}$ allows us to simplify analysis of the MFD method and to prove the following results.

Lemma 1. *Assumptions (A1) and (A2) imply that*

$$0 < |\mathsf{P}|\mathbb{U}_{\mathsf{P}} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} \le \widetilde{\mathsf{K}}^{\theta} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix},$$

where $\widetilde{\mathsf{K}}^{\theta} = \{\widetilde{\mathsf{K}}_{ij}^{\theta}\}_{i,j=1}^{3}$ *with* $\widetilde{\mathsf{K}}_{ii}^{\theta} = \mathsf{K}_{ii}^{\theta}$ *and* $\widetilde{\mathsf{K}}_{ij}^{\theta} = -|\mathsf{K}_{ij}^{\theta}|$ *for* $i \neq j$.

From this and assumptions on $\mathbb{U}_{\mathsf{P}}$, we derive two *necessary conditions* for existence of a monotone MFD method:

$$\widetilde{\mathsf{K}}^{\theta} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} > 0 \quad \text{and} \quad \widetilde{\mathsf{K}}^{\theta} \text{ is SPD}. \tag{11}$$

Conditions (11) impose constraints on the range of values that the coefficients in K and the face areas $|\mathsf{f}_i|$ may attain in order to have a monotone mimetic discretization. If $\widetilde{\mathsf{K}}^{\theta}$ is an SPD matrix, a possible choice for $\mathbb{U}_{\mathsf{P}}$, which maximizes the sparsity its structure, is $\mathbb{U}_{\mathsf{P}} = |\mathsf{P}|^{-1}\widetilde{\mathsf{K}}^{\theta}$.

Let Ω be the unit cube a hole]0.6; 0.8[×]0.438; 0.563[×]0.5; 0.6[. The computational mesh is 10×16×20 with a 2×2×2 hole. A tilted domain and the corresponding mesh are obtained through the linear transformation $x := x + z\cos(\theta)$, $y := y + z\cos(\theta)$, $z := z\sin(\theta)$. The diffusion tensor is

$$\mathsf{K} = \begin{pmatrix} 100 & 0.25 & 0.15 \\ 0.25 & 1 & 0.25 \\ 0.15 & 0.25 & 1 \end{pmatrix}. \tag{12}$$

We set $f = 0$, $g^D = 2$ on the interior boundary (surface of the hole) and $g^D = 0$ on the exterior boundary. The exact solution is not known but must vary between 0 and 2 due to the maximum principle.

In Table 1, we consider a range of parameters θ. The significant violation of the minimum principle is clearly observed in the MFD method [2] where $\mathbb{U}_{\mathsf{P}}$ is set to a scalar matrix. It is due to huge number of negative entries in the inverse of the stiffness matrix (second column). Fig. 2 shows cuts through the element-based

Table 1 Original and monotone MFD methods on tilted parallelepiped meshes

		Original MFD				Monotone MFD	
θ	%	$\min(p_f)$	$\max(p_f)$	$\min(p_P)$	$\max(p_P)$	$\min(p_P)$	$\max(p_P)$
90	36.0	$-3.651\,10^{-1}$	2.083	$-9.371\,10^{-2}$	1.669	$6.304\,10^{-12}$	1.700
80	35.9	$-3.478\,10^{-1}$	2.089	$-9.039\,10^{-2}$	1.659	$5.224\,10^{-11}$	1.694
70	35.6	$-3.657\,10^{-1}$	2.092	$-9.464\,10^{-2}$	1.665	$5.403\,10^{-11}$	1.695
60	35.2	$-4.019\,10^{-1}$	2.084	$-1.028\,10^{-1}$	1.679	$9.995\,10^{-12}$	1.699

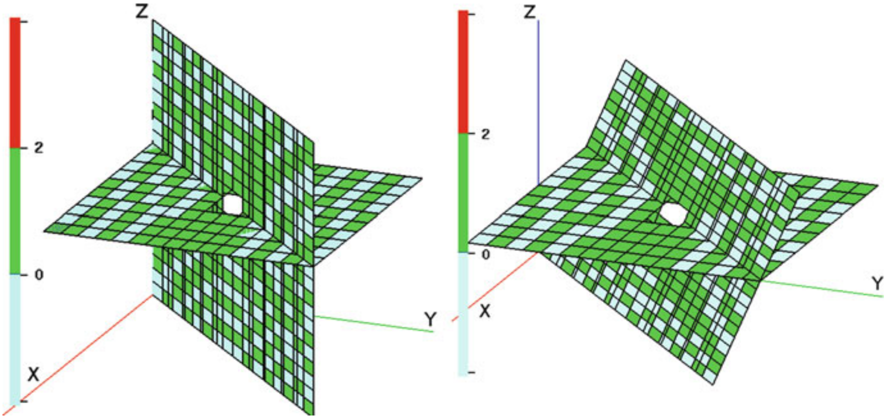


Fig. 2 Original MFD method: undershoots in the element-based numerical solution on the tilted domains with angles $\theta = 90°$ (left) and $61°$ (right)

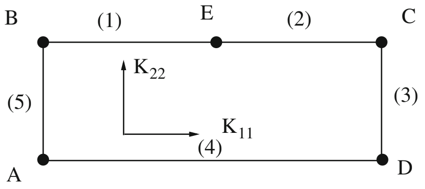


Fig. 3 A rectangular element $ABCD$ with a handing node E

discrete solution for $\theta = 90°$ and $60°$ in the original MFD method. For visualization clarity only two colors are used and the lighter color corresponds to negative solution values. The monotone MFD method satisfies the DMP.

4 Locally refined rectangular meshes

Let us consider a locally refined rectangular meshes (see Fig. 4). In the MFD framework, these meshes are considered as general polygonal meshes; thus, no special treatment of handing nodes is required (see Fig. 3).

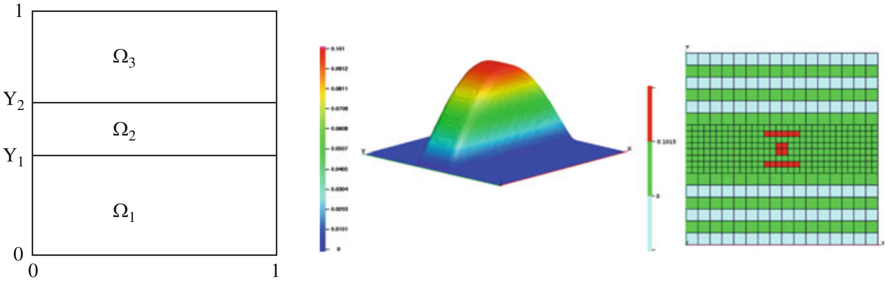


Fig. 4 The computational domain (left), discrete solution calculated with the monotone MFD method (middle), and the locally refined mesh (right)

We assume that the diffusion tensor is diagonal, $\mathsf{K} = \text{diag}\{\mathsf{K}_{11}, \mathsf{K}_{22}\}$. Let $r^2 = \frac{|\mathsf{f}_{AD}|}{|\mathsf{f}_{AB}|}$ be the aspect ratio of pentagon $ABECD$. Analysis of the MFD method leads to the following results.

Lemma 2. *A matrix $\mathbb{W}_{\mathsf{P}}$ satisfying assumptions (A1) and (A2) exists when*

$$r^4 < 4\,\frac{\mathsf{K}_{11}}{\mathsf{K}_{22}}. \tag{13}$$

For each aspect ratio r satisfying (13), we obtain a family of monotone mimetic methods. The closer the aspect ratio to the limiting value $4\mathsf{K}_{11}/\mathsf{K}_{22}$, the narrower this family. Among many of possible choices, we present a member which reduces the number of nonzero elements in the matrix $\mathbb{W}_{\mathsf{P}}$:

$$\mathbb{U}_{\mathsf{P}} = \frac{1}{|\mathsf{P}|}\begin{pmatrix} \frac{\mathsf{K}_{11}}{r^4} + \frac{\mathsf{K}_{22}}{4} & -\frac{\mathsf{K}_{22}}{2} & \frac{\mathsf{K}_{11}}{r^4} \\ -\frac{\mathsf{K}_{22}}{2} & \mathsf{K}_{22} & -\frac{\mathsf{K}_{22}}{2} \\ \frac{\mathsf{K}_{11}}{r^4} & -\frac{\mathsf{K}_{22}}{2} & \frac{\mathsf{K}_{22}}{r^4} + \frac{\mathsf{K}_{11}}{4} \end{pmatrix}, \quad \mathbb{D}_{\mathsf{P}}^T = \begin{pmatrix} 2 & -2 & r^2 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ -2 & 2 & 0 & 0 & r^2 \end{pmatrix}. \tag{14}$$

This choice imposes a stronger condition on the aspect ration, $r^4 < 8\mathsf{K}_{22}/(3\mathsf{K}_{11})$.

Let Ω be the unit square divided into three subdomains $\Omega_1 = (0,1) \times (0, Y_1)$, $\Omega_2 = (0,1) \times (Y_1, Y_2)$, and $\Omega_3 = (0,1) \times (Y_2, 1)$ as shown in Fig. 4. We set the forcing term and the diffusion tensor as follows:

$$f(x,y) = \begin{cases} 0, & (x,y) \in \Omega_1 \cup \Omega_3, \\ 10^3 \sin(\pi x) & (x,y) \in \Omega_2, \end{cases} \qquad \mathsf{K} = \begin{pmatrix} 10^3 & 0 \\ 0 & 1 \end{pmatrix}.$$

In our experiments $Y_1 = 3/8$ and $Y_2 = 5/8$. The exact solution to this problem can be calculated using the separation of variables. It is shown in Fig. 4.

The solution profile has sharp gradients around interfaces between subdomains. Therefore, we refine the subdomain Ω_2 and obtain a set of meshes similar to that shown in Fig. 4. According to the DMP, the solution has to be strictly positive inside the computational domain. The numerical results show that the original MFD method produces numerical solutions that violate the DMP and have large subdomains with overshoots and undershoots. A cell-centered solution is plotted in Fig. 4 using three pseudo-colors. The lighter color represents negative solution, the darker color represents solution overshoot.

The numerical solutions obtained with the original MFD method still violates the DMP after one mesh refined. With one additional refinement, the undershoots become comparable with the solver tolerance. The monotone MFD method uses the parameter matrix in (14) and provides a monotone solution which is bounded by the minimum and maximum of the analytical solution.

5 Monotone MFD methods based on numerical optimization

On more general meshes, analysis of the MFD family of methods becomes too complicated. Therefore, we reformulate the problem of constructing an M-matrix $\mathbb{W}_{\mathsf{P}}$ as a constrained optimization problem:

$$\min_{\mathbb{U}_{\mathsf{P}} \in \mathscr{U}_{\mathsf{P}}} \Phi(\mathbb{W}_{\mathsf{P}}(\mathbb{U}_{\mathsf{P}})),$$

where $\mathscr{U}_{\mathsf{P}}$ is a set of SPD matrices with the smallest eigenvalue bounded from below by $\lambda_{\min}(\mathsf{K}_{\mathsf{P}})/2$ and the functional Φ penalizes positive off-diagonal entries in $\mathbb{W}_{\mathsf{P}}$ as well as violation of the assumption (A1):

$$\Phi(\mathbb{W}_{\mathsf{P}}) = \sum_{i \neq j} (w_{ij} + |w_{ij}|)^2 + \sum_i (s_i - |s_i|)^2, \qquad s_i = w_{ii}|f_i| + \sum_{i \neq j} w_{ij}|f_j|.$$

The functional achieves its minimal value when $w_{ij} \leq 0$ for $i \neq j$ and $s_i \geq 0$. Restriction imposed on the minimal eigenvalue of $\mathbb{U}_{\mathsf{P}}$ guarantees that the matrix $\mathbb{W}_{\mathsf{P}}$ is SPD.

We implemented a simple minimization algorithm based on numerical calculation of the gradient of Φ and functional minimization along this direction. Let us consider again the problem from Sec. 3. Table 2 shows minimal and maximal solution values for two MFD methods. The first method uses a scalar matrix $\mathbb{U}_{\mathsf{P}} = a_{\mathsf{P}}\,\mathbb{I}_{\mathsf{P}}$, where $a_{\mathsf{P}} = \text{trace}(\mathsf{K}_{\mathsf{P}})/3$ lies in a middle of spectrum of K_{P}. The second method uses this matrix as the initial guess for the minimization algorithm. Since for every P, the number of parameters is six, we terminate the algorithm after six steps.

A simple optimization procedure is sensitive to an initial guess. Therefore, in the future, we plan to analyze more advanced optimization strategies as well as different functionals.

Table 2 Original and optimized MFD methods on tilted parallelepiped meshes

	Original MFD ($\mathbb{U}_{\mathsf{P}} = a_{\mathsf{P}}\, \mathbb{I}_{\mathsf{P}}$)		Optimized MFD	
θ	$\min(p_{\mathsf{P}})$	$\max(p_{\mathsf{P}})$	$\min(p_{\mathsf{P}})$	$\max(p_{\mathsf{P}})$
70	$-7.267\, 10^{-2}$	1.577	$5.855\, 10^{-11}$	1.641
60	$-8.320\, 10^{-2}$	1.602	$9.801\, 10^{-12}$	1.648
50	$-8.998\, 10^{-2}$	1.628	$2.378\, 10^{-14}$	1.641

6 Conclusions

In this paper, we present a new methodology for the construction of mimetic discretizations which satisfy the discrete maximum principle. A set of sufficient conditions is derived to ensure that such monotone subfamily exists.

Acknowledgements The work of the first (K.L.) and third author (D.S.) was supported by the Department of Energy (DOE) Advanced Scientific Computing Research (ASCR) program in Applied Mathematics. The work of the second author (G.M.) was partially supported by the Italian MIUR through the program PRIN2008.

References

1. A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences.* Academic Press , New York, 1979. Computer Science and Applied Mathematics.
2. F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Mod. Meth. Appl. Sci.*, 15(10):1533–1551, 2005.
3. P. Grisvard. *Elliptic problems in nonsmooth domains*, volume 24 of *Monographs and Studies in Mathematics*. Pitman (Advanced Publishing Program), Boston, MA, 1985.
4. E. Hopf. Elementare Bemerkungen uber die L osungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. *Sitzungsber. Preuss. Akad. Wiss.*, 19, 1927.
5. K. Lipnikov, G. Manzini, and D. Svyatskiy. Analysis of the monotonicity conditions in the mimetic finite difference method for elliptic problems *J. Comput. Phys.*, 230(7): 2620–2642, 2011.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Discrete Duality Finite Volume Method Applied to Linear Elasticity

Benjamin Martin and Frédéric Pascal

Abstract We present the Discrete Duality Finite Volume method (DDFV) for solving the linear elasticity problem on unstructured mesh applied to solids undergoing mechanical loads. The procedure is described in detail for three dimensional problems and some theoretical results are provided: the discrete problem is well-posed, stable and convergent. A number of numerical test problems demonstrates the ability of this finite volume scheme to approach the solution and some comparisons with the conventional finite element method are provided.

Keywords Finite volume methods, linear elasticity, stability and convergence
MSC2010: 65N08, 73C02, 65N12

1 Motivation

The finite volume method is extensively used in computational fluid dynamics, on its part the finite element method is the conventional tool for solving solid mechanics. However there is a multitude of physical problems combining fluid and solid mechanics where finite volume methods appear to be a pertinent alternative. Let us quote for instance fluid-structure interaction, deformation of geomechanical reservoir, or even the frost heave problem in freezing soils where the moving frozen fringe introduces a discontinuity in the physical parameters. The finite volume approach for elasticity problems has already been discussed and published in [3],

Benjamin Martin and Frédéric Pascal
CMLA, ENS de Cachan, CNRS, 61 Avenue du Pt Wilson, 94235 Cachan, France,
e-mail: benjamin.martin@cmla.ens-cachan.fr, frederic.pascal@cmla.ens-cachan.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_70, © Springer-Verlag Berlin Heidelberg 2011

[15], [17], [18] for cell-vertex formulations, in [12] for cell centered formulation with a decoupled strategy for each component, in [6] and [7] for a coupled cell-center version. In this study, we address the DDFV implementation for solving linear elasticity. Let us recall that the principle of the DDFV discretization consists in integrating the system both over a given primal mesh and a dual mesh built from the primal one. Presentation, convergence analysis and numerical tests of DDFV for diffusion, convection-diffusion and Stokes problems are available in [1], [2], [4], [8], [10], [11], [13], [14].

We limit ourselves to the simplest mathematical model of a linear elastic solid which consists in finding the displacement $\mathbf{u} \in \mathbb{R}^3$ such that

$$-\operatorname{div} \boldsymbol{\sigma}(\mathbf{u}) = \mathbf{f} \quad \text{on} \quad \Omega\,, \qquad \mathbf{u} = \mathbf{g} \quad \text{on} \quad \Gamma_D\,, \qquad \boldsymbol{\sigma}(\mathbf{u}) \cdot \mathbf{n} = \mathbf{h} \quad \text{on} \quad \Gamma_N \tag{1}$$

where $\mathbf{n}$ is the outward normal, $\partial\Omega = \Gamma_D \cup \Gamma_N$ and where the stress tensor $\boldsymbol{\sigma}$ depends on $\mathbf{u}$ by the Hooke relation that links the strain tensor and the trace of the gradient

$$\boldsymbol{\sigma}(\mathbf{u}) = \lambda \mathbb{D}\mathrm{iv}\mathbf{u} + 2\mu \mathbb{D}\mathbf{u} \quad \text{with} \quad \mathbb{D}\mathbf{u} = \frac{\nabla \mathbf{u} + (\nabla \mathbf{u})^T}{2} \quad \text{and} \quad \mathbb{D}\mathrm{iv}\mathbf{u} = \operatorname{div}\mathbf{u}\ \mathrm{Id}\,. \tag{2}$$

For sake of clarity, we assume that Ω is a bounded polyhedral subset of $\mathbb{R}^3$ and that the Lamé coefficients λ and μ are constant.

2 Finite volume discretization

A mesh of Ω is defined by the three sets $\{\mathfrak{M}, \mathfrak{M}^*, \mathfrak{D}\}$, corresponding to the primal, dual and diamond mesh. They form a non overlapping partition of Ω, so that

$$\overline{\Omega} = \bigcup_{\mathrm{D} \in \mathfrak{D}} \mathrm{D} = \bigcup_{\mathrm{K} \in \mathfrak{M}} \mathrm{K} = \frac{1}{2} \bigcup_{\mathrm{K}^* \in \mathfrak{M}^*} \mathrm{K}^*\,.$$

The set $\mathfrak{M}$ is a conforming triangulation of tetraedra. Each element K in $\mathfrak{M}$ is supplied with a center $\mathbf{x}_\mathrm{K}$, in practice the barycenter of K and $\partial\mathfrak{M}$ denotes the set of faces on the boundary of the domain. The elements of $\mathfrak{M}^*$ are polygons K^* corresponding to the primal mesh vertices $\mathbf{x}_{\mathrm{K}^*}$. These polygons are the union of all tetrahedra spanned, for each faces $s = \mathrm{K} \cap \mathrm{L}$ or $s = \mathrm{K} \cap \partial\Omega$ having $\mathbf{x}_{\mathrm{K}^*}$ as vertex, by $\mathbf{x}_{\mathrm{K}^*}$ himself, $\mathbf{x}_\mathrm{K}$ or $\mathbf{x}_\mathrm{L}$ if it exists, $\mathbf{x}_s$ the center of the face s, and one of the other vertices of the face s. In order to take into account the boundary conditions, the dual mesh is splitted into the internal volumes and the boundary ones corresponding to vertices on the boundary: $\mathfrak{M}^* = \mathfrak{M}^{*i} \cup \mathfrak{M}^{*b}$. On its side, diamond cell D in $\mathfrak{D}$ associated to the internal face $s = \mathrm{K} \cap \mathrm{L}$ is the union of the two tetrahedra $\mathrm{D}_{\mathrm{K},s}$ and $\mathrm{D}_{\mathrm{L},s}$ spanned by the face s and respectively by the centers $\mathbf{x}_\mathrm{K}$ and $\mathbf{x}_\mathrm{L}$ (see Fig. 1a). For the boundary face $s = \mathrm{K} \cap \partial\Omega$, the corresponding diamond cell is reduced to

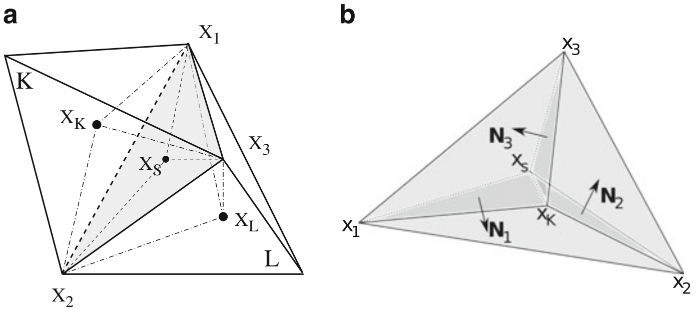


Fig. 1 (a) Primal and diamond cell - (b) Normal orientations in the diamond cell

the tetrahedron $\mathrm{D}_{\mathrm{K},s}$. The number of primal and dual cells is denoted by τ and the number of diamond cell by δ.

2.1 Discrete operators

The idea of the DDFV discretization is to construct gradient and divergence operators that are under discrete duality relation by a formula that mimics the Green fomula for continuous functions (see for instance [5] for a detailed construction). A discrete unknown $\mathbf{u}_{\mathrm{K}}$ (resp. $\mathbf{u}_{\mathrm{K}^*}$) is associated to each volume K (resp. K*) of the primal mesh (resp. dual mesh). They are gathered and denoted by

$$\mathbf{u}^\tau = (\mathbf{u}_{\mathrm{K}}\,,\, \mathbf{u}_{\mathrm{K}^*})_{\mathrm{K}\in\mathfrak{M},\, \mathrm{K}^*\in\mathfrak{M}^*}\,.$$

For a vector field $\mathbf{u}^\tau$ in $(\mathbb{R}^d)^\tau$, we define on each diamond cell a consistent discrete gradient operator $\nabla^{\mathfrak{D}}\mathbf{u}^\tau = (\nabla^{\mathrm{D}}\mathbf{u}^\tau)_{\mathrm{D}\in\mathfrak{D}}$ in $(\mathscr{M}_d(\mathbb{R}))^\delta$ and a consistent discrete divergence operator $\mathrm{div}^{\mathfrak{D}}\mathbf{u}^\tau = (\mathrm{div}^{\mathrm{D}}\mathbf{u}^\tau)_{\mathrm{D}\in\mathfrak{D}}$ in $\mathbb{R}^\delta$ such that on the internal face $s = \mathrm{K}\cap\mathrm{L}$ and for the associated diamond cell $\mathrm{D} = \mathrm{D}_{\mathrm{K},s}\cup\mathrm{D}_{\mathrm{L},s}$, the gradient is given by $\nabla^{\mathrm{D}}\mathbf{u}^\tau = \frac{|\mathrm{D}_{\mathrm{K},s}|}{|\mathrm{D}|}\nabla^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau + \frac{|\mathrm{D}_{\mathrm{L},s}|}{|\mathrm{D}|}\nabla^{\mathrm{D}_{\mathrm{L},s}}\mathbf{u}^\tau$ and the divergence by $\mathrm{div}^{\mathrm{D}}\mathbf{u}^\tau = \frac{|\mathrm{D}_{\mathrm{K},s}|}{|\mathrm{D}|}\mathrm{div}^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau + \frac{|\mathrm{D}_{\mathrm{L},s}|}{|\mathrm{D}|}\mathrm{div}^{\mathrm{D}_{\mathrm{L},s}}\mathbf{u}^\tau$ where for K, we take

$$\nabla^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau = \frac{1}{3\mid \mathrm{D}_{\mathrm{K},s}\mid}(\mathbf{u}_s - \mathbf{u}_{\mathrm{K}})\otimes\mathbf{N}_{\mathrm{K}s} + \frac{1}{3\mid \mathrm{D}_{\mathrm{K},s}\mid}\sum_{i=1}^{d}\mathbf{u}_i\otimes(\mathbf{N}_{i-1}-\mathbf{N}_{i+1}) \tag{3}$$

$$\mathrm{div}^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau = \frac{1}{3\mid \mathrm{D}_{\mathrm{K},s}\mid}(\mathbf{u}_s - \mathbf{u}_{\mathrm{K}})\cdot\mathbf{N}_{\mathrm{K}s} + \frac{1}{3\mid \mathrm{D}_{\mathrm{K},s}\mid}\sum_{i=1}^{d}\mathbf{u}_i\cdot(\mathbf{N}_{i-1}-\mathbf{N}_{i+1})\,. \tag{4}$$

Here $|\cdot|$ denotes the measure and $(\mathbf{x}_i)_{i=1}^d$, respectively $(\mathbf{u}_i)_{i=1}^d$, the vertices of the face s, respectively the corresponding unknowns, with the local numbering convention $\mathbf{x}_0 = \mathbf{x}_d$. The outward normals are defined by (see Fig. 1b)

$$\begin{aligned}
\mathbf{N}_{\mathrm{K}s} &= \sum_{i=1}^{d} \mathbf{N}_{s,i-1,i} \quad \text{with} \quad \mathbf{N}_{s,i-1,i} = \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_s) \wedge (\mathbf{x}_{i-1} - \mathbf{x}_s) \\
\mathbf{N}_i &= \frac{1}{2}(\mathbf{x}_{\mathrm{K}} - \mathbf{x}_s) \wedge (\mathbf{x}_i - \mathbf{x}_s)
\end{aligned} \tag{5}$$

and $\mathbf{u}_s$ is chosen in order to satisfy the continuity of fluxes (see 7). Otherwise, on a boundary face $s \in \partial\mathfrak{M}$ and for the corresponding diamond cell $\mathrm{D} = \mathrm{D}_{\mathrm{K},s}$, the gradient and the divergence are simply $\nabla^{\mathrm{D}}\mathbf{u}^\tau = \nabla^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau$ and $\mathrm{div}^{\mathrm{D}}\mathbf{u}^\tau = \mathrm{div}^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau$ but $\mathbf{u}_s$ depending on the boundary datas is explicited in (8).

2.2 The DDFV scheme

For $\mathbf{u}^\tau$ in $(\mathbb{R}^d)^\tau$, we are now able to define the discrete strain tensor $\mathbb{D}^{\mathfrak{D}}\mathbf{u}^\tau = (\mathbb{D}^{\mathrm{D}}\mathbf{u}^\tau)_{\mathrm{D}\in\mathfrak{D}}$ and the divergence one $\mathbb{D}\mathrm{iv}^{\mathfrak{D}}\mathbf{u}^\tau = (\mathbb{D}\mathrm{iv}^{\mathrm{D}}\mathbf{u}^\tau)_{\mathrm{D}\in\mathfrak{D}}$ by

$$\mathbb{D}^{\mathrm{D}}\mathbf{u}^\tau = \frac{\nabla^{\mathrm{D}}\mathbf{u}^\tau + (\nabla^{\mathrm{D}}\mathbf{u}^\tau)^T}{2} \quad , \quad \mathbb{D}\mathrm{iv}^{\mathrm{D}}\mathbf{u}^\tau = \mathrm{div}^{\mathrm{D}}\mathbf{u}^\tau \ \mathrm{Id} \quad \forall \mathrm{D} \in \mathfrak{D}\,. \tag{6}$$

After extending this definition to each tetrahedron that composes the diamond cell, we can specify that the displacement $\mathbf{u}_s$ at an internal face $s = \mathrm{K} \cap \mathrm{L}$ has to satisfy the continuity of the fluxes

$$(\lambda\mathbb{D}\mathrm{iv}^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau + 2\mu\mathbb{D}^{\mathrm{D}_{\mathrm{K},s}}\mathbf{u}^\tau)\mathbf{N}_{\mathrm{K}s} = -(\lambda\mathbb{D}\mathrm{iv}^{\mathrm{D}_{\mathrm{L},s}}\mathbf{u}^\tau + 2\mu\mathbb{D}^{\mathrm{D}_{\mathrm{L},s}}\mathbf{u}^\tau)\mathbf{N}_{\mathrm{L}s}\,. \tag{7}$$

Now for a tensor field $\xi^{\mathfrak{D}}$ in $(\mathscr{M}_d(\mathbb{R}))^\delta$, we define a consistent approximation of the discrete divergence operator equal to

$$(\mathbf{div}^{\mathfrak{M}}\xi^{\mathfrak{D}}, \mathbf{div}^{\mathfrak{M}^*}\xi^{\mathfrak{D}}) = \Big((\mathbf{div}^{\mathrm{K}}\xi^{\mathfrak{D}})_{\mathrm{K}\in\mathfrak{M}}, (\mathbf{div}^{\mathrm{K}^*}\xi^{\mathfrak{D}})_{\mathrm{K}^*\in\mathfrak{M}^*}\Big)$$

with

$$\mathbf{div}^{\mathrm{K}}\xi^{\mathfrak{D}} = \frac{1}{|\,\mathrm{K}\,|}\sum_{s\in\partial\mathrm{K}} \xi^{\mathrm{D}}\mathbf{N}_{\mathrm{K}s} \quad \text{and} \quad \mathbf{div}^{\mathrm{K}^*}\xi^{\mathfrak{D}} = \frac{1}{|\,\mathrm{K}^*\,|}\sum_{s\ni\mathbf{x}_{\mathrm{K}^*}} \xi^{\mathrm{D}}\mathbf{N}_{\mathrm{K}^*s}$$

and where D is the diamond cell associated to the face s. $\mathbf{N}_{\mathrm{K}^*s}$ is the normal to $\partial\mathrm{K}^*$ pointing outward K^* and it can be explicited using local numbering and applying formula (5):

$$\mathbf{N}_{\mathrm{K}^*s} = \begin{cases} \mathbf{N}^K_{i+1} - \mathbf{N}^K_{i-1} + \mathbf{N}^L_{i+1} - \mathbf{N}^L_{i-1} \text{ for an internal face } s = \mathrm{K}\cap\mathrm{L} \\ \mathbf{N}_{i+1} - \mathbf{N}_{i-1} + \mathbf{N}_{s,i-1,i} + \mathbf{N}_{s,i,i+1} \text{ for a boundary face } s = \mathrm{K}\cap\partial\Omega \end{cases}$$

where we assume that $x_{\mathrm{K}^*} = x_i^{\mathrm{K}}$ (resp. $x_{\mathrm{K}^*} = x_i^{\mathrm{L}}$) in the volume K (resp. L).

Let us now denote $\mathbf{f}^{\mathfrak{M}} = (\mathbf{f}^{\mathrm{K}})_{\mathrm{K}\in\mathfrak{M}}$ and $\mathbf{f}^{\mathfrak{M}^{*i}} = (\mathbf{f}^{\mathrm{K}^*})_{\mathrm{K}^*\in\mathfrak{M}^{*i}}$, where $\mathbf{f}^{\mathrm{K}}$ and $\mathbf{f}^{\mathrm{K}^*}$ are the average of the external force $\mathbf{f}$ on primal and dual cells. Then the DDFV scheme, written here, for sake of simplicity, only for displacement boundary conditions, consists in finding $\mathbf{u}^\tau \in (\mathbb{R}^d)^\tau$ such that

$$\begin{cases} -\mathbf{div}^{\mathfrak{M}}(\lambda \mathbb{D}\mathrm{iv}^{\mathfrak{D}}\mathbf{u}^\tau + 2\mu\mathbb{D}^{\mathfrak{D}}\mathbf{u}^\tau) = \mathbf{f}^{\mathfrak{M}} \\ -\mathbf{div}^{\mathfrak{M}^{*i}}(\lambda \mathbb{D}\mathrm{iv}^{\mathfrak{D}}\mathbf{u}^\tau + 2\mu\mathbb{D}^{\mathfrak{D}}\mathbf{u}^\tau) = \mathbf{f}^{\mathfrak{M}^{*i}} \\ \mathbf{u}_s = \mathbf{g}(\mathbf{x}_s), \quad \forall s \in \partial\mathfrak{M} \\ \mathbf{u}_{\mathrm{K}^*} = \mathbf{g}(\mathbf{x}_{\mathrm{K}^*}), \quad \forall \mathrm{K}^* \in \mathfrak{M}^{*b}\,. \end{cases} \tag{8}$$

2.3 Existence, stability and convergence results

Applying discrete Green formula, Korn and Poincaré inequalities, divergence equality and approximation results on the center value projection operator (see [14]), we prove that the numerical scheme is well-posed, stable and convergent:

Theorem 1. *Under the assumption that $mes(\Gamma_D) \neq 0$, the DDFV scheme for linear elasticity (8) yields to a symmetric positive definite system of linear equations. So it admits exactly one solution $\mathbf{u}^\tau \in (\mathbb{R}^d)^\tau$*

Theorem 2. *Let $\mathbf{u}^\tau \in (\mathbb{R}^d)^\tau$ be the solution of the discrete problem (8). Then there exists a constant C depending only on the regularity of the mesh such that*

$$\mu \mid\mid \nabla^{\mathfrak{D}}\mathbf{u}^\tau \mid\mid_2^2 + \frac{\lambda}{3} \mid\mid \mathbb{D}iv^{\mathfrak{D}}\mathbf{u}^\tau \mid\mid_2^2 \leq C \mid\mid \mathbf{f}^\tau \mid\mid_2^2 \tag{9}$$

Theorem 3. *Assuming that the exact solution of the continuous problem (1) is regular enough then there exists a constant C depending only on the regularity of the mesh, such that*

$$\mid\mid \mathbf{u} - \mathbf{u}^\tau \mid\mid_2 + \mid\mid \nabla\mathbf{u} - \nabla^{\mathfrak{D}}\mathbf{u}^\tau \mid\mid_2 \leq C\ \mathrm{size}(\mathfrak{M}) \tag{10}$$

3 Numerical experiments

The DDFV method has been implemented in two and three dimensions. Free and imposed traction conditions (described in [16]) are also taken into account. Both homogeneous and non homogeneous test cases are considered. Comparisons are made with the analytical solution or with the clasical finite element one.

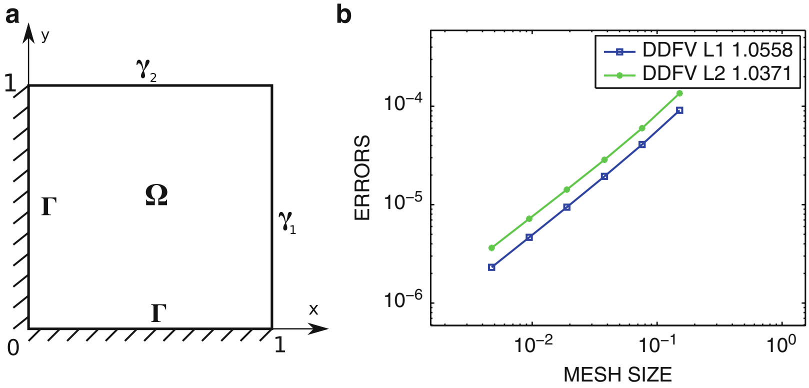


Fig. 2 (a) Geometry and test setup - (b) L^1 and L^2 nors of the error between the analytical and the numerical displacement.

3.1 Two dimensional examples

Following a study of [9], we apply the code to a simple test case with analytical solution in order to study the convergence properties. The geometry of the homogeneous square plate and the specified boundary conditions are shown in Fig. 2(a). Lamé coefficients $(\lambda, \mu) = (2.9\,10^9, 1.9\,10^9)$ correspond to Young modulus and Poisson ratio $(E, \nu) = (5\,10^9, 0.3)$. The displacement $\mathbf{g}$ is null on Γ boundary and the traction is imposed on γ_1 and γ_2 boundaries:

$$\mathbf{g}_{|\gamma_1} = \begin{pmatrix} ((2\mu+\lambda)y - 2\lambda)10^{-2} \\ \mu(1-2y)10^{-2} \end{pmatrix} \qquad \mathbf{g}_{|\gamma_2} = \begin{pmatrix} \mu(x-2)10^{-2} \\ (-2(2\mu+\lambda)x+\lambda)10^{-2} \end{pmatrix}.$$

The external force is equal to $\mathbf{f} = (\mu+\lambda)10^{-2}\,(2,-1)$ and the corresponding exact displacement is $\mathbf{u} = xy10^{-2}\,(1,-2)$. The comparison between the analytical and numerical displacement obtained for various primal meshes are plotted in Fig. 2(b) with an order of convergence of one.

The second example concerns a domain with non homogeneous material properties. The plate (without deformation) is composed of the part $[0,3]\times[0,1]$ with a hole inside and $(\lambda,\mu) = (5.6, 2.6)$ and the part $[3,4]\times[0,1]$ with $(\lambda,\mu) = (10, 8)$. Null displacement is imposed on the left side of the domain, a load of 1 (resp. a displacement of 1) is imposed on the right side for Fig. 3 (resp. for Fig. 4). There is a free traction elsewhere. The deformed domain obtained with the present scheme (above) and with the conventional finite element method (below) are plotted. In both case, solution are similar, the largest differences are observed in the load one.

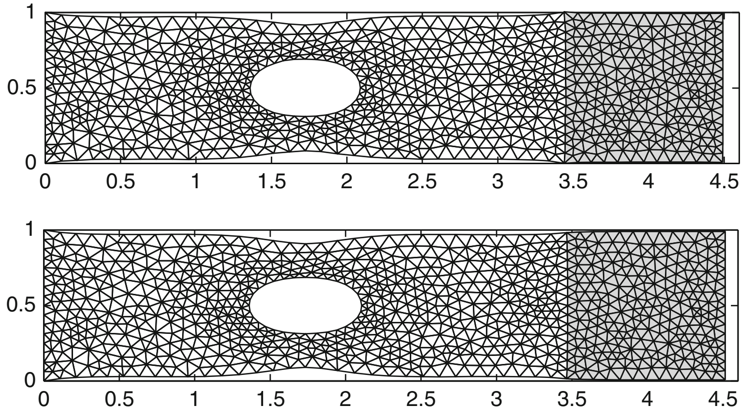


Fig. 3 Deformed domain for the non homogeneous case with an imposed load on the right. DDFV above and FE below

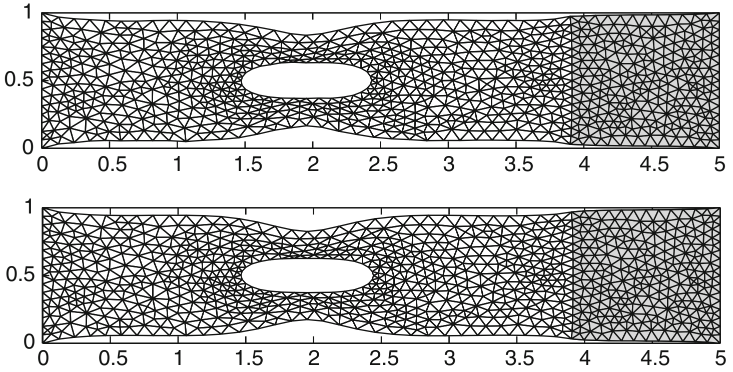


Fig. 4 Deformed domain for the non homogeneous case with an imposed displacement on the right. DDFV above and FE below

3.2 *Three dimensional test*

The domain is the unit cube with an embedding condition on the bottom ($z = 0$), imposed displacement $(0, 0, -0.5)$ on the top ($z = 1$) simulating a compression of the domain (see Fig. 5b) and free traction conditions on the vertical sides of the cube. For Lamé coefficients $(\lambda, \mu) = (28.8, 19.2)$, the solution is compared with the P1 finite element one on a series of meshes: Figure 5a shows the behavior of the error in L^2 and L^1 norms and reveals that the DDFV solution of the linear elasticity problem converges as we expect.

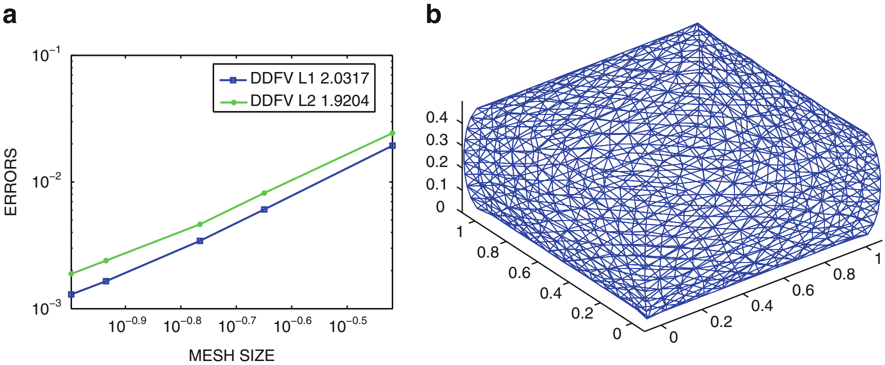


Fig. 5 (a) Differences with the finite element solution - (b) DDFV deformed domain

References

1. Andreianov, B., Bendahmane, M., Karlsen, K., Pierre, C.: Convergence of discrete duality finite volume schemes for the cardiac bidomain model. Arxiv preprint arXiv:1010.2718 (2010)
2. Andreianov, B., Boyer, F., Hubert, F.: Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes. Numer. Methods Partial Differential Equations **23**(1), 145–195 (2007)
3. Bailey, C., Cross, M.: A finite volume procedure to solve elastic solid mechanics problems in three dimensions on an unstructured mesh. Int. J. Numer. Methods Eng. **38**(10), 1757–1776 (1995)
4. Coudière, Y., Manzini, G.: The discrete duality finite volume method for convection-diffusion problems. SIAM J. Numer. Anal. **47**(6), 4163–4192 (2010)
5. Coudière, Y., Pierre, C., Rousseau, O., Turpault, R.: A 2D/3D discrete duality finite volume scheme. Application to ECG simulation. Int. J. Finite Vol. **6**(1), 24 (2009)
6. Demirdžić, I., Muzaferija, S.: Finite volume method for stress analysis in complex domains. Int. J. Numer. Methods Eng. **37**(21), 3751–3766 (1994)
7. Demirdžić, I., Muzaferija, S., Perić, M.: Benchmark solutions of some structural analysis problems using finite-volume method and multigrid acceleration. Int. J. Numer. Methods Eng. **40**(10), 1893–1908 (1997)
8. Domelevo, K., Omnes, P.: A finite volume method for the laplace equation on almost arbitrary two-dimensional grids. M2AN Math. Model. Numer. Anal. **39**, 1203–1249 (2005)
9. Figueiredo, J., Viano, J.: Finite elements q1-lagrange for the linear elasticity problem. Tech. rep., Universidade de Santiago de Compostela (2005)
10. 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)
11. Hermeline, F.: A finite volume method for the approximation of diffusion operators on distorted meshes. J. Comput. Phys. **160**(2), 481–499 (2000)
12. Jasak, H., Weller, H.: Application of the finite volume method and unstructured meshes to linear elasticity. Int. J. Numer. Methods Eng. **48**(2), 267–287 (2000)
13. Krell, S.: Stabilized DDFV schemes for stokes problem with variable viscosity on general 2d schemes. Numer. Methods Partial Differential Equations (2010)
14. Krell, S., Manzini, G.: The discrete duality finite volume method for the stokes equations on 3-d polyhedral meshes (2010). URL http://hal.archives-ouvertes.fr/hal-00448465/en/

15. Maitre, J.F., Rezgui, A., Souhail, H., Zine, A.M.: High order finite volume schemes. Application to non-linear elasticity problems. In: Finite volumes for complex applications, III (Porquerolles, 2002), pp. 391–398. Hermes Sci. Publ., Paris (2002)
16. Martin, B.: Résolution du problème de l'élasticité linéaire en volumes finis. Ph.D. thesis, ENS de Cachan (2011)
17. Souhail, H.: Schéma volumes finis : Estimation d'erreur a posteriori hiérarchique par éléments finis mixtes. Résolution de problèmes d'élasticité non-linéaire. Ph.D. thesis, Ecole Centrale de Lyon (2004). URL http://tel.archives-ouvertes.fr/tel-00005418/en/
18. Wenke, P., Wheel, M.: A finite volume method for solid mechanics incorporating rotational degrees of freedom. Computers & Structures **81**(5), 321–329 (2003)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Model Adaptation for Hyperbolic Systems with Relaxation

Hélène Mathis and Nicolas Seguin

Abstract We address the numerical coupling of two hyperbolic systems, a relaxation model and the associated equilibrium model, separated by spatial interfaces that automatically evolve in time, the whole being approximated by finite volume schemes. The criterion to choose where each model has to be used results of the Chapman–Enskog expansion of the relaxed model, both on a continuous and a discrete view point. Numerical tests illustrate the good behavior of the algorithm.

Keywords Hyperbolic system, adaptation, relaxation
MSC2010: 65M08, 65M15, 35L45, 35L65

1 Introduction

In the framework of the modeling of problems coming from complex phenomena, it is common to handle different scales of modeling depending on the accuracy we need. It leads to the use of a hierarchy of models based on the different scales brought into play both in the spatial and time sides. The problem of spatial coupling of different hyperbolic models has been the topic of numerous papers, see for instance [1]. So far the theoretical and numerical techniques developed lied on the hypothesis that the spatial domains where each model has to be applied is initially prescribed and fixed in time. We aim at developing analytical and numerical tools to determine automatically the space–time domains in which each model has to be used, taking into account the local accuracy and the characteristics of the flow.

Hélène Mathis and Nicolas Seguin
UPMC Univ Paris 06, UMR 7598, LJLL, F-75005, Paris, France; CNRS, UMR 7598, LJLL, F-75005, Paris, France. e-mail: mathis@ann.jussieu.fr, nicolas.seguin@upmc.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_71, © Springer-Verlag Berlin Heidelberg 2011

The whole procedure must allow:

- to increase the accuracy in the domain where the phenomenon scales are small, by means of using a fine model,
- to improve execution time, by the use of a coarse model elsewhere.

It consists thus in constructing local error estimates of modeling and developing adapted numerical schemes. In the sequel we concentrate on an academic problem, involving: a fine model by means of a *hyperbolic system with relaxation* (the information being contained in the relaxation source term) and a coarse model corresponding to the *associated equilibrium* model. To automatically handle the dynamical decomposition of the computational domain into two sub-domains, we have to provide a criterion. It consists in using an intermediate model from which an error estimate is deduced. Since we are dealing with hyperbolic models with relaxation, we propose to consider the first order corrector resulting from the *Chapman–Enskog expansion* of the relaxation model around an equilibrium state. At the interfaces between fine and coarse models, the coupling strategies we use are based on techniques for thin coupling interfaces developed in [1] (see also references therein).

Section 2 is devoted to the structural study of the hyperbolic relaxation system and its associated equilibrium system of conservation laws. The Chapman–Enskog expansion is recalled for such systems. Section 3 addresses the finite volume schemes we use, the derivation of the discrete estimator and the global algorithm for adaptation. Numerical tests are provided in Section 4 and Section 5 is the conclusion.

Let us emphasize that this work is still under development. Most of our attention is paid in this note to the relevance of the developed error estimator but since our framework is still academic (rather simple 1D models), the CPU time saving is not significant (see Sections 4 and 5). More results will be provided at the conference.

2 Hyperbolic systems with relaxation

In order to simplify the presentation, we consider systems of hyperbolic equations which comply with the form

$$\partial_t U + \partial_x f(U,v) = 0, \tag{1}$$

$$\partial_t v + \partial_x g(U,v) = \frac{1}{\varepsilon}(h(U) - v), \tag{2}$$

where the fluxes $f: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $g: \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^n \to \mathbb{R}$ are smooth functions, defining the evolution of the state vector $U: \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}^n$ and the variable $v: \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$. In the following, we will also use the condensed notations $W = (U,v)^T$, $\mathscr{F} = (f,g)^T$ and $R = (0, h(U) - v)^T$. Note that we only focus on scalar-valued functions v in order to make the notations clearer in the following.

When the relaxation parameter ε tends to 0, the relaxation model (1)–(2) reduces to the following system of n conservation laws:

$$\partial_t U + \partial_x f_e(U) = 0, \quad \text{where } f_e(U) = (U, h(U)). \tag{3}$$

The relaxation process thus determines a local equilibrium state $W_e(U) = (U, h(U))$. Several stability conditions exist to justify the asymptotics $\epsilon \to 0$ [2, 4]. Let $(\lambda_{e,k})_{1\le k\le n}$ be the (ordered) eigenvalues of $\nabla_U f_e(U)$ (i.e. of the equilibrium system (3)) and $(\lambda_k)_{1\le k\le n+1}$ be the (ordered) eigenvalues of $\nabla_W \mathscr{F}(W_e)$ (i.e. of the relaxation system (1)–(2) restricted to equilibrium states). Here, we assume that the so-called subcharacteristic condition is satisfied:

$$\lambda_k \le \lambda_{e,k} \le \lambda_{k+1}, \quad \forall 1 \le k \le n. \tag{4}$$

For more details on the different stability conditions for hyperbolic systems with relaxation, see [2, 4].

2.1 Chapman–Enskog expansion

With the aim of coupling the relaxed model and the equilibrium one in an adaptive way, we want to determine a criterion to move from one model to the other. A natural choice is to consider the first order error resulting from the Chapman–Enskog expansion of the relaxed model around an equilibrium state. The Chapman–Enskog method amounts to considering *smooth* solutions of (1)–(2) near equilibrium which we assume to satisfy

$$v = h(U) + \varepsilon v_1 + \mathscr{O}(\varepsilon^2). \tag{5}$$

Plugging (5) into (1)–(2) leads to

$$\partial_t U + \partial_x(f(U, h(U))) + \varepsilon\partial_x \left(\nabla_2 f(U, h(U))v_1\right) = \mathscr{O}(\varepsilon^2), \tag{6}$$

$$\partial_t h(U) + \partial_x g(U, h(U)) = -v_1 + \mathscr{O}(\varepsilon), \tag{7}$$

where $\nabla_\alpha q$ denotes the derivative of the vector field q w.r.t. its α-th variable, $\alpha = 1, 2$. Multiplying (6) by $\nabla h(U)^T$ and combining with (7) leads to

$$v_1 = \nabla h(U)^T \partial_x \left(f(U, h(U))\right) - \partial_x g(U, h(U)) + \mathscr{O}(\varepsilon). \tag{8}$$

Finally, dropping second order terms with respect to ε yields:

$$\begin{aligned} \partial_t U + \partial_x f(U, h(U)) = -\varepsilon\partial_x\big[&\nabla_2 f(U, h(U)) \\ &\big(\nabla h(U)^T \partial_x f(U, h(U)) - \partial_x g(U, h(U))\big)\big]. \end{aligned} \tag{9}$$

This parabolic system can be interpreted as an intermediate model between the relaxation model (1)–(2) and the equilibrium model (3). Indeed, smooth solutions of (1)–(2) solve (9) up to $\mathscr{O}(\epsilon^2)$ and if the right-hand side in (9) vanishes, one recover (3).

Remark 1. Note that the equivalence between the subcharacteristic condition (4) and the dissipativity of the second order term in (9) only holds in rather classical cases. We refer once again to [2,4] for more details.

2.2 Finite volume methods

We depict now the finite volume schemes used to approximate the equilibrium and the relaxed models. We first consider the equilibrium model (3). We introduce the equilibrium state vector $W_{e,i}^n = (U_i^n, h(U_i^n))$ within each cell $C_i = [x_{i-1/2}, x_{i+1/2}]$. The classical finite volume formulation reads

$$\frac{1}{\Delta x}(U_i^{n+1} - U_i^n) + \frac{1}{\Delta t}(\varphi(U_i^n, U_{i+1}^n) - \varphi(U_{i-1}^n, U_i^n)) = 0. \tag{10}$$

The 2-point numerical flux $\varphi : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}^n$ is consistent with the flux f_e in the sense of finite volume methods, i.e. $\varphi(U,U) = f_e(U)$. We now address the numerical scheme to approximate the relaxation system (1)–(2), for which a splitting strategy between the convective part and the source term has been adopted. We introduce the two numerical fluxes F and G respectively consistent with the fluxes f and g. In a first step the convective part is approximated by

$$\frac{1}{\Delta x}(U_i^{n+1,-} - U_i^n) + \frac{1}{\Delta t}(F(W_i^n, W_{i+1}^n) - F(W_{i-1}^n, W_i^n)) = 0. \tag{11}$$

$$\frac{1}{\Delta x}(v_i^{n+1,-} - v_i^n) + \frac{1}{\Delta t}(G(W_i^n, W_{i+1}^n) - G(W_{i-1}^n, W_i^n)) = 0. \tag{12}$$

Then the value $U_i^{n+1,-}$ is taken as the initial data for solving the source term:

$$U_i^{n+1} = U_i^{n+1,-}, \tag{13}$$

$$v_i^{n+1} = v_i^{n+1,-} + \frac{\Delta t}{\varepsilon}(h(U_i^{n+1}) - v_i^{n+1}). \tag{14}$$

Here, the classical implicit Euler scheme has been chosen in order to ensure the unconditional stability of the second step.

Note that it is natural, at least from the academic viewpoint, to impose the compatibility condition between the numerical fluxes $\varphi(U_l, U_r) = F(W_e(U_l), W_e(U_r))$ and it has been done for the numerical results of Section 4.

3 Model adaptation

This section is devoted to the description of the adaptive coupling procedure from the numerical point of view. First we detail the dynamical cutting of the space–time computational domain. Following Section 2.1 the criterion we choose to realize the cutting corresponds to the first order error coming from the discrete Chapman–Enskog expansion. The global algorithm is described in 3.3.

3.1 Basic principles

We want to provide a criterion which enables to automatically determine the space domains $\mathscr{D}^R(t)$ and $\mathscr{D}^E(t)$ where the fine (that is the relaxation system (1–2)) and the coarse (that is the equilibrium system (3)) models have to be used respectively. These two domains evolve as time t increases, without overlapping in such way that their intersection corresponds to the interfaces where the coupling is performed.

We propose to make the cutting of the space into the sub-domains $\mathscr{D}^E(t)$ and $\mathscr{D}^R(t)$ depend on the first order error ϵv_1 which results from the Chapman–Enskog expansion (9). Let θ being a threshold arbitrarily chosen. We then have:

- The region where $|\epsilon v_1| \leq \theta$ is chosen to be $\mathscr{D}^E(t)$. In that domain the error between the equilibrium model and the relaxed one is assumed to be negligible, so that the coarse model (3) is applied.
- The domain $\mathscr{D}^R(t)$ corresponds to the region where $|\varepsilon v_1| \geq \theta$ and the relaxation model (1)–(2) is solved inside.
- At the interfaces separating the sub-domains $\mathscr{D}^E(t)$ and $\mathscr{D}^R(t)$, a numerical coupling method as those developed in [1,3] is used.

Let us note that several strategies can be applied; in the sequel we give a preference to the state coupling, that is only the value v is transmitted through the interface at each time step. Since the interfaces of coupling are always located in a region where the two models are very close to each other, the different methods of coupling should provide very similar results (see [1,3]).

It is important to note that thanks to (8), the estimator ϵv_1 can be computed from the solution of each model, (1)–(2) and (3), since it only depends on U.

3.2 Estimators for adaptation

We now give the discrete estimator we use to perform the adaptive coupling, following the strategy of the Chapman–Enskog method. First, we take the ansatz

$$v_i = h(U_i) + \varepsilon v_{1,i} + \mathscr{O}(\varepsilon^2)$$

and denote $W_{e,i}^n = W_e(U_i^n)$. Plugging the ansatz in (11)–(13) and (12)–(14) and dropping high order terms yields

$$\begin{aligned} U_i^{n+1} - U_i^n &+ \frac{\Delta t}{\Delta x}\big(F(W_{e,i}^n, W_{e,i+1}^n) - F(W_{e,i-1}^n, W_{e,i}^n)\big) \\ &+ \frac{\varepsilon \Delta t}{\Delta x}\big[\nabla_1 F(W_{e,i}^n, W_{e,i+1}^n)\cdot(0, v_{1,i}^n) + \nabla_2 F(W_{e,i}^n, W_{e,i+1}^n)\cdot(0, v_{1,i+1}^n) \\ &\qquad -\nabla_1 F(W_{e,i-1}^n, W_{e,i}^n)\cdot(0, v_{1,i-1}^n) + \nabla_2 F(W_{e,i}^n, W_{e,i+1}^n)\cdot(0, v_{1,i}^n)\big] = 0. \end{aligned} \tag{15}$$

$$h(U_i^{n+1}) - h(U_i^n) + \frac{\Delta t}{\Delta x}(G(W_{e,i}^n, W_{e,i+1}^n) - G(W_{e,i-1}^n, W_{e,i}^n)) = -\Delta t\ v_{1,i}^{n+1}, \tag{16}$$

Since h is smooth, there exists $\overline{U}(.,.)$ such that $\nabla h(\overline{U}(U^n, U^{n+1}))^T(U^{n+1} - U^n) = h(U^{n+1}) - h(U^n)$. Multiplying (16) by $\nabla h(\overline{U}(U_i^n, U_i^{n+1}))^T$ leads to

$$h(U_i^{n+1}) - h(U_i^n) + \frac{\Delta t}{\Delta x}\nabla h(\overline{U})^T(F(W_{e,i}^n, W_{e,i+1}^n) - F(W_{e,i-1}^n, W_{e,i}^n)) = 0.$$

Combining with (16) provides the following expression of $v_{1,i}^{n+1}$:

$$\begin{aligned} v_{1,i}^{n+1} = \nabla h(\overline{U})\frac{1}{\Delta x}(F(W_{e,i}^n, W_{e,i+1}^n) - F(W_{e,i-1}^n, W_{e,i}^n)) \\ - \frac{1}{\Delta x}(G(W_{e,i}^n, W_{e,i+1}^n) - G(W_{e,i-1}^n, W_{e,i}^n)) \end{aligned} \tag{17}$$

which is the discretisation of (8). Replacing the terms $v_{1,i}^n$ and $v_{1,i-1}^n$ into (15) allows us to determine the discrete counterpart of (9). Note that in practice, the term $\nabla h(\overline{U}(U_i^n, U_i^{n+1}))$ can be approximated by $\nabla h(U_i^n)$ so that the estimator $\epsilon v_{1,i}^{n+1}$, at time t^{n+1}, is an explicit function of the discrete solution $(U_i^n)_{i\in\mathbb{Z}}$, at time t^n.

3.3 The general algorithm

We now detail the general algorithm of the dynamical coupling between the fine and the coarse models. Let $(W_i^n)_{i\in\mathbb{Z}}$ be a sequence known at time t^n to be advanced to time t^{n+1}. The algorithm follows the steps:

- For all $i \in \mathbb{Z}$, compute in cell C_i the numerical error $e_i^{n+1} := \epsilon v_{1,i}^{n+1}$ using (17)
- For all $i \in \mathbb{Z}$, if $[|e_i^{n+1}| > \theta]$ then
  $C_i \in \mathscr{D}^R(t^n)$

```
    Else
        $C_i \in \mathscr{D}^E(t^n)$
• At this stage, $\mathscr{D}^R(t^n) \cup \mathscr{D}^E(t^n) = \mathbb{R}$. For all $i \in \mathbb{Z}$:
    – If $[C_{i_1}, C_i, C_{i+1} \in \mathscr{D}^R(t^n)]$ (resp. $\in \mathscr{D}^E(t^n)$) then
        Compute $W_i^{n+1}$ using the numerical scheme (11–14)
        (resp. compute $W_i^{n+1} = W_e(U_i^{n+1})$ using the numerical scheme (10))
    – Else
        Compute $W_i^{n+1}$ using the state coupling method described in [1,3]
```

Besides let us note that the estimator we use is not exactly an *a posteriori* error estimate since the adaptation process add a numerical error. The study of this error estimate is an ongoing work.

4 Numerical experiments

We now present some numerical results in order to illustrate the reliability of the coupling procedure, using the Rusanov scheme for the approximation of each model. The problem we address corresponds to a fluid flow problem governed by the relaxation system

$$\partial_t \tau - \partial_x u = 0, \tag{18}$$

$$\partial_t u + \partial_x \Pi = 0, \tag{19}$$

$$\partial_t \mathscr{T} = \frac{1}{\varepsilon}(\tau - \mathscr{T}) \tag{20}$$

which is derived from the works of Suliciu [7] but also corresponds to Chaplygin gas (see for instance [6]). The state variable τ and u stand for the specific volume and the velocity while $\mathscr{T}$ is a perturbed specific volume. The extended pressure law Π is defined by $\Pi(\tau, \mathscr{T}) = p(\mathscr{T}) + a^2(\mathscr{T} - \tau)$ where p follows a perfect gas law $p = p(\tau) = p^{-\gamma}$, $\gamma > 1$. The associated equilibrium system is obtained setting $\mathscr{T} = \tau$ and is the so-called p-system

$$\partial_t \tau - \partial_x u = 0, \tag{21}$$

$$\partial_t u + \partial_x p = 0. \tag{22}$$

The constant a is assumed to satisfy the Whitham's condition $a^2 > \max_s(-p'(s))$, which ensures that the subcharacteristic condition (4) is satisfied. Plugging the expansion $\mathscr{T} = \tau + \varepsilon\mathscr{T}_1 + \mathscr{O}(\varepsilon^2)$ into (19) yields the parabolic equation, which corresponds to (9),

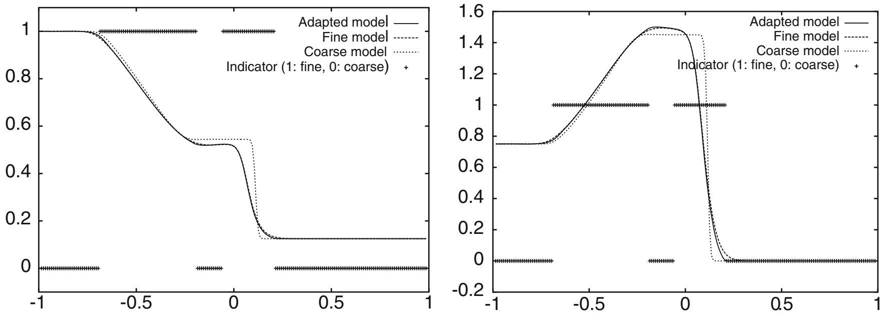


Fig. 1 Density $1/\tau$ (left) and velocity (right) with 200 cells at time $T = 0.5$. The indicator corresponds to the characteristic function of $\mathscr{D}^R(T)$

$$\partial_t u + \partial_x p(\tau) = \varepsilon \partial_x \big(\partial_x u (p'(\tau) + a^2)\big). \tag{23}$$

Figure 1 presents the solution of test case with $\gamma = 1.4$ and $a = 1.5$. The initial data are $\tau_L = 1$, $u_L = 0.75$, $\tau_R = 8$, $u_R = 0$ and the discontinuity is applied at $x = 0$. The relaxation parameter is $\varepsilon = 10^{-6}$. The mesh contains 200 cells and threshold for the adaptation is $\theta = 0.5$. One may check that our method of adaptation only uses the fine model in the regions of large variations of the solution. The results are very close to those with the fine model. One may also note that the results of coarse model are sensibly different: less diffusion and a different intermediate state.

5 Conclusion

In the frame of hyperbolic systems with relaxation, we have proposed a new algorithm for dynamical adaptation of models, based on the Chapman–Enskog expansion. It enables to quantify the difference between a fine model and a coarse model, from the continuous and the discrete points of view. The global algorithm of adaptation is based on a series of works on interface coupling of hyperbolic models and is easy to implement. This is a preliminary work but the first results are encouraging. We are aware that the presented test case is very academic, but it only aims at illustrating the relevance of our estimator. Besides, the fine model (18–20) and its numerical resolution are rather classical and since the space domain is 1D, comparison of CPU times between a computation with the fine model in the whole domain and a computation with our algorithm of adaptation would be meaningless. More complex models (such that nonlinear relaxation terms coming from models of phase transition) and 2D computations will be presented during the conference.

Acknowledgements The first author is supported by the LRC Manon (Modélisation et Approximation Numérique Orientées pour l'énergie Nucléaire — CEA/DM2S-LJLL). The authors would like to thank C. Cancès, F. Coquel and E. Godlewski for their useful comments. The authors also thank the referees who gave them the opportunity to deeply improve the quality of this paper.

References

1. Ambroso, A., Chalons, C., Coquel, F., Godlewski, E., Lagoutière, F., Raviart, P.A., Seguin, N.: The coupling of homogeneous models for two-phase flows. Int. J. Finite Volumes **4**(1), 1–39 (2007)
2. Bouchut, F.: A reduced stability condition for nonlinear relaxation to conservation laws. J. Hyperbolic Differ. Equ. **1**(1), 149–170 (2004)
3. Caetano, F.: Sur certains problèmes de linéarisation et de couplage pour les systèmes hyperboliques non linéaires. Ph.D. thesis, Université Pierre et Marie Curie-Paris6, France (2006)
4. Chen, G.Q., Levermore, C.D., Liu, T.P.: Hyperbolic conservation laws with stiff relaxation terms and entropy. Comm. Pure Appl. Math. **47**(6), 787–830 (1994)
5. Rusanov, V.V.: The calculation of the interaction of non-stationary shock waves with barriers. Ž. Vyčisl. Mat. i Mat. Fiz. **1**, 267–279 (1961)
6. Serre, D.: Multidimensional shock interaction for a Chaplygin gas. Arch. Ration. Mech. Anal. **191**(3), 539–577 (2009)
7. Suliciu, I.: On the thermodynamics of rate-type fluids and phase transitions. I. Rate-type fluids. Internat. J. Engrg. Sci. **36**(9), 921–947 (1998)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Inflow-Implicit/Outflow-Explicit Scheme for Solving Advection Equations

Karol Mikula and Mario Ohlberger

Abstract We present new method for solving non-stationary advection equations based on the finite volume space discretization and the semi-implicit discretization in time. Its basic idea is that outflow from a cell is treated explicitly while inflow is treated implicitly. Since the matrix of the system in this new I^{2}OE method is determined by the inflow fluxes it is an M-matrix yielding favourable solvability and stability properties. The method allows large time steps at a fixed spatial grid without losing stability and not deteriorating precision which makes it attractive for practical applications. Our new method is exact for any choice of a discrete time step on uniform rectangular grids in the case of constant velocity transport of quadratic functions in any dimension. We show that it is formally second order accurate in space and time for 1D advection problems with variable velocity and numerical experiments indicates its second order accuracy for smooth solutions in general.

Keywords Advection equation, semi-implicit scheme, finite volume method
MSC2010: 35L04, 65M08, 65M12

1 Introduction

In this paper we present the inflow-implicit/outflow-explicit (I^{2}OE) method for solving variable velocity advection equations of the form

$$u_t + \mathbf{v} \cdot \nabla u = 0 \tag{1}$$

Karol Mikula
Department of Mathematics, Faculty of Civil Engineering, Slovak University of Technology, Radlinského 11, 81368 Bratislava, Slovakia, e-mail: mikula@math.sk

Mario Ohlberger
Institut für Numerische und Angewandte Mathematik, Universität Münster, Einsteinstr. 62, D-48149 Münster, Germany, e-mail: mario.ohlberger@uni-muenster.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_72, © Springer-Verlag Berlin Heidelberg 2011

where $u \in \mathbb{R}^d \times [0, T]$ is the unknown function and $\mathbf{v}(x)$ is a vector field. The basic idea of our new method is that outflow from a cell is treated explicitly while inflow is treated implicitly. Such an approach is natural, since we know what is flowing out from a cell at an old time step $n-1$ but we leave the method to resolve a system of equations determined by the inflows to obtain a new value in the cell at time step n. Since the matrix of the system is determined by the inflow fluxes it is an M-matrix for Voronoi like grids and thus it has favourable discrete minimum-maximum properties. Consequently, the method allows large time steps at a fixed spatial grid without losing stability. Interestingly, the new I^{2}OE scheme is exact on rectangular grids for constant velocity transport of quadratic polynomials in any dimension and for any length of a time step. In general, it is second order accurate for smooth solutions, both for variable velocity and nonlinear advection problems [5]. A comparison with the second order Lax-Wendroff method for variable velocity shows good properties of the new scheme with respect to precision and CPU times. In [5], the I^{2}OE method was introduced in more general settings where $\mathbf{v} = \mathbf{v}(x, u, \nabla u)$. The semi-implicit forward-backward diffusion level set approach for motion in normal direction [4] is its special case. The variable and nonlinear velocity fields to which our method can be successfully applied arise in many applications, e.g. in level set methods and other transports with non-divergence free velocities and nonlinear conservation laws or in image segmentation by the active contours.

2 The inflow-implicit/outflow-explicit scheme

Let us consider equation (1) in a bounded polygonal domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and time interval $[0, T]$. Let $\mathcal{Q}_h$ denote a primal polygonal partition of Ω. Let p be a finite volume (cell) of a corresponding dual Voronoi tessellation $\mathcal{T}_h$ with measure m_p and let e_{pq} be an edge between p and q, $q \in N(p)$, where $N(p)$ is a set of neighbouring finite volumes (i.e. $\bar{p} \cap \bar{q}$ has nonzero $(d-1)$-dimensional measure). Let c_{pq} be the length of e_{pq} and n_{pq} be the unit outer normal vector to e_{pq} with respect to p. We shall consider $\mathcal{T}_h$ to be an admissible mesh in the sense of [1], i.e., there exists a representative point x_p in the interior of every finite volume p such that the joining line between x_p and x_q, $q \in N(p)$, is orthogonal to e_{pq}. We denote by x_{pq} the intersection of this line segment with the edge e_{pq}. The length of this line segment is denoted by d_{pq}, i.e. $d_{pq} := |x_q - x_p|$. As we have build $\mathcal{T}_h$ based on the primal mesh $\mathcal{Q}_h$, we assume that the points x_p coincide with the vertices of $\mathcal{Q}_h$. Let us denote by u_p a (constant) value of the solution in a finite volume p computed by the scheme. For the solution representation inside the finite volume p we use either this value u_p or a reconstructed (but again constant) value denoted by $\overline{u}_p$. A constant value of the solution assigned to the edge e_{pq} (given again by a reconstruction) is denoted by $\overline{u}_{pq}$. Let us rewrite (1) in the formally equivalent form with conserving and non-conserving parts [2]

$$u_t + \nabla \cdot (\mathbf{v}u) - u\nabla \cdot \mathbf{v} = 0. \tag{2}$$

Integrating (2) over a finite volume p then yields

$$\int_p u_t\, dx + \int_p \nabla \cdot (\mathbf{v}u)\, dx - \int_p u\nabla \cdot \mathbf{v}\, dx = 0.$$

Applying the divergence theorem and using constant representations of the solution on the cell p, denoted by $\overline{u}_p$, and on the cell interfaces e_{pq}, denoted by $\overline{u}_{pq}$, we get

$$\int_p u_t\, dx + \sum_{q\in N(p)} \overline{u}_{pq} \int_{e_{pq}} \mathbf{v}\cdot n_{pq}\, ds - \overline{u}_p \sum_{q\in N(p)} \int_{e_{pq}} \mathbf{v}\cdot n_{pq}\, ds = 0.$$

If we denote the fluxes in the inward normal direction to the finite volume p by

$$\bar{v}_{pq} = -\int_{e_{pq}} \mathbf{v}\cdot n_{pq}\, ds, \tag{3}$$

we finally arrive at the equation

$$\int_p u_t\, dx + \sum_{q\in N(p)} \bar{v}_{pq}(\overline{u}_p - \overline{u}_{pq}) = 0. \tag{4}$$

The novelty of our scheme is to split the resulting fluxes into the corresponding inflow and outflow parts to the cell p. This is done by defining

$$a_{pq}^{in} = \max(\bar{v}_{pq}, 0), \quad a_{pq}^{out} = \min(\bar{v}_{pq}, 0). \tag{5}$$

We then approximate u_t by the time difference $\frac{u_p^n - u_p^{n-1}}{\tau}$, where τ is a uniform time step size, and take the inflow parts implicitly and the outflow parts explicitly in (4). This yields the following system of equations for the finite volume solution u_p^n, $p \in \mathscr{T}_h$ at the n-th discrete time step, representing the general I^{2}OE scheme:

$$m_p u_p^n + \tau \sum_{q\in N(p)} a_{pq}^{in}(\overline{u}_p^n - \overline{u}_{pq}^n) = m_p u_p^{n-1} - \tau \sum_{q\in N(p)} a_{pq}^{out}(\overline{u}_p^{n-1} - \overline{u}_{pq}^{n-1}). \tag{6}$$

The most natural choice for reconstructions $\overline{u}_p^n$ and $\overline{u}_{pq}^n$ at any time step n (i.e. old and new time steps) is given by $\overline{u}_p^n = u_p^n$, $\overline{u}_{pq}^n = \frac{1}{2}(u_p^n + u_q^n)$ and leads to the basic I^{2}OE scheme:

$$m_p u_p^n + \frac{\tau}{2} \sum_{q\in N(p)} a_{pq}^{in}(u_p^n - u_q^n) = m_p u_p^{n-1} - \frac{\tau}{2} \sum_{q\in N(p)} a_{pq}^{out}(u_p^{n-1} - u_q^{n-1}). \tag{7}$$

The equation (4) has the form of a discretization of a diffusion equation, where $\bar{v}_{pq}$ would represent the so-called transmissive coefficients (integrated diffusion fluxes divided by distances between cell centers). In standard forward diffusion all these coefficients are strictly positive which leads to a weighted averaging of the solution and the implicit schemes are natural in this case. On the other hand the negative coefficients would correspond to backward diffusion in which case information propagates outside the cell and explicit schemes are thus natural. In our case the sign of the coefficients is given by the inflow or outflow character of the cell boundary and the inflow-implicit/outflow-explicit approach is thus natural. It is also well-known that in the second order schemes for solving advection problems one can identify the "forward diffusion" part (like the first order upwinding) and the "backward diffusion" part given by the additional sharpening terms coming (sometimes surprisingly) from the second order Taylor's expansions, cf. the Lax-Wendroff scheme [3]. In our method this splitting arises naturally, gives second order accuracy and when treating it semi-implicitly it brings significant improvements in stability of computations.

Let us present the I^{2}OE scheme for 1D variable velocity equation $u_t + v(x) u_x = 0$, which will be used in numerical computations of Section 4. Let p_i be the cell with the spatial index i, length h, center point x_i, left border $x_{i-\frac{1}{2}}$ and right border $x_{i+\frac{1}{2}}$. Let us denote u_i^n the value of the numerical solution at time step n and $\overline{u}_i^n, \overline{u}_{i-\frac{1}{2}}^n$ the reconstructed values. We define

$$a_{i-\frac{1}{2}}^{in} = \max(v(x_{i-\frac{1}{2}}), 0), \;\; a_{i-\frac{1}{2}}^{out} = \min(v(x_{i-\frac{1}{2}}), 0),$$
$$a_{i+\frac{1}{2}}^{in} = \max(-v(x_{i+\frac{1}{2}}), 0), \;\; a_{i+\frac{1}{2}}^{out} = \min(-v(x_{i+\frac{1}{2}}), 0)\,,$$

and if we use the reconstructions $\overline{u}_i^n = u_i^n, \overline{u}_{i-\frac{1}{2}}^n = \frac{1}{2}(u_i^n + u_{i-1}^n)$ in both new and old time steps, the basic one-dimensional I^{2}OE scheme has the following form

$$\begin{aligned} u_i^n + \frac{\tau}{2h} a_{i-\frac{1}{2}}^{in}(u_i^n - u_{i-1}^n) + \frac{\tau}{2h} a_{i+\frac{1}{2}}^{in}(u_i^n - u_{i+1}^n) = u_i^{n-1} \\ - \frac{\tau}{2h} a_{i-\frac{1}{2}}^{out}(u_i^{n-1} - u_{i-1}^{n-1}) - \frac{\tau}{2h} a_{i+\frac{1}{2}}^{out}(u_i^{n-1} - u_{i+1}^{n-1})\,. \end{aligned} \tag{8}$$

The scheme (8) requires to solve a tridiagonal system in every time step which is done by using the standard tridiagonal solver (also called the Thomas algorithm). In practice, the I^{2}OE scheme allows to use much larger time steps without losing L$_\infty$-stability than given by a standard CFL condition for explicit schemes, cf. Section 3. However, the "backward diffusion" (outflow) explicit part is not necessarily always dominated by the implicit part in the basic form of the scheme (8). Some oscillations (not unboundedly growing in time) may arise e.g. on coarse grids or in solutions tending to a shock. One possibility is to leave the method with oscillations and remove them at the end of computations using e.g. some edge preserving filters. Another approach is to supress the oscillations during the computation. In our scheme, one can use an averaging (by a larger stencil) in the reconstruction of $\overline{u}_p^{n-1}$, similarly to the FBD schemes from [4], or to modify the "backward diffusion" part on the right hand side of (8) by using the standard limiters, for details see [5].

Theorem 1. *Let us consider the equation* (1) *in 1D with constant velocity* v *and* I^{2}OE *scheme* (8) *on uniform grid. If the initial condition is given by a second order polynomial, then the scheme gives the exact solution for any choice of time step.*

Proof. The initial condition has the form $u_0(x) = ax^2 + bx + c$ and the exact solution is given by $u(x, \tau) = u^0(x - v\tau)$. For $v > 0$ the scheme (8) takes the form

$$u_i^n + \frac{\tau v}{2h}(u_i^n - u_{i-1}^n) = u_i^{n-1} - \frac{\tau(-v)}{2h}(u_i^{n-1} - u_{i+1}^{n-1}) \tag{9}$$

One can easily check that if we plug the exact values in grid points x_i, x_{i-1}, x_{i+1} at time steps $n = 1$ and $n - 1 = 0$, namely

$$u_i^{n-1} = ax_i^2 + bx_i + c, \;\; u_{i+1}^{n-1} = a(x_i + h)^2 + b(x_i + h) + c, \tag{10}$$
$$u_i^n = a(x_i - v\tau)^2 + b(x_i - v\tau) + c, \; u_{i-1}^n = a(x_i - h - v\tau)^2 + b(x_i - h - v\tau) + c,$$

into the scheme (9), we get true identity, and the same we obtain for $v < 0$. □

It is also possible to make similar considerations as above in higher dimensional case for uniform rectangular grids and constant velocity vector field . One can plug a general 2D or 3D quadratic polynomial as initial condition and the corresponding exact solution at time τ into the I^{2}OE scheme (7), use a symbolic computational software like the Mathematica, and check that the scheme is exact in such situations.

Theorem 2. *Let us consider the equation* (1) *in 1D with variable velocity* $v(x) \geq 0$ *(or* $v(x) \leq 0$*) and the* I^{2}OE *scheme* (8) *on a uniform grid. Then the scheme is formally second order and the consistency error is of order* $\mathcal{O}(h^2) + \mathcal{O}(\tau h) + \mathcal{O}(\tau^2)$.

Proof. We write our transport equation as $\partial_t u + f(v, \partial_x u) = 0$ with $f(v, \partial_x u) := v(x)\partial_x u$ and let $v(x) \geq 0$. We will use notations $u^n := u(t^n)$, $f^n := f(v, \partial_t u^n)$. The Taylor expansion in time yields

$$u^n = u^{n-1} + \tau\partial_t u^{n-1} + \frac{\tau^2}{2}\partial_t^2 u^{n-1} + \mathcal{O}(\tau^3), \quad u^{n-1} = u^n - \tau\partial_t u^n + \frac{\tau^2}{2}\partial_t^2 u^n + \mathcal{O}(\tau^3).$$

Subtracting these two equations we derive relation

$$u^n - u^{n-1} = \frac{\tau}{2}(\partial_t u^n + \partial_t u^{n-1}) + \frac{\tau^2}{4}(\partial_t^2 u^{n-1} - \partial_t^2 u^n) + \mathcal{O}(\tau^3). \tag{11}$$

We can see that the second term on the right hand side is also $\mathcal{O}(\tau^3)$ and using the equation $\partial_t u + f(v, \partial_x u) = 0$, we get for the first term of the right hand side

$$I = \frac{\tau}{2}(\partial_t u^n + \partial_t u^{n-1}) = -\frac{\tau}{2}(f^n + f^{n-1})\,. \tag{12}$$

Using the notation $f_i := f(x_i) = v(x_i)\partial_x u(x_i)$, by the Taylor expansion in space we have (for $v(x) \geq 0$)

$$f_{i-1/2}^n = f_i^n - \frac{h}{2}\partial_x f_i^n + \mathcal{O}(h^2), \quad f_{i+1/2}^{n-1} = f_i^{n-1} + \frac{h}{2}\partial_x f_i^{n-1} + \mathcal{O}(h^2) \tag{13}$$

or (for $v(x) \leq 0$)

$$f_{i-1/2}^{n-1} = f_i^{n-1} - \frac{h}{2}\partial_x f_i^{n-1} + \mathcal{O}(h^2), \quad f_{i+1/2}^n = f_i^n + \frac{h}{2}\partial_x f_i^n + \mathcal{O}(h^2)\,. \tag{14}$$

We continue (for $v(x) \geq 0$) and using (12)-(13) we derive

$$I_i = -\frac{\tau}{2}(f_i^n + f_i^{n-1}) = -\frac{\tau}{2}\Big(f_{i-1/2}^n + f_{i+1/2}^{n-1} + \frac{h}{2}(\partial_x f_i^n - \partial_x f_i^{n-1}) + \mathcal{O}(h^2)\Big).$$

The second term in the brackets on the right hand side is of order $\mathcal{O}(\tau h)$ and we shall analyse the first one. We know that

$$\partial_x u_{i-1/2}^n = \frac{1}{h}(u_i^n - u_{i-1}^n) + \mathcal{O}(h^2), \quad \partial_x u_{i+1/2}^{n-1} = \frac{1}{h}(u_{i+1}^{n-1} - u_i^{n-1}) + \mathcal{O}(h^2)$$

and resubstituting for $f_{i-1/2}^n = v_{i-1/2}\partial_x u_{i-1/2}^n$ and $f_{i+1/2}^{n-1} = v_{i+1/2}\partial_x u_{i+1/2}^{n-1}$ we get

$$I_i = -\frac{\tau}{2}\Big(v_{i-1/2}\frac{1}{h}(u_i^n - u_{i-1}) + v_{i+1/2}\frac{1}{h}(u_{i+1}^{n-1} - u_i^{n-1})\Big) + \mathcal{O}(\tau^2 h) + \mathcal{O}(\tau h^2). \tag{15}$$

From (11) and (15) we finally get

$$\begin{aligned} u_i^n - u_i^{n-1} = &- \frac{\tau}{2}\Big(\frac{v_{i-1/2}}{h}(u_i^n - u_{i-1}^n) + \frac{v_{i+1/2}}{h}(u_{i+1}^{n-1} - u_i^{n-1})\Big) \\ &+ \mathcal{O}(\tau^2 h) + \mathcal{O}(\tau h^2) + \mathcal{O}(\tau^3) \end{aligned}$$

where we recognize the scheme (8) for $v(x) \geq 0$, cf. also (9), and dividing by τ we get the consistency error of the I^{2}OE scheme stated in the theorem. □

3 Numerical experiments

First, let us consider 1D equation (1) with $v(x) \equiv 1$ in interval $\Omega = (-1, 1)$ and time interval $I = (0, T)$, $T = 1$. Let the initial condition u_0 be given by a quadratic polynomial $u_0(x) = 1 - \frac{1}{2}(x^2 - x)$. The exact solution is given $u(x,t) = u_0(x - vt)$. We solve this problem numerically using the exact Dirichlet boundary conditions and compare the results of the I^{2}OE method (8), the standard Lax-Wendroff and explicit up-wind schemes [3] with the exact solution. In all experiments we used

Table 1 Report on the $L_2(I, L_2)$ errors of the I^{2}OE method, the Lax-Wendroff scheme, and the explicit up-wind scheme for the initial quadratic polynomial and for various choices of time step. We note that all the methods are exact for $\tau = h$

n	$\tau = h/2$	NTS	I^{2}OE	Lax-Wendroff	Up-wind
20	0.05	20	$3.7\ 10^{-16}$	$5.1\ 10^{-17}$	$1.83\ 10^{-2}$
40	0.025	40	$8.0\ 10^{-16}$	$7.5\ 10^{-17}$	$8.99\ 10^{-3}$
80	0.0125	80	$1.1\ 10^{-15}$	$8.3\ 10^{-17}$	$4.45\ 10^{-3}$
160	0.00625	160	$2.4\ 10^{-15}$	$9.9\ 10^{-17}$	$2.22\ 10^{-3}$
n	$\tau = 2h$	NTS	I^{2}OE	Lax-Wendroff	Up-wind
20	0.2	5	$2.1\ 10^{-16}$	$1.1\ 10^{-11}$	$5.02\ 10^{-2}$
40	0.1	10	$2.1\ 10^{-16}$	$1.4\ 10^{-9}$	0.641
80	0.05	20	$3.9\ 10^{-16}$	0.466	$3.8\ 10^{+3}$
160	0.025	40	$5.7\ 10^{-16}$	$1.6\ 10^{+16}$	$1.3\ 10^{+12}$
160	$\tau = 10h = 0.125$	8	$2.5\ 10^{-15}$	–	–
160	$\tau = 40h = 0.5$	2	$1.7\ 10^{-15}$	–	–
160	$\tau = 80h = 1$	1	$2.6\ 10^{-15}$	–	–

increasing number n of finite volumes discretizing Ω, $h = 2/n$, and we consider various choices of time step τ and corresponding number of time steps NTS. In Table 1 we report the errors in $L_2(I, L_2)$ norm for all the methods. As one can see, the I^{2}OE method is exact for any relation between space and time step, see Theorem 1, and one can use extremely large (e.g. just one time step $\tau = T$) without any deterioration of the numerical result. Here the errors are comparable to machine precision, they are not exact zeros because we have to solve a tridiagonal system in every time step yielding some rounding errors which, however, do not propagate even in a long run. The Lax-Wendroff method, as the second order, is exact for any quadratic initial function whenever it is stable, i.e. $\tau \le h$. For Courant numbers larger than 1, one can see instabilities in the third and 4th rows of Table 1, when $\tau = 2h$ and grid is refined. The explicit upwind scheme is the first order and exact for any initial data only if the relation $\tau = h$ is fulfilled. Its first order accuracy can be seen for $\tau = h/2$, and oscillations occur soon for $\tau > h$ as documented in Table 1.

Next, let us consider an example with variable velocity field $v(x) = -\sin(x)$ and let the initial profile be given by $u_0(x) = \sin(x)$, $\Omega = (-1, 1)$ and $I = (0, T)$, $T = 1$. The exact solution can be derived by the method of characteristics and is given as $u(x,t) = u_0(\frac{2}{\pi}\mathrm{arctg}(e^{\pi t}\mathrm{tg}(\frac{\pi x}{2})))$. We compare the precision and CPU-time of the I^{2}OE and the Lax-Wendroff scheme [3]. In the solutions a strong peak is formed at $T = 1$, see Fig. 1. Both schemes are stable with slight overshoot and undershoot in the result by the Lax-Wendroff scheme on coarser grids. No overshoot or undershoot is observed for the I^{2}OE scheme, cf. Fig. 1. Figure 2 shows log-log plots of CPU time versus error of the schemes. We can see superior behavior of the I^{2}OE scheme in this example with considerable speed-up when using larger time steps up to 4-8 times exceeding the CFL condition, which must be respected in the

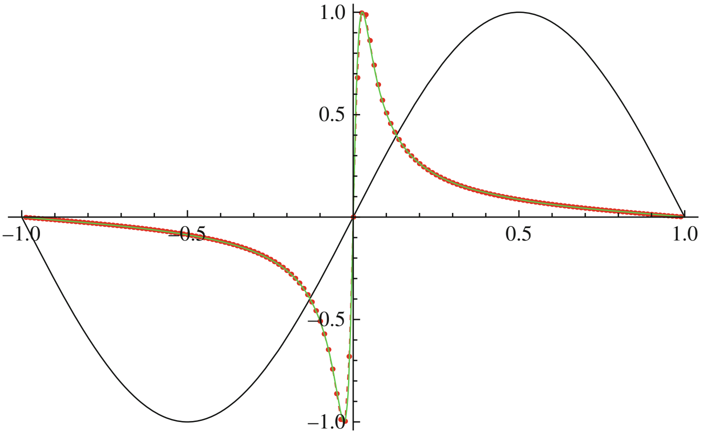


Fig. 1 The result of the I^{2}OE scheme (up, red points) at time $T = 1$, computed with $n = 160$ and $\tau = h$. By green line we plot the exact solution at T and by black line the initial condition

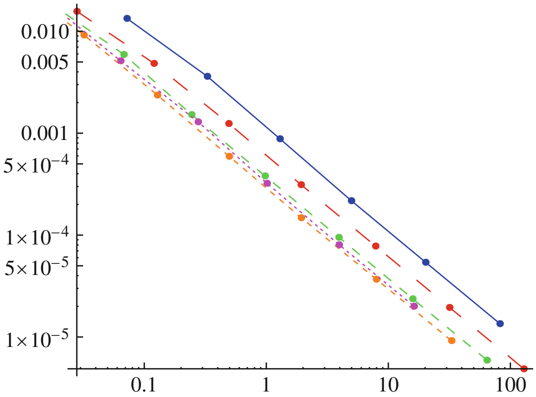


Fig. 2 CPU versus L$_2$(I,L$_2$)-error for the Lax-Wendroff method (blue solid line) and for the I^{2}OE scheme with CFL=1 (red large dashing, $\tau = h$), CFL=2 (green medium dashing, $\tau = 2h$), CFL=4 (orange small dashing, $\tau = 4h$) and CFL=8 (magenta tiny dashing, $\tau = 8h$) for the experiment from Fig. 1. The plots indicate that I^{2}OE scheme is about 4–times faster in order to get the same L$_2$(I,L$_2$)-error

Lax-Wendroff scheme. In this case both schemes are second order accurate which holds true for any time step size of the I^{2}OE scheme.

Further 1D and 2D numerical experiments are reported in [5] showing the second order convergence of the I^{2}OE method for any choice of the time steps. This is the main advantage of the new scheme when comparing with standard explicit second order methods, or, when using limiters, in comparison with the so-called high resolution methods for solving advection equations.

Acknowledgements The work of the first author was supported by the grants APVV-0351-07 and VEGA 1/0269/09.

References

1. Eymard, R., Gallouet, T., & Herbin R.: The finite volume methods, Handbook for Numerical Analysis, 2000, Vol. 7 (Ph. Ciarlet, J. L. Lions, eds.), Elsevier.
2. Frolkovič, P., Mikula, K.: Flux-based level set method: a finite volume method for evolving interfaces, Applied Numerical Mathematics, Vol. 57, No. 4 (2007) pp. 436-454.
3. LeVeque R.J., Finite Volume Methods for Hyperbolic Problems, Cambridge Texts in Applied Mathematics. Cambridge University Press, 2002.
4. Mikula, K., Ohlberger, M.: A new level set method for motion in normal direction based on a semi-implicit forward-backward diffusion approach, SIAM J. Scientific Computing, Vol. 32 , No. 3 (2010) pp. 1527-1544.
5. Mikula, K., Ohlberger, M.: A New Inflow-Implicit/Outflow-Explicit Finite Volume Method for Solving Variable Velocity Advection Equations, Preprint 01/10 - N, Angewandte Mathematik und Informatik, Universität Münster, June 2010, pp.1-20

The paper is in final form and no similar paper has been or is being submitted elsewhere.

4D Numerical Schemes for Cell Image Segmentation and Tracking

K. Mikula, N. Peyriéras, M. Remešíková, and M. Smíšek

Abstract The paper introduces new techniques for 4D (space-time) segmentation and tracking in time sequences of 3D images of zebrafish embryogenesis. Instead of treating each 3D image individually, we consider the whole time sequence as a single 4D image and perform the extraction of objects and cell tracking in four dimensions. The segmentation of the spatiotemporal objects corresponding to the time evolution of the individual cells is realized by using the generalized subjective surface model [1], that is discretized by a new 4D finite volume scheme. Afterwards, we use the distance functions to the borders of the segmented spatiotemporal objects and to the initial cell center positions in order to backtrack the cell trajectories that can be understood as 4D parametrized curves. The distance functions are obtained by numerical solution of the time relaxed eikonal equation.

Keywords Cell tracking, image segmentation, finite volume method
MSC2010: 65M08, 65-06, 92B05

1 Introduction

Cell tracking, i.e. finding the space-time trajectories and moments of divisions of the cells of a developing organism, is one of the most interesting challenges of modern biology. A reliable backward tracking can answer a lot of questions concerning the origin and formation of cell structures and organs, the global and local movement

Karol Mikula, Mariana Remešíková, and Michal Smíšek
Slovak University of Technology, Radlinského 11, 81368 Bratislava, Slovakia
e-mail: mikula@math.sk,remesikova@math.sk,michal.smisek@gmail.com

Nadine Peyriéras
CNRS-DEPSN, Avenue de la Terasse, 91198, Gif sur Yvette, France
e-mail: nadine.peyrieras@inaf.cnrs-gif.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_73, © Springer-Verlag Berlin Heidelberg 2011

of the cells, the cell division rate and localization etc. They all are fundamental questions of developmental biology.

In this paper, we introduce the basic concepts of a novel technique that can be used for the cell tracking from time sequences of 3D images of embryogenesis. A cell can be represented by the surface of its nucleus or by its membrane, depending on the type of images we have at disposal. The time evolution of a cell can be seen as spatiotemporal tube whose cross-section by a chosen time hyperplane corresponds to the 3D representation of the cell at the selected time. This 4D tube is bifurcated in the time moments when the cell undergoes division. Thus, we get a tree-like object corresponding to any cell present at the beginning of the time sequence. In order to track a cell, we need to descend from its current position to the root of the tree in which it is situated. This implies that the tracking procedure consists in solving the following two problems:

1. Segmentation of the 4D cell evolution trees from the spatiotemporal image.
2. Finding the way to the root of a tree from any of its inner points.

In our paper, we discuss the solution of both of these problems. We test our methods on artificial data and on time sequences of 3D images corresponding to the zebrafish embryonic development obtained by a confocal microscope. In order to be able to apply the described methods, we need to have at disposal the approximate positions of cell centers for all cells visible in the images. For the artificial data, these points are known by construction and for the zebrafish images, the approximate cell centers are computed by a level set object detection technique [1].

In order to solve the first problem, we apply the generalized subjective surface model [1, 6]

$$u_t - w_a \nabla g \cdot \nabla u - w_c g |\nabla u| \nabla . \left(\frac{\nabla u}{|\nabla u|} \right) = 0 \,, \tag{1}$$

solved in the domain $[0, T_S] \times \Omega$ where $\Omega \subset R^4$ is the spatiotemporal image domain, i.e. the whole time sequence of 3D images. We set $u(0, x) = u_0(x)$ and we consider the zero Dirichlet boundary condition on $\partial\Omega$. The edge detector function $g = g(|\nabla G_\sigma * I_0|)$, I_0 being the 4D image intensity function, and w_a and w_c are the advection and curvature parameters of the model that determine the way the function u is evolving [1]. The desired cell evolution tree segmentation is represented by a selected isosurface of the function $u(x, T_S)$. We would like to point out the importance of performing this segmentation in 4D. Although the cell evolution tree object could be more easily composed of less time and memory consuming 3D cell segmentations, this could lead to spurious interruptions of the cell trajectories in the points where the cell center and consequently the corresponding cell segmentation is missing for some reason. Looking for a whole spatiotemporal structure rather than a composition of 3D objects makes the procedure more robust and resistant to the possible errors of the center detection technique.

Having segmented the tree object, we now want to find a way down to its root from any of its inner points. Since the root can be represented by the center of the root cell, a reasonable descend direction indicator could be the gradient of the distance function d_1 to this center computed inside the segmented 4D object. However, this might not be sufficient. In real data containing a large number of cells, we can observe that the trees corresponding to different root cells are not always perfectly isolated. In order to prevent dropping into a wrong tree, it is desirable to descend along the center line of the tree branches. For this purpose, we compute the distance function to the border of the 4D tree, denoted by d_2, whose negative gradient leads us towards the center line that we want to follow. The distance function to a set Ω_0 can be computed by solving the time relaxed eikonal equation

$$d_t + |\nabla d| = 1 \tag{2}$$

in the domain $[0, T_D] \times \Omega_D$. In our case, Ω_D is the inner part of the segmented tree object, i.e. the part where $u(x, T_S) > V$, V being the isosurface value chosen to represent the segmentation result. The equation (2) has to be coupled with a Dirichlet type condition

$$d(x,t) = 0, \quad x \in \Omega_0 \tag{3}$$

where Ω_0 can represent the root point of the tree or its boundary, i.e. the set of points where $u(x, T_S) = V$.

The descend to the root of the tree is performed as follows. Given an arbitrary point (doxel center) $[x_1, x_2, x_3, x_4]$ inside the tree, we move to the center of the nearest doxel in the direction given by ∇d_1. Supposing that x_4 represents the time dimension of the 4D data, we repeat this step until we drop to the level $x_4 - 1$. After, we move in the direction of $-\nabla d_2$ until we find the nearest ridge point of d_2. Thus we are situated on the center line of the current branch of the tree. From there we repeat the whole procedure until we descend to the level $x_4 = 0$, resp. to the root of the tree.

2 Discretization of the models

The time discretization of the generalized subjective surface model (1) is semi-implicit since in this way we can guarantee unconditional stability of the curvature term. Let τ_S be the time discretization step, $\tau_S = T_S/N_S$. Then for any $n = 1 \ldots N_S$ we get

$$\frac{u^n - u^{n-1}}{\tau_S} - w_a \nabla g \cdot \nabla u^{n-1} - w_c\, g |\nabla u^{n-1}| \nabla \cdot \frac{\nabla u^n}{|\nabla u^{n-1}|} = 0\,. \tag{4}$$

where u^n represents the numerical solution on the nth time level.

The space discretization is realized by applying the finite volume strategy where one doxel of the 4D image corresponds to one volume of the discretization. Let us suppose that the volumes are 4D cubes of side length h and let $V_{\mathbf{i}}$ denote the volume with index vector $\mathbf{i} = (i, j, k, l)$ and $u^n_{\mathbf{i}}$ the value of the numerical solution u^n in the center $c_{\mathbf{i}}$ of this volume. Further, let $\mathbf{e}_p$, $p = 1 \ldots 4$ represent the standard basis vectors in R^4, $F_{\mathbf{i}}^{+p}$ and $F_{\mathbf{i}}^{-p}$ the two faces of $V_{\mathbf{i}}$ orthogonal to $\mathbf{e}_p$, $\nu_{\mathbf{i}}^{\pm p}$ the normal of the face $F_{\mathbf{i}}^{\pm p}$ and $m(F_{\mathbf{i}}^{\pm p})$ its measure.

Now let us integrate (4) over $V_{\mathbf{i}}$. We get

$$\int\limits_{V_{\mathbf{i}}} \frac{u^n - u^{n-1}}{\tau_S} dx - \int\limits_{V_{\mathbf{i}}} w_a \nabla g \cdot \nabla u^{n-1} dx - \int\limits_{V_{\mathbf{i}}} w_c\, g |\nabla u^{n-1}| \nabla \cdot \frac{\nabla u^n}{|\nabla u^{n-1}|} dx = 0. \quad (5)$$

The time derivative term is approximated by

$$\int\limits_{V_{\mathbf{i}}} \frac{u^n - u^{n-1}}{\tau_S}\, dx \approx m(V_{\mathbf{i}}) \frac{u^n_{\mathbf{i}} - u^{n-1}_{\mathbf{i}}}{\tau_S} \cdot \quad (6)$$

The advection term is approximated by the upwind approach, i.e.

$$\int\limits_{V_{\mathbf{i}}} (-w_a \nabla g \cdot \nabla u)\, dx \approx \quad (7)$$

$$w_a m(V_{\mathbf{i}}) \sum_{p=1}^{4} \left(\max\left(-D^p_{\mathbf{i}} g, 0\right) \frac{u^{n-1}_{\mathbf{i}} - u^{n-1}_{\mathbf{i}-\mathbf{e}_p}}{h} + \min\left(-D^p_{\mathbf{i}} g, 0\right) \frac{u^{n-1}_{\mathbf{i}+\mathbf{e}_p} - u^{n-1}_{\mathbf{i}}}{h} \right)$$

where $D^p_{\mathbf{i}} g = (g_{\mathbf{i}+\mathbf{e}_p} - g_{\mathbf{i}-\mathbf{e}_p})/(2h)$ and $g_{\mathbf{i}}$ is the average value of g in $V_{\mathbf{i}}$. For the curvature term we get the approximation

$$\int\limits_{V_{\mathbf{i}}} w_c g |\nabla u^{n-1}| \nabla . \frac{\nabla u^n}{|\nabla u^{n-1}|} dx = w_c g_{\mathbf{i}} \bar{Q}^{n-1}_{\mathbf{i}} \sum_{p=1}^{4} \sum_{q=-p,+p} \int\limits_{F^q_{\mathbf{i}}} \frac{\nabla u^n}{|\nabla u^{n-1}|} . \nu^q_{\mathbf{i}}\, d\gamma, \quad (8)$$

where $\bar{Q}^{n-1}_{\mathbf{i}}$ is the average value of $|\nabla u^{n-1}|$ in $V_{\mathbf{i}}$. Further

$$\int\limits_{F^{\pm p}_{\mathbf{i}}} \frac{\nabla u^n}{|\nabla u^{n-1}|} . \nu^{\pm p}_{\mathbf{i}}\, d\gamma \approx \frac{m(F^{\pm p}_{\mathbf{i}})}{Q^{\pm p;n-1}_{\mathbf{i}}} \frac{u^n_{\mathbf{i}\pm\mathbf{e}_p} - u^n_{\mathbf{i}}}{h} \quad (9)$$

where $Q^{\pm p;n-1}_{\mathbf{i}}$ is the average value of $|\nabla u^{n-1}|$ on the face $F^{\pm p}_{\mathbf{i}}$.

As we can see, in order to properly perform the approximations indicated in (7) and (8), we need to find an appropriate approximation of the average value of $|\nabla u^{n-1}|$ in both $V_{\mathbf{i}}$ and on the faces $F^{\pm p}$ and the average modulus of $g(|\nabla I_\sigma|)$, $I_\sigma = G_\sigma * I_0$, in $V_{\mathbf{i}}$. There are various possibilities how to do that [4].

Let us first consider the approximation of ∇u^{n-1} in the barycenter $c_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p}$ of the doxel face $F_{\mathbf{i}}^{\pm p}$. The component corresponding to the direction of $\mathbf{e}_p$ is simply approximated by

$$D^{\pm p} u_{\mathbf{i}}^{n-1} = \pm \frac{u_{\mathbf{i}\pm\mathbf{e}_p}^{n-1} - u_{\mathbf{i}}^{n-1}}{h}\,. \tag{10}$$

The other components corresponding to the directions of $\mathbf{e}_q$, $q = 1 \ldots 4$, $q \neq p$, can be approximated as follows. The doxel face $F_{\mathbf{i}}^{\pm p}$ is a 3D cube with faces denoted by $F_{\mathbf{i}}^{\pm p,\pm q}$. The barycenter of $F_{\mathbf{i}}^{\pm p,\pm q}$ can be expressed as $c_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p\pm\frac{1}{2}\mathbf{e}_q}$. Thus, the value of u^{n-1} at this point can be approximated as

$$u_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p\pm\frac{1}{2}\mathbf{e}_q}^{n-1} = \frac{1}{4}\left(u_{\mathbf{i}}^{n-1} + u_{\mathbf{i}\pm\mathbf{e}_p}^{n-1} + u_{\mathbf{i}\pm\mathbf{e}_q}^{n-1} + u_{\mathbf{i}\pm\mathbf{e}_p\pm\mathbf{e}_q}^{n-1}\right). \tag{11}$$

The partial derivatives of u^{n-1} at $c_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p}$ are then approximated as

$$D^{\pm p,q} u_{\mathbf{i}}^{n-1} = \frac{u_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p+\frac{1}{2}\mathbf{e}_q}^{n-1} - u_{\mathbf{i}\pm\frac{1}{2}\mathbf{e}_p-\frac{1}{2}\mathbf{e}_q}^{n-1}}{h} \tag{12}$$

Finally, we can define the required approximations

$$Q_{\mathbf{i}}^{\pm p;n-1} = \sqrt{(D^{\pm p} u_{\mathbf{i}}^{n-1})^2 + \sum_{q\neq p}\left(D^{\pm p,q} u_{\mathbf{i}}^{n-1}\right)^2},\ \bar{Q}_{\mathbf{i}}^{n-1} = \frac{1}{8}\sum_{p=1}^{4}\sum_{q=-p,+p} Q_i^{q;n-1} \tag{13}$$

$$G_{\mathbf{i}}^{\pm p} = \sqrt{(D^{\pm p} I_{\sigma;\mathbf{i}})^2 + \sum_{q\neq p}\left(D^{\pm p,q} I_{\sigma;\mathbf{i}}\right)^2},\ g_{\mathbf{i}} = \frac{1}{8}\sum_{p=1}^{4}\left(G_{\mathbf{i}}^{-p} + G_{\mathbf{i}}^{+p}\right) \tag{14}$$

Combining (6)–(14), we get the finite volume scheme for solving the problem (1).

The eikonal equation (2) is discretized by the Rouy-Tourin scheme [5]. Let $\tau_D = T_D/N_D$ be the time discretization step and $d_{\mathbf{i}}^n$ the value of the numerical solution in the barycenter of the doxel $V_{\mathbf{i}}$ on the nth time level. Let us define for $p = 1 \ldots 4$

$$D_{\mathbf{i}}^{\pm p} = \left(\min\left(d_{\mathbf{i}\pm\mathbf{e}_p}^{n-1} - d_{\mathbf{i}}^{n-1}\right)\right)^2, \quad M_{\mathbf{i}}^{p} = \max\left(D_{\mathbf{i}}^{-p}, D_{\mathbf{i}}^{+p}\right)$$

Then the numerical scheme is written as follows

$$d_{\mathbf{i}}^{n} = d_{\mathbf{i}}^{n-1} + \tau_D - \frac{\tau_D}{h}\sqrt{\sum_{p=1}^{4} M_{\mathbf{i}}^{p}} \tag{15}$$

This scheme is stable for $\tau_D \leq h/4$ and it produces monotonically increasing updates that gradually approach a steady state. This leads to an efficient implementation of the scheme that uses a fixing strategy [2].

3 Experiments

Before we proceed to the experiments concerning the actual segmentation and tracking, we test the experimental order of convergence of the finite volume scheme presented above on a simple regularized mean curvature flow equation

$$\partial_t u = |\nabla u| \, \nabla \cdot \left(\frac{\nabla u}{|\nabla u|} \right) \tag{16}$$

with the exact solution $u(x_1, x_2, x_3, x_4, t) = \frac{x_1^2+x_2^2+x_3^2+x_4^2-1}{6} + t$. We use the Dirichlet boundary condition and the initial condition given by this analytical solution. The problem was solved in the domain $[-1.25, 1.25]^4 \times [0, 0.08]$. The spatial domain consisted of n^4 doxels with $h = 2.5/n$ and $\tau \sim h^2$. The error of the numerical solution was measured in $L_\infty(I, L_2(\Omega))$ norm. The result of this test is displayed in Table 1.

Table 1 The experimental order of convergence of the finite volume scheme described in Sec. 2

n	τ	error	EOC
10	0.04	5.531426e-3	
20	0.01	7.276024e-4	2.926
40	0.0025	1.407815e-4	2.370
80	0.000625	3.264185e-5	2.109

The second experiment illustrates the segmentation of artificial 4D data. The 4D image was constructed as an analogy of the cell nuclei evolution. The cell nuclei are more or less spherical objects, so we started with two spheres. In each time slice of the 4D image, these two spheres are situated at different positions but not far from their positions in the previous time slice. We construct 25 time slices. At time $x_4 = 9$, one of the spheres divides and from then on, we have 3 spheres in the image. To make the situation more general, the radii of the spheres change in time. The centers of these spheres are used to construct the initial segmentation function. We place a 4D ellipsoid with radii a, b, c, d in each of these centers and we set $u_0(x) = 1$ inside these ellipsoids and $u_0(x) = 0$ outside. The model parameters were set as follows: $g(|\nabla I_0|) = 1/(1 + K|\nabla I_0|^2)$, $K = 1.0$, $h = 1.0$, $\tau_S = 0.1$, $w_a = 5.0$, $w_c = 0.1$, $T_D = 30$. Instead of $|\nabla u|$ we use its regularization $\sqrt{\varepsilon + |\nabla u|^2}$ with $\varepsilon = 10^{-6}$. The procedure is illustrated in Fig. 1. In order to visualize a 4D discrete function $u(x_1, x_2, x_3, x_4)$ with m slices in x_4-coordinate, we construct its 3D representation by setting the value in each 3D voxel (x_1, x_2, x_3) to $\max\limits_{x_4=1\ldots m} u(x_1, x_2, x_3, x_4)$. Then we visualize an isosurface of this representation.

Another experiment shows the segmentation of the zebrafish embryogenesis data. We segmented a sequence of 20 3D cell nuclei images preprocessed (denoised) by the geodesic mean curvature flow filter [3]. The initial segmentation function was constructed in the same way as in the case of the artificial data. Further, we set $g(|\nabla I_\sigma|) = 1/(1 + K|\nabla I_\sigma|^2)$, $I_\sigma = G_\sigma * I_0$, $K = 100.0$, $\sigma = 0.01$, $h = 1.0$,

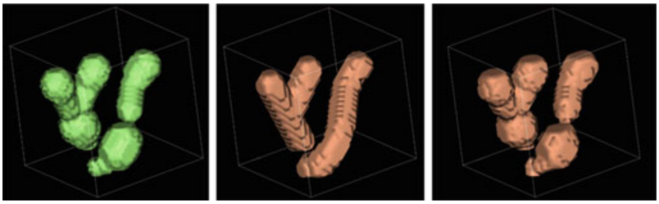


Fig. 1 Segmentation of artificial 4D data. Left, the isosurface $V = 128$ of the 3D representation of the data. Middle, the isosurface $V = 15$ of the 3D representation of the initial segmentation function. Right, the isosurface $V = 15$ of the 3D representation of the segmentation result. This isosurface was chosen as the best representation of the segmented object

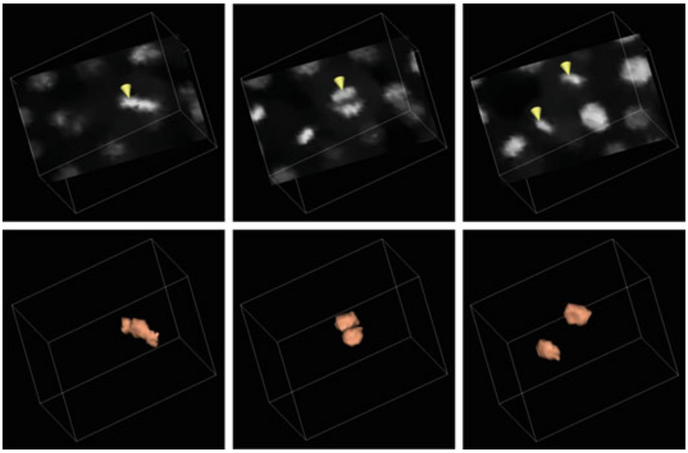


Fig. 2 Segmentation of the zebrafish embryogenesis data. On the top, we display 2D slices of the 4D image corresponding to different x_4 (time) values with indication of the position of the segmented object. On the bottom, we provide the corresponding segmentation result in the form of isosurface $V = 128$ of x_4-slices of the 4D segmentation function

$\tau_S = 0.1$, $w_a = 10.0$, $w_c = 1.0$, $T_D = 50$, $\varepsilon = 10^{-6}$. Fig. 2 displays 2D slices of the 4D data (more precisely, 2D slices of x_4-slices of the 4D data). The object that we tried to segment was a simple cell evolution tree containing one cell division. Together with the image slices, we provide the segmentation result, now displayed as isosurfaces of x_4 (time) slices of the segmentation function.

Fig. 3 shows the result of the cell tracking performed on the artificial data described above. We backtrack the cells (spheres) from the positions of their centers at the end of the time sequence. Both distance functions d_1 and d_2 were computed by setting $h = 1.0$, $\tau_D = 0.25$. The result of the tracking is a set of 4D points characterizing the cell position on the individual time levels. At each time level, we get one point that represents the intersection of the time hyperplane with the ridge of the 4D distance function d_2 (note that these points in general do not correspond

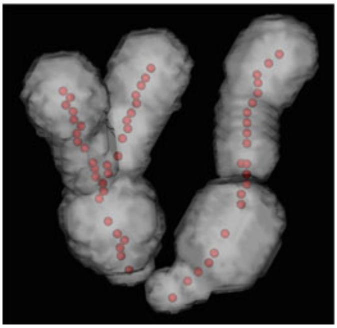


Fig. 3 The result of the cell tracking performed on artificial 4D data. We can see the points characterizing the positions of the cells at each time level visualized by neglecting their x_4 coordinate. The starting points for the tracking are situated on the top of the point sequences

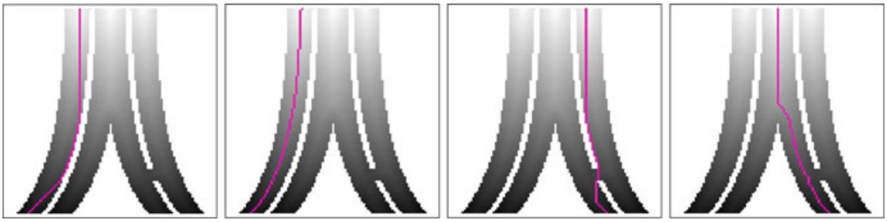


Fig. 4 The effect of using the distance function d_2. From the left: first, the tracking line in an isolated branch obtained by using only d_1, second, the tracking line in the same branch when using d_1 and d_2, third, the tracking line in a branch interconnected with a neighboring branch drops into a wrong branch if only d_1 and not d_2 is applied, fourth, by applying both d_1 and d_2, the line remains in the correct branch. The grey level shading of the branches represents the values of d_1

to the geometrical centers of the individual 3D spheres). The points are visualized by neglecting their x_4 coordinate.

Finally, we present a test illustrating the effect of using the distance function d_2. In Fig. 4, we can see four branches of 2D cell evolution trees. As we can observe, if using only the distance function d_1, the tracking lines tend to go along the borders of their

4 Conclusions

To conclude, we presented the main ideas of a new cell tracking technique and we illustrated the validity of the procedure on several test examples. The method is now prepared to be applied to long time sequences of biological data.

Acknowledgements This work was supported by the grant VEGA 1/0733/10.

References

1. Bourgine, P., Čunderlík, R., Drblíková-Stašová, O., Mikula, O. , Peyriéras, N., Remešíková, M., Rizzi, M., Sarti, A.: 4D embryogenesis image analysis using PDE methods of image processing. Kybernetika **46** **(2)**, 226–259 (2010).
2. Bourgine, P., Frolkovič, P., Mikula, K., Peyriéras, N., Remešíková, M.: Extraction of the intercellular skeleton from 2D microscope images of early embryogenesis. In Lecture Notes in Computer Science **5567** (Proceeding of the 2nd International Conference on Scale Space and Variational Methods in Computer Vision, Voss, Norway, June 2009) (Springer, 2009), p. 38–49.
3. Krivá, Z., Mikula, K., Peyriéras, N., Rizzi, B., Sarti, A., Stašová, O.: Zebrafish early embryogenesis 3D image filtering by nonlinear partial differential equations. Medical Image Analysis, **14** **(4)**, 510–526 (2010).
4. Mikula, K., Remešíková, M.: Finite volume schemes for the generalized subjective surface equation in image segmentation. Kybernetika **45** **(4)**, 646–656 (2009).
5. Rouy, E., Tourin, A.: Viscosity solutions approach to shape-from-shading. SIAM Journal on Numerical Analysis **29** **(3)**, 867–884 (1992).
6. Zanella, C., Campana, M., Rizzi, B., Melani, C., Sanguinetti, G., Bourgine, P., Mikula, K., Peyriéras, N., Sarti, A.: Cells Segmentation from 3-D Confocal Images Of Early Zebrafish Embryogenesis. IEEE Transactions on Image Processing **19** **(2)**, (2010).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Rhie-Chow interpolation for low Mach number flow computation allowing small time steps

Yann Moguen, Tarik Kousksou, Pascal Bruel, Jan Vierendeels, and Erik Dick

Abstract Low Mach number flow computation in co-located grid arrangement requires pressure-velocity coupling in order to prevent the checkerboard phenomenon. A Rhie-Chow interpolation technique can be formulated with such a coupling involving an explicit time step dependence, suitable for unsteady computations. Following this approach, it is observed that unphysical pressure oscillations arise again for sufficiently small time steps. Some remedies have been proposed for incompressible flows. A simple adaptation of these remedies for low Mach number flow computation is numerically investigated. A slight departure from the original approach appears to be suitable.

Keywords Pressure correction, pressure-velocity coupling, co-located grid
MSC2010: 65M08, 65N22, 76G25, 76N15

Yann Moguen, Jan Vierendeels, and Erik Dick
Ghent University - Department of Flow, Heat and Combustion Mechanics,
Sint-Pietersnieuwstraat, 41 - 9 000 Gent, Belgium, e-mail: yann.moguen@free.fr, jan.vierendeels@ugent.be, erik.dick@ugent.be

Tarik Kousksou
Université de Pau et des Pays de l'Adour - Laboratoire des Sciences de l'Ingénieur Appliquées à la Mécanique et au Génie Electrique, ENSGTI, rue Jules Ferry - 64 075 Pau, France,
e-mail: tarik.kousksou@univ-pau.fr

Pascal Bruel
CNRS and Université de Pau et des Pays de l'Adour - Laboratoire de Mathématiques et de leurs Applications, UMR 5142 CNRS-UPPA, avenue de l'Université, BP 1155 - 64 013 Pau, France,
e-mail: pascal.bruel@univ-pau.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_74, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

Coupling between velocity and pressure difference on cell faces is necessary in low Mach number flow computations on grids with co-located arrangement. This allows to avoid the checkerboard phenomenon, which means unphysical pressure oscillations, increasing as a Mach number representative of the flow goes to zero. A pressure-velocity coupling that involves an explicit time step dependence, which appears to be advantageous with unsteady computations, can be obtained by Rhie-Chow interpolation method [5]. Unfortunately, with this choice, some pressure oscillations arise again for 'small' time steps. This may lead to useless computations. Some remedies have been proposed, but they concern incompressible flows [6]. In the present contribution, we investigate numerically this issue in the case of low Mach number flow computations.

For simplicity, a one-dimensional flow of a perfect and ideal gas in a nozzle with a variable section is considered. In the following, x denotes the coordinate in the flow direction. The flow model is given by the Euler equations:

$$\partial_t(\rho S) + \partial_x(\rho v S) = 0 \tag{1a}$$

$$\partial_t(\rho v S) + \partial_x((\rho v^2 + p)S) = p\mathrm{d}_x S \tag{1b}$$

$$\partial_t(\rho E S) + \partial_x(\rho v H S) = 0 \tag{1c}$$

$$E = e + \frac{1}{2}v^2 \tag{1d}$$

$$\rho H = \rho E + p \tag{1e}$$

$$\rho e = \frac{p}{\gamma - 1} \tag{1f}$$

where t, ρ, p, v, e, E and H represent time, density, pressure, velocity, internal energy, total energy and total enthalpy per unit mass, respectively. Furthermore, γ denotes the specific heats ratio and S the cross-section area of the nozzle.

The x axis along the nozzle is divided into a number N of cells of length Δx. A finite volume formulation in co-located arrangement is applied.

2 Pressure correction algorithm

To solve the set (1) of equations, the energy-based pressure correction algorithm that we consider takes the following form, where the superscripts $\star$ and $'$ denote estimated and correction quantities of each iteration (first iteration: $k = n$), respectively:

1. Pre-estimation step: Generate a transporting velocity v^T at the cell-faces, that will be used in the following two steps.
2. Estimation step: With $p_i^\star = p_i^k$, calculate $\rho_i^\star$ and $(\rho v)_i^\star$ using

$$\frac{1}{2}(3\rho_i^\star - 4\rho_i^n + \rho_i^{n-1}) + \frac{\tau}{S_i}\{[\rho_i^\star + \frac{1}{2}\psi_i^k(\rho)(\rho_i^k - \rho_{i-1}^k)]v_{i+1/2}^T S_{i+1/2} - [\rho_{i-1}^\star + \frac{1}{2}\psi_{i-1}^k(\rho)(\rho_{i-1}^k - \rho_{i-2}^k)]v_{i-1/2}^T S_{i-1/2}\} = 0 \quad (2)$$

   where τ is formally defined as $\Delta t/\Delta x$ and practically calculated as $\text{CFL}_v/v_{\max}$, and v^T is positive. A similar equation holds for the momentum. Here ψ denotes a slope limiter, for instance MinMod, allowing to reach second-order accuracy in space, while the same order of accuracy in time is obtained by using the second-order backward discretization. From the estimated density, momentum and pressure, calculate the estimated total energy and total enthalpy.
3. Correction step: Calculate the pressure correction p' by solving the energy equation in second-order accurate backward discretization form in time. Flux terms are expanded as

$$(\rho v H)_{i+1/2}^{k+1} = (\rho H)_{i+1/2}^\star v_{i+1/2}^\star + H_{i+1/2}^\star(\rho v)_{i+1/2}' + (\rho H)_{i+1/2}' v_{i+1/2}^\star \quad (3)$$

   where $H_{i+1/2}^\star$ and $(\rho H)_{i+1/2}^\star$ are upwinded in second-order accurate form, as convected quantities. Furthermore, neglecting the kinetic energy contribution, $(\rho H)_{i+1/2}' = \frac{\gamma}{\gamma-1}p_{i+1/2}'$ and the calculation of $(\rho v)_{i+1/2}'$ is derived from the momentum equation.
4. Updates: $p_i^{k+1} = p_i^\star + p_i'$, $\rho_i^{k+1} = \rho_i^\star + (\partial_p \rho)_i^\star p_i'$, $(\rho v)_i^{k+1} = (\rho v)_i^\star + (\rho v)_i'$, where $(\rho v)_i'$ is derived from the momentum equation in accordance with the derivation of $(\rho v)_{i+1/2}'$. The total energy and the cell-face pressure and velocity are finally updated.

3 Calculation of cell-face quantities

Let us provide some details on the AUSM/Rhie-Chow combination, which we consider as suitable for future unsteady computations.

3.1 AUSM interpolation

For explanations on the AUSM$^+$ and AUSM$^+$-up schemes, we refer to [2] and focus only on a low Mach number adaptation of AUSM$^+$, by using a simple scaling function suggested in this reference for the construction of the AUSM$^+$-up scheme.

The notation L or R, which refers to the left or right side of the face $i + 1/2$, is adopted since an extrapolation technique will be used in the following. Thus, a Mach number on the side S is defined as

$$M_S = \frac{v_S}{c_{i+1/2}} \quad , \quad S = L, R \tag{4}$$

where $c_{i+1/2}$ is the speed of sound at the face $i + 1/2$ (*cf.* Ref. [2]). A mean Mach number at the face $i + 1/2$ is also defined,

$$\overline{M}_{i+1/2} = \sqrt{\frac{(v_L)^2 + (v_R)^2}{2c^2_{i+1/2}}} \tag{5}$$

and a reference Mach number $M_{0,i+1/2}$ by

$$M^2_{0,i+1/2} = \min\{1, \max\{\overline{M}^2_{i+1/2}, \mathrm{Ma}^2_{\mathrm{co}}\}\} \tag{6}$$

where $\mathrm{Ma}_{\mathrm{co}}$ is a cut-off Mach number value such that: $\mathrm{Ma}_{\mathrm{co}} = \mathscr{O}(\mathrm{Ma}_\infty)$. A scaling function suggested in Ref. [2] is

$$f(M) = M(2 - M) \tag{7}$$

The use of this function permits the proper asymptotic behaviour of the pressure dissipation term for M ↘ 0 in the face velocity (see Ref. [2]), defined by the following construction:

$$M^{\pm}_{(1)}(M) = \frac{1}{2}(M \pm |M|) \tag{8}$$

$$M^{\pm}_{(4)}(M) = \pm\frac{1}{4}(M \pm 1)^2 \pm \frac{1}{8}(M^2 - 1)^2 \tag{9}$$

$$P^{\pm}_{(0)}(M) = M^{\pm}_{(1)}(M)/M \tag{10}$$

$$P^{\pm}_{(5)}(M) = \frac{1}{4}(M \pm 1)^2(2 \mp M) \pm \frac{3}{16}(5(f(M_0))^2 - 4)M(M^2 - 1)^2 \tag{11}$$

$$\mathscr{M}^{\pm}(M) = \begin{cases} M^{\pm}_{(1)}(M) \ , & |M| \geq 1 \\ M^{\pm}_{(4)}(M) \ , & |M| < 1 \end{cases} \tag{12}$$

$$\mathscr{P}^{\pm}(M) = \begin{cases} P^{\pm}_{(0)}(M) \ , & |M| \geq 1 \\ P^{\pm}_{(5)}(M) \ , & |M| < 1 \end{cases} \tag{13}$$

$$p_{i+1/2} = \mathscr{P}^{+}(M_L)p_L + \mathscr{P}^{-}(M_R)p_R \tag{14}$$

$$M_{i+1/2} = \mathscr{M}^{+}(M_L) + \mathscr{M}^{-}(M_R) \tag{15}$$

$$v_{i+1/2} = c_{i+1/2}\, M_{i+1/2} \tag{16}$$

To reach second-order accuracy in space, the primitive variables p, ρ and v, which are used in the AUSM$^+$ scheme, are extrapolated at the face $i + 1/2$ according to

$$\phi_L = \phi_i + \frac{1}{2}\psi_i(\phi)(\phi_i - \phi_{i-1}) \quad , \quad \phi_R = \phi_{i+1} - \frac{1}{2}\psi_{i+1}(\phi)(\phi_{i+1} - \phi_i)$$

where ψ denotes a slope limiter. Practically, we choose

$$\psi_i(\theta) = \text{MinMod}(\theta_i - \theta_{i-1}, \theta_{i+1} - \theta_i)\;/\;(\theta_i - \theta_{i-1})$$

where

$$\text{MinMod}(a, b) = \frac{\text{sign}(a) + \text{sign}(b)}{2}\min\{|a|, |b|\}$$

3.2 Rhie-Chow interpolation

The pressure-velocity coupling, especially needed at low Mach number, is achieved through the construction of a transporting velocity with a Rhie-Chow interpolation technique. According to the 'classical' Rhie-Chow approach (see *e.g.* Ref. [1]), the preceding face velocities are interpolated when assembling the current transporting velocity. In this case, steady results which are not time step dependent can be ascertained when a certain interpolation practice is satisfied (see Ref. [4]). However, the numerical dissipation associated with the pressure-velocity coupling arising with this choice can lead to unphysical unsteady computations [3]. An alternative way to allow 'small' time step computations consists to avoid interpolation by directly using the preceding transporting velocities. This approach has been proposed in Ref. [6] for incompressible flows. Its application for low Mach number flow will be addressed in the rest of this paper.

First, an auxiliary density, that can be thought as a 'pre-predicted' one, is defined by solving the continuity equation, as

$$\begin{aligned}\frac{1}{2}(3\rho_i^{\star\star} - 4\rho_i^n + \rho_i^{n-1}) + \frac{\tau}{S_i}\{[\rho_i^{\star\star} + \frac{1}{2}\psi_i^k(\rho)(\rho_i^k - \rho_{i-1}^k)]v_{i+1/2}^k S_{i+1/2}\\ - [\rho_{i-1}^k + \frac{1}{2}\psi_{i-1}^k(\rho)(\rho_{i-1}^k - \rho_{i-2}^k)]v_{i-1/2}^k S_{i-1/2}\} = 0\end{aligned} \tag{17}$$

where the cell-face velocity, which is positive, is given at the last known iteration, and calculated by the scheme described in subsection 3.1. Then, similarly, an auxiliary pre-predicted velocity $v^{\star\star}$ is defined from the momentum equation in which the pressure gradient influence has been removed, by

$$
\begin{aligned}
a_i\rho_i^{\star\star}v_i^{\star\star} = -\frac{\tau}{S_i}\Big[\frac{1}{2}\psi_i^k(\rho v)[(\rho v)_i^k - (\rho v)_{i-1}^k]v_{i+1/2}^k S_{i+1/2} \\
- \{(\rho v)_{i-1}^k + \frac{1}{2}\psi_{i-1}^k(\rho v)[(\rho v)_{i-1}^k - (\rho v)_{i-2}^k]\}v_{i-1/2}^k S_{i-1/2}\Big] \\
+ (1-\varepsilon)[2(\rho v)_i^n - \frac{1}{2}(\rho v)_i^{n-1}]
\end{aligned}
\tag{18}
$$

where $0 \leq \varepsilon \leq 1$ and $a_i = \frac{3}{2} + \frac{\tau}{S_i}v_{i+1/2}^k S_{i+1/2}$. With a similar equation for $v_{i+1}^{\star\star}$, a transporting velocity is defined according to a Rhie-Chow interpolation, as

$$
\begin{aligned}
v_{i+1/2}^T = \frac{1}{2}(v_i^{\star\star} + v_{i+1}^{\star\star}) - \frac{\tau}{2}(\frac{1}{a_i\rho_i^{\star\star}} + \frac{1}{a_{i+1}\rho_{i+1}^{\star\star}})(p_{i+1}^k - p_i^k) \\
+ \frac{\varepsilon}{2}(\frac{1}{a_i} + \frac{1}{a_{i+1}})[2(v_{i+1/2}^T)^n - \frac{1}{2}(v_{i+1/2}^T)^{n-1}]
\end{aligned}
\tag{19}
$$

Now, the pressure correction-mass flux correction coupling is defined accordingly, as

$$
(\rho v)'_{i+1/2} = -\frac{\tau}{2}(\frac{1}{a_i} + \frac{1}{a_{i+1}})(p'_{i+1} - p'_i) \tag{20}
$$

and then, the momentum correction is written as

$$
(\rho v)'_i = -\frac{\tau}{a_i}(p'_{i+1/2} - p'_{i-1/2}) \tag{21}
$$

Taking $\varepsilon = 1$ in expressions (18) and (19) corresponds to the remedy suggested in Ref. [6] to the non-physical pressure oscillations that occur when Rhie-Chow interpolation is used with small time steps for incompressible flows computation. As a preliminary discussion on the suitable choice of ε, the rest of this paper is devoted to numerical experiments illustrating some numerical difficulties encountered and remedies.

4 Numerical experiments

A low Mach number flow in a converging-diverging nozzle is considered. The prescribed inlet velocity is such that the throat Mach number is 10^{-3} at convergence. The cut-off Mach number $\mathrm{Ma}_{\mathrm{co}}$ in expression (6) is set as 10^{-3}. At the inlet, a constant target value of the density is fixed as $\rho_{\mathrm{in}} = 1.2086\ \mathrm{kg/m}^3$. At the outlet, a constant target value of the pressure is fixed as $p_{\mathrm{out}} = 101\ 300$ Pa. Target values are mentioned since a non-reflecting treatment of the boundaries is applied, but this does not relate to the discussion in the present paper. Let us point out that a well-known complete analytical solution is available for the steady flow under consideration.

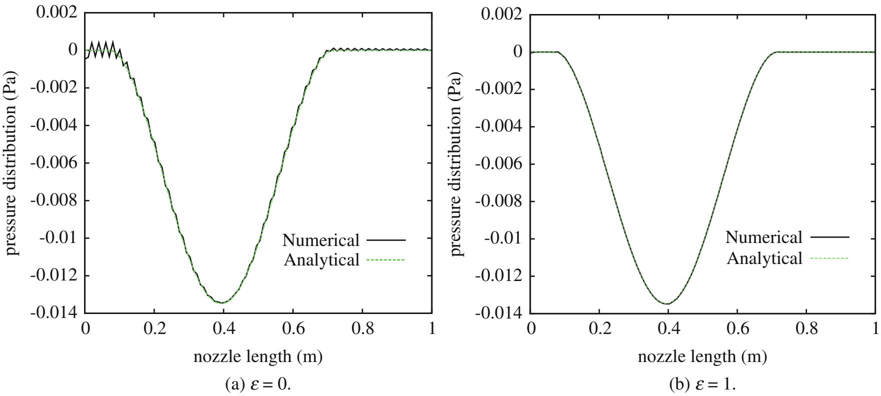


Fig. 1 Pressure distribution (Pa) along the nozzle. $\Delta t = 1.75\ 10^{-5}$ s ; $\varepsilon = 1$: Ref. [6]. Number of cells: 100

First, unphysical pressure oscillations arising with $\varepsilon = 0$ are shown in Fig. 1, left. The oscillations are more pronounced on the left side of the nozzle, which is the less constraining for the pressure variable – in the sense that no target value is imposed at the inlet –, but they are present more or less on the totality of the nozzle. As shown in Fig. 1, right, a simple remedy consists in the choice of $\varepsilon = 1$ in Eqs. (18) and (19), as suggested in Ref. [6].

Between 0 and 1, an intermediate value for ε is also possible, that can be thought as a parameter that manages the deferred treatment of the temporal terms in the momentum equation at the 'pre-prediction' stage of the algorithm (see section 3.2). In Fig. 2, the plot of the error in the pressure versus the parameter ε reveals that the optimal value of ε is not 1, as far as the accuracy is the criterion. The optimal value ε_{opt} that minimizes the pressure error is slightly less than 1. Let us notice that, as $\Delta t \searrow 0$, the transporting velocity is such that

$$v^T_{i+1/2} \to \frac{1}{3}(1-\varepsilon)\big[2\left(\frac{(\rho v)^n_i}{\rho^{\star\star}_i} + \frac{(\rho v)^n_{i+1}}{\rho^{\star\star}_{i+1}}\right) - \frac{1}{2}\left(\frac{(\rho v)^{n-1}_i}{\rho^{\star\star}_i} + \frac{(\rho v)^{n-1}_{i+1}}{\rho^{\star\star}_{i+1}}\right)\big] + \frac{2}{3}\varepsilon\big[2(v^T_{i+1/2})^n - \frac{1}{2}(v^T_{i+1/2})^{n-1}\big] \quad (22)$$

With $\varepsilon \neq 1$ but close to 1, Eq. (22) corresponds to a slight departure from the original approach of Ref. [6]. The observed sensitivity of ε_{opt} to Δt suggests that a comprehensive study of the ε_{opt} dependency on CFL$_v$ should be carried out.

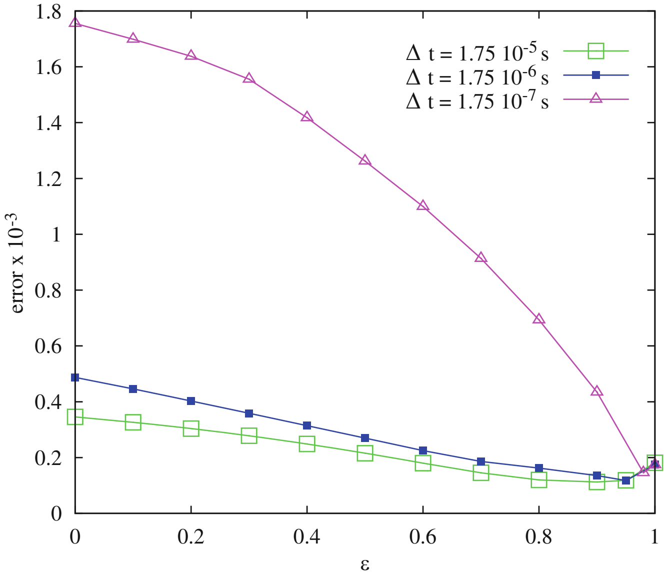


Fig. 2 Error in pressure (L2 norm) *vs.* ε of Eqs. (18) and (19). $\varepsilon = 1$: Ref. [6]. Number of cells: 100

5 Conclusion

The Rhie-Chow stabilisation method with a pressure-velocity coupling that involves an explicit time step dependence does not avoid unphysical oscillations when the time step is sufficiently small. According to Ref. [6], these oscillations originate from the interpolation of the previous velocities in the transporting velocity construction. However, in the steady case considered here, a slight departure from the approach of Ref. [6] is suitable concerning the deferred treatment of the temporal terms in the momentum equation.

Since the problem of oscillations also occurs for unsteady computations, the next step of this study will be to examine how the considered correction of the Rhie-Chow interpolation works in the case of unsteady low Mach number flow computations. Last but not least, the appropriate choice of the parameter ε if the exact solution is unknown is also an issue to be investigated.

References

1. Lien F.S. and Leschziner M.A.: A general non-orthogonal collocated finite volume algorithm for turbulent flow at all speeds incorporating second-moment turbulence-transport closure, Part 1: Computational implementation. Comput. Methods Appl. Mech. Engrg. **114**, 123–148 (1994)

2. Liou M.-S.: A Sequel to AUSM, part II: AUSM$^+$-up for all speeds. J. Comp. Phys. **214**, 137–170 (2006)
3. Moguen Y., Kousksou T., Dick E. and Bruel P.: On the role of numerical dissipation in unsteady low Mach number flow computations. Proceedings of the Sixth International Conference on Computational Fluid Dynamics. Springer (2011). To appear.
4. Pascau A.: Cell face velocity alternatives in a structured colocated grid for the unsteady Navier-Stokes equations. Int. J. Numer. Meth. Fluids **65**, 812-833 (2011)
5. Rhie C.M. and Chow W.L.: Numerical Study of the Turbulent Flow Past an Airfoil with Trailing Edge Separation. AIAA J. **21**(11), 1525–1532 (1983)
6. Shen W.Z., Michelsen J.A. and Sørensen J.N.: Improved Rhie-Chow Interpolation for Unsteady Flow Computations. AIAA J. **39**(12), 2406–2409 (2001)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Study and Approximation of IMPES Stability: the CFL Criteria

C. Preux and F. McKee

Abstract Whether it is for the recovery of hydrocarbons or the injection and storage of CO2, the industry uses numerical simulation. The first stage of study consists in the construction of a geologic model. In view of the field size, the fine model thus built contains several tens of million cells. Numerical fluid flow simulations may require a lot of CPU time and are generally impossible to achieve. To reduce the simulation cost, reservoir engineers use an upscaling of the fine mesh in a coarse one. If the upscaling of absolute permeabilities was already the object of detailed research, it is not the case for polyphasics flows. These flows are described by relative permeabilities and capillary pressure curves defined by limit points and normalized forms. The curves are different according to the facies of the model and in this way, according to the cells. The subject of this paper is to study the IMPES scheme (Implicit Pressure, Explicit Saturation [1], [2], [3]) and to simulate a diphasic flow in order to prepare the upscaling step. A stability analysis of this scheme can highlight a CFL condition [4]. This paper proposes a study of this CFL number for the case of reservoir simulation.

Keywords Diphasics, reservoir simulation, IMPES scheme, stability, CFL
MSC2010: 76T99, 76S05, 65N08, 65N12, 65N12

C. Preux and F. McKee
IFP Energies Nouvelles, 1 et 4 avenue de Bois-Préau - 92852 Rueil-Malmaison Cedex - France,
e-mail: christophe.preux@ifpenergiesnouvelles.fr, francois.mckee@ifpenergiesnouvelles.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_75, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

The acronym IMPES was used in 1968 in a description of a numerical model for simulating black oil reservoir behavior. The IMPES method was generalised in 1980 to apply to simulation models involving any number n of conservation equations. The basic principle is the elimination of differences in non-pressure variables from the model's set of conservation equations to obtain a single pressure equation. Stone [1], Sheldon, Harris, Bavly [5], and Martin [6] used the same principle in deriving the total compressibility of multiphase black oil systems. For Coats [7], the first black oil IMPES model was published by Fagin and Stewert in 1966 [8]. In this work, we deal with the flow of two immiscible and incompressible fluids. We present the system gouverning a two phase flow and proceed to simplifications. We consider the case where there is no gravity and no capillary pressure. This system is based on Darcy's Law generalized for two-phase flow.

1.1 Fully coupled formulation

The most commonly used description for macro-scale two-phase flow in porous media uses a phenomenological extension of Darcy's law introducing the saturation-dependent parameter: relative permeability. This extension was proposed, in the thirties, by Leverett [9] and Wyckoff and Botset [10] and the authors have validated this generalization experimentally.

$$v_i = -\frac{k_{ri}}{\mu_i} K(\nabla p_i - \rho_i g) \tag{1}$$

i represents the considered phase. This equation must be formulated for each phase (for details, see [11–14]). We note k_{ri} the relative permeability, μ_i the dynamic fluid viscosity, p_i the pressure, ρ_i the density of the phase i, and g is the gravity vector. We can define $\lambda_i = \frac{k_{ri}}{\mu_i}$ the mobility of the phase i. We can now write the mass balance equation for the phase i of a multiphase system:

$$\partial_t(\Phi \rho_i S_i) + \nabla.(\rho_i v_i) - \rho_i q_i = 0 \tag{2}$$

where Φ is the porosity, S_i the saturation of the phase i, q_i the source/sink term. Inserting the generalised Darcy law (1) into the mass balance equation (2) and considering the two phases flow system leads to this system of equations:

$$\partial_t(\Phi \rho_w S_w) - \nabla.(\rho_w \lambda_w K(\nabla p_w - \rho_w g)) - \rho_w q_w = 0 \tag{3}$$

$$\partial_t(\Phi \rho_n S_n) - \nabla.(\rho_n \lambda_n K(\nabla p_n - \rho_n g)) - \rho_n q_n = 0 \tag{4}$$

where w represents the wetting phase and n the not-wetting phase. In our case, we consider that Φ is constant in time. Moreover, we suppose that we don't have mass transfer between the two phases. The supplementary constraints to close the system of equations are: the sum of saturations is equal to one ($S_w + S_n = 1$) and the capillary pressure between the two phases (dependent on the saturations) is defined as $p_c(S) = p_n - p_w$. So, if we add the two equations and neglect the source/sink terms, we can present this system with the primary variables p_n et S_w:

$$\phi \partial_t(S_w) - \nabla.(\lambda_w K(\nabla p_n - \nabla p_c - \rho_w g)) = 0 \tag{5}$$

$$-\nabla.(\lambda_w K(\nabla p_n - \nabla p_c - \rho_w g)) - \nabla(\lambda_n K(\nabla p_n - \rho_n g)) = 0 \tag{6}$$

1.2 Simplifications: no capillary pressure, no gravity

In this paper, the problem is simplified and the effects of gravity and capillary pressure are not considered. This may be the case in large homogeneous porous media at high capillary number: viscous forces dominate capillary forces. So, $p_n = p_w$ and the indice n for pressure is neglected. Furthermore, if the total mobility $\lambda_T = \lambda_w + \lambda_n$ is introduced, we can define the total velocity v_T with a form similar to the one of the Darcy law:

$$v_T = -\lambda_T K(\nabla p) \tag{7}$$

In this step the fractional flow function $f_i = \frac{\lambda_i}{\lambda_T}$ is introduced (notice that the fractional flow function depends on the saturation: $f_i = f_i(S_w)$) and if we write the system in term of v_T, the equations (5) and (6) become:

$$\phi \partial_t(S_w) + \nabla.(f_w v_T)) = 0 \tag{8}$$

$$\nabla v_T = 0 \tag{9}$$

2 Discretisation of the fully coupled formulation

In this section, the system is discretized in time on a mesh. We introduce the flux $\mathbf{F}(x, t, S_w) = f_w(S_w)v_T$ and we note:

$$F_{w_{I/J}}(S^*_{w_I}, S^*_{w_J}) \approx \frac{1}{\Delta t} \int_{t^m}^{t^{m+1}} \int_{I/J} f_{w_{I/J}}(S_w, x, t) v_T(x,t) n_{I/J}(s) ds dt \tag{10}$$

where $n_{I/J}(s)$ is the normal vector to the I/J cell interface and where $*$ is a indecision: if $* = m$ then an explicit scheme is obtained and if $* = m + 1$ then the scheme is implicit. We get:

$$\phi \frac{S_w^{m+1} - S_w^m}{\Delta t} + \sum_{J \in Neighbor(I)} F_{w_{I/J}}(S^*_{w_I}, S^*_{w_J}) = 0$$

Moreover, a monotonic flux is needed, i.e: $F_{w_{I/J}}(X,Y)$ increasing with X et decreasing with Y and $F_{w_{I/J}}(X,Y) = -F_{w_{I/J}}(Y,X)$. Also, since a fully upwind scheme is used, the mobility at the cell interface is determined in the following way (we note $\mathbf{V_T} = \int_{t^m}^{t^{m+1}} \int_{I/J} \lambda_T(S_w) K(\nabla p) \mathbf{n}(s)_{I/J} ds dt$):

$$f_{w_{I/J}}(S_w) = f_w(S_{w_I}) \quad if \quad \mathbf{V_T} > 0 \tag{11}$$

$$f_{w_{I/J}}(S_w) = f_w(S_{w_J}) \quad if \quad \mathbf{V_T} < 0 \tag{12}$$

The IMPES scheme is obtained for a two phase flow: an implicit treatment for the pressure terms and an explicit model for the saturation:

$$V\phi \frac{S_{w_I}^{m+1} - S_{w_I}^m}{t^{m+1} - t^m} + \sum_{J \in Neighbor(I)} f_w(S_{w_I}) \left(\mathbf{V_T^{m+1}}\right)^+ + f_w(S_{w_J}) \left(\mathbf{V_T^{m+1}}\right)^- = 0 \tag{13}$$

$$\sum_{J \in Neighbor(I)} \left(\mathbf{V_T^{m+1}}\right)^+ + \left(\mathbf{V_T^{m+1}}\right)^- = 0 \tag{14}$$

3 The CFL criteria

We present here how the CFL criteria of this scheme can be computed using the boundary conditions.

3.1 Consideration of the boundary conditions

In this paper, a waterflooding is considered: the porous media is subjected to a difference of pressure, as shown on Fig. 1.

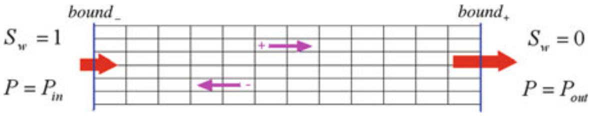


Fig. 1 Pressure imposed

On the left side $P = P_{in}$ and $S_w = 1$. On the right side, we have $P = P_{out}$ and $S_w = 0$. We must now differentiate the surfaces σ between the cells (interior of the media) or between a cell and the edge. Taking into account these conditions and multiplying by $f_w(S_{w_I})$, the equation 14 becomes:

$$\sum_{\sigma=IJ\in\partial I\cap\Sigma_{int}} f_w(S_{w_I})\left(\mathbf{V_{T,I|J}}^{int,m+1}\right)^+ + f_w(S_{w_I})\left(\mathbf{V_{T,I|J}}^{int,m+1}\right)^- + \sum_{\sigma=I\in\partial I\cap\Sigma_{bound}} f_w(S_{w_I})\left(\mathbf{V}^{bound,m+1}_{\mathbf{T,I}|\sigma}\right) = 0 \tag{15}$$

where:

$$\mathbf{V_{T,I|J}}^{int,m+1} = -\mathbf{V_{T,J|I}}^{int,m+1} = \lambda^{int,m}_{T,\sigma} T^{int}_{\sigma}(p_I^{m+1} - p_J^{m+1}), \sigma = I|J \in \Sigma_{int}, \tag{16}$$

$$\mathbf{V}^{bound,m+1}_{\mathbf{T,I}|\sigma} = \lambda^{bound,m}_{T,\sigma} T^{bound}_{\sigma}(p_I^{m+1} - p_\sigma^{m+1}), \sigma = \Sigma_{bound}. \tag{17}$$

with $T^{int}_{\sigma} = \frac{S_\sigma K_{I|J}}{D_{I|J}}$ and $T^{bound}_{\sigma} = \frac{S_\sigma K_I}{D_{I(\sigma)}}$. Taking into account the boundary conditions in the right and left side, we obtain for the equation 13:

$$V_I\phi\frac{S^{m+1}_{w_I} - S^m_{w_I}}{\Delta t} + \sum_{\sigma=IJ\in\partial I\cap\Sigma_{int}} f_w(S^m_{w_I})\left(\mathbf{V_{T,I|J}}^{int,m+1}\right)^+ + f_w(S^m_{w_J})\left(\mathbf{V_{T,I|J}}^{int,m+1}\right)^-$$
$$+ \sum_{\sigma=IJ\in\partial I\cap\Sigma_{bound+}} f_w(S^m_{w_I})\left(\mathbf{V_{T,I|œ}}^{bound,m+1}\right)^+ + f_w(0)\left(\mathbf{V}^{bound,m+1}_{\mathbf{T,I}|\sigma}\right)^- \tag{18}$$
$$+ \sum_{\sigma=IJ\in\partial I\cap\Sigma_{bound-}} f_w(1)\left(\mathbf{V}^{bound,m+1}_{\mathbf{T,I}|\sigma}\right)^+ + f_w(S^m_{w_I})\left(\mathbf{V}^{bound,m+1}_{\mathbf{T,I}|\sigma}\right)^- = 0$$

The last term corresponds to the velocity at the left boundary, where $S_w = 1$ is imposed. Thus we introduce the rates of change and we note:

$$d_{I|J} = \frac{f_w(S_{w_I}) - f_w(S_{w_J})}{S_{w_I} - S_{w_J}}; \quad d_{I|1} = \frac{f_w(S_{w_I}) - f_w(1)}{S_{w_I} - 1}; \quad d_{I|0} = \frac{f_w(S_{w_I}) - f_w(0)}{S_{w_I}} \tag{19}$$

The subtraction of these two equations (15 , 18) gives:

$$S^{m+1}_{w_I} = S^m_{w_I}\left(1 + \frac{\Delta t}{V_I\phi}\left[\sum_{\sigma=IJ\in\partial I\cap\Sigma_{int}} \left(\mathbf{V_{T,I|J}}^{int,m+1}\right)^- d_{I|J}\right.\right. \tag{20}$$
$$\left.\left. + \sum_{\sigma=I\in\partial I\cap\Sigma_{bound-}} \left(\mathbf{V}_{\mathbf{T,I}|\sigma^{bound,m+1}}\right)^- d_{I|1} + \sum_{\sigma=I\in\partial I\cap\Sigma_{bound+}} \left(\mathbf{V}_{\mathbf{T,I}|\sigma^{bound,m+1}}\right)^- d_{I|0}\right]\right)$$

$$-\frac{\Delta t}{V_I\phi}\sum_{\sigma=IJ\in\partial I\cap\Sigma_{int}}\left(\mathbf{V}_{\mathbf{T},\mathbf{I}|\mathbf{J}}{}^{int,m+1}\right)^{-}d_{I|J}S_{w_J}-\frac{\Delta t}{V_I\phi}\sum_{\sigma=I\in\partial I\cap\Sigma_{bound-}}\left(\mathbf{V}_{\mathbf{T},\mathbf{I}|\sigma^{bound,m+1}}\right)^{-}d_{I|1}$$

If we introduce the notation:

$$V_{CFL}=\sum_{\sigma=IJ\in\partial I\cap\Sigma_{int}}-\left(\mathbf{V}_{\mathbf{T},\mathbf{I}|\mathbf{J}}{}^{int,m+1}\right)^{-}d_{I|J}+\sum_{\sigma=I\in\partial I\cap\Sigma_{bound-}}-\left(\mathbf{V}_{\mathbf{T},\mathbf{I}|\sigma^{bound,m+1}}\right)^{-}d_{I|1}$$
$$+\sum_{\sigma=I\in\partial I\cap\Sigma_{bound_+}}-\left(\mathbf{V}_{\mathbf{T},\mathbf{I}|\sigma^{bound,m+1}}\right)^{-}d_{I|0} \tag{21}$$

To ensure the scheme stability, we must satisfy:

$$\Delta t\leq\frac{\inf_I(V_I)\phi}{\sup\left[V_{CFL}\right]\sup_{0\leq S_w\leq 1}f_w'(S_w)} \tag{22}$$

So, the CFL number is thereby defined: $C=\frac{\inf_I(V_I)\phi}{\sup[V_{CFL}]\sup_{0\leq S_w\leq 1}f_w'(S_w)}$

3.2 *The fractional flow function: Brooks and Corey [15] [16]*

The goal of this section is to describe the terms used in the CFL condition formula. Classically, in reservoir multiphase simulations, models are used to define relative permeability. The most known formula was developed by Brooks and Corey and is based on a power law:

$$k_{rw}(S_w)=k_{rw_{max}}\left(\frac{S_w-S_{wi}}{1-S_{wi}-S_{nr}}\right)^{a_w}\quad;\quad k_{rn}(S_w)=k_{rn_{max}}\left(\frac{1-S_w-S_{nr}}{1-S_{wi}-S_{nr}}\right)^{a_n} \tag{23}$$

where a_w and a_n are the Corey exponent. Note that S_w varies between S_{wi} and $1-S_{or}$. In the inequation 22, the fractional flow function $f_w(S_w)$ is present:

$$f_w(S_w)=\frac{\lambda_w}{\lambda_w+\lambda_n}=\frac{\frac{k_{rw}}{\mu_w}}{\frac{k_{rw}}{\mu_w}+\frac{k_{rn}}{\mu_n}}=\mu_n\frac{k_{rw}}{\mu_nk_{rw}+\mu_wk_{rn}}. \tag{24}$$

Its derivative is computed:

$$f_w'(S_w)=\mu_n\frac{k_{rw}'\left[\mu_nk_{rw}+\mu_wk_{rn}\right]-k_{rw}[\mu_nk_{rw}'+\mu_wk_{rn}']}{[\mu_nk_{rw}+\mu_wk_{rn}]^2} \tag{25}$$

3.3 Numerical test case

The aim of this test is to compare the theoretical value C_{Th} of the CFL condition (inequation 22) and C_{Ma} the one found experimentally by slowly raising Δt until the results stop being physically acceptable. Water is injected in a oil-saturated pipe. Its dimensions are: $[L_x, L_y, L_z] = [100, 10, 10]$ m. The built mesh has 100 elements. Each cell I has therefore a size of $1 \times 10 \times 10$ m and $V_I = 100\,\text{m}^3$. We choose these parameters for the test case:

- Independent intrinsic permeability K= 100×10^{-15} m^2 and porosity $\Phi = 0.2$
- Residual oil saturation $S_{nr} = 0.5$ and critical water saturation $S_{wi} = 0.25$
- Oil relative permeability at connate water saturation $k_{rn_{max}} = 1$
- Water relative permeability at the residual oil saturation $k_{rw_{max}} = 0.5$
- Exponents of relative permeability curves $a_n = 2$ and $a_w = 2$
- Entry pressure $P_{in} = 20 \times 10^5$ Pa and exit pressure $P_{out} = 10 \times 10^5$ Pa
- Non-wetting fluid (oil) viscosity $\mu_n = 0.01$ Pa.s
- Wetting fluid (water) viscosity $\mu_w = 0.001$ Pa.s
- Time spent ($\Delta t \times$ iterations number) $T = 173.61$ days

The result of this simulation is drawn in Fig. 2: the water saturation front is apparent.

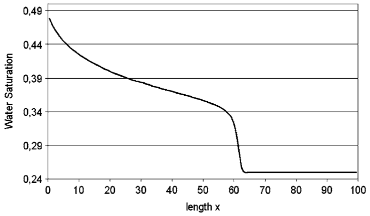


Fig. 2 Water Saturation in the porous media after simulation T=173 days

In the first step, we take care of the theoretical value of C. Regarding the values of the parameters, $\inf_I V_I = 100m^3$ and $\Phi = 0.2$. The derivative of f_w is drawn in Fig. 3 and shows a single maximum: resolving $f_w^{''} = 0$ for these parameters, we find easily $\sup_{0.25 \leq S_w \leq 0.5} f_w^{'}(S_w) \simeq 9.81$

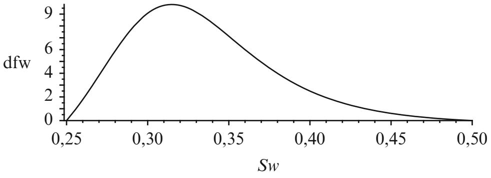


Fig. 3 Derivative of f_w in response of S_w

During a simulation, the value of V_{CFL} is calculated at each iteration and we find $\sup[V_{CFL}] \simeq 1.53 \times 10^{-5}$ m$^3.s^{-1}$. With the equation 22, we can then compute the theorical value of C:

$$C_{Th} \simeq 1.33 \times 10^5 \text{ s.}$$

In a second step, the experimental value of C is computed. Many simulations are performed while Δt is increased manually. For any time during one simulation, if the water saturation curve (Fig. 2) stops being monotonic, Δt has reached its maximum and the result is $C_{Ma} = \Delta t_{max}$. We find:

$$C_{Ma} \simeq 1.33 \times 10^5 \text{ s.}$$

C_{Th} and C_{Ma} are very similar, so we can conclude that if Δt doesn't exceed C_{Th}, the CFL condition won't be violated and the results will be physically acceptable.

3.4 Conclusion

In this paper, we have presented the IMPES scheme for two immiscible and incompressible fluids. For this two-phase flow system, gravity and capillary pressure are neglected. We also consider that there is neither structural evolution nor mass transfer. This CFL condition is a necessary condition for convergence. The approximation was applied to a short example in order to confront the theoretical value with the one found experimentally. More intricate cases must be studied by adding spatial dimensions: a complex mesh with numerous cells will set a significant limit for a heterogeneous test case. Besides, the contribution of gravity and capillary pressure can be studied.

References

1. H.L.Stone, A.O. Jr. Garder, Analysis of gas-Cap or dissolved gas drive reservoirs, Trans., AIME, (1961) 222.
2. L.C.Young, R.E. Stephenson, A generalized compositionnal approach for reservoir simulation, SPEJ (October 1983) 727.
3. K. H. Coats, A note of IMPES and some IMPES-based simulation models, SPEJ (September 2000) 245.
4. K. H. Coats, IMPES stability, the CFL limit, SPEJ, Vol. 8, No. 3, (September 2003).
5. J. W. Sheldon, C. D. Harris, D Bavly, A method for general reservoir behavior simulation on digital computers, paper SPE 1521-G, 1960 SPE annual fall meeting, Denver, Colorado, 2-5 October (1960).
6. J. C. Martin, Simplified equations of flow in gas drive reservoirs and the theorical foundation of multiphase pressure buildup analyses, Trans. AIME 216 (1959).
7. K. H. Coats, Computer simulation of three-phase flow in reservoir, U. Of Austin, Texas (1968).
8. R. G. Fagin, C. H. Stewart, A new approach to the two-dimensional multiphase reservoir simulator, SPEJ 175, Trans., AIME, 237 (June 1966).

9. M. C. Leverett, Flow of oil-water mixture through unconsolidated sands, Trans. AIME 132, 149,(1938).
10. R. D. Wyckoff, H. G. Botset, The flow of gas-liquid mixtures through unconsolidated sands, Physics 7, 325 (1936).
11. J. Wolf, Comparison of mathematical and numerical models for twophase flow in porous media. Diplomarbeit, Institut für Wasserbau, Universität Stuttgart,(2008).
12. R. Helmig, Multiphase flow and transport processes in the subsurface. Springer, (1997).
13. A. Scheidegger, The physics of flow through porous media. In: University of Toronto Press. Toronto and Buffalo, 3rd edition, (1974).
14. J. Niessner, S. Berg, S. Majid Hassanizadeh, Comparison of two-phase Darcy's law with Thermodynamically consistent approach. Transp. Porous Med, DOI 10.1007/s11242-011-9730-0, (2011).
15. R. J. Brooks, A. T. Corey, Hydraulic properties of porous media. In Hydrol. Pap. 3, Colo. State Univ., Fort Collins, (1964).
16. A. Corey, Mechanics of heterogeneous fluids in porous media. In Water Resour., 150 pp., Publ., Fort Collins, Colo., (1977).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Numerical Solution of 2D and 3D Atmospheric Boundary Layer Stratified Flows

Jiří Šimonek, Karel Kozel, and Zbyněk Jaňour

Abstract The work deals with the numerical solution of the 3D turbulent stratified flows in atmospheric boundary layer over the "sinus hill". Mathematical model for the turbulent stratified flows in atmospheric boundary layer is the Boussinesq model - Reynolds averaged Navier-Stokes equations (RANS) for incompressible turbulent flows with addition of the density change equation. The artificial compressibility method and the finite volume method have been used in all computed cases and Lax-Wendroff scheme (MacCormack form) has been used together with the Cebecci-Smith algebraic turbulence model. Computations have been performed with Reynold's number $10^8 \approx u_\infty = 1.5\,\frac{m}{s}$ and with density range $\rho \in [1.2;\ 1.1]\,\frac{kg}{m^3}$.

Keywords CFD, Finite Volume Method, Variable density Flows, Atmospheric Boundary Layer Flows
MSC2010: 65N08, 65N40

1 Mathematical model

Reynolds averaged Navier-Stokes equations for incompressible flows with addition of the equation of density change (Boussinesq model) have been used as a mathematical model for flows in atmospheric boundary layer:

Jiří Šimonek and Karel Kozel
Czech Technical University, Faculty of Mechanical Engineering (Department of Technical Mathematics), Karlovo nám. 13, 121 35 Praha 2, Czech Rep., e-mail: jiri.simonek@fs.cvut.cz, karel.kozel@fs.cvut.cz

Zbyněk Jaňour
Institute of Thermomechanics - Academy of Sciences of the Czech Republic Dolejškova 1402/5, 182 00 Praha 8, Czech Rep. e-mail: janour@it.cas.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_76, © Springer-Verlag Berlin Heidelberg 2011

$$u_x + v_y + w_z = 0 \tag{1}$$

$$u_t + (u^2 + p)_x + (u \cdot v)_y + (u \cdot w)_z = (\nu_e u_x)_x + (\nu_e u_y)_y + (\nu_e u_z)_z \tag{2}$$

$$v_t + (u \cdot v)_x + (v^2 + p)_y + (w \cdot v)_z = (\nu_e v_x)_x + (\nu_e v_y)_y + (\nu_e v_z)_z \tag{3}$$

$$w_t + (u \cdot w)_x + (v \cdot w)_y + (w^2 + p)_z = (\nu_e w_x)_x + (\nu_e w_y)_y + (\nu_e w_z)_z - \frac{\rho}{\rho_0} g \tag{4}$$

$$\rho_t + u \cdot \rho_x + v \cdot \rho_y + w \cdot \rho_z = 0, \tag{5}$$

where (u, v, w) is a velocity vector, $p = \frac{P}{\rho_0}$ (P- static pressure, ρ_0 - initial maximal density), ρ - density, $\nu_e = \nu + \nu_t$, ν - laminar kinematic viscosity, ν_t - turbulent kinematic viscosity computed by the Cebecci-Smith algebraic turbulence model and g - gravity acceleration. Upstream density and pressure are changing depending on height (z-axis) according to the hydrostatic equilibrium pressure function:

$$\rho_\infty(z) = -\frac{\rho_0 - \rho_h}{h} \cdot z + \rho_0 \tag{6}$$

$$\frac{\partial p_\infty}{\partial z} = -\frac{\rho_\infty(z)}{\rho_0} \cdot g \tag{7}$$

The (6) is the linear decreasing function of density and the (7) is the hydrostatic pressure function.

It is possible to separate $p = p_\infty + p'$, where the term p_∞ is the initial state of pressure, the term p' is the pressure disturbance. Using this substitution and $\rho = \rho_\infty + \rho'$ in the system (1) - (5) leads to:

$$u_x + v_y + w_z = 0 \tag{8}$$

$$u_t + (u^2 + p')_x + (u \cdot v)_y + (u \cdot w)_z = (\nu_e u_x)_x + (\nu_e u_y)_y + (\nu_e u_z)_z \tag{9}$$

$$v_t + (u \cdot v)_x + (v^2 + p')_y + (w \cdot v)_z = (\nu_e v_x)_x + (\nu_e v_y)_y + (\nu_e v_z)_z \tag{10}$$

$$w_t + (u \cdot w)_x + (v \cdot w)_y + (w^2 + p')_z = (\nu_e w_x)_x + (\nu_e w_y)_y + (\nu_e w_z)_z - \frac{\rho'}{\rho_0} g \tag{11}$$

$$\rho_t + u \cdot \rho_x + v \cdot \rho_y + w \cdot \rho_z = 0, \tag{12}$$

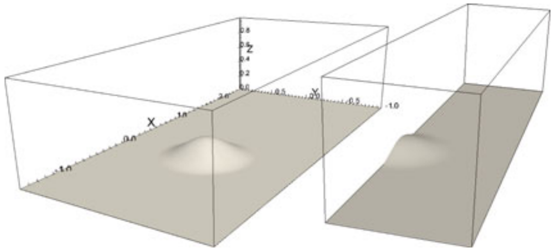


Fig. 1 Computational domains

2 Boundary conditions for 3D computations

Inlet boundary conditions $u = u_\infty = 1.5$, $v = v_\infty = 0$, $w = w_\infty = 0$, $\rho = \rho_\infty(z)$, where $\rho_\infty(z)$ is a linear function which is decreasing with increasing z:

$$\rho_\infty(z) = -\frac{\rho_0 - \rho_h}{h} \cdot z + \rho_0,$$

where $\rho_0 = 1.2\frac{kg}{m^3}$ is a lower (maximal) density and $\rho_h = 1.1\frac{kg}{m^3}$ is a upper (minimal) density.
Outlet boundary conditions: $p' = 0$
Boundary conditions on the wall: $u = 0,\ v = 0,\ w = 0,\ \rho = \rho_0$.
Boundary conditions on the upper domain boundary: $p' = 0$, $\frac{\partial u}{\partial n} = 0$, $\frac{\partial v}{\partial n} = 0$, $\frac{\partial w}{\partial n} = 0$, $\rho = \rho_h$
Boundary conditions on side-walls of the domain: symmetry boundary conditions.

3 Boundary conditions for 2D computations

Inlet boundary conditions $u = u_\infty = 1.0$, $w = w_\infty = 0$, $\rho = \rho_\infty(z)$, where $\rho_\infty(z)$ is a linear function which is decreasing with increasing z:

$$\rho_\infty(z) = -\frac{\rho_0 - \rho_h}{h} \cdot z + \rho_0,$$

where $\rho_0 = 1.2\frac{kg}{m^3}$ is a lower (maximal) density and $\rho_h = 0.6\frac{kg}{m^3}$ is a upper (minimal) density.
Outlet boundary conditions: $p' = 0$
Boundary conditions on the wall: $u = 0,\ w = 0,\ \rho = \rho_0$
Boundary conditions on the upper domain boundary: $p' = 0$, $u = 1.0$, $\frac{\partial w}{\partial n} = 0$, $\rho = \rho_h$

4 Numerical solution

In all cases the artificial compressibility method has been used, i.e. continuity equation is completed by term $\frac{p'_t}{\beta^2}$, $\beta^2 \in R^+$ - then the modified RANS system is valid only for steady state solutions in which $\frac{p'_t}{\beta^2} = 0$. Modified equations can be expressed in a vector form as follows:

$$W_t + F_x + G_y + H_z = R_x + S_y + T_z + K \tag{13}$$

$$W = \left\| \begin{matrix} \frac{p'}{\beta^2} \\ u \\ v \\ w \\ \rho \end{matrix} \right\| \quad F = \left\| \begin{matrix} u \\ u^2 + p' \\ u \cdot v \\ u \cdot w \\ u \cdot \rho \end{matrix} \right\| \quad G = \left\| \begin{matrix} v \\ v \cdot u \\ v^2 + p' \\ v \cdot w \\ v \cdot \rho \end{matrix} \right\| \quad H = \left\| \begin{matrix} w \\ w \cdot u \\ w \cdot v \\ w^2 + p' \\ w \cdot \rho \end{matrix} \right\| \tag{14}$$

$$R = \nu_e \left\| \begin{matrix} 0 \\ u_x \\ v_x \\ w_x \\ 0 \end{matrix} \right\| \quad S = \nu_e \left\| \begin{matrix} 0 \\ u_y \\ v_y \\ w_y \\ 0 \end{matrix} \right\| \quad T = \nu_e \left\| \begin{matrix} 0 \\ u_z \\ v_z \\ w_z \\ 0 \end{matrix} \right\| \quad K = -\frac{\rho - \rho_\infty}{\rho_0} \left\| \begin{matrix} 0 \\ 0 \\ 0 \\ g \\ 0 \end{matrix} \right\| \tag{15}$$

Where W is the vector of conservative variables, F, G, H are convective fluxes, R, S, T are diffusive fluxes and K is the source term, $\nu_e = \nu_{laminar} + \nu_{turbulent}$.

The finite volume method has been used on structured grid of hexahedral cells (uniform in x and y direction, refined near walls in z direction up to $\Delta z = 10^{-5}$, 200x100x80 cells) in 3D and the grid of quadrilateral cells (uniform in x and refined near walls in z direction $\Delta z = 10^{-5}$, 100x40 cells) in 2D.

Lax-Wendroff predictor-corrector scheme (MacCormack form) has been used in a following form:

$$W_i^{n+\frac{1}{2}} = W_i^n - \frac{\Delta t}{V_i}\left(\sum_{k=1}^{6}(\tilde{F} - \tilde{R},\ \tilde{G} - \tilde{S},\ \tilde{H} - \tilde{T})_{i,k}^{n}\mathbf{n}_{i,k}^{0}\Delta S_{i,k}\right) + \Delta t K_i^n \tag{16}$$

$$W_i^{n+1} = \frac{1}{2}(W_i^{n+\frac{1}{2}} + W_i^n) - \frac{\Delta t}{2V_i}\left(\sum_{k=1}^{6}(\tilde{F} - \tilde{R},\ \tilde{G} - \tilde{S},\ \tilde{H} - \tilde{T})_{i,k}^{n+\frac{1}{2}}\mathbf{n}_{i,k}^{0}\Delta S_{i,k}\right) + \frac{\Delta t}{2}K_i^{n+\frac{1}{2}} \tag{17}$$

Convective fluxes have been taken in a forward direction in a predictor step and in a backward direction in a corrector step (see Fig. (2)). Viscous fluxes have been computed centrally (see Fig. (3)).

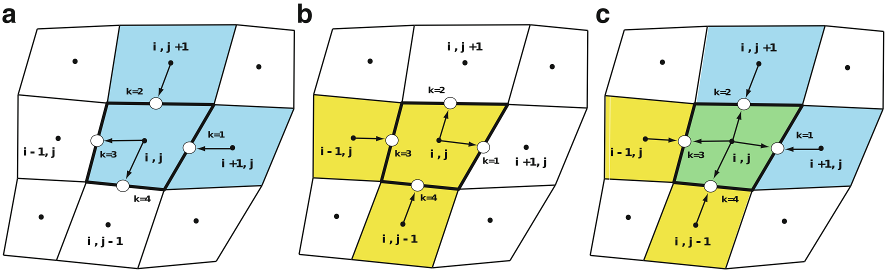


Fig. 2 Stencil for inviscid fluxes computation, (a) predictor step, (b) corrector step, (c) predictor + corrector

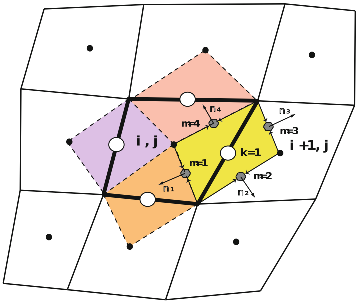


Fig. 3 Stencil for viscous fluxes computation (dual cells)

Jameson's artificial dissipation has been used to stabilize numerical solution.

Cebecci-Smith algebraic turbulence model has been used to compute the turbulent viscosity ν_t. Domain Ω is divided into two subdomains. In the inner subdomain (near walls) the inner turbulent viscosity ν_{Ti} is computed. In the outer subdomain the outer turbulent viscosity ν_{To} is computed. Most common procedure is to compute both turbulent viscosities and use the minimal one:

$$\nu_T = \min\left(\nu_{Ti},\ \nu_{To}\right). \tag{18}$$

For turbulent viscosity computing is necessary to use local systems of coordinates $(X,\ Y)$, where X is parallel with the profile and Y is normal of the profile. In inner subdomain the turbulent viscosity is defined as follows:

$$\nu_{Ti} = \rho l^2 \left|\frac{\partial U}{\partial Y}\right|, \tag{19}$$

where ρ is the density of fluid, $(U,\ V)$ are components of velocity vector in direction of $(X,\ Y)$ and l is given by equation:

$$l = \kappa Y F_D, \tag{20}$$

where:

$$F_D = 1 - \exp\left(-\frac{1}{A^+} u_r Y Re\right), \tag{21}$$

$$u_r = \left(\nu \left|\frac{\partial U}{\partial Y}\right|\right)_{\omega}^{\frac{1}{2}}. \tag{22}$$

In outer subdomain the turbulent viscosity is defined by Clauser‘s equation:

$$\nu_{To} = \rho \alpha \delta^* U_e F_k, \tag{23}$$

$$F_k = \left[1 + 5.5\left(\frac{Y}{\delta}\right)^6\right]^{-1}, \; U_e = U(\delta) \tag{24}$$

where δ is the thickness of boundary layer and

$$\delta^* = \int_0^\delta \left(1 - \frac{U}{U_e}\right) dY. \tag{25}$$

Following values of the constants were used:

$$\kappa = 0.4, \; \alpha = 0.0168, \; A^+ = 26. \tag{26}$$

5 3D Numerical results

Following cases of stratified turbulent flows in atmospheric boundary layer have been computed (see Figs. 7 - 10). Authors consider flows over a geometry with the "sinus hill" with the height 10% of its basis length - half domain symmetrical case and the general 3D geometry and the "hill" with the height 15% of its basis length - general 3D geometry (see Fig. (1)). All the computations have been solved with $Re = 10^8 \approx u_\infty = 1.5\,\frac{m}{s}$ and with density change $\rho_\infty \in [1.2;\; 1.1]$.

6 2D Numerical results

One case with $Re = 6.67 \cdot 10^7 \approx u_\infty = 1.0\,\frac{m}{s}$ and with density change $\rho_\infty \in [1.2;\; 0.6]$ has been solved. One can see in the Figs. (4) (5) (6) the waving character of the flow field. These waves are so called Lee waves which should be seen in the results of the stratified computations. Lee waves were only computed in the 2D case with a coarser mesh (uniform in x and refined near walls in z direction up to $\Delta z = 10^{-5}$, 100x40 cells).

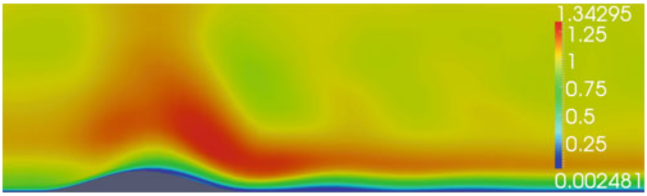


Fig. 4 2D case - Velocity magnitude $\left[\frac{m}{s}\right]$

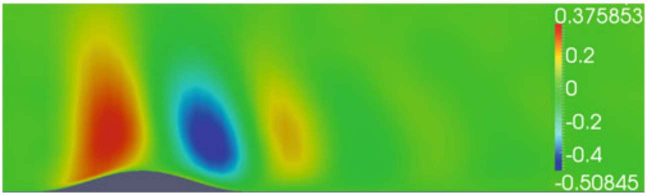


Fig. 5 2D case - Z-velocity $\left[\frac{m}{s}\right]$

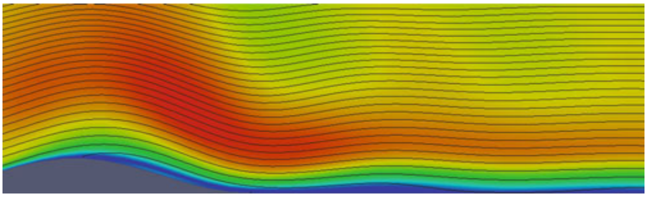


Fig. 6 2D case - stream lines

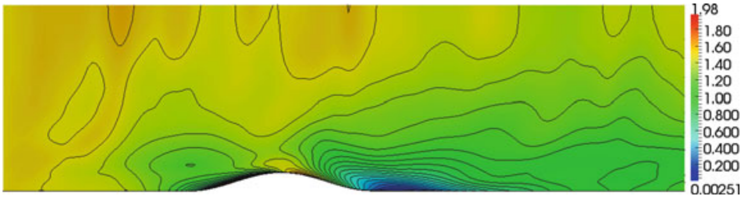


Fig. 7 Half domain symmetrical solution - y-slice in the middle; Velocity mag. $[\frac{m}{s}]$

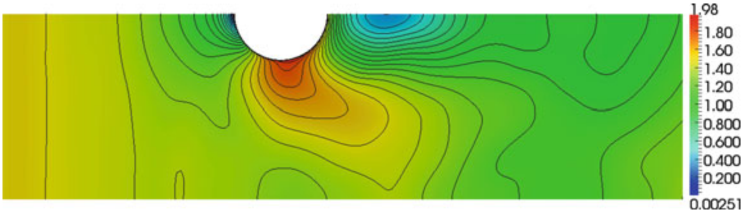


Fig. 8 Half domain symmetrical solution - z-slice in the middle of the hill; Velocity mag. $[\frac{m}{s}]$

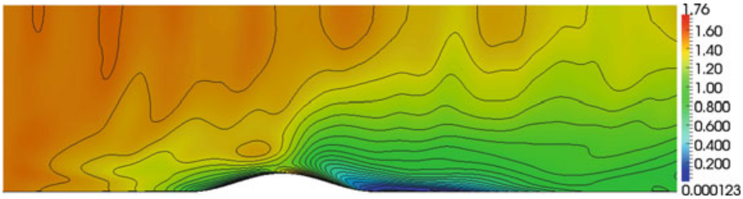


Fig. 9 Full domain solution - y-slice in the middle of the hill; Velocity mag. $[\frac{m}{s}]$

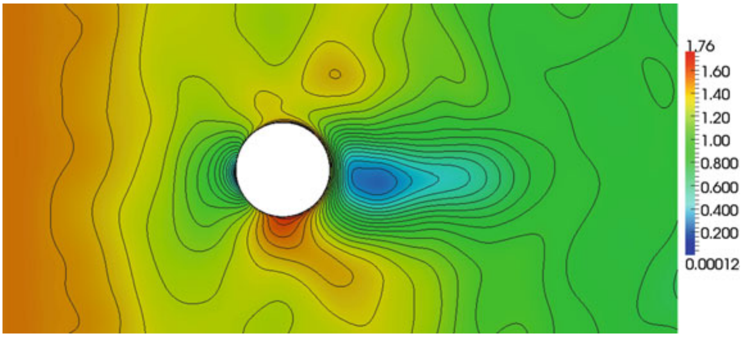


Fig. 10 Full domain solution - z-slice in the middle of the hill; Velocity mag. [$\frac{m}{s}$]

7 Conclusions

Three results of the 3D incompressible turbulent stratified flows in atmospheric boundary layer over the "sinus hill" with Reynolds number $Re = 10^8 \approx u_\infty = 1.5\,\frac{m}{s}$ and with range of density change $\rho \in [1.2;\ 1.1]\,\frac{kg}{m^3}$ have been presented. As one can see in the Figs. (8) and (10) the solution is not symmetrical and therefore it is necessary to perform only the full domain computations in the future. Lee waves were only computed in the 2D case with a coarser mesh.

The future work will be to extend this model for more complex geometries in 3D and to make a comparison with other numerical methods and mathematical models for variable density flows.

Acknowledgements This work was partially supported by Research Plan MSM 6840070010 and MSM 6840770003 and grant GA ČR no. 201/08/0012.

References

1. Eidsvik, K., Utnes, T.: Flow separation and hydrostatic transition over hills modeled by the Reynolds equations, Journal of Wind Engineering and Industrial Aerodynamics, Issues 67 - 68 (1997), p. 408–413
2. Feistauer, M., Felcman, J., Straškraba, I. Mathematical and Computational Methods for Compressible Flow, Clarendon Press, Oxford 2003.
3. Hirsh, C.: Numerical Computation of Internal and External Flows I and II, John Willey and Sons, New York 1991.
4. Fletcher, C., A., J.: Computational Techniques for Fluid Dynamics I and II, Springer Verlag, Berlin 1996.
5. Šimonek, J., Kozel, K., Fraunié, Ph., Jaňour, Z.: Numerical Solution of 2D Stratified Flows in Atmospheric Boundary Layer, Topical Problems of Fluid Mechanics 2008, Prague 2008.
6. Uchida, T., Ohya, Y.: Numerical Study of Stably Stratified Flows over a Two Dimensional Hill in a Channel of Finite Depth, Fluid Dynamics Research 29/2001 (p. 227 - 250).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

On The Numerical Validation Study of Stratified Flow Over 2D–Hill Test Case

Sládek Ivo, Kozel Karel, and Janour Zbynek

Abstract The paper deals with flow validation study performed using our in–house 3D–code which implements mathematical and numerical model capable to compute stratified atmospheric boundary layer flows over hills or terrain obstacles. The objectives of the paper are at first to formulate the applied mathematical/numerical model and at second to present some results from the validation study of thermally stratified flow over an isolated 2D–hill test case. The mathematical model is based on system of RANS equations closed by a two–equation high–Reynolds number k–ε turbulence model. The finite volume method and the explicit Runge–Kutta time integration method are utilized for numerical procedure.

Keywords Turbulent Boundary Layers, Stratification effects, k–ε modeling, Finite Volume Method, Runge–Kutta method
MSC2010: 76F40, 76F45, 76F60, 65N08, 65L06

1 Mathematical model

The flow itself is assumed to be turbulent, viscous, incompressible, stationary and neutrally/stably stratified in general. The mathematical model is based on the Reynolds–averaged Navier–Stokes equations (RANS) modified by the Boussinesq approximation according to which the following decomposition is utilized for pressure p, density ρ and potential temperature Θ

Sládek Ivo and Janour Zbynek
Institute of Thermomechanics, Dolejskova 5, ZIP 182 00, Prague 8, CZ, e-mail: islad@tiscali.cz, janour@it.cas.cz

Kozel Karel
Faculty of Mechanical Engineering, U12101, Karlovo námestí 13, ZIP 121 35, Prague 2, CZ
e-mail: karel.kozel@fs.cvut.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_77, © Springer-Verlag Berlin Heidelberg 2011

$$p = p_0 + p', \quad \rho = \rho_0 + \rho', \quad \Theta = \Theta_0 + \Theta'$$

where $_0$ denotes synoptic large scale part and $'$ concerns the small scale deviation from the synoptic part due to local conditions. The potential temperature Θ is defined as temperature of the atmospheric air after adiabatic compression or expansion to the reference pressure $p_{ref} = 1\,bar$, so $\Theta = T(p_{ref}/p)^{\kappa}$.

The governing equations can be re–casted in the conservative and vector form as follows, [3], [4]

$$(\mathbf{F})_x + (\mathbf{G})_y + (\mathbf{H})_z = (\mathbf{R})_x + (\mathbf{S})_y + (\mathbf{T})_z + \mathbf{f}, \tag{1}$$

where the terms **F**, **G**, **H** represent the physical inviscid fluxes and **R**, **S**, **T** denote the viscous fluxes. The system (1) is then modified in order to be solved by the artificial compressibility method

$$\mathbf{W}_t + \begin{pmatrix} u \\ u^2 + \frac{p'}{\rho_0} \\ uv \\ uw \\ u\Theta' \end{pmatrix}_x + \begin{pmatrix} v \\ vu \\ v^2 + \frac{p'}{\rho_0} \\ vw \\ v\Theta' \end{pmatrix}_y + \begin{pmatrix} w \\ wu \\ wv \\ w^2 + \frac{p'}{\rho_0} \\ w\Theta' \end{pmatrix}_z = \begin{pmatrix} 0 \\ Ku_x \\ Kv_x \\ Kw_x \\ \tilde{K}\Theta'_x \end{pmatrix}_x + \begin{pmatrix} 0 \\ Ku_y \\ Kv_y \\ Kw_y \\ \tilde{K}\Theta'_y \end{pmatrix}_y + \begin{pmatrix} 0 \\ Ku_z \\ Kv_z \\ Kw_z \\ \tilde{K}\Theta'_z \end{pmatrix}_z + \mathbf{f} \tag{2}$$

where

$$\mathbf{W} = (p'/\beta^2,\, u,\, v,\, w,\, \Theta')^T, \quad \mathbf{f} = (0,\, 0,\, 0,\, +g\frac{\Theta'}{\Theta_0},\, -w\Theta'_z)^T \tag{3}$$

where **W** is vector of unknown variables and **f** is the buoyancy force due to the thermal stratification.

The velocity vector components read u, v, w, the term g is the gravitational acceleration, the parameters K, $\tilde{K}$ refer to the turbulent diffusion coefficients for the velocity components and for the potential temperature deviation and β is related to the artificial sound speed.

The synoptic scale part of the potential temperature is taken as $\Theta_0 = \Theta_w + \gamma z$ where Θ_w is the wall potential temperature and γ refers to the wall–normal gradient to be > 0 for the stable stratification and $= 0$ for the neutral stratification.

The system (2) is solved in the computational domain Ω under a stationary boundary conditions for $t \rightarrow \infty$ (t is an artificial time variable) to obtain the expected steady–state solution for all the unknown variables involved in the vector **W**.

2 Turbulence model

Closure of the system of governing equations (2) is achieved by a standard k–ε turbulence model without damping functions [7], [1]. Two additional transport equations are added to the system (2), one for the turbulent kinetic energy abbreviated by

k and one for the rate of dissipation of turbulent kinetic energy denoted by ε. The thermal stratification is taken into account

$$\big(ku\big)_x + \big(kv\big)_y + \big(kw\big)_z = \big(K^{(k)}\,k_x\big)_x + \big(K^{(k)}\,k_y\big)_y + \big(K^{(k)}\,k_z\big)_z + P + G - \varepsilon \quad (4)$$

$$\begin{aligned}\big(\varepsilon u\big)_x + \big(\varepsilon v\big)_y + \big(\varepsilon w\big)_z = \big(K^{(\varepsilon)}\,\varepsilon_x\big)_x + \big(K^{(\varepsilon)}\,\varepsilon_y\big)_y + \big(K^{(\varepsilon)}\,\varepsilon_z\big)_z + \\ C_{\varepsilon 1}(1 + C_{\varepsilon 3}\,R_f)\frac{\varepsilon}{k}\,(P+G) - C_{\varepsilon 2}\frac{\varepsilon^2}{k}\end{aligned} \quad (5)$$

where $G = \beta_\Theta\, g\, \frac{\nu_T}{\sigma_\Theta}\, \frac{\partial\Theta}{\partial z}$ abbreviates the buoyancy term, $R_f = -\frac{G}{P}$ where $P = \tau_{ij}\,\frac{\partial v_i}{\partial x_j}$ denotes the turbulent production term for the Reynolds stress written as

$$\tau_{ij} = -\frac{2}{3}\,k\,\delta_{ij} + \nu_T\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right) \quad (6)$$

and the terms $K^{(k)}$, $K^{(\varepsilon)}$, $\tilde{K}$ stand for the diffusion coefficients and ν_T for the turbulence viscosity

$$K^{(k)} = \nu + \frac{\nu_T}{\sigma_k}, \quad K^{(\varepsilon)} = \nu + \frac{\nu_T}{\sigma_\varepsilon}, \quad \tilde{K} = \nu + \frac{\nu_T}{\sigma_\Theta}, \qquad \nu_T = C_\mu\frac{k^2}{\varepsilon}. \quad (7)$$

The model constants are as follows

$$C_\mu = 0.09,\ \sigma_k = 1.0,\ \sigma_\varepsilon = 1.3,\ C_{\varepsilon 1} = 1.44,\ C_{\varepsilon 2} = 1.92,\ C_{\varepsilon 3} = 0.7. \quad (8)$$

Note that the buoyancy term $G = 0$ in case of neutral stratification.

3 Numerical model

The cell–centered type of finite volume method is applied on structured non–orthogonal grid made of hexahedral control cells Ω_{ijk}. The system of equations (2)+(4)+(5) is integrated over each control cell using the divergence theorem and the mean value theorem, [2]

$$\mathbf{W}_t\Big|_{ijk} = -\frac{1}{\mu_{ijk}}\oint_{\partial\Omega_{ijk}}\Big[(\mathbf{F} - K\cdot\mathbf{R})\,dS_1 + (\mathbf{G} - K\cdot\mathbf{S})\,dS_2 + (\mathbf{H} - K\cdot\mathbf{T})\,dS_3\Big], \quad (9)$$

where $\mathbf{W}_t\Big|_{ijk}$ is the mean value of $\mathbf{W}_t$ over the control cell and $\mu_{ijk} = \int_{\Omega_{ijk}} dV$. The right hand side of (9) is approximated by

$$\mathbf{W}_t\Big|_{ijk} \approx -\frac{1}{\mu_{ijk}} \sum_{l=1}^{6} \Big[(\tilde{\mathbf{F}}_l - K_l \cdot \tilde{\mathbf{R}}_l)\, \Delta S_1^l + (\tilde{\mathbf{G}}_l - K_l \cdot \tilde{\mathbf{S}}_l)\, \Delta S_2^l + (\tilde{\mathbf{H}}_l - K_l \cdot \tilde{\mathbf{T}}_l)\, \Delta S_3^l \Big]. \tag{10}$$

Space discretization of the convective terms in (10) is performed using central differencing while the dual control volumes of octahedral shape is utilized for computation of the viscous terms in (10) at each face of Ω_{ijk}. The resulting semi–discrete system of ordinary differential equations is then integrated in time using the (3)–stage explicit Runge–Kutta method, [6], [9], [5].

The numerical method is second order accurate both in time and space on orthogonal grids. Also the artificial viscosity term of the 4th order is applied due to central differencing of the convective terms in (10) which effectively removes a spurious, high frequency oscillations generated in the computed flow–field.

4 Boundary conditions

The system (2)+(4)+(5) is solved with the following boundary conditions [1], [7]

- Inlet: $u = \frac{u^*}{\kappa} \ln\left(\frac{z}{z_0}\right)$, $v = 0$, $w = 0$, $k = \frac{u^{*2}}{\sqrt{C_\mu}}\left(1 - \frac{z}{D}\right)^2$, $\varepsilon = \frac{C_\mu^{3/4} \cdot k^{3/2}}{\kappa \cdot z}$, $\Theta' = 0$ where the expression for u velocity component is used to cover the boundary layer depth D while constant value $u = U_0$ is prescribed above the boundary layer depth up to the top of computational domain.
- Outlet: homogeneous Neumann conditions for all quantities
- Top: $u = U_0$, $v = 0$, $\frac{\partial w}{\partial z} = 0$, $\frac{\partial k}{\partial z} = 0$, $\frac{\partial \varepsilon}{\partial z} = 0$, $\frac{\partial C}{\partial z} = 0$, $\frac{\partial \Theta'}{\partial z} = 0$
- Wall: standard wall functions are applied and $\frac{\partial C}{\partial n} = 0$ for the concentration and $\Theta' = 0$ for the potential temperature deviation which is equivalent to $\Theta_0 = 300\,K$

where U_0 represents the free–stream velocity magnitude, u^* is the friction velocity, $\kappa = 0.40$ denotes the von Karman constant, z_0 represents the roughness parameter and the parameter D refers to the boundary layer depth.

The wall–function approach enables to apply a wall–coarser grid where near–wall profiles of computed quantities are reconstructed using the algebraic relations, [8].

5 Validation case

The reference numerical results due to Eidsvik and Utnes [10] have been used for comparison. The computational domain extended distance $-15H$ up and $+25H$ downwind of the hill summit and to vertical height $10H$, where $H = 1000\,m$ is the

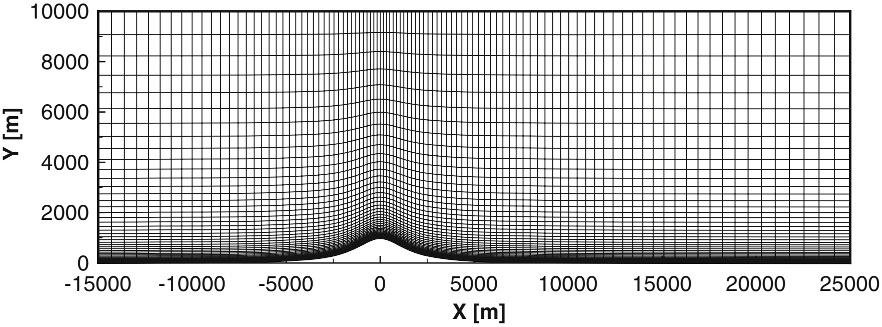


Fig. 1 The whole computational domain and grid 100x40 cells

hill height, see Fig. 1. The integration was performed on grid having 100x40 cells non–uniformly expanding upwind, downwind from the hill summit and vertically from wall using the expansion ratio parameters $ax = 1.04$ and $ay = 1.10$ leading to minimum space increments $\Delta x_{min} = 165\,m$ and $\Delta y_{min} = 20\,m$. Details regarding the grid spacing used by Eidsvik are not available in the reference paper [10].

The flow–field input data used in [10]: the free–stream air velocity $U_0 = 10.5\,m/s$, boundary layer depth of $D = 100\,m$, the friction velocity $u^* = 0.406\,m/s$, the roughness parameters $z_0 = 5\,mm$ and the Reynolds number based on U_0, hill height H and the air kinematic viscosity $\nu = 1.5 \cdot 10^{-5}\,m^2/s$ is $Re = 6.7 \cdot 10^8$.

The inlet profiles for velocity vector components u, v, w, turbulence quantities k, ε as well as for potential temperature deviation Θ' were constructed as described in the Sect. 4.

Totally four different computations have been performed using the same labeling as in [10], N0, N1, N2 and N3. Specifically, the following thermal stratifications of the atmospheric boundary layer were tested

- N0–case: neutral stratification conditions $\gamma = \frac{\partial \Theta_0}{\partial z} = 0\,K/m$
- N1–case: weak stable stratification conditions $\gamma = \frac{\partial \Theta_0}{\partial z} = 3.09 \cdot 10^{-3}\,K/m$
- N2–case: middle stable stratification conditions $\gamma = \frac{\partial \Theta_0}{\partial z} = 12.36 \cdot 10^{-3}\,K/m$
- N3–case: strong stable stratification conditions $\gamma = \frac{\partial \Theta_0}{\partial z} = 27.80 \cdot 10^{-3}\,K/m$

5.1 Numerical results

Separation zone behind hill was found in N0–case under neutral stratification conditions having separation point at $x_1 = 0.9H$ and reattachment point at $x_2 = 3.3H$ downstream from the hill top, see Fig. 2. The recirculation zone in our case is smaller compared to Eidsvik [10] under the same flow conditions where his

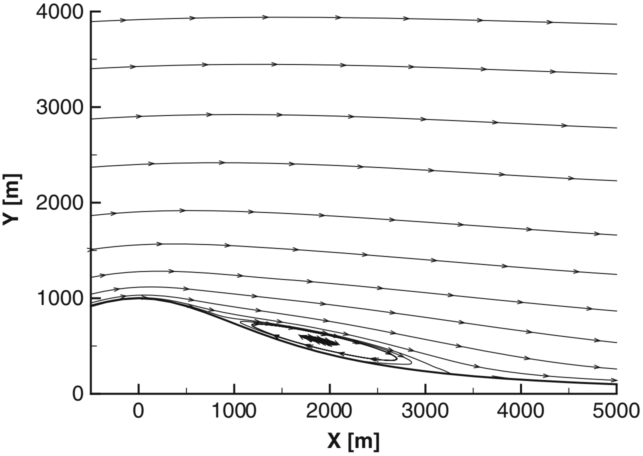


Fig. 2 Zoom to separation zone in N0–case under neutral stratification conditions

separation, reattachment points are $x_1 = 0.8H, x_2 = 5.3H$, respectively. Contours of the wall–normal velocity component w are shown in the following four Figs. 3–6 corresponding to N0–, N1–, N2– and N3–case under neutral, weak, middle and strong stratification conditions, respectively. All contours are labeled using levels of w wall–normal velocity component in $[m/s]$. The lee–waves in cases N1, N2 and N3 are well captured as closed isolines of the wall–normal w velocity component changing sign from ”+” zone where the wave has an increasing slope to ”-” zones where it has a decreasing slope.

According to theory of the internal gravitational waves [11], it is possible to estimate the wavelength of the lee–waves depending on selected stratification conditions. The relation can be written as

$$\lambda_{theoretical} = 2\pi U_0 \left(\frac{g}{\Theta_0} \frac{\partial \Theta_0}{\partial z} \right)^{-1/2} \tag{11}$$

Our prediction of the wavelength is compared to the theory and also to the predictions by Eidsvik [10]

- N0–case: no lee-waves present
- N1–case: $\lambda_{computed} = 6.5\,km$, $\lambda_{Eidsvik} = 6.5\,km$, $\lambda_{theoretical} = 6.3\,km$
- N2–case: $\lambda_{computed} = 4.0\,km$, $\lambda_{Eidsvik} = 3.7\,km$, $\lambda_{theoretical} = 3.1\,km$
- N3–case: $\lambda_{computed} = 2.8\,km$, $\lambda_{Eidsvik} = 2.5\,km$,$\lambda_{theoretical} = 2.1\,km$.

There is a good matching between $\lambda_{computed}$ and $\lambda_{Eidsvik}$ in the N1–case, however there is a difference about $0.3\,km$ in the other two stratification cases N2 and N3. The reason is not clearly known for different wavelength predictions in N2 and N3 cases. It can be attributed to a stretched nature of the computational grid applied

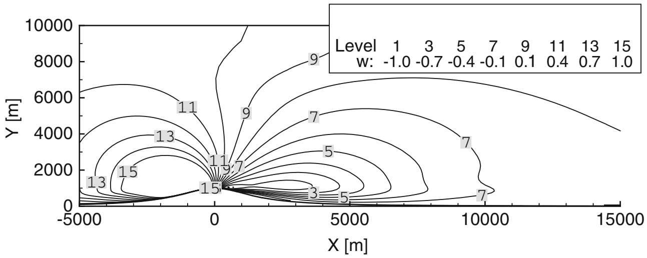


Fig. 3 Contours of the wall–normal *w* velocity component in N0–case under neutral stratification conditions

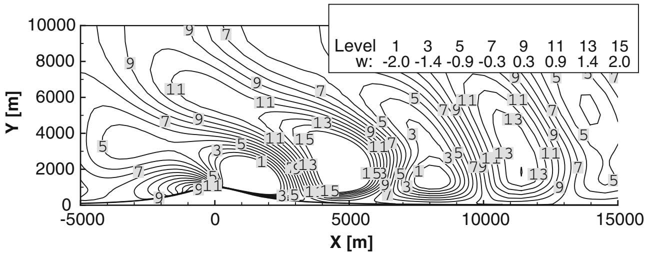


Fig. 4 Contours of the wall–normal *w* velocity component in N1–case under weak stratification conditions

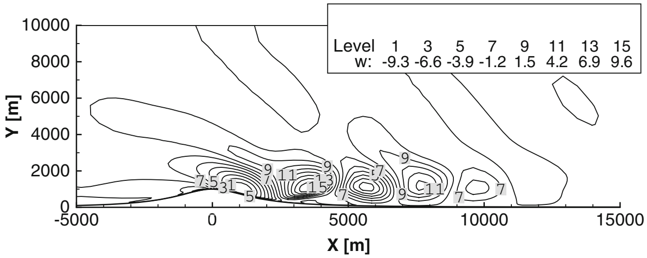


Fig. 5 Contours of the wall–normal *w* velocity component in N2–case under middle stratification conditions

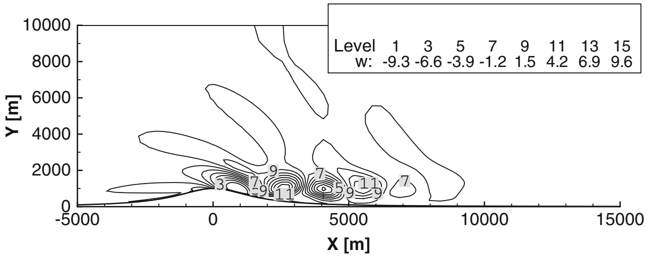


Fig. 6 Contours of the wall–normal *w* velocity component in N3–case under strong stratification conditions

mainly in the vertical direction along the wall and also to the turbulence model. It will be further investigated.

It is also possible to observe a decreasing tendency of the lee–wave amplitude as moving further downstream from the hill summit due to a viscous nature of the flow. Significantly increased flow velocity magnitude was found close to wall on lee–side of the hill mainly for N2 and N3 cases.

6 Conclusion

The above formulated mathematical/numerical model is capable to simulate the atmospheric boundary layer flow problems under different thermal stratification conditions.

The presented validation test case was related to the thermal stratification 2D study where the reference numerical data are due to Eidsvik [10]. The computed lee–waves were observed in all thermally stratified cases. The wavelength was found to be decreasing for increasing thermal stratification conditions. Matching between our predictions of lee wavelength and the reference numerical data is quite good for the weak stratification N1–case. However, there is difference about 0.3 km in the other two cases N2 and N3. Further numerical tests will follow to clarify the differences. The presented work was supported by the Research Plan VZ6840770010.

References

1. Janour Z. (2006): On the mathematical modelling of stratified atmosphere, Institute of Thermodynamics, Report T-470/06, Prague. (in Czech)
2. Sládek I., Kozel K., Janour Z., Gulíková E. (2004): On the Mathematical and Numerical Investigation of the Atmospheric Boundary Layer Flow with Pollution Dispersion, In: COST Action C14 “Impact of Wind and Storm on City Life and Built Environment”, von Karman Institute for Fluid Dynamics, p. 233–242, ISBN 2-930389-11-7.
3. Benes L., Sládek I., Janour Z. (2004): On the Numerical Modelling of 3D – Atmospheric Boundary Layer Flow, In: “Harmonization within Atmospheric Dispersion Modelling for Regulatory Purposes”, Garmisch–Parten Kirchen, Germany, p. 340–344, Vol. 1, ISBN 3-923704-44-5.
4. Bodnár T., Kozel K., Sládek I., Fraunié Ph.: Numerical Simulation of Complex Atmospheric Boundary Layer Flows Problems, In: ERCOFTAC bulletin No. 60: Geophysical and Environmental Turbulence Modeling, p. 5–12, 2004.
5. Sládek I., Bodnár T., Kozel K. (2007): On a numerical study of atmospheric 2D- and 3D-flows over a complex topography with forest including pollution dispersion, Journal of Wind Engineering and Industrial Aerodynamics, Vol. 95, Issues 9–11, p. 1422-1444.
6. Sládek, I. (2005) Mathematical modelling and numerical solution of some 2D– and 3D–cases of atmospheric boundary layer flow, PhD thesis, Czech Technical University, Prague.
7. Castro I.P., Apsley P.P. (1996): Flow and dispersion over topography: A comparison between numerical and laboratory data for two-dimensional flows, Atmospheric Environment, Vol.31, No.6, p.839–850.
8. Wilcox D.C. (1993): Turbulence modeling for CFD, DCW Industries, Inc.

9. Sládek I., Kozel K., Janour Z. (2008): On the 2D-validation study of the atmospheric boundary layer flow model including pollution dispersion, Engineering Mechanics, Vol.16, No.5, p. 323-333, 2009.
10. Eidsvik K.J., Utnes T.: Flow separation and hydraulic transitions over hills modelled by the Reynolds equations, Journal of wind engineering and industrial aerodynamics, Vol. 67 & 68, p.403–413, 1997
11. Holton James: An introduction to dynamic meteorology, Academic Press INC., ISBN 0-12-354360-6, 1979.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Multipoint Flux Approximation Finite Volume Scheme for Solving Anisotropic Reaction–Diffusion Systems in 3D

Pavel Strachota and Michal Beneš

Abstract In [15], our DT–MRI visualization algorithm based on anisotropic texture diffusion is introduced. The diffusion is modeled mathematically by the problem for the Allen–Cahn equation with a space–dependent anisotropic diffusion operator. To preserve its anisotropic properties in the discretized version of the problem, an appropriate numerical treatment is necessary, reducing the isotropic numerical diffusion. The first part of this contribution is concerned with the design and investigation of the finite volume scheme with multipoint flux approximation. Its desirable properties are investigated by means of our technique based on total variation measurement. The second part presents the recent achievements in applying the same scheme to the phase field model of dendritic crystal growth.

Keywords Anisotropic diffusion, finite volume method, multipoint flux approximation, phase field model, microstructure growth
MSC2010: 35K55, 65M08, 65Z05

1 Introduction

The phase field formulation of the Stefan problem [11] describing phase interface evolution during material solidification involves the Allen–Cahn equation [1]. Besides its original purpose, this equation can also be applied in image processing and mathematical visualization [6, 14]. In particular, in order to visualize the streamlines of a given tensor field in 3D, an initial boundary value problem for the modified Allen–Cahn equation with incorporated anisotropy can be used [15], yielding similar results to the LIC method [9]. We begin with the problem

P. Strachota and M. Beneš
Department of Mathematics, Faculty of Nuclear Sciences and Physical Engineering, Czech Technical University in Prague, e-mail: pavel.strachota@fjfi.cvut.cz, michal.benes@fjfi.cvut.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_78, © Springer-Verlag Berlin Heidelberg 2011

formulation and describe its numerical solution using several flux approximation schemes on a rectangular grid. The schemes suffer from an undesired numerical dissipation effect which demonstrates itself as an additional isotropic diffusion of the solution. Hence, we proceed with the development of a measurement technique that would provide for assessing the amount of the numerical diffusion produced by the schemes. A quantitative scheme comparison criterion is thereby created, indicating a clear advantage of our *multipoint flux approximation* (MPFA) scheme. This scheme is then used for the discretization of the complete phase field model of dendritic crystal growth in 3D.

2 Allen–Cahn Equation in Tensor Field Visualization

2.1 Formulation

Assume there is a symmetric positive definite tensor field $\mathbf{D} : \bar{\Omega} \mapsto \mathbb{R}^{3\times 3}$ where $\Omega \subset \mathbb{R}^3$ is a block shaped domain. On the time interval $\mathcal{J} = (0, T)$, the initial boundary value problem for the anisotropic Allen–Cahn equation reads

$$\xi \frac{\partial p}{\partial t} = \xi \nabla \cdot \mathbf{D} \nabla p + \frac{1}{\xi} f_0(p) \qquad \text{in } \mathcal{J} \times \Omega, \tag{1}$$

$$\left.\frac{\partial p}{\partial n}\right|_{\partial\Omega} = 0 \qquad \text{on } \bar{\mathcal{J}} \times \partial\Omega, \tag{2}$$

$$p|_{t=0} = I \qquad \text{in } \Omega \tag{3}$$

where $f_0(p) = p(1-p)\left(p - \frac{1}{2}\right)$. Let $x \in \Omega$. Thanks to $\mathbf{D}(x)$ in the diffusion term on the right hand side of (1), the diffusion of p at x has a directional distribution described by the ellipsoid $\left\{\eta \in \mathbb{R}^3 \middle| \eta^{\mathrm{T}} \mathbf{D}(x)^{-1} \eta = 1\right\}$. In terms of tensor field visualization, we choose the initial condition I in (3) as a noisy texture, preferably an impulse noise. Due to the anisotropic diffusion process carried out by solving (1)–(3), the solution p changes in time from noise to an organized structure. Streamlines of the field of principal eigenvectors of $\mathbf{D}$ can be recognized there as parts with locally similar value of p. The term f_0 efficiently increases contrast of the resulting 3D image provided that the parameter ξ and the final time T are chosen appropriately (in our case by experiment). In order to actually view the resulting 3D image $p(\cdot, T)$, 2D slices through Ω can be helpful.

2.2 Numerical Solution

For numerical solution, the *method of lines* is utilized. Applying a finite volume discretization scheme in space, the problem (1)–(3) is converted to a semidiscrete scheme in the form

$$\xi \frac{\mathrm{d}}{\mathrm{d}t} p_K(t) = \xi \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma}(t) + \frac{1}{\xi} f_{0,K}(t) \qquad \forall K \in \mathscr{T} \tag{4}$$

where $\mathscr{T}$ is an admissible finite volume mesh [7], $K \in \mathscr{T}$ is one particular control volume (cell) and $\mathscr{E}_K$ is the set of all faces of the cell K. $F_{K,\sigma}(t)$ represent the respective numerical fluxes at the time t, which contain difference quotients approximating the derivatives $\partial_x p$, $\partial_y p$, $\partial_z p$ at the center of the face σ. To solve (4), we employ the 4th order Runge–Kutta–Merson solver with adaptive time stepping.

2.3 Numerical Diffusion and Finite Volume Scheme Design

As indicated in the introduction, all schemes introduce a certain amount of *numerical isotropic diffusion* depending on the exact form of $F_{K,\sigma}$. This phenomenon caused by high frequency structures in the solution deteriorates the visual quality of the resulting images by *blurring*. It needs to be suppressed as much as possible e.g. by using difference operators of a sufficient order in $F_{K,\sigma}$ [10].

We have assembled and investigated numerical schemes using the following approximations of the derivatives in the flux term:

- second order central difference approximation with linear interpolation of the missing points in the difference stencil;
- fourth order *multipoint flux approximation* (MPFA) central difference scheme with linear interpolation;
- fourth order MPFA central difference scheme with *cubic* interpolation.

Thereto, a classical forward–backward first order finite difference (FD) scheme has been added for comparison. In the MPFA scheme, the numerical flux $F_{K,\sigma}$ is obtained using the rules below:

- The difference quotient approximating the derivative in the direction perpendicular to the face σ uses a non–equidistant point distribution in order to avoid redundant interpolation (Fig. 1a). Its 1–dimensional analog for a function $u \in \mathrm{C}^1(\mathbb{R})$ can be represented by the formula

  $$\left.\frac{\mathrm{d}u}{\mathrm{d}x}\right|_{x_{i+\frac{1}{2}}} \approx \frac{1}{24h}\left(u_{i-1} - 27u_i + 27u_{i+1} - u_{i+2}\right) \tag{5}$$

  where $x_j = j \cdot h$, $u_j = u\left(x_j\right)$ for $j \in \mathbb{Z}, h > 0$.
- The remaining derivatives are approximated using a uniform 5–point stencil. Again, its 1D analog can be written as

  $$\left.\frac{\mathrm{d}u}{\mathrm{d}x}\right|_{x_i} \approx \frac{1}{12h}\left(u_{i-2} - 8u_{i-1} + 8u_{i+1} - u_{i+2}\right). \tag{6}$$

Moreover, the stencil points (the crosses along the dashed line in Fig. 1b) are interpolated from the neighboring grid nodes using 1–dimensional cubic interpolation.

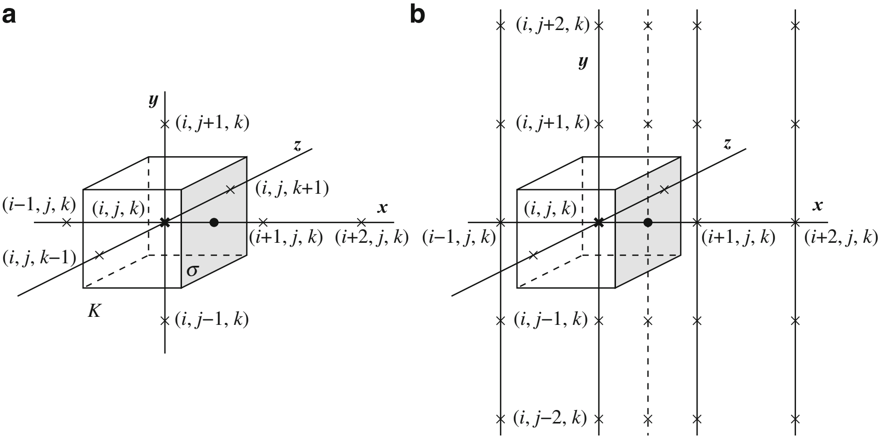


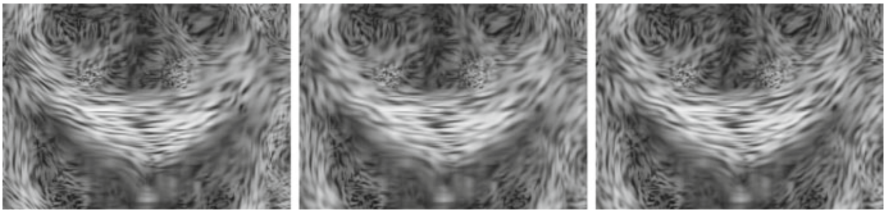
Fig. 1 Point stencils of difference quotients for derivative approximations in the MPFA finite volume scheme

3 Numerical Diffusion Measurement

Having the results available obtained by using different schemes but based on identical input settings, one can try to compare them visually to decide on the scheme with the least artificial diffusion. In Fig. 2, an example of such comparison is demonstrated on a real–data DT–MRI neural tract visualization. In the center part of the images, a major neural tract in the shape of U is displayed in the form of streamlines. It can be observed that the FD scheme produces undesired isotropic diffusion greatly dependent on the prescribed direction of diffusion. This is related to the asymmetry of the difference stencil. The 2nd order central difference flux approximation used in the FV scheme is already symmetric. However, it is clearly outperformed by the scheme based on MPFA which causes significantly weaker blurring.

3.1 Scheme Assessment by Total Variation

In this part we introduce a quantitative measure of the artificial diffusion in the schemes. For this purpose, the total variation of the numerical solution $p^h = p^h(t)$



FD FV 2nd order MPFA cubic 4th order

Fig. 2 Artificial diffusion in different numerical schemes. Crops from colorized DT–MRI visualizations based on real data, transverse plane slice *(Input data: Courtesy of IKEM, Prague)*

finds its rather unusual application. It is defined as

$$TV\left(p^h\right) = \sum_{K\in\mathcal{T}} \left|\nabla_h p_K^h\right| m\left(K\right) \tag{7}$$

where $\nabla^h p_K^h$ represents the discrete approximation of the gradient and $m\left(K\right)$ is the measure of the cell K. From the image processing point of view, the value of TV is proportional to both the number of edges in the image p^h and its contrast. Both these quantities assume their maxima for the noisy initial condition and change in time along with the diffuse evolution of the numerical solution. Performing two computations with identical settings except for the choice of the numerical scheme, it is possible to directly compare the TV values of the results. The scheme producing an image with a greater value of TV exhibits less artificial diffusion as it maintains more edges, more contrast, or both.

3.2 Scheme Comparison Methodology

We have performed extensive testing with phantom input tensor fields to investigate the behavior of the schemes depending on the prescribed direction of diffusion. For each triple of spherical coordinates $(r = 1, \varphi, \theta)$ where $\varphi \in [0, 360°]$, $\theta \in [-90°, 90°]$, let a unit vector $\mathbf{v}_1\left(\varphi,\theta\right) = \left(\cos\varphi\cos\theta, \sin\varphi\cos\theta, \sin\theta\right)$ represent the principal eigenvector of a uniform tensor field $\mathbf{D}\left(\varphi,\theta\right)$, corresponding to the eigenvalue $\lambda_1 = 100$. The remaining eigenvalues are $\lambda_2 = \lambda_3 = 1$ and the eigenvectors $\mathbf{v}_2, \mathbf{v}_3$ complete the orthonormal basis of R^3. Afterwards, a computation is carried out using $\mathbf{D}\left(\varphi,\theta\right)$ as input data and subsequently, TV is evaluated from the resulting datasets. The TV values alone are not of particular interest since they depend on both the grid dimensions and the size of the domain Ω. However, the relative differences of TV between schemes provide the desired information.

The results of the procedure described above performed for all the four schemes in several time levels are shown in Fig. 3. In all graphs, TV is normalized so that

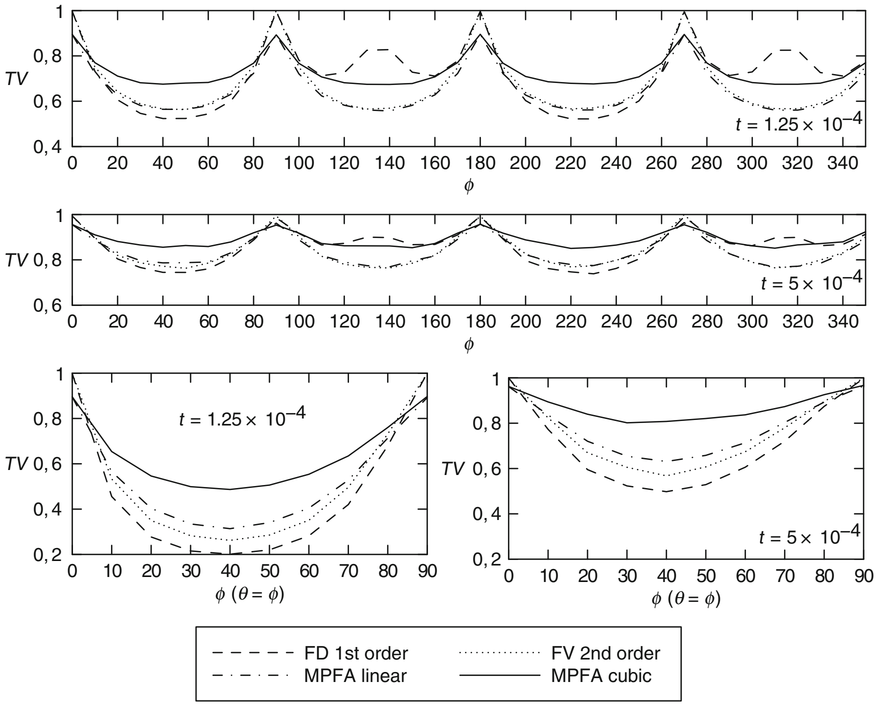


Fig. 3 Comparison of num. schemes based on TV in 2 time levels, $\xi = 5\times10^{-3}$. The investigated angles are $\theta = 0$, $\varphi \in [0°, 350°]$ in the upper two graphs and $\theta \in [0°, 90°]$, $\varphi = \theta$ in the lower two

the maximum in each chart is 1. In the upper two graphs, the latitude θ is fixed to 0 and the longitude φ traverses the angles from 0° to 350° with the step 10°. The lower two graphs depict the "diagonal" cut through the space (φ, θ) in the range from 0° to 90°, including the worst situation for all schemes where $\varphi = \theta = 45°$. Observations from Fig. 3 can be summed up as follows:

- Artificial diffusion clearly depends on $\mathbf{v}_1$ and occurs least when the direction $\mathbf{v}_1$ is aligned with coordinate axes. (In the degenerate case $\lambda_2 = \lambda_3 \to 0$, the equation systems for different rows of grid nodes along $\mathbf{v}_1$ become independent.)
- The performance of all schemes improves (i.e. TV rises) with growing time as the ongoing diffusion gradually limits the frequency spectrum of the solution.
- The FD scheme exhibits an asymmetric behavior; FV schemes are symmetric.
- The FV scheme with MPFA and cubic interpolation outperforms all other schemes in the comparison.

4 The full Phase Field Model for Crystal Growth

We apply the MPFA scheme to the phase field formulation of the simplified Stefan Problem, as studied in [4] and extended to the anisotropic case in [3]. Given a domain Ω and time interval $\mathscr{J}$ as in Sect. 2, the full system of phase field equations reads

$$\frac{\partial u}{\partial t} = \Delta u + L\frac{\partial p}{\partial t} \qquad \text{in } \mathscr{J} \times \Omega, \tag{8}$$

$$\alpha\xi^2\frac{\partial p}{\partial t} = \xi^2\nabla\cdot T^0\left(\nabla p\right) + af_0\left(p\right) - \beta\xi^2\phi^0\left(\nabla p\right)\left(u - u^*\right) \qquad \text{in } \mathscr{J} \times \Omega, \tag{9}$$

$$u|_{t=0} = u_{ini}, \;\; p|_{t=0} = p_{ini} \qquad \text{in } \Omega, \tag{10}$$

with either Dirichlet or Neumann boundary conditions. u represents the temperature field and p the phase field implicitly determining the phase interface Γ by the relation $\Gamma\left(t\right) = \left\{\mathbf{x} \in \mathbb{R}^3 \middle| \, p\left(\mathbf{x}, t\right) = \frac{1}{2}\right\}$. The model parameters involve the melting point of the material u^*, the latent heat L, the attachment kinetics coefficient α, a positive constant a and the parameter ξ controlling the recovery of the sharp interface model [5]. The anisotropic operator T^0 (see [2]) is derived from the dual Finsler metric $\phi^0\left(\eta^*\right)$ as $T^0\left(\eta^*\right) = \phi^0\left(\eta^*\right)\phi^0_\eta\left(\eta^*\right)$ where $\phi^0_\eta = \left(\partial_{\eta_1^*}\phi^0, \partial_{\eta_2^*}\phi^0, \partial_{\eta_3^*}\phi^0\right)^{\mathrm{T}}$. Putting $\phi^0\left(\eta^*\right) = \left|\eta^*\right|\psi\left(-\frac{\eta^*}{\left|\eta^*\right|}\right)$, ψ has the meaning of the anisotropic surface energy [8, 12] and assumes different forms depending on the degree of anisotropy.

The modifications of the MPFA numerical scheme compared to (4) consist in:

1. discretizing the components of ∇p in the last term of (9) by the equidistant stencil (6) at the cell centers,
2. expressing the term $\frac{\partial p}{\partial t}$ in (8) from the equation (9) and using the already computed discretization of its right hand side.

As seen in Fig. 4, the solution of the model can form nontrivial dendritic shapes. It has been confirmed by early comparisons with the standard 2nd order flux

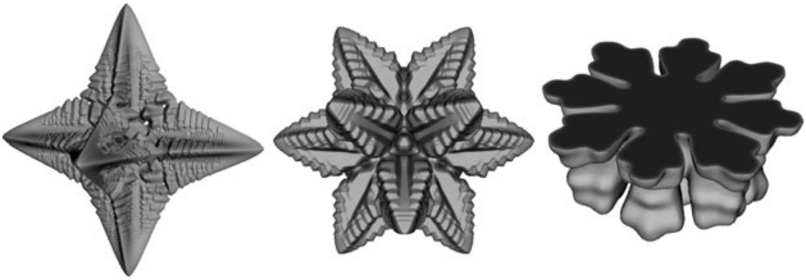


Fig. 4 Sample simulations of dendritic crystal growth with (from left to right) 4–fold, 6–fold *crystalline* anisotropy and a cut through an 8–fold crystal

approximation that the MPFA scheme inclines to the development of dendritic structures more easily, capturing the shape complexity even on lower resolution meshes. This feature was expected due to its low numerical diffusion.

5 Conclusion and Further Research

We have developed an antidiffusive finite volume scheme based on MPFA combined with higher order interpolation. Its properties are demonstrated by our method for measuring the amount of numerical isotropic diffusion. Thorough computational studies based on phantom input data confirm that this technique fulfills the given objective and produces results in agreement with an intuitive notion of blurring observable in images obtained by solving (4). The experimental order of convergence [13] of the MPFA scheme has also been measured and found to be equal to 2. However, the details are beyond the scope of this contribution. Recently, we have finished a MPFA–based parallel numerical algorithm solving the phase field model for crystal growth. Despite the promising results, further investigation of the advantages and verification of the convergence of the numerical solution need to be performed.

Acknowledgements This work was partially supported by the following projects: The project of the Ministry of Education of the Czech Republic MSM6840770010 ”Applied Mathematics in Technical and Physical Sciences”. The Grant Agency of the Czech Technical University in Prague, grant No. SGS10/086/OHK4/1T/14. The project ”Jindřich Nečas Center for Mathematical Modeling”, No. LC06052.

References

1. Allen, S., Cahn, J.W.: A microscopic theory for antiphase boundary motion and its application to antiphase domain coarsening. Acta Metall. **27**, 1084–1095 (1979)
2. Bellettini, G., Paolini, M.: Anisotropic motion by mean curvature in the context of Finsler geometry. Hokkaido Math. J. **25**(3), 537–566 (1996)
3. Beneš, M.: Anisotropic phase-field model with focused latent-heat release. In: FREE BOUNDARY PROBLEMS: Theory and Applications II, *GAKUTO International Series in Mathematical Sciences and Applications*, vol. 14, pp. 18–30 (2000)
4. Beneš, M.: Mathematical and computational aspects of solidification of pure substances. Acta Math. Univ. Comenianae **70**(1), 123–151 (2001)
5. Beneš, M.: Diffuse-interface treatment of the anisotropic mean-curvature flow. Appl. Math-Czech. **48**(6), 437–453 (2003)
6. Beneš, M., Chalupecký, V., Mikula, K.: Geometrical image segmentation by the Allen-Cahn equation. Appl. Numer. Math. **51**(2), 187–205 (2004)
7. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: P.G. Ciarlet, J.L. Lions (eds.) Handbook of Numerical Analysis, vol. 7, pp. 715–1022. Elsevier (2000)
8. Gurtin, M.E.: Thermomechanics of Evolving Phase Boundaries in the Plane. Oxford Mathematical Monographs. Oxford University Press (1993)

9. Hsu, E.: Generalized line integral convolution rendering of diffusion tensor fields. In: Proc. Intl. Soc. Mag. Reson. Med, vol. 9, p. 790 (2001)
10. Lomax, H., Pulliam, T.H., Zingg, D.W.: Fundamentals of Computational Fluid Dynamics. Springer (2001)
11. Meirmanov, A.M.: The Stefan Problem. De Gruyter Expositions in Mathematics. Walter de Gruyter (1992)
12. R. E. Napolitano, S.L.: Three-dimensional crystal-melt Wulff-shape and interfacial stiffness in the Al-Sn binary system. Phys. Rev. B **70**(21), 214,103 (2004)
13. Rice, J.R., Mu, M.: An experimental performance analysis for the rate of convergence of 5-point star on general domains. Tech. rep., Department of Computer Sciences, Purdue University (1988)
14. Strachota, P.: Vector field visualization by means of anisotropic diffusion. In: M. Beneš, M. Kimura, T. Nakaki (eds.) Proceedings of Czech Japanese Seminar in Applied Mathematics 2006, *COE Lecture Note*, vol. 6, pp. 193–205. Faculty of Mathematics, Kyushu University Fukuoka (2007)
15. Strachota, P.: Implementation of the MR tractography visualization kit based on the anisotropic Allen-Cahn equation. Kybernetika **45**(4), 657–669 (2009)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Higher Order Chimera Grid Interface for Transonic Turbomachinery Applications

Petr Straka

Abstract In this paper a higher-order accuracy chimera mesh interface for transonic flow in linear turbine blade cascades is described. Proposed method for calculation of the flow in a transonic blade cascade is applied. Conservation of mass flux through the blade cascade is evaluated. Results of calculation are compared with experimental data.

Keywords Finite-volume method, chimera grid
MSC2010: 35L65, 76H05, 76F40, 65N08

1 Introduction

It is possible to cover a computational domain with structured mesh, even in cases of complex geometry, using the structured chimera mesh. In transonic turbomachinery applications, shock waves structures are formed. The interface between overlapped meshes must operate correctly even if the shock wave intersects. Using of standard interpolation methods (linear, bilinear, polynomial) as well as conservative interpolation methods proposed for subsonic flow [1–3] leads to non-physical reflections of the shock waves at the chimera mesh interface in the case of supersonic flow. A gradient limiting technique is used in this contribution for suppression of the non-physical reflection of shock wave at the chimera mesh interface.

Petr Straka
Aeronautical Research and Test Institute, Plc, Beranových 130, 199 05 Prague - Letňany, Czech Republic, e-mail: straka@vzlu.cz

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_79, © Springer-Verlag Berlin Heidelberg 2011

2 Govering equation

The linear blade cascade is a simple model of an axial turbine stator or rotor wheel. Flow in the linear turbine blade cascade is modeled as 2D compressible viscous turbulent flow of perfect gas. This model is described by the Favre-averaged Navier–Stokes equation

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}_i}{\partial x_i} = \mathbf{Q}\,, \tag{1}$$

where $\mathbf{W} = [\rho,\ \rho u_1,\ \rho u_2,\ e]^{\mathrm{T}}$ is conservative variable vector, ρ is density, u_1 and u_2 are velocity vector components, e is total energy per unit volume, $\mathbf{F}_i = \mathbf{F}_i^{inv} - \mathbf{F}_i^{vis}$ $(i = 1,\ 2)$ stands for flux vectors and $\mathbf{Q}$ is source term. In our case is $\mathbf{Q} = \mathbf{0}$. System (1) is closed by two-equation TNT $k - \omega$ turbulence model [4] which can be formulated in vector form as follows:

$$\frac{\partial \mathbf{W}_t}{\partial t} + \frac{\partial \mathbf{F}_{t,i}}{\partial x_i} = \mathbf{Q}_t\,, \tag{2}$$

where $\mathbf{W}_t = [\rho k,\ \rho \omega]^{\mathrm{T}}$ is vector of turbulent quantities, k is turbulent kinetic energy, ω is specific dissipation rate, $\mathbf{F}_{t,i} = \mathbf{F}_{t,i}^{inv} - \mathbf{F}_{t,i}^{vis}$ $(i = 1,\ 2)$ are flux vectors of turbulent quantities and $\mathbf{Q}_t$ is production and dissipation source term for turbulent quantities.

3 Numerical solution method

The govering equations are discretized on the structured multiblock mesh with quadrilateral elements using a cell-centered finite-volume technique and solved through a time-marching scheme. Both, the mean flow and the turbulence equations are integrated over a control volume D_i and some area integrals are transformed into line integrals along its boundary ∂D_i by the Green-Gauss theorem. Thus

$$\int\!\!\int_{D_i} \frac{\partial \mathbf{W}}{\partial t}\,\mathrm{d}x_1\,\mathrm{d}x_2 + \oint_{\partial D_i} \mathbf{F}_n\,\mathrm{d}s = \int\!\!\int_{D_i} \mathbf{Q}\,\mathrm{d}x_1\,\mathrm{d}x_2\,, \tag{3}$$

where $\mathbf{F}_n = n_i\,\mathbf{F}_i/|\mathbf{n}|$ ($\mathbf{n} = [n_1,\ n_2]$ is the boundary outward normal vector). The line integrals take the following discrete forms

$$\oint_{\partial D_i} \mathbf{F}_n^{inv}\,\mathrm{d}s = \sum_{j=1}^{4} \Phi^{inv}(\mathbf{W}_j^L,\ \mathbf{W}_j^R,\ \mathbf{n}_j)\,s_j\,, \tag{4}$$

$$\oint_{\partial D_i} \mathbf{F}_n^{vis}\,\mathrm{d}s = \sum_{j=1}^{4} \Phi^{vis}(\mathbf{W}_j^C,\ (\nabla\mathbf{W})_{D_j^{dual}},\ \mathbf{n}_j)\,s_j\,. \tag{5}$$

In the mean flow equations, the inviscid numerical fluxes Φ^{inv} are computed by means of the Osher-Solomon flux splitting scheme [5]. Higher order accuracy is achieved through the 2D linear reconstruction method which will be discussed later. In eq. (4) $\mathbf{W}_j^L$ and $\mathbf{W}_j^R$ denote the *left* and *right* states in the corresponding Riemann problem [6]. In the turbulence equations, the inviscid numerical fluxes are computed by the first-order upwind flux-splitting scheme, based on the local convective velocity normal to the cell boundary. The numerical viscous fluxes Φ^{vis} are computed using second-order central scheme, where $\nabla\mathbf{W}$ is approximated through the Green-Gauss formula on a dual cell (Fig. 1) and the local conservative variables vector at the cell boundary is computed as $\mathbf{W}_j^C = (\mathbf{W}_j^L + \mathbf{W}_j^R)/2$.

The time integration is performed using first-order backward Euler scheme

$$\left(\mathbf{I} + \Delta t\,\frac{\partial \mathbf{R}_i^{low}}{\partial \mathbf{W}_i}\right)\Delta\mathbf{W}_i^{n+1/2} + \Delta t\sum_{j=1}^{4}\frac{\partial \mathbf{R}_i^{low}}{\partial \mathbf{W}_j}\Delta\mathbf{W}_j^{n+1/2} = -\Delta t\,\mathbf{R}_i^n\ , \qquad (6)$$

where $\Delta\mathbf{W}^{n+1/2} = \mathbf{W}^{n+1} - \mathbf{W}^n$, and residual approximation $\mathbf{R}_i$ is given as

$$\mathbf{R}_i = \frac{1}{|D_i|}\sum_{j=1}^{4}[\Phi^{inv}(\mathbf{W}_j^L,\ \mathbf{W}_j^R,\ \mathbf{n}_j) - \Phi^{vis}(\mathbf{W}_j^C,\ (\nabla\mathbf{W})_{D_j^{dual}},\ \mathbf{n}_j)]\,s_j\ . \qquad (7)$$

$\mathbf{R}_i^{low}$ denotes first-order approximation in eq. (6).

3.1 Linear reconstruction technique

As mentioned above, higher order accuracy is achived through the 2D linear reconstruction method, which is used for extrapolation of state vector $\mathbf{W}$ at the cell boundary. A piecewise linear function is used for reconstruction of the components

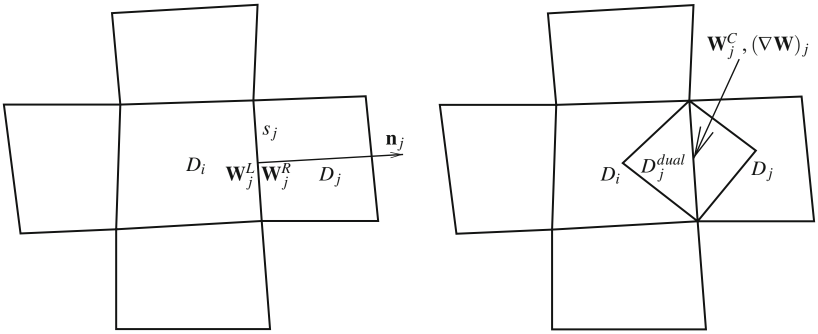


Fig. 1 Scheme of the structured quadrilateral mesh with dual volume

w_k $(k = 1, \dots 4)$ of vector $\mathbf{W}$

$$w_k = f_k(x_1,\ x_2) = a_k + b_k\, x_1 + c_k\, x_2. \tag{8}$$

Coefficients a_k, b_k and c_k are given by supposition

$$w_{k,l} = \frac{1}{|D_l|} \int \int_{D_l} f_k(x_1,\ x_2)\, \mathrm{d}x_1\, \mathrm{d}x_2, \tag{9}$$

where l denotes index of three neighbouring cells. We define four linear functions $f_k^1, \dots, f_k^4$ for reconstruction of state vector components $w_{k\,\mathrm{C}}$ $(k = 1, \dots 4)$ from centre of cell denoted C (Fig. 2) to centre of boundary with cell denoted R, where index l in eq. (9) is for function f_k^1: l = C, R T, for function f_k^2: l = C, T, L, for function f_k^3: l = C, L, B and for function f_k^4: l = C, B, R (R, T, L, B are designations of cells adjacent to cell C - Fig. 2). Components $w^L_{k,\ \mathrm{CR}}$ of reconstructed state vector $\mathbf{W}^L_{\mathrm{CR}}$ (where L means left side in outward normal direction) are given as

$$w^L_{k,\ \mathrm{CR}} = w_{k\,\mathrm{C}} + \delta_{k\,\mathrm{CR}}\ , \tag{10}$$

where $w_{k\,\mathrm{C}}$ are components of state vector in centre of cell C and $\delta_{k\,\mathrm{CR}}$ is defined as

$$\delta_{k\,\mathrm{CR}} = \delta^{min}_{k\,\mathrm{CR}}\ \psi(r(\delta^{min}_{k\,\mathrm{CR}},\ \delta^{max}_{k\,\mathrm{CR}}))\ , \tag{11}$$

$$\delta^{min}_{k\,\mathrm{CR}} = \min_m \{f_k^m(x_{\mathrm{CR},1},\ x_{\mathrm{CR},2}) - w_{k\,\mathrm{C}}\},\ m = 1, \dots, 4\ , \tag{12}$$

$$\delta^{max}_{k\,\mathrm{CR}} = \max_m \{f_k^m(x_{\mathrm{CR},1},\ x_{\mathrm{CR},2}) - w_{k\,\mathrm{C}}\},\ m = 1, \dots, 4\ . \tag{13}$$

There is $r(\delta^{min},\ \delta^{max}) = \delta^{max}/\delta^{min}$ and ψ stands for the limiting function enforcing monotonicity to the solution in eq. (11). The limiters of Van Albada, Van Leer and the super-bee limiter are used in this work.

$$\psi_{VA}(r) = \begin{cases} 0, & r \le 0\ , \\ (r^2 + r)/(r^2 + 1), & r > 0, \end{cases} \tag{14}$$

$$\psi_{VL}(r) = \begin{cases} 0, & r \le 0\ , \\ 2\,r/(r + 1), & r > 0, \end{cases} \tag{15}$$

$$\psi_{SB}(r) = \begin{cases} 0, & r \le 0\ , \\ 2\,r, & 0 \le r \le 1/2\ , \\ 1, & 1/2 \le r \le 1\ , \\ r, & 1 \le r \le 2\ , \\ 2, & r \ge 2\ . \end{cases} \tag{16}$$

3.2 Chimera grid interface

For numerical solution of flow in the turbine blade cascade, O-type mesh is used around the blade profile, H-type mesh covers the chanel between blades as shown in Fig. 3. For simple implementation of the chimera mesh the cells of mesh are classified into three categories: category C0 refers to the regular cell in which the conservative variables vector **W** is solved, category C1 refers to the hidden cell which is skiped during the solution procedure, category C2 refers to the interpolation cell in which the vector **W** is interpolated from overlapped mesh. The same method, as described in paragraph 3.1, is used for interpolation of state vector **W** into the centre of the cell type C2. We need to have state vector $\mathbf{W}^R$ for calculation of flux through boundary between cells type C0 and C2 (Fig. 4 left), which is obtained using the linear reconstruction in cell type C2. One can see, that for correct reconstruction of vector $\mathbf{W}^R$ at the boundary between cells type C0 and C2, we need to have two layers of interpolation cells type C2 (Fig. 4 right).

Proposed chimera mesh interface is very simple for implementation, is higher-order of accuracy and is robust for transonic flow calculation. The mass flux conservation error will be discussed later.

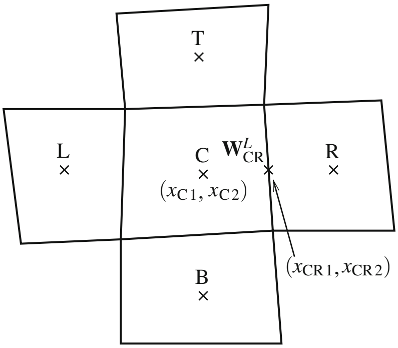

T
L
C
$\mathbf{W}^L_{\mathrm{CR}}$
R
$(x_{\mathrm{C}1}, x_{\mathrm{C}2})$
$(x_{\mathrm{CR}1}, x_{\mathrm{CR}2})$
B

Fig. 2 Scheme of the linear reconstruction on the structured quadrilateral mesh

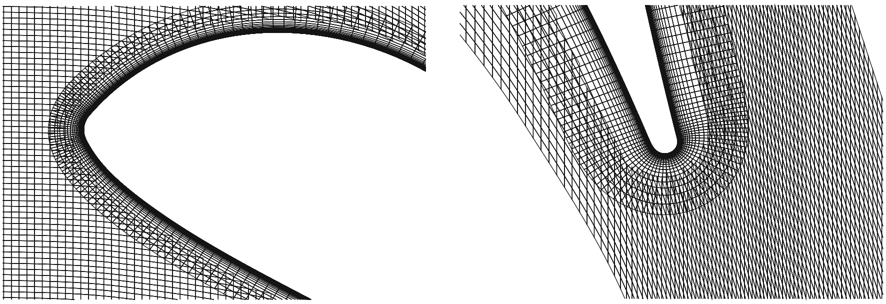


Fig. 3 Detail of chimera mesh around leading and trailing edge

4 Application

The numerical method described in sec. 3 is used for solution of flow in the linear transonic turbine blade cascade VS33R. The computational domain with set boundary conditions types is shown in Fig. 5 (left). The solution was calculated for isentropic output Mach number $0.5 < M_{is,out} < 1.3$, isentropic output Reynolds number $Re_{is,out} = 8.5 \times 10^5$, zero angle of attack and 2 % of inlet turbulence intensity. Transonic flow field in Mach number isolines form is shown in Fig. 5 right. Error of conservation of the mass flux through the blade cascade given as $(1 - q_{in}/q_{out}) \cdot 100$ (where q is the mass flux) is shown in Fig. 7. Further distribution of the total pressure loss coefficient $\eta = 1 - p_{tot,out}/p_{tot,in}$ is compared with the experimental data [7] in Fig. 6.

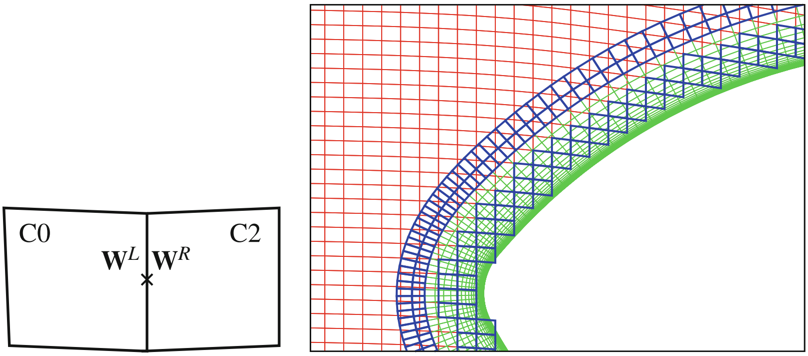


Fig. 4 Left: detail of interface between regular and interpolation cell. Right: two layers of interpolation cells of C2 type (blue)

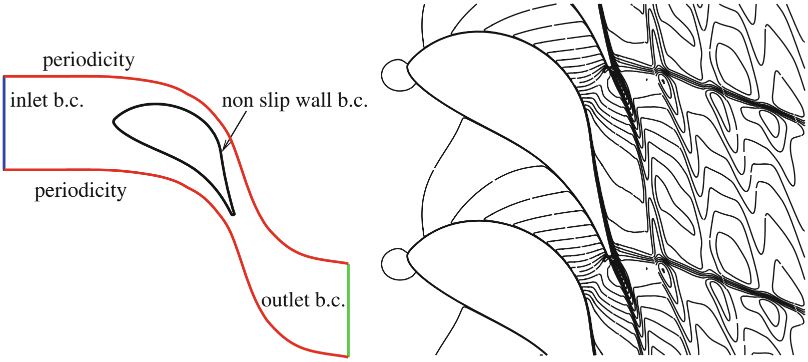


Fig. 5 Left: scheme of computational domain in linear blade cascade. Right: Mach number isolines ($M_{is,out} = 1.3$)

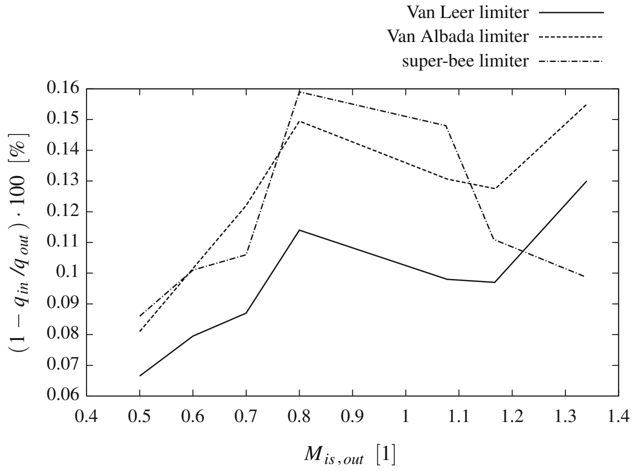


Fig. 6 Distribution of the mass flux conservation error

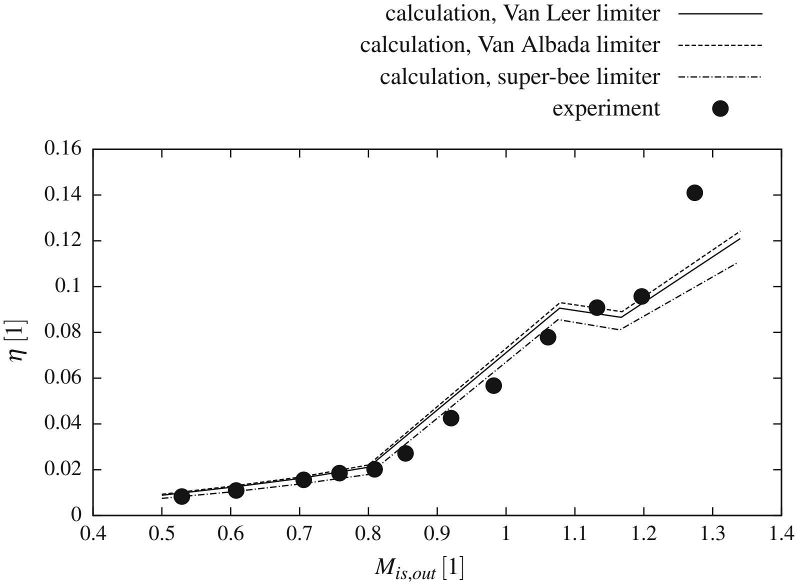


Fig. 7 Distribution of the total pressure loss coefficient

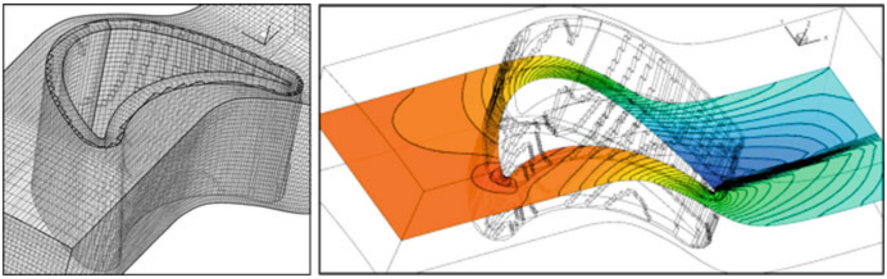


Fig. 8 Chimera mesh and the pressure distribution in 3D problem

4.1 Extension for 3D problem

It is simple to extend the method described in sec. 3 for 3D problems. Two-blocks structured chimera mesh with the hexahedral elements was used for solution of 3D transonic inviscid flow (described by Euler equation: $\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}_i^{inv}}{\partial x_i}$, where $\mathbf{W} = [\rho, \rho u_1, \rho u_2, \rho u_3, e]^{\mathrm{T}}$ and $\mathbf{F}_i^{inv}$ $(i = 1, 2, 3)$ stands for inviscid flux vector) in an axial tubine cascade ST6 [8] for isentropic output Mach number $M_{is,out} = 1.3$, isentropic output Reynolds number $Re_{is,out} = 7.5 \times 10^5$ and angle of attack $\alpha = 45°$. Distribution of pressure is shown in Fig. 8.

5 Conclusion

The chimera mesh interface described in this contribution is simple for the implementation and robust for the transonic turbomachinery applications. Proposed method was applied for the calculation of transonic flow through the linear blade cascade VS33R. The results are in good agreement with the experimental data. Although the condition of conservation is not directly included in the chimera mesh interface, evaluation of the mass flux conservation error (Fig. 6) shows reasonably good conservation.

References

1. Kangle, X., Gang, S.: Assessment of an interface conservative algorithm MFBI in a chimera grid flow solver for multi-element airfoils. Proceedings of the World Congress on Engineering 2009 Vol II, London (2009)
2. Emmert, T., Lafon, P., Bailly, C.: Numerical study of self-induced transonic ow oscillations behind a sudden duct enlargement. Physics of Fluids, **21**, 106105 (2009)
3. Tang, H.S.: Chimera grid method for incompressible flows and its applications in actual problems. 10th Symposium on Overset Composite Grids and Solution Technology, NASA Ames Research Center, CA (2010)

4. Kok, J.C.: Resolving the dependence on freestream values for the $k - \omega$ turbulence model. AIAA Journal. **38**, 1292–1295 (2000)
5. Osher, S., Solomon, F.: Upwind diference schemes for hyperbolic system of conservation laws. Mathematics of Computation, **38**, 339-374 (1982)
6. Toro, E.F.: Riemann solvers and numerical methods for fluid dynamics, A practical introduction, 2nd edn. (Springer, Berlin, 1999)
7. Benetka, J., Kladrubský, M., Valenta, R., Vích, K.: Measurement of turbine blade cascade VS33R. Research report of Aeronautical research and test institute, R-3435/02, (Prague, 2002) (in Czech)
8. Straka, P.: Calculation of 3D unsteady inviscid flow in turbine stage ST6. Research report of Aeronautical research and test institute, R-4910, (Prague, 2010) (in Czech)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Application of Nonlinear Monotone Finite Volume Schemes to Advection-Diffusion Problems

Yuri Vassilevski, Alexander Danilov, Ivan Kapyrin, and Kirill Nikitin

Abstract Two conservative schemes for the nonstationary advection-diffusion equation featuring nonlinear monotone finite volume methods (FVMON) are considered. The first one is an operator-splitting scheme which uses discontinuous finite elements for the advection operator discretization and FVMON for the diffusion operator. The second one introduces another type of FVMON and is implicit second-order BDF in time. A brief description of the schemes and their properties is given. A numerical study is conducted in order to check their convergence and to compare them with conventional methods.

Keywords Monotone finite volumes, advection-diffusion.
MSC2010:65M08

1 Formulation of the methods

1.1 Model Problem

Let Ω be a bounded polyhedral domain in $\mathbb{R}^3$ with a boundary $\partial\Omega$. Consider the following model advection-diffusion problem (for simplicity, with homogeneous Dirichlet boundary conditions):

Yuri Vassilevski, Alexander Danilov, Ivan Kapyrin, and Kirill Nikitin
Institute of Numerical Mathematics RAS, 8 Gubkina, Moscow 119333, Russia, e-mail: vasilevs@dodo.inm.ras.ru, danilov@dodo.inm.ras.ru, kapyrin@dodo.inm.ras.ru, nikitink@dodo.inm.ras.ru

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_80, © Springer-Verlag Berlin Heidelberg 2011

$$\frac{\partial C}{\partial t} - \nabla \cdot D \nabla C + \mathbf{b} \cdot \nabla C = F \quad \text{in } \Omega \times (0, T], \tag{1a}$$

$$C = 0 \quad \text{on } \partial\Omega \times (0, T], \tag{1b}$$

$$C = C_0(x) \text{ in } \Omega \text{ at } t = 0. \tag{1c}$$

Here, C is the contaminant concentration, $\mathbf{b} = \mathbf{b}(x)$ is a conservative convective flux field, $F = F(x)$ is the function of sources or sinks, and $D = D(x)$ is a symmetric positive definite 3×3 diffusion tensor.

1.2 Operator-splitting scheme: DFEM+FVMON

The operator-splitting scheme is designed for tetrahedral grids. It is explained in details in [9], here we give only a brief description. Let a conformal tetrahedral mesh ε_h be introduced in the computational domain Ω. Denote the mesh cells by E_i, $i = 1, \dots, N_E$, the nodes by $O_i = (x_i, y_i, z_i)$, $i = 1, \dots, N_P$. We define the space of discontinuous piecewise linear functions on ε_h

$$W_h = \{v \in L_2(\Omega),\ v_{|E} \in P_1(E), v_{|\partial E \cap \partial\Omega} = 0\ \forall E \in \varepsilon_h\}.$$

The concentration is approximated by piecewise linear discontinuous functions from W_h. The scheme involves splitting over physical components, and the diffusion and convection operators are handled at different substeps. More specifically, at each substep, we solve the incomplete equation (see [8]). The time step of the scheme is defined as follows:

$$I. \int\limits_E \frac{C_h^{n+\frac{1}{2}} - C_h^n}{\Delta t/2} w_h dx - \int\limits_E \mathbf{b} C_h^n \cdot \nabla w_h dx + \int\limits_{\partial E} \mathbf{b} C_{h,in}^n \cdot \mathbf{n} w_h ds = \int\limits_E F^n w_h dx$$

$$\forall w_h \in W_h(E), \quad \forall E \in \varepsilon_h, \tag{2a}$$

$$II. \int\limits_E \frac{C_h^{\star,ad} - C_h^n}{\Delta t} w_h dx - \int\limits_E \mathbf{b} C_h^{n+\frac{1}{2}} \cdot \nabla w_h dx + \int\limits_{\partial E} \mathbf{b} C_{h,in/out}^{n+\frac{1}{2}} \cdot \mathbf{n} w_h ds =$$

$$= \int\limits_E F^{n+\frac{1}{2}} w_h dx \quad \forall w_h \in W_h(E), \quad \forall E \in \varepsilon_h, \tag{2b}$$

$III.$ Slope limiter: $C_h^{*,ad} \longrightarrow C_h^{n+1,ad}$,

$$IV. \int\limits_E \frac{\bar{C}_{h,E}^{n+1} - \bar{C}_{h,E}^{n+1,ad}}{\Delta t} dx = -\sum_{i=1}^{4} r_{E,i}^{n+1} = -\int\limits_{\partial E} \mathbf{r}_E^{n+1} \cdot \mathbf{n} ds \quad \forall E \in \varepsilon_h, \tag{2c}$$

$$V.\ C_h^{n+1} = C_h^{n+1,ad} + (\bar{C}_h^{n+1} - \bar{C}_h^{n+1,ad}). \tag{2d}$$

The convection operator is approximated by an explicit predictor-corrector scheme with an upwind regularization in the corrector. The intermediate concentration $C_h^{n+\frac{1}{2}}$ is calculated in predictor (2a) , while $C_h^{\star,ad}$ in the corrector is calculated from the convective fluxes at the intermediate time level. In the integral over the boundary, $C_{h,in/out}^{n+\frac{1}{2}}$ is taken on the tetrahedron lying upstream. The slope-limiting procedure (2c) is applied to $C_h^{\star,ad}$. Next, implicit scheme (2d) is used to calculate the addition to the mean concentration, $\bar{C}_h$ due to the diffusive fluxes $r_{E,i}^{n+1}$ through the i-th faces of E. The values of $\bar{C}_h^{n+1}$ and $r_{E,i}^{n+1}$ are determined by the nonlinear monotone finite-volume method (FVMON) [3]. Its goal is to derive an as sparse as possible monotone approximation matrix by forming two-point diffusive flux approximations. Then, the solution $\bar{C}_h^{n+1}$ remains nonnegative for nonnegative $\bar{C}_h^{n+1,ad}$. The idea of the two-dimensional FVMON for diffusion problems was set forth in [4]. The details of the present scheme formulation can be found in [9]. The main idea of the scheme construction is based on the following steps:

1. Define the collocation points bearing the degrees of freedom inside each tetrahedron. For cell E we define the point X_E.
2. For two neighbouring tetrahedra E_+, E_- and the corresponding degrees of freedom C_{X_+}, C_{X_-} we define the diffusion flux through the common face e:

$$\mathbf{r}_e \cdot \mathbf{n}_e = K_+(\mathbf{C_X})C_{X_+} - K_-(\mathbf{C_X})C_{X_-}. \tag{3}$$

   Here $\mathbf{C_X}$ is the global vector of unknowns, $\mathbf{n}_e$ is the normal vector to face e. The flux defined in (3) has a two-point approximation stencil with coefficients $K_+(\mathbf{C_X}), K_-(\mathbf{C_X})$ depending on the vector of unknown concentrations. The algorithm of their calculation guarantees positivity of coefficients in case of non-negative vector $\mathbf{C_X}$.
3. Assemble the global nonlinear system and solve it.

To implement step (2c), we find the projection $\hat{c}$ of the solution $C_h^{n+1,ad}$ onto the set $\mathbb{B}$ of the collocation points in cells and solve the FVMON problem for the desired concentrations $\hat{c}^{diff}$ at the points of $\mathbb{B}$:

$$\left(\mathbf{V} + \mathbf{A}(\hat{c}^{diff})\Delta t\right)\hat{c}^{diff} = \mathbf{V}\hat{c}. \tag{4}$$

Here, V is a diagonal matrix of element volumes and $\mathbf{A}(\hat{c}^{diff})$ is an asymmetric matrix whose elements depend on $\hat{c}^{diff}$. All the off-diagonal and diagonal nonzero elements of $\mathbf{A}(\hat{c}^{diff})$ are negative and positive, respectively, for nonnegative $\hat{c}^{diff}$. Moreover, the transpose $(\mathbf{A}(\hat{c}^{diff}))^T$ is row diagonally dominant. Therefore, $(\mathbf{A}(\hat{c}^{diff}))^T$ is an M-matrix and $([(\mathbf{A}(\hat{c}^{diff}))^T]^{-1})_{ij} \geq 0$. Since $(\mathbf{A}^{-1})^T = (\mathbf{A}^T)^{-1}$, the matrix $\mathbf{A}(\hat{c}^{diff})$ is monotone. Nonlinear system (4) is solved by the Picard iteration algorithm

$$\left(\mathbf{V} + \mathbf{A}(\hat{c}^{diff,k})\Delta t\right)\hat{c}^{diff,k+1} = \mathbf{V}\hat{c}$$

with the initial approximation $\hat{c}^{diff,0} = \hat{c} \geq 0$. Since the matrix $\mathbf{V} + \mathbf{A}(\hat{c}^{diff,k})\Delta t$ is monotone for any nonnegative $\hat{c}^{diff,k}$, all the iterative approximations $\hat{c}^{diff,k+1}$ are nonnegative as well; i.e., scheme (4) is monotone.

After $\hat{c}^{diff}$ is determined, we use the formula

$$\bar{C}^{n+1}_{h,E} - \bar{C}^{n+1,ad}_{h,E} = \hat{c}^{diff}_E - \hat{c}_E \quad \forall E \in \varepsilon_h$$

and find the addition to the mean concentrations due to diffusive fluxes, as required in (2e).

1.3 Implicit FVMON scheme

The idea of the implicit nonlinear monotone finite volume scheme is to derive a discretization for the total advective-diffusive flux $\mathbf{r} = -D\nabla C + C\mathbf{b}$ and use the implicit second-order BDF discretization in time. The method is applicable to arbitrary conformal meshes with polyhedral cells and jumping full anisotropic diffusion tensors as well as variable convection fields.

For each cell E, we assign one degree of freedom, C_E, for concentration C. If two cells E_+ and E_- have a common face f and the normal $\mathbf{n}_f$ is exterior to E_+, the two-point flux approximation is as follows:

$$\mathbf{r}_f \cdot \mathbf{n}_f = M_f^+ C_{E_+} - M_f^- C_{E_-}, \tag{5}$$

where M_f^+ and M_f^- are some coefficients. In a linear FV method, these coefficients are equal and fixed. In the nonlinear FV method, they may be different and depend on concentrations in surrounding cells.

Diffusive flux $\mathbf{r}_d = -D\nabla C$ is discretized using the nonlinear two-point flux approximation [1,5] with non-negative coefficients $K_f^{\pm}(C) \geq 0$:

$$(-D\nabla C)_f \cdot \mathbf{n}_f = K_f^+(C)C_{T_+} - K_f^-(C)C_{T_-}. \tag{6}$$

Advective flux $\mathbf{r}_a = C\mathbf{b}$ is approximated via an upwinded linear reconstruction $\mathscr{R}_T$ of the concentration over cell T [6,7]:

$$\mathbf{r}_{f,a} \cdot \mathbf{n}_f = b_f^+ \mathscr{R}_{E_+}(\mathbf{x}_f) + b_f^- \mathscr{R}_{E_-}(\mathbf{x}_f), \tag{7}$$

where

$$b_f^+ = \frac{1}{2}(b_f + |b_f|), \qquad b_f^- = \frac{1}{2}(b_f - |b_f|), \qquad b_f = \frac{1}{|f|}\int_f \mathbf{b} \cdot \mathbf{n}_f \, \mathrm{d}s.$$

We define the reconstruction $\mathscr{R}_E$ as a linear function

$$\mathscr{R}_E(\mathbf{x}) = C_E + \mathbf{g}_E \cdot (\mathbf{x} - \mathbf{x}_E), \qquad \forall \mathbf{x} \in E, \tag{8}$$

with a gradient vector $\mathbf{g}_E$. Since C_E is collocated at the barycenter of E, this reconstruction preserves the mean value of the concentration for any choice of $\mathbf{g}_E$.

The gradient vector $\mathbf{g}_E$ is the solution to the following constrained minimization problem:

$$\mathbf{g}_E = \arg \min_{\tilde{\mathbf{g}}_E \in \mathscr{G}_E} \mathscr{J}_E(\tilde{\tilde{g}}_E), \tag{9}$$

where the functional

$$\mathscr{J}_E(\tilde{\mathbf{g}}_E) = \frac{1}{2} \sum_{\mathbf{x}_k \in \Sigma_E} [C_E + \tilde{\mathbf{g}}_E \cdot (\mathbf{x}_k - \mathbf{x}_E) - C_k]^2$$

measures deviation of the reconstructed function from the targeted values C_k collocated at points $\mathbf{x}_k$ from a set Σ_E of the neighbouring collocation points.

The set of admissible gradients $\mathscr{G}_E$ is defined via three constraints surpressing non-physical oscillations (see [7] for more details).

As the result, we represent the advective flux as the sum of a linear part (the first-order approximation) and a nonlinear part (the second-order correction):

$$\mathbf{r}_{f,a} \cdot \mathbf{n}_f = A_f^+(C)C_+ - A_f^-(C)C_-, \tag{10}$$

where

$$A_f^{\pm}(C) = \pm b_f^{\pm}(1 + \mathbf{g}_{\pm} \cdot (\mathbf{x}_f - \mathbf{x}_{\pm})C_{\pm}^{-1}) \geq 0, \tag{11}$$

subscript ± stands for $E_{\pm}$ and $\mathbf{g}_{\pm} = \mathbf{g}_{E_{\pm}}$.

The resulting nonlinear system is solved using the Picard iterations method. The matrix is monotone on each iteration (see [7]) providing a nonnegative solution.

2 Results of numerical experiments

2.1 Smooth analytical solution

In the first test case the computational domain Ω is a unit cube $[0;1]^3$, the advection field $\mathbf{b} = (0.1; z/10; y/10)$. Two diffusion tensors and the corresponding analitycal solutions are considered:

1. $D = I, \quad C(x,y,z,t) = \left(1-x^2\right) sin(y) e^{-z} sin(t)$
2. $D = 10^{-5} I \quad C(x,y,z,t) = x^2 sin(y) e^{-z} sin(t)$

The first test case features dominating diffusion, the second - dominating advection. The choice of analytical solutions is explained by the desire to obtain nonnegative right-hand sides in the discretization of Eq. (1) in order to verify the monotonicity of the schemes. Recall that only the FVMON guarantees the absence of negative concentrations in this case (although it is unsuitable for problems admitting negative concentrations). Three uniform structured tetrahedral meshes were used in the

Table 1 Solution and flux L_2-errors for DFEM+FVMON scheme

Mesh	case $D = I$		case $D = 10^{-5}I$	
	e_C	e_r	e_C	e_r
1	$1.4 \cdot 10^{-3}$	$2.8 \cdot 10^{-2}$	$4 \cdot 10^{-4}$	$3.8 \cdot 10^{-7}$
2	$4 \cdot 10^{-4}$	$1.3 \cdot 10^{-2}$	$1 \cdot 10^{-4}$	$1.8 \cdot 10^{-7}$
3	$1.2 \cdot 10^{-4}$	$6 \cdot 10^{-3}$	$2.5 \cdot 10^{-5}$	$8.3 \cdot 10^{-8}$

Table 2 Solution L_2-error (e_C)for the BDF FVMON scheme

Mesh	case $D = I$	case $D = 10^{-5}I$
1	$2.2 \cdot 10^{-3}$	$5.3 \cdot 10^{-3}$
2	$5.7 \cdot 10^{-4}$	$1.5 \cdot 10^{-3}$
3	$1.4 \cdot 10^{-4}$	$4.0 \cdot 10^{-4}$

computations. The coarsest of them consisted of 3072 tetrahedra (mesh 1). The other two were obtained by uniformly refining the first and contained 24 576 (mesh 2) and 196 608 (mesh 3) elements, respectively (the mesh size was halved in each refinement procedure). In all the schemes, the time steps used in the tests were 0.025 for mesh 1, 0.0125 for mesh 2, and 0.00625 for mesh 3. The errors were calculated for the solution at the time $T = 1$ and can be seen in tables 1 and 2.

For both schemes we calculate the discrete L_2-error for the solution e_C. For the splitting scheme the diffusion flux L_2-error e_r is computed as well (not implemented yet for the implicit scheme). In both cases we observe second order convergence for the solution, the splitting scheme shows first order convergence for the diffusion fluxes.

2.2 Sharp front resolution

Consider the front of concentration propagating from a constant source occupying a section on the boundary of the domain $\Omega = (0;1) \times (-0.5;0.5) \times (-0.5;0.5)$. More specifically, the following inhomogeneous boundary conditions are set at $x = 0$:

$$C(0,y,z) = \begin{cases} 1 \text{ if } |y| < \frac{1}{4}, |z| < \frac{1}{4}, \\ 0 \qquad \text{elsewhere.} \end{cases}$$

The initial concentration is zero in the entire domain Ω, and the convective flux is $\mathbf{b} = (1,0,0)$. For the solution to have a sharp front, the diffusion tensor is chosen to be small with respect to convection: $D = 10^{-4}I$.

The analytical solution to this problem in the half-space $x \geq 0$ was found in [2]. Passing to the bounded domain Ω, we set Dirichlet conditions on all its boundaries. A non-uniform tetrahedral grid is used for the domain discretization. The numerical solutions are compared with the analytical one at the time $T = 0.5$ and with the

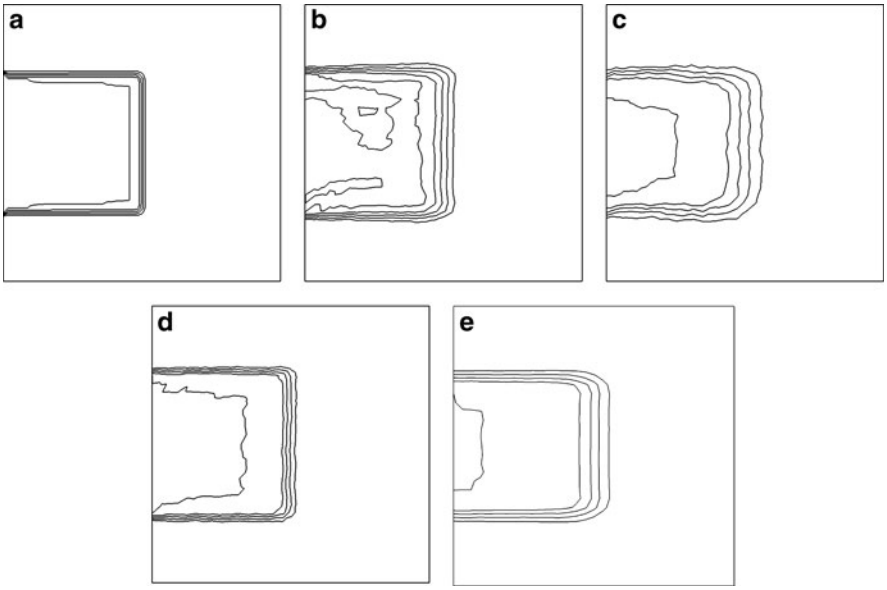


Fig. 1 Analytical and numerical solutions of the front propagation problem:a—analytical; b—implicit BDF P_1-FEM with SUPG; c— BDF implicit HMFEM scheme; d— operator-splitting scheme DFEM+FVMON; e— BDF implicit FVMON scheme.

Table 3 Minima of mean cell concentrations

P_1-FEM	MFEM	DFEM+FVMON	impl. FVMON
$-1.8 \cdot 10^{-1}$	$-6.4 \cdot 10^{-2}$	0	0

solutions obtained by conventional methods: BDF implicit schemes of P_1-FEM with SUPG and HMFEM with upwinding.

Figure 1 displays the exact (a) and approximate (b-e) solutions at T $=0.5$ in the plane $y=0$. The contour lines correspond to the concentration values 0.2, 0.4, 0.6, 0.8, and 1. The conventional methods are nonmonotone, so the solution takes negative values (Table 3), whereas the considered monotone schemes guarantee non-negativity of the solution. Figure 1b shows that FEM with SUPG exhibits strong oscillations. Since the FEM is strongly dispersive, a concentration contour line corresponding to 1 appears in Fig.1b in the area where the solution must be the identical unit. Hybrid MFEM demonstrates high numerical dissipation in Fig.1c. The operator-splitting scheme shows the lowest numerical diffusion (rf.Fig.1d). The implicit FVMON scheme exhibits numerical diffusion comparable to that of FEM with SUPG method.

Conclusions

The two schemes featuring the nonlinear monotone finite volumes prove to be a good alternative to conventional methods especially in cases when monotonicity (in the sense of non-negative concentrations) is important. The operator-splitting scheme makes use of discontinuous finite elements applied for the advection operator discretization. It produces low numerical diffusion. In this scheme diffusion is treated implicitly, and advection explicitly. Thus the time step of the scheme depends on the CFL number. An efficient solution to accelerate its performance is to use different time steps for advection and diffusion. The extra computational burden due to nonlinearity seems to be admissible: the scheme is approximately 20% slower than a linear similar splitting scheme [9].

The BDF implicit scheme of FVMON also shows second order convergence on analytical solutions both for advection and diffusion dominated problems. While suffering from higher numerical diffusion, the scheme has no time step restriction and thus can be more suitable in terms of computational efficiency. Also the scheme is applicable to arbitrary polyhedral cells. Both schemes guarantee non-negativity of the solution in case of non-negative source terms and proper boundary conditions.

Acknowledgements This work has been supported in part by RFBR grants 09-01-00115-a, 11-01-00971-a and the federal program "Scientific and scientific-pedagogical personnel of innovative Russia"

References

1. Danilov A., Vassilevski Yu. A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. *Russian J. Numer. Anal. Math. Modelling*, No.24, pp.207–227, 2009.
2. Feike J.L., Dane J.H. Analitical solutions of the one-dimensional advection equation and two- or three-dimensional dispersion equation. *Water Resources Research*, vol.26, No.7, pp.1475–1482, 1990.
3. Kapyrin I.V. A Family of Monotone Methods for the Numerical Solution of Three-Dimensional Diffusion Problems on Unstructured Tetrahedral Meshes. *Doklady Mathematics*, Vol. 76, No. 2, pp. 734–738, 2007.
4. Le Potier 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, pp.787–792, 2005.
5. Lipnikov K., Svyatskiy D., Vassilevski Yu. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *J. Comp. Phys.* Vol.228, No.3, pp.703–716, 2009.
6. Lipnikov K., Svyatskiy D., Vassilevski Yu. A monotone finite volume method for advection-diffusion equations on unstructured polygonal meshes. *J. Comp. Phys.* Vol.229, pp.4017–4032, 2010.
7. Nikitin K., Vassilevski Yu. A monotone nonlinear finite volume method for advection-diffusion equations on unstructured polyhedral meshes in 3D. *Russian J. Numer. Anal. Math. Modelling*, Vol.25, pp.335–358, 2010.

8. Siegel P., Mose R., Ackerer Ph. and Jaffre J. Solution of the advection-diffusion equation using a combination of discontinuous and mixed finite elements. *International Journal for Numerical Methods in Fluids*, Vol.24, p.595–613, 1997.
9. Vassilevski Yu.V., Kapyrin I.V. Two Splitting Schemes for Nonstationary Convection-Diffusion Problems on Tetrahedral Meshes. *Computational Mathematics and Mathematical Physics*, Vol.48, No. 8, pp. 1349-1366, 2008.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Scale-selective Time Integration for Long-Wave Linear Acoustics

Stefan Vater, Rupert Klein, and Omar M. Knio

Abstract In this note, we present a new method for the numerical integration of one dimensional linear acoustics with long time steps. It is based on a scale-wise decomposition of the data using standard multigrid ideas and a scale-dependent blending of basic time integrators with different principal features. This enables us to accurately compute balanced solutions with slowly varying short-wave source terms. At the same time, the method effectively filters freely propagating compressible short-wave modes. The selection of the basic time integrators is guided by their discrete-dispersion relation. Furthermore, the ability of the schemes to reproduce balanced solutions is shortly investigated. The method is meant to be used in semi-implicit finite volume methods for weakly compressible flows.

Keywords linear acoustics, implicit time discretization, large time steps, balanced modes, multiscale time integration
MSC2010: 35L05, 65M06, 86-08, 86A10

1 Introduction

General circulation models (GCMs) currently used for planetary flow simulations, are based on the Hydrostatic Primitive Equations. This approximation of the full compressible flow equations suppresses vertically propagating sound waves, but it still admits horizontally traveling long wave acoustics, so called "Lamb waves". These and other effects of compressibility are increasingly considered to be non-negligible for planetary-scale dynamics [1, 6]. On the other hand, modern

Stefan Vater and Rupert Klein
Institut für Mathematik, Freie Universität Berlin, Berlin, Germany, e-mail: stefan.vater@math.fu-berlin.de, rupert.klein@math.fu-berlin.de

Omar M. Knio
Dept. of Mechanical Eng., Johns Hopkins University, Baltimore, USA, e-mail: knio@jhu.edu

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_81, © Springer-Verlag Berlin Heidelberg 2011

high-performance computing hardware is beginning to allow the usage of grids with horizontal spacings of merely a few kilometers in such applications (see e.g. [5]). This development introduces considerable numerical difficulties. For explicit time integration schemes, the propagation of sound perturbations introduced by compressibility require very small time steps $\Delta t \sim \Delta x/c$, where Δx is the typical computational grid size, and c a characteristic sound speed. Alternatively, the application of implicit time discretizations solves the problem of the severe time step restriction, but it introduces potentially undesirable numerical dispersion: Most – if not all – existing implicit schemes slow down modes with high wave numbers. Furthermore, there are quite popular schemes, such as the implicit trapezoidal scheme, which preserve the amplitude for all wave numbers. Being a desirable feature at the first glance, it is a potential source of nonlinear instabilities in practice.

In the present work, a new discretization of the linearized acoustic equations is introduced, which overcomes some of the disadvantages of standard implicit discretizations with respect to the representation of compressibility. This means that the scheme should represent the "slaved" dynamics of short-wave solution components induced by slow forcing or arising in the form of high-order corrections to long-wave modes with second-order accuracy. Furthermore, it should eliminate freely propagating compressible short-wave modes that are under-resolved in time, while minimizing dispersion for resolved modes. Here, we describe first successful steps to achieve our goals.

Governing equations. The equations for one dimensional linear acoustics are given by the system

$$\begin{aligned} m_t + p_x &= 0 \\ p_t + c^2 m_x &= q(t, \tfrac{x}{\varepsilon}) \,, \end{aligned} \tag{1}$$

where $p = p(t,x)$ and $m = m(t,x)$ are the pressure and momentum fields. The speed of sound is specified by c, and the source term $q(t, \frac{x}{\varepsilon})$, $\varepsilon \ll 1$, is assumed to be slowly varying in time with small scale variations in space. This source term could simulate the release of latent heat from localized condensation, for example.

For traveling waves $(m, p)(t,x) = (m_0, p_0)\exp(i(\omega t - \kappa x))$, the dispersion relation of (1) is

$$\omega^2 - \kappa^2 c^2 = 0 \,. \tag{2}$$

Thus $\omega(\kappa) = \pm c\kappa$, so that in the continuous system all waves travel with the same velocity, $c = \pm\omega/\kappa$, without dispersion. Also, one can show, that the system preserves a global pseudo energy.

2 Implicit second-order staggered grid schemes

Before the new time integration scheme is introduced, we investigate two standard implicit second-order discretizations. These are the implicit trapezoidal rule and the BDF(2) scheme, which are commonly used in meteorological applications [3].

Their ability to compute reliable approximations to solutions of (1) is discussed with respect to the *discrete-dispersion relations* of these schemes (see [3,9] for details).

Furthermore, the capability of the schemes to reproduce balanced modes is discussed. In the case of slow, short-wave forcing the balance is described by

$$c^2 m_x = q\left(t, \tfrac{x}{\varepsilon}\right) \quad \text{and} \quad p \equiv 0 \tag{3}$$

up to small perturbations introduced by the variation in time of the source term. The schemes should be able to essentially keep this balance. Furthermore, they should reproduce the balanced state in one time step by letting the step going to infinity.

Considering a semi-discretization in time, we leave the choice of a spatial discretization open for the moment. In the subsequent numerical experiments we choose a staggered grid with central differences for simplicity only.

Implicit trapezoidal rule. The implicit trapezoidal rule is derived by integrating the differential equation from t^n to t^{n+1}. The time integral on the right-hand side is then approximated by the trapezoidal quadrature rule. For the system of linear acoustics (1) this results into a Helmholtz problem for p^{n+1}, which is given by

$$p^{n+1} - \frac{c^2 \Delta t^2}{4} \frac{\partial^2 p^{n+1}}{\partial x^2} = p^n - c^2 \Delta t \frac{\partial m^n}{\partial x} + \frac{c^2 \Delta t^2}{4} \frac{\partial^2 p^n}{\partial x^2} + \Delta t\, q^{n+1/2}\,. \tag{4}$$

The update for m is then obtained by

$$m^{n+1} = m^n - \frac{\Delta t}{2} \left(\frac{\partial p^n}{\partial x} + \frac{\partial p^{n+1}}{\partial x} \right)\,. \tag{5}$$

The method is symplectic and A-stable [4]. The discrete-dispersion relation results in a frequency-wave number relationship of the form

$$\omega_{\mathrm{r}} = \pm \frac{2}{\Delta t} \arctan\left(\mathrm{cfl} \cdot \sin\left(\frac{k \Delta x}{2} \right) \right) \tag{6}$$

where $\mathrm{cfl} = \frac{c \Delta t}{\Delta x}$ is the Courant–Friedrichs–Lewy (CFL) number, and the amplification factor per time step is given by $|A| \equiv 1$. Thus, essentially, the frequency ω_{r} depends not only on the wave number k, as in the continuous case, but it is also a function of the CFL number.

Figure 1 shows the discrete-dispersion relation for the trapezoidal rule (dashed line) applied to the linear acoustic equations for a CFL number $\mathrm{cfl} = 1$. The scheme slows down modes at almost all wave numbers, and this behavior is amplified the higher the wave number and the higher the CFL number are. Additionally, the trapezoidal rule is free of numerical dissipation. By letting $\Delta t \to \infty$, one obtains the relations

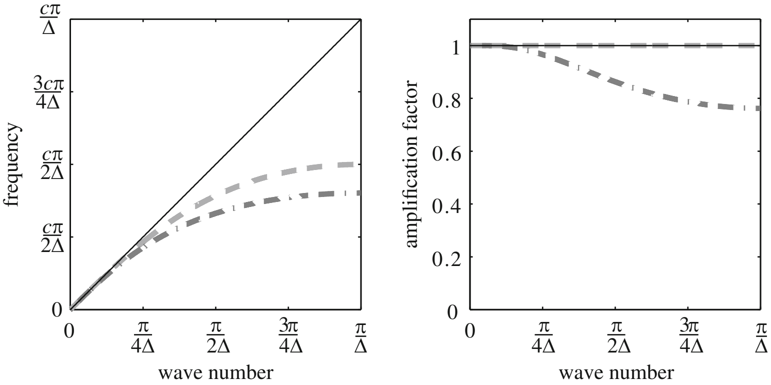


Fig. 1 Discrete-dispersion relations for the trapezoidal (dashed) and the BDF(2) rules (dot-dashed) applied to the linear acoustic equations using cfl = 1. Dispersion relation for continuous system is displayed as black line

$$\frac{c^2}{2}\left(\frac{\partial m^n}{\partial x}+\frac{\partial m^{n+1}}{\partial x}\right)=q^{n+1/2} \quad \text{and} \quad \frac{\partial p^{n+1}}{\partial x}=-\frac{\partial p^n}{\partial x}\,. \tag{7}$$

This reflects the inability to reproduce balanced modes of the trapezoidal rule, and any perturbation of the system cannot dissipate.

BDF(2) scheme. The BDF(2) scheme is a two-step method from the family of the so called *Backward Differentiation Formulas (BDF)*. Here, the left-hand side is approximated by the derivative of a parabola at t^{n+1}, which interpolates the solution at times t^{n-1}, t^n and t^{n+1}. For the acoustic system, this discretization results again in a Helmholtz problem for p^{n+1}, which is

$$p^{n+1}-\frac{4c^2\Delta t^2}{9}\frac{\partial^2 p^{n+1}}{\partial x^2}=\frac{4}{3}p^n-\frac{1}{3}p^{n-1}-c^2\Delta t\left(\frac{8}{9}\frac{\partial m^n}{\partial x}-\frac{2}{9}\frac{\partial m^{n-1}}{\partial x}\right)+\frac{2}{3}\Delta t\, q^{n+1}\,. \tag{8}$$

The update for m is obtained by

$$m^{n+1}=\frac{4}{3}m^n-\frac{1}{3}m^{n-1}-\frac{2}{3}\Delta t\frac{\partial p^{n+1}}{\partial x}\,. \tag{9}$$

The method is A- and L-stable [4].

The discrete-dispersion relation for the BDF(2) scheme is given again in Fig. 1 (dot-dashed line). Concerning the phase error, it shows the same behavior as the trapezoidal rule, although it is considerably amplified. On the other hand, the scheme introduces dissipation for almost all modes. The damping is amplified for high wave and CFL numbers. In the limit $\Delta t \to \infty$ one obtains

$$c^2 \frac{\partial m^{n+1}}{\partial x} = q^{n+1} \quad \text{and} \quad \frac{\partial p^{n+1}}{\partial x} = 0\,,$$

and the scheme achieves balance in a single, sufficiently large, time step. This behavior is characteristic to backward differences formulas by construction [2].

This analysis reveals the dichotomy a practitioner is faced with when having to choose between the two time integrators: Either he could choose to minimize dispersion and preserve the amplitude of well resolved modes by using the trapezoidal rule, or he could ensure that the solution rapidly relaxes to the balanced mode in case of short wave number forcing by the application of the BDF(2) scheme.

3 Multilevel method for long-wave linear acoustics

As described above, the ultimate goal is to filter out all acoustic short wave modes, which are not resolved in time, while sufficiently long wave data is integrated as accurate as possible. Here, we present a strategy for combining the two aspects into one single, scale-dependent numerical time integrator. It is exemplified by using the implicit trapezoidal rule and the BDF(2) scheme as base schemes. One could also use other time integrators (see [9] for a more general presentation), the only restriction is that they are linear in p^{n+1} and m^{n+1}.

Assume that we have scale dependent splittings of the pressure and momentum fields, i.e.

$$p = \sum_{\nu=0}^{\nu_m} p^{(\nu)} \quad \text{and} \quad m = \sum_{\nu=0}^{\nu_m} m^{(\nu)} \tag{10}$$

which could be a quasi-spectral or wavelet decomposition, splitting p and m into (local) high and low wave number components. The idea is to use for each scale component $(p^{(\nu)}, m^{(\nu)})$ a scale dependent blending of the two time integrators. Taking the μ-dependent convex combination with $\mu \in [0,1]$ of the two equations (4) and (8), and summing over the scales results in the Helmholtz problem

$$p^{n+1} - c^2 \Delta t^2 \sum_{\nu=0}^{\nu_M} \left(\frac{\mu_\nu}{4} + \frac{4(1-\mu_\nu)}{9} \right) p_{xx}^{(\nu),n+1} = \sum_{\nu=0}^{\nu_M} \left(\mu_\nu \mathrm{R}_{\mathrm{TRA}}^{p,(\nu)} + (1-\mu_\nu) \mathrm{R}_{\mathrm{BDF2}}^{p,(\nu)} \right)\,, \tag{11}$$

where

$$\begin{aligned} \mathrm{R}_{\mathrm{TRA}}^{p,(\nu)} &= p^{(\nu),n} - c^2 \Delta t\, m_x^{(\nu),n} + \frac{c^2 \Delta t^2}{4} p_{xx}^{(\nu),n} + \Delta t\, q^{(\nu),n+1/2}\,, \\ \mathrm{R}_{\mathrm{BDF2}}^{p,(\nu)} &= \frac{4}{3} p^{(\nu),n} - \frac{1}{3} p^{(\nu),n-1} - c^2 \Delta t \left(\frac{8}{9} m_x^{(\nu),n} - \frac{2}{9} m_x^{(\nu),n-1} \right) + \frac{2}{3} \Delta t\, q^{(\nu),n+1}\,. \end{aligned} \tag{12}$$

The momentum update is derived from the blending of (5) and (9), which is

$$m^{n+1} = \sum_{\nu=0}^{\nu_M} \mu_\nu \left[m^{(\nu),n} - \frac{\Delta t}{2}\left(p_x^{(\nu),n} + p_x^{(\nu),n+1}\right)\right] + \\ (1-\mu_\nu)\left[\frac{4}{3} m^{(\nu),n} - \frac{1}{3} m^{(\nu),n-1} - \frac{2}{3}\Delta t\; p_x^{(\nu),n+1}\right] . \tag{13}$$

The scale splitting is obtained by the application of restriction and prolongation operators used in standard multigrid algorithms. Let $\varphi = \sum \varphi^{(\nu)}$ be a grid function, which is decomposed into parts $\varphi^{(\nu)}$ living on the associated grid levels. Then, the grid function on the coarsest level is obtained by the operation

$$\varphi^{(0)} = \left(R^{(0)} \circ R^{(1)} \circ \cdots \circ R^{(\nu_M-1)}\right) \varphi \tag{14}$$

and the grid functions on finer levels are computed by

$$\varphi^{(\nu)} = \left(I - P^{(\nu-1)} \circ R^{(\nu-1)}\right) \circ \left(R^{(\nu)} \circ R^{(\nu+1)} \circ \cdots \circ R^{(\nu_M-1)}\right) \varphi \,. \tag{15}$$

In our current approach the pressure is decomposed using the *full weighting* (restriction) and the *linear interpolation* (prolongation) operators [7]. They can be defined by their stencil, which are

$$R^{(\nu)} = \frac{1}{4}\left[1\ 2\ 1\right] \quad \text{and} \quad P^{(\nu)} = \frac{1}{2}\left[1\ 2\ 1\right] . \tag{16}$$

On a staggered grid the matching splitting in the momentum field is then defined by (for further details, see [9])

$$R^{(\nu)} = \frac{1}{8}\left[1\ 3\ 3\ 1\right] \quad \text{and} \quad P^{(\nu)} = \left[1\ 1\right]. \tag{17}$$

The description of the scheme is completed by the definition of the weighting function $\mu(\nu)$. In the subsequent tests, it is chosen such that the scheme in (11) and (13) associates the standard implicit trapezoidal scheme with all pressure modes corresponding to coarse grids with grid-CFL number cfl ≤ 1, while we nudge the discretization towards BDF(2) for pressure modes living on grids with cfl > 1.

4 Numerical Results

Here, we shortly describe a test case with "multiscale" initial data in a periodic domain $x \in [0, 1]$. Pressure and momentum fields are chosen in such a way that one obtains a right running acoustic simple wave with a sound speed of $c = 1$. The initial conditions are displayed in Fig. 2 (top row). No source term is present

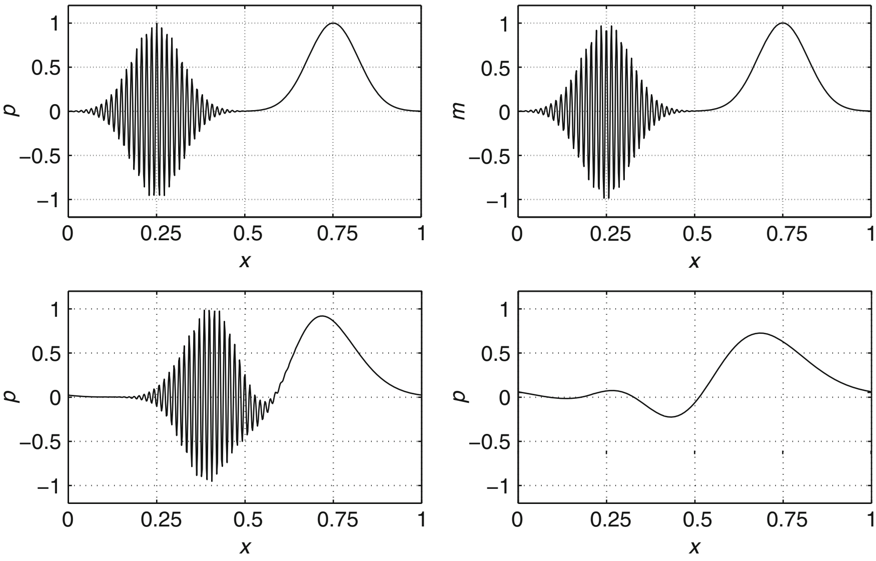


Fig. 2 Top row: "Multiscale" initial data. Bottom row: Numerical solution (pressure) with cfl $=$ 10 at time $t_{\text{end}} = 3$ using the trapezoidal rule (left) and the BDF(2) scheme (right). Grid with 512 cells

(for further details see [9]). We use a grid with 512 cells (i.e. $\Delta x = 1/512$) and a CFL number cfl $= 10$. The results are compared at a final time $t_{\text{end}} = 3.0$, which is equivalent to 154 time steps. At this time the exact solution is identical to the initial data, and the wave has traveled three times across the domain.

The implicit trapezoidal rule produces the results in Fig. 2 (bottom left). Here, and in the following, only pressure is displayed, since the momentum field is essentially the same. The results show what has already been revealed by the discrete-dispersion relation, i.e., the scheme achieves large-CFL stability by slowing down the short wave components of the solution. While the long-wave pulse has traveled at nearly correct speed, the short-wave oscillations have essentially stayed in place. Furthermore, their amplitude has not diminished.

A different behavior is displayed by the BDF(2) scheme (Fig. 2, bottom right). It has considerably more dispersion than the trapezoidal rule, and the damping of the scheme results in a smaller final amplitude, even for the long wave data. On the short scales, the diffusion is so high that at the final time this part of the solution has essentially vanished. Thus, the scheme is able to balance the short-wave modes that are not resolved in time, but it pays the price of simultaneously damping and dispersing the long scales.

The result of the simulation using the new blended scheme with five grid levels is displayed in Fig. 3. For comparison, the result of the trapezoidal rule applied only to long wave data is also shown (dashed line). As one can see, the two results are nearly identical: The short wave data is filtered in such a way that only the long wave

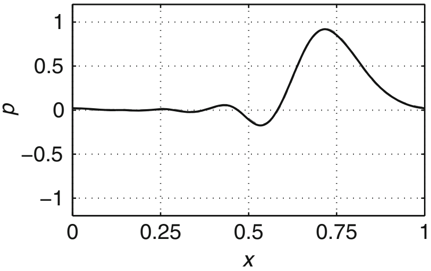


Fig. 3 Numerical solution (pressure) using the blended scheme on a grid with 512 cells and cfl $=$ 10 at time $t_{\mathrm{end}} = 3$ (black line). For comparison, the result of trapezoidal rule obtained with only long wave initial data is plotted as gray dashed line

data is left after some time. On the other hand, the long wave data is integrated as well as one could hope when using a second-order method.

5 Conclusion

The presented scheme effectively filters freely propagating compressible short-wave components, which cannot be accurately represented at long time steps. At the same time, dispersion and the amplitude errors for long-wave modes are minimized. Further tests show that in the presence of a source term, which slowly varies in time but has rapid spatial variations, solutions relax to an asymptotic balanced state (see [9]).

One of the next goals is to apply this scheme into a semi-implicit scheme for weakly compressible flows. The latter is an extension of a second-order projection method for incompressible flows as described in [8]. By using the trapezoidal rule in the implicit part of the scheme, one is faced with instabilities near shocks. This can partly be cured by so called off-centering. However, it also decreases the order of the scheme to one. The authors hope to obtain a second-order version by using the new scheme described in this note.

Acknowledgements R.K. thanks Piotr Smolarkiewicz for stimulating and challenging discussions on the theme of this paper. The authors thank Deutsche Forschungsgemeinschaft for its partial support of this work through grants KL 611/14 and KL 611/22 (MetStröm), and the Leibniz Association for partial support through its PAKT program.

References

1. Davies, T., Staniforth, A., Wood, N., Thuburn, J.: Validity of anelastic and other equation sets as inferred from normal-mode analysis. Q. J. R. Meteorolog. Soc. **129**(593), 2761–2775 (2003)
2. Deuflhard, P., Bornemann, F.: Scientific Computing with Ordinary Differential Equations, *Texts in Applied Mathematics*, vol. 42. Springer (2002)

3. Durran, D.R.: Numerical Methods for Fluid Dynamics: With Applications to Geophysics, 2 edn. No. 32 in Texts in Applied Mathematics. Springer (2010)
4. Hairer, E., Lubich, C., Wanner, G.: Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations. Springer (2006)
5. Ohfuchi, W., Nakamura, H., Yoshioka, M., Enomoto, T., Takaya, K., Peng, X., Yamane, S., Nishimura, T., Kurihara, Y., Ninomiya, K.: 10-km mesh meso-scale resolving simulations of the global atmosphere on the Earth Simulator: Preliminary outcomes of AFES. J. Earth Simulator **1**, 8–34 (2004)
6. Smolarkiewicz, P.K., Dörnbrack, A.: Conservative integrals of adiabatic Durran's equations. Int. J. Numer. Methods Fluids **56**(8), 1513–1519 (2008)
7. Trottenberg, U., Oosterlee, C., Schüller, A.: Multigrid. Academic Press (2001)
8. Vater, S., Klein, R.: Stability of a Cartesian grid projection method for zero Froude number shallow water flows. Numer. Math. **113**(1), 123–161 (2009)
9. Vater, S., Klein, R., Knio, O.M.: A scale-selective multilevel method for long-wave linear acoustics. Acta Geophysica (2011). Submitted

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Nonlocal Second Order Vehicular Traffic Flow Models And Lagrange-Remap Finite Volumes

Florian De Vuyst, Valeria Ricci, and Francesco Salvarani

Abstract In this paper a second order vehicular macroscopic model is derived from a microscopic car–following type model and it is analyzed. The source term includes nonlocal anticipation terms. A Finite Volume Lagrange–remap scheme is proposed.

Keywords Vehicular traffic flow, modeling, car–following, microscopic, macroscopic, finite volumes, Lagrange–remap
MSC2010: 65M08, 65M22, 65P40, 65Y20, 65Z05

1 Motivation and introduction

There are many ways to describe and model a vehicular traffic flow. Microscopic models e.g. [3] describe the interaction between two successive vehicles. It is known that car–following models may have a complex dynamics (see for example [8, 9]) and are able to reproduce all the flow regimes. In the macroscopic models, conservation laws and balance equations on mean quantities are searched. Since the pioneer works by Lighthill, Whitham and Richards (LWR model), numerous improvements and contributions have been proposed. In 2000, Aw and Rascle [1] derived an interesting second order model that fixed the drawbacks of Payne's

Florian De Vuyst
Centre de Mathématiques et de leurs Applications, École Normale Supérieure de Cachan, 61 avenue du Président Wilson, 94235 Cachan France, e-mail: devuyst@cmla.ens-cachan.fr

Valeria Ricci
Dipartimento di Metodi e Modelli Matematici Universita' di Palermo, Viale delle Scienze, Edificio 8, 90128 Palermo, Italy, e-mail: valeria.ricci@unipa.it

Francesco Salvarani
Dipartimento di Matematica, Università degli Studi di Pavia, Via Ferrata 1 - I-27100 Pavia, Italy, e-mail: francesco.salvarani@unipv.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_82, © Springer-Verlag Berlin Heidelberg 2011

model, emphasized by Daganzo [4]. More recently, Aw et al. [2] derived the Aw–Rascle model from microscopic follow–the–leader models. Illner et al. [7] were also able to retrieve the Aw–Rascle model from a kinetic Vlasov description. For related works, see for example [5,6]. In this paper, a continuum traffic flow model is derived from a more complex car–following model.

2 Car–following rule and microscopic model

Let us consider a vehicular traffic flow made of N vehicles, indexed by i, $i = 1, \dots, N$. For simplicity, we will assume that all the vehicles are identical, of length ℓ. The car indexed by i follows the car $(i+1)$. At time t, the vehicle i is located at position $x_i(t)$ with speed $\dot{x}_i = v_i$. The spatial gap between the two vehicles i and $(i+1)$ is then given by $x_{i+1}(t) - x_i(t) - \ell$ (see Fig. 1). The maximum (permitted) speed will be denoted by v_M. Let us also denote by $g^s(v_{,i}, v_{i+1})$ the safety spatial gap for the vehicle i, depending on the vehicle speeds i and $(i+1)$. A simple relaxation rule for the spatial gap is

$$\frac{d}{dt}(x_{i+1} - x_i - \ell) = \frac{g^s(v_i, v_{i+1}) - (x_{i+1} - x_i - \ell)}{x_{i+1} - x_i - \ell}\, a_{i,i+1} \qquad (1)$$

where $a_{i,i+1}$ is a local characteristic speed. The denominator forbids the collision between the two vehicles. Then we get a target speed v_i^{target} equal to

$$v_i^{target} = v_{i+1} + \left(1 - \frac{g^s(v_i, v_{i+1})}{x_{i+1} - x_i - \ell}\right) a_{i,i+1}. \qquad (2)$$

A simple acceleration rule toward the target speed is given by the relaxation scheme

$$\frac{dv_i}{dt} = \frac{v_i^{target} - v_i}{\lambda} = \frac{v_{i+1} - v_i}{\lambda} + \left(1 - \frac{g^s(v_i, v_{i+1})}{x_{i+1} - x_i - \ell}\right) \frac{a_{i,i+1}}{\lambda} \qquad (3)$$

using a characteristic relaxation time $\lambda > 0$. Let us comment three interesting cases. If $0 < x_{i+1} - x_i - \ell \ll 1$, then there is a strong breaking in order not to collide. If $x_{i+1} - x_i - \ell \equiv g^s(v_i, v_{i+1})$, the vehicle i is at the right safe distance, and in that case we have the simple car–following rule $\dot{v}_i = (v_{i+1} - v_i)/\lambda$. If $x_{i+1} - x_i \gg 1$,

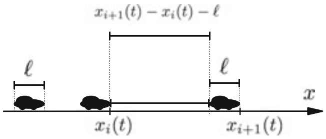


Fig. 1 Microscopic description of the vehicular traffic

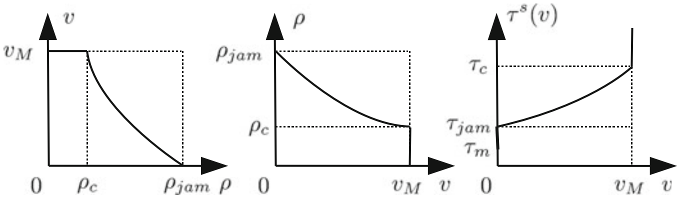


Fig. 2 Fundamental diagram of traffic flow and link with the spatial safety gap

the vehicle's driver i should not be worried about vehicle $(i+1)$ because it is too far from him. In that case, the driver i should accelerate up to the limit speed v_M according to the rule $\dot{v}_i = (v_M - v_i)/\lambda$. This suggests us to choose $a_{i,i+1} = v_M - v_{i+1}$. To summarize, we get the microscopic model

$$\frac{dv_i}{dt} = \frac{v_{i+1} - v_i}{\lambda} + \left(1 - \frac{g^s(v_i, v_{i+1})}{x_{i+1} - x_i - \ell}\right) \frac{v_M - v_{i+1}}{\lambda}. \tag{4}$$

3 Macroscopic quantities and spatial safety gap

From the microscopic quantities, one can define some macroscopic ones. The specific volume $\tau_{i+1/2}(t) := x_{i+1}(t) - x_i(t)$ has the dimension of a length. The density $\rho_{i+1/2}(t) = (\tau_{i+1/2}(t))^{-1}$ returns the local number of vehicles per unit length. The quantity $\rho_M = \ell^{-1}$ represents the maximum density (nose–to–tail vehicles) and $\tau_m = \ell$ is the minimum specific volume. Now in (4), we need a closure for the safety gap function g^s. From g^s one can define a safety specific volume τ^s such that $g^s = \tau^s - \ell = \tau^s - \tau_m$ and a safety density $\rho^s = (\tau^s)^{-1}$. The density ρ^s can be identified to the fundamental diagram of traffic flow which gives a relation between the density and the equilibrium (safe) speed (see Fig. 2). We shall here consider

$$g^s(v_i, v_{i+1}) = \tau^s(\frac{v_i + v_{i+1}}{2}) - \tau_m.$$

4 Macroscopic model

In order to derive a macroscopic model, let us introduce some interpolation functions $v(x,t)$ and $\tau(x,t)$ such that

$$v(x_i(t), t) = v_i(t),\ \tau(x_{i+1/2}(t), t) = x_{i+1}(t) - x_i(t) \quad \forall i = 1, \dots, N.$$

A Taylor expansion allows us to write

$$v_{i+1}(t)-v_i(t) = v(x_{i+1}(t),t)-v(x_i(t),t) = \left(\tau\frac{\partial v}{\partial x}\right)(x_{i+1/2}(t),t)+o((x_{i+1}-x_i)^2).$$

From the motion equation $\dot{x}_i = v_i$, one can write $\frac{d}{dt}(x_{i+1}(t) - x_i(t)) = v_{i+1}(t) - v_i(t)$. Then we have

$$\frac{D\tau}{Dt}(x_{i+1/2}(t),t)) = \frac{\partial v}{\partial x}(x_{i+1/2}(t),t)\,\tau(x_{i+1/2}(t),t) + o((x_{i+1}(t) - x_i(t))^2).$$

We omit the remaining term and consider that the expression holds almost everywhere, then we get the continuity equation

$$\rho\frac{D\tau}{Dt} - \frac{\partial v}{\partial x} = 0 \quad \Leftrightarrow \quad \frac{\partial \rho}{\partial t} + \frac{\partial}{\partial x}(\rho v) = 0. \tag{5}$$

Consider now the acceleration equation. First remark that

$$\left(\frac{\partial v}{\partial x}\tau\right)(x_{i+1/2}(t),t) = \left(\frac{\partial v}{\partial x}\tau\right)(x_i(t),t) + \frac{\tau}{2}\frac{\partial}{\partial x}\left(\tau\frac{\partial v}{\partial x}\right)(x_i(t),t) + o(x_{i+1} - x_i).$$

One can also write $v_{i+1}(t) = v(x_{i+1}(t),t) = v\left(x_i(t) + \tau(x_{i+1/2}(t),t),t\right)$, which allows us to derive the balance equation in Lagrangian form

$$\rho\frac{Dv}{Dt} = \frac{1}{\lambda}\frac{\partial v}{\partial x} + \frac{1}{2\lambda}\frac{\partial}{\partial x}\left(\tau\frac{\partial v}{\partial x}\right) + \left(1 - \frac{g^s(v(x+\tau/2,t))}{\tau - \tau_m}\right)\rho\frac{v_M - v(x+\tau,t)}{\lambda}. \tag{6}$$

i.e. in Eulerian form

$$\frac{\partial}{\partial t}(\rho v) + \frac{\partial}{\partial x}\left(\rho v^2 - \frac{1}{\lambda}v\right) - \frac{1}{2\lambda}\frac{\partial}{\partial x}\left(\tau\frac{\partial v}{\partial x}\right) = \left(1 - \frac{g^s(v(x+\tau/2,t))}{\tau - \tau_m}\right)\rho\frac{v_M - v(x+\tau,t)}{\lambda}. \tag{7}$$

By multiplying formally equation (6) by v and using the continuity equation we get

$$\frac{\partial}{\partial t}(\rho v^2/2) + \frac{\partial}{\partial x}\left(\rho v^3/2 - \frac{1}{\lambda}v^2/2\right) - \frac{1}{2\lambda}\frac{\partial}{\partial x}\left(\tau\frac{\partial(v^2/2)}{\partial x}\right)$$
$$-\left(1 - \frac{g^s(v(x+\tau/2,t))}{\tau - \tau_m}\right)\rho v\frac{v_M - v(x+\tau,t)}{\lambda} = -\frac{1}{2\lambda}\tau\left(\frac{\partial v}{\partial x}\right)^2. \tag{8}$$

This shows that $S = \rho v^2/2$ is an entropy for the system. It is easy to show that S is convex with respect to the conservative variables $(\rho, \rho v)$ (but not strictly convex). More generally, for any $\mathscr{C}^2$ strictly convex function $h : \mathbb{R}^+ \to \mathbb{R}^+$, the function $S = \rho\,h(v)$ is a (non strictly) convex entropy for the system.

Properties Let us consider the first order homogeneous part of the system, i.e.

$$\partial_t \rho + \partial_x(\rho v) = 0, \quad \partial_t(\rho v) + \partial_x(\rho v^2 - \frac{1}{\lambda} v) = 0. \tag{9}$$

In primitive variables (τ, v) we get

$$\partial_t(\tau, v)^T + \begin{pmatrix} v & -\tau \\ 0 & (v - \frac{\tau}{\lambda}) \end{pmatrix} \partial_x(\tau, v)^T = 0.$$

The system is strictly hyperbolic in the admissible space $\Omega_\varepsilon^{ad} = \{(\rho, v),\ \rho \in [\varepsilon, \rho_M],\ v \in [0, v_M]\}$ for any $\varepsilon > 0$. The characteristic speeds are $\lambda_1 = v$ and $\lambda_2 = v - \tau/\lambda$. It easy to check that the two characteristic fields are both linearly degenerate (LD) so that the eigenvalues of the system λ_i, $i = 1, 2$ are the Riemann invariants. One gets a straightforward structure of the solutions of the Riemann problem made of two contact discontinuities.

5 Finite volume scheme

For the sake of simplicity, we shall only deal with the inviscid part of the system above. Let us consider a uniform discretization of the spatial domain (with constant mesh step h) made of discrete points $(x_j)_{j\in\mathbb{Z}}$, $x_{j+1} = x_j + h$ and cells $I_j = (x_{j-1/2}, x_{j+1/2})$, $x_{j+1/2} = (x_j + x_{j+1})/2$. From time t^n, the time advance is performed using a time step Δt^n subject to stability constraints that will be detailed later on. The numerical discretization here follows ideas from Billot et al. [5].

Homogeneous Part Because of the structure of the eigenwaves in (9), a Lagrange–remap conservative FV approach is particularly well suited. Initially the discrete solution is piecewise constant on each control volume I_j with density ρ_j^n, specific volume $\tau_j^n = (\rho_j^n)^{-1}$, and speed v_j^n. In the Lagrange step, the computational grid moves according to the flow; the states into each cell evolve according to the Lagrangian equations. For an initial volume $I_j = (x_{j-1/2}, x_{j+1/2})$, the interface points $x_{j-1/2}$ are moved according to the motion equations $\dot{x}_{j+1/2} = v_{j+1/2}^n$ over a time step Δt^n: this gives $x_j^{n+1,-} = x_j^n + \Delta t^n\, v_{j+1/2}^n$. The choice $v_{j+1/2}^n = v_{j+1}^n$ is compatible with the structure of the solutions of the local Riemann problems, leading to a stable upwind process. After a time step, the cell sizes $h_j^{n+1,-}$ become

$$h_j^{n+1,-} = h + \Delta t^n \left(v_{j+1}^n - v_j^n \right). \tag{10}$$

The continuity equation shows that the number of vehicles m_j into each Lagrangian cell I_j is conserved , i.e. $m_j^n = \rho_j^n\, h = \rho_j^{n+1,-}\, h_j^{n+1,-} = m_j^{n+1,-}$. Combining (10) and mass conservation, we get the equivalent script

$$\tau_j^{n+1,-} = \tau_j^n + \frac{\Delta t^n}{h}\left(v_{j+1}^n - v_j^n\right). \tag{11}$$

The CFL–like condition forbids the 1–waves to interact with the moving interfaces:

$$\frac{\Delta t^n}{h}\sup_{j\in\mathbb{Z}}\left[v_j^n - \min(0, v_j^n - \frac{\tau_j^n}{\lambda})\right] \le 1. \tag{12}$$

By defining

$$v_j^{n+1,-} = \frac{\int_{I_j^{n+1,-}} \rho^{n+1,-}(x)v^{n+1,-}(x)\,dx}{\int_{I_j^{n+1,-}} \rho^{n+1,-}(x)\,dx} = \frac{\int_{I_j^{n+1,-}} \rho^{n+1,-}(x)v^{n+1,-}(x)\,dx}{m_j^n}$$

the speed in the cell $I_j^{n+1,-}$ before projection, we get the following scheme

$$v_j^{n+1,-} = v_j^n + \frac{\Delta t^n}{m_j^n\,\lambda}\left(v_{j+1}^n - v_j^n\right). \tag{13}$$

The Lagrange phase is followed by a conservative projection onto the initial uniform mesh. Denoting by $\alpha_{j+1/2}^n = v_{j+1}^n \Delta t^n/h$ the local Courant number related to the flow speed, the projection of the density in the cell I_j reads

$$\rho_j^{n+1} = \alpha_{j-1/2}^n\,\rho_{j-1}^{n+1,-} + \left(1 - \alpha_{j-1/2}^{n+1,-}\right)\,\rho_j^{n+1,-} \tag{14}$$

as soon as the time step Δt^n satisfies the additional CFL condition $\frac{\Delta t^n}{h}\sup_{j\in\mathbb{Z}} v_j^n \le 1$. Similarly, the projection of the conservative quantity (ρv) writes

$$(\rho v)_j^{n+1} = \alpha_{j-1/2}^n\,(\rho v)_{j-1}^{n+1,-} + \left(1 - \alpha_{j-1/2}^{n+1,-}\right)\,(\rho v)_j^{n+1,-} \tag{15}$$

and gives v_j^{n+1}. It is easy to prove that this numerical scheme fulfills a discrete entropy inequality for the family of entropy functions $S = \rho h(v)$.

Source Term Integration The second equation has a source term that acts as a speed relaxation toward the maximum speed v_m in the case a free flow regime. The differential problem to solve is

$$\frac{dv}{dt} = \left(1 - \frac{g^s(v(x+\tau/2,t))}{\tau - \tau_m}\right)\frac{v_M - v(x+\tau,t)}{\lambda}. \tag{16}$$

When spatially discretized, we have to solve the differential problem

$$\frac{dv_j}{dt} = \left(1 - \frac{g^s(v(x_j+\tau_j/2,t))}{\tau_j - \tau_m}\right)\frac{v_M - v(x_j+\tau_j,t)}{\lambda}, \quad v_j(0) = v_j^0. \tag{17}$$

The problem (17) is nonstandard because of the presence of delays, nonlocal terms (due to the anticipation by the drivers) but also the coupling between the space variable x and the specific volume τ. A computational approach for (17) requires an interpolation of the function v, such as piecewise linear interpolation for example. If, from the discrete point of view, one expects a local influence of the anticipation, we have to assume that h is "large enough" to fulfill the inequality

$$\inf_{j\in\mathbb{Z}} \rho_j^n \geq h^{-1}. \tag{18}$$

The condition (18) may appear surprising, but actually it expresses that the spatial discretization must be compatible with the maximum space headway. As example, consider a road section of length $L = 200$ km and a uniform mesh made of $M = 1000$ points. Then $h = L/M = 0.2$ km and $h^{-1} = 5$ km^{-1}. The discretization of the source term may be local as soon as the vehicle density does not go below 5 veh/km.

Whole Fractional Step Method A consistent second–order accurate time splitting of the full inhomogeneous system may be achieved using the Strang fractional step approach. Each time iteration is made of three substeps: (i) a time integration of the source term over a time step $\Delta t^n/2$; (ii) a time advance of the homogeneous system over a time step Δt^n, (iii) a time integration of the source term over Δt^n as in (i).

6 Numerical experiments

In this experiment we use $v_M = 130$, a section length $L = 200$, a uniform mesh composed of 500 points, $\lambda = 4/3600$ and $\rho_M = 260$. We use periodic boundary conditions. The safety density is chosen as

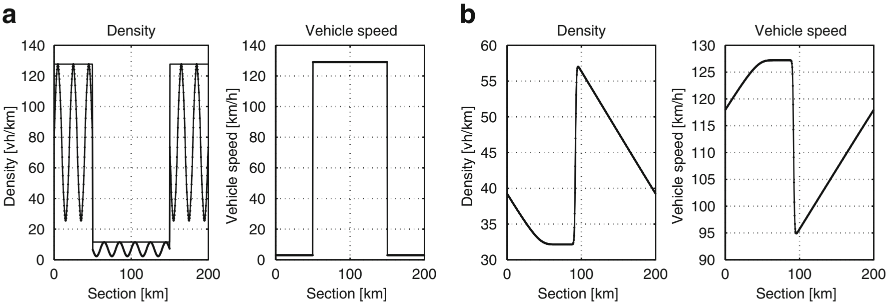


Fig. 3 (a) Initial condition: density (left) and speed (right). (b) Discrete solution at final time: density (left) and speed (right)

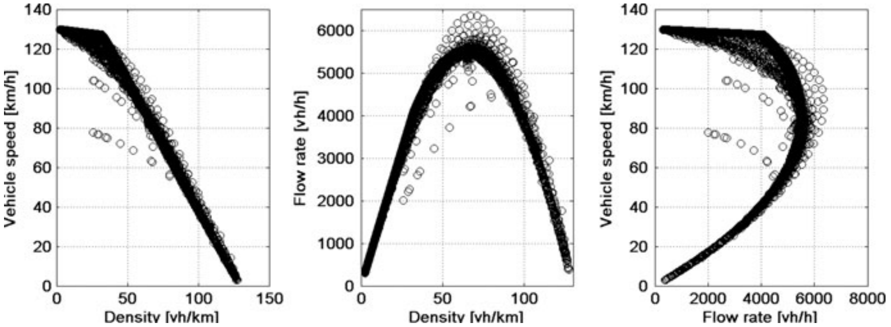


Fig. 4 Discrete solution in the phase space. From left to right: (ρ, v), $(\rho, \rho v)$ and $(\rho v, v)$ diagram

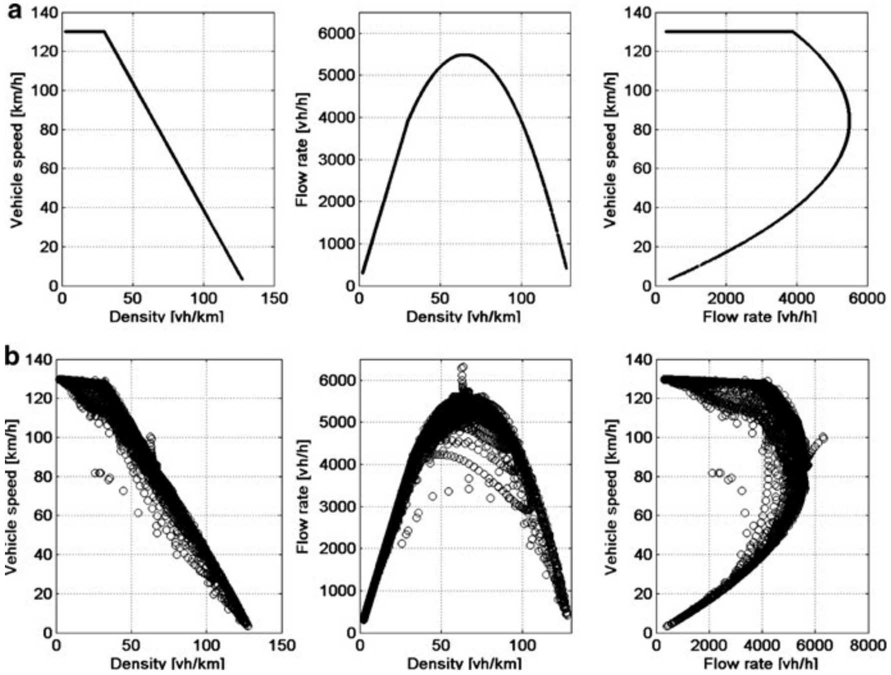


Fig. 5 Numerical Fundamental Diagram computed with: (a) LWR model, (b) Aw–Rascle model

$$\rho^s(v) = \min\left(1500\,\frac{v_M - v}{v_M},\ \rho_{jam} + (\rho_c - \rho_{jam})\frac{v}{v_M}\right)$$

with $\rho_c = 30$ and $\rho_{jam} = 130$. The initial velocity field is a piecewise constant function equal to 3 on $[0, L/4] \cup [3L/4, L]$ and equal to 129 on the interval $(L/4, 3L/4)$. The initial density profile $\rho^0(x) = (0.6 + 0.4\,\sin(20\pi x/L))\ \rho^s(v)$ mimics some nonequilibrium and traffic instabilities (see Fig. 3 (a)). On Fig. 3 (b), the discrete solution at final simulation time $t = 2.22$ is plotted and shows a very good behaviour of the numerical scheme with strong numerical stability, particularly through shock waves. Figure 4 shows the discrete solution for all discrete times in the phase space. One can observe a very good agreement with what is physically expected. The computed numerical discrete fundamental diagram is compared to those obtained with the LWR and Aw–Rascle models, respectively. For the Aw–Rascle model $\partial_t \rho + \partial_x(\rho v) = 0$, $\partial_t v + (v - \rho p'(\rho))\partial_x v = \frac{A}{T}(v^{eq}(\rho) - v)$, we used $v^{eq}(\rho) = \min\Big((v_M\,(1 - \frac{\rho}{1500}), v_M - v_M\frac{\rho - \rho_c}{\rho_{jam} - \rho_c}\Big)$, $p(\rho) = v_M - v^{eq}(\rho)$, $A = 1$, $T = \lambda$.

References

1. A. Aw and M. Rascle, *Resurrection of "second order" models of traffic flow.* SIAM J. Appl. Math., Vol. 60 **(3)**,(2000), 916-938.
2. A. Aw, A. Klar, T. Materne, and M. Rascle. Derivation of continuum traffic flow models from microscopic follow-the-leader models, SIAM J. Applied Math., 63 (**1**), 259–278 (2002).
3. M. Bando, K. Hasebe, A. Nakayama, A. Shibata, Y. Sugiyama, Phys. Rev. E 51, 1035 (1995).
4. C. F. Daganzo, *Requiem for second order fluid approximations of traffic flow*, Transp. Research B, 29, (1995), 277–286.
5. R. Billot, C. Chalons, F. De Vuyst, N. E. El Faouzi, J. Sau, A conditionally linearly stable second-order traffic model derived from a Vlasov kinetic description, Comptes Rendus Mécanique, Volume 338 **(9)** (2010), 529–537.
6. D. Helbing and A. Johansson, *On the controversy around Daganzos requiem for and Aw-Rascle's resurrection of 2nd-order traffic flow models*, Eur. Phys. J. B 69(4), (2009), 549–562.
7. R. Illner, C. Kirchner and R. Pinnau, *A Derivation of the Aw-Rascle traffic models from the Fokker-Planck type kinetic models*, Quart. Appl. Math. 67, (2009), 39–45.
8. B.S. Kerner, Springer, Berlin, New York (2009).
9. E. Tomer, L. Safonov and S. Havlin, Presence of Many Stable Nonhomogeneous States in an Inertial Car-Following Model, Phys. Rev. Lett. 84 **(2)**, 382385 (2000).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Unsteady Numerical Simulation of the Turbulent Flow around an Exhaust Valve

M. Žaloudek, H. Deconinck, and J. Fořt

Abstract The article presents numerical results of the flow which is exhausted from the combustion chamber of a four-stroke engine. The unsteady simulations shown correspond to one working cycle of an exhaust valve.

The flow has been described by the set of Reynolds–averaged Navier–Stokes equations. The working medium has been assumed an ideal gas. The numerical solution has been acquired with an in-house numerical code, *COOLFluiD*, based on a finite volume method (FVM). The numerical code is being developed by the team of engineers with wide range of specialization. Our major contribution has been connected to the implementation of the advanced turbulence models for both steady and unsteady simulations on moving grids.

The current work focuses on the turbulence modelling and on the simulation of the real valve movement. The flow structure and the mass flow rate are observed.

Due to a lack of experimental data, the computations are performed in a stepwise manner, validating each implementation step on the testcases known, before being applied to the valve geometry. The results presented therefore correspond to a planar model. The article focuses on the implementation of turbulence models and their application to complex geometry problems, rather than exploring new numerical methods.

Keywords numerical simulation, exhaust systems, turbulence
MSC2010: 76N15, 76F55, 76H05

Milan Žaloudek and Jaroslav Fořt
Dept. of Technical Mathematics, Czech Technical University, Karlovo nám. 13, CZ-12135
Praha 2, e-mail: Milan.Zaloudek@fs.cvut.cz, Jaroslav.Fort@fs.cvut.cz

Herman Deconinck
von Kármán Institute for Fluid Dynamics, Chaussée de Waterloo 72, B-1640
Rhode-Saint-Genèse, e-mail: deconinck@vki.ac.be

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_83, © Springer-Verlag Berlin Heidelberg 2011

1 RANS Equations

The flow is governed by conservation laws of mass, momentum and energy and two transport equations of the turbulence model.

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}_i^I}{\partial x_i} = \frac{\partial \mathbf{F}_i^V}{\partial x_i} + \mathbf{Q}\,, \tag{1}$$

with t representing time, $\mathbf{x}$ the Cartesian coordinates, $\mathbf{W}$ the vector of conservative unknowns, $\mathbf{F}^I/\mathbf{F}^V$ the convective/viscous fluxes and $\mathbf{Q}$ the source term.

$$\mathbf{W} = |\rho\,,\, \rho w_1\,,\, \rho w_2\,,\, e\,,\, \rho k\,,\, \rho\omega|^T \tag{2}$$

$$\begin{aligned}
\mathbf{F}_i^I &= w_i\, |\rho, \rho w_1 + \tilde{p}\delta_{i1}, \rho w_2 + \tilde{p}\delta_{i2}, e + \tilde{p}, \rho k, \rho\omega|^T \\
\mathbf{F}_i^V &= \left| 0, \tau_{i1}, \tau_{i2}, \tau_{ij} w_j - q_i - q_i^t, (\mu + \sigma_k \mu_t)\frac{\partial k}{\partial x_i}, (\mu + \sigma_\omega \mu_t)\frac{\partial \omega}{\partial x_i} \right|^T \\
\mathbf{Q} &= \left| 0, 0, 0, 0, P - \beta^* \rho k \omega, \frac{\gamma}{\nu_t} P - \beta \rho \omega^2 + (1 - F_1)\, \rho \frac{2\sigma_2}{\omega} \frac{\partial k}{\partial x_j} \frac{\partial \omega}{\partial x_j} \right|^T
\end{aligned}$$

The unknows ρ, $\mathbf{w} = (w_1; w_2)$, e, p, T, k, ω denote in turns the density, the velocity components, the total energy, the pressure, the temperature, the turbulent kinetic energy and the specific dissipation rate. The stress tensor τ_{ij} is expressed as

$$\tau_{ij} = (\mu + \mu_t)\, S_{ij}\ ,\ S_{ij} = \frac{1}{2}\left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_j}{\partial x_i}\right) - \frac{2}{3} \cdot \delta_{ij} \cdot \frac{\partial w_k}{\partial x_k}\,, \tag{3}$$

with δ_{ij} the Kronecker delta and μ, μ_t the molecular and turbulent dynamic viscosity

$$\mu = \frac{C_1 T^{3/2}}{T + S}\ ,\ \mu_t = \gamma^* \rho \frac{k}{\omega}\,. \tag{4}$$

The heat flux and the production term read

$$q_i = -\frac{\lambda}{\mathrm{Pr}} \frac{\partial T}{\partial x_i}\ ,\ q_i^t = q_i \frac{\mathrm{Pr}}{\mu} \frac{\mu_t}{\mathrm{Pr}_t}\ ,\ P = \mu_t S_{ij} S_{ij}\,. \tag{5}$$

Unknowns $\sigma_k, \sigma_\omega, \beta, \beta^*, \gamma, \gamma^*, \sigma_2, C_1, S, \lambda$ represent various constants to be found in the literature [7] and Pr stands for the Prandtl number. The function F_1 provides a blending between the $k-\epsilon$ model in freestream regions and the $k-\omega$ model near the wall surfaces. The system is completed with the state equation. The next turbulence models presented, have used a similar formulation as (1) and their specifics have been published in [10] (EARSM model) and [11] (Wilcox $k-\omega$, rev. 2008).

ALE Formulation. For unsteady simulations with a moving valve the arbitrary Lagrangian–Eulerian formulation of the RANS equations has been used, see [9]. The relative velocity $\mathbf{w}_R$ is defined as

$$\mathbf{w}_R = \mathbf{w} - \mathbf{w}_V \,, \tag{6}$$

with $\mathbf{w}$ the flow velocity and $\mathbf{w}_V$ the velocity of the valve (given by the movement imposed, see fig. 6). The convective flux $\mathbf{F}^I$ is then updated to a form

$$\mathbf{F}_i^{I,ALE} = w_{iR} \left| \rho, \rho w_1 + \tilde{p}\delta_{i1}, \rho w_2 + \tilde{p}\delta_{i2}, e + 2\tilde{p}, \rho k, \rho\omega \right|^T \tag{7}$$

2 Mathematical Formulation

The system (1) is solved upon the computational domain, see the Fig. 1. Although the real configuration is fully 3D, the computational domain has been considered symmetric with respect to the valve axis. Hence, only a half of the domain has been solved. A mathematic solution fulfils the equation (1) upon the domain interior, the *initial condition* at $t = 0$ and the following *boundary conditions* on the domain borders:

inlet total pressure, total temperature, incidence angle, turbulent variables according to the paper [8]:

$$\omega^{in} = \frac{\left|\mathbf{w}^{in}\right|}{L_{ref}}, \quad k^{in} = \omega^{in} \cdot \frac{\mu_\infty}{100} \,. \tag{8}$$

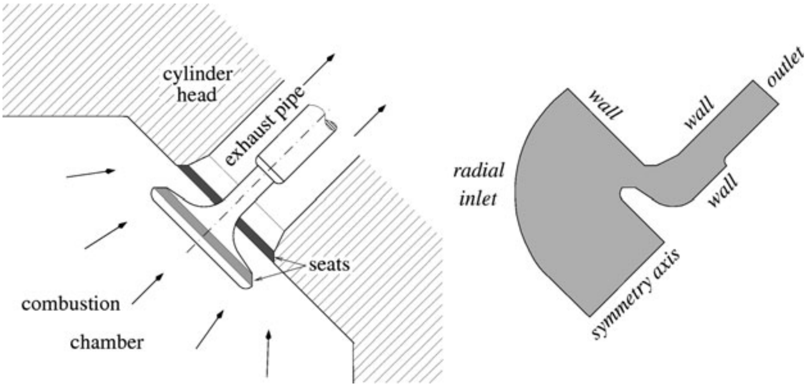


Fig. 1 Detail of the exhaust valve (left), scheme of the computational domain (right)

outlet pressure, velocity, temperature and turbulent variables

$$p = p^{out}\ ,\ \frac{\partial w_i}{\partial n} = \frac{\partial T}{\partial n} = \frac{\partial k}{\partial n} = \frac{\partial \omega}{\partial n} = 0 \tag{9}$$

wall the adiabatic no-slip condition. The turbulent variables use the expressions suggested at [8]

$$\mathbf{w} \equiv 0,\ \frac{\partial T}{\partial n} = 0,\ k^w = 0,\ \omega^w = \frac{60\nu}{\beta_1 y_0^2} \tag{10}$$

3 Discretization and Numerical Method

The computational domain has been discretized by a structured triangular grid, see the Fig. 2. The steady flow computations were achieved with the time marching method based on a finite volume method, discretizing the equations (1) as

$$\frac{W_i^{n+1} - W_i^n}{\Delta t} = \frac{1}{\mu_i} \sum_{k=1}^{\#faces} \left(-\tilde{F}_k^I \cdot \mathbf{n}_k + \tilde{F}_k^V \cdot \mathbf{n}_k\right)\ , \tag{11}$$

with μ_i the area of the i-th volume, $\tilde{F}_k^I/\tilde{F}_k^V$ the numerical approximation of the advection/viscous fluxes and $\mathbf{n}_k$ the unit outward normal vector to the k-th face of the volume i. The fully implicit time integration has been used

$$W_i^{n+1} = W_i^n + \frac{\Delta t}{\mu_i} \sum_{k=1}^{\#faces} \left(-\tilde{F}^I \left(W^n, W^{n+1}\right)_k \cdot \mathbf{n}_k + \tilde{F}^V \left(W^n, W^{n+1}\right)_k \cdot \mathbf{n}_k\right)\ , \tag{12}$$

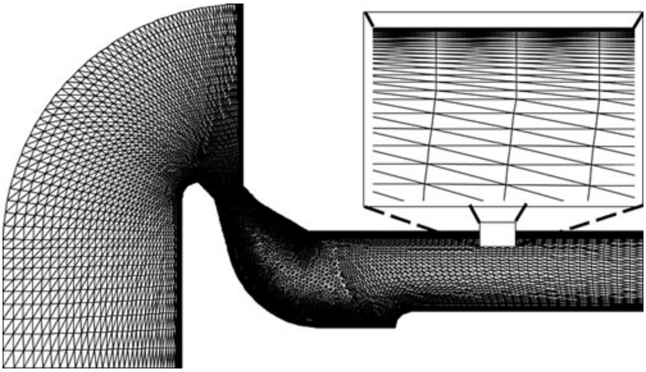


Fig. 2 Overview of the computational grid with the detail of its structure

as it is described in [5]. The linear system has been solved numerically by the GMRES iterative solver, provided by the PETSc library. As the flowfield contains both regions with the gas of negligible velocity (inside the chamber, Mach number $\approx$ 0.05) and regions with the gas of supersonic velocity (between the seats, $M \approx$ 2.0) the numerical scheme $AUSM^{+up}$ able to capture all the velocity scales has been used. The algorithm is based on a solution of the Riemann problem (flux over a discontinuous step between two states) and thanks to the pressure and Mach number correction terms it improves the convergence also for the low velocity regions. The scheme has been published in [6]. The viscous fluxes have been computed as a central approximation, using a diamond dual cell approach.

The spatial accuracy has been improved by a piecewise linear reconstruction that has been built by the least squares interpolation method, complemented with the Barth limiter [2].

Unsteady Flow. The computational domain (and grid) changes with the advancing time. The solution is therefore based on the ALE formulation for moving grids. In order to avoid the situation of two disjunct subdomains with no flow between them for the closed valve a minimal valve opening (treshold) has always been used. Later, due to the significant grid deformation the domain has been remeshed and the current solutions interpolated in a conservative way. The series of three meshes (initial valve lift: 0.5 mm, 2.5 mm, 7.0 mm) have been used to resolve one working cycle of the exhaust valve.

The steady computations algorithm is modified to ensure the accuracy and consistency also for the unsteady flow. The time accurate solution has been obtained with the dual time stepping technique, consisting of an *outer time stepping loop* for a real time-accurate time step Δt and an *inner time stepping loop* with a fictitious time step τ to solve the system at each real time step. For the initiation phase the single step Crank–Nicholson method has been used, followed by the backward differentiation formula BDF2

$$\frac{W^{n+1,\alpha+1} - W^{n+1,\alpha}}{\tau} + \frac{3W^{n+1,\alpha+1} - 4W^{n} + W^{n-1}}{2\Delta t} = \tag{13}$$

$$\frac{1}{\mu_i} \sum_{k=1}^{\#faces} \left[\tilde{F}^{V}\left(W^{n}, W^{n+1,\alpha},, W^{n+1,\alpha+1}\right)_k \cdot \mathbf{n}_k - \tilde{F}^{I}\left(W^{n}, W^{n+1,\alpha},, W^{n+1,\alpha+1}\right)_k \cdot \mathbf{n}_k \right]$$

4 Steady Flow Numerical Results

The Fig. 3 reveals the steady solutions with different turbulence models. The computations have been stated by the parameters: valve opening 4 mm, temperature 500 K, pressure ratio $\frac{p_{inlet}}{p_{outlet}} = 2.5$, with the outlet pressure 100 kPa, corresponding to the exhaust to the atmosphere. The flow topology is similar for all models, consisting of a main beam (in approximately same position), surrounded by separation zones on both sides. The differences are visible on the pressure distribution along a streamline that passes the middle of a channel throat, see the Fig. 4. The BSL and

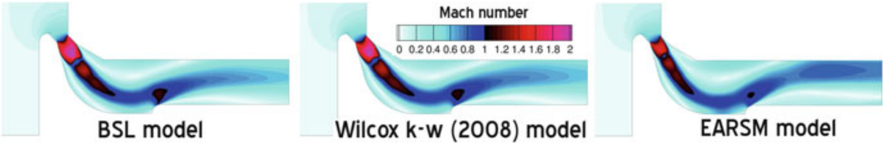


Fig. 3 Contours of Mach number for various turbulence models

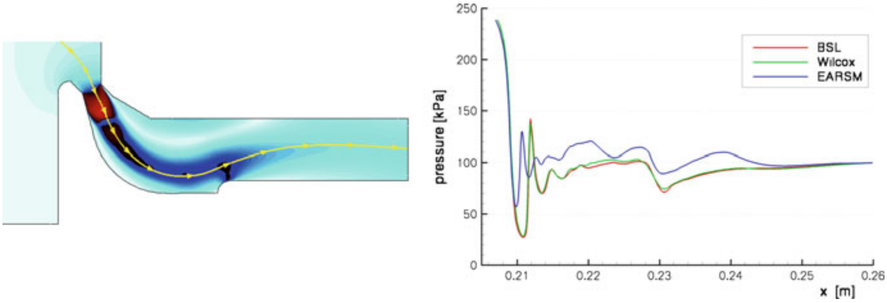
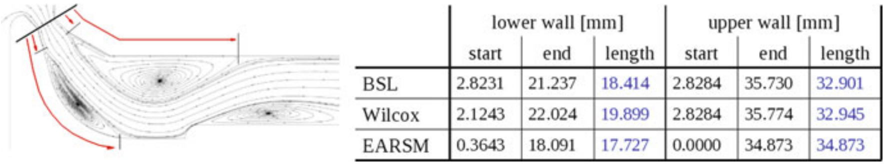


Fig. 4 The streamline for extracting the flow characteristics (left), comparison of the pressure through the exhaust pipe (right)



	lower wall [mm]			upper wall [mm]		
	start	end	length	start	end	length
BSL	2.8231	21.237	18.414	2.8284	35.730	32.901
Wilcox	2.1243	22.024	19.899	2.8284	35.774	32.945
EARSM	0.3643	18.091	17.727	0.0000	34.873	34.873

Fig. 5 Position of separation zones meassured from the channel throat

Wilcox models have a similar nature, which justifies similar results achieved by these models. By the contrary, the EARSM model allows anisotropic turbulence, see [10], leading to the milder peaks predicted and higher outlet velocity. However, the qualitative agreement is observed across all the models. Similar behaviour can be seen also on the comparison of the separation zone positions in the Table 5.

The next expansion is allowed due to the separations which form an artificial nozzle-like channel inside the exhaust pipe. These separations are described in the Fig. 5.

5 Unsteady Flow Numerical Results

The movement of the exhaust valve is shown in the Fig. 6a (valve lift vs. time). The next graph, Fig. 6b, shows the time evolution of the inlet pressure for a spark-ignition (SI), a compression-ignition (CI) engine and the outlet pressure. Values are taken from [4] and represent the boundary conditions for the unsteady simulations.

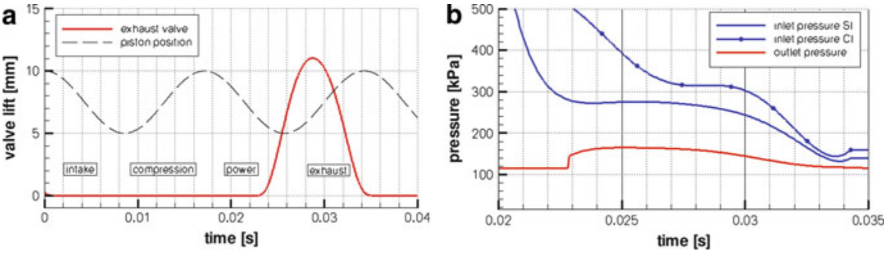


Fig. 6 The movement of the exhaust valve (left), the operating conditions at the inlet and outlet (right)

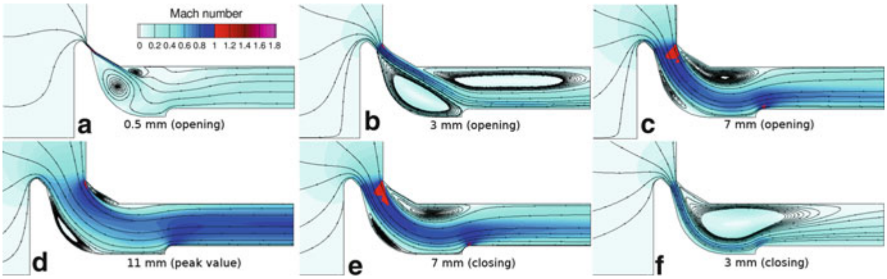


Fig. 7 SI engine. Contours of Mach number, velocity streamlines

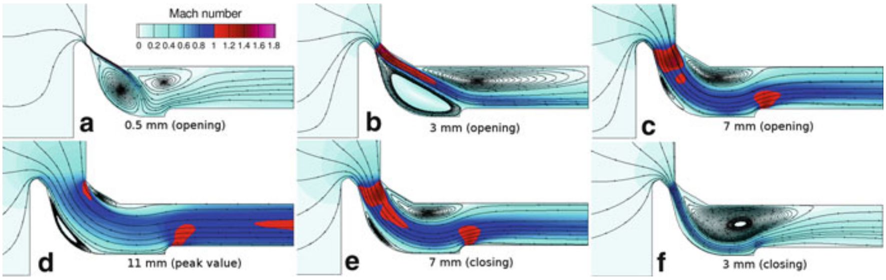


Fig. 8 CI engine. Contours of Mach number, velocity streamlines

The computations start at $t = 0.022$ (see Fig. 6) and the exhaust valve cycle lasts approximately 0.015 seconds. This interval has been resolved with the timestep $\Delta t = 10^{-6}\,s$ and with a valve lift treshold 0.5 mm. The results of the unsteady computations correspond to the valve lifts: 0.5 → 3 → 7 → 11 → 7 → 3 mm for both SI and CI inlet pressure evolutions.

The lift 7 mm has also been supplied by a pair of steady computations at boundary condition of the CI engine for the given lift, see the Fig. 9. The last Fig. 10 shows the mass flow rate over the valve cycle for the SI and CI engines, the steady solutions are mapped by two points. The last graph compares the pressure along the streamline (see Fig. 4) for the unsteady (Fig. 8e) and steady (Fig. 9b) computations at the same valve lift 7 mm.

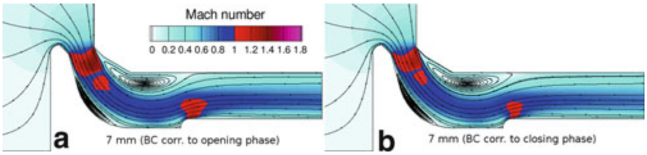

Mach number
0 0.2 0.4 0.6 0.8 1 1.2 1.4 1.6 1.8
a
7 mm (BC corr. to opening phase)
b
7 mm (BC corr. to closing phase)

Fig. 9 Steady results. Boundary conditions correspond to CI engine at valve lift 7 mm in the opening (left) and the closing phase. Contours of Mach number, velocity streamlines

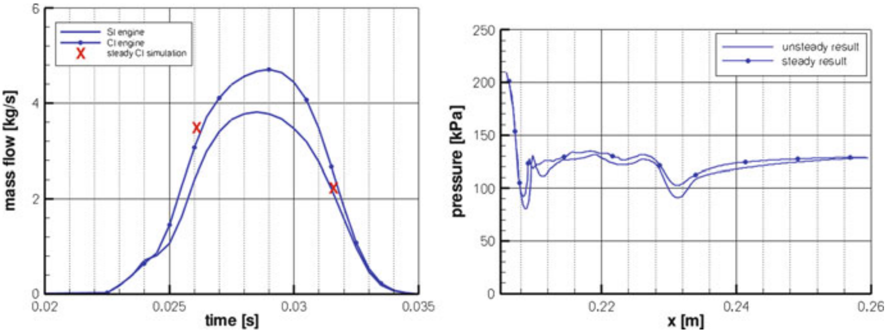


Fig. 10 Comparison of the SI and CI engine model: mass flow rate (left). Comparison of the unsteady and steady model: pressure development in the exhaust pipe (right)

6 Conclusions

The steady results have shown similar behaviour for all the turbulence models tested.

The flowfield of unsteady results are in qualitative agreement with equivalent steady solutions, however, the mass flow rate can differ up to approximately 10% (see Fig. 10). Also the pressure development along the mean streamline behind the channel throat differs from the steady state.

In case of the CI engine (due to higher inlet pressure) one observes the aerodynamical choking and larger supersonic regions, compared to the SI engine. The SI model is choking-free in the dominant time of the valve cycle. The negligible mass flow in the early and late stages of the valve cycle also justifies the use of grids with minimal (non-zero) valve opening. The oncoming work will be aimed at more advanced turbulence models for the unsteady simulations, flow characteristics at different rpm and mainly on 3D unsteady simulations.

Acknowledgements This work has been supported by the grant of the Czech Science Foundation No. P101/10/1329 and by the project 1M6840770002 Josef Božek Research Center of the Ministry of Education of the Czech Republic.

References

1. COOLFluiD homepage [on-line], http://coolfluidsrv.vki.ac.be, Cited 17 Feb 2011
2. Barth, T. J., Jesperson, D. C.: The design and application of upwind schemes on unstructured meshes. AIAA Paper **89(0366)** (1989)
3. Favre, A.: Equations des gaz turbulents compressibles. J. de Mecanique **4**, 361–421 (1965)
4. Heywood, J. B.: Internal Combustion Engine Fundamentals. McGraw-Hill, Inc. USA (1988)
5. Lani, A.: An Object Oriented and High Performance Platform for Aerothermodynamics Simulation. Doctoral thesis, VKI, Belgium (2009)
6. Liou M. S.: A sequel to AUSM, Part II: AUSM$^+$up for all speeds, J. of Computational Physics **214**, 137–170 (2006)
7. Menter, F. R., Rumsey, C. L.: Assessment of Two-Equation Turbulence Models for Transonic Flows. AIAA 25th Fluid Dynamics Conference, Colorado Springs, USA (1994)
8. Menter, F. R.: Two-Equation Eddy-Viscosity Turbulence Models for Engineering Applications, AIAA Journal **32-8** (1994)
9. Michler, C.: Development of an ALE Formulation for Unsteady Flow Computations on Moving Meshes using RD Schemes, Project Report **2000-13**, VKI, Belgium (2000)
10. S. Wallin: Engineering turbulence modelling for CFD with focus on explicit algebraic Reynolds stress models. Dissertation Thesis, Norsteds Tryckeri AB, Sweden (2000)
11. D. C. Wilcox, Formulation of the $k - \omega$ Turbulence Model Revisited, AIAA J. **46-11** (2008)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Part II
Invited Papers

Lowest order methods for diffusive problems on general meshes: A unified approach to definition and implementation

Daniele A. Di Pietro and Jean-Marc Gratien

Abstract In this work we propose an original point of view on lowest order methods for diffusive problems which lays the pillars of a C++ multi-physics, FreeFEM-like platform. The key idea is to regard lowest order methods as (Petrov)-Galerkin methods based on possibly incomplete, broken polynomial spaces defined from a gradient reconstruction. After presenting some examples of methods entering the framework, we show how implementation strategies common in the finite element context can be extended relying on the above definition. Several examples are provided throughout the presentation, and programming details are often omitted to help the reader unfamiliar with advanced C++ programming techniques.

Keywords Lowest-order methods, Domain specific embedded language, Petrov-Galerkin methods, cell centered Galerkin methods, hybrid finite volume methods
MSC2010: 65Y99, 65N08, 65N30

1 Introduction

An increasing amount of attention has recently been given to the discretization of diffusive problems on general meshes. Lowest order methods possibly featuring conservation of physical quantities are traditionally employed in industrial applications where computational cost is a crucial issue. In this context, the main interest of handling general meshes is to reduce the number of elements required to represent complicate domains. In sedimentary basin modeling, non-standard elements may also appear due to the erosion of geological layers. Different ways to adapt finite volume and finite element methods to general, possibly non-conforming

Daniele A. Di Pietro and Jean-Marc Gratien
IFP Energies nouvelles, e-mail: dipietrd@ifpenergiesnouvelles.fr,
j-marc.gratien@ifpenergiesnouvelles.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_84, © Springer-Verlag Berlin Heidelberg 2011

polyhedral meshes have been proposed. In the context of cell centered finite volume methods, we recall, in particular, the classical works of Aavatsmark, Barkve, Bøe and Mannseth [1] and Edwards and Rogers [15] on multipoint fluxes. More recently, two ways of extending the mixed finite element philosophy to general meshes have been proposed independently by Brezzi, Lipnikov, Shashkov and Simoncini [5, 6] (mimetic finite difference methods) and by Droniou and Eymard [13] (mixed/hybrid finite volume methods). Yet another perspective is considered by Eymard, Gallouët and Herbin [17], who show, in particular, that face unknowns can be selectively used as Lagrange multipliers for the flux continuity constraint or be eliminated using a consistent interpolator (SUSHI scheme). The strong link between the strategies above has been highlighted by Droniou, Eymard, Gallouët and Herbin [14]. A slightly different approach based on the analogy between lowest order methods in variational formulation and discontinuous Galerkin methods has been proposed by the author in [8–10] (cell centered Galerkin methods). The key advantage of this approach is that it largely benefits from the well-established theory for discontinuous Galerkin methods applied to diffusive problems [11]. All of the methods above have been (or can be) extended to several classical problems for which the discretization of second order diffusive terms is central.

In this work we present a unified implementation covering a wide range of lowest order methods and applications based on similar experiences in the context of finite element methods. Finite element libraries have nowadays reached a good level of maturity, and user-friendly front-ends are provided in several cases. Just to mention a few, we recall `Feel++` [20] (formerly known as `Life`), `FEniCS` [19], `FreeFEM++` [7]. Our goal is to show that similar tools can be conceived and implemented for lowest order methods. The starting point is to reformulate the method at hand as a (Petrov)-Galerkin scheme based on possibly incomplete broken affine spaces. This new unified perspective, drawing on the lines of [9], allows, in particular, to recycle many ideas originally developed for finite elements. A major difference, however, is that the lowest order methods considered herein are often based on reconstructions of first order differential operators which may depend on problem data such as the diffusion coefficient or the boundary condition. As a consequence, the classical approach based on a table of degrees of freedom computed from a mesh and a finite element (see, e.g., [16, Chapters 7–8]) is no longer adequate. This issue is solved by introducing the programming counterpart of tensor-valued linear combinations of (globally numbered) degrees of freedom. This concept allows, in particular, to reproduce a finite element-like matrix assembly with local contributions stemming from integrals over mesh elements and faces. A further layer of abstraction is added by defining a domain-specific language (DSL) for variational formulations. The DSL is closely inspired by that of `Feel++`, the most noticeable differences being the type-based identification of test and trial functions and the possibility to store the expressions defining linear and bilinear forms independently of their algebraic representation. Another novelty is the introduction of tensor-like notation for systems of PDEs. Domain-specific languages and generative programming are an established tool to break down the complexity of industrial applications by distinguishing the actors that tackle different aspects of the problem, and providing each of them with means of expression as close as possible to his/her technical jargon. An important advantage of the DSL is that it potentially allows to combine lowest order methods with more standard discretizations techniques in a seamless way. In the presentation we try to avoid all technicalities and to pinpoint the main difficulties as well as the proposed solutions. Although the language of choice is `C++`, the listings are rather to be intended as pseudo-code since simplifications are often made to improve readability. The actual implementation is based on the `Arcane` framework [18], a proprietary platform conjointly developed at CEA-DAM and IFP Energies nouvelles which takes care of technical aspects such as memory management, parallelism and post-processing.

The material is organized as follows. In §2 we propose a unified perspective and show how several lowest order methods can fit in there for a simple diffusion problem. In §3 we discuss the implementation. More specifically, we first discuss the solutions to the issues that arise when trying to mimic the finite element approach and then present a DSL which allows to conceal the related technicalities.

2 Definition

2.1 Discrete setting

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, denote a bounded connected polyhedral domain. The first ingredient in the definition of lowest order methods is a suitable discretization of Ω. We denote by $\mathscr{T}_h$ a finite collection of nonempty, disjoint open polyhedra $\mathscr{T}_h = \{T\}$ forming a partition of Ω such that $h = \max_{T\in\mathscr{T}_h} h_T$ and h_T denotes the diameter of the element $T \in \mathscr{T}_h$. Admissible meshes include general polyhedral discretizations with possibly nonconforming interfaces; see Fig. 1. Mesh nodes are collected in the set $\mathscr{N}_h$ and, for all $T \in \mathscr{T}_h$, $\mathscr{N}_T$ contains the nodes that lie on the boundary of T. We say that a hyperplanar closed subset F of $\overline{\Omega}$ is a mesh face if it has positive $(d-1)$-dimensional measure and if either there exist $T_1, T_2 \in \mathscr{T}_h$ such that $F \subset \partial T_1 \cap \partial T_2$ (and F is called an *interface*) or there exists $T \in \mathscr{T}_h$ such that $F \subset \partial T \cap \partial\Omega$ (and F is called a *boundary face*). Interfaces are collected in the set $\mathscr{F}_h^{\mathrm{i}}$, boundary faces in $\mathscr{F}_h^{\mathrm{b}}$ and we let $\mathscr{F}_h := \mathscr{F}_h^{\mathrm{i}} \cup \mathscr{F}_h^{\mathrm{b}}$. For all $T \in \mathscr{T}_h$ we set

$$\mathscr{F}_T := \{F \in \mathscr{F}_h \mid F \subset \partial T\}. \tag{1}$$

Symmetrically, for all $F \in \mathscr{F}_h$, we define

$$\mathscr{T}_F := \{T \in \mathscr{T}_h \mid F \subset \partial T\}.$$

The set $\mathscr{T}_F$ consists of exactly two mesh elements if $F \in \mathscr{F}_h^{\mathrm{i}}$ and of one if $F \in \mathscr{F}_h^{\mathrm{b}}$. For all mesh nodes $P \in \mathscr{N}_h$, $\mathscr{F}_P$ denotes the set of mesh faces sharing P, i.e.

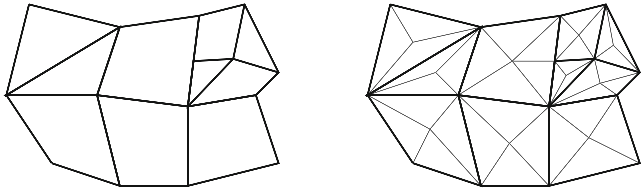


Fig. 1 *Left.* Mesh $\mathcal{T}_h$ *Right.* Pyramidal submesh $\mathcal{P}_h$

$$\mathcal{F}_P := \{F \in \mathcal{F}_h \mid P \in F\}. \tag{2}$$

The diameter of a face $F \in \mathcal{F}_h$ is denoted by h_F. For every interface $F \in \mathcal{F}_h^{\mathrm{i}}$ we introduce an arbitrary but fixed ordering of the elements in $\mathcal{T}_F$ and let $\mathbf{n}_F = \mathbf{n}_{T_1,F} = -\mathbf{n}_{T_2,F}$, where $\mathbf{n}_{T_i,F}$, $i \in \{1,2\}$, denotes the unit normal to F pointing out of $T_i \in \mathcal{T}_F$. On a boundary face $F \in \mathcal{F}_h^{\mathrm{b}}$, $\mathbf{n}_F$ denotes the unit normal pointing out of Ω. The barycenter of a face $F \in \mathcal{F}_h$ is denoted by $\overline{\mathbf{x}}_F := \int_F \mathbf{x}/|F|_{d-1}$. For each $T \in \mathcal{T}_h$ we identify a point $\mathbf{x}_T \in T$ (the *cell center*) such that T is star-shaped with respect to $\mathbf{x}_T$. For all $F \in \mathcal{F}_T$ we let

$$d_{T,F} := dist(\mathbf{x}_T, F).$$

It is assumed that, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, $d_{T,F} > 0$ is comparable to h_T. Starting from cell centers we can define a pyramidal submesh of $\mathcal{T}_h$ as follows:

$$\mathcal{P}_h := \{\mathcal{P}_{T,F}\}_{T \in \mathcal{T}_h,\, F \in \mathcal{F}_T},$$

where, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, $\mathcal{P}_{T,F}$ denotes the open pyramid of apex $\mathbf{x}_T$ and base F, i.e.,

$$\mathcal{P}_{T,F} := \{\mathbf{x} \in T \mid \exists \mathbf{y} \in F \setminus \partial F,\ \exists \theta \in (0,1) \mid \mathbf{x} = \theta \mathbf{y} + (1-\theta)\mathbf{x}_T\}.$$

The pyramids $\{\mathcal{P}_{T,F}\}_{T \in \mathcal{T}_h,\, F \in \mathcal{F}_T}$ are nondegenerate by assumption. Let $\mathcal{S}_h$ be such that

$$\mathcal{S}_h = \mathcal{T}_h \text{ or } \mathcal{S}_h = \mathcal{P}_h. \tag{3}$$

For all $k \geq 0$, we define the broken polynomial spaces of total degree $\leq k$ on $\mathcal{S}_h$,

$$\mathbb{P}_d^k(\mathcal{S}_h) := \{v \in L^2(\Omega) \mid \forall S \in \mathcal{S}_h,\ v_{|S} \in \mathbb{P}_d^k(S)\},$$

with $\mathbb{P}_d^k(S)$ given by the restriction to $S \in \mathcal{S}_h$ of the functions in $\mathbb{P}_d^k$.

Remark 1 (Admissible mesh sequence). In the context of *a priori* convergence analysis for vanishing mesh size h it is necessary to bound some quantities uniformly with respect to h. This leads to the concept of *admissible mesh sequence*. This topic is not addressed in detail herein since our focus is mainly on implementation. For a comprehensive discussion we refer to [5, 6, 9, 13, 17]; see also [11, Chapter 1].

We close this section by introducing trace operators which are of common use in the context of nonconforming finite element methods. Let v be a scalar-valued function defined on Ω smooth enough to admit on all $F \in \mathscr{F}_h$ a possibly two-valued trace. To any interface $F \subset \partial T_1 \cap \partial T_2$ we assign two nonnegtive real numbers $\omega_{T_1,F}$ and $\omega_{T_2,F}$ such that

$$\omega_{T_1,F} + \omega_{T_2,F} = 1,$$

and define the jump and weighted average of v at F for a.e. $\mathbf{x} \in F$ as

$$[\![v]\!]_F(\mathbf{x}) := v_{|T_1} - v_{|T_2}, \qquad \{\!\{ v \}\!\}_{\omega,F}(\mathbf{x}) := \omega_{T_1,F} v_{|T_1}(\mathbf{x}) + \omega_{T_2,F} v_{|T_2}(\mathbf{x}). \tag{4}$$

If $F \in \mathscr{F}_h^{\mathrm{b}}$ with $F = \partial T \cap \partial\Omega$, we conventionally set $\{\!\{ v \}\!\}_{\omega,F}(\mathbf{x}) = [\![v]\!]_F(\mathbf{x}) = v_{|T}(\mathbf{x})$. The subscript ω is omitted from the average operator when $\omega_{T_1,F} = \omega_{T_2,F} = \frac{1}{2}$. The dependence on $\mathbf{x}$ and on the face F is also omitted if no ambiguity arises.

2.2 An abstract perspective

The key idea to gain a unifying perspective is to regard lowest order methods as nonconforming methods based on incomplete broken affine spaces that are defined starting from the space of degrees of freedom (DOFs) $\mathbb{V}_h$. More precisely, we let

$$\mathbb{T}_h := \mathbb{R}^{\mathscr{T}_h}, \qquad \mathbb{F}_h := \mathbb{R}^{\mathscr{F}_h},$$

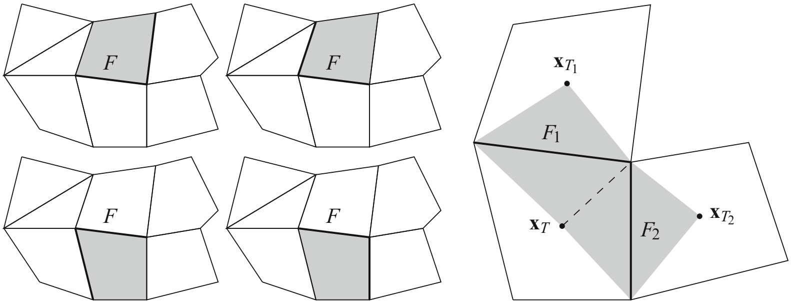
and consider the following choices:

$$\mathbb{V}_h = \mathbb{T}_h \text{ or } \mathbb{V}_h = \mathbb{T}_h \times \mathbb{F}_h.$$

In every case the elements of $\mathbb{V}_h$ are indexed with respect to the mesh entity they belong to. Other choices for $\mathbb{V}_h$ are possible but are not considered herein for the sake of conciseness. To fix the ideas, one can assume that the choice $\mathbb{V}_h = \mathbb{T}_h$ corresponds to cell centered finite volume (CCFV) and cell centered Galerkin (CCG) methods, while the choice $\mathbb{V}_h = \mathbb{T}_h \times \mathbb{F}_h$ leads to mimetic finite difference (MFD) and mixed/hybrid finite volume (MHFV) methods.

The key ingredient in the definition of the broken affine space is a piecewise constant linear gradient reconstruction $\mathfrak{G}_h : \mathbb{V}_h \to [\mathbb{P}_d^0(\mathscr{S}_h)]^d$ (the linearity of $\mathfrak{G}_h$ is a founding assumption for the implementation discussed in §3). Starting from $\mathfrak{G}_h$, we can define the linear operator $\mathfrak{R}_h : \mathbb{V}_h \to \mathbb{P}_d^1(\mathscr{S}_h)$ such that, for all $\mathbf{v}_h \in \mathbb{V}_h$,

$$\forall S \in \mathscr{S}_h,\ S \subset T_S \in \mathscr{T}_h,\ \forall \mathbf{x} \in S, \quad \mathfrak{R}_h(\mathbf{v}_h)_{|S} = v_{T_S} + \mathfrak{G}_h(\mathbf{v}_h)_{|S} \cdot (\mathbf{x} - \mathbf{x}_{T_S}) \in \mathbb{P}_d^1(\mathscr{S}_h). \tag{5}$$



(a) $\mathscr{G}_F =$ L-groups containing the face F

(b) L-construction

Fig. 2 L-construction

The operator $\mathfrak{R}_h$ maps every vector of DOFs onto a piecewise affine function belonging to $\mathbb{P}_d^1(\mathscr{S}_h)$. Hence, we can define a broken affine space as follows:

$$V_h = \mathfrak{R}_h(\mathbb{V}_h) \subset \mathbb{P}_d^1(\mathscr{S}_h). \tag{6}$$

The operator $\mathfrak{R}_h$ is additionally assumed to be injective, so that a bijective operator can be obtained by restricting its codomain. The next section presents some examples covering the methods listed above.

2.3 Examples

In this section we focus on the model problem

$$-\nabla\cdot(\boldsymbol{\kappa}\nabla u) = f, \qquad u = 0, \tag{7}$$

where $f \in L^2(\Omega)$ and $\boldsymbol{\kappa} \in [\mathbb{P}_d^0(\mathscr{T}_h)]^d$ is a piecewise constant, uniformly elliptic tensor field (possibly resulting from a homogeneization process). Problem (7) provides the paradigm to illustrate how selected lowest order methods can be recast in the framework of §2.2.

The G-method As a first example we consider the special instance of CCFV methods analyzed in [3]. A preliminary step consists in presenting the so-called L-construction introduced in [2]. The key idea of the L-construction is to use d cell and boundary face values (provided, in this case, by the homogeneous boundary condition) to express a continuous piecewise affine function with continuous diffusive fluxes. The values are selected using d neighboring faces belonging to a cell and sharing a common vertex. More precisely, we define the set of L-groups (see Fig. 2) as follows:

$$\mathscr{G} := \{\mathfrak{g} \subset \mathscr{F}_T \cap \mathscr{F}_P,\ T \in \mathscr{T}_h,\ P \in \mathscr{N}_T \mid card(\mathfrak{g}) = d\},$$

with $\mathscr{F}_T$ and $\mathscr{F}_P$ given by (1) and (2) respectively. It is useful to introduce a symbol for the set of cells concurring in the L-construction: For all $\mathfrak{g} \in \mathscr{G}$, we let

$$\mathscr{T}_{\mathfrak{g}} := \{T \in \mathscr{T}_h \mid T \in \mathscr{T}_F,\ F \in \mathfrak{g}\}.$$

Let now $\mathfrak{g} \in \mathscr{G}$ and denote by $T_{\mathfrak{g}}$ an element $T_{\mathfrak{g}}$ such that $\mathfrak{g} \subset \mathscr{F}_{T_{\mathfrak{g}}}$ (this element may not be unique). For all $\mathbf{v}_h \in \mathbb{V}_h$ we construct the function $\xi^{\mathfrak{g}}_{\mathbf{v}_h}$ piecewise affine on the family of pyramids $\{\mathscr{P}_{T,F}\}_{F\in\mathfrak{g},\, T\in\mathscr{T}_{\mathfrak{g}}}$ such that: (i) $\xi^{\mathfrak{g}}_{\mathbf{v}_h}(\mathbf{x}_T) = v_T$ for all $T \in \mathscr{T}_{\mathfrak{g}}$ and $\xi^{\mathfrak{g}}_{\mathbf{v}_h}(\overline{\mathbf{x}}_F) = 0$ for all $F \in \mathfrak{g} \cap \mathscr{F}^{\mathrm{b}}_h$; (ii) $\xi^{\mathfrak{g}}_{\mathbf{v}_h}$ is affine inside $T_{\mathfrak{g}}$ and is continuous across every interface in the group: For all $F \in \mathfrak{g} \cap \mathscr{F}^{\mathrm{i}}_h$ such that $F \subset \partial T_1 \cap \partial T_2$,

$$\forall \mathbf{x} \in F, \qquad \xi^{\mathfrak{g}}_{\mathbf{v}_h|T_1}(\mathbf{x}) = \xi^{\mathfrak{g}}_{\mathbf{v}_h|T_2}(\mathbf{x});$$

(iii) $\xi^{\mathfrak{g}}_{\mathbf{v}_h}$ has continuous diffusive flux across every interface in the group: For all $F \in \mathfrak{g} \cap \mathscr{F}^{\mathrm{i}}_h$ such that $F \subset \partial T_1 \cap \partial T_2$,

$$(\boldsymbol{\kappa}\nabla\xi^{\mathfrak{g}}_{\mathbf{v}_h})_{|T_1}\cdot\mathbf{n}_F = (\boldsymbol{\kappa}\nabla\xi^{\mathfrak{g}}_{\mathbf{v}_h})_{|T_2}\cdot\mathbf{n}_F.$$

For further details on the L-construction including an explicit formula for $\xi^{\mathfrak{g}}_{\mathbf{v}_h}$ we refer to [3]. For every face $F \in \mathscr{F}_h$ we define the set $\mathscr{G}_F$ of L-groups containing F,

$$\mathscr{G}_F := \{\mathfrak{g} \in \mathscr{G} \mid F \in \mathfrak{g}\}, \tag{8}$$

and introduce the set of nonnegative weights $\{\varsigma_{\mathfrak{g},F}\}_{\mathfrak{g}\in\mathscr{G}_F}$ such that $\sum_{\mathfrak{g}\in\mathscr{G}_F} \varsigma_{\mathfrak{g},F} = 1$. The trial space for the G-method is obtained as follows: (i) let $\mathscr{S}_h = \mathscr{P}_h$ and $\mathbb{V}_h = \mathbb{T}_h$; (ii) let $\mathfrak{G}_h = \mathfrak{G}^{\mathrm{g}}_h$ with $\mathfrak{G}^{\mathrm{g}}_h$ such that

$$\forall \mathbf{v}_h \in \mathbb{T}_h,\ \forall T \in \mathscr{T}_h,\ \forall F \in \mathscr{F}_T, \qquad \mathfrak{G}^{\mathrm{g}}_h(\mathbf{v}_h)_{|\mathscr{P}_{T,F}} = \sum_{\mathfrak{g}\in\mathscr{G}_F} \varsigma_{\mathfrak{g},F}\nabla\xi^{\mathfrak{g}}_{\mathbf{v}_h|\mathscr{P}_{T,F}}.$$

We denote by $\mathfrak{R}^{\mathrm{g}}_h$ the reconstruction operator defined as in (5) with $\mathfrak{G}_h = \mathfrak{G}^{\mathrm{g}}_h$ and let $V^{\mathrm{g}}_h := \mathfrak{R}^{\mathrm{g}}_h(\mathbb{V}_h)$. The G-method of [3] is then equivalent to the following Petrov-Galerkin method:

$$\text{Find } u_h \in V^{\mathrm{g}}_h \text{ s.t. } a^{\mathrm{g}}_h(u_h, v_h) = \int_\Omega f v_h \text{ for all } v_h \in \mathbb{P}^0_d(\mathscr{T}_h),$$

where $a^{\mathrm{g}}_h(u_h, v_h) := -\sum_{F\in\mathscr{F}_h}\int_F \{\!\{\boldsymbol{\kappa}\nabla_h u_h\}\!\}\cdot\mathbf{n}_F [\![v_h]\!]$ with ∇_h broken gradient on $\mathscr{S}_h$.

Remark 2 (An unconditionally stable method). The main drawback of the G-method is that stability can only be proven under quite stringent conditions; see, e.g., [3, Lemma 3.4]. A possible way to circumvent this difficulty has recently been proposed by one of the authors [10] in the context of CCG methods. The key idea is to use V^{g}_h both as a trial and test space, and modify the discrete bilinear form to recover both consistency and stability. Since the discrete functions in V_h^{g} are discontinuous across the lateral faces of the pyramids in $\mathscr{P}_h$, least-square penalization of the jumps is required to assert stability in terms of coercivity. The resulting method also enters the present framework, but is not detailed here for the sake of conciseness.

A cell centered Galerkin method The L-construction is used to define a trace reconstruction in the CCG method of [8, 10]. More specifically, for all $F \in \mathscr{F}_h^{\mathrm{i}}$, we select one group $\mathfrak{g}_F \in \mathscr{G}_F$ with $\mathscr{G}_F$ defined by (8) and introduce the linear trace operator $\mathbf{T}_h^{\mathrm{g}} : \mathbb{T}_h \to \mathbb{F}_h$ which maps every vector of cell centered DOFs $\mathbf{v}_h \in \mathbb{T}_h$ onto a vector $(v_F)_{F \in \mathscr{F}_h} \in \mathbb{F}_h$ such that

$$v_F = \begin{cases} \xi_{\mathbf{v}_h}^{\mathfrak{g}_F}(\overline{\mathbf{x}}_F) & \text{if } F \in \mathscr{F}_h^{\mathrm{i}}, \\ 0 & \text{if } F \in \mathscr{F}_h^{\mathrm{b}}. \end{cases} \tag{9}$$

The trace operator $\mathbf{T}_h^{\mathrm{g}}$ is then employed in a gradient reconstruction based on Green's formula and inspired from [17]. More precisely, we introduce the linear gradient operator $\mathfrak{G}_h^{\mathrm{green}} : \mathbb{T}_h \times \mathbb{F}_h \to [\mathbb{P}_d^0(\mathscr{T}_h)]^d$ such that, for all $(\mathbf{v}^{\mathscr{T}}, \mathbf{v}^{\mathscr{F}}) \in \mathbb{T}_h \times \mathbb{F}_h$ and all $T \in \mathscr{T}_h$,

$$\mathfrak{G}_h^{\mathrm{green}}(\mathbf{v}^{\mathscr{T}}, \mathbf{v}^{\mathscr{F}})_{|T} = \frac{1}{|T|_d} \sum_{F \in \mathscr{F}_T} |F|_{d-1}(v_F - v_T)\mathbf{n}_{T,F}. \tag{10}$$

The discrete space for the CCG method under examination can then be obtained as follows: (i) let $\mathscr{S}_h = \mathscr{T}_h$ and $\mathbb{V}_h = \mathbb{T}_h$; (ii) let $\mathfrak{G}_h = \mathfrak{G}_h^{\mathrm{ccg}}$ with $\mathfrak{G}_h^{\mathrm{ccg}}$ such that

$$\forall \mathbf{v}_h \in \mathbb{V}_h, \qquad \mathfrak{G}_h^{\mathrm{ccg}}(\mathbf{v}_h) = \mathfrak{G}_h^{\mathrm{green}}(\mathbf{v}_h, \mathbf{T}_h^{\mathrm{g}}(\mathbf{v}_h)). \tag{11}$$

The reconstruction operator defined taking $\mathfrak{G}_h = \mathfrak{G}_h^{\mathrm{ccg}}$ in (5) is denoted by $\mathfrak{R}_h^{\mathrm{ccg}}$, and the corresponding discrete space by $V_h^{\mathrm{ccg}} := \mathfrak{R}_h^{\mathrm{ccg}}(\mathbb{T}_h)$. We define the weights in the average operator as follows: For all $F \in \mathscr{F}_h^{\mathrm{i}}$ such that $F \subset \partial T_1 \cap \partial T_2$,

$$\omega_{T_1,F} = \tfrac{\lambda_{T_2,F}}{\lambda_{T_1,F}+\lambda_{T_2,F}}, \qquad \omega_{T_2,F} = \tfrac{\lambda_{T_1,F}}{\lambda_{T_1,F}+\lambda_{T_2,F}},$$

where $\lambda_{T_i,F} := \boldsymbol{\kappa}_{|T_i}\mathbf{n}_F\cdot\mathbf{n}_F$ for $i \in \{1,2\}$. Set, for all $(u_h, v_h) \in V_h^{\mathrm{ccg}} \times V_h^{\mathrm{ccg}}$,

$$\begin{aligned} a_h^{\mathrm{ccg}}(u_h, v_h) := \int_\Omega \boldsymbol{\kappa}\nabla_h u_h\cdot\nabla_h v_h - \sum_{F \in \mathscr{F}_h} \int_F [\{\!\!\{\boldsymbol{\kappa}\nabla_h u_h\}\!\!\}_\omega\cdot\mathbf{n}_F [\![v_h]\!] + [\![u_h]\!]\{\!\!\{\boldsymbol{\kappa}\nabla v_h\}\!\!\}_\omega\cdot\mathbf{n}_F] \\ + \sum_{F \in \mathscr{F}_h} \eta\frac{\gamma_F}{h_F}\int_F [\![u_h]\!][\![v_h]\!], \end{aligned} \tag{12}$$

with ∇_h broken gradient on $\mathscr{T}_h$, $\gamma_F = \frac{2\lambda_{T_1,F}\lambda_{T_2,F}}{\lambda_{T_1,F}+\lambda_{T_2,F}}$ on internal faces $F \subset \partial T_1 \cap \partial T_2$ and $\gamma_F = \boldsymbol{\kappa}_{|T}\mathbf{n}_F{\cdot}\mathbf{n}_F$ on boundary faces $F \subset \partial T \cap \partial\Omega$. The user-dependent parameter η should be chosen large enough to ensure stability. The CCG method reads

$$\text{Find } u_h \in V_h^{\text{ccg}} \text{ s.t. } a_h^{\text{ccg}}(u_h, v_h) = \int_\Omega f v_h \text{ for all } v_h \in V_h^{\text{ccg}}. \tag{13}$$

The bilinear form a_h^{ccg} has been originally introduced by Di Pietro, Ern and Guermond [12] in the context of dG methods for degenerate advection-diffusion-reaction problems. For $\boldsymbol{\kappa} = \mathbf{1}_d$, the bilinear form a_h^{ccg} becomes

$$\begin{aligned} a_h^{\text{sip}}(u_h, v_h) = &\int_\Omega \nabla_h u_h{\cdot}\nabla_h v_h - \sum_{F\in\mathscr{F}_h}\int_F \left[\{\!\{\nabla_h u_h\}\!\}{\cdot}\mathbf{n}_F [\![v_h]\!] + [\![u_h]\!]\{\!\{\nabla_h v_h\}\!\}{\cdot}\mathbf{n}_F\right] \\ &+ \sum_{F\in\mathscr{F}_h}\frac{\eta}{h_F}\int_F [\![u_h]\!][\![v_h]\!], \end{aligned} \tag{14}$$

and a_h^{sip} is the bilinear form yielding the Symmetric Interior Penalty (SIP) method of Arnold [4]. For further details on the link between CCG and discontinuous Galerkin methods we refer to [8–10].

A hybrid finite volume method As a last example we consider a variant of the SUSHI scheme of [17]; see also [14] for a discussion on the link with the MFD methods of [5, 6]. This method is based on the gradient reconstruction (10), but stabilization is achieved in a rather different manner with respect to (12). More precisely, we define the linear residual operator $\mathfrak{r}_h : \mathbb{T}_h \times \mathbb{F}_h \to \mathbb{P}_d^0(\mathscr{P}_h)$ as follows: For all $T \in \mathscr{T}_h$ and all $F \in \mathscr{F}_T$,

$$\mathfrak{r}_h(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}})_{|\mathscr{P}_{T,F}} = \frac{d^{\frac{1}{2}}}{d_{T,F}}\left[v_F - v_T - \mathfrak{G}_h^{\text{green}}(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}})_{|T}{\cdot}(\mathbf{x}_F - \mathbf{x}_T)\right].$$

We observe, in passing, that the factor $d^{\frac{1}{2}}$ can in general be replaced by a user-defined stabilization parameter $\eta > 0$. The advantage of taking $\eta = d^{\frac{1}{2}}$ is that it yields the classical two-point method on $\boldsymbol{\kappa}$-orthogonal meshes. The discrete space for SUSHI method with hybrid unknowns is obtained as follows: (i) let $\mathscr{S}_h = \mathscr{P}_h$ and $\mathbb{V}_h = \mathbb{T}_h \times \mathbb{F}_h$; (ii) let $\mathfrak{G}_h = \mathfrak{G}_h^{\text{hyb}}$ with $\mathfrak{G}_h^{\text{hyb}}$ such that, for all $(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}}) \in \mathbb{T}_h \times \mathbb{F}_h$, all $T \in \mathscr{T}_h$ and all $F \in \mathscr{F}_T$,

$$\mathfrak{G}_h^{\text{hyb}}(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}})_{|\mathscr{P}_{T,F}} = \mathfrak{G}_h^{\text{green}}(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}})_{|T} + \mathfrak{r}_h(\mathbf{v}_h^{\mathscr{T}}, \mathbf{v}_h^{\mathscr{F}})_{|\mathscr{P}_{T,F}}\mathbf{n}_{T,F}. \tag{15}$$

Denote by $\mathfrak{R}_h^{\text{hyb}}$ the reconstruction operator defined by (5) with $\mathfrak{G}_h = \mathfrak{G}_h^{\text{hyb}}$. The SUSHI method with hybrid unknowns reads

$$\text{Find } u_h \in V_h^{\text{hyb}} \text{ s.t. } a_h^{\text{sushi}}(u_h, v_h) = \int_\Omega f v_h \text{ for all } v_h \in V_h^{\text{hyb}},$$

with $a_h^{\text{sushi}}(u_h, v_h) := \int_\Omega \boldsymbol{\kappa} \nabla_h u_h \cdot \nabla_h v_h$ and ∇_h broken gradient on $\mathscr{P}_h$. Alternatively, one can obtain a cell centered version by setting $\mathbb{V}_h = \mathbb{T}_h$ and replacing $\mathfrak{G}_h^{\text{hyb}}$ defined by (15) by $\mathfrak{G}_h = \mathfrak{G}_h^{\text{cc}}$ with $\mathfrak{G}_h^{\text{cc}}$ such that

$$\forall \mathbf{v}_h \in \mathbb{T}_h, \qquad \mathfrak{G}_h^{\text{cc}}(\mathbf{v}_h) = \mathfrak{G}_h^{\text{hyb}}(\mathbf{v}_h, \mathbf{T}_h^{\text{g}}(\mathbf{v}_h)), \tag{16}$$

and trace operator $\mathbf{T}_h^{\text{g}}$ defined by (9). This variant coincides with the version proposed in [17] for homogeneous $\boldsymbol{\kappa}$, but it has the advantage to reproduce piecewise affine solutions of (7) on $\mathscr{T}_h$ when $\boldsymbol{\kappa}$ is heterogeneous. The discrete space obtained taking $\mathfrak{G}_h = \mathfrak{G}_h^{\text{cc}}$ in (6) is labeled V_h^{cc}.

3 Implementation

The goal of this section is to lay the foundations for a DSL embedded in the C++ language which transposes the mathematical concepts of §2 into practical implementations. To illustrate the capabilities of the DSL in a nutshell, compare Listing 1 with the expression of the bilinear form a_h^{sip} (14). The material is organized as follows: §3.1 introduces the algebraic back-end aiming at replacing the table of DOFs in the context of a element-like assembly procedure; §3.2 deals with more abstract concepts that mimic function spaces, linear and bilinear forms to offer a functional front-end.

3.1 Algebraic back-end

In this section we focus on the elementary ingredients to build the terms appearing in the linear and bilinear forms of §2, which constitute the back-end of the DSL presented in §3.2.

Linear combination The point of view presented in §2 naturally leads to finite element-like assembly of local contributions stemming from integrals over elements or faces. However, a few major differences have to be taken into account: (i) the stencil of the local contributions may vary from term to term; (ii) the stencil may be data-dependent, as is the case for the methods of §2 based on the L-construction; (iii) the stencil may be non-local, as DOFs from neighboring elements may enter in local reconstructions. All of the above facts invalidate the classical approach based on a global table of DOFs inferred from a mesh and a finite element in the sense of Ciarlet. Our approach to meet the above requirements is to (i) drop the concept

Listing 1 Implementation of the bilinear form a_h^{sip} defined by (14) using the DSL of §3

```
// Define discrete spaces, test and trial functions; c.f. Table 1
typedef FunctionSpace<span<Polynomial<d, 1> >,
                      gradient<GreenFormula<LInterpolator> >
                      >::type CCGSpace;
CCGSpace Vh($\mathcal{T}_h$);
Vh.gradientReconstruction().trace().set(DiffusionCoefficient, $\boldsymbol{\kappa}$);
CCGSpace::TrialFunction uh(Vh, "uh");
CCGSpace::TestFunction  vh(Vh, "vh");
// Define the bilinear form
Form2 ah =
  integrate(All<Cell>::items($\mathcal{T}_h$), dot(grad(uh),grad(vh)))
 -integrate(All<Face>::items($\mathcal{T}_h$), dot(N(),avg(grad(uh)))*jump(vh)
                                         +dot(N(),avg(grad(vh)))*jump(uh))
 +integrate(All<Face>::items($\mathcal{T}_h$), $\eta$/H()*jump(uh)*jump(vh));
// Evaluate the bilinear form
MatrixContext context($\mathbf{A}$);
evaluate(ah, context);
```

of local element, and to refer to DOFs by a unique global index; (ii) introduce the concept of `LinearCombination` (with template parameters to be specified in what follows), which realizes a linear application from $\mathbb{V}_h$ onto the space $\mathbb{T}_r$ of real tensors of order $r \leq 2$.

In practice, a `LinearCombination` is an efficient mapping of the DOFs in $\mathbb{V}_h$ onto the corresponding coefficients in $\mathbb{T}_r$. A `LinearCombination` $\mathtt{l}^r$ can indeed be thought of as a list of couples $(I, \tau_{\mathtt{l},I})_{I\in\mathbb{I}_\mathtt{l}}$ where $\mathbb{I}_\mathtt{l} \subset \mathbb{V}_h$ is the stencil (i.e., a vector of global DOFs) and $\tau_{\mathtt{l},I} \in \mathbb{T}_r$, $I \in \mathbb{I}_\mathtt{l}$, are the corresponding coefficients. To account for strongly enforced boundary conditions, `LinearCombination` also contains a constant coefficient $\tau_{\mathtt{l},0}$, so that the evaluation at $\mathbf{v}_h \in \mathbb{V}_h$ (obtained by calling the function `LinearCombination.eval(`$\mathbf{v}_h$`)`) actually returns

$$\mathtt{l}^r(\mathbf{v}_h) = \sum_{I\in\mathbb{I}_\mathtt{l}} \tau_{\mathtt{l},I} v_I + \tau_{\mathtt{l},0} \in \mathbb{T}_r.$$

It is useful to define efficient operations such as the sum and subtraction of linear combinations, as well as different kinds of products by constants. This allows, e.g., to implement the gradient $\mathfrak{G}_h^{\mathrm{green}}$ defined by (10) as described in line 6 of Listing 2. When needed, each DOF I can be represented as a `LinearCombination` containing only the couple $(I, 1)$. As a result, both the hybrid version with face unknowns (15) and the cell centered version (11) of the gradient reconstruction can be obtained from Listing 2 by simply changing the value returned by the trace interpolator $\mathbf{T}_h$`.eval(`F`)` in line 5. We also pinpoint that the tensor order is a template parameter of `LinearCombination` to reduce the need for dynamic allocation.

Listing 2 Implementation of the gradient reconstruction $\mathfrak{G}_h^{\text{green}}$ (10) for an element $T \in \mathscr{T}_h$. The gradient $\mathfrak{G}_h^{\text{ccg}}$ can be obtained by changing the value of the `LinearCombination` returned by $\mathbb{T}_h$ in line 5. Bufferization is used as a means to improve efficiency

```
LinearCombination<0> vT;
vT += LinearCombination<1>::Term($I_T$,1.);
LinearCombination<1, Buffer> buffer;
for($F \in \mathscr{F}_T$) {
  const LinearCombination<0> & vF = $\mathbf{T}_h$.eval($F$);
  buffer += $\frac{|F|_{d-1}}{|T|_d}$(vF-vT)$\mathbf{n}_{T,F}$;
}
LinearCombination<1, Vector> GT;
buffer.compact(GT);
```

In the implementation, particular care must be devoted to expressions containing the sum or subtraction of two linear combinations $\mathtt{l}_1^r$ and $\mathtt{l}_2^r$, since this involves computing the intersection of the corresponding set of DOFs, say $\mathbb{I}_{\mathtt{l}_1}$ and $\mathbb{I}_{\mathtt{l}_2}$ respectively. To overcome this difficulty, complicate expressions are computed in two steps: a first step in which duplicate DOF indices are allowed, followed by a compaction stage where the algebraic sums of the corresponding coefficients are performed. This is obtained by changing the value of the second template parameter of `LinearCombination`. Specifically, in `Buffer`-mode a `LinearCombination` efficiently supports adding terms and can appear in the left-hand side of an assignment operator, while `Vector`-mode (default) only allows to traverse its elements in a fixed order; c.f. lines 3 and 8 of Listing 2.

Linear and bilinear contributions Exploiting the concept of `LinearCombination`, it is possible to devise a unified treatment for local contributions stemming from integrals over elements or faces. We illustrate the main ideas using the an example: For a given $T \in \mathscr{T}_h$ and for $u_h, v_h \in V_h^{\text{ccg}}$, we consider the local contribution $\mathbf{A}_{\text{loc}}$ associated to the term

$$\int_T \boldsymbol{\kappa} \nabla_h u_h \cdot \nabla_h v_h.$$

For the sake of simplicity we focus on the case when the constant coefficient $\tau_{\mathtt{l},0}$ is zero (in the example, this corresponds to the homogeneous Dirichlet boundary condition in problem (7)). The key remark is that both $(\boldsymbol{\kappa}\nabla_h u_h)_{|T} = \boldsymbol{\kappa}_{|T}\nabla(u_{h|T})$ and $(\nabla_h v_h)_{|T} = \nabla(v_{h|T})$ can be represented as objects of type `LinearCombination<1>`, say $\mathtt{l}_u^1 = (J, \tau_{\mathtt{l}_u,J})_{J\in\mathbb{J}}$ and $\mathtt{l}_v^1 = (I, \tau_{\mathtt{l}_v,I})_{I\in\mathbb{I}}$. The associated local contribution reads

$$\mathbf{A}_{\text{loc}} = [|T|_d \tau_{\mathtt{l}_v,I} \cdot \tau_{\mathtt{l}_u,J}]_{I\in\mathbb{I},\, J\in\mathbb{J}}\,. \tag{17}$$

Generalizing the above remark, one can implement local terms in matrix assembly as `BilinearContributions` which can be represented as triplets $(\mathbb{I}, \mathbb{J}, \mathbf{A}_{\text{loc}})$

Listing 3 Assembly of a bilinear and linear contribution (**A** represents here the global matrix **b** the global right-hand side vector)

```
LinearCombination<$r$> $l_u^r$, $l_v^r$;
// Assemble a bilinear contribution into the left-hand side
BilinearContribution<$r$> blc($\gamma$, $l_u^r$, $l_v^r$);
$\mathbf{A}$ << blc;
// Assemble a linear contribution into the right-hand side
LinearContribution<$r$> lc($\gamma$, $l_v^r$);
$\mathbf{b}$ << lc;
```

containing two vectors of DOF indices $\mathbb{I}$ and $\mathbb{J}$ and the local matrix $\mathbf{A}_{\mathrm{loc}}$. Observe, in particular, that $\mathbb{I}$ and $\mathbb{J}$ play the same role as the lines of the table of DOFs corresponding to test and trial functions supported in T in standard finite element implementations. As such, they are related to the lines and columns of the global matrix $\mathbf{A}$ to which $\mathbf{A}_{\mathrm{loc}}$ contributes,

$$\mathbf{A}(\mathbb{I},\mathbb{J}) \leftarrow \mathbf{A}(\mathbb{I},\mathbb{J}) + \mathbf{A}_{\mathrm{loc}}. \tag{18}$$

When the `LinearCombinations` concurring to a local term take values in $\mathbb{T}_r$, the vector inner product in (17) should be replaced by the appropriate tensor contraction. The additional argument γ appearing in Listing 3 serves as a multiplicative factor for the whole expression (in the above example, $\gamma = |T|_d$). More generally, γ can be a function of space and time, and may depend on discrete variables.

Similarly, `LinearContributions` serve to represent right-hand side contributions. `LinearContributions` are not detailed here for the sake of brevity. A typical assembly pattern is described in Listing 3. In particular, line 4 is the programming counterpart of (18).

3.2 Functional front-end

A further level of abstraction can be reached defining a DSL that allows to conceal all technical details and provide only the relevant components in a form as close as possible to the mathematical formulations of §2. We focus here, in particular, on the programming equivalent of discrete spaces and bilinear forms.

Function spaces Incomplete broken polynomial spaces defined by (6) are mapped onto `C++` types conforming to the `FunctionSpace` concept detailed in Listing 4. The actual types are generated by a helper template class parametrized by a containing polynomial space, labeled **`span`**, and a piecewise constant gradient reconstruction, labeled **`gradient`** (labels for template arguments are here defined using the `boost::parameter` library). An example of usage is provided in lines 2–4 in Listing 1. The gradient reconstruction implicitly fixes both the

Listing 4 `FunctionSpace` concept

```
class FunctionSpace {
  // Types for trial and test functions
  class TrialFunction;
  class TestFunction;
  // Constructor
  FunctionSpace(const Mesh &);
  // Constant value of $\mathfrak{G}_{h|S}$ for $S \in \mathscr{S}_h$ as a vector-valued linear combination of DOFs
  const LinearCombination<1> & Gh($S$) const;
  // Value of $\mathfrak{R}_{h|S}(\mathbf{x})$ for $\mathbf{x} \in S$ and $S \in \mathscr{S}_h$ as a scalar-valued linear combination of DOFs
  const LinearCombination<0> & Rh($S$, $\mathbf{x}$) const;
};
```

Table 1 **`span`** and **`gradient`** template parameters for the class `FunctionSpace` corresponding to the discrete spaces of §2

Space	$\mathscr{S}_h$	**span**	**gradient**
$\mathbb{P}_d^0(\mathscr{T}_h)$	$\mathscr{T}_h$	`Polynomial<d, 0>`	`Null`
V_h^{g}	$\mathscr{P}_h$	`Polynomial<d, 1>`	`GFormula`
V_h^{ccg}	$\mathscr{T}_h$	`Polynomial<d, 1>`	`GreenFormula<LInterpolator>`
V_h^{hyb}	$\mathscr{P}_h$	`Polynomial<d, 1>`	`SUSHIFormula<HybridUnknowns>`
V_h^{cc}	$\mathscr{P}_h$	`Polynomial<d, 1>`	`SUSHIFormula<LInterpolator>`

choice (6) for the space of DOFs and the choice (3) for $\mathscr{S}_h$. The programming counterparts of the function spaces appearing in §2 are listed in Table 1.

The key role of a `FunctionSpace` is to bridge the gap between the algebraic representation of DOFs and the functional representation used in the methods of §2. This is achieved by the functions `Gh` and `Rh`, which are the `C++` counterpart of the linear operators $\mathfrak{G}_h$ and $\mathfrak{R}_h$ respectively; see §2.1. More specifically, (i) for all $S \in \mathscr{S}_h$, `Gh(`S`)` returns a vector-valued linear combination corresponding to the (constant) restriction $\mathfrak{G}_{h|S}$; (ii) for all $S \in \mathscr{S}_h$ and all $\mathbf{x} \in S$, `Rh(`S`,` $\mathbf{x}$`)` returns a scalar-valued linear combination corresponding to $\mathfrak{R}_{h|S}(\mathbf{x})$ defined according to (5). The linear combinations returned by `Gh` and `Rh` can be used to generate `LinearContributions` and `BilinearContributions` to build linear and bilinear terms as described above. A `FunctionSpace` also defines the types `TestFunction` and `TrialFunction` that correspond to the mathematical notions of test and trial functions in variational formulations. The main difference between a `TestFunction` and a `TrialFunction` is that the latter is associated to a vector of DOFs which is stored in memory. In addition, when used to define bilinear contributions, test and trial functions are associated to the lines and columns of the local matrix respectively. We conclude by observing that the choice of identifying test and trial functions by their type is in contrast with the approach of [20, §3.4], where special keywords accomplish this task.

Linear and bilinear forms Linear and bilinear forms are obtained as sums of linear and bilinear terms resulting from the composition of `TestFunctions` and

Listing 5 CCG discretization of the Stokes problem

```
CCGSpace::VectorTrialFunction uh($d$);
CCGSpace::VectorTestFunction vh($d$);
P0Space::TrialFunction ph;
P0Space::TestFunction qh;
Range::Index i(Range(0,dim-1));
Form2 ah, bh, sh;
ah = integrate(All<Cell>::items($\mathscr{T}_h$),
               sum(i)(dot(grad(uh(i)),grad(vh(i))) ))
    +integrate(Internal<Face>::items($\mathscr{T}_h$),
               sum(i)(-dot(fn,avg(grad(uh(i)))))*jump(vh(i))
                      -jump(uh(i))*dot(N(),avg(grad(vh(i))))
                      +$\eta$/H()*jump(uh(i))*jump(vh(i))));
bh =-integrate(Internal<Face>::items($\mathscr{T}_h$),
               jump(ph)*dot(N(),avg(vh)));
sh = integrate(Internal<Face>::items($\mathscr{T}_h$),H()*jump(ph)*jump(qh));
```

`TrialFunctions` (or unary modifications thereof) by suitable tensor contractions. Examples of tensor contractions in Listing 1 are the **`dot`** and `*` operators. Products by functions of space, time and possibly discrete variables are also allowed. In Listing 1 we also display examples of geometric operators such as **`N`**`()` and **`H`**`()`, which allow to access face normals and diameters respectively. Unary modifiers encountered in Listing 1 are **`grad`**, **`avg`** and **`jump`**, corresponding, respectively, to the broken gradient on $\mathscr{S}_h$ and to the average and jump operators defined by (4). When applied to a test or trial function, a unary modifier is an object capable of returning a `LinearCombination` at evaluation.

By default, linear and bilinear forms are represented by vectors and (sparse) matrices, but other representations are possible resulting, e.g., in matrix-free implementations. In contrast with [20], the expression corresponding to a linear (resp. bilinear) form is stored as a property of an object **`Form1`** (resp. **`Form2`**) instead of being evaluated on-the-fly. This allows, in particular, to change the operations actually performed at evaluation according to a context. Changing the representation of linear and bilinear forms thus amounts to changing the context of evaluation. An example of evaluation is provided in lines 16–17 of Listing 1, where the global matrix $\mathbf{A}$ is assembled according to the expression of `ah` and to the procedure defined in `MatrixContext`. During the evaluation, each term in the expression of `ah` generates a corresponding `BilinearContribution`, which is in turn assembled as described in §3.1.

To conclude, we present a more complicate example involving a system of PDEs. More specifically, we consider the Stokes problem:

$$-\triangle u + \nabla p = f \text{ in } \Omega, \quad \nabla{\cdot}u = 0 \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega,$$

with $\langle p \rangle_\Omega = 0$ to ensure well-posedness. Let $X_h := [V_h^{\text{ccg}}]^d \times \mathbb{P}_d^0(\mathscr{T}_h)/\mathbb{R}$. In Listing 5 we present the implementation of the CCG method of [9, §3]: Find

$(u_h, p_h) \in X_h$ such that

$$a_h(u_h, v_h) + b_h(v_h, p_h) - b_h(u_h, q_h) + s_h(p_h, q_h) = \int_\Omega f \cdot v_h, \qquad \forall (v_h, q_h) \in X_h$$

where $a_h(u_h, v_h) := \sum_{i=1}^{d} a_h^{\mathrm{sip}}(u_{h,i}, v_{h,i})$, $b_h(p_h, v_h) := -\sum_{F \in \mathscr{F}_h^{\mathrm{i}}} \int_F [\![p_h]\!] \{\!\{v_h\}\!\} \cdot \mathbf{n}_F$ and $s_h(p_h, q_h) := \sum_{F \in \mathscr{F}_h^{\mathrm{i}}} h_F \int_F [\![p_h]\!][\![q_h]\!]$. Notice the use of the **`sum`** keyword.

Acknowledgements Fruitful discussions with Christophe Prud'homme (Laboratoire Jean Kuntzmann, University of Grenoble) are gratefully acknowledged. We also wish to thank all the contributors to the `Arcane` platform.

References

1. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media, Part I: Derivation of the methods. *SIAM J. Sci. Comput.*, 19(5):1700–1716, 1998.
2. I. Aavatsmark, G. T. Eigestad, B. T. Mallison, and J. M. Nordbotten. A compact multipoint flux approximation method with improved robustness. *Numer. Methods Partial Differ. Eq.*, 24:1329–1360, 2008.
3. L. Agélas, D. A. Di Pietro, and J. Droniou. The G method for heterogeneous anisotropic diffusion on general meshes. *M2AN Math. Model. Numer. Anal.*, 44(4):597–625, 2010.
4. D. N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.
5. F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of mimetic finite difference methods for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 43(5):1872–1896, 2005.
6. F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *M3AS*, 15:1533–1553, 2005.
7. I. Danaila, F. Hecht, and O. Pironneau. *Simulation numérique en C++*. Dunod, Paris, 2003. http://www.freefem.org.
8. D. A. Di Pietro. Cell centered Galerkin methods. *C. R. Acad. Sci. Paris, Ser. I*, 348:31–34, 2010.
9. D. A. Di Pietro. Cell centered Galerkin methods for diffusive problems. Submitted. Preprint available at http://hal.archives-ouvertes.fr/hal-00511125/en/, September 2010.
10. D. A. Di Pietro. A compact cell-centered Galerkin method with subgrid stabilization. *C. R. Acad. Sci. Paris, Ser. I.*, 348(1–2):93–98, 2011.
11. D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*. Mathematics & Applications. Springer-Verlag, Berlin, 2010. To appear.
12. D. A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection. *SIAM J. Numer. Anal.*, 46(2):805–831, 2008.
13. J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Num. Math.*, 105(1):35–71, 2006.
14. J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *M3AS, Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.
15. M. G. Edwards and C. F. Rogers. Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Comput. Geosci.*, 2:259–290, 1998.

16. A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004.
17. R. Eymard, Th. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 4(30):1009–1043, 2010.
18. G. Grospellier and B. Lelandais. The Arcane development framework. In *Proceedings of the 8th workshop on Parallel/High-Performance Object-Oriented Scientific Computing*, pages 4:1–4:11, New York, NY, USA, 2009. ACM.
19. A. Logg and G. N. Wells. DOLFIN: Automated finite element computing. *ACM TOMS*, 37, 2010.
20. C. Prud'homme. A domain specific embedded language in C++ for automatic differentiation, projection, integration and variational formulations. *Sci. Prog.*, 14(2):81–110, 2006.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A Unified Framework for a posteriori Error Estimation in Elliptic and Parabolic Problems with Application to Finite Volumes

Alexandre Ern and Martin Vohralík

Abstract We present a unified framework based on potential and flux reconstruction for guaranteed and efficient a posteriori error estimation. We consider as model problems the Laplace equation, the singularly perturbed convection-diffusion-reaction equation, and the heat equation. The analysis is performed for a wide class of space discretization schemes. Three simple conditions need to be verified, which we do for cell- and vertex-centered finite volumes for all model problems.

Keywords a posteriori error estimation, guaranteed upper bound, efficiency, robustness, elliptic and parabolic problems
MSC2010: 65M08, 65M15, 65M50, 65N08, 65N15, 65N50

1 Introduction

A posteriori error estimation is an important tool in practical computations for error control and computational efficiency by adapting the discretization parameters. In the context of finite element methods, residual-based a posteriori error estimation has been initiated by Babuška and Rheinboldt [2] over three decades ago. The application to finite volume (FV) schemes is more recent; we refer, among others, to Achdou, Bernardi, and Coquel [1], Nicaise [19], and Ohlberger [20, 21].

Alexandre Ern
Université Paris-Est, CERMICS, Ecole des Ponts, 77455 Marne la Vallée, France, e-mail: ern@cermics.enpc.fr

Martin Vohralík
UPMC Univ. Paris 06, UMR 7598, Laboratoire J.-L. Lions, 75005, Paris, France & CNRS, UMR 7598, Laboratoire J.-L. Lions, 75005, Paris, France, e-mail: vohralik@ann.jussieu.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_85, © Springer-Verlag Berlin Heidelberg 2011

The purpose of this work is to present some recent results (and extensions thereof) by the authors [9, 11, 28–30] in a general framework. The salient features of this framework can be summarized as follows. Firstly, the error upper bound is formulated in terms of a *potential* and a *flux reconstruction* which must comply with some basic physical properties related to the model problem at hand. This approach allows one to achieve *guaranteed* error upper bounds, that is, upper bounds *without undetermined constants*, which is a key feature in the context of error control. Flux-based a posteriori error estimation for elliptic problems hinges on the Prager–Synge equality [22] and was first developed, among others, by Ladevèze [18] and Haslinger and Hlaváček [14].

Next, the present approach does not rely on a specific discretization scheme (in space), that is, we bound the difference between the exact solution and an arbitrary approximate solution which is only required to be piecewise smooth. Owing to this generality, the approach encompasses a wide class of schemes including FVs and many other schemes (discontinuous Galerkin, mixed finite elements, etc.) in a *unified setting*. At this stage, quite *general meshes* (e.g., with polygonal elements and so-called hanging nodes) can be considered as well. Turning next to *local efficiency*, that is, to local lower bounds on the error, we still proceed generally without resorting to any specific discretization scheme under two additional assumptions. On the one hand, we suppose that the approximate solution, the potential and flux reconstructions, and the problem data are piecewise polynomials and that the meshes possess some regularity which we formulate by introducing a matching simplicial, shape-regular submesh. On the other hand, we assume that the potential and flux reconstructions satisfy some local approximation properties which are expressed in terms of suitable local residuals of the approximate solution (plus its jumps). Local lower bounds on the error then result from the combination of these two assumptions and the fact that the local residuals provide local lower bounds on the approximation error, as previously shown, e.g., in Verfürth [24].

This paper is organized as follows. In §2, we collect some useful notation and basic ingredients for the analysis. Then, we present our results on three model problems. In §3, we consider the Laplace equation. The aim is to present in detail the key ideas in the context of a simple model problem. In §4, we turn to the convection-diffusion-reaction equation. We focus on singularly perturbed regimes resulting from dominant convection or reaction and show how the present approach can achieve *robustness* with respect to physical parameters. In §5, we consider the heat equation and the backward Euler scheme to discretize in time. The purpose is to show how the present approach handles evolution problems including time-varying meshes. In all cases, we first derive upper and lower bounds on the approximation error in an abstract framework applicable to a wide class of discretization schemes in space. Then, we show how the framework can be applied to cell- and vertex-centered FV schemes. For the sake of simplicity, we only consider model problems with homogeneous Dirichlet boundary conditions. Inhomogeneous Dirichlet and Neumann boundary conditions can be taken into account following [29]. Finally, we observe that some interesting applications of a posteriori error estimates are not covered herein; we mention, in particular, the use of such estimates as adaptive stopping criteria for linear [15] and nonlinear [7] iterative solvers.

2 Basic ingredients

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a polygonal (polyhedral) domain (open, bounded, and connected set). Let $\mathscr{T}_h$ be a partition of Ω into polygonal elements. The elements K can be *nonconvex* or *non star-shaped*. We denote by h_K the diameter of $K \in \mathscr{T}_h$ and by $\mathbf{n}_K$ its outward normal. The partition $\mathscr{T}_h$ can be *nonmatching*, that is, so-called hanging nodes are allowed. We only suppose later on (cf. Assumption 3 below) the existence of a simplicial matching and shape-regular submesh $\mathscr{S}_h$. We say that σ is a mesh side if σ has positive $(d-1)$-dimensional measure and if there are distinct $K, L \in \mathscr{T}_h$ such that $\sigma = \partial K \cap \partial L$ or if there is $K \in \mathscr{T}_h$ such that $\sigma = \partial K \cap \partial \Omega$. Mesh sides are collected in the set $\mathscr{E}_h$. We denote by h_σ the diameter of $\sigma \in \mathscr{E}_h$, we fix a unit normal to σ denoted by $\mathbf{n}_\sigma$, and define the jump across σ as the difference following the direction of $\mathbf{n}_\sigma$. Besides the usual Sobolev spaces $H^1(\Omega)$ and $H_0^1(\Omega)$, we consider the so-called broken Sobolev space $H^1(\mathscr{T}_h)$ spanned by those functions whose restriction to each element $K \in \mathscr{T}_h$ belongs to $H^1(K)$ and the so-called broken gradient operator ∇_h acting elementwise on functions in $H^1(\mathscr{T}_h)$. Additionally, we need the space $\mathbf{H}(\mathrm{div}, \Omega)$ spanned by those functions in $[L^2(\Omega)]^d$ with square-integrable weak divergence. The notation $\mathbb{P}_k(\mathscr{T}_h)$ stands for the space of piecewise polynomials of total degree $\leq k$ on $\mathscr{T}_h$, whereas, for $\mathscr{T}_h$ simplicial and matching, $\mathbf{RTN}(\mathscr{T}_h) \subset \mathbf{H}(\mathrm{div}, \Omega)$ stands for the (lowest-order) Raviart–Thomas–Nédélec finite element space [3]. For all $\mathbf{v}_h \in \mathbf{RTN}(\mathscr{T}_h)$, $\mathbf{v}_h \cdot \mathbf{n}_\sigma$ is constant on all sides $\sigma \in \mathscr{E}_h$, the univalued side fluxes $\langle \mathbf{v}_h \cdot \mathbf{n}_\sigma, 1 \rangle_\sigma$ representing the degrees of freedom.

Let $D \subset \Omega$ be a polygon or polyhedron. The Poincaré inequality states that

$$\|\varphi - \varphi_D\|_D^2 \leq C_{\mathrm{P},D} h_D^2 \|\nabla \varphi\|_D^2 \qquad \forall \varphi \in H^1(D), \tag{1}$$

where φ_D is the mean of φ over D given by $\varphi_D := (\varphi, 1)_D / |D|$. When D is convex, the constant $C_{\mathrm{P},D}$ can be evaluated as $1/\pi^2$. The constant $C_{\mathrm{P},D}$ can also be evaluated for nonconvex D, cf. [12, Lemma 10.2] or [5, §2]. Let now $K \subset \Omega$ be a simplex and let σ be one of its sides. The trace inequality states that

$$\|\varphi\|_\sigma^2 \leq C_{\mathrm{t},K,\sigma} (h_K^{-1} \|\varphi\|_K^2 + \|\varphi\|_K \|\nabla \varphi\|_K) \qquad \forall \varphi \in H^1(K). \tag{2}$$

It follows from [23, Lemma 3.12] that the constant $C_{\mathrm{t},K,\sigma}$ can be evaluated as $|\sigma| h_K / |K|$, see also [5, Theorem 4.1] for $d = 2$.

3 Laplace equation

We consider the second-order elliptic problem

$$-\Delta p = f \qquad \text{in } \Omega, \tag{3a}$$

$$p = 0 \qquad \text{on } \partial\Omega, \tag{3b}$$

with $f \in L^2(\Omega)$. The weak formulation consists in finding $p \in H_0^1(\Omega)$ such that

$$(\nabla p, \nabla\varphi) = (f, \varphi) \qquad \forall \varphi \in H_0^1(\Omega). \tag{4}$$

The scalar-valued function $p \in H_0^1(\Omega)$ is called the *potential* and the vector-valued function $\mathbf{t} := -\nabla p \in \mathbf{H}(\mathrm{div}, \Omega)$ the (diffusive) *flux*.

3.1 Abstract framework

The purpose of this section is to present a unified abstract framework for a posteriori error estimation in problem (3a)–(3b). In order to proceed generally, without the specification of the numerical scheme at hand, we merely suppose that we are given a function $p_h \in H^1(\mathscr{T}_h)$ (which will represent the discrete solution later on). We define the energy (semi-)norm as $|||v||| := \|\nabla_h v\|$ for all $v \in H^1(\mathscr{T}_h)$. The a posteriori estimate for the energy error $|||p - p_h|||$ is formulated in terms of a *potential reconstruction* s_h and a *flux reconstruction* $\mathbf{t}_h$. These reconstructions must comply with the following assumption.

Assumption 1 (Potential and flux reconstruction for (3a)–(3b)**)** *There holds* $s_h \in H_0^1(\Omega)$*,* $\mathbf{t}_h \in \mathbf{H}(\mathrm{div}, \Omega)$*, and*

$$(\nabla{\cdot}\mathbf{t}_h, 1)_K = (f, 1)_K \qquad \forall K \in \mathscr{T}_h. \tag{5}$$

Remark 1 (Assumption 1). Assumption 1 is concerned with basic physical *constraints* and *local conservation*. For the exact solution, $p \in H_0^1(\Omega)$ and $\mathbf{t} \in \mathbf{H}(\mathrm{div}, \Omega)$ (physical constraints); moreover, $\nabla{\cdot}\mathbf{t} = f$ (conservation). The potential and flux reconstructions mimic these continuous properties.

We can now state and prove our main result concerning the error upper bound, see [27, Theorem 4.2] and [30, Theorem 4.5].

Theorem 2 (A posteriori estimate for (3a)–(3b)**).** *Let* p *be the solution of* (4) *and let* $p_h \in H^1(\mathscr{T}_h)$ *be arbitrary. Let Assumption 1 be satisfied. Then,*

$$|||p - p_h||| \le \left\{ \sum_{K \in \mathscr{T}_h} \eta_{\mathrm{NC},K}^2 + (\eta_{\mathrm{R},K} + \eta_{\mathrm{DF},K})^2 \right\}^{1/2},$$

where, for all $K \in \mathscr{T}_h$*, the* diffusive flux estimator*, the* nonconformity estimator*, and the* residual estimator *are respectively given by*

$$\eta_{\mathrm{DF},K} := \|\nabla p_h + \mathbf{t}_h\|_K, \tag{6a}$$

$$\eta_{\mathrm{NC},K} := \|\nabla(p_h - s_h)\|_K, \tag{6b}$$

$$\eta_{\mathrm{R},K} := C_{\mathrm{P},K}^{1/2} h_K \|f - \nabla{\cdot}\mathbf{t}_h\|_K. \tag{6c}$$

Proof. Following [17, Lemma 4.4], we obtain using $s_h \in H_0^1(\Omega)$,

$$|||p - p_h|||^2 \le |||p_h - s_h|||^2 + \left\{ \sup_{\varphi \in H_0^1(\Omega), |||\varphi|||=1} (\nabla_h(p - p_h), \nabla\varphi) \right\}^2.$$

The first term equals the Hilbertian sum of the nonconformity estimators, and we are thus left with bounding the second term. Using (4) and $\mathbf{t}_h \in \mathbf{H}(\mathrm{div}, \Omega)$, we obtain

$$\begin{aligned}(\nabla_h(p - p_h), \nabla\varphi) &= (f, \varphi) - (\nabla_h p_h, \nabla\varphi) = (f, \varphi) - (\nabla_h p_h + \mathbf{t}_h, \nabla\varphi) + (\mathbf{t}_h, \nabla\varphi) \\ &= (f - \nabla{\cdot}\mathbf{t}_h, \varphi) - (\nabla_h p_h + \mathbf{t}_h, \nabla\varphi).\end{aligned}$$

We now bound the two above terms separately. For all $K \in \mathscr{T}_h$, let φ_K be the mean value of φ over K. Then, using (5), the Poincaré inequality (1), and the Cauchy–Schwarz inequality, we infer

$$|(f - \nabla{\cdot}\mathbf{t}_h, \varphi)_K| = |(f - \nabla{\cdot}\mathbf{t}_h, \varphi - \varphi_K)_K| \le \eta_{\mathrm{R},K} |||\varphi|||_K.$$

Moreover, bounding $|(\nabla p_h + \mathbf{t}_h, \nabla\varphi)_K| \le \eta_{\mathrm{DF},K} |||\varphi|||_K$ is immediate using the Cauchy–Schwarz inequality. The conclusion is straightforward. □

We now address local efficiency and we still proceed generally, without any notion of a particular numerical scheme. We make two more assumptions.

Assumption 3 (Local efficiency) *We suppose that*

- *there exists a shape-regular matching simplicial submesh* $\mathscr{S}_h$ *of* $\mathscr{T}_h$ *such that, for each* $K \in \mathscr{T}_h$*, the number of subelements* $L \subset K$*,* $L \in \mathscr{S}_h$*, is uniformly bounded;*
- *for a fixed integer* $k \ge 1$*, the approximate solution* p_h *and the datum* f *are in* $\mathbb{P}_k(\mathscr{T}_h)$*, and the flux reconstruction* $\mathbf{t}_h$ *is in* $[\mathbb{P}_k(\mathscr{S}_h)]^d$*;*

Henceforth, we use $A \lesssim B$ when there exists a positive constant C, that can only depend on the space dimension d, the shape-regularity parameter of the mesh $\mathscr{S}_h$, and the polynomial degree k, such that $A \leq CB$. For all $K \in \mathscr{T}_h$, let $\mathfrak{T}_K$ denote all the elements in $\mathscr{T}_h$ having a nonempty intersection with K, $\mathfrak{E}_K$ all the sides in $\mathscr{E}_h$ having a nonempty intersection with K, and $\mathfrak{E}_K^{\mathrm{int}}$ the subset of $\mathfrak{E}_K$ collecting those sides lying in the interior of Ω. We introduce the *classical residual estimators* for problem (3a)–(3b) (cf. [24] for conforming methods and [1, 6] for nonconforming methods) given by

$$\eta_{\mathrm{res},K} := h_K \|f + \Delta p_h\|_{\mathfrak{T}_K} + h_K^{1/2} \|[\![\nabla_h p_h \cdot \mathbf{n}]\!]\|_{\mathfrak{E}_K^{\mathrm{int}}}, \tag{7a}$$

$$|p_h|_{\mathrm{J},K} := h_K^{-1/2} \|[\![p_h]\!]\|_{\mathfrak{E}_K}. \tag{7b}$$

Assumption 4 (Approximation property for (3a)–(3b)) *We assume that, for all $K \in \mathscr{T}_h$,*

$$\|\nabla(p_h - s_h)\|_K + \|\nabla p_h + \mathbf{t}_h\|_K \lesssim \eta_{\mathrm{res},K} + |p_h|_{\mathrm{J},K}. \tag{8}$$

We can now state and prove our main result concerning efficiency.

Theorem 5 (Efficiency of the estimate of Theorem 2). *Let p be the solution of (4) and let Assumptions 3 and 4 be satisfied. Then, for all $K \in \mathscr{T}_h$,*

$$\eta_{\mathrm{NC},K} + \eta_{\mathrm{R},K} + \eta_{\mathrm{DF},K} \lesssim |||p - p_h|||_{\mathfrak{T}_K} + |p_h|_{\mathrm{J},K}.$$

Proof. Our first step is to observe that $\eta_{\mathrm{NC},K} + \eta_{\mathrm{R},K} + \eta_{\mathrm{DF},K} \lesssim \eta_{\mathrm{res},K} + |p_h|_{\mathrm{J},K}$. This bound is immediate for $\eta_{\mathrm{NC},K}$ and $\eta_{\mathrm{DF},K}$ owing to Assumption 4, while for $\eta_{\mathrm{R},K}$, the triangle and inverse inequalities yield $\eta_{\mathrm{R},K} \lesssim h_K\|f + \Delta p_h\|_K + \|\nabla p_h + \mathbf{t}_h\|_K \lesssim \eta_{\mathrm{res},K} + |p_h|_{\mathrm{J},K}$, owing to Assumptions 3 and 4. Our second step is to observe that $\eta_{\mathrm{res},K} \lesssim |||p - p_h|||_{\mathfrak{T}_K}$, as can be derived using suitable bubble functions [24]. □

Remark 2 (Equivalence result). If p_h is in $H_0^1(\Omega)$, the jump seminorm $|p_h|_{\mathrm{J},K}$ vanishes. If the jumps of p_h have zero mean on each side, proceeding as in [1, Theorem 10] yields $|p_h|_{\mathrm{J},K} \lesssim |||p - p_h|||_{\mathfrak{T}_K}$. Finally, in the general case, an equivalence result is achieved by adding the jump seminorm $|p - p_h|_{\mathrm{J},K} = |p_h|_{\mathrm{J},K}$ to both the error measure and the nonconformity estimator.

3.2 Application to finite volumes

We apply here the framework of §3.1 to cell- and vertex-centered finite volume schemes, i.e., we specify s_h and $\mathbf{t}_h$, and we verify Assumptions 1, 3, and 4.

3.2.1 Cell-centered finite volumes

Definition 1 (Cell-centered FVs for (3a)–(3b)). A cell-centered FV scheme for discretizing (3a)–(3b), cf. [12], reads: find $\bar{p}_h \in \mathbb{P}_0(\mathscr{T}_h)$ such that

$$\sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma} = (f, 1)_K \qquad \forall K \in \mathscr{T}_h. \tag{9}$$

Here, $\mathscr{E}_K$ collects the sides of K and $F_{K,\sigma}$ is the diffusive flux through the side σ, which depends on $\bar{p}_h$. A simple example is the so-called "two-point" scheme. In what follows, we do not need the specific form of $F_{K,\sigma}$, but only the conservation property $F_{K,\sigma} = -F_{L,\sigma}$ for all interior sides σ shared by the elements K and L.

Let us first suppose that $\mathscr{T}_h$ is simplicial and matching. Following [13], let $\mathbf{t}_h \in \mathbf{RTN}(\mathscr{T}_h)$ be prescribed on all $K \in \mathscr{T}_h$ by the fluxes $F_{K,\sigma}$ as

$$(\mathbf{t}_h|_K \cdot \mathbf{n}_K)|_\sigma := F_{K,\sigma}/|\sigma|. \tag{10}$$

Since $\bar{p}_h$ is piecewise constant, the energy error $|||p - \bar{p}_h||| = \|\nabla p\|$ is not relevant. Instead, following [28, §3.2], we first postprocess $\bar{p}_h$ locally into $p_h \in \mathbb{P}_2(\mathscr{T}_h)$ such that, for all $K \in \mathscr{T}_h$,

$$-\nabla p_h|_K = \mathbf{t}_h|_K, \qquad (p_h, 1)_K/|K| = \bar{p}_h|_K. \tag{11}$$

The potential s_h is constructed by applying an averaging operator $\mathscr{I}_{\mathrm{av}} : \mathbb{P}_k(\mathscr{T}_h) \to \mathbb{P}_k(\mathscr{T}_h) \cap H_0^1(\Omega)$ to p_h. This operator sets the Lagrangian degrees of freedom inside Ω to the average of the values and sets 0 on $\partial\Omega$. Theorem 2 can now used to bound the error $|||p - p_h|||$ observing that (5) in Assumption 1 results from $(\nabla \cdot \mathbf{t}_h, 1)_K = \langle \mathbf{t}_h \cdot \mathbf{n}_K, 1 \rangle_{\partial K} = \sum_{\sigma \in \mathscr{E}_K} F_{K,\sigma} = (f, 1)_K$. Note that $\eta_{\mathrm{DF},K} = 0$ from (11), which is typical for cell-centered finite volumes. To apply Theorem 5, we verify Assumptions 3 and 4. Assumption 3 is straightforward with $\mathscr{S}_h = \mathscr{T}_h$, whereas Assumption 4 is trivial for $\mathbf{t}_h$ since $\|\nabla p_h + \mathbf{t}_h\|_K = 0$, while the bound on $\|\nabla(p_h - \mathscr{I}_{\mathrm{av}}(p_h))\|_K$ results from [1, 4, 16].

When $\mathscr{T}_h$ is not simplicial or is nonmatching, the submesh $\mathscr{S}_h$ needs to be introduced. We can then proceed as in [28, §5] and [10]. The averaging operator for potential reconstruction maps into $\mathbb{P}_k(\mathscr{S}_h) \cap H_0^1(\Omega)$, while the flux is reconstructed in $\mathbf{RTN}(\mathscr{S}_h)$ either by direct prescription of its degrees of freedom or by solving local Neumann problems.

3.2.2 Vertex-centered finite volumes

We suppose here that $\mathscr{T}_h$ is simplicial and matching. Let $\mathscr{D}_h$ be a dual mesh with dual volumes D associated with the vertices of $\mathscr{T}_h$. We refer to Fig. 1, left, for an illustration. We decompose $\mathscr{D}_h$ into $\mathscr{D}_h^{\mathrm{int}}$ and $\mathscr{D}_h^{\mathrm{ext}}$, with $\mathscr{D}_h^{\mathrm{int}}$ associated with interior vertices and $\mathscr{D}_h^{\mathrm{ext}}$ with boundary ones.

Definition 2 (Vertex-centered FVs for (3a)–(3b)). A vertex-centered FV scheme for discretizing (3a)–(3b), cf. [12], reads: find $p_h \in \mathbb{P}_1(\mathscr{T}_h) \cap H_0^1(\Omega)$ such that

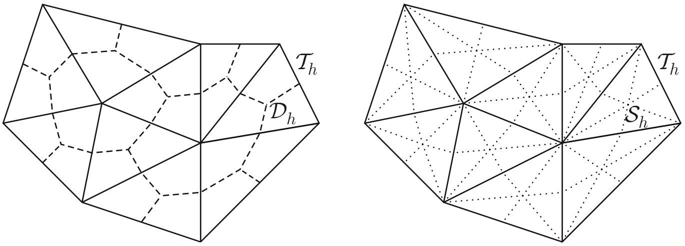


Fig. 1 Simplicial mesh $\mathscr{T}_h$ and the dual mesh $\mathscr{D}_h$ (left); simplicial submesh $\mathscr{S}_h$ (right)

$$-\langle \nabla p_h \cdot \mathbf{n}_D, 1\rangle_{\partial D} = (f, 1)_D \qquad \forall D \in \mathscr{D}_h^{\mathrm{int}}. \tag{12}$$

To apply the framework of §3.1, we first note that, since $p_h \in H_0^1(\Omega)$, we can set $s_h = p_h$. Consequently, $\eta_{\mathrm{NC},K} = 0$ in Theorem 2, which is typical for vertex-centered finite volumes. To construct the flux $\mathbf{t}_h$, we introduce a matching simplicial submesh $\mathscr{S}_h$, cf. Fig. 1, right. Such $\mathscr{S}_h$ is a refinement of both $\mathscr{T}_h$ and $\mathscr{D}_h$. The flux $\mathbf{t}_h$ is reconstructed in $\mathbf{RTN}(\mathscr{S}_h)$ such that, at all interior sides σ of $\mathscr{S}_h$ which lie on the boundary of some $D \in \mathscr{D}_h$, $\mathbf{t}_h \cdot \mathbf{n}_\sigma := -\nabla p_h \cdot \mathbf{n}_\sigma$. Owing to the Green theorem, $(\nabla \cdot \mathbf{t}_h, 1)_D = (f, 1)_D$ for all $D \in \mathscr{D}_h^{\mathrm{int}}$. There are various ways of prescribing the remaining degrees of freedom of $\mathbf{t}_h$. We can merely prescribe them directly, but better computational results are obtained if a local Neumann or Neumann/Dirichlet problem is solved using mixed finite elements in each $D \in \mathscr{D}_h$ [30, §4.3]. Verifying Assumptions 1 and 3 is immediate, while Assumption 4 is proven as in [30, §5].

4 Convection-diffusion-reaction equation

We consider the convection-diffusion-reaction equation

$$-\nabla \cdot (\varepsilon \nabla p - \mathbf{w} p) + r p = f \quad \text{in } \Omega, \tag{13a}$$

$$p = 0 \quad \text{on } \partial\Omega, \tag{13b}$$

with $\varepsilon > 0$, $r \in L^\infty(\Omega)$, $\mathbf{w} \in [W^{1,\infty}(\Omega)]^d$, and $f \in L^2(\Omega)$. We assume that $\mathbf{w}$ is divergence-free with piecewise polynomial components and that r is piecewise constant taking nonnegative values. We introduce the bilinear form $\mathscr{B} := \mathscr{B}_{\mathrm{S}} + \mathscr{B}_{\mathrm{A}}$ on $H_0^1(\Omega) \times H_0^1(\Omega)$ such that

$$\mathscr{B}_{\mathrm{S}}(p, \varphi) := \varepsilon(\nabla p, \nabla \varphi) + (r p, \varphi), \tag{14a}$$

$$\mathscr{B}_{\mathrm{A}}(p, \varphi) := -(\mathbf{w} p, \nabla \varphi). \tag{14b}$$

The weak formulation consists in finding $p \in H_0^1(\Omega)$ such that

$$\mathscr{B}(p,\varphi) = (f,\varphi) \qquad \forall \varphi \in H_0^1(\Omega). \tag{15}$$

The vector-valued functions $\mathbf{t} := -\varepsilon \nabla p$ and $\mathbf{q} := \mathbf{w} p$ are in $\mathbf{H}(\mathrm{div},\Omega)$ and are, respectively, called the *diffusive* and *convective flux*.

4.1 Abstract framework

We present here a unified abstract framework for a posteriori error estimation in problem (13a)–(13b). Extending the above bilinear forms to $H^1(\mathscr{T}_h) \times H^1(\mathscr{T}_h)$ using broken gradients, we now define the energy (semi-)norm as

$$|||v||| := \mathscr{B}_{\mathrm{S}}(v,v)^{1/2} = \left(\|\varepsilon^{1/2}\nabla_h v\|^2 + \|r^{1/2}v\|^2\right)^{1/2} \qquad \forall v \in H^1(\mathscr{T}_h). \tag{16}$$

To achieve robustness of the a posteriori error estimates in the singularly perturbed regime resulting from dominant convection, we introduce, following Verfürth [26], the augmented (semi-)norm defined as

$$|||v|||_{\oplus} := |||v||| + \sup_{\varphi \in H_0^1(\Omega),\, |||\varphi|||=1} \mathscr{B}_{\mathrm{A}}(v,\varphi) \qquad \forall v \in H^1(\mathscr{T}_h). \tag{17}$$

The a posteriori error estimate for $|||p - p_h|||_{\oplus}$ is formulated in terms of a *potential reconstruction* s_h, a *diffusive flux reconstruction* $\mathbf{t}_h$, and a *convective flux reconstruction* $\mathbf{q}_h$. These reconstructions must comply with the following assumption.

Assumption 6 (Potential and flux reconstruction for (13a)–(13b)**)** *There holds* $s_h \in H_0^1(\Omega)$, $\mathbf{t}_h, \mathbf{q}_h \in \mathbf{H}(\mathrm{div},\Omega)$, *and*

$$(\nabla{\cdot}\mathbf{t}_h + \nabla{\cdot}\mathbf{q}_h + r p_h, 1)_K = (f,1)_K \qquad \forall K \in \mathscr{T}_h. \tag{18}$$

We can now state and prove our main result concerning the error upper bound. For simplicity, we assume that the mesh $\mathscr{T}_h$ is matching and simplicial so as to use the trace inequality (2). The general case can be treated by resorting to a matching simplicial submesh.

Theorem 7 (A posteriori estimate for (13a)–(13b)**).** *Let p be the solution of* (15) *and let $p_h \in H^1(\mathscr{T}_h)$ be arbitrary. Let Assumption 6 be satisfied. Assume that $\mathscr{T}_h$ is matching and simplicial. Then,*

$$|||p - p_h|||_{\oplus} \le \eta := 2\left\{\sum_{K\in\mathscr{T}_h} \eta_{\mathrm{NC},K}^2\right\}^{1/2} + \left\{\sum_{K\in\mathscr{T}_h} \widetilde{\eta}_{\mathrm{NC},K}^2\right\}^{1/2}$$
$$+\, 3\left\{\sum_{K\in\mathscr{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{CDF},K})^2\right\}^{1/2}.$$

For all $K \in \mathscr{T}_h$, *the* convective-diffusive flux estimator *is given by*

$$\eta_{\mathrm{CDF},K} := \min(\eta_{\mathrm{CDF},1,K}, \eta_{\mathrm{CDF},2,K}), \tag{19a}$$

$$\eta_{\mathrm{CDF},1,K} := \varepsilon^{-1/2} \|\mathbf{a}_h\|_K, \tag{19b}$$

$$\eta_{\mathrm{CDF},2,K} := m_K \|(I - \Pi_0)\nabla{\cdot}\mathbf{a}_h\|_K + \widetilde{m}_K^{1/2} \sum_{\sigma \in \mathscr{E}_K} C_{\mathrm{t},K,\sigma}^{1/2} \|\mathbf{a}_h{\cdot}\mathbf{n}_\sigma\|_\sigma, \tag{19c}$$

with $\mathbf{a}_h := \mathbf{t}_h + \mathbf{q}_h + \varepsilon\nabla_h p_h - \mathbf{w}s_h$ *and* Π_0 *the* L^2*-orthogonal projector onto* $\mathbb{P}_0(\mathscr{T}_h)$*, the* nonconformity estimators *by*

$$\eta_{\mathrm{NC},K} := |||p_h - s_h|||_K, \qquad \widetilde{\eta}_{\mathrm{NC},K} := \min\left(\widetilde{\eta}_{\mathrm{NC},1,K}, \widetilde{\eta}_{\mathrm{NC},2,K}\right), \tag{20a}$$

$$\widetilde{\eta}_{\mathrm{NC},1,K} := \varepsilon^{-1/2} \|\mathbf{b}_h\|_K, \tag{20b}$$

$$\widetilde{\eta}_{\mathrm{NC},2,K} := m_K \|(I - \Pi_0)\nabla{\cdot}\mathbf{b}_h\|_K + \widetilde{m}_K^{1/2} \sum_{\sigma \in \mathscr{E}_K} C_{\mathrm{t},K,\sigma}^{1/2} \|\mathbf{b}_h{\cdot}\mathbf{n}_\sigma\|_\sigma, \tag{20c}$$

with $\mathbf{b}_h := \mathbf{w}(p_h - s_h)$*, and the* residual estimator *by*

$$\eta_{\mathrm{R},K} := m_K \|f - \nabla{\cdot}\mathbf{t}_h - \nabla{\cdot}\mathbf{q}_h - r p_h\|_K. \tag{21}$$

Here $m_K := \min(C_{\mathrm{P},K}^{1/2} \varepsilon^{-1/2} h_K, r_K^{-1/2})$ *and* $\widetilde{m}_K := 2(1 + C_{\mathrm{P},K}^{1/2})\varepsilon^{-1/2} m_K$.

Proof. Following [27, Lemma 7.1] and [8, Lemma 3.1], we infer

$$|||p - p_h||| \le |||p_h - s_h||| + \sup_{\varphi \in H_0^1(\Omega), |||\varphi|||=1} \left\{\mathscr{B}(p - p_h, \varphi) + \mathscr{B}_{\mathrm{A}}(p_h - s_h, \varphi)\right\},$$

and proceeding as in [9, Lemma 4.2] yields

$$\begin{aligned} |||p - p_h|||_{\oplus} \le 2|||p_h - s_h||| &+ \sup_{\varphi \in H_0^1(\Omega), |||\varphi|||=1} \mathscr{B}_{\mathrm{A}}(p_h - s_h, \varphi) \\ &+ 3 \sup_{\varphi \in H_0^1(\Omega), |||\varphi|||=1} \left\{\mathscr{B}(p - p_h, \varphi) + \mathscr{B}_{\mathrm{A}}(p_h - s_h, \varphi)\right\}. \end{aligned}$$

For the second term on the right-hand side, we obtain

$$\mathscr{B}_{\mathrm{A}}(p_h - s_h, \varphi) = -(\mathbf{b}_h, \nabla\varphi) \le \sum_{K \in \mathscr{T}_h} \widetilde{\eta}_{\mathrm{NC},K} |||\varphi|||_K.$$

Indeed, for all $K \in \mathscr{T}_h$, the Cauchy–Schwarz inequality on the one hand yields $-(\mathbf{b}_h, \nabla\varphi)_K \le \varepsilon^{-1/2}\|\mathbf{b}_h\|_K |||\varphi|||_K = \widetilde{\eta}_{\mathrm{NC},1,K} |||\varphi|||_K$, while integrating by parts on K leads to

$$-(\mathbf{b}_h, \nabla\varphi)_K = ((I - \Pi_0)\nabla\cdot\mathbf{b}_h, \varphi - \varphi_K)_K - \sum_{\sigma\in\mathscr{E}_K} (\mathbf{b}_h\cdot\mathbf{n}_\sigma, \varphi - \varphi_K)_\sigma \leq \widetilde{\eta}_{\mathrm{NC},2,K}|||\varphi|||_K,$$

owing to the Poincaré inequality (1) and the trace inequality (2). For the third term on the right-hand side, we observe that

$$\begin{aligned}\mathscr{B}(p - p_h, \varphi) + \mathscr{B}_{\mathrm{A}}(p_h - s_h, \varphi) &= (f - \nabla\cdot\mathbf{t}_h - \nabla\cdot\mathbf{q}_h - rp_h, \varphi) - (\mathbf{a}_h, \nabla\varphi) \\ &\leq \sum_{K\in\mathscr{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{CDF},K})|||\varphi|||_K,\end{aligned}$$

using Assumption 6 for the residual term and proceeding for $\mathbf{a}_h$ as for $\mathbf{b}_h$. □

We now address the efficiency of the estimate of Theorem 7. In what follows, $\lesssim$ can include factors depending on the maximal ratio m_K/m_L for K, L having a nonempty intersection. We introduce the *classical residual estimators* for problem (13a)–(13b) given by

$$\eta_{\mathrm{res},K} := m_K\|f + \nabla\cdot(\varepsilon\nabla p_h - \mathbf{w}p_h) - rp_h\|_{\mathfrak{T}_K} + m_K^{1/2}\varepsilon^{-1/4}\|[\![\varepsilon\nabla_h p_h]\!]\cdot\mathbf{n}\|_{\mathfrak{E}_K^{\mathrm{int}}}, \tag{22a}$$

$$|p_h|_{\mathrm{J},K} := (\varepsilon^{1/2}h_K^{-1/2} + m_K^{1/2}\varepsilon^{-1/4}\|\mathbf{w}\|_{[L\infty(K)]^d} + r_K^{1/2}h_K^{1/2})\|[\![p_h]\!]\|_{\mathfrak{E}_K}. \tag{22b}$$

We also set $|v|_{\mathrm{J}} := \left\{\sum_{K\in\mathscr{T}_h} |v|^2_{\mathrm{J},K}\right\}^{1/2}$ for all $v \in H^1(\mathscr{T}_h)$.

Assumption 8 (Approximation property for (13a)–(13b)) *We assume that, for all $K \in \mathscr{T}_h$, with $\mathbf{c}_h = \mathbf{a}_h$ or $\mathbf{b}_h$,*

$$m_K\|(I - \Pi_0)\nabla\cdot\mathbf{c}_h\|_K + m_K^{1/2}\varepsilon^{-1/4}\sum_{\sigma\in\mathscr{E}_K} \|\mathbf{c}_h\cdot\mathbf{n}_\sigma\|_\sigma \lesssim \eta_{\mathrm{res},K} + |p_h|_{\mathrm{J},K}.$$

Proceeding as in [9, Theorems 3.2 and 3.4] leads to the following lower bound, which is global in space owing to the use of a dual norm.

Theorem 9 (Efficiency of the estimate of Theorem 7). *Let p be the solution of (15) and let Assumption 8, and the second item of Assumption 3, be satisfied. Then,*

$$\eta \lesssim |||p - p_h|||_\oplus + |p - p_h|_{\mathrm{J}}. \tag{23}$$

Remark 3 (Fully robust equivalence result). Adding the jump seminorm $|\cdot|_{\mathrm{J}}$ to the error measure, a fully robust equivalence result is finally achieved in the form

$$|||p - p_h|||_\oplus + |p - p_h|_{\mathrm{J}} \leq \eta + |p_h|_{\mathrm{J}} \lesssim |||p - p_h|||_\oplus + |p - p_h|_{\mathrm{J}}. \tag{24}$$

4.2 Application to finite volumes

We apply here the framework of §4.1 to cell- and vertex-centered finite volume schemes, i.e., we specify s_h, $\mathbf{t}_h$, and $\mathbf{q}_h$, and we verify Assumption 6, and, at least in some cases, Assumption 8.

4.2.1 Cell-centered finite volumes

Definition 3 (Cell-centered FVs for (13a)–(13b)). A cell-centered FV scheme for discretizing (13a)–(13b), cf. [12], reads: find $\bar{p}_h \in \mathbb{P}_0(\mathcal{T}_h)$ such that

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + \sum_{\sigma\in\mathcal{E}_K} W_{K,\sigma} + r_K \bar{p}_h|_K = (f,1)_K \qquad \forall K \in \mathcal{T}_h. \tag{25}$$

In addition to the diffusive fluxes $F_{K,\sigma}$, $W_{K,\sigma}$ are the convective fluxes, also depending on $\bar{p}_h$. We do not need the precise form of the fluxes, but only $F_{K,\sigma} = -F_{L,\sigma}$ and $W_{K,\sigma} = -W_{L,\sigma}$ for all interior sides σ shared by the elements K and L.

Following the ideas exposed in §3.2.1, we first define $\mathbf{t}_h, \mathbf{q}_h \in \mathbf{RTN}(\mathcal{T}_h)$ by

$$(\mathbf{t}_h|_K\cdot\mathbf{n}_K)|_\sigma := F_{K,\sigma}/|\sigma|, \qquad (\mathbf{q}_h|_K\cdot\mathbf{n}_K)|_\sigma := W_{K,\sigma}/|\sigma|. \tag{26}$$

Define p_h similarly to (11). It is immediate to see using the Green theorem that (26) and (25) yield (18). A reasonable condition on $W_{K,\sigma}$ in the context of upwind or centered convective fluxes is that

$$\|W_{K,\sigma}/|\sigma| - \mathbf{w}\cdot\mathbf{n}_K p_h|_K\|_\sigma \lesssim \|\mathbf{w}\|_{[L^\infty(K)]^d} \|[\![\bar{p}_h]\!]\|_\sigma. \tag{27}$$

Then, Assumption 8 holds, up to the oscillation terms $m_K\|(I - \Pi_0)\nabla\cdot(\mathbf{w}p_h)\|_K$, when additionally including $|\bar{p}_h|_{\mathrm{J},K}$ on the right-hand side, and the efficiency result (23) holds when additionally including $|p - \bar{p}_h|_{\mathrm{J}}$ on the right-hand side.

4.2.2 Vertex-centered finite volumes

Definition 4 (Vertex-centered FVs for (13a)–(13b)). A vertex-centered FV scheme for discretizing (13a)–(13b), cf. [12], reads: find $p_h \in \mathbb{P}_1(\mathcal{T}_h) \cap H_0^1(\Omega)$ such that

$$-\langle\varepsilon\nabla p_h\cdot\mathbf{n}_D, 1\rangle_{\partial D} + \langle\mathbf{w}\cdot\mathbf{n}_D p_h, 1\rangle_{\partial D} + (rp_h, 1)_D = (f,1)_D \qquad \forall D \in \mathcal{D}_h^{\mathrm{int}}. \tag{28}$$

Note that we only consider a centered convective flux.

As in §3.2.2, we set $s_h = p_h$ in Assumption 6. Consequently, $\eta_{\mathrm{NC},K} = \widetilde{\eta}_{\mathrm{NC},K} = 0$ in Theorem 7. For the convective flux reconstruction, we simply set $\mathbf{q}_h := \mathbf{w} p_h$. For the diffusive flux reconstruction, we introduce the mesh $\mathscr{S}_h$ (cf. Fig. 1, right) and we define $\mathbf{t}_h \in \mathbf{RTN}(\mathscr{S}_h)$ such that $\mathbf{t}_h \cdot \mathbf{n}_\sigma := -\varepsilon \nabla p_h \cdot \mathbf{n}_\sigma$ at all interior sides σ of $\mathscr{S}_h$ which lie on the boundary of some $D \in \mathscr{D}_h$. As in §3.2.2, local problems can be solved to fulfill Assumption 6. Assumption 8 can be verified as in §3.2.2 for the diffusive part, while the convective part is trivial owing to the choice of $\mathbf{q}_h$.

5 Heat equation

We consider the heat equation

$$\partial_t p - \Delta p = f \quad \text{in } \Omega \times (0,T), \tag{29a}$$

$$p = 0 \quad \text{on } \partial\Omega \times (0,T), \tag{29b}$$

$$p(\cdot,0) = p_0 \quad \text{in } \Omega, \tag{29c}$$

with $f \in L^2(\Omega \times (0,T))$, initial condition $p_0 \in L^2(\Omega)$, and final time $T > 0$. The exact solution is in the space $Y := \{y \in X; \partial_t y \in X'\}$, with $X := L^2(0,T;H_0^1(\Omega))$ and $X' = L^2(0,T;H^{-1}(\Omega))$, satisfies (29c) in $L^2(\Omega)$, and is such that, for a.e. $t \in (0,T)$,

$$\langle \partial_t p, \varphi \rangle(t) + (\nabla p, \nabla \varphi)(t) = (f,\varphi)(t) \qquad \forall \varphi \in H_0^1(\Omega). \tag{30}$$

The space-time energy norm is given by $\|y\|_X := \left\{\int_0^T \|\nabla y\|^2(t)\,\mathrm{d}t\right\}^{1/2}$ for all $y \in X$. Following Verfürth [25], we augment the energy norm by a dual norm of the time derivative as $\|y\|_Y := \|y\|_X + \|\partial_t y\|_{X'}$ with $\|\partial_t y\|_{X'} := \left\{\int_0^T \|\partial_t y\|_{H^{-1}}^2(t)\,\mathrm{d}t\right\}^{1/2}$.

5.1 Abstract framework

We consider an increasing sequence of discrete times $\{t^n\}_{0 \le n \le N}$ such that $t^0 = 0$ and $t^N = T$ and introduce the time intervals $I_n := (t^{n-1}, t^n]$ and the time steps $\tau^n := t^n - t^{n-1}$ for all $1 \le n \le N$. The meshes are allowed to vary in time; we denote by $\mathscr{T}_h^n$ the mesh used to march in time from t^{n-1} to t^n, for all $1 \le n \le N$, and by $\mathscr{T}_h^0$ the initial mesh. We suppose that the approximate solution on t^n, denoted by $p_{h\tau}^n$, is in $H^1(\mathscr{T}_h^n)$, and we let $p_{h\tau}$ be the space-time approximate solution, given by $p_{h\tau}^n$ at each discrete time t^n and piecewise affine and continuous in time. We denote the space of such functions by $P_\tau^1(H^1(\mathscr{T}_h))$. We also denote by $P_\tau^1(H_0^1(\Omega))$ the space of functions that are piecewise affine and continuous in time and $H_0^1(\Omega)$ in space and by $P_\tau^0(\mathbf{H}(\mathrm{div},\Omega))$ the space of functions that are piecewise constant in time and $\mathbf{H}(\mathrm{div}, \Omega)$ in space. For all $1 \le n \le N$, we set $\widetilde{f}^n := \frac{1}{\tau^n} \int_{I_n} f(\cdot, t)\,\mathrm{d}t$, and, for $\varphi_{h\tau} \in P^1_\tau(H^1(\mathscr{T}_h))$, $\partial_t p^n_{h\tau} := \frac{1}{\tau^n}(\varphi^n_{h\tau} - \varphi^{n-1}_{h\tau})$.

We aim at measuring the error $(p - p_{h\tau})$ in the $\|\cdot\|_Y$-norm using the broken gradient operator in the energy norm. The a posteriori error estimate is formulated in terms of a *space-time potential reconstruction* $s_{h\tau}$ and a *space-time flux reconstruction* $\mathbf{t}_{h\tau}$. These reconstructions must comply with the following assumption.

Assumption 10 (Potential and flux reconstruction for (29a)–(29c)**)** *There holds* $s_{h\tau} \in P^1_\tau(H^1_0(\Omega))$, $\mathbf{t}_{h\tau} \in P^0_\tau(\mathbf{H}(\mathrm{div}, \Omega))$, *and, for all* $1 \le n \le N$ *and for all* $K \in \mathscr{T}^n_h$,

$$(\partial_t s^n_{h\tau}, 1)_K = (\partial_t p^n_{h\tau}, 1)_K, \tag{31a}$$

$$(\widetilde{f}^n - \partial_t p^n_{h\tau} - \nabla\cdot\mathbf{t}^n_{h\tau}, 1)_K = 0. \tag{31b}$$

We can now state our main result concerning the error upper bound, see [11, Theorem 3.6] and also [11, Theorem 3.2] for a slightly sharper bound.

Theorem 11 (A posteriori estimate for (29a)–(29c)**).** *Let* p *be the solution of* (30) *and let* $p_{h\tau} \in P^1_\tau(H^1(\mathscr{T}_h))$ *be arbitrary. Let Assumption 10 be satisfied. Then,*

$$\|p - p_{h\tau}\|_Y \le \left\{\sum_{n=1}^{N} (\eta^n_{\mathrm{sp}})^2\right\}^{1/2} + \left\{\sum_{n=1}^{N} (\eta^n_{\mathrm{tm}})^2\right\}^{1/2} + \eta_{\mathrm{IC}} + 3\|f - \widetilde{f}\|_{X'}, \tag{32}$$

with, for all $1 \le n \le N$, *the* space *and* time error estimators *given by*

$$(\eta^n_{\mathrm{sp}})^2 := \sum_{K \in \mathscr{T}^n_h} 3 \left\{ \tau^n (9(\eta^n_{\mathrm{R},K} + \eta^n_{\mathrm{DF},K})^2 + (\eta^n_{\mathrm{NC},2,K})^2) + \int_{I_n} (\eta^n_{\mathrm{NC},1,K})^2(t)\,\mathrm{d}t \right\}, \tag{33a}$$

$$(\eta^n_{\mathrm{tm}})^2 := \sum_{K \in \mathscr{T}^n_h} 3\tau^n \|\nabla(s^n_{h\tau} - s^{n-1}_{h\tau})\|^2_K. \tag{33b}$$

For all $K \in \mathscr{T}^n_h$, *the* residual estimator*, the* diffusive flux estimator*, and the* nonconformity estimators *are given by*

$$\eta^n_{\mathrm{R},K} := C^{1/2}_{\mathrm{P},K} h_K \|\widetilde{f}^n - \partial_t s^n_{h\tau} - \nabla\cdot\mathbf{t}^n_{h\tau}\|_K, \tag{34a}$$

$$\eta^n_{\mathrm{DF},K} := \|\nabla s^n_{h\tau} + \mathbf{t}^n_{h\tau}\|_K, \tag{34b}$$

$$\eta^n_{\mathrm{NC},1,K}(t) := \|\nabla^n_h(s_{h\tau} - p_{h\tau})(t)\|_K, \quad \forall t \in I_n, \tag{34c}$$

$$\eta^n_{\mathrm{NC},2,K} := C^{1/2}_{\mathrm{P},K} h_K \|\partial_t(s_{h\tau} - p_{h\tau})^n\|_K. \tag{34d}$$

Finally, the initial condition estimator *is given by* $\eta_{\mathrm{IC}} := 2^{1/2}\|s^0_{h\tau} - p_0\|$.

We next turn to the efficiency of the estimate of Theorem 11. We introduce the *classical residual estimators* for problem (29a)–(29c) given by

$$\eta^n_{\mathrm{res},K} := h_K \|\widetilde{f}^n - \partial_t p^n_{h\tau} + \Delta p^n_{h\tau}\|_{\mathfrak{T}_K} + h_K^{1/2} \|[\![\nabla^n_h p^n_{h\tau} \cdot \mathbf{n}]\!]\|_{\mathfrak{E}^{\mathrm{int}}_K}, \tag{35a}$$

$$|p^n_{h\tau}|_{\mathrm{J},K} := h_K^{-1/2} \|[\![p^n_{h\tau}]\!]\|_{\mathfrak{E}_K}. \tag{35b}$$

Assumption 12 (Approximation property for (29a)–(29c)) *We assume that for all $1 \le n \le N$ and for all $K \in \mathcal{T}^n_h$,*

$$\|\nabla^n_h (p^n_{h\tau} - s^n_{h\tau})\|_K + \|\nabla^n_h p^n_{h\tau} + \mathbf{t}^n_{h\tau}\|_K \lesssim \eta^n_{\mathrm{res},K} + |p^n_{h\tau}|_{\mathrm{J},K}. \tag{36}$$

We can now state our efficiency result, see [11, Theorem 3.9]. As in [25], the lower bound is local in time, but global in space.

Theorem 13 (Efficiency of the estimate of Theorem 11). *Let Assumption 12 hold, let Assumption 3 hold at all discrete times, let both the refinement and coarsening in time be not too abrupt, and let, for all $1 \le n \le N$, $(h^n)^2 \lesssim \tau^n$. Then, for all $1 \le n \le N$,*

$$\eta^n_{\mathrm{sp}} + \eta^n_{\mathrm{tm}} \lesssim \|p - p_{h\tau}\|_{Y(I_n)} + \mathcal{J}^n(p_{h\tau}) + \|f - \widetilde{f}\|_{X'(I_n)}, \tag{37}$$

where $\mathcal{J}^n(p_{h\tau}) := \left\{\tau^n \sum_{K \in \mathcal{T}^{n-1}_h} |p^{n-1}_{h\tau}|^2_{\mathrm{J},K} + \tau^n \sum_{K \in \mathcal{T}^n_h} |p^n_{h\tau}|^2_{\mathrm{J},K}\right\}^{1/2}$.

Remark 4 (Equivalence result). We refer to [11, Remark 3.10] for bounding the jumps $\mathcal{J}^n(p_{h\tau})$, see also Remark 2.

5.2 *Application to finite volumes*

We apply here the framework of §5.1 to cell- and vertex-centered finite volume schemes, i.e., we specify $s_{h\tau}$ and $\mathbf{t}_{h\tau}$, and we verify Assumptions 10 and 12. For simplicity, we only discuss matching simplicial meshes.

5.2.1 Cell-centered finite volumes

Definition 5 (Cell-centered FVs for (29a)–(29c)). A cell-centered FV scheme for (29a)–(29c), cf. [12], reads: for all $1 \le n \le N$, find $\bar{p}^n_{h\tau} \in \mathbb{P}_0(\mathcal{T}^n_h)$ s. t.

$$\frac{1}{\tau^n}(\bar{p}^n_{h\tau} - p^{n-1}_{h\tau}, 1)_K + \sum_{\sigma \in \mathcal{E}_K} F^n_{K,\sigma} = (\widetilde{f}^n, 1)_K \qquad \forall K \in \mathcal{T}^n_h. \tag{38}$$

As in §3.2.1, the fluxes $\mathbf{t}^n_{h\tau}$ are constructed from the side fluxes $F^n_{K,\sigma}$ by an equivalent of (10). An elementwise postprocessing as (11) is applied to obtain $p^n_{h\tau}$ from $\bar{p}^n_{h\tau}$. The potential is reconstructed at each discrete time from a modification of the averaging operator of §3.1 where local bubble functions are used to satisfy (31a) (cf. [11]). Then, owing to the construction of $\mathbf{t}^n_{h\tau}$, (31b) is also satisfied, whence Assumption 10 follows. Finally, we set $\mathscr{S}^n_h = \mathscr{T}^n_h$; Assumption 12 is trivial for $\mathbf{t}_{h\tau}$ since $\|\nabla^n_h p^n_{h\tau} + \mathbf{t}^n_{h\tau}\|_K = 0$ and is proven for $s^n_{h\tau}$ in [11].

5.2.2 Vertex-centered finite volumes

Definition 6 (Vertex-centered FVs for (29a)–(29c)). A vertex-centered FV scheme for (29a)–(29c), cf. [12], reads: for all $1 \le n \le N$, find $p^n_{h\tau} \in \mathbb{P}_1(\mathscr{T}^n_h) \cap H^1_0(\Omega)$ s. t.

$$(\partial_t p^n_{h\tau}, 1)_D - \langle \nabla p^n_{h\tau} \cdot \mathbf{n}_D, 1 \rangle_{\partial D} = (\widetilde{f}^n, 1)_D \qquad \forall D \in \mathscr{D}^{\mathrm{int},n}_h. \tag{39}$$

As in §3.2.2, $p^n_{h\tau} \in H^1_0(\Omega)$ for all $1 \le n \le N$, so that we set $s^n_{h\tau} = p^n_{h\tau}$. Consequently, $\eta^n_{\mathrm{NC},1,K} = \eta^n_{\mathrm{NC},2,K} = 0$ in Theorem 11. The fluxes $\mathbf{t}_{h\tau}$ are constructed as in §3.2.2, using the simplicial submeshes $\mathscr{S}^n_h$. Assumptions 10 and 12 are then verified by proceeding as in §3.2.2.

Acknowledgements This work was partly supported by the Groupement MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EdF, IRSN).

References

1. Achdou, Y., Bernardi, C., Coquel, F.: A priori and a posteriori analysis of finite volume discretizations of Darcy's equations. Numer. Math. **96**(1), 17–42 (2003)
2. Babuška, I., Rheinboldt, W.C.: Error estimates for adaptive finite element computations. SIAM J. Numer. Anal. **15**(4), 736–754 (1978)
3. Brezzi, F., Fortin, M.: Mixed and hybrid finite element methods, *Springer Series in Computational Mathematics*, vol. 15. Springer-Verlag, New York (1991)
4. Burman, E., Ern, A.: Continuous interior penalty hp-finite element methods for advection and advection-diffusion equations. Math. Comp. **76**(259), 1119–1140 (2007)
5. Carstensen, C., Funken, S.A.: Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods. East-West J. Numer. Math. **8**(3), 153–175 (2000)
6. Dari, E., Durán, R., Padra, C., Vampa, V.: A posteriori error estimators for nonconforming finite element methods. RAIRO Modél. Math. Anal. Numér. **30**(4), 385–400 (1996)
7. El Alaoui, L., Ern, A., Vohralík, M.: Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. Comput. Methods Appl. Mech. Engrg. (2010). DOI 10.1016/j.cma.2010.03.024
8. Ern, A., Stephansen, A.F.: A posteriori energy-norm error estimates for advection-diffusion equations approximated by weighted interior penalty methods. J. Comp. Math. **26**(4), 488–510 (2008)

9. 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)
10. Ern, A., Vohralík, M.: Flux reconstruction and a posteriori error estimation for discontinuous Galerkin methods on general nonmatching grids. C. R. Math. Acad. Sci. Paris **347**(7-8), 441–444 (2009)
11. Ern, A., Vohralík, M.: A posteriori error estimation based on potential and flux reconstruction for the heat equation. SIAM J. Numer. Anal. **48**(1), 198–223 (2010)
12. Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Handbook of Numerical Analysis, Vol. VII, pp. 713–1020. North-Holland, Amsterdam (2000)
13. Eymard, R., Gallouët, T., Herbin, R.: Finite volume approximation of elliptic problems and convergence of an approximate gradient. Appl. Numer. Math. **37**(1-2), 31–53 (2001)
14. Haslinger, J., Hlaváček, I.: Convergence of a finite element method based on the dual variational formulation. Apl. Mat. **21**(1), 43–65 (1976)
15. Jiránek, P., Strakoš, Z., Vohralík, M.: A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. SIAM J. Sci. Comput. **32**(3), 1567–1590 (2010)
16. Karakashian, O.A., Pascal, F.: A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems. SIAM J. Numer. Anal. **41**(6), 2374–2399 (2003)
17. Kim, K.Y.: A posteriori error analysis for locally conservative mixed methods. Math. Comp. **76**(257), 43–66 (2007)
18. Ladevèze, P.: Comparaison de modèles de milieux continus. Ph.D. thesis, Université Pierre et Marie Curie (Paris 6) (1975)
19. Nicaise, S.: A posteriori error estimations of some cell-centered finite volume methods. SIAM J. Numer. Anal. **43**(4), 1481–1503 (2005)
20. Ohlberger, M.: A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection–diffusion equations. Numer. Math. **87**(4), 737–761 (2001)
21. Ohlberger, M.: A posteriori error estimates for vertex centered finite volume approximations of convection–diffusion–reaction equations. M2AN Math. Model. Numer. Anal. **35**(2), 355–387 (2001)
22. Prager, W., Synge, J.L.: Approximations in elasticity based on the concept of function space. Quart. Appl. Math. **5**, 241–269 (1947)
23. Stephansen, A.F.: Méthodes de Galerkine discontinues et analyse d'erreur a posteriori pour les problèmes de diffusion hétérogène. Ph.D. thesis, Ecole Nationale des Ponts et Chaussées (2007)
24. Verfürth, R.: A review of a posteriori error estimation and adaptive mesh-refinement techniques. Teubner-Wiley, Stuttgart (1996)
25. Verfürth, R.: A posteriori error estimates for finite element discretizations of the heat equation. Calcolo **40**(3), 195–212 (2003)
26. Verfürth, R.: Robust a posteriori error estimates for stationary convection-diffusion equations. SIAM J. Numer. Anal. **43**(4), 1766–1782 (2005)
27. Vohralík, M.: A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. SIAM J. Numer. Anal. **45**(4), 1570–1599 (2007)
28. Vohralík, M.: Residual flux-based a posteriori error estimates for finite volume and related locally conservative methods. Numer. Math. **111**(1), 121–158 (2008)
29. Vohralík, M.: Two types of guaranteed (and robust) a posteriori estimates for finite volume methods. In: Finite Volumes for Complex Applications V, pp. 649–656. ISTE and John Wiley & Sons, London, UK and Hoboken, USA (2008)
30. Vohralík, M.: Guaranteed and fully robust a posteriori error estimates for conforming discretizations of diffusion problems with discontinuous coefficients. J. Sci. Comput. **46**, 397–438 (2011)

The paper is in final form and has not been or is not being submitted elsewhere.

Staggered discretizations, pressure correction schemes and all speed barotropic flows

L. Gastaldo, R. Herbin, W. Kheriji, C. Lapuerta, and J.-C. Latché

Abstract We present in this paper a class of schemes for the solution of the barotropic Navier-Stokes equations. These schemes work on general meshes, preserve the stability properties of the continuous problem, irrespectively of the space and time steps, and boil down, when the Mach number vanishes, to discretizations which are standard (and stable) in the incompressible framework. Finally, we show that they are able to capture solutions with shocks to the Euler equations.

Keywords barotropic Navier-Stokes, staggered discretizations
MSC2010: 65M12

1 Introduction

The problem addressed in this paper is the system of the so-called barotropic compressible Navier-Stokes equations, which reads:

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

$$\partial_t (\bar\rho \bar u) + \operatorname{div}(\bar\rho \bar u \otimes \bar u) + \nabla \bar p - \operatorname{div}(\tau(\bar u)) = 0, \tag{1b}$$

$$\bar\rho = \wp(\bar p), \tag{1c}$$

where t stands for the time, $\bar\rho$, $\bar u$ and $\bar p$ are the density, velocity and pressure in the flow, and $\tau(\bar u)$ stands for the shear stress tensor. The function $\wp(\cdot)$ is the equation

L. Gastaldo, W. Kheriji, C. Lapuerta, and J.-C. Latché
Institut de Radioprotection et de Sûreté Nucléaire (IRSN)
e-mail: [laura.gastaldo,walid.kheriji, celine.lapuerta, jean-claude.latche]@irsn.fr

R. Herbin
Université de Provence
e-mail: herbin@cmi.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_86, © Springer-Verlag Berlin Heidelberg 2011

of state used for the modelling of the particular flow at hand, which may be the actual equation of state of the fluid or may result from assumptions concerning the flow; typically, laws as $\wp(\bar{p}) = \bar{p}^{1/\gamma}$, where $\gamma > 1$ is a coefficient which is specific to the considered fluid, are obtained by making the assumption that the flow is isentropic. This system of equations is posed over $\Omega \times (0, T)$, where Ω is a domain of $\mathbb{R}^d$, $d \leq 3$ supposed to be polygonal ($d = 2$) or polyhedral ($d = 3$), and the final time T is finite. We suppose that the boundary of Ω is split into $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, and we suppose that the velocity and density are prescribed on $\partial\Omega_D$, while Neumann boundary conditions are prescribed on $\partial\Omega_N$. The flow is assumed to enter the domain through $\partial\Omega_D$ and to leave it through Ω_N. This system must be supplemented by initial conditions for $\bar{\rho}$ and $\bar{u}$.

The objective of this paper is to present a class of schemes which enjoy three essential features. First, these schemes work on quite general two and three dimensional meshes, including locally refined non-conforming (*i.e.* with hanging nodes) discretizations. Second, they respect the (expected) stability properties of the continuous problem at hand, irrespectively of the space and time steps: positivity of the density, conservation of mass, energy inequality. Third, they boil down, for vanishing Mach numbers, to usual stable coupled or pressure correction schemes, which means that the discretization enjoys a discrete *inf-sup* condition. Even if this is beyond the scope of this paper, we remark that this latter property allows a control of the pressure to be obtained through a control of its gradient; this property is used as a central argument to obtain convergence results on model problems [5, 6, 8].

This paper is organized as follows. First, we describe the general form of the schemes (Sect. 2). Then we show how stability requirements are taken into account to design the discretization of the velocity convection term (Sect. 3). The final expression for the schemes is given in Sect. 4, and their stability properties are stated. Finally, we discuss their capability to capture solutions of the Euler equations with shocks (Sect. 5).

2 The schemes: general form

2.1 Meshes and unknowns

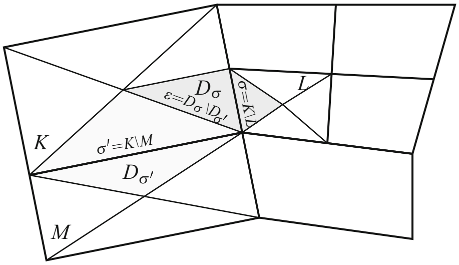

D_σ
$\varepsilon = D_\sigma | D_{\sigma'}$
$\sigma = K|L$
L
K
$\sigma' = K \backslash M$
$D_{\sigma'}$
M

Fig. 1 Notations for primal and dual cells

A finite volume mesh of Ω is defined by a set $\mathcal{M}$ of non–empty convex open disjoint subsets K of Ω (the control volumes), such that $\bar{\Omega} = \bigcup_{K \in \mathcal{M}} \bar{K}$. We denote by $\mathcal{E}$ the set of edges (in 2D) or faces (in 3D), by $\mathcal{E}(K) \subset \mathcal{E}$ the set of faces of the cell $K \in \mathcal{M}$, by $\mathcal{E}_{\text{ext}}$ and $\mathcal{E}_{\text{int}}$ the set of boundary and interior faces, respectively. The set of external faces $\mathcal{E}_{\text{ext}}$ is split in $\mathcal{E}_N$ and $\mathcal{E}_D$, which stand for the set of the faces included in $\partial\Omega_N$ and $\partial\Omega_D$, respectively. Each internal face, denoted by $\sigma \in \mathcal{E}_{\text{int}}$, is supposed to have exactly two neighboring cells, say $K,\ L \in \mathcal{M}$, and $\bar{K} \cap \bar{L} = \bar{\sigma}$ which we denote by $\sigma = K|L$. By analogy, we write $\sigma = K|$ext for an external face σ of K, even if this notation is somewhat incorrect, since K may have more than one external face. The mesh $\mathcal{M}$ will be referred to hereafter as the ”primal mesh”.

The outward normal vector to a face σ of K is denoted by $n_{K,\sigma}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}$, we denote by $|K|$ the measure of K and by $|\sigma|$ the $(d-1)$-measure of the face σ.

Then, for $\sigma \in \mathcal{E}$ and $K \in \mathcal{M}$ such that $\sigma \in \mathcal{E}(K)$ (in fact, the only cell if $\sigma \in \mathcal{E}_{\text{ext}}$ and one among the two possible cells if $\sigma \in \mathcal{E}_{\text{int}}$), we denote by $D_{K,\sigma}$ a subvolume of K having σ as a face (see Fig. 1), and by $|D_{K,\sigma}|$ the measure of $D_{K,\sigma}$. For $\sigma \in \mathcal{E}_{\text{int}},\ \sigma = K|L$, we set $D_\sigma = D_{K,\sigma} \cup D_{L,\sigma}$, so $|D_\sigma| = |D_{K,\sigma}| + |D_{L,\sigma}|$, and for $\sigma \in \mathcal{E}_{\text{ext}}, \sigma = K|$ext, $D_\sigma = D_{K,\sigma}$, so $|D_\sigma| = |D_{K,\sigma}|$. The set of faces of the dual cell D_σ is denoted by $\bar{\mathcal{E}}(D_\sigma)$, and the face separating two adjacent dual cells D_σ and $D_{\sigma'}$ is denoted by $\varepsilon = \sigma|\sigma'$.

For $1 \leq i \leq d$, the degree of freedom for the i^{th} component of the velocity are assumed to be associated to a subset of $\mathcal{E}$, denoted by $\mathcal{E}^{(i)} \subset \mathcal{E}$, and are denoted by:

$$\left\{u_{\sigma,i},\ \sigma \in \mathcal{E}^{(i)}\right\}.$$

The sets of internal, external, Neumann and Dirichlet faces associated to the component i are denoted by $\mathcal{E}^{(i)}_{\text{int}}$, $\mathcal{E}^{(i)}_{\text{ext}}$, $\mathcal{E}^{(i)}_{N}$ and $\mathcal{E}^{(i)}_{D}$ (so, for instance, $\mathcal{E}^{(i)}_{\text{int}} = \mathcal{E}_{\text{int}} \cap \mathcal{E}^{(i)}$). We consider the following assumption:

(H1) for $1 \leq i \leq d,\ \forall K \in \mathcal{M}$,

$$\cup_{\sigma \in \mathcal{E}^{(i)} \cap \mathcal{E}(K)} \overline{D}_{K,\sigma} = \overline{K} \quad \text{and} \quad \sum_{\sigma \in \mathcal{E}^{(i)} \cap \mathcal{E}(K)} |D_{K,\sigma}| = |K|,$$

which means that the volumes $D_{K,\sigma},\ \sigma \in \mathcal{E}^{(i)}$, are disjoint, and that, for $1 \leq i \leq d$, $(D_\sigma)_{\sigma \in \mathcal{E}^{(i)}}$ is a partition of Ω. The sets of faces, internal faces and Neumann faces of this dual mesh are denoted by $\bar{\mathcal{E}}^{(i)}$, $\bar{\mathcal{E}}^{(i)}_{\text{int}}$ and $\bar{\mathcal{E}}^{(i)}_{N}$ respectively.

We suppose that the degrees of freedom for the pressure and the density are associated to the primal cells, so they read

$$\left\{p_K,\ K \in \mathcal{M}\right\}, \quad \left\{\rho_K,\ K \in \mathcal{M}\right\}.$$

We denote by V the approximation space for the velocity, by $V^{(i)}$, $1 \leq i \leq d$, the approximation spaces for the velocity components and by Q the approximation space for the pressure and the density, and we identify the discrete functions to their degrees of freedom:

$$\forall v \in V,\ v_i \in V^{(i)},\ 1 \leq i \leq d \text{ and } v_i = (v_{\sigma,i})_{\sigma \in \mathcal{E}^{(i)}}; \quad \forall q \in Q,\ q = (q_K)_{K \in \mathcal{M}}.$$

For the velocity, since the concerned degrees of freedom are located on the boundary, the Dirichlet boundary conditions are enforced in the approximation space:

$$\text{for } 1 \le i \le d,\ \forall v_i \in V^{(i)},\ \forall \sigma \in \mathcal{E}_D^{(i)}, \quad v_{\sigma,i} = \frac{1}{|\sigma|} \int_\sigma \bar{u}_{D,i}\, \mathrm{d}\gamma,$$

where $\bar{u}_{D,i}$ stands for the i^{th} component of the prescribed velocity.

2.2 The schemes

We now introduce the following notations and assumptions:

- for $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}(K)$, we denote by $u \cdot n_{K,\sigma}$ an approximation of the normal velocity to the face σ outward K,
- for $v \in V$, $1 \le i \le d$ and $\sigma \in \mathcal{E}^{(i)}$, we denote by $(\mathrm{div}\tau(v))_\sigma^{(i)}$ an approximation of the viscous term associated to σ and to the component i, and we suppose that the following assumption is satisfied:

$$\text{(H2)} \qquad \sum_{i=1}^{d} \sum_{\sigma \in \mathcal{E}^{(i)}} |D_\sigma|\, (\mathrm{div}\tau(v))_\sigma^{(i)}\, v_{\sigma,i} \ge 0.$$

- for $q \in Q$, $1 \le i \le d$ and $\sigma \in \mathcal{E}^{(i)}$, we denote by $(\nabla q)_\sigma^{(i)}$ the component i of the discrete gradient of q at the face σ, and we suppose that the following assumption is satisfied for any $q \in Q$ and $v \in V$:

$$\text{(H3)} \qquad \sum_{i=1}^{d} \sum_{\sigma \in \mathcal{E}^{(i)}} |D_\sigma|\, (\nabla q)_\sigma^{(i)}\, v_{\sigma,i} = \sum_{K \in \mathcal{M}} q_K \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, v \cdot n_{K,\sigma}.$$

With these notations, we are able to write the general form of the implicit scheme:

$$\forall K \in \mathcal{M}, \qquad \frac{|K|}{\delta t}(\rho_K - \rho_K^*) + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma} = 0. \tag{2a}$$

$$\text{For } 1 \le i \le d,\ \forall \sigma \in \mathcal{E}_{\mathrm{int}}^{(i)} \cup \mathcal{E}_N^{(i)},$$

$$\frac{|D_\sigma|}{\delta t}(\rho_\sigma u_{\sigma,i} - \rho_\sigma^* u_{\sigma,i}^*) + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon} u_{\varepsilon,i} \tag{2b}$$
$$+ |D_\sigma|\, (\nabla p)_\sigma^{(i)} + |D_\sigma|\, (\mathrm{div}\tau(u))_\sigma^{(i)} = 0,$$

$$\forall K \in \mathcal{M}, \qquad \rho_K = \wp(p_K), \tag{2c}$$

where the * superscript denotes the beginning-of-step quantities, $F_{K,\sigma}$ stands for the mass flux leaving K through σ, ρ_σ stands for an approximation of the density at the face, and $F_{\sigma,\varepsilon}$ is a mass flux leaving D_σ through ε. For the flux $F_{K,\sigma}$ at the internal face $\sigma = K|L$, we choose an upwind approximation of the density:

$$F_{K,\sigma} = |\sigma|\, u \cdot n_{K,\sigma}\, \rho_\sigma^{\mathrm{up}}, \quad \text{with } \rho_\sigma^{\mathrm{up}} = \rho_K \text{ if } F_{K,\sigma} \geq 0,\ \rho_\sigma^{\mathrm{up}} = \rho_L \text{ otherwise.} \tag{3}$$

On $\sigma \in \mathcal{E}_D$, the density $\rho_\sigma^{\mathrm{up}}$ is given by the boundary condition, and, on $\sigma \in \mathcal{E}_N$, $\sigma = K|\mathrm{ext}$, $\rho_\sigma^{\mathrm{up}} = \rho_K$, which is indeed an upwind choice, since the flow is supposed to enter the domain through $\partial\Omega_D$ and to leave it through $\partial\Omega_N$. For the velocity components at the dual faces, $u_{\varepsilon,i}$, we choose either the centred or upwind approximation on the internal faces, and the value at the face for the outflow faces.

A pressure correction scheme is obtained from (2) by splitting the resolution in two steps:

1- Velocity prediction step – Solve for $\tilde{u} \in V$ the momentum balance equation with the beginning-of-step pressure:

$$\begin{aligned}&\text{For } 1 \leq i \leq d,\ \forall \sigma \in \mathcal{E}_{\mathrm{int}}^{(i)} \cup \mathcal{E}_N^{(i)},\\ &\qquad \frac{|D_\sigma|}{\delta t}(\rho_\sigma \tilde{u}_{\sigma,i} - \rho_\sigma^* u_{\sigma,i}^*) + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon} \tilde{u}_{\varepsilon,i}\\ &\qquad\qquad + |D_\sigma|\, (\nabla p^*)_\sigma^{(i)} + |D_\sigma|\, (\mathrm{div}\tau(\tilde{u}))_\sigma^{(i)} = 0,\end{aligned} \tag{4}$$

2 - Correction step – Solve for $u \in V$ and $p \in Q$:

$$\forall K \in \mathcal{M}, \qquad \frac{|K|}{\delta t}(\rho_K - \rho_K^*) + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma} = 0. \tag{5a}$$

$$\begin{aligned}&\text{For } 1 \leq i \leq d,\ \forall \sigma \in \mathcal{E}_{\mathrm{int}}^{(i)} \cup \mathcal{E}_N^{(i)},\\ &\qquad \frac{|D_\sigma|}{\delta t}\, \rho_\sigma\, (u_{\sigma,i} - \tilde{u}_{\sigma,i}) + |D_\sigma| \left(\nabla(p - p^*)\right)_\sigma^{(i)} = 0,\end{aligned} \tag{5b}$$

$$\forall K \in \mathcal{M}, \qquad \rho_K = \wp(p_K). \tag{5c}$$

The equations of the correction step are combined to produce a nonlinear parabolic problem for the pressure, which reads, $\forall K \in \mathcal{M}$:

$$\begin{aligned}\frac{|K|}{\delta t}\left(\wp(p_K) - \rho_K^*\right) + \sum_{\sigma = K|L} \frac{\rho_\sigma^{\mathrm{up}}}{\rho_\sigma}\, \frac{|\sigma|^2}{|D_\sigma|}(\phi_K - \phi_L) + \sum_{\sigma \in \mathcal{E}(K) \cap \mathcal{E}_N} \frac{\rho_\sigma^{\mathrm{up}}}{\rho_\sigma}\, \frac{|\sigma|^2}{|D_\sigma|}\phi_K\\ = \frac{1}{\delta t} \sum_{\sigma \in \mathcal{E}(K)} |\sigma|\, \rho_\sigma^{\mathrm{up}} \tilde{u} \cdot n_{K,\sigma},\end{aligned} \tag{6}$$

where $\phi \in Q$ is defined by $\phi = p - p^*$. Note that the second and third terms at the left-hand side look like a finite volume discretization of a diffusion operator, with homogeneous Neumann boundary conditions on $\mathcal{E}_D$ and Dirichlet boundary conditions on $\mathcal{E}_N$ for the pressure increment, as usual in pressure correction schemes (see [4] for a discussion on the effects of these spurious boundary conditions).

The standard discretizations entering the present framework are either low-degree non-conforming finite elements, namely the Crouzeix-Raviart element [3] for simplicial meshes or the Rannacher-Turek element [23] for quadrangles and hexahedra, or, for structured cartesian grids, the MAC scheme [13, 14]. We describe here the construction of the diffusion and pressure gradient terms for the finite element schemes, supposing for short that the velocity obeys homogeneous Dirichlet boundary conditions on $\partial\Omega$. Let $\sigma \in \mathcal{E}_{\rm int}$ and φ_σ be the finite element shape function associated to σ. In Rannacher-Turek or Crouzeix-Raviart elements, a degree of freedom for each component of the velocity is associated to each face, so $\mathcal{E}_{\rm int}^{(i)} = \mathcal{E}_{\rm int}$, for $1 \leq i \leq d$. Let $1 \leq i \leq d$ be given, let $e^{(i)}$ be the i^{th} vector of the canonical basis of $\mathbb{R}^d$ and let us define $\varphi_\sigma^{(i)}$ by:

$$\varphi_\sigma^{(i)} = \varphi_\sigma \, e^{(i)}.$$

Then the usual finite element discretization of the diffusion term reads, for a constant viscosity Newtonian fluid (that is supposing $\mathrm{div}\tau(u) = \mu\Delta u + (\mu/3)\nabla\mathrm{div}(u)$, with μ the viscosity):

$$|D_\sigma| \; (\mathrm{div}\tau(u))_\sigma^{(i)} = \sum_{K\in\mathcal{M}} \mu \int_K \nabla u : \nabla\varphi_\sigma^{(i)} \, \mathrm{d}x + \frac{\mu}{3}\int_K \mathrm{div}u \; \mathrm{div}\varphi_\sigma^{(i)} \, \mathrm{d}x.$$

The pressure gradient term at the internal face $\sigma = K|L$ reads:

$$|D_\sigma| \; (\nabla p)_\sigma^{(i)} = \sum_{K\in\mathcal{M}} \int_K p \; \mathrm{div}\varphi_\sigma^{(i)} \, \mathrm{d}x = |\sigma| \; (p_L - p_K) \, n_{K,\sigma} \cdot e^{(i)}.$$

3 The stability issue and consequences

3.1 A stability result for the convection

At the continuous level, let us assume that the mass balance $\partial_t\rho + \mathrm{div}(\beta) = 0$ holds, with β a regular vector-valued function. Then, for all scalar regular functions u and v, we have:

$$\int_\Omega \big[\partial_t(\rho u) + \mathrm{div}(u\beta)\big] \, v \, \mathrm{d}x = \int_\Omega \big[\partial_t(\rho u) - \frac{1}{2}\,(\partial_t\rho)\, u\big] \, v \, \mathrm{d}x + s(u,v) + \frac{1}{2}\int_{\partial\Omega} u\, v\, \beta\cdot n \, \mathrm{d}\gamma \quad (7)$$

where s is the following skew-symmetric bilinear form:

$$s(u,v) = \frac{1}{2}\int_\Omega v\beta \cdot \nabla u \,\mathrm{d}x - \frac{1}{2}\int_\Omega u\beta \cdot \nabla v \,\mathrm{d}x.$$

Taking $u = v = u_i$ and summing over i, the first term gives the time derivative of the kinetic energy, the second term vanishes and the last term corresponds to the kinetic energy flux through the boundary of the domain. The following Lemma, proven in [20], states a discrete counterpart of this computation in the case where the (possible) Dirichlet boundary conditions are homogeneous (see also [1] and [9] for a direct estimate of the kinetic energy, for an implicit and explicit scheme respectively).

Lemma 1. *Let us suppose that, for an index i, $1 \le i \le d$, the following discrete mass balance holds over the dual cells associated to the i^{th} component of the velocity:*

$$\forall \sigma \in \mathcal{E}^{(i)}_{\mathrm{int}} \cup \mathcal{E}^{(i)}_N, \qquad \frac{|D_\sigma|}{\delta t}(\rho_\sigma - \rho^*_\sigma) + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon} = 0. \tag{8}$$

Let $u, v \in V^{(i)}$, and let us suppose that these discrete functions obey homogeneous Dirichlet boundary. Then we have:

$$\sum_{\sigma \in \mathcal{E}^{(i)}_{\mathrm{int}} \cup \mathcal{E}^{(i)}_N} v_\sigma \Big[\frac{|D_\sigma|}{\delta t}(\rho_\sigma u_\sigma - \rho^*_\sigma u^*_\sigma) + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma)} F_{\sigma,\varepsilon} u_\varepsilon\Big] \\ \ge T_{\Omega,\mathrm{k}}(u,v) + T_{\Omega,\mathrm{s}}(u,v) + T_{\partial\Omega}(u,v), \tag{9}$$

with:

$$\begin{aligned} T_{\Omega,\mathrm{k}}(u,v) &= \sum_{\sigma \in \mathcal{E}^{(i)}_{\mathrm{int}} \cup \mathcal{E}^{(i)}_N} \frac{|D_\sigma|}{\delta t}\,(\rho_\sigma u_\sigma - \rho^*_\sigma u^*_\sigma)\, v_\sigma - \frac{1}{2}\,(\rho_\sigma - \rho^*_\sigma)\, u_\sigma\, v_\sigma, \\ T_{\Omega,\mathrm{s}}(u,v) &= S(u,v) - S(v,u), \qquad S(u,v) = \frac{1}{2} \sum_{\varepsilon \in \bar{\mathcal{E}}^{(i)}_{\mathrm{int}},\ \varepsilon = D_\sigma | D_{\sigma'}} F_{\sigma,\varepsilon}\, v_\varepsilon\, (u_{\sigma'} - u_\sigma), \\ T_{\partial\Omega}(u,v) &= \frac{1}{2} \sum_{\varepsilon \in \bar{\mathcal{E}}^{(i)}_N,\ \sigma = D_\sigma|\mathrm{ext}} F_{\sigma,\varepsilon}\, u_\varepsilon\, v_\varepsilon. \end{aligned}$$

Inequality (9) becomes an equality for a centred choice of the discretization of the face values u_ε.

Of course, $T_{\Omega,\mathrm{s}}(u,u) = 0$, and an easy computation shows that:

$$T_{\Omega,\mathrm{k}}(u,u) \ge \frac{1}{2\delta t} \sum_{\sigma \in \mathcal{E}^{(i)}_{\mathrm{int}} \cup \mathcal{E}^{(i)}_N} |D_\sigma| \left[\rho_\sigma u_\sigma^2 - \rho^*_\sigma (u^*_\sigma)^2\right].$$

Applying Lemma 1 to each component of the velocity, the obtained term is thus the discrete time-derivative of the kinetic energy, and may be used to obtain stability estimates for the scheme (see Sect. 4).

Remark 1 (Non-homogeneous Dirichlet boundary conditions). The limitation to homogeneous Dirichlet boundary conditions may be seen, from the proof, to stem from the fact that no balance equation is written on the dual cells associated to faces lying on $\partial\Omega_D$. The problem may thus be fixed by keeping these degrees of freedom and using a penalization technique.

Remark 2 (Artificial boundary conditions). Lemma 1 may be used to derive artificial boundary conditions allowing the flow to enter the domain through $\partial\Omega_N$, by first collecting the boundary terms in the variational form of the momentum balance equation (*i.e.* adding to $T_{\partial\Omega}(u, v)$ the terms issued from the diffusion and the pressure gradient) and then imposing that the result may be written as a linear form acting on the test function (see [2] for a similar development in the incompressible case). The so-built boundary condition is observed in practice to give quite good results when modelling external flows [20].

3.2 Discretization of the convection term

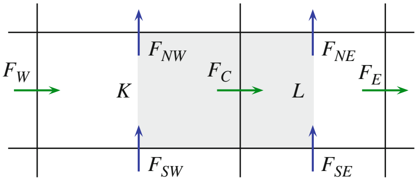


Fig. 2 Local notations for the definition of the mass fluxes at the dual edges with the MAC scheme

The problem to tackle is now the following one: on one side, the discrete mass balance over the dual cells (8) is necessary for the stability of the scheme; on the other side, the mass balance is only written by the scheme(s) for the primal cells (Equation (2a) or (5a)). We are thus lead to express the mass fluxes ($F_{\sigma,\varepsilon}$) through the dual faces as a function of the mass fluxes ($F_{K,\sigma}$) through the primal faces, in such a way that the discrete balance over the primal cells yields a discrete balance over the dual cells. We describe in this section how this may be done, first for the MAC (structured) mesh (see also [15]) and, second, for the Rannacher-Turek element on general quadrangles.

3.2.1 MAC scheme

For the MAC scheme, in two space dimensions and with the local notations introduced on Fig. 2, the mass balance on the primal cells reads:

$$K: \frac{|K|}{\delta t} (\rho_K - \rho_K^*) - F_W - F_{SW} + F_C + F_{NW} = 0,$$
$$L: \frac{|L|}{\delta t} (\rho_L - \rho_L^*) - F_C - F_{SE} + F_E + F_{NE} = 0.$$

Multiplying both equations by $1/2$ and summing them yields, for $\sigma = K|L$:

$$\begin{aligned}&\frac{|D_\sigma|}{\delta t} (\rho_\sigma - \rho_\sigma^*)\\ &-\frac{1}{2}\big[F_W + F_C\big] - \frac{1}{2}\big[F_{SW} + F_{SE}\big] + \frac{1}{2}\big[F_C + F_E\big] + \frac{1}{2}\big[F_{NW} + F_{NE}\big] = 0,\end{aligned} \tag{10}$$

with the usual definition of the dual cell D_σ, which implies that $|D_{K,\sigma}| = |K|/2$ and $|D_{L,\sigma}| = |L|/2$, and with the following definition of the density on the face:

$$|D_\sigma|\, \rho_\sigma = |D_{K,\sigma}|\, \rho_K + |D_{L,\sigma}|\, \rho_L. \tag{11}$$

Equation (10) thus suggests the following definition for the mass fluxes at the dual faces:

left face: $F_{\sigma,\varepsilon} = -\frac{1}{2}\big[F_W + F_C\big]$; right face: $F_{\sigma,\varepsilon} = \frac{1}{2}\big[F_C + F_E\big]$;

bottom face: $F_{\sigma,\varepsilon} = -\frac{1}{2}\big[F_{SW} + F_{SE}\big]$; top face: $F_{\sigma,\varepsilon} = \frac{1}{2}\big[F_{NW} + F_{NE}\big]$.

Note that this definition is rather non-standard: for instance, the flux at the left face of D_σ, which is included in K, may involve densities of the neighbouring primal cells. The extension of the above construction to the three-dimensional case is straightforward.

3.2.2 Rannacher-Turek element

A construction similar to that of the MAC scheme may be performed for rectangular meshes. For K and L two neighbouring cells of $\mathcal{M}$, the half-diamond cell $D_{K,\sigma}$ (resp. $D_{L,\sigma}$) associated to the common face $\sigma = K|L$ is defined as the cone with vertex the mass center of K (resp. L) and with basis σ, the density ρ_σ is defined by the weighted average (11), and the dual mass fluxes are obtained by multiplying the mass

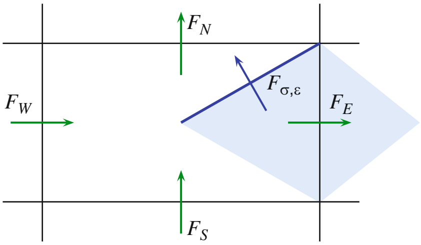


Fig. 3 Local notations for the definition of the mass fluxes at the dual edges with the Rannacher-Turek element

balances over K and L by $1/4$ and summing. With the local notations of Fig. 3, this yields, for the dual mass flux $F_{\sigma,\varepsilon}$, an expression of the form:

$$F_{\sigma,\varepsilon} = -\frac{1}{8} F_W + \frac{3}{8} F_N - \frac{3}{8} F_E + \frac{1}{8} F_S. \tag{12}$$

We now explain how to extend this formulation to general meshes.

Let us suppose that, for any cell $K \in \mathcal{M}$, we are able to define the fluxes through the dual faces included in K in such a way that:

(A1) The mass balance over the half-diamond cells is proportional to the mass balance over K, in the following sense:

$$\forall \sigma \in \mathcal{E}(K), \qquad F_{K,\sigma} + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma),\ \varepsilon \subset K} F_{\sigma,\varepsilon} = \xi_K^\sigma \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma},$$

with $\sum_{\sigma \in \mathcal{E}(K)} \xi_K^\sigma = 1$ and, for any $\sigma \in \mathcal{E}(K)$, $\xi_K^\sigma \geq 0$.

(A2) The dual fluxes are conservative, *i.e.*, for any $\varepsilon = D_\sigma | D'_\sigma$, $F_{\sigma,\varepsilon} = -F_{\sigma',\varepsilon}$.

(A3) The dual fluxes are bounded with respect to the $(F_{K,\sigma})_{\sigma \in \mathcal{E}(K)}$:

$$\forall \sigma \in \mathcal{E}(K),\ \forall \epsilon \in \bar{\mathcal{E}}(D_\sigma) \subset K \quad |F_{\sigma,\epsilon}| \leq C \ \max \Big\{ |F_{K,\sigma}|,\ \sigma \in \mathcal{E}(K) \Big\}.$$

In addition, let us define $|D_{K,\sigma}|$ as:

$$|D_{K,\sigma}| = \xi_K^\sigma \ |K|, \tag{13}$$

and ρ_σ, once again, by the weighted average (11). Then the dual fluxes satisfy the required mass balance. Indeed, for $\sigma \in \mathcal{E}_{\mathrm{int}}$, $\sigma = K|L$, we have:

$$\begin{aligned}
&\frac{|D_\sigma|}{\delta t} (\rho_\sigma - \rho_\sigma^*) + \sum_{\varepsilon \in \mathcal{E}(D_\sigma)} F_{\sigma,\varepsilon} \\
&= \frac{|D_{K,\sigma}|}{\delta t} (\rho_K - \rho_K^*) + F_{K,\sigma} + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma),\ \varepsilon \subset K} F_{\sigma,\varepsilon} \\
&\qquad + \frac{|D_{L,\sigma}|}{\delta t} (\rho_L - \rho_L^*) + F_{L,\sigma} + \sum_{\varepsilon \in \bar{\mathcal{E}}(D_\sigma),\ \varepsilon \subset L} F_{\sigma,\varepsilon} \\
&= \xi_K^\sigma \left[\frac{|K|}{\delta t} (\rho_K - \rho_K^*) + \sum_{\sigma \in \mathcal{E}(K)} F_{K,\sigma} \right] + \xi_L^\sigma \left[\frac{|L|}{\delta t} (\rho_L - \rho_L^*) + \sum_{\sigma \in \mathcal{E}(L)} F_{L,\sigma} \right] = 0.
\end{aligned}$$

A similar computation leads to the same conclusion for the (half-)dual cells associated to the Neumann boundary faces.

The next issue is to check whether Assumptions (A1)-(A3) are sufficient for the consistency of the scheme. In this respect, the following lemma [16] brings a decisive argument.

Lemma 2. *Let Assumptions (A1)-(A3) hold. For $v \in V$ and $K \in \mathcal{M}$, let v_K be defined by $v_K = \sum_{\sigma\in\mathcal{E}(K)} \xi_K^\sigma\, v_\sigma$. Let $u \in V$, and $R(u,v)$ be the quantity defined by:*

$$R(u,v) = \sum_{\sigma\in\mathcal{E}_{\rm int}} v_\sigma \sum_{\substack{\varepsilon\in\bar{\mathcal{E}}(D_\sigma),\\ \varepsilon=D_\sigma|D'_\sigma}} F_{\sigma,\varepsilon}\, \frac{u_\sigma + u_{\sigma'}}{2} - \sum_{K\in\mathcal{M}} v_K \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\, u_\sigma.$$

Let us suppose that the primal fluxes are associated to a convection momentum field β, i.e. $\forall K \in \mathcal{M},\ \forall\sigma \in \mathcal{E}(K),\quad F_{K,\sigma} = |\sigma|\ \beta_\sigma \cdot n_{K,\sigma}$. *(For the schemes used here, of course, β depends the density and the velocity, see (3).) Then there exists C depending only on the regularity of the mesh such that:*

$$|R(u,v)| \le C\ h\ \|\beta\|_{l^\infty}\ \|u\|_1\ \|v\|_1,$$

with $\|\beta\|_{l^\infty} = \max_{\sigma\in\mathcal{E}} |\beta_\sigma|$ and the discrete H^1*-norm on the dual mesh is defined by:*

$$\forall v \in V,\quad \|v\|_1 = \sum_{K\in\mathcal{M}} h_K^{d-2} \sum_{\sigma,\sigma'\in\mathcal{E}(K)} (v_\sigma - v_{\sigma'})^2.$$

The quantity $R(u,v)$ compares two discrete analogues to $\int_\Omega v\,{\rm div}(u\beta)\,{\rm d}x$; the first analogue is defined with the divergence taken over the dual meshes while the second analogue is defined with the divergence over the primal cells. Let us suppose that the discrete H^1-norm of the solution is controlled thanks to the diffusion term. Then, in a convergence or error analysis study in the linear case (*i.e.* with a given regular convection field β), Lemma 2 allows to replace the first discrete analogue by the second one, thus substituting well defined quantities to quantities only defined through (A1)-(A3). It is used in [16] to prove that the scheme is first-order for the stationary convection-diffusion equation. The convergence for the constant density Navier-Stokes equations (that is with $\beta = u$) was also proven, controlling now $\|u\|_{l^\infty}$ by $\|u\|_1$ thanks to an inverse inequality.

The last task is now to build fluxes satisfying (A1)-(A3); this is easily done by choosing $\xi_K^\sigma = 1/4$, and keeping for the expression of the dual fluxes as a function of the primal fluxes the same linear combination (12) as in the rectangular case. Note that this implicitly implies that the geometrical definition of the dual cells has been generalized, since it is not possible in general to split a quadrangle in four simplices of same measure (even if the quadrangle is convex) . The extension to three dimensions only needs to deal with the rectangular parallelepipedic case, which is quite simple [1]. Finding directly a solution to (A1)-(A3) may also be an alternative route, to deal with more complex cases, as done in [16] to extend the scheme to locally refined non-conforming grids.

4 Schemes and stability estimates

In order to obtain the complete formulation of the considered schemes, we now have to fix the time-marching procedure. This is straightforward for the implicit scheme, and we concentrate here on the pressure correction scheme. The problem which we face in this case is that the mass balance is not yet solved when performing the prediction step. In our implementations in the ISIS computer code [18] developed at IRSN on the basis of the software component library PELICANS [22], it is circumvented by just shifting in time the density ρ_σ; the mass balance on the dual cells is recovered from the mass balance on the primal cells at the previous time step. This has essentially two drawbacks. First, the trick indeed works only if the time step is constant; for a variable time step, one has to choose between loosing stability or consistency (locally in time, so fortunately, without observed impact in practice). Second, the scheme is only first order in time.

In addition, stability seems to require an initial pressure renormalization step, which is an algebraic variant of the one introduced in [12]. It seems however that this step may be omitted in practice.

The algorithm (keeping in this presentation the pressure renormalization step) reads, assuming that u^n, p^n, ρ^n and the family $(F^n_{K,\sigma})$ are known:

1- Pressure renormalization step – Let $(\lambda_\sigma)_{\sigma\in\mathcal{E}_{\rm int}}$ be a family of positive real numbers, and let $-{\rm div}(\lambda\nabla)_{\mathcal{M}}$ be the discrete elliptic operator from Q to Q defined by, $\forall K \in \mathcal{M}$ and $q \in Q$:

$$\big[-{\rm div}(\lambda\nabla)_{\mathcal{M}}(q)\big]_K = \sum_{\sigma=K|L} \lambda_\sigma \, \frac{|\sigma|^2}{|D_\sigma|}(q_K - q_L) + \sum_{\sigma\in\mathcal{E}_N,\sigma=K|{\rm ext}} \lambda_\sigma \, \frac{|\sigma|^2}{|D_\sigma|} q_K.$$

Then $\tilde{p}^{n+1} \in Q$ is given by:

$$-\,{\rm div}(\frac{1}{\rho^n}\nabla)_{\mathcal{M}}\;(\tilde{p}^{n+1}) = -{\rm div}(\frac{1}{[\rho^n\;\rho^{n-1}]^{1/2}}\nabla)_{\mathcal{M}}\;(p^n), \qquad (14)$$

the weights $(\rho^n_\sigma)_{\sigma\in\mathcal{E}_{\rm int}\cup\mathcal{E}_N}$ and $(\rho^{n-1}_\sigma)_{\sigma\in\mathcal{E}_{\rm int}\cup\mathcal{E}_N}$ being the densities involved in the time-derivative term of the momentum balance equation.

2- Velocity prediction step – Solve for $\tilde{u}^{n+1} \in V$, for $1 \leq i \leq d$ and $\forall\sigma \in \mathcal{E}^{(i)}_{\rm int} \cup \mathcal{E}^{(i)}_N$:

$$\frac{|D_\sigma|}{\delta t}(\rho^n_\sigma \tilde{u}^{n+1}_{\sigma,i} - \rho^{n-1}_\sigma u^n_{\sigma,i}) + \sum_{\varepsilon\in\bar{\mathcal{E}}(D_\sigma)} F^n_{\sigma,\varepsilon}\tilde{u}^{n+1}_{\varepsilon,i}$$
$$+\,|D_\sigma|\,(\nabla\tilde{p}^{n+1})^{(i)}_\sigma + |D_\sigma|\,({\rm div}\tau(\tilde{u}^{n+1}))^{(i)}_\sigma = 0, \quad (15)$$

where the $(F^n_{\sigma,\varepsilon})_{\varepsilon\in\tilde{\mathcal{E}}(D_\sigma)}$ are built as explained in the previous section, from the primal fluxes at time t^n.

3 - Correction step – Solve for $u^{n+1} \in V$ and $p^{n+1} \in Q$:

$$\forall K \in \mathcal{M}, \qquad \frac{|K|}{\delta t}(\rho_K^{n+1} - \rho_K^n) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}^{n+1} = 0. \tag{16a}$$

For $1 \leq i \leq d$, $\forall \sigma \in \mathcal{E}^{(i)}_{\rm int} \cup \mathcal{E}^{(i)}_N$,

$$\frac{|D_\sigma|}{\delta t}\,\rho^n_\sigma\,(u^{n+1}_{\sigma,i} - \tilde{u}^{n+1}_{\sigma,i}) + |D_\sigma|\,\big(\nabla(p^{n+1} - \tilde{p}^{n+1})\big)^{(i)}_\sigma = 0, \tag{16b}$$

$$\forall K \in \mathcal{M}, \qquad \rho_K^{n+1} = \wp(p_K^{n+1}). \tag{16c}$$

The algorithm must be initialized by the data $u^0 \in V$, $\rho^{-1} \in Q$ and $\rho^0 \in Q$ satisfying the discrete mass balance equation, and with the corresponding mass fluxes $(F^0_{K,\sigma})$. A possible way to obtain these quantities is to evaluate u^0 and ρ^{-1} from the initial conditions, and, as a preliminary step, to solve for ρ^0 the mass balance equation.

The upwinding in the discretization of the mass balance equation has for consequence that any density appearing in the algorithm is positive (provided that the initial density is positive). The existence and uniqueness of a solution to Steps 1 and 2 is then clear: these are linear problems with coercive operators (for Step 2, thanks to the stability of the convection term). The existence of a solution to Step 3 may be obtained by a Brouwer fixed point argument, using the fact that the conservativity of the mass balance yields an estimate for ρ, so for p, and finally for u (in any norm, since we work on finite dimensional spaces). The algorithm is thus well-posed.

Let us now turn to the energy estimate. At the continuous level, this relation is obtained for the barotropic Navier-Stokes equations by choosing the velocity u as a test function in the variational form of the momentum balance equation, writing the convection term as the time derivative of the kinetic energy, and setting the pressure work, namely $-\int_\Omega p\,\mathrm{div}(u)\,\mathrm{d}x$, under a convenient form. This latter step is done by the following formal computation. Let $b(\cdot)$ be a regular function from $(0,+\infty)$ to $\mathbb{R}$, and let us multiply the mass balance by $b'(\rho)$. Using:

$$\begin{aligned} b'(\rho)\mathrm{div}(\rho u) = b'(\rho)[u\cdot\nabla\rho + \rho\mathrm{div}(u)] &= u\cdot\nabla b(\rho) + \rho b'(\rho)\mathrm{div}(u) \\ &= \mathrm{div}(b(\rho)u) + \big[\rho b'(\rho) - b(\rho)\big]\,\mathrm{div}(u), \end{aligned}$$

we get:

$$\partial_t\big[b(\rho)\big] + \mathrm{div}\big[b(\rho)\,u\big] + \big[\rho b'(\rho) - b(\rho)\big]\,\mathrm{div}(u) = 0.$$

Choosing now the function $b(\cdot)$ in such a way that $\rho b'(\rho) - b(\rho) = \wp^{-1}(p)$, integrating over Ω and supposing homogeneous Dirichlet boundary conditions over $\partial\Omega$ yields:

$$-\int_\Omega p \operatorname{div}(u)\,\mathrm{d}x = \frac{d}{dt}\int_\Omega b(\rho)\,\mathrm{d}x.$$

The following lemma [7] states a discrete counterpart of this computation.

Lemma 3. *Let us suppose that the velocity field obeys homogeneous Dirichlet boundary conditions. Let $b(\cdot)$ be a regular convex function from $(0,+\infty)$ to $\mathbb{R}$, and $(\rho^\star_K)_{K\in\mathcal{M}}$ be a positive family of real numbers. Then, with the upwind discretization (3) of the mass balance equation, the family $(\rho_K)_{K\in\mathcal{M}}$ is also positive, and we get:*

$$\sum_{K\in\mathcal{M}} b'(\rho_K)\Big[\frac{|K|}{\delta t}(\rho_K - \rho^*_K) + \sum_{\sigma\in\mathcal{E}(K)} F_{K,\sigma}\Big] \geq$$
$$\frac{1}{\delta t}\sum_{K\in\mathcal{M}} |K|\,\big[b(\rho_K) - b(\rho^*_K)\big] + \sum_{K\in\mathcal{M}} \big[\rho_K b'(\rho_K) - b(\rho_K)\big] \sum_{\sigma\in\mathcal{E}(K)} |\sigma|\, u_\sigma \cdot n_{K,\sigma}.$$

We are now in position to state the following stability result.

Theorem 1. *Let us suppose that the velocity field obeys homogeneous Dirichlet boundary conditions. The scheme (14)-(16) satisfies the following energy identity, for $1 \leq n \leq N$:*

$$\frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}^{(i)}_{\mathrm{int}}} |D_\sigma|\,\rho^{n-1}_\sigma\,(u^n_{\sigma,i})^2 + \delta t\sum_{k=1}^{n}\sum_{\sigma\in\mathcal{E}^{(i)}} |D_\sigma|\,(\operatorname{div}\tau(u^k))^{(i)}_\sigma\, u^k_{\sigma,i}$$
$$+ \sum_{K\in\mathcal{M}} |K|\, b(\rho^n_K) \leq \frac{1}{2}\sum_{i=1}^{d}\sum_{\sigma\in\mathcal{E}^{(i)}_{\mathrm{int}}} |D_\sigma|\,\rho^{(-1)}_\sigma\,(u^0_{\sigma,i})^2 + \sum_{K\in\mathcal{M}} |K|\, b(\rho^0_K).$$

The proof of this theorem is based on Lemma 1 and Lemma 3, and may be found, for the essential arguments, in [7].

Remark 3. Let us suppose that the equation of state reads $p = \rho^\gamma$, with $\gamma \in (1,+\infty)$. Then an easy computation yields $b(\rho) = \rho^\gamma/(\gamma-1) = p/(\gamma-1)$. Theorem 1 thus yields an estimate for the pressure in $\mathrm{L}^\infty(0,T;\mathrm{L}^1)$-norm. Note that this estimate is however not sufficient to ensure that a sequence of pressures obtained as discrete solutions converges to a function; in fact, in convergence studies of numerical schemes [5, 6, 8] as well as in mathematical analysis of the continuous problem [21], the pressure has to be controlled from estimates of its gradient.

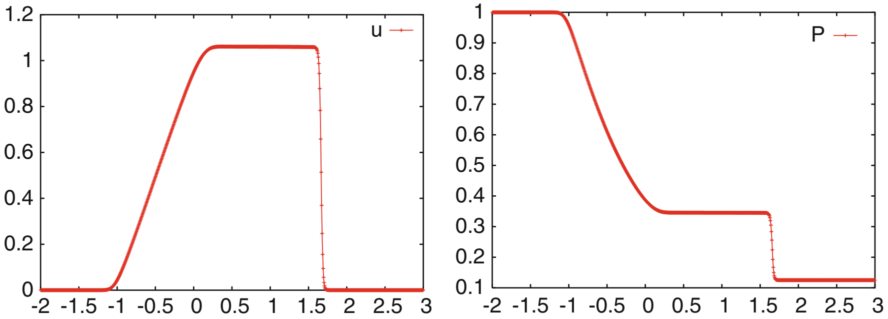


Fig. 4 Solution for the Sod shock-tube problem, obtained with a uniform mesh of 800 cells, with a residual viscosity – *left:* velocity, *right:* pressure

5 Euler equations and solutions with shocks

In this section we briefly discuss the capability of the considered numerical schemes to compute irregular (*i.e.* with discontinuities) solutions of inviscid flows.

The results obtained with the above described pressure correction scheme for the so-called one-dimensional Sod shock-tube problem are displayed on Fig. 4 (see [19] for a more detailed presentation). From numerical experiments, it seems that this scheme converges when the velocity space translates are controlled, either by upwinding the discretization of the velocity convection term, or by keeping a residual viscosity in the (discrete) momentum balance equation. Numerical experiments reported in [19] (addressing also an extension of this algorithm to the barotropic homogeneous two-phase flow model [11]) confirm the stability of the scheme, and show that the qualitative behaviour of the solution is captured up to very large values of the CFL number (typically, in the range of 50).

From the theoretical point of view, for Euler equations (*i.e.*, precisely speaking, with a diffusion vanishing with the space step), the control that we are able to prove on the solution of course does not yield (weak or strong) convergence in strong enough norms to pass to the limit in the scheme. We can however prove the following result: supposing convergence for the density in $\mathrm{L}^p(\Omega)$, $p \in [1, +\infty)$ and for the velocity in $\mathrm{L}^r(\Omega)$, $r \in [1, 3]$, it is possible to pass to the limit in the discrete equations, provided that the viscosity vanishes as h^α, $\alpha \in (0, 2)$ for both the implicit and the pressure correction scheme. In this case, the limit of a sequence of discrete solutions is proven to satisfy the weak form of the Euler equations, and so, in particular, the Rankine-Hugoniot conditions at the shocks.

6 Discussion and perspectives

The theoretical analysis of the schemes presented here has been undertaken for model stationary problems: in [5, 8], we prove the convergence for the Crouzeix-Raviart discretization of the Stokes equations (with the additional stabilization term needed for purely technical reasons); in [6], we prove the same result for the (standard) MAC scheme. An extension, still for the MAC discretization, to the stationary Navier-Stokes equations is underway.

From a practical point of view, a next step for the barotropic Navier-Stokes equations should be to derive an upwind explicit version of the scheme presented here; in this direction, an extension of Lemma 1 (stability of the velocity convection term) to the explicit case may be found in [9].

The main objective is however to deal with the full (*i.e.* non barotropic, therefore including an energy balance) Navier-Stokes equations. An unconditionally stable pressure correction scheme has been derived for this problem [17], but extensive tests of this scheme remain to be done. In particular, stability requires that the internal energy remains non-negative (in practice, positive); the way we obtained this property was to solve the internal energy balance, with a scheme able to preserve the sign of the unknown. However, it is commonly agreed that, for the scheme to converge toward the correct weak solution, a conservative discretization of the total energy balance should be used. The actual occurrence of this problem, and the possibility to circumvent it, possibly by adding stabilizing viscous terms, will deserve investigations in the near future; a preliminary step on this route may be found in [10].

References

1. G. Ansanay-Alex, F. Babik, J.-C. Latché, D. Vola: An L^2-stable approximation of the Navier-Stokes convection operator for low-order non-conforming finite elements. IJNMF, online (2010).
2. C.H. Bruneau, P. Fabrie: Effective downstream boundary conditions for incompressible Navier-Stokes equations. IJNMF, **99**, 693–705 (1994).
3. M. Crouzeix, P.-A. Raviart: Conforming and nonconforming finite element methods for solving the stationary Stokes equations I, Revue Française d'Automatique, Informatique et Recherche Opérationnelle (R.A.I.R.O.), **R-3**, 33–75 (1973).
4. F. Dardalhon, J.-C. Latché, S. Minjeaud: Analysis of a projection method for low-order non-conforming finite elements. Submitted (2011).
5. R. Eymard, T. Gallouët, R. Herbin, J.-C. Latché: A convergent Finite Element-Finite Volume scheme for the compressible Stokes problem. Part II: the isentropic case. Mathematics of Computation, **79**, 649–675 (2010).
6. R. Eymard, T. Gallouët, R. Herbin, J.-C. Latché: Convergence of the MAC scheme for the compressible Stokes equations. SIAM Journal on Numerical Analysis, **48**, 2218–2246 (2010).
7. T. Gallouët, L. Gastaldo, R. Herbin, J.-C. Latché: An unconditionally stable pressure correction scheme for compressible barotropic Navier-Stokes equations. Mathematical Modelling and Numerical Analysis, **42**, 303–331 (2008).
8. T. Gallouët, R. Herbin, J.-C. Latché: A convergent Finite Element-Finite Volume scheme for the compressible Stokes problem. Part I: the isothermal case. Mathematics of Computation, **78**, 1333–1352 (2009).
9. T. Gallouët, R. Herbin, J.-C. Latché: Kinetic energy control in explicit Finite-Volume discretizations of the incompressible and compressible Navier-Stokes equations. International Journal of Finite Volumes **2** (2010).

10. T. Gallouët, R. Herbin, J.-C. Latché, T.T. Nguyen: Playing with Burgers equation. Finite Volumes for Complex Applications VI (FVCA VI), these proceedings.
11. L. Gastaldo, R. Herbin, J.-C. Latché: An unconditionally stable Finite Element-Finite Volume pressure correction scheme for the drift-flux model. Mathematical Modelling and Numerical Analysis, **44**, 251–287 (2010).
12. J.-L. Guermond, L. Quartapelle: A Projection FEM for Variable Density Incompressible Flows. Journal of Computational Physics, **165**, 167–188 (2000).
13. F.H. Harlow, J.E. Welsh: Numerical calculation of time-dependent viscous incompressible flow of fluid with free surface. Physics of Fluids, **8**, 2182–2189 (1965).
14. F.H. Harlow, A.A. Amsden: A numerical fluid dynamics calculation method for all flow speeds. Journal of Computational Physics, **8**, 197–213 (1971).
15. R. Herbin, J.-C. Latché: A kinetic energy control in the MAC discretization of compressible Navier-Stokes equations. International Journal of Finite Volumes **2** (2010).
16. R. Herbin, J.-C. Latché, B. Piar: A finite-element finite-volume face centred scheme with non-conforming local refinement. I – Convection-diffusion equation. In preparation (2011).
17. R. Herbin, W. Kheriji, J.-C. Latché: An unconditionally stable pressure correction scheme for the compressible Navier-Stokes equations. In preparation (2011).
18. ISIS: a CFD computer code for the simulation of reactive turbulent flows, https://gforge.irsn.fr/gf/project/isis.
19. W. Kheriji, R. Herbin, J.-C. Latché: Numerical tests of a new pressure correction scheme for the homogeneous model. ECCOMAS CFD 2010, Lisbon, Portugal, June 2010.
20. C. Lapuerta, J.-C. Latché: Discrete artificial boundary conditions for compressible external flows. In preparation (2011).
21. P.-L. Lions: Mathematical Topics in Fluid Mecanics. Volume 2. Compressible Models. Oxford Lecture Series in Mathematics and its Applications, vol. 10 (1998).
22. PELICANS: Collaborative Development Environment. https://gforge.irsn.fr/gf/project/pelicans.
23. R. Rannacher, S. Turek: Simple Nonconforming Quadrilateral Stokes Element. Numerical Methods for Partial Differential Equations, **8**, 97–111 (1992).

The paper is in final form and no similar paper has been or is being submitted elsewhere.

ALE Method for Simulations of Laser-Produced Plasmas

Liska R., Kuchařík M., Limpouch J., Renner O., Váchal P., Bednárik L., and Velechovský J.

Abstract Simulations of laser-produced plasmas are essential for laser-plasma interaction studies and for inertial confinement fusion (ICF) technology. Dynamics of such plasmas typically involves regions of large scale expansion or compression, which requires to use the moving Lagrangian coordinates. For some kind of flows such as shear or vortex the moving Lagrangian mesh however tangles and such flows require the use of arbitrary Lagrangian Eulerian (ALE) method. We have developed code PALE (Prague ALE) for simulations of laser-produced plasmas which includes Lagrangian and ALE hydrodynamics complemented by heat conductivity and laser absorption. Here we briefly review the numerical methods used in PALE code and present its selected applications to modeling of laser interaction with targets.

Keywords Lagrangian coordinates, ALE method, laser-produced plasma
MSC2010: 35L65, 35K05, 65M08

1 Introduction

Understanding of laser-produced plasma behavior and evolution is crucial for studies of intense laser interaction with targets and for inertial confinement fusion (ICF) technology employing high-power laser beams to ignite the fusion reaction of deuterium-tritium fuel. We model the complex problem of laser-produced plasma by a set of hydrodynamical conservation laws for mass, momentum and energy

Liska R., Kuchařík M., Limpouch J., Váchal P., Bednárik L, and Velechovský J.
Czech Technical University, Faculty of Nuclear Sciences and Physical Engineering, Břehová 7, 115 19 Prague 1, Czech Republic, e-mail: liska@siduri.fjfi.cvut.cz

Renner O.
Institute of Physics, v.v.i., Academy of Sciences Czech Rep., Na Slovance 2, 182 21 Prague, Czech Republic

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_87, © Springer-Verlag Berlin Heidelberg 2011

of compressible fluid, complemented by laser absorption and heat transfer, which written in Lagrangian coordinates have the form

$$\frac{1}{\rho}\frac{\mathrm{d}\rho}{\mathrm{d}t} = -\mathrm{div}\,\mathbf{U}, \tag{1}$$

$$\rho\frac{\mathrm{d}\mathbf{U}}{\mathrm{d}t} = -\mathrm{grad}\,p, \tag{2}$$

$$\rho\frac{\mathrm{d}\varepsilon}{\mathrm{d}t} = -p\,\mathrm{div}\,\mathbf{U} + \mathrm{div}(\kappa\,\mathrm{grad}\,T) - \mathrm{div}\,\mathbf{I}, \tag{3}$$

where ρ is density, $\mathbf{U}$ velocity, p pressure, ε specific internal energy (energy per unit mass), T temperature, κ heat conductivity, $\mathbf{I}$ laser energy flux density (Poynting vector) and $\mathrm{d}/\mathrm{d}\,t = \partial/\partial\,t + \mathbf{u}\cdot\mathrm{grad}$ is the total Lagrangian time derivative including convective terms. The system is closed by the equation of state coupling density, internal energy, pressure and temperature. Laser-produced plasma is usually modeled in the Lagrangian coordinates moving with the fluid, which are able to deal with moving boundaries and large scale deformation like compression or expansion appearing typically in laser-produced plasmas. Some types of plasma flows such as shear or vortex can result in tangling of the computational mesh. This problem can be avoided by Arbitrary Lagrangian-Eulerian (ALE) method [8], which smooths (rezones) the computational mesh and interpolates (remaps) conservative variables to the smoothed mesh after several Lagrangian time steps. Standard Lagrangian numerical hydrodynamics employs staggered method [3,4,6], however one can use composite schemes [18,26] and recently much attention has been attracted by the cell-centered methods [19,22,23].

We have developed a 2D ALE code PALE (Prague ALE) for laser-produced plasma simulations, which uses a 2D quadrilateral, logically rectangular computational mesh, We shortly outline the numerical methods employed in the PALE code for ALE hydrodynamics, heat conductivity and laser absorption. The PALE code capabilities are demonstrated on simulations of laser interaction with targets.

2 Hydrodynamics

The hydrodynamical ALE method consists from Lagrangian, rezone and remap phases. Rezone and remap is applied either regularly after fixed number of Lagrangian time steps or adaptively when quality of the moving mesh becomes bad.

2.1 Staggered Lagrangian Method

We consider a 2D staggered location of physical variables: velocity vector is defined at point (node) p of the computational mesh and is denoted $\mathbf{U}_p = (u_p, v_p)$, specific internal energy ε_c is defined at the center of the cell c and density ρ_{pc}

is defined at the center of the subcell Ω_{pc}. The subcell Ω_{pc} is the quadrilateral whose vertexes are point p, center of cell c and two midpoints of two edges of cell c originating at point p. In the Lagrangian gas dynamics the nodes move with the local fluid velocity, and the mass of a subcell is assumed to be constant in time. Conservative variables are the mass m, momentum $m\mathbf{U}$ and total energy $E = m\left(\varepsilon + \frac{1}{2}||\mathbf{U}||^2\right)$. This discretization is based on the philosophy of compatible hydrodynamics algorithms introduced in [4]. The Lagrangian phase is conservative, that is, some discrete form of mass, momentum, and total energy is conserved [4]. The mass of any subcell is given by $m_{pc} = \rho_{pc} V_{pc}$, where V_{pc} is the volume of the subcell. Then masses of the cell and node are defined by summation of subcell masses

$$m_c = \sum_{n \in \mathscr{P}(c)} m_{pc} \quad \text{and} \quad m_p = \sum_{c \in \mathscr{C}(p)} m_{pc}$$

over set of points $\mathscr{P}(c)$ being vertexes of the cell c and set of cells $\mathscr{C}(p)$ sharing the node p. All these masses participate in the Lagrangian phase of an ALE method, subcell mass m_{pc} is assumed to be Lagrangian, so it does not change with time, therefore $\rho_{pc}(t) = m_{pc}/V_{pc}(t)$, which can be considered as a definition of subcell density for given constant subcell mass. Masses of cells and nodes are also Lagrangian because they are sums of subcell masses. As in the subcell density definition, one gets

$$\rho_c(t) = \frac{m_c}{V_c(t)} = \frac{\sum_{p \in \mathscr{P}(c)} m_{pc}}{\sum_{p \in \mathscr{P}(c)} V_{pc}(t)}. \tag{4}$$

The total mass $\mathscr{M}$, which is conserved during the Lagrangian phase, is $\mathscr{M} = \sum_{pc} m_{pc} = \sum_c m_c = \sum_p m_p$. In this part we show how momentum and specific internal energy can be discretized in such a way that mass, momentum and total energy are conserved.

Assume a general force $\mathbf{F}_{pc}$ modeling the action of subcell pc on point p is given, then a general force for point p can be assembled as

$$\mathbf{F}_p = \sum_{c \in \mathscr{C}(p)} \mathbf{F}_{pc}. \tag{5}$$

Then spatial discretization of the momentum equation is defined by

$$m_p \frac{\mathrm{d}\mathbf{U}_p}{\mathrm{d}t} = \mathbf{F}_p. \tag{6}$$

The discrete total energy over the whole domain is given by the sum of internal energy and kinetic energy

$$\mathscr{E} = \sum_c m_c \varepsilon_c + \sum_p \frac{1}{2} m_p ||\mathbf{U}_p||^2, \tag{7}$$

and conservation implies $\mathrm{d}\mathscr{E}/\mathrm{d}t) = 0$. However the differentiation of (7) with respect to time formally gives

$$
\begin{aligned}
\frac{\mathrm{d}\mathscr{E}}{\mathrm{d}t} &= \sum_c m_c \frac{\mathrm{d}\varepsilon_c}{\mathrm{d}t} + \sum_p \underbrace{m_p \frac{\mathrm{d}\mathbf{U}_p}{\mathrm{d}t}}_{=\mathbf{F}_p} \cdot \mathbf{U}_p, \\
&= \sum_c \left(m_c \frac{\mathrm{d}\varepsilon_c}{\mathrm{d}t} + \sum_{p\in\mathscr{P}(c)} \mathbf{F}_{pc} \cdot \mathbf{U}_p \right) = 0. \qquad (8)
\end{aligned}
$$

If the sum over cells is set to zero, for each cell this gives an expression for the change in internal energy as

$$
m_c \frac{\mathrm{d}\varepsilon_c}{\mathrm{d}t} = - \sum_{p\in\mathscr{P}(c)} \mathbf{F}_{pc} \cdot \mathbf{U}_p, \qquad (9)
$$

such that (8) is true and total energy conservation is preserved. This is the semi-discrete form of internal energy equation which was derived from the total energy conservation.

Provided that the subcell force $\mathbf{F}_{pc}$ is known, the numerical scheme is defined by equations for the velocity (6), specific internal energy (9) and density (4) defined from the mesh motion. The mesh motion is modeled by the set of ordinary differential equations $\mathrm{d}\mathbf{X}_p/\mathrm{d}t = \mathbf{U}_p$ being solved at each mesh point p for its position $\mathbf{X}_p(t)$. Remark that whatever subcell force $\mathbf{F}_{pc}$ one wishes to consider (pressure force, viscosity, elastic-plastic contribution, etc.), the conservation of discrete momentum and total energy as defined by (7) is fulfilled. In other words, the mechanism responsible for the conservativeness is independent of the way the forces are constructed.

The subcell force is an combination of three forces: a pressure force $\mathbf{F}^p_{pc}$ that approximates grad p in the momentum equation (2), a subzonal pressure force $\mathbf{F}^{\delta p}_{pc}$ designed to prevent the Hourglass mesh motion and an artificial viscosity force $\mathbf{F}^v_{pc}$ designed to treat shock waves

$$
\mathbf{F}_{pc} = \mathbf{F}^p_{pc} + \mathbf{F}^{\delta p}_{pc} + \mathbf{F}^v_{pc}. \qquad (10)
$$

The pressure force in subcell Ω_{pc} with boundary $\partial\Omega_{pc}$ is given by

$$
\mathbf{F}^p_{pc} = - \int_{\Omega_{pc}} \operatorname{grad} p \, \mathrm{d}V = - \int_{\partial\Omega_{pc}} p \, \mathbf{N} \, \mathrm{d}l. \qquad (11)
$$

The subzonal pressure force $\mathbf{F}^{\delta p}_{pc}$ [6,20], given by the difference between the subcell pressure and the cell pressure $p_{pc} - p_c$ and the geometry of the cell, acts against the Hourglass mode motion, which might invert the cell (moving cell is being inverted when its node crosses another edge of the cell).

The last part of the subcell force is the artificial viscosity devoted to deal with shock waves. The simplest viscosity in cell c, across which the velocity has a difference $\varDelta\mathbf{U}$, is in the compression regime $\varDelta\mathbf{U} < 0$ [3]

$$Q_c = c_1\rho_c a_c|\varDelta\mathbf{U}| + c_2\rho_c(\varDelta\mathbf{U})^2. \tag{12}$$

Constants c_1, c_2 are of the order of unity and a_c is the sound speed. In the expansion regime $\varDelta\mathbf{U} \geq 0$ viscosity is set to zero. The artificial viscosity has the dimension of pressure and is generally added to the classical cell pressure producing the viscosity force $\mathbf{F}^v_{pc}$ in the same way as pressure force (11), usually preventing spurious numerical oscillations on shock waves. Many formulations of artificial viscosity in more than one dimension use the form given above with multidimensional modifications as edge artificial viscosity [3, 5] or tensor artificial viscosity [2].

2.2 *Rezoning Phase*

The rezoning phase of the ALE method consists in moving the nodes of the Lagrangian mesh to improve the geometric quality of the grid while keeping the rezoned grid as close as possible to the Lagrangian grid. This constraint must be taken into account to maintain the accuracy of the computation gained by the Lagrangian phase and to minimize the error of the remap phase. If the Lagrangian phase produces non-valid (inverted) cells then we have to use an untangling procedure [27]. The rezoning phase of the ALE method covers mesh smoothing and untangling.

The mesh resulting from the Lagrangian step can be of low quality and smoothing process changes the mesh in a way to improve it. One of the simplest smoothing methods is Winslow smoothing method [28]. The new positions of the mesh nodes are computed (with possible iteration over l starting at the old Lagrangian mesh) in case of logically rectangular mesh as

$$\begin{aligned}\mathbf{X}^{l+1}_{i,j} = \frac{1}{2\,(\alpha^l + \gamma^l)} \Big(&\alpha^l\,(\mathbf{X}^l_{i,j+1} + \mathbf{X}^l_{i,j-1}) + \gamma^l\,(\mathbf{X}^l_{i+1,j} + \mathbf{X}^l_{i-1,j}) \\ &-\frac{1}{2}\,\beta^l\,(\mathbf{X}^l_{i+1,j+1} - \mathbf{X}^l_{i-1,j+1} + \mathbf{X}^l_{i-1,j-1} - \mathbf{X}^l_{i+1,j-1})\Big),\end{aligned}$$

where the coefficients $\alpha^l = x_\xi^2 + y_\xi^2$, $\beta^l = x_\xi\,x_\eta + y_\xi\,y_\eta$, $\gamma^l = x_\eta^2 + y_\eta^2$, and (ξ, η) are logical coordinates $\xi_i = i/n_x$, $\eta_j = j/n_y$ for $i = 0, \ldots, n_x$ and $j = 0, \ldots, n_y$. The derivatives x_ξ, x_η are approximated by the central differences $(x_\xi)_{i,j} \approx (x_{i+1,j} - x_{i-1,j})/(2\,\varDelta\xi)$, $(x_\eta)_{i,j} \approx (x_{i,j+1} - x_{i,j-1})/(2\,\varDelta\eta)$ and similarly for y.

Further rezoning methods include the condition number smoothing [10] and the Reference Jacobian Method [11].

2.3 Remapping Phase

The remapping stage of the ALE method is in fact a conservative interpolation of the discrete conserved quantities from the old Lagrangian mesh to the new smoother one. The remapping stage consists of three steps: reconstruction, integration and repair. First, the remapped conservative function g (e.g. density ρ) is reconstructed from the discrete values by a piecewise linear function on each old cell, e.g. with the Barth-Jespersen limiter [1]. Then, the reconstructed piecewise linear function is integrated over each new cell $\tilde{c}$ (objects related to the new mesh are accented by a tilde here) to get the total value $G_{\tilde{c}} = \int_{\tilde{c}} g \,\mathrm{d}x \,\mathrm{d}y$ of the conserved quantity (e.g. mass of the cell) inside the new cell, which defines the remapped density of conserved quantity $g_{\tilde{c}} = G_{\tilde{c}} / V_{\tilde{c}}$, where $V_{\tilde{c}}$ is the volume of the cell $\tilde{c}$.

The natural exact integration of the piecewise linear function over the new cell requires computing intersections of the new cell with all neighboring old cells. For example on logically rectangular mesh see Fig. 1 (a) where the new cell $\tilde{c}_{i,j} = [\tilde{P}_{i,j}, \tilde{P}_{i+1,j}, \tilde{P}_{i+1,j+1}, \tilde{P}_{i,j+1}]$ intersects with nine (3 × 3 patch) old cells $c_{k,l}, k = i-1, i, i+1, l = j-1, j, j+1$. The linear reconstruction at the old cell c'

$$g(x, y) = g_{c'} + \left(\frac{\partial g}{\partial x}\right)_{c'} (x - x_{c'}) + \left(\frac{\partial g}{\partial y}\right)_{c'} (y - y_{c'}),$$

(where $(x_{c'}, y_{c'})$ is the centroid of the cell c') inside each such intersection $I_{c'}^{\tilde{c}} = \tilde{c} \cap c'$ results in the contribution

$$\begin{aligned} G_{I_{c'}^{\tilde{c}}} = g_{c'} \int_{I_{c'}^{\tilde{c}}} \mathrm{d}x \,\mathrm{d}y + \left(\tfrac{\partial g}{\partial x}\right)_{c'} \left(\int_{I_{c'}^{\tilde{c}}} x \,\mathrm{d}x \,\mathrm{d}y - x_{c'} \int_{I_{c'}^{\tilde{c}}} \mathrm{d}x \,\mathrm{d}y \right) \\ + \left(\tfrac{\partial g}{\partial y}\right)_{c'} \left(\int_{I_{c'}^{\tilde{c}}} y \,\mathrm{d}x \,\mathrm{d}y - y_{c'} \int_{I_{c'}^{\tilde{c}}} \mathrm{d}x \,\mathrm{d}y \right) \end{aligned} \tag{13}$$

to the whole integral $G_{\tilde{c}}$. The integrals in this contribution over the polygonal intersection are transformed using Green's theorem into integrals over the edges of the polygonal intersection and computed analytically. The exact integration is computationally rather expensive because it requires finding all cell intersections.

The approximate integration over swept regions [14], which are the regions swept by the cell edges moving from the old mesh to the new position in the new mesh (see Fig. 1 (b)), is much faster. The contribution from each of the four swept regions has similar form as (13) with the intersection $I_{c'}^{\tilde{c}}$ replaced by the swept region. Green's theorem again transforms integrals over polygons into integrals over the edges of the polygon, which can be exactly evaluated. In the swept region method the integrals of the reconstructed function over the swept regions can be interpreted as remap fluxes through the mesh edges and the remapping formula can be written in a conservative flux form.

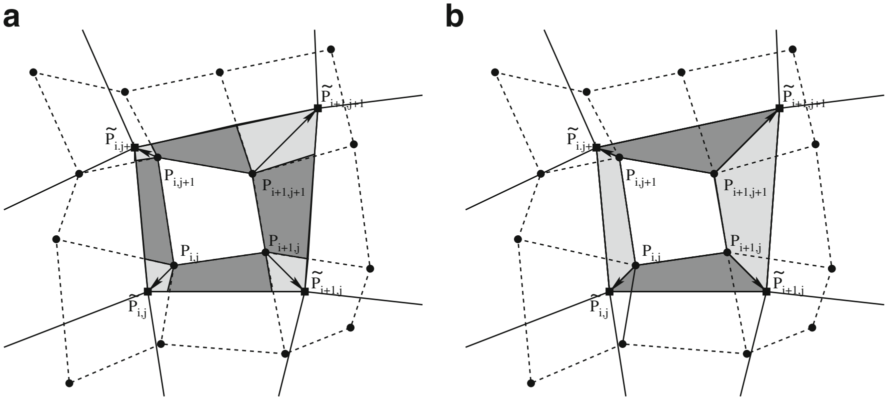


Fig. 1 Old (dashed) and new (solid segments) mesh with intersection regions for the exact integration (a) and swept regions for the approximate integration (b)

The last step of the remapping phase is repair [21] which conservatively redistributes conserved quantities in such a way that the remapping does not introduce any new local extrema. The repair is a post-processing *adhoc* correction. Better treatment, based on flux corrected transport (FCT) and called flux corrected remap [16, 17], guarantees that new local extrema are not introduced.

For the staggered scheme the remapping is however more complicated than what is outlined above as thermodynamical quantities are cell or subcell centered and velocity (thus also momentum) is centered at mesh nodes. This means that the internal energy is defined at cells and the kinetic energy at nodes and one has to be careful to conserve the total energy during the remapping step. The basic idea to treat this issue is to transform all quantities to subcells, remap on the subcell mesh and transform the quantities back to staggered form (density in subcells, internal energy in cells and velocity in nodes). The most accurate method employs the rigorously derived, matrix based, invertible transformation between the nodal and subcell velocities [20].

3 Heat Conductivity

The parabolic part of energy equation (3) is treated separately by splitting from the hyperbolic part of the whole system (1)-(3). It is transformed to the heat equation for temperature $a\partial T/\partial t = \mathrm{div}(\kappa \,\mathrm{grad}\, T)$ where $a = \rho\partial \epsilon/\partial T$. We write the heat equation as the first order system, in so-called flux form

$$a\frac{\partial T}{\partial t} - \mathrm{div}\,\mathbf{w} = 0, \quad \mathbf{w} = -\kappa\,\mathrm{grad}\,T, \tag{14}$$

introducing the heat flux $\mathbf{w}$. The heat equation is solved on domain V with boundary ∂V with Neumann boundary conditions.

We treat the space discretization of the heat equation by the mimetic method [25], which has been generalized to unstructured triangular meshes in [7]. The mimetic method introduces operators of generalized gradient $\mathbf{G}$ and extended divergence $\mathbf{D}$

$$\mathbf{G}T = -\kappa \,\mathrm{grad}\, T, \quad \mathbf{D}\,\mathbf{w} = \begin{cases} \mathrm{div}\,\mathbf{w} & \text{on} \quad V \\ -(\mathbf{w}, \mathbf{n}) & \text{on} \quad \partial V \end{cases}.$$

The integral properties of these operators are given by divergence Green formula and Gauss theorem. The divergence Green formula

$$\int_V \mathrm{div}\,\mathbf{w}\,\mathrm{d}\,V - \oint_{\partial V} (\mathbf{w}, \mathbf{n})\,\mathrm{d}\,S = 0 \tag{15}$$

can be restated as $(\mathbf{D}\,\mathbf{w}, 1)_H = 0$ where we use the inner product on space H of scalar functions

$$(u, v)_H = \int_V u\, v\,\mathrm{d}\,V + \oint_{\partial V} u\, v\,\mathrm{d}\,S. \tag{16}$$

Gauss theorem

$$\int_V T\,\mathrm{div}\,\mathbf{w}\,\mathrm{d}\,V - \oint T(\mathbf{w}, \mathbf{n})\,\mathrm{d}\,S + \int_V (\mathbf{w}, \kappa^{-1}\kappa\,\mathrm{grad}\,T)\mathrm{d}\,V = 0$$

can be restated as $(\mathbf{D}\mathbf{w}, T)_H = (\mathbf{w}, \mathbf{G}T)_{\mathbf{H}}$ where we use also the inner product on space $\mathbf{H}$ of vector functions

$$(\mathbf{A}, \mathbf{B})_{\mathbf{H}} = \int_V (\kappa^{-1}\mathbf{A}, \mathbf{B})\mathrm{d}\,V. \tag{17}$$

Gauss theorem states that the generalized gradient is the adjoint operator of the extended divergence $\mathbf{G} = \mathbf{D}^*$. The basic idea of the support operator mimetic method [25] is to mimic these two integral properties also in the discrete case on spaces of discrete functions. We discretize the temperature T inside each computational cell and the vector heat flux $\mathbf{w}$ at the center of each edge by its projections on the normal of the edge. This discretization of vector heat flux guarantees the continuity of normal flux through each edge. On the spaces of discrete scalar and vector functions the discrete analogs of the inner products (16) and (17) are defined. The discrete divergence is derived in a standard way from the discrete analog of (15) on a computational cell. This gives the discrete operator of extended divergence D inside computational domain, while on the boundary it is given by the discrete heat flux (up to sign). Now the discrete extended gradient G is constructed as the adjoint (in the discrete inner products on the whole domain) of the discrete extended divergence $G = D^*$. The discrete gradient constructed this way has a global stencil.

Now in the heat equation (14) we use this mimetic spatial discretization, i.e. the discrete extended divergence D and the discrete generalized gradient operators G. We employ the implicit scheme written in flux form

$$a\frac{T^{n+1}-T^{n}}{\Delta t}+D\mathbf{W}^{n+1}=0,\quad \mathbf{W}^{n+1}-GT^{n+1}=0. \tag{18}$$

We express $T^{n+1}=T^{n}-D\mathbf{W}^{n+1}$ and eliminate it from the second equation which gives us

$$(I+\Delta t/aGD)\mathbf{W}^{n+1}=GT^{n}.$$

This system with a global stencil can be transformed into a system with local stencil [25] having symmetric, positive definite matrix. The conjugate gradient method, preconditioned by the altered direction implicit method, is applied as effective iterative solver for this system resulting in the numerical heat fluxes.

For laser plasma these heat fluxes often produce physically unrealistic (too big) heat fluxes which cannot be carried by electrons carrying most of heat energy. Direct decrease of the heat flux magnitude (where needed) leads to temperature oscillations and checker board patterns, thus the heat flux limiting has to be performed differently. In the regions where unlimited heat flux violates physical limits the heat conductivity is decreased by the ratio of the unlimited flux magnitude and the heat flux limit. The heat equation it then solved again with the updated heat conductivity giving the final limited heat fluxes. Finally, having fluxes $\mathbf{W}^{n+1}$ the temperature T^{n+1} is computed from the first equation of (18). The presented numerical method for heat equation works well on bad quality meshes appearing often in Lagrangian simulations and it allows discontinuous diffusion coefficient.

4 Laser Absorption and Cylindrical Geometry

Laser absorption in plasma is modeled by the term div $\mathbf{I}$ in the internal energy equation (3). Important notion for laser absorption is a critical density, which for laser with wavelength λ is proportional to $1/\lambda^2$. The critical density defines the critical surface which is the surface of electron density being equal to critical density. The laser can propagate only in the sub-critical regions of plasma with electron density less than critical density. Typically most of the laser energy is absorbed into plasma around the critical surface. Laser absorption on critical surface assumes that laser propagates without damping and refraction till the critical surface where it is absorbed. The absorption term div $\mathbf{I}$ with absorption coefficient is evaluated in cells at the critical surface and is zero everywhere else.

Ray tracing is a more complicated method for laser absorption modeling. The laser beam is split into many laser rays carrying initially appropriate energy depending on radius and radial laser profile. Propagating of each ray is computed (traced) independently. Inside a cell through which the ray propagates it does not change direction and deposits a part of its energy into plasma internal energy by inverse bremsstrahlung. On the edge the ray refracts according to Snell law with refraction plane being orthogonal to the electron density gradient. A special case is full reflection near the critical surface when the ray on the edge reflects back.

Laser beam has cylindrical symmetry and most simulated problems are cylindrically symmetric (here all problems except oblique incidence on thin foil studied in Sect. 5.1), so one has to include cylindrical $r - z$ geometry. All numerical methods, initially designed in Cartesian geometry, have been generalized into cylindrical geometry with a special boundary condition on the symmetry axis z. In cylindrical geometry for Lagrangian method we employ control volume method [4], rezoning methods need only to change boundary treatment on the symmetry axis z, cylindrical remapping [13] requires additional factor r in the integrals (13, the mimetic method for heat conductivity has been generalized to cylindrical geometry and also laser absorption module supports both cylindrical and Cartesian geometries.

5 Interaction of Laser with Targets

In this section we present selected simulations of laser beam interaction with targets modeled by our ALE code PALE. All simulations correspond to particular experiments performed at PALS laser facility in Prague, which provides a laser beam on the first harmonics with wavelength $\lambda = 1.315$ nm or on the third harmonics with wavelength $\lambda = 438$ nm.

5.1 *Oblique Incidence on Thin Foil*

We start with oblique incidence of laser beam on a 0.8 μm thin Aluminum foil which is reasonably simple and provides initial insight into laser interactions with matter. This simulation is an initial study to double foil targets which are used for investigation of plasma-wall interactions [24] and which are subject of the next Sect. 5.2. The third harmonics Gaussian laser pulse with energy 36 J, full width half maximum (FWHM) length 250 ps and focal spot radius 40 μm interacts with 30° oblique thin foil (the angle between laser beam axis and normal to the foil is 30°). The simulation starts at time $t = 0$, which is 250 ps before the laser maximum at $t = 250$ ps, and uses Cartesian geometry as the setup is not cylindrically symmetric. Density of the developing laser plasma at three times 150, 200 and 250 ps is presented in Fig. 2 in a logarithmic scale with computational mesh and a magenta curves of the critical surface. Laser is coming from above with the beam axis on the z axis $r = 0$. It propagates through the sub-critical plasma until the critical surface and is absorbed on the critical surface. At time 150 ps in Fig. 2 (a) the laser does not penetrate the foil, while at time 200 ps in Fig. 2 (b) the laser has already burned through the foil and only a small part far from the z axis is still being

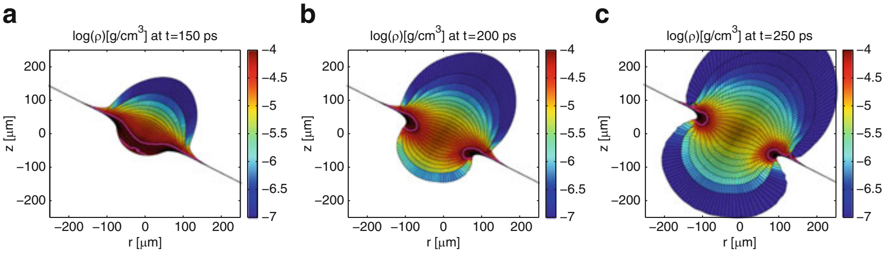


Fig. 2 Density for interaction of oblique laser beam with a thin Aluminum foil at time: (a) 150 ps, (b) 200 ps and (c) 250 ps. Magenta curves denote the position of the critical surface

absorbed at the critical surface. In the beginning of the interaction laser energy is being deposited close to the upper boundary of the foil which starts to expand in the upper right direction creating plasma plume (corona). Before time 150 ps the whole foil in an area around the z axis is heated and secondary plume starts to expand in the lower left direction. This simulation provides an example of a large scale change of computational domain (initial 0.8 μm thin foil expands to plumes of the size around 500 μm at Fig. 2 (c) at time of laser maximum, which is still not the end of simulation), which dictates the use of Lagrangian coordinates moving with the moving plasma. The simulation justifies that even with oblique laser incidence the plasma plumes propagation is orthogonal to the foil. This is going to be used for oblique incidence double foil targets where the laser going through the upper foil does not hit the shorter lower foil, which interacts directly with the plasma plume from the upper foil.

5.2 *Double Foil Target*

The double foil target shown in Fig. 3 (a) is composed from two parallel foils located at distance $L = 600$ μm. The thickness of the upper Aluminum foil is $d_u = 0.8$ μm and the thickness of the lower Magnesium foil is $d_l = 2$ μm. The double foil target is irradiated by the third harmonics Gaussian laser beam (orthogonal to the foils) with energy 115 J, FWHM length 300 ps, focal spot radius 40 μm and angular beam divergence 15°. The beam is focused on the lower foil. Laser absorption has been modeled in this simulation by ray tracing. Results are presented in Fig. 3(b) and (c) by density and pressure color-maps at time 600 ps. Fig. 3 (b) shows density color-map with selected laser rays in the left part. The thickness of the rays is proportional to the energy they carry, so when the energy goes below a threshold the thickness goes to zero and the ray curve ends. Rays are refracted around the critical surface, some rays are reflected from the z axis. In the beginning the upper foil expands in two plumes similarly as the oblique foil in Fig. 2. The lower foil remains static until the laser burns through the upper foil. After burning through the upper foil the laser

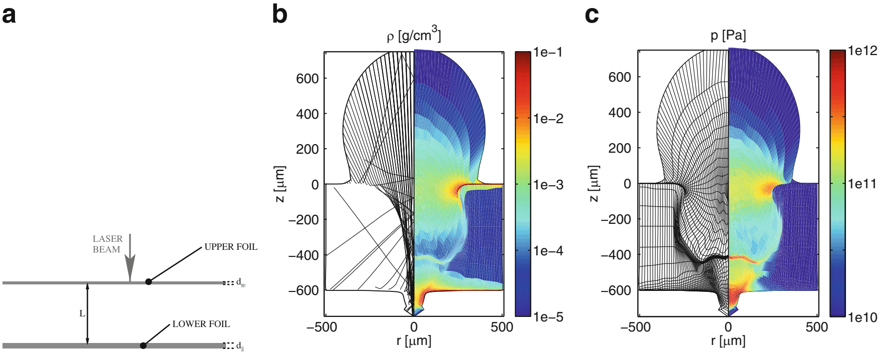


Fig. 3 Experimental setup for double foil target (a), density ρ with laser rays (b) and pressure p with computational mesh (c) for double foil target at $t = 500$ ps

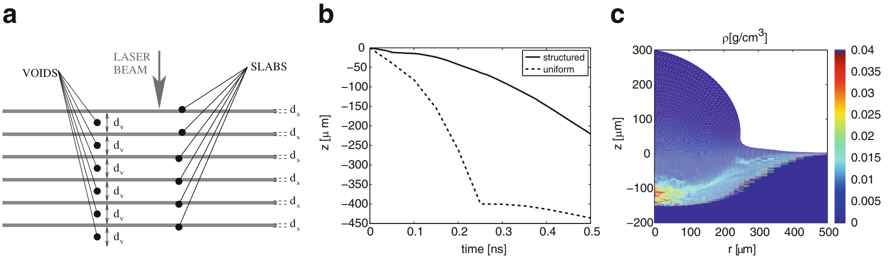


Fig. 4 Structured model of the foam (a), burning of laser through the foam target (b) and density from simulation with the structured target at $t = 400$ ps (c)

reaches also the lower foil from which at first the upper Magnesium plume starts to develop. The lower Aluminum plume collides with the upper Magnesium plume around time 500 ps producing high density and pressure at the colliding area.

5.3 Foam Target

Foam layers are used in the ICF targets for smoothing inhomogeneities in the laser beam by its propagation through the low density foam. Interaction of laser with low density foam is difficult to model because for low density homogeneous material one gets too high speed of laser burning through the foam. This problem can be avoided by introducing structured model of the foam [9] shown in Fig. 4 (a) consisting from a series of parallel high-density slabs separated by low-density voids. When the laser burns through this structured model of foam it is delayed on each slab as it needs some time to burn through the slab. Here, we simulate the interaction of the third harmonic Gaussian laser pulse of 320 ps FWHM duration, energy of 170 J and focal spot radius of 300 μm with 400 μm-thick layer of TAC foam of density 9.1 mg/cm^3

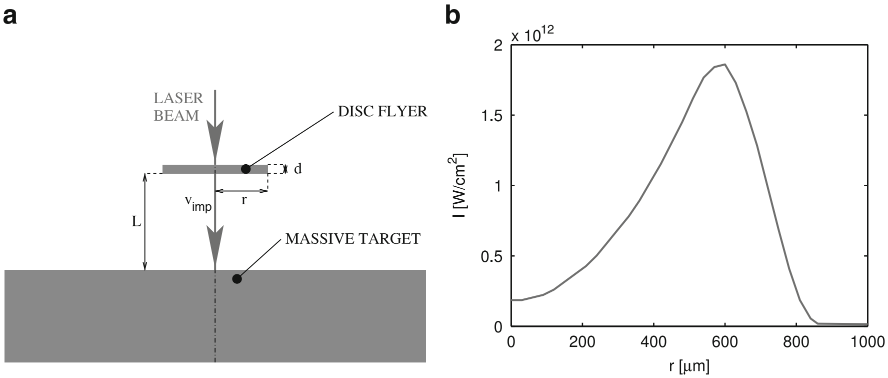


Fig. 5 Experimental setup for high velocity impact (a) and annular radial laser beam intensity profile at time of laser maximum for jet formation (b)

with 2 μm pores. The foam is modeled by uniform density 9.1 mg/cm^3 material and by structured model consisting from a sequence of $d_s = 0.018$ μm thick dense slabs of density $\rho_s = 1$ g/cm^3 separated by $d_v = 1.982$ μm thick voids with density $\rho_v = 1$ mg/cm^3. The time evolution of the depth of the burned region of foam on the z axis is plotted in Fig. 4 (b) for uniform and structured foam model (with density at $t = 400$ ps in Fig. 4 (c)). The speed of burning through for the structured model is reasonably close to the experimental measurement, while this speed for uniform model is more than twice higher.

5.4 High Velocity Impact

The target setup for the high velocity impact problem is shown in Fig. 5(a). A cylindrical Aluminum disc flyer with radius $r = 150$ μm and thickness $d = 11$ μm is placed at distance $L = 200\,\mu$m above an Aluminum massive target, parallel to it. The laser irradiated disc flyer is ablatively accelerated up to very high velocity (40-190 km/s) and impacts the massive target creating a crater in it. Here the third harmonics laser pulse of energy 390 J, FWHM length 400 ps and focal spot radius 125 μm interacts with the target. The simulation is split into two parts: ablative disc flyer acceleration by laser beam and the impact of disc flyer into the massive target. The density for the final time of disc acceleration is presented in Fig. 6 (a) and (b). In the presented case the final time of acceleration is about 1.1 ns, when the high density region, which contains most of the disc mass, momentum and energy and is seen in Fig. 6 (b), approaches the upper boundary of the massive target located at $z = -200$ μm. The average vertical velocity of the impacting disc is 187 km/s. A new mesh is constructed containing this region and the whole massive target. Conservative quantities at the final time of acceleration are remapped to this new

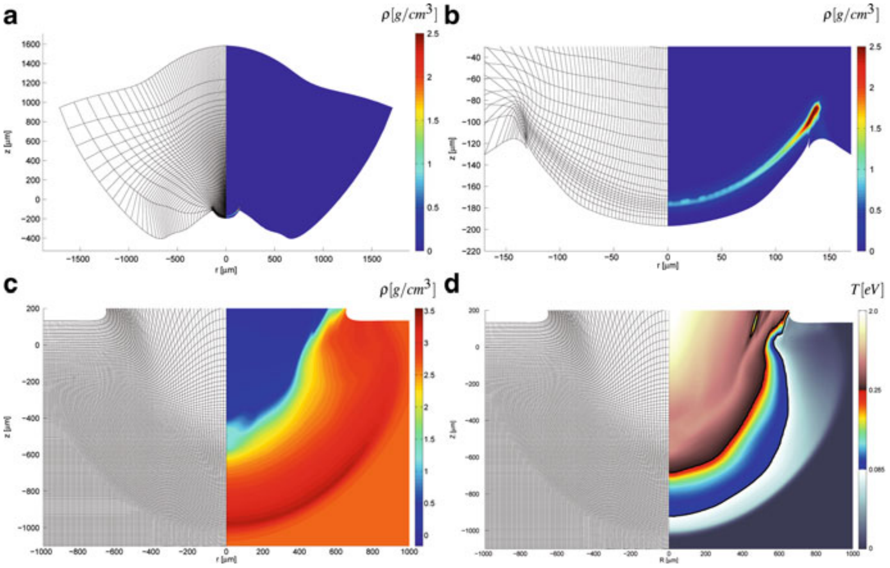


Fig. 6 High velocity impact problem: density (a) and zoomed density (b) of accelerated disc before the impact, density (c) and temperature (d) at the final time 80 ns after the impact (zoomed to crater) – three different color scales in temperature color-map distinguish solid (gray), liquid (blue-red) and gaseous (brown-pink) phase of Aluminum

mesh and serve as initial conditions for the second part of the simulation, disc flyer impact. The impact creates circular shock wave propagating into the massive target and visible in density and temperature color-maps in Figs. 6 (c),(d). A large hot low-density plasma plume is being reflected from the massive target. The impact melts and evaporates part of massive target creating a crater. Solid, liquid and gas phase of Aluminum are distinguished by three color-maps in Fig. 6 (d). Solid-liquid and liquid-gas phase interfaces, given by temperature isolines (corresponding to Aluminum melting and boiling temperature), are visible in Fig. 6 (d) as two white-black (in bottom-up direction on the z axis) interfaces. The crater is defined by gas - liquid phase interface. Simulated craters size and shape correspond reasonably well to experimental data also for other laser energies and other disc flyers [12].

5.5 Jet Formation

In this section we investigate formation of plasma jets by interaction of annular laser beam with a massive Aluminum target. We use Gaussian in time laser pulse on 3-rd harmonics with FWHM length 400 ps and energy 10 J. The radial intensity profile of the annular beam is presented in Fig. 5(b). It has 10% minimum on the z axis at $r = 0$, it is proportional to r^2 for small r and has a smooth maximum

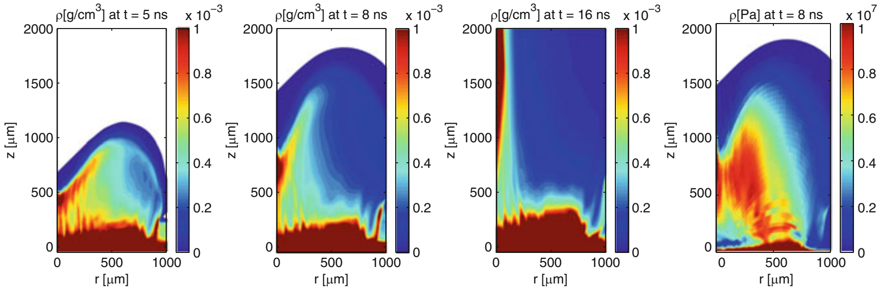


Fig. 7 Plasma jet formation by annular laser beam: (a), (b), (c) density evolution at times 5, 8 and 16 ns; (d) pressure at 8 ns

around $r = 600\ \mu$m. The density evolution at times 5, 8 and 16 ns is presented in Fig. 7 (a),(b),(c). Thanks to the annular radial laser profile the plasma plume develops and expands faster around the radial maximum of intensity at $r = 600\mu$m, than around the z axis at $r = 0$. Such plume development leads to cone profile of higher density region visible in Fig. 7 (a) and (b) at 5 and 8 ns. The cone moves up in z direction and left in r direction towards the z axis and collides on the symmetry axis creating a plasma jet which can be seen in Fig. 7 (c) at 16 ns as high density, high pressure region along the z axis propagating up. Important is the radial pressure gradient on the cone directed inwards towards the z axis which drives the negative radial velocity towards the z axis, which can be seen in Fig. 7 (d). The outlined dynamics of the plasma plume created by annular laser provides pure hydrodynamical mechanism for plasma jets generation [15]. Plasma jets appear not only on the laser plasma micro-scale presented here, but also astrophysics deals with giant jets on macro-scale.

6 Conclusion

Numerical methods for Lagrangian and ALE hydrodynamics, heat conductivity and laser absorption used in our code PALE have been shortly presented. PALE code have been applied to simulate selected problems of laser interaction with targets. For most simulations we have to use the ALE method as pure Lagrangian computation (without any rezoning and remapping) fails due to severe distortion of moving computational mesh. Only the first simulation of oblique incidence on a thin foil presented in section 5.1 has been computed by pure Lagrangian method. PALE code is regularly used for simulations of experiments at PALS laser facility. The simulations provide theoretical backgound for interpretation of experimetal results.

Acknowledgements This research has been supported in part by the Czech Ministry of Education grants MSM6840770022 and LC528, Czech Science Foundation grant GAP205/10/0814 and Czech Technical University grant SGS10/299/OHK4/3T/14.

The authors thank M. Shashkov, B. Wendroff, R. Loubere, P.-H. Maire, V. Kmetik, R. Garimella, M. Berndt for fruitful discussions and constructive comments.

References

1. Barth, T., Jespersen, D.: The design and application of upwind schemes on unstructured meshes. Tech. Rep. AIAA-89-0366, AIAA, NASA Ames Research Center (1989)
2. Campbell, J., Shashkov, M.: A tensor artificial viscosity using a mimetic finite difference algorithm. J. Comput. Phys. **172**(2), 739–765 (2001)
3. Caramana, E., Shashkov, M.J., Whalen, P.: Formulations of artificial viscosity for multi-dimensional shock wave computations. J. Comput. Phys. **144**, 70–97 (1998)
4. Caramana, E.J., Burton, D.E., Shashkov, M.J., Whalen, P.P.: The construction of compatible hydrodynamics algorithms utilizing conservation of total energy. J. Comput. Phys. **146**(1), 227–262 (1998)
5. Caramana, E.J., Loubère, R.: "curl-q": A vorticity damping artificial viscosity for Lagrangian hydrodynamics calculations. J. Comput. Phys. **215**(2), 385–391 (2006)
6. Caramana, E.J., Shashkov, M.J.: Elimination of artificial grid distortion and hourglass-type motions by means of Lagrangian subzonal masses and pressures. J. Comput. Phys. **142**, 521–561 (1998)
7. Ganzha, V., Liska, R., Shashkov, M., Zenger, C.: Mimetic finite difference methods for diffusion equations on unstructured triangular grid. In: M. Feistauer, V. Dolejší, P. Knobloch, K. Najzar (eds.) Numerical Mathematics and Advanced Applications ENUMATH 2003, pp. 368–377. Springer-Verlag, Berlin (2004)
8. Hirt, C., Amsden, A., Cook, J.: An arbitrary Lagrangian-Eulerian computing method for all flow speeds. J. Comput. Phys. **14**, 227–253 (1974). Reprinted in vol. 135(2), 203–216, 1997.
9. Kapin, T., Kuchařík, M., Limpouch, J., Liska, R.: Hydrodynamic simulations of laser interactions with low-density foams. Czechoslovak Journal of Physics **56**, B493–B499 (2006)
10. Knupp, P.: Achieving finite element mesh quality via optimization of the Jacobian matrix norm and associated quantities. Part I – a framework for surface mesh optimization. Int. J. Numer. Meth. Eng. **48**, 401–420 (2000)
11. Knupp, P., Margolin, L., Shashkov, M.: Reference Jacobian optimization-based rezone strategies for arbitrary Lagrangian Eulerian methods. J. Comput. Phys. **176**, 93–128 (2002)
12. Kuchařík, M., Limpouch, J., Liska, R.: Laser plasma simulations by arbitrary Lagrangian Eulerian method. J. de Physique IV **133**, 167–169 (2006)
13. Kuchařík, M., Liska, R., Loubere, R., Shashkov, M.: Arbitrary Lagrangian-Eulerian (ALE) method in cylindrical coordinates for laser plasma simulations. In: S. Benzoni-Gavage, D. Serre (eds.) Hyperbolic Problems: Theory, Numerics, Applications, pp. 687–694. Springer (2008)
14. Kuchařík, M., Shashkov, M., Wendroff, B.: An efficient linearity-and-bound-preserving remapping method. J. Comput. Phys. **188**(2), 462–471 (2003)
15. Limpouch, J., Liska, R., Kuchařík, M., Váchal, P., Kmetík, V.: Laser-driven collimated plasma flows studied via ALE code. In: 37th EPS Conference on Plasma Physics, pp. P4.222, 1–4. European Physical Society, Mulhouse (2010)
16. Liska, R., Shashkov, M., Váchal, P., Wendroff, B.: Optimization-based synchronized flux-corrected conservative interpolation (remapping) of mass and momentum for arbitrary Lagrangian-Eulerian methods. J. Comput. Phys. **229**(5), 1467–1497 (2010)
17. Liska, R., Shashkov, M., Váchal, P., Wendroff, B.: Synchronized flux corrected remapping for ale methods. Computers and Fluids (2011). DOI: 10.1016/j.compfluid.2010.11.013
18. Liska, R., Shashkov, M., Wendroff, B.: Lagrangian composite schemes on triangular unstructured grids. In: M. Kočandrlová, V. Kelar (eds.) Mathematical and Computer Modelling in Science and Engineering, pp. 216–220. Prague (2003)

19. Loubere, R., Ovadia, J., Abgrall, R.: A Lagrangian discontinuous Galerkin type method on unstructured meshes to solve hydrodynamics problems. Int. J. Numer. Meth. Fluids **44**(6), 645–663 (2004)
20. Loubère, R., Shashkov, M.: A subcell remapping method on staggered polygonal grids for arbitrary-Lagrangian-Eulerian methods. J. Comput. Phys. **209**(1), 105–138 (2005)
21. Loubère, R., Staley, M., Wendroff, B.: The repair paradigm: New algorithms and applications to compressible flow. J. Comput. Phys. **211**(2), 385–404 (2006)
22. Maire, P.H.: A high-order cell-centered Lagrangian scheme for two-dimensional compressible fluid flows on unstructured meshes. J. Comput. Phys. **228**(7), 2391–2425 (2009)
23. Maire, P.H., Abgrall, R., Breil, J., Ovadia, J.: A cell-centered Lagrangian scheme for two-dimensional compressible flow problems. SIAM Journal on Scientific Computing **29**(4), 1781–1824 (2007)
24. Renner, O., Liska, R., Rosmej, F.: Laser-produced plasma-wall interaction. Laser and Particle Beams **27**(4), 725–731 (2009)
25. Shashkov, M., Steinberg, S.: Solving diffusion equation with rough coefficients in rough grids. J. Comput. Phys. **129**, 383–405 (1996)
26. Shashkov, M., Wendroff, B.: A composite scheme for gas dynamics in Lagrangian coordinates. J. Comput. Phys. **150**, 502–517 (1999)
27. Váchal, P., Garimella, R., Shashkov, M.: Untangling of 2D meshes in ALE simulations. J. Comput. Phys. **196**(2), 627–644 (2004)
28. Winslow, A.: Equipotential zoning of two-dimensional meshes. Tech. Rep. UCRL-7312, Lawrence Livermore National Laboratory (1963)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

A two-dimensional finite volume solution of dam-break hydraulics over erodible sediment beds

Fayssal Benkhaldoun, Imad Elmahi, Saïda Sari, and Mohammed Seaïd

Abstract Two-dimensional dam-break hydraulics over erodible sediment beds are solved using a well-balanced finite volume method. The governing equations consist of three coupled model components: (i) the shallow water equations for the hydrodynamical model, (ii) a transport equation for the dispersion of suspended sediments, and (iii) an Exner equation for the morphological model. These coupled models form a hyperbolic system of conservation laws with source terms. The proposed finite volume method consists of a predictor stage for the discretization of gradient terms and a corrector stage for the treatment of source terms. The gradient fluxes are discretized using a modified Roe's scheme using the sign of the Jacobian matrix in the coupled system. A well-balanced discretization is used for the treatment of source terms. In this paper, we also describe an adaptive procedure in the finite volume method by monitoring the concentration of suspended sediments in the computational domain during its transport process. The method uses unstructured meshes, incorporates upwinded numerical fluxes and slope limiters to provide sharp resolution of steep sediment concentration and bed-load gradients that may form in the approximate solution.

Keywords Sediment transport, shallow water equations, finite volume method
MSC2010: 35L65, 65L08, 65C20

Fayssal Benkhaldoun and Saïda Sari
LAGA, Université Paris 13, 99 Av J.B. Clement, 93430 Villetaneuse, France,
e-mail: fayssal@math.univ-paris13.fr, sari@math.univ-paris13.fr

Imad Elmahi
ENSAO Complex Universitaire, B.P. 669, 60000 Oujda, Morocco, e-mail: ielmahi@ensa.univ-oujda.ac.ma

Mohammed Seaïd
School of Engineering and Computing Sciences, University of Durham, South Road, Durham
DH1 3LE, UK, e-mail: m.seaid@durham.ac.uk

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_88, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

The main concern of the sediment transport (or morphodynamics) is to determine the evolution of bed levels for hydrodynamics systems such as rivers, estuaries, bays and other nearshore regions where water flows interact with the bed geometry. Example of applications include among others, beach profile changes due to severe wave climates, seabed response to dredging procedures or imposed structures, and harbour siltation. The ability to design numerical methods able to predict the morphodynamics evolution of the coastal seabed has a clear mathematical and engineering relevances. In practice, morphodynamics involve coupling between a hydrodynamics model, which provides a description of the flow field leading to a specification of local sediment transport rates, and an equation for bed level change which expresses the conservative balance of sediment volume and its continual redistribution with time. Here, the hydrodynamic model is described by the shallow water equations, the bed-load is modelled by the Exner equation, and the suspended sediment transport is modelled by an advection equation accounting for erosion and deposition effects. The coupled models form a hyperbolic system of conservation laws with a source term. Nowadays, much effort has been devoted to develop numerical schemes for morphodynamics models able to resolve all hydrodynamics and morphodynamics scales. In the current study, a class of finite volume methods is proposed for numerical simulation of transient flows involving erosion and deposition of sediments. The method consists of a predictor stage where the numerical fluxes are constructed and a corrector stage to recover the conservation equations. The sign matrix of the Jacobian matrix is used in the reconstruction of the numerical fluxes. Most of these techniques have been recently investigated in [1,2] for solving sediment transport models without accounting for erosion and deposition effects. The current study presents an extension of this method to transient flows involving erosion and deposition of sediments. A detailed formulation of the sign matrix and the numerical fluxes is presented. The proposed method also satisfies the property of well-balancing flux-gradient and source-term in the system. Numerical results and comparisons will be shown for several suspended sediment transport problems.

2 The governing equations

In the current study, the sediment transport model consists of three parts: A hydraulic variables describing the motion of water, a concentration variable describing the dispersion of suspended sediments, and a morphology variable which describes the deformation of the bed-load. In the present work we assume that the flow is almost horizontal, the vertical component of the acceleration is vanishingly small, the pressure is taken to be hydrostatic, the free-surface gravity waves are long with respect to the mean flow depth and wave amplitude, and the water-species mixture is vertically homogeneous and non-reactive. The governing equations are obtained by balancing the net inflow of mass, momentum and species through boundaries of a control volume during an infinitesimal time interval while accounting for the accumulation of mass, resultant forces and species within the control volume, compare for example [1, 17] among others. Thus, the equations for mass conservation and momentum flux balance are given by

$$\begin{aligned}
&\frac{\partial h}{\partial t} + \frac{\partial (hu)}{\partial x} + \frac{\partial (hv)}{\partial y} = \frac{E-D}{1-p},\\
&\frac{\partial (hu)}{\partial t} + \frac{\partial}{\partial x}\left(hu^2 + \frac{1}{2}gh^2\right) + \frac{\partial}{\partial y}(huv) = gh\left(-\frac{\partial Z}{\partial x} - S_f^x\right) - \frac{(\rho_s-\rho_w)}{2\rho}gh^2\frac{\partial c}{\partial x}\\
&\qquad\qquad - \frac{(\rho_0-\rho)(E-D)}{\rho(1-p)}u, \qquad (1)\\
&\frac{\partial (hv)}{\partial t} + \frac{\partial}{\partial x}(huv) + \frac{\partial}{\partial y}\left(hv^2 + \frac{1}{2}gh^2\right) = gh\left(-\frac{\partial Z}{\partial y} - S_f^y\right) - \frac{(\rho_s-\rho_w)}{2\rho}gh^2\frac{\partial c}{\partial y}\\
&\qquad\qquad - \frac{(\rho_0-\rho)(E-D)}{\rho(1-p)}v,
\end{aligned}$$

where t is the time variable, $\mathbf{x} = (x, y)^T$ the space coordinates, $\mathbf{u} = (u, v)^T$ the depth-averaged water velocity, h the water depth, Z the bottom topography, g the gravitational acceleration, p the porosity, ρ_w the water density, ρ_s the sediment density, c is the depth-averaged concentration of the suspended sediment, E and D represent the entrainment and deposition terms in upward and downward directions, respectively. In (1), ρ and ρ_0 are respectively, the density of the water-sediment mixture and the density of the saturated bed defined by

$$\begin{aligned}
\rho &= \rho_w(1-c) + \rho_s c,\\
\rho_0 &= \rho_w p + \rho_s(1-p).
\end{aligned} \qquad (2)$$

The friction slopes S_f^x and S_f^y are defined, using the Manning roughness coefficient n_b, as

$$S_f^x = \frac{n_b^2}{h^{4/3}}u\sqrt{u^2+v^2}, \qquad S_f^y = \frac{n_b^2}{h^{4/3}}v\sqrt{u^2+v^2}. \qquad (3)$$

The equation for mass conservation of species is modeled by

$$\frac{\partial (hc)}{\partial t} + \frac{\partial}{\partial x}(huc) + \frac{\partial}{\partial y}(hvc) = E - D. \qquad (4)$$

To determine the entrainment and deposition terms in the above equations we assume a non-cohesive sediment and we use empirical relations reported in [8]. Thus,

$$D = w(1 - C_a)^m C_a, \tag{5}$$

where w is the settling velocity of a single particle in tranquil water

$$\omega = \frac{\sqrt{(36\nu/d)^2 + 7.5\rho_s gd - 36\nu/d}}{2.8}, \tag{6}$$

with ν is the kinematic viscosity of the water, d the averaged diameter of the sediment particle, m an exponent indicating the effects of hindered settling due to high sediment concentrations, C_a the near-bed volumetric sediment concentration, $C_a = \alpha_c c$, where α_c is a coefficient larger than unity. To ensure that the near-bed concentration does not exceed $(1 - p)$, the coefficient α_c is computed by [10]

$$\alpha_c = \min\left(2, \frac{1-p}{c}\right).$$

For the entrainment of a cohesive material the following relation is used

$$E = \begin{cases} \varphi \dfrac{\theta - \theta_c}{h} \overline{u} d^{-0.2}, & \text{if } \theta \geq \theta_c, \\ 0, & \text{otherwise,} \end{cases} \tag{7}$$

where

$$\overline{u} = \sqrt{u^2 + v^2},$$

and φ is a coefficient to control the erosion forces, θ_c is a critical value of Shields parameter for the initiation of sediment motion and θ is the Shields coefficient defined by

$$\theta = \frac{u_*^2}{sgd}, \tag{8}$$

with u_* is the friction velocity defined using the Darcy-Weisbach friction factor f as

$$u_*^2 = \sqrt{\frac{f}{8}}\overline{u}.$$

In (8), s is the submerged specific gravity of sediment given by

$$s = \frac{\rho_s}{\rho_w} - 1.$$

To update the bedload, we consider the Exner equation proposed in [14]

$$\frac{\partial Z}{\partial t} + \frac{A_s}{1-p}\frac{\partial}{\partial x}\big(u(u^2 + v^2)\big) + \frac{A_s}{1-p}\frac{\partial}{\partial y}\big(v(u^2 + v^2)\big) = -\frac{E - D}{1-p}, \tag{9}$$

where A_s is a coefficient usually obtained from experiments taking into account the grain diameter and the kinematic viscosity of the sediments. For simplicity in the presentation, let us rewrite the equations (1), (4) and (9) in the following vector form

$$\frac{\partial \mathbf{W}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{W})}{\partial x} + \frac{\partial \mathbf{G}(\mathbf{W})}{\partial y} = \mathbf{S}(\mathbf{W}) + \mathbf{Q}(\mathbf{W}), \tag{10}$$

where $\mathbf{W}$ is the vector of conserved variables, $\mathbf{F}$ and $\mathbf{G}$ are the physical fluxes in x- and y-direction, $\mathbf{S}$ and $\mathbf{Q}$ are the source terms. These variables are defined as

$$\mathbf{W} = \begin{pmatrix} h \\ hu \\ hv \\ hc \\ Z \end{pmatrix}, \quad \mathbf{F}(\mathbf{W}) = \begin{pmatrix} hu \\ hu^2 + \frac{1}{2}gh^2 \\ huv \\ huc \\ \frac{A_s}{1-p}u(u^2+v^2) \end{pmatrix}, \quad \mathbf{G}(\mathbf{W}) = \begin{pmatrix} hv \\ huv \\ hv^2 + \frac{1}{2}gh^2 \\ hvc \\ \frac{A_s}{1-p}v(u^2+v^2) \end{pmatrix},$$

$$\mathbf{S} = \begin{pmatrix} 0 \\ -gh\frac{\partial Z}{\partial x} - \frac{(\rho_s - \rho_w)}{2\rho}gh^2\frac{\partial c}{\partial x} \\ -gh\frac{\partial Z}{\partial y} - \frac{(\rho_s - \rho_w)}{2\rho}gh^2\frac{\partial c}{\partial y} \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{Q} = \begin{pmatrix} \frac{E-D}{1-p} \\ -ghS_f^x - \frac{(\rho_0 - \rho)(E-D)}{\rho(1-p)}u \\ -ghS_f^y - \frac{(\rho_0 - \rho)(E-D)}{\rho(1-p)}v \\ E-D \\ -\frac{E-D}{1-p} \end{pmatrix}.$$

It is worth emphasizing that, using the Exner equation (9) to model the bed-load transport, the nonhomegenuous terms in the right-hand side in (10) are not standard source terms but nonconservative products, since they include derivatives of two of the variables. The presence of these terms in sediment transport system can cause sever difficulties in their numerical approximations. In principle, the nonhomegenuous term in these equations can be viewed as a source term and/or a nonconservative term. In the approach presented in this study these terms are considered and discretized as source terms.

3 The finite volume method

The governing sediment transport equations (10) are formulated in Cartesian coordinates and will be discretized into the unstructured grids by the finite volume method. The unstructured grids are polygons and the number of edges of the grids is not limited in theory, but only triangular grids are considered in the current study. Hence, we divide the time interval into sub-intervals $[t_n, t_{n+1}]$ with stepsize Δt and discretize the spatial domain in conforming triangular elements $\mathscr{T}_i$. Each triangle represents a control volume and the variables are located at the geometric centres of the cells. Hence, using the control volume depicted in Fig. 1, a finite volume discretization of (10) yields

$$\mathbf{W}_i^{n+1} = \mathbf{W}_i^n - \frac{\Delta t}{|\mathscr{T}_i|} \sum_{j \in N(i)} \int_{\Gamma_{ij}} \mathscr{F}(\mathbf{W}^n; \mathbf{n}) \, d\sigma + \frac{\Delta t}{|\mathscr{T}_i|} \int_{\mathscr{T}_i} \mathbf{S}(\mathbf{W}^n) \, dV \\ + \frac{\Delta t}{|\mathscr{T}_i|} \int_{\mathscr{T}_i} \mathbf{Q}(\mathbf{W}^n) \, dV, \tag{11}$$

where $N(i)$ is the set of neighboring triangles of the cell $\mathscr{T}_i$, $\mathbf{W}_i^n$ is an averaged value of the solution $\mathbf{W}$ in the cell $\mathscr{T}_i$ at time t_n,

$$\mathbf{W}_i = \frac{1}{|\mathscr{T}_i|} \int_{\mathscr{T}_i} \mathbf{W} \, dV,$$

where $|\mathscr{T}_i|$ denotes the area of $\mathscr{T}_i$ and $\mathscr{S}_i$ is the surface surrounding the control volume $\mathscr{T}_i$. Here, Γ_{ij} is the interface between the two control volumes $\mathscr{T}_i$ and $\mathscr{T}_j$, $\mathbf{n} = (n_x, n_y)^T$ denotes the unit outward normal to the surface $\mathscr{S}_i$, and

$$\mathscr{F}(\mathbf{W}; \mathbf{n}) = \mathbf{F}(\mathbf{W}) n_x + \mathbf{G}(\mathbf{W}) n_y.$$

To deal with the source terms $\mathbf{Q}$, a standard splitting procedure (see for instance [2]) is employed for the discrete system (11) as

$$\begin{aligned} \mathbf{W}_i^{*} &= \mathbf{W}_i^n - \frac{\Delta t}{|\mathscr{T}_i|} \sum_{j \in N(i)} \int_{\Gamma_{ij}} \mathscr{F}(\mathbf{W}^n; \mathbf{n}) \, d\sigma + \frac{\Delta t}{|\mathscr{T}_i|} \int_{\mathscr{T}_i} \mathbf{S}(\mathbf{W}^n) \, dV, \\ \mathbf{W}_i^{n+1} &= \mathbf{W}_i^{*} + \frac{\Delta t}{|\mathscr{T}_i|} \int_{\mathscr{T}_i} \mathbf{Q}(\mathbf{W}^{*}) \, dV. \end{aligned} \tag{12}$$

Note that the time splitting (12) is only first-order accurate. A second-order splitting for the system (11) can be derived analogously using the Strang method [16]. The finite volume discretization (11) is complete once the gradient fluxes $\mathscr{F}(\mathbf{W}; \mathbf{n})$ and a discretization of source terms $\mathbf{Q}(\mathbf{W}^n)$ and $\mathbf{S}(\mathbf{W}^n)$ are well defined.

For the discretization of the gradient fluxes we consider a modified Roe's method studied in [3–6] among others. The method consists of the predictor-corrector

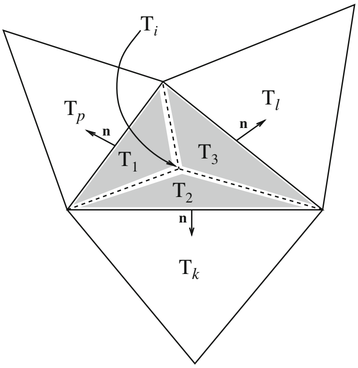


Fig. 1 A generic control volume $\mathscr{T}_i$ and notations

procedure

$$\begin{aligned} \mathbf{U}_{ij}^n &= \frac{1}{2}\left(\mathbf{U}_i^n + \mathbf{U}_j^n\right) - \frac{1}{2}\,\mathrm{sgn}\Big[\mathbf{A}_\eta\left(\overline{\mathbf{U}}\right)\Big]\left(\mathbf{U}_j^n - \mathbf{U}_i^n\right), \\ \mathbf{W}_i^{n+1} &= \mathbf{W}_i^n - \frac{\Delta t}{|\mathscr{T}_i|}\sum_{j\in N(i)} \mathscr{F}\left(\mathbf{W}_{ij}^n;\eta_{ij}\right)\left|\Gamma_{ij}\right| + \Delta t\mathbf{S}_i^n, \end{aligned} \tag{13}$$

where

$$\mathbf{U} = \begin{pmatrix} h \\ u_\eta \\ u_\tau \\ c \\ Z \end{pmatrix}, \quad \mathbf{A}_\eta(\mathbf{U}) = \begin{pmatrix} u_\eta & h & 0 & 0 & 0 \\ g & u_\eta & 0 & \dfrac{(\rho_s-\rho_w)}{2\rho}gh & g \\ 0 & 0 & u_\eta & 0 & 0 \\ 0 & 0 & 0 & u_\eta & 0 \\ 0 & \dfrac{A_s}{1-p}(3\,u_\eta^2+u_\tau^2) & 2\dfrac{A_s}{1-p}u_\eta\,u_\tau & 0 & 0 \end{pmatrix}.$$

the normal velocity $u_\eta = un_x + vn_y$ and tangential velocity $u_\tau = un_y - vn_x$. The sign matrix of the Jacobian is defined as

$$\mathrm{sgn}\Big[\nabla\mathbf{F}_\eta\left(\overline{\mathbf{U}}\right)\Big] = \mathscr{R}(\overline{\mathbf{U}})\,\mathrm{sgn}\Big[\Lambda(\overline{\mathbf{U}})\Big]\mathscr{R}^{-1}(\overline{\mathbf{U}}),$$

with $\Lambda(\overline{\mathbf{U}})$ is the diagonal matrix of eigenvalues, and $\mathscr{R}(\overline{\mathbf{U}})$ is the right eigenvector matrix. These matrices can be explicitly expressed using the associated eingenvalues of $\mathbf{A}_\eta(\mathbf{U})$. The sign matrix can be formulated in the same manner as in [3–6] and details are omitted here. In (13), $\overline{\mathbf{U}}$ is the Roe's average state given by

$$\overline{\mathbf{U}} = \begin{pmatrix} \dfrac{h_i + h_j}{2} \\ \dfrac{u_i\sqrt{h_i} + u_j\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}\eta_x + \dfrac{v_i\sqrt{h_i} + v_j\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}\eta_y \\ -\dfrac{u_i\sqrt{h_i} + u_j\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}\eta_y + \dfrac{v_i\sqrt{h_i} + v_j\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}}\eta_x \\ \dfrac{c_i\sqrt{h_i} + c_j\sqrt{h_j}}{\sqrt{h_i} + \sqrt{h_j}} \\ \dfrac{Z_i + Z_j}{2} \end{pmatrix}. \tag{14}$$

Next we discuss the treatment of source terms $\mathbf{S}_i^n$ in the proposed finite volume scheme and also the extension of the scheme to a second-order accuracy. An adaptive procedure is also described in this section.

3.1 Treatment of the source term

The treatment of the source terms in the shallow water equations presents a challenge in many numerical methods. In our scheme, the source term approximation $\mathbf{S}_i^n$ in the corrector stage is reconstructed such that the still-water equilibrium (C-property) is satisfied. Here, a numerical scheme is said to satisfy the C-property for the equations (10) if the condition

$$E - D = 0, \qquad u = 0, \qquad Z = \bar{Z}(x), \qquad h + Z = H, \qquad \rho = C, \tag{15}$$

holds for stationary flows at rest. In (15), H and C are nonnegative constants. Therefore, the treatment of source terms in (13) is reconstructed such that the condition (15) is preserved at the discretized level. Remark that the last condition in (15) means that at the equilibrium the sediment medium is assumed to be saturated. Furthermore, from the density equation (2), a constant density is equivalent to a constant concentration c. Hence, $\mathbf{S}_i^n$ should be consistent discretization of the source term in (13) defined as

$$\mathbf{S}_i^n = \begin{pmatrix} 0 \\ -g\bar{h}_{xi}^n \displaystyle\sum_{j\in N(i)} Z_{ij} n_{xij} \left|\Gamma_{ij}\right| - \dfrac{(\rho_s - \rho_w)}{2\rho} g\left(\bar{h}_{xi}^n\right)^2 \displaystyle\sum_{j\in N(i)} c_{ij} n_{xij} \left|\Gamma_{ij}\right| \\ -g\bar{h}_{yi}^n \displaystyle\sum_{j\in N(i)} Z_{ij} n_{yij} \left|\Gamma_{ij}\right| - \dfrac{(\rho_s - \rho_w)}{2\rho} g\left(\bar{h}_{yi}^n\right)^2 \displaystyle\sum_{j\in N(i)} c_{ij} n_{yij} \left|\Gamma_{ij}\right| \\ 0 \\ 0 \end{pmatrix}. \tag{16}$$

The approximations $\bar{h}^n_{xi}$ and $\bar{h}^n_{yi}$ are reconstructed using a technique recently developed in [3] for the proposed finite volume method to satisfy the well-known C-property. In this section we briefly describe the formulation of this procedure and more details can be found in [3]. Hence, at the stationary state, the numerical flux in the corrector stage yields

$$\sum_{j\in N(i)} \mathscr{F}\left(\mathbf{W}^n_{ij};\mathbf{n}_{ij}\right) = \begin{pmatrix} 0 \\ -g\displaystyle\int_{T_i} h\frac{\partial Z}{\partial x}\,dV - g\frac{(\rho_s-\rho_w)}{2\rho}\int_{T_i} h^2\frac{\partial c}{\partial x}\,dV \\ -g\displaystyle\int_{T_i} h\frac{\partial Z}{\partial y}\,dV - g\frac{(\rho_s-\rho_w)}{2\rho}\int_{T_i} h^2\frac{\partial c}{\partial y}\,dV \\ 0 \\ 0 \end{pmatrix},$$

which is equivalent to

$$\begin{pmatrix} 0 \\ \displaystyle\sum_{j\in N(i)} \frac{1}{2}g\left(h^n_{ij}\right)^2 N_{xij} \\ \displaystyle\sum_{j\in N(i)} \frac{1}{2}g\left(h^n_{ij}\right)^2 N_{yij} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ -g\displaystyle\int_{T_i} h\frac{\partial Z}{\partial x}\,dV - g\frac{(\rho_s-\rho_w)}{2\rho}\int_{T_i} h^2\frac{\partial c}{\partial x}\,dV \\ -g\displaystyle\int_{T_i} h\frac{\partial Z}{\partial y}\,dV - g\frac{(\rho_s-\rho_w)}{2\rho}\int_{T_i} h^2\frac{\partial c}{\partial y}\,dV \\ 0 \\ 0 \end{pmatrix}. \tag{17}$$

where $N_{xij} = n_{xij}\,\left|\Gamma_{ij}\right|$ and $N_{yij} = n_{yij}\,\left|\Gamma_{ij}\right|$. Next, to approximate the source terms we proceed as follows. First we decompose the triangle T_i into three sub-triangles as depicted in Fig. 1. Then, the source term is approximated as

$$\int_{T_i} h\frac{\partial Z}{\partial x}\,dV = \int_{T_1} h\frac{\partial Z}{\partial x}\,dV + \int_{T_2} h\frac{\partial Z}{\partial x}\,dV + \int_{T_3} h\frac{\partial Z}{\partial x}\,dV, \tag{18}$$

where

$$\int_{T_1} h\frac{\partial Z}{\partial x}\,dV = h_1\int_{T_1} \frac{\partial Z}{\partial x}\,dV,$$

with h_1 is an average value of h on the sub-triangle T_1. Hence,

$$\begin{aligned}\int_{T_1} h\frac{\partial Z}{\partial x}\, dV &= h_1 \sum_{j\in N(1)} \int_{\Gamma_{1j}} Z n_x \, d\sigma,\\ &= h_1 \sum_{j\in N(1)} Z_{1j}\ N_{x1j},\\ &= h_1 \sum_{j\in N(1)} \frac{Z_1 + Z_j}{2}\ N_{x1j}. \qquad (19)\end{aligned}$$

Again, using the stationary flow condition $h_1 + Z_1 = h_j + Z_j = H = constant$, one gets

$$h_1 + Z_1 + h_j + Z_j = 2H \qquad \text{and} \qquad \frac{Z_1 + Z_j}{2} = H - \frac{h_1 + h_j}{2}.$$

Thus, (19) gives

$$\int_{T_1} h\frac{\partial Z}{\partial x}\, dV = h_1 \sum_{j\in N(1)} \left(H - \frac{h_1 + h_j}{2}\right) N_{x1j}.$$

Using the fact that $\sum_{j\in N(1)} N_{x1j} = 0$,

$$\begin{aligned}\int_{T_1} h\frac{\partial Z}{\partial x}\, dV &= -\frac{h_1}{2} \sum_{j\in N(1)} h_j\ N_{x1j},\\ &= -\frac{h_1}{2}\left(h_p N_{x1p} + h_2 N_{x12} + h_3 N_{x13}\right).\end{aligned}$$

A similar procedure leads to the following approximations of the other terms in (18)

$$\begin{aligned}\int_{T_2} h\frac{\partial Z}{\partial x}\, dV &= -\frac{h_2}{2}\left(h_k N_{x2k} + h_1 N_{x21} + h_3 N_{x23}\right),\\ \int_{T_3} h\frac{\partial Z}{\partial x}\, dV &= -\frac{h_3}{2}\left(h_l N_{x3l} + h_1 N_{x31} + h_2 N_{x32}\right).\end{aligned}$$

Notice that h_p, h_k and h_l are the average values of h respectively, on the triangle T_p, T_k and T_l, see Fig. 1. Summing up, the discretization (18) gives

$$\int_{T_i} h\frac{\partial Z}{\partial x}\, dV = -\frac{h_1}{2} h_p N_{x1p} - \frac{h_2}{2} h_k N_{x2k} - \frac{h_3}{2} h_l N_{x3l}.$$

For this reconstruction, the source terms in (16) result in

$$\begin{aligned}\sum_{j\in N(i)}\left(h_{ij}^n\right)^2 N_{xij} &= h_1\left(h_p N_{x1p}\right)+h_2\left(h_k N_{x2k}\right)+h_3\left(h_l N_{x3l}\right),\\ \sum_{j\in N(i)}\left(h_{ij}^n\right)^2 N_{yij} &= h_1\left(h_p N_{y1p}\right)+h_2\left(h_k N_{y2k}\right)+h_3\left(h_l N_{y3l}\right).\end{aligned} \tag{20}$$

Here, (20) forms a linear system of two equations for the three unknowns h_1, h_2 and h_3. To complete the system we add the natural conservation equation

$$h_1+h_2+h_3=3h_i.$$

Analogously, the bottom values Z_j, $j=1,2,3$ are reconstructed in each subtriangle of T_i as

$$Z_j+h_j^n=Z_i+h_i^n, \qquad j=1,2,3.$$

Finally, the source terms in (18) are approximated as

$$\begin{aligned}h_1\int_{T_1}\frac{\partial Z}{\partial x}\,dV &= h_1\left(\frac{Z_1+Z_p}{2}N_{x1p}+\frac{Z_1+Z_2}{2}N_{x12}+\frac{Z_1+Z_3}{2}N_{x13}\right),\\ h_2\int_{T_1}\frac{\partial Z}{\partial x}\,dV &= h_2\left(\frac{Z_2+Z_k}{2}N_{x2k}+\frac{Z_2+Z_1}{2}N_{x21}+\frac{Z_2+Z_3}{2}N_{x23}\right),\\ h_3\int_{T_1}\frac{\partial Z}{\partial x}\,dV &= h_3\left(\frac{Z_3+Z_l}{2}N_{x3l}+\frac{Z_3+Z_1}{2}N_{x31}+\frac{Z_3+Z_2}{2}N_{x32}\right),\end{aligned} \tag{21}$$

with a similar equation for the other source terms in the y-direction.

4 Numerical results

We present numerical results for a test problem of partial dam-break over erodible bed. In all the computations reported herein, the Courant number Cr is set to 0.8 and the time stepsize Δt is adjusted at each step according to the stability condition

$$\Delta t=Cr\min_{\Gamma_{ij}}\left(\frac{|T_i|+|T_j|}{2\left|\Gamma_{ij}\right|\max_p\left|(\lambda_p)_{ij}\right|}\right),$$

where Γ_{ij} is the edge between two triangles T_i and T_j. The water density $\rho_w = 1000\ kg/m^3$ and the gravitational acceleration is fixed to $g=9.81\ m/s^2$.

We consider a 200 m long and 200 m wide flat reservoir with two different constant levels of water separated by a dam. At $t=0$ part of the dam breaks instantaneously. The dam is 10 m thick and the breach is assumed to be 75 m wide, as shown in Fig. 2. Initially, $u(x,y,0)=v(x,y,0)=0\,m/s$

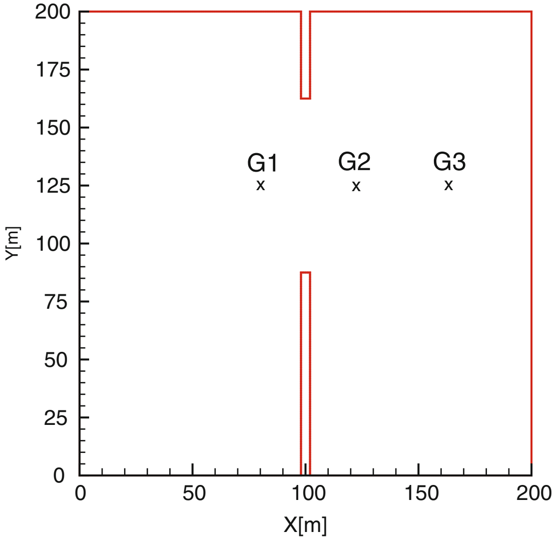


Fig. 2 Computational domain for the partial dam-break over erodible bed

$$h(x,y,0)=\begin{cases}10\,m, & \text{if} \quad x<100\,m,\\ 1\,m, & \text{otherwise},\end{cases} \qquad c(x,y,0)=\begin{cases}0.01, & \text{if} \quad x<100\,m,\\ 0, & \text{otherwise.}\end{cases}$$

The selected values for the evaluation of the present finite volume model are summarized in Table 1. At $t = 0$ the dam collapses and the flow problem consists of a shock wave traveling downstream and a rarefaction wave traveling upstream.

Table 1 Reference parameters used for the dam-break problem

Quantity	Reference value	Quantity	Reference value
ρ_s	$1500\ kg/m^3$	ν	$1.2\times10^{-6}\ m^2/s$
p	0.28	n_b	$0.015\ s/m^{1/3}$
φ	$0.015\ m^{1.2}$	θ_c	0.045
m	2	d	$1\ mm$

In Fig. 3 we present the water free-surface and bed-load, the adapted meshes and snapshots of the water depth obtained for the partial dam-break over fixed bed at times $t = 2$, 4, 6 and 8 s. The results obtained for the partial dam-break over erodible bed are presented in Fig. 4. By using adaptive meshes, high resolution is automatically obtained in those regions where the gradients of the water depth are steep such as the moving fronts. Apparently, the overall flow pattern for this

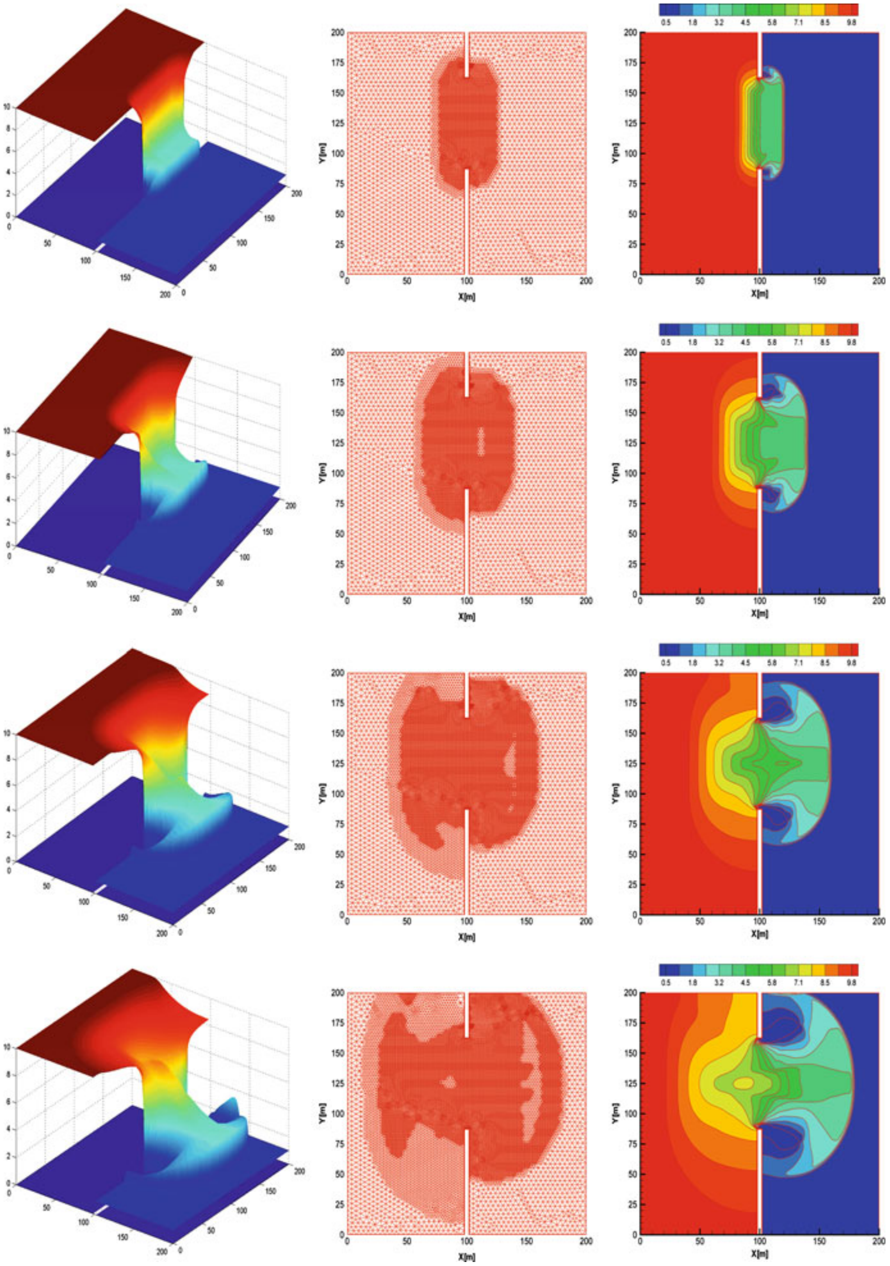


Fig. 3 Water free-surface and bed-load (first column), adapted meshes (second column) and water free-surface contours (third column) for the partial dam-break over fixed bed at different simulation times. From top to bottom $t = 2$, 4, 6 and 8 s

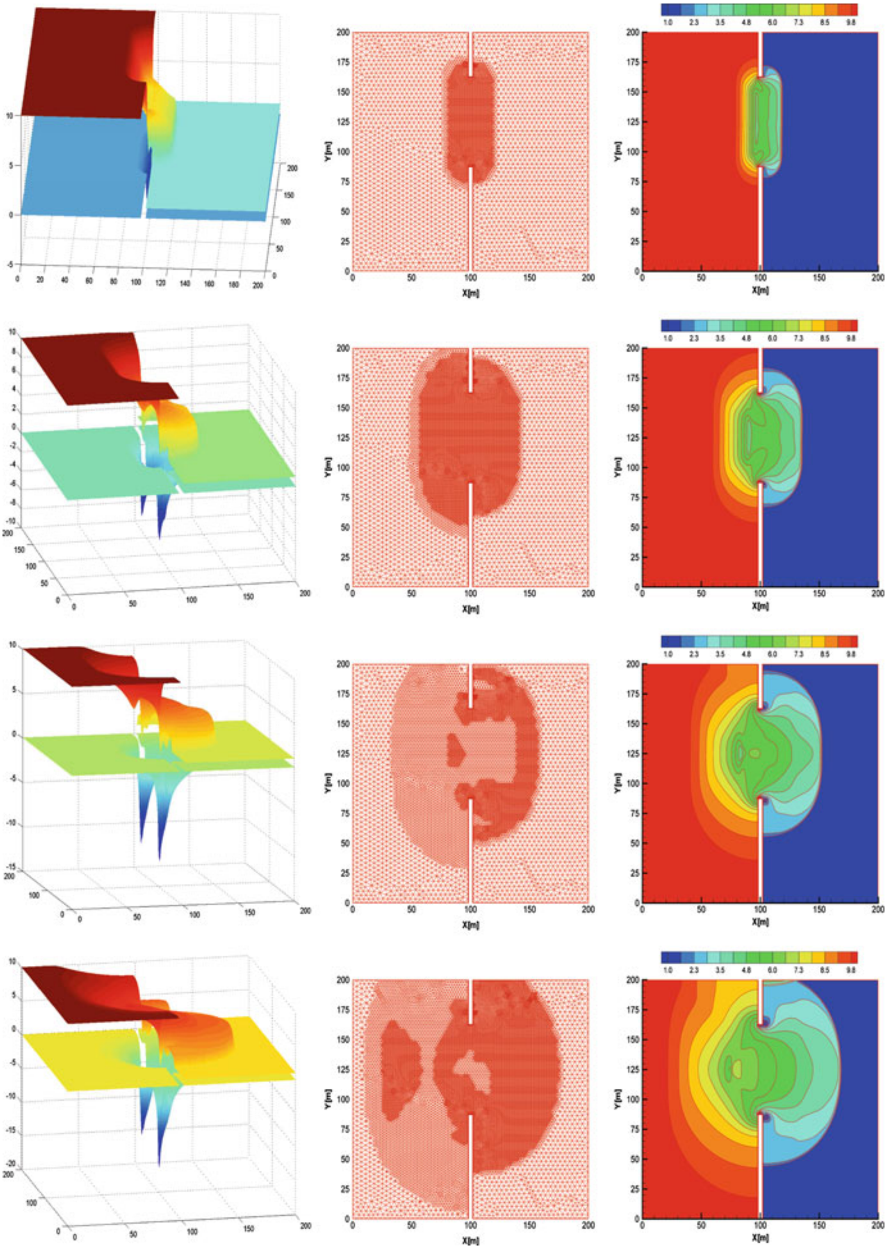


Fig. 4 Water free-surface and bed-load (first column), adapted meshes (second column) and water free-surface contours (third column) for the partial dambreak over erodible bed at different simulation times. From top to bottom $t = 2$, 4, 6 and 8 s

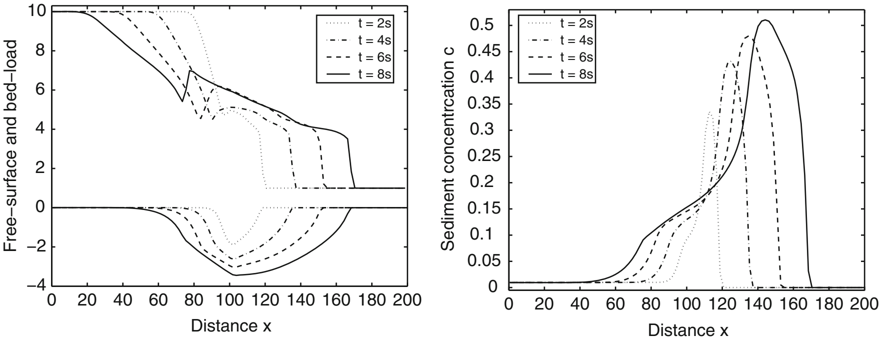


Fig. 5 Cross sections at $y = 125\ m$ of the water free-surface and bed-load (left plot) and sediment concentration (right plot) at four instants

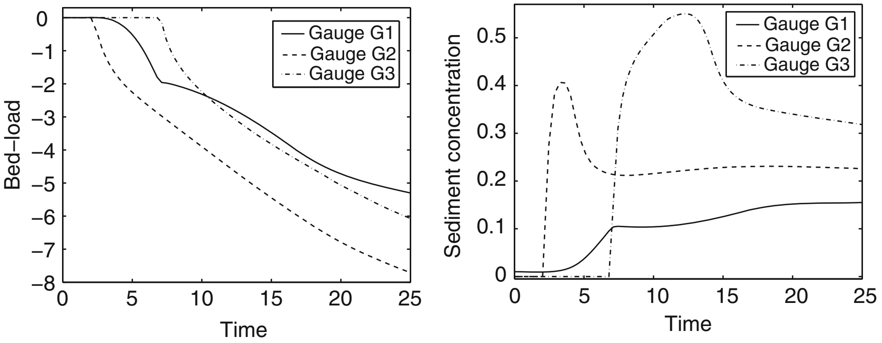


Fig. 6 Time evolution of the water free-surface and bed-load (left plot) and sediment concentration (right plot) at the three gauges G1, G2 and G3 presented in Fig. 2

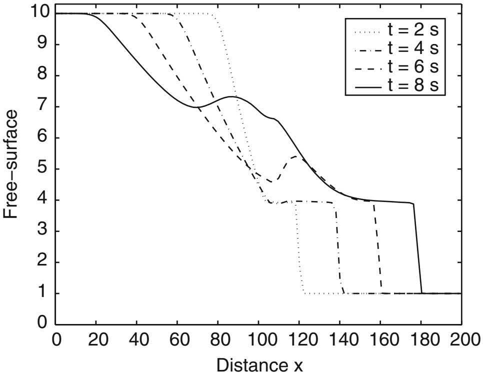


Fig. 7 Cross sections at $y = 125\ m$ of the water free-surface for the partial dam-break over fixed bed

example is preserved with no excessive numerical diffusion in the results by finite volume method using adaptive mesh. The adaptive finite volume method performs well for this test problem since it does not diffuse the moving fronts and no spurious oscillations have been observed when the water flows over the movable bed.

In order to quantify the results for this test example we display in Fig. 5 cross sections at $y = 125\ m$ of the water free-surface and bed-load and sediment concentration at four instants shown in Fig. 4. The results for the partial dam-break over fixed bed are depicted in Fig. 7. Figure 6 exhibites the time evolution of the water free-surface, bed-load and sediment concentration at the three gauges G1, G2 and G3 presented in Fig. 2. As can be observed from these results, the erosion effects on the bed are clearly visible for the considered sediment conditions. The inclusion of Exner equation in the model creates a very active sediment exchange between the water flow and the bed load, and also produces a sharp spatial gradient of sediment concentration, which justifies its incorporation in the momentum equations (10). Apparently, the overall flow and sediment features for this example are preserved with no spurious oscillations appearing in the results obtained using the adaptive finite volume method. Obviously, the computed results verify the stability and the shock capturing properties of the proposed finite volume method.

References

1. M.B. Abbott, Computational hydraulics: Elements of the theory of free surface flows, Fearon-Pitman Publishers, London, 1979.
2. S.J. Billett, E.F. Toro, On WAF-Type Schemes for Multidimensional Hyperbolic Conservation Laws, J. Comp. Physics. **130**, 1–24 (1997).
3. F. Benkhaldoun, I. Elmahi, M. Seaïd, “A new finite volume method for flux-gradient and source-term balancing in shallow water equations”, Computer Methods in Applied Mechanics and Engineering. **199** pp:49-52 (2010).
4. F. Benkhaldoun, S. Sahmim, M. Seaïd, A two-dimensional finite volume morphodynamic model on unstructured triangular grids. Int. J. Num. Meth. Fluids. 63 (2010) 1296–1327.
5. F. Benkhaldoun, S. Sahmim, M. Seaïd, Solution of the sediment transport equations using a finite volume method based on sign matrix. SIAM J. Sci. Comp. 31 (2009) 2866–2889.
6. F. Benkhaldoun, I. Elmahi, M. Seaïd, “Well-balanced finite volume schemes for pollutant transport by shallow water equations on unstructured meshes”, J. Comp. Physics. **226** pp:180-203 (2007).
7. K. Bloundi and J. Duplay, Heavy metals distribution in sediments of nador lagoon (Morocco), *Geophysical Research Abstracts*, **5**, pp. 11744, 2003.
8. Z. Cao and P. Carling. Mathematical modelling of alluvial rivers: reality and myth. Part I: General overview. Water Maritime Engineering. **154**, 207-220 (2002)
9. Z. Cao and G. Pender. Numerical modelling of alluvial rivers subject to interactive sediment mining and feeding. Advances in Water Resources. **27**, 533-546 (2004)
10. Z. Cao, G. Pender, S. Wallis and P. Carling. Computational dam-break hydraulics over erodible sediment bed. J. Hydraulic Engineering. **67**, 689-703 (2004)
11. Z. Cao, G. Pender and P. Carling. Shallow water hydrodynamic models for hyperconcentrated sediment-laden floods over erodible bed. Advances in Water Resources. **29**, 546-557 (2006)
12. N.S. Cheng, Simplified settling velocity formula for sediment paticle. J. Hydraulic Engineering ASCE. 123 (1997) 149–152.

13. J. Fredsøe, R. Deigaard, Mechanics of Coastal Sediment Transport. *Advanced Series on Ocean Engineering - Vol. 3*, 1992.
14. A.J. Grass, Sediment Transport by Waves and Currents. (SERC London Cent. Mar. Technol. Report No: FL29, 1981).
15. G. Simpson and S. Castelltort. Coupled model of surface water flow, sediment transport and morphological evolution. Computers & Geosciences. **32**, 1600-1614 (2006)
16. G. Strang. On the Construction and the Comparison of Difference Schemes. SIAM J. Numer. Anal. **5**, 506517 (1968)
17. C.Y. Yang,: Sediment Transport: Theory and Practice. McGraw-Hill, New York (1996)
18. R.L. Soulsby: Dynamics of marine sands, a manual for practical applications. HR Wallingford, Report SR 466 (1997)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Part III
Benchmark Papers

3D Benchmark on Discretization Schemes for Anisotropic Diffusion Problems on General Grids

Robert Eymard, Gérard Henry, Raphaèle Herbin, Florence Hubert, Robert Klöfkorn, and Gianmarco Manzini

Abstract We present a number of test cases and meshes that were designed as a benchmark for numerical schemes dedicated to the approximation of three-dimensional anisotropic and heterogeneous diffusion problems. These numerical schemes may be applied to general, possibly non conforming, meshes composed of tetrahedra, hexahedra and quite distorted general polyhedra. A number of methods were tested among which conforming finite element methods, discontinuous Galerkin finite element methods, cell-centered finite volume methods, discrete duality finite volume methods, mimetic finite difference methods, mixed finite element methods, and gradient schemes. We summarize the results presented by the participants to the benchmark, which range from the number of unknowns, the approximation errors of the solution and its gradient, to the minimum and maximum values and energy. We also compare the performance of several iterative or direct linear solvers for the resolution of the linear systems issued from the presented schemes.

Keywords Anisotropic and heterogeneous medium, diffusion problem, numerical schemes for general polyhedral meshes, non-conforming meshes, 3D benchmark.
MSC2010: 65N08, 65N30, 65Y20, 76S05

Robert Eymard
Université Paris-Est, France, e-mail: Robert.Eymard@univ-mlv.fr

Gérard Henry, Raphaèle Herbin, Florence Hubert
Université Aix-Marseille, France, e-mail: Gerard.Henry@latp.univ-mrs.fr, Raphaele.Herbin@latp.univ-mrs.fr, Florence.Hubert@latp.univ-mrs.fr

Robert Klöfkorn
Universität Freiburg, Germany, e-mail: robertk@mathematik.uni-freiburg.de

Gianmarco Manzini
IMATI-CNR and CeSNA-IUSS Pavia, Italy, e-mail: Marco.Manzini@imati.cnr.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_89, © Springer-Verlag Berlin Heidelberg 2011

1 Introduction

The two-dimensional (2D) anisotropy benchmark organized in 2007-2008 [21] provided a better understanding of the relative properties of a huge number of numerical schemes in terms of robustness, accuracy, problem size (number of degrees of freedom and matrix size), quality of the numerical approximation (maximum/minimum principles), etc. Nonetheless, a direct extrapolation of these results to three-dimensional (3D) problems is not possible because of the much higher complexity of the meshes involved in a 3D calculation and the larger size of the resulting linear systems. Hence, a new benchmark was organized between the end of 2010 and the beginning of 2011 with the additional goal of comparing CPU times versus accuracy.

A number of anisotropic and heterogeneous diffusion problems, associated with general, possibly non-conforming, 3D grids, were proposed in order for the participants to test a variety of numerical schemes. The participants were expected to provide information about the results obtained in these test cases and to use a set of solvers made available by the benchmark organizers for the linear systems arising from the discretization. In order to ensure a fair comparison of CPU times, all linear systems were solved by the same program implemented sequentially on the same computer, located at Université Aix-Marseille, France.

In most test cases the domain Ω is the unit cube; the boundary of Ω is denoted by Γ. We consider the steady diffusion problem with either homogeneous or non-homogeneous Dirichlet conditions on the boundary that is formulated in strong form as:

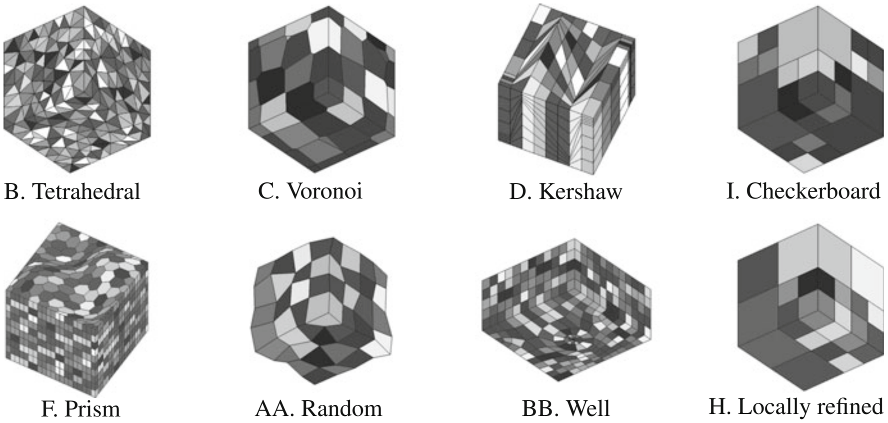
$$-\nabla\cdot(\mathbf{K}\nabla u) = f \quad \text{on } \Omega, \tag{1}$$

$$u = \bar{u} \quad \text{on } \Gamma, \tag{2}$$

where $\mathbf{K} : \Omega \to \mathbb{R}^{3\times 3}$ is the diffusion tensor, f is the source term and $\bar{u}$ is the Dirichlet boundary condition. The tensor fields $\mathbf{K}$ that we consider in the benchmark test cases are, as usual, strongly elliptic in Ω, i.e., each $\mathbf{K}$ is given by a field of symmetric matrices whose eigenvalues are uniformly bounded from above and from below by two strictly positive values. The data f and $\bar{u}$ of the problem are determined in accordance with the given exact solution and the diffusion field of each test case.

The paper is organized as follows. In Sect. 2, we present the five test cases, each one being specified by the shape of the computational domain, the exact solution, the diffusion field, and the set of meshes to be used. In Sect. 3, we briefly describe the linear solvers that were proposed for the resolution of the linear systems issued from the different numerical schemes. In Sect. 4, we list the participants to the benchmark and the numerical method that they used. In Sect. 5, we present the nature of the results obtained from the participants.

Final conclusions are drawn in Sect. 6. The tables and figures of results are given in Sect 7.



B. Tetrahedral C. Voronoi D. Kershaw I. Checkerboard

F. Prism AA. Random BB. Well H. Locally refined

Fig. 1 The different meshes.

2 The test cases and the meshes

The test cases are summarized in Table 1, where we specify, for each test case, the shape of the computational domain Ω, the label of the permeability tensor, the label of the exact solution and the name of the mesh family. For more details about the meshes and other data, see at the URL:

`http://www.latp.univ-mrs.fr/latp_numerique/?q=node/4,`

where mesh data files can be downloaded.

Table 1 The test cases

Test Case	Domain	Permeability $\mathbf{K}(x,y,z)$	Solution $u(x,y,z)$	Meshes
Test 1 Mild anisotropy	Unit cube	$\mathbf{K}_1(x,y,z)$	$u_1(x,y,z)$	Tetrahedral (B) Voronoi (C) Kershaw (D) Checkerboard (I)
Test 2 Heterogeneity and anisotropy	Unit cube	$\mathbf{K}_2(x,y,z)$	$u_2(x,y,z)$	Prism (F)
Test 3 Random meshes	Determined by the mesh	$\mathbf{K}_3(x,y,z)$	$u_3(x,y,z)$	Random (AA)
Test 4 The well	Ω_4	$\mathbf{K}_4(x,y,z)$	$u_4(x,y,z)$	Well (BB)
Test 5 Locally refined	Unit cube	$\mathbf{K}_5(x,y,z)$	$u_5(x,y,z)$	Locally refined (H)

The meshes are presented in Fig. 1.

The data labeled in Table 1 (permeability tensor and exact solution for all test cases and computational domain for Test Cases 4 and 5) are as follows.

1. **Test Case 1**. We consider a constant, anisotropic permeability tensor and a regular solution that implies a non-homogeneous Dirichlet condition on the domain boundary Γ:

$$\mathbf{K}_1(x,y,z) = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{pmatrix}$$

$$u_1(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$$

2. **Test Case 2**. We consider a smoothly variable permeability tensor and a regular solution that implies a non-homogeneous Dirichlet condition on the domain boundary Γ:

$$\mathbf{K}_2(x,y,z) = \begin{pmatrix} y^2+z^2+1 & -xy & -xz \\ -xy & x^2+z^2+1 & -yz \\ -xz & -yz & x^2+y^2+1 \end{pmatrix}$$

$$u_2(x,y,z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)\sin(2\pi z)$$

3. **Test Case 3**. We consider a constant, anisotropic permeability tensor and a regular solution on the domain, whose definition results from each of the considered meshes:

$$\mathbf{K}_3(x,y,z) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^3 \end{pmatrix}$$

$$u_3(x,y,z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$$

Since the meshes which are used for this test case (random meshes) have boundary vertices which are not located exactly on the boundary of the unit cube, the boundary conditions are non-homogeneous Dirichlet boundary conditions.

4. **Test Case 4**. The computational domain is given by $\Omega_4 = P \setminus W$, where P is the parallelepiped $]-15,15[\times]-15,15[\times]-7.5,7.5[$ and W is a slanted circular cylinder with radius $r_w = 0.1$. The axis of this well is a straight line located in the $x0z$ plane, passing by the origin, with an angle (in degrees) $\theta = -70°$ with the x axis, as shown in Fig. 2.

We consider the constant permeability tensor, which is slightly anisotropic in the third coordinate direction, given by

$$\mathbf{K}_4 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.2 \end{pmatrix}.$$

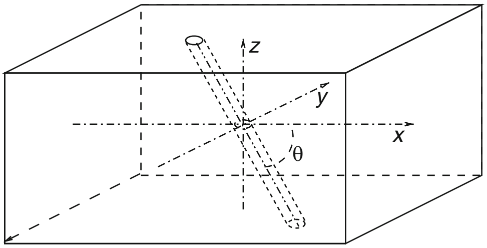
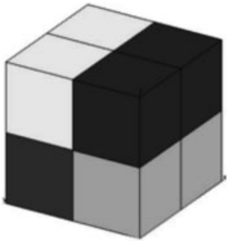


Fig. 2 The circular slanted well

The exact solution $u_4(x, y, z)$ is detailed in [1]: once a stretching of the axes has been performed so as to obtain an isotropic problem, we seek an exact solution that is constant on the well boundary. The solution simulates the pressure field that would be obtained for the same infinite slanted circular well in an infinite domain for a given constant flow rate q across any section of the well.

5. **Test Case 5**. The domain $\Omega = [0, 1]^3$ is split into four subdomains $\Omega = \cup_{i=1}^{4} \Omega_i$, which are given by

$$\Omega_1 = \{(x, y, z) \in [0, 1]^3 \text{ such that } y \leq 0.5,\ z \leq 0.5\}$$
$$\Omega_2 = \{(x, y, z) \in [0, 1]^3 \text{ such that } y > 0.5,\ z \leq 0.5\}$$
$$\Omega_3 = \{(x, y, z) \in [0, 1]^3 \text{ such that } y > 0.5,\ z > 0.5\}$$
$$\Omega_4 = \{(x, y, z) \in [0, 1]^3 \text{ such that } y \leq 0.5,\ z > 0.5\}$$



The permeability tensor and the exact solution are given by:

$$\mathbf{K}_5(x, y, z) = \begin{pmatrix} a_x^i & 0 & 0 \\ 0 & a_y^i & 0 \\ 0 & 0 & a_z^i \end{pmatrix} \quad \text{for } (x, y, z) \in \Omega_i \quad \text{with}$$

i	1	2	3	4
a_x^i	1	1	1	1
a_y^i	10	0.1	0.01	100
a_z^i	0.01	100	10	0.1
α_i	0.1	10	100	0.01

$$u_5(x, y, z) = \alpha_i \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$$

The permeability tensor $\mathbf{K}_5$ is discontinuous across the internal planes separating the unit cube in four subdomains and the exact solution u_5 is designed to be continuous and to ensure the conservation of the normal flux across such planes. Note that the homogeneous Dirichlet boundary condition is imposed in this test case.

3 Linear solvers used for the linear system benchmark

In order to access the different linear solver packages: UMFPACK [17, 18], DUNE-ISTL [7, 12], and PETSc [4, 5], all participants were asked to store their resulting linear systems for each test/mesh using a Compressed Row Storage (CRS) format, using an open source software package, which is available on line. All packages were installed on the 1 node Sun Fire X2270, equipped with 2 Quad-core processors (Intel, X5570, 2.93 GHz) and 24 GB memory (1333 MHz DDR3) and run sequentially.

Let us now briefly describe the available linear solvers and preconditioning methods.

1 **The direct solver library UMFPACK**. Written in ANSI/ISO C, UMFPACK is a set of routines for solving unsymmetric sparse linear systems $Ax = b$, using the Unsymmetric MultiFrontal method (see [17, 18] for details). For the benchmark, version 5.4.0 was used.

2 **The Iterative Solver Template Library – DUNE-ISTL** is a DUNE module [7, 12], which provides C++ programmed iterative solvers of linear systems stemming from finite element discretizations. The efficiency of the solvers is enhanced by taking into account the specific block recursive structure of matrices and vectors. For the benchmark version 2.0 has been used. The following solvers and preconditioning methods are used:

   - *Iterative solvers*: Conjugate Gradient, BiCG-stab, GMRES;
   - *Preconditioning*: Jacobi, ILU-0, ILU-n, $n = 1, ..., 4$, Algebraic Multi Grid.

3 **The Portable, Extensible Toolkit for Scientific Computation – PETSc** [4, 5] is a suite of data structures and routines for the scalable (parallel) solution of scientific applications modeled by partial differential equations. The program code is written in ANSI C. For the benchmark, version 3.1-p5 was used. The following iterative solvers and preconditioning methods are used:

   - *Iterative solvers*: Conjugate Gradient, BiCG-stab, GMRES;
   - *Preconditioning*: Jacobi, ILU-n, $n = 0, ..., 4$.

4 **Condition number calculation**. For the approximate calculation of the condition number of a given matrix, the Krylov-Schur method from the Scalable Library for Eigenvalue Problem Computations (SLEPc) package version 3.1-p4 [22] was used. SLEPc is written in ANSI C and built on top of PETSc.

5 **CPU time measurement**. The measurement of the CPU time spent for the solution process is based on the getrusage routines. The setup of the matrices (for the different solvers) is not included in the CPU time measurement in any case. The CPU time needed for the solution of the system with the iterative solvers (DUNE-ISTL and PETSc) is calculated by adding the time spent for building the preconditioner and the time spent in the linear solver. The CPU time with UMFPACK is not provided because the size of the matrices was too large for a direct solver in several cases.

4 The participating schemes and teams

Even though the benchmark is associated with the FVCA6 conference, the call for submission was by no means restricted to finite volume schemes, and, indeed, many types of schemes were submitted.

Cell-centered schemes

- MPFA-O: a Multi-Point Flux Approximation O-scheme programmed by the benchmark organizers for completeness purposes.
- LS-FVM: *The cell-centered finite volume method using least squares vertex reconstruction (diamond scheme)* , by Y. Coudière and G. Manzini [14].

Discontinuous Galerkin schemes

- CDG2: *The Compact Discontinuous Galerkin 2 Scheme*, R. Klöfkorn, [23].
- SWPG: *Symmetric Weighted Interior Penalty Discontinuous Galerkin Scheme*, by P. Bastian [6].

Discrete duality finite volume schemes

- CEVEDDFV-A: *A version of the DDFV scheme with cell/vertex unknowns on general meshes*, by B. Andreianov, F. Hubert and S. Krell [3].
- CEVEDDFV-B: *CeVe-DDFV, a discrete duality scheme with cell/vertex unknowns*, by Y. Coudière and C. Pierre [15].
- CEVEFE-DDFV: *CeVeFE-DDFV, a discrete duality scheme with cell/vertex/-face+edge unknowns*, by Y. Coudière, F. Hubert and G. Manzini [13].

Finite element schemes

- FEM: *Finite elements of order one* (FEM1) *and two* (FEM2) provided by P. Bastian with the DUNE environment [8,9].
- MELODIE, *A linear finite element solver*, by H. Amor, M. Bourgeois, and G. Mathieu [2].

Mixed or hybrid methods

- MFD-GEN: *Mimetic finite difference method for generalized polyhedral meshes*, by K. Lipnikov and G. Manzini [24].
- MFD-PLAIN: *A mimetic finite difference method*, by P. Bastian, O. Ippisch, and S. Marnach, [10].
- MFMFE: *A multipoint flux mixed finite element method on general hexahedra*, by M. F. Wheeler, G. Xue and I. Yotov [25].
- CHMFE: *A composite hexahedral mixed finite element*, by I. Ben Gharbia, J. Jaffré , N. Suresh Kumar and J. E. Roberts [11].

Gradient schemes

- SUSHI: *The SUSHI scheme*, by R. Eymard, T. Gallouët and R. Herbin, [19].
- VAG and VAGR: *The VAG scheme*, by R. Eymard, C. Guichard and R. Herbin, [20].

Nonlinear schemes The schemes are nonlinear in order to ensure the positivity of the scheme (that is, if the right hand side is positive then the solution is positive) or the discrete maximum principle (that is, if the linear system stems from the discretization of an elliptic equation satisfying the maximum principle, then its solution is also bounded by the bounds of the continuous system).

- FVMON: *A monotone nonlinear finite volume method for diffusion equations on polyhedral meshes*, by A. Danilov and Y. Vassilevski, [16].

The choice of categories that we considered above is neither exhaustive nor unique. In fact, most of these categories intersect: schemes are not so easy to classify, and some schemes are known to be identical in special cases and when using some special meshes. We refer to the above-cited papers for the details of the schemes and their implementation. Our purpose is to give here a synthesis of the results presented by the participants.

5 Results obtained by the participants

5.1 Results provided by the participants

The results obtained by the participants are presented in the contributed papers in several tables.

First table: it reports the data related to the size of the discrete problem produced by a numerical scheme and some information about the quality of the numerical approximation. In particular, the minimum and maximum values of the discrete solution at cell-centers are compared with the same kind of values for the exact solution, and an estimate of `ngrad`$\sim \int_\Omega \|\nabla u\|$ allows us to evaluate possible oscillations of the approximation.

`i`	number of mesh
`nu`	number of unknowns of the linear system
`nmat`	number of non zero terms in the matrix
`umin`	minimum value of the approximate solution at the cell centers
`uemin`	minimum value of the exact solution at the cell centers
`umax`	maximum value of the approximate solution at the cell centers
`uemax`	maximum value of the exact solution at the cell centers
`normg`	L^1 norm of the euclidean norm of the approximate gradient

Second table: it provides information about the accuracy of the schemes, which is measured for all the test cases versus `nu`, the number of unknowns, by the following quantities:

`i`	number of mesh
`nu`	number of unknowns of the linear system
`erl2`	relative L^2 norm of the error with respect to the L^2 norm of the exact solution.
`ratiol2`	order of convergence of the L^2 norm of the error on the solution between mesh `i` and `i-1`.
`ergrad`	relative H^1 semi-norm of the error with respect to the H^1 semi-norm of the exact solution.
`ratiograd`	order of convergence of the H^1 norm of the error on the solution between mesh `i` and `i-1`.
`ener`	relative energy norm of the error with respect to the energy norm of the exact solution.
`ratioener`	order of convergence of the energy norm of the error on the solution between mesh `i` and `i-1`.

where, denoting *err* the numerical error,

- the relative L^2 norm of the error is given by:

$$\texttt{erl2} \approx \left(\int_\Omega |err|^2 / \int_\Omega |u|^2 \right)^{\frac{1}{2}};$$

- the relative L^2 norm of the gradient of the error is given by:

$$\texttt{ergrad} \approx \left(\int_\Omega |\nabla err|^2 / \int_\Omega |\nabla u|^2 \right)^{\frac{1}{2}};$$

- the relative energy norm of the error is given by:

$$\texttt{ener} \approx \left(\int_\Omega \mathbf{K}\nabla err \cdot \nabla err / \int_\Omega \mathbf{K}\nabla u \cdot \nabla u \right)^{\frac{1}{2}}.$$

and the convergence rates are defined, for $i \geq 2$, by:

$$\texttt{ratiol2(i)} = -3 \frac{\log(\texttt{erl2(i)/erl2(i-1)})}{\log(\texttt{nu(i)/nu(i-1)})};$$

$$\texttt{ratiograd(i)} = -3 \frac{\log(\texttt{ergrad(i)/ergrad(i-1)})}{\log(\texttt{nu(i)/nu(i-1)})};$$

$$\texttt{ratioener(i)} = -3 \frac{\log(\texttt{ener(i)/ener(i-1)})}{\log(\texttt{nu(i)/nu(i-1)})}.$$

Matrices and right-hand sides were uploaded by the participants on the computer dedicated to the bench, in order to compare CPU time and memory.

5.2 Comparisons

- **Maximum principle**. For all test cases, we collect the values of u_{min}, u_{max} for the coarsest and finest grids handled by the participants, in Tables 2, 3 and 4 (Test Case 1), 5 (Test Case 2), 6 (Test Case 3), 7 (Test Case 4) and 8 (Test Case 5). We colored in red (resp. purple) the values that are below (resp. above) the minimum value of the exact solution.
- **Accuracy**. In Figs. 3-10, we report the log-log curves of the approximation errors measured by the benchmark participants for their numerical schemes. Each figure refers to a specific combination "test case + mesh family"; the upper left-most plot reports `erl2`, the upper right-most plot reports `ergrad`, the lower left-most plot reports `normg`, and the lower right-most plot reports `ener`. The convergence rates in these log-log plots are reflected by the slopes of the convergence curves.
- **Condition number**. We report the condition number (see Sect. 3) of the matrices involved in the numerical discretizations of first two test cases in Tables 9, 10, 11 and 12 (Test Case 1) and in Table 13 (Test Cases 2). The condition numbers in each table are calculated for the first mesh and the two next mesh refinements. The eigensolver tolerance was set to 10^{-8} for all matrices.
- **Cost of the resolution**. The cost of the resolution of the linear systems is shown in Figs. 11-18, where the L^2 error is plotted with respect to the CPU time and the used memory. The CPU time was measured for the linear system with the right hand side $b = A\mathbf{1}$, where $\mathbf{1}$ is the vector with all components equal to 1. The stopping criterion for all the iterative methods is: residual $\leq 10^{-10}$. For the sake of simplicity, all methods, including conjugate gradient methods, have been applied to symmetric and non-symmetric matrices.

6 Conclusion

This paper proposes a comparison of sixteen numerical schemes (and variants) which were tested on a family of three-dimensional anisotropic diffusion problems. The tests presented here involve both a wide class of diffusion tensors (anisotropic and at time heterogeneous and/or discontinuous) and a wide class of conforming and non-conforming meshes with very general polyhedral cells.

The number of results which were obtained on this benchmark is impressive with respect to the difficulty of the exercise and the time constraint. In fact, additional results are available on the bench web site:

`http://www.latp.univ-mrs.fr/latp_numerique/?q=node/4.`

and will be updated. The benchmark was found to be most useful to the participants to compare their schemes to reference solutions. The participation to the 3D benchmark was an opportunity for several participants to learn more about the efficient implementation of their schemes. Indeed, several variants of the schemes were thus developed. Last but not least, a user-friendly comparison platform was developed for this benchmark, which allows anyone to link to the solver and preconditioner of his choice; this possibility has already been used by other users than the 3D benchmark. The platform which was developed for the 3D benchmark should proof useful for further investigations on numerical schemes for various models.

7 Tables and figures of results

Table 2 Maximum principle for Test 1: mild anisotropy on tetrahedral meshes

Scheme	umin coarse	umax coarse	umin fine	umax fine
CDG2K1	−1.54E-02	2.017	−6.63E-04	2.002
CDG2K2	0.00	1.999	0.00E+00	1.999
CEVEDDFV-A	0.706E-02	1.992	0.140E-02	1.999
CEVEDDFV-B	1.34E-02	1.99	1.30E-03	2.00
CEVEFE-DDFV	6.09E-03	1.988	1.93E-03	1.999
FEM1	8.34E-02	1.932	6.35E-03	1.990
FEM2	2.13E-02	1.989	1.84E-03	1.997
FVMON	0.028	1.997	0.003	1.998
LS-FVM	2.03E-02	1.989	1.83E-03	1.997
MELODIE	7.69E-02	1.935	6.19E-03	1.991
MPFA-O	−1.13E-02	2.01	−1.46E-03	2.00
MFD-PLAIN	2.33E-03	1.994	1.66E-03	1.998
MFD-GEN	2.26E-02	1.986	1.75E-03	1.997
SWPG-1	5.32E-02	1.965	3.69E-03	1.994
SWPG-2	2.11E-02	1.989	1.84E-03	1.997
SWPG-3	2.04E-02	1.989	1.83E-03	1.997
SWPG-4	2.03E-02	1.989	1.83E-03	1.997
SUSHI	3.21E-02	1.98	1.74E-03	2.00
VAG	6.77E-02	1.94	4.62E-03	1.99
VAGR	5.77E-02	1.95	3.63E-03	1.99

Table 3 Maximum principle for Test 1: mild anisotropy on Kershaw meshes

Scheme	umin (coarse)	umax (coarse)	umin(fine)	umax(fine)
CDG2LEGK1	−2.95E-02	2.016	−5.37E-04	2.000
CDG2LEGK2	0.00	1.997	0.00	1.999
CDG2TETK1	−2.81E-02	2.012	−4.65E-04	2.000
CDG2TETK2	0.00	1.995	0.00	1.999
CEVEDDFV-A	2.28E-02	1.989	3.82E-04	2.000
CEVEDDFV-B	7.16E-02	1.94	4.61E-04	2.00
CEVEFE-DDFV	5.67E-02	1.940	6.52E-04	2.000
CHMFE	−0.032	1.94685	−0.008	2.00061
FEM1	1.77E-01	1.786	2.94E-03	1.996
FEM2	3.29E-02	1.941	7.11E-04	1.999
FVMON	0.112	1.942	0.003	1.997
LS-FVM	3.03E-02	1.958	7.14E-04	1.999
MELODIE	1.34E-01	1.833	2.04E-03	1.997
MFD-GEN	−2.52E-02	1.973	2.71E-04	1.999
MFD-PLAIN	−6.03E-01	2.100	1.65E-04	2.000
MFMFE-NS	−1.26E-03	2.01	5.00E-05	2.00
MFMFE-S	4.66E-03	1.97	7.49E-05	2.00
MPFA-O	−3.76E-02	2.05	−1.06E-03	2.00
SWPG-1	9.58E-02	1.850	1.71E-03	1.997
SWPG-2	3.12E-02	1.944	7.11E-04	1.999
SWPG-3	2.91E-02	1.955	1.75E-03	1.997
SWPG-4	3.02E-02	1.958	1.75E-03	1.997
SUSHI	−2.14E-03	1.91	8.51E-04	2.00
VAG	1.43E-01	1.93	1.07E-03	2.00
VAGR	7.80E-02	1.96	−2.64E-04	2.00

Table 4 Maximum principle for Test 1: mild anisotropy on Checkerboard meshes

Scheme	umin (coarse)	umax(coarse)	umin(fine)	umax(fine)
CDG2K1	0.00	1.901	−5.50E-04	2.000
CDG2K2	−3.34E-02	2.050	0.000	1.999
CDG2LEGK1	−7.94E-02	2.081	−3.06E-04	2.000
CDG2LEGK2	0.00	1.998	0.00	1.999
CDG2TETK2	0.00	2.003	0.00	1.999
CEVEDDFV-A	0.341E-01	1.966	0.134E-03	2.000
CEVEDDFV-B	1.46E-01	1.86	5.01E-04	2.00
CEVEFE-DDFV	8.58E-02	1.903	2.88E-04	2.000
FEM1	3.26E-01	1.671	1.54E-03	1.998
FVMON	0.122	1.905	0.001	2.000
LS-FVM	1.54E-01	1.846	6.36E-04	1.999
MFD-GEN	2.91E-01	1.880	2.15E-03	1.999
MFD-PLAIN	1.27E-01	1.883	−3.52E-03	2.004
SWPG-1	2.35E-01	1.784	6.36E-04	1.999
SWPG-2	1.82E-01	1.812	6.37E-04	1.999
SWPG-3	1.61E-01	1.839	6.36E-04	1.999
SWPG-4	1.55E-01	1.845	6.36E-04	1.999
SUSHI	1.05E-01	1.87	3.83E-04	2.00
VAG	−1.95	2.50	−3.06E-02	2.03
VAGR	−9.81E-02	2.08E+00	−4.33E-03	2.00

Table 5 Maximum principle for Test 2: heterogeneous anisotropy on Prismatic meshes

Scheme	umin (coarse)	umax(coarse)	umin(fine)	umax(fine)
CEVEDDFV-A	−.856	1.044	−.862	1.049
CEVEDDFV-B	−8.53E-01	9.85E-01	−8.58E-01	1.03
CEVEFE-DDFV	−8.55E-01	1.014	−8.60E-01	1.040
FVMON	−0.854	1.002	−0.858	1.034
LS-FVM	−8.42E-01	0.978	−8.57E-01	1.033
MFD-GEN	−0.873	0.832	−0.890	0.963
MPFA-O	−9.23E-01	1.07	−8.63E-01	1.05
SUSHI	−8.22E-01	9.82E-01	−8.55E-01	1.03
VAG	−9.49E-01	1.23	−8.53E-01	1.05
VAGR	−8.73E-01	1.10E+00	−8.53E-01	1.04

Table 6 Maximum principle for Test 3: flow on random meshes

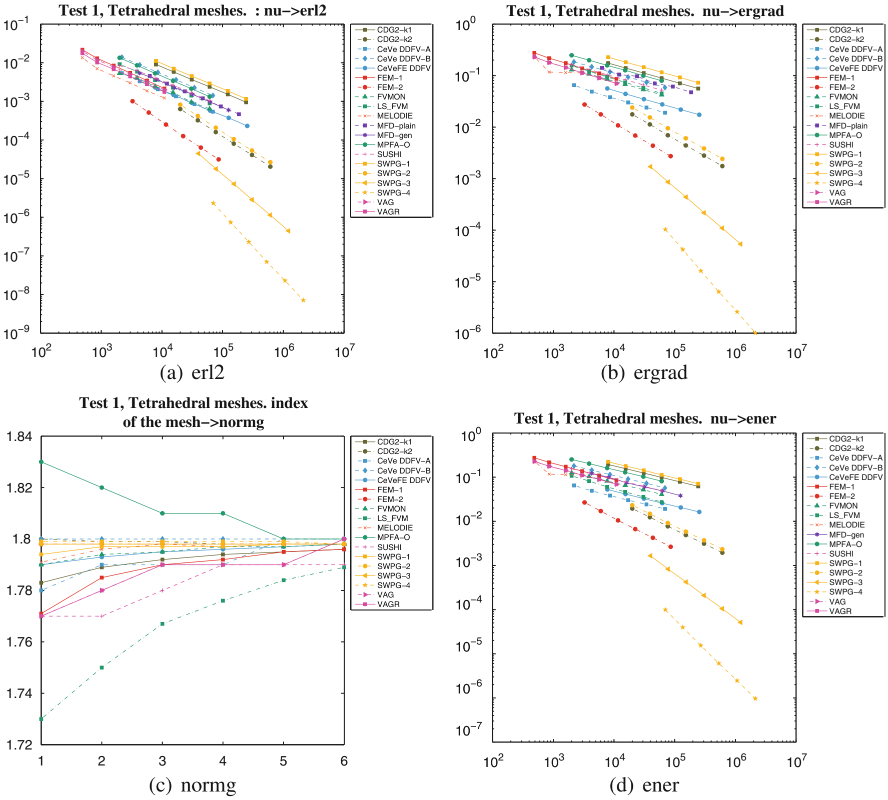
Scheme	umin(coarse)	umax(coarse)	umin(fine)	umax(fine)
CDG2LEGK1	−1.143	1.244	−1.009	1.000
CDG2LEGK2	−1.015	1.034	−1.00E+00	1.00
CDG2TETK1	−1.261	1.167	−1.008	1.002
CDG2TETK2	−1.238	1.295	−1.000	1.000
CEVEDDFV-A	−.202E+01	1.969	−.101E+01	1.014
CEVEDDFV-B	−1.58	1.54	−1.01	1.01
CEVEFE-DDFV	−4.25E+01	49.169	−2.67	2.725
FEM1	−3.73E-01	0.313	−9.90E-01	0.989
FEM2	−7.48E-01	0.679	−9.96E-01	0.996
FVMON	−0.905	0.759	−0.989	1.001
LS-FVM	−7.56E-01	0.711	−9.96E-01	0.996
MELODIE	−0.665	0.685	−0.988	0.991
MFD-GEN	−1.268	1.430	−1.027	1.021
MFD-PLAIN	−1.02E+00	1.045	−1.00	1.000
MFMFE-S	−6.20	5.75	−1.06	1.04
MPFA-O	−9.79	1.22E+01	−2.61E+01	2.44E+01
SUSHI	−7.51E-01	7.58E-01	−9.90E-01	9.89E-01
SWPG-1	−4.34E-01	0.355	−9.90E-01	0.989
SWPG-2	−7.50E-01	0.676	−9.96E-01	0.996
SWPG-3	−7.53E-01	0.684	−9.96E-01	0.996
SWPG-4	−7.59E-01	0.691	−9.85E-01	0.982
VAG	−1.31	1.50	−1.00	1.00
VAGR	−1.51	1.68E+00	−1.01	1.01

Table 7 Maximum principle for Test 4: the flow around the well

Scheme	umin(coarse)	umax(coarse)	umin(fine)	umax(fine)
CDG2LEGK1	0.00	5.406	0.00	5.410
CDG2LEGK2	0.00	5.408	0.00	5.411
CDG2TETK1	0.00	5.406	0.00	5.410
CDG2TETK2	−5.92E-03	5.414	0.00	5.414
CEVEDDFV-A	−.438E-01	5.415	−.198E-02	5.415
CEVEDDFV-B	4.85E-01	5.32	5.80E-02	5.36
CEVEFE-DDFV	3.83E-01	5.317	5.66E-02	5.361
FEM1	3.73E-01	5.317	5.66E-02	5.361
FEM2	4.12E-01	5.317	5.65E-02	5.361
FVMON	0.518	5.318	0.059	5.361
LS-FVM	4.57E-01	5.317	5.75E-02	5.361
MELODIE	0.189	5.360	0.029	5.39
MFD-GEN	5.37E-01	5.317	5.91E-02	5.361
MFD-PLAIN	5.74E-01	5.317	5.91E-02	5.361
MPFA-O	4.36E-01	5.39	−1.49E-03	5.40
SUSHI	4.26E-01	5.32	5.78E-02	5.36
SWPG-1	3.52E-01	5.316	5.55E-02	5.361
SWPG-2	4.13E-01	5.317	5.65E-02	5.361
SWPG-3	4.15E-01	5.317	5.65E-02	5.361
SWPG-4	4.14E-01	5.317	8.99E-02	5.339
VAG	3.89E-01	5.32	5.69E-02	5.36
VAGR	3.89E-01	5.32	5.69E-02	5.36

Table 8 Maximum principle for Test 5: discontinuous anisotropy

Scheme	umin(coarse)	umax(coarse)	umin(fine)	umax(fine)
CDG2LEGK1	−12.747	12.747	−100.241	100.241
CDG2LEGK2	−94.815	94.815	−99.987	99.987
CEVEFE-DDFV	−6.34E+01	64.462	−1.02E+02	102.394
FEM1	−1.87E-02	0.019	−9.78E+01	97.772
FVMON	−246.736	246.736	−99.719	99.719
LS-FVM	−1.00E+02	1.00E+02	−9.86E+01	98.562
MFD-GEN	−1.66E+02	1.66E+02	−9.95E+01	9.95E+01
MFD-PLAIN	−2.51E+02	250.808	−9.89E+01	98.887
SWPG-1	−5.46E+01	54.594	−9.78E+01	97.780
SWPG-2	−1.18E+02	118.325	−9.86E+01	98.563
SWPG-3	−1.05E+02	104.586	−9.86E+01	98.562
SUSHI	−2.49E+02	2.49E+02	−9.89E+01	9.89E+01
VAG	−7.65E+02	7.65E+02	−9.93E+01	9.93E+01
VAGR	−7.39E+02	7.39E+02	−1.00E+02	1.00E+02



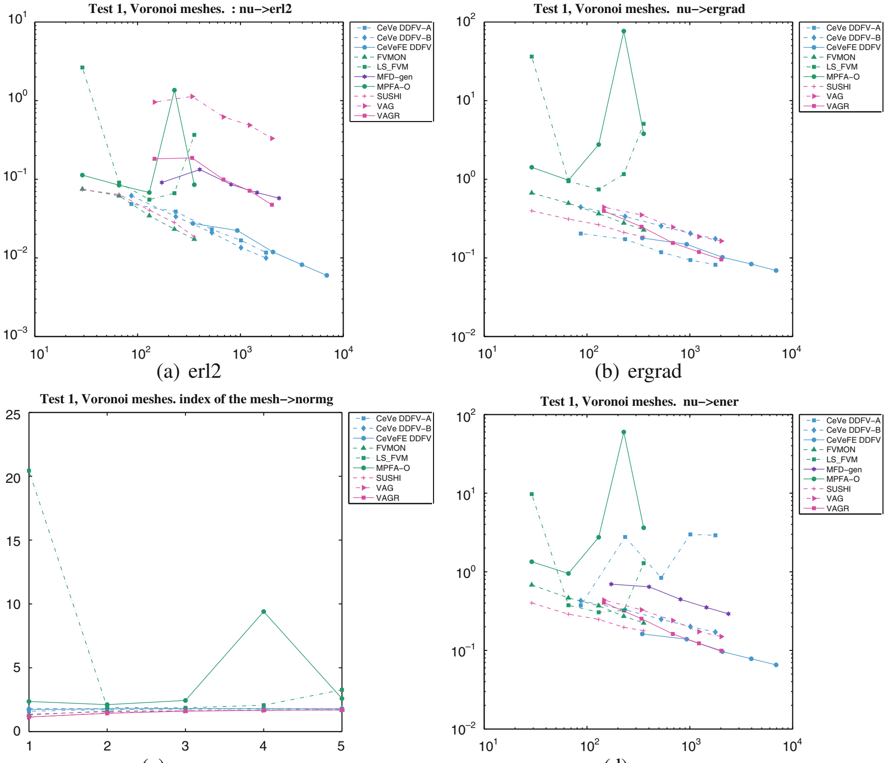
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 3 Accuracy of the schemes for Test Case 1 on tetrahedral meshes. Plot (*a*) shows the relative L^2-norm of the error, plot (*b*) shows the relative H^1-seminorm of the error, plot (*c*) the L^1-norm of the numerical gradient, and (*d*) the energy norm of the error



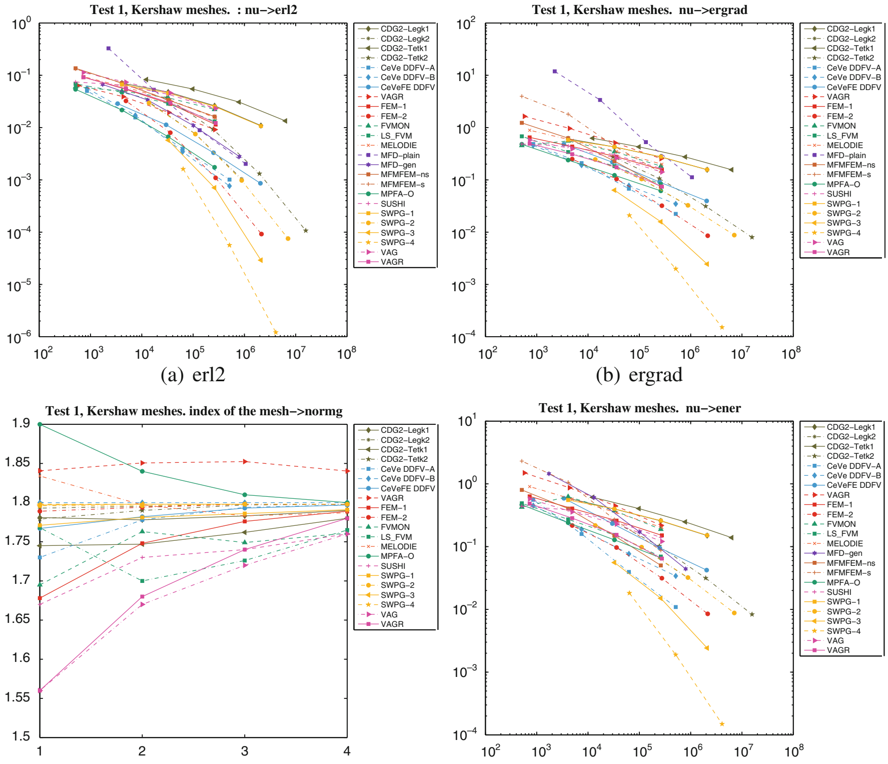
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 4 Accuracy of the schemes for Test Case 1 on Voronoi meshes. Plot (*a*) shows the relative L^2-norm of the error, plot (*b*) shows the relative H^1-seminorm of the error, plot (*c*) the L^1-norm of the numerical gradient, and (*d*) the energy norm of the error



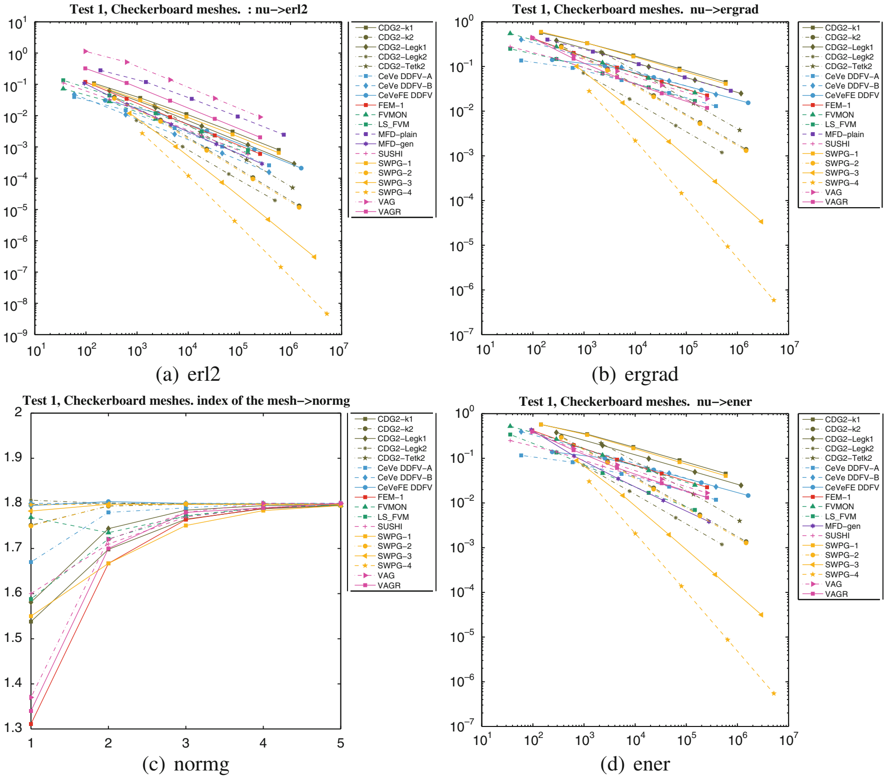
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 5 Accuracy of the schemes for Test Case 1 on Kershaw meshes. Plot (*a*) shows the relative L^2-norm of the error, plot (*b*) shows the relative H^1-seminorm of the error, plot (*c*) the L^1-norm of the numerical gradient, and (*d*) the energy norm of the error



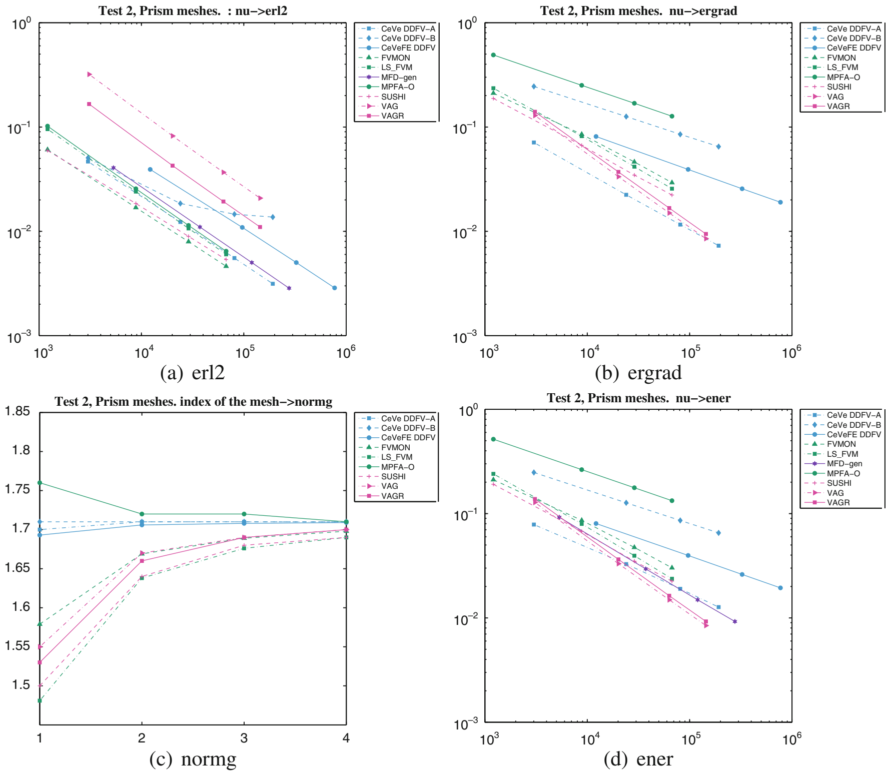
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 6 Accuracy of the schemes for Test Case 1 on Checkerboard meshes. Plot (*a*) shows the relative L^2-norm of the error, plot (*b*) shows the relative H^1-seminorm of the error, plot (*c*) the L^1-norm of the numerical gradient, and (*d*) the energy norm of the error



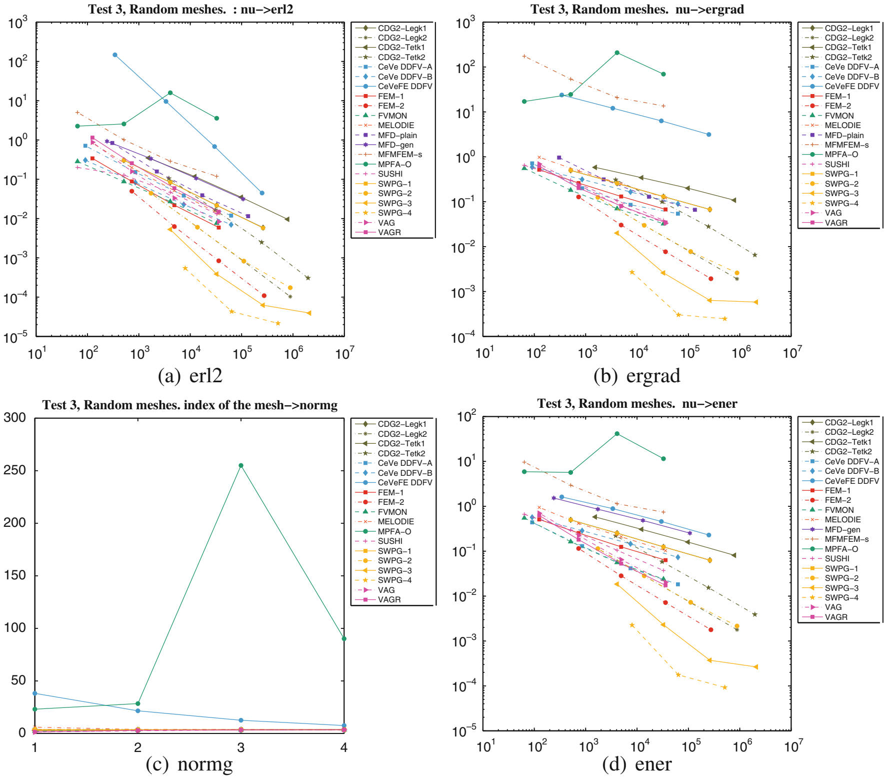
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 7 Accuracy of the schemes for Test Case 2. Plot (a) shows the relative L^2-norm of the error, plot (b) shows the relative H^1-seminorm of the error, plot (c) the L^1-norm of the numerical gradient, and (d) the energy norm of the error



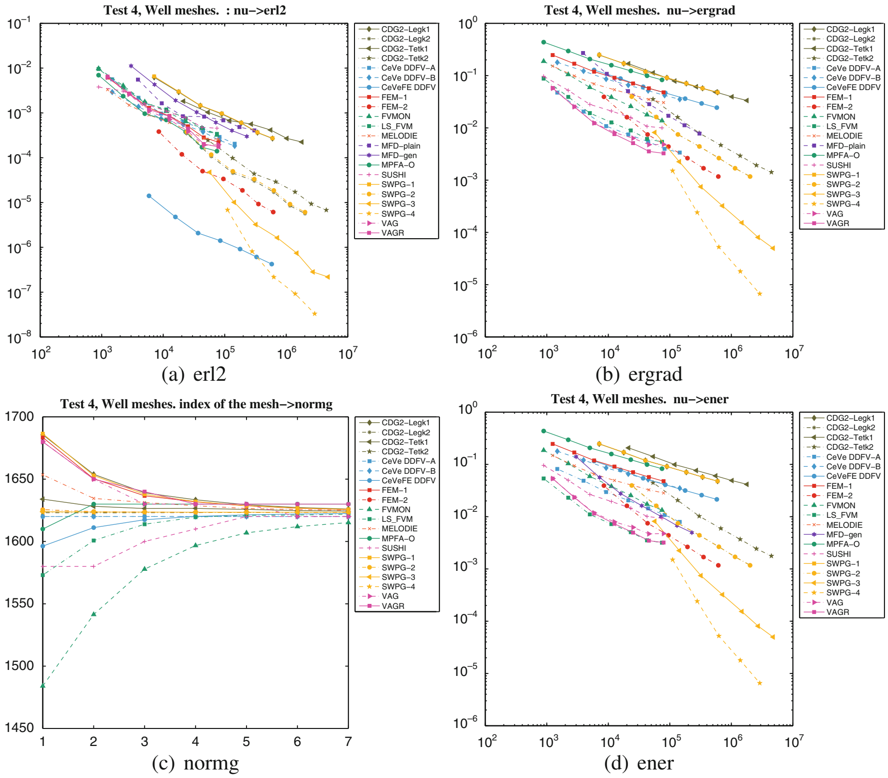
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 8 Accuracy of the schemes for Test Case 3. Plot (a) shows the relative L^2-norm of the error, plot (b) shows the relative H^1-seminorm of the error, plot (c) the L^1-norm of the numerical gradient, and (d) the energy norm of the error



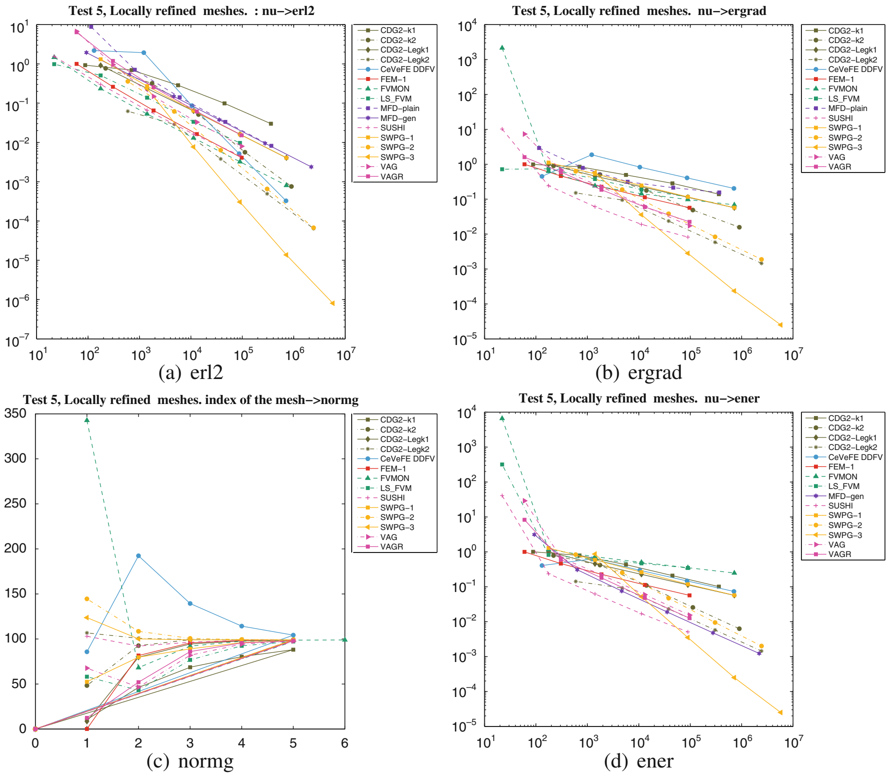
(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 9 Accuracy of the schemes for Test Case 4. Plot (a) shows the relative L^2-norm of the error, plot (b) shows the relative H^1-seminorm of the error, plot (c) the L^1-norm of the numerical gradient, and (d) the energy norm of the error



(a) erl2

(b) ergrad

(c) normg

(d) ener

Fig. 10 Accuracy of the schemes for Test Case 5. Plot (a) shows the relative L^2-norm of the error, plot (b) shows the relative H^1-seminorm of the error, plot (c) the L^1-norm of the numerical gradient, and (d) the energy norm of the error

Table 9 Matrix condition numbers for the first three meshes in the solution of Test Case 1 using Tetrahedral meshes.

Scheme	Condition number		
	i=1	i=2	i=3
CDG2-k1	6.96E+03	8.36E+03	1.81E+04
CDG2-k2	2.37E+04	2.80E+04	5.98E+04
CeVe DDFV-A	2.67E+02	3.72E+02	9.15E+02
CeVe DDFV-B	2.75E+02	3.91E+02	9.34E+02
CeVeFE DDFV	9.36E+02	1.31E+03	3.04E+03
FEM-1	2.68E+01	4.79E+01	7.39E+01
FEM-2	2.24E+02	3.66E+02	5.96E+02
FVMON	8.90E+02	9.56E+02	2.09E+03
LS-FVM	2.80E+02	3.91E+02	9.62E+02
MELODIE	1.18E+02	6.04E+01	1.65E+02
MFD-gen	1.34E+03	1.46E+03	3.32E+03
MFD-plain	1.46E+03	2.47E+03	3.87E+03
MPFA-O	3.28E+01	5.01E+01	8.64E+01
SUSHI	8.34E+02	1.32E+03	2.67E+03
SWPG-1	1.29E+04	1.53E+04	3.33E+04
SWPG-2	6.37E+04	7.42E+04	1.60E+05
SWPG-3	1.82E+05	2.13E+05	4.64E+05
SWPG-4	4.15E+05	4.86E+05	1.06E+06
VAG	2.68E+01	4.79E+01	7.39E+01
VAGR	2.68E+01	4.79E+01	7.39E+01

Table 10 Matrix condition numbers for the first three meshes in the solution of Test Case 1 using Voronoi meshes.

Scheme	Condition number		
	i=1	i=2	i=3
CeVe DDFV-A	9.51E+01	1.24E+02	3.33E+02
CeVe DDFV-B	5.07E+01	9.40E+01	2.05E+02
CeVeFE DDFV	1.05E+03	2.00E+05	1.98E+05
FVMON	1.03E+01	9.97E+00	1.58E+02
MPFA-O	5.78E+01	8.32E+01	–
SUSHI	1.45E+01	1.12E+01	3.07E+01
VAG	6.51E+01	7.95E+02	4.19E+02
VAGR	1.82E+01	3.68E+01	8.36E+01

Table 11 Matrix condition numbers for the first three meshes in the solution of Test Case 1 using Kershaw meshes.

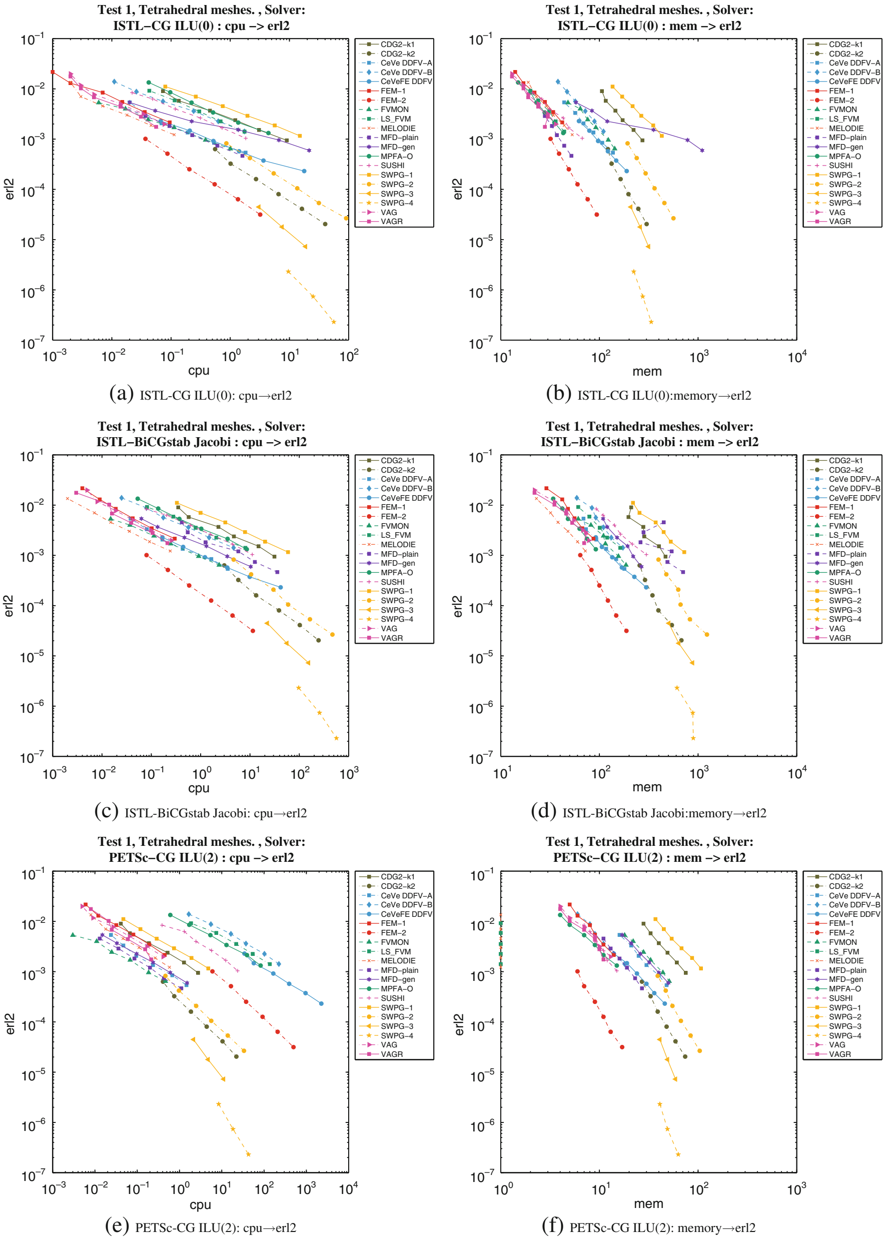
Scheme	Condition number		
	i=1	i=2	i=3
CDG2-Legk1	3.06E+04	1.84E+05	1.01E+06
CDG2-Legk2	1.99E+05	1.04E+06	–
CDG2-Tetk1	1.41E+05	6.14E+05	2.62E+06
CDG2-Tetk2	5.22E+05	2.17E+06	–
CeVe DDFV-A	6.67E+02	3.25E+03	1.54E+04
CeVe DDFV-B	7.08E+02	3.85E+03	1.84E+04
CeVeFE DDFV	3.80E+03	1.99E+04	9.77E+04
FEM-1	1.54E+02	1.12E+03	7.50E+03
FEM-2	2.55E+03	1.55E+04	9.58E+04
FVMON	3.31E+02	2.07E+03	8.65E+03
LS-FVM	2.86E+02	1.37E+03	9.76E+03
MELODIE	5.27E+02	2.27E+03	1.28E+04
MFD-gen	2.10E+03	7.53E+03	4.17E+04
MFD-plain	2.65E+03	1.29E+04	7.47E+04
MFMFEM-ns	1.12E+02	9.19E+02	6.88E+03
MFMFEM-s	2.02E+02	1.25E+03	7.77E+03
MPFA-O	8.19E+01	8.12E+02	5.31E+02
SUSHI	1.08E+03	2.51E+03	1.47E+04
VAG	1.80E+02	1.08E+03	7.28E+03
VAGR	1.76E+02	1.19E+03	7.62E+03

Table 12 Matrix condition numbers for the first three meshes in the solution of Test Case 1 using Checkerboard meshes.

Scheme	Condition number		
	i=1	i=2	i=3
CeVe DDFV-A	1.52E+01	5.20E+01	2.00E+02
CeVe DDFV-B	9.82E+00	3.39E+01	1.29E+02
CeVeFE DDFV	5.72E+01	2.31E+02	9.29E+02
FVMON	8.00E+00	2.62E+01	9.44E+01
MFD-plain	3.06E+01	1.71E+02	8.09E+02
SUSHI	6.96E+00	2.47E+01	9.83E+01
SWPG-1	–	1.50E+02	6.55E+02
VAG	3.41E+00	2.01E+01	1.46E+02
VAGR	2.62E+00	1.83E+01	1.42E+02

Table 13 Matrix condition numbers for the first three meshes in the solution of Test Case 2 using Prismatic meshes.

Scheme	Condition number		
	i=1	i=2	i=3
CeVe DDFV-A	2.08E+02	1.03E+03	2.51E+03
CeVe DDFV-B	1.31E+02	7.16E+02	1.79E+03
CeVeFE DDFV	1.17E+03	5.65E+03	1.35E+04
FVMON	7.23E+01	3.49E+02	8.41E+02
LS-FVM	9.77E+01	5.13E+02	1.29E+03
MPFA-O	8.65E+01	4.90E+02	1.27E+03
SUSHI	1.02E+02	5.26E+02	1.30E+03
VAG	7.44E+01	4.48E+02	1.42E+03
VAGR	9.57E+01	5.41E+02	1.42E+03



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

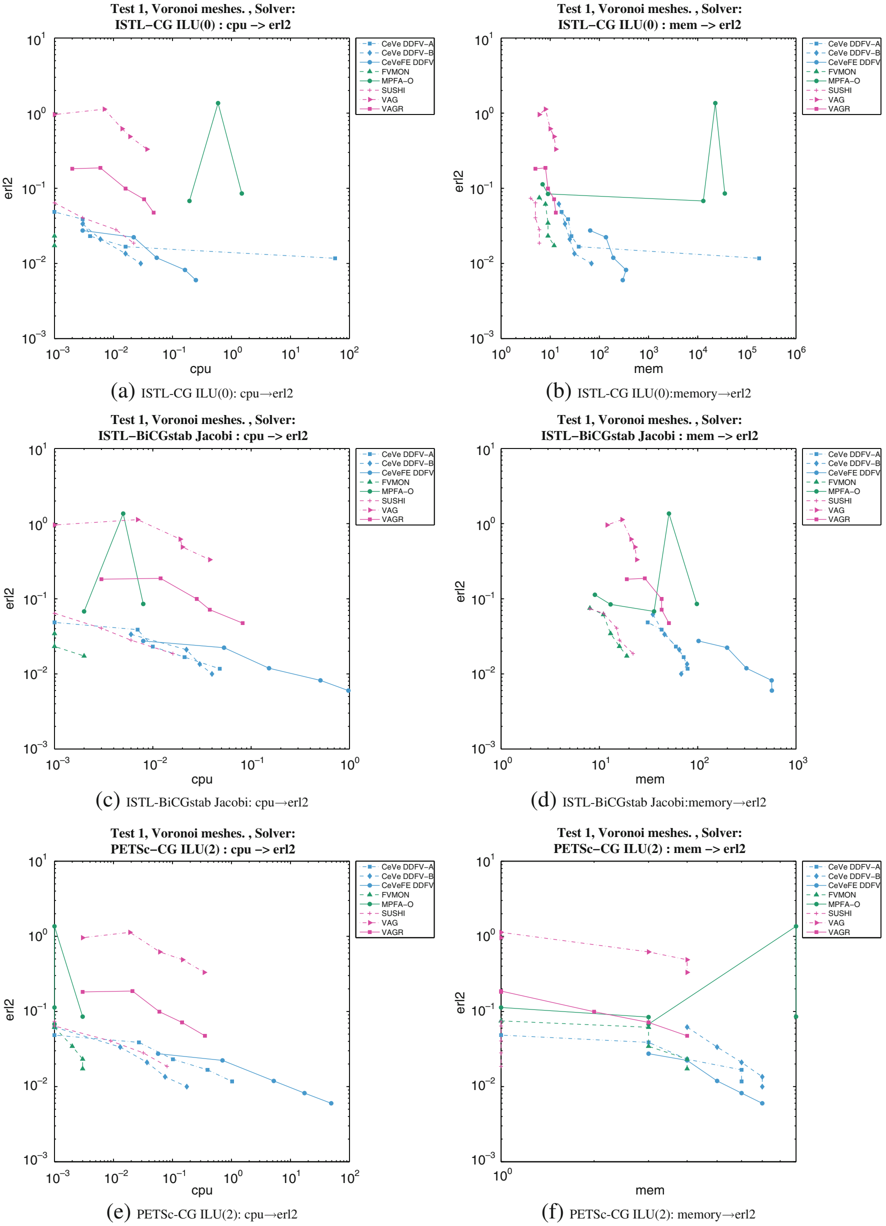
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 11 Test 1-Tetrahedral meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

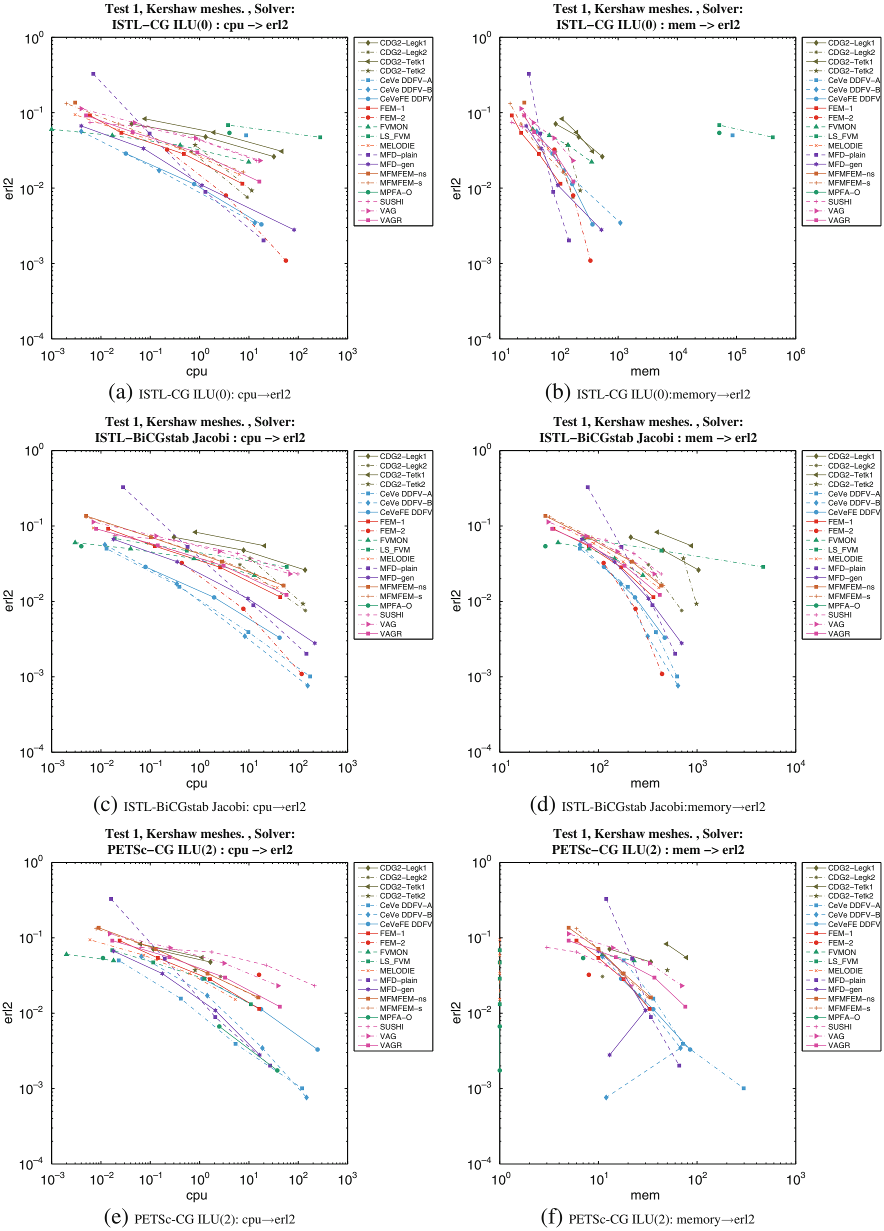
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 12 Test 1-Voronoi meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

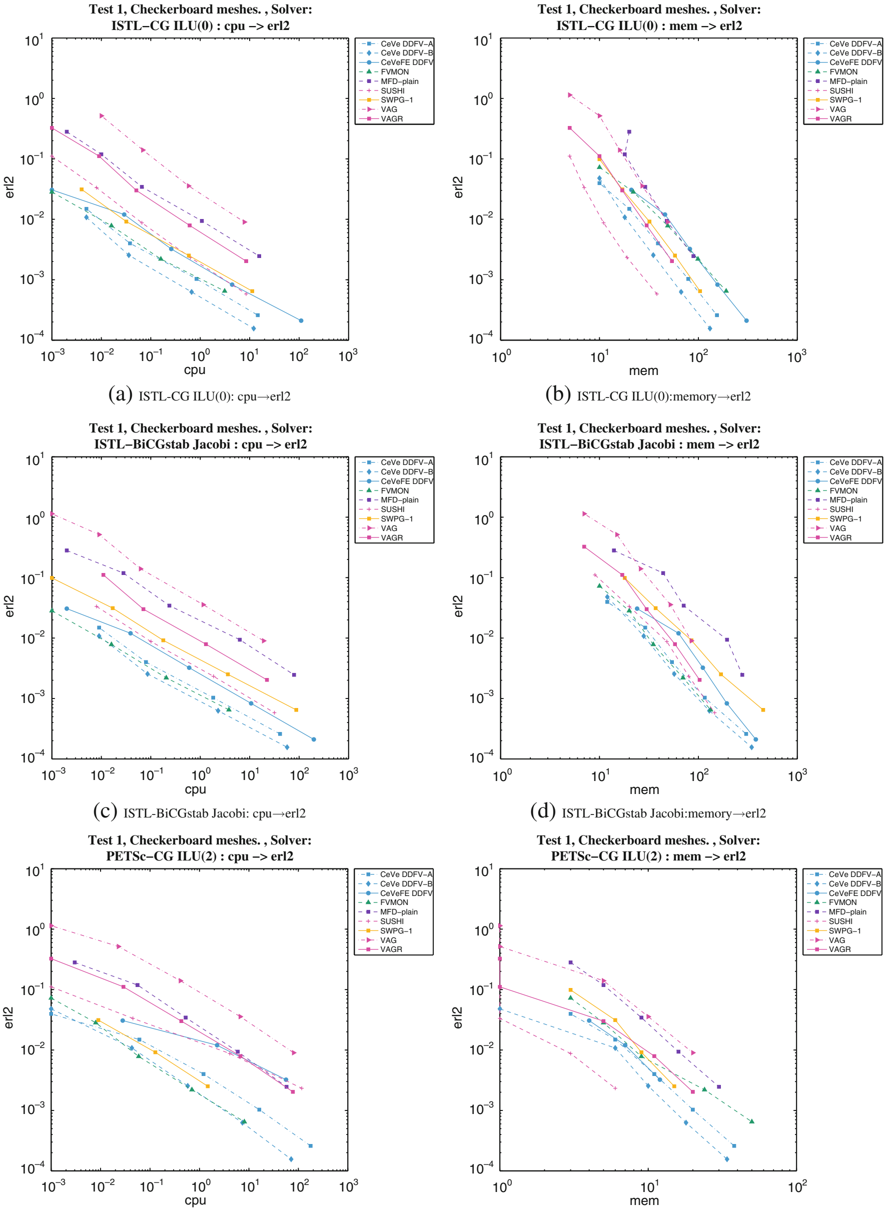
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 13 Test 1-Kershaw meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

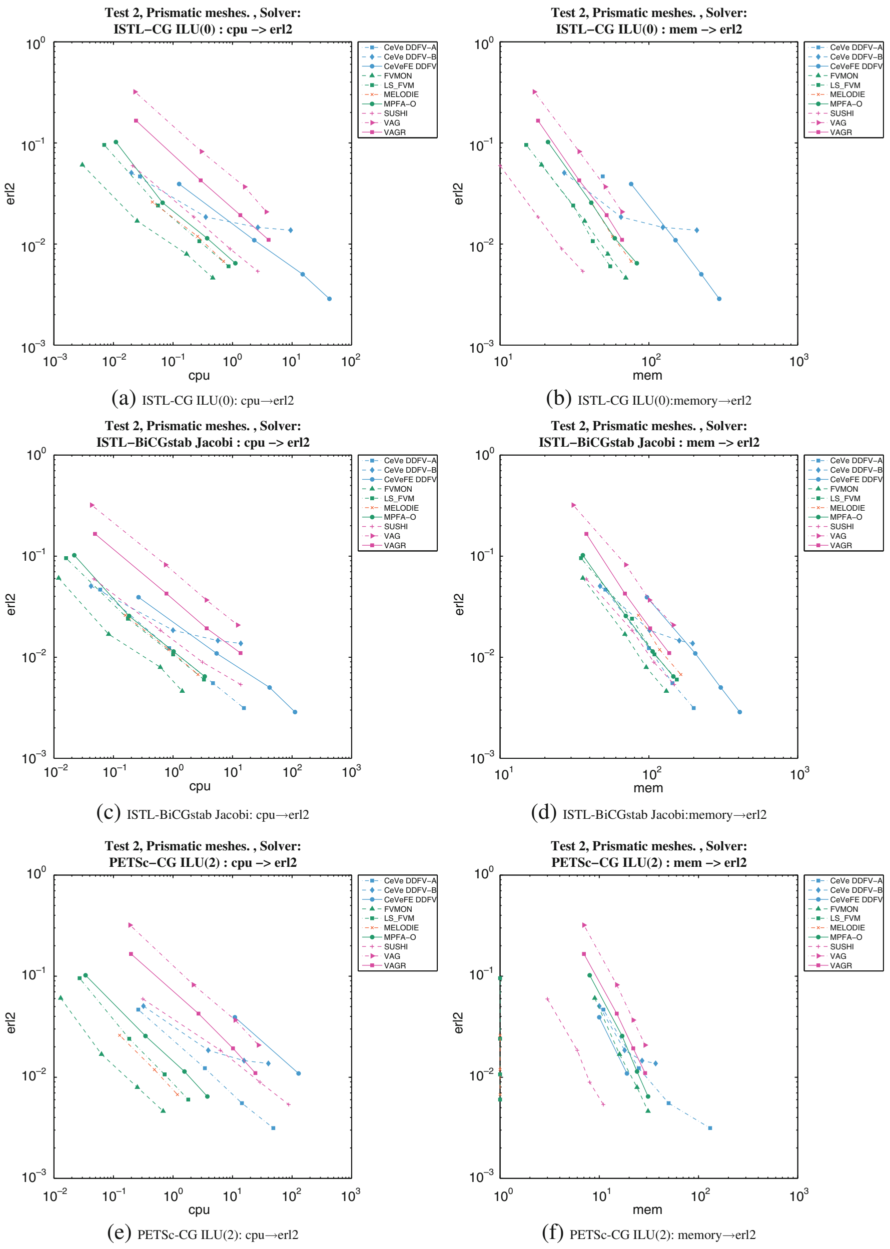
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 14 Test 1-Checkerboard meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

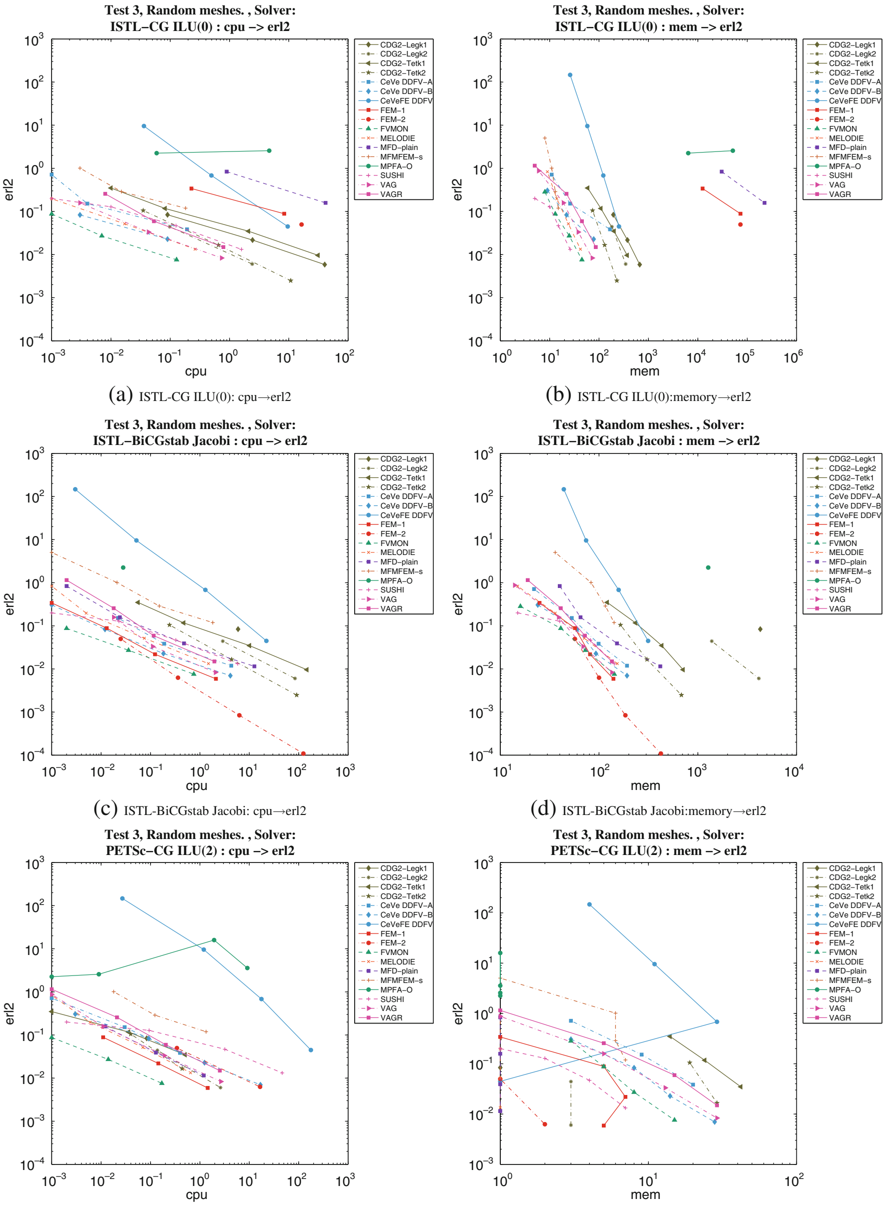
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 15 Test 2-Prismatic meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

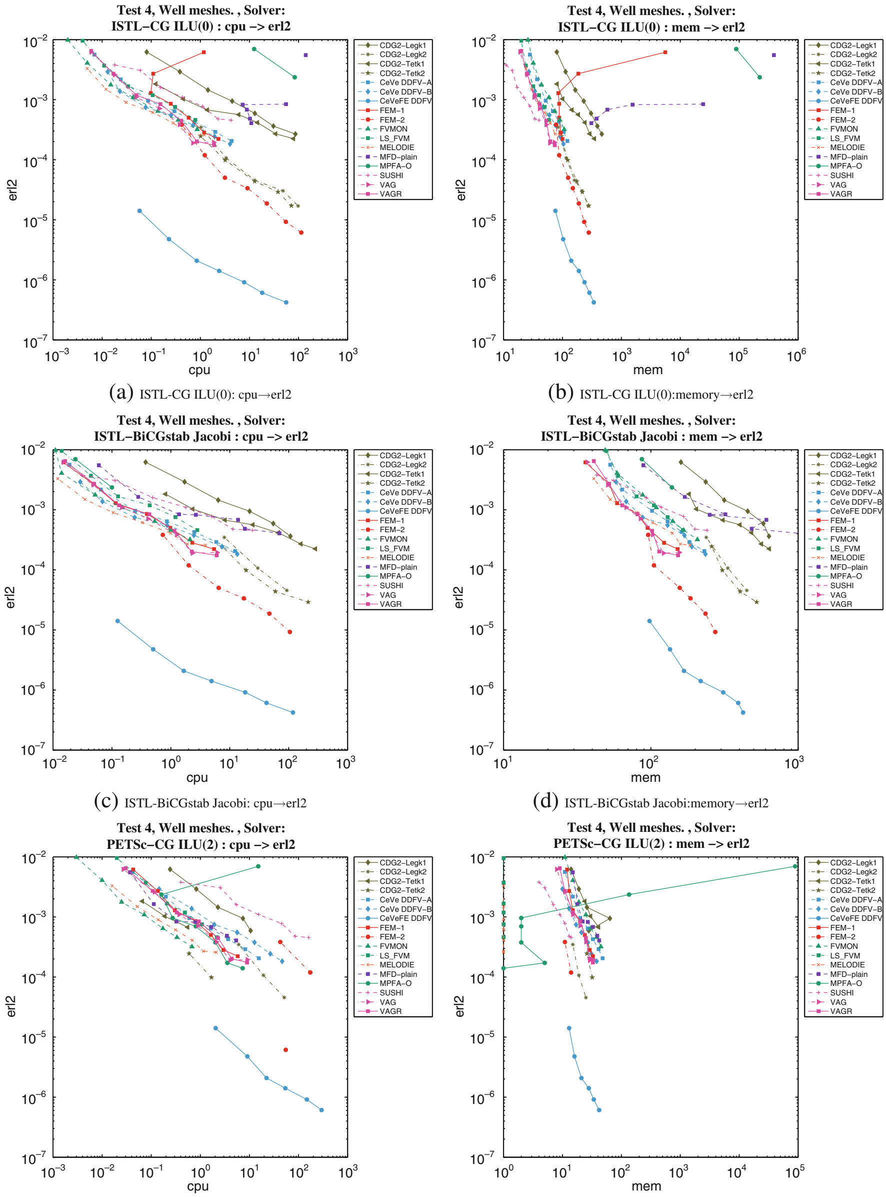
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 16 Test 3-Random meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

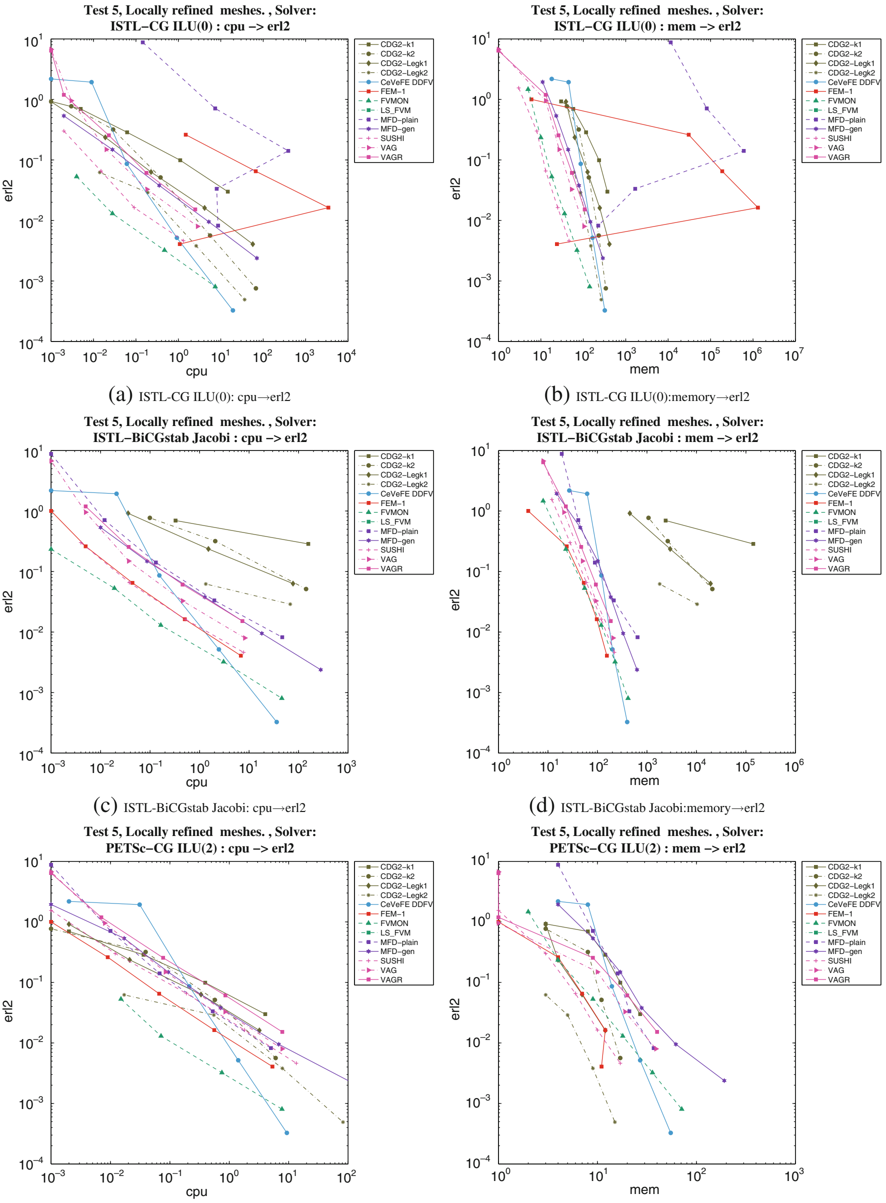
(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 17 Test 4- Well meshes



(a) ISTL-CG ILU(0): cpu→erl2

(b) ISTL-CG ILU(0):memory→erl2

(c) ISTL-BiCGstab Jacobi: cpu→erl2

(d) ISTL-BiCGstab Jacobi:memory→erl2

(e) PETSc-CG ILU(2): cpu→erl2

(f) PETSc-CG ILU(2): memory→erl2

Fig. 18 Test 5-Locally refined grid

Acknowledgements The present paper is a synthesis of the work of several people:

- we first warmly thank all participants for their active collaboration in this benchmark, and for letting us use their results in this synthesis,
- we are also very grateful to P. Bastian, C. Guichard, K. Lipnikov, R. Masson and S. Minjeaud for their help in the benchmark design.

This work was supported by GDR MOMAS, CNRS/PACEN and the ANR VFSitCom project. The work of the last author was also partially supported by the Italian MIUR through the program PRIN2008.

References

1. I. Aavatsmark and R. Klausen. Well index in reservoir simulation for slanted and slightly curved wells in 3D grids. *SPE Journal*, 8:41–48, 2003.
2. H. Amor, M. Bourgeois, and G. Mathieu. Benchmark 3D: a linear finite element solver. In *these proceedings*, 2011.
3. B. Andreianov, F. Hubert, and S. Krell. Benchmark 3D: a version of the DDFV scheme with cell/vertex unknowns on general meshes. In *these proceedings*, 2011.
4. S. Balay, J. Brown, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. C. McInnes, B. F. Smith, and H. Zhang. PETSc Web page, 2011. http://www.mcs.anl.gov/petsc.
5. S. Balay, W. D. Gropp, L. C. McInnes, and B. F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.
6. P. Bastian. Benchmark 3D: Symmetric weighted interior penalty discontinuous Galerkin scheme. In *these proceedings*, 2011.
7. P. Bastian, M. Blatt, A. Dedner, C. Engwer, J. Fahlke, C. Gräser, R. Klöfkorn, M. Nolte, M. Ohlberger, and O. Sander. DUNE Web page, 2011. http://www.dune-project.org.
8. P. Bastian, M. Blatt, A. Dedner, M. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. Part I: abstract framework. *Computing*, 82(2-3):103–119, 2008.
9. P. Bastian, M. Blatt, A. Dedner, M. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. Part II: implementation and tests in DUNE. *Computing*, 82(2-3):121–138, 2008.
10. P. Bastian, O. Ippisch, and S. Marnach. Benchmark 3D: A mimetic finite difference method. In *these proceedings*, 2011.
11. I. Ben Gharbia, J. Jaffré, S. N. Kumar, and J. E. Roberts. Benchmark 3D: a composite hexahedral mixed finite element. In *these proceedings*, 2011.
12. M. Blatt and P. Bastian. The iterative solver template library. In B. Kågström, E. Elmroth, J. Dongarra, and J. Wasniewski, editors, *Applied Parallel Computing. State of the Art in Scientific Computing*, volume 4699 of *Lecture Notes in Computer Science*, pages 666–675. Springer Berlin / Heidelberg, 2007.
13. Y. Coudière, F. Hubert, and G. Manzini. Benchmark 3D: CeVeFE-DDFV, a discrete duality scheme with cell/vertex/face+edge unknowns. In *these proceedings*, 2011.
14. Y. Coudière and G. Manzini. The cell-centered finite volume method using least squares vertex reconstruction (diamond scheme). In *these proceedings*, 2011.
15. Y. Coudière and C. Pierre. Benchmark 3D: CeVe-DDFV, a discrete duality scheme with cell/vertex unknowns. In *these proceedings*, 2011.
16. A. Danilov and Y. Vassilevski. Benchmark 3D: A monotone nonlinear finite volume method for diffusion equations on polyhedral meshes. In *these proceedings*, 2011.

17. T. A. Davis. Algorithm 832: UMFPACK V4.3 – an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):196–199, 2004.
18. T. A. Davis. UMFPACK Web page, 2011. http://www.cise.ufl.edu/research/sparse/umfpack/.
19. R. Eymard, T. Gallouët, and R. Herbin. Benchmark 3D: the SUSHI scheme. In *these proceedings*, 2011.
20. R. Eymard, C. Guichard, and R. Herbin. Benchmark 3D: the VAG scheme. In *these proceedings*, 2011.
21. R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.
22. V. Hernandez, J. E. Roman, and V. Vidal. SLEPc: A scalable and flexible toolkit for the solution of eigenvalue problems. *ACM Transactions on Mathematical Software*, 31(3):351–362, Sept. 2005.
23. R. Klöfkorn. Benchmark 3D: The compact discontinuous Galerkin 2 scheme. In *these proceedings*, 2011.
24. K. Lipnikov. and G. Manzini. Benchmark 3D: Mimetic finite difference method for generalized polyhedral meshes. In *these proceedings*, 2011.
25. M. F. Wheeler, G. Xue, and I. Yotov. Benchmark 3D: A multipoint flux mixed finite element method on general hexahedra. In *these proceedings*, 2011.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: a linear finite element solver

Hanen Amor, Marc Bourgeois, and Gregory Mathieu

1 Introduction

In the present paper[1], we address some of the benchmark problems defined for the Finite Volume for Complex Applications conference (FVCA6 [1]). The tests, which are described in [2], consist in solving the following anisotropic diffusion problem :

$$\begin{cases} -\nabla.(K\nabla u) = \mathrm{f} & \text{on} \quad \Omega \\ \mathrm{u} = \overline{u} & \text{on} \quad \Gamma_d \end{cases} \tag{1}$$

where u is the unknown, Ω is in most cases a unit cube, $K : \Omega \rightarrow \mathbb{R}^{3\times3}$ is the diffusion tensor, f the source term and $\overline{u}$ the Dirichlet boundary conditions.
For this benchmark, the computations were performed with MELODIE (Modèle d'Evaluation à LOng terme des Déchets Irradiants Enterrés) software, which is devoted to simulate the migration of a plume of radionuclides in a 3-dimensional geological media.

2 Presentation of MELODIE

The MELODIE [3] software, is developped by IRSN, and constantly upgraded, to assess the long-term containment capabilities of radioactive waste repositories. This software is designed to model a disposal site taking into account all the main

[1] The model of this paper is provided by the benchmark organization. The results of this benchmark will be detailed and discussed by Florence HUBERT and Raphaële HERBIN in a paper gathering all the contributions.

Hanen Amor, Marc Bourgeois, and Gregory Mathieu
Institut de Radioprotection et de Sûreté Nucléaire - DSU/SSIAD/BERIS - BP 17 - 92262 Fontenay-aux-Roses Cedex,
email: {hanen.amor,marc.bourgeois,gregory.mathieu}@irsn.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_90, © Springer-Verlag Berlin Heidelberg 2011

physical and chemical characteristics of the disposal components. The model is adapted to large scales of time and space required for simulation.
The MELODIE software models water flow and the phenomena involved in the transport of radionuclides in saturated porous media in 2 dimensions and in 3 dimensions; physical and chemical interactions are represented by a retardation factor integrated in the computational equations. These equations are discretised using a so-called FVFE method -Finite Volume Finite Element-, which is based on a Galerkin method to discretise time and variables, together with a finite volume method using the Godunov scheme for the convection term. The FVFE method is used to convert partial differential equations into a finite number of algebraic equations that match the number of nodes in the mesh used to model the considered site. It also serves to stabilise the numerical scheme. The present benchmark adresses only diffusive problems, which are therefore solved by using a standard P1 finite element method.

3 Numerical results

Numerical results presented in that contribution concern the tests 1, 3 and 4. The error generated by the P1 method has been evaluated by defining the quantity: $e = u - u_h$, where u is the analytical solution and u_h is the numerical solution. Then the error can be computed as follows.

3.1 Discrete L^2 and H^1 norms

The continuous L^2 and H^1 norms of a function u are given by

$$\|u\|_{L^2(\Omega)} = \left(\int_\Omega u^2\right)^{1/2} \quad , \quad \|u\|_{H^1(\Omega)} = \|u\|_{L^2(\Omega)} + \|\nabla u\|_{L^2(\Omega)}$$

where Ω is an open bounded in $\mathbf{R}^3$. In most of the test cases, the domain Ω is a unit cube. To compute those norms, we perform the L^2 and H^1 semi-norm of the function u on a tetrahedron T:

$$\|u\|_{L^2(T)} = \left(\int_T u^2\right)^{1/2} \quad \text{and} \quad \|\nabla u\|_{H^1(T)} = \left(\int_T \nabla u^2\right)^{1/2}$$

The numerical quadrature used to approximate this integral, are given by the following formula:

- in the case where the values of the function u are known on the vertices

$$\int_T u^2 dx \simeq \frac{1}{4} V_T \sum_{i=1}^{4} u(\overline{s_i})^2$$

- in the case where values of the gradient of the function u are known on the centre of gravity

$$\int_T \nabla u.\nabla u dx \simeq V_T \nabla u(G_T).\nabla u(G_T)$$

The previous formula are adapted to calculate the relative L^2 norm of the error : erl2, the relative L^2 norm of a gradient of the error : ergrad and the relative L^2 norm of the energy norm : ener.

3.2 Expected results

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	488	6072	7.69E-02	8.29E-02	1.935	1.935	1.791
2	857	11269	2.76E-02	2.83E-02	1.955	1.955	1.796
3	1601	21675	3.07E-02	3.07E-02	1.970	1.969	1.798
4	2997	41839	1.81E-02	1.77E-02	1.984	1.983	1.797
5	5692	81688	1.32E-02	1.37E-02	1.990	1.990	1.798
6	10994	160852	6.19E-03	6.49e-03	1.991	1.991	1.798

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	488	1.35E-02	-	2.32E-01	-	2.29E-01	
2	857	7.01E-03	3.531	1.17E-01	1.370	1.17E-01	1.362
3	1601	4.56E-03	2.052	1.14E-01	1.052	1.14E-01	1.082
4	2997	3.01E-03	1.998	1.13E-01	1.155	1.11E-01	1.170
5	5692	1.87E-03	2.219	9.03E-02	1.067	8.90E-02	1.035
6	10994	1.22E-03	1.941	7.05E-02	1.128	6.92E-02	1.148

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	729	9097	1.34E-01	8.76E-02	1.833	1.883	1.834
2	4913	66961	3.12E-02	1.92E-02	1.955	1.970	1.797
3	35937	513313	8.55E-03	6.64E-03	1.988	1.992	1.783
4	274625	4018753	2.04E-03	1.92E-03	1.997	1.998	1.787

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	729	9.41E-02	-	8.88E-01	-	9.05E-01	-
2	4913	5.88E-02	0.737	5.70E-01	0.696	5.710E-01	0.723
3	35937	3.35E-02	0.849	3.49E-01	0.741	3.47E-01	0.750
4	274625	1.52E-02	1.158	1.99E-01	0.823	2.02E-01	0.793

- **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	125	1333	-0.665	-0.338	0.685	0.363	6.004
2	729	9097	-0.885	-0.784	0.812	0.751	3.867
3	4913	66961	-0.970	-0.943	0.949	0.925	3.666
4	35937	513313	-0.988	-0.982	0.991	0.984	3.613

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	125	8.34E-01	-	9.77E-01	-	9.44E-01	-
2	729	1.97E-01	2.456	4.84E-01	1.193	4.17E-01	1.390
3	4913	5.16E-02	2.107	2.48E-01	1.051	2.10E-01	1.078
4	35937	1.33E-02	2.040	1.22E-01	1.067	1.03E-01	1.066

- **Test 4 Flow around a well, Well meshes,** $\min = 0,\ \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1248	15886	0.189	0.189	5.360	5.360	1653.52
2	2800	37836	0.120	0.119	5.368	5.368	1634.57
3	5889	81531	0.078	0.076	5.345	5.345	1631.27
4	12582	178018	0.060	0.058	5.349	5.349	1628.68
5	25300	363768	0.046	0.045	5.377	5.377	1626.49
6	45668	662730	0.037	0.036	5.380	5.380	1625.64
7	79084	1154172	0.029	0.028	5.39	5.39	1624.98

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1248	3.30E-03	-	1.52E-01	-	1.48E-01	-
2	2800	1.49E-03	2.941	9.83E-02	1.621	9.40E-02	1.703
3	5889	8.99E-04	2.058	6.93E-02	1.409	6.52E-02	1.472
4	12582	6.11E-04	1.522	5.36E-02	1.014	4.93E-02	1.104
5	25300	4.09E-04	1.722	4.26E-02	0.983	3.94E-02	0.960
6	45668	2.69E-04	2.130	3.50E-02	0.997	3.26E-02	0.954
7	79084	2.58E-04	0.224	3.05E-02	0.753	2.84E-02	0.751

4 Comments

The computations for the post-processing purpose of the relative L^2 error norm and the relative H^1 error semi-norm have been done using the continuous solution and a numerical quadrature rule presented in the section 3.1. As can be seen, for the test 1 using the Tetrahedral meshes, for the test 3 using the Random meshes and for the test 4 using the Well meshes, the theoretical results are recovered, since a convergence of order 2 for the L^2-norm and a convergence of order 1 for the H^1-norm are obtained. For the test 1 using the Kershaw meshes, the theoretical results are not recovered. In fact, a convergence of order 1 for the L^2-norm and a convergence of order 1/2 for the H^1-norm are obtained. Those decreases in the rates of convergence orders are due to the characteristics of the Kershaw meshes that present strong anisotropy.

In addition, FVFE method implemented in MELODIE is only available for tetrahedrons and conform meshes. In the benchmark, hexahedral meshes are divided in tetrahedrons without changing the number of vertices. For the tests 2 and 5, that kind of adaptation is not possible due to the specific shape of the meshes. It is the reason why those cases were not considered.

In this benchmark, the system obtained after assembling of the discretized equations on each element is linear. Within MELODIE, this linear system is solved by using a bi-conjugate gradient method with an incomplete Gauss-type preconditioning. That method is specifically suitable for resolution of non-symmetrical system. Thereby, among the solvers proposed in the benchmark, our choice was the Petsc bi-conjugate gradient (using various preconditioning) complying with the implemented method in MELODIE.

References

1. website : http://fvca6.fs.cvut.cz/
2. website : http://www.latp.univ-mrs.fr/latp_numerique/
3. website : http://www.irsn.fr/FR/Larecherche/outils-scientifiques/Codes-de-calcul/Pages/Le-logiciel-MELODIE-3133.aspxl

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: a version of the DDFV scheme with cell/vertex unknowns on general meshes

Boris Andreianov, Florence Hubert, and Stella Krell

1 DDFV methods in 2D and in 3D. A 3D CeVe-DDFV scheme

This paper gives numerical results for a 3D extension of the 2D DDFV scheme. Our scheme is of the same inspiration as the one called CeVe-DDFV ([9]), with a more straightforward dual mesh construction. We sketch the construction in which, starting from a given 3D mesh (which can be non conformal and have arbitrary polygonal faces), one defines a dual mesh and a diamond mesh, reconstructs a discrete gradient, and proves the discrete duality property. Details can be found in [1].

DDFV ("Discrete Duality Finite Volume") scheme was introduced in 2D by Hermeline in [15] and by Domelevo and Omnès in [13]. To handle anisotropic problems or nonlinear problems, or in order to work on general distorted meshes, full gradient reconstruction from point values is a popular strategy. It is well known that reconstruction of a discrete gradient is facilitated by adding unknowns that are new with respect to those of standard cell-centered finite volume schemes. The 2D DDFV method consists in adding new unknowns at the vertices of the initial mesh (this initial mesh is often called the primal one), and in use of new control volumes (called dual cells, or co-volumes) around these points. A family of diamond cells is naturally associated to this construction, each diamond being built on two neighbor cell centers x_K, x_L and the two vertices of the edge $K|L$ that separates them. On a diamond, one can construct a discrete gradient direction per direction (cell-cell and vertex-vertex), following the idea of [8]. It turns out that this discrete gradient

Boris Andreianov
CNRS UMR 6623, Besançon, France, e-mail: boris.andreianov@univ-fcomte.fr

Florence Hubert
LATP, Université de Provence, Marseille, France, e-mail: fhubert@cmi.univ-mrs.fr

Stella Krell
INRIA, Lille, France, e-mail: stella.krell@inria.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_91, © Springer-Verlag Berlin Heidelberg 2011

is related by a discrete analogue of integraton-by-parts formula, called "discrete duality", to the classical discrete finite volume divergence associated with these two families of meshes. This duality property greatly simplifies the theoretical analysis of finite volume schemes based on the DDFV construction, see e.g. [2, 5]. This 2D strategy reveals to be particularly efficient in terms of gradient approximation (see [7, 14]) and has been extended to a wide class of PDE problems (see [1, 5, 6, 18, 19] and references therein).

The 3D CeVe-DDFV scheme we present here also keeps unknowns only at the cell centers and the vertices of the primal mesh, and it uses the primal mesh, a dual mesh and a diamond mesh; as in the 2D case, a diamond is constructed from two neighbor cell centers x_K, x_L and from l vertices of the edge $K|L$ that separates them ($l \geq 3$). The price to pay is that the gradient reconstruction becomes more intricate. As in 2D, one direction per diamond is reconstructed using the two cell center unknowns at the nodes x_K, x_L; two complementary directions of the gradient in $K|L$ are reconstructed simultaneously, by a suitable interpolation of the vertex values in each face $K|L$ of the primal mesh. While the case $l = 3$ (meshes with triangular faces) offers no choice, in general we have to fix a formula for interpolation that is consistent with affine functions and which leads to discrete duality (with respect to appropriately defined dual cells). This was achieved independently in [17] and in [1, 3, 4], with two different approaches (the above description stems from the point of view developed in [1, 3, 4]).

Several 3D DDFV constructions exist. The CeVe-DDFV scheme by Pierre et al. (see [12]) was the pioneering work in 3D; a particular feature of this method was in the double covering of the domain by the dual mesh. This approach led to a method that is only slightly different from ours; we refer to the benchmark paper [9] in the same collection. Next, Hermeline in [16] introduced the important idea to associate additional unknowns with the face centers of the primal mesh. In the subsequent work [17] of Hermeline, elimination of these unknowns eventually led to the same method that the one we describe. Many numerical tests are given in [16, 17]. Finally, Coudière and Hubert in [10] introduced edge unknowns, instead of eliminating face unknowns. This idea assessed a new strategy of 3D DDFV approximation; we call it CeVeFE-DDFV because with respect to the primal mesh, cell, vertex and face+edge unknowns are used. Let us point out the differences with respect to CeVe-DDFV strategies. In [10], each diamond is constructed on two cell centers x_K, x_L, on two vertices x_{K^*}, x_{L^*} in the face $K|L$, and one face center $x_{K|L} \in K|L$ and one edge center $x_{K^*|L^*} \in [x_{K^*}, x_{L^*}]$. Then the gradient is reconstructed per direction (cell-cell, vertex-vertex and face-edge), as in 2D. The edge and face centers are the centers for a new, third mesh. The CeVeFE-DDFV method is the object of the benchmark paper [11] in the same collection.

Let us present the construction of our 3D CeVe-DDFV scheme. The primal mesh needs not be conformal; there is no restriction on number of faces or face edges. For simplicity, let us assume that the primal mesh volumes are convex; that their centers belong to the volumes; and the face centers belong to the faces. These restrictions can be relaxed, see [1]; but let us stress that the edge points must be the middlepoints.

Notation. We use a triple $\mathfrak{T} = \left(\overline{\mathfrak{M}^o}, \overline{\mathfrak{M}^*}, \mathfrak{D}\right)$ of partitions of Ω into polyhedra.

- $\mathfrak{M}^o$ denotes the initial mesh[1], called *primal mesh*. We call $\partial\mathfrak{M}^o$ the set of all faces of this mesh that are included in $\partial\Omega$. These faces are considered as flat *boundary (primal) control volumes*. We denote by $\overline{\mathfrak{M}^o}$ the union $\mathfrak{M}^o \cup \partial\mathfrak{M}^o$.
  - **Center:** To any (primal) control volume $K \in \overline{\mathfrak{M}^o}$, we associate a point $x_K \in K$.
  - **Vertex:** A generic vertex of $\overline{\mathfrak{M}^o}$ is denoted by x_{K^*}.
  - **Neighbors:** given $K \in \mathfrak{M}^o$, all control volumes $L \in \overline{\mathfrak{M}^o}$ such that K and L have a common face (or part of a face) form the set $\mathcal{N}(K)$ of neighbors of K.
  - **Face:** for all $L \in \mathcal{N}(K)$, by $K|L$ we denote $\partial K \cap \partial L$ which is a face (or a part of a face) of the mesh $\mathfrak{M}^o$; it is supplied with a *face center* $x_{K|L} \in K|L$.
  - **Edge:** An egde $[x_{K^*}, x_{L^*}]$ of $\overline{\mathfrak{M}^o}$ is defined by two neighbor vertices x_{K^*}, x_{L^*}; it is marked with the center $x_{K^*|L^*}$ that must be its middlepoint $(x_{K^*} + x_{L^*})/2$.
  - **Element:** An element $T = T^{K;L}_{K^*;L^*}$ is the tetrahedron $(x_K, x_{K|L}, x_{K^*|L^*}, x_{K^*})$: here K is a primal volume ; $K|L$ is a face of K ; and $[x_{K^*}, x_{L^*}]$ is an edge of $K|L$ (see Fig. 1). The set of all elements is denoted by $\mathcal{T}$. If x_K is a vertice of $T \in \mathcal{T}$, then we say that T is associated[2] with the volume K, and we write $T \sim K$.
- $\overline{\mathfrak{M}^*}$ denotes the *dual mesh* constructed as follows. A generic vertex x_{K^*} of $\mathfrak{M}^o$ is associated with the polyhedron $K^* \in \overline{\mathfrak{M}^*}$ made of all elements $T \in \mathcal{T}$ that share the vertex x_{K^*} (we write $T \sim K^*$). If $x_{K^*} \in \Omega$, we say that K^* is a *dual control volume* and write $K^* \in \mathfrak{M}^*$; and if $x_{K^*} \in \partial\Omega$, we say that K^* is a *boundary dual control volume* and write $K^* \in \partial\mathfrak{M}^*$. Thus $\overline{\mathfrak{M}^*} = \mathfrak{M}^* \cup \partial\mathfrak{M}^*$.
- $\mathfrak{D}$ is the *diamond mesh*. For $K \in \mathfrak{M}^o$, $L \in \mathcal{N}(K)$, the union of the convex hull of x_K and $K|L$ with the convex hull of x_L and $K|L$ is called *diamond*, denoted by $D^{K|L}$.

For expression of the discrete operators one needs a convention on diamond orientation, subdiamonds and other objects and notation of [1]; we give them via Fig. 1.

Discrete space and discrete operators; the discrete duality feature.

- A *discrete function on* Ω is a set $w^{\mathfrak{T}} = \left(w^{\mathfrak{M}^o}, w^{\mathfrak{M}^*}\right)$ consisting of two sets of real values $w^{\mathfrak{M}^o} = (w_K)_{K \in \mathfrak{M}^o}$ and $w^{\mathfrak{M}^*} = (w_{K^*})_{K^* \in \mathfrak{M}^*}$.
- A *discrete function on* $\overline{\Omega}$ is a set $w^{\overline{\mathfrak{T}}} = \left(w^{\mathfrak{M}^o}, w^{\mathfrak{M}^*}; w^{\partial\mathfrak{M}^o}, w^{\partial\mathfrak{M}^*}\right) \equiv \left(w^{\mathfrak{T}}; w^{\partial\mathfrak{T}}\right)$, $w^{\mathfrak{M}^o} = (w_K)_{K \in \mathfrak{M}^o}$, $w^{\mathfrak{M}^*} = (w_{K^*})_{K^* \in \mathfrak{M}^*}$, $w^{\partial\mathfrak{M}^o} = (w_K)_{K \in \partial\mathfrak{M}^o}$, $w^{\partial\mathfrak{M}^*} = (w_{K^*})_{K^* \in \partial\mathfrak{M}^*}$.
- A *discrete field on* Ω is a set $\overrightarrow{\mathcal{F}}^{\mathfrak{T}} = \left(\overrightarrow{\mathcal{F}_D}\right)_{D \in \mathfrak{D}}$ of vectors of $\mathbb{R}^3$.
- We write $\mathbb{R}^{\mathfrak{T}}$, $\mathbb{R}^{\overline{\mathfrak{T}}}$, $(\mathbb{R}^3)^{\mathfrak{D}}$, respectively, for the sets of discrete functions/fields.

[1] This means, $\mathfrak{M}^o$ is one of the meshes provided by the benchmark organizers.

[2] Because we have made the assumption that $x_{K|L} \in K|L$, the relation $T \sim K$ simply means that T is included in K. The same observation applies to the notation $T \sim K^*$. See [1] for generalizations.

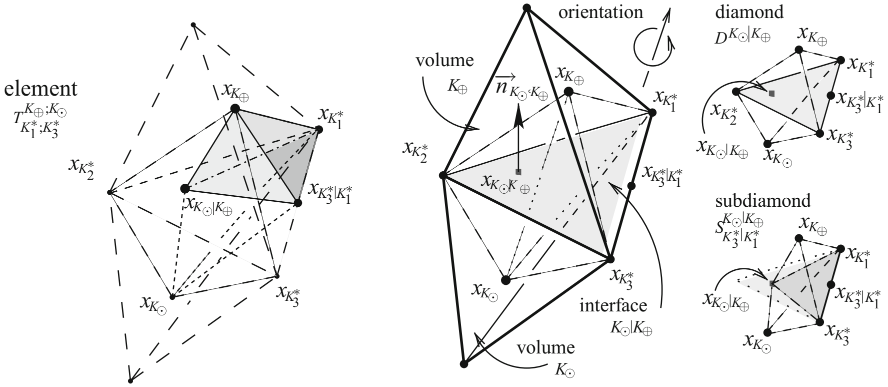


Fig. 1 Element (left). Oriented diamond, subdiamond and related notation, cf. [1] (right)

- *Discrete divergence* is the operator acting from $(\mathbb{R}^3)^{\mathfrak{D}}$ to $\mathbb{R}^{\mathfrak{T}}$, given by

$$\operatorname{div}^{\mathfrak{T}} : \ \overrightarrow{\mathscr{F}}^{\mathfrak{T}} \ \mapsto \ \left(\left(\operatorname{div}_K \overrightarrow{\mathscr{F}}^{\mathfrak{T}}\right)_{K\in\mathfrak{M}^o}, \left(\operatorname{div}_{K^*}\overrightarrow{\mathscr{F}}^{\mathfrak{T}}\right)_{K^*\in\mathfrak{M}^*}\right) =: \ \operatorname{div}^{\mathfrak{T}}\overrightarrow{\mathscr{F}}^{\mathfrak{T}}, \tag{1}$$

  where the entries $\operatorname{div}_K\overrightarrow{\mathscr{F}}^{\mathfrak{T}}$, $\operatorname{div}_{K^*}\overrightarrow{\mathscr{F}}^{\mathfrak{T}}$ of the discrete function $\operatorname{div}^{\mathfrak{T}}\overrightarrow{\mathscr{F}}^{\mathfrak{T}}$ on Ω are

$$\operatorname{div}_K\overrightarrow{\mathscr{F}}^{\mathfrak{T}} = \frac{1}{Vol(K)}\sum_{T\sim K} m_T \overrightarrow{\mathscr{F}}_T\cdot\overrightarrow{n}_T, \quad \operatorname{div}_{K^*}\overrightarrow{\mathscr{F}}^{\mathfrak{T}} = \frac{1}{Vol(K^*)}\sum_{T\sim K^*} m_T^* \overrightarrow{\mathscr{F}}_T\cdot\overrightarrow{n}_T^*, \tag{2}$$

  $\overrightarrow{n}_T$, $\overrightarrow{n}_T^*$ being the exterior normal vectors to ∂K, ∂K^*. Formulae (2) stem from the standard procedure of finite volume discretization, applied on $\mathfrak{M}^o$ and on $\mathfrak{M}^*$.
- *Discrete gradient* is the operator acting from $\mathbb{R}^{\overline{\mathfrak{T}}}$ to $(\mathbb{R}^3)^{\mathfrak{D}}$, given by

$$\overrightarrow{\nabla}^{\mathfrak{T}} : \ w^{\overline{\mathfrak{T}}} \ \mapsto \ \left(\overrightarrow{\nabla}_D\, w^{\overline{\mathfrak{T}}}\right)_{D\in\mathfrak{D}} \ =: \ \overrightarrow{\nabla}^{\mathfrak{T}} w^{\overline{\mathfrak{T}}} \tag{3}$$

  where the entry $\overrightarrow{\nabla}_D\, w^{\overline{\mathfrak{T}}}$ of the discrete field $\overrightarrow{\nabla}^{\mathfrak{T}} w^{\overline{\mathfrak{T}}}$ corresponding to $D = D^{K_\odot|K_\oplus}$ (see Fig. 1) is reconstructed from the values $w_{K_\odot}, w_{K_\oplus}$ at the neighbor centers $x_{K_\odot}, x_{K_\oplus}$ (they give the projection on $\overrightarrow{x_{K_\odot}x_{K_\oplus}}$) and the values $(w_{K^*_i})_{i=1}^{l}$ at the l vertices of the interface $K_\odot|K_\oplus$ (they give the projection on the plane $K_\odot|K_\oplus$)[3] with

[3] When $l = 3$, one simply uses the three-point interpolation in the plane $K_\odot|K_\oplus$ to reconstruct this projection. Clearly, the interpolation is exact for affine functions. In general, the reconstruction (3), which is exact for affine functions, is based upon the 2D identity given in [3] and [1, Appendix].

$$\overrightarrow{\nabla}_{D}\, w^{\overline{\mathfrak{T}}} = \frac{1}{6\,Vol(D)} \sum_{i=1}^{l} \left\{ \frac{\langle\, \overrightarrow{x_{K_\odot} x_{K_\oplus}}\,,\ \overrightarrow{x_{K_\odot|K_\oplus} x_{K_i^*|K_{i+1}^*}}\,,\ \overrightarrow{x_{K_i^*} x_{K_{i+1}^*}}\ \rangle}{\overrightarrow{x_{K_\odot} x_{K_\oplus}} \cdot \overrightarrow{n}_{K_\odot,K_\oplus}} (w_{K_\oplus} - w_{K_\odot})\, \overrightarrow{n}_{K_\odot,K_\oplus} \right.$$
$$\left. + 2(w_{K_{i+1}^*} - w_{K_i^*}) \left[\overrightarrow{x_{K_\odot} x_{K_\oplus}} \times \overrightarrow{x_{K_\odot|K_\oplus} x_{K_i^*|K_{i+1}^*}} \right] \right\}. \qquad (4)$$

- Pick $\left[\!\!\left[w^{\mathfrak{T}}, v^{\mathfrak{T}} \right]\!\!\right] := \frac{1}{3} \sum_{K \in \mathfrak{M}^o} Vol(K)\ w_K v_K + \frac{2}{3} \sum_{K^* \in \mathfrak{M}^*} Vol(K^*)\ w_{K^*} v_{K^*}$ for scalar product on $\mathbb{R}^{\mathfrak{T}}$ and $\left\{\!\!\left\{ \overrightarrow{\mathscr{F}}^{\mathfrak{T}}, \overrightarrow{\mathscr{G}}^{\mathfrak{T}} \right\}\!\!\right\} := \sum_{D \in \mathfrak{D}} Vol(D)\, \overrightarrow{\mathscr{F}_D} \cdot \overrightarrow{\mathscr{G}_D}$ for scalar product on $(\mathbb{R}^3)^{\mathfrak{D}}$.

And now, one can mimic the identity $-\int_{\Omega} (\operatorname{div} \overrightarrow{\mathscr{F}})\, w = \int_{\Omega} \overrightarrow{\mathscr{F}} \cdot \overrightarrow{\nabla} w$ for $w|_{\partial\Omega} = 0$:

Proposition 1 (the *discrete duality* property; see [1, 3], see also [17]). *For all* $\overrightarrow{\mathscr{F}}^{\mathfrak{T}} \in (\mathbb{R}^3)^{\mathfrak{D}}$ *and all* $w^{\overline{\mathfrak{T}}} \in \mathbb{R}^{\overline{\mathfrak{T}}}$ *with* $w^{\partial\mathfrak{T}} = 0$, $\left[\!\!\left[-\operatorname{div}^{\mathfrak{T}} \overrightarrow{\mathscr{F}}^{\mathfrak{T}}, w^{\mathfrak{T}} \right]\!\!\right] = \left\{\!\!\left\{ \overrightarrow{\mathscr{F}}^{\mathfrak{T}}, \overrightarrow{\nabla}^{\mathfrak{T}} w^{\overline{\mathfrak{T}}} \right\}\!\!\right\}$.

The scheme. In this benchmark, one approximates the linear diffusion problem $-\operatorname{div}[\mathbf{A}(\cdot)\overrightarrow{\nabla} u] = f(\cdot)$ with Dirichlet boundary condition $u|_{\partial\Omega} = \bar{u}(\cdot)$, $\mathbf{A}(\cdot)$ being a heterogeneous anisotropic diffusion tensor and $f(\cdot)$ being a source term. Let $\mathbb{P}^{\mathfrak{T}}$ denote the projection on the DDFV mesh $\mathfrak{T}$ (i.e. the components of $\mathbb{P}^{\mathfrak{T}} f$ are the mean values of $f \in L^1(\Omega)$ per primal and per dual volumes); $\mathbb{P}^{\partial\mathfrak{T}}$ is the projection on the boundary part of the mesh. Let $\overrightarrow{\mathbb{P}}^{\mathfrak{T}}$ denote the projection on the diamond mesh $\mathfrak{D}$. For general data, the heterogeneity of the matrix $\mathbf{A}(\cdot)$ is taken into account by using the diamond-wise projection $\mathbf{A}^{\mathfrak{T}} := \overrightarrow{\mathbb{P}}^{\mathfrak{T}}\mathbf{A}(\cdot)$; similarly, we use $f^{\mathfrak{T}} = \mathbb{P}^{\mathfrak{T}} f(\cdot)$ as the discrete source term. The boundary condition is given by the projection $\mathbb{P}^{\partial\mathfrak{T}}\bar{u}(\cdot)$.

For a fully practical discretization of $\mathbf{A}(\cdot)$ and $f(\cdot)$ (which are continuous in all the tests we perform), for every element (recall that diamonds, primal volumes and dual volumes of a DDFV mesh are unions of elements, see Fig. 1) we take the mean value of the four vertices of the element. The point values of the exact solution u_e in the centers of the boundary volumes are used as discrete boundary conditions.

Given a DDFV mesh $\mathfrak{T}$ of Ω the method writes as:

$$\text{Find } u^{\mathfrak{T}} \text{ s.t. } -\operatorname{div}^{\mathfrak{T}}\left[\, \mathbf{A}^{\mathfrak{T}}\, \overrightarrow{\nabla}^{\mathfrak{T}} u^{\overline{\mathfrak{T}}} \right] = f^{\mathfrak{T}} \text{ with } u^{\overline{\mathfrak{T}}} = (u^{\mathfrak{T}}; \mathbb{P}^{\partial\mathfrak{T}}\bar{u}).$$

Convergence. From the discrete duality (Prop. 1) which is a cornerstone of DDFV schemes, and from consistency properties of the projection, gradient and divergence operators (see [2]; cf. [5] for analogous properties in 2D) one easily derives that the scheme is well posed for $l \leq 4$.[4] Given a family $(\mathfrak{T}_h)_h$ of CeVe-DDFV meshes, the associated discrete solutions $u^{\overline{\mathfrak{T}}_h}$ enjoy a uniform discrete H^1 estimate, and they converge to the exact solution u as the size h of the mesh tends to zero. Convergence analysis requires mild proportionality assumptions on the meshes $\mathfrak{T}_h$ in use, see [2].

2 Numerical results

In this section, we describe the results obtained on Tests 1–4 of the benchmark. Notice that, while the method converges for merely L^∞ uniformly elliptic tensor $\mathbf{A}(\cdot)$, it is not designed for a smart handling of a *piecewise* continuous $\mathbf{A}(\cdot)$[5]. Therefore, we skip Test 5 that involves piecewise constant $\mathbf{A}(\cdot)$. We refer to Coudière, Pierre, Rousseau and Turpault [12] and to Hermeline [17] for 3D CeVe-DDFV constructions efficiently taking into account discontinuities of $\mathbf{A}(\cdot)$.

Choice of the cell and face points. We pick for x_K, the isobarycenter of the cell K, and for $x_{K|L}$, the isobarycenter of the face $K|L$.

Measure of errors and convergence orders. To put the discrete and the exact solutions "at the same level", we use the projection $\mathbb{P}^{\mathfrak{T}}u_e$ of the exact solution and the associated discrete gradient reconstruction $\overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e$, where $\mathbb{P}^{\overline{\mathfrak{T}}}\cdot \;=\; \big(\mathbb{P}^{\mathfrak{T}}\cdot\,;\,\mathbb{P}^{\partial\mathfrak{T}}\cdot\big)$. The L^2 norms of the errors $e^{\mathfrak{T}} := u^{\mathfrak{T}} - \mathbb{P}^{\mathfrak{T}}u_e$ and $\overrightarrow{\nabla}^{\mathfrak{T}}e^{\overline{\mathfrak{T}}} := \overrightarrow{\nabla}^{\mathfrak{T}}u^{\overline{\mathfrak{T}}} - \overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e$ are measured in terms of the scalar products $\left\llbracket \cdot\,,\cdot \right\rrbracket$ on $\mathbb{R}^{\mathfrak{T}}$, $\left\{\!\!\left\{ \mathbf{A}^{\mathfrak{T}}\cdot\,,\cdot \right\}\!\!\right\}$ and $\left\{\!\!\left\{ \cdot\,,\cdot \right\}\!\!\right\}$ on $(\mathbb{R}^3)^{\mathfrak{D}}$: the relative error indicators *erl2* and *ener* , *ergrad* we use are defined, respectively, as $\left(\frac{\llbracket e^{\mathfrak{T}}, e^{\mathfrak{T}} \rrbracket}{\llbracket \mathbb{P}^{\mathfrak{T}}u_e, \mathbb{P}^{\mathfrak{T}}u_e \rrbracket}\right)^{1/2}$ and as $\left(\frac{\{\!\!\{ \mathbf{A}^{\mathfrak{T}}\overrightarrow{\nabla}^{\mathfrak{T}}e^{\overline{\mathfrak{T}}}, \overrightarrow{\nabla}^{\mathfrak{T}}e^{\overline{\mathfrak{T}}} \}\!\!\}}{\{\!\!\{ \mathbf{A}^{\mathfrak{T}}\overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e, \overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e \}\!\!\}}\right)^{1/2}$, $\left(\frac{\{\!\!\{ \overrightarrow{\nabla}^{\mathfrak{T}}e^{\overline{\mathfrak{T}}}, \overrightarrow{\nabla}^{\mathfrak{T}}e^{\overline{\mathfrak{T}}} \}\!\!\}}{\{\!\!\{ \overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e, \overrightarrow{\nabla}^{\mathfrak{T}}\mathbb{P}^{\overline{\mathfrak{T}}}u_e \}\!\!\}}\right)^{1/2}$.

[4] The restriction on the number l of face vertices is only needed for justifying a discrete Poincaré inequality; yet this property is immaterial, e.g., for the associated evolution problem. In practice, in the below tests values $l = 3, 4, 6$ were used, and no particular problem for $l = 6$ is reported.

[5] In 2D, a scheme called m-DDFV, specifically designed to handle *discontinuous* diffusion tensors, was designed by Boyer and Hubert in [6]. There is a clear difference in convergence orders for the basic DDFV version [5] and the m-DDFV version [6] (see the 2D benchmark paper [7]).

• **Test 1 Mild anisotropy,** $u_e(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi\left(y + \frac{1}{2}\right)\right) \sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	2187	21287	0.706E-02	0.706E-02	1.992	1.992	0.178E+01
2	4301	44813	0.706E-02	0.706E-02	1.997	1.996	0.179E+01
3	8584	94088	0.278E-02	0.278E-02	1.993	1.993	0.179E+01
4	17102	195074	0.792E-03	0.792E-03	1.997	1.997	0.179E+01
5	34343	405077	0.140E-02	0.140E-02	1.999	1.999	0.180E+01
6	69160	838856	0.140E-02	0.140E-02	1.999	1.999	0.180E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	2187	0.539E-02	-	0.654E-01	-	0.649E-01	-
2	4301	0.331E-02	2.165	0.488E-01	1.297	0.491E-01	1.239
3	8584	0.206E-02	2.069	0.381E-01	1.077	0.383E-01	1.079
4	17102	0.135E-02	1.841	0.301E-01	1.018	0.302E-01	1.026
5	34343	0.846E-03	1.998	0.240E-01	0.973	0.242E-01	0.955
6	69160	0.539E-03	1.934	0.190E-01	1.012	0.191E-01	1.008

• **Test 1 Mild anisotropy,** $u_e(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi\left(y + \frac{1}{2}\right)\right) \sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	87	1433	0.667E-01	0.667E-01	1.904	1.904	0.159E+01
2	235	4393	0.432E-02	0.432E-02	1.997	1.997	0.172E+01
3	527	10777	0.280E-01	0.280E-01	1.990	1.990	0.176E+01
4	1013	21793	0.108E-02	0.108E-02	2.003	1.995	0.177E+01
5	1776	40998	0.113E-01	0.113E-01	2.000	1.996	0.178E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	87	0.484E-01	-	0.204E+00	-	0.374E+00	-
2	235	0.388E-01	0.666	0.173E+00	0.496	0.277E+01	-6.049
3	527	0.231E-01	1.925	0.118E+00	1.402	0.838E+00	4.445
4	1013	0.167E-01	1.484	0.940E-01	1.060	0.299E+01	-5.843
5	1776	0.117E-01	1.937	0.818E-01	0.742	0.291E+01	0.147

- **Test 1 Mild anisotropy,** $u_e(x, y, z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	855	13819	2.28E-02	2.28E-02	1.989	1.989	1.730
2	7471	138691	2.52E-03	2.52E-03	1.994	1.994	1.778
3	62559	1237459	1.99E-03	1.99E-03	1.999	1.999	1.794
4	512191	10443763	3.82E-04	3.82E-04	2.000	2.000	1.797

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	855	0.501E-01	-	0.484E+00	-	0.558E+00	-
2	7471	0.156E-01	1.611	0.209E+00	1.160	0.159E+00	1.735
3	62559	0.392E-02	1.954	0.677E-01	1.594	0.395E-01	1.970
4	512191	0.101E-02	1.936	0.223E-01	1.585	0.109E-01	1.835

- **Test 1 Mild anisotropy,** $u_e(x, y, z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	59	703	0.341E-01	0.341E-01	1.966	1.966	0.167E+01
2	599	9835	0.856E-02	0.856E-02	1.991	1.991	0.178E+01
3	5423	101539	0.214E-02	0.214E-02	1.998	1.998	0.179E+01
4	46175	917395	0.535E-03	0.535E-03	1.999	1.999	0.180E+01
5	381119	7788403	0.134E-03	0.134E-03	2.000	2.000	0.180E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	59	0.396E-01	-	0.136E+00	-	0.116E+00	-
2	599	0.149E-01	1.266	0.928E-01	0.499	0.818E-01	0.449
3	5423	0.400E-02	1.792	0.497E-01	0.849	0.448E-01	0.820
4	46175	0.103E-02	1.905	0.256E-01	0.931	0.232E-01	0.920
5	381119	0.259E-03	1.954	0.130E-01	0.965	0.118E-01	0.961

- **Test 2 Heterogeneous anisotropy,** $\min = -0.862, \ \max = 1.0487$ $u_e(x, y, z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)\sin(2\pi z)$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3010	64158	-.856E+00	-.856E+00	1.044	1.044	0.170E+01
2	24020	555528	-.859E+00	-.859E+00	1.047	1.047	0.171E+01
3	81030	1924098	-.861E+00	-.861E+00	1.049	1.049	0.171E+01
4	192040	4619868	-.862E+00	-.862E+00	1.049	1.049	0.171E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3010	0.467E-01	-	0.711E-01	-	0.785E-01	-
2	24020	0.123E-01	1.931	0.224E-01	1.667	0.328E-01	1.262
3	81030	0.554E-02	1.960	0.116E-01	1.634	0.190E-01	1.348
4	192040	0.314E-02	1.973	0.728E-02	1.607	0.127E-01	1.389

• **Test 3 Flow with strong anisotropy on random meshes,** $\min = 0$, $\max = 1$**,** $u_e(x, y, z) = \sin(\pi x) \sin(\pi y) \sin(\pi z)$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	91	1063	-.202E+01	-.978E+00	1.969	0.931	0.392E+01
2	855	13819	-.116E+01	-.994E+00	1.206	0.982	0.363E+01
3	7471	138691	-.105E+01	-.995E+00	1.029	0.991	0.362E+01
4	62559	1237459	-.101E+01	-.998E+00	1.014	0.998	0.360E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	91	0.713E+00	-	0.716E+00	-	0.439E+00	-
2	855	0.152E+00	2.068	0.199E+00	1.712	0.130E+00	1.633
3	7471	0.384E-01	1.906	0.854E-01	1.174	0.417E-01	1.568
4	62559	0.119E-01	1.656	0.542E-01	0.640	0.183E-01	1.165

• **Test 4 Flow around a well,** $\min = 0$, $\max = 5.415$**, Well meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1482	23942	-.438E-01	-.438E-01	5.415	5.415	0.162E+04
2	3960	70872	-.239E-01	-.239E-01	5.415	5.415	0.162E+04
3	9229	173951	-.132E-01	-.132E-01	5.415	5.415	0.162E+04
4	21156	412240	-.661E-02	-.661E-02	5.415	5.415	0.162E+04
5	44420	882520	-.411E-02	-.411E-02	5.415	5.415	0.162E+04
6	82335	1654893	-.281E-02	-.281E-02	5.415	5.415	0.162E+04
7	145079	2937937	-.198E-02	-.198E-02	5.415	5.415	0.162E+04

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1482	0.564E-02	-	0.473E-01	-	0.817E-01	-
2	3960	0.218E-02	2.897	0.205E-01	2.556	0.487E-01	1.578
3	9229	0.964E-03	2.898	0.108E-01	2.255	0.296E-01	1.770
4	21156	0.645E-03	1.454	0.748E-02	1.344	0.205E-01	1.320
5	44420	0.427E-03	1.664	0.546E-02	1.274	0.144E-01	1.443
6	82335	0.291E-03	1.864	0.396E-02	1.560	0.108E-01	1.391
7	145079	0.205E-03	1.848	0.337E-02	0.858	0.794E-02	1.624

3 Comments

Let us summarize the observations; footnotes provide comments of theoretical order.

Choice of the solvers. The following results have been performed either with the direct solvers given by the UMFPACK library, or with the BiCGStab algorithm with ILU(0) preconditionning delivered in the HSL library. A comparison between ISTL-CG with ILU(0) preconditionning and PETSC-CG with ILU(2) preconditionning shows that, whenever ISTL-CG/ILU(0) algorithm converges, much less CPU time and much less memory is used than for the PETSC-CG/ILU(2) algorithm.

Convergence orders observed[6]. Even if the orders present serious oscillations for some cases (e.g., in Test 3 and in Test 1 on Voronoï meshes), orders slightly below h^2 (superconvergence) for the solution in the L^2 norm are observed quite systematically. One exception is Test 4, where an order intermediate between $h^{3/2}$ and h^2 seems to appear; this may be related to the presence of a singularity in the well center.

Regarding the gradient norm, convergence orders close to h are seen in Test 1 on tetrahedral, Voronoï, checkerboard meshes. On Kershaw meshes in Test 1 and prism meshes of Test 2, more structured though distorted, an $h^{3/2}$ convergence order can be observed. For random meshes of Test 3, orders degrade quickly but the numerical evidence (four meshes only) seems insufficient. The well meshes of Test 4 appear as rather structured but having a singularity; the effect of singularity grows as the mesh becomes finer, and the convergence order falls from h^2 to $h^{3/2}$ and then to h. Yet from Tests 3 and 4 with stronger anisotropy of $\mathbf{A}(\cdot)$, it becomes clear that more adequate norm for measuring gradient convergence is the energy norm. In Test 4 we observe an accurate $h^{3/2}$ convergence and in Test 3, an order $h^{3/2}$ can be conjectured.

Violation (and fulfillment) of the maximum principle[7]. We observe that violation of discrete maximum principle does not occur systematically (or if it occurs, it is of imperceptible magnitude, even on coarse meshes). No overshoot/undershoot is reported on Kershaw, checkerboard and prism meshes for Test 1, nor on the well meshes of Test 4; a very slight overshoot can be seen in Test 1 on tetrahedral meshes. On the contrary, random meshes of Test 3, and also the finest ones among the Voronoï meshes of Test 1, exhibit a perceptible violation of the maximum principle which is nonetheless reduced as the mesh size diminishes[8]. Difficulties on these two

[6] For regular enough $\mathbf{A}(\cdot)$ and u_e, order h can be proved for both solution and its gradient in L^2.

[7] In principle, DDFV methods are not designed in order to respect the discrete maximum principle; and the convergence analysis exploits rather the variational structure, well preserved by the method (this is one of the benefits from the discrete duality of Prop. 1). Let us point out that for isotropic problems on primal meshes satisfying the orthogonality condition (e.g., Delaunay tetrahedral meshes with the choice of circumcenters for the cell centers x_K - note that x_K may fall out of K), the discrete maximum principle is easily shown for the CeVe-DDFV scheme under study ([4]).

[8] In theory, one can prove convergence in L^q for $q < 6$; nothing guarantees convergence in L^∞.

kinds of meshes can be explained by their poor shape regularity (e.g., fine Voronoï meshes in Test 1 present a dramatic contrast of size between neighbor cells).

Influence of the mesh type and quality on convergence orders[9]. Among the different mesh properties that could influence the numerical behavior, restrictions on l appear as immaterial (the best convergence orders are achieved for prism meshes of Test 2 having up to $l = 6$ face vertices). While conformity is not needed for the method, non-conformal meshes bring more distorted cells and diamonds. We have seen that bad shape conditioning may induce violation of the maximum principle. In Test 1, presence of neighbor cells with considerable contrast in size (for Voronoï meshes and for non-conformal checkerboard meshes) degrades convergence orders for the gradient, in contrast to rather gradually distorted Kershaw and prism meshes.

References

1. B. Andreianov, M. Bendahmane, F. Hubert and S. Krell. On 3D DDFV discretization of gradient and divergence operators. I. Meshing, operators and discrete duality. Preprint HAL (2011), http://hal.archives-ouvertes.fr/hal-00355212.
2. B. Andreianov, M. Bendahmane and F. Hubert. On 3D DDFV discretization of gradient and divergence operators. II. Discrete functional analysis tools and applications to degenerate parabolic problems. Preprint HAL (2011), http://hal.archives-ouvertes.fr/hal-00567342.
3. B. Andreianov, M. Bendahmane, and K. Karlsen. A gradient reconstruction formula for finite-volume schemes and discrete duality. In R. Eymard and J.-M. Hérard, editors, *Finite Volume For Complex Applications, Problems And Perspectives. 5th International Conference*, (2008),161–168. London (UK) Wiley.
4. B. Andreianov, M. Bendahmane, and K.H. Karlsen. Discrete duality finite volume schemes for doubly nonlinear degenerate hyperbolic-parabolic equations. *J. Hyperbolic Diff. Equ.* (2010), **7**(1):1–67.
5. B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general 2D-meshes. *Numer. Methods PDE*, (2007), **23**(1):145–195.
6. F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Num. Anal.*, (2008), **46**(6):3032–3070.
7. F. Boyer and F. Hubert. Benchmark on anisotropic problems, the ddfv discrete duality finite volumes and m-ddfv schemes. In R. Eymard and J.-M. Hérard, editors, *Finite Volume For Complex Applications, Problems And Perspectives. 5th International Conference*, (2008), 735–750. London (UK) Wiley.
8. Y. Coudière, J.-P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. *M2AN Math. Modelling Num. Anal.*, (1999), **33**(3):493–516.
9. Y. Coudière and C. Pierre. Benchmark 3D: CeVe-DDFV, a discrete duality scheme with cell/vertex unknowns. *This volume*, (2011).
10. Y. Coudière and F. Hubert. A 3D discrete duality finite volume method for nonlinear elliptic equation. preprint (2009), http://hal.archives-ouvertes.fr/docs/00/45/68/37/PDF/ddfv3d.pdf

[9] Recall that conformity of meshes is not required by the method; and the construction allows for unrestricted number l of face vertices. Yet it is a well-known difficulty for the analysis of the scheme that the discrete Poincaré inequality cannot be proved for $l > 4$, see [2, 17].

11. Y. Coudière, F. Hubert and G. Manzini. Benchmark 3D: CeVeFE-DDFV, a discrete duality scheme with cell/vertex/face+edge unknowns. *This volume*, (2011).
12. Y. Coudière, C. Pierre, O. Rousseau, and R. Turpault. A 2d/3d discrete duality finite volume scheme. Application to ecg simulation. *Int. Journal on Finite Volumes*, (2009), **6**(1).
13. K. Domelevo and P. Omnès. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, (2005), **39**(6):1203–1249.
14. R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In R. Eymard and J.-M. Hérard, editors, *Finite Volume For Complex Applications, Problems And Perspectives. 5th International Conference*, (2008), 659–692. London (UK) Wiley.
15. F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, (2000), **160**(2):481–499.
16. F. Hermeline. Approximation of 2-d and 3-d diffusion operators with variable full tensor coefficients on arbitrary meshes. *Comput. Methods Appl. Mech. Engrg.*, (2007), **196**(21-24): 2497–2526.
17. F. Hermeline. A finite volume method for approximating 3d diffusion operators on general meshes. *Journal of computational Physics*, (2009), **228**(16):5763–5786.
18. S. Krell. *Schémas Volumes Finis en mécanique des fluides complexes*. Ph.D. Thesis, Univ. de Provence, Marseilles, (2010), http://tel.archives-ouvertes.fr/tel-00524509.
19. S. Krell. Stabilized DDFV schemes for Stokes problem with variable viscosity on general 2D meshes. *Num. Meth. PDEs*, (2010), available on-line: http://dx.doi.org/10.1002/num.20603.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: Symmetric Weighted Interior Penalty Discontinuous Galerkin Scheme

Peter Bastian

1 Weighted Interior Penalty Discontinuous Galerkin Schemes

Consider the stationary diffusion equation with Dirichlet boundary conditions

$$-\nabla \cdot (K \nabla u) = f \qquad \text{in } \Omega, \tag{1a}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{1b}$$

with Ω a domain in $\mathbb{R}^n$, $n = 1, 2, 3$, and K a uniformly symmetric positive definite permeability tensor. The weak formulation of (1) consist of finding $u \in u_g + V$, $V = H_0^1(\Omega)$, u_g an extension of the Dirichlet data such that

$$(K \nabla u, \nabla v)_{0,\Omega} = (f, v)_{0,\Omega}$$

where $(.,.)_{0,\omega}$ denotes the L^2-scalar product on a domain ω.

Discontinuous Galerkin (DG) methods are a class of numerical schemes that has been studied extensively in the last two decades, see [8]. We use the weighted interior penalty discontinuous Galerkin (WIPG) schemes introduced in [6].

Let $\{\mathscr{T}_h\}_{h>0}$ denote a family of triangulations of the domain Ω. An element of the triangulation is denoted by T, h_T is its diameter, $|T|$ its volume and n_T its unit outer normal vector. F is called an "interior face" independent of the dimension if there are two elements $T^-(F), T^+(F) \in \mathscr{T}_h$ with $T^-(F) \cap T^+(F) = F$ and F has nonzero measure. All interior faces are collected in the set $\mathscr{F}_h^i$. The intersection of $T \in \mathscr{T}_h$ with the boundary $\partial\Omega$ of non-zero measure is called a boundary face. All boundary faces are collected in the set $\mathscr{F}_h^{\partial\Omega}$ and we set $\mathscr{F} = \mathscr{F}_h^i \cup \mathscr{F}_h^{\partial\Omega}$. The diameter of a face is denoted by h_F and its volume by $|F|$. With each $F \in \mathscr{F}$ we

Peter Bastian
Interdisciplinary Center for Scientific Computing, University of Heidelberg, Im Neuenheimer Feld 368, D-69120 Heidelberg, e-mail: peter.bastian@iwr.uni-heidelberg.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_92, © Springer-Verlag Berlin Heidelberg 2011

associate a unit normal vector n_F (depending on position if F is curved) oriented from $T^-(F)$ to $T^+(F)$ for an interior face and coinciding with the exterior unit normal for a boundary face.

The DG approximation space is

$$V_h = \{v \in L^2(\Omega) \, : \, \forall T \in \mathscr{T}_h, v|_T \in \mathscr{P}\}$$

where $\mathscr{P}$ is either $\mathbb{P}_k=\{p{:}p = \sum_{\|\alpha\|_1 \leq k} c_\alpha x^\alpha\}$ or $\mathbb{Q}_k = \{p : p = \sum_{\|\alpha\|_\infty \leq k} c_\alpha x^\alpha\}$ in the standard multiindex notation. On an interior face F a function $v \in V_h$ is two-valued and its values v^- and v^+ are the restrictions from $T^-(F)$ and $T^+(F)$, respectively. For $F \in \mathscr{F}_h^i$ and $v \in V_h$ we introduce the jump and the weighted average

$$[\![v]\!]_F = v^- - v^+, \qquad \{v\}_\omega = \omega^- v^- + \omega^+ v^+,$$

with the weights satisfying $\omega^- + \omega^+ = 1, \omega^-, \omega^+ \geq 0$.

In the WIPG schemes the discrete solution $u_h \in V_h$ satisfies the variational equation

$$a_h(u_h, v) = l_h(v) \qquad \forall v \in V_h$$

with bilinear and linear forms defined as

$$\begin{aligned} a_h(u,v) = & \sum_{T \in \mathscr{T}_h} (K\nabla u, \nabla v)_{0,T} \\ & - \sum_{F \in \mathscr{F}_h^i} \left[(\{n_F^t K \nabla u\}_\omega, [\![v]\!])_{0,F} - \theta(\{n_F^t K \nabla v\}_\omega, [\![u]\!])_{0,F} - \gamma_F([\![u]\!], [\![v]\!])_{0,F} \right] \\ & - \sum_{F \in \mathscr{F}_h^{\partial\Omega}} \left[(n_F^t K \nabla u, v)_{0,F} - \theta(n_F^t K \nabla v, u)_{0,F} - \gamma_F(u,v)_{0,F} \right], \\ l_h(u,v) = & \sum_{T \in \mathscr{T}_h} (f, v)_{0,T} + \sum_{F \in \mathscr{F}_h^{\partial\Omega}} \left[\theta(n_F^t K \nabla v, g)_{0,F} + \gamma_F(g,v)_{0,F} \right]. \end{aligned}$$

For $\theta = -1$ we obtain the symmetric weighted interior penalty Galerkin method (SWIPG) used below for all tests.

The weights $\omega^\pm$ are defined with respect to the permeability as

$$\omega^- = \frac{\delta_{Kn}^+}{\delta_{Kn}^- + \delta_{Kn}^+}, \qquad \omega^+ = \frac{\delta_{Kn}^-}{\delta_{Kn}^- + \delta_{Kn}^+},$$

with $\delta_{Kn}^\pm = n_F^t K^\pm n_F$ for $F \in \mathscr{F}_h^i$ and $\delta_{Kn} = n_F^t K n_F$ for $F \in \mathscr{F}_h^{\partial\Omega}$.

The choice of the interior penalty parameter γ_F is crucial, as the scheme should be as independent of the problem and mesh parameters as possible. We use the following definition of the penalty parameter:

$$\forall F \in \mathscr{F}_h^i, \qquad \gamma_F = \alpha \, \frac{2\delta_{Kn}^- \delta_{Kn}^+}{\delta_{Kn}^- + \delta_{Kn}^+} \, k(k+n-1) \, \frac{|F|}{\min(|T^-(F)|, |T^+(F)|)}, \tag{2a}$$

$$\forall F \in \mathscr{F}_h^{\partial\Omega}, \quad \gamma_F = \alpha \, \delta_{Kn} \, k(k+n-1) \, \frac{|F|}{|T^-(F)|}. \tag{2b}$$

with α a user-defined parameter. This choice is a combination of different papers: The harmonic average of "normal" permeabilities was introduced and analyzed in [6], the dependence on the polynomial degree was analyzed in [5] and the choice of the h-dependence is taken from [7]. The parameter α was chosen as follows:

Test	1	1	1	3	3	3	3	4	5
Mesh	tetra	kershaw	checkerboard	rand	rand	rand	rand	well	locraf
k	1-4	1-4	1-4	1	2	3	4	1-4	1-4
α	3.0	2.5	1.0	1000	2000	5000	10000	1000	0.7

Unfortunately, the choice of α is heuristic. It should be subject of future research to find a formula (2) that better takes into account the element shape as it was done in [5] for tetrahedral elements.

2 Numerical results

The L^2, H^1 and energy error are computed by numerical integration of order 12.

• Test 1, Mild anisotropy, Tetrahedral meshes

Table 1, $\mathbb{P}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	8012	150576	5.32E-02	2.03E-02	1.965	1.989	1.794
2	15592	297376	1.31E-02	6.84E-03	1.974	1.989	1.797
3	30844	593648	1.71E-02	9.13E-03	1.983	1.994	1.797
4	61064	1184192	1.05E-02	5.52E-03	1.992	1.997	1.797
5	121920	2379936	7.19E-03	1.49E-03	1.994	1.997	1.798
6	244208	4791872	3.69E-03	1.83E-03	1.994	1.997	1.798

Table 1, $\mathbb{P}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	20030	941100	2.11E-02	2.03E-02	1.989	1.989	1.799
2	38980	1858600	6.97E-03	6.84E-03	1.989	1.989	1.799
3	77110	3710300	9.19E-03	9.13E-03	1.993	1.994	1.799
4	152660	7401200	5.56E-03	5.52E-03	1.997	1.997	1.799
5	304800	14874600	1.51E-03	1.49E-03	1.997	1.997	1.799
6	610520	29949200	1.84E-03	1.83E-03	1.997	1.997	1.798

Table 1, $\mathbb{P}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	40060	3764400	2.04E-02	2.03E-02	1.989	1.989	1.798
2	77960	7434400	6.83E-03	6.84E-03	1.989	1.989	1.798
3	154220	14841200	9.14E-03	9.13E-03	1.993	1.994	1.798
4	305320	29604800	5.52E-03	5.52E-03	1.997	1.997	1.798
5	609600	59498400	1.49E-03	1.49E-03	1.997	1.997	1.798
6	1221040	119796800	1.83E-03	1.83E-03	1.997	1.997	1.798

Table 1, $\mathbb{P}_4$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	70105	11528475	2.03E-02	2.03E-02	1.989	1.989	1.798
2	136430	22767850	6.84E-03	6.84E-03	1.989	1.989	1.798
3	269885	45451175	9.13E-03	9.13E-03	1.994	1.994	1.798
4	534310	90664700	5.52E-03	5.52E-03	1.997	1.997	1.798
5	1066800	182213850	1.49E-03	1.49E-03	1.997	1.997	1.798
6	2136820	366877700	1.83E-03	1.83E-03	1.997	1.997	1.798

Table 2, $\mathbb{P}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	8012	1.11E-02	—	2.28E-01	—	2.23E-01	—
2	15592	7.02E-03	2.084	1.82E-01	1.015	1.79E-01	1.004
3	30844	4.52E-03	1.934	1.46E-01	0.961	1.43E-01	0.988
4	61064	2.91E-03	1.931	1.16E-01	1.016	1.13E-01	1.022
5	121920	1.87E-03	1.925	9.23E-02	0.993	9.03E-02	0.979
6	244208	1.16E-03	2.068	7.28E-02	1.021	7.11E-02	1.034

Table 2, $\mathbb{P}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	20030	8.22E-04	—	2.40E-02	—	2.32E-02	—
2	38980	4.19E-04	3.034	1.54E-02	1.986	1.48E-02	2.033
3	77110	2.08E-04	3.079	9.51E-03	2.123	9.17E-03	2.104
4	152660	1.05E-04	3.016	6.05E-03	1.985	5.84E-03	1.980
5	304800	5.34E-05	2.925	3.85E-03	1.962	3.72E-03	1.960
6	610520	2.66E-05	3.015	2.42E-03	2.015	2.33E-03	2.022

Table 2, $\mathbb{P}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	40060	4.46E-05	—	1.71E-03	—	1.66E-03	—
2	77960	1.78E-05	4.152	8.58E-04	3.118	8.28E-04	3.125
3	154220	7.27E-06	3.928	4.38E-04	2.957	4.21E-04	2.969
4	305320	2.84E-06	4.135	2.17E-04	3.076	2.09E-04	3.074
5	609600	1.14E-06	3.969	1.09E-04	3.004	1.05E-04	3.006
6	1221040	4.43E-07	4.067	5.36E-05	3.052	5.15E-05	3.060

Table 2, $\mathbb{P}_4$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	70105	2.31E-06	—	1.03E-04	—	9.93E-05	—
2	136430	7.36E-07	5.154	4.19E-05	4.068	3.97E-05	4.134
3	269885	2.30E-07	5.118	1.62E-05	4.186	1.54E-05	4.161
4	534310	7.00E-08	5.226	6.35E-06	4.110	6.06E-06	4.097
5	1066800	2.29E-08	4.850	2.59E-06	3.900	2.47E-06	3.896
6	2136820	7.08E-09	5.066	1.01E-06	4.047	9.64E-07	4.061

- **Test 1, Mild anisotropy, Kershaw meshes**

Table 1, $\mathbb{Q}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	4096	204800	9.58E-02	3.03E-02	1.850	1.958	1.771
2	32768	1736704	2.40E-02	1.06E-02	1.953	1.993	1.781
3	262144	14286848	5.39E-03	1.75E-03	1.987	1.997	1.786
4	2097152	115867648	1.71E-03	7.14E-04	1.997	1.999	1.791

Table 1, $\mathbb{Q}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	13824	2332800	3.12E-02	3.03E-02	1.944	1.958	1.796
2	110592	19782144	1.04E-02	1.06E-02	1.990	1.993	1.796
3	884736	162736128	1.71E-03	1.75E-03	1.997	1.997	1.798
4	7077888	1319804928	7.11E-04	7.14E-04	1.999	1.999	1.798

Table 1, $\mathbb{Q}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	32768	13107200	2.91E-02	3.03E-02	1.955	1.958	1.797
2	262144	111149056	1.05E-02	1.06E-02	1.992	1.993	1.798
3	2097152	914358272	1.75E-03	1.75E-03	1.997	1.997	1.798

Table 1, $\mathbb{Q}_4$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	64000	50000000	3.02E-02	3.03E-02	1.958	1.958	1.798
2	512000	424000000	1.06E-02	1.06E-02	1.993	1.993	1.798
3	4096000	-806967296	1.75E-03	1.75E-03	1.997	1.997	1.798

Table 2, $\mathbb{Q}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	4096	6.65E-02	—	5.79E-01	—	5.52E-01	—
2	32768	4.57E-02	0.541	4.26E-01	0.441	4.00E-01	0.466
3	262144	2.53E-02	0.856	2.76E-01	0.627	2.59E-01	0.624
4	2097152	1.07E-02	1.246	1.57E-01	0.813	1.51E-01	0.778

Table 2, $\mathbb{Q}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	13824	2.95E-02	—	2.45E-01	—	2.17E-01	—
2	110592	7.51E-03	1.975	1.04E-01	1.236	9.80E-02	1.145
3	884736	9.75E-04	2.945	3.26E-02	1.675	3.22E-02	1.606
4	7077888	7.54E-05	3.693	8.76E-03	1.895	8.77E-03	1.877

Table 2, $\mathbb{Q}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	32768	5.72E-03	—	6.34E-02	—	5.57E-02	—
2	262144	7.05E-04	3.020	1.58E-02	2.002	1.50E-02	1.895
3	2097152	2.91E-05	4.598	2.45E-03	2.689	2.42E-03	2.629

Table 2, $\mathbb{Q}_4$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	64000	1.61E-03	—	2.09E-02	—	1.82E-02	—
2	512000	5.64E-05	4.837	1.99E-03	3.389	1.89E-03	3.266
3	4096000	1.21E-06	5.540	1.51E-04	3.725	1.49E-04	3.664

- **Test 1, Mild anisotropy, Checkerboard meshes**

Table 1, $\mathbb{P}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	144	3648	2.35E-01	1.54E-01	1.784	1.846	1.550
2	1152	35328	6.71E-02	4.01E-02	1.931	1.960	1.667
3	9216	307200	1.47E-02	1.01E-02	1.985	1.990	1.751
4	73728	2555904	2.56E-03	2.54E-03	1.997	1.997	1.784
5	589824	20840448	6.36E-04	6.36E-04	1.999	1.999	1.795

Table 1, $\mathbb{P}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	360	22800	1.82E-01	1.54E-01	1.812	1.846	1.750
2	2880	220800	4.52E-02	4.01E-02	1.954	1.960	1.795
3	23040	1920000	1.05E-02	1.01E-02	1.989	1.990	1.799
4	184320	15974400	2.57E-03	2.54E-03	1.997	1.997	1.799
5	1474560	130252800	6.37E-04	6.36E-04	1.999	1.999	1.798

Table 1, $\mathbb{P}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	720	91200	1.61E-01	1.54E-01	1.839	1.846	1.783
2	5760	883200	4.02E-02	4.01E-02	1.960	1.960	1.798
3	46080	7680000	1.01E-02	1.01E-02	1.990	1.990	1.798
4	368640	63897600	2.54E-03	2.54E-03	1.997	1.997	1.798
5	2949120	521011200	6.36E-04	6.36E-04	1.999	1.999	1.798

Table 1, $\mathbb{P}_4$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1260	279300	1.55E-01	1.54E-01	1.845	1.846	1.797
2	10080	2704800	4.01E-02	4.01E-02	1.960	1.960	1.798
3	80640	23520000	1.01E-02	1.01E-02	1.990	1.990	1.798
4	645120	195686400	2.54E-03	2.54E-03	1.997	1.997	1.798
5	5160960	1595596800	6.36E-04	6.36E-04	1.999	1.999	1.798

Table 2, $\mathbb{P}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	144	9.89E-02	—	5.97E-01	—	5.71E-01	—
2	1152	3.13E-02	1.658	3.33E-01	0.842	3.35E-01	0.769
3	9216	9.17E-03	1.773	1.68E-01	0.984	1.68E-01	0.998
4	73728	2.51E-03	1.871	8.30E-02	1.022	8.22E-02	1.031
5	589824	6.47E-04	1.954	4.10E-02	1.015	4.05E-02	1.020

Table 2, $\mathbb{P}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	360	3.56E-02	—	2.70E-01	—	2.93E-01	—
2	2880	6.55E-03	2.442	8.25E-02	1.713	8.00E-02	1.874
3	23040	7.78E-04	3.075	2.10E-02	1.971	2.04E-02	1.970
4	184320	9.46E-05	3.039	5.26E-03	2.001	5.11E-03	1.998
5	1474560	1.17E-05	3.010	1.31E-03	2.001	1.28E-03	2.000

Table 2, $\mathbb{P}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	720	1.19E-02	—	1.01E-01	—	8.92E-02	—
2	5760	1.02E-03	3.544	1.54E-02	2.712	1.47E-02	2.602
3	46080	7.31E-05	3.803	2.08E-03	2.886	1.95E-03	2.912
4	368640	4.78E-06	3.933	2.67E-04	2.962	2.50E-04	2.968
5	2949120	3.04E-07	3.975	3.37E-05	2.986	3.15E-05	2.988

Table 2, $\mathbb{P}_4$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1260	2.72E-03	—	2.82E-02	—	3.04E-02	—
2	10080	1.19E-04	4.517	2.18E-03	3.693	2.07E-03	3.874
3	80640	4.29E-06	4.789	1.46E-04	3.903	1.37E-04	3.923
4	645120	1.44E-07	4.900	9.34E-06	3.964	8.68E-06	3.977
5	5160960	4.62E-09	4.958	5.90E-07	3.985	5.46E-07	3.992

- **Test 3, Flow on random meshes, Random meshes**

Table 1, $\mathbb{Q}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	512	22528	-4.34E-01	-7.59E-01	0.355	0.691	3.007
2	4096	204800	-8.45E-01	-9.39E-01	0.791	0.923	3.431
3	32768	1736704	-9.58E-01	-9.85E-01	0.946	0.982	3.565
4	262144	14286848	-9.90E-01	-9.96E-01	0.989	0.996	3.588

Table 1, $\mathbb{Q}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1728	256608	-7.50E-01	-7.59E-01	0.676	0.691	3.637
2	13824	2332800	-9.40E-01	-9.39E-01	0.924	0.923	3.569
3	110592	19782144	-9.85E-01	-9.85E-01	0.982	0.982	3.597
4	884736	162736128	-9.96E-01	-9.96E-01	0.996	0.996	3.596

Table 1, $\mathbb{Q}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	4096	1441792	-7.53E-01	-7.59E-01	0.684	0.691	3.655
2	32768	13107200	-9.38E-01	-9.39E-01	0.922	0.923	3.570
3	262144	111149056	-9.85E-01	-9.85E-01	0.982	0.982	3.597
4	2097152	914358272	-9.96E-01	-9.96E-01	0.996	0.996	3.596

Table 1, $\mathbb{Q}_4$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	8000	5500000	-7.59E-01	-7.59E-01	0.691	0.691	3.655
2	64000	50000000	-9.39E-01	-9.39E-01	0.923	0.923	3.570
3	512000	424000000	-9.85E-01	-9.85E-01	0.982	0.982	3.597

Table 2, $\mathbb{Q}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	512	3.04E-01	—	4.99E-01	—	5.00E-01	—
2	4096	8.38E-02	1.858	2.58E-01	0.952	2.53E-01	0.984
3	32768	2.16E-02	1.958	1.30E-01	0.988	1.25E-01	1.017
4	262144	5.77E-03	1.902	6.72E-02	0.953	6.30E-02	0.989

Table 2, $\mathbb{Q}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1728	4.41E-02	—	1.25E-01	—	1.14E-01	—
2	13824	6.01E-03	2.874	2.98E-02	2.066	2.83E-02	2.014
3	110592	8.26E-04	2.864	7.77E-03	1.941	7.27E-03	1.962
4	884736	1.74E-04	2.248	2.60E-03	1.578	2.16E-03	1.750

Table 2, $\mathbb{Q}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	4096	5.29E-03	—	2.00E-02	—	1.85E-02	—
2	32768	3.86E-04	3.776	2.60E-03	2.943	2.30E-03	3.002
3	262144	6.22E-05	2.634	6.36E-04	2.032	3.72E-04	2.629
4	2097152	3.93E-05	0.661	5.81E-04	0.130	2.65E-04	0.489

Table 2, $\mathbb{Q}_4$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	8000	5.44E-04	—	2.69E-03	—	2.26E-03	—
2	64000	4.30E-05	3.662	3.02E-04	3.159	1.76E-04	3.684
3	512000	2.15E-05	0.999	2.47E-04	0.291	9.24E-05	0.930

- **Test 4, Flow around a well, Well meshes**

Table 1, $\mathbb{Q}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	7120	356736	3.52E-01	4.14E-01	5.316	5.317	1686.482
2	17856	931328	2.23E-01	2.44E-01	5.328	5.328	1652.956
3	40128	2139904	1.46E-01	1.54E-01	5.329	5.329	1637.838
4	89760	4857216	1.13E-01	1.18E-01	5.330	5.330	1632.052
5	185680	10136320	8.72E-02	8.99E-02	5.339	5.339	1628.690
6	341064	18717760	7.06E-02	7.23E-02	5.345	5.345	1626.812
7	597432	32900416	5.55E-02	5.65E-02	5.361	5.361	1625.784

Table 1, $\mathbb{Q}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	24030	4063446	4.13E-01	4.14E-01	5.317	5.317	1625.412
2	60264	10608408	2.43E-01	2.44E-01	5.328	5.328	1623.883
3	135432	24374844	1.54E-01	1.54E-01	5.329	5.329	1623.603
4	302940	55326726	1.18E-01	1.18E-01	5.330	5.330	1623.529
5	626670	115459020	8.99E-02	8.99E-02	5.339	5.339	1623.506
6	1151091	213206985	7.23E-02	7.23E-02	5.345	5.345	1623.497
7	2016333	374756301	5.65E-02	5.65E-02	5.361	5.361	1623.491

Table 1, $\mathbb{Q}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	56960	22831104	4.15E-01	4.14E-01	5.317	5.317	1623.772
2	142848	59604992	2.44E-01	2.44E-01	5.328	5.328	1623.611
3	321024	136953856	1.54E-01	1.54E-01	5.329	5.329	1623.546
4	718080	310861824	1.18E-01	1.18E-01	5.330	5.330	1623.514
5	1485440	648724480	8.99E-02	8.99E-02	5.339	5.339	1623.500
6	2728512	1197936640	7.24E-02	7.23E-02	5.345	5.345	1623.493
7	4779456	2105626624	5.65E-02	5.65E-02	5.361	5.361	1623.489

Table 1, $\mathbb{Q}_4$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	111250	87093750	4.14E-01	4.14E-01	5.317	5.317	1623.709
2	279000	227375000	2.44E-01	2.44E-01	5.328	5.328	1623.607
3	627000	522437500	1.54E-01	1.54E-01	5.329	5.329	1623.546
4	1402500	1185843750	1.18E-01	1.18E-01	5.330	5.330	1623.514
5	2901250	2474687500	8.99E-02	8.99E-02	5.339	5.339	1623.500

Table 2, $\mathbb{Q}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	7120	6.54E-03	—	2.50E-01	—	2.48E-01	—
2	17856	2.99E-03	2.551	1.70E-01	1.254	1.70E-01	1.239
3	40128	1.47E-03	2.622	1.19E-01	1.318	1.19E-01	1.310
4	89760	9.71E-04	1.555	9.08E-02	1.010	9.09E-02	1.008
5	185680	6.04E-04	1.959	7.07E-02	1.033	7.08E-02	1.031
6	341064	3.68E-04	2.440	5.71E-02	1.056	5.72E-02	1.056
7	597432	2.72E-04	1.629	4.77E-02	0.962	4.78E-02	0.959

Table 2, $\mathbb{Q}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	24030	3.77E-04	—	3.95E-02	—	3.95E-02	—
2	60264	1.17E-04	3.813	1.61E-02	2.932	1.61E-02	2.931
3	135432	4.96E-05	3.180	7.47E-03	2.843	7.48E-03	2.838
4	302940	3.33E-05	1.487	4.40E-03	1.976	4.40E-03	1.978
5	626670	1.86E-05	2.412	2.64E-03	2.099	2.65E-03	2.098
6	1151091	9.19E-06	3.469	1.69E-03	2.197	1.70E-03	2.194
7	2016333	6.09E-06	2.198	1.17E-03	1.966	1.17E-03	1.964

Table 2, $\mathbb{Q}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	56960	4.74E-05	—	8.16E-03	—	8.15E-03	—
2	142848	1.01E-05	5.050	2.25E-03	4.211	2.24E-03	4.210
3	321024	3.23E-06	4.225	7.44E-04	4.094	7.44E-04	4.090
4	718080	1.63E-06	2.552	3.22E-04	3.122	3.22E-04	3.123
5	1485440	7.41E-07	3.242	1.53E-04	3.070	1.53E-04	3.066
6	2728512	2.82E-07	4.766	8.03E-05	3.177	8.04E-05	3.177
7	4779456	2.18E-07	1.377	4.98E-05	2.560	4.98E-05	2.557

Table 2, $\mathbb{Q}_4$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	111250	6.83E-06	—	1.51E-03	—	1.50E-03	—
2	279000	8.10E-07	6.958	2.41E-04	5.976	2.41E-04	5.975
3	627000	2.17E-07	4.874	5.24E-05	5.661	5.23E-05	5.659
4	1402500	9.17E-08	3.217	1.79E-05	4.011	1.78E-05	4.024
5	2901250	3.30E-08	4.213	6.69E-06	4.050	6.53E-06	4.131

• Test 5, Discontinuous permeability, Locally refined meshes

Table 1, $\mathbb{Q}_1$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	176	7936	-5.46E+01	-1.00E+02	54.594	100.000	52.441
2	1408	71168	-3.09E+01	-3.54E+01	30.857	35.355	79.708
3	11264	600064	-7.05E+01	-7.89E+01	70.515	78.858	89.071
4	90112	4923392	-9.14E+01	-9.43E+01	91.442	94.346	96.089
5	720896	39878656	-9.78E+01	-9.86E+01	97.780	98.562	98.260

Table 1, $\mathbb{Q}_2$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	594	90396	-1.18E+02	-1.00E+02	118.325	100.000	144.541
2	4752	810648	-3.83E+01	-3.54E+01	38.300	35.355	108.493
3	38016	6835104	-7.90E+01	-7.89E+01	78.962	78.858	100.680
4	304128	56080512	-9.44E+01	-9.43E+01	94.354	94.346	99.360
5	2433024	454242816	-9.86E+01	-9.86E+01	98.563	98.562	99.089

Table 1, $\mathbb{Q}_3$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1408	507904	-1.05E+02	-1.00E+02	104.586	100.000	123.703
2	11264	4554752	-3.55E+01	-3.54E+01	35.484	35.355	100.084
3	90112	38404096	-7.88E+01	-7.89E+01	78.828	78.858	99.078
4	720896	315097088	-9.43E+01	-9.43E+01	94.343	94.346	99.013
5	5767168	2552233984	-9.86E+01	-9.86E+01	98.562	98.562	99.010

Table 2, $\mathbb{Q}_1$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	176	1.31E+00	—	1.12E+00	—	1.29E+00	—
2	1408	2.71E-01	2.277	5.30E-01	1.074	5.89E-01	1.126
3	11264	6.42E-02	2.079	2.46E-01	1.109	2.62E-01	1.169
4	90112	1.61E-02	1.992	1.16E-01	1.080	1.18E-01	1.152
5	720896	4.05E-03	1.993	5.70E-02	1.027	5.71E-02	1.044

Table 2, $\mathbb{Q}_2$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	594	3.58E-01	—	6.29E-01	—	8.35E-01	—
2	4752	6.14E-02	2.543	1.88E-01	1.740	2.53E-01	1.720
3	38016	6.35E-03	3.274	3.82E-02	2.302	4.72E-02	2.426
4	304128	6.54E-04	3.278	8.34E-03	2.196	9.44E-03	2.321
5	2433024	6.72E-05	3.284	1.88E-03	2.148	2.01E-03	2.228

Table 2, $\mathbb{Q}_3$

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1408	2.37E-01	—	5.81E-01	—	8.64E-01	—
2	11264	7.67E-03	4.951	3.58E-02	4.023	5.19E-02	4.059
3	90112	3.04E-04	4.656	2.82E-03	3.665	3.51E-03	3.884
4	720896	1.37E-05	4.474	2.35E-04	3.583	2.49E-04	3.816
5	5767168	8.04E-07	4.088	2.50E-05	3.236	2.50E-05	3.316

3 Comments

Tests 1 and 5 were uncritical on all meshes. The penalty parameter can be chosen small and the corresponding symmetric and positive definite systems are easily solved. Either $\mathbb{P}_k$ or $\mathbb{Q}_k$ can be chosen, with $\mathbb{Q}_k$ being slightly more efficient in terms of error with respect to degrees of freedom. Tests 3 and 4 are much more difficult with two consequences: First, only $\mathbb{Q}_k$ did work on these meshes. Secondly, the global penalty parameter α has to be chosen large in order to obtain optimal convergence rates. In test 3 it has to be increased with polynomial degree (and even with these values the convergence rate breaks down on the finest mesh with $k = 4$). Large penalty parameters lead to very ill conditioned matrices which take a large number of iterations to solve. Note that the standard continuous Galerkin finite element method has no difficulties at all with tests 3 and 4. All numerical tests have been performed with the DUNE software framework[1, 2, 4], using the `dune-pdelab` discretization framework described in [3].

Acknowledgements I would like to thank Robert Klöfkorn for providing all the meshes in DUNE grid format and lots of discussions.

References

1. P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. part II: implementation and tests in DUNE. *Computing*, 82(2-3):121–138, 2008.
2. P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. part I: abstract framework. *Computing*, 82(2-3):103–119, 2008.
3. P. Bastian, F. Heimann, and S. Marnach. Generic implementation of finite element methods in the distributed and unified numerics environment (dune). *Kybernetika*, 46(2):294–315, 2010.
4. M. Blatt and P. Bastian. The iterative solver template library. In B. Kagström, E. Elmroth, J. Dongarra, and J. Wasniewski, editors, *Applied Parallel Computing. State of the Art in Scientific Computing*, number 4699 in Lecture Notes in Scientific Computing, pages 666–675, 2007.
5. Y. Epshteyn and B. Rivière. Estimation of penalty parameters for symmetric interior penalty Galerkin methods. *Journal of Computational and Applied Mathematics*, 206:843–872, 2007.
6. A. Ern, A. F. Stephansen, and P. Zunino. A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity. *IMA Journal of Numerical Analysis*, 29:235–256, 2009. doi:10.1093/imanum/drm050.
7. P. Houston and R. Hartmann. An optimal order interior penalty discontinuous Galerkin discretization of the compressible Navier-Stokes equations. *J. Comp. Phys.*, 227:9670–9685, 2008.
8. B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*. Frontiers in Applied Mathematics. SIAM, 2008.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: A Mimetic Finite Difference Method

Peter Bastian, Olaf Ippisch, and Sven Marnach

1 Presentation of the scheme

In the two-dimensional discretisation benchmark session at the FVCA5 conference, we participated with a Mimetic Finite Difference (MFD) method [7]. In this paper, we present results for the three-dimensional case using the same method. Since the previous conference, the equivalence of MFD, Hybrid Finite Volume and Mixed Finite Volume methods has been demonstrated in [6]. Our outline of the method as used in our computations follows the exposition in [5].

First, the diffusion problem is restated as a system of two first order PDEs

$$\begin{aligned} \operatorname{div}\mathbf{v} &= f \quad \text{in } \Omega, \\ \mathbf{v} &= -\mathbf{K}\nabla u \quad \text{in } \Omega, \\ u &= \bar{u} \quad \text{on } \Gamma_D, \\ \mathbf{K}\nabla u \cdot n &= g \quad \text{on } \Gamma_N. \end{aligned} \tag{1}$$

Our aim will be the definition of discrete analogues of the divergence operator div and the flux operator $-\mathbf{K}\nabla$. To this end, we first define the spaces of discrete scalar and vector functions. Let $\mathscr{T}_h$ denote a conforming triangulation of the domain Ω. The elements $E \in \mathscr{T}_h$ are assumed to be polyhedra. For details on the requirements of $\mathscr{T}_h$, see [4].

A *discrete scalar function* u is assumed to be constant on the elements E of the triangulation. The value of u in the element E is denoted by u_E. The dimension of the space Q_h of discrete scalar functions is equal to the number of elements of $\mathscr{T}_h$.

Peter Bastian, Olaf Ippisch, and Sven Marnach
Interdisciplinary Center for Scientific Computing, University of Heidelberg.
Corresponding author Sven Marnach, e-mail: sven.marnach@iwr.uni-heidelberg.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_93, © Springer-Verlag Berlin Heidelberg 2011

A *discrete vector function* $\mathbf{v}$ is given by assigning a real number $\mathbf{v}_E^e$ to each face e of each element E. These numbers are regarded as the normal components of the vector function with respect to the outer normal $\mathbf{n}_E^e$ on the face e. For a face e that is shared by elements E_1 and E_2, we require the compatibility of the normal components

$$\mathbf{v}_{E_1}^e = -\mathbf{v}_{E_2}^e. \tag{2}$$

Thus the dimension of the space X_h of discrete vector functions is equal to the number of faces of $\mathscr{T}_h$.

We now introduce the discrete differential operators. Again, see [4] for the details. The *discrete divergence operator* $\mathrm{div}_h : \ X_h \to Q_h$ is defined to comply with the divergence theorem on each element,

$$(\mathrm{div}_h \mathbf{v})_E = \frac{1}{|E|} \sum_{e \subset \partial E} |e| \; v_E^e, \tag{3}$$

where the sum is over all faces e of the element E. For the definition of the *discrete flux operator*, we introduce additional scalar unknowns on each face e of the triangulation denoted by u^e, and define $(-\mathbf{K}\nabla)_h : \ Q_h \to X_h$ by

$$\big((-\mathbf{K}\nabla)_h u\big)_E^e = \sum_{f \subset \partial E} \mathbb{W}_E^{e,f} |f| (u_E - u^f), \tag{4}$$

where $\mathbb{W}_E$ is a symmetric and positive definite matrix defined below. The scalar unknowns on the faces can be eliminated by compatibility requirement (2) on the inner faces and by boundary conditions (1) on the outer faces.

Now we can give the whole linear system discretising the linear diffusion problem. For each element E, we have the equation

$$\mathrm{div}_h (-\mathbf{K}\nabla)_h u = \frac{1}{|E|} \sum_{e \subset \partial E} |e| \sum_{f \subset \partial E} \mathbb{W}_E^{e,f} |f| (u_E - u^f) = q_E. \tag{5}$$

The equation for an inner face e shared by the elements E_1 and E_2 reads

$$\begin{aligned}\big((-\mathbf{K}\nabla)_h u\big)_{E_1}^e &+ \big((-\mathbf{K}\nabla)_h u\big)_{E_2}^e \\ &= \sum_{f \subset \partial E_1} \mathbb{W}_{E_1}^{e,f} |f| (u_{E_1} - u^f) + \sum_{f \subset \partial E_2} \mathbb{W}_{E_2}^{e,f} |f| (u_{E_2} - u^f) = 0.\end{aligned} \tag{6}$$

For each face e on the Neumann boundary Γ_N, we get

$$\big((-\mathbf{K}\nabla)_h u\big)_E^e = \sum_{f \subset \partial E} \mathbb{W}_E^{e,f} |f| (u_E - u^f) = g^e. \tag{7}$$

Finally, a face e on the Dirichlet boundary provides us with the trivial equation

$$u^e = \bar{u}^e. \tag{8}$$

To obtain a symmetric and positive definite stiffness matrix, we first eliminate the unknowns on the Dirichlet boundaries. Then, we scale (5) by $|E|$ and (6) as well as (7) by $-|e|$, see [8].

We finally give the definition of the matrix $\mathbb{W}_E$ for an element E. Let k_E denote the number of faces of E. Define the $k_E \times 3$ matrices $\mathbb{R}$ and $\mathbb{N}$ by

$$\mathbb{R}_{e,i} = \int_e (x_i - x_{E,i}), \quad \mathbb{N}_{e,i} = \mathbf{e}_i \cdot n_E^e, \tag{9}$$

where e ranges over the faces of E, $i = 1, 2, 3$, x_i is the i-th coordinate function, x_E is the ”centre” of E and $\mathbf{e}_i$ is the unit vector in the direction of the i-th axis. The centre point x_E can be chosen on each cell individually (subject to some restrictions). A good choice is to use the centre of mass, which we used for the tetrahedral mesh. For the hexahedral meshes, we used the image of the centre of the unit cube under the usual trilinear coordinate mappings.

We construct a $k_E \times k_E$ matrix $\mathbb{W}_E$ according to algorithm 1 in [5]. In short, that means the following:

1. Orthonormalise the columns of the matrix $\mathbb{R}$ using the Gram–Schmidt algorithm and call the resulting matrix $\tilde{\mathbb{R}}$.
2. Set $\mathbb{D} = \mathbb{I} - \tilde{\mathbb{R}}\tilde{\mathbb{R}}^{\mathrm{T}}$, where $\mathbb{I}$ denotes the $k_E \times k_E$ unit matrix.
3. Define $\mathbb{W}_E$ by

$$\mathbb{W}_E = \frac{1}{|E|}\mathbb{N}\mathbf{K}\mathbb{N}^{\mathrm{T}} + \omega\mathbb{D}, \tag{10}$$

   where ω is an arbitrary positive real number and $\mathbf{K}$ is simply evaluated at the cell centre x_E.

   We used the common choice for ω

$$\omega = \frac{\operatorname{trace}\mathbf{K}}{|E|}, \tag{11}$$

which was suggested in [5].

2 Numerical results

For estimating the L_2 error, we compared the approximate solution u_E on a cell E with the average

$$u_{E,\text{exact}} = \frac{1}{|E|}\int_E u(x)dx$$

of the exact solution over the cell,

$$\text{erl2}^2 = \frac{\sum_E |E|(u_E - u_{E,\text{exact}})^2}{\sum_E |E|(u_{E,\text{exact}})^2}. \tag{12}$$

The MFD method provides values for the fluxes on the faces, but does not allow the direct computation of approximate gradients of the solution. In some cases it would be possible to get a reconstruction of the gradients in the interior of a cell using the Piola transformation, but this fails, for example, for cells with hanging nodes on some faces. To circumvent these problems, we substituted the H_1 semi-norm by the somewhat unnatural "flux norm"

$$\text{ergrad}^2 = \frac{\sum_e |e|(\mathbf{v}^e - \mathbf{v}^e_{\text{exact}})^2}{\sum_e |e|(\mathbf{v}^e_{\text{exact}})^2}, \tag{13}$$

where the exact average flux over the face e is given by

$$\mathbf{v}^e_{\text{exact}} = \frac{1}{|e|}\int_e \mathbf{n}(x)\cdot A(x)\nabla u(x)dx$$

We did not provide any values for the energy norm E. Though it is possible to give an approximation of the energy norm using the formulation

$$\begin{aligned} E &= \int_\Omega \mathbf{K}\nabla u\cdot\nabla u \\ &= \int_\Omega \mathbf{K}^{-1}(-\mathbf{K}\nabla u)\cdot(-\mathbf{K}\nabla u) \\ &= \sum_{E\in\mathscr{T}_h}\int_E \mathbf{K}^{-1}(-\mathbf{K}\nabla u)\cdot(-\mathbf{K}\nabla u) \end{aligned}$$

and the scalar product on X_h, this would not provide much information, since E would coincide with the exact energy norm up to the accuracy of the linear solver by construction of the method.

All numerical tests have been performed with the DUNE software framework [1,3] using the `dune-pdelab` discretisation framework described in [2].

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	6311	46371	2.26E-02	2.03E-02	1.986	1.989	0.000
2	12146	90106	5.50E-03	6.84E-03	1.989	1.989	0.000
3	23859	178079	8.50E-03	9.13E-03	1.994	1.994	0.000
4	46957	352277	5.10E-03	5.52E-03	1.997	1.997	0.000
5	93267	702867	1.91E-03	1.49E-03	1.996	1.997	0.000
6	186040	1407080	1.75E-03	1.83E-03	1.997	1.997	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	6311	4.55E-03	0.000	1.41E-01	0.000	0.00E+00	0.000
2	12146	2.88E-03	2.102	1.05E-01	1.376	0.00E+00	0.000
3	23859	1.82E-03	2.030	9.73E-02	0.328	0.00E+00	0.000
4	46957	1.20E-03	1.859	7.03E-02	1.440	0.00E+00	0.000
5	93267	7.38E-04	2.120	6.13E-02	0.602	0.00E+00	0.000
6	186040	4.65E-04	2.004	4.75E-02	1.102	0.00E+00	0.000

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	2240	23744	-6.03E-01	3.03E-02	2.100	1.958	0.000
2	17152	189184	-5.83E-03	1.06E-02	2.008	1.993	0.000
3	134144	1510400	-1.11E-03	1.75E-03	2.000	1.997	0.000
4	1060864	12070912	1.65E-04	7.14E-04	2.000	1.999	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	2240	3.27E-01	0.000	1.19E+01	0.000	0.00E+00	0.000
2	17152	5.28E-02	2.685	3.37E+00	1.859	0.00E+00	0.000
3	134144	8.89E-03	2.600	5.26E-01	2.709	0.00E+00	0.000
4	1060864	2.02E-03	2.146	1.12E-01	2.245	0.00E+00	0.000

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi\left(y + \frac{1}{2}\right)\right) \sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	192	2496	1.27E-01	1.54E-01	1.883	1.846	0.000
2	1488	25248	-1.32E-01	4.01E-02	2.150	1.960	0.000
3	11712	225696	-5.37E-02	1.01E-02	2.053	1.990	0.000
4	92928	1905600	-1.40E-02	2.54E-03	2.014	1.997	0.000
5	740352	15655296	-3.52E-03	6.36E-04	2.004	1.999	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	192	2.81E-01	0.000	4.00E-01	0.000	0.00E+00	0.000
2	1488	1.19E-01	1.263	2.15E-01	0.906	0.00E+00	0.000
3	11712	3.44E-02	1.801	1.13E-01	0.936	0.00E+00	0.000
4	92928	9.39E-03	1.879	5.71E-02	0.992	0.00E+00	0.000
5	740352	2.46E-03	1.936	2.86E-02	1.001	0.00E+00	0.000

• **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	304	2992	-1.02E+00	-7.59E-01	1.045	0.691	0.000
2	2240	23744	-9.79E-01	-9.39E-01	1.019	0.923	0.000
3	17152	189184	-1.02E+00	-9.85E-01	1.008	0.982	0.000
4	134144	1510400	-1.00E+00	-9.96E-01	1.000	0.996	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	304	8.38E-01	0.000	9.62E-01	0.000	0.00E+00	0.000
2	2240	1.58E-01	2.502	3.14E-01	1.679	0.00E+00	0.000
3	17152	3.91E-02	2.060	1.29E-01	1.314	0.00E+00	0.000
4	134144	1.15E-02	1.788	6.61E-02	0.975	0.00E+00	0.000

- **Test 4 Flow around a well, Well meshes,** min $= 0$, max $= 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3888	41268	5.74E-01	4.14E-01	5.317	5.317	0.000
2	9464	103208	2.96E-01	2.44E-01	5.328	5.328	0.000
3	20902	231574	1.75E-01	1.54E-01	5.329	5.329	0.000
4	46203	517443	1.30E-01	1.18E-01	5.330	5.330	0.000
5	94885	1069705	9.66E-02	8.99E-02	5.339	5.339	0.000
6	173515	1964101	7.67E-02	7.23E-02	5.345	5.345	0.000
7	303058	3439576	5.91E-02	5.65E-02	5.361	5.361	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3888	5.53E-03	0.000	2.70E-01	0.000	0.00E+00	0.000
2	9464	1.64E-03	4.090	1.08E-01	3.089	0.00E+00	0.000
3	20902	8.43E-04	2.530	5.04E-02	2.888	0.00E+00	0.000
4	46203	8.27E-04	0.074	2.84E-02	2.173	0.00E+00	0.000
5	94885	6.83E-04	0.796	1.70E-02	2.142	0.00E+00	0.000
6	173515	4.84E-04	1.709	1.12E-02	2.060	0.00E+00	0.000
7	303058	4.07E-04	0.932	7.78E-03	1.965	0.00E+00	0.000

- **Test 5 Discontinuous permeability,** $u(x, y, z) = a_i \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$**,** min $= -100$, max $= 100$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	115	1231	-2.51E+02	-1.00E+02	250.808	100.000	0.000
2	812	8972	-4.44E+01	-3.54E+01	44.367	35.355	0.000
3	6064	68272	-8.32E+01	-7.89E+01	83.205	78.858	0.000
4	46784	532160	-9.56E+01	-9.43E+01	95.600	94.346	0.000
5	367360	4201216	-9.89E+01	-9.86E+01	98.887	98.562	0.000

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	115	8.77E+00	0.000	2.92E+00	0.000	0.00E+00	0.000
2	812	7.09E-01	3.860	7.88E-01	2.012	0.00E+00	0.000
3	6064	1.41E-01	2.413	3.15E-01	1.367	0.00E+00	0.000
4	46784	3.33E-02	2.116	2.13E-01	0.576	0.00E+00	0.000
5	367360	8.21E-03	2.038	1.55E-01	0.464	0.00E+00	0.000

References

1. Bastian, P., Blatt, M., Dedner, A., Engwer, C., Klöfkorn, R., Kornhuber, R., Ohlberger, M., Sander, O.: A generic grid interface for parallel and adaptive scientific computing. part ii: Implementation and tests in dune. Computing **82**(2–3), 121–138 (2007)
2. Bastian, P., Heimann, F., Marnach, S.: Generic software components for finite elements (2009). To appear in the proceedings of Algorithmy 2009
3. Blatt, M., Bastian, P.: The iterative solver template library. In: B. Kagström, E. Elmroth, J. Dongarra, J. Wasniewski (eds.) Applied Parallel Computing. State of the Art in Scientific Computing, *Lecture Notes in Scientific Computing*, vol. 4699, pp. 666–675. Springer (2007)
4. Brezzi, F., Lipnikov, K., Shashkov, M.: Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. SIAM Journal on Numerical Analysis **43**(5), 1872–1896 (2005)
5. Brezzi, F., Lipnikov, K., Simoncini, V.: A family of mimetic finite difference methods on polygonal and polyhedral meshes. Mathematical Models and Methods in Applied Sciences **15**(10), 1533–1551 (2005)
6. Droniou, J., Eymard, R., Gallouët, T., Herbin, R.: A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. Mathematical Models and Methods in Applied Sciences **20**(2), 265 – 295 (2010)
7. Marnach, S.: Benchmark on anisotropic problems – a mimetic finite difference method. In: J.M.H. Robert Eymard (ed.) Finite Volumes for Complex Applications V – Problems and Perspectives. ISTE Ltd and John Wiley & Sons Inc (2008)
8. Morel, J.E., Roberts, R.M., Shashkov, M.J.: A local support-operators diffusion discretization scheme for quadrilateral r-z meshes. Journal of Computational Physics **144**, 17–51 (1998)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: A Composite Hexahedral Mixed Finite Element

Ibtihel Ben Gharbia, Jérôme Jaffré, N. Suresh Kumar, and Jean E. Roberts

1 The Numerical Scheme

The numerical method used here (see [6]) is a mixed finite element method based on the weak formulation of the problem:

$$\begin{aligned}
&\text{Find } (p,\boldsymbol{u}) \in L^2(\Omega) \times H(\mathrm{div};\Omega) \text{ such that}\\
&\int_\Omega \boldsymbol{K}^{-1}\boldsymbol{u}\cdot\boldsymbol{v} - \int_\Omega p\,\mathrm{div}\boldsymbol{v} = -\int_{\Gamma_D} \bar{p}\,\boldsymbol{v}\cdot\boldsymbol{n} \quad \forall \boldsymbol{v} \in H(\mathrm{div};\Omega) \qquad (1)\\
&\int_\Omega \mathrm{div}\boldsymbol{v}\,q = \int_\Omega fq \quad \forall q \in L^2(\Omega).
\end{aligned}$$

Straightforward extensions of the Raviart-Thomas-Nédelec mixed finite elements [3, 5] to hexahedral meshes do not converge. Therefore in [6] a composite mixed finite element was introduced and analyzed.

Given a discretization $\mathscr{T}_h$ of Ω into hexahedra (with planar faces) we solve the following system:

Ibtihel Ben Gharbia, Jérôme Jaffré, and Jean E. Roberts
INRIA Paris-Rocquencourt, 78153 LeChesnay, France,
e-mail: ibtihel.ben-gharbia@inria.fr, jerome.jaffre@inria.fr, jean.roberts@inria.fr

N. Suresh Kumar
Department of Mathematics, National Institute of Technology Calicut, India,
e-mail: sureshknsk@gmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_94, © Springer-Verlag Berlin Heidelberg 2011

$$\text{Find } (p_h, \boldsymbol{u}_h) \in M_h \times \boldsymbol{W}_h \text{ such that}$$
$$\int_\Omega \boldsymbol{K}^{-1}\boldsymbol{u}_h \cdot \boldsymbol{v}_h - \int_\Omega p_h \text{div}\boldsymbol{v}_h = -\int_{\Gamma_D} \bar{p}\boldsymbol{v}_h \cdot \boldsymbol{n} \quad \forall \boldsymbol{v}_h \in \boldsymbol{W}_h, \tag{2}$$
$$\int_\Omega \text{div}\boldsymbol{v}_h\, q_h = \int_\Omega f q_h \quad \forall q_h \in M_h,$$

where $M_h \subset L^2(\Omega)$ is the space of piecewise constant functions (just as in the lowest order Raviart-Thomas-Nedelec spaces for tetrahedra or for rectangular solids). The space $\boldsymbol{W}_h \subset H(\text{div}; \Omega)$ is constructed following ideas of Kuznetsov and Repin see [4]. It is a space of composite elements satisfying the following 4 conditions (all of which are satisfied by the Raviart-Thomas-Nédelec elements when the underlying spatial discretization is made up of tetrahedra and/or rectangular solids):

- $\boldsymbol{W}_h \subset H(\text{div}; \Omega)$; i.e. elements of $\boldsymbol{W}_h$ are locally in $H(\text{div}; T)$; $\forall T \in \mathscr{T}_h$, and normal components of elements of $\boldsymbol{W}_h$ are continuous across edges of the hexahedra in $\mathscr{T}_h$.
- normal components of elements of $\boldsymbol{W}_h$ are constant on each face of an element of $\mathscr{T}_h$.
- $\text{div}\boldsymbol{W}_h \subset M_h$; i.e. the divergence of an element of $\boldsymbol{W}_h$ is constant on each hexahedron of $\mathscr{T}_h$.
- an element of $\boldsymbol{W}_h$ is uniquely determined by its flux through the faces of elements of $\mathscr{T}_h$; i.e. $\boldsymbol{W}_h$ has a basis of functions $\{\boldsymbol{v}_F : F \in \mathscr{F}_h\}$, where $\mathscr{F}_h$ is the set of all faces of hexahedra in $\mathscr{T}_h$, not lying on Γ_N, and for $F \in \mathscr{F}_h$, $\boldsymbol{v}_F$ is the unique function in $\boldsymbol{W}_h$ having normal component with flux across the face $E \in \mathscr{F}_h$ equal to $\delta_{E,F}$.

The space $\boldsymbol{W}_h$ is constructed element by element: for an element $T \in \mathscr{T}_h$ we define the space $\boldsymbol{W}_T$ of functions on T, and then $\boldsymbol{W}_h$ is defined to be the subspace of $H(\text{div}; \Omega)$ consisting of those functions whose restriction to T is in $\boldsymbol{W}_T$ for each $T \in \mathscr{T}_h$. To construct $\boldsymbol{W}_T$ for an element $T \in \mathscr{T}_h$, T is subdivided into 5 tetrahedra as follows: starting from any vertex V_1 of T there are 3 vertices (say V_2, V_4, and V_5) of T that can be joined to V_1 by an edge of T, there are 3 other vertices (say V_3, V_6, and V_8) that lie on a face with V_1 (but not on an edge with V_1). The remaining vertex V_7 together with V_2, V_4, and V_5 forms a tetrahedron S_0 having no face lying on the boundary of T. Then $T \setminus S_0$ is made up of 4 tetrahedra S_1, S_2, S_3 and S_4, each of which has 3 faces lying on the surface of T and one face in common with S_0; see Fig. 1.

The collection of tetrahedra $\mathscr{T}_T = \{S_i : i = 0, 1, \cdots, 4\}$ is a discretization of T by tetrahedra, and we denote by $\widetilde{\boldsymbol{W}}_T$ the Raviart-Thomas-Nédelec space of lowest order associated with $\mathscr{T}_T$. We let $\widetilde{\boldsymbol{M}}_T$ denote the set of functions constant on each of the five tetrahedra in $\mathscr{T}_T$, let $\widetilde{\boldsymbol{W}}_{T,0} \subset \widetilde{\boldsymbol{W}}_T$ denote the set of functions in $\widetilde{\boldsymbol{W}}_T$ whose normal traces vanish on all of ∂T, and let $|T|$ denote the volume of T. For each face F of T, letting $|F|$ denote the area of F and letting $\widetilde{\boldsymbol{W}}_{T,F} \subset \widetilde{\boldsymbol{W}}_T$ denote the set of functions in $\widetilde{\boldsymbol{W}}_T$ whose normal traces vanish on all of $\partial T \setminus F$ and are identically

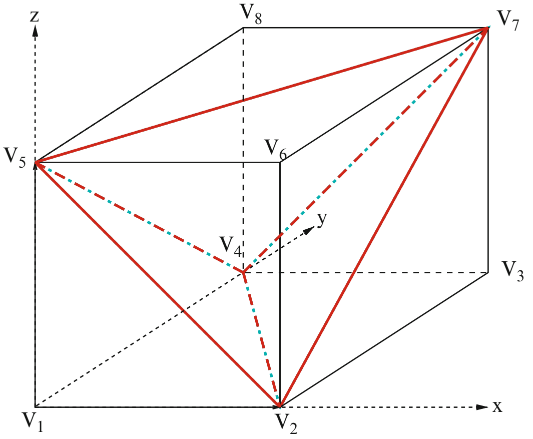


Fig. 1 A partition of the reference hexahedron into 5 tetrahedra: one tetrahedron lies in the interior of T and is determined by the vertices V_2, V_4, V_5, V_7. The four other tetrahedra have each three faces on the surface of T and each contains one of the vertices V_1, V_3, V_6, V_8. There are two possible such constructions depending on which vertex is chosen as V_1

equal to $\frac{1}{|F|}$ on F, we define $\boldsymbol{v}_F$ to be the second component of the solution of the problem

$$\begin{aligned} &\text{Find}(q_F, \boldsymbol{v}_F) \in \widetilde{M}_T \times \widetilde{\boldsymbol{W}}_{T,F} \text{ such that} \\ &\int_T \boldsymbol{v}_F \cdot \boldsymbol{v}_h - \int_T q_F \text{div} \boldsymbol{v}_h = 0, \quad \forall \boldsymbol{v}_h \in \widetilde{\boldsymbol{W}}_{T,0}, \\ &\int_T \text{div} \boldsymbol{v}_F \; q_h = \frac{1}{|T|} \int_T q_h \quad \forall q_h \in \widetilde{M}_T. \end{aligned} \tag{3}$$

The pure Neumann problem (3) has a solution since the compatibility condition - that the integral over ∂T of the Neumann data function be equal to the integral over T of the source term - is satisfied. The second component $\boldsymbol{v}_F$ of the solution is uniquely determined: in the algebraic system associated with problem (3), the four equations corresponding to the four exterior tetrahedra, $S_1, \cdots, S_4$, determine $\boldsymbol{v}_F$, the equation associated with S_0 is redundant but is not a problem since the compatibility condition is satisfied. (The four equations associated with the internal faces, the four faces of S_0, imply that q_F is constant on all of T, but do not determine the value of the constant, but this is not needed here.) Then $\boldsymbol{W}_T \subset \widetilde{\boldsymbol{W}}_T$ is defined to be simply the six-dimensional subspace generated by the basis elements $\{\boldsymbol{v}_F :$ F is a face of $T\}$. Now defining $\boldsymbol{W}_h$ by

$$\boldsymbol{W}_h = \{\boldsymbol{v} \in H(\text{div}; \Omega) \; : \; \boldsymbol{v}_{|T} \in \boldsymbol{W}_T, \quad \forall T \in \mathscr{T}_h\},$$

one can easily check that $\boldsymbol{W}_h$ satisfies the four conditions listed above.

Remark 1. We point out that there are two possible choices for $\mathscr{T}_T$ (and thus for $\boldsymbol{W}_T$) depending on whether (in the notation used above) vertices $\{V_2, V_4, V_5, V_7\}$ or the vertices $\{V_1, V_3, V_6, V_8\}$ are used to form the interior tetrahedron. Also it is not always possible to choose the sets $\mathscr{T}_T$ is such a way that $\cup_{T\in\mathscr{T}_h}\mathscr{T}_T$ forms a finite element decomposition of Ω into tetrahedra.

Remark 2. This method is not appropriate for meshes containing deformed cubes which are not true hexahedra; i.e. for meshes containing deformed cubes with nonplanar "faces". In applications nonplanar "faces" arise when a cube is deformed in such a way that four vertices defining a face of the cube are moved so that they are no longer planar. However any three of the vertices remain planar so for either choice of the decomposition into five tetrahedra the nonplanar "face" is divided into two (planar) triangles so that one obtains a polyhedron (with planar faces) of from seven to twelve sides depending on how many nonplanar "faces" the original "hexahedron" had. Thus the new polyhedron is divided into five tetrahedra and one could generalize the method used here to include this case. However, as mentioned above it may not be possible to choose the divisions of the hexahedra into five tetrahedra in such a way that the resulting collection of tetrahedra forms a finite element mesh; i. e. in such a way that the resulting division of the quadrilateral interior faces into two triangles is compatible for each pair of adjacent "hexahedra". The resulting pair of adjacent polyhedra may then either overlap or leave a void space between the two polyhedra. A new composite mixed finite element is now under development to treat the case of nonplanar faces.

Remark 3. One could in a perhaps more natural way divide each of the hexahedra into 6 tetrahedra (all of equal volume for the reference hexahedron) by adding a central edge between any single pair of vertices not belonging to a common face. The six tetrahedra would all have this edge in common and each would have two internal faces and two external faces. One could form a system similar to (2) for each of the six faces of T. The dimension of $\widetilde{M}_T$ would then be 6 instead of 5 and that of $\widetilde{\boldsymbol{W}}_{T,0}$ would be 6 instead of 4 as there would be 6 interior faces. The six equations of the linear system corresponding to one of the six tetrahedra would each only give a relation between the fluxes through the internal faces of the tetrahedron, so the second component of the solution would be determined only up to a (divergence free) flow going around the central edge. One would then need to impose a condition to make the macro elements rotational free (as are the Raviart-Thomas-Nédelec elements on tetrahedra and on rectangular solids as well as are those defined above on hexahedra using a decomposition into five tetrahedra). We have not further investigated this possibility.

Error estimates

In this paragraph we briefly recall the error estimates obtained in [6]. Following [1] we define the notion of shape regularity for a family of meshes of hexahedra.

Definition 1. For S a tetrahedron let ρ_S and h_S denote respectively the radius of the inscribed sphere of S and the diameter of S. Then for a hexahedron T, as seen earlier, there are two possible ways of decomposing T into five tetrahedra. Let ρ_T

be the smallest of the ρ_S's for these 10 tetrahedra, let h_T be the diameter of T and let $\sigma_T = h_T/\rho_T$ be the shape constant of T. For a mesh $\mathscr{T}_h$ of hexahedra, the shape constant of $\mathscr{T}_h$ is the largest σ_T for $T \in \mathscr{T}_h$. A family $\{\mathscr{T}_h : h \in \mathscr{H}\}$ of meshes $\mathscr{T}_h$ made up of hexahedra is said to be *shape regular* if the shape constants for the meshes can be uniformly bounded.

In [6] it is shown that if $(p, \boldsymbol{u}) \in L^2(\Omega) \times H(\mathrm{div};\Omega)$ is the solution of problem (1) and $(p_h, \boldsymbol{u}_h) \in M_h \times \boldsymbol{W}_h$ is the solution of problem (2) and the family $\{\mathscr{T}_h : h \in \mathscr{H}\}$ of meshes $\mathscr{T}_h$ made up of hexahedra is shape regular then there is a constant C independent of h such that

$$\|p_h - p\|^2_{L^2(\Omega)} + \|\boldsymbol{u}_h - \boldsymbol{u}\|^2_{H(\mathrm{div};\Omega)} \leq C\, h^2 \left(|p|^2_{H^1(\Omega)} + \|\boldsymbol{u}\|^2_{H^1(\Omega)} + \|\mathrm{div}\boldsymbol{u}\|^2_{H^1(\Omega)}\right),$$

provided that p and $\boldsymbol{u}$ are sufficiently regular for the righthand side to be defined.

Mixed-hybrid finite elements and solution of the linear problem

As with the Raviart-Thomas-Nédelec elements for tetrahedra and rectangular solids, the solution $(\boldsymbol{u}_h, p_h)$ is sought in a subspace $M_h \times \boldsymbol{W}_h$ of $L^2(\Omega) \times H(\mathrm{div};\Omega)$ in which the degrees of freedom are the average values of the pressure over the hexahedra of the grid and the fluxes through the faces of the grid. The resulting linear system then has exactly the same form as that for the Raviart-Thomas-Nédelec elements for grids of rectangular solids (when the problem has full tensor coefficients). As in [2] we can relax the condition that the approximate solution be sought in a subspace of $H(\mathrm{div};\Omega)$ and enforce this condition using Lagrange multipliers. We then define the approximation space $\boldsymbol{W}^*_h$ by

$$\boldsymbol{W}^*_h = \{\boldsymbol{v} \in (L^2(\Omega))^3 \;:\; \boldsymbol{v}_{|T} \in \boldsymbol{W}_T, \quad \forall T \in \mathscr{T}_h\},$$

and introduce a space of Lagrange multipliers $\Lambda_h = \{\lambda_h = \{\lambda_F\}_{F\in\mathscr{F}_h} \in R^{n_F}\}$ where n_F is the number of faces in $\mathscr{F}_h$. Then the following problem has a unique solution:

Find $(p^*_h, \boldsymbol{u}^*_h, \lambda_h) \in M_h \times \boldsymbol{W}^*_h \times \Lambda_h$ such that

$$\sum_{T\in\mathscr{T}_h}\int_T \boldsymbol{K}^{-1}\boldsymbol{u}^*_h \cdot \boldsymbol{v}_h - \sum_{T\in\mathscr{T}_h}\int_T p^*_h \mathrm{div}\boldsymbol{v}_h - \sum_{F\in\mathscr{F}_h}\int_F \lambda_F[\boldsymbol{v}_h \cdot \boldsymbol{n}_F] = -\int_{\Gamma_D} \bar{p}\boldsymbol{v}_h \cdot \boldsymbol{n} \quad \forall \boldsymbol{v}_h \in \boldsymbol{W}^*_h,$$

$$\sum_{T\in\mathscr{T}_h} \mathrm{div}\boldsymbol{u}^*_h\, q_h = \int_\Omega f q_h \quad \forall q_h \in M_h,$$

$$\sum_{F\in\mathscr{F}_h}\int_F [\boldsymbol{u}^*_h \cdot \boldsymbol{n}_F]\mu_F = 0, \quad \forall \mu_h \in \Lambda_h,$$

where for $F \in \mathscr{F}_h$, $\boldsymbol{n}_F$ is a unit vector normal to F and for $\boldsymbol{v}_h \in \boldsymbol{W}_h^*$, $[\boldsymbol{v}_h \cdot \boldsymbol{n}_F]$ denotes the jump across F of $\boldsymbol{v}_h \cdot \boldsymbol{n}_F$ in the direction of $\boldsymbol{n}_F$. As with the Raviart-Thomas-Nédelec method it is now easy to eliminate first $\boldsymbol{u}_h^*$ and then p_h^* from the linear system and thus obtain a symmetric positive definite system with λ_h as the only unknown. For $F \in \mathscr{F}_h$ the multiplier λ_F enforcing continuity of $\boldsymbol{u}_h^* \cdot \boldsymbol{n}_F$ across F is in fact an approximation of the trace of the pressure p on F.

Once λ_h is found one can recover $\boldsymbol{u}_h^*$ and p_h^* through local calculations given by the first two equations of system (4). One shows easily that $\boldsymbol{u}_h^*$ is in fact in $\boldsymbol{W}_h$ and is equal to $\boldsymbol{u}_h$ and that $p_h^* = p_h$.

2 Numerical experiments

The data are provided by the *FVCA6 3D anisotropic benchmark* . We have chosen to do the first test case with mild anisotropy and Kershaw grids.Table 1 gives results obtained for the 4 Kershaw meshes which were proposed in the benchmark. The index i, i $= 1, 2, 3, 4$, denotes the mesh index for the $8 \times 8 \times 8$, the $16 \times 16 \times 16$, the $32 \times 32 \times 32$, and the $64 \times 64 \times 64$ Kershaw meshes respectively. As mentioned earlier, the matrix of the linear system associated with the mixed-hybrid finite element after elimination of p_h^* and $\boldsymbol{u}_h^*$ is symmetric and positive definite, and the unknowns are the Lagrangian multipliers λ_h which are approximations of the averages of the trace of the scalar variable (pressure) over the faces. From λ_h local calculations yield the cell pressure unknowns of p_h and the fluxes across the faces of the velocity $\boldsymbol{u}_h$.

In Table 1 nu, the number of unknowns of the linear system, is the number of degrees of freedom for λ_h which is the number of faces. The number of matrix nonzeros, nmat, is given in the table for the full matrix (not the upper or lower halves). umin, uemin, λmin (resp. umax, uemax, λmax) the minimum (resp. maximum) of p_h, p and λ_h.

The function p_h is constant inside each hexahedral cell, so the L^2 error erl2 between p and p_h is calculated as

$$\text{erl2} = \frac{\sqrt{\int_{\Omega} (p - p_h)^2}}{\sqrt{\int_{\Omega} p^2}}$$

where the integrals in the numerator are calculated using on each cell an integration formula exact for polynomials of degree 2 in 3D.

The mixed finite element method calculates also the velocity $\boldsymbol{u}_h$ approximating the vector unknown $\boldsymbol{u} = -\boldsymbol{K}\nabla p$ as a piecewise polynomial vector function. The usual error calculated with the mixed method is the L^2 error for $\boldsymbol{u}_h$ in addition to the L^2 error for p_h. However in this benchmark the errors for p_h in the H^1 seminorm and in the energy norm are asked for. These norms are actually equivalent to the

Table 1 Results obtained for a composite hexahedral mixed finite element on a sequence of Kershaw meshes

i	nu	nmat	umin	uemin	umax	uemax	normg
1	576	2496	-0.03255	0.	1.94685	2.	1.84064
2	4352	32512	-0.04618	0.	1.99488	2.	1.85063
3	33792	310272	-0.03621	0.	2.00028	2.	1.85242
4	266240	2682880	-0.00837	0.	2.00061	2.	1.84036

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	576	0.063751		1.63849		1.49726	
2	4352	0.038971	0.73	0.96309	0.79	0.86755	0.81
3	33792	0.019424	1.02	0.51181	0.92	0.44853	0.97
4	266240	0.009148	1.09	0.25421	1.02	0.21819	1.05

L^2 norm of $\boldsymbol{u}$. Indeed we have $|\nabla p|^2 = |\boldsymbol{K}^{-1}\boldsymbol{u}|^2$ and $(\boldsymbol{K}\nabla p)\cdot\nabla p = (\boldsymbol{K}^{-1}\boldsymbol{u})\cdot\boldsymbol{u}$. Therefore we calculate the error for the gradient and the error in the energy norm with the formula

$$\text{ergrad} = \frac{\sqrt{\int_\Omega |\boldsymbol{K}^{-1}(\boldsymbol{u}-\boldsymbol{u}_h)|^2}}{\sqrt{\int_\Omega |\boldsymbol{K}^{-1}\boldsymbol{u}|^2}}, \quad \text{ener} = \frac{\sqrt{\int_\Omega (\boldsymbol{K}^{-1}(\boldsymbol{u}-\boldsymbol{u}_h))\cdot(\boldsymbol{u}-\boldsymbol{u}_h)}}{\sqrt{\int_\Omega (\boldsymbol{K}^{-1}\boldsymbol{u})\cdot\boldsymbol{u}}}$$

where again the integrals in the numerator were calculated with an integration formula exact for polynomials of degree 2 inside each cell.

Similarly the L^1 norm of the gradient of p_h was calculated as

$$\text{normgrad} = \int_\Omega |\boldsymbol{K}^{-1}\boldsymbol{u}_h|.$$

The rates of convergence ratiol2, ratioener and ratiograd are calculated as required by the benchmark by comparing the errors erl2, ergrad and ener obtained on meshes i and i-1 using the formula

$$\text{ratio(i)} = -3\frac{\log(\text{error(i)}/\text{error(i-1)})}{\log(\text{nu(i)}/\text{nu(i-1)})}.$$

All errors behave as predicted by the theory and show an asymptotic rate of convergence of order 1. The exact solution is such that $0 \leq p \leq 2$ and the calculated solution has small undershoots which become smaller as the meshes are refined.

3 Conclusion

In spite of the bad aspect ratios of some of the hexahedra in the Kershaw meshes, the proposed composite hexahedral mixed finite element shows first order convergence for the pressure as well as for the velocity, as it was predicted by the analysis of the method.

Acknowledgements This work was partially supported by the GNR MoMaS (PACEN/CNRS, ANDRA, BRGM, CEA, EDF, IRSN).

References

1. D. N. Arnold, D. Boffi, and R. S. Falk. Quadrilateral h(div) finite elements. *SIAM J. Numer. Anal.*, 42:2429–2451, 2005.
2. D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods : implementation, postprocessing and error estimates. *M2AN*, 19:7–32, 1985.
3. F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer-Verlag, New York, 1991.
4. Yu. Kuznetzov and S. Repin. Convergence analysis and error estimates for mixed finite element method on distorted meshes. *Russ. J. Numer. Anal. Math. Modelling J. Numer. Math.*, 13:33–51, 2005.
5. J. E. Roberts and J.-M. Thomas. Mixed and hybrid methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of Numerical Analysis Vol.II*, pages 523–639. North Holland, Amsterdam, 1991.
6. A. Sboui, J. Jaffré, and J. E. Roberts. A composite mixed finite element for hexahedral grids. *SIAM J. Sci. Comput.*, 31(3):2623–2645, 2009.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: CeVeFE-DDFV, a discrete duality scheme with cell/vertex/face+edge unknowns

Yves Coudière, Florence Hubert, and Gianmarco Manzini

1 Presentation of the scheme

The method that we investigate in this contribution was proposed by Y. Coudière and F. Hubert in [1] as a three-dimensional (3D) extension of the finite volume scheme previously studied by F. Hermeline in [4] and K. Domelevo and P. Omnès in [3]. This method belongs to the family of Discrete Duality Finite Volume (DDFV) methods, which can naturally handle anisotropic or non-linear problems on general distorted meshes.

In this benchmark paper, we present the results obtained by using the formulation in [1] and the variant for discontinuous permeabilities that is presented in the proceeding paper [2].

The DDFV method that we consider herein makes use of three polyhedral meshes for the solution approximation, denoted by $\mathscr{M}$, $\mathscr{N}$, $\mathscr{F}\mathscr{E}$, and the mesh of diamonds for the solution gradient approximation, denoted by $\mathscr{D}$.

We denote the control volumes of the primal mesh $\mathscr{M}$ by K and L, and with every primal cell we associate an internal point, e.g., $x_{\mathrm{K}} \in \mathrm{K}$. Different choices are possible, which give rise to different versions of the same scheme, such as the barycenters or the arithmetic average of the position vector of cell vertices (also called "iso-barycenters"). For the results shown here, we used the second choice, but apparently there is no significant difference between the two choices mentioned above as far as accuracy and convergence behavior are concerned. The vertices, the

Yves Coudière
Laboratoire de Mathématiques Jean Leray, Nantes, FRANCE, e-mail: Yves.Coudiere@univ-nantes.fr

Florence Hubert
LATP, Université de Provence, Marseille, FRANCE, e-mail: fhubert@cmi.univ-mrs.fr

Gianmarco Manzini
IMATI and CESNA-IUSS, Pavia, ITALY, e-mail: gm.manzini@gmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_95, © Springer-Verlag Berlin Heidelberg 2011

edges, and the faces of mesh $\mathscr{M}$ are denoted by x_{A}, E and F, respectively. Also, we denote the midpoint of E by x_{E} and the barycenter of F by x_{F}. We associate a degree of freedom (the scheme unknowns) with each one of these points; hence, the unknown scalar variable takes the form:

$$u^{\mathscr{T}} = \big((u_{\mathrm{K}})_{K\in\mathscr{M}}, (u_{\mathrm{A}})_{A\in\mathscr{N}}, (u_{\mathrm{E}})_{E\in\mathscr{E}}, (u_{\mathrm{F}})_{F\in\mathscr{F}}\big).$$

We denote the collections of the boundary items (vertices, edges and faces) by $\partial\mathscr{N}$, $\partial\mathscr{F}\mathscr{E}$ and we introduce the set of boundary cells $\partial\mathscr{M}$ which is composed by the boundary faces here considered as degenerated control volumes. Dirichlet boundary conditions are easily introduced into the scheme through the set of boundary data

$$\delta u^{\mathscr{T}} = ((u_{\mathrm{K}})_{x_{\mathrm{K}}\in\partial\mathscr{M}},\ (u_{\mathrm{A}})_{x_{\mathrm{A}}\in\partial\mathscr{N}},\ (u_{\mathrm{E}})_{x_{\mathrm{E}}\in\partial\mathscr{F}\mathscr{E}},\ (u_{\mathrm{F}})_{x_{\mathrm{F}}\in\partial\mathscr{F}\mathscr{E}})\,.$$

The scalar solution field u is approximated by the degrees of freedom $(u^{\mathscr{T}}, \delta u^{\mathscr{T}})$.

The gradient formula is given on each diamond cell $\mathrm{D} \in \mathscr{D}$, which is the convex hull of the points $\mathrm{K}, \mathrm{L}, x_{\mathrm{A}}, x_{\mathrm{B}}, x_{\mathrm{F}}, x_{\mathrm{E}}$, by

$$\nabla^{\mathrm{D}}_{\delta u} u^{\mathscr{T}} = \frac{1}{3|\mathrm{D}|}\big((u_{\mathrm{L}} - u_{\mathrm{K}})N_{\mathrm{KL}} + (u_{\mathrm{B}} - u_{\mathrm{A}})N_{\mathrm{AB}} + (u_{\mathrm{F}} - u_{\mathrm{E}})N_{\mathrm{EF}}\big) \tag{1}$$

using the normal vectors $N_{\mathrm{KL}} = \frac{1}{2}(x_{\mathrm{B}} - x_{\mathrm{A}}) \times (x_{\mathrm{F}} - x_{\mathrm{E}})$, $N_{\mathrm{AB}} = \frac{1}{2}(x_{\mathrm{F}} - x_{\mathrm{E}}) \times (x_{\mathrm{L}} - x_{\mathrm{K}})$ and $N_{\mathrm{EF}} = \frac{1}{2}(x_{\mathrm{L}} - x_{\mathrm{K}}) \times (x_{\mathrm{B}} - x_{\mathrm{A}})$. Gradient formula (1) allows us to define the numerical flux through each interface of the control volumes of the three meshes $\mathscr{M}$, $\mathscr{N}$ and $\mathscr{F}\mathscr{E}$. Let $\mathbf{Q}$ be the linear space of piecewise constant vector fields defined on the mesh of diamonds $\mathscr{D}$ and X be the linear space of triples of piecewise constant scalar fields defined on the three meshes $\mathscr{M}$, $\mathscr{N}$ and $\mathscr{F}\mathscr{E}$. Three finite volume schemes are written by using a discrete divergence operator that maps each vector field in $\mathbf{Q}$ to a triple of scalar functions in X. Formally, we introduce the operator

$$\mathrm{div}^{\mathscr{T}} : \xi = (\xi_{\mathrm{D}})_{\mathrm{D}\in\mathscr{D}} \in \mathbf{Q} \mapsto (\mathrm{div}^{\mathscr{M}}\xi, \mathrm{div}^{\mathscr{N}}\xi, \mathrm{div}^{\mathscr{F}\mathscr{E}}\xi) \in X$$

whose components

$$\mathrm{div}^{\mathscr{M}}\xi = (\mathrm{div}_{\mathrm{K}}\xi)_{\mathrm{K}}, \mathrm{div}^{\mathscr{N}}\xi = (\mathrm{div}_{\mathrm{A}}\xi)_{\mathrm{A}} \text{ and } \mathrm{div}^{\mathscr{F}\mathscr{E}}\xi = \{(\mathrm{div}_{\mathrm{E}}\xi)_{\mathrm{E}}, (\mathrm{div}_{\mathrm{F}}\xi)_{\mathrm{F}}\}$$

are given by

$$|\mathrm{K}|\mathrm{div}_{\mathrm{K}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{K}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{KL}}, \quad |\mathrm{A}|\mathrm{div}_{\mathrm{A}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{A}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{AB}}, \tag{2}$$

$$|\mathrm{E}|\mathrm{div}_{\mathrm{E}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{E}}} \xi_{\mathrm{D}} \cdot N_{\mathrm{EF}}, \quad |\mathrm{F}|\mathrm{div}_{\mathrm{F}}\xi = \sum_{\mathrm{D}\in\mathrm{D}_{\mathrm{F}}} \xi_{\mathrm{D}} \cdot (-N_{\mathrm{EF}})\,. \tag{3}$$

In the previous statements, the symbols D_{K}, D_{A}, D_{E}, D_{F} refer to the diamond cells which overlap the cells labeled by the corresponding subscripted indices K, A, E, and L.

Since each of the $\operatorname{div}_{\mathrm{C}}\xi$ approximates $\frac{1}{|\mathrm{C}|}\int_{\mathrm{C}} \operatorname{div}\xi$ (for $\mathrm{C} = \mathrm{K}, \mathrm{A}, \mathrm{E}, \mathrm{F}$), the right hand side of the discrete problem is given by the piecewise constant projection of the function f onto the space X, $\pi^{\mathscr{T}} f = \{(f_{\mathrm{K}})_{\mathrm{K}\in\mathscr{M}}, (f_{\mathrm{A}})_{\mathrm{A}\in\mathscr{N}}, (f_{\mathrm{E}}, f_{\mathrm{F}})_{\mathrm{E}\in\mathscr{E},\mathrm{F}\in\mathscr{F}}\}$ with $f_{\mathrm{C}} = \frac{1}{|\mathrm{C}|}\int_{\mathrm{C}} f(x)dx$ for any cell $\mathrm{C} = \mathrm{K} \in \mathscr{M}$ or $\mathrm{A} \in \mathscr{N}$ or F or $\mathrm{E} \in \mathscr{F}\mathscr{E}$.

The CeVeFE-DDFV scheme reads:

$$-\operatorname{div}^{\mathscr{T}}(\mathbf{K}_{\mathrm{D}}\nabla^{\mathrm{D}}_{\delta u} u^{\mathscr{T}}) = \pi^{\mathscr{T}} f, \tag{4}$$

where $\mathbf{K}_{\mathrm{D}} = \frac{1}{|\mathrm{D}|}\int_{\mathrm{D}} \mathbf{K}(x)dx$ is a piecewise constant tensor field on the mesh of the diamond cells. The scheme in (4) originates a symmetric and positive-definite linear system of equations (see [1] for a thourough discussion of the other properties). The case of the discontinuous permeability tensor of test 5 deserves a special treatment that is thouroughly discussed in [2].

Mesure on the error

To put the discrete and the exact solutions "at the same level", we use the projection $\pi^{\mathscr{T}} u_e$ of the exact solution and the associated discrete gradient reconstruction $\nabla^{\mathscr{T}}\pi^{\mathscr{T}} u_e$. Approximation errors are evaluated through the following norms:

$$\text{erl2} = \|e^{\mathscr{T}}\|_{L^2}/\|\pi^{\mathscr{T}} u_e\|_{L^2} \text{ with } \|e^{\mathscr{T}}\|^2_{L^2} = \frac{1}{3}\sum_{\mathrm{C}\in\mathscr{M}\cup\mathscr{N}\cup\mathscr{F}\mathscr{E}} |\mathrm{C}||e_{\mathrm{C}}|^2$$

$$\text{ergrad} = \|\nabla^{\mathscr{T}} e^{\mathscr{T}}\|_{L^2}/\|\nabla^{\mathscr{T}}\pi^{\mathscr{T}} u_e\|_{L^2} \text{ with } \|\nabla^{\mathscr{T}} e^{\mathscr{T}}\|^2 = \sum_{\mathrm{D}\in\mathscr{D}} |\mathrm{D}||\nabla^{\mathrm{D}} e^{\mathscr{T}}|^2$$

$$\text{ener} = (\mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}} e^{\mathscr{T}}, \nabla^{\mathscr{T}} e^{\mathscr{T}})_{L^2}/(\mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}}\pi^{\mathscr{T}} u_e, \nabla^{\mathscr{T}}\pi^{\mathscr{T}} u_e)_{L^2}$$

$$\text{with } (\mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}} e^{\mathscr{T}}, \nabla^{\mathscr{T}} e^{\mathscr{T}})_{L^2} = \sum_{\mathrm{D}\in\mathscr{D}} |\mathrm{D}|(\mathbf{K}_{\mathrm{D}}\nabla^{\mathrm{D}} e^{\mathscr{T}}, \nabla^{\mathrm{D}} e^{\mathscr{T}})$$

In the case of the discontinuous tensor of test 5, the diamond cell D is divided in two subdiamond cells, namely, D_{K} and D_{L}. The gradient $\nabla^{\mathrm{D}} u$ is constant on D_{K} (respectively, D_{L}) with value $\nabla^{\mathrm{D}}_{\mathrm{K}} u$ (respectively, $\nabla^{\mathrm{D}}_{\mathrm{L}} u$). The quantities $\|\nabla^{\mathscr{T}} e^{\mathscr{T}}\|^2_{L^2}$ and $(\mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}} e^{\mathscr{T}}, \nabla^{\mathscr{T}} e^{\mathscr{T}})_{L^2}$ become

$$\|\nabla^{\mathscr{T}} e^{\mathscr{T}}\|^2_{L^2} = \sum_{\mathrm{D}\in\mathscr{D}} \Big(|\mathrm{D}_{\mathrm{K}}||\nabla_{\mathrm{D}_{\mathrm{K}}} e^{\mathscr{T}}|^2 + |\mathrm{D}_{\mathrm{L}}||\nabla_{\mathrm{D}_{\mathrm{L}}} e^{\mathscr{T}}|^2\Big)$$

and

$$(\mathbf{K}^{\mathscr{D}}\nabla^{\mathscr{T}} e^{\mathscr{T}}, \nabla^{\mathscr{T}} e^{\mathscr{T}})_{L^2} = \sum_{\mathrm{D}\in\mathscr{D}} \Big(|\mathrm{D}_{\mathrm{K}}|(\mathbf{K}_{\mathrm{D}_{\mathrm{K}}}\nabla_{\mathrm{D}_{\mathrm{K}}} e^{\mathscr{T}}, \nabla_{\mathrm{D}_{\mathrm{K}}} e^{\mathscr{T}}) + |\mathrm{D}_{\mathrm{L}}|(\mathbf{K}_{\mathrm{D}_{\mathrm{L}}}\nabla_{\mathrm{D}_{\mathrm{L}}} e^{\mathscr{T}}, \nabla_{\mathrm{D}_{\mathrm{L}}} e^{\mathscr{T}})\Big).$$

2 Numerical results

The following results were obtained by using a BiCG-stab solver with ILU(0) preconditioner (routine MI26 of HSL implementation).

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	7777	100569	6.09E-03	1.05E-02	1.988	1.980	1.790
2	15495	208527	7.48E-03	9.35E-03	1.995	1.994	1.793
3	31139	431667	3.19E-03	5.93E-03	1.993	1.993	1.795
4	62419	885735	1.48E-03	2.98E-03	1.996	1.996	1.796
5	125993	1823199	1.56E-03	2.28E-03	2.000	1.999	1.797
6	254657	3746829	1.93E-03	2.70E-03	1.999	1.998	1.798

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	7777	0.228E-02	-	0.562E-01	-	0.528E-01	-
2	15495	0.147E-02	1.904	0.441E-01	1.051	0.415E-01	1.054
3	31139	0.916E-03	2.036	0.349E-01	1.011	0.327E-01	1.021
4	62419	0.573E-03	2.025	0.276E-01	1.006	0.258E-01	1.022
5	125993	0.374E-03	1.819	0.219E-01	0.994	0.206E-01	0.969
6	254657	0.231E-03	2.067	0.174E-01	0.983	0.163E-01	0.990

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	345	4559	7.93E-02	1.51E-01	1.875	1.844	1.719
2	933	12811	4.79E-02	4.74E-02	1.989	1.982	1.785
3	2075	29291	5.46E-02	5.64E-02	1.987	1.978	1.794
4	3963	56947	3.25E-02	3.23E-02	2.000	1.996	1.795
5	6909	101229	1.28E-02	3.17E-02	2.000	1.996	1.797

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	345	0.274E-01	-	0.179E+00	-	0.162E+00	-
2	933	0.223E-01	0.622	0.149E+00	0.556	0.139E+00	0.458
3	2075	0.119E-01	2.364	0.102E+00	1.409	0.964E-01	1.373
4	3963	0.819E-02	1.724	0.835E-01	0.933	0.782E-01	0.972
5	6909	0.599E-02	1.694	0.691E-01	1.021	0.655E-01	0.953

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3375	49071	5.67E-02	3.43E-02	1.940	1.974	1.767
2	29791	455895	9.19E-03	7.33E-03	1.988	1.991	1.782
3	250047	3916359	2.42E-03	1.59E-03	1.999	1.998	1.793
4	2048383	32446751	6.52E-04	6.17E-04	2.000	1.999	1.797

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3375	0.287E-01	-	0.481E+00	-	0.589E+00	-
2	29791	0.113E-01	1.289	0.218E+00	1.088	0.233E+00	1.277
3	250047	0.330E-02	1.730	0.904E-01	1.243	0.953E-01	1.260
4	2048383	0.859E-03	1.922	0.395E-01	1.180	0.422E-01	1.161

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	239	2871	8.58E-02	8.40E-02	1.903	1.916	1.795
2	2543	34927	2.90E-02	2.13E-02	1.971	1.979	1.804
3	23135	336735	4.68E-03	5.35E-03	1.995	1.995	1.800
4	196799	2943487	1.69E-03	1.34E-03	1.998	1.999	1.799
5	1622399	24588351	2.88E-04	3.35E-04	2.000	2.000	1.799

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	239	0.307E-01	-	0.141E+00	-	0.139E+00	-
2	2543	0.120E-01	1.190	0.104E+00	0.384	0.101E+00	0.405
3	23135	0.323E-02	1.786	0.571E-01	0.814	0.550E-01	0.827
4	196799	0.830E-03	1.905	0.298E-01	0.909	0.285E-01	0.920
5	1622399	0.210E-03	1.955	0.154E-01	0.937	0.147E-01	0.945

• **Test 2 Heterogeneous anisotropy,** $u(x,y,z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)$ $\sin(2\pi z)$**,** $\min = -0.862,\ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	12179	188089	-8.55E-01	-8.46E-01	1.014	1.009	1.693
2	96759	1545215	-8.55E-01	-8.57E-01	1.026	1.031	1.706
3	325739	5259545	-8.61E-01	-8.59E-01	1.037	1.035	1.708
4	771119	12518433	-8.60E-01	-8.60E-01	1.040	1.041	1.709

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	12179	0.392E-01	-	0.811E-01	-	0.803E-01	-
2	96759	0.109E-01	1.854	0.392E-01	1.054	0.397E-01	1.019
3	325739	0.502E-02	1.917	0.256E-01	1.051	0.261E-01	1.040
4	771119	0.287E-02	1.942	0.190E-01	1.039	0.194E-01	1.034

- **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$**,** $\min = 0,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	343	4447	-4.25E+01	-9.78E-01	49.169	0.931	38.139
2	3375	49855	-2.22E+01	-9.94E-01	21.970	0.982	21.514
3	29791	466111	-6.96E+00	-9.95E-01	7.051	0.993	12.536
4	250047	4019647	-2.67E+00	-9.98E-01	2.725	0.998	7.541

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	343	0.147E+03	-	0.238E+02	-	0.162E+01	-
2	3375	0.956E+01	3.589	0.121E+02	0.892	0.888E+00	0.787
3	29791	0.681E+00	3.640	0.632E+01	0.891	0.459E+00	0.909
4	250047	0.447E-01	3.840	0.314E+01	0.988	0.229E+00	0.979

- **Test 4 Flow around a well, Well meshes,** $\min = 0,\ \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	5868	86728	3.83E-01	4.13E-01	5.317	5.317	1596.292
2	15776	243104	2.37E-01	2.43E-01	5.328	5.328	1611.158
3	36846	580244	1.54E-01	1.54E-01	5.329	5.329	1617.452
4	84546	1350382	1.17E-01	1.18E-01	5.330	5.330	1620.143
5	177590	2860258	8.96E-02	8.98E-02	5.339	5.339	1621.406
6	329236	5329338	7.22E-02	7.23E-02	5.345	5.345	1622.053
7	580190	9422104	5.66E-02	5.64E-02	5.361	5.361	1622.472

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	5868	0.141E-04	-	0.128E+00	-	0.116E+00	-
2	15776	0.476E-05	3.290	0.877E-01	1.144	0.781E-01	1.212
3	36846	0.208E-05	2.924	0.610E-01	1.283	0.542E-01	1.287
4	84546	0.141E-05	1.411	0.466E-01	0.975	0.408E-01	1.033
5	177590	0.914E-06	1.747	0.362E-01	1.021	0.316E-01	1.023
6	329236	0.609E-06	1.976	0.293E-01	1.026	0.258E-01	0.998
7	580190	0.422E-06	1.941	0.244E-01	0.964	0.214E-01	0.971

• **Test 5 Discontinuous permeability,** $u(x, y, z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$**,** $\min = 0,\ \max = 1$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	131	1017	-6.34E+01	-1.00E+02	64.462	100.000	85.763
2	1215	8303	-3.10E+02	-1.00E+02	309.886	100.000	192.379
3	10463	65007	-1.34E+02	-1.00E+02	134.323	100.000	139.345
4	86847	509567	-1.09E+02	-1.00E+02	109.373	100.000	114.251
5	707711	4024287	-1.02E+02	-1.00E+02	102.394	100.000	104.279

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	131	0.218E+01	-	0.450E+00	-	0.406E+00	-
2	1215	0.193E+01	0.159	0.187E+01	-1.917	0.623E+00	-0.578
3	10463	0.862E-01	4.334	0.828E+00	1.134	0.297E+00	1.033
4	86847	0.517E-02	3.989	0.407E+00	1.006	0.147E+00	0.995
5	707711	0.326E-03	3.953	0.203E+00	0.994	0.734E-01	0.994

3 Comments

This finite volume method assigns one degree of freedom to any mesh item (cells, faces, edges, and vertices). For this reason, the scheme has a large number of degrees of freedom if compared to other finite volume methods or similar discretization techniques (such as mimetic finite differences). Nonetheless, the method was proved very effective both for two and three dimensional problems with strong anisotropic coefficients and using meshes with strongly distorted cells. Among the other advantages offered by the method, we mention the coercivity of the method that eases the convergence analysis and the fact that this finite volume method generally shows second order of accuracy in all numerical experiments where the exact solution is sufficiently regular. The results shown in the tables of the previous section confirm this general behavior.

All linear systems were solved efficiently by standard preconditioned Krylov methods as BiCG-stab or GMRES. Direct solvers for general asymmetric systems (UMFPACK) can also be used, but they normally require a huge memory storage, in particular for the biggest problems. In Table 1-2, we see an example of the performance of the different solvers offered by the benchmark site when solving Test 1 on the checkerboard meshes 8 × 8 × 8 and 16 × 16 × 16. The comparison reveals that PETSc implementation of the CG solver is the fastest one, in particular, when combined with the diagonal preconditioner (Jacobi). A good performance in terms of CPU costs is also provided by the ISTL-BiCGstab implementation using Jacobi or ILU(0) preconditioners. CPU times are usually smaller than those obtained by using the direct solver UMFPACK, which is also available in the benchmark site. For example, in the case of 8 × 8 × 8-size mesh we note that UMFPACK requires a CPU time of 3.180 seconds.

Table 1 CeVeFe-DDFV method, test 1 using checkerboard mesh, grid resolution 8 × 8 × 8; CPU times are measured in seconds

solver	precond	CPU time	# iters	Rel. resid.
PETSc-CG	Jacobi	0.209	202	6.368e-11
PETSc-CG	none	0.243	242	1.715e-10
ISTL-BiCGstab	ILU(0)	0.404	53	6.832e-11
ISTL-BiCGstab	none	0.563	167	3.656e-11
ISTL-BiCGstab	Jacobi	0.680	120	4.415e-11
ISTL-GMRES	ILU(0)	0.683	152	4.241e-11

Table 2 CeVeFe-DDFV method, test 1 using checkerboard mesh, grid resolution 16 × 16 × 16; CPU times are measured in seconds

solver	precond	CPU time	# iters	Rel. resid.
PETSc-CG	Jacobi	3.946	369	4.248e-11
PETSc-CG	none	4.989	471	8.038e-11
ISTL-CG	ILU(0)	5.540	166	4.041e-11
ISTL-CG	none	7.319	471	8.038e-11
ISTL-BiCGstab	ILU(0)	8.681	107	3.873e-11
ISTL-CG	Jacobi	10.989	368	4.281e-11

Acknowledgements The work of the first and second author was supported by Groupement de Recherche MOMAS. The work of the third author was partially supported by the Italian MIUR through the program PRIN2008.

References

1. Y. Coudière and F. Hubert. A 3D discrete duality finite volume method for nonlinear elliptic equation. Preprint HAL, URL: http://hal.archives-ouvertes.fr/hal-00456837/fr, 2010.
2. Y. Coudière, F. Hubert, and G. Manzini. A cevefe-ddfv scheme for discontinuous permeability tensors. In *Finite Volume For Complex Applications, Problems And Perspectives. 6th International Conference (this volume)*, 2011.
3. K. Domelevo and P. Omnès. 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. F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, 192(16):1939–1959, 2003.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: The Cell-Centered Finite Volume Method Using Least Squares Vertex Reconstruction (“Diamond Scheme”)

Yves Coudière and Gianmarco Manzini

1 Presentation of the scheme

We consider, for this contribution, the cell-centered finite volume method based on least squares vertex reconstruction. This method, which is also popularly known as *“the diamond scheme”*, was originally presented for the advection-diffusion equation in two-dimensions and then extended in 3-D. The discretization of the diffusive term in 2-D and 3-D is found in in [1, 3–5]. The scalar solution of the diffusion problem u is numerically approximated by a piecewise constant function u_T on the cells K of mesh T. The numerical approximation u_T is defined as $u_T(x) = \sum_{K \in T} u_K \chi_K(x)$ ($\chi_K(x)$ being the characteristic function of cell K) through the values $(u_K)_{K \in T}$. To define the numerical diffusive flux through the interface f of the mesh, a polyhedral cell is built around this interface. This polyhedral cell, which has a quadrilateral shape in two dimensions, is named after its shape as *“diamond cell”*, which also motivates the name of the method. Specifically, let $x_K \in K$ be the center of gravity of the cell K of mesh T. The diamond cell D associated to the interface f between two cells K and L in T is the convex hull $D = \text{hull}(f, x_K, x_L)$. If f is a boundary face, thus defined by $f = \partial K \cap \partial \Omega$ where $\partial \Omega$ is the boundary of the computational domain Ω, then the diamond cell associated to f is the convex hull $D = \text{hull}(f, x_K)$.

The numerical diffusive flux is built by using a constant approximation of the solution gradient on each diamond cell. Let $(x_1, x_2, \dots x_m)$ denote the vertices of face f, x_K and x_L the centers of gravity of the cells K and L that share this face, and D the convex hull of these points. For any function $u \in H^1(D)$, the Green-Gauss formula yields the relation $\int_D \nabla u(x) dx = \int_{\partial D} u(x) n(x) d\sigma(x)$, where n is

Yves Coudière
LMJL, Université de Nantes, France, e-mail: Yves.Coudiere@univ-nantes.fr

Gianmarco Manzini
IMATI-CNR and CESNA-IUSS, Pavia, Italy, e-mail: marco.manzini@imati.cnr.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_96, © Springer-Verlag Berlin Heidelberg 2011

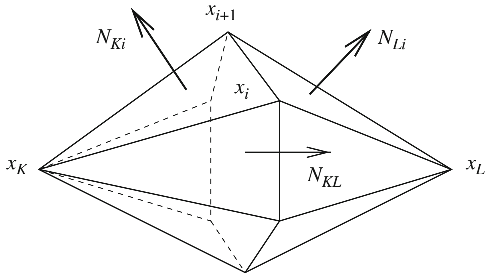


Fig. 1 Geometry of the *diamond* cell

the unit vector orthogonal to ∂D and pointing out of D. If the restriction of u to the face f of ∂D is an affine function, the boundary integral only depends on the values of u at the vertices of D. In this case, the Green-Gauss divergence theorem yields

$$\frac{1}{|D|}\int_D \nabla u(x)dx = \frac{1}{|D|}\int_{\partial D} u(x)n(x)dx = \frac{1}{|D|}\sum_{f\in\partial D}\int_f u(x)n(x)dx$$

$$= \frac{1}{3|D|}\sum_{i=1}^{m}\big(N_{Ki}(u(x_i)+u(x_{i+1})+u(x_K)) + N_{Li}(u(x_i)+u(x_{i+1})+u(x_L))\big)$$

$$= \frac{1}{3|D|}\Big(\big(u(x_L)-u(x_K)\big)\,N_{KL} + \sum_{i=1}^{m} u(x_i)N_i\Big),$$

where N_{Ki} and N_{Li} are the vectors orthogonal to the triangular facets hull(x_K, x_i, x_{i+1}) and hull(x_L, x_i, x_{i+1}), respectively, and having lengths equal to the measure of the facets; specifically, $N_{Ki} = \frac{1}{2}(x_i - x_K)\times(x_{i+1}-x_K)$ and $N_{Li} = -\frac{1}{2}(x_i - x_L)\times(x_{i+1}-x_L)$, see Fig. 1. The vectors N_{KL} and N_i actually used during the computations are

$$N_{KL} = \sum_{i=1}^{m} N_{Li} = -\sum_{i=1}^{m} N_{Ki} \quad \text{and} \quad N_i = N_{Ki} + N_{Li} + N_{Ki-1} + N_{Li-1}.$$

To derive the gradient formula, we consider the right-hand side of the above equality with the unknown u_K and u_L replacing $u(x_K)$ and $u(x_L)$ and some values u_i replacing the values $u(x_i)$. These values, $u_1, u_2, \dots u_m$, are linearly interpolated from the values $(u_K)_{K\in T}$ as follows. For any vertex x_i of the mesh T, we consider $u_i = \sum_{K\in T_i}\omega_{iK}u_K$ where $T_i = \{K : x_i \in K\}$ denotes the subset of the mesh cells which share the vertex x_i. The interpolation weights ω_{iK} are assumed to verify the consistency relations [4]:

$$\sum_{K\in T_i}\omega_{iK} = 1 \quad \text{and} \quad \sum_{K\in T_i}\omega_{iK}(x_i - x_K) = 0.$$

The interpolation weights ω_{iK} are obtained by solving the reconstruction problem that approximates the cell-averaged data set $\{(x_K, u_K)$ for $K \in T_i\}$ by the affine function

$$\tilde{u}_i(x) = \alpha + \beta \cdot (x - x_i) \ \text{ for } x \in \mathscr{V}_i$$

on the co-volume $\mathscr{V}_i = \bigcup_{K \in T_i} K$ and in a least square sense, cf. [2, 4]. The reconstructed value at vertex x_i is now given by taking $u_i = \tilde{u}_i(x_i) = \alpha$. The coefficients $(\alpha, \beta)^T$ are the minimizers of the least squares functional

$$\mathscr{J}(\alpha, \beta^T) = \sum_{K:x_i \in K} \big(\alpha + \beta \cdot (x - x_i) - u_K\big)^2.$$

Imposing the zero gradient condition, i.e., $\nabla_{\alpha,\beta}\mathscr{J}(\alpha, \beta^T) = 0$, yields a linear system for the coefficients (α, β), whose solution returns the interpolation weights. The values u_i at the vertices $x_i \in \partial\Omega$ on the Dirichlet boundary are constrained to the boundary data, for instance $u_i = 0$ for a homogeneous condition. Other kinds of boundary conditions, e.g., Neumann or Robin, can be taken into account by extending to the 3-D case the technique investigated in [2]. Finally, the scheme reads as

$$\forall K \in T, \quad -\sum_{f \subset \partial K} \Lambda_f \nabla_D u_T \cdot N_{KL} = f_K|K| := \int_K f(x)dx,$$

where Λ_f is an arithmetic average of the diffusion tensor Λ over the diamond cell located around face f and N_{KL} is exactly the normal from above.

2 Numerical results

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
0	215	6985	3.02E-02	3.15E-02	1.949	1.948	1.627
1	2003	107331	2.03E-02	1.13E-02	1.989	1.995	1.730
2	3898	227618	6.84E-03	4.21E-03	1.989	1.990	1.750
3	7711	476645	9.13E-03	8.18E-03	1.994	1.995	1.767
4	15266	994892	5.52E-03	4.10E-03	1.997	1.997	1.776
5	30480	2072944	1.49E-03	2.57E-04	1.997	1.999	1.784
6	61052	4292073	1.83E-03	1.20E-03	1.997	1.998	1.789

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
0	215	3.750E-02	-	3.503E-01	-	2.636E-01	-
1	2003	9.173E-03	1.892	1.568E-01	1.081	1.071E-01	1.210
2	3898	5.897E-03	1.991	1.215E-01	1.149	8.159E-02	1.225
3	7711	3.551E-03	2.230	9.410E-02	1.122	6.016E-02	1.339
4	15266	2.255E-03	1.994	7.387E-02	1.063	4.648E-02	1.132
5	30480	1.412E-03	2.032	5.768E-02	1.073	3.565E-02	1.152
6	61052	8.882E-04	2.001	4.502E-02	1.070	2.733E-02	1.147

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	29	257	8.51E-02	-6.18E+00	1.870	7.968	20.435
2	66	660	1.43E-01	2.06E-01	1.854	1.846	1.887
3	130	1410	3.85E-02	7.00E-03	1.925	1.941	1.855
4	228	2620	1.74E-02	2.37E-02	1.914	1.920	2.067
5	356	4424	2.84E-03	-2.18E+00	1.979	3.546	3.274

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	29	2.639E+00	-	3.631E+01	-	9.728E+00	-
2	66	9.077E-02	12.292	9.457E-01	13.307	3.751E-01	11.876
3	130	5.508E-02	2.210	7.434E-01	1.065	3.055E-01	0.907
4	228	6.650E-02	-1.006	1.163E+00	-2.391	3.224E-01	-0.287
5	356	3.674E-01	-11.507	5.071E+00	-9.912	1.287E+00	-9.321

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	512	10648	3.03E-02	8.74E-02	1.958	1.916	1.768
2	4096	97336	1.06E-02	3.00E-02	1.993	1.973	1.700
3	32768	830584	1.75E-03	5.87E-03	1.997	1.991	1.726
4	262144	6859000	7.14E-04	9.88E-04	1.999	1.998	1.765

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	512	6.846E-02	-	6.798E-01	-	4.901E-01	-
2	4096	4.715E-02	0.537	3.403E-01	0.998	2.715E-01	0.851
3	32768	2.866E-02	0.718	1.831E-01	0.894	1.532E-01	0.825
4	262144	1.315E-02	1.123	8.289E-02	1.143	6.942E-02	1.142

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	36	424	1.54E-01	1.33E-01	1.846	1.833	1.588
2	288	4528	4.01E-02	3.47E-02	1.960	1.958	1.721
3	2304	41896	1.01E-02	8.74E-03	1.990	1.990	1.773
4	18432	360280	2.54E-03	1.98E-03	1.997	1.998	1.791
5	147456	2987704	6.36E-04	5.26E-04	1.999	1.999	1.796

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	36	1.356E-01	-	2.488E-01	-	3.406E-01	-
2	288	4.427E-02	1.615	1.471E-01	0.758	1.346E-01	1.339
3	2304	1.191E-02	1.894	7.031E-02	1.065	4.678E-02	1.524
4	18432	3.112E-03	1.936	3.410E-02	1.043	1.687E-02	1.471
5	147456	7.976E-04	1.964	1.695E-02	1.008	7.003E-03	1.268

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

• **Test 2 Heterogeneous anisotropy,** $u(x, y, z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)$ $\sin(2\pi z)$**,** $\min = -0.862,\ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1210	21308	-8.42E-01	-8.57E-01	0.978	0.977	1.481
2	8820	169418	-8.38E-01	-8.41E-01	1.010	1.011	1.638
3	28830	570328	-8.58E-01	-8.60E-01	1.032	1.033	1.676
4	67240	1350038	-8.57E-01	-8.58E-01	1.033	1.034	1.690

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1210	9.551E-02	-	2.356E-01	-	2.404E-01	-
2	8820	2.403E-02	2.084	8.174E-02	1.598	7.974E-02	1.666
3	28830	1.067E-02	2.057	4.167E-02	1.706	3.947E-02	1.781
4	67240	6.013E-03	2.030	2.562E-02	1.722	2.371E-02	1.805

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

• **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = 0, \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	64	1000	-7.56E-01	-7.11E-01	0.711	0.525	1.650
2	512	10648	-9.39E-01	-8.32E-01	0.926	0.933	2.674
3	4096	97336	-9.86E-01	-9.77E-01	0.982	0.978	3.330
4	32768	830584	-9.96E-01	-9.92E-01	0.996	0.990	3.527

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	64	5.548E-01	-	7.060E-01	-	7.651E-01	-
2	512	1.427E-01	1.958	3.067E-01	1.202	3.264E-01	1.229
3	4096	2.967E-02	2.266	9.532E-02	1.685	9.569E-02	1.770
4	32768	7.166E-03	2.049	3.253E-02	1.551	2.529E-02	1.919

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

• **Test 4 Flow around a well, Well meshes,** $\min = 0, \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	890	18876	4.57E-01	5.26E-01	5.317	5.318	1573.020
2	2232	51800	2.61E-01	2.89E-01	5.329	5.329	1600.780
3	5016	121584	1.62E-01	1.73E-01	5.329	5.329	1613.840
4	11220	280868	1.23E-01	1.29E-01	5.330	5.330	1619.520
5	23210	592448	9.28E-02	9.66E-02	5.339	5.339	1620.960
6	42633	1100865	7.42E-02	7.67E-02	5.345	5.345	1621.200
7	74679	1942619	5.75E-02	5.91E-02	5.361	5.361	1621.930

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	890	9.562E-03	-	8.767E-02	-	5.372E-02	-
2	2232	3.699E-03	3.098	3.903E-02	2.640	2.305E-02	2.761
3	5016	1.676E-03	2.932	1.916E-02	2.636	1.104E-02	2.727
4	11220	1.190E-03	1.275	1.270E-02	1.531	7.205E-03	1.589
5	23210	7.545E-04	1.882	8.919E-03	1.458	5.053E-03	1.463
6	42633	4.601E-04	2.439	6.148E-03	1.835	3.522E-03	1.781
7	74679	3.402E-04	1.616	5.400E-03	0.693	3.174E-03	0.556

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

• **Test 5 Discontinuous permeability,** $u(x, y, z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$**,** $\min = 0,\ \max = 1$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	22	252	-1.00E+02	-5.24E+01	100.000	52.359	58.097
2	176	3220	-3.54E+01	-2.62E+01	35.355	26.180	43.055
3	1408	31524	-7.89E+01	-7.30E+01	78.858	73.021	76.757
4	11264	277396	-9.43E+01	-9.25E+01	94.346	92.545	92.422
5	90112	2324532	-9.86E+01	-9.81E+01	98.562	98.089	97.247

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	22	9.831E-01	-	7.196E-01	-	3.176E+02	-
2	176	5.072E-01	0.954	7.376E-01	-0.035	8.184E-01	8.600
3	1408	1.376E-01	1.882	3.770E-01	0.968	6.058E-01	0.433
4	11264	3.347E-02	2.039	1.874E-01	1.008	4.685E-01	0.370
5	90112	9.731E-03	1.782	1.159E-01	0.693	3.448E-01	0.442

Name of the solver: BiCG-stab with Jacobi preconditioner (in-house implementation).

3 Comments

This finite volume method is truly cell-centered and, for this reason, it has a relatively small number of degrees of freedom with respect to other finite volume discretizations which introduce face unknowns to approximate the scalar variable. The coercivity was proved only for simple cases (see [3] for details), so very few can be said from a theoretical standpoint about the convergence properties of this scheme and the literature misses a general convergence analysis. Despite this fact, the resulting finite volume method generally show second order of accuracy in all numerical experiments where the exact solution is enough regular and on

Table 1 LS-FVM method, test 1 using Kershaw mesh with grid resolution 32 × 32 × 32; CPU times are measured in seconds.

solver	precond	CPU time	# iters	Rel. resid.
UMFPACK	none	28.712	0	7.287e-15
ISTL-BiCGstab	Jacobi	36.048	2990	1.126e-10
ISTL-GMRES	Jacobi	64.775	9186	3.931e-10
ISTL-BiCGstab	none	78.880	11417	3.431e-10
ISTL-BiCGstab	ILU(4)	1888.48	2	1.191e-12
ISTL-GMRES	ILU(4)	1692.49	4	9.103e-10

"reasonable" meshes. Moreover, it can be easily applied to complex, distorted meshes and anisotropic permeabilities for which it provides a reliable numerical approximation. It is also generally robust even if a locking phenomenon for the convergence has been reported in the literature [6].

The linear system for the cell-centered unknowns that is originated by this scheme on a general polyhedral mesh leads to an *asymmetric sparse matrix*. Therefore, this system can be solved efficiently by standard preconditioned Krylov methods (BiCG-stab or GMRES) or by direct solvers for general asymmetric systems (UMFPACK). An example of a typical behavior is reported in Table 1 for a subset of the combinations solvers and preconditioners available on the benchmark site. The comparison among these results reveals that the BiCG-stab solver using a diagonal Jacobi preconditioner seems to be the more efficient choice in most of the cases. The performance is usually comparable with that offered by the direct solver (UMFPACK), but the memory storage required by this latter may be from 2 to 60 times greater.

References

1. Bertolazzi, E., Manzini, G.: A cell-centered second-order accurate finite volume method for convection-diffusion problems on unstructured meshes. Math. Models Methods Appl. Sci. **8**, 1235–1260 (2004)
2. Bertolazzi, E., Manzini, G.: On vertex reconstructions for cell-centered finite volume approximations of 2-D anisotropic diffusion problems. Math. Models Methods Appl. Sci. **17**(1), 1–32 (2007)
3. Coudière, Y.: Analyse de schémas volumes finis sur maillages non structurés pour des problèmes linéaires hyperboliques et elliptiques. Ph.D. thesis, Universitè "P. Sabatier" de Toulouse, Toulouse III, Toulouse, France (1999)
4. Coudière, Y., Vila, J.P., Villedieu, P.: Convergence rate of a finite volume scheme for a two-dimensional diffusion convection problem. M2AN, Math. Model. Numer. Anal. **33**(3), 493–516 (1999)
5. Coudière, Y., Villedieu, P.: Convergence of a finite volume scheme for the linear convection-diffusion equation on locally refined meshes. M2AN, Math. Model. Numer. Anal. **34**(6), 1123–1149 (2000)
6. Manzini, G., Putti, M.: Mesh locking effects in the finite volume solution of 2-D anisotropic diffusion equations. J. Comput. Phys. **220**(2), 751–771 (2007)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: A Monotone Nonlinear Finite Volume Method for Diffusion Equations on Polyhedral Meshes

Alexander Danilov and Yuri Vassilevski

1 Presentation of the scheme

We propose a new monotone FV method based on a nonlinear two-point flux approximation scheme. The original idea belongs to C. LePotier [2] who proposed a monotone FV scheme for the discretization of parabolic equations on triangular meshes, which was extended to steady-state diffusion problems with full anisotropic tensors on triangulations or scalar diffusion coefficients on shape regular polygonal meshes [3]. Later a new interpolation-free monotone cell-centered FV method with nonlinear two-point flux approximation was proposed for full diffusion tensors and unstructured conformal polygonal 2D meshes [4]. In this paper, we extend the last approach to the case of 3D conformal polyhedral meshes [1].

Let Ω be a three-dimensional polyhedral domain with boundary Γ. We consider a model diffusion problem for unknown concentration u:

$$\begin{aligned} -\mathrm{div}(\mathbb{K}\nabla u) &= g \quad \text{in} \quad \Omega \\ u &= g_D \quad \text{on} \quad \Gamma \end{aligned} \tag{1}$$

where $\mathbb{K}(\mathbf{x}) = \mathbb{K}^T(\mathbf{x}) > 0$ is an anisotropic diffusion tensor, and g is a source term.

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, and each face is a star-shaped 2D domain with respect to face's barycenter. Let $N_{\mathscr{T}}$ be the number of polyhedral cells and $N_{\mathscr{B}}$ be the number of boundary faces. The tensor function $\mathbb{K}(\mathbf{x})$ is assumed to be smooth for

Alexander Danilov and Yuri Vassilevski
Institute of Numerical Mathematics, Gubkina 8, Moscow, 119333, Russia, e-mail: a.a.danilov@gmail.com, yuri.vassilevski@gmail.com

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_97, © Springer-Verlag Berlin Heidelberg 2011

the sake of simplicity of presentation; however the original method is designed for discontinuous tensor function, which may jump across mesh faces as well as may change orientation of principal directions [1].

We denote by $\mathscr{F}_I$, $\mathscr{F}_B$ disjoint sets of interior and boundary faces, respectively. The cardinality of set $\mathscr{F}_*$ is denoted by $N_{\mathscr{F}_*}$. Let $\mathscr{F}_T$ and $\mathscr{E}_T$ denote the sets of faces and edges of polyhedron T, respectively.

Let $\mathbf{q} = -\mathbb{K}\nabla u$ denote the flux which satisfies the mass balance equation:

$$\operatorname{div} \mathbf{q} = g \quad \text{in} \quad \Omega. \tag{2}$$

We derive a FV scheme with a nonlinear two-point flux approximation. Integrating equation (2) over a polyhedron T and using the Green's formula we get:

$$\int_{\partial T} \mathbf{q} \cdot \mathbf{n}_T \,\mathrm{d}s = \int_T g \,\mathrm{d}x, \tag{3}$$

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 equation (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.

For each cell T, we assign one degree of freedom, U_T, for concentration u. Let U be the vector of all unknown concentrations. If two cells T_+ and T_- have a common face f, the two-point flux approximation is as follows:

$$\mathbf{q}_f^h \cdot \mathbf{n}_f = M_f^+ U_{T_+} - M_f^- U_{T_-}, \tag{5}$$

where M_f^+ and M_f^- are some coefficients. In a linear FV method, these coefficients are equal and fixed. In the nonlinear FV method, they may be different and depend on concentrations in surrounding cells.

For every cell T in $\mathscr{T}$, we define the collocation point $\mathbf{x}_T$ at the barycenter of T. For every face $f \in \mathscr{F}_B$, 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. We assume that for every cell-face pair $T \in \mathscr{T}$, $f \in \mathscr{F}_T$, there exist three points $\mathbf{x}_{f,1}$, $\mathbf{x}_{f,2}$, and $\mathbf{x}_{f,3}$ in set Σ_T such that the following condition holds:

The co-normal vector $\boldsymbol{\ell}_f = \mathbb{K}(\mathbf{x}_f)\mathbf{n}_f$ *started from* $\mathbf{x}_T$ *belongs to the trihedral corner formed by vectors*

$$\mathbf{t}_{f,1} = \mathbf{x}_{f,1} - \mathbf{x}_T, \quad \mathbf{t}_{f,2} = \mathbf{x}_{f,2} - \mathbf{x}_T, \quad \mathbf{t}_{f,3} = \mathbf{x}_{f,3} - \mathbf{x}_T, \tag{6}$$

and

$$\frac{1}{|\boldsymbol{\ell}_f|}\boldsymbol{\ell}_f = \frac{\alpha_f}{|\mathbf{t}_{f,1}|}\mathbf{t}_{f,1} + \frac{\beta_f}{|\mathbf{t}_{f,2}|}\mathbf{t}_{f,2} + \frac{\gamma_f}{|\mathbf{t}_{f,3}|}\mathbf{t}_{f,3}, \tag{7}$$

where $\alpha_f \geq 0$*,* $\beta_f \geq 0$*,* $\gamma_f \geq 0$.

Let f be an internal face. We denote by T_+ and T_- the cells that share f and assume that $\mathbf{n}_f$ is outward for T_+. Let $\mathbf{x}_\pm$ be the collocation point of $T_\pm$. Let $U_\pm$ be the discrete concentrations in $T_\pm$.

Let $T = T_+$ and $\mathbb{K}_f = \mathbb{K}(\mathbf{x}_f)$. Using the above notations, definition of the directional derivative,

$$\frac{\partial u}{\partial \boldsymbol{\ell}_f}|\boldsymbol{\ell}_f| = \nabla u \cdot (\mathbb{K}_f\,\mathbf{n}_f),$$

and assumption (7), we write

$$\mathbf{q}_f \cdot \mathbf{n}_f = -\frac{|\boldsymbol{\ell}_f|}{|f|}\int_f \frac{\partial u}{\partial \boldsymbol{\ell}_f}\,\mathrm{d}s = -\frac{|\boldsymbol{\ell}_f|}{|f|}\int_f \left(\alpha_f \frac{\partial u}{\partial \mathbf{t}_{f,1}} + \beta_f \frac{\partial u}{\partial \mathbf{t}_{f,2}} + \gamma_f \frac{\partial u}{\partial \mathbf{t}_{f,3}}\right)\mathrm{d}s. \tag{8}$$

Replacing directional derivatives by finite differences, we get

$$\int_f \frac{\partial u}{\partial \mathbf{t}_{f,i}}\,\mathrm{d}s = \frac{U_{f,i} - U_T}{|\mathbf{x}_{f,i} - \mathbf{x}_T|}\,|f| + O(h_T^2), \quad i = 1, 2, 3, \tag{9}$$

where h_T is the diameter of cell T. Using the finite difference approximations (9), we transform formula (8) to

$$\mathbf{q}_f^h \cdot \mathbf{n}_f = -|\boldsymbol{\ell}_f|\left(\frac{\alpha_f}{|\mathbf{t}_{f,1}|}(U_{f,1} - U_T) + \frac{\beta_f}{|\mathbf{t}_{f,2}|}(U_{f,2} - U_T) + \frac{\gamma_f}{|\mathbf{t}_{f,3}|}(U_{f,3} - U_T)\right). \tag{10}$$

At the moment, this flux involves four rather than two concentrations. To derive a two-point flux approximation, we consider the cell T_- and derive another approximation of flux through face f. To distinguish between T_+ and T_-, we add subscripts ± and omit subscript f. Since $\mathbf{n}_f$ is the internal normal vector for T_-, we have to change sign of the right hand side:

$$\mathbf{q}_\pm^h \cdot \mathbf{n}_f = \mp|\boldsymbol{\ell}_f|\left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}|}(U_{\pm,1} - U_\pm) + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|}(U_{\pm,2} - U_\pm) + \frac{\gamma_\pm}{|\mathbf{t}_{\pm,3}|}(U_{\pm,3} - U_\pm)\right), \tag{11}$$

where $\alpha_\pm$, $\beta_\pm$ and $\gamma_\pm$ are given by (7) and $U_{\pm,i}$ denote concentrations at points $\mathbf{x}_{\pm,i}$ from $\Sigma_{T_\pm}$. We define a new discrete flux as a linear combination of $\mathbf{q}_\pm^h \cdot \mathbf{n}_f$ with non-negative weights $\mu_\pm$:

$$\begin{aligned}\mathbf{q}^h_f \cdot \mathbf{n}_f &= \mu_+ \mathbf{q}^h_+ \cdot \mathbf{n}_f + \mu_- \mathbf{q}^h_- \cdot \mathbf{n}_f \\ &= \mu_+ |\boldsymbol{\ell}_f| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} + \frac{\gamma_+}{|\mathbf{t}_{+,3}|} \right) U_+ \\ &\quad -\mu_- |\boldsymbol{\ell}_f| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} + \frac{\gamma_-}{|\mathbf{t}_{-,3}|} \right) U_- \\ &\quad -\mu_+ |\boldsymbol{\ell}_f| \left(\frac{\alpha_+}{|\mathbf{t}_{+,1}|} U_{+,1} + \frac{\beta_+}{|\mathbf{t}_{+,2}|} U_{+,2} + \frac{\gamma_+}{|\mathbf{t}_{+,3}|} U_{+,3} \right) \\ &\quad +\mu_- |\boldsymbol{\ell}_f| \left(\frac{\alpha_-}{|\mathbf{t}_{-,1}|} U_{-,1} + \frac{\beta_-}{|\mathbf{t}_{-,2}|} U_{-,2} + \frac{\gamma_-}{|\mathbf{t}_{-,3}|} U_{-,3} \right).\end{aligned} \tag{12}$$

The obvious requirement for the weights is to cancel the terms in the last two rows of (12) which results in a two-point flux formula. The second requirement is to approximate the true flux. These requirements lead us to the following system

$$\begin{cases} -\mu_+ d_+ + \mu_- d_- = 0, \\ \mu_+ + \mu_- = 1, \end{cases} \tag{13}$$

where

$$d_\pm = |\boldsymbol{\ell}_f| \left(\frac{\alpha_\pm}{|\mathbf{t}_{\pm,1}|} U_{\pm,1} + \frac{\beta_\pm}{|\mathbf{t}_{\pm,2}|} U_{\pm,2} + \frac{\gamma_\pm}{|\mathbf{t}_{\pm,3}|} U_{\pm,3} \right).$$

Since coefficients $d_\pm$ depend on both geometry and concentration, the weights $\mu_\pm$ do as well. Thus, the resulting two-point flux approximation is *nonlinear*.

It may happen that concentration $U_{+,i}$, $(U_{-,i})$ $i = 1, 2, 3$, is defined at the same collocation point as U_- (U_+). In this case the terms to be cancelled are changed so that they do not incorporate $U_\pm$. By doing so, for the Laplace operator we recover the classical linear scheme with the 6-1-1-1-1-1-1 stencil on uniform cubic meshes.

The solution of (13) can be written explicitly. In all cases $d_\pm \geq 0$ if $U \geq 0$. If $d_\pm = 0$, we set $\mu_+ = \mu_- = \frac{1}{2}$. Otherwise,

$$\mu_+ = \frac{d_-}{d_- + d_+} \qquad \text{and} \qquad \mu_- = \frac{d_+}{d_- + d_+}.$$

This implies that the weights $\mu_\pm$ are non-negative. Substituting this into (12), we get the two-point flux formula (5) with coefficients

$$M^\pm_f = \mu_\pm |\boldsymbol{\ell}_f| (\alpha_\pm / |\mathbf{t}_{\pm,1}| + \beta_\pm / |\mathbf{t}_{\pm,2}| + \gamma_\pm / |\mathbf{t}_{\pm,3}|). \tag{14}$$

Now we consider the case of Dirichlet boundary face $f \in \mathscr{F}_B$ where we define

$$U_f = \bar{g}_{D,f} = \frac{1}{|f|} \int_f g_D \, \mathrm{d}s. \tag{15}$$

It may be convenient to think about f as the ghost cell with zero volume. Let T be the cell with face f. Replacing U_+ and U_- with U_T and U_f, and Σ_{T+}, Σ_{T-} with Σ_T, $\Sigma_{f,T}$ respectively, we get

$$\mathbf{q}_f^h \cdot \mathbf{n}_f = M_f^+ U_T - M_f^- U_f, \tag{16}$$

where coefficients $M_f^\pm$ are given by (14).

For every T in $\mathscr{T}$, the cell equation (4) is

$$\sum_{f \in \mathscr{F}_T} \chi(T, f)\, \mathbf{q}_f^h \cdot \mathbf{n}_f = \int_T f \,\mathrm{d}x, \tag{17}$$

where $\chi(T, f) = sign(\mathbf{n}_f \cdot \mathbf{n}_T(\mathbf{x}_f))$. Substituting two-point flux formula (5) with non-negative coefficients given by (14) into (17), and using equations (15) and (16) to eliminate concentrations at boundary faces, we get a nonlinear system of $N_{\mathscr{T}}$ equations

$$\mathbb{M}(U)U = G(U). \tag{18}$$

The right hand side vector $G(U)$ is generated by the source and the boundary data:

$$G_T(U) = \int_T g \,\mathrm{d}x + \sum_{f \in \mathscr{F}_B \cap \mathscr{F}_T} M_f^-(U)\bar{g}_{D,f}, \qquad \forall T \in \mathscr{T}. \tag{19}$$

For data functions $g \geq 0$ and $g_D \geq 0$ the components of vector G are non-negative. We use the Picard iterations to solve the nonlinear system (18).

The details of the presented scheme, algorithms, modifications of the scheme for Neumann boundary conditions and discontinuous diffusion tensor coefficients, as well as monotonicity analysis of the scheme are presented in [1].

2 Numerical results

We use discrete L_2-norm to evaluate discretization errors for the concentration u:

$$\text{erl2} = \left[\frac{\sum_{T \in \mathscr{T}} (u(\mathbf{x}_T) - U_T)^2 \, |T|}{\sum_{T \in \mathscr{T}} (u(\mathbf{x}_T))^2 \, |T|} \right]^{1/2}.$$

For each cell T we derive the value of ∇u from the linear reconstruction of the concentration over T introduced in [5]:

$$\mathscr{R}_T(\mathbf{x}) = U_T + \nabla U_T(\mathbf{x} - \mathbf{x}_T).$$

This reconstruction minimizes the deviation of $\mathscr{R}_T(\mathbf{x}_k)$ from targeted values U_k in nearby cells. We denote the approximate gradient in cell T as ∇U_T.

We use the following estimate for L_1-norm of the euclidean norm of the approximate gradient:

$$\text{normg} = \sum_{T\in\mathscr{T}} ||\nabla U_T||\,|T|.$$

The discrete H^1 and energy norms of the error are defined as follows:

$$\text{ergrad} = \left[\frac{\sum\limits_{T\in\mathscr{T}} ||\nabla u(\mathbf{x}_T) - \nabla U_T||^2|T|}{\sum\limits_{T\in\mathscr{T}} ||\nabla u(\mathbf{x}_T)||^2|T|}\right]^{1/2}$$

$$\text{ener} = \left[\frac{\sum\limits_{T\in\mathscr{T}} \mathbb{K}(\nabla u(\mathbf{x}_T) - \nabla U_T)\cdot(\nabla u(\mathbf{x}_T) - \nabla U_T)\,|T|}{\sum\limits_{T\in\mathscr{T}} K\nabla u(\mathbf{x}_T)\cdot\nabla u(\mathbf{x}_T)\,|T|}\right]^{1/2}$$

The proposed method is designed for non-negative solutions, and may behave unexpectedly if the values of solution go below zero. Since several test cases have negative values of exact solution, we added positive constants to exact solutions to force their positivity. We added +1 in tests 2 and 3, and +100 in test 5. We subtracted back the positive constants at the end of test runs. We denote minimum and maximum values for discrete solution as umin and umax respectively, and exact values of u at the cell centers as uemin and uemax respectively.

We use Picard method to solve the nonlinear system (18). The values nu and nmat correspond to the number of unknowns and number of non-zero terms in linearized system.

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0$, $\max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	2003	9411	0.028	0.020	1.997	1.989	1.790
2	3898	18586	0.014	0.007	1.992	1.989	1.794
3	7711	37103	0.014	0.009	1.997	1.994	1.795
4	15266	74012	0.008	0.006	1.998	1.997	1.797
5	30480	148746	0.004	0.001	1.999	1.997	1.797
6	61052	299492	0.003	0.002	1.998	1.997	1.798

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	2003	5.31e-03		1.30e-01		1.26e-01	
2	3898	4.00e-03	1.281	1.07e-01	0.879	1.04e-01	0.858
3	7711	2.44e-03	2.167	8.47e-02	1.039	8.16e-02	1.057
4	15266	1.70e-03	1.592	6.78e-02	0.979	6.52e-02	0.986
5	30480	9.57e-04	2.490	5.36e-02	1.016	5.20e-02	0.986
6	61052	6.42e-04	1.726	4.23e-02	1.026	4.09e-02	1.037

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	29	257	0.011	0.085	1.991	1.870	1.303
2	66	660	0.107	0.143	1.902	1.854	1.614
3	130	1410	0.038	0.038	1.963	1.925	1.609
4	228	2620	0.021	0.017	1.941	1.914	1.685
5	356	4424	0.002	0.003	2.004	1.979	1.686

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	29	7.49e-02		6.70e-01		6.81e-01	
2	66	6.16e-02	0.710	4.96e-01	1.099	4.66e-01	1.384
3	130	3.44e-02	2.579	3.65e-01	1.351	3.70e-01	1.023
4	228	2.32e-02	2.098	2.78e-01	1.470	2.72e-01	1.636
5	356	1.73e-02	1.988	2.27e-01	1.364	2.23e-01	1.341

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	512	3200	0.112	0.030	1.942	1.958	1.695
2	4096	27136	0.037	0.011	1.977	1.993	1.763
3	32768	223232	0.011	0.002	1.989	1.997	1.749
4	262144	1810432	0.003	0.001	1.997	1.999	1.761

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	512	6.02e-02		4.53e-01		4.27e-01	
2	4096	4.98e-02	0.276	5.55e-01	-0.294	6.25e-01	-0.548
3	32768	3.70e-02	0.428	3.45e-01	0.687	3.74e-01	0.740
4	262144	2.22e-02	0.737	1.83e-01	0.910	1.89e-01	0.988

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	36	228	0.122	0.154	1.905	1.846	1.769
2	288	2208	0.053	0.040	1.966	1.960	1.735
3	2304	19200	0.014	0.010	1.992	1.990	1.772
4	18432	159744	0.005	0.003	1.998	1.997	1.790
5	147456	1302528	0.001	0.001	2.000	1.999	1.796

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	36	7.24e-02		5.48e-01		5.21e-01	
2	288	2.83e-02	1.356	2.77e-01	0.988	2.68e-01	0.959
3	2304	7.82e-03	1.854	1.21e-01	1.192	1.17e-01	1.190
4	18432	2.20e-03	1.829	5.48e-02	1.142	5.34e-02	1.139
5	147456	6.49e-04	1.761	2.59e-02	1.085	2.52e-02	1.084

- **Test 2 Heterogeneous anisotropy,** $u(x,y,z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)$ $\sin(2\pi z)$**,** $\min = -0.862,\ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1210	9788	-0.854	-0.842	1.002	0.978	1.579
2	8820	75178	-0.840	-0.838	1.014	1.010	1.669
3	28830	250168	-0.859	-0.858	1.034	1.032	1.689
4	67240	588758	-0.858	-0.857	1.034	1.033	1.698

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1210	6.07e-02		2.11e-01		2.11e-01	
2	8820	1.69e-02	1.928	8.51e-02	1.368	8.62e-02	1.352
3	28830	7.96e-03	1.911	4.63e-02	1.543	4.74e-02	1.517
4	67240	4.62e-03	1.933	2.93e-02	1.626	3.02e-02	1.597

- **Test 3 Flow on random meshes,** $u(x,y,z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	64	352	-0.905	-0.778	0.759	0.702	2.241
2	512	3200	-0.928	-0.937	0.959	0.930	3.167
3	4096	27136	-1.005	-0.985	0.996	0.982	3.492
4	32768	223232	-0.989	-0.996	1.001	0.996	3.568

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	64	2.83e-01		5.55e-01		5.50e-01	
2	512	8.74e-02	1.698	1.81e-01	1.618	1.63e-01	1.756
3	4096	2.71e-02	1.688	7.01e-02	1.366	5.61e-02	1.537
4	32768	7.58e-03	1.839	3.23e-02	1.118	2.39e-02	1.230

- **Test 4 Flow around a well, Well meshes,** min $= 0$, max $= 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	890	5574	0.518	0.458	5.318	5.317	1484.035
2	2232	14552	0.287	0.262	5.329	5.329	1541.433
3	5016	33436	0.173	0.162	5.329	5.329	1577.828
4	11220	75894	0.129	0.123	5.330	5.330	1596.743
5	23210	158380	0.096	0.093	5.339	5.339	1606.920
6	42633	292465	0.077	0.074	5.345	5.345	1611.935
7	74679	514069	0.059	0.058	5.361	5.361	1615.163

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	890	9.82e-03		1.88e-01		1.86e-01	
2	2232	4.07e-03	2.871	1.05e-01	1.892	1.04e-01	1.883
3	5016	1.77e-03	3.081	5.95e-02	2.120	5.90e-02	2.116
4	11220	1.09e-03	1.813	3.86e-02	1.611	3.80e-02	1.642
5	23210	6.44e-04	2.169	2.58e-02	1.656	2.54e-02	1.653
6	42633	4.58e-04	1.680	1.78e-02	1.854	1.73e-02	1.894
7	74679	3.20e-04	1.923	1.37e-02	1.373	1.33e-02	1.409

- **Test 5 Discontinuous permeability,** $u(x, y, z) = a_i \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$**,** min $= -100$, max $= 100$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	22	124	-246.736	-100.000	246.736	100.000	342.699
2	176	1112	-43.618	-35.355	43.618	35.355	68.108
3	1408	9376	-83.040	-78.858	83.040	78.858	92.094
4	11264	76928	-95.567	-94.346	95.567	94.346	97.550
5	90112	623104	-98.880	-98.562	98.880	98.562	98.676
6	720896	5015552	-99.719	-99.639	99.719	99.639	98.928

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	22	1.47e+00		2.14e+03		6.60e+03	
2	176	2.34e-01	2.651	6.29e-01	11.733	1.16e+00	12.475
3	1408	5.30e-02	2.140	2.40e-01	1.390	7.21e-01	0.684
4	11264	1.30e-02	2.034	1.43e-01	0.750	4.99e-01	0.532
5	90112	3.22e-03	2.008	9.78e-02	0.547	3.51e-01	0.507
6	720896	8.04e-04	2.002	6.87e-02	0.510	2.48e-01	0.502

3 Comments

In our experiments the linear systems in Picard method with the non-symmetric matrices were solved by the Bi-Conjugate Gradient Stabilized (BiCGStab) method with the ILU0 preconditioner. The nonlinear iterations are terminated when the relative norm of the residual norm becomes smaller then $\varepsilon_{non} = 10^{-9}$. The convergence tolerance for the linear solver is set to $\varepsilon_{lin} = 10^{-12}$. The number of Picard iterations for different test cases are presented in the table (Test 1 Mild anisotropy: 1B – Tetrahedral meshes, 1C – Voronoi meshes, 1D – Kershaw meshes, 1I – Checkerboard meshes; Test 2 Heterogeneous anisotropy: 2F – Prism meshes; Test 3 Flow on random meshes: 3AA – Random meshes; Test 4 Flow around a well: 4BB – Well meshes; Test 5 Discontinuous permeability: 5H – Locally refined meshes).

i	1B	1C	1D	1I	2F	3AA	4BB	5H
1	37	10	43	14	23	15	18	14
2	47	13	112	27	35	28	18	12
3	41	14	190	37	41	52	20	12
4	50	17	351	41	45	92	21	11
5	57	19		40			23	11
6	58						24	12
7							24	

Acknowledgements This work has been supported in part by RFBR grants 09-01-00115-a, 11-01-00971-a and the federal program "Scientific and scientific-pedagogical personnel of innovative Russia"

References

1. Danilov A., Vassilevski Yu. A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. *Russian J. Numer. Anal. Math. Modelling*, No.24, pp.207-227, 2009.

2. Le Potier 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, pp.787-792, 2005.
3. Lipnikov K., Svyatskiy D., Shashkov M., Vassilevski Yu. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *J. Comp. Phys.* Vol.227, pp.492-512, 2007.
4. Lipnikov K., Svyatskiy D., Vassilevski Yu. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *J. Comp. Phys.* Vol.228, No.3, pp.703-716, 2009.
5. Nikitin K., Vassilevski Yu. A monotone nonlinear finite volume method for advection-diffusion equations on unstructured polyhedral meshes in 3D. *Russian J. Numer. Anal. Math. Modelling*, Vol.25, pp.335-358, 2010.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: the SUSHI Scheme

Robert Eymard, Thierry Gallouët, and Raphaèle Herbin

1 Presentation of the scheme

We present the SUSHI scheme [2] in the case of a general heterogeneous and anisotropic diffusion problem with homogeneous Dirichlet boundary conditions. Let Ω be a bounded open domain of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$, let $f \in L^2(\Omega)$ and let Λ be a measurable function from Ω to the set $\mathscr{M}_d(\mathbb{R})$ of $d \times d$ matrices, such that for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ is symmetric, and such that the set of its eigenvalues is included in $[\underline{\lambda}, \overline{\lambda}]$, where $0 < \underline{\lambda} \le \overline{\lambda}$. We wish to approximate the function u solution of

$$u \in H_0^1(\Omega) \text{ and } \forall v \in H_0^1(\Omega), \int_\Omega \Lambda(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

by the following scheme:

$$U \in X_{\mathscr{D}},\ \forall V \in X_{\mathscr{D}},\ \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathscr{D}}U(\boldsymbol{x}) \cdot \nabla_{\mathscr{D}}V(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_{\mathscr{D}}V(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

where the reconstruction operator $\Pi_{\mathscr{D}}$ and the discrete gradient operator $\nabla_{\mathscr{D}}$ acting on the discrete functional space $X_{\mathscr{D}}$, depending on the discretization $\mathscr{D}$, are now defined, along with some notations:

1. $\mathscr{M}$ is the set of grid cells, that are disjoint open subsets of Ω such that $\bigcup_{K\in\mathscr{M}} \overline{K} = \overline{\Omega}$, $\mathscr{F}$ is the set of the faces of the mesh; note that each non-planar face is decomposed into planar faces without increasing the cost of the method. We assume that Λ is constant on all $K \in \mathscr{M}$, and we denote by Λ_K its value in K; a point $\boldsymbol{x}_K$ is chosen in K such that K is star-shaped with respect to $\boldsymbol{x}_K$;

R. Eymard
Université Paris-Est, France, e-mail: robert.eymard@univ-mlv.fr

T. Gallouët and R. Herbin
Université Aix-Marseille, France, e-mail: Thierry.Gallouet@latp.univ-mrs.fr, Raphaele.Herbin@latp.univ-mrs.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_98, © Springer-Verlag Berlin Heidelberg 2011

2. the set of discrete unknowns $X_{\mathscr{D}}$ is the finite dimensional vector space on $\mathbb{R}$, containing all real families $U = (u_K)_{K\in\mathscr{M}}$;
3. the space step $h_{\mathscr{D}} \in (0, +\infty)$ is the maximum diameter of all control volumes;
4. the mapping $\Pi_{\mathscr{D}} : X_{\mathscr{D}} \to L^2(\Omega)$ is the reconstruction of the approximate function defined by the value u_K in each $K \in \mathscr{M}$;
5. the mapping $\nabla_{\mathscr{D}} : X_{\mathscr{D}} \to L^2(\Omega)^d$ is the reconstruction of the gradient of the function, defined below.

The construction of $\nabla_{\mathscr{D}}$ involves the following steps.

1. for all exterior faces $\sigma \in \mathscr{F}_{\text{ext}}$, a value u_σ is given at the barycentre $\boldsymbol{x}_\sigma$ of σ
2. for each face $\sigma \in \mathscr{F}_{\text{int}}$ and $U = (u_K)_{K\in\mathscr{M}}$, a value u_σ, meant to approximate u at the barycentre $\boldsymbol{x}_\sigma$ of σ, is computed such that

$$u_\sigma = \sum_{K\in\mathscr{M}} \alpha_\sigma^K u_K + \sum_{\tau\in\mathscr{F}_{\text{ext}}} \beta_\sigma^\tau u_\tau, \text{ with } \sum_{K\in\mathscr{M}} \alpha_\sigma^K + \sum_{\tau\in\mathscr{F}_{\text{ext}}} \beta_\sigma^\tau = 1,$$

   where the coefficients α_σ^K and β_σ^τ are chosen as explained below. Note that the interior faces of the mesh cannot be defined as $\partial K \cap \partial L$ for two neighbouring control volumes K and L, since there may exist more than one common face between K and L, in particular, if non-planar faces are split in triangular faces. Indeed, the definition of $\nabla_K U$ below is exact for affine functions only in the case of planar faces (see the comments on the results in the last section of this paper).
3. Denoting by $\mathscr{F}_K$ the subset of $\mathscr{F}$ containing all the faces of $K \in \mathscr{M}$ and, for $\sigma \in \mathscr{F}_K$, by $\boldsymbol{n}_{K,\sigma}$ the unit normal vector to σ outward to K, one defines

$$\nabla_K U = \frac{1}{|K|} \sum_{\sigma\in\mathscr{F}_K} |\sigma|\,(u_\sigma - u_K)\boldsymbol{n}_{K,\sigma},$$

   and, for all $\sigma \in \mathscr{F}_K$, denoting by $d_{K,\sigma}$ the orthogonal distance between $\boldsymbol{x}_K$ and $\sigma \in \mathscr{F}_K$

$$\nabla_{K,\sigma} U = \nabla_K U + \frac{\sqrt{3}}{d_{K,\sigma}}\,(u_\sigma - u_K - \nabla_K U \cdot (\boldsymbol{x}_\sigma - \boldsymbol{x}_K))\,\boldsymbol{n}_{K,\sigma},$$

4. $\nabla_{\mathscr{D}} U$ is given by the constant value $\nabla_{K,\sigma} U$ in the cone with vertex $\boldsymbol{x}_K$ and basis σ.

Let us now turn to the computation of α_σ^K and β_σ^τ. Let K and L be two grid cells separated by a common face σ. Let τ be a face of K or L, which is not common to K and L. We first compute a value w_τ at some point $\boldsymbol{y}_\tau$ by the following method:

1. if $\tau \in \mathscr{F}_{\text{ext}}$, then $\boldsymbol{y}_\tau = \boldsymbol{x}_\tau$ and $w_\tau = u_\tau$;
2. if $\tau \in \mathscr{F}_{\text{int}}$ is a common face to grid cells M and N (with one and one only of them being equal to K or L), then we define

$$\boldsymbol{y}_\tau = \frac{\lambda_N d_{M,\tau} \boldsymbol{y}_N + \lambda_M d_{N,\tau} \boldsymbol{y}_M + d_{M,\tau} d_{N,\tau} (\boldsymbol{\lambda}_N^\sigma - \boldsymbol{\lambda}_M^\sigma)}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}},$$

where, denoting by $\mathscr{P}(\boldsymbol{x},\tau)$ the orthogonal projection on τ of any point $\boldsymbol{x}$ and by $\boldsymbol{n}_{MN}$ the unit normal vector, orthogonal to τ, oriented from M to N, we set

$$\begin{array}{lll} \boldsymbol{y}_M = \mathscr{P}(\boldsymbol{x}_M,\tau), & \lambda_M = \boldsymbol{n}_{MN}\cdot\Lambda_M\boldsymbol{n}_{MN}, & \boldsymbol{\lambda}_M^\tau = \Lambda_M\boldsymbol{n}_{MN} - \lambda_M\boldsymbol{n}_{MN}, \\ \boldsymbol{y}_N = \mathscr{P}(\boldsymbol{x}_N,\tau), & \lambda_N = \boldsymbol{n}_{MN}\cdot\Lambda_N\boldsymbol{n}_{MN}, & \boldsymbol{\lambda}_N^\tau = \Lambda_N\boldsymbol{n}_{MN} - \lambda_N\boldsymbol{n}_{MN}; \end{array}$$

then the following averaging formula is used to define the values w_τ as linear combinations of u_M and u_N:

$$w_\tau = \frac{\lambda_N d_{M,\tau} u_N + \lambda_M d_{N,\tau} u_M}{\lambda_N d_{M,\tau} + \lambda_M d_{N,\tau}};$$

3. two faces $\tau \in \mathscr{F}_K$ and $\tau' \in \mathscr{F}_L$ are then selected so that there exists a unique function w, affine in K and in L, continuous on σ, such that $\Lambda_K(\nabla w)_K \cdot \boldsymbol{n}_{KL} = \Lambda_L(\nabla w)_L\cdot\boldsymbol{n}_{KL}$ and such that $u_K = w(\boldsymbol{x}_K)$, $u_L = w(\boldsymbol{x}_L)$, $w_\tau = w(\boldsymbol{y}_\tau)$ and $w_{\tau'} = w(\boldsymbol{y}_{\tau'})$; we then set $u_\sigma = w(\boldsymbol{x}_\sigma)$, hence defining u_σ as a linear combination of u_K, u_L, w_τ and $w_{\tau'}$; the choice of τ and τ' is done thanks to an invertibility criterion of the 4×4 local linear systems thus obtained.

We refer to [1] for the complete presentation of the mathematical properties of the scheme, which are obtained for a slightly different choice of the coefficients α_σ^K and β_σ^τ from the one presented here. These mathematical properties remain valid in the case of the coefficients chosen here, which present the advantage of preserving exact affine solutions even in the heterogeneous case.

2 Numerical results

In this section, denoting by $|\cdot|$ the Euclidean norm, the norms have been computed by the following formula:

$$\text{normg} = \sum_{K\in\mathscr{M}} |K|\ |\nabla_K U|,$$

$$\text{erl2} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,(u_K - u(\boldsymbol{x}_K))^2\right) / \left(\sum_{K\in\mathscr{M}} |K|\,u(\boldsymbol{x}_K)^2\right)\right)^{1/2},$$

$$\text{ergrad} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,|\nabla_K U - \nabla u(\boldsymbol{x}_K)|^2\right) / \left(\sum_{K\in\mathscr{M}} |K|\,|\nabla u(\boldsymbol{x}_K)|^2\right)\right)^{1/2},$$

$$\text{ener} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,|\nabla_K U - \nabla u(\boldsymbol{x}_K)|_\Lambda^2\right) / \left(\sum_{K\in\mathscr{M}} |K|\,|\nabla u(\boldsymbol{x}_K)|_\Lambda^2\right)\right)^{1/2},$$

setting, for any $K \in \mathscr{M}$, $|\xi|_\Lambda^2 = \Lambda_K\xi\cdot\xi$ for all $\xi \in \mathbb{R}^3$.

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	2003	59943	3.21E-02	2.03E-02	1.98E+00	1.99E+00	1.77E+00
2	3898	122098	1.29E-02	6.84E-03	1.98E+00	1.99E+00	1.77E+00
3	7711	249457	1.30E-02	9.13E-03	1.99E+00	1.99E+00	1.78E+00
4	15266	504716	4.66E-03	5.52E-03	1.99E+00	2.00E+00	1.79E+00
5	30480	1029682	4.03E-03	1.49E-03	2.00E+00	2.00E+00	1.79E+00
6	61052	2102030	1.74E-03	1.83E-03	2.00E+00	2.00E+00	1.79E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	2003	8.39E-03	-	1.63E-01	-	1.55E-01	-
2	3898	6.28E-03	1.31E+00	1.32E-01	9.50E-01	1.25E-01	9.41E-01
3	7711	3.98E-03	2.01E+00	1.04E-01	1.03E+00	9.90E-02	1.04E+00
4	15266	2.63E-03	1.83E+00	8.34E-02	9.77E-01	7.87E-02	1.01E+00
5	30480	1.67E-03	1.96E+00	6.59E-02	1.02E+00	6.26E-02	9.92E-01
6	61052	1.04E-03	2.03E+00	5.21E-02	1.02E+00	4.93E-02	1.03E+00

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	29	765	1.11E-01	1.56E-01	1.90E+00	1.86E+00	1.36E+00
2	66	2934	1.29E-01	1.79E-01	1.85E+00	1.81E+00	1.56E+00
3	130	7598	2.67E-02	2.67E-02	1.94E+00	1.93E+00	1.63E+00
4	228	16210	9.62E-03	1.20E-02	1.93E+00	1.91E+00	1.67E+00
5	356	29820	1.02E-02	3.85E-03	2.00E+00	1.97E+00	1.69E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	29	7.37E-02	-	3.97E-01	-	4.01E-01	-
2	66	6.41E-02	5.08E-01	3.12E-01	8.80E-01	2.89E-01	1.19E+00
3	130	4.05E-02	2.03E+00	2.64E-01	7.42E-01	2.48E-01	6.83E-01
4	228	2.81E-02	1.95E+00	2.11E-01	1.20E+00	1.97E-01	1.22E+00
5	356	1.86E-02	2.79E+00	1.85E-01	8.87E-01	1.78E-01	6.86E-01

- **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi \left(y + \frac{1}{2}\right)\right) \sin\left(\pi \left(z + \frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	512	21422	-2.14E-03	3.03E-02	1.91E+00	1.96E+00	1.67E+00
2	4096	192664	1.58E-02	1.06E-02	1.96E+00	1.99E+00	1.73E+00
3	32768	1618164	4.90E-03	1.75E-03	1.99E+00	2.00E+00	1.74E+00
4	262144	13109746	8.51E-04	7.14E-04	2.00E+00	2.00E+00	1.76E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	512	7.45E-02	-	4.84E-01	-	4.40E-01	-
2	4096	6.48E-02	2.00E-01	4.43E-01	1.28E-01	3.85E-01	1.92E-01
3	32768	4.32E-02	5.84E-01	3.02E-01	5.50E-01	2.56E-01	5.92E-01
4	262144	2.31E-02	9.02E-01	1.66E-01	8.62E-01	1.40E-01	8.74E-01

- **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi \left(y + \frac{1}{2}\right)\right) \sin\left(\pi \left(z + \frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	36	836	1.05E-01	1.54E-01	1.87E+00	1.85E+00	1.60E+00
2	288	15848	3.67E-02	4.01E-02	1.96E+00	1.96E+00	1.71E+00
3	2304	173048	5.91E-03	1.01E-02	1.99E+00	1.99E+00	1.77E+00
4	18432	1560014	1.71E-03	2.54E-03	2.00E+00	2.00E+00	1.79E+00
5	147456	13339482	3.83E-04	6.36E-04	2.00E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	36	1.11E-01	-	2.77E-01	-	2.49E-01	-
2	288	3.34E-02	1.74E+00	1.50E-01	8.86E-01	1.40E-01	8.27E-01
3	2304	8.76E-03	1.93E+00	6.88E-02	1.13E+00	6.63E-02	1.08E+00
4	18432	2.33E-03	1.91E+00	3.34E-02	1.04E+00	3.33E-02	9.95E-01
5	147456	5.82E-04	2.00E+00	1.55E-02	1.10E+00	1.54E-02	1.12E+00

- **Test 2 Heterogeneous anisotropy,** $u(x, y, z) = x^3 y^2 z + x \sin(2\pi xz) \sin(2\pi xy)$ $\sin(2\pi z)$**,** $\min = -0.862, \ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1210	65648	-8.22E-01	-8.41E-01	9.82E-01	9.84E-01	1.50E+00
2	8820	553442	-8.33E-01	-8.39E-01	1.00E+00	1.01E+00	1.64E+00
3	28830	1935862	-8.55E-01	-8.59E-01	1.03E+00	1.03E+00	1.68E+00
4	67240	4710944	-8.55E-01	-8.57E-01	1.03E+00	1.03E+00	1.69E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1210	5.95E-02	-	1.88E-01	-	1.91E-01	-
2	8820	1.85E-02	1.76E+00	6.67E-02	1.56E+00	6.75E-02	1.57E+00
3	28830	8.94E-03	1.85E+00	3.46E-02	1.66E+00	3.50E-02	1.66E+00
4	67240	5.37E-03	1.80E+00	2.23E-02	1.55E+00	2.25E-02	1.56E+00

- **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1, \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	64	2306	-7.51E-01	-7.59E-01	7.58E-01	6.91E-01	1.43E+00
2	512	31576	-8.36E-01	-9.39E-01	8.64E-01	9.23E-01	2.58E+00
3	4096	317246	-9.69E-01	-9.85E-01	9.58E-01	9.82E-01	3.28E+00
4	32768	2819464	-9.90E-01	-9.96E-01	9.89E-01	9.96E-01	3.51E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	64	2.00E-01	-	6.47E-01	-	6.64E-01	-
2	512	1.28E-01	6.48E-01	3.00E-01	1.11E+00	2.98E-01	1.16E+00
3	4096	4.67E-02	1.45E+00	1.06E-01	1.50E+00	1.05E-01	1.50E+00
4	32768	1.32E-02	1.82E+00	3.85E-02	1.46E+00	3.72E-02	1.50E+00

- **Test 4 Flow around a well, Well meshes,** $\min = 0, \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	890	56952	4.26E-01	4.14E-01	5.32E+00	5.32E+00	1.58E+03
2	2232	164566	2.58E-01	2.44E-01	5.33E+00	5.33E+00	1.58E+03
3	5016	394986	1.61E-01	1.54E-01	5.33E+00	5.33E+00	1.60E+03
4	11220	927684	1.23E-01	1.18E-01	5.33E+00	5.33E+00	1.61E+03
5	23210	1980998	9.28E-02	8.99E-02	5.34E+00	5.34E+00	1.62E+03
6	42633	3702759	7.41E-02	7.23E-02	5.35E+00	5.35E+00	1.62E+03
7	74679	6573107	5.78E-02	5.65E-02	5.36E+00	5.36E+00	1.62E+03

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	890	3.79E-03	-	9.69E-02	-	9.56E-02	-
2	2232	3.07E-03	6.86E-01	5.21E-02	2.03E+00	4.96E-02	2.14E+00
3	5016	1.60E-03	2.42E+00	2.81E-02	2.29E+00	2.66E-02	2.31E+00
4	11220	1.10E-03	1.38E+00	2.10E-02	1.08E+00	1.95E-02	1.16E+00
5	23210	7.77E-04	1.45E+00	1.57E-02	1.19E+00	1.45E-02	1.21E+00
6	42633	4.78E-04	2.39E+00	1.07E-02	1.89E+00	1.01E-02	1.81E+00
7	74679	4.56E-04	2.59E-01	9.98E-03	3.72E-01	9.27E-03	4.42E-01

• **Test 5 Discontinuous permeability,** $u(x, y, z) = \sin(\pi x)\sin(\pi y)\sin(\pi z)$**,** $\min = 0,\ \max = 1$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	22	358	-2.49E+02	-1.00E+02	2.49E+02	1.00E+02	1.03E+02
2	176	4570	-4.63E+01	-3.54E+01	4.63E+01	3.54E+01	9.21E+01
3	1408	37730	-8.35E+01	-7.89E+01	8.35E+01	7.89E+01	9.64E+01
4	11264	293666	-9.56E+01	-9.43E+01	9.56E+01	9.43E+01	9.85E+01
5	90112	2309882	-9.89E+01	-9.86E+01	9.89E+01	9.86E+01	9.91E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	22	1.55E+00	-	1.02E+01	-	4.09E+01	-
2	176	3.00E-01	2.37E+00	2.42E-01	5.39E+00	2.35E-01	7.44E+00
3	1408	6.57E-02	2.19E+00	6.08E-02	2.00E+00	6.20E-02	1.92E+00
4	11264	1.63E-02	2.01E+00	1.90E-02	1.68E+00	1.69E-02	1.88E+00
5	90112	4.62E-03	1.82E+00	8.18E-03	1.22E+00	5.10E-03	1.73E+00

3 Comments on the results

All the linear solvers could be solved using the conjugate gradient solver of the PETSC library with ILU(2) preconditioning with tolerance (or reduction factor) set to 10^{-10}. The following results have been obtained:

1. Using the conjugate gradient solver of the PETSC library, the ILU(0) seems to be the fastest preconditioning on some cases. For example, using the fourth Kershaw mesh, for test 1, we obtain the following CPU times: for ILU(2), 178s, for ILU(1), 60s, for ILU(0), 14s and for Jacobi, 20s. We systematically used ILU(2) in order to prevent from any possible failure.
2. The computing times, using the conjugate gradient solver of the PETSC library with ILU(2) preconditioning are the following, for tetrahedral meshes 2 to 6 on test 1: 1.06, 2.45, 5.73, 6.07 and 23.65s. For the conjugate gradient solver of the ISTL library, with ILU(0), we obtain 0.05, 0.10, 0.29, 0.73 and 1.83s, which seems to show that the computing time which can be expected on full scale studies will be acceptable.

A second remark concerns the treatment of non-planar faces. In the above results, we used the possibility to decompose the non-planar faces in triangles, in particular in test3, “Flow on random meshes”, the results which are obtained without using this possibility are the following:

i	nu	nmat	umin	uemin	umax	uemax	normg
1	64	1774	-9.69E-01	-7.55E-01	8.61E-01	6.98E-01	2.57E+00
2	512	21812	-9.30E-01	-9.39E-01	9.97E-01	9.24E-01	3.22E+00
3	4096	210534	-1.02E+00	-9.85E-01	1.01E+00	9.82E-01	3.51E+00
4	32768	1839254	-1.00E+00	-9.96E-01	1.01E+00	9.96E-01	3.58E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	64	2.96E-01	-	3.89E-01	-	3.54E-01	-
2	512	9.59E-02	1.62E+00	1.87E-01	1.06E+00	1.57E-01	1.17E+00
3	4096	3.89E-02	1.30E+00	1.23E-01	6.00E-01	6.93E-02	1.18E+00
4	32768	1.92E-02	1.02E+00	1.13E-01	1.20E-01	4.94E-02	4.87E-01

They show a clear loss of accuracy of the scheme (the order of convergence being around 1 and not 2).

Acknowledgements Work supported by Groupement MOMAS and ANR VFSitCom

References

1. R. Eymard, T. Gallouët, and R. Herbin. 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(4):1009–1043, 2010. see also http://hal.archives-ouvertes.fr/.
2. R. Eymard and R. Herbin. Gradient schemes for diffusion problem. *these proceedings*, 2011.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: the VAG scheme

Robert Eymard, Cindy Guichard, and Raphaèle Herbin

1 Presentation of the scheme

Let Ω be a bounded open domain of $\mathbb{R}^3$, let $f \in L^2(\Omega)$ and let Λ be a measurable function from Ω to the set $\mathscr{M}_3(\mathbb{R})$ of 3×3 matrices, such that for a.e. $\boldsymbol{x} \in \Omega$, $\Lambda(\boldsymbol{x})$ is symmetric, and such that the set of its eigenvalues is included in $[\underline{\lambda}, \overline{\lambda}]$, where $0 < \underline{\lambda} \leq \overline{\lambda}$. We wish to approximate the function u solution of

$$u \in H_0^1(\Omega) \text{ and } \forall v \in H_0^1(\Omega), \int_\Omega \Lambda(\boldsymbol{x})\nabla u(\boldsymbol{x}) \cdot \nabla v(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})v(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad (1)$$

by the approximate gradient scheme [2,4] which reads:

$$U \in X_{\mathscr{D}}, \; \forall V \in X_{\mathscr{D}}, \int_\Omega \Lambda(\boldsymbol{x})\nabla_{\mathscr{D}}U(\boldsymbol{x})\cdot\nabla_{\mathscr{D}}V(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega f(\boldsymbol{x})\Pi_{\mathscr{D}}V(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \quad (2)$$

where $\Pi_{\mathscr{D}}$ is a reconstruction operator and $\nabla_{\mathscr{D}}$ a discrete gradient operator which act on the discrete functional space $X_{\mathscr{D}}$, where the index $\mathscr{D}$ denotes the discretization; these operators are defined as defined as follows:

R. Eymard
Université Paris-Est, e-mail: robert.eymard@univ-mlv.fr

C. Guichard
IFP Energies nouvelles and Université Paris-Est, e-mail: guichard@ifpenergiesnouvelles.fr

R. Herbin
Université Aix-Marseille, e-mail: Raphaele.Herbin@latp.univ-mrs.fr

J. Fovrt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_99, © Springer-Verlag Berlin Heidelberg 2011

1. $\mathcal{M}$ is the set of control volumes, that are disjoint open subsets of Ω such that $\bigcup_{K\in\mathcal{M}}\overline{K}=\overline{\Omega}$, $\mathcal{V}=\mathcal{V}_{int}\cup\mathcal{V}_{ext}$ is the set of vertices of the mesh; any element K of $\mathcal{M}$ is defined by its vertices $s\in\mathcal{V}_K$, its faces $\sigma\in\mathcal{F}_K$; each face is also defined by the set of its vertices $s\in\mathcal{V}_\sigma$, using a suitable geometric definition for the resulting surface in the case of non-planar faces; we assume that Λ is constant on all $K\in\mathcal{M}$, and we denote by Λ_K its value in K;
2. the set of discrete unknowns $X_{\mathcal{D}}$ is the finite dimensional vector space on $\mathbb{R}$, containing all real families $U=((u_K)_{K\in\mathcal{M}},(u_s)_{s\in\mathcal{V}})$, such that $u_s=0$ if $s\in\mathcal{V}_{\text{ext}}$;
3. the mapping $\Pi_{\mathcal{D}}:X_{\mathcal{D}}\to L^2(\Omega)$ maps $U=((u_K)_{K\in\mathcal{M}},(u_s)_{s\in\mathcal{V}})\in X_{\mathcal{D}}$ to the piecewise constant function $u_{\mathcal{D}}\in L^2(\Omega)$ equal to u_K on each cell $K\in\mathcal{M}$;
4. the mapping $\nabla_{\mathcal{D}}:X_{\mathcal{D}}\to L^2(\Omega)^3$ is the reconstruction of a gradient from the values $U=((u_K)_{K\in\mathcal{M}},(u_s)_{s\in\mathcal{V}})\in X_{\mathcal{D}}$; different expressions for this reconstruction are proposed below, which all lead to convergent gradient schemes in the sense of [2, 4]. Their theoretical analysis is related to that of the SUSHI scheme [1, 3]. Detailed numerical results are given in this paper only using the method described in subsection 1.2; the differences obtained using the other expressions are commented in the last section.

The exterior faces are those of the form $\partial K\cap\partial\Omega$ for any boundary control volume K, and the interior faces are those of the form $\partial K\cap\partial L$ for two neighbouring control volumes K and L. For any face σ, we define a point $\boldsymbol{x}_\sigma$, which is a barycentre with non-negative weights $\beta_{\sigma,s}$ of the elements of the set $\mathcal{V}_\sigma$ including all the vertices of the face, and the value u_σ is defined by

$$u_\sigma=\sum_{s\in\mathcal{V}_\sigma}\beta_{\sigma,s}u_s \text{ with } \sum_{s\in\mathcal{V}_\sigma}\beta_{\sigma,s}=1.$$

In the next three subsections, we describe three ways of defining a gradient operator which satisfies the VAG requirements. The first gradient is constructed from the Stokes formula on the cells of the mesh (we call it the primal cell to distinguish it from further constructed cells), and requires a stabilization. The second gradient and third gradients are constructed on tetrahedral or octahedral sub–cells of the primal mesh, and are natively stable.

1.1 Stabilised gradient on the primal mesh cells

For a face $\sigma\in\mathcal{F}$, we denote by τ any triangular sub-face with vertices $\boldsymbol{x}_\sigma$, s and s', where s and s' are two consecutive vertices of σ. The barycentre $\boldsymbol{x}_\tau$ of each sub-face τ may thus be expressed by the following barycentric combination:

$$\boldsymbol{x}_\tau=\sum_{s\in\mathcal{V}_\sigma}\beta_{\tau,s}s \text{ with } \sum_{s\in\mathcal{V}_\sigma}\beta_{\tau,s}=1,$$

where $\beta_{\tau,s} \geq 0$ for all $s \in \mathscr{V}_\sigma$. We then define $\beta_{\tau,s} = 0$ for all $s \in \mathscr{V} \setminus \mathscr{V}_\sigma$. Next, we reconstruct a value u_τ at the point $\boldsymbol{x}_\tau$, by $u_\tau = \sum_{s\in\mathscr{V}_\sigma} \beta_{\tau,s} u_s$. Let $K \in \mathscr{M}$ be an element of the mesh. We denote by $\mathscr{T}_K$ the set of all sub-faces of the faces of K.

We first define, for $U = ((u_K)_{K\in\mathscr{M}}, (u_s)_{s\in\mathscr{V}})$, an approximation of the gradient on cell K:

$$\nabla_K U = \frac{1}{|K|} \sum_{\tau\in\mathscr{T}_K} |\tau|\, (u_\tau - u_K) n_{K,\tau} = \sum_{s\in\mathscr{V}_K} (u_s - u_K) v_{K,s}, \tag{3}$$

where we denote by

$$v_{K,s} = \frac{1}{|K|} \sum_{\tau\in\mathscr{T}_K} \beta_{\tau,s} |\tau|\, n_{K,\tau},$$

where $n_{K,\tau}$ is the unit normal vector to τ, outward to K, and $|\tau|$, $|K|$ are respectively the area and the volume of τ and K. We then define a partition $M_{K,s}$ of K (there is no need to define this partition precisely), such that $|M_{K,s}| = |K|\,/N_K$, where N_K is the number of vertices of K and we introduce

$$R_{K,s} U = u_s - u_K - \nabla_K U \cdot (s - \boldsymbol{x}_K).$$

We then define, for a given $\gamma > 0$, the constant value $\nabla_{K,s} U$ in $M_{K,s}$:

$$\nabla_{K,s} U = \nabla_K u + \gamma R_{K,s} U v_{K,s}.$$

We finally define a piecewise constant gradient by $\nabla_{\mathscr{D}} U(\boldsymbol{x}) = \nabla_{K,s} U$ for a.e. $\boldsymbol{x} \in M_{K,s}$. This scheme is denoted by "VAG" in [5].

1.2 Piecewise constant gradient on octahedral sub-cells

For a given face σ of a control volume K and for any vertex s of σ, we respectively denote by s^- and s^+ the preceding and the following vertices of s in the face σ (defining any orientation on σ), and we consider the (degenerate) octahedron, denoted by $V_{K,\sigma,s}$ and depicted in Fig. 1, whose vertices are $A_1 = \boldsymbol{x}_K$, $A_2 = \boldsymbol{x}_\sigma$, $A_3 = \frac{1}{2}(s^- + s)$, $A_5 = s$, $A_6 = \frac{1}{2}(s^+ + s)$ and $A_4 = \frac{1}{2}(\boldsymbol{x}_\sigma + s)$ (note that all these octahedra are disjoint, and that the union of their closure is $\overline{\Omega}$). The approximate values of U at the vertices of $V_{K,\sigma,s}$ are respectively $u_1 = u_K$, $u_2 = u_\sigma$, $u_3 = \frac{1}{2}(u_{s^-} + u_s)$, $u_5 = u_s$, $u_6 = \frac{1}{2}(u_{s^+} + u_s)$ and $u_4 = \frac{1}{2}(u_\sigma + u_s)$ (the main diagonals of $V_{K,\sigma,s}$ are therefore (A_1, A_4), (A_2, A_5) and (A_3, A_6)). We then define the following approximate gradient:

$$\nabla_{K,\sigma,s} U = \sum_{i=1}^{3} (u_{i+3} - u_i) \frac{\overrightarrow{A_{i+1}A_{i+4}} \wedge \overrightarrow{A_{i+2}A_{i+5}}}{\mathrm{Det}(\overrightarrow{A_{i+1}A_{i+4}}, \overrightarrow{A_{i+2}A_{i+5}}, \overrightarrow{A_i A_{i+3}})}, \tag{4}$$

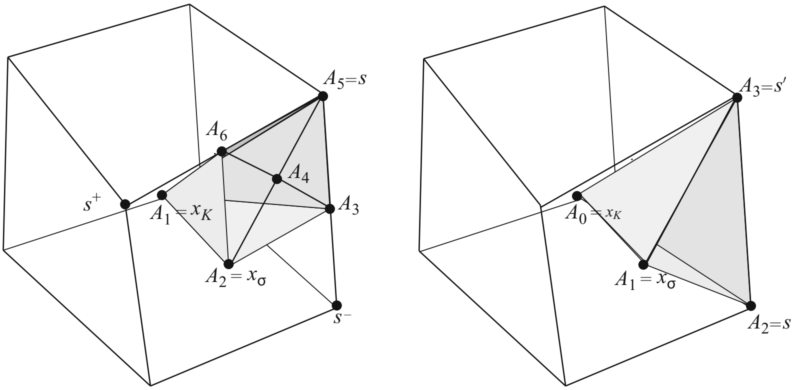


Fig. 1 The octahedral (left) and tetrahedral (right) cells for the definition of the gradient

setting $A_7 = A_1$ and $A_8 = A_2$. We finally define a piecewise constant gradient by $\nabla_{\mathscr{D}} U(\boldsymbol{x}) = \nabla_{K,\sigma,s} U$ for a.e. $\boldsymbol{x} \in V_{K,\sigma,s}$. Remark that, denoting for simplicity V instead of $V_{K,\sigma,s}$, defining $\mathscr{F}_V$ as the set of the 8 triangular faces of V and $\mathscr{V}_\tau$ as the set of the 3 vertices of each triangular face τ of V, one may check that

$$\nabla_{K,\sigma,s} U = \frac{1}{|V|} \sum_{\tau \in \mathscr{F}_V} |\tau| n_{V,\tau} \big(\frac{1}{3} \sum_{s \in \mathscr{V}_\tau} u_s \big). \tag{5}$$

We then set define $\nabla_K U$, used in the tables below, by

$$|K| \nabla_K U = \sum_{\sigma \in \mathscr{F}_K} \sum_{s \in \mathscr{V}_\sigma} |V_{K,\sigma,s}| \; \nabla_{K,\sigma,s} U.$$

This scheme is denoted by "VAGR" in [5].

1.3 Piecewise constant gradient on tetrahedral sub–cells

For a given face σ of a control volume K and for any pair of consecutive vertices (s, s') of σ, we consider the tetrahedron, denoted by $V_{K,\sigma,s,s'}$ and depicted in Fig. 1, whose vertices are $A_0 = \boldsymbol{x}_K$, $A_1 = \boldsymbol{x}_\sigma$, $A_2 = s$ and $A_3 = s'$ (note that all these tetrahedra are disjoint, and that the union of their closure is $\overline{\Omega}$). The approximate values of U at the vertices of $V_{K,\sigma,s,s'}$ are respectively $u_0 = u_K$, $u_1 = u_\sigma$, $u_2 = u_s$ and $u_3 = u_{s'}$. We then define the following approximate gradient:

$$\nabla_{K,\sigma,s,s'} U = \sum_{i=1}^{3} (u_i - u_0) \frac{\overrightarrow{A_0 A_{i+1}} \wedge \overrightarrow{A_0 A_{i+2}}}{\mathrm{Det}(\overrightarrow{A_0 A_{i+1}}, \overrightarrow{A_0 A_{i+2}}, \overrightarrow{A_0 A_i})}, \tag{6}$$

where $A_4 = A_1$ and $A_5 = A_2$. We finally define a piecewise constant gradient by $\nabla_{\mathscr{D}} U(\boldsymbol{x}) = \nabla_{K,\sigma,s,s'} U$ for a.e. $\boldsymbol{x} \in V_{K,\sigma,s,s'}$. Remark that (5) also holds in this case, denoting V instead of $V_{K,\sigma,s,s'}$.

2 Numerical results

We provide the detailed numerical results obtained, using the scheme VAGR for computing the discrete gradient. In the numerical implementation, the values u_K are locally eliminated, and the unknowns of the linear solver are the values u_s. Denoting by $|\cdot|$ denotes the Euclidean norm, the norms used in the bench tables have been computed using the following formulae:

$$\text{normg} = \sum_{K\in\mathscr{M}} |K|\ |\nabla_K U|,$$

$$\text{erl2} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,(u_K - u(\boldsymbol{x}_K))^2\right) \Big/ \left(\sum_{K\in\mathscr{M}} |K|\,u(\boldsymbol{x}_K)^2\right)\right)^{1/2},$$

$$\text{ergrad} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,|\nabla_K U - \nabla u(\boldsymbol{x}_K)|^2\right) \Big/ \left(\sum_{K\in\mathscr{M}} |K|\,|\nabla u(\boldsymbol{x}_K)|^2\right)\right)^{1/2},$$

$$\text{ener} = \left(\left(\sum_{K\in\mathscr{M}} |K|\,|\nabla_K U - \nabla u(\boldsymbol{x}_K)|_\Lambda^2\right) \Big/ \left(\sum_{K\in\mathscr{M}} |K|\,|\nabla u(\boldsymbol{x}_K)|_\Lambda^2\right)\right)^{1/2},$$

setting, for any $K \in \mathscr{M}$, $|\xi|_\Lambda^2 = \Lambda_K \xi \cdot \xi$ for all $\xi \in \mathbb{R}^3$.

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	488	6072	5.77E-02	2.03E-02	1.95E+00	1.99E+00	1.77E+00
2	857	11269	1.88E-02	6.84E-03	1.97E+00	1.99E+00	1.78E+00
3	1601	21675	2.19E-02	9.13E-03	1.98E+00	1.99E+00	1.79E+00
4	2997	41839	1.13E-02	5.52E-03	1.99E+00	2.00E+00	1.79E+00
5	5692	81688	8.73E-03	1.49E-03	1.99E+00	2.00E+00	1.79E+00
6	10994	160852	3.63E-03	1.83E-03	1.99E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	488	1.76E-02	-	2.30E-01	-	2.28E-01	-
2	857	1.02E-02	2.93E+00	1.79E-01	1.35E+00	1.77E-01	1.35E+00
3	1601	6.79E-03	1.94E+00	1.44E-01	1.05E+00	1.42E-01	1.08E+00
4	2997	4.44E-03	2.03E+00	1.13E-01	1.14E+00	1.11E-01	1.17E+00
5	5692	2.79E-03	2.18E+00	9.02E-02	1.06E+00	8.89E-02	1.03E+00
6	10994	1.75E-03	2.13E+00	7.04E-02	1.13E+00	6.92E-02	1.15E+00

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	146	5936	7.54E-02	1.56E-01	2.15E+00	1.86E+00	1.14E+00
2	339	16267	-3.42E-01	1.79E-01	1.95E+00	1.81E+00	1.43E+00
3	684	37194	8.40E-04	2.67E-02	2.02E+00	1.93E+00	1.60E+00
4	1227	71069	-9.54E-02	1.20E-02	2.06E+00	1.91E+00	1.66E+00
5	2023	127883	-1.78E-02	3.85E-03	2.06E+00	1.97E+00	1.70E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	146	1.82E-01	-	3.96E-01	-	4.05E-01	-
2	339	1.87E-01	-8.43E-02	2.49E-01	1.65E+00	2.53E-01	1.68E+00
3	684	9.92E-02	2.70E+00	1.55E-01	2.02E+00	1.62E-01	1.90E+00
4	1227	7.15E-02	1.68E+00	1.19E-01	1.35E+00	1.23E-01	1.42E+00
5	2023	4.74E-02	2.47E+00	9.56E-02	1.33E+00	9.92E-02	1.29E+00

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	729	15625	7.80E-02	3.03E-02	1.96E+00	1.96E+00	1.56E+00
2	4913	117649	1.72E-02	1.06E-02	1.98E+00	1.99E+00	1.68E+00
3	35937	912673	-2.58E-04	1.75E-03	1.99E+00	2.00E+00	1.74E+00
4	274625	7189057	-2.64E-04	7.14E-04	2.00E+00	2.00E+00	1.78E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	729	9.17E-02	-	4.91E-01	-	4.84E-01	-
2	4913	5.53E-02	7.96E-01	3.09E-01	7.28E-01	2.84E-01	8.40E-01
3	35937	2.97E-02	9.38E-01	1.74E-01	8.70E-01	1.54E-01	9.22E-01
4	274625	1.22E-02	1.31E+00	7.40E-02	1.26E+00	6.44E-02	1.29E+00

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	97	2413	-9.81E-02	1.54E-01	2.08E+00	1.85E+00	1.34E+00
2	625	22585	-1.90E-01	4.01E-02	2.19E+00	1.96E+00	1.70E+00
3	4417	188641	-6.12E-02	1.01E-02	2.06E+00	1.99E+00	1.78E+00
4	33025	1529617	-1.70E-02	2.54E-03	2.02E+00	2.00E+00	1.79E+00
5	254977	12295153	-4.33E-03	6.36E-04	2.00E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	97	3.25E-01	-	4.37E-01	-	3.97E-01	-
2	625	1.11E-01	1.73E+00	1.50E-01	1.72E+00	1.52E-01	1.54E+00
3	4417	3.01E-02	2.00E+00	5.73E-02	1.47E+00	6.09E-02	1.41E+00
4	33025	7.92E-03	1.99E+00	2.51E-02	1.23E+00	2.77E-02	1.18E+00
5	254977	2.03E-03	2.00E+00	1.18E-02	1.11E+00	1.32E-02	1.08E+00

- **Test 2 Heterogeneous anisotropy,** $u(x,y,z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)$ $\sin(2\pi z)$**,** $\min = -0.862,\ \max = 1.048$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3080	99634	-8.73E-01	-8.41E-01	1.10E+00	9.84E-01	1.53E+00
2	20160	710894	-8.25E-01	-8.39E-01	1.04E+00	1.01E+00	1.66E+00
3	63240	2301754	-8.52E-01	-8.59E-01	1.05E+00	1.03E+00	1.69E+00
4	144320	5340214	-8.53E-01	-8.57E-01	1.04E+00	1.03E+00	1.70E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3080	1.66E-01	-	1.40E-01	-	1.38E-01	-
2	20160	4.26E-02	2.17E+00	3.71E-02	2.13E+00	3.64E-02	2.13E+00
3	63240	1.93E-02	2.08E+00	1.67E-02	2.10E+00	1.63E-02	2.10E+00
4	144320	1.10E-02	2.05E+00	9.44E-03	2.06E+00	9.25E-03	2.07E+00

- **Test 3 Flow on random meshes,** $u(x,y,z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	125	2197	-1.51E+00	-7.55E-01	1.68E+00	6.98E-01	1.53E+00
2	729	15625	-1.13E+00	-9.39E-01	1.21E+00	9.24E-01	2.99E+00
3	4913	117649	-1.08E+00	-9.85E-01	1.06E+00	9.82E-01	3.44E+00
4	35937	912673	-1.01E+00	-9.96E-01	1.01E+00	9.96E-01	3.56E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	125	1.15E+00	-	6.19E-01	-	6.26E-01	-
2	729	2.56E-01	2.56E+00	2.02E-01	1.90E+00	1.81E-01	2.11E+00
3	4913	5.93E-02	2.30E+00	8.04E-02	1.45E+00	5.30E-02	1.93E+00
4	35937	1.49E-02	2.09E+00	3.45E-02	1.28E+00	1.74E-02	1.68E+00

- **Test 4 Flow around a well, Well meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1248	27072	3.89E-01	4.29E-01	5.32E+00	5.32E+00	1.68E+03
2	2800	65184	2.41E-01	2.50E-01	5.33E+00	5.33E+00	1.65E+03
3	5889	143079	1.55E-01	1.57E-01	5.33E+00	5.33E+00	1.64E+03
4	12582	314964	1.18E-01	1.20E-01	5.33E+00	5.33E+00	1.63E+03
5	25300	645210	9.03E-02	9.09E-02	5.34E+00	5.34E+00	1.63E+03
6	45668	1178094	7.27E-02	7.30E-02	5.34E+00	5.35E+00	1.63E+03
7	79084	2055600	5.69E-02	5.68E-02	5.36E+00	5.36E+00	1.63E+03

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1248	6.47E-03	-	5.78E-02	-	5.35E-02	-
2	2800	2.71E-03	3.23E+00	2.54E-02	3.05E+00	2.34E-02	3.08E+00
3	5889	1.19E-03	3.31E+00	1.23E-02	2.93E+00	1.15E-02	2.85E+00
4	12582	8.42E-04	1.37E+00	7.59E-03	1.91E+00	7.31E-03	1.79E+00
5	25300	4.47E-04	2.72E+00	5.10E-03	1.71E+00	4.95E-03	1.68E+00
6	45668	2.02E-04	4.03E+00	3.55E-03	1.83E+00	3.47E-03	1.80E+00
7	79084	1.75E-04	7.84E-01	3.26E-03	4.76E-01	3.19E-03	4.56E-01

- **Test 5 Discontinuous permeability,** $u(x,y,z)=\alpha_i \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = 0,\ \max = 1$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	60	1148	-7.39E+02	-1.00E+02	7.39E+02	1.00E+02	1.24E+01
2	305	6825	-7.82E+01	-3.54E+01	7.82E+01	3.54E+01	5.20E+01
3	1881	46025	-9.90E+01	-7.89E+01	9.90E+01	7.89E+01	8.60E+01
4	13073	335601	-9.99E+01	-9.43E+01	9.99E+01	9.43E+01	9.56E+01
5	97185	2557793	-1.00E+02	-9.86E+01	1.00E+02	9.86E+01	9.80E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	60	6.39E+00	-	1.60E+00	-	8.27E+00	-
2	305	1.19E+00	3.10E+00	5.97E-01	1.82E+00	6.01E-01	4.84E+00
3	1881	2.55E-01	2.55E+00	1.86E-01	1.92E+00	1.80E-01	1.99E+00
4	13073	6.10E-02	2.21E+00	5.96E-02	1.76E+00	4.78E-02	2.05E+00
5	97185	1.52E-02	2.08E+00	2.24E-02	1.46E+00	1.26E-02	2.00E+00

3 Comments on the results

The results obtained using (3) (VAG) instead of (4) (VAGR) are systematically less precise, except in the test5 case, where we obtained the following tables:

i	nu	nmat	umin	uemin	umax	uemax	normg
1	60	1148	-7.65E+02	-1.00E+02	7.65E+02	1.00E+02	6.76E+01
2	305	6825	-7.73E+01	-3.54E+01	7.73E+01	3.54E+01	4.65E+01
3	1881	46025	-9.02E+01	-7.89E+01	9.02E+01	7.89E+01	8.19E+01
4	13073	335601	-9.72E+01	-9.43E+01	9.72E+01	9.43E+01	9.43E+01
5	97185	2557793	-9.93E+01	-9.86E+01	9.93E+01	9.86E+01	9.77E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	60	6.71E+00	-	7.32E+00	-	2.90E+01	-
2	305	9.53E-01	3.60E+00	6.91E-01	4.36E+00	6.76E-01	6.93E+00
3	1881	1.49E-01	3.05E+00	2.24E-01	1.85E+00	2.20E-01	1.85E+00
4	13073	3.27E-02	2.35E+00	6.17E-02	2.00E+00	5.95E-02	2.03E+00
5	97185	7.98E-03	2.11E+00	1.73E-02	1.90E+00	1.54E-02	2.02E+00

The results using (6) are very similar to those obtained using (4) (VAGR). For both (3) (VAG) and (4) (VAGR), we have chosen the conjugate gradient solver of the ISTL library with ILU(0) preconditioning with tolerance (or reduction factor) set to 10^{-10}. The following observations have been made on the computing times, using (3) (VAG) (we may expect that similar observations could be done with VAGR).

1. On the fourth Kershaw mesh and test 1, we obtain the following CPU times using the conjugate gradient solver of the PETSC library: with ILU(2), 33s, with ILU(1), 17s, with ILU(0), 10s, and with Jacobi, 11s, which shows that the ILU(0) preconditioning seems the fastest one on this case. Note that this computing time is depending on the unknown orderings. For the bench computations, we used the recursive domain decomposition ordering, which is the most efficient for direct solvers, and the respective computing times with PETSC CG+ILU(0) and with ISTL CG+ILU(0) are 10.3 and 11.2 s. Using the reverse Cuthill - McKee ordering, we respectively obtain 4.4 s and 15.3 s with PETSC CG+ILU(0) and ISTL CG+ILU(0).
2. The computing times, for the conjugate gradient solver of the PETSC library with ILU(1) preconditioning, in the test 1 case on tetrahedral meshes 2 to 5, have been approximately equal to 0.01, 0.03, 0.04, 0.08, and 0.16 s, showing the possibility to apply this method on much larger meshes.

Finally, we may not exclude that the systematic choice of computing the L^2 error with respect to the values in the control volumes instead of the vertex values, makes all these results somewhat pessimistic.

Acknowledgements Work supported by Groupement MOMAS and ANR VFSitCom.

References

1. R. Eymard, T. Gallouët, and R. Herbin. 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(4):1009–1043, 2010. see also http://hal.archives-ouvertes.fr/.
2. R. Eymard, C. Guichard, and R. Herbin. Small-stencil 3D schemes for diffusive flows in porous media. *submitted*, 2010. see also http://hal.archives-ouvertes.fr/.
3. R. Eymard, T. Gallouët and R. Herbin. Benchmark 3D: the SUSHI scheme *these proceedings*, 2011.
4. R. Eymard and R. Herbin. Gradient schemes for diffusion problem. *these proceedings*, 2011.
5. R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klöfkorn and G. Manzini. 3D Benchmark on Discretization Schemes for Anisotropic Diffusion Problems on General Grids. *these proceedings*, 2011.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: The Compact Discontinuous Galerkin 2 Scheme

Robert Klöfkorn

1 The Compact Discontinuous Galerkin 2 Method

In this paper we provide results for the *3d Benchmark on Anisotropic Diffusion Problems*. We consider the Compact Discontinuous Galerkin 2 (CDG2) method first presented in [3]. In [3] a detailed stability analysis as well as a numerical investigation showing that the CDG2 method outperforms other DG methods (e.g. Bassi–Rebay 2, symmetric Interior Penalty, or the original Compact Discontinuous Galerkin Method, see [1, 3] and references therein) in terms of L^2–accuracy versus computational time. Furthermore, the CDG2 method is a parameter free method in the sense that all tests have been calculated with the same set of parameters without specific test case tuning.

In this section we derive the flux and primal formulation of the CDG2 method for a scalar diffusion equation in $\mathbb{R}^d$, $d = 1, 2, 3$ of the form

$$\begin{aligned} -\nabla \cdot (\mathbf{K}\nabla u) &= f \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega, \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ is a bounded polygonal area, $\mathbf{K} \in L^\infty(\Omega, \mathbb{R}^{d\times d})$ a positive definite diffusion matrix, and $f \in L^2(\Omega)$.

For a given partition $\mathcal{T}_h = \{E\}$ of Ω into polygons E, we look for an approximation u_h of u such that $u_h \in V_h^l$ and

$$V_h^l = \{\mathbf{v} \in L^\infty(\Omega, \mathbb{R}^l) \;\; : \;\; \mathbf{v}|_E \in [\mathbb{P}_k(E)]^l \;\} \quad \text{for some } l \in \mathbb{N}\,.$$

Robert Klöfkorn
Section of Applied Mathematics, University of Freiburg, Hermann-Herder-Strasse 10, D-79104 Freiburg, Germany, e-mail: robertk@mathematik.uni-freiburg.de

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_100, © Springer-Verlag Berlin Heidelberg 2011

In the following we use the abbreviations $V_h = V_h^1$ and $\Sigma_h = V_h^d$. Let Γ_i be the family of all interior intersections of elements $E_e^+, E_e^- \in \mathscr{T}_h$ with $e = E_e^- \cap E_e^+$ and Hausdorff measure $\mathscr{H}_{d-1}(e) > 0$. Similarly, we define Γ_D to be the family of all intersections of elements E with $\partial\Omega$. We denote $\Gamma = \Gamma_i \cup \Gamma_D$. For $e \in \Gamma_i$, $\varphi \in V_h$, and $\tau \in \Sigma_h$ we introduce operators $[[\varphi]]_e = \varphi_{|E_e^-}\mathbf{n}_{E_e^-} + \varphi_{|E_e^+}\mathbf{n}_{E_e^+}$, $\{\varphi\}_e = \frac{1}{2}(\varphi_{|E_e^-} + \varphi_{|E_e^+})$, $[\tau]_e = \tau_{|E_e^-} \cdot \mathbf{n}_{E_e^-} + \tau_{|E_e^+} \cdot \mathbf{n}_{E_e^+}$, and $\{\{\tau\}\}_e = \frac{1}{2}(\tau_{|E_e^-} + \tau_{|E_e^+})$ and for $e \subset \partial\Omega$ we set $[[\varphi]]_e = \varphi\mathbf{n}$, $\{\varphi\}_e = \varphi$, $[\tau]_e = \tau \cdot \mathbf{n}$, and $\{\{\tau\}\}_e = \tau$.

The DG method in flux formulation is derived by introducing an auxiliary variable σ such that

$$-\nabla \cdot (\mathbf{K}\sigma) = f \quad \text{in } \Omega \qquad \text{and} \qquad \sigma = \nabla u \quad \text{in } \Omega. \tag{2}$$

Multiplying (2) by arbitrary $\varphi \in V_h$ and $\tau \in \Sigma_h$, respectively, integrating over E, and summing up over all $E \in \mathscr{T}_h$ we arrive at the *flux formulation* of the DG method on the whole domain $\Omega = \bigcup_{E \in \mathscr{T}_h} E$:

$$-\int_\Omega \mathbf{K}\sigma_h \cdot \nabla_h\varphi = \int_\Omega \varphi f - \sum_{e \in \Gamma} \int_e \widehat{\sigma}(u_h) \cdot [[\varphi]]_e \qquad \forall\, \varphi \in V_h, \tag{3a}$$

$$\int_\Omega \tau \cdot \sigma_h = -\int_\Omega \nabla_h \cdot \tau u_h + \sum_{e \in \Gamma} \int_e [\tau]_e \widehat{u}(u_h) \qquad \forall\, \tau \in \Sigma_h, \tag{3b}$$

where $\nabla_h v \in \Sigma_h$ is a function whose restriction to each element $E \in \mathscr{T}_h$ is equal to ∇v. Furthermore, $\widehat{u}(u_h)$ and $\widehat{\sigma}(u_h)$ are numerical fluxes, where the second is an approximation of the diffusive flux $\mathbf{K}\sigma$, over the boundaries of E where we allow the solution to be discontinuous. The method is completely described once the physical parameters f and $\mathbf{K}$ are known and appropriate numerical fluxes have been chosen. Here, we present the numerical fluxes for the CDG2 method. Other possible choices can be found in [1, 3]. To describe the numerical fluxes we introduce two kinds of *lifting operators* $r_e : [L^2(\Gamma)]^d \to \Sigma_h$ and $l_e : L^2(\Gamma_i) \to \Sigma_h$ with

$$\int_\Omega r_e(\boldsymbol{\xi}) \cdot \tau = -\int_e \boldsymbol{\xi} \cdot \{\{\tau\}\}_e, \quad \int_\Omega l_e(\phi) \cdot \tau = -\int_e \phi[\tau]_e, \tag{4}$$

for all $\tau \in \Sigma_h$. We extend l_e for $e \subset \partial\Omega$ by setting $l_e(\phi) \equiv 0$ on $\partial\Omega$ for all $\phi \in L^2(\Gamma)$. For convenience we define $L_e(u) := r_e([[u]]_e) + l_e(\boldsymbol{\beta}_e \cdot [[u]]_e)$ on Γ. The parameter $\boldsymbol{\beta}_e$ (frequently denoted by $\mathbf{C}_{12}$ in the literature) is called the *switch function* and is defined by $\boldsymbol{\beta}_e = \frac{1}{2}\mathbf{n}_{E_e^*}$, where $E_e^* \in \{E_e^+, E_e^-\}$ is the element adjacent to e with the smaller volume. For this switch one can show that L_e only has support on either E_e^- or E_e^+ (see [3] for details). The fluxes for the CDG2 method are

$$\widehat{u}(u) = \{u\}_e, \quad \widehat{\sigma}(u) = \{\{\mathbf{K}\nabla_h u\}\}_e + \chi_e\big(\{\{\mathbf{K}L_e(u)\}\}_e + \boldsymbol{\beta}_e[\mathbf{K}L_e(u)]_e\big) - \boldsymbol{\delta}_e(u), \tag{5}$$

with $\boldsymbol{\delta}_e(u) = \eta_e(\mathbf{K})[[u]]_e \geq 0$, $\chi_e > 0$, and the switch function $\boldsymbol{\beta}_e$ as described above. Using this switch the method is proven to be stable for any $\chi_e \geq N_{\mathscr{T}_h}/2$ (cf. [3]), where $N_{\mathscr{T}_h}$ is the maximal number of intersections an element $E \in \mathscr{T}_h$ can have.

Since the computation of σ_h might be expensive in terms of computation time as well as memory consumptions one might be interested in deriving a *primal formulation* of the form

$$B(u_h, \varphi) = L(\varphi) \qquad \forall\, \varphi \in V_h, \tag{6}$$

where any solution $u_h \in V_h$ of equation (6) also solves equation (3) and vice versa. The derivation of the primal formulation is, for example, described in [1, 3]. The basic idea is to express σ through u via

$$\sigma = \sigma(u) := \nabla u + \sum_{e \in \Gamma} r_e([[u]]_e) + l_e(\{u - \widehat{u}(u)\}_e) \overset{(5)}{=} \nabla u + \sum_{e \in \Gamma} r_e([[u]]_e).$$

Using the numerical fluxes and $\sigma(u)$ (for details see [1, 3]) we arrive at the primal form

$$\begin{aligned} B(u_h, \varphi) := & \int_\Omega \mathbf{K}\nabla u_h \cdot \nabla\varphi + \sum_{e \in \Gamma} \chi_e \int_\Omega \mathbf{K} L_e(u_h) \cdot L_e(\varphi) \\ & - \sum_{e \in \Gamma} \int_e \left(\{\{\mathbf{K}\nabla u_h\}\}_e \cdot [[\varphi]]_e + \{\{\mathbf{K}\nabla\varphi\}\}_e \cdot [[u_h]]_e\right) \\ & - \sum_{e \in \Gamma} \int_e \boldsymbol{\delta}_e(u_h) \cdot [[\varphi]]_e \qquad \forall\, \varphi \in V_h\ , \end{aligned} \tag{7a}$$

$$L(\varphi) := \int_\Omega f\varphi + \sum_{e \in \Gamma_D} \int_e \mathbf{n}_e \cdot \Big(\boldsymbol{\delta}_e(g)\varphi - \mathbf{K}\big(g\nabla\varphi + \chi_e L_e(g)\varphi\big)\Big)\ \forall\, \varphi \in V_h \tag{7b}$$

Note that by choosing $\chi_e = 0\ \forall\, e \in \Gamma$ in (7a) and (7b) we obtain the well known symmetric Interior Penalty Galerkin (SIPG) method. Finally, we need to specify $\eta_e\big(\mathbf{K}\big)$. The proof of stability of the CDG2 method given in [3] shows that for simple $\mathbf{K}$ one can choose $\eta_e\big(\mathbf{K}\big) = 0$. However, in case $\mathbf{K}$ jumps across element interfaces, this parameter can be used to increase the accuracy of the method. For all numerical tests presented in this paper we used

$$\eta_e\big(\mathbf{K}\big) := \begin{cases} \Lambda\big(\{\{\mathbf{K}\}\}_e\big)/|e| & \text{if } e \in \Gamma_i \text{ and } \big|\Lambda(\mathbf{K}^+_{|e}) - \Lambda(\mathbf{K}^-_{|e})\big| > 0, \\ 0 & \text{otherwise.} \end{cases}$$

with $\Lambda(\mathbf{K}) = (\lambda^{max}_{\mathbf{K}})^2/\lambda^{min}_{\mathbf{K}}$, where $\lambda^{max}_{\mathbf{K}}$, $\lambda^{min}_{\mathbf{K}}$ denote the maximal and minimal eigenvalues of $\mathbf{K}$, respectively.

2 Numerical results

The implementation of the CDG2 scheme is based on the discretization framework DUNE–FEM (see [4]) which is a module of DUNE [2]. Among other things DUNE provides an interface for implementations of discretization grids.

In the following we present results for Test 1, 3, 4, and 5 of the benchmark. Since there is no grid implementation of the DUNE grid interface that can handle Voronoi cells, Test 1 with Voronoi meshes and Test 2 could not be computed.

We use three different versions of the CDG2 method differing in the choice of the basis functions used to build V_h and Σ_h. The first, and most commonly used, is the CDG2–$\mathbb{P}_k$, k being the polynomial order. On hexahedral meshes we also provide results of two alternative schemes. First, there is the possibility to use tensor product Legendre polynomials as basis functions which is denoted by CDG2–$\mathbb{Q}_k$. Another possibility is to split each hexahedron into 6 tetrahedra resulting in a conforming tetrahedral mesh, if possible. This scheme is denoted CDG2–$\mathbb{P}_k$(tetra). This feature is implemented in DUNE and could be used in this case without a complicated mesh generation procedure. An advantage of this approach is that reference mappings will be linear. A disadvantage is that the number of cells and unknowns are increased by a factor of 6. We use this scheme only for the tests with complicated meshes, i.e. Test 1 (Kershaw), Test 3, and Test 4.

Although basis functions up to order 8 are implemented we restrict ourself to $k = 1, 2$. While $k = 1$ allows a direct comparison with Finite Volume schemes, $k = 2$ shows the potential of a higher order approach. The integrals in (7a) and (7b) have been calculated using quadratures of order $2k$, $2k + 1$ for element and face integrals, respectively. If the reference mapping of the element is nonlinear then the quadrature orders have been increased by 2. The calculation of the relative L^2–error as well as the energy norm is straightforward in the DG context. For the calculation of the H^1–error the auxiliary variable σ_h given by equation (3b) has been used. The quadrature order for evaluation of the errors is $2k + 4$ ($+2$ in case of nonlinear reference mappings). The convergence order of the schemes is $k + 1$ in the L^2–norm and k in the H^1–norm.

In the following computations we used $\chi_e = 2$ for tetrahedral grids and $\chi_e = 3$ for hexahedral grids, also in the non-conforming cases where χ_e would be larger when using the theoretical values presented in Section 1.

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi \left(y + \frac{1}{2}\right)\right) \sin\left(\pi \left(z + \frac{1}{3}\right)\right)$ min $= 0$, max $= 2$**, Tetrahedral meshes**

Table 1, **CDG2–$\mathbb{P}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	8012	150576	-1.54E-02	0.00E+00	2.017	1.989	1.783
2	15592	297376	-1.22E-02	0.00E+00	2.016	1.988	1.789
3	30844	593648	-7.41E-03	0.00E+00	2.005	1.993	1.792
4	61064	1184192	-3.12E-03	0.00E+00	1.999	1.997	1.794
5	121920	2379936	0.00E+00	0.00E+00	2.002	1.997	1.795
6	244208	4791872	-6.63E-04	0.00E+00	2.002	1.997	1.796

Table 1, **CDG2–$\mathbb{P}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	20030	941100	0.00E+00	0.00E+00	1.999	1.989	1.800
2	38980	1858600	0.00E+00	0.00E+00	1.999	1.988	1.799
3	77110	3710300	0.00E+00	0.00E+00	1.999	1.993	1.799
4	152660	7401200	0.00E+00	0.00E+00	1.999	1.997	1.798
5	304800	14874600	0.00E+00	0.00E+00	1.999	1.997	1.798
6	610520	29949200	0.00E+00	0.00E+00	1.999	1.997	1.798

Table 2, **CDG2–$\mathbb{P}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	8012	9.01E-03	—	1.75E-01	—	1.94E-01	—
2	15592	5.78E-03	2.001	1.40E-01	0.999	1.56E-01	0.979
3	30844	3.68E-03	1.981	1.12E-01	0.968	1.24E-01	0.998
4	61064	2.38E-03	1.925	8.94E-02	1.004	9.87E-02	1.009
5	121920	1.52E-03	1.943	7.09E-02	1.005	7.86E-02	0.989
6	244208	9.45E-04	2.049	5.61E-02	1.007	6.20E-02	1.021

Table 2, **CDG2–$\mathbb{P}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	20030	6.32E-04	—	1.77E-02	—	1.95E-02	—
2	38980	3.24E-04	3.012	1.13E-02	1.996	1.24E-02	2.039
3	77110	1.60E-04	3.095	6.94E-03	2.165	7.71E-03	2.097
4	152660	8.03E-05	3.035	4.40E-03	2.001	4.90E-03	1.994
5	304800	4.10E-05	2.913	2.80E-03	1.958	3.12E-03	1.961
6	610520	2.04E-05	3.022	1.75E-03	2.026	1.95E-03	2.023

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0, \ \max = 2$**, Kershaw meshes**

Table 1, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	12288	233472	-2.81E-02	0.00E+00	2.012	1.951	1.745
2	98304	1916928	-7.22E-03	0.00E+00	1.996	1.998	1.747
3	786432	15532032	-1.17E-03	0.00E+00	2.000	1.999	1.762
4	6291456	125042688	-4.65E-04	0.00E+00	2.000	1.999	1.780

Table 1, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	30720	1459200	0.00E+00	0.00E+00	1.995	1.951	1.779
2	245760	11980800	0.00E+00	0.00E+00	1.998	1.998	1.790
3	1966080	97075200	0.00E+00	0.00E+00	1.999	1.999	1.797
4	15728640	781516800	0.00E+00	0.00E+00	1.999	1.999	1.798

Table 1, **CDG2–$\mathbb{Q}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	4096	204800	-2.95E-02	0.00E+00	2.016	1.958	1.781
2	32768	1736704	-8.49E-03	0.00E+00	1.999	1.992	1.778
3	262144	14286848	-2.86E-04	0.00E+00	2.001	1.996	1.783
4	2097152	115867648	-5.37E-04	0.00E+00	2.000	1.999	1.790

Table 1, **CDG2–$\mathbb{Q}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	13824	2332800	0.00E+00	0.00E+00	1.997	1.958	1.793
2	110592	19782144	0.00E+00	0.00E+00	1.999	1.992	1.795
3	884736	162736128	0.00E+00	0.00E+00	1.999	1.996	1.798

Table 2, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	12288	8.29E-02	—	6.25E-01	—	6.12E-01	—
2	98304	5.47E-02	0.600	4.28E-01	0.547	4.04E-01	0.601
3	786432	3.07E-02	0.831	2.74E-01	0.643	2.48E-01	0.702
4	6291456	1.34E-02	1.194	1.56E-01	0.809	1.39E-01	0.841

Table 2, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	30720	3.72E-02	—	2.74E-01	—	2.53E-01	—
2	245760	9.29E-03	2.002	1.06E-01	1.365	1.02E-01	1.310
3	1966080	1.31E-03	2.831	3.12E-02	1.771	3.15E-02	1.695
4	15728640	1.07E-04	3.611	7.98E-03	1.966	8.32E-03	1.919

Table 2, **CDG2–$\mathbb{Q}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	4096	7.09E-02	—	6.08E-01	—	5.88E-01	—
2	32768	4.77E-02	0.574	4.25E-01	0.516	3.97E-01	0.565
3	262144	2.61E-02	0.870	2.74E-01	0.636	2.57E-01	0.630
4	2097152	1.09E-02	1.262	1.56E-01	0.812	1.50E-01	0.774

Table 2, **CDG2–$\mathbb{Q}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	13824	3.06E-02	—	2.44E-01	—	2.15E-01	—
2	110592	7.57E-03	2.012	1.03E-01	1.240	9.73E-02	1.141
3	884736	1.04E-03	2.865	3.22E-02	1.682	3.18E-02	1.615

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

Table 1, **CDG2–$\mathbb{P}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	144	3648	0.00E+00	0.00E+00	1.901	1.846	1.538
2	1152	35328	-8.80E-03	0.00E+00	2.006	1.959	1.698
3	9216	307200	-6.73E-03	0.00E+00	2.005	1.989	1.765
4	73728	2555904	-1.86E-03	0.00E+00	2.001	1.997	1.788
5	589824	20840448	-5.50E-04	0.00E+00	2.000	1.999	1.795

Table 1, **CDG2–$\mathbb{P}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	360	22800	-3.34E-02	0.00E+00	2.050	1.846	1.752
2	2880	220800	-7.34E-04	0.00E+00	2.000	1.959	1.794
3	23040	1920000	0.00E+00	0.00E+00	1.999	1.989	1.798
4	184320	15974400	0.00E+00	0.00E+00	1.999	1.997	1.798
5	1474560	130252800	0.00E+00	0.00E+00	1.999	1.999	1.798

Table 1, **CDG2–$\mathbb{Q}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	288	14592	-7.94E-02	0.00E+00	2.081	1.846	1.581
2	2304	141312	-1.77E-02	0.00E+00	2.017	1.959	1.744
3	18432	1228800	-4.59E-03	0.00E+00	2.004	1.989	1.785
4	147456	10223616	-1.27E-03	0.00E+00	2.001	1.997	1.795
5	1179648	83361792	-3.06E-04	0.00E+00	2.000	1.999	1.797

Table 1, **CDG2–$\mathbb{Q}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	972	166212	0.00E+00	0.00E+00	1.998	1.846	1.807
2	7776	1609632	0.00E+00	0.00E+00	1.999	1.959	1.800
3	62208	13996800	0.00E+00	0.00E+00	1.999	1.989	1.798
4	497664	116453376	0.00E+00	0.00E+00	1.999	1.997	1.798

Table 2, **CDG2–$\mathbb{P}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	144	1.08E-01	—	5.66E-01	—	5.70E-01	—
2	1152	3.70E-02	1.549	3.32E-01	0.770	3.46E-01	0.720
3	9216	1.12E-02	1.719	1.76E-01	0.917	1.79E-01	0.946
4	73728	3.09E-03	1.860	8.95E-02	0.974	9.05E-02	0.987
5	589824	8.01E-04	1.950	4.51E-02	0.990	4.54E-02	0.995

Table 2, **CDG2–$\mathbb{P}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	360	3.93E-02	—	2.88E-01	—	3.07E-01	—
2	2880	7.29E-03	2.432	8.86E-02	1.699	8.56E-02	1.845
3	23040	8.62E-04	3.080	2.24E-02	1.982	2.18E-02	1.974
4	184320	1.05E-04	3.033	5.59E-03	2.004	5.46E-03	1.997
5	1474560	1.31E-05	3.007	1.40E-03	2.001	1.37E-03	1.999

Table 2, **CDG2–$\mathbb{Q}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	288	6.31E-02	—	3.81E-01	—	3.77E-01	—
2	2304	1.83E-02	1.789	1.96E-01	0.960	1.94E-01	0.959
3	18432	4.63E-03	1.980	9.86E-02	0.989	9.82E-02	0.983
4	147456	1.16E-03	1.998	4.96E-02	0.993	4.94E-02	0.991
5	1179648	2.89E-04	2.001	2.49E-02	0.996	2.48E-02	0.995

Table 2, **CDG2–$\mathbb{Q}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	972	7.05E-03	—	7.17E-02	—	7.03E-02	—
2	7776	1.02E-03	2.782	1.87E-02	1.936	1.84E-02	1.935
3	62208	1.36E-04	2.910	4.75E-03	1.981	4.68E-03	1.974
4	497664	1.95E-05	2.803	1.19E-03	1.993	1.18E-03	1.989

• Test 3 Flow on random meshes, $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$, $\min = -1,\ \max = 1$, Random meshes

Table 1, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1536	27648	-1.261	-7.43E-01	1.167	8.36E-01	2.757
2	12288	233472	-1.009	-9.35E-01	1.033	9.33E-01	3.198
3	98304	1916928	-1.016	-9.82E-01	1.017	9.84E-01	3.484
4	786432	15532032	-1.008	-9.95E-01	1.002	9.96E-01	3.568

Table 1, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3840	172800	-1.238	-7.43E-01	1.295	8.36E-01	3.637
2	30720	1459200	-1.042	-9.35E-01	1.028	9.33E-01	3.577
3	245760	11980800	-1.000	-9.82E-01	1.000	9.84E-01	3.598
4	1966080	97075200	-1.000	-9.95E-01	1.000	9.96E-01	3.596

Table 1, **CDG2–$\mathbb{Q}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	512	22528	-1.143	-7.59E-01	1.244	6.91E-01	3.016
2	4096	204800	-1.076	-9.39E-01	1.074	9.23E-01	3.432
3	32768	1736704	-1.026	-9.85E-01	1.021	9.82E-01	3.564
4	262144	14286848	-1.009	-9.96E-01	1.000	9.96E-01	3.587

Table 1, **CDG2–$\mathbb{Q}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1728	256608	-1.015	-7.59E-01	1.034	6.91E-01	3.635
2	13824	2332800	-1.002	-9.39E-01	9.95E-01	9.23E-01	3.568
3	110592	19782144	-1.000	-9.85E-01	9.99E-01	9.82E-01	3.596
4	884736	162736128	-1.000	-9.96E-01	1.000	9.96E-01	3.595

Table 2, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1536	3.49E-01	—	5.82E-01	—	5.80E-01	—
2	12288	1.17E-01	1.572	3.43E-01	0.764	3.07E-01	0.920
3	98304	3.48E-02	1.751	2.00E-01	0.777	1.60E-01	0.935
4	786432	9.59E-03	1.861	1.08E-01	0.892	8.12E-02	0.981

Table 2, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3840	1.05E-01	—	3.08E-01	—	2.21E-01	—
2	30720	1.66E-02	2.659	9.93E-02	1.632	5.82E-02	1.925
3	245760	2.48E-03	2.739	2.80E-02	1.829	1.54E-02	1.917
4	1966080	3.06E-04	3.021	6.50E-03	2.106	3.90E-03	1.980

Table 2, **CDG2–$\mathbb{Q}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	512	3.05E-01	—	4.98E-01	—	5.00E-01	—
2	4096	8.38E-02	1.862	2.58E-01	0.947	2.53E-01	0.982
3	32768	2.17E-02	1.952	1.30E-01	0.988	1.25E-01	1.015
4	262144	5.86E-03	1.885	6.72E-02	0.954	6.31E-02	0.988

Table 2, **CDG2–$\mathbb{Q}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1728	4.41E-02	—	1.25E-01	—	1.14E-01	—
2	13824	6.02E-03	2.875	2.97E-02	2.067	2.82E-02	2.012
3	110592	8.07E-04	2.897	7.58E-03	1.971	7.16E-03	1.980
4	884736	1.03E-04	2.968	1.91E-03	1.993	1.78E-03	2.009

- **Test 4 Flow around a well, Well meshes,** $\min = 0,\ \max = 5.415$

Table 1, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	21360	369664	0.00E+00	0.00E+00	5.406	5.358	1633.939
2	53568	941888	0.00E+00	0.00E+00	5.408	5.366	1628.212
3	120384	2168128	0.00E+00	0.00E+00	5.408	5.368	1626.468
4	269280	4963456	0.00E+00	0.00E+00	5.408	5.371	1626.554
5	557040	10433536	0.00E+00	0.00E+00	5.409	5.376	1625.759
6	1023192	19371936	0.00E+00	0.00E+00	5.409	5.379	1624.751
7	1792296	34218592	0.00E+00	0.00E+00	5.410	5.387	1624.488

Table 1, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	53400	2310400	-5.92E-03	0.00E+00	5.414	5.358	1623.061
2	133920	5886800	-1.61E-03	0.00E+00	5.414	5.366	1622.951
3	300960	13550800	-1.42E-04	0.00E+00	5.414	5.368	1623.127
4	673200	31021600	0.00E+00	0.00E+00	5.414	5.371	1623.224
5	1392600	65209600	0.00E+00	0.00E+00	5.414	5.376	1623.329
6	2557980	121074600	0.00E+00	0.00E+00	5.414	5.379	1623.423
7	4480740	213866200	0.00E+00	0.00E+00	5.414	5.387	1623.472

Table 1, **CDG2–$\mathbb{Q}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	7120	356736	0.00E+00	0.00E+00	5.406	5.316	1685.638
2	17856	931328	0.00E+00	0.00E+00	5.407	5.328	1653.938
3	40128	2139904	0.00E+00	0.00E+00	5.407	5.328	1639.060
4	89760	4857216	0.00E+00	0.00E+00	5.407	5.330	1633.474
5	185680	10136320	0.00E+00	0.00E+00	5.408	5.339	1629.453
6	341064	18717760	0.00E+00	0.00E+00	5.409	5.345	1627.258
7	597432	32900416	0.00E+00	0.00E+00	5.410	5.360	1626.055

Table 1, **CDG2–$\mathbb{Q}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	24030	4063446	0.00E+00	0.00E+00	5.408	5.316	1623.988
2	60264	10608408	-5.25E-04	0.00E+00	5.409	5.328	1623.197
3	135432	24374844	0.00E+00	0.00E+00	5.409	5.328	1623.222
4	302940	55326726	0.00E+00	0.00E+00	5.409	5.330	1623.226
5	626670	115459020	0.00E+00	0.00E+00	5.410	5.339	1623.337
6	1151091	213206985	0.00E+00	0.00E+00	5.410	5.345	1623.423
7	2016333	374756301	0.00E+00	0.00E+00	5.411	5.360	1623.446

Table 2, **CDG2–$\mathbb{P}_1$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	21360	1.83E-03	—	1.69E-01	—	2.07E-01	—
2	53568	1.03E-03	1.878	1.13E-01	1.316	1.41E-01	1.259
3	120384	6.75E-04	1.570	7.87E-02	1.328	9.92E-02	1.302
4	269280	5.63E-04	0.675	6.12E-02	0.938	7.65E-02	0.968
5	557040	4.16E-04	1.253	4.83E-02	0.980	6.02E-02	0.991
6	1023192	2.70E-04	2.119	3.93E-02	1.015	4.91E-02	1.005
7	1792296	2.24E-04	1.014	3.34E-02	0.862	4.15E-02	0.905

Table 2, **CDG2–$\mathbb{P}_2$(tetra)**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	53400	2.46E-04	—	3.53E-02	—	4.46E-02	—
2	133920	9.83E-05	2.993	1.60E-02	2.569	2.06E-02	2.530
3	300960	4.37E-05	3.005	7.94E-03	2.607	1.02E-02	2.606
4	673200	2.91E-05	1.517	4.68E-03	1.973	5.95E-03	1.997
5	1392600	1.72E-05	2.162	2.92E-03	1.944	3.69E-03	1.978
6	2557980	9.30E-06	3.037	1.92E-03	2.075	2.43E-03	2.050
7	4480740	6.79E-06	1.684	1.42E-03	1.598	1.77E-03	1.710

Table 2, **CDG2–$\mathbb{Q}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	7120	6.22E-03	—	2.47E-01	—	2.46E-01	—
2	17856	2.92E-03	2.461	1.70E-01	1.232	1.69E-01	1.220
3	40128	1.45E-03	2.593	1.19E-01	1.312	1.19E-01	1.304
4	89760	9.46E-04	1.595	9.08E-02	1.008	9.09E-02	1.006
5	185680	5.92E-04	1.937	7.07E-02	1.032	7.08E-02	1.030
6	341064	3.63E-04	2.419	5.71E-02	1.055	5.72E-02	1.056
7	597432	2.68E-04	1.610	4.77E-02	0.962	4.78E-02	0.960

Table 2, **CDG2–$\mathbb{Q}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	24030	3.48E-04	—	3.93E-02	—	3.93E-02	—
2	60264	1.07E-04	3.843	1.61E-02	2.912	1.61E-02	2.910
3	135432	4.55E-05	3.175	7.48E-03	2.838	7.49E-03	2.834
4	302940	3.03E-05	1.511	4.40E-03	1.978	4.40E-03	1.977
5	626670	1.70E-05	2.383	2.64E-03	2.101	2.65E-03	2.098
6	1151091	8.46E-06	3.451	1.69E-03	2.194	1.70E-03	2.194
7	2016333	5.65E-06	2.159	1.17E-03	1.968	1.18E-03	1.965

- **Test 5 Discontinuous permeability,**
$u(x,y,z) = a_i \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -100$, $\max = 100$**, Locally refined meshes**

Table 1, **CDG2–$\mathbb{P}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	88	1984	-8.502	-100.000	8.502	100.000	10.584
2	704	17792	-52.177	-35.355	52.177	35.355	46.202
3	5632	150016	-83.767	-78.858	83.767	78.858	68.596
4	45056	1230848	-95.672	-94.345	95.672	94.345	80.340
5	360448	9969664	-98.927	-98.562	98.927	98.562	88.092

Table 1, **CDG2–$\mathbb{P}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	220	12400	-18.447	-100.000	18.447	100.000	48.277
2	1760	111200	-99.735	-35.355	99.735	35.355	92.615
3	14080	937600	-102.184	-78.858	102.184	78.858	99.167
4	112640	7692800	-100.544	-94.345	100.544	94.345	99.221
5	901120	62310400	-100.098	-98.562	100.098	98.562	99.100

Table 1, **CDG2–$\mathbb{Q}_1$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	176	7936	-12.747	-100.000	12.747	100.000	8.513
2	1408	71168	-118.320	-35.355	118.320	35.355	79.760
3	11264	600064	-103.899	-78.858	103.899	78.858	94.412
4	90112	4923392	-100.970	-94.345	100.970	94.345	97.877
5	720896	39878656	-100.241	-98.562	100.241	98.562	98.740

Table 1, **CDG2–$\mathbb{Q}_2$**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	594	90396	-94.815	-100.000	94.815	100.000	106.736
2	4752	810648	-100.376	-35.355	100.376	35.355	100.759
3	38016	6835104	-99.836	-78.858	99.836	78.858	99.387
4	304128	56080512	-99.951	-94.345	99.951	94.345	99.084
5	2433024	454242816	-99.987	-98.562	99.987	98.562	99.024

Table 2, **CDG2–$\mathbb{P}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	88	9.26E-01	—	9.86E-01	—	1.003	—
2	704	6.95E-01	0.412	8.49E-01	0.216	7.90E-01	0.345
3	5632	2.86E-01	1.283	4.95E-01	0.777	4.30E-01	0.879
4	45056	9.88E-02	1.532	2.84E-01	0.802	2.08E-01	1.045
5	360448	2.99E-02	1.724	1.39E-01	1.034	1.01E-01	1.042

Table 2, **CDG2–$\mathbb{P}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	220	7.70E-01	—	8.98E-01	—	7.83E-01	—
2	1760	3.18E-01	1.276	5.07E-01	0.825	4.17E-01	0.909
3	14080	5.14E-02	2.629	1.77E-01	1.519	1.09E-01	1.939
4	112640	5.63E-03	3.191	4.87E-02	1.862	2.56E-02	2.086
5	901120	7.55E-04	2.898	1.57E-02	1.633	6.27E-03	2.031

Table 2, **CDG2–$\mathbb{Q}_1$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	176	9.16E-01	—	1.001	—	1.005	—
2	1408	2.36E-01	1.960	4.57E-01	1.133	4.54E-01	1.146
3	11264	6.32E-02	1.897	2.27E-01	1.007	2.27E-01	1.001
4	90112	1.61E-02	1.970	1.13E-01	1.003	1.13E-01	1.001
5	720896	4.06E-03	1.992	5.67E-02	1.001	5.67E-02	1.000

Table 2, **CDG2–$\mathbb{Q}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	594	6.22E-02	—	1.51E-01	—	1.42E-01	—
2	4752	2.89E-02	1.107	9.46E-02	0.670	9.23E-02	0.626
3	38016	3.80E-03	2.925	2.36E-02	2.002	2.32E-02	1.993
4	304128	4.92E-04	2.949	5.83E-03	2.019	5.78E-03	2.002
5	2433024	6.39E-05	2.945	1.45E-03	2.011	1.44E-03	2.002

3 Comments

For all tests the solver reduction tolerance was taken to be 10^{-10}. Test 1 (except Kershaw) and Test 5 were uncritical for all meshes. The solution was a matter of minutes rather than hours. Test 3 was the one that was most difficult to solve. Here, solving took a long time and the only combination of solver and preconditioning that produced results was GMRES + JACOBI. For all other tests practically any combination of solver and preconditioning from the solver–bench package produced good results. The scheme CDG2–$\mathbb{P}_k$(tetra) seems to be a good alternative to the scheme CDG2–$\mathbb{Q}_k$, since the reference mappings are linear and quadratures of lower order can be used. Also, the solution of the resulting linear system seemed to be much easier (hours rather than days). This is payed by a factor of 6 more elements and, in case of $k = 2$ by a factor of 3 more DoFs. For CDG2–$\mathbb{Q}_2$ we had to skip the last mesh of Test 1 (checkerboard) due to memory limitations.

Using CDG2–$\mathbb{P}_k$ for Test 3 and 4 we were not able to produce satisfying results. The next table shows the L^2**–projection** of the exact solution of **Test 3** onto V_h using the **random mesh** series:

Table 2, **CDG2–$\mathbb{P}_2$**

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	640	1.39E-01	—	3.09E-01	—	3.85E-01	—
2	5120	2.31E-02	2.586	9.51E-02	1.699	1.28E-01	1.591
3	40960	6.30E-03	1.875	4.65E-02	1.033	6.45E-02	0.987
4	327680	2.76E-03	1.189	3.99E-02	0.222	5.26E-02	0.294

As we can see, the space $\mathbb{P}_k$ seems not to be suitable to project the solution properly onto the discrete spaces. The convergence does not show the expected *ratiol2* of $k + 1$.

To conclude, with the CDG2 scheme all test cases (except Test 2, see Section 2 for explanation) could have been computed. The CDG2 provides a higher order alternative to solve anisotropic diffusion problems in 3d without specific parameter tuning required by the user. Finally, I would like to thank Peter Bastian for fruitful discussions about the test cases and for providing the test case implementation which saved a lot of time.

References

1. D.N. Arnold, F. Brezzi, B. Cockburn, and L.D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.
2. P. Bastian, M. Blatt, A. Dedner, C. Engwer, R. Klöfkorn, R. Kornhuber, M. Ohlberger, and O. Sander. A generic grid interface for parallel and adaptive scientific computing. II: Implementation and tests in dune. *Computing*, 82(2-3):121–138, 2008.
3. S. Brdar, A. Dedner, and R. Klöfkorn. Compact and stable Discontinuous Galerkin methods for convection-diffusion problems. Preprint No. 2/2010-15.11-2010, Mathematisches Institut, Universität Freiburg, 2010. submitted to SIAM J. Sci. Comput.
4. A. Dedner, R. Klöfkorn, M. Nolte, and M. Ohlberger. A generic interface for parallel and adaptive discretization schemes: abstraction principles and the DUNE–FEM; module. *Computing*, 90:165–196, 2010.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: Mimetic Finite Difference Method for Generalized Polyhedral Meshes

Konstantin Lipnikov and Gianmarco Manzini

1 Presentation of the scheme

Let Ω be a subset of $\Re^3$ with a Lipschitz continuous boundary. We consider the mixed (velocity-pressure) formulation of the diffusion problem,

$$\mathbf{u} = -\mathbb{K}\,\nabla p \qquad \text{and} \qquad \operatorname{div}\mathbf{u} = b \qquad \text{in} \quad \Omega, \tag{1}$$

subject to Dirichlet boundary conditions on $\partial\Omega$. Here $\mathbb{K}$ is the diffusion tensor and b is the source function.

Let Ω_h be a conformal partition of Ω into generalized polyhedral elements E. We assume that each element E is shape-regular and satisfies assumptions (M2)–(M3) formulated in [1]. Let f denote a face of a generalized polyhedron E and $|f|$ be its area. Furthermore, let $\mathbf{n}_f(\mathbf{x})$ be a unit normal vector to face f at point $\mathbf{x}$. Direction of $\mathbf{n}_f(\mathbf{x})$ is fixed once and for all. Let $\mathbf{n}_{E,f}(\mathbf{x})$ be a unit normal vector external to E, so that $\mathbf{n}_{E,f}(\mathbf{x}) \cdot \mathbf{n}_f(\mathbf{x}) = \pm 1$. We introduce the average normal to face f as

$$\widetilde{\mathbf{n}}_f = \frac{1}{|f|}\int_f \mathbf{n}_f(\mathbf{x})\,\mathrm{d}A.$$

Maximal deviation of the average normal $\widetilde{\mathbf{n}}_f$ from a pointwise normal characterizes deviation of face f from a planar face. More precisely, we say that a face f is *moderately curved* if

Konstantin Lipnikov
Los Alamos National Laboratory, Theoretical Division, MS B284, Los Alamos, NM 87545, USA, e-mail: lipnikov@lanl.gov

Gianmarco Manzini
IMATI-CNR and CESNA-IUSS, Pavia, Italy, e-mail: marco.manzini@imati.cnr.it

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_101, © Springer-Verlag Berlin Heidelberg 2011

$$\max_{\mathbf{x}\in f} \|\mathbf{n}_f(\mathbf{x}) - \widetilde{\mathbf{n}}_f\| \le \sigma_* |f|^{1/2},$$

where σ_* is a positive constant independent of the mesh. Otherwise, we say that the face is *strongly curved*. For example, for a polyhedral mesh with planar faces, all faces are classified as moderately curved.

Integrating the second equation in (1) over element E and using the divergence theorem, we get

$$\sum_{f\in\partial E} u^h_{E,f}|f| = \int_E b \,\mathrm{d}V, \qquad u^h_{E,f} = \frac{1}{|f|}\int_f \mathbf{u}\cdot\mathbf{n}_{E,f}(\mathbf{x})\,\mathrm{d}A. \tag{2}$$

Thus, it is natural to take average normal components of the velocity $\mathbf{u}$ on mesh faces as discrete unknowns. For a moderately curved face, this is the *sole* unknown representing the velocity $\mathbf{u}$ on this face. If face f is shared by two elements E and E', we impose the following continuity condition

$$u^h_{E,f} = -u^h_{E',f}. \tag{3}$$

For a strongly curved face, regardless of the number of its vertices, we introduce two additional velocity degrees of freedom as

$$u^h_{E,f,i} = \frac{1}{|f|}\int_f \mathbf{u}\cdot\mathbf{a}_{f,i}\,\mathrm{d}A, \qquad i = 2,3,$$

where $\mathbf{a}_{f,i}$ are two arbitrary chosen unit vectors orthogonal to $\widetilde{\mathbf{n}}_f$ (see Fig. 1).

If this face is shared by two elements E and E', we impose the following continuity conditions

$$u^h_{E,f,i} = u^h_{E',f,i}, \qquad i = 2,3. \tag{4}$$

For problems with discontinuous coefficients, it is more natural to define additional discrete unknowns as tangential components of the gradient ∇p, rather than velocity $\mathbf{u}$. Fortunately, in practical applications, material interfaces are composed of moderately curved faces; therefore, we did not investigate other definitions of degrees of freedom.

Taken into account continuity conditions, the total number of discrete velocity unknowns is equal to the number of moderately curved faces plus three times the number of strongly curved faces. The threshold σ_* affects the number of strongly curved faces. Smaller value of σ_* results in a more accurate method at a cost of solving larger system of equations.

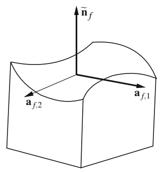


Fig. 1 Local coordinate system associated with a curved face

The scalar (pressure) variable p is represented by its average values over elements E and faces f. For a moderately curved face f, we introduce

$$p_E = \frac{1}{|E|} \int_E p \, \mathrm{d}V \qquad \text{and} \qquad p_{E,f} = \frac{1}{|f|} \int_f (\mathbf{n}_f(\mathbf{x}) \cdot \widetilde{\mathbf{n}}_f) p \, \mathrm{d}A, \tag{5}$$

where $|E|$ denote the volume of element E. For a strongly curved face, we need two additional degrees of freedom to match the number of velocity unknowns:

$$p_{E,f,i} = \frac{1}{|f|} \int_f (\mathbf{n}_f(\mathbf{x}) \cdot \mathbf{a}_{f,i}) p \, \mathrm{d}A, \qquad i = 2, 3. \tag{6}$$

The total number of discrete pressure unknowns is equal to the number of elements plus the number of moderately curved faces plus three times the number of strongly curved faces.

Let us consider an element E with m faces $f_1, \ldots, f_m$. Without loss of generality, we assume that only face f_1 is classified as strongly curved and the other faces are planar, as shown in Fig. 1. We assume that there exists a matrix $\mathbb{W}_E$ and the following linear relations between the discrete unknowns:

$$\begin{bmatrix} u_{E,f_1} \\ u_{E,f_1,2} \\ u_{E,f_1,3} \\ u_{E,f_2} \\ \vdots \\ u_{E,f_m} \end{bmatrix} = \mathbb{W}_E \begin{bmatrix} |f_1| \, (p_{E,f_1} - p_E) \\ |f_1| \, p_{E,f_1,2} \\ |f_1| \, p_{E,f_1,3} \\ |f_2| \, (p_{E,f_2} - p_E) \\ \vdots \\ |f_m| \, (p_{E,f_m} - p_E) \end{bmatrix}. \tag{7}$$

The key of the MFD method is in construction of a proper $(m + 2) \times (m + 2)$ matrix $\mathbb{W}_E$. Let $\mathbb{K}_E$ be a constant tensor approximating tensor $\mathbb{K}$ in element E. In practice, we take $\mathbb{K}_E = \mathbb{K}(\mathbf{x}_E)$, where $\mathbf{x}_E$ is the center of mass of E. We will define the matrix $\mathbb{W}_E$ such that equation (7) is exact for any linear function p and the corresponding constant vector $\mathbf{u}$.

It is trivial for $p = 1$ with $\mathbf{u} = (0, 0, 0)^T$, since the vectors on the left and the right-hand sides are zero vectors. For $p(x, y, z) = x$ with $\mathbf{u} = -\mathbb{K}_E \, (1, 0, 0)^T$, $p(x, y, z) = y$ with $\mathbf{u} = -\mathbb{K}_E \, (0, 1, 0)^T$, and $p(x, y, z) = z$ with $\mathbf{u} = -\mathbb{K}_E \, (0, 0, 1)^T$, we obtain three matrix equations:

$$\mathbb{N}_{E,x} = \mathbb{W}_E \, \mathbb{R}_{E,x}, \quad \mathbb{N}_{E,y} = \mathbb{W}_E \, \mathbb{R}_{E,y} \quad \text{and} \quad \mathbb{N}_{E,z} = \mathbb{W}_E \, \mathbb{R}_{E,z}. \tag{8}$$

The left and right hand-side vectors can be calculated using only geometric data for faces of E which results in relatively simple calculations for an arbitrary-shaped element.

We define $(m + 2) \times 3$ matrices $\mathbb{N} = [\mathbb{N}_{E,x}; \mathbb{N}_{E,y}; \mathbb{N}_{E,z}]$ and $\mathbb{R} = [\mathbb{R}_{E,x}; \mathbb{R}_{E,y}; \mathbb{R}_{E,z}]$. It has been proved in [2] that a particular solution to the matrix equations (8) is

$$\mathbb{W}_{E,0} = \frac{1}{|E|} \mathbb{N}\, \mathbb{K}_E^{-1}\, \mathbb{N}^T.$$

The rank of this matrix is 3 and therefore less than $m + 2$. To build a positive definite $(m + 2) \times (m + 2)$ matrix $\mathbb{W}_E$, we have to add a matrix $\mathbb{W}_{E,1}$ such that $\mathbb{W}_{E,1}\mathbb{R} = 0$. In practice, we take

$$\mathbb{W}_{E,1} = a_E \left(\mathbb{I} - \mathbb{R} \left(\mathbb{R}^T \mathbb{R} \right)^{-1} \mathbb{R}^T \right), \qquad a_E = \frac{\mathrm{trace}(\mathbb{K}_E)}{|E|}.$$

It has been proved in [1] that the matrix $\mathbb{W}_E$ given by

$$\mathbb{W}_E = \mathbb{W}_{E,0} + \mathbb{W}_{E,1} \tag{9}$$

is positive definite. Moreover, its condition number depends only on the anisotropy of tensor $\mathbb{K}_E$ and the shape-regularity of element E.

The mimetic finite difference method is defined by (2), (3), (7), (9), and boundary conditions. The Dirichlet boundary conditions are incorporated in a straightforward manner by prescribing values of integrals in (5) and (6) to corresponding pressure unknowns.

Substituting (7) into (2), (3) and (4), we may easily get an algebraic problem with a sparse symmetric and positive definite matrix. In practice, we also eliminate the cell-based pressure unknowns p_E. The size of the final problem is equal to the number of moderately curved faces plus 3 times the number of strongly curved faces. This number is reported in tables in the next section.

To get a method that is exact for linear solutions, we have to classify all non-planar faces as strongly curved.

Under assumptions of the mesh shape regularity and the solution regularity, the second-order convergence estimate for the pressure p in a discrete norm and the first-order convergence estimate for the velocity $\mathbf{u}$ have been proved in [1].

2 Numerical results

The discrete energy norm is calculated using the inner product of the mass matrices $\mathbb{W}_E^{-1}$ with vectors of discrete velocities $u_{E,f}^h$. Due to the lack of accurate quadrature rules for generalized polyhedra, we used a mid-point quadrature rule. The quadrature points were located at the mass centers of elements. In practice, an error in average pressure values is of greater interest to engineers than an accurate estimate of the exact L^2 error. Instead of the L^1 norm of a discrete gradient, we provide a discrete energy norm, *norme*. Calculation of the discrete gradient is feasible via a post-processing of fluxes; however, such a capability was not available at the moment of writing this paper.

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	4308	28344	2.29E-02	2.03E-02	1.987	1.989	—	1.925
2	8248	55024	2.33E-03	6.84E-03	1.994	1.989	—	1.924
3	16148	108680	7.67E-03	9.13E-03	1.995	1.994	—	1.924
4	31691	214883	3.17E-03	5.52E-03	1.997	1.997	—	1.924
5	62787	428547	2.49E-03	1.49E-03	1.996	1.997	—	1.924
6	124988	857612	1.66E-03	1.83E-03	1.998	1.997	—	1.924

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	4308	5.37E-03		—	—	1.22E-01	
2	8248	3.67E-03	1.758	—	—	9.95E-02	0.942
3	16148	2.26E-03	2.165	—	—	7.48E-02	1.274
4	31691	1.53E-03	1.736	—	—	6.02E-02	0.966
5	62787	9.55E-04	2.068	—	—	4.89E-02	0.912
6	124988	5.96E-04	2.054	—	—	3.81E-02	1.088

• **Test 1 Mild anisotropy,** $u(x, y, z) = 1 + \sin(\pi x)\sin\left(\pi\left(y + \frac{1}{2}\right)\right)\sin\left(\pi\left(z + \frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoi meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	172	2964	-1.36E-02	8.51E-02	1.936	1.870	—	1.906
2	402	7788	-9.00E-02	1.43E-01	1.911	1.854	—	2.128
3	811	17205	1.92E-02	3.85E-02	2.039	1.925	—	2.058
4	1452	32446	-2.92E-02	1.74E-02	1.975	1.914	—	2.046
5	2376	57180	-3.24E-02	2.84E-02	2.055	1.979	—	2.026

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	172	9.11E-02		—	—	6.96E-01	
2	402	1.33E-01	-1.337	—	—	6.44E-01	0.274
3	811	8.59E-02	1.869	—	—	4.46E-01	1.570
4	1452	6.77E-02	1.226	—	—	3.52E-01	1.219
5	2376	5.75E-02	0.995	—	—	2.92E-01	1.138

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	1728	17088	-2.52E-02	3.03E-02	1.973	1.958	—	1.920
2	13056	135936	-5.20E-03	1.06E-02	1.998	1.993	—	1.925
3	101376	1084416	-1.48E-03	1.75E-03	1.998	1.997	—	1.925
4	798720	8663040	2.71E-04	7.14E-04	1.999	1.999	—	1.924

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1728	6.67E-02		—	—	1.45E-00	
2	13056	3.35E-02	1.022	—	—	6.05E-01	1.297
3	101376	1.09E-02	1.643	—	—	1.71E-01	1.850
4	798720	2.79E-03	1.981	—	—	4.42E-02	1.966

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	93	921	2.91E-01	5.17E-01	1.880	1.846	—	2.193
2	636	6588	1.42E-01	1.54E-01	1.968	1.960	—	1.961
3	4656	49584	3.45E-02	4.01E-02	1.992	1.990	—	1.932
4	35520	384192	8.63E-03	1.01E-02	1.998	1.997	—	1.926
5	277248	3023616	2.15E-03	2.54E-03	1.999	1.999	—	1.924

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	93	1.08E-01		—	—	4.34E-01	
2	636	2.26E-02	2.441	—	—	1.13E-01	2.100
3	4656	5.25E-03	2.200	—	—	3.49E-02	1.771
4	35520	1.23E-03	2.143	—	—	1.14E-02	1.652
5	277248	2.89E-04	2.115	—	—	3.83E-03	1.593

- **Test 2 Heterogeneous anisotropy,** $u(x,y,z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)$ $\sin(2\pi z)$, $\min = -0.862,\ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	5331	72291	-0.873	-0.862	0.832	0.831	—	2.794
2	37261	529581	-0.861	-0.861	0.925	0.925	—	2.790
3	119791	1731871	-0.883	-0.883	0.951	0.951	—	2.790
4	276921	4039161	-0.890	-0.890	0.963	0.963	—	2.790

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	5331	4.07E-02		—	—	9.23E-02	
2	37261	1.10E-02	2.019	—	—	2.95E-02	1.760
3	119791	5.02E-03	2.015	—	—	1.49E-02	1.755
4	276921	2.85E-03	2.027	—	—	9.24E-03	1.711

- **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	240	2160	-1.268	-0.756	1.430	0.712	—	
2	1728	17088	-1.184	-0.939	1.397	0.926	—	
3	13056	135936	-1.135	-0.986	1.111	0.982	—	
4	107118	1223876	-1.027	-0.996	1.021	0.996	—	

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	240	9.26E-01		—	—	1.53E+00	
2	1728	3.34E-01	1.550	—	—	8.63E-01	0.870
3	13056	1.06E-01	1.702	—	—	4.84E-01	0.858
4	107118	3.14E-02	1.734	—	—	2.53E-01	0.925

- **Test 4 Flow around a well, Well meshes,** $\min = 0,\ \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	3004	29782	5.37E-01	4.57E-01	5.317	5.317	—	2.08E+01
2	7232	74192	2.90E-01	2.62E-01	5.329	5.329	—	2.16E+01
3	15886	166366	1.74E-01	1.62E-01	5.329	5.329	—	2.18E+01
4	34983	371583	1.29E-01	1.23E-01	5.330	5.330	—	2.19E+01
5	71683	768167	9.66E-02	9.28E-02	5.339	5.339	—	2.20E+01
6	130894	1410160	7.67E-02	7.42E-02	5.345	5.345	—	2.20E+01
7	228463	2471077	5.91E-02	5.75E-02	5.361	5.361	—	2.20E+01

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3004	1.12E-02		—	—	1.40E-01	
2	7232	4.30E-03	3.269	—	—	5.61E-02	3.123
3	15886	1.90E-03	3.114	—	—	2.76E-02	2.704
4	34983	1.04E-03	2.290	—	—	1.62E-02	2.025
5	71683	6.14E-04	2.204	—	—	1.01E-02	1.976
6	130894	4.04E-04	2.086	—	—	6.74E-03	2.015
7	228463	2.99E-04	1.621	—	—	4.95E-03	1.663

• **Test 5 Discontinuous permeability,** $u(x, y, z) = a_i \sin(2\pi x) \sin(2\pi y) \sin(2\pi z)$, $\min = -100$, $\max = 100$**, Locally refined meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg	norme
1	93	921	-1.66E+02	-1.00E+02	1.66E+02	1.00E+02	—	1.43E+03
2	636	6588	-5.43E+01	-3.54E+01	5.43E+01	3.54E+01	—	4.81E+02
3	4656	49584	-9.06E+01	-7.89E+01	9.06E+01	7.89E+01	—	4.14E+02
4	35520	384192	-9.79E+01	-9.44E+01	9.79E+01	9.44E+01	—	3.93E+02
5	277248	3023616	-9.95E+01	-9.86E+01	9.95E+01	9.86E+01	—	3.87E+02
6	2190336	23989248	-9.99E+01	-9.96E+01	9.99E+01	9.96E+01	—	3.86E+02

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	93	1.94E+00		—	—	3.12E+00	
2	636	5.34E-01	2.013	—	—	3.07E-01	3.618
3	4656	1.48E-01	1.934	—	—	7.56E-02	2.112
4	35520	3.78E-02	2.015	—	—	1.89E-02	2.047
5	277248	9.50E-03	2.016	—	—	4.79E-03	2.004
6	2190336	2.38E-03	2.009	—	—	1.23E-03	1.973

3 Comments

All problems have been solved with a preconditioned conjugate gradient method applied to a system for face-based pressure unknowns. We used the diagonal of the stiffness matrix as the preconditioner. Relative reduction of the residual by factor 10^{-12} required 1382 iterations for the last Kershaw mesh. The other tests require less than 679 iterations.

The method is superconvergent on all smooth meshes for both primary variables. We used $\sigma_* = 10$ in Test 3 and $\sigma_* = 0.1$ in Test 4. Our experience shows that the presented meshes are too coarse to see significant impact of this parameter on the asymptotic convergence rate.

References

1. F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of mimetic finite difference method for diffusion problems on polyhedral meshes with curved faces. *Math. Models Methods Appl. Sci.*, 16(2):275–297, 2006.
2. F. Brezzi, K. Lipnikov, M. Shashkov, and V. Simoncini. A new discretization methodology for diffusion problems on generalized polyhedral meshes. *Comput. Methods Appl. Mech. Engrg.*, 196, 3682–3692, 2007.

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: CeVe-DDFV, a Discrete Duality Scheme with Cell/Vertex Unknowns

Yves Coudière and Charles Pierre

1 Presentation of the scheme

1.1 General presentation of DDFV methods

"Discrete Duality" Finite Volumes (*DDFV*) schemes have been specifically designed for anisotropic and/or heterogeneous diffusion problems working on general meshes: distorted, non-conformal and locally refined. They first were introduced in 2D independently by Hermeline [9, 10] and Domelevo and Omnès [8], though the key ideas already appear in the work of Nicolaides [14].

As originally defined in [8], a 2D *DDFV* scheme consists in associating a second mesh (the dual mesh) to the original (primal) mesh by building dual cells around each (primal) mesh vertex. Cell and vertex centered *scalar data* are associated to this *double mesh* framework (one data per primal and dual cell), whereas a *vector data* consists in one vector per (primal) mesh edge. To a scalar data is associated a discrete gradient that is a vector data. A *gradient reconstruction* method is used to define this discrete gradient: precisely using the diamond method [6]. A discrete divergence acts on vector data by averaging their normal component on the primal and dual cell boundaries, which procedure is classical for finite volume methods. The key feature is a duality property between the discrete gradient and the discrete divergence operators of Green formula type.

Yves Coudière
LMJL, Université de Nantes, France, e-mail: Yves.Coudiere@univ-nantes.fr

Charles Pierre
LMA, Université de Pau et des pays de l'Adour, France, e-mail: charles.pierre@univ-pau.fr

J. Fořt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_102, © Springer-Verlag Berlin Heidelberg 2011

Extensions of *DDFV* schemes to 3D [1–3, 11, 12, 15] are of two types.

CV-DDFV. The original 2D *double mesh* framework is conserved, dual cells are built around the primal mesh vertices and scalar data consist in a double set of unknowns associated with the (primal) mesh cells and vertices.
CeVeFE-DDFV. Recently Coudière and Hubert [3, 4] modified the 2D framework by considering a third mesh (triple mesh method), with unknowns associated with cells, faces, edges and vertices of the primal mesh.

The method considered here is of *CV-DDFV* type, *CV* holding for Cell and Vertex centered. Two versions have been developed so far.

(A) A first 3D construction was introduced by Pierre in [15] for anisotropic and/or heterogeneous diffusion problems. The dual cells here do not form a mesh in the classical sense: they recover the domain twice.
(B) A second version, independently introduced by Hermeline [12] and Andreianov & al. [1, 2] differs from the previous one by the dual cell definition that here form a partition of Ω.

For both versions, in presence of heterogeneity, auxiliary (locally eliminated) data are added relatively to faces, as presented in [5, 12]. In case of complex meshes, involving face shapes other than triangles or quadrangles, this local elimination procedure is made difficult enforcing to consider auxiliary data as real unknowns inside the algorithm, which drastically increases the problem size.

We first emphasize the similarities between *(A)* and *(B)*. These two versions are based on the same definition of the discrete gradient. They also induce comparable discrete duality properties. Indeed, after a careful examination of these duality properties in [15] and in [2] it turns out that they do involve exactly the same stiffness and mass matrices. As a result, between these two versions, only the averaging of the source terms on the dual cells will differ.

In this paper, version *(A)* will be considered without auxiliary data on the mesh faces. The fifth test case, including heterogeneity and thus necessitating these auxiliary unknowns per face center, will not be treated here because of a lack of time.

1.2 CV- DDFV version (A), discrete duality

Let the domain $\Omega \subset \mathbb{R}^3$ be a connected open subset, its boundary is assumed to be polyhedral. Let $\mathscr{M}$ be a (general) mesh of Ω, possibly non conformal, and whose (primal) cells (*resp.* faces) are general polyhedral (*resp.* polygonal). The set of cells, faces and vertices of $\mathscr{M}$ are respectively denoted $\mathscr{C}$, $\mathscr{F}$ and $\mathscr{V}$. To any vertex $v \in \mathscr{V}$ is associated a dual cell $v^\star$ and to any face $f \in \mathscr{F}$ is associated a diamond cell D_f. Diamond cells form a partition of Ω, whereas dual cells intersect and recover Ω exactly twice.

A vector data is a piecewise constant vector function on the diamond cells. A scalar data is provided by one scalar per cell and per vertex of $\mathscr{M}$. The space of vector data is denoted $\mathbb{Q}_h$ and the space of scalar data $\mathbb{F}_h$. A discrete function is obtained by supplementing a scalar data with one scalar data per boundary face. The space of discrete functions is denoted $\mathbb{U}_h$. As developed in Sect. 1.3, $u_h \in \mathbb{U}_h$ will be interpreted as a function defined on the diamond cell boundaries:

$$\partial\mathscr{D} := \bigcup_{f\in\mathscr{F}} \partial_{D_f} \ , \quad u_h : \partial\mathscr{D} \longrightarrow \mathbb{R}, \tag{1}$$

that moreover is continuous and piecewise affine on the diamond cell faces.

Two discrete operators will be defined, $\nabla_h : \mathbb{U}_h \longrightarrow \mathbb{Q}_h$ and $\mathrm{div}_h : \mathbb{Q}_h \longrightarrow \mathbb{F}_h$. that satisfy the *discrete duality property* (see [15])

$$\int_\Omega \nabla_h u_h \cdot \mathbf{q}_h dx = -\langle\langle u_h, \mathrm{div}_h \mathbf{q}_h\rangle\rangle + \int_{\partial\Omega} u_h \, \mathbf{q}_h \cdot \mathbf{n} \, ds \tag{2}$$

for any functions $u_h \in \mathbb{U}_h$ and $\mathbf{q}_h \in \mathbb{Q}_h$, with $\mathbf{n}$ the unit normal on $\partial\Omega$ pointing outside Ω, and the pairing:

$$\langle\langle u_h, \mathrm{div}_h \mathbf{q}_h\rangle\rangle = \frac{1}{3}\sum_{c\in\mathscr{C}} u_c \, \mathrm{div}_c \, \mathbf{q}_h |c| + \frac{1}{3}\sum_{v\in\mathscr{V}} u_v \, \mathrm{div}_v \, \mathbf{q}_h |v^\star|. \tag{3}$$

Here, $|c|$ and $|v^\star|$ are the volumes of the primal and dual cells c and $v^\star$, u_c and $\mathrm{div}_c \, \mathbf{q}_h$ are the values associated to the cell c of the two scalar data u_h and $\mathrm{div}_h \mathbf{q}_h$, and similarly u_v and $\mathrm{div}_v \, \mathbf{q}_h$ are the values associated to the vertex v of the two scalar data u_h and $\mathrm{div}_h \mathbf{q}_h$.

In (2) the two integrals are well defined. The first integral is an L^2 product on Ω since both $\mathbf{q}_h$ and $\nabla_h u_h$ are piecewise constant vector functions on the diamond cells.The second integral is an L^2 product on $\partial\Omega$: $\mathbf{q}_h$ is piecewise constant on the boundary faces and its normal component $\mathbf{q}_h \cdot \mathbf{n}$ also, moreover $\partial\Omega \subset \partial\mathscr{D}$ defined in (1) and so u_h has a restriction to $\partial\Omega$ that is continuous.

1.3 Dual and diamond cells

A center x_c (resp. x_f) is associated to each cell $c \in \mathscr{C}$ (resp. $f \in \mathscr{F}$).

Diamond cells. Let $f \in \mathscr{F}$. In case $f \not\subset \partial\Omega$ then f is the interface between two cells $c_1, \ c_2 \in \mathscr{C}$: $f = \overline{c_1} \cap \overline{c_2}$. Denoting x_i the center of c_i, then D_f is the union of the two pyramids with apex x_i and with base f as depicted on Fig. 1. In case $f \subset \partial\Omega$, then $f = \partial\Omega \cap c$ for one cell $c \in \mathscr{C}$. In this case D_f is the pyramid with apex x_c and base f. Still in this cases, f can be considered as a degenerated (flat) pyramid of apex its own center x_f and base f. Thus, in all cases, D_f is the union of two pyramids, and its boundary can be partitioned into triangles. The vertices of

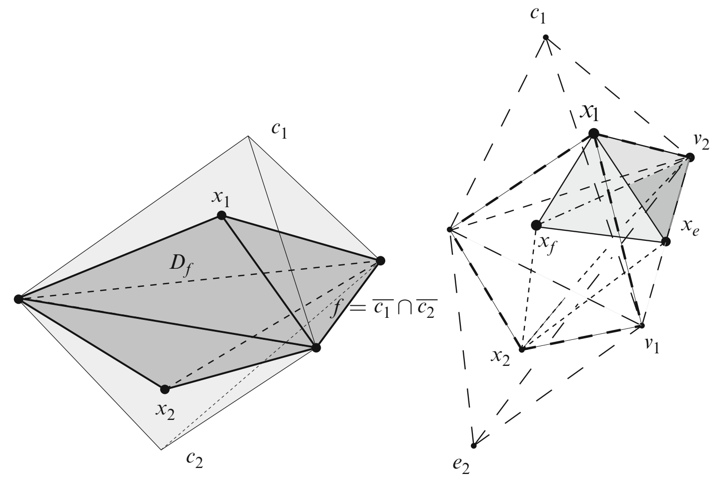


Fig. 1 Left: diamond cell for an internal triangular face f. Right: dual cell construction

these triangles either are: cell centers, vertices or boundary face centers of $\mathscr{M}$. As a result providing a scalar value to each cell, vertex and boundary face of $\mathscr{M}$ defines a unique continuous piecewise affine function $u_h\ :\ \partial\mathscr{D} \mapsto \mathbb{R}$, with $\partial\mathscr{D}$ defined in (1). This is precisely the lift from the discrete function in $\mathbb{U}_h$ presented in Sect. 1.2 into continuous piecewise affine functions on $\partial\mathscr{D}$ in (1).

Dual cells. Let $v \in \mathscr{V}$ and consider a cell $c \in \mathscr{C}$ and a face $f \in \mathscr{F}$ so that v is a vertex of f and f is a face of c. This configuration is denoted by $v \prec f \prec c$. To a triple (v, f, c) so that $v \prec f \prec c$ is associated an element $T_{v,f,c}$. The dual cell $v^\star$ then is defined as $v^\star = \bigcup_{f,c:\ v\prec f\prec c} T_{v,f,c}$. Let us eventually define the element $T_{v,f,c}$, as depicted on Fig. 1. Introduce w_1 and w_2 the two vertices of f such that $[v, w_1]$ and $[v, w_2]$ are two edges of f. Then $T_{v,f,c}$ is the union of the two tetrahedra $vx_cx_fw_i$ for i =1, 2.

As one can see, for a fixed face f and a fixed cell c such that $f \subset \partial c$, considering all elements $T_{v,f,c}$ for all the vertices v of f recovers exactly twice $D_f \cap c$. As a result the dual cells recover the whole domain exactly twice: $\sum_{v\in\mathscr{V}} |v^\star| = 2|\Omega|$.

1.4 Discrete operators

The discrete divergence is classically defined by averaging the normal component of $\mathbf{q}_h \in \mathbb{Q}_h$ on the primal and dual cells, for all $c \in \mathscr{C}$ and all $v \in \mathscr{V}$:

$$\operatorname{div}_c \mathbf{q}_h = \frac{1}{|c|}\int_{\partial c} \mathbf{q}_h \cdot \mathbf{n} ds, \quad \operatorname{div}_v \mathbf{q}_h = \frac{1}{|v^\star|}\int_{\partial v^\star} \mathbf{q}_h \cdot \mathbf{n} ds, \tag{4}$$

for $\mathbf{n}$ the unit normal on ∂c (resp. $\partial v^\star$) pointing outside c (resp. $v^\star$). This definition is well posed since the discontinuity set of $\mathbf{q}_h \in \mathbb{Q}_h$ has a zero 2-dimensional measure intersection with the primal and dual cell boundaries.

The discrete gradient is defined as follows. Let $u_h \in \mathbb{U}_h$, for all $f \in \mathscr{F}$:

$$\nabla_f u_h = \frac{1}{|D_f|} \int_{\partial D_f} u_h \mathbf{n} ds, \tag{5}$$

where $\nabla_f u_h$ is the (vector) value of $\nabla_h u_h$ on D_f and for $\mathbf{n}$ the unit normal on ∂D_f pointing outside D_f.

In practice, definition (5) can always be reformulated in terms of data differences as in the 2D case where (see e.g. [7]):

$$\nabla_f u_h = (u_{c_1} - u_{c_2})\mathbf{N}_f + (u_{v_1} - u_{v_2})\mathbf{M}_f,$$

f is a mesh interface (edge), c_1 and c_2 the two cells on each side of f, v_1 and v_2 the two vertices of f and $\mathbf{N}_f$, $\mathbf{M}_f$ two vectors. We refer to [5,15] for similar expansions in 3D.

1.5 The scheme

The linear diffusion problem $-div(\mathbf{K}\nabla u) = f$ is considered together with a Dirichlet boundary condition $u_{|\partial\Omega} = g$. The tensor $\mathbf{K}$ is discretized into $\mathbf{K}_h$ by averaging $\mathbf{K}$ on each diamond cells and the source term f is discretized as $f_h \in \mathbb{F}_h$ by averaging f over each primal and dual cells. The problem reads: find $u_h \in \mathbb{U}_h$ such that

$$\forall\, c \in \mathscr{C} : \ \operatorname{div}_c(\mathbf{K}_h \nabla_h u_h) = f_c \ , \quad \forall\, v \in \mathscr{V},\ v \notin \partial\Omega : \ \operatorname{div}_v(\mathbf{K}_h \nabla_h u_h) = f_v \tag{6}$$

$$\forall\, v \in \mathscr{V},\ v \in \partial\Omega : \ u_h(v) = g(v)\ , \quad \forall\, f \in \mathscr{F},\ f \subset \partial\Omega : u_h(x_f) = g(x_f), \tag{7}$$

To solve (6) (7), we split $\mathbb{U}_h = \mathbb{U}_{h,0} \oplus \mathbb{B}$ where $\mathbb{U}_{h,0}$ is the subset of discrete functions equal to zero on $\partial\Omega$. Then u_h decomposes as $u_h = u_0 + \tilde{u}$, where $\tilde{u} \in \mathbb{B}$ is uniquely determined by (7). Now $u_0 \in \mathbb{U}_{h,0}$ satisfies $-\operatorname{div}_h(\mathbf{K}_h \nabla_h u_0) = f_h + \operatorname{div}_h(\mathbf{K}_h \nabla_h \tilde{u}) := \tilde{f_h}$ for all primal cells and all interior vertices. This is a square linear system equivalent with: find $u_0 \in \mathbb{U}_{h,0}$ so that for all $v \in \mathbb{U}_{h,0}$ we have:

$$-\langle\!\langle \operatorname{div}_h(\mathbf{K}_h \nabla_h u_0), v \rangle\!\rangle = \langle\!\langle \tilde{f}_h, v \rangle\!\rangle$$

With the help of the discrete duality property (2) it is also equivalent with finding $u_0 \in \mathbb{U}_{h,0}$ so that for all $v \in \mathbb{U}_{h,0}$:

$$\int_{\Omega} \mathbf{K}_h \nabla_h u_0 \cdot \nabla_h v dx = \langle\langle \tilde{f}_h, v \rangle\rangle.$$

In practice, introducing the stiffness matrix S associated to the discrete tensor $\mathbf{K}_h$, this problem is rewritten as the square positive symmetric linear system

$$S U_0 = \tilde{F}, \tag{8}$$

with U_0 (resp. $\tilde{F}$) the vector formed by the values of u_0 (resp. $\tilde{f}_h$) at the cell centers and interior vertices. The stiffness matrix S has the coefficients $S_{ij} = \int_{\Omega} \mathbf{K}_h \nabla_h w_i \cdot \nabla_h w_j dx$, with $w_i \in \mathbb{U}_{h,0}$ the base function having value 1 at one cell center or interior vertex and 0 everywhere else. This matrix is clearly symmetric and positive.

2 Numerical results

The cell centers as well as the face centers are set to their iso-barycenter.

Let us first define the data (source term f and anisotropy tensor $\mathbf{K}$) discretization. Primal, dual and diamond cells are partitioned using a single set of tetrahedra of type $E = x_c x_f v_1 v_2$, with $c \in \mathscr{C}$, f a face of c and v_1, v_2 two vertices of f forming one of its edges. To form the scalar data f_h, f is averaged on the tetrahedra E partitioning each primal and dual cells whereas the discrete tensor $\mathbf{K}_h$ is obtained by averaging $\mathbf{K}$ on the tetrahedra E partitioning the diamond cells. Averaging is made by the mean of Gaussian quadrature on each tetrahedra E using a 15 points quadrature formula of order 5, see e.g. [13]. Assembling the discrete source term f_h and tensor $\mathbf{K}_h$ requires one loop on the mesh faces.

The stiffness matrix S in (8) also is assembled using a loop on the mesh faces. Precisely two base functions w_i and w_j have a non zero interaction (i.e. $S_{ij} = \int_{\Omega} \mathbf{K}_h \nabla_h w_i \cdot \nabla_h w_j dx \neq 0$) in case they are associated to two vertices of a same diamond D_f.

Let us now define the L^2, H^1 and energy errors reported in the following tables as erl2, ergrad and ener respectively. Let u_h denote the discrete solution of one of the test case, and u the solution of the associated continuous problem. The discrete function u_h is lifted to a function $\overline{u}_h \in L^2(\Omega)$ as follows. Consider a face f, u_h provides a value at each vertex of D_f and also at the face center x_f in case of a boundary face. In case of an interior face, a supplementary value u_f is computed at x_f as $u_f = (\sum_{i=1}^{n} u_{v_i})/n$ where the v_i are the n vertices of f, which definition is consistent since x_f is the iso-barycenter of f. With these additional values, scalars are available for every vertices of the tetrahedra E that partition Ω: this defines a unique function $\overline{u}_h$ by P^1 interpolation, which then is continuous piecewise affine on Ω. We define:

$$\text{erl2}^2 = \frac{\int_{\Omega} |\overline{u}_h - u|^2 dx}{\int_{\Omega} |u|^2 dx}.$$

The discrete vector data $\nabla_h u_h$ is a piecewise constant vector function on the diamond cells. Therefore $\nabla_h u_h$ is an L^2 functions on Ω and the H^1 and energy errors reported in the following tables are defined as:

$$\text{ergrad}^2 = \frac{\int_\Omega |\nabla_h u_h - \nabla u|^2 dx}{\int_\Omega |\nabla u|^2 dx}\,, \quad \text{ener}^2 = \frac{\int_\Omega K(\nabla_h u_h - \nabla u)\cdot(\nabla_h u_h - \nabla u)}{\int_\Omega K\nabla u \cdot \nabla u dx}.$$

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Tetrahedral meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	2187	21287	1.34E-02	1.53E-02	1.99E+00	1.99E+00	1.80E+00
2	4301	44813	3.24E-03	6.84E-03	1.99E+00	1.99E+00	1.80E+00
3	8584	94088	8.78E-03	7.44E-03	2.00E+00	1.99E+00	1.80E+00
4	17102	195074	4.74E-03	5.52E-03	2.00E+00	2.00E+00	1.80E+00
5	34343	405077	5.90E-04	1.49E-03	2.00E+00	2.00E+00	1.80E+00
6	69160	838856	1.30E-03	6.19E-04	2.00E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	2187	1.39E-02	–	1.85E-01	–	1.80E-01	–
2	4301	8.80E-03	2.04E+00	1.48E-01	1.01E+00	1.44E-01	9.89E-01
3	8584	5.64E-03	1.93E+00	1.18E-01	9.73E-01	1.15E-01	9.97E-01
4	17102	3.61E-03	1.94E+00	9.36E-02	1.01E+00	9.10E-02	1.01E+00
5	34343	2.26E-03	2.01E+00	7.43E-02	9.92E-01	7.24E-02	9.81E-01
6	69160	1.42E-03	2.00E+00	5.87E-02	1.01E+00	5.70E-02	1.02E+00

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Voronoï meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	87	1433	1.23E-01	1.79E-01	1.91E+00	1.85E+00	1.80E+00
2	235	4393	6.66E-02	2.93E-03	1.87E+00	2.00E+00	1.80E+00
3	527	10777	1.32E-02	9.56E-03	1.93E+00	1.97E+00	1.80E+00
4	1013	21793	-1.76E-03	4.97E-03	1.93E+00	2.00E+00	1.80E+00
5	1776	40998	5.42E-04	4.30E-03	1.98E+00	1.97E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	87	6.19E-02	–	4.43E-01	–	4.29E-01	–
2	235	3.36E-02	1.85E+00	3.37E-01	8.28E-01	3.29E-01	7.96E-01
3	527	2.10E-02	1.74E+00	2.55E-01	1.03E+00	2.49E-01	1.04E+00
4	1013	1.35E-02	2.03E+00	2.05E-01	1.01E+00	2.01E-01	9.85E-01
5	1776	9.99E-03	1.62E+00	1.75E-01	8.38E-01	1.71E-01	8.47E-01

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Kershaw meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	855	13819	7.16E-02	2.88E-02	1.94E+00	1.96E+00	1.80E+00
2	7471	138691	1.26E-02	6.45E-03	1.99E+00	1.99E+00	1.80E+00
3	62559	1237459	1.30E-03	1.75E-03	2.00E+00	2.00E+00	1.80E+00
4	512191	10443763	4.61E-04	5.45E-04	2.00E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	855	5.64E-02	–	4.57E-01	–	4.51E-01	–
2	7471	1.71E-02	1.65E+00	1.91E-01	1.20E+00	1.89E-01	1.21E+00
3	62559	3.45E-03	2.26E+00	7.74E-02	1.28E+00	7.67E-02	1.27E+00
4	512191	7.62E-04	2.15E+00	3.47E-02	1.14E+00	3.41E-02	1.16E+00

• **Test 1 Mild anisotropy,** $u(x,y,z) = 1 + \sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$ $\min = 0,\ \max = 2$**, Checkerboard meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	59	703	1.46E-01	3.41E-02	1.86E+00	1.97E+00	1.80E+00
2	599	9835	3.87E-02	8.56E-03	1.96E+00	1.99E+00	1.80E+00
3	5423	101539	9.24E-03	2.14E-03	1.99E+00	2.00E+00	1.80E+00
4	46175	917395	2.15E-03	5.35E-04	2.00E+00	2.00E+00	1.80E+00
5	381119	7788403	5.01E-04	1.34E-04	2.00E+00	2.00E+00	1.80E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	59	4.79E-02	–	4.01E-01	–	3.94E-01	–
2	599	1.08E-02	1.93E+00	1.95E-01	9.34E-01	1.92E-01	9.31E-01
3	5423	2.55E-03	1.96E+00	9.58E-02	9.66E-01	9.37E-02	9.73E-01
4	46175	6.27E-04	1.96E+00	4.75E-02	9.83E-01	4.63E-02	9.89E-01
5	381119	1.56E-04	1.98E+00	2.36E-02	9.92E-01	2.30E-02	9.95E-01

- **Test 2 Heterogeneous anisotropy,** $u(x, y, z) = x^3y^2z + x\sin(2\pi xz)\sin(2\pi xy)\sin(2\pi z)$**,** $\min = -0.862, \ \max = 1.0487$**, Prism meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	3010	64158	-8.54E-01	-8.41E-01	1.00E+00	1.00E+00	1.71E+00
2	24020	555528	-8.56E-01	-8.59E-01	1.02E+00	1.05E+00	1.71E+00
3	81030	1924098	-8.61E-01	-8.59E-01	1.04E+00	1.04E+00	1.71E+00
4	192040	4619868	-8.59E-01	-8.61E-01	1.04E+00	1.05E+00	1.71E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	3010	5.06E-02	–	2.45E-01	–	2.48E-01	–
2	24020	1.85E-02	1.45E+00	1.26E-01	9.63E-01	1.27E-01	9.66E-01
3	81030	1.46E-02	5.90E-01	8.51E-02	9.63E-01	8.59E-02	9.66E-01
4	192040	1.37E-02	2.08E-01	6.49E-02	9.44E-01	6.53E-02	9.50E-01

- **Test 3 Flow on random meshes,** $u(x, y, z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1, \ \max = 1$**, Random meshes**

i	nu	nmat	umin	uemin	umax	uemax	normg
1	91	1063	-1.58E+00	-9.78E-01	1.54E+00	9.31E-01	3.65E+00
2	855	13819	-1.08E+00	-9.94E-01	1.12E+00	9.82E-01	3.57E+00
3	7471	138691	-1.04E+00	-9.95E-01	1.01E+00	9.91E-01	3.60E+00
4	62559	1237459	-1.01E+00	-9.98E-01	1.01E+00	9.98E-01	3.60E+00

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	91	3.06E-01	–	5.89E-01	–	5.70E-01	–
2	855	8.29E-02	1.75E+00	3.14E-01	8.56E-01	2.87E-01	9.21E-01
3	7471	2.28E-02	1.79E+00	1.65E-01	8.90E-01	1.46E-01	9.28E-01
4	62559	6.98E-03	1.67E+00	8.96E-02	8.58E-01	7.34E-02	9.68E-01

- **Test 4 Flow around a well, Well meshes,** $\min = 0, \ \max = 5.415$

i	nu	nmat	umin	uemin	umax	uemax	normg
1	1482	23942	4.85E-01	-6.02E-06	5.32E+00	5.42E+00	1.62E+03
2	3960	70872	2.71E-01	-5.68E-06	5.33E+00	5.42E+00	1.62E+03
3	9229	173951	1.66E-01	-5.76E-06	5.33E+00	5.42E+00	1.62E+03
4	21156	412240	1.25E-01	-7.39E-06	5.33E+00	5.42E+00	1.62E+03
5	44420	882520	9.37E-02	-6.93E-06	5.34E+00	5.42E+00	1.62E+03
6	82335	1654893	7.48E-02	-6.94E-06	5.35E+00	5.42E+00	1.62E+03
7	145079	2937937	5.80E-02	-8.05E-06	5.36E+00	5.42E+00	1.62E+03

i	nu	erl2	ratiol2	ergrad	ratiograd	ener	ratioener
1	1482	2.92E-03	–	1.79E-01	–	1.78E-01	–
2	3960	1.38E-03	2.29E+00	1.22E-01	1.18E+00	1.21E-01	1.16E+00
3	9229	7.45E-04	2.19E+00	8.57E-02	1.25E+00	8.56E-02	1.24E+00
4	21156	5.53E-04	1.08E+00	6.56E-02	9.71E-01	6.55E-02	9.72E-01
5	44420	3.77E-04	1.55E+00	5.14E-02	9.85E-01	5.13E-02	9.83E-01
6	82335	2.44E-04	2.11E+00	4.18E-02	1.01E+00	4.17E-02	1.01E+00
7	145079	1.83E-04	1.53E+00	3.51E-02	9.27E-01	3.50E-02	9.26E-01

3 Comments

The linear system (8) to be solved is symmetric and positive: a Conjugate Gradient algorithm has been applied, together with a basic Jacobi preconditioner. The sparsity pattern of the stiffness matrix is not compact, especially for matrix lines corresponding to vertex nodes. The stiffness matrix lines corresponding to cell nodes have $1 + n_f + n_s$ nonzero terms with n_f and n_v the number of faces and vertices of the considered cell; for a tetrahedra $1 + n_f + n_v = 9$. The maximum principle is not fulfilled by *DDFV* schemes. In practice it has been violated only once for test one on Voronoï meshes and more significantly on test 3. Meanwhile no oscillation phenomena are observed. Expected order 2 convergence on erl2 is observed for all tests excepted test 2. Order 1 convergence is observed for ergrad and ener on all tests.

References

1. Andreianov, B., Bendahmane, M., Karlsen, K.H.: A gradient reconstruction formula for finite volume schemes and discrete duality. Proceedings of FVCA5 (2008)
2. Andreianov, B., Bendahmane, M., Karlsen, K.H.: Discrete duality finite volume schemes for doubly nonlinear degenerate hyperbolic-parabolic equations. J. Hyperbolic Differ. Equ. **7**(1), 1–67 (2010)
3. Coudière, Y., Hubert, F.: A 3d discrete duality finite volume method for nonlinear elliptic equation. HAL Preprint URL http://hal.archives-ouvertes.fr/docs/00/45/68/37/PDF/ddfv3d.pdf
4. Coudière, Y., Hubert, F.: A 3d duality finite volume method for nonlinear elliptic equations. Proceedings of Algoritmy 2009 pp. 51–60 (2009)
5. Coudiere, Y., Pierre, C., Rousseau, O., Turpault, R.: 2D/3D DDFV scheme for anisotropic-heterogeneous elliptic equations, application to electrograms simulation from medical data. Int. J. Finite Volumes (2009)
6. Coudière, Y., Vila, J.P., Villedieu, P.: Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. M2AN **33**(3), 493–516 (1999)
7. Delcourte, S., Domelevo, K., Omnès, P.: Discrete-duality finite volume method for second order elliptic problems. Proceedings of FVCA4 pp. 447–458 (2005)

8. Domelevo, K., Omnes, P.: A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. M2AN Math. Model. Numer. Anal. **39**(6), 1203–1249 (2005)
9. Hermeline, F.: Une méthode de volumes finis pour les équations elliptiques du second ordre. C. R. Acad. Sci. **326**(12), 1433–1436 (1998)
10. Hermeline, F.: A finite volume method for the approximation of diffusion operators on distorted meshes. J. Comput. Phys. **160**(2), 481–499 (2000)
11. Hermeline, F.: Approximation of 2-D and 3-D diffusion operators with variable full tensor coefficients on arbitrary meshes. Comput. Methods Appl. Mech. Engrg. **196**(21-24), 2497–2526 (2007)
12. Hermeline, F.: A finite volume method for approximating 3D diffusion operators on general meshes. Comput. Meth. Appl. Mech. Engrg. (2009)
13. Jinyun, Y.: Symmetric gaussian quadrature formulae for tetrahedronal regions. Computer Methods in Applied Mechanics and Engineering (1981)
14. Nicolaides, R.: Direct discretization of planar div-curl problems. SIAM J. Numer. Anal. **29**(1), 32–56 (1992)
15. Pierre, C.: Modelling and simulating the electrical activity of the heart embedded in the torso, numerical analysis and finite volumes methods. . PhD Thesis, Université de Nantes (2005)

The paper is in final form and no similar paper has been or is being submitted elsewhere.

Benchmark 3D: A multipoint flux mixed finite element method on general hexahedra

Mary F. Wheeler, Guangri Xue, and Ivan Yotov

1 Presentation of the scheme

In this paper we discuss a family of numerical schemes for modeling Darcy flow, the multipoint flux mixed finite element (MFMFE) methods. The MFMFE methods allow for an accurate and efficient treatment of irregular geometries and heterogeneities such as faults, layers, and pinchouts that require highly distorted grids and discontinuous coefficients. The methods can be reduced to cell-centered discretizations and have convergent pressures and velocities on general hexahedral and simplicial grids.

The development of the MFMFE methods has been motivated by the multipoint flux approximation (MPFA) methods [1, 2, 7, 8]. In the MPFA finite volume framework, sub-edge (sub-face) fluxes are introduced, which allows for local flux elimination and reduction to a cell-centered scheme. Similar elimination is achieved in the MFMFE variational framework, by employing appropriate finite element spaces and special quadrature rules. Our approach is based on the BDM_1 [5] or the BDDF_1 [3] spaces with a trapezoidal quadrature rule applied on the reference element. We refer to [4] for a similar approach on simplicial grids, as well as to [10, 11] for a related work on quadrilateral grids using a broken Raviart-Thomas space. Mortar MFMFE methods on non-matching grids have been developed in [14].

We describe the method for a single phase Darcy flow in a domain $\Omega \subset \mathbb{R}^3$

$$\psi = -K\nabla u, \quad \nabla\cdot\psi = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega,$$

Mary F. Wheeler and Guangri Xue
The University of Texas at Austin, USA, e-mail: mfw@ices.utexas.edu, gxue@ices.utexas.edu

Ivan Yotov
University of Pittsburgh, USA, e-mail: yotov@math.pitt.edu

J. Fovrt et al. (eds.), *Finite Volumes for Complex Applications VI – Problems & Perspectives*, Springer Proceedings in Mathematics 4,
DOI 10.1007/978-3-642-20671-9_103, © Springer-Verlag Berlin Heidelberg 2011

where ψ is the Darcy velocity, u is the pressure, and K is a symmetric, uniformly positive definite tensor representing the rock permeability divided by the fluid viscosity. Other boundary conditions can also be treated. The weak formulation of the problem reads: find $\psi \in H(\mathbf{div}; \Omega)$ and $u \in L^2(\Omega)$, such that

$$(K^{-1}\psi, \mathbf{v}) - (u, \nabla \cdot \mathbf{v}) = 0, \qquad \forall \mathbf{v} \in H(\mathbf{div}; \Omega), \tag{1}$$

$$(\nabla \cdot \psi, w) = (f, w), \qquad \forall w \in L^2(\Omega), \tag{2}$$

where $H(\mathbf{div}; \Omega) := \left\{\mathbf{v} \in (L^2(\Omega))^d : \nabla \cdot \mathbf{v} \in L^2(\Omega)\right\}$ and $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$.

Multipoint flux mixed finite element (MFMFE) methods have been developed and analyzed in [9, 13–15] for simplicial, quadrilateral, and hexahedral grids. The method is defined as follows: find $\psi_h \in V_h$ and $u_h \in W_h$ such that

$$(K^{-1}\psi_h, \mathbf{v})_Q - (u_h, \nabla \cdot \mathbf{v}) = 0, \qquad \forall \mathbf{v} \in V_h, \tag{3}$$

$$(\nabla \cdot \psi_h, w) = (f, w), \qquad \forall w \in W_h \tag{4}$$

In the above V_h and W_h are suitable mixed finite element spaces and $(\cdot, \cdot)_Q$ is a special quadrature rule. Appropriate choices allow for a flux variable defined at a vertex to be expressed by cell-centered pressures surrounding the vertex. This results in a 27 point pressure stencil on logically rectangular 3D grids.

The quadrature rule (9) can be symmetric or non-symmetric. On smooth hexahedral grids, both the symmetric and non-symmetric MFMFE methods give first-order accurate velocities and pressures, as well as second order accurate face fluxes and pressures at the cell centers [9, 13, 15]. On highly distorted hexahedral grids with non-planar faces [13], the convergence of the symmetric MFMFE can deteriorate while the non-symmetric MFMFE still gives a first order accuracy under a mild assumption on the grids and permeability anisotropy. The non-symmetric quadrature rule was first proposed in [10] for quadrilateral grids.

Finite element spaces. Let Ω be a polyhedral domain partitioned into a union of hexahedral finite elements of characteristic size h. Let us denote the partition by $\mathcal{T}_h$ and assume that it is shape-regular and quasi-uniform [6]. The velocity and pressure finite element spaces on any physical grid-block E are defined, respectively, via the Piola transformation

$$\mathbf{v} \leftrightarrow \hat{\mathbf{v}} : \hat{\mathbf{v}} = \frac{1}{J_E} DF_E \hat{\mathbf{v}} \circ F_E^{-1},$$

and the scalar transformation

$$w \leftrightarrow \hat{w} : w = \hat{w} \circ F_E^{-1},$$

where $\hat{E}$ is the reference cube or tetrahedron, F_E denotes a trilinear mapping from $\hat{E}$ to E, DF_E is the Jacobian of F_E, and J_E is its determinant. The Piola transformation preserves the normal components of the vectors. The finite element spaces V_h and

W_h on $\mathscr{T}_h$ are given by

$$
\begin{aligned}
V_h &= \left\{ \mathbf{v} \in H(\mathbf{div};\Omega): \quad \mathbf{v}|_E \leftrightarrow \hat{\mathbf{v}},\ \hat{\mathbf{v}} \in \hat{V}(\hat{E}), \quad \forall E \in \mathscr{T}_h \right\}, \\
W_h &= \left\{ w \in L^2(\Omega): \quad w|_E \leftrightarrow \hat{w},\ \hat{w} \in \hat{W}(\hat{E}), \quad \forall E \in \mathscr{T}_h \right\},
\end{aligned} \tag{5}
$$

where $\hat{V}(\hat{E})$ and $\hat{W}(\hat{E})$ are finite element spaces on the reference element $\hat{E}$.

The spaces on the reference cube are defined by enhancing the BDDF_1 spaces:

$$
\begin{aligned}
\hat{V}(\hat{E}) = {} & \mathrm{BDDF}_1(\hat{E}) + r_2 \mathrm{curl}\,(0,0,\hat{x}^2\hat{z})^T + r_3 \mathrm{curl}\,(0,0,\hat{x}^2\hat{y}\hat{z})^T + s_2 \mathrm{curl}\,(\hat{x}\hat{y}^2,0,0)^T \\
& + s_3 \mathrm{curl}\,(\hat{x}\hat{y}^2\hat{z},0,0)^T + t_2 \mathrm{curl}\,(0,\hat{y}\hat{z}^2,0)^T + t_3 \mathrm{curl}\,(0,\hat{x}\hat{y}\hat{z}^2,0)^T, \\
\hat{W}(\hat{E}) = {} & P_0(\hat{E}),
\end{aligned}
$$

where the $\mathrm{BDDF}_1(\hat{E})$ space is defined as [3]:

$$
\begin{aligned}
\mathrm{BDDF}_1(\hat{E}) = {} & P_1(\hat{E})^3 + r_0 \mathrm{curl}\,(0,0,\hat{x}\hat{y}\hat{z})^T + r_1 \mathrm{curl}\,(0,0,\hat{x}\hat{y}^2)^T + s_0 \mathrm{curl}\,(\hat{x}\hat{y}\hat{z},0,0)^T, \\
& + s_1 \mathrm{curl}\,(\hat{y}\hat{z}^2,0,0)^T + t_0 \mathrm{curl}\,(0,\hat{x}\hat{y}\hat{z},0)^T + t_1 \mathrm{curl}\,(0,\hat{x}^2\hat{z},0)^T.
\end{aligned}
$$

In above equations, r_i, s_i, t_i $(i = 0,\dots,3)$ are real constants, P_k denotes polynomials of degree at most k, and $(\hat{x},\ \hat{y},\ \hat{z})^T$ denotes a point in the reference element. The enhancement of the BDDF_1 space is needed to obtain a space with four degrees of freedom per face, rather than three in the original formulation. This allows to associate a degree of freedom with each vertex of the face, which is needed in the reduction to a cell-centered pressure stencil as described later in this section.

There are four degrees of freedom (DOF) per reference face. The DOF are chosen to be the normal components at the vertices. This choice of DOF guarantees continuity of the normal component of the velocity vector across element faces, which is needed for an $H(\mathbf{div};\Omega)$-conforming velocity space as required by (5).

A quadrature rule. The integration on a physical element is performed by mapping to the reference element and choosing a quadrature rule on $\hat{E}$. Using the Piola transformation, we write $(K^{-1}\cdot,\cdot)$ in (1) as

$$
(K^{-1}\mathbf{q},\mathbf{v})_E = \left(\frac{1}{J_E} DF_E^T K^{-1}(F_E(\hat{x})) DF_E \hat{\mathbf{q}}, \hat{\mathbf{v}} \right)_{\hat{E}} \equiv (\mathscr{M}_E \hat{\mathbf{q}}, \hat{\mathbf{v}})_{\hat{E}},
$$

where

$$
\mathscr{M}_E = \frac{1}{J_E} DF_E^T K^{-1}(F_E(\hat{x})) DF_E. \tag{6}
$$

Define a perturbed $\widetilde{\mathscr{M}}_E$ as

$$
\widetilde{\mathscr{M}}_E = \frac{1}{J_E} DF_E^T(\hat{\mathbf{r}}_{c,\hat{E}}) \overline{K}_E^{-1} DF_E, \tag{7}
$$

where $\hat{\mathbf{r}}_{c,\hat{E}}$ is the centroid of $\hat{E}$ and $\overline{K}_E$ denotes the mean of K on E. In addition, denote the trapezoidal rule on $\hat{E}$ by $\text{Trap}(\cdot,\cdot)_{\hat{E}}$:

$$\text{Trap}(\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{E}} \equiv \frac{|\hat{E}|}{k}\sum_{i=1}^{k}\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)\cdot\hat{\mathbf{v}}(\hat{\mathbf{r}}_i), \tag{8}$$

where $\{\hat{\mathbf{r}}_i\}_{i=1}^k$ are the vertices of $\hat{E}$.

The symmetric quadrature rule is based on the original $\mathscr{M}_E$ while the non-symmetric one is based on the perturbed $\widetilde{\mathscr{M}}_E$. The quadrature rule on an element E is defined as

$$(K^{-1}\mathbf{q},\mathbf{v})_{Q,E} \equiv \begin{cases} \text{Trap}(\mathscr{M}_E\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{E}} = \frac{|\hat{E}|}{k}\sum_{i=1}^{k}\mathscr{M}_E(\hat{\mathbf{r}}_i)\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)\cdot\hat{\mathbf{v}}(\hat{\mathbf{r}}_i), & \text{symmetric,}\\ \text{Trap}(\widetilde{\mathscr{M}}_E\hat{\mathbf{q}},\hat{\mathbf{v}})_{\hat{E}} = \frac{|\hat{E}|}{k}\sum_{i=1}^{k}\widetilde{\mathscr{M}}_E(\hat{\mathbf{r}}_i)\hat{\mathbf{q}}(\hat{\mathbf{r}}_i)\cdot\hat{\mathbf{v}}(\hat{\mathbf{r}}_i), & \text{non-symmetric.}\end{cases} \tag{9}$$

The non-symmetric quadrature rule has certain critical properties on the physical elements that lead to a convergent method on rough hexahedra [13].

Reduction to a cell-centered pressure system. The choice of trapezoidal quadrature rule implies that on each element, the velocity degrees of freedom associated with a vertex become decoupled from the rest of the degrees of freedom. As a result, the assembled velocity mass matrix in (3) has a block-diagonal structure with one block per grid vertex. The dimension of each block equals the number of velocity DOF associated with the vertex. Inverting each local block in the mass matrix in (3) allows for expressing the velocity DOF associated with a vertex in terms of the pressures at the centers of the elements that share the vertex. Substituting these expressions into the mass conservation equation (4) leads to a cell-centered system for the pressures. The stencil is 27 points on logically rectangular hexahedral grids. The local linear systems and the resulting global pressure system are positive definite and therefore invertible for the symmetric MFMFE method and, under a mild restriction on the shape regularity of the grids and/or the anisotropy of the permeability, for the non-symmetric MFMFE method; see (11) below. The reader is referred to [9, 13, 15] for further details on the reduction to a cell-centered pressure system.

Theoretical convergence results. Let $W^{k,\infty}_{\mathscr{T}_h}$ consist of functions ϕ such that $\phi|_E \in W^{k,\infty}(E)$ for all $E \in \mathscr{T}_h$. Here k is a multi-index with integer components and $W^{k,\infty}(E)$ denotes the Sobolev space of functions whose derivatives of order k belong to $L^\infty(E)$. Let $\|\cdot\|_k$ be the norm in the Hilbert space $H^k(\Omega)$ with functions whose derivatives of order k belong to $L^2(\Omega)$. The norm in $L^2(\Omega)$ is denoted by $\|\cdot\|$. Let $X \lesssim (\gtrsim)\ Y$ denote that there exists a constant C, independent of the mesh size h, such that $X \le (\ge)\ CY$. The notation $X \eqsim Y$ means that both $X \lesssim Y$ and $X \gtrsim Y$ hold.

The following convergence results have been established for the symmetric MFMFE method on h^2-perturbed parallelepipeds.

Theorem 1 **([9, 15]).** *If $K^{-1} \in W^{1,\infty}_{\mathscr{T}_h}$, then, the velocity ψ_h and pressure u_h of the symmetric MFMFE method (3)–(4) satisfy*

$$\|\psi - \psi_h\| \lesssim h\|\psi\|_1, \quad \|\nabla\cdot(\psi - \psi_h)\| \lesssim h\|\nabla\cdot\psi\|_1, \quad \|u - u_h\| \lesssim h(\|\psi\|_1 + \|u\|_1).$$

On h^2-perturbed parallelepipeds, the non-symmetric MFMFE method has same order of accuracy as the symmetric method. In addition, the non-symmetric method has first order convergence for the velocity and pressure on general quadrilaterals and for the face flux and pressure on general hexahedra with non-planar faces.

For the analysis of the non-symmetric MFMFE method, we require some properties of the bilinear form $(K^{-1}\cdot,\cdot)_Q$ defined on the space V_h. Note that

$$(K^{-1}\mathbf{q},\mathbf{v})_Q = \sum_{E\in\mathscr{T}_h}(K^{-1}\mathbf{q},\mathbf{v})_{Q,E} = \sum_{c\in\mathscr{C}_h}\mathbf{v}_c^T\mathbf{M}_c\mathbf{q}_c, \tag{10}$$

where $\mathscr{C}_h$ denotes the set of corner or vertex points in $\mathscr{T}_h$, $\mathbf{v}_c := \{(\mathbf{v}\cdot\mathbf{n}_e)(\mathbf{x}_c)\}_{e=1}^{n_c}$, $\mathbf{x}_c$ is the coordinate vector of point c, and n_c is the number of faces (or edges in 2D) that share the vertex point c.

Lemma 1 **([13]).** *Assume that $\mathbf{M}_c$ is uniformly positive definite for all $c \in \mathscr{C}_h$:*

$$h^d\boldsymbol{\xi}^T\boldsymbol{\xi} \lesssim \boldsymbol{\xi}^T\mathbf{M}_c\boldsymbol{\xi}, \quad \forall\boldsymbol{\xi}\in\mathbb{R}^{n_c}. \tag{11}$$

Then the bilinear form $(K^{-1}\cdot,\cdot)_Q$ is coercive in $\mathbf{V}_h$ and induces a norm in V_h equivalent to the L^2-norm:

$$(K^{-1}\mathbf{v},\mathbf{v})_Q \eqsim \|\mathbf{v}\|^2, \quad \forall\,\mathbf{v}\in V_h. \tag{12}$$

If in addition

$$\boldsymbol{\xi}^T\mathbf{M}_c^T\mathbf{M}_c\boldsymbol{\xi} \lesssim h^{2d}\boldsymbol{\xi}^T\boldsymbol{\xi}, \quad \forall\,\boldsymbol{\xi}\in\mathbb{R}^{n_c}, \tag{13}$$

then the following Cauchy-Schwarz type inequality holds:

$$(K^{-1}\mathbf{q},\mathbf{v})_Q \lesssim \|\mathbf{q}\|\|\mathbf{v}\| \quad \forall\mathbf{q},\mathbf{v}\in V_h, \tag{14}$$

Conditions (11) and (13) impose mild restrictions on the element geometry and the anisotropy of the permeability tensor K, see [10, 12].

Theorem 2 **([13]).** *Let $K \in W^{1,\infty}_{\mathscr{T}_h}(\Omega)$ and $K^{-1} \in W^{0,\infty}(\Omega)$. If (11) and (13) hold, then the velocity ψ_h and the pressure u_h of the non-symmetric MFMFE method (3)—(4) satisfy*

$$\|\Pi\psi - \psi_h\| + \|Q_h u - u_h\| \lesssim h(|\psi|_1 + \|u\|_2), \tag{15}$$

where Π is the canonical interpolation operator onto V_h and Q_h is the L^2-orthogonal projection onto W_h.

This result further implies convergence of the computed normal velocity to the true normal velocity on the element faces. First, define a norm for vectors in Ω based on the normal components on the faces of $\mathscr{T}_h$:

$$\|\mathbf{v}\|^2_{\mathscr{F}_h} := \sum_{E\in\mathscr{T}_h}\sum_{e\in\partial E}\frac{|E|}{|e|}\|\mathbf{v}\cdot\mathbf{n}_e\|^2_e, \tag{16}$$

where $|E|$ is the volume of E and $|e|$ is the area of e. This norm gives an appropriate scaling of $|\Omega|^{1/2}$ for a unit vector.

Theorem 3 ([13]). *Let $K \in W^{1,\infty}_{\mathscr{T}_h}(\Omega)$ and $K^{-1} \in W^{0,\infty}_{\mathscr{T}_h}(\Omega)$. If (11) and (13) hold, then the velocity ψ_h of the non-symmetric MFMFE method (3)–(4) satisfies*

$$\|\psi - \psi_h\|_{\mathscr{F}_h} \lesssim h(\|\psi\|_1 + \|u\|_2). \tag{17}$$

2 Numerical results

We note that in all tests we report absolute errors. Both the pressure error $\|u - u_h\|$ and the velocity error $\|\Pi\psi - \psi_h\|$ are approximated by the trapezoidal quadrature rule on the reference unit cube. For the velocity face error $\|\psi - \psi_h\|_{\mathscr{F}_h}$ and the mean velocity face error

$$|||\psi - \psi_h|||^2_{\mathscr{F}_h} \equiv \sum_{E\in\mathscr{T}_h}\sum_{e\in\partial E}|E|\left(\frac{1}{|e|}\int_e \psi\cdot\mathbf{n}_e - \frac{1}{|e|}\int_e \psi_h\cdot\mathbf{n}_e\right)^2,$$

the face integrals are approximated by the 9-point Gaussian quadrature rule on the reference face.

- **Test 1 Mild anisotropy,** $u(x,y,z) = 1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$**,** $\min = 0,\ \max = 2$**, Kershaw meshes**

Symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	512	10648	4.66E-03	3.03E-02	1.97E+00	1.96E+00
2	4096	97336	4.23E-03	1.06E-02	1.99E+00	1.99E+00
3	32768	830584	-2.42E-03	1.75E-03	2.00E+00	2.00E+00
4	262144	6859000	7.49E-05	7.14E-04	2.00E+00	2.00E+00

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate	Iters
1	512	2.08E-01	–	3.01E+00	–	4.74E+00	–	4.30E+00	–	9
2	4096	1.17E-01	0.83	1.11E+00	1.44	2.17E+00	1.13	1.94E+00	1.15	17
3	32768	5.96E-02	0.97	3.95E-01	1.45	7.44E-01	1.54	6.62E-01	1.55	32
4	262144	2.95E-02	1.01	1.54E-01	1.36	2.43E-01	1.61	2.01E-01	1.72	65

Non-symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	512	10648	-1.25E-03	3.03E-02	2.01E+00	1.96E+00
2	4096	97336	-3.35E-03	1.06E-02	2.00E+00	1.99E+00
3	32768	830584	-2.08E-03	1.75E-03	2.00E+00	2.00E+00
4	262144	6859000	5.02E-05	7.14E-04	2.00E+00	2.00E+00

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate	Iters
1	512	2.20E-01	–	2.81E+00	–	2.52E+00	–	2.14E+00	–	8
2	4096	1.19E-01	0.89	9.15E-01	1.62	1.23E+00	1.03	1.07E+00	1.00	16
3	32768	5.95E-02	1.00	3.06E-01	1.58	4.27E-01	1.53	3.66E-01	1.55	33
4	262144	2.94E-02	1.02	1.18E-01	1.37	1.40E-01	1.61	1.00E-01	1.87	73

• **Test 1 Flow on random meshes,** $u(x,y,z)=1+\sin(\pi x)\sin\left(\pi\left(y+\frac{1}{2}\right)\right)\sin\left(\pi\left(z+\frac{1}{3}\right)\right)$**,** $\min=0,\ \max=2$**, Random meshes**

Symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	64	1000	-1.43E-02	4.46E-02	1.90E+00	1.82E+00
2	512	10648	2.03E-02	3.17E-02	1.96E+00	1.95E+00
3	4096	97336	-1.07E-03	2.69E-03	1.99E+00	1.99E+00
4	32768	830584	1.14E-03	1.23E-03	2.00E+00	2.00E+00

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate	Iters
1	64	2.54E-01	–	1.15E+00	–	1.01E+00	–	4.26E-01	–	5
2	512	1.25E-01	1.02	6.14E-01	0.91	4.91E-01	1.04	1.79E-01	1.25	6
3	4096	6.29E-02	0.99	3.82E-01	0.68	2.86E-01	0.78	1.00E-01	0.84	7
4	32768	3.15E-02	1.00	2.96E-01	0.37	2.34E-01	0.29	8.84E-02	0.18	8

Non-symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	64	1000	-2.17E-02	4.46E-02	1.90E+00	1.82E+00
2	512	10648	1.52E-02	3.17E-02	1.96E+00	1.95E+00
3	4096	97336	-1.42E-03	2.69E-03	1.99E+00	1.99E+00
4	32768	830584	5.59E-04	1.23E-03	2.00E+00	2.00E+00

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate	Iters
1	64	2.54E-01	–	1.19E+00	–	1.01E+00	–	3.83E-01	–	5
2	512	1.25E-01	1.02	5.54E-01	1.10	4.32E-01	1.23	1.23E-01	1.64	7
3	4096	6.30E-02	0.99	2.78E-01	0.99	2.05E-01	1.08	4.63E-02	1.41	7
4	32768	3.15E-02	1.00	1.39E-01	1.00	1.09E-01	0.91	2.32E-02	1.00	8

- **Test 3 Flow on random meshes,** $u(x,y,z) = \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -1,\ \max = 1$**, Random meshes**

Symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	64	1000	-6.20E+00	-7.59E-01	5.75E+00	6.91E-01
2	512	10648	-1.93E+00	-9.39E-01	2.05E+00	9.23E-01
3	4096	97336	-1.20E+00	-9.85E-01	1.19E+00	9.82E-01
4	32768	830584	-1.06E+00	-9.96E-01	1.04E+00	9.96E-01

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate	Iters
1	64	1.88E+00	–	1.67E+03	–	1.76E+03	–	1.22E+03	–	17
2	512	4.27E-01	2.14	5.84E+02	1.52	5.31E+02	1.73	3.34E+02	1.87	36
3	4096	1.48E-01	1.53	2.97E+02	0.98	2.32E+02	1.19	1.19E+02	1.49	59
4	32768	6.71E-02	1.14	2.02E+02	0.56	1.59E+02	0.55	7.57E+01	0.65	77

Non-symmetric MFMFE

i	nu	nmat	umin	uemin	umax	uemax
1	64	1000	-1.20E+02	-7.59E-01	3.76E+01	6.91E-01
2	512	10648	-5.01E+02	-9.39E-01	6.34E+02	9.23E-01
3	4096	97336	-3.34E+01	-9.85E-01	4.97E+01	9.82E-01
4	32768	830584	-2.16E+03	-9.96E-01	4.12E+03	9.96E-01

i	nu	$\|u-u_h\|$	rate	$\|\Pi\psi-\psi_h\|$	rate	$\|\psi-\psi_h\|_{\mathscr{F}_h}$	rate	$\|\|\|\psi-\psi_h\|\|\|_{\mathscr{F}_h}$	rate
1	64	2.94E+01	–	1.07E+05	–	9.36E+04	–	3.45E+04	–
2	512	8.96E+01	< 0	3.45E+05	< 0	2.78E+05	< 0	1.41E+05	< 0
3	4096	5.84E+00		3.39E+04		2.50E+04		1.11E+04	
4	32768	3.56E+02	< 0	3.23E+06	< 0	2.53E+06	< 0	1.08E+06	< 0

- **Test 5 Discontinuous permeability,**
$u(x,y,z) = a_i \sin(2\pi x)\sin(2\pi y)\sin(2\pi z)$**,** $\min = -100,\ \max = 100$**, Locally refined meshes**

The locally refined grids are treated by introducing mortar finite elements on the subdomain interfaces to approximate the interface pressure and impose weakly continuity of flux; for details see [14]. Here we take the mortar grid to be the trace of the coarser subdomain grid and choose the mortar space to consist of discontinuous piecewise bilinear functions. It is easy to check that this results in forcing on each interface element the four fine grid normal velocities to be equal to the coarse grid normal velocity.

For this test we also report the velocity error $\|\psi - \psi_h\|$, approximated by 27-point Gaussian quadrature rule on the reference unit cube, as well as the norm $|||\psi - \Pi^{RT}\psi_h|||$ defined as follows. For a scalar function $\phi(x_1, x_2, x_3)$ in a cubic element E, let $|||\phi|||_{i,E}$ denote an approximation integral of $|\phi|^2$ using exact integration rule in x_i and midpoint rule in the other directions. Then, for $\mathbf{q} = (q_1, q_2, q_3)^T$, let

$$|||\mathbf{q}|||^2 = \sum_{E \in \mathscr{T}_h} \sum_{i=1}^{3} |||q_i|||_{i,E}^2.$$

In the reported error norm, Π^{RT} is the canonical interpolation operator in the lowest order Raviart-Thomas space.

i	nu	umin	uemin	umax	uemax
2	176	-4.36E+01	-3.54E+01	4.36E+01	3.54E+01
3	1408	-8.30E+01	-7.89E+01	8.30E+01	7.89E+01
4	11264	-9.56E+01	-9.43E+01	9.56E+01	9.43E+01
5	90112	-9.89E+01	-9.86E+01	9.89E+01	9.86E+01

i	$\|u - u_h\|$	rate	$\|\Pi\psi - \psi_h\|$	rate	$\|\psi - \psi_h\|$	rate	$|||\psi - \Pi^{RT}\psi_h|||$	rate	CGiter
2	2.28E+01	–	1.77E+03	–	9.49E+02	–	3.38E+02	–	12
3	1.19E+01	0.94	8.80E+02	1.00	4.96E+02	0.94	8.18E+01	2.05	18
4	6.02E+00	0.98	4.38E+02	1.00	2.51E+02	0.98	2.03E+01	2.01	21
5	3.02E+00	1.00	2.19E+02	1.00	1.26E+02	0.99	5.06E+00	2.00	31

3 Comments

In Test 1 and Test 3, the resulting linear algebraic system is solved using the software HYPRE (high performance preconditioners) developed by researchers at Lawrence Livermore National Laboratory[1]. Specifically, we use the generalized minimum residual (GMRES) method with one V-cycle of algebraic multigrid method as a preconditioner. The stopping criteria for GMRES is relative residual less than 10^{-9}. The number of iterations is reported in each table.

[1] https://computation.llnl.gov/casc/hypre/software.html

In Test 5, the problem is reduced to an interface problem in terms of mortar variables. We use the conjugate gradient (CG) method and the stopping criteria is the relative residual less than 10^{-9}. The number of CG iterations is given in the table.

In Test 1 Mild anisotropy, both the symmetric and non-symmetric methods are first order accurate for the pressure and the velocity, as well as superconvergent of order approaching $O(h^2)$ for the face velocities.

In Test 1 Flow on random meshes, the pressure is first order for both methods. However, the velocity convergence of the symmetric method deteriorates due to the element distortion, while the non-symmetric method maintains first order accuracy in the velocity. These results are consistent with Theorems 1, 2, and 3.

In Test 3 Flow on random meshes, the symmetric method is first order convergent for the pressure and approximately $O(h^{1/2})$ convergent for the velocity, as expected by the theory. For the non-symmetric method, the severe anisotropy in the permeability combined with element distortion leads to near violation of conditions (11) and (13). As a result, the algebraic system is very ill-conditioned and the method fails to converge.

In Test 5 Discontinuous permeability, the two methods are identical, since the elements are cuboids. We observe first order convergence for the pressure and velocity, as well as second order superconvergence for the error $|||\psi - \Pi^{RT}\psi_h|||$, as predicted by the theory from [14].

The paper is in final form and no similar paper has been or is being submitted elsewhere.

References

1. I. Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Comput. Geosci.*, 6:405–432, 2002.
2. I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media, part ii: Discussion and numerical results. *SIAM J. Sci. Comput.*, 19(5):1717–1736, 1998.
3. F. Brezzi, J. Douglas, R. Duran, and M. Fortin. Mixed finite elements for second order elliptic problems in three variables. *Numer. Math.*, 51:237–250, 1987.
4. F. Brezzi, M. Fortin, and L. D. Marini. Error analysis of piecewise constant pressure approximations of Darcy's law. *Comput. Methods Appl. Mech. Eng.*, 195:1547–1559, 2006.
5. Franco Brezzi, Jim Douglas, and L. D. Marini. Two families of mixed finite elements for second order elliptic problems. *Numer. Math.*, 47(2):217–235, 1985.
6. P. G. Ciarlet. *The Finite Element Method for Elliptic Problems*. Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978; reprinted, SIAM, Philadelphia, 2002.
7. M. G. Edwards. Unstructured control-volume distributed, full-tensor finite-volume schemes with flow based grids. *Comput. Geosci.*, 6:433–452, 2002.
8. M. G. Edwards and C. F. Rogers. Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Comput. Geosci.*, 2:259–290, 1998.
9. R. Ingram, M. F. Wheeler, and I. Yotov. A multipoint flux mixed finite element method on hexahedra. *SIAM J. Numer. Anal.*, 48:1281–1312, 2010.

10. R. A. Klausen and R. Winther. Robust convergence of multi point flux approximation on rough grids. *Numer. Math.*, 104:317–337, 2006.
11. Runhild A. Klausen and Ragnar Winther. Convergence of multipoint flux approximations on quadrilateral grids. *Numer. Methods Partial Differential Equations*, 22(6):1438–1454, 2006.
12. K. Lipnikov, M. Shashkov, and I. Yotov. Local flux mimetic finite difference methods. *Numer. Math.*, 112(1):115–152, 2009.
13. M. F. Wheeler, G. Xue, and I. Yotov. A multipoint flux mixed finite element method on distorted quadrilaterals and hexahedra. *ICES REPORT 10-34, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, Submitted*, 2010.
14. M. F. Wheeler, G. Xue, and I. Yotov. A multiscale mortar multipoint flux mixed finite element method. *ICES REPORT 10-33, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, Submitted*, 2010.
15. M. F. Wheeler and I. Yotov. A multipoint flux mixed finite element method. *SIAM. J. Numer. Anal.*, 44(5):2082–2106, 2006.