

SEMA SIMAI Springer Series 27

Daniele Antonio Di Pietro
Luca Formaggia
Roland Masson *Eds.*

Polyhedral Methods in Geosciences

SEMA

SIMAI
SOCIETÀ ITALIANA DI MATEMATICA
APPLICATA E INDUSTRIALE



Springer

SEMA SIMAI Springer Series

Volume 27

Editors-in-Chief

Luca Formaggia, MOX–Department of Mathematics, Politecnico di Milano, Milano, Italy

Pablo Pedregal, ETSI Industriales, University of Castilla–La Mancha, Ciudad Real, Spain

Series Editors

Mats G. Larson, Department of Mathematics, Umeå University, Umeå, Sweden

Tere Martínez-Seara Alonso, Departament de Matemàtiques, Universitat Politècnica de Catalunya, Barcelona, Spain

Carlos Parés, Facultad de Ciencias, Universidad de Málaga, Málaga, Spain

Lorenzo Pareschi, Dipartimento di Matematica e Informatica, Università degli Studi di Ferrara, Ferrara, Italy

Andrea Tosin, Dipartimento di Scienze Matematiche “G. L. Lagrange”, Politecnico di Torino, Torino, Italy

Elena Vázquez-Cendón, Departamento de Matemática Aplicada, Universidade de Santiago de Compostela, A Coruña, Spain

Paolo Zunino, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

As of 2013, the SIMAI Springer Series opens to SEMA in order to publish a joint series aiming to publish advanced textbooks, research-level monographs and collected works that focus on applications of mathematics to social and industrial problems, including biology, medicine, engineering, environment and finance. Mathematical and numerical modeling is playing a crucial role in the solution of the complex and interrelated problems faced nowadays not only by researchers operating in the field of basic sciences, but also in more directly applied and industrial sectors. This series is meant to host selected contributions focusing on the relevance of mathematics in real life applications and to provide useful reference material to students, academic and industrial researchers at an international level. Interdisciplinary contributions, showing a fruitful collaboration of mathematicians with researchers of other fields to address complex applications, are welcomed in this series.

THE SERIES IS INDEXED IN SCOPUS

More information about this series at <http://www.springer.com/series/10532>


Daniele Antonio Di Pietro ·
Luca Formaggia · Roland Masson
Editors

Polyhedral Methods in Geosciences

 Springer

Editors

Daniele Antonio Di Pietro
Institut Montpelliérain Alexander
Grothendieck
University of Montpellier
Montpellier, France

Luca Formaggia 
MOX-Laboratory for Modeling and
Scientific Computing
Dipartimento di Matematica
Politecnico di Milano
Milan, Italy

Roland Masson
Laboratoire de Mathématiques
J. A. Dieudonné
Université Côte d'Azur
Nice, France

ISSN 2199-3041

ISSN 2199-305X (electronic)

SEMA SIMAI Springer Series

ISBN 978-3-030-69362-6

ISBN 978-3-030-69363-3 (eBook)

<https://doi.org/10.1007/978-3-030-69363-3>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

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

Preface

The numerical simulation of physical systems in geoscience applications entails several difficulties. Sources of complexity in the models include the presence of strongly heterogeneous coefficients (possibly leading to local degeneration), anisotropy, internal boundaries and strong nonlinearities. Realistic modelling additionally requires, in most circumstances, the coupling of different physics accounting, e.g., for the interaction between the fluid flow and the mechanical behaviour of the subsoil. The corresponding partial differential equations typically have to be solved in non-trivial domains, following sedimentation layers, and accounting for faults, fractures, or geometric features with very different scales.

The design and analysis of numerical technologies capable of handling such complex situations is an extremely active research field. In the last few years, one of the most relevant advances has been the development of numerical methods for linear and nonlinear problems supporting polyhedral meshes. The first endeavours to design numerical schemes supporting polyhedral meshes were independently undertaken in the context of finite volume and mimetic methods, focusing on low-order versions. Successful outcomes include Multi-Point Flux Approximation (MPFA) methods [1, 2, 20] (see also [3]), Hybrid Mixed Mimetic (HMM) methods [10, 19], and Vertex Approximate Gradient (VAG) methods [21]. Comprehensive reviews of low-order polyhedral methods can be found in [16, 17]; see also the introduction of [12] for a discussion of broader scope. The possibility to obtain high-order approximations on general meshes was explored later on. Among the first high-order polyhedral technologies, we can cite polytopal Discontinuous Galerkin (PolyDG) methods [4, 6]; see also [13, 14], where a comprehensive set of analysis tools was first developed. More recently, novel paradigms have resulted from the hybridization of the finite element and PolyDG paradigms with mimetic technologies. Particularly successful families of methods in this context include Virtual Element methods (VEM) (in their standard [8], mixed [9], and nonconforming [5] flavours) and Hybrid High-Order (HHO) methods [12, 15].

The need for general meshes in the numerical simulation of geological systems arises in several situations: in petroleum reservoir modelling, general polyhedral elements can appear, e.g., when transitioning from the radial mesh around wells to

the (structured) mesh used elsewhere; in petroleum basin modelling, fractures are typically incorporated into the numerical models by the mutual sliding of two portions of a corner-point grid along the fracture plane, resulting in highly non-conforming meshes; in this same context, significant mesh distortion can occur when coupled poromechanical models are considered; in the modelling of geological CO₂ storage, nonconforming meshes can appear, e.g., when local mesh adaptation is performed or in the presence of fractures, which can have a sizeable impact on the flow patterns. Polyhedral meshes may also ease up the grid generation in the presence of internal interfaces, like faults or the ones between sedimentation layers. Polyhedral mesh generation algorithms have been devised resorting to Voronoi tessellation [11, 22], or by modifying hexahedra meshing procedures, like in [23]. In the above and other situations, classical discretisation methods are either not viable, or require ad hoc modifications which can possibly add to the implementation complexity.

Polyhedral discretisation methods additionally pave the way to new computational strategies. In nuclear waste storage modelling, e.g., an accurate representation of the storage site traditionally requires the use of small elements, which can significantly add to the computational burden. With polyhedral meshes, on the other hand, one can incorporate small geometric features into larger agglomerated elements, thus achieving a significant cost reduction without compromising the accuracy; see, e.g., [4]. Polyhedral meshes can also be exploited to perform mesh adaptation by locally coarsening an underlying fine mesh, as in [6, 7].

This monograph collects state-of-the-art contributions on polyhedral methods for geoscience applications from top-level research groups. The methods and applications considered provide a wide overview of the subject, covering a significant portion of the most up to date research topics. Different points of view are represented, leading to a balanced mix of theoretical results, overviews of the state-of-the-art and numerical experimentation. The target audience of the book are graduate students and academics active in the field of numerical analysis and scientific computing, as well as researchers working in industry, environmental agencies, or research centres who have to deal with the complex endeavour of numerical simulation in geosciences.

An overview of the content of the monograph is provided in what follows:

Chapters 1–4 mainly focus on low-order polyhedral methods and their applications to nonlinear and/or coupled problems. Specifically, Chap. 1 introduces the novel Locally Enriched Polytopal Non-Conforming (LEPNC) method, first in the context of locally degenerate elliptic problems, then of more general, possibly nonlinear models covered by the Gradient Discretization framework [18]. The latter makes the object of Chap. 2, where its extension to degenerate parabolic equations of porous medium type is considered. The simulation of two-phase Darcy flows in fractured porous media is considered in Chap. 3, where the authors focus on VAG schemes. Finally, Chap. 4 contains a general overview of MPFA and of the derived Multi-Point Stress Approximation schemes followed by an application to thermo-poroelasticity.

Chapters 5–8 illustrate some applications of high-order methods of nonconforming, hybrid and mixed type; see [12, Chap. 5] for a broad discussion on the links among these and related methods. Specifically, Chap. 5 hinges on the use of PolyDG methods for the numerical modelling of seismic wave propagation and fractured reservoir simulations, with particular focus on their efficient implementation. In Chap. 6, the authors develop and analyse an HHO method for multiple-network poroelasticity relevant, e.g., in the simulation of fissured porous media. Chapter 7 focuses on a mixed VEM discretization of the Richards equations, modelling water flow in an unsaturated soil under the effect of gravity and the action of capillarity. Mixed VEM are also considered in Chap. 8, this time for single-phase flows in fractured porous media.

Montpellier, France
 Milan, Italy
 Nice, France

Daniele Antonio Di Pietro
 Luca Formaggia
 Roland Masson

References

1. I. Aavatsmark, T. Barkve, Ø. Bøe, T. Mannseth, Discretization on unstructured grids for inhomogeneous, anisotropic media. I. Derivation of the methods. *SIAM J. Sci. Comput.* **19**(5), 1700–1716 (1998)
2. I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth, Discretization on unstructured grids for inhomogeneous, anisotropic media. II. Discussion and numerical results. *SIAM J. Sci. Comput.* **19**(5), 1717–1736 (1998)
3. L. Agélas, D.A. Di Pietro, J. Droniou, The G method for heterogeneous anisotropic diffusion on general meshes. *ESAIM: Math. Model Numer. Anal.* **44**(4), 597–625 (2010)
4. P.F. Antonietti, S. Giani, P. Houston, *hp*-version composite discontinuous Galerkin methods for elliptic problems on complicated domains. *SIAM J. Sci. Comput.* **35**(3), A1417–A1439 (2013)
5. B. Ayuso de Dios, K. Lipnikov, G. Manzini, The nonconforming virtual element method. *ESAIM: Math. Model Numer. Anal.* **50**(3), 879–904 (2016)
6. F. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, P. Tesini, On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *J. Comput. Phys.* **231**(1), 45–65 (2012)
7. F. Bassi, L. Botti, A. Colombo, Agglomeration-based physical frame dG discretizations: an attempt to be mesh free. *Math. Models Methods Appl. Sci.* **24**(8), 1495–1539 (2014)
8. L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, A. Russo, Basic principles of virtual element methods. *Math. Models Methods Appl. Sci. (M3AS)* **199**(23), 199–214 (2013)
9. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, Mixed virtual element methods for general second order elliptic problems on polygonal meshes. *ESAIM: Math. Model. Numer. Anal.* **50**(3), 727–747 (2016)
10. L. Beirão da Veiga, K. Lipnikov, G. Manzini, *The mimetic finite difference method for elliptic problems*, volume 11 of *MS&A. Modeling, Simulation and Applications*, (Springer, Cham, 2014)
11. R.L. Berge, Ø.S. Klemetsdal, K.-A. Lie, Unstructured voronoi grids conforming to lower dimensional objects. *Comput. Geosci.* **23**(1), 169–188 (2019)

12. D.A. Di Pietro, J. Droniou, *The Hybrid High-Order method for polytopal meshes*. Number 19 in Modeling, Simulation and Application. (Springer International Publishing, 2020)
13. D.A. Di Pietro, A. Ern, Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations. *Math. Comp.* **79**(271), 1303–1330 (2010)
14. D.A. Di Pietro, A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*, (Springer, Heidelberg, 2012)
15. D.A. Di Pietro, A. Ern, A hybrid high-order locking-free method for linear elasticity on general meshes. *Comput. Meth. Appl. Mech. Engrg.* **283**, 1–21 (2015)
16. D.A. Di Pietro, M. Vohralík, A review of recent advances in discretization methods, a posteriori error analysis, and adaptive algorithms for numerical modeling in geosciences. *Oil Gas Sci. Technol.* **69**(4), 701–730 (2014)
17. J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci.* **24**(8), 1575–1619 (2014)
18. J. Droniou, R. Eymard, T. Gallouët, C. Guichard, R. Herbin, *The gradient discretisation method*, volume 82 of *Mathematics & Applications*, (Springer, 2018)
19. J. Droniou, R. Eymard, T. Gallouët, R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci. (M3AS)* **20**(2), 1–31 (2010)
20. M.G. Edwards, C.F. Rogers, Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Comput. Geosci.* **2**(4), 259–290 (1999), (1998)
21. R. Eymard, C. Guichard, R. Herbin, Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM Math. Model. Numer. Anal.* **46**(2), 265–290 (2012)
22. R. Merland, G. Caumon, B. Lévy, P. Collon-Drouaillet. Voronoi grids conforming to 3d structural features. *Comput. Geosci.* **18**(3–4), 373–383 (2014)
23. W. Oaks, S. Paoletti, Polyhedral mesh generation. In *IMR*, pages 57–67 (2000)

Contents

1 Non-conforming Finite Elements on Polytopal Meshes	1
Jérôme Droniou, Robert Eymard, Thierry Gallouët, and Raphaèle Herbin	
2 Error Estimates for the Gradient Discretisation Method on Degenerate Parabolic Equations of Porous Medium Type	37
Clément Cancès, Jérôme Droniou, Cindy Guichard, Gianmarco Manzini, Manuela Bastidas Olivares, and Iuliu Sorin Pop	
3 Nodal Discretization of Two-Phase Discrete Fracture Matrix Models	73
Konstantin Brenner, Julian Hennicker, and Roland Masson	
4 An Introduction to Multi-point Flux (MPFA) and Stress (MPSA) Finite Volume Methods for Thermo-poroelasticity	119
Jan Martin Nordbotten and Eirik Keilegavlen	
5 High-order Discontinuous Galerkin Methods on Polyhedral Grids for Geophysical Applications: Seismic Wave Propagation and Fractured Reservoir Simulations	159
Paola F. Antonietti, Chiara Facciola, Paul Houston, Ilario Mazzieri, Giorgio Pennesi, and Marco Verani	
6 A Hybrid High-Order Method for Multiple-Network Poroelasticity	227
Lorenzo Botti, Michele Botti, and Daniele A. Di Pietro	
7 The Mixed Virtual Element Method for the Richards Equation	259
Dibyendu Adak, Gianmarco Manzini, and Sundararajan Natarajan	

**8 Performances of the Mixed Virtual Element Method
on Complex Grids for Underground Flow** 299
Alessio Fumagalli, Anna Scotti, and Luca Formaggia

Index 331

Editors and Contributors

About the Editors

Daniele Antonio Di Pietro has been full professor since 2012 and is presently Director of Institut Montpelliérain Alexander Grothendieck at University of Montpellier (France). He is author of 3 research monographs published by Springer and of over 80 scientific papers published in refereed international journals or conference proceedings. He currently serves as associate editor for Numerical Algorithms (Springer). His research fields include the development and analysis of advanced numerical methods for partial differential equations, with applications to fluid and solid mechanics and porous media. Over his career, he has supervised 10 Ph.D. students and 6 post-doctoral fellows.

Luca Formaggia is Full Professor of Numerical Analysis since 2006 at Politecnico di Milano, Italy. He is author of more than 100 publications and Editor of 5 scientific Books published by Springer-Nature and Birkhauser. His scientific work addresses the study of numerical methods for partial differential equations, scientific computing, computational fluid dynamics with applications to computational geosciences, biomedicine and industrial problems. Currently, he is the President of the Italian Society of Applied and Industrial Mathematics and was the Head of the MOX Laboratory of Politecnico di Milano from 2012 to 2016. Over his career, he has supervised around 40 among Ph.D. students and post-docs. He is the co-Char of the 19th edition of the SIAM Conference on Mathematical and Computational Issues in the Geosciences.

Roland Masson is full professor of Mathematics at the University Côte d'Azur since 2011, member of the Jean-Alexandre Dieudonné department of Mathematics and of the Inria team Coffee. He was previously head of the Applied Mathematics department of the Institut Français du Pétrole (IFP) from 2000 to 2011. He received

his Ph.D. in Mathematics from Sorbonne University in 1999 and his Habilitation from Paris East University in 2006. He is an expert in numerical methods and scientific computing for subsurface flow and transport problems.

Contributors

Dibyendu Adak Department of Mechanical Engineering, Indian Institute of Technology Madras, Chennai, India

Paola F. Antonietti MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Lorenzo Botti Department of Engineering and Applied Sciences, University of Bergamo, Bergamo, Italy

Michele Botti MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Konstantin Brenner Université Côte d'Azur, CNRS, Inria, LJAD, Nice, France

Clément Cancès Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, Lille, France

Daniele A. Di Pietro IMAG, University of Montpellier, CNRS, Montpellier, France

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

Robert Eymard LAMA, Université Gustave Eiffel, UPEM, Marne-la-Vallée, France

Chiara Facciola MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Luca Formaggia MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Alessio Fumagalli MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Thierry Gallouët Institut de Mathématiques de Marseille, Aix-Marseille Université, Marseille, France

Cindy Guichard Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Université and Université de Paris, CNRS, Inria, Paris, France

Julian Hennicker University of Geneva, Geneva, Switzerland

Raphaèle Herbin Institut de Mathématiques de Marseille, Aix-Marseille Université, Marseille, France

Paul Houston School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham, UK

Eirik Keilegavlen Department of Mathematics, University of Bergen, Bergen, Norway

Gianmarco Manzini Istituto di Matematica Applicata e Tecnologie Informatiche “E. Magenes”, Pavia, Italy

Roland Masson Université Côte d’Azur, CNRS, Inria, LJAD, Nice, France

Ilario Mazzieri MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Sundararajan Natarajan Department of Mechanical Engineering, Indian Institute of Technology Madras, Chennai, India

Jan Martin Nordbotten Department of Mathematics, University of Bergen, Bergen, Norway

Manuela Bastidas Olivares Hasselt University, Diepenbeek, Belgium

Giorgio Pennesi MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Iuliu Sorin Pop Hasselt University, Diepenbeek, Belgium

Anna Scotti MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milan, Italy

Marco Verani MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica, Politecnico di Milano, Milano, Italy

Chapter 1

Non-conforming Finite Elements on Polytopal Meshes



Jérôme Droniou, Robert Eymard, Thierry Gallouët, and Raphaële Herbin

Abstract In this work we present a generic framework for non-conforming finite elements on polytopal meshes, characterised by elements that can be generic polygons/polyhedra. We first present the functional framework on the example of a linear elliptic problem representing a single-phase flow in porous medium. This framework gathers a wide variety of possible non-conforming methods, and an error estimate is provided for this simple model. We then turn to the application of the functional framework to the case of a steady degenerate elliptic equation, for which a mass-lumping technique is required; here, this technique simply consists in using a different –piecewise constant– function reconstruction from the chosen degrees of freedom. A convergence result is stated for this degenerate model. Then, we introduce a novel specific non-conforming method, dubbed Locally Enriched Polytopal Non-Conforming (LEPNC). These basis functions comprise functions dedicated to each face of the mesh (and associated with average values on these faces), together with functions spanning the local \mathbb{P}^1 space in each polytopal element. The analysis of the interpolation properties of these basis functions is provided, and mass-lumping techniques are presented. Numerical tests are presented to assess the efficiency and the accuracy of this method on various examples. Finally, we show that generic polytopal non-conforming methods, including the LEPNC, can be plugged into the

J. Droniou

School of Mathematics, Monash University, Melbourne, Australia

e-mail: jerome.droniou@monash.edu

R. Eymard (✉)

LAMA, Université Gustave Eiffel, UPEM, 77447 Marne-la-Vallée, France

e-mail: robert.eynard@univ-eiffel.fr

T. Gallouët · R. Herbin

Institut de Mathématiques de Marseille, Aix-Marseille Université, Marseille, France

e-mail: thierry.gallouet@univ-amu.fr

R. Herbin

e-mail: raphaele.herbin@univ-amu.fr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

D. A. Di Pietro et al. (eds.), *Polyhedral Methods in Geosciences*,

SEMA SIMAI Springer Series 27,

https://doi.org/10.1007/978-3-030-69363-3_1

gradient discretization method framework, which makes them amenable to all the error estimates and convergence results that were established in this framework for a variety of models.

Keywords Nonconforming finite elements · General meshes

1.1 Introduction

Problems involving elliptic partial differential equations are often efficiently approximated by the Lagrange finite element method, yielding an approximation of the unknown functions at the nodes of the mesh. In some cases, it may however be more interesting to approximate the unknown functions at the centre of the faces of the mesh. This is for example the case for the Stokes and Navier-Stokes problems, where an approximation of the velocity of a fluid at the faces of the mesh leads to an easy way to take into account the conservation of fluid mass in each element. This property is the basis of the success of the Crouzeix-Raviart approximation for the incompressible Stokes and Navier-Stokes equations; see the seminal paper by Crouzeix and Raviart [5], and recent extensions including linear elasticity [7].

Another situation for which approximating functions at the face centre is highly relevant is found in underground flows in heterogeneous porous media. Several coupled models require to simultaneously solve an elliptic equation associated with the pressure of the fluid, and equations associated with the transport of species by different mechanisms including convection with the displacement of the fluid, diffusion/dispersion mechanisms, and chemical and thermodynamic reactions. In such cases, the accuracy of the model on relatively coarse meshes can only be obtained if the elements of the mesh are homogeneous, in order to compute the flows in the high permeability zones as precisely as possible, without integrating in these zones some porous volume belonging to low permeability zones. Non-conforming methods with unknowns at the face naturally lead to finite volume properties on the elements, which are useful for the discretisation of such coupled equations. Note that non-conforming methods are in some way strongly linked with mixed finite elements on the same mesh, in the sense that the matrix resulting from the mixed hybrid condensed formulation for the Raviart-Thomas finite element is the same as the non conforming P1 finite element [3, 18].

The aim of this chapter is twofold.

On one hand, we wish to provide a general framework for the functional basis of non-conforming methods on polytopal meshes. Polytopal meshes have elements that can be generic polygons or polyhedra; they have gained considerable interest because they allow to mesh complex geometries or match specific underground features. For example, in the framework of petroleum engineering, general hexahedra have been used for several years; numerical developments for the computation of porous flows on such grids may be found in [1], for multi-point flux approximation finite volume methods for instance, in [19] for multi-point mixed approximations, or in [14] for

mimetic finite difference methods. The use of polytopal meshes for underground flows has motivated so many papers that it is impossible to give an exhaustive list; we refer the reader to the introduction of [6] for a thorough literature review on the topic.

Let us focus on the non-conforming finite element method for second order differential forms, described on simplicial meshes for example in [4, 20]. By non-conforming finite element method we refer to a method such that:

- the restriction to each element of the approximate solution belongs to H^1 ,
- the approximate solution can be discontinuous at the common face between two elements everywhere, but some weak (averaged or at a certain point on the face) continuity is imposed,
- the approximate gradient is defined as the broken gradient, which is locally (i.e. on each cell) the gradient of the function.

The mathematical properties behind the nature of the continuity conditions at the faces, needed for the convergence of the method, are sometimes called the “patch test” [16]. In Sect. 1.2, we revisit these properties, plugging all the non-conforming methods into a broken continuous H^1 space defined on a general polytopal mesh. We thus obtain in Sect. 1.2.2, a general error estimate in the case of a linear elliptic equation in heterogeneous and anisotropic cases. Section 1.2 can be read as a simple introduction, using a basic linear model as illustration, to generic non-conforming finite-element methods on polytopal meshes.

In Sect. 1.3, we explore the use of these methods on a more challenging model, which is however very relevant to applications in geosciences: a nonlinear degenerate elliptic equation of the Stefan or porous medium equation type. We introduce in Sect. 1.3.2 a mass lumping technique, which is mandatory for designing robust numerical schemes for this model.

We then focus, in Sect. 1.4, on a new specific non-conforming approximation on general polytopal meshes, called the Locally Enriched Polytopal Non-Conforming (LEPNC) method. This method is based on the H^1 piecewise approximation, imposing the continuity of the mean value on the interfaces. The advantage of the method presented here is its robustness, which is not the case for other possible simpler methods, such as choosing on each cell polynomials of degree k with $\dim \mathbb{P}^k(\mathbb{R}^d)$ larger than or equal to the number of faces of the polytopal cell (this condition is necessary to obtain a decent approximation, see e.g. the hexagonal example of Sect. 1.4, but it is not sufficient to solve robustness issues, see Remark 1.7). In particular, the LEPNC method allows for hanging nodes which frequently occur when meshing two different zones such as in domain decomposition methods. Another important feature of the finite element method presented here is that it can be used together with \mathbb{P}^1 nonconforming finite elements on simplicial parts of the mesh. The LEPNC basis functions are described in Sects. 1.4.1–1.4.2, and the approximation properties of the method are detailed in Sect. 1.4.3. The convergence theorems for the LEPNC method are given in Sect. 1.4.5. Various numerical tests are then proposed in Sect. 1.4.6, showing the accuracy and the efficiency of this method on problems presenting some complex features.

Section 1.5 covers the generic analysis of the convergence of non-conforming methods, which is encompassed in the framework of the Gradient Discretization method [9]. Some perspectives are then drawn in Sect. 1.6.

1.2 Principles of Polytopal Non-conforming Approximations

1.2.1 The Model: Linear Single-Phase Incompressible Flows in Porous Media

The principles of a generic polytopal non-conforming method are first presented on the following linear model of pressure for a single-phase incompressible flow in a porous medium:

$$\begin{cases} -\operatorname{div}(\Lambda \nabla \bar{u}) = f + \operatorname{div}(\mathbf{F}) & \text{in } \Omega, \\ \bar{u} = 0 & \text{on } \partial\Omega, \end{cases} \quad (1.1)$$

with the following assumptions on the data:

- Ω is a polytopal open subset of \mathbb{R}^d ($d \in \mathbb{N}^*$), (1.2a)

- Λ is a measurable function from Ω to the set of $d \times d$ symmetric matrices and there exists $\underline{\lambda}, \bar{\lambda} > 0$ such that, for a.e. $\mathbf{x} \in \Omega$, $\Lambda(\mathbf{x})$ has eigenvalues in $[\underline{\lambda}, \bar{\lambda}]$, (1.2b)

- $f \in L^2(\Omega)$, $\mathbf{F} \in L^2(\Omega)^d$. (1.2c)

We note in passing that a polytopal open set is simply a bounded polygon (if $d = 2$) or polyhedron (if $d = 3$) without slit, that is, it lies everywhere on one side of its boundary; see [9, Section 7.1.1] for a more formal definition.

The solution to (1.1) is to be understood in the standard weak sense:

$$\begin{aligned} &\text{Find } \bar{u} \in H_0^1(\Omega) \text{ such that, } \forall v \in H_0^1(\Omega), \\ &\int_{\Omega} \Lambda \nabla \bar{u} \cdot \nabla v \, d\mathbf{x} = \int_{\Omega} f v \, d\mathbf{x} - \int_{\Omega} \mathbf{F} \cdot \nabla v \, d\mathbf{x}. \end{aligned} \quad (1.3)$$

1.2.2 Polytopal Non-conforming Method

A polytopal non-conforming scheme for (1.3) is obtained by replacing the continuous space $H_0^1(\Omega)$ in this weak formulation by a finite-dimensional subspace of a “non-conforming Sobolev space”. Let us first give the definition of polytopal mesh we will be working with; this definition is a simplified version of [9, Definition 7.2].

Definition 1.1 (*Polytopal mesh*) Let Ω satisfy Assumption (1.2a). A polytopal mesh of Ω is a triplet $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P})$, where:

1. \mathcal{M} is a finite family of non empty connected polytopal open disjoint subsets of Ω (the “cells”) such that $\overline{\Omega} = \cup_{K \in \mathcal{M}} \overline{K}$. For any $K \in \mathcal{M}$, $\partial K = \overline{K} \setminus K$ is the boundary of K , $|K| > 0$ is the measure of K and h_K denotes the diameter of K , that is the maximum distance between two points of \overline{K} .
2. $\mathcal{F} = \mathcal{F}_{\text{int}} \cup \mathcal{F}_{\text{ext}}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the “faces” of the mesh – “edges” in 2D), such that any $\sigma \in \mathcal{F}_{\text{int}}$ is contained in Ω and any $\sigma \in \mathcal{F}_{\text{ext}}$ is contained in $\partial\Omega$. Each $\sigma \in \mathcal{F}$ is assumed to be a non empty open subset of a hyperplane of \mathbb{R}^d , with a strictly positive $(d - 1)$ -dimensional measure $|\sigma|$, and a relative interior $\overline{\sigma} \setminus \sigma$ of zero $(d - 1)$ -dimensional measure. We denote by \overline{x}_σ the centre of mass of σ . Furthermore, for all $K \in \mathcal{M}$, there exists a subset \mathcal{F}_K of \mathcal{F} such that $\partial K = \cup_{\sigma \in \mathcal{F}_K} \overline{\sigma}$. We set $\mathcal{M}_\sigma = \{K \in \mathcal{M} : \sigma \in \mathcal{F}_K\}$ and assume that, for all $\sigma \in \mathcal{F}$, either \mathcal{M}_σ has exactly one element and then $\sigma \in \mathcal{F}_{\text{ext}}$, or \mathcal{M}_σ has exactly two elements and then $\sigma \in \mathcal{F}_{\text{int}}$. For $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, $\mathbf{n}_{K,\sigma}$ is the (constant) unit vector normal to σ outward to K .
3. $\mathcal{P} = (\mathbf{x}_K)_{K \in \mathcal{M}}$ is a family of points of Ω such that $\mathbf{x}_K \in K$ for all $K \in \mathcal{M}$. We denote by $d_{K,\sigma}$ the signed orthogonal distance between \mathbf{x}_K and $\sigma \in \mathcal{F}_K$ (see Fig. 1.1), that is:

$$d_{K,\sigma} = (\mathbf{x} - \mathbf{x}_K) \cdot \mathbf{n}_{K,\sigma}, \text{ for all } \mathbf{x} \in \sigma. \quad (1.4)$$

(Note that $(\mathbf{x} - \mathbf{x}_K) \cdot \mathbf{n}_{K,\sigma}$ is constant for $\mathbf{x} \in \sigma$.) We then assume that each cell $K \in \mathcal{M}$ is strictly star-shaped with respect to \mathbf{x}_K , that is $d_{K,\sigma} > 0$ for all $\sigma \in \mathcal{F}_K$. This implies that for all $\mathbf{x} \in K$, the line segment $[\mathbf{x}_K, \mathbf{x}]$ is included in K .

For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_K$, we denote by $D_{K,\sigma}$ the pyramid with vertex \mathbf{x}_K and basis σ , that is

$$D_{K,\sigma} = \{t\mathbf{x}_K + (1 - t)\mathbf{y} : t \in (0, 1), \mathbf{y} \in \sigma\}. \quad (1.5)$$

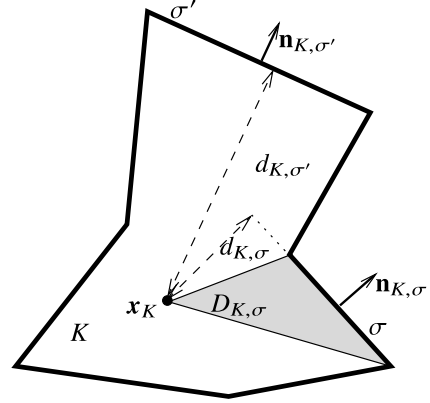
We denote, for all $\sigma \in \mathcal{F}$, $D_\sigma = \bigcup_{K \in \mathcal{M}_\sigma} D_{K,\sigma}$ (this set is called the “diamond” associated with the face σ , and for obvious reasons $D_{K,\sigma}$ is also referred to as an “half-diamond”).

The size of the polytopal mesh is defined by $h_{\mathcal{M}} = \sup\{h_K : K \in \mathcal{M}\}$ and the mesh regularity parameter $\gamma_{\mathfrak{T}}$ is defined by:

$$\gamma_{\mathfrak{T}} = \max_{K \in \mathcal{M}} \left(\max_{\sigma \in \mathcal{F}_K} \frac{h_K}{d_{K,\sigma}} + \text{Card}(\mathcal{F}_K) \right) + \max_{\sigma \in \mathcal{F}_{\text{int}}, \mathcal{M}_\sigma = \{K,L\}} \left(\frac{d_{K,\sigma}}{d_{L,\sigma}} + \frac{d_{L,\sigma}}{d_{K,\sigma}} \right). \quad (1.6)$$

We can now define the notion of non-conforming Sobolev space, which is built from the standard broken Sobolev space on a mesh by imposing some weak continuity property between the cells.

Fig. 1.1 A cell K of a polytopal mesh



Definition 1.2 (*Non-conforming $H_0^1(\Omega)$ space*) Let $\mathfrak{T} = (\mathcal{M}, \mathcal{F}, \mathcal{P})$ be a polytopal mesh of Ω in the sense of Definition 1.1. The non-conforming $H_0^1(\Omega)$ space on \mathfrak{T} , denoted by $H_{\mathfrak{T},0}^1$, is the space of all functions $w \in L^2(\Omega)$ such that:

1. [*H^1 -regularity in each cell*] For all $K \in \mathcal{M}$, the restriction $w|_K$ of w to K belongs to $H^1(K)$. The trace of $w|_K$ on $\sigma \in \mathcal{F}_K$ is denoted by $w|_{K,\sigma}$.
2. [*Continuity of averages on internal faces*] For all $\sigma \in \mathcal{F}_{\text{int}}$ with $\mathcal{M}_\sigma = \{K, L\}$,

$$\int_{\sigma} w|_{K,\sigma} = \int_{\sigma} w|_{L,\sigma}. \quad (1.7)$$

3. [*Homogeneous Dirichlet BC for averages on external faces*] For all $\sigma \in \mathcal{F}_{\text{ext}}$ with $\mathcal{M}_\sigma = \{K\}$,

$$\int_{\sigma} w|_{K,\sigma} = 0. \quad (1.8)$$

If $w \in H_{\mathfrak{T},0}^1$, its “broken gradient” $\nabla_{\mathcal{M}} w$ is defined by

$$\forall K \in \mathcal{M}, \quad \nabla_{\mathcal{M}} w = \nabla(w|_K) \text{ in } K$$

and we set $\|w\|_{H_{\mathfrak{T},0}^1} := \|\nabla_{\mathcal{M}} w\|_{L^2(\Omega)^d}$.

It can easily be checked that $\|\cdot\|_{H_{\mathfrak{T},0}^1}$ is indeed a norm on $H_{\mathfrak{T},0}^1$. The continuity (1.7) is a “0-degree patch test”, and some functions in $H_{\mathfrak{T},0}^1(\Omega)$ are therefore not conforming (they do not belong to $H_0^1(\Omega)$). Actually, disregarding the boundary condition (1.8), the non-conforming Sobolev space strictly lies between the classical Sobolev space $H^1(\Omega)$ and the fully broken Sobolev space $H^1(\mathcal{M}) = \{v \in L^2(\Omega) : v|_K \in H^1(K) \text{ for all } K \in \mathcal{M}\}$.

A polytopal non-conforming approximation of (1.3) is obtained by selecting a finite-dimensional subspace $V_{\mathfrak{T},0} \subset H_{\mathfrak{T},0}^1$, by replacing, in this weak formulation,

the infinite-dimensional space $H_0^1(\Omega)$ by $V_{\mathcal{T},0}$, and by using broken gradients instead of standard gradients:

$$\begin{aligned} & \text{Find } u \in V_{\mathcal{T},0} \text{ such that, } \forall v \in V_{\mathcal{T},0}, \\ & \int_{\Omega} \Lambda \nabla_{\mathcal{M}} u \cdot \nabla_{\mathcal{M}} v \, dx = \int_{\Omega} f v \, dx - \int_{\Omega} \mathbf{F} \cdot \nabla_{\mathcal{M}} v \, dx. \end{aligned} \quad (1.9)$$

Since $\|\cdot\|_{H_{\mathcal{T},0}^1}$ is a norm on $V_{\mathcal{T},0}$, the Lax-Milgram theorem immediately gives the existence and uniqueness of the solution to (1.9). The following error estimate is a straightforward consequence of the analysis carried out in Sect. 1.5 (see in particular Theorem 1.6 and Proposition 1.1).

Theorem 1.1 (Error estimates for polytopal non-conforming methods) *We assume that the solution \bar{u} of (1.3) and the data Λ and \mathbf{F} in Hypotheses (1.2) are such that $\Lambda \nabla \bar{u} + \mathbf{F} \in H^1(\Omega)^d$. Let $V_{\mathcal{T},0}$ be a finite-dimensional subspace of $H_{\mathcal{T},0}^1$ and let u be the solution of the non-conforming scheme (1.9). Then, there exists $C > 0$ depending only on Ω , $\underline{\lambda}$, $\bar{\lambda}$ in (1.2b) and increasingly depending on $\gamma_{\mathcal{T}}$ such that*

$$\|\bar{u} - u\|_{L^2(\Omega)} + \|\nabla \bar{u} - \nabla_{\mathcal{M}} u\|_{L^2(\Omega)^d} \leq Ch_{\mathcal{M}} \|\Lambda \nabla \bar{u} + \mathbf{F}\|_{H^1(\Omega)^d} + C \min_{v \in V_{\mathcal{T},0}} \|\bar{u} - v\|_{H_{\mathcal{T},0}^1}. \quad (1.10)$$

Remark 1.1 (Role of the terms in (1.10)) The term $Ch_{\mathcal{M}} \|\Lambda \nabla \bar{u} + \mathbf{F}\|_{H^1(\Omega)^d}$ in the right-hand side of (1.10) comes from the non-conformity of the space $V_{\mathcal{T},0}$, and from the fact that an exact Stokes formula is not satisfied in this space (as measured by $W_{\mathcal{D}}$ in Sect. 1.5.1). The minimum appearing in (1.10) measures the approximation properties of the space $V_{\mathcal{T},0}$, as in the second Strang lemma [15] (see $S_{\mathcal{D}}$ in Sect. 1.5.1).

1.3 Application to a Non-linear Model: Mass-Lumping

1.3.1 Model: Stationary Stefan/porous Medium Equation

We now consider the polytopal non-conforming approximation of a more challenging model, which encompasses the stationary versions of both the Stefan model and the porous medium equation:

$$\begin{cases} \bar{u} - \operatorname{div}(\Lambda \nabla \zeta(\bar{u})) = f + \operatorname{div}(\mathbf{F}) & \text{in } \Omega, \\ \zeta(\bar{u}) = 0 & \text{on } \partial\Omega. \end{cases} \quad (1.11)$$

A related unsteady problem is treated in Chap. 2; see, in particular, (2.1). We still assume that (1.2) holds and, additionally, that

$$\begin{aligned} & \zeta : \mathbb{R} \rightarrow \mathbb{R} \text{ is non-decreasing, } \zeta(0) = 0 \text{ and} \\ & \exists C_1, C_2 > 0 \text{ such that } |\zeta(s)| \geq C_1 |s| - C_2 \text{ for all } s \in \mathbb{R}. \end{aligned} \quad (1.12)$$

The weak form of (1.11) is

$$\begin{aligned} \text{Find } \bar{u} \in L^2(\Omega) \text{ such that } \zeta(\bar{u}) \in H_0^1(\Omega) \text{ and, } \forall v \in H_0^1(\Omega), \\ \int_{\Omega} (\bar{u}v + \Lambda \nabla \zeta(\bar{u}) \cdot \nabla v) dx = \int_{\Omega} f v dx - \int_{\Omega} \mathbf{F} \cdot \nabla v dx. \end{aligned} \quad (1.13)$$

1.3.2 Mass-Lumping

As explained in the introduction of [8] (see also Appendix B therein), using a standard (conforming or non-conforming) Galerkin approximation for (1.13) leads to a numerical scheme whose properties are difficult to establish. In particular, no convergence result seems attainable if $\mathbf{F} \neq 0$ and, in the case $\mathbf{F} = 0$, only weak convergence can be obtained in general. Instead, a modified approximation must be considered that uses a mass-lumping operator for the reaction term.

Specifically, let $V_{\mathfrak{T},0}$ be a subspace of $H_{\mathfrak{T},0}^1$; we select a basis $(\chi_i)_{i \in I}$ of $V_{\mathfrak{T},0}$ and disjoint subsets $(U_i)_{i \in I}$ of Ω , and we define the mass-lumping operator $\Pi_{\mathfrak{T}} : V_{\mathfrak{T},0} \rightarrow L^\infty(\Omega)$ by:

$$\forall v = \sum_{i \in I} v_i \chi_i, \quad \Pi_{\mathfrak{T}} v = \sum_{i \in I} v_i \mathbf{1}_{U_i}, \quad (1.14)$$

where $\mathbf{1}_{U_i}(\mathbf{x}) = 1$ if $\mathbf{x} \in U_i$ and $\mathbf{1}_{U_i}(\mathbf{x}) = 0$ otherwise. Note that the design of $\Pi_{\mathfrak{T}}$ actually depends on $V_{\mathfrak{T},0}$, and not just on the polytopal mesh \mathfrak{T} , but the natural notation $\Pi_{V_{\mathfrak{T},0}}$ has been simplified to $\Pi_{\mathfrak{T}}$ for legibility.

The function $\Pi_{\mathfrak{T}} v$ is piecewise constant and can be considered a good substitute of v , provided that each v_i represents some approximate value of v on U_i . In this setting, it also makes sense to define $\zeta(v) \in V_{\mathfrak{T},0}$ by applying the non-linear function ζ component-wise:

$$\forall v = \sum_{i \in I} v_i \chi_i, \quad \zeta(v) = \sum_{i \in I} \zeta(v_i) \chi_i.$$

Remark 1.2 (Mass-lumping of the non-conforming \mathbb{P}^1 method) Let us illustrate the mass-lumping process on the non-conforming \mathbb{P}^1 method on a simplicial mesh. A basis of its space is given by $(\chi_\sigma)_{\sigma \in \mathcal{F}_{\text{int}}}$, where each χ_σ is piecewise linear in each element, with value 1 at the centre of σ and 0 at the centres of all other faces. A mass-lumping operator $\Pi_{\mathfrak{T}}$ for this method is constructed in the following way: for each $v = \sum_{\sigma \in \mathcal{F}_{\text{int}}} v_\sigma \chi_\sigma$, let $\Pi_{\mathfrak{T}} v$ be the piecewise constant function equal to v_σ on each diamond D_σ , $\sigma \in \mathcal{F}_{\text{int}}$, (and $\Pi_{\mathfrak{T}} v = 0$ on the half-diamonds around boundary faces), see Fig. 1.2 for an illustration.

A non-conforming approximation of (1.13) is then obtained replacing $H_0^1(\Omega)$ by $V_{\mathfrak{T},0}$, ∇ with $\nabla_{\mathcal{M}}$ and using Π_V in the reaction and source terms:

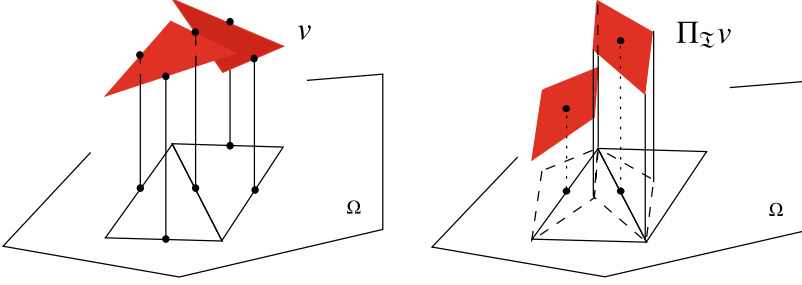


Fig. 1.2 Example of a non-conforming \mathbb{P}^1 function (left) and its mass-lumped version (right)

$$\begin{aligned} & \text{Find } u \in V_{\mathcal{T},0} \text{ such that, } \forall v \in V_{\mathcal{T},0}, \\ & \int_{\Omega} (\Pi_{\mathcal{T}} u \Pi_{\mathcal{T}} v + \Lambda \nabla_{\mathcal{M}} \zeta(u) \cdot \nabla_{\mathcal{M}} v) dx = \int_{\Omega} f \Pi_{\mathcal{T}} v dx - \int_{\Omega} \mathbf{F} \cdot \nabla_{\mathcal{M}} v dx. \end{aligned} \quad (1.15)$$

Remark 1.3 (Computing the source and reaction terms) In practice, the right-hand side in (1.15) is never computed exactly, but through a low order quadrature rule on f , assuming that f is approximated by a piecewise constant function on each U_i . If f is continuous, for example, one can take

$$\int_{\Omega} f \Pi_{\mathcal{T}} v dx \approx \sum_{i \in I} |U_i| f(\mathbf{x}_i) v_i$$

where \mathbf{x}_i is a point selected in or close to U_i . The reaction term in (1.15) is trivial to (exactly) compute:

$$\int_{\Omega} \Pi_{\mathcal{T}} u \Pi_{\mathcal{T}} v dx = \sum_{i \in I} |U_i| u_i v_i.$$

The matrix associated with this term in the scheme is therefore diagonal, as expected. These considerations show that only the measures of $(U_i)_{i \in I}$ are actually needed to implement (1.15).

The following convergence theorem results from the analysis in Sect. 1.5—see Theorems 1.7 and 1.8 together with Lemma 1.4. Error estimates could also be stated, but they are more complicated to present and require stronger assumptions on the solution to the Stefan equation; we therefore refer the interested reader to [8] for details, in which a partial uniqueness result is also stated for the solution of (1.15). We also mention in passing that error estimates for transient Stefan/porous medium equations are established in Chap. 2; these estimates are stated in the generic framework of the Gradient Discretisation Method, which covers polytopal non-conforming methods.

Theorem 1.2 (Convergence of polytopal non-conforming methods for the Stefan problem) *Let $\gamma > 0$ be a fixed number, and let $(\mathcal{T}_m)_{m \in \mathbb{N}}$ be a sequence of polytopal*

meshes such that $\gamma_{\mathcal{T}_m} \leq \gamma$ for all $m \in \mathbb{N}$ and such that $h_{\mathcal{M}_m} \rightarrow 0$ as $m \rightarrow \infty$. For each $m \in \mathbb{N}$, take a finite-dimensional subspace $V_{\mathcal{T}_m,0}$ of $H_{\mathcal{T}_m,0}^1$ and a mass-lumping operator $\Pi_{\mathcal{T}_m} : V_{\mathcal{T}_m,0} \rightarrow L^\infty(\Omega)$ as in (1.14), and assume the following:

$$\min_{v \in V_{\mathcal{T}_m,0}} \|\phi - v\|_{H_{\mathcal{T}_m,0}^1} \rightarrow 0 \text{ as } m \rightarrow \infty, \quad \forall \phi \in H_0^1(\Omega), \quad (1.16)$$

$$\max_{v \in V_{\mathcal{T}_m,0} \setminus \{0\}} \frac{\|v - \Pi_{\mathcal{T}_m} v\|_{L^2(\Omega)}}{\|\nabla_{\mathcal{M}_m} v\|_{L^2(\Omega)^d}} \rightarrow 0 \text{ as } m \rightarrow \infty. \quad (1.17)$$

Then, for all $m \in \mathbb{N}$ there exists $u_m \in V_{\mathcal{T}_m,0}$ solution of (1.15) and, as $m \rightarrow \infty$, $\Pi_{\mathcal{T}_m} \zeta(u_m) \rightarrow \zeta(\bar{u})$ strongly in $L^2(\Omega)$, $\nabla_{\mathcal{M}_m} \zeta(u_m) \rightarrow \nabla \zeta(\bar{u})$ strongly in $L^2(\Omega)^d$, and $\Pi_{\mathcal{T}_m} u_m \rightarrow \bar{u}$ weakly in $L^2(\Omega)$, where \bar{u} is a solution to (1.13).

1.4 A Locally Enriched Polytopal Non-conforming Finite Element Scheme

We describe here a non-conforming method that can be applied to almost any polytopal mesh as per Definition 1.1. Actually, the only additional assumption we make on the mesh is the following:

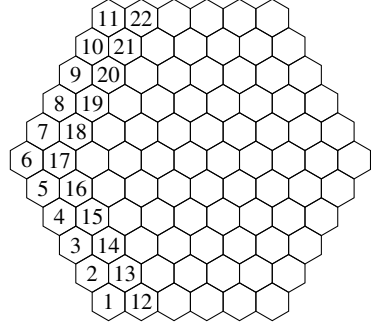
$$\forall \sigma \in \mathcal{F}, \quad \sigma \text{ is convex.} \quad (1.18)$$

This convexity assumption on the face is rather weak, and the cells themselves can be non-convex – which is often the case in 3D.

Let us first describe the underlying idea. To ensure the consistency of the method, a basic requirement would be for the local spaces (restriction of $V_{\mathcal{T},0}$ to a cell $K \in \mathcal{M}$) to contain $\mathbb{P}^1(K)$. Denoting by $\mathbb{P}^1(\mathcal{M})$ the space of piecewise linear functions on the mesh, without continuity conditions, this means that we should have $\mathbb{P}^1(\mathcal{M}) \cap H_{\mathcal{T},0}^1 \subset V_{\mathcal{T},0}$. This suggests to take $\mathbb{P}^1(\mathcal{M}) \cap H_{\mathcal{T},0}^1$ as our non-conforming finite-dimensional space. However, if the number of faces of most of the elements is greater than $d + 1$, the constraints of continuity at the faces will impede a correct interpolation. For instance, on a domain Ω that can be meshed by uniform hexagons (see Fig. 1.3), the space $\mathbb{P}^1(\mathcal{M}) \cap H_{\mathcal{T},0}^1$ is reduced to $\{0\}$. Indeed, the three boundary conditions on the exterior edges of element 1 imply that the constant gradient vanishes in element 1. Therefore the mean values at the three interior edges of element 1 also vanish, so that the same reasoning holds in element 2. By induction, the gradient vanishes in all the elements of the mesh.

We therefore enrich this initial space with functions associated with the faces, that we use to ensure the proper continuity conditions by “localising” the basis of \mathbb{P}^1 inside each element. The resulting global basis is made of functions associated with the faces and of additional local functions on the cell. As a consequence, we call the corresponding method the Locally Enriched Polytopal Non-Conforming finite element method (LEPNC for short).

Fig. 1.3 Hexagonal mesh



Remark 1.4 (Link with the non conforming \mathbb{P}^1 finite element method) Note that, when applied to a triangular mesh in 2D, the LEPNC yields 6 degrees of freedom on each triangle, while the classical non conforming \mathbb{P}^1 finite element (NCP1FE) method has only 3. However, when performing static condensation (see Remark 1.12) on the LEPNC scheme on triangles, only the 3 degrees of freedom pertaining to the faces remain, so that the computational cost is close to that of the NCP1FE scheme. In fact, the precision of the methods are close. Moreover, in the case of an elliptic equation with non homogeneous Dirichlet boundary conditions and a zero right hand side, the approximate solutions given by the NCP1FE and the condensed LEPNC schemes are identical.

1.4.1 Local Space

We first describe the local spaces and shape functions. Let $K \in \mathcal{M}$, for $\sigma \in \mathcal{F}_K$, the pyramid $D_{K,\sigma}$ has σ as one of its faces, as well as faces τ that are internal to K , and gathered in the set $\mathcal{F}_{K\sigma,\text{int}}$; see Fig. 1.4 for an illustration.

Let $\phi_{K,\sigma} : K \rightarrow \mathbb{R}$ be the piecewise-polynomial function such that, inside $D_{K,\sigma}$, $\phi_{K,\sigma}$ is the product of the distances to each internal face $\tau \in \mathcal{F}_{K\sigma,\text{int}}$, and outside $D_{K,\sigma}$ we set $\phi_{K,\sigma} = 0$. Additionally, $\phi_{K,\sigma}$ is scaled in order to have an average equal to one on σ . The function $\phi_{K,\sigma}$ vanishes on all the faces of $D_{K,\sigma}$ except σ . Under the convexity assumption (1.18) and letting $\mathbf{n}_{K\sigma,\tau}$ be the outer unit normal to $D_{K,\sigma}$ on $\tau \in \mathcal{F}_{K\sigma,\text{int}}$, we therefore set

$$\phi_{K,\sigma}(\mathbf{x}) = c_{K,\sigma} \prod_{\tau \in \mathcal{F}_{K\sigma,\text{int}}} [(\mathbf{x}_K - \mathbf{x}) \cdot \mathbf{n}_{K\sigma,\tau}]^+ \quad \forall \mathbf{x} \in K, \quad (1.19)$$

where $s^+ = \max(s, 0)$ is the positive part of $s \in \mathbb{R}$. As previously mentioned, $c_{K,\sigma} > 0$ is chosen to ensure that $\phi_{K,\sigma}$ has an average of one on σ ; since this function vanishes outside $D_{K,\sigma}$, this means that we have

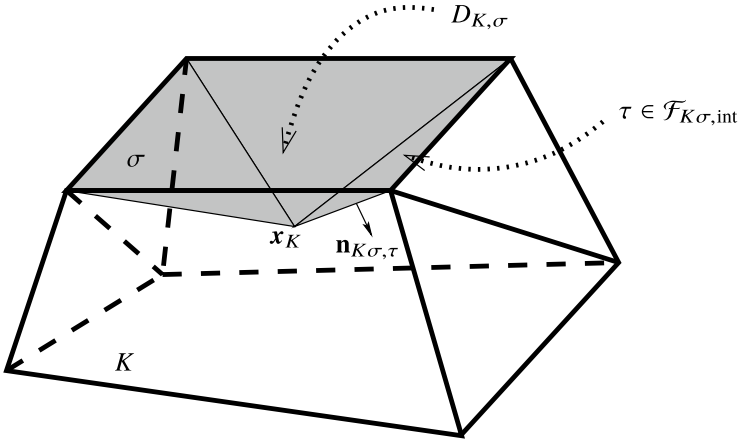


Fig. 1.4 Notations for the design of the local polytopal non-conforming space of Sect. 1.4.1

$$\frac{1}{|\sigma|} \int_{\sigma} \phi_{K,\sigma} = 1, \text{ and } \int_{\sigma'} \phi_{K,\sigma} = 0 \quad \forall \sigma' \in \mathcal{F}_K \setminus \{\sigma\}. \quad (1.20)$$

We then define the local space on K of the LEPNC method by

$$V_K^{\text{LEPNC}} := \text{span}(\mathbb{P}^1(K) \cup \{\phi_{K,\sigma} : \sigma \in \mathcal{F}_K\}). \quad (1.21)$$

The component $\mathbb{P}^1(K)$ will be responsible for the approximation properties of the global space, whereas the face-based basis functions will be used to glue local spaces together and ensure (1.7).

Remark 1.5 (Nature of the functions in the local space) The functions of V_K^{LEPNC} are continuous on K , and polynomial in each pyramid $D_{K,\sigma}$ for $\sigma \in \mathcal{F}_K$. The maximal polynomial degree of functions in V_K^{LEPNC} is $\max_{\sigma \in \mathcal{F}_K} \text{Card}(\mathcal{E}_{\sigma})$, where \mathcal{E}_{σ} is the set of edges of σ (vertices in 2D, in which case the maximal degree is 2).

A practical implementation of any non-conforming method requires to integrate the local functions and their gradients on each cell. For V_K^{LEPNC} , this is very easy: one simply has to select quadrature rules in K that are constructed by assembling quadrature rules on each pyramid. This is actually a standard way of constructing quadrature rules on polytopal cells, these pyramids being then cut into tetrahedra on which quadrature rules are known.

1.4.2 Global LEPNC Space and Basis of Functions

The global non-conforming space of the Locally Enriched Polytopal Non-Conforming method is

$$V_{\mathcal{T},0}^{\text{LEPNC}} = \{v \in H_{\mathcal{T},0}^1 : v|_K \in V_K^{\text{LEPNC}} \quad \forall K \in \mathcal{M}\}. \quad (1.22)$$

By construction of $(V_K^{\text{LEPNC}})_{K \in \mathcal{M}}$, an explicit and local basis of $V_{\mathcal{T},0}^{\text{LEPNC}}$ can be constructed thanks to the functions $(\phi_{K,\sigma})_{K \in \mathcal{M}, \sigma \in \mathcal{F}_K}$. For each $\sigma \in \mathcal{F}$, first define the function $\phi_\sigma : \Omega \rightarrow \mathbb{R}$ by patching the local functions, in the cells on each side of σ , associated with σ :

$$(\phi_\sigma)|_K = \phi_{K,\sigma} \quad \forall K \in \mathcal{M}_\sigma, \quad (\phi_\sigma)|_L = 0 \text{ if } L \notin \mathcal{M}_\sigma. \quad (1.23)$$

The properties (1.20) ensure that ϕ_σ satisfies 1. and 2. in Definition 1.2 (it also satisfies 3. if $\sigma \in \mathcal{F}_{\text{int}}$). We also note that each ϕ_σ is a sort of bubble function on the diamond D_σ , as it vanishes on all its faces (but, contrary to standard bubble functions, ϕ_σ is not in $H^1(D_\sigma)$).

We then select, for each $K \in \mathcal{M}$, $d + 1$ vertices (s_0, \dots, s_d) of K which maximise the volume of their convex hull, that is, maximise their determinant; in fact the determinant only needs to be non-zero, but maximising it leads to better conditioned matrices. We then define the nodal basis $(\psi_{K,i})_{i=0,\dots,d}$ of $\mathbb{P}^1(K)$ associated to these vertices, that is, the basis that satisfies $\psi_{K,i}(s_j) = 1$ if $i = j$ and 0 if $i \neq j$. We will see in Sect. 1.4.4 that this choice is relevant for mass lumping techniques. For each $i = 0, \dots, d$, we set

$$\phi_{K,i} = \psi_{K,i} - \sum_{\sigma \in \mathcal{F}_K} \bar{\psi}_{K,i,\sigma} \phi_{K,\sigma} \quad \text{with} \quad \bar{\psi}_{K,i,\sigma} = \frac{1}{|\sigma|} \int_\sigma \psi_{K,i}. \quad (1.24)$$

This choice ensures that

$$\int_\sigma \phi_{K,i} = 0 \quad \forall \sigma \in \mathcal{F}_K. \quad (1.25)$$

Extended by 0 outside K , each $\phi_{K,i}$ therefore belongs to $H_{\mathcal{T},0}^1$. It can also easily be checked that $\{\phi_{K,i} : i = 0, \dots, d\} \cup \{\phi_{K,\sigma} : \sigma \in \mathcal{F}_K\}$ spans V_K^{LEPNC} (the basis $(\psi_{K,i})_{i=0,\dots,d}$ of $\mathbb{P}^1(K)$ can be obtained by linear combinations of these functions). As shown in the following lemma, a basis of $V_{\mathcal{T},0}^{\text{LEPNC}}$ is then obtained by gathering all the functions (1.23) (for internal faces) and (1.24).

Lemma 1.1 (Basis of the LEPNC global space) *The following family forms a basis of $V_{\mathcal{T},0}^{\text{LEPNC}}$ defined by (1.22):*

$$\{\phi_{K,i} : K \in \mathcal{M}, i = 0, \dots, d\} \cup \{\phi_\sigma : \sigma \in \mathcal{F}_{\text{int}}\}. \quad (1.26)$$

Moreover, for any $v \in V_{\mathcal{T},0}^{\text{LEPNC}}$ we have

$$v = \sum_{K \in \mathcal{M}} \sum_{i=0}^d v_{K,i} \phi_{K,i} + \sum_{\sigma \in \mathcal{F}_{\text{int}}} v_\sigma \phi_\sigma, \quad (1.27)$$

with

$$v_\sigma = \frac{1}{|\sigma|} \int_\sigma v \quad \forall \sigma \in \mathcal{F}_{\text{int}}. \quad (1.28)$$

and, for all $K \in \mathcal{M}$,

$$v_{K,i} = v|_K(s_i) \quad \forall i = 0, \dots, d. \quad (1.29)$$

Remark 1.6 (Single-valuedness of v_σ) We note that, since $v \in H_{\mathbb{T},0}^1$, the condition (1.7) ensures that v_σ is uniquely defined by (1.28) (it depends only on σ , not on the choice of a cell in \mathcal{M}_σ in which we would consider the values of v).

Proof Proving (1.27)–(1.29) for a generic $v \in V_{\mathbb{T},0}^{\text{LEPNC}}$ shows that (1.26) spans this space, and also that it is a linearly independent family since all coefficients in the right-hand side of (1.27) vanish when the left-hand side v vanishes.

Let us take $v \in V_{\mathbb{T},0}^{\text{LEPNC}}$. It suffices to show that (1.27) holds on each cell $K \in \mathcal{M}$. Since $\{\phi_{K,i} : i = 0, \dots, d\} \cup \{\phi_{K,\sigma} : \sigma \in \mathcal{F}_K\}$ spans $V_K^{\text{LEPNC}} \ni v|_K$, there are coefficients $(\lambda_{K,i})_{i=0,\dots,d}$ and $(\lambda_{K,\sigma})_{\sigma \in \mathcal{F}_K}$ such that

$$v|_K = \sum_{i=0}^d \lambda_{K,i} \phi_{K,i} + \sum_{\sigma \in \mathcal{F}_K} \lambda_{K,\sigma} \phi_\sigma. \quad (1.30)$$

Taking the average over one face $\sigma \in \mathcal{F}_K$ and using (1.20) and (1.25), we obtain

$$\lambda_{K,\sigma} = \frac{1}{|\sigma|} \int_\sigma v|_K.$$

Hence, by Remark 1.6, $\lambda_{K,\sigma} = v_\sigma$ defined by (1.28). Applying now (1.30) at one of the vertices s_i , recalling the definition (1.24), the fact that $(\psi_{K,j})_{j=0,\dots,d}$ is the nodal basis associated with $(s_j)_{j=0,\dots,d}$, and noticing that all functions $\phi_{K,\sigma}$ vanish at the vertices of K (consequence of (1.19) and of the fact that each vertex either does not belong to $D_{K,\sigma}$, or belongs to one face in $\mathcal{F}_{K\sigma,\text{int}}$), we see that $v|_K(s_i) = \lambda_{K,i}$. To summarise, (1.30) is written

$$v|_K = \sum_{i=0}^d v_{K,i} \phi_{K,i} + \sum_{\sigma \in \mathcal{F}_K \cap \mathcal{F}_{\text{int}}} v_\sigma \phi_\sigma, \quad (1.31)$$

the restriction of the last sum to internal edges coming from $\int_\sigma v = 0$ whenever $\sigma \in \mathcal{F}_{\text{ext}}$, see (1.8). Since all functions $\phi_{L,i}$ vanish on K whenever $L \neq K$, and all ϕ_σ vanish on K whenever $\sigma \notin \mathcal{F}_K$, (1.31) proves that (1.27) holds on K . \square

Let $C(\mathcal{M})$ denote the functions whose restriction to each $K \in \mathcal{M}$ is continuous on \overline{K} . Lemma 1.1 shows us how to define a natural interpolator $\mathcal{I}_{\mathbb{T}} : H^1(\Omega) \cap C(\mathcal{M}) \rightarrow V_{\mathbb{T},0}^{\text{LEPNC}}$: for all $u \in H^1(\Omega) \cap C(\mathcal{M})$:

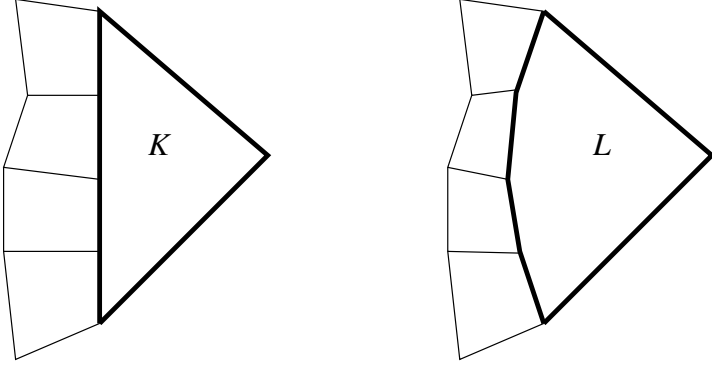


Fig. 1.5 Hexagons with aligned (left) and almost aligned (right) edges

$$\mathcal{I}_{\mathfrak{T}} u = \sum_{K \in \mathcal{M}} \sum_{i=0}^d u_{K,i} \phi_{K,i} + \sum_{\sigma \in \mathcal{F}_{\text{int}}} u_{\sigma} \phi_{\sigma} \quad (1.32a)$$

where $(u_{\sigma})_{\sigma \in \mathcal{F}_{\text{int}}}$ and $(u_{K,i})_{K \in \mathcal{M}, i=0, \dots, d}$ are defined by

$$u_{\sigma} = \frac{1}{|\sigma|} \int_{\sigma} u \quad \forall \sigma \in \mathcal{F}_{\text{int}}, \quad (1.32b)$$

$$u_{K,i} = u|_K(s_i) \quad \forall K \in \mathcal{M}, \quad \forall i = 0, \dots, d. \quad (1.32c)$$

Remark 1.7 (The need to enrich the bubble functions) As the above construction shows (see in particular (1.24)), the design of a finite-dimensional subspace of the non-conforming space $H_{\mathfrak{T},0}^1$ requires access, for each face σ of each cell K , to a local basis function that has average 1 on σ and 0 on all other faces of K . Instead of using the bubble functions (1.19), an alternative idea is to use a rich enough space of polynomial functions. The question of “how rich” this space should be (which degree the polynomials should have) is however not easy to answer, when considering generic polytopal meshes.

Consider for example the cell K on the left of Fig. 1.5, an hexagon with 4 aligned edges. Since it has a total of 6 edges, the minimum local space of polynomial should be $\mathbb{P}^2(K)$, which has dimension 6. However, the restrictions of functions in $\mathbb{P}^2(K)$ on the line of the aligned edges are polynomials of degree 2 in dimension 1, and form therefore a space of dimension 3. This space is not large enough to contain, for each of the 4 edges, a function with average 1 on this edge and 0 on all other edges. This shows that we should at least consider $\mathbb{P}^3(K)$ as the local polynomial space on K ; note that this argument only discusses the space dimension: it would still have to be fully established that $\mathbb{P}^3(K)$ is indeed rich enough.

The situation is perhaps more severe, from the robustness point of view, for the hexagon L on the right of Fig. 1.5. Since its edges are not aligned, from the pure dimensional point of view it might be sufficient to consider $\mathbb{P}^2(L)$ as the local poly-

nomial space on L . However, because L has *almost aligned* edges, the basis functions we would construct (with average 1 on one edge and 0 on all other edges) would form an “almost dependent” set of functions – even more so as the edges become more and more aligned, e.g. along a sequence of refined meshes. The practical consequence is that, in an implementation of the scheme using these basis functions, some local mass or stiffness matrices would be close to singular, which would lead to an ill-conditioned global system and a poor numerical resolution.

On the contrary, the usage of the (piecewise-polynomial) basis functions (1.19) solves these two issues: the local space is always defined as the span of \mathbb{P}^1 and the bubble functions, independently of the cell geometry, and, even when edges become aligned, the basis functions remain well independent (recall that the vertices (s_0, \dots, s_d) are chosen in each cell to maximise the volume they encompass and thus, in Fig. 1.5, they would be chosen as the three leftmost vertices in each case and would not become aligned or close to aligned).

1.4.3 Approximation Properties of the LEPNC Space

The approximation properties of the LEPNC space require a slightly more stringent, but still very flexible, regularity condition on the meshes than the boundedness of $\gamma_{\mathfrak{T}}$ (see (1.6)).

Definition 1.3 (*ρ -regular polytope and polytopal mesh*) A polytopal open set $K \subset \mathbb{R}^d$ is said to be a ρ -regular polytope, where $\rho > 0$, if:

1. There exists $\mathbf{x}_K \in K$ and open disjoint simplices $(K_i)_{i=1, \dots, n}$ such that $\overline{K} = \bigcup_{i=1}^n \overline{K}_i$, and, for $i = 1, \dots, n$, \mathbf{x}_K is a vertex of K_i , exactly one face of K_i is included in ∂K and all the other faces of K_i are common with a neighbouring simplex K_j .
2. There exists $\mathbf{x}_{K_i} \in K_i$ such that $B(\mathbf{x}_{K_i}, \rho h_K) \subset K_i$.

A ρ -regular polytopal mesh of Ω is a polytopal mesh \mathfrak{T} as per Definition 1.1, such that any cell $K \in \mathcal{M}$ is a ρ -regular polytope and if, for any simplex K_i as above, there exists $\sigma \in \mathcal{F}_K$ such that one face of K_i is included in σ .

Remark 1.8 (*ρ -regular polytope and polytopal mesh*) The number n in Definition 1.3 is always bounded by $1/\rho^d$, the ratio of the measure of $B(\mathbf{x}_K, h_K)$ and that of $B(\mathbf{x}_{K_i}, \rho h_K)$. As a consequence, it can be easily checked that $\gamma_{\mathfrak{T}}$ (defined by (1.6)) is bounded above by a real number depending only on ρ .

The additional requirement, for a polytopal mesh, that one face of K_i is included in one of the mesh face prevents the situation where the face of K_i that lies in ∂K is actually split between two mesh faces (the mesh faces could be different from the geometrical faces of its elements, e.g. in case of non-conforming meshes with hanging nodes).

To state approximation properties of the global non-conforming space (1.22), we first define an alternate interpolator, which does not require the functions to be continuous on each cell and therefore enjoys boundedness properties for a larger class of functions. For all $K \in \mathcal{M}$, let $\mathcal{J}_K : H^1(K) \rightarrow V_K^{\text{LEPNC}}$ be such that

$$\mathcal{J}_K u = \mathcal{J}_{\mathcal{F}_K} u + P_K(u - \mathcal{J}_{\mathcal{F}_K} u) \quad \forall u \in H^1(K), \quad (1.33)$$

where

$$\mathcal{J}_{\mathcal{F}_K} u = \sum_{\sigma \in \mathcal{F}_K} u_\sigma \phi_{K,\sigma} \quad \text{with } (u_\sigma)_{\sigma \in \mathcal{F}_K} \text{ given by (1.32b)}, \quad (1.34)$$

and $P_K : L^2(K) \rightarrow V_K^{\text{LEPNC}}$ is the L^2 -orthogonal projector on $\text{span}\{\phi_{K,i} : i = 0, \dots, d\}$. The global interpolator $\mathcal{J}_{\mathfrak{T}} : H_0^1(\Omega) \rightarrow V_{\mathfrak{T},0}^{\text{LEPNC}}$ is obtained patching the local ones:

$$(\mathcal{J}_{\mathfrak{T}} u)|_K = \mathcal{J}_K(u|_K) \quad \forall u \in H_0^1(\Omega), \quad \forall K \in \mathcal{M}.$$

Using (1.20) and (1.25), it is easily verified that $\mathcal{J}_{\mathfrak{T}} u$ indeed belongs to $V_{\mathfrak{T},0}^{\text{LEPNC}}$.

Theorem 1.3 (Approximation properties of $V_{\mathfrak{T},0}^{\text{LEPNC}}$) *Assume that \mathfrak{T} is a ρ -regular polytopal mesh. Then, there exists C depending only on ρ such that*

$$\|u - \mathcal{J}_{\mathfrak{T}} u\|_{L^2(\Omega)} + h_{\mathcal{M}} \|\nabla_{\mathcal{M}}(u - \mathcal{J}_{\mathfrak{T}} u)\|_{L^2(\Omega)} \leq Ch_{\mathcal{M}}^2 |u|_{H^2(\Omega)} \quad \forall u \in H_0^1(\Omega) \cap H^2(\Omega), \quad (1.35)$$

where $|\cdot|_{H^2(\Omega)}$ denotes the $H^2(\Omega)$ -seminorm.

Remark 1.9 (Approximation properties in generic Sobolev spaces) Using the results of [6, Chap. 1], a straightforward adaptation of the proof below shows that the approximation property (1.35) also holds with L^2 , H_0^1 and H^2 replaced by L^p , $W_0^{1,p}$ and $W^{2,p}$, for any $p \in [1, \infty)$.

Before proving this theorem, let us establish the boundedness of the local interpolator \mathcal{J}_K .

Lemma 1.2 (Boundedness of \mathcal{J}_K) *Assume that K is a ρ -regular polytope. Then, there exists $C > 0$ depending only on ρ such that, for all $u \in H^1(K)$,*

$$\|\mathcal{J}_K u\|_{L^2(K)} \leq C(\|u\|_{L^2(K)} + h_K \|\nabla u\|_{L^2(K)^d}), \quad (1.36)$$

$$\|\nabla \mathcal{J}_K u\|_{L^2(K)^d} \leq C \|\nabla u\|_{L^2(K)^d}. \quad (1.37)$$

Proof In this proof, $C > 0$ denotes a generic real number, that can change from one line to the next but depends only on ρ .

Step 1: *Polynomial invariance of \mathcal{J}_K and estimates on the basis functions.*

The definitions (1.24) and (1.34) show that $\phi_{K,i} = \psi_{K,i} - \mathcal{J}_{\mathcal{F}_K} \psi_{K,i}$ for all $i = 0, \dots, d$. Hence, $P_K(\psi_{K,i} - \mathcal{J}_{\mathcal{F}_K} \psi_{K,i}) = P_K \phi_{K,i} = \phi_{K,i}$ and $\mathcal{J}_K \psi_{K,i} = \mathcal{J}_{\mathcal{F}_K} \psi_{K,i} + \phi_{K,i} = \psi_{K,i}$. Since $\mathbb{P}^1(K) = \text{span}\{\psi_{K,i} : i = 0, \dots, d\}$ this establishes the following polynomial invariance of \mathcal{J}_K :

$$\mathcal{J}_K q = q \quad \forall q \in \mathbb{P}^1(K). \quad (1.38)$$

The definition (1.19) and the ρ -regularity of K imply that $\phi_{K,\sigma} \geq c_{K,\sigma} Ch_\sigma^{n_\sigma}$ on a ball B_σ in σ of diameter Ch_σ , where h_σ is the diameter of σ and $n_\sigma = \text{Card}(\mathcal{F}_{K\sigma,\text{int}})$. Integrating this relation over B_σ , using (1.20) and noticing that $|\sigma| \leq C|B_\sigma|$, we infer $c_{K,\sigma} \leq Ch_\sigma^{-n_\sigma}$ and thus, since $h_K \leq Ch_\sigma$ by ρ -regularity of K ,

$$|\phi_{K,\sigma}| \leq C \quad \text{on } K. \quad (1.39)$$

The same definition (1.19) also yields $|\nabla\phi_{K,\sigma}| \leq c_{K,\sigma} Ch_K^{n_\sigma-1}$ on K , and therefore

$$|\nabla\phi_{K,\sigma}| \leq Ch_K^{-1} \quad \text{on } K. \quad (1.40)$$

Step 2: Estimate on $\nabla\mathcal{J}_K u$.

By (1.38), $\mathcal{J}_K 1 = 1$ and thus $\nabla\mathcal{J}_K u = \nabla\mathcal{J}_K(u - \bar{u}_K)$, where $\bar{u}_K = \frac{1}{|K|} \int_K u$, which implies

$$\nabla\mathcal{J}_K u = \nabla\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K) + \nabla P_K[(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)]. \quad (1.41)$$

Let us first estimate $\nabla\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)$. By [9, Est. (B.11)] we have

$$|u_\sigma - u_K|^2 \leq \frac{Ch_K}{|\sigma|} \int_K |\nabla u|^2 dx \quad \forall \sigma \in \mathcal{F}_K,$$

from which we deduce

$$|\nabla\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)| \leq C \sum_{\sigma \in \mathcal{F}_K} \frac{h_K}{(|\sigma|h_K)^{1/2}} \|\nabla u\|_{L^2(K)^d} |\nabla\phi_{K,\sigma}|.$$

The estimate (1.40) yields $\|\nabla\phi_{K,\sigma}\|_{L^2(K)^d} \leq Ch_K^{-1}|K|^{1/2}$ and thus, since $|K| \leq C|\sigma|h_K$ and $\text{Card}(\mathcal{F}_K) \leq C$ (consequence of Remark 1.8),

$$\|\nabla\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)^d} \leq C \|\nabla u\|_{L^2(K)^d}. \quad (1.42)$$

The same arguments with $\phi_{K,\sigma}$ instead of $\nabla\phi_{K,\sigma}$ and (1.39) instead of (1.40) yields

$$\|\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} \leq Ch_K \|\nabla u\|_{L^2(K)^d}. \quad (1.43)$$

We now turn to the second term in the right-hand side of (1.41). The range of P_K is contained in a space of piecewise polynomials, with uniformly bounded degree, on a regular subdivision of K . The inverse inequality of [6, Lemma 1.28 and Remark 1.33] therefore gives

$$\|\nabla P_K[(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)]\|_{L^2(K)^d} \leq Ch_K^{-1} \|P_K[(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)]\|_{L^2(K)}.$$

Since P_K is an L^2 -orthogonal projection, we infer

$$\begin{aligned} \|\nabla P_K[(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)]\|_{L^2(K)^d} &\leq Ch_K^{-1} \|(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} \\ &\leq Ch_K^{-1} \|u - \bar{u}_K\|_{L^2(K)} + Ch_K^{-1} \|\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} \\ &\leq C \|\nabla u\|_{L^2(K)^d}, \end{aligned} \quad (1.44)$$

where we have used $\|u - u_K\|_{L^2(K)} \leq Ch_K \|\nabla u\|_{L^2(K)^d}$ (see [9, Est. (B.12)]) and (1.43) in the last line. Combined with (1.42) and (1.41), this proves (1.37).

Step 3: Estimate on $\mathcal{J}_K u$.

We use the triangle inequality together with $\mathcal{J}_K \bar{u}_K = \bar{u}_K$ (see (1.38)) to write

$$\begin{aligned} \|\mathcal{J}_K u\|_{L^2(K)} &\leq \|\mathcal{J}_K(u - \bar{u}_K)\|_{L^2(K)} + \|\bar{u}_K\|_{L^2(K)} \\ &\leq \|\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} + \|P_K[(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)]\|_{L^2(K)} + \|u\|_{L^2(K)} \\ &\leq \|\mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} + \|(u - \bar{u}_K) - \mathcal{J}_{\mathcal{F}_K}(u - \bar{u}_K)\|_{L^2(K)} + \|u\|_{L^2(K)} \\ &\leq Ch_K \|\nabla u\|_{L^2(K)^d} + \|u\|_{L^2(K)}, \end{aligned}$$

where we have used the definition (1.33) of \mathcal{J}_K together with Jensen's inequality (to write $\|\bar{u}_K\|_{L^2(K)} \leq \|u\|_{L^2(K)}$) in the second line, and the same arguments that led to (1.44) to conclude. The proof of (1.36) is complete. \square

We can now complete the proof of Theorem 1.3.

Proof (Theorem 1.3) As in the proof of Lemma 1.2, C denotes here a generic constant that can change from one line to the other but depends only on ρ . Let $K \in \mathcal{M}$ and denote by q_1 the L^2 -orthogonal projection of $u|_K$ on $\mathbb{P}^1(K)$. By [6, Theorem 1.45], we have that

$$\|u - q_1\|_{L^2(K)} + h_K \|\nabla(u - q_1)\|_{L^2(K)^d} \leq Ch_K^2 |u|_{H^2(K)}. \quad (1.45)$$

Using the polynomial invariance (1.38) and the triangle inequality, we write, for $s = 0, 1$,

$$|u - \mathcal{J}_K u|_{H^s(K)} = |(u - q_1) - \mathcal{J}_K(u - q_1)|_{H^s(K)} \leq |u - q_1|_{H^s(K)} + |\mathcal{J}_K(u - q_1)|_{H^s(K)}.$$

The boundedness properties (1.36) and (1.37) together with the approximation property (1.45) then yield

$$|u - \mathcal{J}_K u|_{H^s(K)} \leq C(\|u - q_1\|_{L^2(K)} + h_K^{1-s} \|\nabla(u - q_1)\|_{L^2(K)^d}) \leq Ch_K^{2-s} |u|_{H^2(K)}.$$

Squaring, for each $s = 0, 1$, this inequality and summing over $K \in \mathcal{M}$ yields the estimate on each term in the left-hand side of (1.35). \square

1.4.4 Mass-Lumping of the LEPNC Method

As discussed in Sect. 1.3.2, approximating non-linear models such as (1.11) requires the usage of mass-lumping, which necessitates to identify a basis of $V_{\mathfrak{T},0}^{\text{LEPNC}}$ such that the coefficients of $v \in V_{\mathfrak{T},0}^{\text{LEPNC}}$ on this basis represent approximate values of v in some portions of Ω .

Definition 1.4 (*Mass-lumping operator for the LEPNC method*) Let $\varpi \in [0, 1]$ be a weight, representing the fraction of mass allocated to the faces. For each $K \in \mathcal{M}$, create a partition $((K_i)_{i=0,\dots,d}, (K_\sigma)_{\sigma \in \mathcal{F}_K})$ of K into $(d + 1) + \text{Card}(\mathcal{F}_K)$ sets, such that, for all $i = 0, \dots, d$ and $\sigma \in \mathcal{F}_K$,

$$s_i \in \overline{K_i}, \quad \bar{x}_\sigma \in \overline{K_\sigma}, \quad (1.46)$$

$$|K_i| = (1 - \varpi) \frac{|K|}{d + 1}, \quad |K_\sigma| = \varpi \frac{|K|}{\text{Card}(\mathcal{F}_K)}. \quad (1.47)$$

The mass-lumping operator $\Pi_{\mathfrak{T}}^{\text{LEPNC}} : V_{\mathfrak{T},0}^{\text{LEPNC}} \rightarrow L^\infty(\Omega)$ is then defined by: for all $v \in V_{\mathfrak{T},0}^{\text{LEPNC}}$,

$$\Pi_{\mathfrak{T}}^{\text{LEPNC}} v = \sum_{K \in \mathcal{M}} \sum_{i=0}^d v_{K,i} \mathbf{1}_{K_i} + \sum_{\sigma \in \mathcal{F}_{\text{int}}} v_\sigma \mathbf{1}_{K_\sigma},$$

with $(v_\sigma)_{\sigma \in \mathcal{F}_{\text{int}}}$ and $(v_{K,i})_{K \in \mathcal{M}, i=0,\dots,d}$ given by (1.28)–(1.29).

Remark 1.10 (*Shape of the partition of K*) Fig. 1.6 illustrates possible choices of regions K_i and K_σ . In practice, due to the usage of quadrature rules for source terms (see Remark 1.3), the precise shapes of these region are irrelevant. Only their measures are required to implement the scheme (1.15).

The following lemma shows that the above designed mass-lumping technique preserves the approximation properties of the LEPNC, see Lemma 1.4.

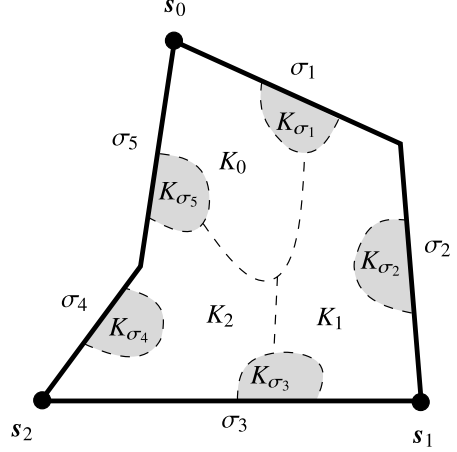
Lemma 1.3 (*Estimate for the mass-lumping operator of the LEPNC*) Let \mathfrak{T} be a ρ -regular polytopal mesh in the sense of Definition 1.3, and let $\Pi_{\mathfrak{T}}^{\text{LEPNC}}$ be given by Definition 1.4. Then, there exists $C > 0$ depending only on ρ and d such that

$$\|v - \Pi_{\mathfrak{T}}^{\text{LEPNC}} v\|_{L^2(\Omega)} \leq Ch_{\mathcal{M}} \|\nabla_{\mathcal{M}} v\|_{L^2(\Omega)^d} \quad \forall v \in V_{\mathfrak{T}}^{\text{LEPNC}}.$$

Proof In this proof, C is a real number that may vary, but depends only on ρ and d . Let $v \in V_{\mathfrak{T}}^{\text{LEPNC}}$. For all $K \in \mathcal{M}$, the function $v|_K$ is Lipschitz-continuous on K and the ρ -regularity of K together with the mean value theorem gives, for all $i = 0, \dots, d$ and $\sigma \in \mathcal{F}_K$,

$$|v_{K,i} - v| = |v|_K(s_i) - v| \leq Ch_K \|\nabla v|_K\|_{L^\infty(K)^d} \text{ on } K$$

Fig. 1.6 Regions for mass-lumping of the LEPNC method in dimension $d = 2$. Here, ϖ is small and most of the weight has been put on the three chosen vertices (s_0, s_1, s_2)



and

$$|v_\sigma - v| \leq Ch_K \|\nabla v|_K\|_{L^\infty(K)^d} \text{ on } K.$$

Writing $v|_K = \sum_{i=0}^d v \mathbf{1}_{K_i} + \sum_{\sigma \in \mathcal{F}_K} v \mathbf{1}_{K_\sigma}$ and subtracting the definition of $\Pi_{\mathfrak{T}}^{\text{LEPNC}} v$ we infer

$$|v|_K - (\Pi_{\mathfrak{T}}^{\text{LEPNC}} v)|_K \leq \sum_{i=0}^d Ch_K \|\nabla v|_K\|_{L^\infty(K)^d} \mathbf{1}_{K_i} + \sum_{\sigma \in \mathcal{F}_K} Ch_K \|\nabla v|_K\|_{L^\infty(K)^d} \mathbf{1}_{K_\sigma}.$$

Since $\nabla v|_K$ is piecewise polynomial on a regular subdivision of K , with a degree bounded above by a positive real number depending only on ρ , the inverse Lebesgue inequalities of [6, Lemma 1.25 and Remark 1.33] yield $\|\nabla v|_K\|_{L^\infty(K)^d} \leq C|K|^{-\frac{1}{2}} \|\nabla v|_K\|_{L^2(K)^d}$. Plugging this estimate into the above relation and using $\sum_{i=0}^d \mathbf{1}_{K_i} + \sum_{\sigma \in \mathcal{F}_K} \mathbf{1}_{K_\sigma} = 1$ on K , we infer

$$|v|_K - (\Pi_{\mathfrak{T}}^{\text{LEPNC}} v)|_K \leq Ch_K |K|^{-\frac{1}{2}} \|\nabla v|_K\|_{L^2(K)^d}.$$

The proof is complete by taking the $L^2(K)$ -norm of this estimate, squaring, summing over $K \in \mathcal{M}$ and taking the square root. \square

1.4.5 Convergence Results

Together with the above analysis of the LEPNC properties, the general nonconforming framework of Sect. 1.2 yields the following results. We first give an error estimate for the LEPNC approximation of the linear problem (1.1).

Theorem 1.4 (Error estimates for the LEPNC approximation) *We assume that the solution \bar{u} of (1.3) and the data Λ and \mathbf{F} in Hypotheses (1.2) are such that $\Lambda \nabla \bar{u} + \mathbf{F} \in H^1(\Omega)^d$ and $\bar{u} \in H^2(\Omega)$. Let \mathfrak{T} be a ρ -regular polytopal mesh in the sense of Definition 1.3. Let u be the solution of the non-conforming scheme (1.9), letting $V_{\mathfrak{T},0} = V_{\mathfrak{T},0}^{\text{LEPNC}}$ defined by (1.22). Then, there exists $C > 0$ depending only on Ω , ρ and $\underline{\lambda}, \bar{\lambda}$ in (1.2b) such that*

$$\|\bar{u} - u\|_{L^2(\Omega)} + \|\nabla \bar{u} - \nabla_{\mathcal{M}} u\|_{L^2(\Omega)^d} \leq Ch_{\mathcal{M}}(\|\Lambda \nabla \bar{u} + \mathbf{F}\|_{H^1(\Omega)} + |u|_{H^2(\Omega)}), \quad (1.48)$$

where $|\cdot|_{H^2(\Omega)}$ denotes the $H^2(\Omega)$ -seminorm.

Proof The result is an immediate consequence of Theorems 1.1 and 1.3. \square

Turning to the nonlinear problem (1.13), the following theorem states the convergence of the LEPNC method.

Theorem 1.5 (Convergence of the LEPNC method for the Stefan problem) *Let $\rho > 0$ be a fixed number, and let $(\mathfrak{T}_m)_{m \in \mathbb{N}}$ be a sequence of ρ -regular polytopal mesh polytopal meshes, in the sense of Definition 1.3, such that $h_{\mathcal{M}_m} \rightarrow 0$ as $m \rightarrow \infty$.*

Then, for all $m \in \mathbb{N}$, letting $V_{\mathfrak{T}_m,0} = V_{\mathfrak{T}_m,0}^{\text{LEPNC}}$ defined by (1.22) and $\Pi_{\mathfrak{T}_m} = \Pi_{\mathfrak{T}_m}^{\text{LEPNC}}$ from Definition 1.4, there exists u_m solution of (1.15) and, as $m \rightarrow \infty$, $\Pi_{\mathfrak{T}_m}^{\text{LEPNC}} \zeta(u_m) \rightarrow \zeta(\bar{u})$ strongly in $L^2(\Omega)$, $\nabla_{\mathcal{M}_m} \zeta(u_m) \rightarrow \nabla \zeta(\bar{u})$ strongly in $L^2(\Omega)^d$, and $\Pi_{\mathfrak{T}_m}^{\text{LEPNC}} u_m \rightarrow \bar{u}$ weakly in $L^2(\Omega)$, where \bar{u} is a solution to (1.13).

Proof We apply Theorem 1.2. Property (1.16) is a consequence of Theorem 1.3, and of the density of $H^2(\Omega) \cap H_0^1(\Omega)$ in $H_0^1(\Omega)$. Property (1.17) is proven by Lemma 1.3. \square

1.4.6 Numerical Tests

We present here some numerical results obtained by the LEPNC method on the linear single-phase incompressible flow (1.1) and on the Stefan/porous medium equation problem (1.11), on $\Omega = (0, 1)^2$ and with the diffusion tensor $\Lambda = \text{Id}$. The schemes we consider are therefore (1.9) and (1.15) with the space $V_{\mathfrak{T},0}^{\text{LEPNC}}$ and the mass-lumping operator $\Pi_{\mathfrak{T}}^{\text{LEPNC}}$. The tests below were run using the LEPNC implementation available in the HArDCore2D library [12]. We note that some of the tests here involve non-homogeneous Dirichlet boundary conditions; adapting the LEPNC scheme to this case is straightforward, and done as for standard non-conforming \mathbb{P}^1 finite elements. We also refer the interested reader to Chap. 2 for a numerical assessment of the LEPNC (and comparison with other methods) on the transient porous medium equation.

Let us first make some remarks relative to the practical implementation of these LEPNC schemes.

Remark 1.11 (Choice of implementation unknown for the Stefan model) Owing to Lemma 1.1, the unknowns for the implementation of the LEPNC represent function values $X_{K,i}$ at the chosen vertices s_i inside each cell $K \in \mathcal{M}$, and function values X_σ at the center of mass of each face $\sigma \in \mathcal{F}$ (such values are order 2 approximations of the averages appearing in (1.28)). When considering the scheme (1.15) for the Stefan problem and because of the plateaux of ζ , however, these values may not be values of u , but sometimes of $\zeta(u)$. Specifically, if $\varpi = 0$, then the face values of the unknowns u do not appear in the mass-matrix in each Newton iteration on (1.15); if we were to use these face values as unknown X_σ for the implementation, they would be multiplied in the stiffness matrix by $\zeta'(X_\sigma^{k-1})$, where X_σ^{k-1} is the face value at the previous Newton iteration; this factor $\zeta'(X_\sigma^{k-1})$ could vanish, leading to a zero line in the complete linear system. For this reason, when $\varpi = 0$, each X_σ should represent the value on σ of $\zeta(u)$, not u ; this way, when writing Newton iterations, no linearisation is performed on this unknown in the stiffness matrix, which ensures that it remains invertible. For the same reason, if $\varpi = 1$, each unknown $X_{K,i}$ should represent values at s_i of $\zeta(u)$, not u . We refer the reader to [8, Remark 3.1] for more on this topic.

Remark 1.12 (Static condensation of cell-based degrees of freedom) For each $K \in \mathcal{M}$, the basis functions $\{\tilde{\phi}_{K,i} : i = 0, \dots, d\}$ have support in K . In the linear systems to be solved (at each iteration of the Newton algorithm in the case of non-linear problems), the stencil of their associated unknowns therefore only contains the unknowns of the other basis functions related to K , and of the basis functions related to the faces of K . A static condensation process can thus be applied, exactly as in Hybrid High-Order methods (see [6, Appendix B.3.2]), to eliminate the cell-based unknowns. The resulting globally coupled linear system then only involves face-based unknowns, and two faces are in a stencil of this matrix only if they share a cell.

Remark 1.13 (Alternate construction of the basis functions) Instead of using the nodal basis functions $(\psi_{K,i})_{i=0,\dots,d}$ in (1.24), one can instead take the scaled and translated monomial basis functions: $\psi_{K,0} = 1$ and $\psi_{K,i}(\mathbf{x}) = \frac{x_i - \bar{x}_{K,i}}{h_K}$, where x_i is the i -th coordinate of \mathbf{x} and $x_{K,i}$ is the i -th coordinate of the centre of mass of K . The obtained basis $(\phi_{K,i})_{i=0,\dots,d}$ can afterwards be transformed by linear combinations into a nodal basis (ensuring that (1.27)–(1.29) holds). This implementation is the choice made in the HARDCore library.

When an analytical solution is available, we present error estimates in the following relative norms:

$$E_{L^2} := \frac{\|u - \mathcal{I}_{\mathcal{T}} \bar{u}\|_{L^2(\Omega)}}{\|\mathcal{I}_{\mathcal{T}} \bar{u}\|_{L^2(\Omega)}} \quad \text{and} \quad E_{H^1} := \frac{\|\nabla_{\mathcal{M}}(u - \mathcal{I}_{\mathcal{T}} \bar{u})\|_{L^2(\Omega)^d}}{\|\nabla_{\mathcal{M}} \mathcal{I}_{\mathcal{T}} \bar{u}\|_{L^2(\Omega)^d}}$$

for the linear model, and

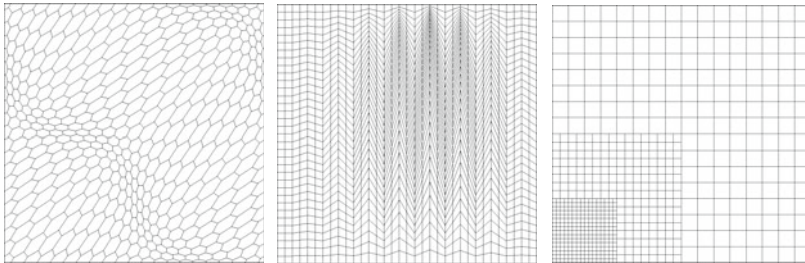


Fig. 1.7 Examples of members from the mesh families used in numerical tests: hexagonal (left), Kershaw (centre) and locally refined Cartesian (right)

$$E_{L^2, \text{ml}} := \frac{\|\Pi_{\mathcal{T}}^{\text{LEPNC}}(u - \mathcal{I}_{\mathcal{T}}\bar{u})\|_{L^2(\Omega)}}{\|\Pi_{\mathcal{T}}^{\text{LEPNC}}\mathcal{I}_{\mathcal{T}}\bar{u}\|_{L^2(\Omega)}} \quad \text{and} \quad E_{H^1, \zeta} := \frac{\|\nabla_{\mathcal{M}}(\zeta(u) - \mathcal{I}_{\mathcal{T}}\zeta(\bar{u}))\|_{L^2(\Omega)^d}}{\|\nabla_{\mathcal{M}}\mathcal{I}_{\mathcal{T}}\zeta(\bar{u})\|_{L^2(\Omega)^d}}$$

for the non-linear model; here \bar{u} is the exact analytical solution to (1.11), u is the solution to the LEPNC scheme, $\mathcal{I}_{\mathcal{T}}$ is the interpolator defined by (1.32), and $\Pi_{\mathcal{T}}^{\text{LEPNC}}$ is the mass-lumping operator given by Definition 1.4.

The tests have been run using three families of meshes, an example of each is represented in Fig. 1.7: (mostly) hexagonal meshes, Kershaw meshes and locally refined Cartesian meshes. The last two are taken from the FVCA5 Benchmark [13]. In all the tests we have chosen a mass-lumping weight ϖ of 0 on the edges; tests (not reported here) with other weights show similar results, except that the Newton iterations converge sometimes more slowly when mass is allocated to the edges.

1.4.6.1 Linear Single-Phase Incompressible Flow

We first test the LEPNC method on (1.1) with $\Lambda = \text{Id}$ and exact solution $\bar{u}(x, y) = \sin(\pi x) \sin(\pi y)$. For comparison, we also present the results obtained with the HHO(k, ℓ) method detailed in [6, Sect. 5.1], with degree of edge unknowns $k = 0$ and degree of element unknowns $\ell = 1$. The reason for choosing these particular (k, ℓ) is that the HHO(0, 1) method has (whether before or after static condensation) the same number of degrees of freedom as the LEPNC method. The results for the three families of meshes are presented in Fig. 1.8. Note that for the the HHO(0, 1) method, the error E_{H^1} is measured using the discrete H^1 -norm defined in [6, Eq. (2.35)], and E_{L^2} is computed from the L^2 -norm of the element unknowns.

As expected from Theorem 1.4, the rate of convergence of the LEPNC scheme in H^1 -norm is 1 on all three families of meshes. An improved rate of order 2 is observed in L^2 -norm and, even though it is not stated in Theorem 1.4, it is also quite expected since LEPNC is close to a lowest-order finite element method (we note that improved L^2 estimates can be obtained, using a Nitsche argument, in the context of the GDM [11]).

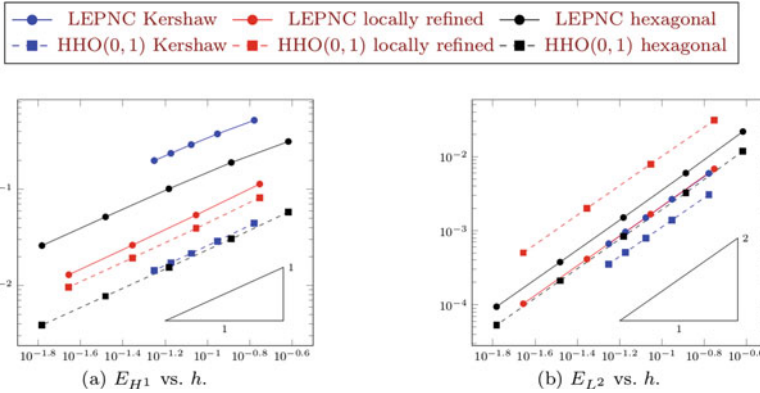


Fig. 1.8 Errors versus mesh size for the linear equation

In terms of H^1 -error, HHO(0, 1) seems to over-perform LEPNC on all meshes, especially on distorted ones (Kershaw, hexagonal) where the difference is a full order of magnitude; the difference is less perceptible on more regular meshes like the locally refined ones. This is also the case, although much less pronounced (factor 2 instead of a full order of magnitude), in L^2 -norm on hexagonal and Kershaw meshes; interestingly, the trend is actually reversed on locally refined meshes, with LEPNC providing an L^2 -error about five times smaller than HHO(0, 1), indicating that LEPNC seems to produce a better approximation of the solution itself (if not its gradient) on regular meshes. Of course, all these comparisons must be taken with a grain of salt since they do not exactly use the same norms. Additionally, it should be noted that the HHO(0, 1) scheme does not readily produce an explicit function that embeds all the methods' design (it is, in this sense, more of a *virtual* method), whereas LEPNC does.

1.4.6.2 Stefan Problem

We consider the problem (1.13) with the following Stefan non-linearity:

$$\zeta(s) = \begin{cases} s & \text{if } s \leq 0, \\ 0 & \text{if } 0 \leq s \leq 1, \\ s - 1 & \text{if } s \geq 1. \end{cases}$$

TEST S1. For this test, we take an exact smooth solution \bar{u} such that $\zeta(\bar{u})$ is also smooth, but not trivial (the solution \bar{u} crosses the value 0 at which ζ is not differentiable). Setting $s(x, y) = \frac{x+y}{\sqrt{2}}$ the coordinate along the first diagonal, the exact solution is $\bar{u}(x, y) = (s(x, y) - 0.5)^3$. The functions \bar{u} and $\zeta(\bar{u})$ are represented in Fig. 1.9

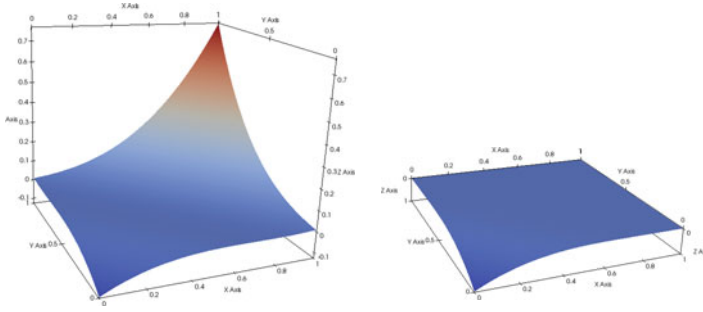


Fig. 1.9 Exact solution \bar{u} (left) and $\zeta(\bar{u})$ (right) for test S1

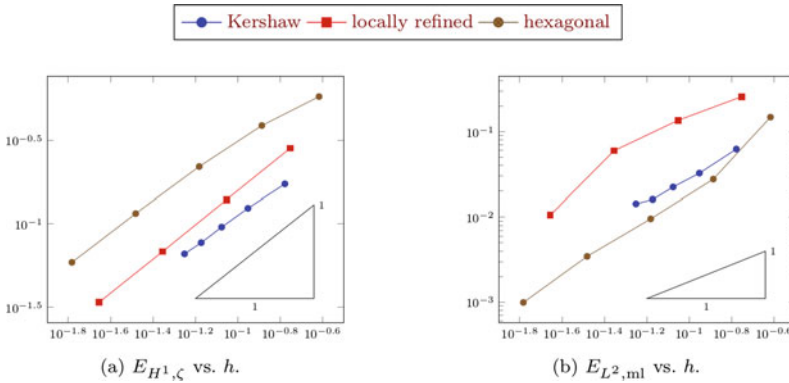


Fig. 1.10 Errors versus mesh size for test S1

The convergence graphs are given in Fig. 1.10. For solutions that are piecewise smooth on the mesh, the analysis of [8] shows that, for a low-order scheme as the LEPNC, the expected rate of convergence in energy error $E_{H^1, \zeta}$ for the regular variable $\zeta(u)$ is $\mathcal{O}(h)$, which corresponds to the rate observed for all three families in Fig. 1.10. The convergence rate in mass-lumped L^2 -norm on the u variable is always larger than one: it is almost 2 for the hexagonal and locally refined mesh families, and around 1.5 for the Kershaw family. This convergence is however less regular than the convergence on the variable $\zeta(u)$.

TEST S2. The previous test is not representative of the typical behaviour of solutions to Stefan problems. In the general case, and in particular with null source terms, these solutions \bar{u} are discontinuous in the range of values where ζ remains constant, which therefore does not prevent $\zeta(\bar{u})$ from being continuous. This next test case, taken from [8], displays such a behaviour. Setting $\gamma = \frac{1}{3}$, the exact solution is

$$\bar{u}(x, y) = \cosh(s(x, y) - \gamma) \text{ if } s(x, y) \geq \gamma, \quad \bar{u}(x, y) = 0 \text{ if } s(x, y) < \gamma,$$

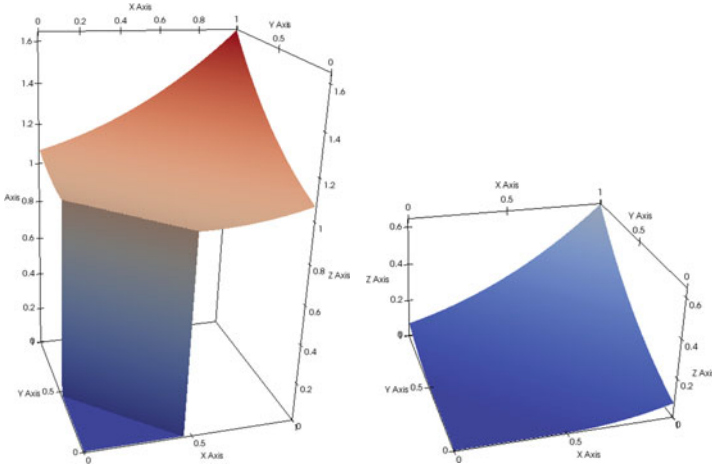


Fig. 1.11 Exact solution \bar{u} (left) and $\zeta(\bar{u})$ (right) for test S2

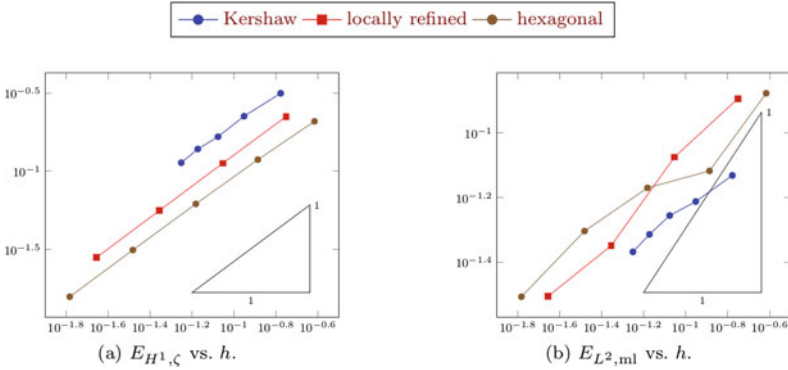


Fig. 1.12 Errors versus mesh size for test S2

where, as in Test S1, $s(x, y) = \frac{x+y}{\sqrt{2}}$ is the coordinate along the first diagonal. This solution is discontinuous along the line $s(x, y) = \gamma$, but $\zeta(\bar{u})$ is continuous (and even in $H^2(\Omega)$); see Fig. 1.11. This function corresponds to a zero source term in (1.11).

The convergence results are presented in Fig. 1.12. As expected from the results of [8], we observe in Fig. 1.12, left, an estimate of the kind $E_{H^1, \zeta} = \mathcal{O}(h)$. The convergence rate in mass-lumped L^2 error $E_{L^2, ml}$ for the variable u is however much lower (and, as in Test S1, rather irregular), which is expected since u is discontinuous; the overall convergence rate of $E_{L^2, ml}$ is about $\mathcal{O}(h^{0.6})$ for all mesh families. Figure 1.13 shows the approximate variables u and $\zeta(u)$ obtained on the second hexagonal mesh in the family; the discontinuity of \bar{u} , typical in Stefan’s problems, clearly impacts the convergence on this variable.

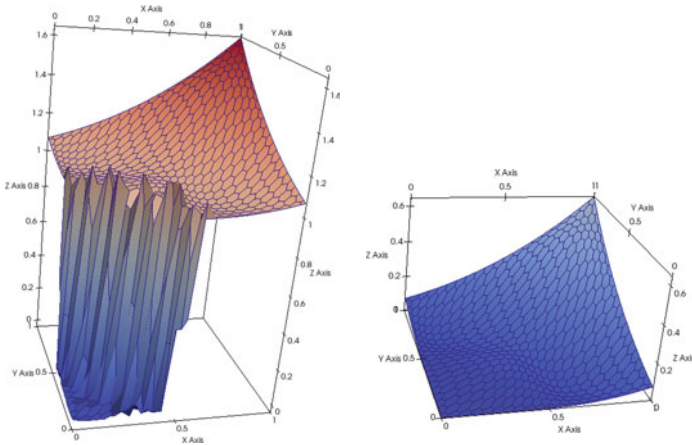


Fig. 1.13 Approximate solution u (left) and $\zeta(u)$ (right) obtained on the second member of the hexagonal mesh family in test S2

1.4.6.3 Porous Medium Equation

We now consider the stationary porous medium equation, corresponding to (1.11) with non-linearity

$$\zeta(s) = |s|^{m-1}s \text{ with } m \geq 1.$$

TEST P1. For this test, the exact solutions \bar{u} and $\zeta(\bar{u})$ are both smooth. We take $\bar{u}(x, y) = \sin(\pi x) \sin(\pi y)$, and $m \in \{1, 2, 3, 4\}$. Note that the case $m = 1$ actually corresponds to $\zeta(s) = s$, so (1.11) is the linear equation (1.1) with an added reaction term u . The results of the test, on the same Kershaw, locally refined and hexagonal meshes as in Tests S1 and S2, are presented in Fig. 1.14.

Looking first at the case $m = 1$, we notice that the results are worse on the Kershaw meshes; despite the smoothness of the solution, the distortion of these meshes impact the approximation error negatively. We still see an order $\mathcal{O}(h)$ convergence in both energy and mass-lumped L^2 norm; this is expected for the energy error given that LEPNC is a low-order scheme, but one could have hoped to see a super-convergence effect in the L^2 -norm. On the contrary, for locally refined and hexagonal meshes, this super-convergence is visible and the L^2 -norm error decays as $\mathcal{O}(h^2)$, while the energy norm decays as $\mathcal{O}(h)$.

Considering now the nonlinear cases $m = 2, 3, 4$, we see that the energy error still decays as h for the locally refined and hexagonal meshes. However, the L^2 -norm error no longer super-converges with an order 2, but rather with an order 1.5. The results for the Kershaw meshes show much lower convergence rates. For $m = 2$ rate for the L^2 -norm error is still close to 1, but the energy error only decays as about $\mathcal{O}(h^{0.5})$. For $m = 3, 4$, the rates in L^2 -norm and energy error are respectively 0.5 and 0.3 – at least at the considered mesh sizes. Looking at the pictures it seems

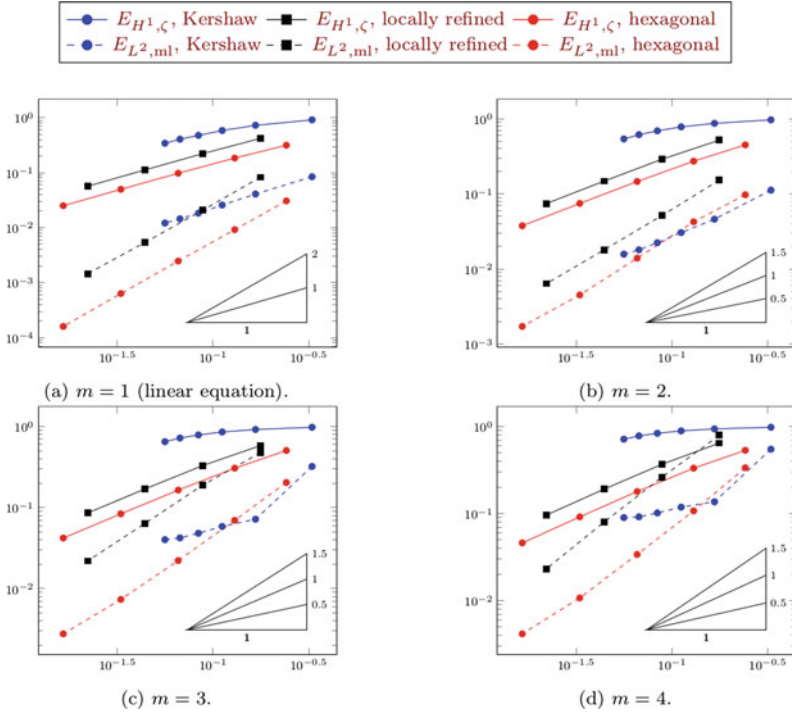


Fig. 1.14 Errors versus mesh size for test P1

that the rate in energy norm has a tendency to increase towards the last meshes in the Kershaw family. It should be mentioned here that for certain cases (typically, the finest hexagonal or Kershaw meshes, with $m = 3, 4$), a straightforward Newton algorithm does not converge and relaxation has to be applied.

TEST P2. This test features a less regular exact solution \bar{u} . We take $\bar{u}(x, y) = \max(\rho^2 - r(x, y)^2, 0)$, where $\rho = 0.3$ and $r(x, y)^2 = (x - 0.5)^2 + (y - 0.5)^2$. In the domain Ω , the graph of \bar{u} is the tip of a paraboloid; this solution belongs to $H^1(\Omega)$ but not to $H^2(\Omega)$. We take $m = 2$, so $\zeta(\bar{u}) \in H^2(\Omega)$. For this value of m , the singularity of \bar{u} at the circle $r(x, y)^2 = \rho^2$ is typical of the singularity exhibited by the Barenblatt solution in the transient setting [2, 17]. The results are presented in Fig. 1.15. As in Test P1, we see that the energy error decays as $\mathcal{O}(h)$, except for the very distorted Kershaw meshes for which a rate of about 0.3 is achieved with the last two meshes (further refinement might improve that rate). In terms of the L^2 -error, all three mesh families lead to a rate of convergence of about 1. Even for the relatively regular mesh families (hexahedral, locally refined), no super-convergence is observed. This is somehow expected given that the exact solution is not H^2 -regular.

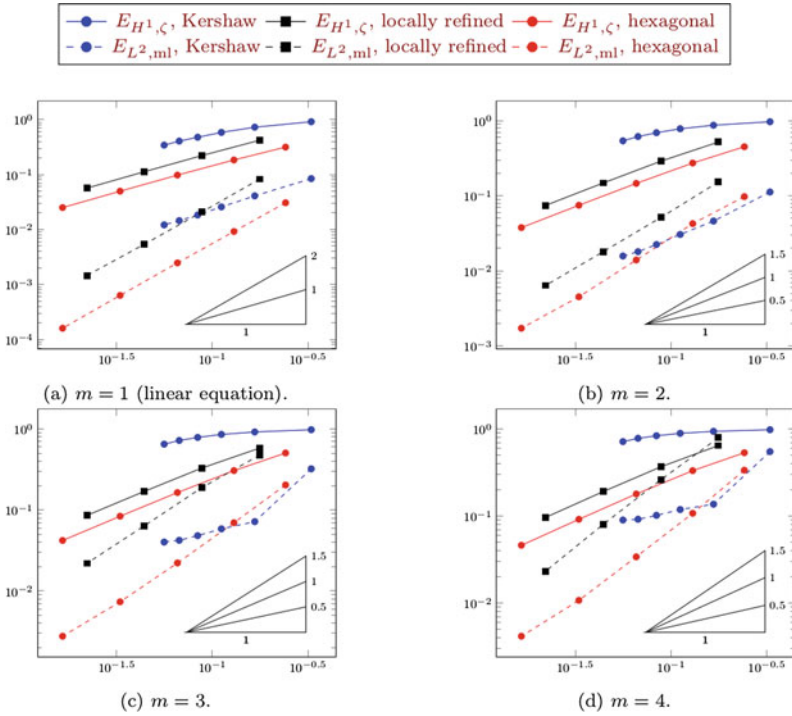


Fig. 1.15 Errors versus mesh size for test P2

1.5 Analysis of Polytopal Non-conforming Finite Element Schemes

Polytopal non-conforming finite element schemes are gradient discretisation methods (GDM) and, as such, enjoy all the error estimates and convergence results of GDMs. We recall here the notion of GDM and associated results, which yield in particular the Theorems 1.1 and 1.2. Most of the following material is taken from [9, Sect. 9.1].

1.5.1 Gradient Discretisation Method

The GDM is a generic framework for designing and analysing numerical schemes for elliptic and parabolic problems (although extensions to linear advection is also possible [10]). It consists in replacing, in the weak formulation of the model, the continuous space and operator by their discrete analogues given by a gradient discretisation (GD).

Definition 1.5 (*Gradient discretisation for homogeneous Dirichlet boundary conditions*) A gradient discretisation for homogeneous Dirichlet boundary conditions is a triplet $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}})$ where

- $X_{\mathcal{D},0}$ is a finite-dimensional space of unknowns, that encodes the homogeneous boundary conditions,
- $\Pi_{\mathcal{D}} : X_{\mathcal{D},0} \rightarrow L^2(\Omega)$ is a linear operator that reconstructs a function from a vector of unknowns,
- $\nabla_{\mathcal{D}} : X_{\mathcal{D},0} \rightarrow L^2(\Omega)^d$ is a linear operator that reconstructs a “gradient” from a vector of unknowns; it must be chosen such that $\|\nabla_{\mathcal{D}} \cdot\|_{L^2(\Omega)^d}$ is a norm on $X_{\mathcal{D},0}$.

A gradient discretisation \mathcal{D} is said to have a piecewise constant reconstruction if there exists a basis $(\mathbf{e}_i)_{i \in I}$ of $X_{\mathcal{D},0}$ and disjoint subsets $(U_i)_{i \in I}$ of Ω such that

$$\Pi_{\mathcal{D}}v = \sum_{i \in I} v_i \mathbf{1}_{U_i} \quad \forall v = \sum_{i \in I} v_i \mathbf{e}_i \in X_{\mathcal{D},0}, \quad (1.49)$$

where $\mathbf{1}_{U_i}$ is the characteristic function of U_i (equal to 1 in this set and to 0 elsewhere).

Once a GD \mathcal{D} is chosen, a gradient scheme (GS) for the linear diffusion problem (1.3) is obtained by writing:

$$\begin{aligned} &\text{Find } u \in X_{\mathcal{D},0} \text{ such that, } \forall v \in X_{\mathcal{D},0}, \\ &\int_{\Omega} \Lambda \nabla_{\mathcal{D}} u \cdot \nabla_{\mathcal{D}} v \, d\mathbf{x} = \int_{\Omega} f \Pi_{\mathcal{D}} v \, d\mathbf{x} - \int_{\Omega} \mathbf{F} \cdot \nabla_{\mathcal{D}} v \, d\mathbf{x}. \end{aligned} \quad (1.50)$$

If \mathcal{D} has a piecewise constant reconstruction, then it makes sense, for a generic function $g : \mathbb{R} \rightarrow \mathbb{R}$ and $v \in X_{\mathcal{D},0}$, to define $g(v) \in X_{\mathcal{D},0}$ component-by-component: if $v = \sum_{i \in I} v_i \mathbf{e}_i$, then $g(v) = \sum_{i \in I} g(v_i) \mathbf{e}_i$. This definition is justified by the following commutation property, coming from (1.49):

$$\Pi_{\mathcal{D}}g(v) = g(\Pi_{\mathcal{D}}v) \quad \forall v \in X_{\mathcal{D},0}.$$

Then, a GS for the non-linear model (1.13) is obtained writing

$$\begin{aligned} &\text{Find } u \in X_{\mathcal{D},0} \text{ such that, } \forall v \in X_{\mathcal{D},0}, \\ &\int_{\Omega} (\Pi_{\mathcal{D}}u \Pi_{\mathcal{D}}v + \Lambda \nabla_{\mathcal{D}} \zeta(u) \cdot \nabla_{\mathcal{D}} v) \, d\mathbf{x} = \int_{\Omega} f \Pi_{\mathcal{D}} v \, d\mathbf{x} - \int_{\Omega} \mathbf{F} \cdot \nabla_{\mathcal{D}} v \, d\mathbf{x}. \end{aligned} \quad (1.51)$$

The accuracy and convergence of a GS is assessed through the following quantities and notions.

1. *Coercivity.* The discrete Poincaré constant of a GD \mathcal{D} is

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

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ is *coercive* if $(C_{\mathcal{D}_m})_{m \in \mathbb{N}}$ is bounded.

2. *Consistency.* The interpolation error of a GD \mathcal{D} is

$$S_{\mathcal{D}}(\phi) := \min_{v \in X_{\mathcal{D},0}} (\|\Pi_{\mathcal{D}}v - \phi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}}v - \nabla\phi\|_{L^2(\Omega)^d}) \quad \forall \phi \in H_0^1(\Omega).$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ is *consistent* if $S_{\mathcal{D}_m}(\phi) \rightarrow 0$ as $m \rightarrow \infty$, for all $\phi \in H_0^1(\Omega)$.

3. *Limit-conformity.* The defect of conformity of a GD \mathcal{D} is

$$W_{\mathcal{D}}(\psi) := \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{1}{\|\nabla_{\mathcal{D}}v\|_{L^2(\Omega)^d}} \left| \int_{\Omega} \Pi_{\mathcal{D}}v \operatorname{div} \psi + \nabla_{\mathcal{D}}v \cdot \psi x \right| \quad \forall \psi \in H_{\operatorname{div}}(\Omega).$$

A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ is *limit-conforming* if $W_{\mathcal{D}_m}(\psi) \rightarrow 0$ as $m \rightarrow \infty$, for all $\psi \in H_{\operatorname{div}}(\Omega)$.

4. *Compactness.* A sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ is *compact* if, for any $(v_m)_{m \in \mathbb{N}}$ such that $v_m \in X_{\mathcal{D}_m,0}$ for all $m \in \mathbb{N}$ and $(\|\nabla_{\mathcal{D}_m}v_m\|_{L^2(\Omega)^d})_{m \in \mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathcal{D}_m}v_m)_{m \in \mathbb{N}}$ is relatively compact in $L^2(\Omega)$.

We then recall an error estimate for the linear model and a convergence result for the non-linear model.

Theorem 1.6 (Error estimate for the linear model [9, Theorem 2.28]) *Let \bar{u} be the solution to (1.3), \mathcal{D} be a GD, and u be the solution to the gradient scheme (1.50). Then, there exists C depending only on Ω and $\underline{\lambda}, \bar{\lambda}$ in (1.2b) such that*

$$\|\bar{u} - \Pi_{\mathcal{D}}u\|_{L^2(\Omega)} + \|\nabla\bar{u} - \nabla_{\mathcal{D}}u\|_{L^2(\Omega)^d} \leq C(1 + C_{\mathcal{D}})(W_{\mathcal{D}}(\Lambda\nabla\bar{u} + \mathbf{F}) + S_{\mathcal{D}}(\bar{u})).$$

Theorem 1.7 (Convergence for the nonlinear model [8, Theorem 2.9]) *Let $(\mathcal{D}_m)_{m \in \mathbb{N}}$ be a sequence of GDs which is consistent, limit-conforming and compact (which implies its coercivity [9, Lemma 2.10]), and such that each \mathcal{D}_m has a piecewise constant reconstruction. Then, for any $m \in \mathbb{N}$ there exists a solution to (1.51) with $\mathcal{D} = \mathcal{D}_m$ and there exists a solution \bar{u} to (1.13) such that, as $m \rightarrow \infty$, the following convergences hold:*

$$\begin{aligned} \Pi_{\mathcal{D}_m}u_m &\rightarrow \bar{u} \text{ weakly in } L^2(\Omega), \\ \Pi_{\mathcal{D}_m}\zeta(u_m) &\rightarrow \zeta(\bar{u}) \text{ strongly in } L^2(\Omega), \\ \nabla_{\mathcal{D}_m}\zeta(u_m) &\rightarrow \nabla\zeta(\bar{u}) \text{ strongly in } L^2(\Omega)^d. \end{aligned}$$

The following lemma is particularly useful when considering mass-lumping of a given gradient discretisation. It shows that, under a simple assumption comparing the original and mass-lumped reconstructions, the properties of gradient discretisations that ensure the convergence of the gradient scheme are preserved.

Lemma 1.4 (Mass-lumping preserves the approximation properties [9, Theorem 7.50]) *Let $(\mathcal{D}_m)_{m \in \mathbb{N}}$ be a sequence of gradient discretisations that is coercive, consistent, limit-conforming and compact. For each $m \in \mathbb{N}$ let $\mathcal{D}_m^* = (X_{\mathcal{D}_m,0}, \Pi_{\mathcal{D}_m}^*, \nabla_{\mathcal{D}_m})$ be a gradient discretisation that differs from \mathcal{D}_m only through its function reconstruction. Assume the existence of a sequence $(\omega_m)_{m \in \mathbb{N}}$ of positive numbers such that*

$\omega_m \rightarrow 0$ as $m \rightarrow \infty$ and, for all $m \in \mathbb{N}$,

$$\|\Pi_{\mathcal{D}_m} v - \Pi_{\mathcal{D}_m}^* v\|_{L^2(\Omega)} \leq \omega_m \|\nabla_{\mathcal{D}_m} v\|_{L^2(\Omega)^d} \quad \forall v \in X_{\mathcal{D}_m,0}.$$

Then, the sequence $(\mathcal{D}_m^*)_{m \in \mathbb{N}}$ is also coercive, consistent, limit-conforming and compact.

1.5.2 Non-conforming Gradient Discretisations

We recall here that polytopal non-conforming methods, as defined in Sect. 1.2, are gradient discretisation methods for gradient discretisations that satisfy the properties required for the error estimates/convergence of the scheme.

Let $V_{\mathfrak{T},0}$ be a finite-dimensional subspace of $H_{\mathfrak{T},0}^1$, and define the gradient discretisation \mathcal{D} by:

$$X_{\mathcal{D},0} = V_{\mathfrak{T},0}, \quad \Pi_{\mathcal{D}} v = v \text{ and } \nabla_{\mathcal{D}} v = \nabla_{\mathcal{M}} v \quad \forall v \in X_{\mathcal{D},0}. \quad (1.52)$$

Then, the non-conforming scheme (1.9), for the linear model, based on $V_{\mathfrak{T},0}$ is the gradient scheme (1.50) based on \mathcal{D} . Likewise, if $\Pi_{\mathfrak{T}} : V_{\mathfrak{T},0} \rightarrow L^\infty(\Omega)$ is a piecewise-constant reconstruction of the form (1.14) and $\mathcal{D}^* = (V_{\mathfrak{T},0}, \Pi_{\mathfrak{T}}, \nabla_{\mathcal{M}})$, then the non-conforming scheme (1.15) for the Stefan/PME model is the gradient scheme (1.51) with \mathcal{D}^* instead of \mathcal{D} .

Proposition 1.1 (Estimates for non-conforming methods [9, Proposition 9.5]) *Let \mathfrak{T} be a polytopal mesh and assume that $\gamma_{\mathfrak{T}} \leq \gamma$. Let $V_{\mathfrak{T},0}$ be a finite-dimensional subspace of $H_{\mathfrak{T},0}^1$ and define the GD \mathcal{D} by (1.52). Then, there exists $C > 0$ depending only on Ω and γ such that*

$$C_{\mathcal{D}} \leq C \quad (1.53)$$

$$S_{\mathcal{D}}(\phi) \leq C \min_{v \in V_{\mathfrak{T},0}} \|v - \phi\|_{H_{\mathfrak{T},0}^1} \quad \forall \phi \in H_0^1(\Omega), \quad (1.54)$$

$$W_{\mathcal{D}}(\psi) \leq C h_{\mathcal{M}} \|\psi\|_{H^1(\Omega)^d} \quad \forall \psi \in H^1(\Omega)^d. \quad (1.55)$$

Theorem 1.8 (Properties of polytopal non-conforming methods [9, Theorem 9.6]) *Let $(\mathfrak{T}_m)_{m \in \mathbb{N}}$ be a sequence of polytopal meshes such that $h_{\mathcal{M}_m} \rightarrow 0$ as $m \rightarrow \infty$ and $(\gamma_{\mathfrak{T}_m})_{m \in \mathbb{N}}$ is bounded. For each $m \in \mathbb{N}$ let $V_{\mathfrak{T}_m,0}$ be a finite-dimensional subspace of $H_{\mathfrak{T}_m,0}^1$ and assume that*

$$\min_{v \in V_{\mathfrak{T}_m,0}} \|v - \phi\|_{H_{\mathfrak{T}_m,0}^1} \rightarrow 0 \text{ as } m \rightarrow \infty, \quad \forall \phi \in H_0^1(\Omega).$$

Then, the sequence $(\mathcal{D}_m)_{m \in \mathbb{N}}$ defined from $(V_{\mathfrak{T}_m,0})_{m \in \mathbb{N}}$ as in (1.52) is coercive, consistent, limit-conforming, and compact.

Remark 1.14 (Mass-lumped non-conforming method) Combining this theorem with Lemma 1.4 shows that mass-lumped versions of polytopal non-conforming methods, such as the one presented in Sect. 1.4.4, usually also inherits the coercivity, consistency, limit-conformity and compactness properties.

1.6 Perspectives

The LEPNC presented here is a low-order method. It is possible to extend this method into an arbitrary order approximation method. Let $k \geq 1$ be a sought approximation degree. For $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_K$ and $q \in \mathbb{P}^{k-1}(\sigma)$, by the Riesz representation theorem in $L^2(\sigma)$ for the Lebesgue measure weighted by $\phi_{K,\sigma}$ (which is strictly positive on σ), there exists a unique $q_K \in \mathbb{P}^{k-1}(\sigma)$ such that

$$\int_{\sigma} (\phi_{K,\sigma})|_{\sigma} q_K r = \int_{\sigma} q r, \quad \forall r \in \mathbb{P}^{k-1}(\sigma). \quad (1.56)$$

Set $\phi_{K,\sigma,q} = \phi_{K,\sigma} \hat{q}_K$, where $\hat{q}_K \in \mathbb{P}^{k-1}(K)$ is defined by $\hat{q}_K(\mathbf{x}) = q_K(\pi_{\sigma}(\mathbf{x}))$ with $\pi_{\sigma} : \mathbb{R}^d \rightarrow H_{\sigma}$ the orthogonal projection on the hyperspace H_{σ} spanned by σ . Then, the local k -degree LEPNC space is

$$V_K^{\text{LEPNC},k} := \text{span}(\mathbb{P}^k(K) \cup \{\phi_{K,\sigma,q} : \sigma \in \mathcal{F}_K, q \in \mathbb{P}^{k-1}(\sigma)\}).$$

For any set of moments of degree $\leq k-1$ on σ , there exists $q \in \mathbb{P}^{k-1}(\sigma)$ that has the same moments and thus, in virtue of (1.56), $\phi_{K,\sigma,q}$ also has these same moments on σ . Let $(\psi_{K,i})_{i=1,\dots,n_k}$ be a basis of $\mathbb{P}^k(K)$. For each $i = 1, \dots, n_k$ we can find a linear combination $\sum_{\sigma \in \mathcal{F}_K} \phi_{K,\sigma,q_i}$ that has the same moments of degree $\leq k-1$ as $\psi_{K,i}$ on each $\sigma \in \mathcal{F}_K$. The function $\psi_{K,i} - \sum_{\sigma \in \mathcal{F}_K} \phi_{K,\sigma,q_i}$ therefore has zero moments of degree $\leq k-1$ on each face and, extended by 0 outside K , satisfies the $(k-1)$ -degree patch test: its moments on each face coincide when viewed from each side of the faces.

When $\{K, L\} = \mathcal{M}_{\sigma}$, for a given $q \in \mathbb{P}^{k-1}(\sigma)$, by (1.56) the functions $\phi_{K,\sigma,q}$ and $\phi_{L,\sigma,q}$ have the same moments of degree $\leq k-1$ on σ . Hence, in a similar way as in (1.23), we can glue $\phi_{K,\sigma,q}$ and $\phi_{L,\sigma,q}$ to obtain a global function that satisfies the $(k-1)$ -degree patch test.

The family of these extended functions span a non-conforming space that has approximation properties of order k (that is, (1.35) holds with $\mathcal{O}(h_{\mathcal{M}}^{k+1})$ instead of $\mathcal{O}(h_{\mathcal{M}}^2)$ in the right-hand side). The only caveat is the following: letting $(q_j)_{j=1,\dots,\ell_k}$ be a basis of $\mathbb{P}^{k-1}(\sigma)$, the family $\{\psi_{K,i} : i = 1, \dots, n_k\} \cup \{\phi_{K,\sigma,q_j} : \sigma \in \mathcal{F}_K, j = 1, \dots, \ell_k\}$ spans the local space $V_K^{\text{LEPNC},k}$; however, it is not clear if, in general, this family is linearly independent. Hence, describing a space of the local space (and, in consequence, the global space) requires to actually solve local linear problems, extracting a basis from a generating family.

References

1. I. Aavatsmark, An introduction to multipoint flux approximations for quadrilateral grids. *Comput. Geosci.* **6**(3–4), 405–432 (2002). Locally conservative numerical methods for flow in porous media
2. G.I. Barenblatt, On some unsteady motions of a liquid and gas in a porous medium. *Akad. Nauk SSSR. Prikl. Mat. Meh.* **16**, 67–78 (1952)
3. Z. Chen, Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems. *East-West J. Numer. Math.* **4**(1), 1–33 (1996)
4. P.G. Ciarlet, The finite element method for elliptic problems, in *Studies in Mathematics and its Applications*, vol. 4, pp. xix+530. North-Holland Publishing Co., Amsterdam-New York-Oxford (1978)
5. M. Crouzeix, P.-A. Raviart, Conforming and nonconforming finite element methods for solving the stationary Stokes equations. I. *Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge* **7**(R–3), 33–75 (1973)
6. D.A. Di Pietro, J. Droniou, *The Hybrid High-Order Method for Polytopal Meshes: Design, Analysis, and Applications*, vol. 19 of *Modeling, Simulation and Applications* (Springer International Publishing, 2020)
7. D.A. Di Pietro, S. Lemaire, An extension of the Crouzeix-Raviart space to general meshes with application to quasi-incompressible linear elasticity and Stokes flow. *Math. Comp.* **84**(291), 1–31 (2015)
8. J. Droniou, R. Eymard, High-order mass-lumped schemes for nonlinear degenerate elliptic equations. *SIAM J. Numer. Anal.* **58**(1), 153–188 (2020)
9. J. Droniou, R. Eymard, T. Gallouët, C. Guichard, R. Herbin, *The Gradient Discretisation Method*, vol. 82 of *Mathematics & Applications* (Springer, Berlin, 2018)
10. J. Droniou, R. Eymard, T. Gallouët, R. Herbin, The gradient discretisation method for linear advection problems. *Comput. Methods Appl. Math.* **20**(3), 437–458 (2020)
11. J. Droniou, N. Nataraj, Improved L^2 estimate for gradient schemes and super-convergence of the TPFA finite volume scheme. *IMA J. Numer. Anal.* **38**(3), 1254–1293 (2018)
12. HArDCore2D—Hybrid Arbitrary Degree::Core 2D. <https://github.com/jdroniou/HArDCore2D-release>, Version 2.0.2
13. R. Herbin, F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids, in *Finite Volumes for Complex Applications V*, pp. 659–692 (ISTE, London, 2008)
14. K. Lipnikov, G. Manzini, M. Shashkov, Mimetic finite difference method. *J. Comput. Phys.* **257**(Part B), 1163–1227 (2014)
15. G. Strang, G. Fix, *An Analysis of the Finite Element Method*, 2nd edn. (Wellesley-Cambridge Press, Wellesley, MA, 2008)
16. F. Stummel, The generalized patch test. *SIAM J. Numer. Anal.* **16**(3), 449–471 (1979)
17. J. Vázquez, *The Porous Medium Equation: Mathematical Theory* (Oxford Mathematical Monographs, The Clarendon Press, Oxford University Press, 2007)
18. M. Vohralík, J. Maryška, O. Severýn, Mixed and nonconforming finite element methods on a system of polygons. *Appl. Numer. Math.* **57**(2), 176–193 (2007)
19. M.F. Wheeler, I. Yotov, A multipoint flux mixed finite element method. *SIAM J. Numer. Anal.* **44**(5), 2082–2106 (2006)
20. O.C. Zienkiewicz, R.L. Taylor, D.D. Fox, *The Finite Element Method for Solid and Structural Mechanics*, 7th edn. (Amsterdam, Elsevier/Butterworth Heinemann, 2014)

Chapter 2

Error Estimates for the Gradient Discretisation Method on Degenerate Parabolic Equations of Porous Medium Type



Clément Cancès, Jérôme Droniou, Cindy Guichard, Gianmarco Manzini, Manuela Bastidas Olivares, and Iuliu Sorin Pop

Abstract The gradient discretisation method (GDM) is a generic framework for the spatial discretisation of partial differential equations. The goal of this contribution is to establish an error estimate for a class of degenerate parabolic problems, obtained under very mild regularity assumptions on the exact solution. Our study covers well-known models like the porous medium equation and the fast diffusion equations, as well as the strongly degenerate Stefan problem. Several schemes are then compared in a last section devoted to numerical results.

Keywords Gradient discretisation method · Porous medium equation · Slow diffusion · Fast diffusion · Error estimates · Numerical tests · Hybrid mimetic mixed method · Virtual element method · Vertex approximate gradient method · Discontinuous Galerkin method · Polytopal methods

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

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

C. Guichard
Laboratoire Jacques-Louis Lions (LJLL), Sorbonne Université and Université de Paris, CNRS, Inria, F-75005 Paris, France
e-mail: cindy.guichard@sorbonne-universite.fr

G. Manzini
Istituto di Matematica Applicata e Tecnologie Informatiche - CNR, via Ferrata 1, Pavia, Italy
e-mail: marco.manzini@imati.cnr.it

M. B. Olivares · I. S. Pop
Hasselt University, Campus Diepenbeek, Agoralaan Gebouw D, 3590 Diepenbeek, Belgium
e-mail: manuela.bastidas@uhasselt.be

I. S. Pop
e-mail: sorin.pop@uhasselt.be

2.1 Introduction

Degenerate parabolic equations appear as mathematical models for numerous real-life applications, like reactive solute transport in porous media, water infiltration in the vadose zone, geological CO₂ sequestration, oil recovery, biological systems, or phase transition problems. In the simplest form, one has

$$\begin{aligned} \partial_t \bar{u} - \Delta \zeta(\bar{u}) &= f && \text{in } (0, T) \times \Omega, \\ \zeta(\bar{u}) &= 0 && \text{on } (0, T) \times \partial\Omega, \\ \bar{u}(0, \cdot) &= u_{\text{ini}} && \text{on } \Omega. \end{aligned} \quad (2.1)$$

With L^∞ denoting the space of essentially bounded functions and $\|\cdot\|_\infty$ the corresponding norm, throughout this chapter we assume the following.

- (A0) $T > 0$ and Ω is a bounded connected open set of \mathbb{R}^d ($d \in \mathbb{N}^*$) with Lipschitz continuous boundary $\partial\Omega$.
- (A1) $\zeta : \mathbb{R} \rightarrow \mathbb{R}$ is continuous, non-decreasing and satisfies $\zeta(0) = 0$.
- (A2) $u_{\text{ini}} \in L^\infty(\Omega)$ with $M_0 := \|u_{\text{ini}}\|_\infty$.
- (A3) $f \in L^\infty((0, T] \times \Omega)$ with $M_f := \|f\|_\infty$.

As follows from (A1), ζ' may become zero, or unbounded for certain arguments \bar{u} . Consequently, the equation may degenerate from a parabolic equation into an elliptic or an ordinary one. The degeneracy regions are not known a-priori, but depend on the solution itself and may change in time.

One of the most representative example in this sense, the porous medium equation (PME), appeared in the last century as a mathematical model for the flow of an ideal gas in a porous medium. In this case one has

$$\zeta(\bar{u}) = |\bar{u}|^{m-1} \bar{u} \text{ for some } m > 1. \quad (2.2)$$

Compared to the heat equation, which is obtained for $m = 1$ and in which the equation is linear and parabolic everywhere regardless of the data, if $m > 1$ the nonlinear diffusive term vanishes if $\bar{u} = 0$, and the equation degenerates. In particular, this leads to the occurrence of free boundaries separating regions in Ω where $\bar{u} > 0$ from those where $\bar{u} \leq 0$. These free boundaries have an a-priori unknown location and move in time with a finite speed, which is the reason for calling such cases as “slow diffusion” ones.

Another remarkable example in the category of “slow diffusion” equations is the Stefan problem, which models phase transition problems like melting or solidification. In this case

$$\zeta(\bar{u}) = \begin{cases} \bar{u}, & \text{if } \bar{u} < 0, \\ \max\{0, \bar{u} - 1\}, & \text{if } \bar{u} \geq 1. \end{cases} \quad (2.3)$$

Though bounded, ζ' is vanishing on the entire interval $(0, 1)$.

A different situation appears when ζ is as in (2.2), but with $m \in (0, 1)$. In this case no free boundaries occur, but $\zeta' \rightarrow \infty$ whenever $\bar{u} \rightarrow 0$ so the diffusion coefficient becomes unbounded. This equation is also known as generalized porous medium equation (GPME), and one speaks about a “fast diffusion”. It can appear as a mathematical model for reactive transport in porous media, for equilibrium kinetics (see [5]). We also refer to Chap. 3 for the description of a more complex degenerate parabolic model of two-phase flows. See also Chap. 7 for a numerical analysis of the Richard’s equation, which is strongly related to the Stefan model.

The degeneracy has direct impact on the regularity of the solutions. Unlike the regular parabolic case, the solutions to degenerate parabolic problems have lower regularity, and the singularities are not smoothed out but may even develop in time. Such effects are particularly encountered at the free boundaries. The lack of regularity motivates the introduction of a notion of weak solution.

We use standard notations and function spaces in the functional analysis: $L^2(\Omega)$, $L^\infty(\Omega)$, $H_0^1(\Omega)$, or its dual $H^{-1}(\Omega)$. Whenever obvious, the domain Ω is left out. With X being one of the spaces before, $L^2(0, T; X)$ is the space of X -valued measurable functions that are square integrable in the sense of Bochner. We let (\cdot, \cdot) stand for the inner product on $L^2(\Omega)$, or the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$, and $\|\cdot\|$ for the norm in $L^2(\Omega)$, or the straightforward extension to $L^2(\Omega)^d$, and $\|\cdot\|_\infty$ is the L^∞ norm in Ω or in $(0, T] \times \Omega$. We will often write u or $u(t)$ instead of $u(t, \mathbf{x})$ and use C to denote a generic positive constant independent of the discretisation parameters or the function itself.

We start by defining a weak solution for (2.1):

Definition 2.1 A weak solution to (2.1) is a measurable function $\bar{u} : (0, T) \times \Omega \rightarrow \mathbb{R}$ such that $\bar{u} \in H^1(0, T; H^{-1}(\Omega))$, $\zeta(\bar{u}) \in L^2(0, T; H_0^1(\Omega))$, $\bar{u}(0) = u_{\text{ini}}$ in $H^{-1}(\Omega)$ and, for a.e. $t \in (0, T]$ and for all $v \in H_0^1(\Omega)$, it holds

$$(\partial_t \bar{u}(t), v) + (\nabla \zeta(\bar{u}(t)), \nabla v) = (f(t), v). \quad (2.4)$$

The existence and uniqueness of a weak solution to (2.1) is proved e.g. in [1] and [42] in the case where ζ is increasing. If ζ is merely nondecreasing, existence and uniqueness still hold, see e.g. [14], as well as [44]. As already suggested, the degenerate aspect of the problem makes the usual regularity theory for parabolic problems (see for instance [35]) fail. What is kept is mainly the following:

- *Maximum principle*: the solution \bar{u} belongs to $L^\infty((0, T] \times \Omega)$, with

$$\|\bar{u}\|_\infty \leq M_0 + T M_f. \quad (2.5)$$

- *Energy estimate*: Consider the primitive of ζ defined by $\Xi : \mathbb{R} \rightarrow \mathbb{R}$, $\Xi(v) = \int_0^v \zeta(z) dz$. Ξ is convex and positive and one has

$$\int_\Omega \Xi(\bar{u}(t)) + \frac{1}{2} \int_0^t \int_\Omega |\nabla \zeta(s)|^2 \leq \int_\Omega \Xi(u_{\text{ini}}) + \frac{1}{2} \|f\|_{L^2(0, T; H^{-1}(\Omega))}^2. \quad (2.6)$$

- *Continuity of $\zeta(\bar{u})$* : it is shown in [54] under quite general assumptions on ζ (including cases where ζ is constant on an interval) that $\zeta(\bar{u})$ belongs to $C((0, T] \times \Omega)$. In the case where ζ is increasing (thus invertible), one gets that $\bar{u} \in C((0, T] \times \Omega)$ too. Because of the degeneracy of the problem, this estimate is not enough to initiate a bootstrap to recover the usual parabolic regularity theory.
- *Time continuity of \bar{u}* : even if ζ is not invertible, one can still give a (weaker) sense to $\bar{u}(t)$ as a function (and not only as a distribution in H^{-1} as suggested by Definition 2.1). Indeed, $\bar{u} \in C([0, T]; L^p(\Omega))$ for all $p \in [1, +\infty)$ thanks to [10].

Further regularity results in the PME case where $\zeta(u) = |u|^{m-1}u$ (or more generally when ζ is increasing) can be found in the monographs [49, 50] (see also [36] for the local Hölder continuity), while the Stefan problem is extensively discussed in [39].

The literature on the numerical approximation of degenerate parabolic equations is extremely rich. Often, the numerical scheme are including a regularisation step, that is used to deal with the lack of regularity of the solution to degenerate problems. Whenever regularisation is involved, this is mostly obtained through a perturbation ζ_ε of ζ , of which derivative is bounded away from 0 from below and from infinity from above (see e.g. [41]). Alternatively, one can exploit the maximum principle and perturb the boundary and initial data in such a way that the solution stays away from values at which degeneracy is encountered.

Concerning various specific numerical schemes, we mention that often the time stepping is of first order. In particular Euler implicit or semi-implicit methods are popular, and this is due to the lack in regularity of the solution. For the spatial discretisation we mention the conformal finite element schemes analysed e.g. in [41] for the slow diffusion, or in [5] for the fast diffusion. The convergence of the mixed finite element discretisation is proved in [4, 52] for the slow diffusion case, and for a range allowing for both kind of degeneracies in [48]. We also mention [53] for the analysis of a scheme combining mortars with mixed finite elements. These papers are proving the convergence of the scheme by obtaining a-priori error estimates rigorously. The convergence of finite volume schemes is proved in [2, 3, 28, 30] by means of compactness arguments, and in [29] for a finite volume phase-by-phase upstream weighting. Error estimates are obtained in [34] for a multipoint flux approximation scheme by using the equivalence with a mixed finite element scheme, and in [45] for the simplest two-point approximation in the slow diffusion case, but under minimal regularity assumptions. Discontinuous Galerkin schemes for porous media flow models leading to degenerate parabolic equations are analysed e.g. in [25, 26]. To conclude this paragraph, we mention that a-posteriori error estimates for degenerate problems related to porous media flows are derived in [13, 51].

The goal of this chapter is to study in a general way a large class of numerical approximation of (2.1), entering the framework of the so-called *Gradient Discretisation Method (GDM)* [22]. This general framework is detailed in Sect. 2.3. The ideas used in the numerical analysis below apply to methods which are energetically stable (i.e., a discrete counterparts of (2.6) holds). Our approach does not require any monotonicity properties of the approximation, like the maximum principle. This choice is due to the fact that proving the maximum principle (2.5), as well as the

(time-)continuity for $\zeta(\bar{u})$ and \bar{u} , relies strongly on the monotonicity of Eq. (2.1), which also extends to the time-discrete case but not to the fully discrete case provided by the GDM. We mention that the convergence of the GDM for a class of problems covering (2.1) when ζ is Lipschitz-continuous is obtained in [20, 27] by means of compactness arguments; this convergence was extended in [24] to the cases of slow and fast diffusion porous medium equations (for which ζ is not Lipschitz continuous). The aim here is to extend such results by providing a-priori error estimates. Note that the GDM for the stationary version of (2.1) is analysed in Chap. 1.

Remark 2.1 The present results can be adapted without any particular further difficulty to the case of problems with anisotropy of the form

$$\partial_t \bar{u} - \nabla \cdot (\Lambda \nabla \zeta(\bar{u})) = f \quad \text{in } (0, T) \times \Omega,$$

where $\Lambda \in L^\infty(\Omega; \mathbb{R}^{d \times d})$ is a symmetric definite positive tensor field, i.e., $\Lambda(\mathbf{x}) = \Lambda(\mathbf{x})^T$ and there exists $\lambda_m, \lambda_M > 0$ such that

$$\lambda_m |v|^2 \leq \Lambda(\mathbf{x})v \cdot v \leq \lambda_M |v|^2, \quad \mathbf{x} \in \Omega, v \in \mathbb{R}^d.$$

The paper is organised as follows. Section 2.2 is introducing the sequence of time discrete in time problems. The time discretisation relies on the backward Euler scheme and is thus very standard. The a-priori error estimates for the time discretisation are deeply inspired from [38], and do not require any regularity assumption on the exact solution. Section 2.3 is devoted to the fully discrete setting. This encompasses the definition of the notions of *Gradient Discretisation* and *Gradient Scheme*, which were introduced in [23] and further developed in the monograph [22]. The main result is an error estimate for any scheme entering this general framework of the GDM. To this purpose, reasonable extra regularity slightly overpassing the aforementioned regularity results rigorously established in the literature will be assumed on the solution \bar{u} . Finally, several numerical schemes are compared in Sect. 2.4; these schemes consist of the Locally Enriched Non-Conforming Polytopal scheme, the Hybrid Mimetic Mixed method, two versions of the Vertex Approximate Gradient scheme, the mass-lumped \mathbb{P}^1 Finite Element scheme, the Hybridizable Discontinuous Galerkin scheme, and the Conforming Virtual Element Method.

2.2 Time Discrete Problem

Our purpose in this section is to show how to derive an error estimate using only minimal regularity assumptions for time-discrete approximations of (2.1). To this end, we first establish some a-priori estimates on the time-discrete solution. This section shall be seen as a first step towards the derivation of the fully discrete error estimate of Theorem 2.3.

2.2.1 The Time Discretisation

In view of the low regularity of the solution, we only consider first order time discretisation schemes. To this aim we consider a sequence of times $0 = t^{(0)} < t^{(1)} < \dots < t^{(N)} = T$ ($N \in \mathbb{N}^*$) and define the time steps $\delta t^{(n+\frac{1}{2})} = t^{(n+1)} - t^{(n)}$ ($n \in \{0, \dots, N-1\}$). We let $\bar{u}^{(n)}$ be a time discrete approximation of $\bar{u}(t^{(n)})$. To define a weak solution to the time-discrete problems we use the set $X_0 := \{u \in L^2(\Omega) : \zeta(u) \in H_0^1(\Omega)\}$. The Euler implicit discretisation of (2.4) consists in finding a sequence of solutions to the time discrete problems, as defined in

Definition 2.2 (Time discrete problem) Set $\bar{u}^{(0)} = u_{\text{ini}}$. With $n \in \{0, \dots, N-1\}$, given $\bar{u}^{(n)} \in X_0$, a weak solution $\bar{u}^{(n+1)} \in X_0$ to the time discrete problem at time step $t^{(n+1)}$ satisfies, for all $v \in H_0^1(\Omega)$,

$$(\bar{u}^{(n+1)}, v) + \delta t^{(n+\frac{1}{2})} (\nabla \zeta(\bar{u}^{(n+1)}), \nabla v) = (\bar{u}^{(n)}, v) + \delta t^{(n+\frac{1}{2})} (f^{(n+1)}, v), \quad (2.7)$$

where $f^{(n+1)}(\mathbf{x}) = \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} f(s, \mathbf{x}) ds$.

Theorem 2.1 (Existence and uniqueness of a solution to the time discrete problem) There exists a unique family $(\bar{u}^{(n)})_{n=0, \dots, N}$ solution to the time discrete problem in the sense of Definition 2.2.

Proof Follows by applying [21, Theorem A.1] to solve, at each step, the non-linear elliptic problem $w - \delta t^{(n+\frac{1}{2})} \Delta \zeta(w) = \bar{u}^{(n)} + \delta t^{(n+\frac{1}{2})} f^{(n+1)}$. \square

2.2.2 A-Priori Estimates

Our goal is to provide a fully discrete error analysis for numerical schemes for (2.1). For ease of legibility, we start by discussing some properties of the time discrete, Euler implicit discretisation in (2.7). In doing so, we follow the ideas in [38], where the convergence of a linear, time discrete scheme is proved for a class of problems that includes (2.1).

We start with a remark on the essential boundedness of a solution. This property is physically justified for many of the applications that can be modelled mathematically in the form of (2.1) (e.g. the gas flow in porous media flows, or the reactive transport). In this context, the essential boundedness is inherited by the time discrete solutions, which satisfy a maximum principle. However, since this property does not extend to the fully discrete cases excepting some particular finite element or finite volume discretisation, we will avoid using it below.

Assuming that the initial data and the source term are both essentially bounded, as stated in Assumptions (A2) and (A3), one has

Lemma 2.1 Assume $\bar{u}^{(n)} \in X_0$ is such that $\|\bar{u}^{(n)}\|_\infty \leq M_0 + M_f t^{(n)}$. Then the solution $\bar{u}^{(n+1)}$ of (2.7) satisfies $\|\bar{u}^{(n+1)}\|_\infty \leq M_0 + M_f t^{(n+1)}$.

These estimates are obtained straightforwardly by testing in (2.7) with $v = [\zeta(\bar{u}^{(n+1)}) - \zeta(M_0 + M_{ft}^{(n+1)})]_+$, and with $v = [\zeta(\bar{u}^{(n+1)}) + \zeta(M_0 + M_{ft}^{(n+1)})]_-$ (with $[s]_+ = \max(s, 0)$ and $[s]_- = \min(s, 0)$). We omit the details.

We state some elementary results that are used below, and which are valid for all set of vectors $\mathbf{a}_n, \mathbf{b}_n \in \mathbb{R}^d$ ($d \geq 1$), $n \in \{0, \dots, m\}$.

$$2 \sum_{n=1}^m \mathbf{a}_n \cdot (\mathbf{a}_n - \mathbf{a}_{n-1}) = |\mathbf{a}_m|^2 - |\mathbf{a}_0|^2 + \sum_{n=1}^m |\mathbf{a}_n - \mathbf{a}_{n-1}|^2, \quad (2.8)$$

$$2 \sum_{n=0}^m \sum_{j=0}^n \mathbf{a}_n \cdot \mathbf{a}_j = \left| \sum_{n=0}^m \mathbf{a}_n \right|^2 + \sum_{n=0}^m |\mathbf{a}_n|^2, \quad (2.9)$$

$$\sum_{n=1}^m \mathbf{a}_n \cdot (\mathbf{b}_n - \mathbf{b}_{n-1}) = \mathbf{a}_m \cdot \mathbf{b}_m - \mathbf{a}_0 \cdot \mathbf{b}_0 - \sum_{n=1}^m (\mathbf{a}_n - \mathbf{a}_{n-1}) \cdot \mathbf{b}_{n-1}. \quad (2.10)$$

Further, with the convex, positive primitive Ξ of ζ appearing in (2.6), a classical convexity relation yields

$$(b - a)\zeta(b) \geq \Xi(b) - \Xi(a), \quad \forall a, b \in \mathbb{R}. \quad (2.11)$$

Finally, we state for completeness the Young inequality, valid for any $a, b \in \mathbb{R}$ and $\varepsilon > 0$,

$$ab \leq \frac{1}{2\varepsilon} a^2 + \frac{\varepsilon}{2} b^2. \quad (2.12)$$

The stability of the time discrete scheme is stated in

Lemma 2.2 *Let $(\bar{u}^{(n+1)})_{n=0, \dots, N-1}$ be the sequence of time discrete solutions introduced in Definition 2.2. Then,*

$$\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\nabla \zeta(\bar{u}^{(n+1)})\|^2 \leq C. \quad (2.13)$$

Proof Taking in (2.7) $v = \zeta(\bar{u}^{(n+1)})$ gives

$$\begin{aligned} & (\bar{u}^{(n+1)} - \bar{u}^{(n)}, \zeta(\bar{u}^{(n+1)})) + \delta t^{(n+\frac{1}{2})} \|\nabla \zeta(\bar{u}^{(n+1)})\|^2 \\ &= \delta t^{(n+\frac{1}{2})} (f^{(n+1)}, \zeta(\bar{u}^{(n+1)})). \end{aligned} \quad (2.14)$$

For the first term one uses (2.11) to obtain

$$(\bar{u}^{(n+1)} - \bar{u}^{(n)}, \zeta(\bar{u}^{(n+1)})) \geq \int_{\Omega} \Xi(\bar{u}^{(n+1)}) - \Xi(\bar{u}^{(n)}) \, dx. \quad (2.15)$$

The second term needs no further discussion, whereas for the term on the right one obtains

$$\begin{aligned} & \delta t^{(n+\frac{1}{2})} |(f^{(n+1)}, \zeta(\bar{u}^{(n+1)}))| \\ & \leq \frac{\delta t^{(n+\frac{1}{2})}}{2} \|f^{(n+1)}\|_{H^{-1}(\Omega)}^2 + \frac{\delta t^{(n+\frac{1}{2})}}{2} \|\nabla \zeta(\bar{u}^{(n+1)})\|^2. \end{aligned} \quad (2.16)$$

Using (2.15) and (2.16) into (2.14) and summing the resulting relation over $n \in \{1, \dots, N-1\}$ yields

$$\begin{aligned} & \int_{\Omega} \Xi(\bar{u}^{(n+1)}) + \frac{1}{2} \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\nabla \zeta(\bar{u}^{(n+1)})\|^2 \\ & \leq \int_{\Omega} \Xi(\bar{u}^{(0)}) + \sum_{n=0}^{N-1} \frac{\delta t^{(n+\frac{1}{2})}}{2} \|f^{(n+1)}\|_{H^{-1}(\Omega)}^2. \end{aligned} \quad (2.17)$$

The first term on the right is bounded due to Assumption (A2). For the second term on the right one uses Assumption (A3) and the fact that $L^\infty((0, T) \times \Omega)$ is continuously embedded into $L^2(0, T; H^{-1}(\Omega))$, to obtain a uniform bound w.r.t. N . Since Ξ is a positive function, the first term on the left is positive, which provides (2.13). \square

Remark 2.2 (*Other estimates*) Other a-priori estimates can be obtained if further assumptions are made on ζ and on the initial data. For example, if ζ is Lipschitz and $\zeta(u_{\text{ini}}) \in H_0^1(\Omega)$, then one can prove that $\|\nabla \zeta(\bar{u})\|$ and $\|\nabla \zeta(\bar{u}^{(n)})\|$ are uniformly bounded (w.r.t. t , respectively n) and obtain L^2 estimates for $\partial_t \zeta(\bar{u})$ or its time discrete counterpart.

2.2.3 Error Estimates, Time Discrete Case

We now establish error estimates for the time discrete scheme. The proof follows the lines of [38, Sect. 4]. We use the following notations for the errors

$$\begin{aligned} e_u(t) & := \bar{u}(t) - \bar{u}^{(n+1)}, \\ e_\zeta(t) & := \zeta(\bar{u}(t)) - \zeta(\bar{u}^{(n+1)}), \end{aligned}$$

for $t \in (t^{(n)}, t^{(n+1)})$ and $n \in \{0, \dots, N-1\}$. With this, one has

Theorem 2.2 *Let \bar{u} be the solution in Definition 2.1, and $(\bar{u}^{(n+1)})_{n=0, \dots, N-1}$ be the sequence of solutions to the time discrete problems in Definition 2.2. Setting $\delta t = \max_{n \in \{0, \dots, N-1\}} \delta t^{(n+\frac{1}{2})}$, one has*

$$\max_{n \in \{0, \dots, N-1\}} \|e_u(t^{(n+1)})\|_{H^{-1}(\Omega)}^2 + \int_0^T (e_u(t), e_\zeta(t)) dt \leq C \delta t. \quad (2.18)$$

Remark 2.3 Although the second term in (2.18) is not a proper norm, it can generate an error estimate in a certain norm whenever the particular form of ζ is taken into account. This idea is exploited, in the fully discrete setting, in Corollary 2.1 below. Moreover, as for the a-priori estimates, under additional assumptions on ζ one can get L^2 error estimates for either $\zeta(\bar{u})$ (if ζ is Lipschitz continuous, as appearing in the slow diffusion case) or \bar{u} (if ζ is bijective and its inverse Lipschitz, as appearing in the fast diffusion case).

Proof Before giving the proof we observe that, due to the monotonicity of ζ , the second term in (2.18) is positive.

With $j \in \{0, \dots, N-1\}$ we integrate (2.4) in time for $t \in (t^{(j)}, t^{(j+1)})$ and subtract (2.7) for $n = j$ to obtain

$$(e_u(t^{(j+1)}) - e_u(t^{(j)}), v) + \left(\nabla \int_{t^{(j)}}^{t^{(j+1)}} e_\zeta(t) dt, \nabla v \right) = 0,$$

for all $v \in H_0^1(\Omega)$. After summation over $j \in \{0, \dots, p\}$ for some $p \in \{0, \dots, N-1\}$, this yields

$$(e_u(t^{(p+1)}), v) + \left(\nabla \int_0^{t^{(p+1)}} e_\zeta(t) dt, \nabla v \right) = 0, \quad (2.19)$$

for all $v \in H_0^1(\Omega)$. Taking $v = \int_{t^{(p)}}^{t^{(p+1)}} e_\zeta(t) dt$ in the above equation provides

$$\begin{aligned} & \left(e_u(t^{(p+1)}), \int_{t^{(p)}}^{t^{(p+1)}} e_\zeta(t) dt \right) \\ & + \sum_{j=0}^p \left(\nabla \int_{t^{(j)}}^{t^{(j+1)}} e_\zeta(t) dt, \nabla \int_{t^{(p)}}^{t^{(p+1)}} e_\zeta(t) dt \right) = 0. \end{aligned} \quad (2.20)$$

Fixing $n \in \{0, \dots, N-1\}$, we sum (2.20) over $p \in \{0, \dots, n\}$ and obtain

$$\begin{aligned} & \overbrace{\sum_{p=0}^n \left(e_u(t^{(p+1)}), \int_{t^{(p)}}^{t^{(p+1)}} e_\zeta(t) dt \right)}^{=: I_1} \\ & + \underbrace{\sum_{p=0}^n \sum_{j=0}^p \left(\nabla \int_{t^{(j)}}^{t^{(j+1)}} e_\zeta(t) dt, \nabla \int_{t^{(p)}}^{t^{(p+1)}} e_\zeta(t) dt \right)}_{=: I_2} = 0. \end{aligned} \quad (2.21)$$

The first term can be rewritten as

$$\begin{aligned}
 I_1 &= \underbrace{\int_0^{t^{(n+1)}} (e_u(t), e_\zeta(t)) dt}_{=: I_{11}} \\
 &+ \underbrace{\sum_{p=0}^n \int_{t^{(p)}}^{t^{(p+1)}} (\bar{u}(t^{(p+1)}) - \bar{u}(t), e_\zeta(t)) dt}_{=: I_{12}}. \tag{2.22}
 \end{aligned}$$

Being positive, I_{11} needs no further handling. For I_{12} we write

$$\bar{u}(t^{(p+1)}) - \bar{u}(t) = \int_t^{t^{(p+1)}} \partial_s \bar{u} ds$$

to obtain

$$\begin{aligned}
 |I_{12}| &= \left| \sum_{p=0}^n \int_{t^{(p)}}^{t^{(p+1)}} \left(\int_t^{t^{(p+1)}} \partial_s \bar{u} ds, e_\zeta(t) \right) dt \right| \\
 &\leq \left(\max_{n \in \{0, \dots, N-1\}} \delta t^{(n+\frac{1}{2})} \right) \|\partial_t \bar{u}\|_{L^2(0, T; H^{-1}(\Omega))} \|\nabla e_\zeta\|_{L^2(0, T; L^2(\Omega))}.
 \end{aligned}$$

The regularity of the weak solution prescribed in Definition 2.1 and the a-priori estimate (2.13) ensure the existence of a $C > 0$ not depending on the time discretisation so that

$$\|\partial_t \bar{u}\|_{L^2(0, T; H^{-1}(\Omega))} \leq C, \quad \|\nabla e_\zeta\|_{L^2(0, T; L^2(\Omega))} \leq C.$$

As a consequence, we obtain that

$$|I_{12}| \leq C \delta t. \tag{2.23}$$

Finally, using (2.9), I_2 is nonnegative and can be underestimated by

$$I_2 \geq \frac{1}{2} \left\| \nabla \int_0^{t^{(n+1)}} e_\zeta(t) dt \right\|^2. \tag{2.24}$$

Since n was chosen arbitrarily, using (2.22)–(2.24) in (2.21) yields

$$\int_0^T (e_u(t), e_\zeta(t)) dt + \max_{n \in \{0, \dots, N-1\}} \left\| \nabla \int_0^{t^{(n+1)}} e_\zeta(t) dt \right\|^2 \leq C \delta t. \tag{2.25}$$

To complete the proof of Theorem 2.2 one needs to estimate $\|e_u\|_{H^{-1}(\Omega)}$. This follows straightforwardly from (2.19). For all $v \in H_0^1(\Omega)$ such that $\|v\|_{H_0^1(\Omega)} \leq 1$, the Cauchy–Schwarz inequality yields

$$(e_u(t^{(p+1)}), v) = - \left(\nabla \int_0^{t^{(p+1)}} e_\zeta(t) dt, \nabla v \right) \leq \left\| \nabla \int_0^{t^{(p+1)}} e_\zeta(t) dt \right\|.$$

Taking the supremum over such v , squaring and using (2.25) we infer

$$\|e_u(t^{(p+1)})\|_{H^{-1}(\Omega)}^2 \leq C \delta t$$

and the proof is complete. \square

2.3 Gradient Discretisation Method and Generic Error Estimate

2.3.1 Definition of the Gradient Scheme

The principle of the GDM is to replace, in the weak formulation of the problem, the continuous space and differential operators by discrete ones. To this aim a discrete space and function/gradient reconstructions on this space are used. Altogether these form a gradient discretisation (GD), denoted by \mathcal{D} . In general, very few assumptions are made on the GD [22]. However, to deal with the non-linearity in (2.1) we will need, in a similar way as in [21], to consider *nodal* gradient discretisations with *piecewise constant* reconstructions, and that also contain the definition of an interpolator. Therefore, we take $\mathcal{D} = (X_{\mathcal{D},0}, \Pi_{\mathcal{D}}, \nabla_{\mathcal{D}}, I_{\mathcal{D}})$ with

- (*Space*) $X_{\mathcal{D},0} = \{v = (v_i)_{i \in I} : v_i = 0 \text{ for all } i \in I_\partial\}$, where I is a finite set and $I_\partial \subset I$ identifies the boundary degrees of freedom.
- (*Function reconstruction*) There is a partition $(U_i)_{i \in I}$ of Ω such that, for all $v = (v_i)_{i \in I} \in X_{\mathcal{D},0}$, the reconstructed function $\Pi_{\mathcal{D}}v \in L^\infty(\Omega)$ is defined by

$$\Pi_{\mathcal{D}}v = \sum_{i \in I} v_i \mathbf{1}_{U_i}, \quad (2.26)$$

where $\mathbf{1}_{U_i}$ is the characteristic function of U_i .

- (*Gradient reconstruction*) $\nabla_{\mathcal{D}} : X_{\mathcal{D},0} \rightarrow L^2(\Omega)^d$ is a linear operator such that $v \mapsto \|\nabla_{\mathcal{D}}v\|$ is a norm on $X_{\mathcal{D},0}$.
- (*Interpolator*) The components $(v_i)_{i \in I}$ of $v \in X_{\mathcal{D},0}$ represent values at points $(\mathbf{x}_i)_{i \in \bar{\Omega}}$, with $\mathbf{x}_i \in \bar{U}_i$ for all $i \in I$ and $\mathbf{x}_i \in \partial\Omega$ whenever $i \in I_\partial$. With $C_{\text{pw},0}(\Omega)$ the set of piecewise continuous functions on $\bar{\Omega}$ that have a zero limit on $\partial\Omega$, we define the interpolator $I_{\mathcal{D}} : C_{\text{pw},0}(\Omega) \rightarrow X_{\mathcal{D},0}$ such that

$$(I_{\mathcal{D}}\phi)_i = \operatorname{ess-limsup}_{x \rightarrow x_i, x \in \bar{U}_i} \phi(\mathbf{x}), \quad \text{for all } i \in I, \quad (2.27)$$

where

$$\operatorname{ess-limsup}_{x \rightarrow x_i, x \in \bar{U}_i} \phi(\mathbf{x}) = \lim_{\epsilon \rightarrow 0} \operatorname{ess-sup}_{B(x_i, \epsilon) \cap \bar{U}_i} \phi.$$

The fact that the function reconstruction is piecewise constant enables us to define, for $g : \mathbb{R} \rightarrow \mathbb{R}$ with $g(0) = 0$ and $v \in X_{\mathcal{D},0}$, the element $g(v) \in X_{\mathcal{D},0}$ by applying g to each nodal value: if $v = (v_i)_{i \in I}$, we set $g(v) = (g(v_i))_{i \in I}$. It then holds

$$\Pi_{\mathcal{D}}g(v) = g(\Pi_{\mathcal{D}}v), \quad \forall v \in X_{\mathcal{D},0}. \quad (2.28)$$

The subtle choice for the definition of the interpolator is motivated by the following points. Since our study covers the Stefan problem, whose solution might be discontinuous, there is a real need to define an interpolator that allows for merely piecewise continuous functions. If ϕ is continuous, say $\phi \in C_0(\Omega)$, then

$$\operatorname{ess-limsup}_{x \rightarrow x_i, x \in \bar{U}_i} \phi(\mathbf{x}) = \phi(x_i), \quad \text{for all } i \in I.$$

Therefore, for any continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$ and $\phi \in C_0(\Omega)$,

$$I_{\mathcal{D}}g(\phi) = g(I_{\mathcal{D}}\phi). \quad (2.29)$$

When ϕ is only piecewise continuous, then the definition (2.27) of the interpolator ensures that (2.29) still holds as soon as g is continuous and nondecreasing.

We can now define the gradient scheme for (2.1) with implicit time stepping. It is obtained from (2.7) using the discrete space for trial and test functions, and replacing the functions and gradients by the corresponding reconstructions. This gives a sequence of fully discrete, nonlinear algebraic problems, obtained for $n \in \{0, \dots, N-1\}$ and starting with $u^{(0)} = I_{\mathcal{D}}u_{\text{ini}}$.

Problem $\mathbf{P}_{\mathcal{D}}^{(n+1)}$: *Given $u^{(n)} \in X_{\mathcal{D},0}$, find $u^{(n+1)} \in X_{\mathcal{D},0}$ such that*

$$\begin{aligned} & \int_{\Omega} \Pi_{\mathcal{D}}(u^{(n+1)} - u^{(n)}) \Pi_{\mathcal{D}}v \\ & + \delta t^{(n+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}}\zeta(u^{(n+1)}) \cdot \nabla_{\mathcal{D}}v = \delta t^{(n+\frac{1}{2})} \int_{\Omega} f^{(n+1)} \Pi_{\mathcal{D}}v \end{aligned} \quad (2.30)$$

for all $v \in X_{\mathcal{D},0}$, where $f^{(n+1)}$ is introduced in Definition 2.2.

Proposition 2.1 (Existence and uniqueness of a solution to the gradient scheme [21, Lemma 2.7]) There exists a solution $u = (u^{(n)})_{n=0, \dots, N}$ to the gradient scheme and, if $u_1, u_2 \in X_{\mathcal{D},0}^{N+1}$ are two solutions to this scheme, then $\zeta(u_1) = \zeta(u_2)$ and $\Pi_{\mathcal{D}}u_1 = \Pi_{\mathcal{D}}u_2$.

Remark 2.4 (Limit to the uniqueness) If ζ is strictly increasing, then we have complete uniqueness of the solution: $u_1 = u_2$. However, when ζ has a plateau this uniqueness may fail [21, Remark 2.8].

The accuracy of a GD is measured through three quantities: a discrete Poincaré constant $C_{\mathcal{D}}$ (yielding the coercivity of the method), a measure of the defect of the discrete Stokes formula $W_{\mathcal{D}}$ (associated with the limit-conformity of the method), and a measure of the interpolation error $S_{\mathcal{D}}$ (which, when it tends to zero, yields the consistency of the method). The discrete Poincaré constant is

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{\|\Pi_{\mathcal{D}}v\|}{\|\nabla_{\mathcal{D}}v\|}. \quad (2.31)$$

The measure of the defect of the discrete Stokes formula is $W_{\mathcal{D}} : H_{\text{div}}(\Omega) \rightarrow [0, \infty)$ where, for all $\boldsymbol{\psi} \in H_{\text{div}}(\Omega)$ (that is, $\boldsymbol{\psi} \in L^2(\Omega)^d$ and $\text{div}\boldsymbol{\psi} \in L^2(\Omega)$),

$$W_{\mathcal{D}}(\boldsymbol{\psi}) := \max_{v \in X_{\mathcal{D},0} \setminus \{0\}} \frac{1}{\|\nabla_{\mathcal{D}}v\|} \left| \int_{\Omega} \Pi_{\mathcal{D}}v \text{div}\boldsymbol{\psi} + \nabla_{\mathcal{D}}v \cdot \boldsymbol{\psi} \right|. \quad (2.32)$$

In the GDM, the interpolation error usually involves L^2 -errors in both function and gradient approximation. However, for time-dependent problems such as (2.1), it will be more efficient to use a weaker norm for the function approximation. We define the discrete H^{-1} -seminorm by: for $\phi \in L^2(\Omega)$,

$$|\phi|_{\mathcal{D},*} := \max \left\{ \int_{\Omega} \phi \Pi_{\mathcal{D}}v : v \in X_{\mathcal{D},0}, \|\nabla_{\mathcal{D}}v\| \leq 1 \right\}. \quad (2.33)$$

We then set, for $\phi \in C_{\text{pw},0}(\Omega)$ and $\boldsymbol{\psi} \in C_0(\Omega) \cap H_0^1(\Omega)$,

$$S_{\mathcal{D}}^{\Pi,*}(\phi) = |\Pi_{\mathcal{D}}I_{\mathcal{D}}\phi - \phi|_{\mathcal{D},*} \quad \text{and} \quad S_{\mathcal{D}}^{\nabla}(\boldsymbol{\psi}) = \|\nabla_{\mathcal{D}}I_{\mathcal{D}}\boldsymbol{\psi} - \nabla\boldsymbol{\psi}\|. \quad (2.34)$$

2.3.2 A-Priori Estimates

We start with the observation that some properties of the solution to the original problem (2.1), or its time discrete counterpart, are not preserved by the gradient scheme (2.30). In particular we refer to the maximum principle (see Lemma 2.1),

which in the spatially-continuous case is obtained by testing with a cut-off function. However, in the spatially-discrete case the cut-off of an element in the finite dimensional space may not belong to that space any more, since the transition from negative to positive values does not necessarily happen at edges or nodes. In the generic GDM framework, to obtain a-priori estimates we are therefore restricted to using in (2.30) test functions that are affine functions of $\zeta(u^{(n+1)})$; we however mention that schemes allowing to take nonlinear functions are designed and analysed in [9, 11, 12].

The following lemma extends the estimates in Lemma 2.2 to the fully discrete case.

Lemma 2.3 *For the sequence of fully discrete solutions of (2.30) it holds*

$$\begin{aligned} & \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\nabla_{\mathcal{D}} \zeta(u^{(n+1)})\|^2 \\ & \leq C_{\mathcal{D}}^2 \|f\|_{L^2(0,T;L^2(\Omega))}^2 + 2\|\Xi(\Pi_{\mathcal{D}}u^{(0)})\|_{L^1(\Omega)}. \end{aligned} \quad (2.35)$$

Proof Choosing $v = \zeta(u^{(n+1)})$ in (2.30) leads, for all $n \in \{0, \dots, N-1\}$, to

$$\begin{aligned} & \int_{\Omega} \Pi_{\mathcal{D}}(u^{(n+1)} - u^{(n)}) \Pi_{\mathcal{D}} \zeta(u^{(n+1)}) \\ & + \delta t^{(n+\frac{1}{2})} \int_{\Omega} |\nabla_{\mathcal{D}} \zeta(u^{(n+1)})|^2 = \delta t^{(n+\frac{1}{2})} \int_{\Omega} f^{(n+1)} \Pi_{\mathcal{D}} \zeta(u^{(n+1)}). \end{aligned} \quad (2.36)$$

Using (2.28) and the convexity inequality (2.11), we have

$$\begin{aligned} & \int_{\Omega} \Pi_{\mathcal{D}}(u^{(n+1)} - u^{(n)}) \Pi_{\mathcal{D}} \zeta(u^{(n+1)}) = \int_{\Omega} (\Pi_{\mathcal{D}}u^{(n+1)} - \Pi_{\mathcal{D}}u^{(n)}) \zeta(\Pi_{\mathcal{D}}u^{(n+1)}) \\ & \geq \int_{\Omega} (\Xi(\Pi_{\mathcal{D}}u^{(n+1)}) - \Xi(\Pi_{\mathcal{D}}u^{(n)})). \end{aligned} \quad (2.37)$$

The right-hand side of (2.36) can be estimated thanks to Young and discrete Poincaré inequalities as follows:

$$\begin{aligned} & \int_{\Omega} f^{(n+1)} \Pi_{\mathcal{D}} \zeta(u^{(n+1)}) \leq \frac{C_{\mathcal{D}}^2}{2} \|f^{(n+1)}\|^2 + \frac{1}{2C_{\mathcal{D}}^2} \|\Pi_{\mathcal{D}} \zeta(u^{(n+1)})\|^2 \\ & \leq \frac{C_{\mathcal{D}}^2}{2} \|f^{(n+1)}\|^2 + \frac{1}{2} \|\nabla_{\mathcal{D}} \zeta(u^{(n+1)})\|^2. \end{aligned} \quad (2.38)$$

Combining (2.37) and (2.38) in (2.36), summing over $n \in \{0, \dots, N-1\}$, and using $\Xi \geq 0$, the proof of (2.35) is complete. \square

2.3.3 Error Estimate

With $I_{\mathcal{D}}^n \bar{u} = I_{\mathcal{D}} \bar{u}(t^{(n)})$ for $n \in \{0, \dots, N\}$, we define the errors in $X_{\mathcal{D},0}$:

$$\begin{aligned} e_{\mathcal{D},u}^{(n)} &:= u_{\mathcal{D}}^{(n)} - I_{\mathcal{D}}^n \bar{u}, \\ e_{\mathcal{D},\zeta}^{(n)} &:= \zeta(u_{\mathcal{D}}^{(n)}) - I_{\mathcal{D}}^n \zeta(\bar{u}), \end{aligned}$$

as well as, for $n \in \{1, \dots, N\}$,

$$\varepsilon_{\mathcal{D},\zeta}^{(n)} := \sum_{p=1}^n \delta t^{(p-\frac{1}{2})} e_{\mathcal{D},\zeta}^{(p)}.$$

The error estimates for the fully discrete approximation is stated in the following theorem, whose proof is carried out in Sect. 2.3.4.

Theorem 2.3 (GDM error estimate for degenerate parabolic problem) *Assume that the solution \bar{u} of (2.1) satisfies $\bar{u}(t, \cdot) \in C_{pw,0}(\Omega)$ for all $t \in [0, T]$, $\zeta(\bar{u}) \in C([0, T]; C_0(\Omega) \cap H_0^1(\Omega))$, and $\nabla \zeta(\bar{u}) \in C([0, T]; H_{\text{div}}(\Omega))$. Then, there exists a universal constant K , depending neither on the data of the continuous problem nor on the discretisation parameters, such that*

$$\begin{aligned} \max_{1 \leq n \leq N} \left| \Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)} \right|_{\mathcal{D},*}^2 + \max_{1 \leq n \leq N} \left\| \nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)} \right\|^2 \\ + \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} (\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n+1)}, \Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(n+1)}) \leq K(1+T) E_{\mathcal{D}}(\bar{u})^2, \end{aligned} \quad (2.39)$$

where

$$E_{\mathcal{D}}(\bar{u})^2 = \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} E_{\mathcal{D}}^n(\bar{u})^2 \quad (2.40)$$

with, for $n \in \{0, \dots, N-1\}$,

$$\begin{aligned} E_{\mathcal{D}}^n(\bar{u}) &:= \left| \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \Delta \zeta(\bar{u}(s)) ds - \Delta \zeta(\bar{u}(t^{(n+1)})) \right|_{\mathcal{D},*} \\ &+ S_{\mathcal{D}}^{\Pi,*} \left(\frac{\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})}{\delta t^{(n+\frac{1}{2})}} \right) + S_{\mathcal{D}}^{\nabla}(\zeta(\bar{u}(t^{(n+1)}))) \\ &+ W_{\mathcal{D}}(\nabla \zeta(\bar{u}(t^{(n+1)}))). \end{aligned} \quad (2.41)$$

Due to the non-decreasing property of ζ and to (2.28) and (2.29), each term in the sum in the left-hand side of (2.39) is non-negative.

Remark 2.5 (*Expected rates of convergence*) Under regularity assumptions on \bar{u} , a rate of convergence in terms of time and mesh sizes can be obtained on $E_{\mathcal{D}}(\bar{u})$. Specifically, if $\zeta(\bar{u}) \in C([0, T]; H^2(\Omega))$, $\partial_t \bar{u} \in L^\infty(0, T; H_0^1(\Omega))$ and $\Delta \zeta(\bar{u}) \in W^{1,\infty}(0, T; L^2(\Omega))$, following the techniques in [22, Sect. 7.4] and using $|\phi|_{\mathcal{D},*} \leq C_{\mathcal{D}} \|\phi\|$ it can be proved, for all usual low-order gradient discretisations based on meshes of maximum size h (which include all schemes used in Sect. 2.4 except VAG-b), that $E_{\mathcal{D}}^n(\bar{u}) \leq C_{\bar{u}}((1 + C_{\mathcal{D}})\delta t^{(n+\frac{1}{2})} + h)$, where $C_{\bar{u}}$ only depends on \bar{u} . Hence, in this situation and setting $\delta t = \max_{n \in \{0, \dots, N-1\}} \delta t^{(n+\frac{1}{2})}$, we have

$$E_{\mathcal{D}}(\bar{u}) \leq T^{\frac{1}{2}} C_{\bar{u}}((1 + C_{\mathcal{D}})\delta t + h).$$

Remark 2.6 For the slow diffusion case and under the regularity stated in Definition 2.1, error estimates for a simple, two-point flux approximation scheme (which fits in the GDM framework) are obtained in [45]. The approach there consists in using a discrete Green function to estimate the error for the fully discrete approximation of the sequence of time discrete approximations (Definition 2.2). This approach involves a regularisation step, which we avoid here by using a different strategy.

For the nonlinearities appearing in the Stefan and porous medium equations, (2.39) leads to the following error estimates on more natural quantities.

Corollary 2.1 (*Estimate for the Stefan equation and the PME*) *Under the assumptions and notations in Theorem 2.3, the following holds.*

- (Stefan equation) *Assume that ζ is Lipschitz-continuous with Lipschitz constant L_ζ . Then,*

$$\begin{aligned} & \left(\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\zeta(\Pi_{\mathcal{D}} u^{(n+1)}) - \zeta(\Pi_{\mathcal{D}} I_{\mathcal{D}}^{n+1} \bar{u})\|^2 \right)^{\frac{1}{2}} \\ & \leq (K(1+T)L_\zeta)^{\frac{1}{2}} E_{\mathcal{D}}(\bar{u}). \end{aligned} \quad (2.42)$$

- (Slow diffusion PME) *Let $\zeta(s) = |s|^{m-1}s$ with $m \geq 1$. Then, there exists $C_m > 0$ depending only on m such that*

$$\begin{aligned} & \left(\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{D}} u^{(n+1)} - \Pi_{\mathcal{D}} I_{\mathcal{D}}^{n+1} \bar{u}\|_{L^{m+1}(\Omega)}^{m+1} \right)^{\frac{1}{m+1}} \\ & \leq C_m T^{\frac{1}{m+1}} E_{\mathcal{D}}(\bar{u})^{\frac{2}{m+1}}. \end{aligned} \quad (2.43)$$

- (Fast diffusion PME) *Let $\zeta(s) = |s|^{m-1}s$ with $m < 1$. Then, there exists $C_m > 0$ depending only on m such that*

$$\left(\sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \|\zeta(\Pi_{\mathcal{D}} u^{(n+1)}) - \zeta(\Pi_{\mathcal{D}} I_{\mathcal{D}}^{n+1} \bar{u})\|_{L^{\frac{m}{m+1}}(\Omega)} \right)^{\frac{m}{m+1}} \leq C_m T^{\frac{m}{m+1}} E_{\mathcal{D}}(\bar{u})^{\frac{2m}{m+1}}. \quad (2.44)$$

Proof (Corollary 2.1) *Stefan equation.* Since ζ is a Lipschitz-continuous and non-decreasing function, we have

$$|\zeta(a) - \zeta(b)|^2 \leq |\zeta(a) - \zeta(b)| L_{\zeta} |a - b| = L_{\zeta} (\zeta(a) - \zeta(b))(a - b), \quad \forall a, b \in \mathbb{R}.$$

Used in (2.39), this relation proves (2.42).

Slow diffusion PME. We first prove that, for some $k_m > 0$,

$$|a - b|^{m+1} \leq k_m (|a|^{m-1} a - |b|^{m-1} b)(a - b), \quad \forall a, b \in \mathbb{R}. \quad (2.45)$$

The case for $b = 0$ is trivial and reduces to $k_m \geq 1$. Consider $b \neq 0$ and set $s = a/b$. To establish (2.45), we have to prove that $|s - 1|^{m+1} \leq k_m (|s|^{m-1} s - 1)(s - 1)$, which reduces to $|s - 1|^m \leq c_m |s|^{m-1} s - 1$. The function $s \mapsto \frac{|s-1|^m}{|s|^{m-1}s-1}$ is continuous on \mathbb{R} (use a Taylor expansion about $s = 1$ to deal with the singularity) and has limit 1 at $\pm\infty$. It is therefore bounded, which proves the required estimate.

Using (2.45) in (2.39), the estimate (2.43) follows.

Fast diffusion PME. Let $a', b' \in \mathbb{R}$ and apply (2.45) with $\frac{1}{m} > 1$ instead of m and $a = \zeta(a') = |a'|^{m-1} a'$, $b = \zeta(b') = |b'|^{m-1} b'$. Noting that $|a|^{\frac{1}{m}-1} a = a'$ and $|b|^{\frac{1}{m}-1} b = b'$, we infer

$$|\zeta(a') - \zeta(b')|^{\frac{1}{m}+1} \leq k_{1/m} (a' - b') (\zeta(a') - \zeta(b')).$$

Used in (2.39) this establishes (2.44). □

2.3.4 Proof of Theorem 2.3

We follow the approach of [17] which consists in identifying an error equation on the discrete solution and the interpolate of the continuous solution, and estimating a consistency error.

To identify the error equation we introduce the interpolates of the exact solution and use (2.30) to obtain, for all $j \in \{0, \dots, N-1\}$,

$$\int_{\Omega} \Pi_{\mathcal{D}} \left(e_{\mathcal{D},u}^{(j+1)} - e_{\mathcal{D},u}^{(j)} \right) \Pi_{\mathcal{D}} v + \delta t^{(j+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(j+1)} \cdot \nabla_{\mathcal{D}} v = \mathfrak{E}_{\mathcal{D}}^j(v), \quad (2.46)$$

and all $v \in X_{\mathcal{D},0}$, where the consistency error is defined by

$$\begin{aligned} \mathfrak{E}_{\mathcal{D}}^j(v) := & \delta t^{(j+\frac{1}{2})} \int_{\Omega} f^{(j+1)} \Pi_{\mathcal{D}} v - \int_{\Omega} \left[\Pi_{\mathcal{D}} I_{\mathcal{D}}^{j+1} \bar{u} - \Pi_{\mathcal{D}} I_{\mathcal{D}}^j \bar{u} \right] \Pi_{\mathcal{D}} v \\ & - \delta t^{(j+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} \zeta(I_{\mathcal{D}}^{j+1} \bar{u}) \cdot \nabla_{\mathcal{D}} v. \end{aligned} \quad (2.47)$$

This is a linear form $\mathfrak{E}_{\mathcal{D}}^n : X_{\mathcal{D},0} \rightarrow \mathbb{R}$. Its boundedness is established in

Lemma 2.4 *For all $v \in X_{\mathcal{D},0}$ and $n \in \{0, \dots, N-1\}$, there holds*

$$|\mathfrak{E}_{\mathcal{D}}^n(v)| \leq \delta t^{(n+\frac{1}{2})} E_{\mathcal{D}}^n(\bar{u}) \|\nabla_{\mathcal{D}} v\|. \quad (2.48)$$

Proof Recalling (2.34), $\Pi_{\mathcal{D}} I_{\mathcal{D}}^{n+1} \bar{u} - \Pi_{\mathcal{D}} I_{\mathcal{D}}^n \bar{u}$ can be replaced by $\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})$ in (2.47) with a cost measured by $S_{\mathcal{D}}^{\Pi,*}$. Specifically, by the definition of the discrete H^{-1} -seminorm in (2.33) we have

$$\left| \int_{\Omega} \phi \Pi_{\mathcal{D}} v \right| \leq |\phi|_{\mathcal{D},*} \|\nabla_{\mathcal{D}} v\| \quad \forall \phi \in L^2(\Omega), \quad \forall v \in X_{\mathcal{D},0}, \quad (2.49)$$

and thus, since $I_{\mathcal{D}}^k \bar{u} = I_{\mathcal{D}}(\bar{u}(t^{(k)}))$ for $k = n, n+1$,

$$\begin{aligned} & \left| \int_{\Omega} [\Pi_{\mathcal{D}} I_{\mathcal{D}}^{n+1} \bar{u} - \Pi_{\mathcal{D}} I_{\mathcal{D}}^n \bar{u}] \Pi_{\mathcal{D}} v - \int_{\Omega} [\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})] \Pi_{\mathcal{D}} v \right| \\ &= \left| \int_{\Omega} [\Pi_{\mathcal{D}} I_{\mathcal{D}}(\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})) - (\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)}))] \Pi_{\mathcal{D}} v \right| \\ &\leq S_{\mathcal{D}}^{\Pi,*}(\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})) \|\nabla_{\mathcal{D}} v\|. \end{aligned}$$

Similarly, replacing $\nabla_{\mathcal{D}} \zeta(I_{\mathcal{D}}^{n+1} \bar{u}) = \nabla_{\mathcal{D}} I_{\mathcal{D}} \zeta(\bar{u}(t^{(n+1)}))$ (see (2.29)) with $\nabla \zeta(\bar{u}(t^{(n+1)}))$ incurs a cost measured by $S_{\mathcal{D}}^{\nabla}(\zeta(\bar{u}(t^{(n+1)})))$:

$$\left| \int_{\Omega} \nabla_{\mathcal{D}} \zeta(I_{\mathcal{D}}^{n+1} \bar{u}) \cdot \nabla_{\mathcal{D}} v - \int_{\Omega} \nabla \zeta(\bar{u}(t^{(n+1)})) \cdot \nabla_{\mathcal{D}} v \right| \leq S_{\mathcal{D}}^{\nabla}(\zeta(\bar{u}(t^{(n+1)}))) \|\nabla_{\mathcal{D}} v\|.$$

Hence,

$$\begin{aligned} \mathfrak{E}_{\mathcal{D}}^n(v) = & \delta t^{(n+\frac{1}{2})} \int_{\Omega} f^{(n+1)} \Pi_{\mathcal{D}} v \\ & - \int_{\Omega} [\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})] \Pi_{\mathcal{D}} v - \delta t^{(n+\frac{1}{2})} \int_{\Omega} \nabla \zeta(\bar{u}(t^{(n+1)})) \cdot \nabla_{\mathcal{D}} v \\ & + \mathcal{O}_1 \left[S_{\mathcal{D}}^{\Pi,*}(\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})) + \delta t^{(n+\frac{1}{2})} S_{\mathcal{D}}^{\nabla}(\zeta(\bar{u}(t^{(n+1)}))) \right] \|\nabla_{\mathcal{D}} v\| \end{aligned}$$

where, here and in the following, $\mathcal{O}_1(X)$ denotes a generic function such that $|\mathcal{O}_1(X)| \leq |X|$. Now, by definition of $W_{\mathcal{D}}(\nabla\zeta(\bar{u}(t^{(n+1)})))$,

$$\left| \int_{\Omega} \nabla\zeta(\bar{u}(t^{(n+1)})) \cdot \nabla_{\mathcal{D}}v + \Delta\zeta(\bar{u}(t^{(n+1)}))\Pi_{\mathcal{D}}v \right| \leq W_{\mathcal{D}}(\nabla\zeta(\bar{u}(t^{(n+1)})))\|\nabla_{\mathcal{D}}v\|$$

and thus, writing $\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)}) = \int_{t^{(n)}}^{t^{(n+1)}} \partial_t \bar{u}(s) ds$ (which is valid since the equation (2.1) and the regularity $\nabla\zeta(\bar{u}) \in C([0, T]; H_{\text{div}}(\Omega))$ imply $\partial_t \bar{u} \in C([0, T]; L^2(\Omega))$) and recalling the definition of $f^{(n+1)}$,

$$\begin{aligned} \mathfrak{E}_{\mathcal{D}}^n(v) &= \int_{\Omega} \left[\int_{t^{(n)}}^{t^{(n+1)}} (f - \partial_t \bar{u})(s) ds + \delta t^{(n+\frac{1}{2})} \Delta\zeta(\bar{u}(t^{(n+1)})) \right] \Pi_{\mathcal{D}}v \\ &\quad + \mathcal{O}_1 \left[\mathcal{S}_{\mathcal{D}}^{\Pi,*}(\bar{u}(t^{(n+1)}) - \bar{u}(t^{(n)})) + \delta t^{(n+\frac{1}{2})} \mathcal{S}_{\mathcal{D}}^{\nabla}(\zeta(\bar{u}(t^{(n+1)}))) \right. \\ &\quad \left. + \delta t^{(n+\frac{1}{2})} W_{\mathcal{D}}(\nabla\zeta(\bar{u}(t^{(n+1)}))) \right] \|\nabla_{\mathcal{D}}v\|. \end{aligned} \quad (2.50)$$

Since $f - \partial_t \bar{u} = -\Delta\zeta(\bar{u})$, the property (2.49) yields

$$\begin{aligned} &\left| \int_{\Omega} \left[\int_{t^{(n)}}^{t^{(n+1)}} (f - \partial_t \bar{u})(s) ds + \delta t^{(n+\frac{1}{2})} \Delta\zeta(\bar{u}(t^{(n+1)})) \right] \Pi_{\mathcal{D}}v \right| \\ &\leq \delta t^{(n+\frac{1}{2})} \left| \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} \Delta\zeta(\bar{u}(s)) ds - \Delta\zeta(\bar{u}(t^{(n+1)})) \right|_{\mathcal{D},*} \|\nabla_{\mathcal{D}}v\|. \end{aligned}$$

Plugging this estimate into (2.50) and recalling the definition (2.41) of $E_{\mathcal{D}}^n(v)$, this shows (2.48). \square

With Lemma 2.4 at hand, the next proposition is the main step towards the error estimate in the fully discrete setting.

Proposition 2.2 *For all $n \in \{1, \dots, N\}$, there holds*

$$\begin{aligned} &\max_{1 \leq n \leq N} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^n\|^2 + 4 \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} (\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n+1)}, \Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(n+1)}) \\ &\leq 24T \exp(1) E_{\mathcal{D}}(\bar{u}). \end{aligned} \quad (2.51)$$

Proof Following the lines of the proof in the time discrete case, we sum (2.46) over $j \in \{0, \dots, p-1\}$ with $p \in \{1, \dots, N\}$, leading to

$$\int_{\Omega} \Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(p)} \Pi_{\mathcal{D}}v + \sum_{j=0}^{p-1} \delta t^{(j+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(j+1)} \cdot \nabla_{\mathcal{D}}v = \sum_{j=0}^{p-1} \mathfrak{E}_{\mathcal{D}}^j(v). \quad (2.52)$$

Here, we used the fact that $e_{\mathcal{D},u}^0 = 0$ thanks to the definition of $u^0 = I_{\mathcal{D}}u_{\text{ini}}$. Choosing $v = \delta t^{(p-\frac{1}{2})} e_{\mathcal{D},\zeta}^{(p)}$ in the above equation and summing over $p \in \{1, \dots, n\}$ for some $n \in \{1, \dots, N\}$ provides

$$\mathfrak{J}_1^{(n)} + \mathfrak{J}_2^{(n)} = \mathfrak{R}^{(n)}, \quad (2.53)$$

where

$$\begin{aligned} \mathfrak{J}_1^{(n)} &:= \sum_{p=1}^n \delta t^{(p-\frac{1}{2})} \int_{\Omega} \Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(p)} \Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(p)}, \\ \mathfrak{J}_2^{(n)} &:= \sum_{p=1}^n \sum_{j=1}^p \delta t^{(p-\frac{1}{2})} \delta t^{(j-\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(j)} \cdot \nabla_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(p)}, \\ \mathfrak{R}^{(n)} &:= \sum_{p=1}^n \sum_{j=0}^{p-1} \delta t^{(p-\frac{1}{2})} \mathfrak{E}_{\mathcal{D}}^j(e_{\mathcal{D},\zeta}^{(p)}). \end{aligned}$$

$\mathfrak{J}_1^{(n)}$ corresponds to the third term in (2.39). On the other hand, the identity (2.9) ensures that

$$\mathfrak{J}_2^{(n)} \geq \frac{1}{2} \int_{\Omega} \left| \nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)} \right|^2. \quad (2.54)$$

The term $\mathfrak{R}^{(n)}$ can be reorganized as

$$\mathfrak{R}^{(n)} = \sum_{j=0}^{n-1} \mathfrak{E}_{\mathcal{D}}^j(\varepsilon_{\mathcal{D},\zeta}^{(n)}) - \sum_{j=1}^{n-1} \mathfrak{E}_{\mathcal{D}}^j(\varepsilon_{\mathcal{D},\zeta}^{(j)}) := \mathfrak{R}_1^{(n)} + \mathfrak{R}_2^{(n)}.$$

Owing to Lemma 2.4, the first contribution $\mathfrak{R}_1^{(n)}$ can be estimated as follows:

$$\left| \mathfrak{R}_1^{(n)} \right| \leq \sum_{j=0}^{n-1} \left| \mathfrak{E}_{\mathcal{D}}^j(\varepsilon_{\mathcal{D},\zeta}^{(n)}) \right| \leq \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} E_{\mathcal{D}}^j(\bar{u}) \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|.$$

Applying Cauchy–Schwarz inequality and then the Young inequality (2.12) provides

$$\left| \mathfrak{R}_1^{(n)} \right| \leq T \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} E_{\mathcal{D}}^j(\bar{u})^2 + \frac{1}{4} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2 \leq T E_{\mathcal{D}}(\bar{u})^2 + \frac{1}{4} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2,$$

so that, in view of (2.54), there holds

$$\mathfrak{J}_2^{(n)} - \mathfrak{R}_1^{(n)} \geq \frac{1}{4} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2 - T E_{\mathcal{D}}(\bar{u})^2. \quad (2.55)$$

Using once again Lemma 2.4, one can bound the term $\mathfrak{R}_2^{(n)}$ by

$$\mathfrak{R}_2^{(n)} \leq \sum_{j=1}^{n-1} \left| \mathfrak{E}_{\mathcal{D}}^j(\varepsilon_{\mathcal{D},\zeta}^{(j)}) \right| \leq \sum_{j=1}^{n-1} \delta t^{(j+\frac{1}{2})} E_{\mathcal{D}}^j(\bar{u}) \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(j)}\|.$$

Using the Cauchy–Schwarz and Young inequality as above then yields

$$\mathfrak{R}_2^{(n)} \leq 2T E_{\mathcal{D}}(\bar{u})^2 + \frac{1}{8T} \sum_{j=1}^{n-1} \delta t^{(j+\frac{1}{2})} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(j)}\|^2. \quad (2.56)$$

Combining (2.55)–(2.56) in (2.53) provides

$$\begin{aligned} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2 + 4 \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} (\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(j+1)}, \Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(j+1)}) \\ \leq \frac{1}{2T} \sum_{j=1}^{n-1} \delta t^{(j+\frac{1}{2})} \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(j)}\|^2 + 12T E_{\mathcal{D}}(\bar{u})^2. \end{aligned}$$

The generalized Gronwall Lemma [32, Lemma 5.1] then yields

$$\|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2 + 4 \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} (\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(j+1)}, \Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(j+1)}) \leq 12T \exp(1) E_{\mathcal{D}}(\bar{u}).$$

Choosing $n = N$, and then $n = \operatorname{argmax}_j \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(j)}\|^2$, we obtain (2.51). \square

With Proposition 2.2 we have estimated $\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)}$, $\Pi_{\mathcal{D}} e_{\mathcal{D},\zeta}^{(n)}$ and $\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}$. The proof of Theorem 2.3 is concluded by establishing the discrete $L^\infty(0, T; H^{-1}(\Omega))$ estimate on $(\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)})_{1 \leq n \leq N}$. This is obtained in the lemma below. Together with Proposition 2.2, this completes the proof of Theorem 2.3.

Lemma 2.5 *For all $n \in \{1, \dots, N\}$, there holds*

$$\left| \Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)} \right|_{\mathcal{D},*}^2 \leq 2 \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\|^2 + 2T E_{\mathcal{D}}(\bar{u})^2.$$

Proof Applying (2.52) for $p = n$ and with $v \in X_{\mathcal{D},0}$ such that $\|\nabla_{\mathcal{D}} v\| \leq 1$, using the Cauchy–Schwarz inequality and recalling (2.48), we have

$$\int_{\Omega} \Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)} \Pi_{\mathcal{D}} v \leq \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\| + \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} E_{\mathcal{D}}^j(\bar{u}).$$

Taking the supremum over such v gives

$$\|\Pi_{\mathcal{D}} e_{\mathcal{D},u}^{(n)}\|_{\mathcal{D},*} \leq \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\| + \sum_{j=0}^{n-1} \delta t^{(j+\frac{1}{2})} E_{\mathcal{D}}^j(\bar{u}) \leq \|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D},\zeta}^{(n)}\| + \sqrt{T} E_{\mathcal{D}}(\bar{u}),$$

and the proof follows straightforwardly. \square

2.4 Numerical Examples

2.4.1 Numerical Results

For the numerical tests, we consider the porous medium equation in dimension 2, corresponding to (2.1) with $\zeta(\bar{u}) = |\bar{u}|^{m-1}\bar{u}$ for $m \in \{2, 3, 4\}$. The computational domain is given by $T = 1$ and $\Omega = (0, 1)^2$, and the exact solution is $\bar{u}(t, x) = \mathcal{B}(t_0 + t, x - x_0)$, where $t_0 = 0.1$, $x_0 = (0.5, 0.5)$ and

$$\mathcal{B}(t, x) = t^{-\frac{1}{m}} \left\{ \left[C_{\mathcal{B}} - \frac{m-1}{4m^2} \left(\frac{|x|}{t^{\frac{1}{2m}}} \right)^2 \right]_+ \right\}^{\frac{1}{m-1}}, \quad (2.57)$$

is the Barenblatt–Pattle solution. The initial solution is fixed by $u_{\text{ini}} = \bar{u}(0, \cdot)$. We choose $C_{\mathcal{B}} = 0.005$, so that \mathcal{B} remains equal to 0 on $\partial\Omega$ during the entire simulation $t \in [0, 1]$. Note that by the offset t_0 in \mathcal{B} , the singularity of this function at $t = 0$ is avoided, and the initial condition satisfies Assumption (A2).

The simulations are run over three different mesh families: a family of (mostly) hexagonal meshes, a family of locally refined Cartesian meshes, and a family of triangular meshes. Examples of members of each family are provided in Fig. 2.1. We consider uniform time steps. For the coarsest mesh in each family, the time step is $\delta t^{(n+\frac{1}{2})} = 0.1$ for all n ; then, for each mesh refinement, the time step is divided by 4. Since we use implicit Euler time stepping, this means that the truncation error in time decay as $\mathcal{O}(h^2)$, where h is the mesh size; as our spatial methods (see below) are low order, in the best possible situation (linear equations, smooth exact solution), the maximal approximation rates are $\mathcal{O}(h)$ on gradients and $\mathcal{O}(h^2)$ on functions. The choice of time steps thus ensures that the spatial truncation error is the leading term in the estimate. The following schemes will be used for the tests.

- LEPNC (Locally Enriched Polytopal Non-Conforming scheme), see Chap. 1 of this book: applicable on generic polytopal meshes, one unknown per internal edge (after static condensation), based on broken polynomial functions with weak continuity properties across the edges. We have taken a zero weight ϖ on the edge unknowns, so $\Pi_{\mathcal{D}}$ is only computed from the cell unknowns.

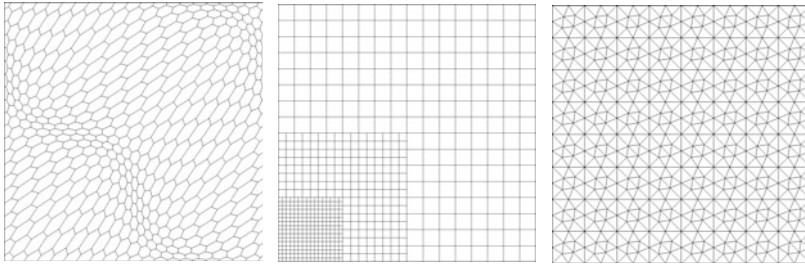


Fig. 2.1 Examples of meshes used in numerical tests: hexagonal (left), locally refined Cartesian (centre), and triangular (right)

- HMM (Hybrid Mimetic Mixed scheme) [22, Chap. 13]: applicable on generic polytopal meshes, one unknown per internal edge (after static condensation), based on local reconstruction of piecewise constant functions and gradients. HMM are the lowest-order version of the Hybrid High-Order (HHO) method, described in [18] (see also Chap. 6 for an application of HHO to poroelasticity problems).
- MLP1 (Mass-Lumped \mathbb{P}^1 finite element) [22, Sect. 8.4]: only applicable on triangular meshes, one unknown per vertex, based on standard \mathbb{P}^1 shape functions for the gradient and piecewise constant reconstruction around each vertex.
- VAG-a (Vertex Approximate Gradient, first presentation) [22, Sect. 8.5]: applicable on generic polytopal meshes, one unknown per internal vertex and one unknown per cell, based on standard \mathbb{P}^1 on a triangular subdivision of the cells (using the center of the cell as additional vertex), with a mass-lumping that equally distributes the available area between cell and vertex unknowns. A local algebraic elimination (static condensation) of cell unknowns is also performed, leading to a globally coupled system on the vertex unknowns only. See also Chap. 3 for an application of the VAG method to a two-phase flow model.
- VAG-b (Vertex Approximate Gradient, second presentation) [12]: as above, but applied after writing the diffusion term as $\operatorname{div}(m|u|^{m-1}\nabla u)$. Note that this scheme does not present itself as a gradient scheme.
- CFVEM (Conforming Virtual Element) [6]: applicable on generic polytopal meshes, one unknown per internal vertex, based on the elliptic projection of virtual shape functions, with algebraic mass-lumping.
- HDG (Hybridizable Discontinuous Galerkin, order 1) [16]: applicable on generic polytopal meshes (but the results are presented here only on triangular meshes), based on modal Legendre–Dubiner basis functions with one polynomial of degree 1 per cell and edge. The degrees of freedom are reduced to only edge polynomials after static condensation.

The LEPNC and HMM tests are based on the code available in the `HArDCore2D` library [31] (based on the implementation principles of Hybrid High-Order methods [18]), while MLP1 tests were conducted using the code at the following URL: github.com/jdroniou/matlab-PME.

Remark 2.7 (CFVEM and \mathbb{P}^1 finite elements on triangular meshes) On triangular meshes and for the standard Laplace problem, CFVEM coincides with the conforming \mathbb{P}^1 finite element method. The results presented below however show different behaviour of CFVEM and MLP1; the main reason can be found in the mass-lumping strategy adopted for each method: for MLP1, a geometrical mass-lumping was used, allocating to each vertex a mass corresponding to $1/3$ of the sum of the areas of the triangles it belongs to; for CFVEM, an algebraic mass-lumping was used, reducing the standard mass matrix to a diagonal one by summing all elements on each row.

The accuracy of the schemes are provided through the following quantities, all measured at the final time:

- Relative error in L^2 -norm between the (reconstructed) gradients of the approximation of $\zeta(\bar{u})$ and the interpolate of $\zeta(\bar{u})$:

$$E_{H^1, \zeta} = \frac{\|\nabla_{\mathcal{D}}(\zeta(u^N) - I_{\mathcal{D}}\zeta(\bar{u})(T, \cdot))\|}{\|\nabla_{\mathcal{D}}I_{\mathcal{D}}\zeta(\bar{u})(T, \cdot)\|}. \quad (2.58)$$

- Relative error in L^{m+1} -norm between the (reconstructed) functions of the approximate solution and the interpolate of the exact solution \bar{u} :

$$E_{L^{m+1}} = \frac{\|\Pi_{\mathcal{D}}(u^N - I_{\mathcal{D}}\bar{u}(T, \cdot))\|_{L^{m+1}(\Omega)}}{\|\Pi_{\mathcal{D}}I_{\mathcal{D}}\bar{u}(T, \cdot)\|_{L^{m+1}(\Omega)}}. \quad (2.59)$$

- Fraction of negative mass over total mass:

$$\text{nMass} = \frac{\int_{\Omega} (\Pi_{\mathcal{D}}u^N)_-}{\int_{\Omega} |\Pi_{\mathcal{D}}u^N|}, \quad (2.60)$$

where $s_- = \max(-s, 0)$ is the negative part of $s \in \mathbb{R}$.

2.4.1.1 Rates of Convergence Versus Mesh Size

We first present the relative errors versus the mesh sizes, for all considered schemes and mesh families. The outputs are given in log-log graphs in Figs. 2.2, 2.3 and 2.4. In these figures, the chosen reference slopes correspond to an estimate of the overall behaviour of the schemes, drawn from the tables as well as computed rates of convergence from one mesh to the other. Combining estimates (2.39), (2.43) and Remark 2.5, and considering averaged-in-time norms (which are less stringent than the final time norms (2.58) and (2.59)), we would expect for a smooth enough exact solution a rate of convergence $\mathcal{O}(h^{\frac{2}{m+1}})$ in L^{m+1} -norm on \bar{u} and $\mathcal{O}(h)$ in L^2 -norm on (the time integral of) $\nabla\zeta(\bar{u})$.

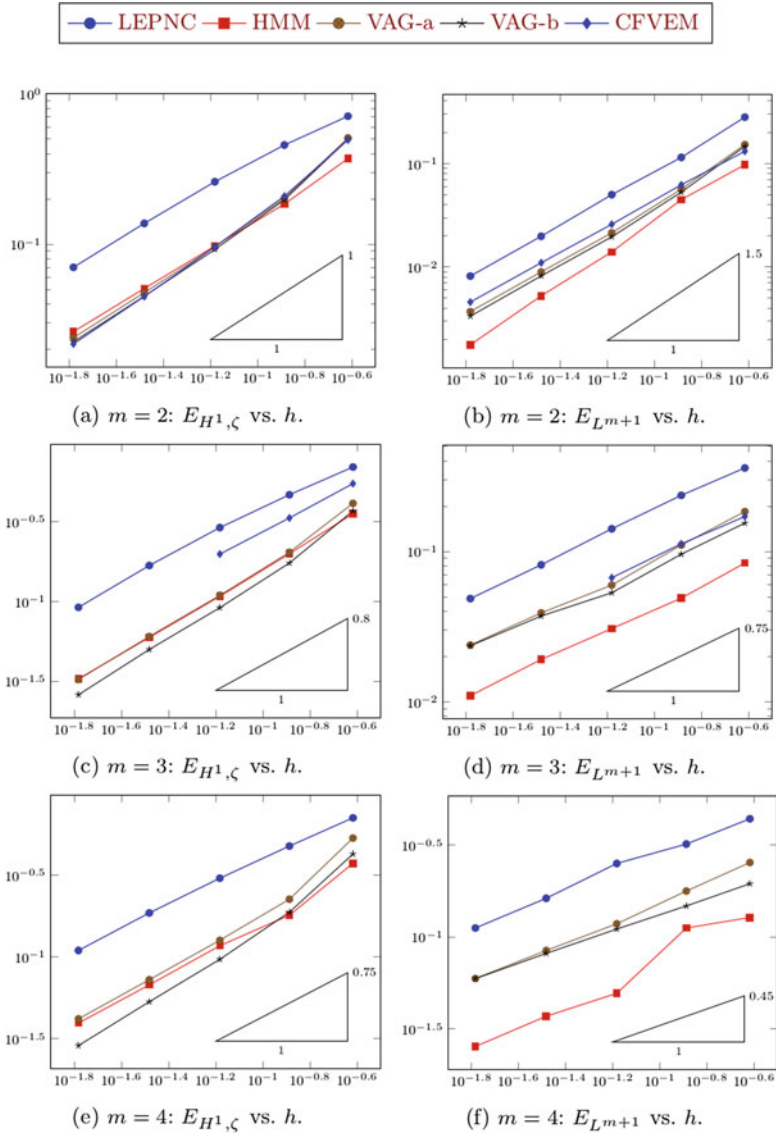


Fig. 2.2 Hexagonal meshes: errors versus mesh size

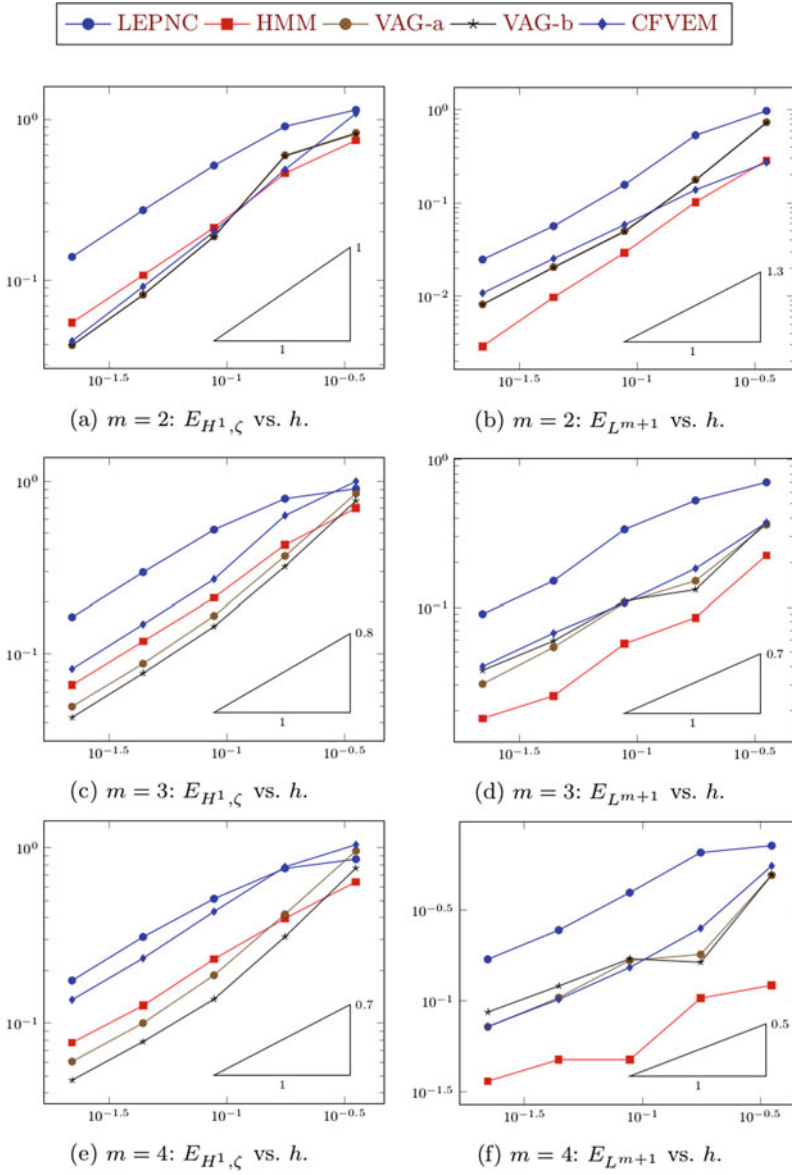


Fig. 2.3 Locally refined Cartesian meshes: errors versus mesh size

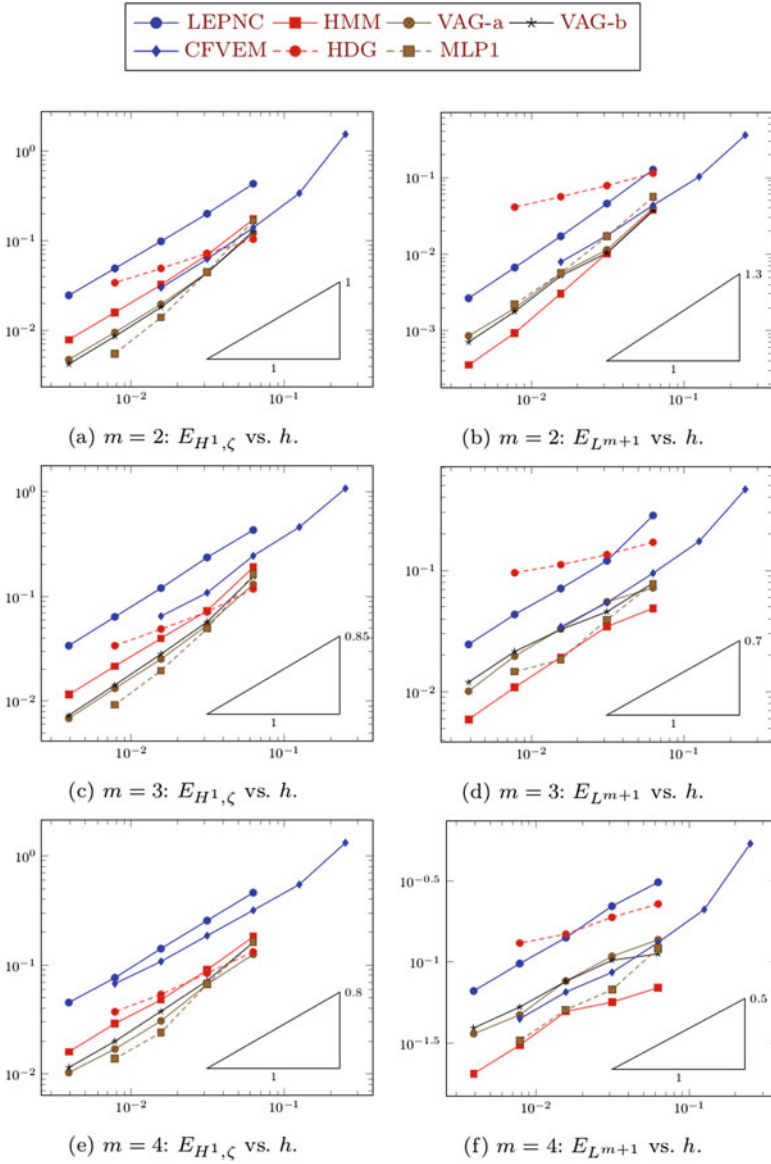


Fig. 2.4 Triangular meshes: errors versus mesh size

As can be seen in the numerical results, for the considered tests, the theoretical rates for $E_{L^{m+1}}$ are sub-optimal for the lower values of m but tend to be close to the observed results for $m = 4$. The general trend, seen in both the theoretical and numerical results, is that the rates deteriorate as m increases. Two reasons can be found for that: as m increases, the L^{m+1} -norm becomes more constrained, while the regularity of the exact solution \bar{u} decays (for example, it is H^1 in space for $m = 2$, but no longer for $m > 2$).

Interestingly, but not surprisingly, the rates for $E_{H^1, \zeta}$ seem to resist a little bit better as m increases, although they appear to be slightly below 1 for $m > 2$ (which is not surprising since, as mentioned, (2.58) is a more constraining norm than $\|\nabla_{\mathcal{D}} \varepsilon_{\mathcal{D}, \zeta}^{(n)}\|$). Although $E_{H^1, \zeta}$ measures an approximation of the gradient, which can be expected to be of lower order than that of a function, it measures this in a norm that is independent of m and relates to $\zeta(\bar{u})$, a function that has better regularity properties than \bar{u} (for example, it is H^1 in space irrespective of the value of m).

Comparing the various schemes, they all seem to adopt similar rates of convergence for $E_{H^1, \zeta}$ on hexagonal and locally refined Cartesian meshes; the differences mostly lie in the multiplicative constants, with the largest factor between these multiplicative constants of order 10. More variation in the rates is observed on these mesh families for $E_{L^{m+1}}$, which is probably due to the variation of m and reduced regularity of \bar{u} , as discussed above. The rates on triangular meshes seem to depend much more on the chosen scheme. Focusing on $E_{H^1, \zeta}$, which is a more stable measure, we see that MLP1 outperforms the other schemes, at least for $m = 2, 3$; of course, the drawback of MLP1 is that it can only be applied on triangular meshes. The other outlier is HDG, whose rates are much lower than the other schemes; the reason for that might be found in the total number of degrees of freedom, after static condensation, which is lower for HDG than some other schemes (see discussion in Sect. 2.4.1.2), and which therefore prevents this scheme from achieving optimal rates with respect to the mesh size.

It can also be noticed that some schemes produce a better $E_{H^1, \zeta}$ error than others, but that the “ranking” between the schemes can be reversed if we look at the error $E_{L^{m+1}}$.

2.4.1.2 Algebraic Complexity

We now briefly discuss the performance of the schemes relative to their algebraic complexity, measured primarily here in terms of their number NDOFs of degrees of freedom. Focusing only on $m = 4$ (the most severe case), and hexagonal and triangular grids, we plot in Fig. 2.5 the energy error $E_{H^1, \zeta}$ of each scheme versus its degrees of freedom.

A first remark is that this measure is slightly more favourable to HDG than in the previous section. Its rate remains a bit lower than that of the other schemes, at least for the considered meshes, but perhaps less so than when comparing the error versus the mesh size. We also notice that, on triangular meshes, the (mostly) vertex-centered methods VAG-a, VAG-b, and MLP1 outperform the other schemes,

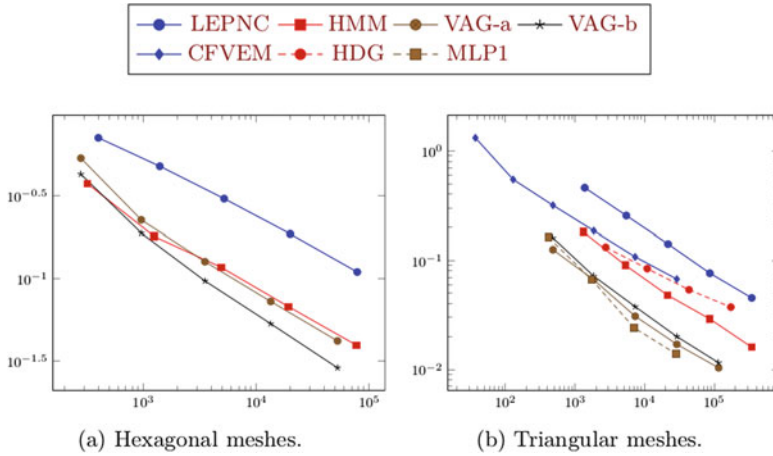


Fig. 2.5 Energy error $E_{H^1, \zeta}$ versus number of degrees of freedom NDOFs, $m = 4$

which is expected: on triangular meshes, vertex-based methods have much fewer degrees of freedom than edge-based methods such as LENCPC or HMM. Curiously, using the same argument, we would expect CFVEM, which is also a vertex-based method, to behave better than it does; this could be explained by the different kind of mass-lumping applied for these two methods.

On hexagonal meshes, except for LEPNC, all schemes presented here have a comparable error vs. complexity. The advantage of vertex-based methods is less perceptible on such meshes than on triangular meshes, which is not surprising: triangular meshes roughly have three times more edges than vertices, while hexagonal meshes have 1.5 times more edges than vertices on average.

2.4.1.3 Positivity

Finally, we look at the positivity properties of the schemes. As the standard linear heat equation, the porous medium equation satisfies a maximum principle: if the initial solution and the source terms are positive, then the solution remains positive for all times. Maintaining this property at the discrete level is particularly challenging, especially for schemes designed for generic polygonal/polyhedral meshes [19]. Except for MLP1 (which is restricted to triangular meshes), none of the schemes presented here satisfy this property in general.

In Figs. 2.6, 2.7 and 2.8, we present the log-log graphs of the relative negative masses N_{Mass} versus the mesh sizes. In most situations, the schemes produce some negative mass, but it decays as the mesh is refined and is rather small relative to the total mass of the solution at the final time. On hexagonal and locally refined Cartesian meshes, the VAG and CFVEM schemes—which are (mostly) vertex-centered—have better positivity properties than the edge-centered schemes LEPNC and HMM, with

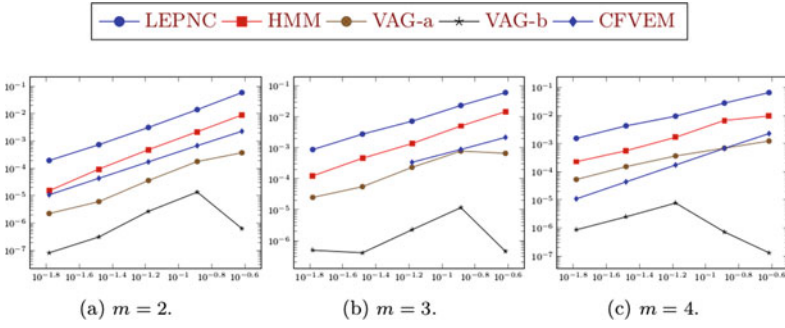


Fig. 2.6 Hexagonal meshes: fraction of negative mass NMass versus mesh size

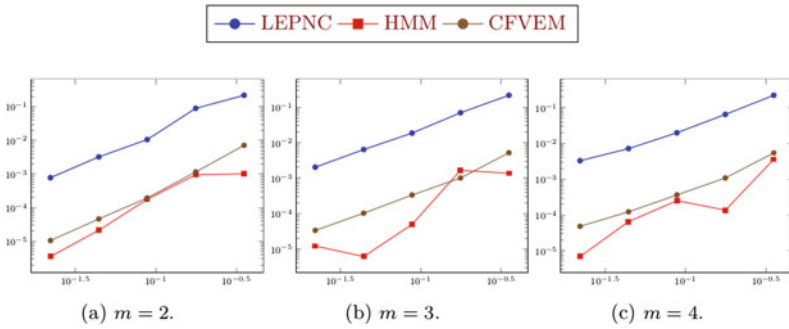


Fig. 2.7 Locally refined Cartesian meshes: fraction of negative mass NMass versus mesh size

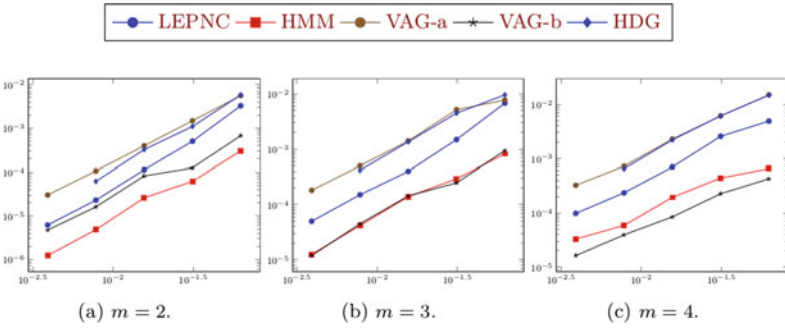


Fig. 2.8 Triangular meshes: fraction of negative mass NMass versus mesh size

VAG-b outperforming all the other schemes. Surprisingly, perhaps, VAG-a and VAG-b do not even produce negative values at the final time on locally refined meshes (and are thus absent from Fig. 2.7). On triangular meshes, however, VAG-a produces more negative mass than the other schemes (and an amount comparable to HDG), with HMM and VAG-b being much closer to each other (with relative performance depending on m), and LEPNC in between these clusters. MLP1 is known to preserve positive values and is therefore not represented in Fig. 2.8. CFVEM was also found to preserve positive values on these triangular meshes.

These results demonstrate a strong interaction between scheme design and mesh geometries when it comes to preserving the maximum principle of the continuous model.

2.4.2 A Word on Non-linear Iterative Schemes

The time discrete problems (2.7) or the fully discrete counterparts (2.30) are nonlinear. To determine an approximation of the solution one needs to employ an iterative method. The common choice is the Newton method (see, e.g., [7]), which converges quadratically. However, this convergence is guaranteed only if the initial guess is close enough to the solution. A choice at hand being the solution computed at the previous time step, this means that the convergence is guaranteed if the differences between the solutions at two successive times are small enough. This induces restrictions on the time step.

For (2.1), the Newton method can fail to converge due to the singularities of ζ , and in particular, for the fast diffusion case. To overcome this, one can regularise ζ to avoid degeneracy, but even in this case, the convergence is only guaranteed under severe restrictions on the time step. To address these shortcomings, alternative iterative schemes have been designed. We mention here the relaxation scheme in [33], which shows to be more stable w.r.t. the choice of the initial condition, and the modified Picard scheme in [15], which is a simplified version of the Newton method. Both schemes are converging linearly. For these, as for the Newton scheme, the convergence is guaranteed rigorously under severe restrictions for the time step, as proved in [47].

A fixed point (contraction) scheme exploiting the monotonicity of ζ has been proposed in [43] for the fast diffusion case and extended to more general situations in [46]. Though linear, the convergence is guaranteed under mild restrictions on the time step, regardless of the initial guess, and for any spatial discretisation. Moreover, as shown in [37], this scheme can be used to obtain a good initial guess for the Newton scheme, which leads to a stable and fast convergent iterative method. We also mention the scheme in [40], where the fixed point approach is combined with the Picard or Newton method by adding a stabilisation term. This leads to a scheme with the stability of the fixed point scheme and converging like the Picard scheme.

Finally, we refer to [8], where both \bar{u} and $\zeta(\bar{u})$ are expressed in terms of a different unknown, based on a properly chosen parametrisation. This allows reformulating the

Newton method in such a way that the quadratic convergence is unaffected, but the stability is significantly improved, so that larger time steps are allowed.

Here we discuss a simple iterative scheme that is inspired by the fixed point approach in [37, 43]. We restrict to the case where ζ is Lipschitz continuous and let L_ζ denote the Lipschitz constant, but the idea can be extended to more general situations as well, e.g., by applying the idea in [8]. For the ease of presentation, we present the scheme in the time-discrete case, the fully discrete one being analogous. With $n \in \{0, \dots, N-1\}$ fixed we start by observing that (2.7) can be rewritten as the system

$$\begin{aligned} (\bar{u}^{(n+1)}, \varphi) + \delta t^{(n+\frac{1}{2})} (\nabla \bar{w}^{(n+1)}, \nabla \varphi) &= (\bar{u}^{(n)}, \varphi) + \delta t^{(n+\frac{1}{2})} (f^{(n+1)}, \varphi), \\ (\bar{w}^{(n+1)}, \psi) &= (\zeta(\bar{u}^{(n+1)}), \psi), \end{aligned} \quad (2.61)$$

for all $\varphi \in H_0^1(\Omega)$ and $\psi \in L^2(\Omega)$. For a given $L \geq \frac{L_\zeta}{2}$, the iterative scheme consists in finding the pairs $(\bar{u}^i, \bar{w}^i) \in L^2(\Omega) \times H_0^1(\Omega)$ ($i \in \mathbb{N}^*$) solving the linear systems

$$\begin{aligned} (\bar{u}^i, \varphi) + \delta t^{(n+\frac{1}{2})} (\nabla \bar{w}^i, \nabla \varphi) &= (\bar{u}^{(n)}, \varphi) + \delta t^{(n+\frac{1}{2})} (f^{(n+1)}, \varphi), \\ (\bar{w}^i, \psi) &= L(\bar{u}^i - \bar{u}^{i-1}, \psi) + (\zeta(\bar{u}^{i-1}), \psi), \end{aligned} \quad (2.62)$$

for all $\varphi \in H_0^1(\Omega)$ and $\psi \in L^2(\Omega)$. A natural choice for the initial guess is $\bar{u}^0 = \bar{u}^{(n)}$ (the solution at the previous time step), but, as will be seen below, the convergence is guaranteed for any starting point. To prove this, we define the iteration errors

$$e_u^i = \bar{u}^{(n+1)} - \bar{u}^i, \quad \text{and} \quad e_w^i = \zeta(\bar{u}^{(n+1)}) - \bar{w}^i. \quad (2.63)$$

From (2.61) and (2.62), the errors satisfy

$$\begin{aligned} (e_u^i, \varphi) + \delta t^{(n+\frac{1}{2})} (\nabla e_w^i, \nabla \varphi) &= 0, \\ (e_w^i, \psi) &= L(e_u^i - e_u^{i-1}, \psi) + (\zeta(\bar{u}^{(n+1)}) - \zeta(\bar{u}^{i-1}), \psi), \end{aligned} \quad (2.64)$$

for all $\varphi \in H_0^1(\Omega)$ and $\psi \in L^2(\Omega)$. With this, the convergence result is

Lemma 2.6 *The iterative scheme in (2.62) is convergent regardless of the initial guess. More precisely, one has $\bar{w}^i \rightarrow \zeta(\bar{u}^{(n+1)})$ in $H^1(\Omega)$ and $\bar{u}^i \rightarrow \bar{u}^{(n+1)}$ in $L^2(\Omega)$ as $i \rightarrow \infty$.*

Proof Taking $\varphi = e_u^i$ and $\psi = e_w^i$ into (2.64) and subtracting the result gives

$$L \|e_u^i\|^2 + \delta t^{(n+\frac{1}{2})} \|\nabla e_w^i\|^2 = (L e_u^{i-1} - (\zeta(\bar{u}^{(n+1)}) - \zeta(\bar{u}^{i-1})), e_u^i).$$

Since ζ is Lipschitz, by the choice of L one has $|L e_u^{i-1} - (\zeta(\bar{u}^{(n+1)}) - \zeta(\bar{u}^{i-1}))| \leq L |e_u^{i-1}|$. This, together with the Cauchy–Schwarz inequality leads to

$$L \|e_u^i\|^2 + \delta t^{(n+\frac{1}{2})} \|\nabla e_w^i\|^2 \leq L \|e_u^{i-1}\| \|e_u^i\|.$$

Applying now (2.12) and multiplying the resulting by 2 yields

$$L\|e_u^i\|^2 + 2\delta t^{(n+\frac{1}{2})}\|\nabla e_w^i\|^2 \leq L\|e_u^{i-1}\|^2. \quad (2.65)$$

Adding (2.65) for $i = 1, \dots, k$ (k being arbitrary) leads to

$$L\|e_u^k\|^2 + 2\delta t^{(n+\frac{1}{2})}\sum_{i=1}^k\|\nabla e_w^i\|^2 \leq L\|e_u^0\|^2. \quad (2.66)$$

This shows that the second term above is a convergent series, implying the first convergence result. Using this convergence in the first equation of (2.64) completes the proof. \square

The convergence extends straightforwardly to the fully discrete case. We use the notations in Sect. 2.3 and apply the GDM to the time discrete system in (2.62). Given $u^{(n)} \in X_{\mathcal{D},0}$, for $i \geq 1$ we seek $u^i, w^i \in X_{\mathcal{D},0}$ such that, for all $\phi, \psi \in X_{\mathcal{D},0}$,

$$\begin{aligned} & \int_{\Omega} \Pi_{\mathcal{D}} u^i \Pi_{\mathcal{D}} \phi + \delta t^{(n+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} w^i \cdot \nabla_{\mathcal{D}} v \\ &= \int_{\Omega} \Pi_{\mathcal{D}} u^{(n)} \Pi_{\mathcal{D}} v + \delta t^{(n+\frac{1}{2})} \int_{\Omega} f^{(n+1)} \Pi_{\mathcal{D}} v \\ & \int_{\Omega} \Pi_{\mathcal{D}} w^i \Pi_{\mathcal{D}} \psi = L \int_{\Omega} \Pi_{\mathcal{D}} (u^i - u^{i-1}) \Pi_{\mathcal{D}} \psi + \int_{\Omega} \Pi_{\mathcal{D}} \zeta(u^{i-1}) \Pi_{\mathcal{D}} \psi. \end{aligned} \quad (2.67)$$

As before, a good starting point is $u^0 = u^{(n)}$ but this choice is not required for the convergence. Using the errors

$$e_{\mathcal{D},u}^i = u^{(n+1)} - u^i \quad \text{and} \quad e_{\mathcal{D},w}^i = \zeta(u^{(n+1)}) - w^i, \quad (2.68)$$

from (2.30) and (2.67) one obtains

$$\begin{aligned} & \int_{\Omega} \Pi_{\mathcal{D}} e_{\mathcal{D},u}^i \Pi_{\mathcal{D}} \phi + \delta t^{(n+\frac{1}{2})} \int_{\Omega} \nabla_{\mathcal{D}} e_{\mathcal{D},w}^i \cdot \nabla_{\mathcal{D}} \phi = 0 \\ & \int_{\Omega} \Pi_{\mathcal{D}} e_{\mathcal{D},w}^i \Pi_{\mathcal{D}} \psi = L \int_{\Omega} \Pi_{\mathcal{D}} (e_{\mathcal{D},u}^i - e_{\mathcal{D},u}^{i-1}) \Pi_{\mathcal{D}} \psi \\ & \quad + \int_{\Omega} \Pi_{\mathcal{D}} (\zeta(u^{(n+1)}) - \zeta(u^{i-1})) \Pi_{\mathcal{D}} \psi. \end{aligned} \quad (2.69)$$

The convergence of the fully discrete iteration is obtained by taking in the above $\phi = e_{\mathcal{D},w}^i$ and $\psi = e_{\mathcal{D},u}^i$, and following the steps of the proof for Lemma 2.6. More precisely, we obtain the L^2 -convergences $\nabla_{\mathcal{D}} w^i \rightarrow \nabla_{\mathcal{D}} \zeta(u^{(n+1)})$ and $\Pi_{\mathcal{D}} u^i \rightarrow \Pi_{\mathcal{D}} u^{(n+1)}$, as $i \rightarrow \infty$. We omit the details of the proof as they are similar to those developed in the semi-discrete case above in this section.

Acknowledgements CC acknowledges support from the Labex CEMPI (ANR-11-LABX-0007-01). JD and GM were partially supported by the Australian Government through the Australian Research Council's Discovery Projects funding scheme (project DP170100605); GM was also partially supported by the ERC Project CHANGE, which has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No 694515). MBO and ISP are supported by the the Research Foundation-Flanders (FWO), Belgium through the Odysseus programme project G0G1316N.

References

1. H.W. Alt, S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183**, 311–341 (1983)
2. B. Andreianov, C. Cancès, A. Moussa, A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. *J. Funct. Anal.* **273**(12), 3633–3670 (2017)
3. O. Angelini, K. Brenner, D. Hilhorst, A finite volume method on general meshes for a degenerate parabolic convection-reaction-diffusion equation. *Numer. Math.* **123**(2), 219–257 (2013)
4. T. Arbogast, M.F. Wheeler, N.-Y. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.* **33**(4), 1669–1687 (1996)
5. J.W. Barrett, P. Knabner, Finite element approximation of the transport of reactive solutes in porous media. II. Error estimates for equilibrium adsorption processes. *SIAM J. Numer. Anal.* **34**(2), 455–479 (1997)
6. L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D. Marini, A. Russo, Basic principles of virtual element methods. *Math. Models Methods Appl. Sci. (M3AS)* **199**(23), 199–214 (2013)
7. L. Bergamaschi, M. Putti, Mixed finite elements and Newton-type linearizations for the solution of Richards' equation. *Internat. J. Numer. Methods Engrg.* **45**(8), 1025–1046 (1999)
8. K. Brenner, C. Cancès, Improving Newton's method performance by parametrization: the case of the Richards equation. *SIAM J. Numer. Anal.* **55**(4), 1760–1785 (2017)
9. C. Cancès, Energy stable numerical methods for porous media flow type problems. *Oil Gas Sci. Technol. Rev. IFP Energ. Nouv.* **73** (2018)
10. C. Cancès, T. Gallouët, On the time continuity of entropy solutions. *J. Evol. Equ.* **11**(1), 43–55 (2011)
11. C. Cancès, C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.* **85**(298), 549–580 (2016)
12. C. Cancès, C. Guichard, Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found. Comput. Math.* **17**(6), 1525–1584 (2017)
13. C. Cancès, I.S. Pop, M. Vohralík, An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow. *Math. Comp.* **83**(285), 153–188 (2014)
14. J. Carrillo, Entropy solutions for nonlinear degenerate problems. *Arch. Ration. Mech. Anal.* **147**(4), 269–361 (1999)
15. M. Celia, E. Bouloutas, R. Zarba, A general mass-conservative numerical-solution for the unsaturated flow equation. *Water Resour. Res.* **26**(7), 1483–1496 (1990)
16. B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.* **47**(2), 1319–1365 (2009)

17. D.A. Di Pietro, J. Droniou, A third Strang lemma and an Aubin-Nitsche trick for schemes in fully discrete formulation. *Calcolo* **55**(3), Art. 40, 39 (2018)
18. D.A. Di Pietro, J. Droniou, The hybrid high-order method for polytopal meshes: design, analysis, and applications, in *Modeling, Simulation and Applications* (Springer International Publishing, 2020)
19. J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci. (M3AS)* **24**(8), 1575–1619 (2014). Special issue on Recent Techniques for PDE Discretizations on Polyhedral Meshes
20. J. Droniou, R. Eymard, Uniform-in-time convergence of numerical methods for non-linear degenerate parabolic equations. *Numer. Math.* **132**(4), 721–766 (2016)
21. J. Droniou, R. Eymard, High-order mass-lumped schemes for nonlinear degenerate elliptic equations. *SIAM J. Numer. Anal.* **58**(1), 153–188 (2020)
22. J. Droniou, R. Eymard, T. Gallouët, C. Guichard, R. Herbin, The gradient discretisation method. *Math. Appl.* (2018)
23. J. Droniou, R. Eymard, T. Gallouët, R. Herbin, Gradient schemes: a generic framework for the discretisation of linear, nonlinear and nonlocal elliptic and parabolic equations. *Math. Models Methods Appl. Sci. (M3AS)* **23**(13), 2395–2432 (2013)
24. J. Droniou, K.-N. Le, The gradient discretisation method for slow and fast diffusion porous media equations. *SIAM J. Numer. Anal.* **58**(3), 1965–1992 (2020). <https://doi.org/10.1137/19M1260165>
25. Y. Epshteyn, B. Rivière, Analysis of hp discontinuous Galerkin methods for incompressible two-phase flow. *J. Comput. Appl. Math.* **225**(2), 487–509 (2009)
26. A. Ern, I. Mozolevski, Discontinuous Galerkin method for two-component liquid-gas porous media flows. *Comput. Geosci.* **16**(3), 677–690 (2012)
27. R. Eymard, P. Féron, T. Gallouët, C. Guichard, R. Herbin, Gradient schemes for the Stefan problem. *Int. J. Finite Vol.* **13**, 1–37 (2013)
28. R. Eymard, T. Gallouët, D. Hilhorst, Y. Naït Slimane, Finite volumes and nonlinear diffusion equations. *RAIRO Modél. Math. Anal. Numér.* **32**(6), 747–761 (1998)
29. R. Eymard, R. Herbin, A. Michel, Mathematical study of a petroleum-engineering scheme. *M2AN Math. Model. Numer. Anal.* **37**(6), 937–972 (2003)
30. R. Eymard, D. Hilhorst, M. Vohralík, A combined finite volume-nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numer. Math.* **105**(1), 73–131 (2006)
31. HARDCore2D—Hybrid Arbitrary Degree::Core 2D. <https://github.com/jdroniou/HARDCore2D-release>, Version 2.0.2
32. J.G. Heywood, R. Rannacher, Finite-element approximation of the nonstationary Navier-Stokes problem. IV. Error analysis for second-order time discretization. *SIAM J. Numer. Anal.* **27**(2), 353–384 (1990)
33. W. Jäger, J. Kačur, Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *RAIRO Modél. Math. Anal. Numér.* **29**(5), 605–627 (1995)
34. R.A. Klausen, F.A. Radu, G.T. Eigestad, Convergence of MPFA on triangulations and for Richards' equation. *Internat. J. Numer. Methods Fluids* **58**(12), 1327–1351 (2008)
35. O.A. Ladyženskaja, V.A. Solonnikov, N.N. Ural'ceva, Linear and quasilinear equations of parabolic type. Translated from the Russian by S. Smith. *Translations of Mathematical Monographs*, Vol. 23. (American Mathematical Society, Providence, R.I., 1967)
36. N. Liao, A unified approach to the holder regularity of solutions to degenerate and singular parabolic equations. *J. Differential Equations* **268**(10), 5704–5750 (2020)
37. F. List, F.A. Radu, A study on iterative methods for solving Richards' equation. *Comput. Geosci.* **20**(2), 341–353 (2016)
38. E. Magenes, R.H. Nochetto, C. Verdi, Energy error estimates for a linear scheme to approximate nonlinear parabolic problems. *ESAIM: Math. Modell. Numer. Anal.* **21**(4), 655–678 (1987)
39. A.M. Meirmanov, The Stefan, problem, volume 3 of de Gruyter Expositions in Mathematics. Walter de Gruyter & Co., Berlin, *Translated from the Russian by Marek Niezgodka and Anna Crowley* (With an appendix by the author and I. G. Götz, 1992)

40. K. Mitra, I.S. Pop, A modified L-scheme to solve nonlinear diffusion problems. *Comput. Math. Appl.* **77**(6), 1722–1738 (2019)
41. R.H. Nochetto, C. Verdi, Approximation of degenerate parabolic problems using numerical integration. *SIAM J. Numer. Anal.* **25**(4), 784–814 (1988)
42. F. Otto, L^1 -contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differential Equations* **131**, 20–38 (1996)
43. I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure. *J. Comput. Appl. Math.* **168**(1–2), 365–373 (2004)
44. I.S. Pop, B. Schweizer, Regularization schemes for degenerate Richards equations and outflow conditions. *Math. Models Methods Appl. Sci.* **21**(8), 1685–1712 (2011)
45. I.S. Pop, M. Sepúlveda, F.A. Radu, O.P. Vera Villagrán. Error estimates for the finite volume discretization for the porous medium equation. *J. Comput. Appl. Math.* **234**(7), 2135–2142 (2010)
46. F.A. Radu, K. Kumar, J.M. Nordbotten, I.S. Pop, A robust, mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities. *IMA J. Numer. Anal.* **38**(2), 884–920 (2018)
47. F.A. Radu, I.S. Pop, P. Knabner, Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations, in *Numerical Mathematics and Advanced Applications* (Springer, Berlin, 2006), pp. 1192–1200
48. F.A. Radu, I.S. Pop, P. Knabner, Error estimates for a mixed finite element discretization of some degenerate parabolic equations. *Numer. Math.* **109**(2), 285–311 (2008)
49. J.L. Vázquez, Smoothing and decay estimates for nonlinear diffusion equations: Equations of porous medium type, in *Oxford Lecture Series in Mathematics and its Applications* (Oxford University Press, Oxford, 2006)
50. J.L. Vázquez. The porous medium equation: Mathematical theory, in *Oxford Mathematical Monographs* (The Clarendon Press Oxford University Press, Oxford, 2007)
51. M. Vohralík, M.F. Wheeler, A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows. *Comput. Geosci.* **17**(5), 789–812 (2013)
52. C.S. Woodward, C.N. Dawson, Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media. *SIAM J. Numer. Anal.* **37**(3), 701–724 (2000)
53. I. Yotov, A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arising in porous media flow. *East-West J. Numer. Math.* **5**(3), 211–230 (1997)
54. W.P. Ziemer, Interior and boundary continuity of weak solutions of degenerate parabolic equations. *Trans. Amer. Math. Soc.* **271**, 733–748 (1982)

Chapter 3

Nodal Discretization of Two-Phase Discrete Fracture Matrix Models



Konstantin Brenner, Julian Hennicker, and Roland Masson

Abstract This chapter reviews the nodal Vertex Approximate Gradient (VAG) discretization of two-phase Darcy flows in fractured porous media for which the fracture network is represented as a manifold of co-dimension one with respect to the surrounding matrix domain. Different types of models and their discretizations are considered depending on the transmission conditions set at matrix fracture interfaces accounting for fractures acting either as drains or both as drains or barriers. Difficulties raised by nodal discretizations in heterogeneous media are investigated and solutions to solve these issues are discussed. It includes the adaptation of the porous volumes at nodal unknowns and discontinuous saturations accounting for the jumps induced by the discontinuity in space of the capillary pressure functions. A new Multi-Point upwind scheme is also introduced for the approximation of the mobilities at matrix fracture interfaces to address the issue of fluxes not defined at faces. The most accurate approach is based on the extension of the discontinuous pressure model to two-phase Darcy flows taking into account the discontinuities of both the pressures and saturations at matrix fracture interfaces. As opposed to single phase flows, It improves the accuracy even in the case of fracture acting as drains. On the other hand this approach can still exhibit a robustness issue in terms of nonlinear convergence.

Keywords Two-phase Darcy flows · Heterogeneous media · Discrete fracture matrix models · Nodal discretization · Finite volume · Vertex approximate gradient · Discontinuous capillary pressures

K. Brenner · R. Masson (✉)
Université Côte d'Azur, CNRS, Inria, LJAD, Parc Valrose, 06108 Nice, France
e-mail: roland.masson@univ-cotedazur.fr

K. Brenner
e-mail: konstantin.brenner@univ-cotedazur.fr

J. Hennicker
University of Geneva, Geneva, Switzerland
e-mail: julian.hennicker@unige.ch

3.1 Introduction

Many real life applications in the geosciences like oil and gas recovery, basin modelling, energy storage, geothermal energy or hydrogeology involve two-phase Darcy flows in heterogeneous porous media. Such models are governed by nonlinear partial differential equations typically coupling elliptic and degenerate parabolic equations. Next to the inherent difficulties posed by such equations, further challenges are due to the heterogeneity of the medium and the presence of discontinuities like fractures. This has a strong impact on the complexity of the models, challenging the development of efficient simulation tools.

This work focuses on the numerical modelling of two-phase Darcy flows in fractured porous media, for which the fracture network is represented as a manifold of co-dimension one with respect to the matrix domain. These reduced models are obtained by averaging the physical unknowns as well as the conservation equations along the fracture width. They are termed hybrid-dimensional or also Discrete Fracture Matrix (DFM) Darcy flow models. Given the high geometrical complexity of real life fracture networks, the main advantages of these hybrid-dimensional compared with equi-dimensional models are both to facilitate the mesh generation and the discretisation of the model, and to reduce the computational cost of the resulting schemes. This type of hybrid-dimensional models is the object of intensive researches since the last 15 years due to the ubiquity of fractures in geology and their considerable impact on the flow and transport in the porous medium.

DFM models are closed with appropriate transmission conditions at matrix fracture (mf) interfaces which differ for fractures acting as drains or as barriers. For single-phase flows there are two major approaches. The first, designed for modelling highly conductive fractures and referred to as continuous pressure model [7, 17], assumes the continuity of the fluid pressure at the mf interfaces. The second approach, referred to as discontinuous pressure model [10, 15, 24, 32, 33, 39, 41], allows to represent fractures acting as permeability barriers by imposing Robin-type transmission conditions at mf interfaces.

When the modelling of two-phase flow is concerned, three major types of models can be distinguished. The first and most common type is based on a straightforward adaptation of the single-phase continuous pressure model to the two-phase setting (see [13, 14, 20, 38, 43, 44]), it assumes the continuity of each phase pressure at mf interfaces which allows to capture the saturation jump for fractures acting as drains and matrix as barrier. As for single-phase flow, this approach cannot account for fractures acting as barriers. In contrast to the single-phase context, let us stress that, due to heterogeneous capillary pressures, fractures having a large absolute permeability may still act as barriers for a given phase, typically for the wetting phase for fractures filled by the non-wetting phase (see [1]). Another existing type of models, accounting for both drains or permeability barriers, is based on the linear (without mobility but including gravity) single-phase Darcy flux conservation equation imposed at mf interfaces for each phase. It is usually combined with Two-Point [1, 41] or Multi-Point [4, 5, 36, 46, 51, 52] cell-centred finite volume schemes for which

the interfacial discontinuous pressures are eliminated when building the single phase Darcy flux transmissibilities. These models account for the discontinuity of the pressures but not of the mobilities at mf interfaces. Both previous types of models are based on linear mf transmission conditions. The last type of models considers nonlinear mf transmission conditions which are based on the nonlinear (including mobility) two-phase normal flux continuity equations at mf interfaces. This type of models is considered in [1, 2, 6, 16, 25, 26] using a two-point flux approximation in the fracture width with upwinding of the mobilities, and in [3, 40] using a global pressure formulation. Such nonlinear transmission conditions account for the discontinuity of both the phase pressures and the mobilities at mf interfaces. A comparison of these three types of models using reference equi-dimensional solutions can be found in [1, 16].

Having in mind that tetrahedral meshes are commonly used to cope with the geometrical complexity of fracture networks, nodal discretizations of DFM two-phase Darcy flow models have a clear advantage over cell-centred or face based discretizations thanks to their much lower number of degrees of freedom (d.o.f.). This is in particular the case when considering fully coupled implicit time integration which are necessary to avoid severe time step restrictions in high velocity regions such as fractures and to account for the strong coupling between the pressure and saturation unknowns at mf interfaces [9]. Alternatively, cell centred discretizations have been considered for DFM two-phase flow models using the Two-Point Flux Approximation (TPFA) as in [1, 6, 41] or Multi-Point Flux Approximations (MPFA) as in [5, 36, 52]. Face based discretizations have been considered in [3, 38] using the Mixed Hybrid Finite Element (MHFE) method and in [2, 37] using the Hybrid Finite Volume (HFV) scheme. Non conforming discretizations have also been developed for this type of models using XFEM discretizations as in [34] or Embedded Discrete Fracture Models as in [50].

Nodal discretizations, such as the Control Volume Finite Element (CVFE) method, have been first introduced in [20, 35, 43, 44] for DFM two-phase Darcy flow models with continuous pressures at mf interfaces accounting for fractures acting as drains. In this work, we review the Vertex Approximate Gradient (VAG) discretization introduced in [13, 14, 53] for continuous pressure models and in [16, 25] for discontinuous pressure models. The VAG scheme is based on nodal d.o.f. like CVFE methods but it also includes the cell d.o.f. which are eliminated at the linear algebra level at each Newton iteration without any fill-in. These cell d.o.f. provide an additional flexibility in the design of the discretization allowing to cope with traditional issues raised at mf interfaces by nodal discretizations of the transport equation. On practical meshes, for which the cell sizes at mf interfaces are much larger than the fracture width, these issues are induced by the use of dual control volumes combined with heterogeneous petrophysical and hydrodynamical properties defined on the primal mesh.

The outline of the remaining of this article is as follows. Section 3.2 describes the DFM continuous and discontinuous pressure two-phase Darcy flow models as introduced in [13, 16]. Section 3.3 presents the VAG discretizations of DFM continuous pressure two-phase Darcy flow models. Several techniques to cope with the issues raised by nodal discretizations at mf interfaces are discussed, including the adaptation

of the control volumes at mf interfaces, a new Multi-Point upwind approximation of the mobilities in Sect. 3.3.3, and taking into account the saturation jump for general capillary pressure curves in Sect. 3.3.5. Section 3.4 reviews the VAG discretizations of the three types of DFM discontinuous pressure two-phase Darcy flow models as presented in [16, 25]. For each type of model and its VAG discretization, numerical experiments are exhibited on 2D and 3D DFM models including comparisons of the VAG discretizations to a face based scheme, as well as the comparison between the hybrid-dimensional DFM models and the reference equi-dimensional model.

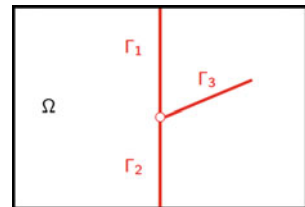
3.2 Two-Phase DFM Discontinuous and Continuous Pressure Models

Let Ω be a bounded domain of \mathbb{R}^d , $d = 2, 3$ assumed to be polyhedral for $d = 3$ and polygonal for $d = 2$. To fix ideas, the dimension will be fixed to $d = 3$ when it needs to be specified, for instance in the naming of the geometrical objects or for the space discretization. The adaptations to the case $d = 2$ are straightforward. Let $\bar{\Gamma} = \bigcup_{i \in I} \bar{\Gamma}_i$ denotes the network of fractures $\Gamma_i \subset \Omega$, $i \in I$, such that each Γ_i is a planar polygonal simply connected open domain included in some plane of \mathbb{R}^d (see Fig. 3.1).

In the matrix domain Ω , we denote by $\phi_m(\mathbf{x})$ the porosity and by $\Lambda_m(\mathbf{x})$ the permeability tensor. Along the fracture network $\mathbf{x} \in \Gamma$, we denote by $\phi_f(\mathbf{x})$ the porosity averaged on the fracture width and by $d_f(\mathbf{x})$ the fracture aperture. The permeability tensor is assumed constant along the width of the fracture and the normal vector to the fracture is assumed to be a principal direction. It results that we can define along the fracture network $\mathbf{x} \in \Gamma$, the tangential permeability tensor $\Lambda_f(\mathbf{x})$ and the normal permeability $\lambda_{n,f}(\mathbf{x})$.

It is assumed, for the sake of simplicity, that the matrix (resp. the fracture network) has a single rock type. Hence, for each phase $\alpha \in \{nw, w\}$ (where nw stands for the non-wetting phase and w for the wetting phase) we denote by $M_m^\alpha(s^\alpha)$ (resp. $M_f^\alpha(s^\alpha)$), the matrix (resp. fracture network) phase mobility, and by $P_{c,m}(s^{nw})$ (resp. $P_{c,f}(s^{nw})$), the matrix (resp. fracture network) capillary pressure function. The inverse of the monotone graph extension of the matrix (resp. fracture network) capillary pressure

Fig. 3.1 Example of a 2D DFM with the matrix domain Ω and 3 intersecting fractures Γ_i , $i = 1, 2, 3$



is denoted by $S_m^{nw}(p)$ (resp. $S_f^{nw}(p)$). We will also denote by ρ^α the phase density which for the sake of simplicity is assumed constant for both phases $\alpha \in \{nw, w\}$.

Let $\alpha \in \{nw, w\}$, we denote by u_m^α (resp. u_f^α) the phase pressure and by s_m^α (resp. s_f^α) the phase saturation in the matrix (resp. the fracture network) domain. The Darcy velocity of phase $\alpha \in \{nw, w\}$ in the matrix domain is defined by

$$\mathbf{q}_m^\alpha = -M_m^\alpha(s_m^\alpha)\Lambda_m(\nabla u_m^\alpha - \rho^\alpha \mathbf{g}),$$

where $\mathbf{g} = -g\nabla z$ stands for the gravity vector with g the gravitational acceleration constant. The flow in the matrix domain is described by the volume balance equation

$$\phi_m \partial_t s_m^\alpha + \operatorname{div}(\mathbf{q}_m^\alpha) = 0, \quad (3.1)$$

for $\alpha \in \{nw, w\}$, and the closure laws defined by the macroscopic capillary pressure law together with the sum to one of the phase saturations

$$s_m^{nw} = S_m^{nw}(p_{c,m}), \quad p_{c,m} = u_m^{nw} - u_m^w, \quad s_m^w = 1 - s_m^{nw}. \quad (3.2)$$

On the fracture network Γ , we denote by ∇_τ the tangential gradient and by div_τ the tangential divergence. In addition, we can define the two sides \pm of the fracture network Γ in $\Omega \setminus \bar{\Gamma}$ and the corresponding unit normal vectors \mathbf{n}^\pm at Γ inward to the sides \pm . Let $\gamma_{\mathbf{n}^\pm}$ (resp. γ^\pm) formally denote the normal trace (resp. trace) operators at both sides of the fracture network Γ for vector fields in $H_{\operatorname{div}}(\Omega \setminus \bar{\Gamma})$ (resp. scalar fields in $H^1(\Omega \setminus \bar{\Gamma})$). The Darcy tangential velocity of phase $\alpha \in \{nw, w\}$ in the fracture network Γ integrated over the width of the fracture is defined by

$$\mathbf{q}_f^\alpha = -d_f M_f^\alpha(s_f^\alpha)\Lambda_f(\nabla_\tau u_f^\alpha - \rho^\alpha \mathbf{g}_\tau),$$

with $\mathbf{g}_\tau = \mathbf{g} - (\mathbf{g} \cdot \mathbf{n}^+) \mathbf{n}^+$. The flow in the fracture network Γ is described, for each phase $\alpha \in \{nw, w\}$, by the volume balance equation

$$d_f \phi_f \partial_t s_f^\alpha + \operatorname{div}_\tau(\mathbf{q}_f^\alpha) + \gamma_{\mathbf{n}^+} \mathbf{q}_m^\alpha + \gamma_{\mathbf{n}^-} \mathbf{q}_m^\alpha = 0, \quad (3.3)$$

and by the closure laws

$$s_f^{nw} = S_f^{nw}(p_{c,f}), \quad p_{c,f} = u_f^{nw} - u_f^w, \quad s_f^w = 1 - s_f^{nw}. \quad (3.4)$$

3.2.1 Two-Phase DFM Discontinuous Pressure Model

We consider the transmission conditions introduced in [16]. They are based on a two-point approximation of each phase normal flux within the fracture combined with a phase potential upwinding of the phase mobility taking into account the phase

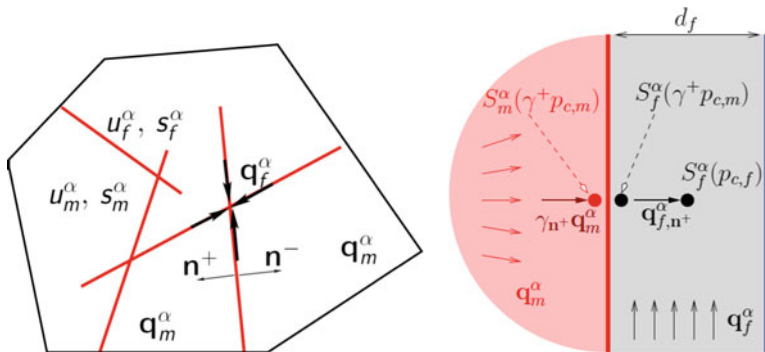


Fig. 3.2 (Left): example of a 2D DFM discontinuous pressure model with the normal vectors \mathbf{n}^\pm at both sides of a fracture, the matrix phase pressure and saturation u_m^α, s_m^α , the fracture phase pressure and saturation u_f^α, s_f^α , the matrix Darcy phase velocity \mathbf{q}_m^α and the fracture network tangential Darcy phase velocity \mathbf{q}_f^α . (Right): illustration of the coupling condition $\mathbf{q}_{f,n^+}^\alpha = \gamma_{n^+}^\alpha \mathbf{q}_m^\alpha$ for the hybrid-dimensional discontinuous pressure model

saturation jump at the mf interface. Let us first define, for both phases $\alpha \in \{nw, w\}$, the “single” phase normal flux in the fracture network

$$V_{f,n}^{\alpha,\pm} = \lambda_{f,n} \left(\frac{\gamma^\pm u_m^\alpha - u_f^\alpha}{d_f/2} - \rho^\alpha \mathbf{g} \cdot \mathbf{n}^\pm \right), \quad (3.5)$$

which does not include the phase mobility. For any $a \in \mathbb{R}$, let us set $a^+ = \max\{0, a\}$ and $a^- = \min\{0, a\}$. The conditions coupling the matrix and fracture unknowns then read, for $\alpha \in \{nw, w\}$ (see the right Fig. 3.2):

$$\gamma_{n^\pm}^\alpha \mathbf{q}_m^\alpha = \mathbf{q}_{f,n^\pm}^\alpha, \quad \mathbf{q}_{f,n^\pm}^\alpha = M_f^\alpha(S_f^\alpha(\gamma^\pm p_{c,m}))(V_{f,n}^{\alpha,\pm})^+ + M_f^\alpha(s_f^\alpha)(V_{f,n}^{\alpha,\pm})^-. \quad (3.6)$$

The hybrid dimensional two-phase flow discontinuous pressure model looks for $u_m^\alpha, u_f^\alpha, s_m^\alpha, s_f^\alpha, \alpha \in \{nw, w\}$, satisfying (3.1)–(3.2) and (3.3)–(3.4) together with the transmission conditions (3.6).

3.2.2 Two-Phase DFM Continuous Pressure Model

In the case of pervious fractures, for which the ratio of the transversal permeability of the fracture to the width of the fracture is large compared with the ratio of the permeability of the matrix to the size of the domain, it is classical to assume that the phase pressures are continuous at the interfaces between the fractures and the matrix domain. Let us also mention that in the context of two-phase flows the continuous pressure DFM models have to be used with caution. It has been shown in [1, 16] that

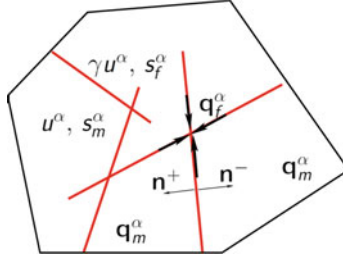


Fig. 3.3 Example of a 2D DFM continuous pressure model with the normal vectors \mathbf{n}^\pm at both sides of a fracture, the phase pressure u^α and its trace γu^α on the fracture network Γ , the matrix phase saturation s_m^α , the fracture phase saturation s_f^α , the matrix Darcy phase velocity \mathbf{q}_m^α and the fracture network tangential Darcy phase velocity \mathbf{q}_f^α

even highly pervious fractures may still act as barriers. This is due to the potential degeneracy of the mobilities in the transmission condition (3.6) and to the saturation jumps resulting from the high contrast of the capillary pressure curves across mf interface. Typically a fracture filled with the non-wetting phase would act as a barrier for the wetting phase, and therefore would induce a discontinuity of the wetting phase's pressure. We refer to [1, 16] for a detailed comparison of continuous and discontinuous pressure models in case of very pervious fractures.

The continuous pressure model replaces the transmission condition (3.6) by the following phase pressure continuity conditions at mf interfaces:

$$\gamma^+ u_m^\alpha = \gamma^- u_m^\alpha = u_f^\alpha \text{ on } \Gamma, \alpha \in \{nw, w\}. \quad (3.7)$$

It results that we can denote by u^α the matrix pressure of phase $\alpha \in \{nw, w\}$ and by γu^α the fracture pressure of phase $\alpha \in \{nw, w\}$, where γ is the trace operator on Γ for functions in $H^1(\Omega)$ (Fig. 3.3).

The hybrid dimensional two-phase flow continuous pressure model looks for s_m^α , s_f^α , and u^α , $\alpha = nw, w$ satisfying (3.1)–(3.2) and (3.3)–(3.4).

For both continuous and discontinuous pressure models, a no-flux boundary conditions is prescribed at the tips of the immersed fractures, that is to say on $\partial\Gamma \setminus \partial\Omega$, and the volume conservation and pressure continuity conditions are imposed at the fracture intersections. We refer to [13, 16] for more details on those conditions.

Finally, one should provide some appropriate initial and boundary data. To fix ideas, we consider in a non homogeneous Dirichlet boundary conditions on the matrix boundary $\partial\Omega_{\text{Dir}} \subset \partial\Omega$ and on the fracture boundary $\Sigma_{\text{Dir}} \subset \partial\Gamma \cap \partial\Omega$. Homogeneous Neumann boundary conditions are set on $\partial\Omega_N = \partial\Omega \setminus \overline{\partial\Omega_{\text{Dir}}}$ and on $\Sigma_N = (\partial\Gamma \cap \partial\Omega) \setminus \overline{\Sigma_{\text{Dir}}}$.

3.3 Vertex Approximate Gradient (VAG) Discretization of Two-Phase DFM Continuous Pressure Models

The VAG discretization of hybrid dimensional two-phase Darcy flows introduced in [13] considers generalised polyhedral meshes of Ω in the spirit of [29]. Let us briefly recall some notations related to the space discretization. We denote by \mathcal{M} the set of disjoint open polyhedral cells, by \mathcal{F} the set of faces and by \mathcal{V} the set of nodes of the mesh. For each cell $K \in \mathcal{M}$ we denote by $\mathcal{F}_K \subset \mathcal{F}$ the set of its faces and by \mathcal{V}_K the set of its nodes. Similarly, we will denote by \mathcal{V}_σ the set of nodes of $\sigma \in \mathcal{F}$. The set \mathcal{M}_σ denotes the two cells sharing an interior face σ or the single cell to which the boundary face σ belongs. The set \mathcal{M}_s (resp. \mathcal{F}_s) is the subset of cells (resp. faces) sharing the node $s \in \mathcal{V}$.

Let \mathcal{E}_σ denote the set of edges of the face $\sigma \in \mathcal{F}$. It is then assumed that for each face $\sigma \in \mathcal{F}$, there exists a so-called ‘‘centre’’ of the face $\mathbf{x}_\sigma \in \sigma \setminus \bigcup_{e \in \mathcal{E}_\sigma} e$ such that $\mathbf{x}_\sigma = \sum_{s \in \mathcal{V}_\sigma} \beta_{\sigma,s} \mathbf{x}_s$, with $\sum_{s \in \mathcal{V}_\sigma} \beta_{\sigma,s} = 1$, and $\beta_{\sigma,s} \geq 0$ for all $s \in \mathcal{V}_\sigma$. The face σ is not necessarily planar, hence the term generalised polyhedral mesh. More precisely, each face σ is assumed to be defined by the union of the triangles $T_{\sigma,e}$ defined by the face centre \mathbf{x}_σ and each edge $e \in \mathcal{E}_\sigma$.

The mesh is supposed to be conforming w.r.t. the fracture network Γ in the sense that there exists a subset \mathcal{F}_Γ of \mathcal{F} such that $\bar{\Gamma} = \bigcup_{\sigma \in \mathcal{F}_\Gamma} \bar{\sigma}$. We set

$$\mathcal{V}_\Gamma = \bigcup_{\sigma \in \mathcal{F}_\Gamma} \mathcal{V}_\sigma,$$

and, for $s \in \mathcal{V}_\Gamma$, we define $\mathcal{F}_{\Gamma,s} = \mathcal{F}_s \cap \mathcal{F}_\Gamma$ as the subset of faces in \mathcal{F}_Γ sharing the node s .

The VAG discretization proposed in [13] is based upon the following set of degrees of freedom (d.o.f.)

$$\mathcal{D} = \mathcal{M} \cup \mathcal{V} \cup \mathcal{F}_\Gamma$$

and the corresponding vector space:

$$X_{\mathcal{D}} = \{v_\nu \in \mathbb{R}, \nu \in \mathcal{D}\}.$$

The d.o.f. are exhibited in Fig. 3.4 for a given cell K with one fracture face σ in bold. Let us denote by

$$\mathcal{V}_{\text{Dir}} = \{s \in \mathcal{V} \mid \mathbf{x}_s \in \bar{\partial\Omega}_{\text{Dir}} \cup \bar{\Sigma}_{\text{Dir}}\},$$

the subset of Dirichlet nodes.

A finite element discretization is built from the vector space of d.o.f. $X_{\mathcal{D}}$ using a tetrahedral sub-mesh of \mathcal{M} and a second order interpolation at the face centres \mathbf{x}_σ , $\sigma \in \mathcal{F} \setminus \mathcal{F}_\Gamma$ defined by the operator $I_\sigma : X_{\mathcal{D}} \rightarrow \mathbb{R}$ such that

$$I_\sigma(v) = \sum_{\mathbf{s} \in \mathcal{V}_\sigma} \beta_{\sigma, \mathbf{s}} v_{\mathbf{s}}.$$

The tetrahedral sub-mesh is defined by

$$\mathcal{T} = \{T_{K, \sigma, e}, e \in \mathcal{E}_\sigma, \sigma \in \mathcal{F}_K, K \in \mathcal{M}\}, \quad (3.8)$$

where $T_{K, \sigma, e}$ is the tetrahedron joining the cell centre \mathbf{x}_K to the triangle $T_{\sigma, e}$. For a given $v_{\mathcal{D}} \in X_{\mathcal{D}}$, we define the function $\pi_{\mathcal{T}} v_{\mathcal{D}}$ as the continuous piecewise affine function on each tetrahedron of \mathcal{T} such that $\pi_{\mathcal{T}} v_{\mathcal{D}}(\mathbf{x}_K) = v_K$, $\pi_{\mathcal{T}} v_{\mathcal{D}}(\mathbf{x}_{\mathbf{s}}) = v_{\mathbf{s}}$, $\pi_{\mathcal{T}} v_{\mathcal{D}}(\mathbf{x}_\sigma) = v_\sigma$, and $\pi_{\mathcal{T}} v_{\mathcal{D}}(\mathbf{x}_{\sigma'}) = I_{\sigma'}(v)$ for all $K \in \mathcal{M}$, $\mathbf{s} \in \mathcal{V}$, $\sigma \in \mathcal{F}_\Gamma$, and $\sigma' \in \mathcal{F} \setminus \mathcal{F}_\Gamma$. The nodal basis of this finite element discretization will be denoted by $\eta_K, \eta_{\mathbf{s}}, \eta_\sigma$, for $K \in \mathcal{M}$, $\mathbf{s} \in \mathcal{V}$, $\sigma \in \mathcal{F}_\Gamma$.

The VAG scheme is a control volume scheme in the sense that it results, for each d.o.f. not located at the Dirichlet boundary and each phase, in a volume balance equation. The two main ingredients are therefore the conservative fluxes and the porous volumes. The VAG matrix and fracture fluxes are exhibited in Fig. 3.4. They are derived from the variational formulation on the finite element subspace. For $u_{\mathcal{D}} \in X_{\mathcal{D}}$, the matrix fluxes $F_{K, v}(u_{\mathcal{D}})$ connect the cell $K \in \mathcal{M}$ to all the d.o.f. located at the boundary of K , namely $v \in \Xi_K = \mathcal{V}_K \cup (\mathcal{F}_K \cap \mathcal{F}_\Gamma)$. They are defined by

$$F_{K, v}(u_{\mathcal{D}}) = \int_K -\Lambda_m(\mathbf{x}) \nabla \pi_{\mathcal{T}} u_{\mathcal{D}}(\mathbf{x}) \cdot \nabla \eta_v(\mathbf{x}) d\mathbf{x} = \sum_{v' \in \Xi_K} \mathbb{T}_K^{v, v'}(u_K - u_{v'}),$$

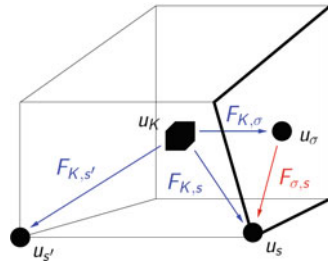
with the cell transmissibilities

$$\mathbb{T}_K^{v, v'} = \int_K \Lambda_m(\mathbf{x}) \nabla \eta_{v'}(\mathbf{x}) \cdot \nabla \eta_v(\mathbf{x}) d\mathbf{x}.$$

The fracture fluxes $F_{\sigma, \mathbf{s}}(u_{\mathcal{D}})$ connect each fracture face $\sigma \in \mathcal{F}_\Gamma$ to its nodes $\mathbf{s} \in \mathcal{V}_\sigma$ and are defined by

$$F_{\sigma, \mathbf{s}}(u_{\mathcal{D}}) = \int_\sigma -d_f \Lambda_f \nabla_\tau \gamma \pi_{\mathcal{T}} u_{\mathcal{D}}(\mathbf{x}) \cdot \nabla_\tau \gamma \eta_{\mathbf{s}}(\mathbf{x}) d\sigma(\mathbf{x}) = \sum_{\mathbf{s}' \in \mathcal{V}_\sigma} \mathbb{T}_\sigma^{\mathbf{s}, \mathbf{s}'}(u_\sigma - u_{\mathbf{s}'}),$$

Fig. 3.4 For a cell K and a fracture face σ (in bold), examples of VAG d.o.f. $u_K, u_{\mathbf{s}}, u_\sigma, u_{\mathbf{s}'}$ and VAG fluxes $F_{K, \sigma}, F_{K, \mathbf{s}}, F_{K, \mathbf{s}'}, F_{\sigma, \mathbf{s}}$



with the fracture face transmissibilities

$$\mathbb{T}_{\sigma}^{\mathbf{s},\mathbf{s}'} = \int_{\sigma} d_f(\mathbf{x}) \Lambda_f(\mathbf{x}) \nabla_{\tau} \gamma \eta_{\mathbf{s}'}(\mathbf{x}) \cdot \nabla_{\tau} \gamma \eta_{\mathbf{s}}(\mathbf{x}) d\sigma(\mathbf{x}),$$

where $d\sigma(\mathbf{x})$ denotes the Lebesgue $d - 1$ dimensional measure on Γ .

The porous volumes are obtained by distributing the porous volumes of each cell $K \in \mathcal{M}$ and fracture face $\sigma \in \mathcal{F}_{\Gamma}$ to the d.o.f. located on their respective boundaries. For each $K \in \mathcal{M}$ we define a set of non-negative volume fractions $(\alpha_{K,v})_{v \in \Xi_K \setminus \mathcal{V}_{\text{Dir}}}$ satisfying $\sum_{v \in \Xi_K \setminus \mathcal{V}_{\text{Dir}}} \alpha_{K,v} \leq 1$, and we set

$$\phi_{K,v} = \alpha_{K,v} \int_K \phi_m(\mathbf{x}) d\mathbf{x}.$$

Similarly, for all $\sigma \in \mathcal{F}_{\Gamma}$ we set

$$\phi_{\sigma,s} = \alpha_{\sigma,s} \int_{\sigma} \phi_f(\mathbf{x}) d_f(\mathbf{x}) d\sigma(\mathbf{x}),$$

with the non-negative volume fractions $(\alpha_{\sigma,s})_{s \in \mathcal{V}_{\sigma} \setminus \mathcal{V}_{\text{Dir}}}$ satisfying $\sum_{s \in \mathcal{V}_{\sigma} \setminus \mathcal{V}_{\text{Dir}}} \alpha_{\sigma,s} \leq 1$.

Then, we set for all $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_{\Gamma}$:

$$\phi_K = \int_K \phi_m(\mathbf{x}) d\mathbf{x} - \sum_{v \in \Xi_K \setminus \mathcal{V}_{\text{Dir}}} \phi_{K,v},$$

$$\phi_{\sigma} = \int_{\sigma} \phi_f(\mathbf{x}) d_f(\mathbf{x}) d\sigma(\mathbf{x}) - \sum_{s \in \mathcal{V}_{\sigma} \setminus \mathcal{V}_{\text{Dir}}} \phi_{\sigma,s}.$$

On practical meshes with cell sizes at mf interfaces much larger than the fracture width, the flexibility in the choice of the weights $\alpha_{K,s}$ and $\alpha_{\sigma,s}$ is shown in [13] (see also [30]) to be a crucial asset compared with usual CVFE approaches, allowing to improve significantly the accuracy of the scheme. As exhibited in Fig. 3.5, and in contrast with the usual CVFE approaches, the fracture porous volumes can be defined with no contribution of the matrix porous volume, thus avoiding to enlarge artificially the flow path in the fractures and to slow down the front speed. This is achieved by choosing the volume fractions such that

$$\begin{aligned} \alpha_{K,\sigma} &= 0 \quad \text{for all } \sigma \in \mathcal{F}_{\Gamma}, K \in \mathcal{M}_{\sigma}, \\ \alpha_{K,s} &= 0 \quad \text{for all } s \in \mathcal{V}_{\Gamma}, K \in \mathcal{M}_s. \end{aligned}$$

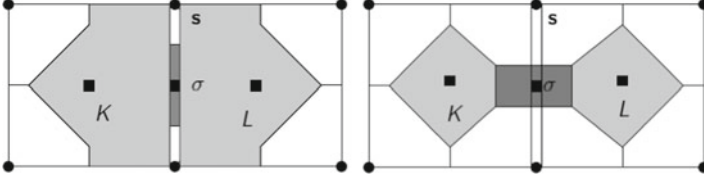


Fig. 3.5 Example of control volumes at cells, fracture face, and nodes, in the case of two cells K and L splitted by one fracture face σ (the width of the fracture has been enlarged in this Figure). (left): VAG choice of the porous volumes avoiding mixing between fracture and matrix porous volumes. (right): CVFE like choice of the porous volumes mixing fracture and matrix porous volumes leading to a considerable enlargement of the fracture drain on practical meshes

3.3.1 VAG Phase Potential Two-Point (TP) Upwind Formulation

We consider in the following of Sect. 3.3, the usual approach (termed f-upwind model) for which a single rock type is assigned to each d.o.f. Quite naturally, the fracture rock type is associated with d.o.f. located on Γ , while the matrix rock type is associated to the remaining d.o.f., that is we set

$$P_{c,v}(s) = \begin{cases} P_{c,m}(s) & \text{if } v \notin (\mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma), \\ P_{c,f}(s) & \text{if } v \in (\mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma), \end{cases}$$

and

$$M_v^\alpha(s) = \begin{cases} M_m^\alpha(s) & \text{if } v \notin (\mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma), \\ M_f^\alpha(s) & \text{if } v \in (\mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma), \end{cases} \quad \alpha \in \{nw, w\}.$$

The set of discrete unknowns is defined by the set of phase pressure $u_{\mathcal{D}}^\alpha \in X_{\mathcal{D}}$ and phase saturation $s_{\mathcal{D}}^\alpha \in X_{\mathcal{D}}$ for each phase $\alpha \in \{nw, w\}$.

The “single” phase VAG Darcy fluxes, not including the phase mobility, are defined, for each phase $\alpha \in \{nw, w\}$, by

$$F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha) = F_{K,v}(u_{\mathcal{D}}^\alpha) + \rho^\alpha g F_{K,v}(z_{\mathcal{D}}),$$

$$F_{\sigma,s}^\alpha(u_{\mathcal{D}}^\alpha) = F_{\sigma,s}(u_{\mathcal{D}}^\alpha) + \rho^\alpha g F_{\sigma,s}(z_{\mathcal{D}}),$$

with $z_{\mathcal{D}} = (\mathbf{x}_v)_{v \in \mathcal{D}}$, and for $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_\Gamma$, $v \in \Xi_K$, $\mathbf{s} \in \mathcal{V}_\sigma$. They are combined with the usual Two-Point (TP) phase potential upwinding of the mobilities [8, 23], leading to the following two-phase Darcy VAG fluxes

$$q_{K,v}^\alpha = M_K^\alpha(s_K^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_v^\alpha(s_v^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^-,$$

$$q_{\sigma,s}^\alpha = M_\sigma^\alpha(s_\sigma^\alpha)(F_{\sigma,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_s^\alpha(s_s^\alpha)(F_{\sigma,s}^\alpha(u_{\mathcal{D}}^\alpha))^-.$$

Let us define the accumulation terms by

$$\left\{ \begin{array}{ll} \mathcal{A}_K^\alpha = \phi_K s_K^\alpha, & K \in \mathcal{M}, \\ \mathcal{A}_\sigma^\alpha = (\phi_\sigma + \sum_{K \in \mathcal{M}_\sigma} \phi_{K,\sigma}) s_\sigma^\alpha, & \sigma \in \mathcal{F}_\Gamma, \\ \mathcal{A}_s^\alpha = (\sum_{K \in \mathcal{M}_s} \phi_{K,s} + \sum_{\sigma \in \mathcal{F}_{\Gamma,s}} \phi_{\sigma,s}) s_s^\alpha, & s \in \mathcal{V} \setminus \mathcal{V}_{\text{Dir}}. \end{array} \right.$$

Note that neither the accumulation terms $\mathcal{A}_\sigma^\alpha$ and \mathcal{A}_s^α nor the mf fluxes take into account the discontinuity of the saturations across mf interface. In other terms, the discrete problem does not involve quantities such as $P_{c,m}(s_f)$. An alternative approach is described in Sect. 3.3.5.

For $N \in \mathbb{N}^*$, let us consider the time discretization $t^0 = 0 < t^1 < \dots < t^{n-1} < t^n \dots < t^N = T$ of the time interval $[0, T]$. We denote the time steps by $\Delta t^n = t^n - t^{n-1}$ for all $n = 1, \dots, N$. The superscript n will be used to denote the unknowns at time t^n . To reduce the amount of notation, only the previous time step superscript $n - 1$ will be specified in the following, while the superscript n will not be specified by default.

The set of discrete equations couples the volume balance equations at each d.o.f. excluding the Dirichlet nodes

$$\left\{ \begin{array}{l} \frac{\mathcal{A}_K^\alpha - \mathcal{A}_K^{\alpha,n-1}}{\Delta t^n} + \sum_{v \in \Xi_K} q_{K,v}^\alpha = 0, \quad K \in \mathcal{M}, \quad \alpha = nw, w, \\ \frac{\mathcal{A}_\sigma^\alpha - \mathcal{A}_\sigma^{\alpha,n-1}}{\Delta t^n} + \sum_{s \in \mathcal{V}_\sigma} q_{\sigma,s}^\alpha - \sum_{K \in \mathcal{M}_\sigma} q_{K,\sigma}^\alpha = 0, \quad \sigma \in \mathcal{F}_\Gamma, \quad \alpha = nw, w, \\ \frac{\mathcal{A}_s^\alpha - \mathcal{A}_s^{\alpha,n-1}}{\Delta t^n} + \sum_{K \in \mathcal{M}_s} -q_{K,s}^\alpha + \sum_{\sigma \in \mathcal{F}_{\Gamma,s}} -q_{\sigma,s}^\alpha = 0, \quad s \in \mathcal{V} \setminus \mathcal{V}_{\text{Dir}}, \quad \alpha = nw, w, \end{array} \right. \quad (3.9)$$

combined with the closure laws

$$\left\{ \begin{array}{l} s_v^{nw} + s_v^w = 1, \quad v \in \mathcal{D}, \\ u_v^{nw} - u_v^w = P_{c,v}(s_v^{nw}), \quad v \in \mathcal{D}, \end{array} \right. \quad (3.10)$$

and the Dirichlet boundary conditions

$$s_s^{nw} = s_{\text{Dir},s}^{nw} \quad u_s^{nw} = u_{\text{Dir},s}^{nw}, \quad s \in \mathcal{V}_{\text{Dir}}, \quad (3.11)$$

for given $s_{\text{Dir},s}^{nw} \in [0, 1]$, $u_{\text{Dir},s}^{nw}$, $s \in \mathcal{V}_{\text{Dir}}$.

To solve the discrete nonlinear system (3.9), one first uses the closure equations (3.10) to eliminate the unknowns s_v^w and u_v^w for $v \in \mathcal{D}$ reducing the system to the primary unknowns u_v^{nw} , s_v^{nw} , $v \in \mathcal{D}$ coupled by the set of equations (3.9) and the Dirichlet boundary conditions (3.11). A Newton's method is used to solve this nonlinear system at each time step of the simulation. At each Newton step, the Jacobian

matrix is assembled and the cell unknowns $u_K^{nw}, s_K^{nw}, K \in \mathcal{M}$ are eliminated without any fill-in using the linearized cell volume balance equations reducing the system to the node and fracture face primary unknowns only. This elimination results in a huge gain in terms of system size in particular for tetrahedral meshes. The reduced linear system is solved using a Krylov subspace solver preconditioned by a CPR-AMG preconditioner. This preconditioner combines multiplicatively an AMG preconditioner on a pressure block (elliptic part of the system) with a zero fill-in incomplete factorization of the full system. Let us refer to [42, 47] for its detailed description. In the following numerical experiments, the pressure block is simply obtain as the sum over both phases of the volume balance equations on each fracture face and non Dirichlet node.

3.3.2 What Is Wrong with Two-Point Upwinding at *mf* Interfaces

In this Section, we discuss one particular difficulty that the nodal discretizations have in regard of the discrete fluxes reconstruction. As shown below, due to the dual control volumes at *mf* interfaces, nodal schemes may result in fluxes having an opposite sign compared to the fluxes computed at the physical *mf* interfaces. Using Two-Point upwinding, this results in an artificial diffusion of the saturation toward an upstream direction. To avoid this drawback we propose below an alternative Multi-Point upwinding technique.

For a given constant velocity \mathbf{q} , let us choose $u_{\mathcal{D}} \in X_{\mathcal{D}}$ such that $u_v = -\Lambda_m^{-1} \mathbf{q} \cdot \mathbf{x}_v$ for all $v \in \mathcal{D}$. From $-\Lambda_m \nabla \pi_{\mathcal{T}} u_{\mathcal{D}} = \mathbf{q}$, we obtain

$$F_{K,s}(u_{\mathcal{D}}) = \mathbf{q} \cdot \int_K \nabla \eta_s(\mathbf{x}) d\mathbf{x},$$

at a given fracture node $s \in \mathcal{V}_{\Gamma}$, and cell $K \in \mathcal{M}_s$.

For the sake of simplicity, let us assume the geometrical configuration illustrated in Fig. 3.6. It results that

$$\begin{aligned} F_{K,s}(u_{\mathcal{D}}) &= -\frac{1}{2} |\mathbf{s}_2 \sigma_1| \mathbf{q} \cdot \mathbf{n}_1, \\ F_{J,s}(u_{\mathcal{D}}) &= -\frac{1}{2} |\mathbf{s}_3 \sigma_2| \mathbf{q} \cdot \mathbf{n}_3 = -F_{K,s}(u_{\mathcal{D}}), \\ F_{I,s}(u_{\mathcal{D}}) &= +\frac{1}{2} |\mathbf{s}_2 \mathbf{s}_3| \mathbf{q} \cdot \mathbf{n}, \\ F_{K,\sigma_1}(u_{\mathcal{D}}) &= +\frac{1}{2} |\mathbf{s} \mathbf{s}_1| \mathbf{q} \cdot \mathbf{n}, \\ F_{J,\sigma_2}(u_{\mathcal{D}}) &= +\frac{1}{2} |\mathbf{s} \mathbf{s}_4| \mathbf{q} \cdot \mathbf{n} = F_{I,\sigma_1}(u_{\mathcal{D}}). \end{aligned}$$

We remark that, whatever the velocity \mathbf{q} , either the flux $F_{K,s}(u_{\mathcal{D}})$ or the flux $F_{J,s}(u_{\mathcal{D}})$ have the opposite sign as the one of $\mathbf{q} \cdot \mathbf{n}$. Assuming, to fix ideas that $\mathbf{q} \cdot \mathbf{n} > 0$, it results from the Two-Point upwinding used for the transport scheme of a given phase with Darcy velocity \mathbf{q} , that the phase propagates from the fracture either to the upstream cell K or the upstream cell J .

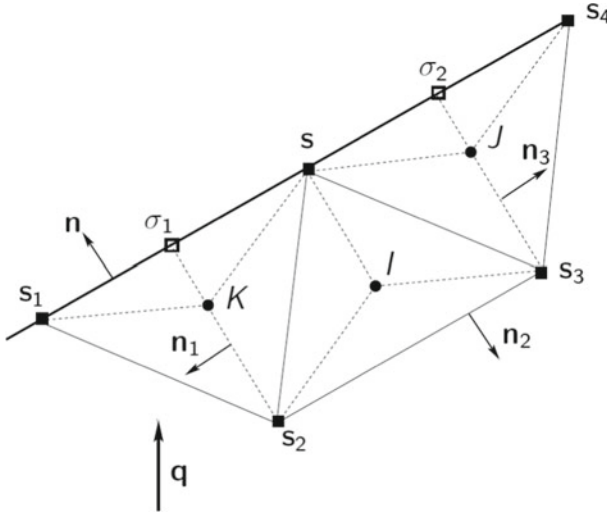


Fig. 3.6 Example of a 2D mesh with three isosceles triangular cells at the interface with a fracture in bold. It is assumed that the unit normal vectors are such that $\mathbf{n}_2 = -\mathbf{n}$, $\mathbf{n}_1 = -\mathbf{n}_3$ and that $\mathbf{n}_1 \cdot \mathbf{n} = 0$. The cell centers are chosen as the isobarycenters of their 3 nodes

On the other hand, let us remark that the ill-orientated discrete fluxes cancel out when summing over the cells connected to the node s and located on the same side with respect to the planar fracture, that is we have

$$F_{K,s}(u_{\mathcal{D}}) + F_{I,s}(u_{\mathcal{D}}) + F_{J,s}(u_{\mathcal{D}}) = \frac{1}{2}|\sigma_1\sigma_2|\mathbf{q} \cdot \mathbf{n}. \quad (3.12)$$

This property actually holds for an arbitrary number of polygonal cells sharing the node s and whatever the choice of the cell centers. In the three-dimensional case, this property also holds for tetrahedral meshes.

In the following Subsection, this property on the sum of the fluxes is exploited to avoid the artificial diffusion of the phase toward an upstream direction.

3.3.3 Multi-Point (MP) Upwind Fluxes at *mf* Interfaces

We first define an equivalence relation on each subset \mathcal{M}_s of cells, for any fixed node $s \in \mathcal{V}$, by

$$K \equiv_{\mathcal{M}_s} L \iff \text{there exists } n \in \mathbb{N} \text{ and a sequence } (\sigma_i)_{i=1,\dots,n} \text{ in } \mathcal{F}_s \setminus \mathcal{F}_\Gamma, \\ \text{such that } K \in \mathcal{M}_{\sigma_1}, L \in \mathcal{M}_{\sigma_n} \text{ and } \mathcal{M}_{\sigma_{i+1}} \cap \mathcal{M}_{\sigma_i} \neq \emptyset \\ \text{for } i = 1, \dots, n-1.$$

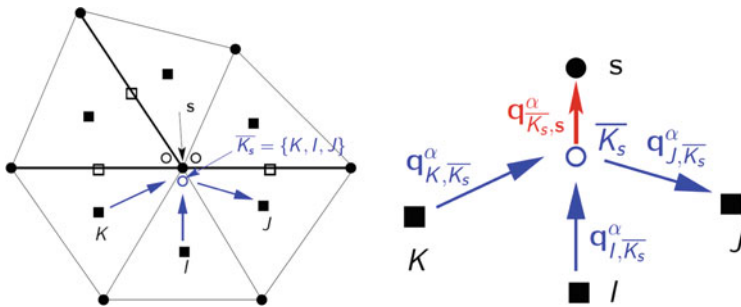


Fig. 3.7 (left) 2D mesh with 3 fracture faces in bold and the 3 d.o.f. in $\bar{\mathcal{M}}_s$ at the node $s \in \mathcal{V}_\Gamma$, (right) Darcy fluxes joining each cell $L \in \bar{\mathcal{K}}_s$ to the new d.o.f. $\bar{\mathcal{K}}_s$, and joining the new d.o.f. $\bar{\mathcal{K}}_s$ to the node s (the node s and $\bar{\mathcal{K}}_s$ are located at the same point s but they have been separated for the sake of clarity of the Figure)

Let us then denote by $\bar{\mathcal{M}}_s$ the set of all classes of equivalence of \mathcal{M}_s and by $\bar{\mathcal{K}}_s$ the element of $\bar{\mathcal{M}}_s$ containing $K \in \mathcal{M}$. Obviously $\bar{\mathcal{M}}_s$ might have more than one element only if $s \in \mathcal{V}_\Gamma$. Note that $\bar{\mathcal{K}}_s$ is both considered as a subset of cells of \mathcal{M}_s as well as an additional d.o.f. located at the same point than the node s i.e. we set $\mathbf{x}_{\bar{\mathcal{K}}_s} = \mathbf{x}_s$ (Fig. 3.7).

Let us define the phase mobilities

$$M_{\bar{\mathcal{K}}_s}^\alpha, \quad s \in \mathcal{V}_\Gamma, \quad \bar{\mathcal{K}}_s \in \bar{\mathcal{M}}_s, \quad \alpha \in \{nw, w\}, \quad (3.13)$$

as matrix fracture additional unknowns. Then, using phase potential upwinding of the mobilities, let us define for all $L \in \bar{\mathcal{K}}_s$ the half Darcy fluxes between L and $\bar{\mathcal{K}}_s$ by

$$q_{L, \bar{\mathcal{K}}_s}^\alpha = M_L^\alpha (s_L^\alpha) (F_{L,s}^\alpha (u_D^\alpha))^+ + M_{\bar{\mathcal{K}}_s}^\alpha (F_{L,s}^\alpha (u_D^\alpha))^-,$$

as well as the half Darcy flux between $\bar{\mathcal{K}}_s$ and s by

$$q_{\bar{\mathcal{K}}_s, s}^\alpha = M_s^\alpha (s_s^\alpha) (F_{\bar{\mathcal{K}}_s, s}^\alpha)^- + M_{\bar{\mathcal{K}}_s}^\alpha (F_{\bar{\mathcal{K}}_s, s}^\alpha)^+,$$

where we set by flux conservation

$$F_{\bar{\mathcal{K}}_s, s}^\alpha = \sum_{L \in \bar{\mathcal{K}}_s} F_{L, s}^\alpha (u_D^\alpha).$$

The flux continuity equation

$$q_{\bar{\mathcal{K}}_s, s}^\alpha = \sum_{L \in \bar{\mathcal{K}}_s} q_{L, \bar{\mathcal{K}}_s}^\alpha, \quad (3.14)$$

is used to eliminate the mobility unknown $M_{\overline{K}_s}^\alpha$ leading to the following convex linear combination of the cells $L \in \overline{K}_s$ and node s mobilities:

$$M_{\overline{K}_s}^\alpha = \frac{\sum_{L \in \overline{K}_s} (F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ M_L^\alpha(s_L^\alpha) - M_s^\alpha(s_s^\alpha) (\sum_{L \in \overline{K}_s} F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^-}{\sum_{L \in \overline{K}_s} (F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ - (\sum_{L \in \overline{K}_s} F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^-}. \quad (3.15)$$

We deduce the definition of the new Multi-Point upwind flux

$$q_{K,s}^\alpha = q_{K,\overline{K}_s}^\alpha = M_K^\alpha(s_K^\alpha)(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_{\overline{K}_s}^\alpha(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^-,$$

denoted by VAG MP in the following and to be used in the conservation equations (3.9). Compared with the Two-Point upwind flux

$$q_{K,s}^\alpha = M_K^\alpha(s_K^\alpha)(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_s^\alpha(s_s^\alpha)(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^-,$$

denoted by VAG TP, the VAG MP flux uses the fracture node saturation s_s^α only if $\sum_{L \in \overline{K}_s} F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha) < 0$, which, in view of (3.12), ensures that the phase will not go out from the fracture on the wrong side in the case of a linear phase pressure field.

Note also that, if \overline{K}_s contains only one cell, both the VAG TP and VAG MP fluxes match, this is why the fluxes $q_{K,\sigma}^\alpha$, $\sigma \in \mathcal{F}_K \cap \mathcal{F}_\Gamma$ do not need to be modified.

The matrix fracture mobility unknowns (3.13) and flux continuity equations (3.14) can be kept in the nonlinear system and solve simultaneously with the other unknowns and equations. Let us recall that the CPR-AMG preconditioner combines multiplicatively an AMG preconditioner on a pressure block (elliptic part of the system) with a zero fill-in incomplete factorization of the full system. The matrix fracture mobility unknowns $M_{\overline{K}_s}^\alpha$ and the flux continuity equations (3.14), $s \in \mathcal{V}_\Gamma$, $\overline{K}_s \in \overline{\mathcal{M}}_s$, are not included in the definition of the pressure block due to their hyperbolic nature. It results that the pressure block has the same number of unknowns and sparsity pattern as the one of the usual VAG TP scheme. Since the AMG step is the most expensive part of the CPR-AMG two stage preconditioner, this explains why keeping the matrix fracture mobility unknowns is quite efficient.

On the other hand, the elimination of the matrix fracture mobility unknowns together with the flux continuity equations in (3.15) leads to a rather large fill-in of the Jacobian (depending on the density of the fracture network) and also prevents the elimination of the cell unknowns connected to the fractures. The following numerical experiments confirm that it is much more efficient in terms of CPU time to keep the matrix fracture mobility unknowns in the linear system.

3.3.4 Numerical Experiments

The Objectives of this Subsection is to compare the solutions obtained with the following schemes:

- the CVFE like VAG scheme with rock type mixture and Two-Point upwinding of the mobilities at mf interfaces (VAG CVFE),
- the VAG scheme with no rock type mixture and Two-Point upwinding of the mobilities at mf interfaces (VAG TP),
- the VAG scheme with no rock type mixture and Multi-Point upwinding of the mobilities at mf interfaces (VAG MP). The VAG MP scheme is implemented either with elimination of the interface mobility unknowns (VAG MP) or without elimination of these unknowns (VAG MP no elim).
- the Hybrid Finite Volume (HFV) scheme with cell, face and fracture edge unknowns as described in [37] (HFV).

All these schemes are implemented in the same code using the Fortran 90 programming language combined with the gfortran compiler. The linear systems are solved using the Slatec library [48] for the GMRes iterative solver and the ILU0 preconditioner as well as the AMG1R5 library for the Algebraic MultiGrid preconditioner [45].

Tables 3.1 and 3.2 exhibit the following entries:

- mesh: number of cells,
- dof : number of degrees of freedom of each scheme (with 2 physical primary unknowns per d.o.f.),
- dof_{lin} : number of degrees of freedom in the linear system after reduction. Let us recall that the cell unknowns are eliminated for VAG CVFE, VAG TP, HFV, and VAG MP no elim, while the interface mobilities together with the cell unknowns not connected to the fractures are eliminated for VAG MP,
- N_z : number of nonzero elements in the reduced Jacobian (with 2×2 matrix elements).

Note that for the VAG MP no elim implementation, the pressure block is stored separately after reduction with a lower number of d.o.f. and nonzero elements than the remaining part of the Jacobian. This is a key point to lower the CPU time when the CPR-AMG preconditioner is used. Then, the first dof_{lin} (resp. N_z) entry corresponds to the pressure block, and the second entry to the remaining part.

These tables also include the following entries:

- $N_{\Delta t}$: number of successful time steps,
- N_{chop} : number of time step chops,
- N_{Newton} : average number of Newton iterations per successful time step,
- N_{GMRes} : average number of GMRes iteration per Newton step,
- CPU (s): CPU time in seconds.

The CPU time takes into account the full time loop including the outputs in ensight format files at each time step but excluding the preprocessing computations (mesh reading, mesh connectivity, VAG transmissibilities, CSR format of the Jacobian) which are negligible in terms of CPU time compared with the time loop.

3.3.4.1 Tracer DFM Model with a Single Fracture

Let us denote by (x, y) the Cartesian coordinates of \mathbf{x} and let us set $\Omega = (0, 1 \text{ m})^2$, $\mathbf{x}_1 = (0, \frac{1}{4})$, $\mathbf{x}_2 = (1, 0.875)$. We consider a single fracture defined by $\Gamma = (\mathbf{x}_1, \mathbf{x}_2)$ with tangential permeability $\Lambda_f = 200 \text{ m}^2$ and width $d_f = 10^{-3} \text{ m}$. The matrix permeability is isotropic and set to $\Lambda_m = 1 \text{ m}^2$. The matrix and fracture porosities are set to $\phi_m = \phi_f = 1$. Let us set

$$\mathbf{t} = \begin{pmatrix} 1 \\ 0.625 \end{pmatrix}, \quad \mathbf{q} = \begin{pmatrix} 1 \\ \frac{1}{3} \end{pmatrix}.$$

We consider the hybrid-dimensional tracer model obtained from the two-phase DFM model by setting $M_m^\alpha(s) = M_f^\alpha(s) = s$ for $\alpha \in \{nw, w\}$, $P_{c,m}(s) = P_{c,f}(s) = 0$, $g = 0$. The pressure analytical solution is defined for $\alpha \in \{nw, w\}$ by

$$u^\alpha(\mathbf{x}, t) = 1 - \Lambda_m^{-1} \mathbf{x} \cdot \mathbf{q},$$

leading to the matrix Darcy velocity

$$\mathbf{q}_m^\alpha = \mathbf{q},$$

and the tangential fracture velocity integrated over the width

$$\mathbf{q}_f^\alpha = d_f \Lambda_f \frac{(\mathbf{t} \cdot \Lambda_m^{-1} \mathbf{q})}{|\mathbf{t}|^2} \mathbf{t},$$

This pressure solution is exactly solved by the VAG scheme using Dirichlet condition at the boundary of the domain. An input Dirichlet boundary condition is imposed for the non-wetting phase saturation (tracer) with zero value at the matrix boundary and a value of 1 at the fracture boundary \mathbf{x}_1 . The initial condition is defined by a zero non-wetting phase saturation both in the fracture and matrix domains. Figure 3.8 illustrates that the tracer VAG TP solution goes out on the wrong side of the fracture on a few layers of cells, while it is not the case for the HFV and VAG MP solutions as expected. The VAG CVFE stationary tracer solution is not plotted since it is the same than the VAG TP stationary tracer solution. Figure 3.9 exhibits the stationary solutions along the fracture showing that the HFV and VAG MP solutions match on both meshes while the VAG TP solution is not fully converged even on the fine mesh.

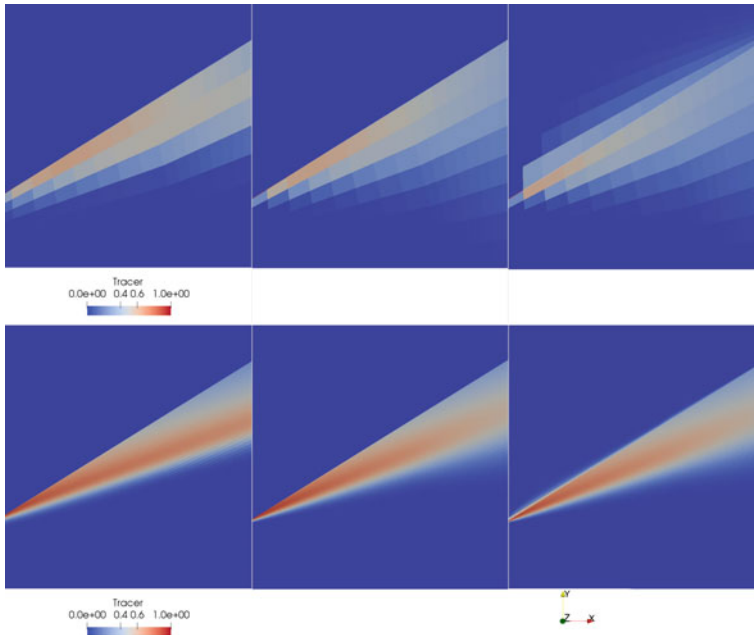


Fig. 3.8 Stationary solution for the non-wetting phase saturation (tracer) in the matrix and in the fracture obtained by, from left to right, the HVF, VAG MP and VAG TP schemes, and, from top to bottom, on the 16×16 and 128×128 topologically Cartesian meshes

Fig. 3.9 Stationary non-wetting phase saturation along the fracture as a function of x obtained by the HVF, VAG MP, VAG TP schemes on the 16×16 and 128×128 topologically Cartesian meshes

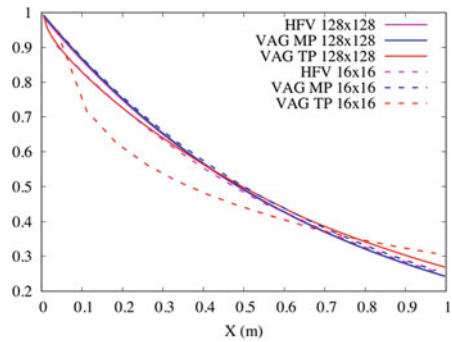
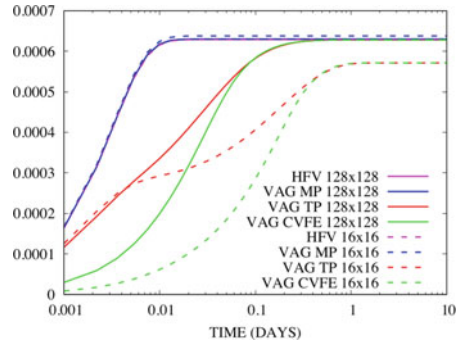


Figure 3.10 exhibits the tracer volume in the fracture as a function of time. Again, the HVF and VAG MP solutions match on both meshes, while the tracer front in the fracture is clearly slow down for the VAG TP solution on both meshes. This is much worse for the VAG CVFE solution due to the fracture enlargement resulting from the rock type mixture at mf interfaces.

Fig. 3.10 Volume of the non-wetting phase in the fracture as a function of time for the HVF, VAG MP, VAG TP, VAG CVFE scheme solutions on the 16×16 and 128×128 topologically Cartesian meshes



3.3.4.2 Large 2D DFM Model

This test case considers the DFM model with the matrix domain $\Omega = (0, 100 \text{ m}) \times (0, 186.5 \text{ m})$ and a fracture network including 581 connected components both exhibited in Fig. 3.11. The fracture width is $d_f = 1 \text{ cm}$ and the fracture network is homogeneous and isotropic with $\Lambda_f = 10^{-11} \text{ m}^2$, $\phi_f = 0.2$. The matrix is homogeneous and isotropic with $\Lambda_m = 10^{-14} \text{ m}^2$, $\phi_m = 0.4$.

The relative permeabilities are given by $k_{r,f}^\alpha(s^\alpha) = s^\alpha$ and $k_{r,m}^\alpha(s^\alpha) = (s^\alpha)^2$, $\alpha \in \{nw, w\}$ and the capillary pressure is fixed to $P_{c,m}(s^{nw}) = -10^4 \ln(1 - s^{nw}) \text{ Pa}$ in the matrix and to $P_{c,f}(s^{nw}) = 0 \text{ Pa}$ in the fracture network. The fluid properties are defined by their dynamic viscosities $\mu^{nw} = 5 \cdot 10^{-3}$, $\mu^w = 10^{-3} \text{ Pa s}$ and their mass densities $\rho^w = 1000$ and $\rho^{nw} = 700 \text{ kg m}^{-3}$.

The reservoir is initially saturated with the wetting phase. Dirichlet boundary conditions are imposed at the top boundary with a wetting phase pressure of 1 MPa and $s_m^w = 1$, as well as at the bottom boundary with $s_m^{nw} = 0.9$ and $u^w = 4 \text{ MPa}$. The remaining boundaries are assumed impervious and the final simulation time is fixed to $t_f = 1800 \text{ days}$.

The time stepping is defined by $\Delta t^1 = \Delta t_{init} = 10 \text{ days}$, and for all $n \geq 1$ by

$$\Delta t^{n+1} = \max(\Delta t_{max}, 1.2\Delta t^n) \text{ with } \Delta t_{max} = 10 \text{ days}, \quad (3.16)$$

in case of a successful time step Δt^n , and $\Delta t^{n+1} = \frac{\Delta t^n}{2}$, in case of non convergence of the Newton algorithm in $Newton_{max} = 30$ iterations. This last value is chosen not too small to avoid too many time step failures even on the finest mesh but also not too large to avoid increased CPU time in case of time step failures induced by residual oscillations.

The criterion of convergence for the Newton algorithm is based on a relative residual in l_1 norm smaller than Res_{max} or on a Newton step in l_∞ norm (scaled by 10^{-6} for the primary pressure unknown and by 1 for the other primary unknown) smaller than dx_{max} with

$$Res_{max} = 10^{-5}, \quad dx_{max} = 10^{-4}. \quad (3.17)$$

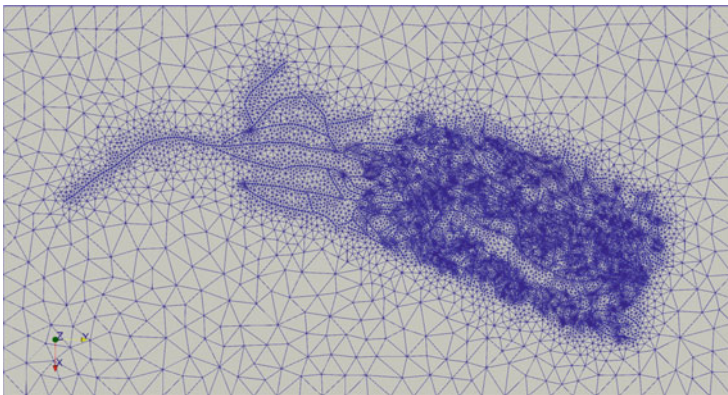


Fig. 3.11 Triangular mesh of the DFM model with 32340 (32k) cells and 5344 fracture faces (Courtesy of M. Karimi-Fard, Stanford, and A. Lapène, Total). This mesh is refined uniformly to obtain the 129k and 517k cells meshes

Note also that the Newton step is relaxed such that its l_∞ norm (scaled by 10^{-6} for the primary pressure unknown and by 1 for the other primary unknown) is smaller than dx_{obj} with

$$dx_{\text{obj}} = 1. \quad (3.18)$$

The non-wetting phase saturation is exhibited at final simulation time in Fig. 3.12 in the matrix and in the fracture network, and the volume of the non-wetting phase as a function of time is presented in Fig. 3.13. We clearly see in Figs. 3.12 and 3.13 that the VAG CVFE discretization considerably slows down the non-wetting phase front in the fracture network due to the drain enlargement induced by the mixing of matrix and fracture porous volumes at mf interfaces. The VAG TP discretization does a better job but still underestimates the front speed in the fracture network. As clearly exhibited by Fig. 3.12, this is due to the fact that the VAG TP scheme propagates the non-wetting phase on the wrong side of the fractures as explained in Sect. 3.3.2. From Fig. 3.13, the VAG TP solution gets very close to the VAG MP solution after two level of refinement of the coarse mesh, while the VAG CVFE solution has not yet converged on the finest mesh. The comparison between the VAG MP and HFV solutions shows that they are in good agreement for all meshes. It appears in Fig. 3.13 that the HFV scheme converges more slowly than the VAG MP scheme.

The numerical behavior of the four schemes is reported in Table 3.1 with CPU time is in seconds on Intel E5-2670 2.6GHz. We remark that the average number of Newton iterations is in all cases quite smaller than $Newton_{\text{max}}$ due to significant variations in the number of Newton iterations during the simulation. This can be explained typically by a higher number of Newton iterations when the non-wetting phase reaches the tips of the fracture network.

For this large 2D network, the VAG MP implementation with elimination of the matrix fracture mobilities leads to a twice large CPU time than the VAG MP

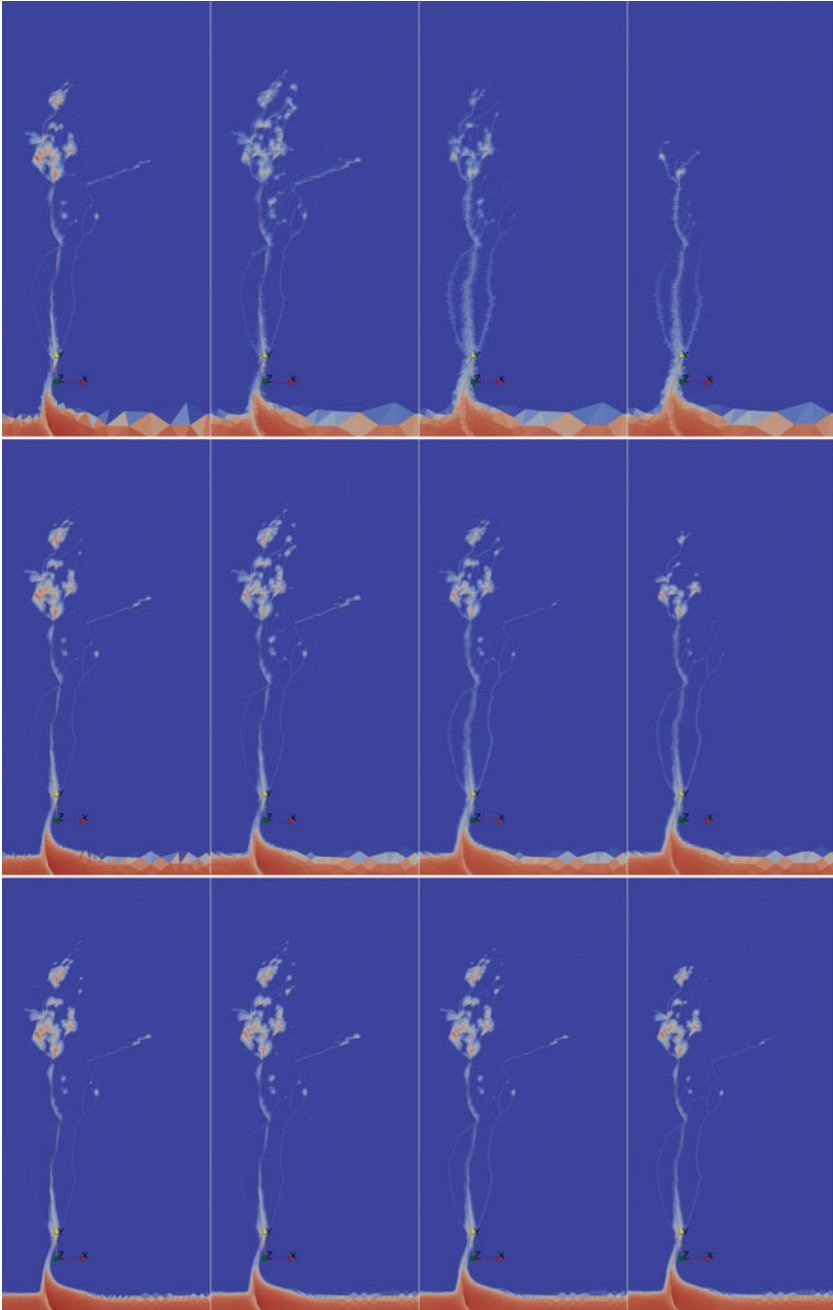


Fig. 3.12 Non-wetting phase saturation in the matrix and fracture network at time $t_f = 1800$ days for the HVF, VAG MP, VAG TP, VAG CVFE schemes from left to right, and the 32k, 129k, 517k cells meshes from top to bottom

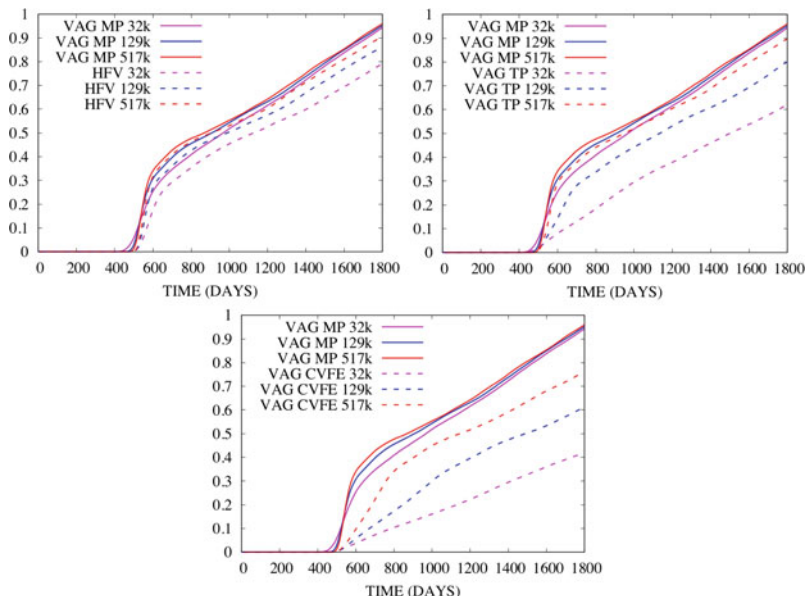


Fig. 3.13 Non-wetting phase volume in the fracture network as a function of time for the VAG MP, HFV, VAG TP and VAG CFVE schemes on the 3 meshes of sizes 32k, 129k and 517k cells

implementation with no elimination. Regarding the comparison between VAG MP and VAG TP, we notice a twice larger CPU time, which is a rather good result for such a large network. The comparison between HFV and VAG MP shows for this 2D test case that HFV is competitive on the coarse mesh due to the additional matrix fracture unknowns for VAG MP, but becomes more expensive on the two refined meshes. We will see in the next test case that the situation is much more in favor of the VAG schemes on tetrahedral 3D meshes.

3.3.4.3 3D DFM Model

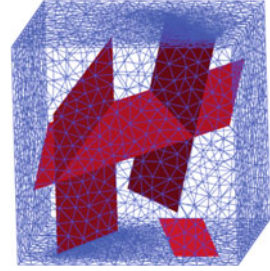
The DFM model of matrix domain $\Omega = (0, 100 \text{ m})^3$ and its coarsest tetrahedral mesh conforming to the fracture network are illustrated in Fig. 3.14. The fracture network is assumed to be of constant aperture $d_f = 1 \text{ cm}$. The matrix and fracture porosities, permeabilities, relative permeabilities and capillary pressures are the same as in the previous test case. The fluid properties are also the same than in the previous test case.

At initial time, the reservoir is fully saturated with the wetting phase. Then, non-wetting phase is injected from below, which is managed by imposing Dirichlet conditions at the bottom and at the top of the reservoir. We impose at the bottom boundary either an overpressure $\Delta p = 2 \text{ MPa}$ or no overpressure $\Delta p = 0 \text{ MPa}$ w.r.t. the hydrostatic distribution of the water pressure. The remaining boundaries are assumed

Table 3.1 Numerical behavior of the simulation for the 2D DFM test case on the 3 meshes and for the HFV, VAG CVFE, VAG TP and VAG MP schemes. The VAG MP scheme is implemented either with elimination (VAG MP) or without elimination (VAG MP no elim) of the mf interface mobility unknowns. We refer to the beginning of Sect. 3.3.4 for the description of the entries

Scheme	mesh	dof	dof _{lin}	N_z	$N_{\Delta t}$	N_{chop}	N_{Newton}	N_{GMRes}	CPU (s)
HFV	32k	87k	54k	280k	180	0	5.9	16.0	465
VAG CVFE	32k	54k	21k	161k	180	0	3.0	8.1	165
VAG TP	32k	54k	21k	161k	180	0	3.7	14.3	224
VAG MP	32k	65k	43k	459k	188	4	5.2	19.5	794
VAG MP no elim.	32k	65k	21/32k	161/356k	180	0	4.8	20.4	491
HFV	129k	335k	205k	1045k	180	0	8.5	56	7747
VAG CVFE	129k	205k	75k	549k	180	0	4.8	26	1301
VAG TP	129k	205k	75k	549k	180	0	6.7	55	2995
VAG MP	129k	226k	122k	1143k	261	37	9.3	54	11880
VAG MP no elim.	129k	226k	75/96k	549/916k	182	1	8.0	64	5417
HFV	517k	1315k	798k	4.0M	182	1	12.6	104	145704
VAG CVFE	517k	798k	280k	2.0M	180	0	9.7	59	25403
VAG TP	517k	798k	280k	2.0M	180	0	10.9	103	50472
VAG MP no elim	517k	840k	280/322k	2.0/2.7M	199	12	12.8	122	98390

Fig. 3.14 Geometry of the domain $\Omega = 100 \text{ m} \times 100 \text{ m} \times 100 \text{ m}$ with the fracture network in red (left), coarsest tetrahedral mesh with 47670 cells (right)



impervious and the final simulation time is fixed to $t_f = 360$ days for $\Delta p = 2$ MPa and to $t_f = 3600$ days for $\Delta p = 0$ MPa. The time stepping is defined as in (3.16) using $\Delta t_{init} = 0.1$ days, $Newton_{max} = 30$, and either $\Delta t_{max} = 10$ days for $\Delta p = 2$ MPa or $\Delta t_{max} = 100$ days for $\Delta p = 0$ MPa. The criterion of convergence for the Newton algorithm is defined as in (3.17) with $Res_{max} = 10^{-6}$ and $dx_{max} = 10^{-5}$, and the relaxation of the Newton step is controlled as in (3.18) by the parameter $dx_{obj} = 1$.

From Figs. 3.15 and 3.16, we observe that the VAG TP and VAG CVFE schemes are far from convergence even on the finest mesh with 450k cells while the solution provided by the VAG MP scheme is quite close to the one of the HFV scheme. The discrepancy between, on the one hand, the VAG TP and VAG CVFE, and, on the other hand, the VAG MP and HFV schemes is even more striking on the coarse mesh for the no-overpressure gravity dominant test case exhibited in Figs. 3.17 and 3.18. In terms

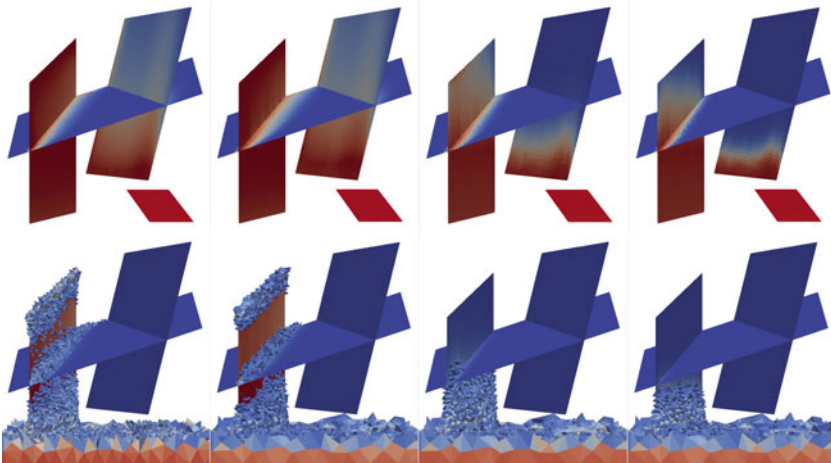


Fig. 3.15 Non-wetting phase saturation solutions obtained with the HFV, VAG MP, VAG TP, VAG CVFE schemes from left to right, at time $t_f = 360$ days (top), and at time $t = 100$ days (bottom), with overpressure $\Delta p = 2$ MPa, and the mesh of size 450k cells. The threshold in the matrix is $S_m^{nw} > 0.1$ (bottom)

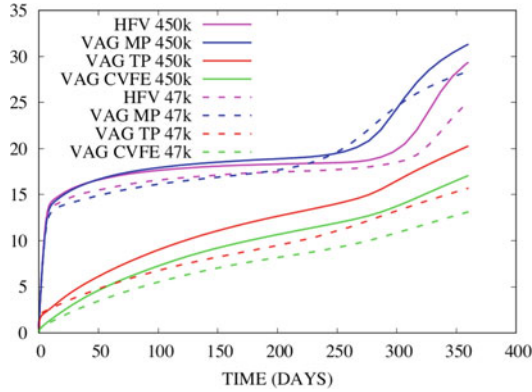


Fig. 3.16 Non-wetting phase volume in the fracture network as a function of time for the 3D DFM test case with the overpressure $\Delta p = 2$ MPa using the VAG MP, HFV, VAG TP and VAG CFVE schemes on the 2 meshes of sizes 47k and 450k cells

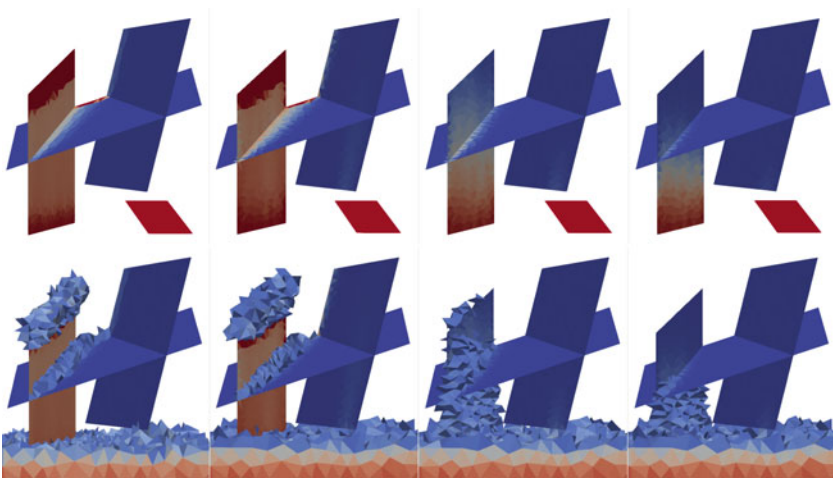
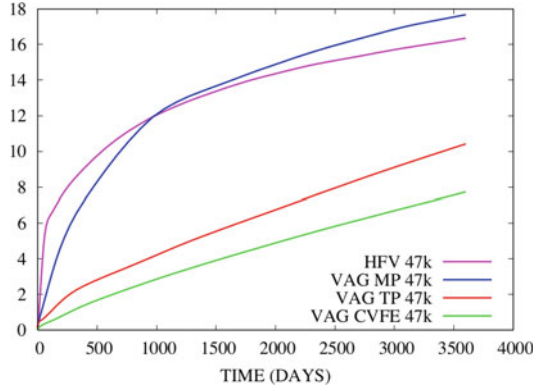


Fig. 3.17 Non-wetting phase saturation solutions obtained with the HVE, VAG MP, VAG TP, VAG CVFE schemes from left to right, at time $t_f = 3600$ days with no overpressure $\Delta p = 0$ MPa, and the mesh of size 47k cells. The threshold in the matrix is $S_m^{nw} > 0.1$ (bottom)

of CPU time, as exhibited in Table 3.2, the VAG MP scheme implemented with no elimination of the matrix fracture mobility unknowns is competitive compared with the VAG TP scheme. It is also much cheaper than the HFV scheme which leads to a much larger number of d.o.f. and requires both more Newton and GMRes iterations than the VAG schemes. Note that the HFV scheme cannot be run in a reasonable CPU time for the finest mesh of size 1600k cells.

Fig. 3.18 Non-wetting phase volume in the fracture network as a function of time for the 3D DFM test case with no overpressure $\Delta p = 0$ MPa using the VAG MP, HFV, VAG TP and VAG CFVE schemes on the 47k cells mesh



3.3.5 Capturing the Saturation Jumps at mf Interfaces

Given cellwise and fracture facewise constant rock types, the idea introduced in [20, 43, 44] for CVFE methods and in [14, 17, 31] for the VAG scheme is to define as many saturations as rock types shared at a given node or fracture face. This allows to capture the saturation jumps at rock type interfaces resulting from the continuity of the capillary pressure in the graphical sense [18, 21, 22, 27, 28].

The choice of the primary unknowns may greatly affect the convergence of Newton's method used to solve the nonlinear system at each time step of the simulation. For the cells and the nodal d.o.f. associated with a single rock the choice of the primary unknowns does not change compared to Sect. 3.3.1. That is we use the non-wetting phase's pressure and saturation as pair of primary unknowns. In contrast the d.o.f. located at rock type interfaces require a special treatment. For such d.o.f. $v \in \mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma$ we set again the pressure of the non-wetting phase as the first primary unknown, while the second primary unknown is chosen based on the variable switching strategy introduced in [14]. For a given rock type $rt \in \mathcal{RT} = \{m, f\}$ let $\tilde{P}_{c,rt}$ denote the monotone graph extension of $P_{c,rt}$ as introduced in [21, 22]. For each subset $\chi \in \{\{m\}, \{m, f\}\}$ of \mathcal{RT} , non-decreasing continuous functions

$$\begin{cases} P_{c,\chi}(\tau), \\ S_{\chi,rt}^{nw}(\tau), \text{ for all } rt \in \chi, \end{cases} \quad (3.19)$$

are built such that

$$P_{c,\chi}(\tau) \in \tilde{P}_{c,rt}(S_{\chi,rt}^{nw}(\tau)), \text{ for all } \tau \text{ and } rt \in \chi,$$

and such that $P_{c,\chi}(\tau) + \sum_{rt \in \chi} S_{\chi,rt}^{nw}(\tau)$ is strictly increasing. Then, we set

$$S_{\chi,rt}^w(\tau) = 1 - S_{\chi,rt}^{nw}(\tau).$$

The variable τ is going to be used as the second primary unknown.

Table 3.2 Numerical behavior of the simulation for the 3D DFM test case with the overpressure $\Delta p = 2$ MPa on the three meshes of sizes 47k, 450k and 1600k cells. The VAG MP scheme is implemented either with elimination (VAG MP) or without elimination (VAG MP no elim) of the mf interface mobility unknowns. We refer to the beginning of Sect. 3.3.4 for the description of the entries

Scheme	mesh	dof	dof _{lin}	N_z	$N_{\Delta t}$	N_{chop}	N_{Newton}	N_{GMRes}	CPU (s)
HFV	47k	147k	98k	682k	57	0	4.0	28.5	862
VAG CVFE	47k	58k	9k	131k	57	0	2.91	9.0	97
VAG TP	47k	58k	9k	131k	57	0	2.92	10.6	104
VAG MP	47k	60k	23k	590k	57	0	3.07	13.8	254
VAG MP no elim.	47k	60k	9/11k	211/131k	57	0	3.03	13.5	137
HFV	450k	1357k	920k	6.4M	57	0	8.0	99	57900
VAG CVFE	450k	535k	81k	1.2M	57	0	3.63	14.4	1601
VAG TP	450k	535k	81k	1.2M	57	0	3.82	19.3	1750
VAG MP	450k	547k	152k	3.7M	57	0	3.86	24.2	3320
VAG MP no elim.	450k	547k	1.2/1.7M	450k	57	0	3.86	24.8	2306
HFV	1600k	4812k	3217k	22.6M	x	x	x	x	Too long
VAG CVFE	1600k	1866k	274k	4.2M	57	0	3.95	20.4	6994
VAG TP	1600k	1866k	274k	4.2M	57	0	4.42	26.7	8230
VAG MP no elim.	1600k	1896k	274/304k	4.2/5.3M	57	0	4.35	34	10270

The main advantage of this framework, which applies to an arbitrary number of rock types, is to incorporate in the construction of the functions (3.19) the saturation jump condition at different rock type interfaces and to apply to general capillary pressure functions. In practice, we use $\tau = s^{nw}$ for $\chi = \{m\}$ and the parametrization defined in [14] for $\chi = \{m, f\}$. This parametrization is based on a generalization of variable switch approaches (see also [43]) between s_f^{nw} , s_m^{nw} , p_c and applies to general, including non invertible, capillary functions (see numerical section for an example and Fig. 3.20).

Let us set

$$\begin{cases} \text{rt}_K = m, & K \in \mathcal{M}, \\ \text{rt}_\sigma = f, & \sigma \in \mathcal{F}_\Gamma, \end{cases} \quad \begin{cases} \chi_\nu = \{m\}, & \nu \in \mathcal{M} \cup (\mathcal{V} \setminus \mathcal{V}_\Gamma), \\ \chi_\nu = \{m, f\}, & \nu \in \mathcal{V}_\Gamma \cup \mathcal{F}_\Gamma. \end{cases}$$

Using the above framework, given the primary unknowns $u_D^{nw} = (u_\nu^{nw})_{\nu \in \mathcal{D}}$ and $\tau_D = (\tau_\nu)_{\nu \in \mathcal{D}}$, we set $u_D^w = (u_\nu^w)_{\nu \in \mathcal{D}}$ with $u_\nu^w = u_\nu^{nw} - P_{c,\chi_\nu}(\tau_\nu)$ for all d.o.f. $\nu \in \mathcal{D}$, and we define the discrete values of the saturation as follows. For all cells $K \in \mathcal{M}$ and the nodes $\mathbf{s} \in \mathcal{V} \setminus \mathcal{V}_\Gamma$ associated with the single matrix rock type, we set

$$\begin{aligned} s_K^\alpha &= S_{\chi_K, \text{rt}_K}^\alpha(\tau_K) = S_{\{m\}, m}^\alpha(\tau_K) = \tau_K \\ s_{K, \mathbf{s}}^\alpha &= S_{\chi_{\mathbf{s}}, \text{rt}_K}^\alpha(\tau_{\mathbf{s}}) = S_{\{m\}, m}^\alpha(\tau_{\mathbf{s}}) = \tau_{\mathbf{s}}, \quad K \in \mathcal{M}_s. \end{aligned}$$

For the fracture faces $\sigma \in \mathcal{F}_\Gamma$, we set

$$\begin{aligned} s_\sigma^\alpha &= S_{\chi_\sigma, \text{rt}_\sigma}^\alpha(\tau_\sigma) = S_{\{m, f\}, f}^\alpha(\tau_\sigma) \\ s_{K, \sigma}^\alpha &= S_{\chi_\sigma, \text{rt}_K}^\alpha(\tau_\sigma) = S_{\{m, f\}, m}^\alpha(\tau_\sigma), \quad K \in \mathcal{M}_\sigma. \end{aligned}$$

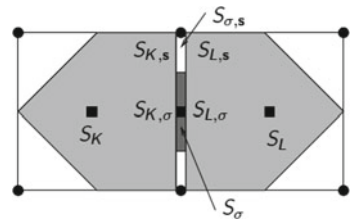
For the nodes $\mathbf{s} \in \mathcal{V}_\Gamma$, located at the mf interface, we set

$$\begin{cases} s_{K, \mathbf{s}}^\alpha = S_{\chi_{\mathbf{s}}, \text{rt}_K}^\alpha(\tau_{\mathbf{s}}) = S_{\{m, f\}, m}^\alpha(\tau_{\mathbf{s}}), \\ s_{\sigma, \mathbf{s}}^\alpha = S_{\chi_{\mathbf{s}}, \text{rt}_\sigma}^\alpha(\tau_{\mathbf{s}}) = S_{\{m, f\}, f}^\alpha(\tau_{\mathbf{s}}), \quad \sigma \in \mathcal{F}_{\Gamma, \mathbf{s}}. \end{cases}$$

As exhibited in Fig. 3.19, the above definition of the saturations at the mf interfaces takes into account the jump of the saturations induced by the different rock types.

Let us remark that, in our specific example, since the matrix domain is homogeneous in terms of capillary pressure-saturation relation, the variables $s_{K, \mathbf{s}}^\alpha$, $K \in \mathcal{M}_s$ (resp. $s_{K, \sigma}^\alpha$, $K \in \mathcal{M}_\sigma$) refer to the same nodal (resp. facial) saturation values. Sim-

Fig. 3.19 Saturations inside the cells K and L , the fracture face σ and at the mf interfaces taking into account the saturation jumps induced by the different rock types



ilarly, the values $s_{\sigma,s}^\alpha$, $\sigma \in \mathcal{F}_{\Gamma,s}$ are identical. This is however not true for general heterogeneous matrix and fracture domains.

We define the accumulation terms by

$$\left\{ \begin{array}{ll} \mathcal{A}_K^\alpha = \phi_K s_K^\alpha, & K \in \mathcal{M}, \\ \mathcal{A}_\sigma^\alpha = \phi_\sigma s_\sigma^\alpha + \sum_{K \in \mathcal{M}_\sigma} \phi_{K,\sigma} s_{K,\sigma}^\alpha, & \sigma \in \mathcal{F}_\Gamma, \\ \mathcal{A}_s^\alpha = \sum_{K \in \mathcal{M}_s} \phi_{K,s} s_{K,s}^\alpha + \sum_{\sigma \in \mathcal{F}_{\Gamma,s}} \phi_{\sigma,s} s_{\sigma,s}^\alpha, & s \in \mathcal{V} \setminus \mathcal{V}_{\text{Dir}}, \end{array} \right.$$

and the VAG fluxes with TP phase potential upwinding of the mobilities by

$$\begin{aligned} q_{K,v}^\alpha &= M_{\text{rt}_K}^\alpha(s_K^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_K}^\alpha(s_{K,v}^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^- , \\ q_{\sigma,s}^\alpha &= M_{\text{rt}_\sigma}^\alpha(s_\sigma^\alpha)(F_{\sigma,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_\sigma}^\alpha(s_{\sigma,s}^\alpha)(F_{\sigma,s}^\alpha(u_{\mathcal{D}}^\alpha))^- , \end{aligned}$$

for all $\alpha \in \{nw, w\}$ and $K \in \mathcal{M}$, $\sigma \in \mathcal{F}_\Gamma$, $v \in \Xi_K$, $s \in \mathcal{V}_\sigma$.

The VAG TP discretization capturing the saturation jumps at rock type interfaces looks for $u_{\mathcal{D}}^{nw}$ and $\tau_{\mathcal{D}}$ satisfying the conservation equations (3.9) together with the Dirichlet boundary conditions

$$\tau_s = \tau_{\text{Dir},s} \quad u_s^{nw} = u_{\text{Dir},s}^{nw}, \quad s \in \mathcal{V}_{\text{Dir}}. \quad (3.20)$$

It will be termed VAG TP m-upwind discretization in the following. The VAG MP m-upwind discretization can also be defined as previously using the MP upwind flux

$$q_{K,s}^\alpha = q_{K,\overline{K}_s}^\alpha = M_{\text{rt}_K}^\alpha(s_K^\alpha)(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_{\overline{K}_s}^\alpha(F_{K,s}^\alpha(u_{\mathcal{D}}^\alpha))^- ,$$

for $s \in \mathcal{V}_\Gamma$ with the interface mobility

$$M_{\overline{K}_s}^\alpha = \frac{\sum_{L \in \overline{K}_s} (F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ M_{\text{rt}_L}^\alpha(s_L^\alpha) - M_{\text{rt}_K}^\alpha(s_{K,s}^\alpha) (\sum_{L \in \overline{K}_s} F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^-}{\sum_{L \in \overline{K}_s} (F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^+ - (\sum_{L \in \overline{K}_s} F_{L,s}^\alpha(u_{\mathcal{D}}^\alpha))^-},$$

assuming that $\text{rt}_L = \text{rt}_K$ for all $L \in \overline{K}_s$. This assumption can always be verified by setting new interface face(s) between the different rock types in \overline{K}_s . This discretization will be termed VAG MP m-upwind discretization in the following.

A comparison of the f-upwind and m-upwind models with reference equidimensional solutions can be found in [1, 16]. Basically, it concludes that, thanks to the saturation jump capturing at mf interfaces, the m-upwind model provides a better approximation than the f-upwind model as long as the fractures are not fully

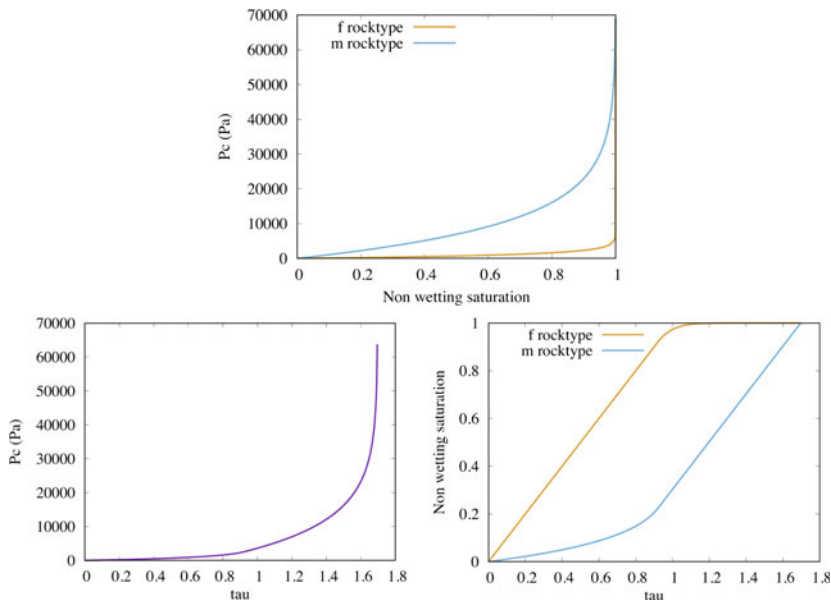


Fig. 3.20 (Top): capillary pressure as a function of the non-wetting phase saturation for both the fracture (f) and matrix (m) rock types with $b_m = 10^4$ and $b_f = 10^3$ Pa. (Bottom): capillary pressure and fracture and matrix non-wetting phase saturations as functions of the parameter $\tau \in [0, \tau_2)$

filled with the non-wetting phase. When the fractures are filled, the m-upwind model overestimates the fracture capillary pressure and underestimates the capillary barrier effect. In that case the f-upwind model provides a better approximation.

In the following numerical section, the VAG TP and MP m-upwind discretizations are compared both in terms of solutions and CPU times.

3.3.5.1 Numerical Experiments

In this subsection, we compare the m-upwind version of the VAG TP and VAG MP schemes using the same code implementation as described in the beginning of Sect. 3.3.4. The test case considers the large DFM model exhibited in Fig. 3.21 with domain $\Omega = (0, 85) \times (0, 60) \times (0, 140)$ m kindly provided by the authors of the Benchmark [11, 12].

The fracture width is $d_f = 1$ cm and the fracture network is homogeneous and isotropic with $\Lambda_f = 10^{-11}$ m², $\phi_f = 0.2$. The matrix is homogeneous and isotropic with $\Lambda_m = 10^{-14}$ m², $\phi_m = 0.4$.

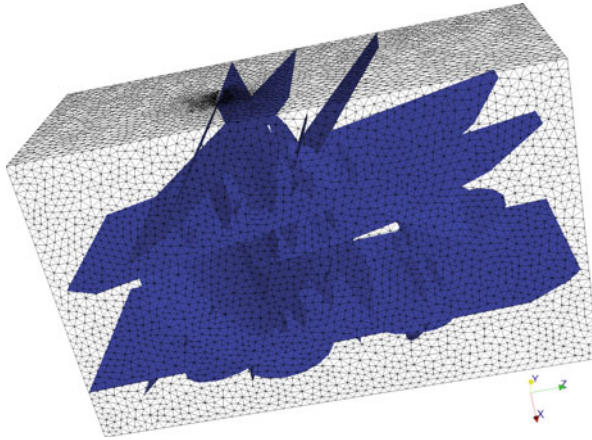


Fig. 3.21 Large DFM model with its mesh of size 495233 tetrahedral cells and 66908 fracture faces provided by the authors of the Benchmark [11]

The relative permeabilities are given by $k_{r,f}^\alpha(s_f^\alpha) = s_f^\alpha$ and $k_{r,m}^\alpha(s_m^\alpha) = (s_m^\alpha)^2$, $\alpha \in \{nw, w\}$ and the capillary pressure is fixed to $P_{c,m}(s_m^{nw}) = -b_m \ln(1 - s_m^{nw})$ Pa in the matrix and to $P_{c,f}(s_f^{nw}) = -b_f \ln(1 - s_f^{nw})$ Pa in the fracture network, with $b_f = 10^3$ Pa, and $b_m = 10^4$ Pa. The fluid properties are defined by their dynamic viscosities $\mu^{nw} = 5 \cdot 10^{-3}$, $\mu^w = 10^{-3}$ Pa s and their mass densities $\rho^w = 1000$ and $\rho^{nw} = 700$ kg m $^{-3}$.

The parametrization τ at mf interfaces introduced in [14] is recalled below and illustrated in Fig. 3.20 for the convenience of the reader.

$$S_{\{m,f\},f}^{nw}(\tau) = \begin{cases} \tau, & \tau \in [0, \tau_1), \\ 1 - (\tau_1 + (1 - \tau_1)^{\frac{b_f}{b_m}} - \tau)^{\frac{b_m}{b_f}}, & \tau \in [\tau_1, \tau_2), \end{cases} \quad (3.21)$$

$$S_{\{m,f\},m}^{nw}(\tau) = \begin{cases} 1 - (1 - \tau)^{\frac{b_f}{b_m}}, & \tau \in [0, \tau_1), \\ \tau - \tau_1 + 1 - (1 - \tau_1)^{\frac{b_f}{b_m}}, & \tau \in [\tau_1, \tau_2), \end{cases} \quad (3.22)$$

and

$$P_{c,\{m,f\}}(\tau) = \begin{cases} -b_f \ln(1 - \tau), & \tau \in [0, \tau_1), \\ -b_m \ln\left(\tau_1 + (1 - \tau_1)^{\frac{b_f}{b_m}} - \tau\right), & \tau \in [\tau_1, \tau_2), \end{cases} \quad (3.23)$$

where $\tau_1 = 1 - \left(\frac{b_f}{b_m}\right)^{\frac{b_m}{b_m - b_f}}$ and $\tau_2 = \tau_1 + (1 - \tau_1)^{\frac{b_f}{b_m}}$.

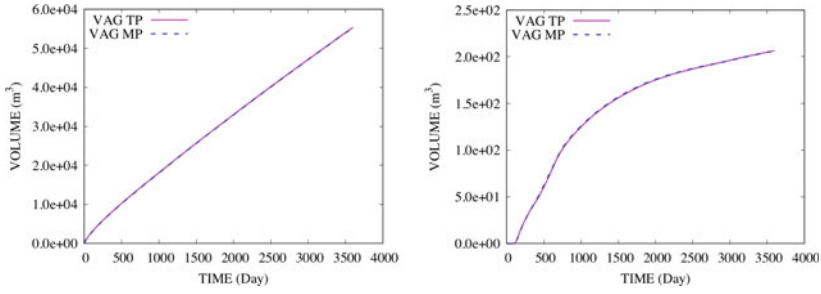


Fig. 3.22 Non-wetting phase saturation volumes in the matrix (left) and in the fracture network (right) as a function of time obtained for the VAG TP and the VAG MP m-upwind schemes

The reservoir is initially saturated with the wetting phase. Output Dirichlet boundary conditions are imposed at the boundary $\{0, 85\} \times (0, 20) \times (110, 140)$ with a wetting phase pressure of 1 MPa and $s_m^w = 1$, and input Dirichlet boundary conditions are set at the boundary $\{0\} \times (40, 60) \times (0, 30) \cup (0, 30) \times (40, 85) \times \{0\}$ with $s_m^{nw} = 0.9$ and $u^w = 4$ MPa. The remaining boundaries are assumed impervious and the final simulation time is fixed to $t_f = 3600$ days. The time stepping is defined as in (3.16) using $\Delta t_{init} = 0.01$ days, $\Delta t_{max} = 100$ days and $Newton_{max} = 25$. The criterion of convergence for the Newton algorithm is defined as in (3.17) with $Res_{max} = 10^{-5}$ and $dx_{max} = 10^{-4}$, and the relaxation of the Newton step is controlled as in (3.18) by the parameter $dx_{obj} = 1$.

The same issue at mf interfaces as for the VAG TP f-upwind approximation can be noticed in Fig. 3.23 for the VAG TP m-upwind discretization in the sense that the non-wetting phase can go out from the fractures on the wrong side for the VAG TP approximation. Nevertheless, thanks to the rather large saturation jump captured by the m-upwind model in this test case, it involves small amounts of the non-wetting phase and does not have visible effects on overall quantities (see Fig. 3.22) nor on the non-wetting phase saturation front (see Fig. 3.23). In terms of CPU time, as exhibited in Table 3.3, a factor of roughly 1.7 is observed in favor of the TP discretization due to the additional mf interface unknowns on this rather large fracture network and to the slightly larger number of Newton iterations for the MP scheme.

Let us refer to [13] for a numerical comparison between the m-upwind VAG TP scheme and the m-upwind VAG CVFE scheme (i.e. without adaptive distribution of the porous volumes at mf interfaces). It shows that the m-upwind VAG CVFE scheme still slows down the transport in the fractures in particular for a high matrix fracture permeability ratio.

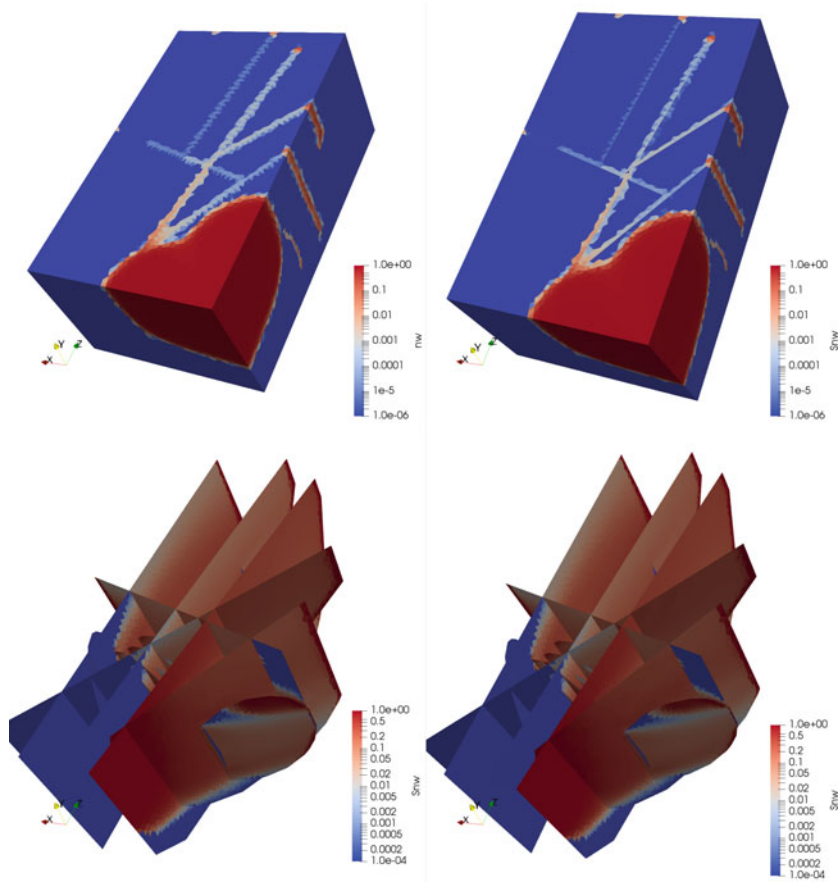


Fig. 3.23 Non-wetting phase saturation in the matrix (top) and in the fracture network (bottom) obtained for the VAG TP (left) and VAG MP (right) m-upwind schemes at time $t = 350$ days

Table 3.3 Numerical behavior of the simulation field test case for the VAG TP and MP m-upwind schemes. We refer to the beginning of Sect. 3.3.4 for the description of the entries

Scheme	mesh	dof	dof _{lin}	N_z	$N_{\Delta t}$	N_{chop}	N_{Newton}	N_{GMRes}	CPU (s)
VAG TP	495k	648k	150k	2.0M	80	4	6.8	28	8200
VAG MP	495k	718k	150/220k	2.0/4.8M	79	3	8.2	30	14190

3.4 Vertex Approximate Gradient (VAG) Discretization of Two-Phase DFM Discontinuous Pressure Models

Discontinuous pressure models are required to account for fractures acting as barriers. Such barriers are usually induced by a low fracture normal permeability combined with a capillary barrier effect. Note that even in the case of a high normal fracture permeability, a barrier behavior can still be observed for a given phase due to the degeneracy of the phase mobility when the fracture is filled by the other phase (see [1, 16]). Compared to the single phase flow models the possibility of such capillary barriers constitutes an additional motivation for the use of discontinuous pressure models.

VAG discrete unknowns: as exhibited in Fig. 3.24, the discrete unknowns are defined by the matrix d.o.f.

$$\mathcal{D}_m = \mathcal{M} \cup \{\bar{K}_s \mid \bar{K}_s \in \bar{\mathcal{M}}_s, \mathbf{s} \in \mathcal{V} \setminus \mathcal{V}_\Gamma\} \cup \mathcal{D}_{mf}$$

and by the fracture d.o.f.

$$\mathcal{D}_f = \mathcal{F}_\Gamma \cup \mathcal{V}_\Gamma,$$

where $\mathcal{D}_{mf} \subset \mathcal{D}_m$ are the mf interface d.o.f.

$$\mathcal{D}_{mf} = \bar{\mathcal{M}}_\Gamma \cup \bar{\mathcal{F}}_\Gamma,$$

with

$$\bar{\mathcal{M}}_\Gamma = \{\bar{K}_s \mid \bar{K}_s \in \bar{\mathcal{M}}_s, \mathbf{s} \in \mathcal{V}_\Gamma\}, \quad \bar{\mathcal{F}}_\Gamma = \{K_\sigma \mid K \in \mathcal{M}_\sigma, \sigma \in \mathcal{F}_\Gamma\}.$$

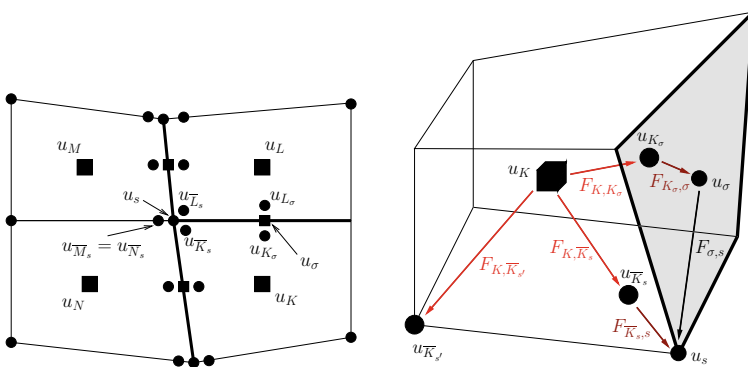


Fig. 3.24 Single phase VAG discretization of the discontinuous pressure hybrid-dimensional model: example of discrete unknowns in 2D with 3 fracture faces intersecting at node \mathbf{s} (left), and VAG fluxes (matrix fluxes in red, fracture fluxes in black and matrix fracture fluxes in dark red) in a 3D cell K with a fracture face σ in bold (right)

Let us set $\mathcal{D} = \mathcal{D}_m \cup \mathcal{D}_f$ and let us remark that for $\mathbf{s} \in \mathcal{V} \setminus \mathcal{V}_\Gamma$, $\overline{\mathcal{M}}_{\mathbf{s}}$ is reduced to the set of cells around \mathbf{s} and the d.o.f. $\overline{K}_{\mathbf{s}} \in \overline{\mathcal{M}}_{\mathbf{s}}$ is considered to match with the node \mathbf{s} .

For each cell $K \in \mathcal{M}$, let us also define the following subset of d.o.f. located at the boundary of the cell:

$$\Xi_K = \{\overline{K}_{\mathbf{s}}, \mathbf{s} \in \mathcal{V}_K, K_\sigma, \sigma \in \mathcal{F}_\Gamma \cap \mathcal{F}_K\}.$$

The subset of Dirichlet d.o.f. is denoted by $\mathcal{D}_{\text{Dir}} \subset \mathcal{D}$.

As in Sect. 3.3.5, the definition of the primary and secondary unknowns at the d.o.f. located at the rock type interfaces is based on the parametrization of the capillary pressure graphs (3.19). To fix ideas, we assume the presence of 3 rock types $\mathcal{RT} = \{m, f_d, f_b\}$ where f_d is a fracture drain rock type and f_b is a fracture barrier rock type while m denote again the matrix rock type. Let the fracture network Γ be partitioned into the networks Γ_d of fractures acting as drains and the network Γ_b of fractures acting as barriers. In order to simplify the presentation of the numerical scheme, we will assume that $\overline{\Gamma}_d \cap \overline{\Gamma}_b = \emptyset$. Then, the collection χ of rock types associated with any given d.o.f. take values in

$$\{\{m\}, \{f_d\}, \{f_b\}, \{m, f_d\}, \{m, f_b\}\}.$$

corresponding to assume no intersections between fractures acting as drain and barrier. In practice, we use the parametrization $\tau = s^{nw}$ for $\chi = \{m\}, \{f_d\}, \{f_b\}$ and the parametrizations defined in [14] for $\chi = \{m, f_d\}$ or $\{m, f_b\}$. More precisely, let us set for $\star = b, d$

$$\left\{ \begin{array}{l} \text{rt}_K = m, \quad K \in \mathcal{M}, \\ \text{rt}_\sigma = f_\star, \quad \{\sigma \in \mathcal{F}_\Gamma \mid \mathbf{x}_\sigma \in \Gamma_\star\}, \end{array} \right\} \left\{ \begin{array}{l} \chi_v = \{m\}, \quad v \in \mathcal{D}_m \setminus \mathcal{D}_{mf}, \\ \chi_v = \{f_\star\}, \quad \{v \in \mathcal{D}_f \mid \mathbf{x}_v \in \overline{\Gamma}_\star\} := \mathcal{D}_f^\star, \\ \chi_v = \{m, f_\star\}, \quad \{v \in \mathcal{D}_{mf} \mid \mathbf{x}_v \in \overline{\Gamma}_\star\} := \mathcal{D}_{mf}^\star, \end{array} \right.$$

Using the above framework, given the primary unknowns $u_{\mathcal{D}}^{nw} = (u_v^{nw})_{v \in \mathcal{D}}$ and $\tau_{\mathcal{D}} = (\tau_v)_{v \in \mathcal{D}}$, we set $u_{\mathcal{D}}^w = (u_v^w)_{v \in \mathcal{D}}$ with $u_v^w = u_v^{nw} - P_{c, \chi_v}(\tau_v)$ for all d.o.f. $v \in \mathcal{D}$, and we define the discrete values of the saturation as follows. For all d.o.f. associated with a single rock type, that is $K \in \mathcal{M}$ and $\sigma \in \mathcal{F}_\Gamma$ we set

$$s_K^\alpha = S_{\chi_K, \text{rt}_K}^\alpha(\tau_K) = \tau_K, \quad s_\sigma^\alpha = S_{\chi_\sigma, \text{rt}_\sigma}^\alpha(\tau_\sigma) = \tau_K,$$

for all $v \in \Xi_K \cap \mathcal{D}_m \setminus \mathcal{D}_{mf}$, $K \in \mathcal{M}$, we set

$$s_{K,v}^\alpha = S_{\chi_v, \text{rt}_K}^\alpha(\tau_v) = S_{\{m\}, m}^\alpha(\tau_v) = \tau_v,$$

for all $\mathbf{s} \in \mathcal{V}_s$, $\sigma \in \mathcal{F}_\Gamma$, we set

$$s_{\sigma, \mathbf{s}}^\alpha = S_{\chi_{\mathbf{s}}, \text{rt}_\sigma}^\alpha(\tau_{\mathbf{s}}) = \tau_{\mathbf{s}},$$

while for all mf interface d.o.f. from \mathcal{D}_{mf} , and with $\star = b, d$, we impose

$$\begin{cases} s_{K,v}^\alpha = S_{\chi_{v,rt_K}}^\alpha(\tau_v) = S_{\{m,f_\star\},m}^\alpha(\tau_v), & v \in \Xi_K \cap \mathcal{D}_{mf}^\star, K \in \mathcal{M}, \\ s_{\sigma,K_\sigma}^\alpha = S_{\chi_{K_\sigma,rt_\sigma}}^\alpha(\tau_{K_\sigma}) = S_{\{m,f_\star\},f_d}^\alpha(\tau_{K_\sigma}), & K_\sigma \in \overline{\mathcal{F}}_\Gamma \cap \mathcal{D}_{mf}^\star, \\ s_{\sigma,\overline{K}_s}^\alpha = S_{\chi_{\overline{K}_s,rt_\sigma}}^\alpha(\tau_{\overline{K}_s}) = S_{\chi_{\{m,f_\star\},f_d}}^\alpha(\tau_{\overline{K}_s}), & \overline{K}_s \in \overline{\mathcal{M}}_\Gamma \cap \mathcal{D}_{mf}^\star, \sigma \in \mathcal{F}_{\Gamma,\overline{K}_s}, \end{cases}$$

where $\mathcal{F}_{\Gamma,\overline{K}_s} = \mathcal{F}_{\Gamma,s} \cap (\bigcup_{K \in \overline{K}_s} \mathcal{F}_K)$.

Discrete fluxes: the VAG fluxes connect each cell K (resp. each fracture face σ) to its boundary d.o.f. $v \in \Xi_K$ (resp. $\mathbf{s} \in \mathcal{V}_\sigma$) using the same transmissibility coefficients as for the continuous pressure model

$$F_{K,v}(u_{\mathcal{D}_m}) = \sum_{v' \in \Xi_K} \mathbb{T}_K^{v,v'}(u_K - u_{v'}), \quad F_{\sigma,\mathbf{s}}(u_{\mathcal{D}_f}) = \sum_{\mathbf{s}' \in \mathcal{V}_\sigma} \mathbb{T}_\sigma^{\mathbf{s},\mathbf{s}'}(u_\sigma - u_{\mathbf{s}'}).$$

Additionally, two-point matrix fracture fluxes are defined by

$$F_{\overline{K}_s,\mathbf{s}}(u_{\overline{K}_s}, u_{\mathbf{s}}) = T_{\overline{K}_s,\mathbf{s}}(u_{\overline{K}_s} - u_{\mathbf{s}}), \quad F_{K_\sigma,\sigma}(u_{K_\sigma}, u_\sigma) = T_{K_\sigma,\sigma}(u_{K_\sigma} - u_\sigma),$$

for $\mathbf{s} \in \mathcal{V}_\Gamma$, $\overline{K}_s \in \overline{\mathcal{M}}_s$ and $\sigma \in \mathcal{F}_\Gamma$, $K \in \mathcal{M}_\sigma$, with

$$T_{\overline{K}_s,\mathbf{s}} = \frac{1}{3} \sum_{T \in \Delta | \mathbf{s} \in \overline{T}} \int_T \frac{2\lambda_{f,n}}{df} d\sigma(\mathbf{x}), \quad T_{K_\sigma,\sigma} = \int_\sigma \frac{2\lambda_{f,n}}{df} d\sigma(\mathbf{x}),$$

where Δ is the triangular submesh of Γ defined as the trace on Γ of the tetrahedral submesh \mathcal{T} introduced in (3.8) (see [15]) for details).

Setting $z_{\mathcal{D}_m} = (z_v)_{v \in \mathcal{D}_m}$ and $z_{\mathcal{D}_f} = (z_v)_{v \in \mathcal{D}_f}$, the two-phase VAG fluxes combine the VAG single phase Darcy fluxes including gravity

$$F_{K,v}^\alpha(u_{\mathcal{D}_m}^\alpha) = F_{K,v}(u_{\mathcal{D}_m}^\alpha) + \rho^\alpha g F_{K,v}(z_{\mathcal{D}_m}), \quad F_{\sigma,\mathbf{s}}^\alpha(u_{\mathcal{D}_f}^\alpha) = F_{\sigma,\mathbf{s}}(u_{\mathcal{D}_f}^\alpha) + \rho^\alpha g F_{\sigma,\mathbf{s}}(z_{\mathcal{D}_f}),$$

$$F_{\overline{K}_s,\mathbf{s}}^\alpha(u_{\overline{K}_s}^\alpha, u_{\mathbf{s}}^\alpha) = F_{\overline{K}_s,\mathbf{s}}(u_{\overline{K}_s}^\alpha, u_{\mathbf{s}}^\alpha) - \frac{1}{3} \rho^\alpha \sum_{T \in \Delta | \mathbf{s} \in \overline{T}} \int_T \lambda_{f,n} \mathbf{g} \cdot \mathbf{n}_{\overline{K}_s,T} d\sigma(\mathbf{x}),$$

$$F_{K_\sigma,\sigma}^\alpha(u_{K_\sigma}^\alpha, u_\sigma^\alpha) = F_{K_\sigma,\sigma}(u_{K_\sigma}^\alpha, u_\sigma^\alpha) - \rho^\alpha \int_\sigma \lambda_{f,n} \mathbf{g} \cdot \mathbf{n}_{K_\sigma,\sigma} d\sigma(\mathbf{x}),$$

with the usual Two-Point phase potential upwinding of the mobilities, leading to define

$$q_{K,v}^\alpha = M_{\text{rk}}^\alpha(s_K^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^+ + M_{\text{rk}}^\alpha(s_{K,v}^\alpha)(F_{K,v}^\alpha(u_{\mathcal{D}}^\alpha))^-,$$

for all $K \in \mathcal{M}$, $v \in \Xi_K$,

$$q_{\sigma,s}^\alpha = M_{\Gamma_\sigma}^\alpha (s_\sigma^\alpha) (F_{\sigma,s}^\alpha (u_{\mathcal{D}}^\alpha))^+ + M_{\Gamma_\sigma}^\alpha (s_{\sigma,s}^\alpha) (F_{\sigma,s}^\alpha (u_{\mathcal{D}}^\alpha))^- ,$$

for all $\sigma \in \mathcal{F}_\Gamma$, $\mathbf{s} \in \mathcal{V}_\sigma$,

$$q_{\overline{K}_s,s}^\alpha = \frac{1}{\text{Card}(\mathcal{F}_{\Gamma,\overline{K}_s})} \sum_{\sigma \in \mathcal{F}_{\Gamma,\overline{K}_s}} \left(M_{\Gamma_\sigma}^\alpha (s_{\sigma,\overline{K}_s}^\alpha) (F_{\overline{K}_s,s}^\alpha (u_{\overline{K}_s}^\alpha, u_{\mathbf{s}}^\alpha))^+ \right. \\ \left. + M_{\Gamma_\sigma}^\alpha (s_{\sigma,s}^\alpha) (F_{\overline{K}_s,s}^\alpha (u_{\overline{K}_s}^\alpha, u_{\mathbf{s}}^\alpha))^- \right)$$

for all $\mathbf{s} \in \mathcal{V}_\Gamma$, $\overline{K}_s \in \overline{\mathcal{M}}_s$, and

$$q_{K_\sigma,\sigma}^\alpha = M_{\Gamma_\sigma}^\alpha (s_{\sigma,K_\sigma}^\alpha) (F_{K_\sigma,\sigma}^\alpha (u_{K_\sigma}^\alpha, u_\sigma^\alpha))^+ + M_{\Gamma_\sigma}^\alpha (s_\sigma^\alpha) (F_{K_\sigma,\sigma}^\alpha (u_{K_\sigma}^\alpha, u_\sigma^\alpha))^-$$

for all $\sigma \in \mathcal{F}_\Gamma$, $K \in \mathcal{M}_\sigma$.

Control volumes and accumulation terms: as for the continuous pressure model, porous volumes $\phi_{K,v}$, $v \in \Xi_K \setminus \mathcal{D}_{\text{Dir}}$ (resp. $\phi_{\sigma,s}$, $\mathbf{s} \in \mathcal{V}_\sigma \setminus \mathcal{D}_{\text{Dir}}$) are obtained by distribution of the cell $K \in \mathcal{M}$ (resp. fracture face $\sigma \in \mathcal{F}_\Gamma$) porous volume. A porous volume $\phi_{\sigma,\overline{K}_s}$ (resp. ϕ_{σ,K_σ}) is also distributed from the fracture face σ to the interface d.o.f. \overline{K}_s (resp. K_σ) for $\sigma \in \mathcal{F}_{\Gamma,\overline{K}_s}$, $\overline{K}_s \in \overline{\mathcal{M}}_\Gamma$ (resp. for $K_\sigma \in \overline{\mathcal{F}}_\Gamma$). These interface porous volumes are required to avoid the singularity of the linear systems obtained after Newton linearization. Their influence on the solution is small provided that they are chosen small enough (see [25]). Then we set

$$\left\{ \begin{array}{l} \phi_K = \int_K \phi_m(\mathbf{x}) d\mathbf{x} - \sum_{v \in \Xi_K \setminus \mathcal{D}_{\text{Dir}}} \phi_{K,v}, \quad K \in \mathcal{M}, \\ \phi_\sigma = \int_\sigma d_f(\mathbf{x}) \phi_f(\mathbf{x}) d\mathbf{x} - \sum_{\mathbf{s} \in \mathcal{V}_\sigma \setminus \mathcal{D}_{\text{Dir}}} \phi_{\sigma,s} - \sum_{K \in \mathcal{M}_\sigma} \phi_{\sigma,K_\sigma} \\ \quad - \sum_{\overline{K}_s \in \overline{\mathcal{M}}_\Gamma \setminus \mathcal{D}_{\text{Dir}} \mid \sigma \in \mathcal{F}_{\Gamma,\overline{K}_s}} \phi_{\sigma,\overline{K}_s}, \quad \sigma \in \mathcal{F}_\Gamma, \end{array} \right.$$

and we define the accumulations terms by

$$\left\{ \begin{array}{l} \mathcal{A}_K^\alpha = \phi_K s_K^\alpha, \quad K \in \mathcal{M}, \\ \mathcal{A}_{\overline{K}_s}^\alpha = \sum_{K \in \mathcal{M}_s} \phi_{K,\overline{K}_s} s_{K,\overline{K}_s}^\alpha, \quad \mathbf{s} \in \mathcal{V} \setminus (\mathcal{D}_{\text{Dir}} \cup \mathcal{V}_\Gamma), \\ \mathcal{A}_\sigma^\alpha = \phi_\sigma s_\sigma^\alpha, \quad \sigma \in \mathcal{F}_\Gamma, \\ \mathcal{A}_s^\alpha = \sum_{\sigma \in \mathcal{F}_{\Gamma,s}} \phi_{\sigma,s} s_{\sigma,s}^\alpha, \quad \mathbf{s} \in \mathcal{V}_\Gamma \setminus \mathcal{D}_{\text{Dir}}, \\ \mathcal{A}_{K_\sigma}^\alpha = \phi_{\sigma,K_\sigma} s_{\sigma,K_\sigma}^\alpha + \phi_{K,K_\sigma} s_{K,K_\sigma}^\alpha, \quad K_\sigma \in \overline{\mathcal{F}}_\Gamma, \\ \mathcal{A}_{\overline{K}_s}^\alpha = \sum_{K \in \overline{K}_s} \phi_{K,\overline{K}_s} s_{K,\overline{K}_s}^\alpha + \sum_{\sigma \in \mathcal{F}_{\Gamma,\overline{K}_s}} \phi_{\sigma,\overline{K}_s} s_{\sigma,\overline{K}_s}^\alpha, \quad \overline{K}_s \in \overline{\mathcal{M}}_\Gamma \setminus \mathcal{D}_{\text{Dir}}, \end{array} \right.$$

Conservation equations: the VAG discretization of the discontinuous pressure model solves for $u_{\mathcal{D}}^{nw}$ and $\tau_{\mathcal{D}}$ such that

$$\left\{ \begin{array}{l} \frac{\mathcal{A}_K^\alpha - \mathcal{A}_K^{\alpha, n-1}}{\Delta t^n} + \sum_{v \in \mathfrak{E}_K} q_{K,v}^\alpha = 0, \quad K \in \mathcal{M}, \\ \frac{\mathcal{A}_{\bar{K}_s}^\alpha - \mathcal{A}_{\bar{K}_s}^{\alpha, n-1}}{\Delta t^n} - \sum_{K \in \mathcal{M}_s} q_{K, \bar{K}_s}^\alpha = 0, \quad \mathbf{s} \in \mathcal{V} \setminus (\mathcal{V}_\Gamma \cup \mathcal{D}_{\text{Dir}}), \\ \frac{\mathcal{A}_\sigma^\alpha - \mathcal{A}_\sigma^{\alpha, n-1}}{\Delta t^n} + \sum_{s \in \mathcal{V}_\sigma} q_{\sigma, s}^\alpha - \sum_{K \in \mathcal{M}_\sigma} q_{K, \sigma}^\alpha = 0, \quad \sigma \in \mathcal{F}_\Gamma, \\ \frac{\mathcal{A}_s^\alpha - \mathcal{A}_s^{\alpha, n-1}}{\Delta t^n} - \sum_{\sigma \in \mathcal{F}_{\Gamma, s}} q_{\sigma, s}^\alpha - \sum_{\bar{K}_s \in \bar{\mathcal{M}}_s} q_{\bar{K}_s, s}^\alpha = 0, \quad \mathbf{s} \in \mathcal{V}_\Gamma \setminus \mathcal{D}_{\text{Dir}}, \\ \frac{\mathcal{A}_{K_\sigma}^\alpha - \mathcal{A}_{K_\sigma}^{\alpha, n-1}}{\Delta t^n} - q_{K, K_\sigma}^\alpha + q_{K_\sigma, \sigma}^\alpha = 0, \quad K_\sigma \in \bar{\mathcal{F}}_\Gamma, \\ \frac{\mathcal{A}_{\bar{K}_s}^\alpha - \mathcal{A}_{\bar{K}_s}^{\alpha, n-1}}{\Delta t^n} - \sum_{K \in \bar{\mathcal{K}}_s} q_{K, \bar{K}_s}^\alpha + q_{\bar{K}_s, s}^\alpha = 0, \quad \bar{K}_s \in \bar{\mathcal{M}}_\Gamma \setminus \mathcal{D}_{\text{Dir}}, \\ \tau_v = \tau_{\text{Dir}, v}, \quad u_v^{nw} = u_{\text{Dir}, v}^{nw}, \quad v \in \mathcal{D}_{\text{Dir}}. \end{array} \right. \quad (3.24)$$

f and m-upwind discontinuous pressure models: the above discontinuous pressure model, termed mf nonlinear model in the following, leads to difficulties to solve the nonlinear system (3.24) due to the combination of highly contrasted matrix and fracture rock types and to the small pore volumes at mf interface d.o.f. One possibility to solve this issue, still preserving the ability to take into account fractures acting as drains or barriers, is to linearize the matrix fracture transmission conditions w.r.t. the mf interface unknowns and to apply a f or m-upwind approximation of the mobilities. This idea, developed in [16] for the VAG discretization and in [1] for the TPFA discretization, replaces the primary unknowns u_v^{nw} , τ_v at matrix fracture d.o.f. $v \in \mathcal{D}_{mf}$ by both phase pressures u_v^{nw} , u_v^w , $v \in \mathcal{D}_{mf}$, and the conservation equations at matrix fracture d.o.f. by

$$F_{\bar{K}_s, s}^\alpha(u_{\bar{K}_s}^\alpha, u_s^\alpha) - \sum_{K \in \bar{\mathcal{K}}_s} F_{K, \bar{K}_s}^\alpha(u_{\mathcal{D}_m}^\alpha) = 0, \quad F_{K_\sigma, \sigma}^\alpha(u_{K_\sigma}^\alpha, u_\sigma^\alpha) - F_{K, K_\sigma}^\alpha(u_{\mathcal{D}_m}^\alpha) = 0,$$

for $\bar{K}_s \in \bar{\mathcal{M}}_\Gamma$ and $K_\sigma \in \bar{\mathcal{F}}_\Gamma$. Note that the pore volumes ϕ_{σ, \bar{K}_s} and ϕ_{σ, K_σ} are set to zero. Since phase saturations are no longer defined at matrix fracture d.o.f., one need to modify the upwind mobilities in the definition of the fluxes q_{K, \bar{K}_s}^α , $\bar{K}_s \in \bar{\mathcal{M}}_\Gamma$ now connecting directly the cell K and the fracture d.o.f. s , and in the definition of q_{K, K_σ}^α , $K_\sigma \in \bar{\mathcal{F}}_\Gamma$ now connecting the cell K and the fracture d.o.f. σ . These new connectivities modify the fracture conservations equations for $\sigma \in \mathcal{F}_\Gamma$ and $s \in \mathcal{V}_\Gamma \setminus \mathcal{D}_{\text{Dir}}$ as follows:

$$\begin{cases} \frac{\mathcal{A}_\sigma^\alpha - \mathcal{A}_\sigma^{\alpha, n-1}}{\Delta t^n} + \sum_{\mathbf{s} \in \mathcal{V}_\sigma} q_{\sigma, \mathbf{s}}^\alpha - \sum_{K \in \mathcal{M}_\sigma} q_{K, K_\sigma}^\alpha = 0, \sigma \in \mathcal{F}_\Gamma, \\ \frac{\mathcal{A}_\mathbf{s}^\alpha - \mathcal{A}_\mathbf{s}^{\alpha, n-1}}{\Delta t^n} - \sum_{\sigma \in \mathcal{F}_{\Gamma, \mathbf{s}}} q_{\sigma, \mathbf{s}}^\alpha - \sum_{\bar{K}_s \in \bar{\mathcal{M}}_s} q_{K, \bar{K}_s}^\alpha = 0, \mathbf{s} \in \mathcal{V}_\Gamma \setminus \mathcal{D}_{\text{Dir}}, \end{cases} \quad (3.25)$$

The modified fluxes are defined by

$$\begin{cases} q_{K, \bar{K}_s}^\alpha = M_{\text{rt}_K}^\alpha (s_K^\alpha) (F_{K, \bar{K}_s}^\alpha (u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_K}^\alpha (S_{\chi_{\bar{K}_s, \text{rt}_K}}^\alpha (\tau_s)) (F_{K, \bar{K}_s}^\alpha (u_{\mathcal{D}}^\alpha))^- , \\ q_{K, K_\sigma}^\alpha = M_{\text{rt}_K}^\alpha (s_K^\alpha) (F_{K, K_\sigma}^\alpha (u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_K}^\alpha (S_{\chi_{K_\sigma, \text{rt}_K}}^\alpha (\tau_\sigma)) (F_{K, K_\sigma}^\alpha (u_{\mathcal{D}}^\alpha))^- , \end{cases} \quad (3.26)$$

for the m-upwind discontinuous pressure model, and by

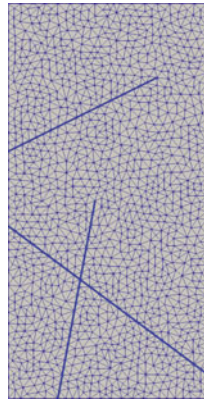
$$\begin{cases} q_{K, \bar{K}_s}^\alpha = M_{\text{rt}_K}^\alpha (s_K^\alpha) (F_{K, \bar{K}_s}^\alpha (u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_s}^\alpha (S_{\chi_{\mathbf{s}, \text{rt}_s}}^\alpha (\tau_s)) (F_{K, \bar{K}_s}^\alpha (u_{\mathcal{D}}^\alpha))^- , \\ q_{K, K_\sigma}^\alpha = M_{\text{rt}_K}^\alpha (s_K^\alpha) (F_{K, K_\sigma}^\alpha (u_{\mathcal{D}}^\alpha))^+ + M_{\text{rt}_\sigma}^\alpha (S_{\chi_{K_\sigma, \text{rt}_\sigma}}^\alpha (\tau_\sigma)) (F_{K, K_\sigma}^\alpha (u_{\mathcal{D}}^\alpha))^- , \end{cases} \quad (3.27)$$

for the f-upwind discontinuous pressure model, where a fracture rock type $\text{rt}_\mathbf{s}$ has been assigned to the node \mathbf{s} . As for the continuous pressure model, a Multi-Point upwinding can also be introduced for these fluxes using the additional mobility unknowns $M_{\bar{K}_s}^\alpha$, and $M_{K_\sigma}^\alpha$, $\alpha \in \{nw, w\}$. Note that, for fracture acting as drains, these f and m-upwind discontinuous pressure models provide basically the same solutions than respectively the f and m-upwind continuous pressure models. As already mentioned, this is not the case of the mf nonlinear discontinuous pressure model (3.24) due to the possible degeneracy of the phase mobilities appearing in the matrix fracture transmission conditions.

3.4.1 Numerical Experiments

In this subsection, we compare on the following test case, the mf nonlinear, the m-upwind and the f-upwind models using a reference solution obtained by the equi-dimensional model. The code implementation is the same for all models and described in the beginning of Sect. 3.3.4. The m-upwind and f-upwind models would require the design of specific preconditioners due to the two independent elliptic pressure unknowns at mf interfaces combined with a single independent elliptic unknown at cells and fracture faces. This explains the use for these two models of the direct linear solver SuperLU from the library [49]. The GMRes iterative solver combined with the CPR-AMG preconditioner is still used for the mf nonlinear and equi-dimensional models. It results that the overall numbers of Newton iterations $N_{\text{Newton}} N_{\Delta t}$ are more relevant for performance comparison than the CPU times which are not reported for this test case.

Fig. 3.25 Coarse mesh over the domain under consideration, which contains two intersecting fractures with high permeability and low capillarity and one upper fracture with low permeability and high capillarity. The size of the domain is $4\text{ m} \times 8\text{ m}$ and the fractures have an aperture of 4 cm



We consider a fractured domain as defined in Fig. 3.25. The matrix permeability is isotropic of 0.1 Darcy and matrix porosity is 0.2. The two lower fractures are drains (f_d) of isotropic permeability 100.0 Darcy and porosity 0.4. In the upper fracture, acting as a barrier (f_b), the permeability is isotropic of 0.001 Darcy and the porosity is 0.2. The capillary pressures are the same than in Sect. 3.3.5.1 with the Corey parameters $b_m = 1$ bar in the matrix, $b_{f_b} = 10$ bar in the barrier fracture and $b_{f_d} = 0.1$ bar in the drain fractures. Initially, the reservoir is saturated with water (density 1000 kg/m^3 , viscosity 0.001 Pa s) and oil (density 700 kg/m^3 , viscosity 0.005 Pa s) is injected in the bottom fracture, which is managed by imposing non-homogeneous Neumann conditions at the injection location. The oil then rises by gravity, thanks to its lower density compared to water and by the overpressure induced by the imposed injection rate. Also, Dirichlet boundary conditions are imposed at the upper boundary of the domain. Elsewhere, we have homogeneous Neumann conditions.

The tests are driven on triangular meshes, extended to 3D prismatic meshes by adding a second layer of nodes as a translation of the original nodes in normal direction to the plane of the original 2D domain (cf. Fig. 3.25). The equi-dimensional mesh contains two layers of cells in the fractures. In order to focus on modelling errors, the meshes are chosen to be fine with cell sizes of the same order as the fracture aperture. The final simulation time is fixed to $t_f = 54$ days. The time stepping is defined as in (3.16) using $\Delta t_{init} = 0.01$ days and $\Delta t_{max} = 0.1$ days for the equi-dimensional and hybrid dimensional mf nonlinear models, and $\Delta t_{init} = 0.002$ days and $\Delta t_{max} = 0.27$ days for the hybrid-dimensional m-upwind and f-upwind models. The maximum number of Newton iterations per time step is fixed as $Newton_{max} = 35$. The criterion of convergence for the Newton algorithm is defined as in (3.17) with $Res_{max} = 10^{-6}$ and $dx_{max} = 10^{-4}$, and the relaxation of the Newton step is controlled as in (3.18) by the parameter $dx_{obj} = 0.5$.

The hybrid dimensional mf nonlinear and m-upwind models make use of the parametrization (3.21)–(3.23) at the mf interfaces.

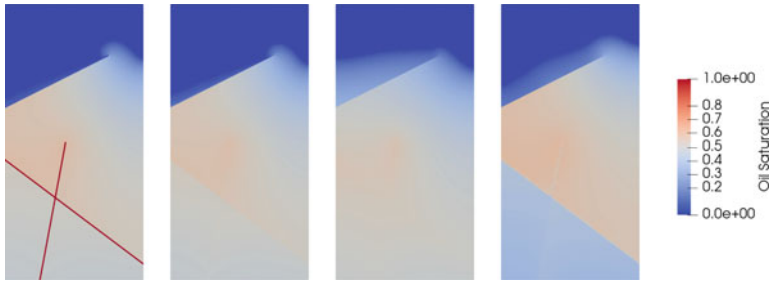


Fig. 3.26 Comparison of the equi-dimensional model and of the mf nonlinear, m-upwind and f-upwind discontinuous pressure DFM models (from left to right) numerical solutions for non-wetting phase saturation at final time $t = 54$ days

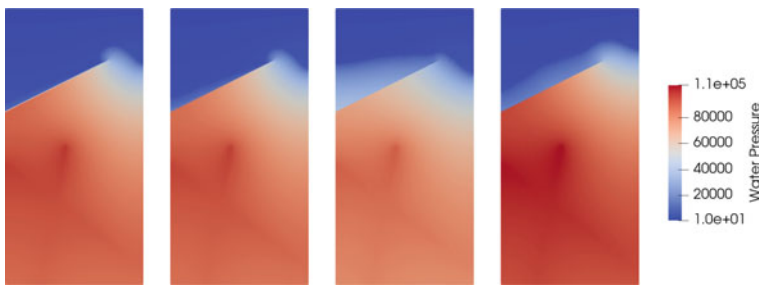


Fig. 3.27 Comparison of the equi-dimensional model and of the mf nonlinear, m-upwind and f-upwind discontinuous pressure DFM models (from left to right) numerical solutions for water overpressure at final time $t = 54$ days

In this test case, we study the presence of a fracture, which acts as a barrier, both by its low permeability and by its high capillarity compared to the rock matrix. As a result of the higher capillarity, the sign of the matrix-fracture non-wetting phase saturation jump $S_m^{nw}(\gamma^\pm p_{c,m}) - S_f^{nw}(\gamma^\pm p_{c,m})$ at the mf interfaces is non negative.

Figures 3.26, 3.27 and 3.28 compare the above mf nonlinear, m-upwind and f-upwind discontinuous pressure models to a reference equi-dimensional model. For the f-upwind and m-upwind models, mass transfer of the non-wetting phase from the matrix to the barrier is overestimated, since in this direction, saturation jumps are not accounted for. The assumption of constant saturation across the fracture for these models consequently leads to an overestimation of the non-wetting phase leaving the barrier. This overestimation is most severe for the m-upwind model, which takes into account saturation jumps for fluxes directed from the fracture to the matrix. Again, the mf nonlinear model does not suffer from the difficulties described above, since it provides mass transport that passes by the mf interfaces and takes into account the saturation jumps. Table 3.4 compares the numerical behavior of the different models on this test case.

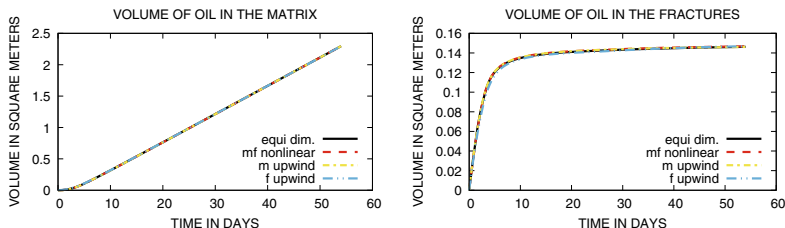


Fig. 3.28 Matrix and fracture volumes occupied by the non-wetting phase as a function of time for the equi-dimensional model and for the mf nonlinear, m-upwind and f-upwind discontinuous pressure DFM models

Table 3.4 Numerical behavior of the simulation obtained by the VAG scheme for the equi-dimensional model and for the mf nonlinear, m-upwind and f-upwind discontinuous pressure DFM models, as presented in Sect. 3.4. We refer to the beginning of Sect. 3.3.4 for the description of the entries with dof_{lin} and N_z accounting for the elimination of the cell unknowns in the linear systems

Scheme	mesh	dof	dof_{lin}	N_z	$N_{\Delta t}$	N_{chop}	N_{Newton}
equi dim.	22k	45k	23k	317k	589	2	4.1
mf nonlinear	17k	35k	18k	261k	585	1	3.4
m-upwind	17k	35k	18k	266k	255	0	4.8
f-upwind	17k	35k	18k	266k	255	0	4.6

3.5 Conclusions and Perspectives

This article reviews the nodal VAG discretization of DFM two-phase Darcy flow models. For linear transmission conditions, the adaptation of the control volumes combined with a Multi-Point upwind approximation of the mobilities for f-upwind models or taking into account the saturation jump for m-upwind models, allows to obtain a similar accuracy as face based discretizations with a much lower CPU time on tetrahedral meshes. Nonlinear mf transmission conditions provide a more accurate DFM model than linear transmission conditions. As discussed in [1, 16], they can account for a large range of physical processes at mf interfaces which cannot be captured by linear mf transmission conditions even in the case of fractures acting as drains. It is typically the case for fractures acting as capillary barriers, or for highly permeable fractures filled with a given phase acting as a barrier for the other phase. The VAG discretization of DFM models with nonlinear mf transmission conditions still raises the issue of numerical efficiency regarding the nonlinear convergence due to the combination of highly nonlinear transmission conditions with tiny volumes at mf interfaces. Improving the numerical efficiency for this type of DFM models is the object of ongoing researches in two directions. The first is to go back to face based discretizations allowing the elimination of the mf interface unknowns with a local nonlinear interface solver as in [1] using TPFA discretization on orthogonal meshes

and in [2] using an HFV discretization. The second perspective is to use the more robust Hybrid Upwinding approximation of the mobilities to define the two-phase Darcy fluxes at mf interfaces as proposed in [6] for TPFA schemes and in [19] for the VAG discretization.

References

1. J. Aghili, K. Brenner, R. Masson, L. Trenty, Two-phase discrete fracture matrix models with linear and nonlinear transmission conditions. *GEM Int. J. Geomath.* **10**, 1 (2019)
2. J. Aghili, K. Brenner, J. Hennicker, R. Masson, L. Trenty, Hybrid finite volume discretization of two-phase discrete fracture matrix models with nonlinear interface solver, in *ECMOR XVI—16th European Conference on the Mathematics of Oil Recovery*, Sept 2018
3. E.M. Ahmed, J. Jaffré, J.E. Roberts, A reduced fracture model for two-phase flow with different rock types. *Math. Comput. Simul.* **7**, 49–70 (2017)
4. R. Ahmed, M.G. Edwards, S. Lamine, B.A.H. Huisman, Control-volume distributed multi-point flux approximation coupled with a lower-dimensional fracture model. *J. Comput. Phys.* **284**, 462–489 (2015)
5. R. Ahmed, Y. Xie, M.G. Edwards, A cell-centred CVD-MPFA finite volume method for two-phase fluid flow problems with capillary heterogeneity and discontinuity. *Transp. Porous Media* **127**, 35–52 (2019)
6. A.H. Alali, F.P. Hamon, B.P. Mallison, H.A. Tchelepi, Finite-volume simulation of capillary-dominated flow in matrix-fracture systems using interface conditions (2019). [arXiv:1907.03747v1](https://arxiv.org/abs/1907.03747v1). math.NA
7. C. Alboin, J. Jaffré, J.E. Roberts, C. Serres, Modeling fractures as interfaces for flow and transport in porous media, in *Fluid Flow and Transport in Porous Media*, vol. 295, ed. by E. Chen (American Mathematical Society, 2002), pp. 13–24
8. K. Aziz, A. Settari, *Petroleum Reservoir Simulation* (Elsevier, London, 1979)
9. B. Andreianov, K. Brenner, C. Cancès, Approximating the vanishing capillarity limit of two-phase flow in multi-dimensional heterogeneous porous medium. *ZAMM J. Appl. Math. Mech./Zeitschrift für Angewandte Mathematik und Mechanik* **94**(7–8), 655–667 (2014)
10. P. Angot, F. Boyer, F. Hubert, Asymptotic and numerical modeling of flows in fractured porous media. *M2AN* **43**(2), 239–275 (2009)
11. I. Berre, W. Boon, B. Flemisch, A. Fumagalli, D. Gläser, E. Keilegavlen, A. Scotti, I. Stefansson, A. Tatomir, Call for participation: verification benchmarks for single-phase flow in three-dimensional fractured porous media (2018). [arXiv:1809.06926](https://arxiv.org/abs/1809.06926)
12. I. Berre, W.M. Boon, B. Flemisch, A. Fumagalli, D. Gläser, E. Keilegavlen, A. Scotti, I. Stefansson, A. Tatomir, K. Brenner, S. Burbulla, P. Devloo, O. Duran, M. Favino, J. Hennicker, I.-H. Lee, K. Lipnikov, R. Masson, K. Mosthaf, M.C.G. Nestola, C.-F. Ni, K. Nikitin, P. Schädle, D. Svyatskiy, R. Yanbarisov, P. Zulian, Verification benchmarks for single-phase flow in three-dimensional fractured porous media (2020). [arXiv:2002.07005](https://arxiv.org/abs/2002.07005)
13. K. Brenner, M. Groza, C. Guichard, R. Masson, Vertex approximate gradient scheme for hybrid-dimensional two-phase Darcy flows in fractured porous media. *ESAIM Math. Model. Numer. Anal.* **49**, 303–330 (2015)
14. K. Brenner, M. Groza, L. Jeannin, R. Masson, J. Pellerin, Immiscible two-phase Darcy flow model accounting for vanishing and discontinuous capillary pressures: application to the flow in fractured porous media. *Comput. Geosci.* **21**, 1075–1094 (2017)
15. K. Brenner, J. Hennicker, R. Masson, P. Samier, Gradient discretization of hybrid-dimensional Darcy flow in fractured porous media with discontinuous pressures at matrix-fracture interfaces. *IMA J. Numer. Anal.* **37**(3), 1551–1585 (2016)

16. K. Brenner, J. Hennicker, R. Masson, P. Samier, Hybrid-dimensional modeling of two-phase flow through fractured porous media with enhanced matrix fracture transmission conditions. *J. Comput. Phys.* **357**, 100–124 (2018)
17. K. Brenner, M. Groza, C. Guichard, G. Lebeau, R. Masson, Gradient discretization of hybrid-dimensional Darcy flows in fractured porous media. *Numer. Math.* **134**(3), 569–609 (2016)
18. K. Brenner, C. Cancès, D. Hilhorst, Finite volume approximation for an immiscible two-phase flow in porous media with discontinuous capillary pressure. *Comput. Geosci.* **17**(3), 573–597 (2013)
19. K. Brenner, R. Masson, E.H. Quenjel, Positivity-preserving vertex approximate gradient discretization of two-phase Darcy flows in heterogeneous porous media. *J. Comput. Phys.* **409** (2020)
20. I. Bogdanov, V. Mourzenko, J.-F. Thovert, P.M. Adler, Two-phase flow through fractured porous media. *Phys. Rev. E* **68**, 026703 (2003)
21. C. Cancès, Finite volume scheme for two-phase flows in heterogeneous porous media involving capillary pressure discontinuities. *Math. Model. Numer. Anal.* **43**, 973–1001 (2009)
22. C. Cancès, M. Pierre, An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.* **44**, 966–992 (2012)
23. G. Chavent, J. Jaffré, *Mathematical Models and Finite Elements for Reservoir Simulation: Single Phase, Multiphase and Multicomponent Flows Through Porous Media* (North-Holland, Amsterdam, stud. math. appl., 1986)
24. F. Chave, D. Di Pietro, L. Formaggia, A hybrid high-order method for Darcy flows in fractured porous media. *SIAM J. Sci. Comput.* **40**(2), 1063–1094 (2018)
25. J. Droniou, J. Hennicker, R. Masson, Numerical analysis of a two-phase flow discrete fracture model. *Numer. Math.* **141**(1), 21–62 (2019)
26. J. Droniou, J. Hennicker, R. Masson, Uniform-in-time convergence of numerical schemes for a two-phase discrete fracture model. *Finite Volumes for Complex Applications VIII-Methods and Theoretical Aspects*. Springer Proceedings in Mathematics & Statistics, vol. 199 (Springer, Cham, 2017), pp. 275–283
27. C.J. Van Duijn, J. Molenaar, M.J. De Neef, The effect of capillary forces on immiscible two-phase flow in heterogeneous porous media. *Transp. Porous Media* **21**(1), 71–93 (1995)
28. G. Enchéry, R. Eymard, A. Michel, Numerical approximation of a two-phase flow problem in a porous medium with discontinuous capillary forces. *SIAM J. Numer. Anal.* **43**(6), 2402–2422 (2006)
29. R. Eymard, C. Guichard, R. Herbin, Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM Math. Model. Numer. Anal.* **46**, 265–290 (2010)
30. R. Eymard, C. Guichard, R. Herbin, R. Masson, Vertex centred discretization of two-phase Darcy flows on general meshes. *ESAIM Proc.* **35**, 59–78 (2012)
31. R. Eymard, C. Guichard, R. Herbin, R. Masson, Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. *ZAMM J. Appl. Math. Mech.* **94**(7–8), 560–585 (2014)
32. E. Flauraud, F. Nataf, I. Faille, R. Masson, Domain decomposition for an asymptotic geological fault modeling. *Comptes Rendus à l’Académie des Sciences, Mécanique* **331**, 849–855 (2003)
33. L. Formaggia, A. Fumagalli, A. Scotti, P. Ruffo, A reduced model for Darcy’s problem in networks of fractures. *ESAIM: M2AN* **48**(4), 1089–1116 (2014)
34. A. Fumagalli, A. Scotti, A numerical method for two-phase flow in fractured porous media with non-matching grids. *Adv. Water Resour.* **62**, 454–464 (2013)
35. S. Geiger, S. Matthäi, J. Niessner, R. Helmig, Black-oil simulations for three-component, three-phase flow in fractured porous media. *SPE J.* **6**, 338–354 (2009)
36. D. Gläser, R. Helmig, B. Flemish, H. Class, A discrete fracture model for two-phase flow in fractured porous media. *Adv. Water Resour.* **110**, 335–348 (2017)
37. M. Groza, Modelization and discretization of two-phase flows in porous media with discrete fracture networks. PhD, Nov 2016, <https://tel.archives-ouvertes.fr/tel-01466743/document>
38. J. Hoteit, A. Firoozabadi, An efficient numerical model for incompressible two-phase flow in fracture media. *Adv. Water Resour.* **31**, 891–905 (2008)

39. J. Jaffré, V. Martin, J.E. Roberts, Modeling fractures and barriers as interfaces for flow in porous media. *SIAM J. Sci. Comput.* **26**(5), 1667–1691 (2005)
40. J. Jaffré, M. Mnejja, J.E. Roberts, A discrete fracture model for two-phase flow with matrix-fracture interaction. *Procedia Comput. Sci.* **4**, 967–973 (2011)
41. M. Karimi-Fard, L.J. Durlofsky, K. Aziz, An efficient discrete-fracture model applicable for general-purpose reservoir simulators. *SPE-88812-PA* **9**(2) (2004)
42. S. Lacroix, Y.V. Vassilevski, M.F. Wheeler, Decoupling preconditioners in the implicit parallel accurate reservoir simulator (IPARS). *Numer. Linear Algebr. Appl.* **8**, 537–549 (2001)
43. J. Monteagudu, A. Firoozabadi, Control-volume model for simulation of water injection in fractured media: incorporating matrix heterogeneity and reservoir wettability effects. *SPE-98108-PA* **12**(3), 355–366 (2007)
44. V. Reichenberger, H. Jakobs, P. Bastian, R. Helmig, A mixed-dimensional finite volume method for multiphase flow in fractured porous media. *Adv. Water Resour.* **29**(7), 1020–1036 (2006)
45. J.W. Ruge, K. Stüben, Algebraic multigrid (AMG), in *Multigrid Methods, Frontiers in Applied Mathematics*, vol. 5, ed. by S.F. McCormick (SIAM, Philadelphia, 1986)
46. T.H. Sandve, I. Berre, J.M. Nordbotten, An efficient multi-point flux approximation method for discrete fracture-matrix simulations. *J. Comput. Phys.* **231**, 3784–3800 (2012)
47. R. Scheichl, R. Masson, J. Wendebourg, Decoupling and block preconditioning for sedimentary basin simulations. *Comput. Geosci.* **7**, 295–318 (2003)
48. SLATEC Common Mathematical Library, Version 4.1, July 1993, <http://www.netlib.org/slatec/index.html>
49. X.S. Li, J.W. Demmel, J.R. Gilbert, L. Grigori, M. Shao, I. Yamazaki, Technical report LBNL-44289. Lawrence Berkeley National Laboratory, SuperLU Users' Guide, Sept 1999, <http://crd.lbl.gov/~xiaoye/SuperLU>
50. M. Tene, S. Bosma, M.S. Al Kobaisi, H. Hajibeygi, Projection-based embedded discrete fracture model (pEDFM). *Adv. Water Resour.* **105**, 205–216 (2017)
51. X. Tunc, I. Faille, T. Gallouët, M.C. Cacas, P. Havé, A model for conductive faults with non matching grids. *Comput. Geosci.* **16**, 277–296 (2012)
52. Y. Xie, M.G. Edwards, Unstructured CVD-MPFA reduced-dimensional DFM models for two-phase flow, coupled with higher resolution hybrid upwind methods. *Soc. Pet. Eng. SPE-193886-MS* (2019)
53. F. Xing, R. Masson, S. Lopez, Parallel numerical modeling of hybrid-dimensional compositional non-isothermal Darcy flows in fractured porous media. *J. Comput. Phys.* **345**, 637–664 (2017)

Chapter 4

An Introduction to Multi-point Flux (MPFA) and Stress (MPSA) Finite Volume Methods for Thermo-poroelasticity



Jan Martin Nordbotten and Eirik Keilegavlen

Abstract In this chapter, we give a unified introduction to the MPFA- and MPSA-type finite volume methods for Darcy flow and poro-elasticity, applicable to general polyhedral grids. This leads to a more systematic perspective of these methods than has been exposed in previous texts, and we therefore refer to this discretization family as the MPxA methods. We apply this MPxA framework to also define a consistent finite-volume discretization of thermo-poro-elasticity. The present chapter introduces the general theory and state-of-the-art of MPFA-type methods, leaving the more technical results to the provided references. We close the chapter by a section containing applications to problems with complex geometries and non-linear physics.

Keywords Polyhedral grids · Finite volume methods · Elliptic equations · Coupled flow and mechanics

4.1 Introduction and Historical Context

The first so-called Multi-Point Finite Volumes methods were developed in the first half of the 1990s, within the context of numerical discretizations for multi-phase flow in geological porous media [1–4]. In particular, these methods are constructed to solve the so-called pressure equation, which is a second-order elliptic partial differential equation where it is understood that the material parameter may have very low regularity in space. This equation is best presented as a system of first order equations, consisting of balancing the flux q with a source r

$$\nabla \cdot q = r. \quad (4.1.1)$$

On occasion of the 70th anniversary of Ivar Aavatsmark.

J. M. Nordbotten (✉) · E. Keilegavlen
Department of Mathematics, University of Bergen, Bergen, Norway
e-mail: jan.nordbotten@math.uib.no

E. Keilegavlen
e-mail: Eirik.Keilegavlen@uib.no

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021
D. A. Di Pietro et al. (eds.), *Polyhedral Methods in Geosciences*,
SEMA SIMAI Springer Series 27,
https://doi.org/10.1007/978-3-030-69363-3_4

Complemented by the constitutive law that the flux is derived from a fluid potential p :

$$q + \kappa \nabla p = g. \quad (4.1.2)$$

Here κ is the material tensor, which is essentially the permeability to flow. The permeability may be both anisotropic and vary strongly as a function of space due to the complex nature of natural rocks. In order to keep the presentation simple, we have simplified terms. Thus it is understood that in applications, the conservation statement is for the mass flux, while the right-hand side of Eq. (4.1.2) is the product of the permeability and gravity, etc. For a detailed physical exposition of Eqs. (4.1.1–4.1.2), see e.g. [5–8].

For problems on the form of Eqs. (4.1.1–4.1.2), favorable attributes of a numerical discretization method can be summarized as follows (acknowledging that no list of this form is complete):

- (A) *Flux balance*: An exact local representation of fluid flux balance is considered essential for stability of multi-phase flow simulations. This is made precise in the sense that Stokes' theorem must hold exactly for some volumes $\omega \in \mathcal{T}$, comprising a reasonably fine partitioning \mathcal{V} of the domain:

$$\int_{\partial\omega} q \cdot n \, dS = \int_{\omega} \psi \, dV. \quad (4.1.3)$$

- (B) *Accuracy on coarse grids*: In geological porous media, the regularity of coefficients is very low, and thus accuracy is to a large extent equated with accurate handling of material discontinuities, in particular when the discontinuities coincide with the boundaries $\partial\omega$.
- (C) *Flexible grids*: While many early simulation studies were conducted on regular grids, both anisotropic coefficients, as well as complex geological features, motivates discretizations suitable for complex grids.
- (D) *Symmetric and positive definite discretization matrix*: A symmetric and positive definite (SPD) matrix allows for application of Conjugate Gradient solvers, which have good performance, in particular with respect to memory usage.
- (E) *Local flux stencils*: The size of the discretization stencil directly impacts both memory usage, but also floating point operations associated with matrix-vector multiplication. A local expression for the flux (as opposed to a post-processed flux), allows the use of automatic differentiation software for constructing the Jacobian for non-linear problems.
- (F) *Monotonicity of solution*: The continuous problem has the property that for a positive source term ψ , and zero-pressure boundary conditions, the pressure p that solves Eqs. (4.1.1–4.1.2) is guaranteed to be positive everywhere in the interior of the domain. Monotonicity is closely related to spurious oscillations, which is a major problem for multi-phase simulations.

- (G) *Accuracy on fine grids*: As the discretization grid is refined, the truncation error of the discrete approximation should vanish, and the discrete approximation should converge to the continuous solution.

It is perhaps intuitive that all these properties cannot be achieved optimally by any linear discretization. By the late 1980s, it was well understood that none of the existing methods at the time achieved all the favorable properties [9]. These were standard Galerkin finite elements (P1-P1 finite elements or similar), Petrov-Galerkin finite elements (P1-P0 finite elements on staggered grids, also known as Control Volume Finite elements), Mixed Finite Elements (lowest-order Raviart-Thomas for flux and P0 for pressure), or Two-Point Finite Volume methods (still the industry standard for practical simulation). We will make a quick summary of the weakness of each of these discretization methods, to better understand the relative advantages (and disadvantages) of the Multi-Point Finite Volume methods.

Galerkin finite elements is perhaps the most common discretization method available in the field of computational mathematics (for an introduction, see textbooks [10, 11]). This discretization method is well-suited for simplicial and Cartesian grids, but for more complex grids the definition of the elements becomes more complicated. While Galerkin finite elements have both local stencils as well as lead to SPD matrices, they need post-processing to obtain a local flux balance [12], and are not particularly well suited to discontinuous permeability coefficients [13].

Petrov-Galerkin finite elements, or Control-Volume Finite Elements (CVFE) as we will refer to the method, attempts to improve over the standard finite element methods by introducing a dual grid around each vertex of the primal grid [14]. On this dual grid, piecewise constant test functions are chosen, so that the local Stokes' equation holds exactly. Nevertheless, the pressure solution p is still represented by finite element functions, so the primal grid must still be relatively simple, and no accuracy is gained over finite element methods with respect to discontinuous permeability coefficients. Furthermore, due to the different choice of elements for the trial and solution spaces, the symmetry of the discretization matrix is lost.

Mixed finite elements (MFE) is another way of generalizing finite element methods [15]. In this approach, the first-order structure indicated in Eqs. (4.1.1–4.1.2) is retained explicitly, where the pressure is represented as piecewise constant, while the flux is in a relatively simple space whose divergence is piecewise constant (for relatively simple grids this is the lowest-order Raviart-Thomas space, but defining this space becomes non-trivial even for perturbations of Cartesian grids [16, 17]). The mixed-finite element method is accurate for material contrasts, and has an explicit flux balance. On the other hand, it does not immediately lead to an SPD matrix (without hybridization) and has relatively poor monotonicity properties [18].

Two-Point Finite Volume (TPFV) methods in are a sense the simplest methods satisfying the flux balance. The methods consist of imposing Eq. (4.1.3) on any polyhedral partition \mathcal{T} of the domain, and then constructing an approximation to $q \cdot n$ using the pressure values in the two neighbors of any face of the polyhedral partitioning. This simplicity leads to a method satisfying all desired properties (A–F) above, and one could ask if it is the perfect method. Unfortunately, the method is

indeed too simple—and in contrast to the three preceding methods discussed—the truncation error only vanishes on a quite restrictive class of grids, and thus in general one can observe convergence to the wrong solution (see e.g. [19]).

The above summary gives some impression of the state-of-the-art when the multi-point methods were developed. As the name suggests, this family of methods attempts to develop a discretization with favorable properties, not by improving on finite element methods, but rather with basis in the TPFV method. More recently, it has been shown how this development ties back to developments also in the finite element literature, a topic that we will return to at several points in later sections of the chapter.

The first multi-point methods were introduced in two independent papers at the ECMOR conference at Røros, Norway in 1994 [1, 2], and an excellent introduction to the Multi-Point Flux Approximation (MPFA), and references to the early literature, can be found by Aavatsmark [20]. However, since that introductory text was written, these methods have seen significant development, both in terms of applicability to complex problems, but also in terms of a maturing of our understanding of the multi-point methods as a general discretization approach. Our goal with this chapter is therefore to provide a contemporary account of these methods. With concrete reference to Aavatsmark [20], the current text covers a consistent treatment of right-hand-side terms in the constitutive laws, more general continuity conditions, discretization of elasticity and poro-elasticity, and a review of the mathematical analysis of these methods. Moreover, our presentation of the method is based on a more abstract construction than in the introduction by Aavatsmark, more suited to general polyhedral grids.

We preempt some of the later discussion by already announcing some of the main features of the multi-point methods. They are developed to have local flux balance, (relatively) small stencils, and be accurate for challenging grids, including polyhedral grids, and handle accurately heterogeneous permeability fields. It has also been shown that the convergence properties of the methods are good, both for smooth and non-smooth data. The cost of these advantages is that the discretization matrix is only symmetric for simplicial grids, although it is in general positive definite. Monotonicity of the discretization holds subject to conditions which are not prohibitively harsh, but still strict enough to affect some realistic cases.

The chapter is subdivided as follows. In Sect. 4.2, we will develop the general principles of multi-point finite volume methods, which we refer to as MPxA methods. We will see that these general principles imply a family of methods for elliptic problems with conservation structure. Building on this, we will in Sect. 4.3 apply the general principles to three concrete problems: First, fluid flow in porous media, as is the classical motivation for these methods, and leads to the MPFA methods. Secondly, to momentum balance in elastic solids, which leads to the so-called Multi-Point Stress Approximation (MPSA) methods. The MPSA methods are naturally suited combined problem of fluid flow in elastically deformable materials, also known as poroelasticity. Moreover, we also consider the case of thermo-poroelasticity, which includes an advective term in addition to the coupling between heat, flow, and deformation. Having developed the discretization methods for these concrete applications, we will review the mathematical and numerical properties of these methods, as has been

reported in literature in Sect. 4.4, together with applications to real-world data-sets in Sect. 4.5.

4.2 Multi-point Finite Volume Methods

We will structure our presentation of the general construction of multi-point finite volume methods in two parts. In Sect. 4.2.1, we will present the primal grid and the conservation structure, which is common to all finite volume methods. In Sect. 4.2.2 we will detail the particular choices which give rise to the so-called multi-point finite volume methods. Our goal throughout the exposition is to be both general yet pedagogical. As a result, the presentation, in particular in Sect. 4.2.2, deviates significantly from the presentation of these methods found in research articles. In Sect. 4.2.3, we will discuss aspects related to efficient and stable implementation of these methods.

All the derivations in this section are agnostic to the conservation law of interest (mass or momentum). In order to emphasize this generality, we will in this section refer to MPFA or MPSA methods by the generic acronym MPxA. On the other hand, in order to allow for a streamlined presentation, we will present a rather general concept of the so-called O-methods, thereby excluding the less common variants of the MPFA methods, namely the so-called L-, U-, and Z-methods [3, 21–23].

4.2.1 Finite Volume Methods

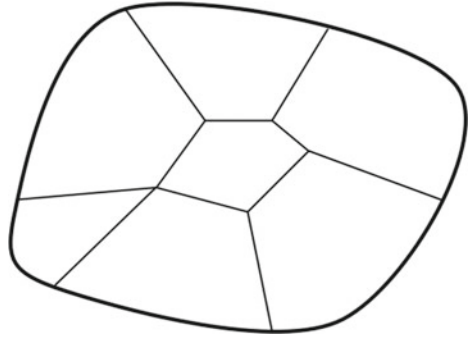
This section gives the basic notion of a finite volume method for a conservation law, following e.g. [24, 25]. As alluded to in the introduction, a conservation law is a statement of the form

$$\frac{d}{dt} \int_{\omega} u \, dV + \int_{\partial\omega} n \cdot \tau \, dS = \int_{\omega} r \, dV. \quad (4.2.1)$$

Given that we have a domain of interest $\Omega \subset \mathbb{R}^n$, where n is the dimension of the problem, the conservation law is interpreted as follows. We are concerned with a conserved quantity u (e.g. mass, momentum or energy) within *any* measurable subdomain $\omega \subset \Omega$ with external normal vector n . The conservation law asserts that the accumulation of u within ω , is determined by a flux field τ , which may represent mass flux, energy flux or stress, and volumetric sources r .

As we are concerned with spatial discretization, we will in the remainder of this section disregard the temporal term, and only consider the steady state of Eq. (4.2.1). We note that when the variables are sufficiently regular, Eqs. (4.1.2) and (4.2.1) are equivalent due to Stokes' theorem. In absence of such regularity, Eq. (4.2.1) is a

Fig. 4.1 The domain Ω shown in thick solid black line together with the finite volume grid ω_i (thinner solid black lines corresponding to faces between cells). Note that the cells may be polyhedral, and that more than three cells may meet at a vertex



more general statement than (4.1.2), and this is the motivation for discretizing Equation (4.2.1) directly. Discretization methods that are developed from this viewpoint are known as *finite volume methods*.

In order to construct a numerical method from Eq. (4.2.1), we consider a non-overlapping partitioning of Ω into a finite set of N subdomains $\omega_k \in \mathcal{T}$, for $k = 1 \dots N$. An example of such a partitioning for $N = 7$ is given by the solid lines in Fig. 4.1.

The subdomains ω_k are referred to as control volumes, or simpler, *cells*. For any two cells ω_{k_1} and ω_{k_2} that are neighbors, in the intersection of their boundaries is measurable, $\text{meas}(\partial\omega_{k_1} \cap \partial\omega_{k_2}) > 0$, we refer to this intersection as a *face*, and the collection of faces is denoted \mathcal{F} . We extend the definition of a face to also account for intersections with the boundary, such that if $\text{meas}(\partial\omega_{k_1} \cap \partial\Omega) > 0$, this also defines a face, and is included in \mathcal{F} . In particular, we recognize that all faces of, say, ω_k is a subset of \mathcal{F} , and we denote this subset as \mathcal{F}_k . These definitions allow us to rewrite (the steady state of) Eq. (4.2.1) as

$$\sum_{\sigma \in \mathcal{F}_k} \int_{\sigma} n_{\sigma,k} \cdot \tau \, dS = \int_{\omega_k} r \, dV. \quad (4.2.2)$$

Equation (4.2.2) must hold for any k , due to Eq. (4.2.1). Moreover, we recognize that it is tempting to define the *normal flux out of ω_k through σ* as

$$q_{\sigma,k} \equiv \int_{\sigma} n_{\sigma,k} \cdot \tau \, dS. \quad (4.2.3)$$

A *finite volume method* is then any method that can be written on the form

$$\sum_{\sigma \in \mathcal{F}_k} q_{\sigma,k} = \int_{\omega_k} r \, dV \quad \text{for all } \omega_k \in \mathcal{T} \quad (4.2.4)$$

The finite volume method has local flux balance if for any $\sigma = \partial\omega_{k_1} \cap \partial\omega_{k_2}$, it holds that

$$q_{\sigma,k_1} = -q_{\sigma,k_2}. \quad (4.2.5)$$

Since we will only consider methods with local flux balance, we therefore identify the face flux as the flux from the cell with the lower index, i.e. for $k_1 < k_2$, then we define

$$n_\sigma \equiv n_{\sigma,k_1} \quad \text{and} \quad q_\sigma \equiv q_{\sigma,k_1}.$$

4.2.2 MPxA Finite Volume Methods

The basic construction of a finite volume method is agnostic to how the numerical flux field q_σ is obtained, and indeed is common for hyperbolic, parabolic and elliptic conservation laws. As stated in the introduction, this chapter deals with methods for problems where there is a proportionality between q and ∇u , as indicated in Eq. (4.1.2). In the absence of the time-derivative, such conservation laws are referred to as elliptic, and include Fourier, Fick, Darcy, Hooke and other constitutive laws.

To be precise, we will thus consider constitutive laws on the form

$$\tau = \mathbb{C}\nabla u + g. \quad (4.2.6)$$

Here \mathbb{C} is understood to be a local linear operator from the space of functions spanned ∇u , to the space associated with the flux τ . The residual g is in practice derived from a known external force, we will consider it as such. The precise definition of the function spaces depends on the regularity imposed on u and q , but also whether one considers scalar or vector equations. As this precision will not be important for introducing the numerical methods, we will omit these details here (for a detailed exposition of the function spaces, see e.g. [24, 26, 27]).

4.2.2.1 Grid Structure

The MPxA methods are a family of methods for approximating the normal flux q_σ from Eq. (4.2.6), based on a core set of foundational principles. To construct an MPxA approximation, additional structure must be introduced relative to the bare-bones finite volume structure given in Sect. 4.2.1. In particular, we associate with each cell ω_k a point x_k , which we will refer to as its center. The point x_k should be chosen such that ω_k is *star-shaped* relative to x_k (this is always possible for simplexes, but may not be possible for non-convex polyhedral). Moreover, we identify that the partition \mathcal{T} gives rise to vertexes of the grid (intersection points of multiple cells), which we will refer to as \mathcal{V} . We will denote the subset of \mathcal{V} that are logical vertexes of ω_k as \mathcal{V}_k , such that every point $s \in \mathcal{V}_k$ satisfies $s \in \partial\omega_k$. Conversely, we will denote the subset of \mathcal{T} that meet at a vertex $s \in \mathcal{V}$ as \mathcal{T}_s , such that for every subdomain

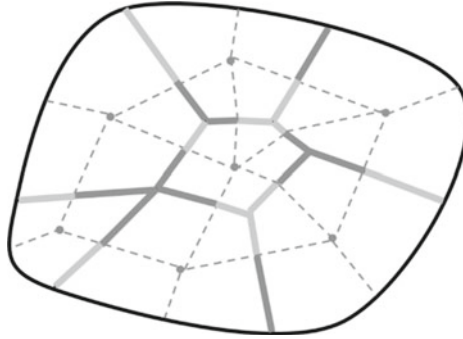


Fig. 4.2 This figure provides an illustration of the extra grid structure used for MPxA methods relative to the basic finite volume grid shown in Fig. 4.1. The division of faces into subfaces is indicated by two nuances of grey, while cell centers are indicated by dots. The dual grid is indicated by dashed lines, and the subgrid is thus obtained as the quadrilaterals having two dashed and two solid-grey boundaries

$\omega_k \in \mathcal{T}_s$, it again holds that $s \in \partial\omega_k$. The definitions of \mathcal{F}_s and \mathcal{V}_σ , for all $\sigma \in \mathcal{F}$, are analogous.

With the preceding definitions, we introduce a refinement of the finite volume grid structure, as shown in Fig. 4.2. First, we refine the faces of the grid as follows: Let every face $\sigma \in \mathcal{F}$ be partitioned into *subfaces* $\tilde{\sigma} \in \mathcal{S}_\sigma$, such that each subface contains exactly one vertex of σ . Thus, if the set of all subfaces is denoted \mathcal{S} , then for any pair of a face $\sigma \in \mathcal{F}$ and a vertex $s \in \mathcal{V}_\sigma$, there is a unique element of $\mathcal{S}_{\sigma,s}$. Extending the notational convention above, we denote the subfaces of ω_k meeting at a vertex $s \in \mathcal{V}_k$ as $\mathcal{S}_{k,s}$.

We introduce the following definition of a dual grid: For each vertex $s \in \mathcal{V}$, let the dual cell $\omega_s^* \in \mathcal{T}_s^*$ be defined such that subfaces in $\tilde{\sigma} \in \mathcal{S}_s$ are contained in ω_s^* , and the cell-centers x_k of the cells $\omega_k \in \mathcal{T}_s$ are on the boundary of the dual cell, i.e. $x_k \in \partial\omega_s^*$. Finally, let the dual cells be a non-overlapping partitioning of the domain Ω . The intersection of the primal and dual grids creates an even finer grid $\tilde{\mathcal{T}}$, elements of which are uniquely defined by a cell and a corner. Thus, the subcell $\tilde{\omega}_{k,s} \in \tilde{\mathcal{T}}$ is defined as $\tilde{\omega}_{k,s} = \omega_k \cap \omega_s^*$. Again, we retain the same conventions on subscripts, so that in particular $\tilde{\mathcal{T}}_s$ are the subcells adjacent to the corner s .

4.2.2.2 Approximation Spaces

The MPxA approximations do not attempt to construct the numerical flux q_σ over a face $\sigma \in \mathcal{F}$ directly, but instead construct approximations over the subfaces $\tilde{\sigma}$. The subface normal fluxes $\tilde{q}_{\tilde{\sigma}}$ are then subsequently assembled such that

$$q_\sigma = \sum_{\tilde{\sigma} \in \mathcal{S}} \tilde{q}_{\tilde{\sigma}}. \quad (4.2.7)$$

The MPxA methods for the fluxes over the subfaces $\tilde{\sigma} \in \mathcal{S}_s$ are based on the following common seven ingredients:

- (i) A linear approximation $u_{k,s}(x)$ to the potential field within each subcell $\tilde{\omega}_{k,s} \in \tilde{\mathcal{T}}$.
- (ii) A constant approximation $\tau_{k,s}$ to the flux field within each subcell $\tilde{\omega}_{k,s} \in \tilde{\mathcal{T}}$.
- (iii) A constant approximation g_k to the external force field within each cell $\omega_k \in \mathcal{T}$.
- (iv) A relation between the potential fields $u_{k,s}(x)$, flux fields $\tau_{k,s}$, and force field g_k , consistent with (4.2.6).
- (v) Local flux balance over each subface $\tilde{\sigma} \in \mathcal{S}_s$ in the sense of (4.2.3) and (4.2.5):

$$\int_{\tilde{\sigma}} \tau_{k_1,s} \cdot n_{\tilde{\sigma}} dS = \int_{\tilde{\sigma}} \tau_{k_2,s} \cdot n_{\tilde{\sigma}} dS \equiv \tilde{q}_{\tilde{\sigma}}. \quad (4.2.8)$$

- (vi) Continuity between the linear potential field approximations at the cell centers x_k , i.e. for a given $\omega_k \in \mathcal{T}$, and any $s_1, s_2 \in \mathcal{V}_k$

$$u_{k,s_1}(x) = u_{k,s_2}(x) \equiv u_k. \quad (4.2.9)$$

- (vii) A minimization of a quadratic penalty function \mathcal{M}_s , measuring the discontinuity of the linear potential fields across subfaces $\tilde{\sigma} \in \mathcal{S}_s$. The precise choice of the penalty function used in this minimization gives rise to variations in the method, as detailed in the next section.

If one of the subfaces $\tilde{\sigma} \in \mathcal{S}_s$ is on the boundary of the domain Ω , additional conditions apply, as discussed in Sect. 4.2.2.3. We emphasize that the method formulation given below applies independent of the boundary condition assigned.

The core MPxA ingredients are perhaps most intuitively understood by the following interpretation: The potential field is piecewise linear function relative to the fine grid obtained from the intersection of the primal and dual grid, with a minimum of continuity imposed in order to allow for a compromise between a consistent discretization and flexible grids, while always allowing for a static condensation in terms of cell-center potentials alone. The numerical flux is a derived quantity from the potential field.

The continuity requirements on the piecewise linear potential field are chosen to have a very particular structure. With reference to Fig. 4.2, continuity conditions are essentially imposed at cell centers (points in the figure), and across subfaces (various thick grey lines). Thus *no continuity is explicitly enforced over the edges of the dual grid*. This observation justifies the claim that all degrees of freedom in the construction can be locally eliminated with respect to the cell-center potentials u_k . This is the key to an efficient numerical implementation, and also implies that the resulting discretization matrix has minimum size (potential variables in the cell centers).

To make the above claims more precise, we now introduce discrete operators that allow for an efficient presentation of MPxA methods in general. A more detailed discussion with focus on implementation is provided in Sect. 4.2.3, while application of the general framework, that is, identification of the discrete operators for specific equations is considered in Sect. 4.3

We denote vectors of variables by bold letters, and matrixes by capitals. First, let the finite volume method, Eq. (4.2.4), be represented in matrix form in terms of a divergence matrix \mathbf{D} (simply a summation over fluxes, accounting for sign convention). Similarly, we represent the summation over subfluxes, weighted by the area of the subcells, as defined in Eq. (4.2.7) as $\Sigma_{\mathcal{F}}$. Then Eqs. (4.2.4) and (4.2.7) are equivalent to

$$\mathbf{D}\mathbf{q} = \mathbf{D}\Sigma_{\mathcal{F}}\tilde{\mathbf{q}} = \mathbf{r}. \quad (4.2.10)$$

Furthermore, let the vector of cell center potentials be denoted \mathbf{u} , and the vector containing the degrees of freedom for the linear pressure variations in each subcell $\tilde{\mathbf{u}}$. We denote the operator that extracts \mathbf{u} from $\tilde{\mathbf{u}}$ as \mathbf{E} , such that

$$\mathbf{u} = \mathbf{E}\tilde{\mathbf{u}}. \quad (4.2.11)$$

The (continuous) gradient induces a map from $\tilde{\mathbf{u}}$ to piece-wise constant vector fields on each subcell, and we denote the matrix representation of this map as \mathbf{G} . We denote the discrete constitutive law by the matrix \mathbf{B} , such that Eq. (4.2.6) becomes

$$\boldsymbol{\tau} = \mathbf{B}\mathbf{G}\tilde{\mathbf{u}} + \mathbf{E}^*\mathbf{g}. \quad (4.2.12)$$

Here \mathbf{E}^* is the matrix that maps cell values to the individual subcells (in a sense dual to \mathbf{E}).

We furthermore denote by \mathbf{F} the matrix that extracts normal subface fluxes $\tilde{\mathbf{q}}$ from the fluxes $\boldsymbol{\tau}$ on the side of the face with the *lower* index (similar to definition used in (4.2.5b)), and conversely we denote by $\hat{\mathbf{F}}$ the matrix that extracts normal subface normal fluxes $\tilde{\mathbf{q}}$ from the fluxes $\boldsymbol{\tau}$ on the side of the face with the *higher* index. The flux balance and the definition of the subface fluxes is summarized in matrix form as

$$\mathbf{F}\boldsymbol{\tau} = \hat{\mathbf{F}}\tilde{\mathbf{q}}, \quad (4.2.13)$$

$$\tilde{\mathbf{q}} = \mathbf{F}\boldsymbol{\tau}. \quad (4.2.14)$$

Finally, let the penalty function $\mathcal{M}(\tilde{\mathbf{u}}) = \sum_{s \in \mathcal{V}} \mathcal{M}_s(\tilde{\mathbf{u}})$ be the (still quadratic) measure of the total discontinuity of $\tilde{\mathbf{u}}$ across subfaces $\tilde{\sigma} \in \mathcal{S}$.

Then the MPxA method can then be explicitly defined as follows:

Definition 4.2.1 (*generalized MPxA (global)*) Let $\mathcal{N}_{\mathbf{u},\mathbf{g}}$ be the null-space of the constraints given in Eqs. (4.2.11–4.2.13), subject to a given potential \mathbf{u} and an external

field \mathbf{g} . Then the MPxA method is defined by the pair $(\tilde{\mathbf{u}}, \boldsymbol{\tau}) \in \mathcal{N}_{\mathbf{u}, \mathbf{g}}$ such that

$$(\tilde{\mathbf{u}}, \boldsymbol{\tau}) = \arg \min_{(\tilde{\mathbf{u}}', \boldsymbol{\tau}') \in \mathcal{N}_{\mathbf{u}, \mathbf{g}}} \mathcal{M}(\tilde{\mathbf{u}}'), \quad (4.2.15)$$

and the numerical flux is defined as $\mathbf{q} = \mathbf{Q}_u \mathbf{u} + \mathbf{Q}_g \mathbf{g} \equiv \Sigma_{\mathcal{F}} \mathbf{F} (\mathbf{B} \mathbf{G} \tilde{\mathbf{u}} + \mathbf{E}^* \mathbf{g})$.

We emphasize that \mathcal{M} is a sum of local quadratic measures \mathcal{M}_s for each $s \in \mathcal{V}$, and moreover that the constraints (4.2.11–4.2.13) are all local expressions relative to subcells $\tilde{\omega} \in \tilde{\mathcal{T}}_s$. That is to say that the matrices $\mathbf{B}, \mathbf{E}, \mathbf{E}^*, \mathbf{F}, \hat{\mathbf{F}}$ and \mathbf{G} can all be written as a sum of local matrices for each vertex, e.g. $\mathbf{B} = \sum_{s \in \mathcal{V}} \mathbf{B}_s$, where the local matrices such as \mathbf{B}_s are in terms of degrees of freedom only associated with the subcells in $\tilde{\mathcal{T}}_s$. This gives rise to the local formulation of MPxA, which is defined as

Definition 4.2.2 (*generalized MPxA (local)*) For a vertex $s \in \mathcal{V}$, and for a given potential \mathbf{u} , let $\mathcal{N}_{\mathbf{u}, \mathbf{g}, s}$ be the null-space of the constraints

$$\mathbf{u} = \mathbf{E}_s \tilde{\mathbf{u}}_s, \boldsymbol{\tau}_s = \mathbf{B}_s \mathbf{G}_s \tilde{\mathbf{u}}_s + \mathbf{E}_s^* \mathbf{g}, \quad \text{and} \quad \mathbf{F}_s \boldsymbol{\tau}_s = \hat{\mathbf{F}}_s \boldsymbol{\tau}_s. \quad (4.2.16)$$

in terms of the local degrees of freedom $\tilde{\mathbf{u}}_s$ and $\boldsymbol{\tau}_s$ on subcells in $\tilde{\mathcal{T}}_s$. Then the local problem for the MPxA method is defined by the pair $(\tilde{\mathbf{u}}_s, \boldsymbol{\tau}_s) \in \mathcal{N}_{\mathbf{u}, \mathbf{g}, s}$ such that

$$(\tilde{\mathbf{u}}_s, \boldsymbol{\tau}_s) = \arg \min_{(\tilde{\mathbf{u}}'_s, \boldsymbol{\tau}'_s) \in \mathcal{N}_{\mathbf{u}, \mathbf{g}, s}} \mathcal{M}_s(\tilde{\mathbf{u}}'_s), \quad (4.2.17)$$

and the numerical flux is assembled as $\mathbf{q} = \mathbf{Q}_u \mathbf{u} + \mathbf{Q}_g \mathbf{g} \equiv \Sigma_{\mathcal{F}} \sum_{s \in \mathcal{V}} \mathbf{F}_s (\mathbf{B}_s \mathbf{G}_s \tilde{\mathbf{u}}_s + \mathbf{E}_s^* \mathbf{g})$.

The local formulation of MPxA is clearly equivalent to the global formulation. As a consequence, the minimization problems (4.2.15) are local linear saddle-point problems of modest size for each vertex of the grid, and can be solved efficiently (and in parallel, if desired), using any standard explicit linear solver. We shall return to the structure of the local problems in Sect. 4.2.3.3.

Since the minimization problem is quadratic, the numerical flux is a linear function of the potential \mathbf{u} . The MPxA finite volume discretization matrix is obtained by combining the numerical flux and the finite volume method, Eq. (4.2.10), to obtain the linear system

$$\boxed{D\mathbf{Q}_u \mathbf{u} = \mathbf{r} - D\mathbf{Q}_g \mathbf{g}} \quad (4.2.18)$$

We will return to the properties of the matrix $D\mathbf{Q}_u$ in Sect. 4.4.

4.2.2.3 Penalty Functions

An attractive feature of the MPxA methods is that the penalty functions \mathcal{M}_s used in minimization problems (4.2.15) can be chosen to enhance various properties of the MPxA methods.

The natural starting point for developing quadratic minimization problems to penalize the discontinuities in the linear pressure approximation, is to consider the norm of the discontinuities across subfaces [28]. Thus, for every subface $\tilde{\sigma} \in \mathcal{S}$, we define the penalty function

$$\mathcal{M}_{\tilde{\sigma}}(\mathbf{u}) \equiv \int_{\tilde{\sigma}} (u_{k_1,s}(x) - u_{k_2,s}(x))^2 dS. \quad (4.2.19)$$

As previously, k_1 , k_2 and s are the indexes such that $\tilde{\omega}_{k_1,s}$ and $\tilde{\omega}_{k_2,s}$ are the two subcells sharing the subface $\tilde{\sigma}$. Any positive linear combination of the subface discontinuity measure will be a new measure of discontinuity, and thus it follows that for any vertex, we make the natural definition

$$\mathcal{M}_s(\mathbf{u}) \equiv \sum_{\tilde{\sigma} \in \mathcal{S}_s} c_{\tilde{\sigma}} \mathcal{M}_{\tilde{\sigma}}(\mathbf{u}). \quad (4.2.20)$$

The weights $c_{\tilde{\sigma}}$ can in principle be chosen to optimize the method, although the simple choice $c_{\tilde{\sigma}} = 1$ appears sufficient in practice.

Since the potentials $u_{k,s}(x)$ are approximated as linear, the integral in Eq. (4.2.19) is a quadratic function on the subface $\tilde{\sigma}$, and can be exactly evaluated using only a low number of quadrature points (two in 2D and four in 3D). The majority of MPxA literature simplify the minimization problem further, and consider only a single quadrature point. We will for historic reasons denote this minimization with the Greek letter η , and introduce the simplified penalty functions

$$\mathcal{M}_{\tilde{\sigma}}^{\eta}(\mathbf{u}) \equiv (u_{k_1,s}(x_{\tilde{\sigma}}^{\eta}) - u_{k_2,s}(x_{\tilde{\sigma}}^{\eta}))^2. \quad (4.2.21)$$

The definition of the simplified penalty functions is completed by specifying the points $x_{\tilde{\sigma}}^{\eta}$. The common choice is obtained if the face σ subdivided into subfaces relative to a central point x_{σ} . Then let $\eta \in [0, 1)$, and define

$$x_{\tilde{\sigma}}^{\eta} = x_{\sigma} + \eta \frac{x_s - x_{\sigma}}{|x_s - x_{\sigma}|}. \quad (4.2.22)$$

In this expression, we have used x_s to denote the coordinate of vertex s . Given $\mathcal{M}_{\tilde{\sigma}}^{\eta}(\mathbf{u})$, the full expression for minimization $\mathcal{M}_s^{\eta}(\mathbf{u})$ is defined analogously to Eq. (4.2.20).

The main advantage of the simplified penalty functions \mathcal{M}_s^{η} , is that it can be shown that for many common grid types (all grids in 2D, and e.g. Cartesian or simplicial grids in 3D, but not grids containing pyramids), the optimal value of the minimization

problems (4.2.17) is indeed 0. That is to say, that the minimization problem can be omitted, and be replaced by the direct condition that

$$\mathcal{M}_\sigma^\eta(\mathbf{u}) = 0 \quad (4.2.23)$$

for all subfaces $\tilde{\sigma} \in \mathcal{S}$. When this condition holds, the pressure is indeed continuous across the subface exactly at the point x_σ^η , and this point is then referred to as a continuity point in the literature [20].

Two particular choices of x_σ^η are particularly appealing and common in practice: $\eta = 0$ leads to a simple method that has the best monotonicity properties on quadrilaterals [29]. $\eta = \frac{1}{3}$ leads to a method which has a symmetric discretization method on simplexes [30, 31]. Another possible choice is to take $\eta = \frac{1}{2}$, which gives a high-order method on smooth problems on quadrilaterals [32]. We will return to this topic in more detail in Sect. 4.4 of the chapter.

4.2.3 Implementation Aspects

To further explore the approximation properties and implementation of the MPxA methods, it is instructive to consider the local problem (4.2.2) in some more detail. As discussed above, the approximation spaces on the subcells are not rich enough to allow full continuity over subfaces $\tilde{\sigma} \in \mathcal{S}_s$, thus MPxA can be interpreted as a discontinuous Galerkin method with a particular set of continuity constraints on potentials and normal fluxes over the subfaces. Critical for efficiency and implementation, we exploit the two-scale approximation, in that the (fine scale) degrees of freedom associated with the potential gradients on the subcells can be eliminated by static condensation around each vertex s . This leaves a method where only the (coarse scale) cell center degrees of freedom enter into the global problem.

4.2.3.1 Local Minimization Problem

To be concrete, we make the choice of representing the linear potential field, $u_{k,s}$ in a subcell by its value in the cell center, u_k , and the (constant) components of its gradient, which we denote $h_{k,s}$. To understand an efficient implementation of the local linear system set in Definition 4.2.2, it is instructive to discuss the size of the matrices that form the problem. To that end, let $n_f = |\mathcal{S}_s|$ be the number of subfaces meeting in s , and similarly $n_c = |\tilde{\mathcal{T}}_s|$ be the number of cells that has s as vertex. The number of faces with Neumann and Dirichlet boundary conditions are denoted $n_{\tilde{\sigma},N}$ and $n_{\tilde{\sigma},D}$, respectively. Let d be the dimension of the potential field u , this will be 1 for scalar equations and the spatial dimension n for vector equations, and $n_{\tilde{\sigma},q}$ be the number of quadrature points on subface $\tilde{\sigma}$.

As degrees of freedom in the local linear system, we use the cell center potentials \mathbf{u}_s in \mathcal{T}_s and the components of the gradients in the subcells $\tilde{\mathcal{T}}_s$, represented by \mathbf{h}_s , so that the full vector of local unknowns is $\tilde{\mathbf{u}}_s = (\mathbf{u}_s, \mathbf{h}_s)^T$. This representation has the advantage that the matrices \mathbf{E} and \mathbf{G} take the particularly simple form

$$\begin{aligned} \mathbf{E}_s \tilde{\mathbf{u}}_s &= (\mathbf{I} \mathbf{0}) \tilde{\mathbf{u}}_s = \mathbf{u}_s, \\ \mathbf{G}_s \tilde{\mathbf{u}}_s &= (\mathbf{0} \mathbf{I}) \tilde{\mathbf{u}}_s = \mathbf{h}_s. \end{aligned}$$

The flux field can therefore be recovered directly from \mathbf{h}_s using (4.2.12), where we see that the (local) matrix that contains the constitutive law, \mathbf{B}_s is of size $(n_c \cdot d \cdot n) \times (n_c \cdot d \cdot n)$.

The computation of subface normal fluxes is split into internal and boundary faces: The internal faces are covered by the matrices \mathbf{F}_s^I and $\hat{\mathbf{F}}_s^I$, both of size $(d \cdot (n_f - n_{\tilde{\sigma},N} - n_{\tilde{\sigma},D})) \times (n_c \cdot d \cdot n)$, which represent the multiplication by subface normal vectors of the flux on the neighboring cells of lower and higher index, respectively. The normal flux over faces with Neumann and Dirichlet boundary conditions is computed from the matrices \mathbf{F}_s^N and \mathbf{F}_s^D , of size $(d \cdot n_{\tilde{\sigma},N}) \times (n_c \cdot d \cdot n)$ and $(d \cdot n_{\tilde{\sigma},D}) \times (n_c \cdot d \cdot n)$, respectively. The evaluation of the potential at the subface quadrature points is similarly split: For internal subfaces, the matrix \mathbf{M}_s^I of size $\left(d \cdot \sum_{\tilde{\sigma} \in \mathcal{S}_s^i} n_{\tilde{\sigma},q}\right) \times (n_c \cdot d \cdot n)$ has elements composed of the distance from cell centers to subface quadrature points; here \mathcal{S}_s^i denotes the internal subfaces of vertex s . $\hat{\mathbf{M}}_s^I$ is the corresponding matrix for neighboring cells of higher index, while \mathbf{M}_s^D and \mathbf{M}_s^N are assigned for subfaces with Dirichlet and Neumann boundary conditions, respectively. Finally, we similarly define the matrix $\hat{\mathbf{E}}_s^*$ relative to \mathbf{E}_s^* , and the internal and boundary components.

With the above definitions, the penalty term is stated in terms of the local variables as

$$\mathcal{M}_s(\tilde{\mathbf{u}}_s) = \mathcal{M}_s(\mathbf{u}_s, \mathbf{h}_s) = \boldsymbol{\Sigma} \left\| \left((\mathbf{E}_s^{*,I} - \hat{\mathbf{E}}_s^{*,I}) \mathbf{u}_s + (\mathbf{M}_s^I - \hat{\mathbf{M}}_s^I) \mathbf{h}_s \right) \right\|^2, \quad (4.2.24)$$

where the norm is the sum of the integral (4.2.19) taken over all subfaces $\tilde{\sigma} \in \mathcal{S}_s$, and where $\boldsymbol{\Sigma}$ a diagonal matrix containing the quadrature weights for the integrals. The minimization problem is subject to the constraints that \mathbf{u}_s and \mathbf{g} are given, as well as

$$\boldsymbol{\tau}_s = \mathbf{B}_s \mathbf{h}_s + \mathbf{E}_s^* \mathbf{g}, \mathbf{F}_s \boldsymbol{\tau}_s = \hat{\mathbf{F}}_s \boldsymbol{\tau}_s, \mathbf{F}_s^N \boldsymbol{\tau}_s = \mathbf{q}_s^N, \mathbf{E}_s^{*,D} \mathbf{u}_s + \mathbf{M}_s^D \mathbf{h}_s = \mathbf{u}_s^D. \quad (4.2.25)$$

Here, we have retained the explicit dependency of the constraints on $\boldsymbol{\tau}_s$, and introduced the Neumann and Dirichlet conditions as \mathbf{q}_s^N and \mathbf{u}_s^D , respectively. These conditions are void if none of the subfaces around vertex s are located on the domain boundary. We have assumed that the number of quadrature points on subfaces with

Dirichlet conditions is sufficiently low for the relevant constraint to be fulfilled exactly; as an alternative, the condition $\mathbf{M}_s^D \mathbf{u}_s = \mathbf{u}_s^D$ can be incorporated into the minimization problem. While the matrices \mathbf{F}_s^D and \mathbf{M}_s^N are not used in the method constructions, they can be used in post-processing to calculate the normal flux through Dirichlet faces and potential on Neumann faces, respectively.

We pause to consider the size of the local problem (4.2.24)–(4.2.25), and specifically compare the number of degrees of freedom and constrains. We limit ourselves here to internal vertexes, similar reasoning applies to vertexes on the domain boundary. The number of subcell gradient degrees of freedom is $n_k \cdot d \cdot n$, while there are $n_{\bar{\sigma}} \cdot d$ equations for flux continuity, and $\left(d \cdot \sum_{\bar{\sigma} \in \mathcal{S}_s} n_{\bar{\sigma},q} \right)$ quadrature points for potential continuity. If $n = 2$, $n_{\bar{\sigma}} = n_c$ independent of the cell type thus if each subfaces is assigned a single quadrature point, $n_{\bar{\sigma},q} = 1$, then the number of equations equals the number of gradient unknowns. In 3d, the situation is more nuanced: For simplex and logically Cartesian grids, $\frac{n_{\bar{\sigma}}}{n_c} = \frac{3}{2}$, thus with a single quadrature point on each subface, the number of equations and gradient unknowns still match. For general cell shapes, notably pyramids, this is no longer the case, and the method with a single quadrature point fails.

4.2.3.2 Expression in Terms of Coarse Degrees of Freedom

To arrive at a discretization in terms of the coarse scale, cell center, degrees of freedom, the next step is to eliminate the subcell gradients \mathbf{h}_s . Which approach is practical here depends on the size of the minimization problem. When the number of equations and gradient unknowns match, it turns out that the value of the minimization problem is in fact zero, and the problem can be formulated as a linear system on the form (with $\boldsymbol{\tau}_s$ eliminated)

$$\begin{pmatrix} \mathbf{M}_s^I - \hat{\mathbf{M}}_s^I \\ \mathbf{F}_s \mathbf{B}_s - \hat{\mathbf{F}}_s \mathbf{B}_s \\ \mathbf{F}_s^N \mathbf{B}_s \\ \mathbf{M}_s^D \end{pmatrix} \mathbf{h}_s = - \begin{pmatrix} \mathbf{E}_s^{*,I} - \hat{\mathbf{E}}_s^{*,I} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{E}_s^{*,D} \end{pmatrix} \mathbf{u}_s + \begin{pmatrix} \mathbf{0} \\ (\mathbf{E}_s^{*,I} - \hat{\mathbf{E}}_s^{*,I}) \mathbf{g}_s \\ \mathbf{q}_s^N \\ \mathbf{u}_s^D \end{pmatrix}. \quad (4.2.26)$$

This system can be solved to express \mathbf{h}_s as a linear function of \mathbf{u}_s and the right-hand side terms. We write the respective solutions as $\mathbf{h}_s^u = \mathbf{S}_u \mathbf{u}_s$, $\mathbf{h}_s^g = \mathbf{S}_g \mathbf{g}_s$, $\mathbf{h}_s^N = \mathbf{S}_N \mathbf{q}_s^N$ and $\mathbf{h}_s^D = \mathbf{S}_D \mathbf{u}_s^D$, where the matrices \mathbf{S}_* are computed from the left and right hand sides of (4.2.26). The solvability of (4.2.26) depends on the grid types and problem under consideration, and problems can arise in special cases such as some non-matching grids in 3D. However, for regular grids (simplicial and Cartesian) for the problems considered in Sect. 4.4 and 4.5 (with the exception of Sect. 4.4.3), Eq. (4.2.26) is solvable. The size of the left-hand side matrix is relatively small (see Sect. 4.2.3.3), and in our experience an explicit construction of its inverse, drawing

upon LAPACK routines for inversion of dense matrices, is an efficient option. With the explicit inverse available, the matrices $\mathbf{S}_{\{u,g,N,D\}}$ can be constructed by matrix multiplications.

The more general case with multiple quadrature points leads to a true minimization problem, and thus resolves many of the cases where (4.2.26) is not suitable. Since this is a quadratic minimization problem with linear constraints, the minimum can be found in the standard way as the solution to a linear system of equation obtained via a Lagrange multiplier vector λ . For completeness, we state this system for internal cells (i.e. with no boundary cells), for which the constrained system is:

$$\begin{aligned} & \begin{pmatrix} (\mathbf{M}_s^I - \hat{\mathbf{M}}_s^I)^T \Sigma^T \Sigma (\mathbf{M}_s^I - \hat{\mathbf{M}}_s^I) & (\mathbf{F}_s \mathbf{B}_s - \hat{\mathbf{F}}_s \mathbf{B}_s)^T \\ \mathbf{F}_s \mathbf{B}_s - \hat{\mathbf{F}}_s \mathbf{B}_s & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{h}_s \\ \lambda \end{pmatrix} \\ &= - \begin{pmatrix} 2(\mathbf{M}_s^I - \hat{\mathbf{M}}_s^I)^T \Sigma^T \Sigma (\mathbf{E}_s^{*,I} - \hat{\mathbf{E}}_s^{*,I}) \\ \mathbf{0} \end{pmatrix} \mathbf{u}_s + \begin{pmatrix} \mathbf{0} \\ (\mathbf{E}_s^{*,I} - \hat{\mathbf{E}}_s^{*,I}) \mathbf{g}_s \end{pmatrix}. \end{aligned} \quad (4.2.27)$$

Thus the local system is still linear, and has the advantage that unique solvability can be proved for many classes of grids due to the relationship to the minimization problem [28]. However, Eq. (4.2.27) contains $n_{\bar{\sigma}} \cdot d$ extra unknowns, corresponding to the Lagrange multipliers for flux continuity.

4.2.3.3 Computational Cost

We make the following comments on the local problems: Their size depends on the number of cells that share the vertex s , the dimension of the potential field, and the number of quadrature points assigned on the subfaces. In practice, 12 gradients are eliminated for scalar problems on a Cartesian 2d grid with a single quadrature points, up to on the order of a few hundred degrees of freedom for vector problems on 3d unstructured grids with multiple quadrature points [33]. The choice of quadrature rule thus has a significant impact on the overall cost of discretization, and count in favor of using few quadrature points when this is feasible. Independent of which strategy is chosen, the local problems can be solved in parallel.

4.3 Multi-point Methods for Thermo-poroelasticity

In this section, we will apply concretely the discretization concepts presented in Sect. 4.2 to the problem of thermo-poroelasticity. As in Sect. 4.2, our aim is to be pedagogical, and we will defer the discussion of mathematical properties to Sect. 4.4.

We will address the discretization for thermo-poroelasticity through four steps, following the natural progression from flow in porous media in Sect. 4.3.1, via elasticity in Sect. 4.3.2, and then combining the concepts to poroelasticity in Sect. 4.3.3. Finally, the full thermo-poroelastic discretization is presented in Sect. 4.3.4.

4.3.1 Flow in Porous Media

The basic equations for flow in porous media, as far as this exposition is concerned, are captured by the (steady state) conservation law for fluid flow, Eq. (4.2.1), and Darcy's law relating pressure gradients to fluid flux. In preparation for poroelasticity later, we will denote the pressure potential as p , and the fluid flux as τ_p , and re-state conservation and Darcy's law in terms of these variables as

$$\int_{\partial\omega} n \cdot \tau_p dS = \int_{\omega} r_p dV, \quad (4.3.1)$$

$$\tau_p = -\kappa \nabla p + g. \quad (4.3.2)$$

In a slight abuse of language, will refer to the 2nd order tensor κ as the permeability, and g as gravity.

The Multi-Point Flux Approximation (MPFA) follows exactly the general structure of MPxA methods, with Eq. (4.3.2) imposed exactly in order to define the discrete constitutive law \mathbf{B} . We will avoid restating the equations of Sect. 4.2.2, and summarize that a numerical fluid normal flux \mathbf{q} is defined by the MPFA method as a linear function of pressure, i.e. the finite volume scheme for Eqs. (4.3.1–4.3.2) is given as

$$\mathbf{D}_p \mathbf{q} = r_p, \quad (4.3.3)$$

$$\mathbf{q} = \mathbf{Q}_p p + \mathbf{Q}_g g. \quad (4.3.4)$$

Here we use the subscript p on the discrete divergence operator for this scalar problem, in order to distinguish it from the divergence operator for vector problems in the next sections.

With the choice of the simplified penalty function \mathcal{M}_s^η and thus replacing the minimization problem by Eq. (4.2.23), the method is referred to simply as the MPFA-O (η), and represents one of the two original MPFA methods [3]. When the full penalty function \mathcal{M}_s is used, we refer to this as the generalized MPFA-O method.

It is not a priori obvious whether we should consider the pressure p as representing a cell-center variable or a mean value for the cell. However, when reviewing the conservation law, we note that the conserved quantity is actually the integrated mass density over the cell, which is related to pressure via a constitutive law. As such, in

most implementations, it is most natural to consider the pressure as a mean value for the cell.

For the scalar problem, several specialized variants of the MPFA methods can be derived [3, 21–23]. However, each of these variants utilize a separate calculation for each of the subface fluxes. As such, these variants cannot be interpreted as having a unique piecewise linear pressure field, and are thus more complex to describe, implement and analyze. Their usage is limited in practice.

4.3.2 Elasticity

The equations of elasticity have the same basic elliptic structure as those for flow, however they have significant differences in the details. First, we note that the deformation vector u takes the role of potential, while the stress tensor π takes the role of a flux. The steady state of conservation of momentum is the balance equation for forces

$$\int_{\partial\omega} n \cdot \pi \, dS = \int_{\omega} r_u \, dV. \quad (4.3.5)$$

Note that this is a vector equation. Elastic materials satisfy Hooke's law, which can be written as

$$\pi = \mathbb{C} : \varepsilon(\nabla u) + \chi. \quad (4.3.6)$$

Here $\varepsilon(\nabla u)$ denotes the material strain, which in the regime of small deformations can be linearized as

$$\varepsilon(\nabla u) = \frac{\nabla u + \nabla u^T}{2}. \quad (4.3.7)$$

The external tensor field χ can arise from an existing stress state in the material, or as we will see below, from interactions with a separate process in composite materials. We will assume that the external tensor field is always symmetric.

The strain tensor $\varepsilon(\nabla u)$ is symmetric by definition, and we therefore refer to Eqs. (4.3.6)–(4.3.7) as Hooke's law with *strong symmetry*. In general, the gradient of the deformation ∇u need not be symmetric, and as a consequence, the compound action of \mathbb{C} and ε does not have a unique inverse when Hooke's law is written on the form (4.3.6)–(4.3.7). This has consequences for the stability of numerical methods, as we will see below.

We therefore consider also an alternative formulation of Hooke's law, known as Hooke's law with *weak symmetry*. Equations (4.3.6) and (4.3.7) can then be equivalently stated as

$$\pi = \mathbb{C} : (\nabla u + b) + \chi, \quad (4.3.8)$$

where b is the asymmetry of the gradient of deformation, which we will at the moment treat as unknown. In order to determine b , we enforce that the stress is symmetric. It turns out that it is sufficient to impose symmetry of the stress tensor weakly [33, 34]. Pre-empting that we will impose symmetry on the dual grid, we state the weak symmetry as follows: For all subdomains $\omega^* \in \Omega$, it holds that

$$\int_{\omega^*} as(\pi) dS = 0, \quad (4.3.9)$$

where the asymmetry of a tensor is defined as

$$as(\pi) = \frac{\pi + \pi^T}{2}. \quad (4.3.10)$$

It is straight-forward to verify, by setting $b = -as(\nabla u)$, that Eqs. (4.3.8)–(4.3.10) are satisfied by the solution of Eqs. (4.3.6)–(4.3.7), and have also been recently considered in the mixed finite element context (see e.g. [34]).

4.3.2.1 MPSA with Strong Symmetry

The MPxA finite volume method can be applied directly to Eqs. (4.3.5–4.3.7), in exact analogy to the scalar case, and we refer to this as the generalized MPSA-O method. As with the fluid flow, we can use the constitutive law (3.6) directly to define the discrete constitutive law \mathbf{B} .

Again we will avoid restating the equations of Sect. 4.2.2, and summarize that a numerical normal stress (i.e. traction) \mathbf{w} is defined by the MPSA method as a linear function of pressure, i.e. the finite volume scheme for Eqs. (4.3.5–4.3.7) is given as

$$\mathbf{D}_u \mathbf{w} = \mathbf{r}_u, \quad (4.3.11)$$

$$\mathbf{w} = \mathbf{W}_u \mathbf{u} + \mathbf{W}_\chi \chi. \quad (4.3.12)$$

Note that while the action of \mathbf{D}_p and \mathbf{D}_u are logically similar, the matrixes have slightly different structure as Eq. (4.3.11) represent n times as many degrees of freedom due to the vector nature of the elasticity equations.

It turns out that the simplified penalty function \mathcal{M}_s^η is not suitable for elasticity with strong symmetry. Indeed, the symmetry of the stress tensor reduces the number of constraints imposed by the local balance stated in Eq. (4.2.13), and the minimization problem given by (4.2.26) for the simplified penalty functions fail to have a unique solution. On the other hand, it can be shown that Eq. (4.2.27) does have a solution, and as such, only the generalized MPSA-O method is applicable elasticity with strong

symmetry [35]. While the generalized MPSA-O method is well suited for polyhedral grids in 2D and 3D, it is deficient on simplicial meshes [28, 33]. As is the case with mixed finite elements [34], it turns out that the formulation with weakly imposed symmetry is preferable.

4.3.2.2 MPSA with Weak Symmetry

When we consider the constitutive law with weak symmetry, we quickly note that the MPxA framework needs an adaptation in order to accommodate the condition Eq. (4.3.10). Indeed, by imposing Eq. (4.3.10) on each dual cell, it is equivalent to stating that for all $s \in \mathcal{V}$, it holds that

$$\sum_{\tilde{\omega} \in \tilde{\mathcal{T}}_s} \int_{\tilde{\omega}} a_s(\pi) dS = 0. \quad (4.3.13)$$

Since the stress π is approximated as constant on each subcell, Eq. (4.3.13) can easily be represented in terms of degrees of freedom as the matrix equation

$$\mathbf{S}\boldsymbol{\pi} = \mathbf{0}, \quad (4.3.14)$$

where again the matrix \mathbf{S} can be written as a sum of local matrices for each vertex, e.g. $\mathbf{S} = \sum_{s \in \mathcal{V}} \mathbf{S}_s$. With this tool in hand, the MPxA framework can be used directly to obtain a discretization, which we refer to as MPSA-W (W signifying weak symmetry), and state to be precise as:

Definition 4.3.1 (*MPSA-W (elasticity)*) Let \mathcal{N}_u be the null-space of the vector extension of the constraints given in Eqs. (4.2.11)–(4.2.13), as well as (4.3.14), subject to a given displacement \mathbf{u} . Then the MPSA-W method is defined by the pair $(\tilde{\mathbf{u}}, \boldsymbol{\pi}) \in \mathcal{N}_u$ such that

$$(\tilde{\mathbf{u}}, \boldsymbol{\pi}) = \arg \min_{(\tilde{\mathbf{u}}', \boldsymbol{\pi}') \in \mathcal{N}_p} \mathcal{M}(\tilde{\mathbf{u}}'), \quad (4.3.16)$$

and the numerical normal stress is defined as $\mathbf{w} = \mathbf{W}_u \mathbf{u} + \mathbf{W}_\chi \boldsymbol{\chi} \equiv \boldsymbol{\Sigma}_{\mathcal{F}} \mathbf{F} \boldsymbol{\pi}$.

Clearly, due to the choice of imposing the asymmetry of the stress on the dual grid, the MPSA-W method reduces to local calculations in the same way as other MPxA methods.

The MPSA-W discretization is now obtained by combining the normal stress from the MPSA-W method with the momentum balance, Eq. (4.3.11), in exactly the same manner as for the generalized MPSA-O method.

The MPSA-W method behaves qualitatively analogously to the MPFA methods, and can be used together with either the full penalty functions or the simplified

penalty functions. In contrast to the generalized MPSA-O method, the MPSA-W method is equally applicable to polyhedral as well as simplicial grids [33].

4.3.3 Poroelasticity

We will consider the linearized equations for poro-elasticity, after an implicit discretization over a time-step length θ . Then the linear system for pressure and displacement consists of two conservation laws for the fluid and solid [27]:

$$\int_{\omega} \alpha : \nabla u + cp dV + \theta \int_{\partial\omega} n \cdot \tau_p dS = \int_{\omega} r_p dV, \quad (4.3.17)$$

$$\int_{\partial\omega} n \cdot \pi dS = \int_{\omega} r_u dV, \quad (4.3.18)$$

as well as the constitutive laws for fluid flow (Darcy) and stress in poroelastic materials (Biot), stated in the form with weak symmetry:

$$\tau_p = -\kappa \nabla p + g, \quad (4.3.19)$$

$$\pi = \mathbb{C} : (\nabla u + b) - \alpha p, \quad (4.3.20)$$

$$\int_{\omega^*} as(\pi) dS = 0. \quad (4.3.21)$$

Relative to the previous sections, we have introduced the Biot coupling coefficient α , which is in general a symmetric second-order tensor (but often approximated as a scalar times an isotropic tensor in practice), as well as the effective compressibility term c , containing contributions from both bulk and fluid compressibility. Note also that the information from the previous time-step is integrated into the right-hand side term r_p .

In order to obtain a numerical stress function for poroelasticity, we follow the MPxA framework outlined in Sect. 4.2.2. From the perspective of the mechanics, the pressure is an external stress in the constitutive law, while conversely, from the perspective of flow, the mechanics affects the conservation statement.

We will therefore for the mechanical calculation consider the fluid pressure as the (previously external) imposed stress for the cell. The MPxA framework can then be applied directly, as in the case of Sect. 4.3.2. To incorporate the new material constants arising from the coupling terms, let \mathbb{C} be the application of the compressibility factor c at the cell level. Moreover, let \mathbf{A}_1 and \mathbf{A}_2 be the application of the Biot coefficient α at the subcell level, where \mathbf{A}_1 acts as the double inner-product on tensors (confer Equation (4.3.17)), while \mathbf{A}_2 acts a tensor-scalar product (confer Equation (4.3.20)).

Finally, similarly to $\Sigma_{\mathcal{F}}$ we let $\Sigma_{\mathcal{T}}$ be the summation over subcells, weighted by the subcell volumes.

With these definitions, the MPSA method can be directly applied to the linearized equations for poroelasticity. For completeness, we state its definition as:

Definition 4.3.2 (*MPSA-W (poroelasticity)*) Let $\mathcal{N}_{u,p}$ be the null-space of the vector extension of the constraints given in Eqs. (4.2.11)–(4.2.13), as well as (4.3.14), subject to a given displacement \mathbf{u} and a pressure \mathbf{p} . Then the MPSA-W method for poroelasticity is defined by the pair $(\tilde{\mathbf{u}}, \boldsymbol{\pi}) \in \mathcal{N}_{u,p}$ such that

$$(\tilde{\mathbf{u}}, \boldsymbol{\pi}) = \arg \min_{(\tilde{\mathbf{u}}', \boldsymbol{\pi}') \in \mathcal{N}_{u,p}} \mathcal{M}(\tilde{\mathbf{u}}'), \quad (4.3.24)$$

and the numerical normal stress is defined as $\mathbf{w} = \mathbf{W}(\mathbf{u}, \mathbf{p}) = \mathbf{W}_u \mathbf{u} + \mathbf{W}_p \mathbf{p} \equiv \Sigma_{\mathcal{F}} \mathbf{F} \boldsymbol{\pi} - \Sigma_{\mathcal{F}} \mathbf{F} \mathbf{A}_2 \mathbf{E}^* \mathbf{p}$. Moreover, the impact of displacement on the fluid mass conservation law is given by $\mathbf{J}(\mathbf{u}, \mathbf{p}) = \mathbf{J}_u \mathbf{u} + \mathbf{J}_p \mathbf{p} \equiv \Sigma_{\mathcal{F}} \mathbf{A}_1 \mathbf{G} \tilde{\mathbf{u}}$.

We note that the linear discretization matrices \mathbf{W}_u , \mathbf{W}_p , \mathbf{J}_u and \mathbf{J}_p are implicitly defined, since $\tilde{\mathbf{u}}$ and $\boldsymbol{\pi}$ are linear functions of \mathbf{u} and \mathbf{p} . As previously, all the minimization problems can be solved in parallel for each vertex of the grid. Moreover, the same points about simplified penalty functions as discussed in Sect. 4.3.2.2 are applicable.

We close this section by stating the full discrete system for the poroelastic Eqs. (4.3.17–4.3.21):

$$\mathbf{J}_u \mathbf{u} + \mathbf{J}_p \mathbf{p} + \mathbf{C} \mathbf{p} + \theta \mathbf{D}_p \mathbf{q} = \mathbf{r}_p, \quad (4.3.25)$$

$$\mathbf{D}_u \mathbf{w} = \mathbf{r}_u, \quad (4.3.26)$$

$$\mathbf{q} = \mathbf{Q}_p \mathbf{p} + \mathbf{Q}_g \mathbf{g}, \quad (4.3.28)$$

$$\mathbf{w} = \mathbf{W}_u \mathbf{u} + \mathbf{W}_p \mathbf{p}. \quad (4.3.29)$$

By eliminating the flux and normal stress, we get a system of matrix equations only in terms of cell-center pressure and displacement, given as

$$\begin{pmatrix} \mathbf{D}_u \mathbf{W}_u & \mathbf{D} \mathbf{W}_p \\ \mathbf{J}_u \mathbf{u} & \mathbf{C} + \theta \mathbf{D}_p \mathbf{Q}_p + \mathbf{J}_p \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{r}_u \\ \mathbf{r}_p - \theta \mathbf{D}_p \mathbf{Q}_g \mathbf{g} \end{pmatrix}. \quad (4.3.30)$$

As expected, Eqs. (4.3.25–4.3.29), and consequently also (4.3.30), has essentially the same structure as the continuous problem. The exception is the presence of the term \mathbf{J}_p , which appears implicitly in the pressure conservation equation, since the discrete displacement gradient $\nabla \mathbf{u}$ calculated by the MPSA method also depends on the pressure. This dependence is weak and can be interpreted as the expansion of a

(sub)cell due to an increment of pressure in that cell. As such, it has the structure of a Laplacian operator, scaled by the bulk modulus of the solid \mathbb{C} , and the square of the characteristic length scale of the cell Δx , i.e., the following spectral equivalence holds [27]:

$$J_p \sim \frac{(\Delta x)^2}{c} \mathbf{D}_p \mathbf{D}_p^T. \quad (4.3.31)$$

4.3.4 Thermo-poroelasticity

As the final application of the MPFA framework, we will consider the transport of heat in a poroelastic medium. Again, we will consider the equations subject to an implicit discretization over a time-step length θ . One has a choice in which variable to use to represent heat, however we will for simplicity consider temperature ϕ , as we are primarily interested in the spatial discretization of the linearized equations. Then the linear system for pressure, displacement and temperature then consists of three conservation laws for the fluid, entropy, and solid [8, 36]:

$$\int_{\omega} \alpha_p : \nabla u + c_{p,p} p + c_{p,\phi} \phi \, dV + \theta \int_{\partial\omega} n \cdot \tau_p \, dS = \int_{\omega} r_p \, dV, \quad (4.3.32)$$

$$\int_{\omega} \alpha_{\phi} : \nabla u + c_{\phi,p} p + c_{\phi,\phi} \phi \, dV + \theta \int_{\partial\omega} n \cdot \tau_{\phi} \, dS = \int_{\omega} r_{\phi} \, dV, \quad (4.3.33)$$

$$\int_{\partial\omega} n \cdot \pi \, dS = \int_{\omega} r_u \, dV, \quad (4.3.34)$$

as well as the constitutive laws for fluid flow (Darcy), heat transfer (Fourier's law and advection), and stress in thermo-poroelastic materials, the latter stated in the form with weak symmetry:

$$\tau_p = -\kappa_p \nabla p + g, \quad (4.3.35)$$

$$\tau_{\phi} = -\kappa_{\phi} \nabla \phi + \phi \tau_p, \quad (4.3.36)$$

$$\pi = \mathbb{C} : (\nabla u + b) - \alpha_p p - \alpha_{\phi} \phi, \quad (4.3.37)$$

$$\int_{\omega^*} a s(\pi) \, dS = 0. \quad (4.3.38)$$

Relative to the previous sections, we have for each of the fluid and thermal conservation laws separate linearized constitutive laws $c_{\phi,p}$, $c_{\phi,\phi}$, $c_{p,p}$, and $c_{p,\phi}$, Biot

coupling coefficients α_p and α_ϕ , and constitutive laws κ_p and κ_ϕ . Furthermore, we notice that the constitutive law for heat flux, given in Eq. (4.3.36), is non-linear due to the presence of the product $\phi\tau_p$, representing heat advection with the fluid flux.

With exception of the head advection term, the coupled problem represented by Eqs. (4.3.32–4.3.38) presents no new challenges relative to the poroelastic problem considered in Sect. 4.3.3, and the application of the MPxA method for the problem is equivalent. Thus we have the following discrete system for thermo-poroelasticity, which is the discrete analog of Eqs. (4.3.32–4.3.38). The conservation laws take the form:

$$\mathbf{J}_{p,u}\mathbf{u} + \mathbf{J}_{p,p}\mathbf{p} + \mathbf{C}_{p,p}\mathbf{p} + \mathbf{C}_{p,\phi}\phi + \theta\mathbf{D}_p\mathbf{q}_p = \mathbf{r}_p, \quad (4.3.39)$$

$$\mathbf{J}_{\phi,u}\mathbf{u} + \mathbf{J}_{\phi,\phi}\phi + \mathbf{C}_{\phi,p}\mathbf{p} + \mathbf{C}_{\phi,\phi}\phi + \theta\mathbf{D}_p\mathbf{q}_\phi = \mathbf{r}_\phi, \quad (4.3.40)$$

$$\mathbf{D}_u\mathbf{w} = \mathbf{r}_u. \quad (4.3.41)$$

While the constitutive laws take the form:

$$\mathbf{q}_p = \mathbf{Q}_{p,p}\mathbf{p} + \mathbf{Q}_{p,g}\mathbf{g}, \quad (4.3.42)$$

$$\mathbf{q}_\phi = \mathbf{Q}_{\phi,\phi}\phi + \phi^*\mathbf{q}_p, \quad (4.3.43)$$

$$\mathbf{w} = \mathbf{W}_u\mathbf{u} + \mathbf{W}_p\mathbf{p} + \mathbf{W}_\phi\phi. \quad (4.3.44)$$

The discrete matrixes are constructed exactly as in Definition 4.3.2, with the notational convention that (say) \mathbf{W}_ϕ is calculated using the coupling coefficient α_ϕ , while \mathbf{W}_p is calculated using the coupling coefficient α_p . Similarly, the discrete flux stencil $\mathbf{Q}_{\phi,\phi}$ is calculated using the MPFA method of Sect. 4.3.1, with the coefficient tensor κ_ϕ .

It remains to define the temperature ϕ^* on faces of the grid. We denote the matrix with these temperatures on the main diagonal as ϕ^* , for which the simplest and most commonly choice is obtained via the so-called upstream weighting [7, 37]. Thus, for any face $\sigma \in \mathcal{F}$, with neighboring cells ω_{k_1} and ω_{k_2} where $k_1 < k_2$, then

$$\phi_{\sigma,\sigma}^* = \begin{cases} \phi_{k_1} & \text{if } \mathbf{q}_{p,\sigma} \geq \mathbf{0}, \\ \phi_{k_2} & \text{if } \mathbf{q}_{p,\sigma} < \mathbf{0}. \end{cases} \quad (4.3.45)$$

The entries of ϕ^* are typically taken as zero away from the main diagonal.

A compact and simple discretization for coupled thermo-poromechanics is then obtained in terms of cell-center variables as

$$\begin{pmatrix} D_u W_u & D_u W_p & D_u W_\phi \\ J_{p,u} u & C_{p,p} + \theta D_p Q_{p,p} + J_{p,p} & C_{p,\phi} \\ J_{\phi,u} u & C_{\phi,p} + \theta D_p \phi^* Q_{p,p} & C_{\phi,\phi} + \theta D_p Q_{\phi,\phi} + J_{\phi,\phi} \end{pmatrix} \begin{pmatrix} u \\ p \\ \phi \end{pmatrix} = \begin{pmatrix} r_u \\ r_p - \theta D_p Q_{p,g} g \\ r_\phi - \theta D_p \phi^* Q_{p,g} g \end{pmatrix}. \quad (4.3.46)$$

Note that this discretization inherits the non-linearity of the original problem (4.3.32–4.3.38), due to the presence of the advective term, which explicitly becomes $\theta D_p \phi^* Q_{p,p}$.

4.4 Mathematical Properties of MPxA Methods

Since their inception, the MPFA (and later MPSA) methods have been intensely studied. Various viewpoints have been considered, using both analysis frameworks building on theory of finite volume and mixed finite element methods, as well as numerical validations. As a whole, these studies provide a comprehensive perspective on not just the properties of the MPFA finite volume discretization for the model problem from the introduction, Eqs. (4.1.2–4.1.2), but also the performance for elasticity and coupled problems as discussed in Sects. 4.3.2–4.3.4. We will summarize some of the main aspects below.

4.4.1 Analysis of Consistency and Convergence

Already in the earliest papers on MPFA methods, the consistency of the discretization was validated on parallelogram grids [3, 4]. More general analysis followed a decade later, and the first proof of convergence was established by Klausen and Winther, considering perturbations of parallelogram grids [38]. That analysis explicitly constructed a mixed finite element method which is algebraically equivalent to the MPFA method with simplified quadrature, by using the local problems detailed in Sect. 4.2.2 and 4.2.3 to define so-called “broken” finite element spaces for the flux.

While the analysis of Klausen, Winther provides both convergence as well as rates of convergence, it is quite restrictive, and does not apply to polyhedral grids nor non-smooth coefficients. The analysis was later extended to general polyhedral grids by Klausen and Stephansen by exploiting a link to mimetic finite differences [39]. A different approach was pursued by Agelas et al. [26], where they considered a formulation of the MPFA method in terms of the finite volume analysis framework [24]. This yielded convergence proofs through compactness arguments for quite general grids, and with minimal assumptions on the coefficients. On the other hand, this generality reduces the regularity of the exact solution, and thus explicit rates of convergence cannot be considered in this framework. Furthermore, the proofs suffered from an a priori assumption that the local formulation of MPFA

was uniformly coercive (with respect to all corners of the grid and all grid refinement). Such an assumption automatically holds for self-similar grid refinements, but was not proved.

The approach of Agelas was extended to show the convergence of MPSA for elasticity, and later also to show the convergence of MPFA + MPSA for the poroelastic problem of Sect. 4.3.3 [27, 28]. In these proofs, the general case of penalty functions with multiple quadrature points, as introduced in Sect. 4.2.2.3, was first considered. Considering the full penalty formulation had the further advantage of avoiding the local coercivity assumptions of Agelas, as the local coercivity could be proved based on the structure of the minimization problems. Moreover, the convergence proofs were shown to hold even for degenerate coefficients, such as incompressible materials and near-zero time-step size.

It is worth noting the related development of so-called Multipoint Flux Mixed finite Element (MFME) [40] and Multipoint Stress Mixed Finite Element (MFSE) [41] methods. These methods are obtained from mixed-finite element methods with BDM1 elements for flux (or stress) and P0 elements for pressure (or displacement), using various quadrature rules to eliminate the flux variables. The resulting methods, for which detailed analysis is possible based on standard theory of mixed finite elements, are close cousins of the MPxA finite volume methods described herein. However, these methods are less suited for geometrically complex problems, since the quadrature rules lead to reduced rates of convergence for rough grids [42], and the underlying finite element spaces preclude the applications to polyhedral grids.

4.4.2 Monotonicity

The question of monotonicity is essentially a translation of Hopf's lemma from the continuous problem to the discrete problem. For the scalar case, Hopf's lemma can be stated as the property that for a zero right-hand side $r_p = 0$, then the maximum (and minimum) value of the solution p should be found on the boundary of the domain [43].

Several numerical methods preserve the monotonicity property, in particular those that lead to discretization matrices on the form of M -matrices such as TPFA and FE. On the other hand, this property is in no way guaranteed, and as an example, the MFE method is in general not monotone. As the MPFA discretization does not guarantee an M -matrix, the question of monotonicity is subtle.

Sufficient and necessary conditions for any finite volume discretization to satisfy a discrete maximum principle can be established in the case of quadrilateral grids [29, 44]. As a result, it is now known that there are essentially three categories of grid (cells): (1) Those for which essentially any finite volume methods will lead to monotone discretization, (2) Those for which it is possible by a judicious choice to construct a monotone discretization, and (3) Those for which no linear finite volume discretization (with a relatively compact stencil) exists.

Clearly, point (3) above means that there are certain grids which are sufficiently bad that the performance of a MPxA discretization cannot be guaranteed, and these are in general grids combining high aspect ratios with a high degree of skewness. Point (2) above furthermore inspired research into constructing MPxA methods that are optimal with respect to monotonicity. Such methods can be constructed either by optimizing the location of quadrature points in the penalty functions [29, 45], or by allowing for more general formulation of the MPxA methods than that outlined in Sect. 4.2.2. As a result of the latter approach, the MPxA-Z method with a larger stencil [22], and the MPxA-L method with a smaller stencil [21, 23], were developed.

4.4.3 Numerical Investigations of Convergence

Complementing to the analysis summarized above, it is worth noting that the convergence properties of the MPFA and MPSA methods have been extensively studied numerically. These numerical investigations also consider problems not covered by analysis, due to either challenging coefficients [46], non-linearities [30], or grids [47]. We will review some of these results here, emphasizing the results that give a most comprehensive understanding of the general features of the MPxA methods.

4.4.3.1 Convergence Rates for Smooth Solutions

For problems with smooth coefficients on regular domains, the analysis of MPFA methods indicates that one can expect 2nd order convergence of the potential and 1st order convergence of the fluxes [38]. In practice 2nd order convergence of fluxes has been observed in numerical calculations for the flow problem, and what appears to be 1.5 order convergence for the elasticity and Biot problems. We will revisit some of these results here [27].

The problem under consideration is the poroelastic equations as presented in Sect. 4.3.3, with the MPFA and MPSA methods using full penalty functions as given in Eq. (4.2.20), and with the elasticity discretized with strong symmetry.

With the L^2 norms defined as

$$\|u\|_{\mathcal{T},0} = \left(\sum_{k \in \mathcal{T}} m_k u_k^2 \right)^{1/2} \quad \text{and} \quad \|q\|_{\mathcal{F},0} = \left(\sum_{\sigma \in \mathcal{F}} m_\sigma^2 q_\sigma^2 \right)^{1/2}. \quad (4.4.1)$$

We can define errors using the following L^2 type metrics, where variables in plain type are the exact analytical solution, and variables in bold are the discrete solutions, as in the preceding sections. The error in primary variables is then measured as relative to the projection $\Pi_{\mathcal{T}}$ which returns cell-center values (i.e. $(\Pi_{\mathcal{T}}p)_k = p(x_k)$):

$$\epsilon_u = \frac{\|\mathbf{u} - \Pi_{\mathcal{T}}\mathbf{u}\|_{\mathcal{T},0}}{\|\Pi_{\mathcal{T}}\mathbf{u}\|_{\mathcal{T},0}} \quad \text{and} \quad \epsilon_p = \frac{\|\mathbf{p} - \Pi_{\mathcal{T}}p\|_{\mathcal{T},0}}{\|\Pi_{\mathcal{T}}p\|_{\mathcal{T},0}} \quad (4.4.2)$$

and the error in secondary variables as relative to the projection $\Pi_{\mathcal{F}}$ which returns face-center fluxes (i.e. $(\Pi_{\mathcal{F}}\boldsymbol{\tau})_{\sigma} = \boldsymbol{\tau}(x_{\sigma}) \cdot \mathbf{n}_{\sigma}$):

$$\epsilon_{\pi} = \frac{\|\boldsymbol{w} - \Pi_{\mathcal{F}}\boldsymbol{\pi}\|_{\mathcal{F},0}}{\|\Pi_{\mathcal{F}}\boldsymbol{\pi}\|_{\mathcal{F},0}} \quad \text{and} \quad \epsilon_q = \frac{\|\mathbf{q} - \Pi_{\mathcal{F}}\boldsymbol{\tau}_p\|_{\mathcal{F},0}}{\|\Pi_{\mathcal{F}}\boldsymbol{\tau}_p\|_{\mathcal{F},0}}. \quad (4.4.3)$$

Finally, we also consider the error based on the L^2 seminorm of pressure, which discards the datum value, defined as

$$\epsilon_{p,|} = \inf_{p_0 \in \mathbb{R}} \frac{\|\mathbf{p} - \Pi_{\mathcal{T}}p + p_0\|_{\mathcal{T},0}}{\|\Pi_{\mathcal{T}}p\|_{\mathcal{T},0}}. \quad (4.4.4)$$

In order to illustrate the numerical convergence rate of the primary variables, we give the *primary error* associated with the primary variables displacement and pressure as

$$\epsilon_{u,p} = \epsilon_u + c\epsilon_p. \quad (4.4.5)$$

Furthermore, it can be shown that the MPSA-MPFA discretization is stable even for degenerate timestep size $\tau \rightarrow 0$ and compressibility $c \rightarrow 0$, subject to a weighted combination of the norms above [27]. Thus, we introduce the so-called *stable error*

$$\epsilon_{\Sigma} = \epsilon_u + \epsilon_{\pi} + (\theta + c)\epsilon_p + \theta\epsilon_q + \epsilon_{p,|}. \quad (4.4.6)$$

The numerical convergence rates for a smooth manufactured solution on irregular simplicial, irregular quadrilateral, and unstructured polyhedral grids, as illustrated in Fig. 4.3. The calculations are based on seven levels of refinement for each grid type, for which the finest grid level has a characteristic cell diameter of $h \sim 2^{-7}$, the results of which are summarized in Table 4.1. We note that as expected, 2nd order

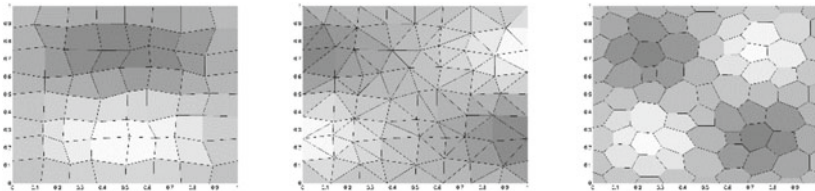


Fig. 4.3 From left to right the figures illustrate grid types **a** (quadrilaterals), **b** (triangles), **c** (unstructured grids). Furthermore, in grey-scale, the figures indicate the structure of the analytical solution used for this example

Tables 4.1 Asymptotic convergence rate of stable primary error $\epsilon_{u,p}$ and stable error ϵ_Σ for grids of types A, B, and C. We recall that τ is the dimensionless time-step, and c is the dimensionless compressibility

$\epsilon_{u,p}$	$\tau = 1$			$\tau = 10^{-6}$		
Grid	A	B	C	A	B	C
$c = 1$	2.00	1.97	1.99	1.99	1.95	1.98
$c = 10^{-2}$	2.00	1.97	1.99	1.98	1.94	1.98
$c = 10^{-6}$	2.00	1.97	1.99	1.98	1.94	1.98
ϵ_Σ	$\tau = 1$			$\tau = 10^{-6}$		
Grid	A	B	C	A	B	C
$c = 1$	1.36	1.36	1.27	1.09	1.14	1.16
$c = 10^{-2}$	1.32	1.32	1.23	1.20	1.29	1.28
$c = 10^{-6}$	1.32	1.32	1.23	1.20	1.29	1.29

convergence is observed for primary variables, and better-than-1st order convergence is observed for fluxes and stresses.

4.4.3.2 Convergence Rates for Singular Solutions

In order to assess the convergence rates for non-smooth problems, Eigestad and Klausen considered domains with discontinuous permeability coefficients, such as illustrated in Fig. 4.4 [46].

For such domains, analytical solutions can be defined on using polar coordinates around the center point, on the form

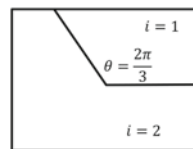
$$p(r, \theta) = r^\alpha (a_i \cos(\alpha\theta) + b_i \sin(\alpha\theta)). \tag{4.4.7}$$

The constants α , a_i and b_i , for $i = 1, 2$, depend on the permeability contrast chosen, and in particular, the exponent α also determines the regularity of the solution, i.e.

$$p \in H^{1+\alpha}(\Omega). \tag{4.4.8}$$

For such problems, they report a loss of convergence rate, such that one observes that the pressure converges at a rate of $\epsilon_p \sim h^{\min(2, 2\alpha)}$ while the flux converges at

Fig. 4.4 Partitioning of domain such that a non-trivial material discontinuity can be defined



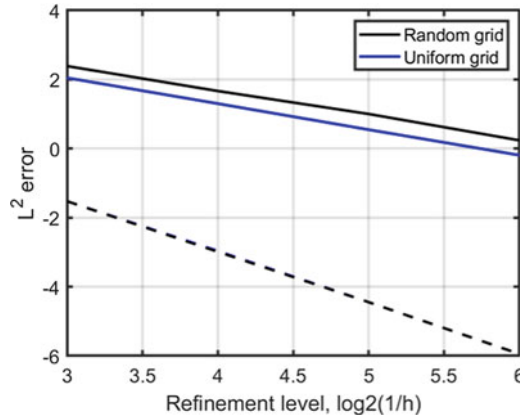


Fig. 4.5 Convergence of pressure (dashed lines) and flux (solid lines) for the MPFA method for the non-smooth problem of Sect. 4.4.3.1, on both random (black) and uniform grids (blue). The observed convergence rates for this problem, where the exact solution is order 1.5 for pressure and order 0.75 for flux. For comparison, the TPFA method was included in the original study, which does not converge (not shown in figure). Note that the blue dashed line is not visible behind the black dashed line

a rate $\epsilon_q \sim h^{\min(1, \alpha)}$. For the particular choice of $\theta = 2\pi/3$, and a permeability contrast $\frac{k_1}{k_2} = 100$, the resulting analytical solution has the exponent $\alpha \approx 0.75$. Figure 4.5 illustrates the convergence for this case, based on the MPFA method with simplified penalty functions ($\eta = 0$), and both regular and perturbed grid sequences.

4.4.3.3 Robustness on Degenerate Grids

Contrasting the previous two studies, Nilsen et al. emphasized degeneracies of the grid (as opposed to regularity-preserving refinements) [47]. To this end, they considered a series of cases with polyhedral grids, grids of high aspect ratio, and unusual refinement strategies. All of their calculations considered the MPSA discretization with strong symmetry, applied to either elasticity or coupled with MPFA for Biot.

An illustrative example from that study, considers a problem of non-matching grids, meeting at a thin layer, as illustrated in Fig. 4.6 (left). The thin layer is discretized by a finer grid which has roughly isotropic shape, as shown in Fig. 4.6 (right). The color scale in that figure indicates the approximation error relative to a smooth reference solution, which can be seen to be less than 4% in displacement (left figure) and as much as around 50% in the grid cells immediately adjacent to the thin layer for the volumetric strain (right figure).

To study the robustness of the method the ratio of the thin layer as compared to the external grid cells was varied from a factor 1 to a factor 20 (for comparison, Fig. 4.6. illustrates a factor 7 difference in grids). Recall that the grid cells in the thin layer are nearly isotropic in shape, thus when the thickness of the thin layer is reduced,

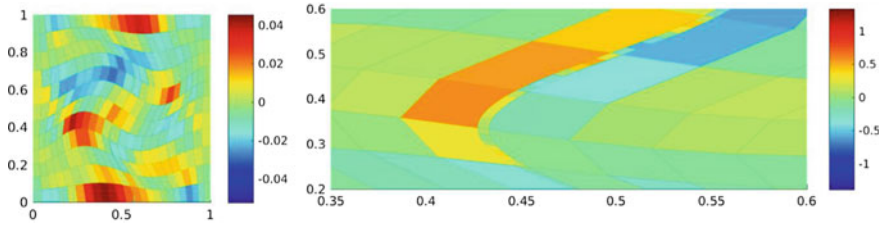


Fig. 4.6 Illustration of the grid used for robustness study (left). Note there is a vertical section of thin cells in the middle of the domain, mimicking a thin geological layer, as shown in the zoom (right). The discretization on the right-hand side of the thin layer is intentionally chosen to be slightly coarser than the left-hand side to ensure that the inner layer of grid cells is always non-matching relative to the surroundings. The color map on the left indicates the relative error in the x-component of displacement, while the color map on the right indicates the relative error in volumetric strain, both as compared to a manufactured analytical solution for this problem

the number of cells in the layer is simultaneously increased, thus introducing an increasing number of hanging nodes between. The study can thus be seen both as a study of robustness to an abrupt change in grid sizes in the discretization, as well as a study in the robustness to hanging nodes.

The results are shown in Fig. 4.7, where a comparison is also made to a Virtual Element Discretization for the same grid [48]. As can be seen, the approximation quality of the MPSA method is essentially unaffected by the presence of the thin

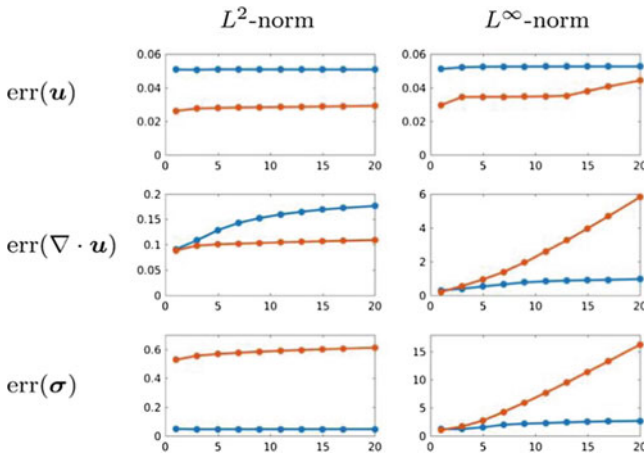


Fig. 4.7 Error in numerical approximation relative to analytical solution for grid types as shown in this figure. Blue lines are the MPSA method with strong symmetry from Sect. 4.3.2.1, while for comparison, an elasticity discretization using the virtual element method [49] is also shown. Errors are shown for displacement, volumetric strain, and stress in the x-axis, respectively, and using the L^2 and maximum norms in the columns. The x-axis of all figures denotes the aspect ratio between the thin layer and the outer grid, thus the right-most data-point corresponds to a factor 20 finer grid in the thin layer

layer, both in the L^2 and L^∞ norms. In particular, the stability in the L^∞ norm shows that spurious oscillations are not introduced in the transition between the grids.

4.4.3.4 Convergence for Thermo-poroelasticity

We close this section with a convergence study for the MPxA discretization of the full thermo-poroelastic problem. To our knowledge, results for this problem have not been reported before. The domain is the unit square, and the grid is formed by quadrilaterals that are roughly perturbed on all refinement levels. As in the study discussed in Sect. 4.4.3.1, we consider a single time step for the system, with time discretization by a backward Euler approach. All parameters are assigned unit values in this case. The manufactured solution is given by

$$u = \begin{pmatrix} \sin(2\pi x)y(1-y) \\ \sin(2\pi x)\sin(2\pi y) \end{pmatrix}, \quad p = \sin(2\pi x)y(1-y), \quad \phi = xy(1-x)(1-y).$$

The thermo-poroelastic system was discussed with MPSA/MPFA as discussed in Sect. 4.3.4, while a single point upstream approach was applied for the temperature advection term.

The convergence behavior is shown in Fig. 4.8. Displacement and pressure retain the second order convergence observed on the comparable test for the poro-elastic system considered in Sect. 4.4.3.1. For the temperature, the first order scheme for advection makes the convergence deteriorate to first order as the grid is refined, as expected from the theory of hyperbolic conservation laws [37]. Without the advective term, temperature also showed second order convergence. Finally, the mechanical stress and the fluid fluxes both are first order convergent on the perturbed grids. The test case thus confirms that the combined MPxA schemes can be applied successfully applied to problems including a non-linear advection term.

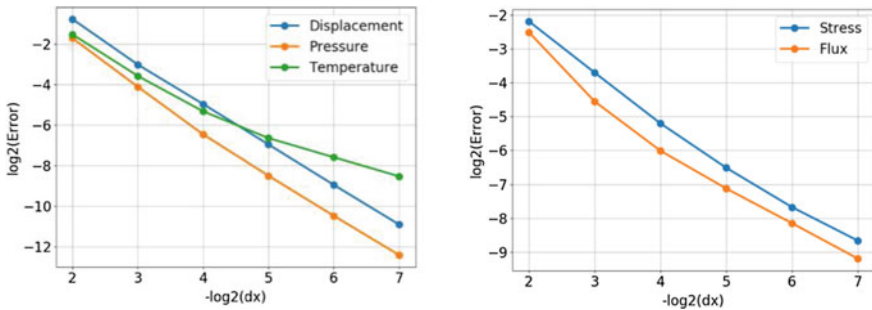


Fig. 4.8 Convergence plot for the thermo-poroelasticity problem described in Sect. 4.4.3.4. The convergence for the primary variables is shown to the left, to the right is shown the convergence results for mechanical stresses and fluid fluxes. The error is plotted in terms of the characteristic grid size dx

4.5 Applications to Complex Problems

Having reviewed the mathematical properties of MPxA methods in the previous section, we here show application-motivated scenarios where the methods can be applied. We present three setups: Poro-elastic deformation during fluid injection, thermo-poroelastic response to cooling, and flow through a fractured porous media. The cases are designed to showcase the applicability of MPxA methods on a wide range of grids, and the cells in the three cases are respectively perturbed hexahedra, prismatically extended polygons and simplexes. All simulations use the open source simulation tool PorePy [50], which provides an implementation of MPxA method that follow the principles discussed in this chapter, see [50, 51] for details.

4.5.1 Poro-elastic Response to Fluid Injection

We consider the poro-elastic response to fluid injection a domain of $10 \times 10 \times 1.8$ km, covered by $71 \times 71 \times 40 = 201640$ cells forming a Cartesian grid. The test case is motivated by CO₂ storage, although only single-phase flow is considered, with alternating layers of high and low-permeable domains that act as storage formation and trap, respectively [5]. The permeability contrast is four to five orders of magnitude, while the elastic moduli are heterogeneous, though of comparable size. The height of the layers varies, so that the computational cells are perturbed from their original hexahedral form, as indicated in Fig. 4.9.

The system is discretized with MPSA/MPFA, and fluid injection in the middle of storage layer was simulated. Figure 4.9 shows the fluid pressure and the vertical displacement in a cut domain. The pressure solution adapts to the permeability contrast, while the displacement varies smoothly throughout the domain. As expected, there are no signs of pressure oscillations due to the stabilization term J_p , see Sect. 4.3.3, despite the presence of strong permeability contrasts, which is known to cause problems for many discretization schemes [52]. The example thus illustrates the robustness of the MPxA discretizations of poro-elastic problems with strongly heterogeneous parameters.

4.5.2 Thermo-poroelastic Response to Cooling

To illustrate MPxA applied to the full thermo-poroelastic system, we consider a 3D unit cube that undergoes cooling. Specifically: The domain is cooled at the bottom by fixing a temperature lower than the initial state. Fluid is allowed to leave through the top, the bottom is impermeable for fluid flow, while the lateral sides are assigned homogeneous Neumann conditions for both fluid and temperature. The domain is fixed on all sides except the bottom, which is free to move. The domain is meshed

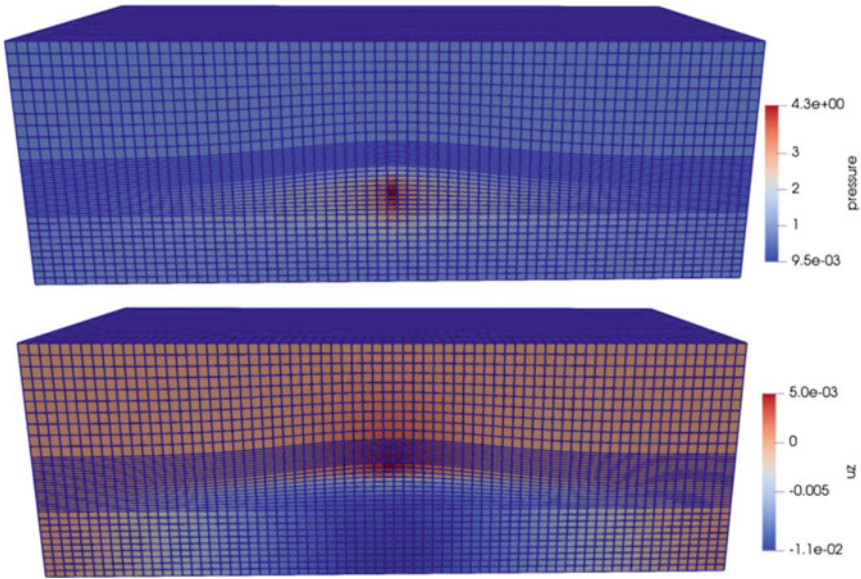


Fig. 4.9 Simulation of fluid injection into a poro-elastic cube. Fluid is injected into the middle of the domain, and the domain is cut to show the effects near the injection point. Top: Increase in fluid pressure due to injection, measured in bar. Bottom: Vertical displacement, measured in meters

with polyhedral cells formed by first taking the Voronoi diagram of a 2d triangulation, and then extruding the grid in the third direction. The resulting grid has 4275 cells, with a mixture of 6, 7 and 8 faces per cell.

On this mesh, the full thermo-poroelastic system is discretized as described in Sect. 4.3.4. Snapshots of the time evolution of temperature, pressure and displacement are shown in Fig. 4.10. At an early stage, the couplings in the system lead to noticeable 3d effects towards the bottom of the domain and significant displacements. The snapshots at later stages reflect the gradual cooling of the domain, and a decrease in pressure and displacement gradients. Note also that the pressure has low regularity at early time, as is expected since the elliptic term for the pressure in Eq. (4.3.30) scales with θ . This is consistent with the use of a weighted norm in Eq. (4.4.6).

From this simulation, we conclude that the MPxA family of method can handle problems with strong multi-physics couplings with no stability issues. Moreover, the example illustrates the schemes' applicability to general polyhedral grids; indeed, the implementation employed herein is agnostic both to spatial dimension and grid type.

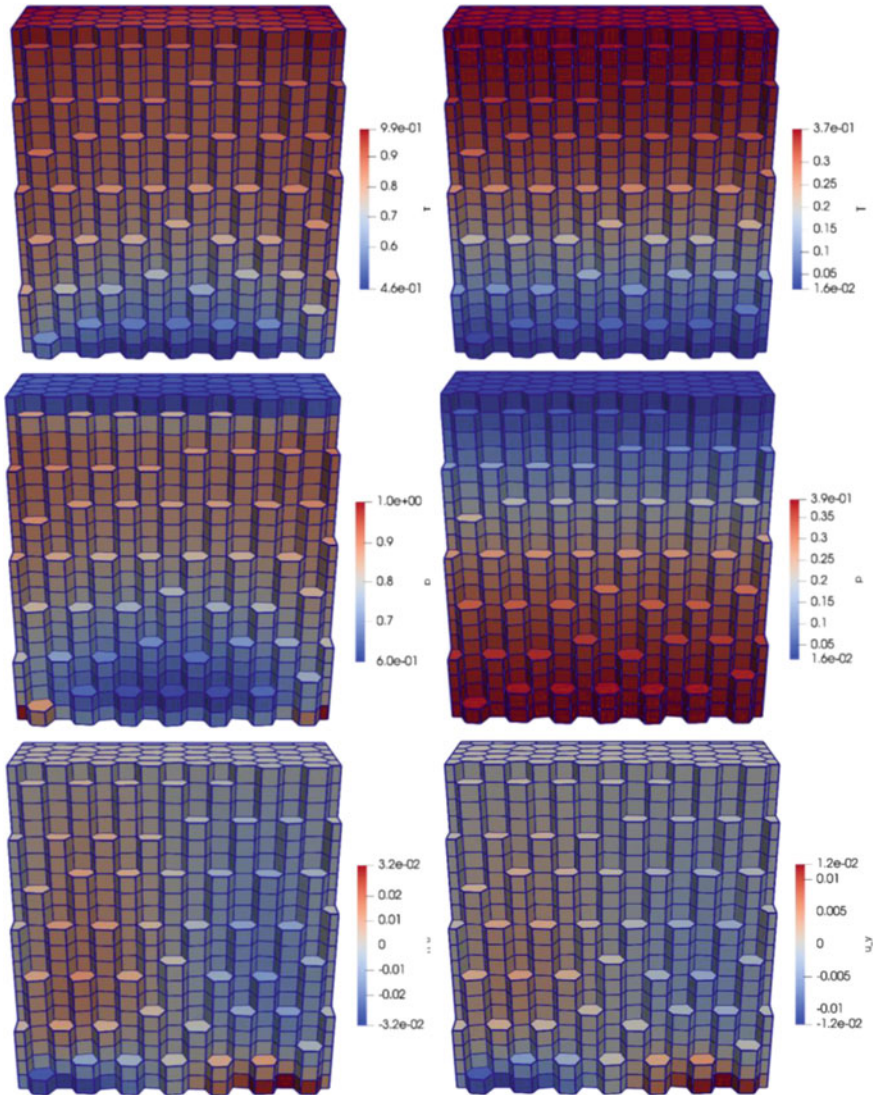


Fig. 4.10 Thermo-poroelastic deformation of a domain meshed by a polyhedral grid. The domain is cut to expose the 3d structure of the grid cells. The figure shows temperature (top), pressure (middle) and displacement in a direction approximately parallel to the cutting plane (bottom) at an early (left) and late (right) stage

4.5.3 Flow in Fractured Porous Media

Our final example considers simulation of flow in a 3d domain that contains a network of intersecting fractures. The fractures are modeled as manifolds of co-dimension 1 that are embedded in the host medium. Intersections between fractures form lines of co-dimension 2, while the intersection of intersection lines define intersection points. The fracture network and its host medium thus together define a hierarchy of domains with decreasing dimensions, which we refer to as a mixed-dimensional geometry [53]. Following the model defined e.g. in [51] flow in each of the subdomains is modeled by Eqs. (4.1.1)–(4.1.2), with the modification that (4.1.2) is void in 0d domains.

To define the coupling between subdomains, let Ω_h and Ω_l be two domains so that a part of the boundary $\partial\Omega_h$ geometrically coincides with Ω_l , and let Γ be an interface between the subdomains. The flow over Γ is then governed by the Darcy-like flux law

$$\lambda = \kappa(\text{tr } p_h - p_l)$$

Here, λ is the interface flux, κ is the interface permeability, $\text{tr } p_h$ denotes the trace of the pressure in Ω_h , evaluated on the relevant part of $\partial\Omega_h$, and p_l is the pressure in Ω_l . The interface flux can be considered a mortar variable, which is represented as a Neumann boundary condition to Ω_h and a source term for Ω_l .

Following the principles outlined in [51], the MPFA discretization can readily be adapted to mixed-dimensional flow problems. As an illustration we consider the final test case in the benchmark study proposed in [54]. The case consists of a 3d domain with 52 fractures that further form 106 intersection lines as indicated in Fig. 4.11. Boundary conditions are set up to drive flow through the host domain and the fracture network, with an inlet in the upper left corner referring to Fig. 4.11, and outlets in the two corners of the domain that are in the lower left part of the figure.

The computational mesh is constructed to conform with all fractures and fracture intersection lines. The resulting mesh consists of almost 260K 3d cells, 52k 2d cells (on fracture planes), 1.6k 1d cells (intersection lines) and 105k mortar variables. The MPFA discretization of the full problem produce almost 420K degrees of freedom with almost 23M non-zero matrix elements. The resulting pressure profile is shown in Fig. 4.11. The results obtained with MPFA are in good agreement with other methods applied to this benchmark, see [55]. The test case thus illustrates the applicability of the MPFA method also to non-standard problems such as flow in mixed-dimensional geometries.

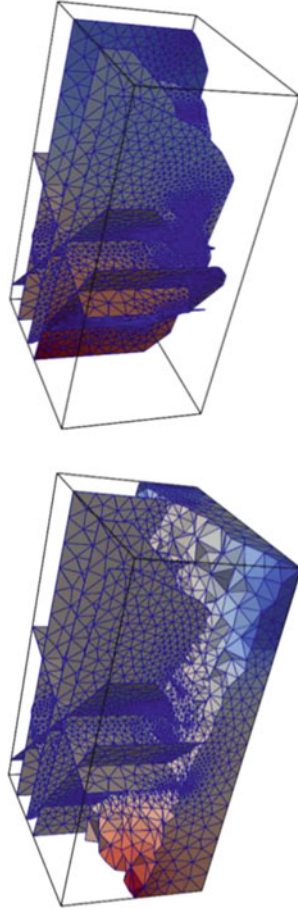


Fig. 4.11 MPFA pressure solution for the fracture flow benchmark problem presented in Sect. 4.5.3. The left figure shows the pressure in the host medium and the fracture network, while the pressure variations internal to the fracture network are illustrated in the right figure

Acknowledgements We wish to thank Ivar Aavatsmark, Inga Berre, Geir Terje Eigestad and Ivar Stefansson, for fruitful discussions, and for contributions to various aspects of the numerical implementation. This work forms part of Norwegian Research Council projects 250223 and 267908.

References

1. I. Aavatsmark, T. Barkve, Ø. Bøe, T. Mannseth, Discretization on non-orthogonal, curvilinear grids for multi-phase flow, in *Proceedings of the fourth European Conference on the Mathematics of Oil Recovery* (Røros, Norway, 1994)
2. M.G. Edwards, C.F. Rogers, A flux continuous scheme for the full tensor pressure equation, in *Proceedings of the fourth European Conference on the Mathematics of Oil Recovery* (Røros, 1994)
3. I. Aavatsmark, T. Barkve, Ø. Bøe, T. Mannseth, Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.* **127**, 2–14 (1996)
4. M.G. Edwards, C.F. Rogers, Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Comput. Geosci.* **2**(4), 259–290 (1998)
5. J.M. Nordbotten, M.A. Celia, *Geological storage of CO₂: Modeling approaches for large-scale simulation* (Wiley, Hoboken, NJ, 2011)
6. J. Bear, *Hydraulics of Groundwater* (McGraw-Hill, 1979)
7. Z. Chen, G. Huan, Y. Ma, *Computational Methods for Multiphase Flows in Porous Media* (SIAM, 2006)
8. O. Coussy, *Poromechanics* (Wiley, 2003)
9. T.F. Russell, M.F. Wheeler, Finite element and finite difference methods for continuous flows in porous media, in *Mathematics of Reservoir Simulation*, ed. by R.E. Ewing (SIAM, 1983), pp. 35–106
10. D. Braess, *Finite Elements* (Cambridge, 2007)
11. T. Hughes, *The Finite Element Method* (Dover, 2000)
12. T. Hughes, G. Engel, L. Mazzei, M.G. Larson, The continuous Galerkin method is locally conservative. *J. Comput. Phys.* **163**(2), 467–488 (2000)
13. T.Y. Hou, X.H. Wu, A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.* **134**(1), 169–189 (1997)
14. L.J. Durlofsky, Accuracy of mixed and control volume finite element approximations to Darcy velocity and related quantities. *Water Resour. Res.* **30**(4), 965–973 (1994)
15. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods*. Springer Series in Computational Mathematics (1991)
16. J. Nordbotten, H. Hægland, On reproducing uniform flow exactly on general hexahedral cells using one degree of freedom per surface. *Adv. Water Resour.* **32**(2), 264–267 (2009)
17. T. Arbogast, M.R. Correa, Two families of (div) mixed finite elements on quadrilaterals of minimal dimension. *SIAM J. Numer. Anal.* **54**(6), 3332–3356 (2016)
18. R. Herbin, F. Hubert, Benchmark on discretization schemes for anisotropic diffusion problems on general grids, in *Finite Volumes for Complex Applications V* (Wiley-ISTE, 2008), p. 750
19. I. Aavatsmark, Interpretation of a two-point flux stencil for skew parallelogram grids. *Comput. Geosci.* **11**(3), 199–206 (2007)
20. I. Aavatsmark, An introduction to the multipoint flux approximations for quadrilateral grids. *Comput. Geosci.* **6**(3–4), 405–432 (2002)
21. I. Aavatsmark, G.T. Eigestad, B.T. Mallison, J.M. Nordbotten, A compact multipoint flux approximation method with improved robustness. *Numer. Methods Part. Different. Eqs.* **24**(5), 1329–1360 (2008)

22. J.M. Nordbotten, G.T. Eigestad, Discretization on quadrilateral grids with improved monotonicity properties. *J. Comput. Phys.* **203**(2), 744–760 (2005)
23. I. Aavatsmark, G.T. Eigestad, B.O. Heimsund, B.T. Mallison, J.M. Nordbotten, E. Oian, A new finite-volume approach to efficient discretization on challenging grids. *SPE J.* **15**(3), 658–669 (2010)
24. R. Eymard, T. Gallouët, R. Herbin, Finite volume methods, in *Handbook of Numerical Analysis*, vol. VII (Elsevier, 2006), pp. 713–1020
25. J.M. Nordbotten, Finite volume methods, in *Encyclopedia of Applied and Computational Mathematics* (Springer, 2015)
26. L. Agelas, C. Guichard, R. Masson, Convergence of finite volume MPFA O type schemes for heterogeneous anisotropic diffusion problems on general meshes. *Int. J. Finite Vols.* (2010)
27. J.M. Nordbotten, Stable cell-centered finite volume discretization for Biot equations. *SIAM J. Numer. Anal.* **54**(2), 942–968 (2016)
28. J.M. Nordbotten, Convergence of a cell-centered finite volume discretization for linear elasticity. *SIAM J. Numer. Anal.* **53**(6), 2605–2625 (2016)
29. J.M. Nordbotten, I. Aavatsmark, G.T. Eigestad, Monotonicity of control volume methods. *Numer. Math.* **106**(2), 255–288 (2007)
30. R. Klausen, F. Radu, G.T. Eigestad, Convergence of MPFA on triangulations and for Richard’s equation. *Int. J. Numer. Method Fluids* **58**(12), 1327–1351 (2008)
31. H.A. Friis, M.G. Edwards, J. Mykkeltveit, Symmetric positive definite flux-continuous full-tensor finite-volume schemes on unstructured cell-centered triangular grids. *SIAM J. Sci. Comput.* **31**(2), 1192–1220 (2008)
32. M. Edwards, Cross flow tensors and finite volume approximation with by deferred correction. *Comput. Methods Appl. Mech. Eng.* **151**(1–2), 143–161 (1998)
33. E. Keilegavlen, J.M. Nordbotten, Finite volume methods for elasticity with weak symmetry. *Int. J. Numer. Meth. Eng.* **112**(8), 939–962 (2017)
34. D. Arnold, R. Winther, Mixed finite elements for elasticity. *Numer. Math.* **92**, 401–419 (2002)
35. J.M. Nordbotten, Cell-centered finite volume discretizations for deformable porous media. *Int. J. Numer. Methods Eng.* **100**(6), 399–418 (2014)
36. A.P.S. Selvadurai, A.P. Suvorov, *Thermo-Poroelasticity and Geomechanics* (Cambridge University Press, 2016)
37. R. J. LeVeque, *Numerical Methods for Conservation Laws* (Birkhäuser, 1992)
38. R. Klausen, R. Winther, Robust convergence of multi point flux approximation on rough grids. *Numer. Math.* **104**, 317–337 (2006)
39. R.A. Klausen, A.F. Stephansen, Convergence of the multi-point flux approximations on general grids and media. *Int. J. Numer. Anal. Model.* **9**(3), 584–606 (2012)
40. M.F. Wheeler, I. Yotov, A multipoint flux mixed finite element method. *SIAM J. Numer. Anal.* **44**(5), 2082–2106 (2006)
41. I. Ambartsumyan, E. Khattatov, J.M. Nordbotten, I. Yotov, A multipoint stress mixed finite element method for elasticity on simplicial grids. *SIAM J. Numer. Anal.* (in press, 2020)
42. I. Aavatsmark, G.T. Eigestad, R.A. Klausen, M.F. Wheeler, I. Yotov, Convergence of a symmetric MPFA method on quadrilateral grids. *Comput. Geosci.* **11**(4), 333–345 (2007)
43. E. Hopf, Elementare Bemerkungen über die Lösungen partieller Differentialgleichungen zweiter Ordnung vom elliptischen Typus. *Sitzungsber. Preuß. Akad. Wiss.* **19**, 147–152 (1927)
44. E. Keilegavlen, J.M. Nordbotten, I. Aavatsmark, Sufficient criteria are necessary for monotone control volume methods. *Appl. Math. Lett.* **22**(8), 1178–1180 (2009)
45. M. Edwards, H. Zheng, Double-families of quasi-positive Darcy-flux approximations with highly anisotropic tensors on structured and unstructured grids. *J. Comput. Phys.* **3**(1), 594–625 (2010)
46. G.T. Eigestad, R. Klausen, On the convergence of the multi-point flux approximation O-method: numerical experiments for discontinuous permeability. *Numer. Methods Partial Different. Eqs.* **21**, 1079–1098 (2005)
47. H.M. Nilsen, J.M. Nordbotten, X. Raynaud, Comparison between cell-centered and nodal-based discretization schemes for linear elasticity. *Comput. Geosci.* **22**(1), 233–260 (2018)

48. L. Beirão da Veiga, F. Brezzi, L.D. Marini, Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal.* **51**(2), 794–812 (2013)
49. A. Gain, C. Talischi, G. Paulino, On the virtual element method for three-dimensional linear elasticity problems on arbitrary polyhedral meshes. *Comput. Methods Appl. Mech. Eng.* **282**, 132–160 (2014)
50. E. Keilegavlen, R. Berge, A. Fumagalli, J. Starnoni, I. Stefansson, J. Varela, I. Berre, PorePy: an open-source software for simulation of multiphysics processes in fractured porous media. *Comput. Geosci.* **25**, 243–265 (2021)
51. J. Nordbotten, W. Boon, A. Fumagalli, E. Keilegavlen, Unified approach to discretization of flow in fractured porous media. *Comput. Geosci.* **23**(2), 225–237 (2019)
52. J. Haga, H. Osnes, H. Langtangen, On the causes of pressure oscillations in low-permeable and low-compressible porous media. *Int. J. Numer. Anal. Meth. Geomech.* **36**(12), 1507–1522 (2012)
53. W.M. Boon, J.M. Nordbotten, J.E. Vatne, Functional analysis and exterior calculus on mixed-dimensional geometries. *Annali di Matematica* **200**, 757–789 (2021)
54. I. Berre, W. Boon, B. Flemisch, A. Fumagalli, D. Gläser, E. Keilegavlen, A. Scotti, I. Stefansson, A. Tatomir, Call for participation: Verification benchmarks for single-phase flow in three-dimensional fractured porous media. [arXiv:1809.06926](https://arxiv.org/abs/1809.06926) (2018)
55. I. Berre, W.M. Boon, B. Flemisch, A. Fumagalli, D. Gläser, E. Keilegavlen, A. Scotti, I. Stefansson, A. Tatomir, et al., Verification benchmarks for single-phase flow in three-dimensional fractured porous media. *Adv. Water. Res.* **147**, 103759 (2021)

Chapter 5

High-order Discontinuous Galerkin Methods on Polyhedral Grids for Geophysical Applications: Seismic Wave Propagation and Fractured Reservoir Simulations



Paola F. Antonietti, Chiara Facciola, Paul Houston, Ilario Mazzieri, Giorgio Pennesi, and Marco Verani

Abstract We present a comprehensive review of the current development of discontinuous Galerkin methods on polytopic grids (PolyDG) methods for geophysical applications, addressing as paradigmatic applications the numerical modeling of seismic wave propagation and fracture reservoir simulations. We first recall the theoretical background of the analysis of PolyDG methods and discuss the issue of its efficient implementation on polytopic meshes. We address in detail the issue of numerical quadrature and recall the new *quadrature free* algorithm for the numerical evaluation of the integrals required to assemble the mass and stiffness matrices introduced in [22]. Then we present PolyDG methods for the approximate solution of the elastodynamics equations on computational meshes consisting of polytopic ele-

P.F.A. and M.V. acknowledge the financial support of MIUR through the PRIN grant n. 201744KLJL. P.F.A., M.V. and I.M. have also been funded by INdAM-GNCS.

P. F. Antonietti (✉) · C. Facciola · I. Mazzieri · G. Pennesi · M. Verani
MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica,
Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy
e-mail: paola.antonietti@polimi.it

C. Facciola
e-mail: chiara.facciola@polimi.it

I. Mazzieri
e-mail: ilario.mazzieri@polimi.it

G. Pennesi
e-mail: giorgio.pennesi@polimi.it

M. Verani
e-mail: marco.verani@polimi.it

P. Houston
School of Mathematical Sciences, The University of Nottingham, University Park, Nottingham
NG7 2RD, UK
e-mail: paul.houston@nottingham.ac.uk

ments. We review the well-posedness of the numerical formulation and *hp*-version *a priori* stability and error estimates for the semi-discrete scheme, following [10]. The computational performance of the fully-discrete approximation obtained based on employing the PolyDG method for the space discretization coupled with the leap-frog time marching scheme are demonstrated through numerical experiments. Next, we address the problem of modeling the flow in a fractured porous medium and we review the unified construction and analysis of PolyDG methods following [16]. We show, in a unified setting, the well-posedness of the numerical formulations and *hp*-version *a priori* error bounds, that are then validated through numerical tests. We also briefly discuss the extendability of our approach to handle networks of partially immersed fractures and networks of *intersecting* fractures, recently proposed in [15].

Keywords High-order Discontinuous Galerkin · Polygonal and polyhedral meshes · Fast implementation · Quadrature free · Seismic wave propagation · Fractured reservoir simulations

5.1 Introduction

Many geophysical and engineering applications, including, for example, fluid-structure interaction, crack and wave propagation phenomena, and flow in fractured porous media, are characterized by a strong complexity of the physical domain, possibly involving faults and/or fractures, heterogeneous media, moving geometries/interfaces and complex topographies. Whenever classical finite element methods are employed to discretize the underlying differential model, the process of mesh generation can represent a severe bottleneck for the simulation process, as classical finite element methods (in three-dimensions) typically only support computational grids composed of tetrahedral/hexahedral/prismatic/pyramidal elements. To overcome this limitation, in the last decade a wide strand of literature has focused on the design of numerical methods that support computational meshes composed of general polygonal and polyhedral (polytopic, for short) elements. In the conforming setting, we mention, for example, the Composite Finite Element Method, see, e.g., [103, 104], the Mimetic Finite Difference (MFD) method, see, e.g., [7, 18, 42, 58–60, 106], the Polygonal Finite Element Method, see, e.g., [140], the eXtended Finite Element Method, see, e.g., [88, 97, 141], and, more recently, the Virtual Element Method (VEM), see, e.g., [8, 9, 39–41, 43–48]. In the setting of non-conforming/discontinuous polygonal methods, we mention, for example, Composite Discontinuous Galerkin Finite Element methods [19, 20], Hybridizable Discontinuous Galerkin methods [75–78], the Hybrid High-Order (HHO) method [1, 53–55, 71, 72, 84–87], the non-conforming VEM [24, 35, 68], and Gradient Schemes [90]. This article focuses on discontinuous Galerkin (DG) methods on polytopic grids (PolyDG), which represent the natural extension of the classical discontinuous Galerkin method on tetrahedral/hexahedral grids to meshes composed of arbitrarily-shaped polytopic elements. Due to the fact that the discrete space is constructed based

on employing piecewise *discontinuous* polynomials, DG methods are naturally suited to robustly support polytopic meshes. In fact, in the last few years intensive research has been undertaken on this topic; in particular, we refer here to the pioneering works [12, 36–38, 67], the more recent results [5, 11, 21, 22, 26, 30, 63], and refer to [13, 66], and the references therein, for a comprehensive review.

This article focuses on two challenging applications in geophysics, namely, seismic wave propagation and fractured reservoir simulations and presents a review of PolyDG methods for this class of problems, as well a detailed discussion on the development of efficient quadrature rules on polytopic elements that allows a massively-parallel implementation of PolyDG methods on parallel architectures. From the mathematical and modeling viewpoints, these two paradigmatic applications share a number of challenges. For example, they both require, at the same time (i) a flexible description of the domain involving multiple scales, interfaces, network of fractures, and strongly heterogeneous media; and (ii) an accurate representation of the solution field, particularly for wave propagation phenomena, where a sufficiently high number of nodes per wavelength is needed to keep numerical dispersion and dissipation errors low. PolyDG methods are perfectly suited to tame all these mathematical and numerical challenges, indeed (i) they are naturally oriented towards high-order approximations, in any space dimension, and feature a high-level of intrinsic parallelism; (ii) the dimension of the local approximation space only depends on the local approximation order, and is independent of the shape of an element and the number of faces/edges of an element. As a consequence, in contrast to other polytopic finite element methods, on agglomeration based meshes the dimension of the local space remains under control; (iii) they can handle mesh elements with possibly an unbounded number of faces and face/edge degeneration can be supported. We point out that the last feature is very important in practical applications, since it allows for hybrid mesh algorithms that efficiently deal with heterogeneous media, localized geological/topographic irregularities, faults and fractures characterizing geophysical applications. The main idea consists in generating an initial (hexahedral/tetrahedral in three dimensions, for example) mesh, based on employing standard mesh generators; then elements intersecting the geological irregularities are suitably cut and/or agglomerated, thus generating polytopes, while keeping a regular structure elsewhere, cf. Figure 5.1 for an illustrative example. Beyond the simplicity of generating the computational hybrid grids based on a convenient combination of hexahedral/tetrahedral/polyhedral elements, one of the other advantages of polytopic decompositions over standard simplicial/hexahedral grids is that, even on relatively simple geometries, the average number of elements needed to discretize complicated domains is substantially smaller [19, 20], without enforcing any domain approximation. This advantage becomes even more evident whenever the domain contains complex geometrical features (large number of fractures, fractures intersecting with small angles, etc.) and the underlying grid is chosen to be matching with the interfaces.

In the following we provide a brief description of the contents of each of the following sections, and highlight their scientific importance within the community. In Sect. 5.2 we introduce the notation and the key theoretical results needed to analyze

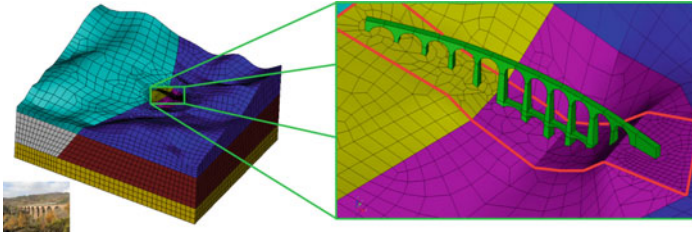


Fig. 5.1 A three-dimensional example of hybrid hexahedral/polyhedral grid of the Acquasanta railway bridge on the Genoa-Ovada railway (Italy). The mesh is obtained by exploiting the flexibility of polyhedral elements near the Acquasanta bridge (cf. the region delimited by the red line of the zoom) while keeping a regular structure elsewhere

PolyDG approximations. In particular, we summarize the main theoretical results concerning this class of methods outlined in [13, 63, 66, 67]. Following [63, 66], we start from the generalization of the standard shape-regularity property to polytopic elements and introduce some trace and inverse inequalities and polynomial approximation properties of the underlying discrete spaces. These results represent the main tools for handling elements with a degenerating and/or unlimited number of faces/edges. The contents of this section form the basis for the theoretical analysis of the discretization schemes for seismic wave propagation and flow in fractured porous media presented in the second part of the manuscript.

Section 5.3 focuses on the construction and outline of a new technique for the efficient computation of integrals of polynomial functions over convex and non-convex polytopic domains, and its application to the numerical computation of the terms appearing in the weak formulation of PolyDG methods. The classical (and most widely employed) approach for the integration of polynomial functions over polytopes is the so called *sub-tessellation* method: here, the domain of integration is sub-divided (sub-tessellated) into d -simplices, whereby standard quadrature rules are employed, cf. [101, 117, 122, 137, 139, 150] and also [115, 151] for a similar quadrature approach where the polytopic domain of integration is sub-tessellated into d -parallelograms. However, the *sub-tessellation* method is generally computationally expensive as it leads to a very large number of function evaluations, particularly when the integrand is a high order polynomial. For this reason, the development of quadrature rules that avoid sub-tessellation and optimize the number of function evaluations is an active research field. Several approaches have been proposed; in particular, we mention [105, 126, 145, 146], for example. Other approaches are represented by the *Moment Fitting Equation* technique, firstly proposed in [116], for the construction of quadrature rules on polygons featuring the same symmetry as the regular hexagon. The key idea here is, starting from a quadrature rule on the integration domain which integrates exactly a class of basis functions for a desired function space, an iterative node elimination procedure is then applied under an exactness constraint. This leads to the definition of a new quadrature rule where the number of function evaluations is optimized. Further improvements of the moment

fitting equation algorithm can also be found in [123] and [138], see also [124] for a generalization to more general convex and non-convex polytopic domains. The main drawback of the moment fitting approach is the need to store the resulting nodes and weights for every polytope, which severely affects memory efficiency when applied to finite element approximations. An alternative approach designed to overcome the limitations of the sub-tessellation and the moment fitting equation methods is based on employing the generalized version of Stokes' theorem; with this approach, the integral over a generic domain is reduced to an integration over its boundary; we refer to [143] for further details. Following this idea, Sommariva and Vianello proposed in [135] a quadrature rule where, if an x - or y -primitive of the integrand is available (as for bivariate polynomial functions), the integral over the polygon is reduced to a sum of line integrations over its edges, each of which is then computed exactly with a Gaussian one dimensional quadrature rule. The authors also generalized this approach to the more general case when the primitive is not known. While this algorithm does not directly require a sub-tessellation of the polygon, a careful choice of the parameters in the proposed formula leads to a quadrature rule that can be viewed as a particular sub-partition of the polygon itself. Moreover, in this case it is not possible to guarantee that all of the quadrature points lie inside the domain of integration. An alternative approach, proposed by Lassere in [114], provides a very efficient formula for the integration of *homogeneous functions* over convex polytopes. Here, the essential idea is to exploit the generalized Stokes' theorem together with Euler's homogeneous function theorem, cf. [134], in order to reduce the integration over a polytope only to boundary evaluations. The main difference with respect to the work presented in [135] is the possibility to apply the same idea recursively, leading to a quadrature formula which exactly evaluates integrals over a polygon/polyhedron by employing only point-evaluations of the integrand and its partial derivatives at the vertices of the polytope. This technique has been recently extended to general convex and non-convex polytopes in [74]. In Sect. 5.3 we present an efficient *quadrature free* algorithm for the numerical approximation of integrals of polynomial functions over general polygonal/polyhedral elements that do not require an explicit construction of a sub-tessellation. The method extends the idea of [74, 114] and is based on successive application of Stokes' theorem; thereby, the underlying integral may be evaluated using only the values of the integrand and its derivatives at the vertices of the polytopic domain. To demonstrate the practical performance of this *quadrature free* method we present some numerical results obtained by the numerical computation of the stiffness and mass matrices arising from hp -version PolyDG discretization of second-order elliptic partial differential equations.

Section 5.4 focuses on the analysis of PolyDG methods for the numerical discretization of seismic wave propagation; that is the ground motion phenomenon induced by the passage of body waves radially from the source of earthquake energy released into the surrounding soil medium. In the context of numerical modeling of direct and inverse wave propagation phenomena, many contributions can be found in the literature, stimulated not only by geophysical problems but also from vibroacoustics, aeroacoustic, acoustics, and electromagnetics engineering applications [50,

51, 70, 80, 82, 89, 102, 111, 144, 149]. Here, our target are large-scale seismological phenomena and ground-motion induced by seismic events. Seismic waves are elastic waves propagating within the Earth and along its surface as a result of an earthquake, or of an explosion. Seismic waves induce a vibratory ground-motion in the area surrounding the seismic source. From the mathematical viewpoint, the propagation of seismic waves in a (visco)elastic heterogeneous material can be modeled by means of the elastodynamics equation. In order to solve the elastodynamics equation based on employing a finite element based numerical scheme, a number of distinguishing challenges have to simultaneously be taken into account which reflect the key features required by the numerical scheme: *accuracy*, *geometric flexibility* and *scalability*. High-order *accuracy* is mandatory in order to correctly approximate wave velocities, i.e., to keep as low as possible both the numerical dissipation and dispersion. *Geometric flexibility* is mandatory since within earthquake engineering the computational domain usually features complicated geometrical details, as well as sharp contrasts in the media. Finally, for real earthquake models, the size of the excited body is very large compared to the wave lengths of interest: this typically leads to numerical models featuring hundred of millions of unknowns, and therefore massively parallel *scalable* algorithms are required. Within the context of numerical methods for the approximation of the elastodynamics equation in computational seismology, spectral element methods are one of the most successfully employed tools, in particular for large scale applications; see, for example, [52, 93, 100, 112, 113]. To enhance the flexibility of spectral element methods, in recent years DG and DG spectral element (DGSE) methods have been extensively used for elastic waves propagation, see e.g. [6, 10, 11, 25, 29, 83, 91, 92, 110, 121, 125, 131, 132], and [17] for an overview on the numerical modeling of seismic waves by DGSE methods. Given their local nature, DG methods are particularly well suited to deal with highly heterogeneous media, or in soil-structure interaction problems, where local refinements are needed to resolve the different spatial scales [120]. In the context of time integration of the (second-order) ordinary differential systems stemming from spatial discretization of second-order hyperbolic partial differential equations we also mention the DG time-integration scheme of [27]. Very recently, also PolyDG methods have been shown to be perfectly suited to reduce the complexity of modelling wave propagation problems. Indeed, on the one hand they further enhance the geometric flexibility offered by ‘classical’ DG schemes on simplicial/tensor-product meshes, allowing for grids composed by arbitrarily shaped elements, with possibly degenerating faces, thus reducing the computational costs related to the process of grid generation, while maintaining the same degree of accuracy. On the other hand, they guarantee lower dispersion errors compared to classical DG schemes on simplicial/tensor-product grids of comparable granularity, see [26].

Section 5.5 is concerned with the numerical approximation of Darcy flows through porous media enclosing networks of fractures. The focus is on presenting a unified design and analysis of PolyDG methods on general polytopic meshes with possibly degenerating edges/faces. The problem of developing efficient numerical methods for fractured reservoir simulations has received increasing attention in the past decades, being fundamental in many energy and environmental engineering applications, such

as water resources management, oil migration tracing, isolation of radioactive waste and groundwater contamination, for example. Fractures are regions of the porous medium featuring a different porous structure, so that they usually have a strong impact on the flow, possibly acting as barriers for the fluid (when they are filled with low permeable material), or as preferential paths (when their permeability is higher than that of the surrounding medium). Moreover, fractures are characterised by a very small width compared to their length and to the size of the domain. For this reason, one popular modelling choice consists in treating them as $(d - 1)$ -dimensional interfaces between d -dimensional porous matrices, $d = 2, 3$. The development of this kind of reduced models has been addressed for single-phase flows in several works, see, e.g., [2, 3, 98, 118]. We will refer mainly to the model described in [118], see also [81], which considers the simplified case of a single, non-immersed fracture. Here, the flow in the porous medium (bulk) is assumed to be governed by Darcy's law and a suitable reduced version of the law is formulated also on the surface modelling the fracture. Physically consistent coupling conditions are then added (in strong form) to account for the exchange of fluid between the fracture and the porous medium. We remark that this model is able to handle both fractures with low and large permeability. Even if the use of this kind of dimensionally reduced models avoids the need for extremely refined grids inside the fracture domains, in realistic cases, the construction of a computational grid aligned with the fractures is still a major issue. For example, fractured oil reservoirs can feature thousands of fractures, which are often intersecting with small angles or nearly coincident [96]. In line with the discussion above, our aim is then to take advantage of the intrinsic geometric flexibility of PolyDG methods for the approximation of the coupled bulk-fracture problem, thus avoiding the limitations imposed by standard finite element methods. We also point out that various other numerical methods supporting polytopic elements have been employed in the literature for the approximation of this problem. In particular, we mention [18, 96], where a mixed approximation based on Mimetic Finite Differences has been explored; the works [47, 48], where a framework for treating flows in Discrete Fracture Networks based on the Virtual Element Method has been introduced, and [71], where the Hybrid High-Order method has been employed. We also mention that an alternative strategy consists in the use of non-conforming discretizations. Here, the bulk grid can be chosen fairly regular since fractures are allowed to arbitrarily cut it. We refer to [81, 94, 99] for the use of the eXtended Finite Element Method and to [61] for the Cut Finite Element Method. Notice that the geometric flexibility of PolyDG methods illustrated above is not the only motivation to employ these kinds of techniques for addressing this problem. Another important issue is that the discontinuous nature of the solution at the matrix-fracture interface is intrinsically captured in the choice of the discrete spaces. Moreover, coupling conditions between bulk and fracture can be easily reformulated using jump and average operators (basic tools for the construction of DG methods) and then naturally embedded in the variational formulation. Furthermore, employing the abstract setting, based on the *flux-formulation*, introduced in [33] for the *unified* analysis of all DG methods present in the literature, it is possible to introduce a unified framework where, according to the desired approximation properties of the

model, one may resort to either a primal or mixed approximation for the problem in the bulk, as well as to a primal or mixed approximation for the problem in the fracture network. In particular, the primal discretizations are obtained using the Symmetric Interior Penalty DG method [32, 148], whereas the mixed discretizations are based on employing the Local DG (LDG) method of [79], both in their generalization to polytopic grids. Finally, we point out that, even if not addressed here, our formulation can be extended to the case of networks of *intersecting* fractures, cf. [15] and Sect. 5.5.4.

5.2 Theoretical Framework of PolyDG Methods

In this section we introduce the necessary notation and key analytical results required for the definition and analysis of PolyDG approximations. In particular, we summarize the main theoretical results concerning this class of methods contained in [13, 63, 65, 67], where an *hp*-version interior penalty PolyDG method for the numerical approximation of elliptic problems on polytopic meshes has been proposed and analysed. The exploitation of grids consisting of general polytopic elements poses a number of key challenges. Indeed, in contrast to the case when standard-shaped elements are employed, polytopes may admit an arbitrary number of faces/edges and the measure of these faces/edges may potentially be much smaller than the measure of the element itself. In [12, 13, 67] it is assumed that the number of edges/faces of each mesh element is uniformly bounded. In [63, 65] this assumption is no longer required (i.e., elements with an arbitrary number of possibly degenerating faces/edges are admitted). However, this comes at the cost of adding an assumption (see Sect. 5.2 below) that may be regarded as the natural generalization to polytopic grids of the classical shape-regularity assumption [65]. For ease of presentation, we adopt the setting of [63, 65]; for the generalization to other classes of polytopic meshes, we refer to the recent article [62]. In particular, in Sect. 5.2.1, we introduce the notation related to the discretization of domains using polytopic elements and state the regularity assumptions on the meshes. In Sect. 5.2.2 we define the DG discrete spaces and introduce standard jump and average operators. Finally, in Sect. 5.2.3, starting from the mesh assumptions of Sect. 5.2.1, we state trace inverse inequalities and approximation results for general polytopic elements that are sensitive to the type of edge/face degeneracy described above. We also remark that the capability of the method of handling faces with arbitrarily small measure is intimately related to the correct choice of the discontinuity-penalization function, which will be introduced in the following sections.

We will employ the following notation. For an open, bounded domain $D \subset \mathbb{R}^d$, $d = 2, 3$, we denote by $H^s(D)$ the standard Sobolev space of order s , for a real number $s \geq 0$. For $s = 0$, we write $L^2(D)$ in lieu of $H^0(D)$. The usual norm on $H^s(D)$ is denoted by $\|\cdot\|_{H^s(D)}$ and the usual seminorm by $|\cdot|_{H^s(D)}$. We denote the corresponding Sobolev spaces of vector-valued functions and symmetric tensors by $\mathbf{H}^m(\Omega) = [H^m(D)]^d$, $\mathcal{H}^m(D) = [H^m(D)]_{\text{sym}}^{d \times d}$, $d = 2, 3$, respectively. We also intro-

duce the standard space $H_{div}(D) = \{\mathbf{v} : D \rightarrow \mathbb{R}^d : \|\mathbf{v}\|_{L^2(D)} + \|\nabla \cdot \mathbf{v}\|_{L^2(D)} < \infty\}$. Given a decomposition of the domain into a computational mesh \mathcal{T}_h , we denote by $H^s(\mathcal{T}_h)$ the standard *broken* Sobolev space, equipped with the broken norm $\|\cdot\|_{s,\mathcal{T}_h}$. Furthermore, we denote by $\mathbb{P}_k(D)$ the space of polynomials of *total* degree less than or equal to $k \geq 1$ on D . The symbols \lesssim and \gtrsim will signify that the inequalities hold up to multiplicative constants that are independent of the discretization parameters, but might depend on the physical parameters of the underlying problem.

5.2.1 Grid Assumptions

Following [13, 65, 67], we introduce the notation related to the subdivision of the computational domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, by means of polytopic meshes. We consider classes of meshes \mathcal{T}_h made of disjoint open *polygonal/polyhedral* elements E . For each element $E \in \mathcal{T}_h$, we denote by $|E|$ its measure, h_E its diameter and we set $h = \max_{E \in \mathcal{T}_h} h_E$. With the aim of handling hanging nodes, we introduce the concept of mesh *interfaces*, which are defined as the intersection of the $(d - 1)$ -dimensional facets of two neighbouring elements. We need now to distinguish between the case when $d = 3$ and $d = 2$:

- when $d = 3$, each interface consists of a general polygon, which we assume may be decomposed into a set of co-planar triangles. We assume that a sub-triangulation of each interface is provided and we denote the set of all these triangles by \mathcal{F}_h . We then use the terminology *face* to refer to one of the triangular elements in \mathcal{F}_h ;
- when $d = 2$, each interface simply consists of a line segment, so that the concepts of face and interface are in this case coincident; however, we still denote by \mathcal{F}_h the set of all faces.

Here, we note that \mathcal{F}_h is always defined as a set of $(d - 1)$ -dimensional simplices (triangles or line segments).

In order to introduce the PolyDG formulation, it is useful to further subdivide the set \mathcal{F}_h into

$$\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B,$$

where \mathcal{F}_h^I is the set of interior faces and \mathcal{F}_h^B is the set of faces lying on the boundary of the domain $\partial\Omega$. Moreover, if $\partial\Omega$ is split into the Dirichlet boundary Γ_D and the Neumann boundary Γ_N , we will further decompose the set $\mathcal{F}_h^B = \mathcal{F}_h^D \cup \mathcal{F}_h^N$, where \mathcal{F}_h^D and \mathcal{F}_h^N are the boundary faces contained in Γ_D and Γ_N , respectively. Implicit in this definition is the assumption that the mesh \mathcal{T}_h conforms to the partition of $\partial\Omega$. Next, we outline the key assumptions that the underlying polytopic mesh \mathcal{T}_h needs to satisfy in order to derive suitable inverse inequalities and approximation results. To this end, we write \mathcal{F}_E^F to denote a d -dimensional simplex contained in E which shares a specific face $F \subset \partial E$, $F \in \mathcal{F}_h$. With this notation we introduce the following definition.

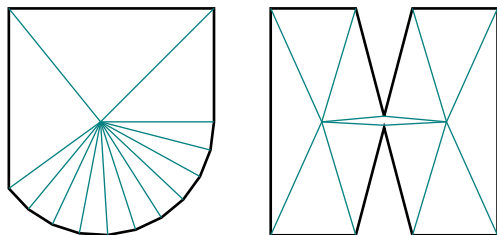
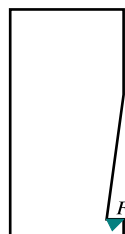


Fig. 5.2 Two examples of polytopic-regular elements as in Definition 5.1. Here, all the triangles S_E^F (coloured in teal) have height of size comparable to the diameter h_E . Note also that the element on the right is not covered by the union of the simplices

Fig. 5.3 Example of an element that violates polytopic-regularity: the shape of the polygon does not allow for the definition of a triangle S_E^F with base F whose height is comparable to the diameter h_E



Definition 5.1 A family of meshes $\{\mathcal{T}_h\}_h$ is said to be *polytopic-regular* if, for any h and for any $E \in \mathcal{T}_h$, there exists a set of non-overlapping (not necessarily shape-regular) d -dimensional simplices $\{S_E^F\}_{F \subset \partial E}$ contained in E , such that for all faces $F \subset \partial E$, the following condition holds

$$h_E \lesssim \frac{d|S_E^F|}{|F|}, \tag{5.1}$$

where the hidden constant is independent of the discretization parameters, the number of faces of the element, and the face measure.

We remark that the union of simplices $\{S_E^F\}_{F \subset \partial E}$ does *not* have to cover, in general, the whole element E , that is $\cup_{F \subset \partial E} S_E^F \subseteq \bar{E}$, see Fig. 5.2 for an example. We also stress that this definition does not require any restriction on either the number of faces per element or their relative measure. In particular, it allows the size of a face $|F|$, $F \subset \partial E$, to be arbitrarily small compared to the diameter of the element h_E , provided that the height of the corresponding simplex S_E^F is comparable to h_E . Figure 5.2 shows two examples of elements belonging to a polytopic-regular mesh, while Fig. 5.3 shows an element which may not satisfy the definition, for example, when the length of the vertical section of the boundary in the lower right-hand corner tends to zero at a faster rate than the mesh size h . We refer to [65] for more details.

Assumption 5.1 We assume that the family of meshes $\{\mathcal{T}_h\}_h$ is uniformly *polytopic-regular*. □

This assumption will allow us to state the inverse trace estimate (5.1) below. The next definition and assumption are instrumental for the validity of the approximation results (5.2) below.

Definition 5.2 [13, 63, 65, 67] A covering $\mathcal{T}_\# = \{T_E\}$ related to the polytopical mesh \mathcal{T}_h is a set of shape-regular d -dimensional simplices T_E , such that for each $E \in \mathcal{T}_h$, there exists a $T_E \in \mathcal{T}_\#$ such that $E \subsetneq T_E$.

Assumption 5.2 [13, 63, 65, 67] There exists a covering $\mathcal{T}_\#$ of \mathcal{T}_h (see Definition 5.2) and a positive constant O_Ω , independent of the mesh parameters, such that

$$\max_{E \in \mathcal{T}_h} \text{card}\{E' \in \mathcal{T}_h : E' \cap T_E \neq \emptyset, T_E \in \mathcal{T}_\# \text{ s.t. } E \subset T_E\} \leq O_\Omega,$$

and $h_{T_E} \lesssim h_E$ for each pair $E \in \mathcal{T}_h$ and $T_E \in \mathcal{T}_\#$, with $E \subset T_E$. \square

Assumption 5.2 implies that, when the computational mesh \mathcal{T}_h is refined, the amount of overlap present in the covering $\mathcal{T}_\#$ remains bounded.

5.2.2 PolyDG Discrete Spaces

Given a polytopical mesh partition \mathcal{T}_h of the domain Ω , the corresponding scalar, vector-valued and symmetric tensor-valued discontinuous finite element spaces are defined as

$$Q_h^{DG} = \{q_h \in L^2(\Omega) : q|_E \in \mathbb{P}_{p_E}(E) \forall E \in \mathcal{T}_h\}, \quad (5.2)$$

$$\mathbf{W}_h^{DG} = \{\mathbf{w} \in [L^2(\Omega)]^d : \mathbf{w}|_E \in [\mathbb{P}_{p_E}(E)]^d \forall E \in \mathcal{T}_h\}, \quad (5.3)$$

$$\mathbf{W}_h^{DG} = \{\mathbf{w} \in [L^2(\Omega)]_{\text{sym}}^{d \times d} : \mathbf{w}|_E \in [\mathbb{P}_{p_E}(E)]_{\text{sym}}^{d \times d} \forall E \in \mathcal{T}_h\}, \quad (5.4)$$

where we assume that $p_E \geq 1$ for all $E \in \mathcal{T}_h$. To avoid technicalities, for the analysis, we assume that a *local bounded variation* property holds for both the polynomial approximation degrees and the local mesh sizes, cf. [127].

Remark 5.1 From the implementation point of view, an essential feature of DG methods is that the local elemental polynomial spaces can be defined in the physical space, without the need to introduce a mapping to a reference element, as is typically necessary for classical finite element methods. This allows DG methods to naturally deal with general polytopical elements with polynomial degrees varying from one element to the other. A possible approach for the definition of the basis functions was first proposed in [67], based on the definition of the polynomial spaces over suitably defined bounding boxes of each polytopical element. More precisely, given an element $E \in \mathcal{T}_h$, we can define its (for example) Cartesian bounding box B_E , such that the sides of B_E are aligned with the Cartesian axes and $\bar{E} \subseteq \bar{B}_E$. On the Cartesian bounding box B_E , we can then define a standard polynomial space, employing, for

example, tensor-product Legendre polynomials. Finally, the polynomial basis over the general polytopic element may be defined by simply restricting the support of the basis functions to E ; we refer to [65] for further details. We also mention that another key aspect related to the implementation of DG methods is the design of efficient numerical integration schemes over polytopic elements; this issue will be addressed in detail in the forthcoming (Sect. 5.3.1), where a *quadrature-free* approach for the efficient integration of polynomial functions over polytopic domains will be discussed, following the recent work [22].

In order to efficiently deal with discontinuous functions, we now introduce average and jump operators on a face, which play a central role in the design and analysis of all DG methods [33]. Let $F \in \mathcal{F}_h^I$ be an interior face shared by the elements E^\pm . We define \mathbf{n}^\pm to be the unit normal vectors on F pointing exterior to E^\pm , respectively. Then, for sufficiently regular scalar-valued, vector-valued, and tensor-valued functions q , \mathbf{v} , and $\boldsymbol{\tau}$, respectively, we define the standard *average* $\{\cdot\}$ and *jump* $[[\cdot]]$ operators on F as

$$\begin{aligned} \{q\} &= \frac{1}{2}(q^+ + q^-), & [[q]] &= q^+ \mathbf{n}^+ + q^- \mathbf{n}^-, \\ \{\mathbf{v}\} &= \frac{1}{2}(\mathbf{v}^+ + \mathbf{v}^-), & [[\mathbf{v}]] &= \mathbf{v}^+ \cdot \mathbf{n}^+ + \mathbf{v}^- \cdot \mathbf{n}^-, \\ \{\boldsymbol{\tau}\} &= \frac{1}{2}(\boldsymbol{\tau}^+ + \boldsymbol{\tau}^-), & [[\boldsymbol{\tau}]] &= \boldsymbol{\tau}^+ \mathbf{n}^+ + \boldsymbol{\tau}^- \mathbf{n}^-, \end{aligned} \quad (5.5)$$

where the subscript \pm on q , \mathbf{v} , and $\boldsymbol{\tau}$ denote the respective traces of the functions on F restricted to E^\pm , respectively. To tackle elastic wave propagation phenomena, we also need the following jump operator for a sufficiently regular vector-valued function \mathbf{v} :

$$[[[\mathbf{v}]]] = \mathbf{v}^+ \odot \mathbf{n}^+ + \mathbf{v}^- \odot \mathbf{n}^-,$$

where $\mathbf{v} \odot \mathbf{n} = (\mathbf{v} \mathbf{n}^\top + \mathbf{n} \mathbf{v}^\top)/2$. Notice that with the above definition $[[[\mathbf{v}]]]$ is a $d \times d$ symmetric tensor. On a boundary face $F \in \mathcal{F}_h^B$ we set analogously $\{q\} = q$, $[[q]] = q \mathbf{n}$, $\{\mathbf{v}\} = \mathbf{v}$, $[[\mathbf{v}]] = \mathbf{v} \cdot \mathbf{n}$, $[[[\mathbf{v}]]] = \mathbf{v} \odot \mathbf{n}$, $\{\boldsymbol{\tau}\} = \boldsymbol{\tau}$, and $[[\boldsymbol{\tau}]] = \boldsymbol{\tau} \mathbf{n}$, where \mathbf{n} is the outward unit normal vector on $\partial\Omega$, cf. [33, 34]. For future use, we remark that on every $F \in \mathcal{F}_h^I$ we can use the definition of jump and average operators to write

$$[[q\mathbf{v}]] = [[\mathbf{v}]]\{q\} + \{\mathbf{v}\} \cdot [[q]]. \quad (5.6)$$

We also recall the identity:

$$\sum_{E \in \mathcal{T}_h} \int_{\partial E} q \mathbf{v} \cdot \mathbf{n}_E = \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^B} \{\mathbf{v}\} \cdot [[q]] + \int_{\mathcal{F}_h^I} [[\mathbf{v}]]\{q\}, \quad (5.7)$$

cf. [32], where we have used the compact notation $\int_{\mathcal{F}_h} = \sum_{F \in \mathcal{F}_h} \int_F$.

5.2.3 Trace Inverse Estimates on Polytopic Elements

Trace inverse estimates are one of the key tools employed to study the stability and error analysis of DG-methods: they bound the norm of a polynomial on an element's face/edge by the norm on the element itself. In particular, Lemma 5.1 is required to establish the stability of the PolyDG approximation of second-order elliptic partial differential equations. Trace inverse estimates on polytopic elements are obtained under the polytopic-regular Assumption 5.1 as in [63], Lemma 4.1, and [21, 65]; the proof is reported here for completeness.

Lemma 5.1 *Let E be a polytope satisfying Assumption 5.1 and let $q \in \mathbb{P}_{p_E}(E)$. Then, we have*

$$\|q\|_{L^2(\partial E)}^2 \lesssim \frac{p_E^2}{h_E} \|q\|_{L^2(E)}^2, \quad (5.8)$$

where the hidden constant depends on the dimension d , but it is independent of the discretization parameters, i.e., the local mesh size h_E and the local polynomial approximation degree p_E , and the number of faces that the element possesses.

Proof The proof follows immediately if we apply “classical” hp -version inverse estimate valid for generic simplices, see, e.g., [147], to each simplex $S_E^F \subset E$, cf. Assumption 5.1, together with (5.1), i.e.,

$$\begin{aligned} \|q\|_{L^2(\partial E)}^2 &= \sum_{F \subset \partial E} \|q\|_{L^2(F)}^2 \lesssim p_E^2 \sum_{F \subset \partial E} \frac{|F|}{|S_E^F|} \|q\|_{L^2(S_E^F)}^2 \\ &\lesssim \frac{p_E^2}{h_E} \|q\|_{L^2(\cup_{F \subset \partial E} S_E^F)}^2 \leq \frac{p_E^2}{h_E} \|q\|_{L^2(E)}^2. \end{aligned}$$

5.2.4 Polynomial Approximation over Polytopic Elements

A crucial mathematical tool needed to study the *a priori* error analysis of PolyDG methods are hp -interpolation estimates. In [13, 65, 67] standard results on simplices are extended to polytopic elements, based on considering appropriate coverings and submeshes consisting of d -dimensional simplices (where standard results can be applied) and using an appropriate extension operator. In [63] these results are further extended in order to be successfully applied also in the case when the number of edges/faces is unbounded. Here, we recall the results contained in [13, 63, 65, 67].

Let $\mathcal{E} : H^s(\Omega) \rightarrow H^s(\mathbb{R}^d)$, $s \geq 0$, be the continuous extension operator introduced by Stein in [136], such that $\mathcal{E}(q)|_\Omega = q$ and $\|\mathcal{E}q\|_{H^s(\mathbb{R}^d)} \lesssim \|q\|_{H^s(\Omega)}$. Based on the existence of a suitable covering of the polytopic mesh (see Definition 5.2), we can state the following approximation result.

Lemma 5.2 [13, 63, 65, 67] *Assume that Assumptions 5.1 and 5.2 are satisfied. Given $E \in \mathcal{T}_h$, let $T_E \in \mathcal{T}_\#$ be the corresponding simplex such that $E \subset T_E$ (see Definition 5.2). For $q \in L^2(\Omega)$, such that $\mathcal{E}q|_{T_E} \in H^{r_E}(T_E)$, for some $r_E \geq 0$, there exists a sequence of approximations $\Pi_E^{p_E} q \in \mathbb{P}_{p_E}(E)$, $p_E = 0, 1, 2, \dots$, of q satisfying*

$$\|q - \Pi_E^{p_E} q\|_{H^m(E)} \lesssim \frac{h_E^{s_E - m}}{p_E^{r_E - m}} \|\mathcal{E}q\|_{H^{r_E}(T_E)}, \quad 0 \leq m \leq r_E. \quad (5.9)$$

Moreover, if $r_E \geq 1 + d/2$,

$$\|q - \Pi_E^{p_E} q\|_{L^2(\partial E)} \lesssim \frac{h_E^{s_E - 1/2}}{p_E^{r_E - 1/2}} \|\mathcal{E}q\|_{H^{r_E}(T_E)}. \quad (5.10)$$

Here, $s_E = \min(p_E + 1, r_E)$ and the hidden constants depend on the shape-regularity of T_E , but are independent of q , h_E , p_E and the number of faces per element. \square

Proof See [67] for a detailed proof of (5.9) and [63] for the proof of (5.10).

We note that the inequalities (5.8) and (5.10) hold on the whole boundary of E , and not just on one of its edges/faces; this is of fundamental importance in the analysis when considering elements that contain an arbitrary number of faces.

5.3 Computing Integrals over Polytopic Mesh Elements and Mesh Interfaces

In this section we review the *quadrature free* approach for the efficient computation of the volume/face integral terms appearing in PolyDG methods. We point out that our approach is completely general and can be directly applied to other discretization schemes, such as VEM, HHO, Hybridizable DG, and MFD, for example. We present the main idea of the algorithm and show that our integration approach leads to a considerable improvement in the computational performance compared to classical quadrature algorithms based on sub-tessellation, in both two- and three-dimensions.

5.3.1 Quadrature Free Algorithm

First, we recall the idea introduced by Chin, Lasserre, and Sukumar in [74] for the integration of homogeneous function g over a polytopic domain \mathcal{P} , where

- $\mathcal{P} \subset \mathbb{R}^d$, $= 2, 3$, is a closed polytope, whose boundary $\partial\mathcal{P}$ is defined by m $(d - 1)$ -dimensional faces F_i , $i = 1, \dots, m$, cf. Fig. 5.4. Each face F_i lies in a hyperplane \mathcal{H}_i identified by a vector $\mathbf{a}_i \in \mathbb{R}^d$ and a scalar number b_i , such that

$$\mathbf{x} \in \mathcal{H}_i \iff \mathbf{a}_i \cdot \mathbf{x} = b_i, \quad i = 1, \dots, m. \quad (5.11)$$

We observe that $\mathbf{a}_i, i = 1, \dots, m$, can be chosen as the unit outward normal vector to $F_i, i = 1, \dots, m$, respectively, relative to \mathcal{P} .

- $g : \mathcal{P} \rightarrow \mathbb{R}$ is a *homogeneous function* of degree $q \in \mathbb{R}$, i.e., for all $\lambda > 0, g(\lambda \mathbf{x}) = \lambda^q g(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{P}$.

Euler's homogeneous function theorem [134] states that, if g is a *homogeneous function* of degree $q \geq 0$, then the following identity holds:

$$q g(\mathbf{x}) = \nabla g(\mathbf{x}) \cdot \mathbf{x} \quad \forall \mathbf{x} \in \mathcal{P}. \quad (5.12)$$

We point out that, in view of the application to polygonal/polyhedral finite element methods, we are interested in the integration of a particular class of *homogeneous functions*, namely *polynomial homogeneous functions* of the form

$$g(\mathbf{x}) = x_1^{k_1} x_2^{k_2} \cdots x_d^{k_d}, \quad \text{where } k_n \in \mathbb{N}_0, \text{ for } n = 1, \dots, d, \quad (5.13)$$

that is a *homogeneous function* of degree $q = k_1 + \cdots + k_d$, and the general partial derivative $\frac{\partial g}{\partial x_n}$ is still a *homogeneous function* of degree $q - 1$.

Next we recall the generalized Stokes' theorem, cf. [143]: given a generic vector field $\mathbf{X} : \mathcal{P} \rightarrow \mathbb{R}^d$, we have that

$$\int_{\mathcal{P}} (\nabla \cdot \mathbf{X}(\mathbf{x})) g(\mathbf{x}) + \int_{\mathcal{P}} \nabla g(\mathbf{x}) \cdot \mathbf{X}(\mathbf{x}) = \int_{\partial \mathcal{P}} \mathbf{X}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) g(\mathbf{x}), \quad (5.14)$$

where \mathbf{n} is the unit outward normal vector to \mathcal{P} . Selecting $\mathbf{X} = \mathbf{x}$ in (5.14), and employing (5.12), gives

$$\int_{\mathcal{P}} g(\mathbf{x}) = \frac{1}{d+q} \int_{\partial \mathcal{P}} \mathbf{x} \cdot \mathbf{n}(\mathbf{x}) g(\mathbf{x}) = \frac{1}{d+q} \sum_{i=1}^m b_i \int_{F_i} g(\mathbf{x}). \quad (5.15)$$

Equation (5.15) states that the integral of a homogeneous function g over a polytope \mathcal{P} can be computed by integrating the same function over the boundary faces $F_i \subset \partial \mathcal{P}, i = 1, \dots, m$. By recursion, we can further reduce each term $\int_{F_i} g(\mathbf{x}), i = 1, \dots, m$, to the integration over $\partial F_i, i = 1, \dots, m$, respectively. To this end, Stokes' theorem needs to be applied on the hyperplane $\mathcal{H}_i, i = 1, \dots, m$, in which each $F_i, i = 1, \dots, m$, lies, respectively. In order to proceed, let $\boldsymbol{\gamma} : \mathbb{R}^{d-1} \rightarrow \mathbb{R}^d$ be the function which expresses a generic point $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_{d-1})^\top \in \mathbb{R}^{d-1}$ as a point in \mathbb{R}^d that lies on $\mathcal{H}_i, i = 1, \dots, m$, i.e.,

$$\tilde{\mathbf{x}} \longmapsto \boldsymbol{\gamma}(\tilde{\mathbf{x}}) = \mathbf{x}_{0,i} + \sum_{n=1}^{d-1} \tilde{x}_n \mathbf{e}_{in}, \quad \text{with } \mathbf{e}_{in} \in \mathbb{R}^d, \quad \mathbf{e}_{in} \cdot \mathbf{e}_{im} = \delta_{nm}. \quad (5.16)$$

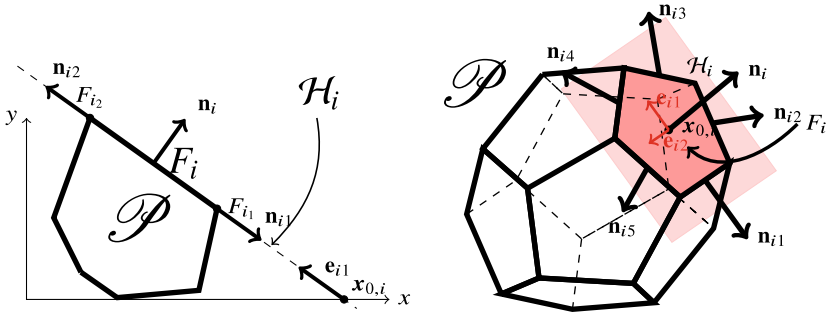


Fig. 5.4 Left: Example of a two-dimensional polytope \mathcal{P} and its face F_i . The hyperplane \mathcal{H}_i is defined by the local origin $\mathbf{x}_{0,i}$ and the vector \mathbf{e}_{i1} . Right: the dodecahedron \mathcal{P} with pentagonal faces and the face $F_i \subset \partial \mathcal{P}$ with unit outward normal vector \mathbf{n}_i . Here, F_i has five edges F_{ij} , $j = 1, \dots, 5$, and five unit outward normal vectors \mathbf{n}_{ij} , $j = 1, \dots, 5$, lying on the plane \mathcal{H}_i . The hyperplane \mathcal{H}_i is identified by the local origin $\mathbf{x}_{0,i}$ and the orthonormal vectors $\mathbf{e}_{i1}, \mathbf{e}_{i2}$. Figure taken from [22]

Here, $\mathbf{x}_{0,i} \in \mathcal{H}_i$, $i = 1, \dots, m$, is an arbitrary point which represents the origin of the coordinate system on \mathcal{H}_i , and $\{\mathbf{e}_{in}\}_{n=1}^{d-1}$ is an orthonormal basis on \mathcal{H}_i , $i = 1, \dots, m$; see Fig. 5.4. We observe that $\mathbf{x}_{0,i}$ does not have to lie inside F_i , $i = 1, \dots, m$. By defining $\tilde{F}_i \subset \mathbb{R}^{d-1}$ such that $\boldsymbol{\gamma}(\tilde{F}_i) = F_i$, $i = 1, \dots, m$, we obtain

$$\int_{F_i} g(\mathbf{x}) = \int_{\tilde{F}_i} g(\boldsymbol{\gamma}(\tilde{\mathbf{x}})), \quad i = 1, \dots, m. \tag{5.17}$$

It is easy to prove that, writing $F_{ij} \subset \partial F_i$, $j = 1, \dots, m_i$, to denote the vertices/edges of F_i , $i = 1, \dots, m$, for $d = 2, 3$, respectively, the following identity holds

$$\tilde{\mathbf{n}}_{ij} = \mathbf{E}^\top \mathbf{n}_{ij}, \quad i = 1, \dots, m, \quad j = 1, \dots, m_i, \tag{5.18}$$

where \mathbf{n}_{ij} is the unit outward normal vectors to F_{ij} lying in \mathcal{H}_i , $\mathbf{E} \in \mathbb{R}^{d \times (d-1)}$, whose columns are the vectors $\{\mathbf{e}_{in}\}_{n=1}^{d-1}$, $i = 1, \dots, m$, $\tilde{F}_{ij} \subset \partial \tilde{F}_i$ is the preimage of F_{ij} with respect to the map $\boldsymbol{\gamma}$, and $\tilde{\mathbf{n}}_{ij}$ are the corresponding unit outward normal vectors; we refer to [22] for more details. Next we recall the following result.

Proposition 5.1 [22, Proposition 1] *Let F_i , $i = 1, \dots, m$, be a face of the polytope \mathcal{P} , and let F_{ij} , $j = 1, \dots, m_i$, be the planar/straight faces/edges such that $\partial F_i = \cup_{j=1}^{m_i} F_{ij}$ for some $m_i \in \mathbb{N}$. Then, for any homogeneous function g , of degree $q \geq 0$, the following identity holds*

$$\int_{F_i} g(\mathbf{x}) = \frac{1}{d-1+q} \left(\sum_{j=1}^{m_i} d_{ij} \int_{F_{ij}} g(\mathbf{x}) + \int_{F_i} \mathbf{x}_{0,i} \cdot \nabla g(\mathbf{x}) \right), \tag{5.19}$$

where d_{ij} denotes the Euclidean distance between F_{ij} and $\mathbf{x}_{0,i}$, $\mathbf{x}_{0,i} \in \mathcal{H}_i$, is arbitrary, $i = 1, \dots, m$.

Algorithm 5.1 $I(N, \mathcal{E}, k_1, \dots, k_d) = \int_{\mathcal{E}} x_1^{k_1} \dots x_d^{k_d}$

if $N = 0$ ($\mathcal{E} = (v_1, \dots, v_d) \in \mathbb{R}^d$ is a point)

$$\text{return } I(N, \mathcal{E}, k_1, \dots, k_d) = v_1^{k_1} \dots v_d^{k_d}; \quad (5.20)$$

else if $1 \leq N \leq d - 1$ (\mathcal{E} is a point if $d = 1$ or an edge if $d = 2$ or a face if $d = 3$)

$$\begin{aligned} I(N, \mathcal{E}, k_1, \dots, k_d) &= \frac{1}{N + \sum_{n=1}^d k_n} \left(\sum_{i=1}^m d_i I(N - 1, \mathcal{E}_i, k_1, \dots, k_d) \right. \\ &\quad + x_{0,1} k_1 I(N, \mathcal{E}, k_1 - 1, k_2, \dots, k_d) \\ &\quad \left. + \dots + x_{0,d} k_d I(N, \mathcal{E}, k_1, \dots, k_d - 1) \right); \end{aligned}$$

else if $N = d$ (\mathcal{E} is an interval if $d = 1$ or a polygon if $d = 2$ or a polyhedron if $d = 3$)

$$I(N, \mathcal{E}, k_1, \dots, k_d) = \frac{1}{N + \sum_{n=1}^d k_n} \left(\sum_{i=1}^m b_i I(N - 1, \mathcal{E}_i, k_1, \dots, k_d) \right).$$

end if

Using Proposition 5.1, together with equation (5.15), we obtain the following identity

$$\int_{\mathcal{P}} g(\mathbf{x}) = \frac{1}{d+q} \sum_{i=1}^m \frac{b_i}{d-1+q} \left(\sum_{j=1}^{m_i} d_{ij} \int_{F_{ij}} g(\mathbf{x}) + \int_{F_i} \mathbf{x}_{0,i} \cdot \nabla g(\mathbf{x}) \right), \quad (5.21)$$

where we recall that $\partial \mathcal{P} = \cup_{i=1}^m F_i$ and $\partial F_i = \cup_{j=1}^{m_i} F_{ij}$, for $i = 1, \dots, m$. We point out that in two-dimensions, i.e., $d = 2$, then F_{ij} is a point and (5.21) states that the integral of g on \mathcal{P} can be computed by vertex-evaluations of the integrand plus a line integration of the partial derivative of g . If $d = 3$ we can apply Stokes' Theorem recursively to $\int_{F_{ij}} g(\mathbf{x})$. We point out that, whenever g is a *homogeneous polynomial function* of the form (5.13), so that the derivatives of $g(\cdot)$ are *homogeneous polynomial functions* as well, it is possible to recursively apply formula (5.21) to the terms involving the integration of the derivatives of g . With this observation in mind, we define the function that returns the integral of the polynomial $x_1^{k_1} \dots x_d^{k_d}$ over \mathcal{E} as

$$I(N, \mathcal{E}, k_1, \dots, k_d) = \int_{\mathcal{E}} x_1^{k_1} \dots x_d^{k_d}, \quad (5.22)$$

where $\mathcal{E} \subset \mathbb{R}^d$, $d = 2, 3$, can be a N -polytopic domain of integration, with $N = 1, \dots, d$, $\partial \mathcal{E} = \cup_{i=1}^m \mathcal{E}_i$, where each $\mathcal{E}_i \subset \mathbb{R}^d$ is a $(N - 1)$ -polytopic domain. When $N = d$ and $d = 2, 3$, \mathcal{E}_i , $i = 1, \dots, m$, will be an edge or a face, respectively; see Table 5.1 for details. According to Proposition 5.1, the recursive definition of the function $I(\cdot, \cdot, \dots, \cdot)$ is given in Algorithm 5.1. We point out that the computational

Table 5.1 Polytopical domains of integration \mathcal{E} as a function of the dimension d , cf. Algorithm 5.1

	$N = 3$	$N = 2$	$N = 1$	$N = 0$
$d = 3$	$\mathcal{E} = \mathcal{P}$ is a polyhedron	$\mathcal{E} = F_i \subset \partial\mathcal{P}$ is a polygon	$\mathcal{E} = F_{ij} \subset \partial F_i$ is an edge	$\mathcal{E} = F_{ijk} \subset \partial F_{ij}$ is a point
$d = 2$		$\mathcal{E} = \mathcal{P}$ is a polygon	$\mathcal{E} = F_i \subset \partial\mathcal{P}$ is an edge	$\mathcal{E} = F_{ij} \subset \partial F_i$ is a point
$d = 1$			$\mathcal{E} = \mathcal{P}$ is an interval	$\mathcal{E} = F_i \subset \partial\mathcal{P}$ is a point

complexity of Algorithm 5.1 depends in general on the number of recursive calls of the function $\mathcal{I}(\cdot, \cdot, \dots, \cdot)$; a detail discussion on the FLOPS required by Algorithm 5.1 and on optimization strategies to improve the computational complexity of Algorithm 5.1 are discussed in [22]. Here, we just remark that in the context of employing the *quadrature free* approach within a polygonal finite element method, we are not interested in integrating a single monomial function, but instead an entire family of monomials, which, for example, form a basis for the space of polynomials of a given degree over a given polytopical element $E \in \mathcal{T}_h$. For example, when $d = 2$, let us consider the evaluation of

$$\int_E x^{k_1} y^{k_2} \quad \forall k_1, k_2 \geq 0, \quad k_1 + k_2 \leq p. \tag{5.23}$$

As shown in [22], when employing Algorithm 5.1 with an with an optimal choice of the points which define the origin of the coordinate system on each element facet, the total number of FLOPs required for the computation of (5.23) is approximately $\mathcal{O}(p^3)$, as p increases. To improve efficiency, an alternative approach, cf. Algorithms 5.2 and 5.3, are based on the observation that, using the notation of Algorithm 5.1, if the values of $\mathcal{I}(N - 1, \mathcal{E}_j, k_1, \dots, k_d)$, $j = 1, \dots, m$, $\mathcal{I}(N, \mathcal{E}, k_1 - 1, \dots, k_d) \dots \mathcal{I}(N, \mathcal{E}, k_1, \dots, k_d - 1)$, for $1 \leq N \leq d - 1$, in Algorithm 5.1, have already been computed, then the computation of $\mathcal{I}(N, \mathcal{E}, k_1, \dots, k_d)$ is extremely cheap. Indeed, since we must store the integrals of all the monomials on E anyway, we can start by computing and storing $\int_E x^{k_1} y^{k_2}$ related to the lower degrees k_1, k_2 and $N = 1$, then exploit these values in order to compute the integrals with higher degrees k_1, k_2 and higher dimension N of the integration domain \mathcal{E} . We remark that, in Algorithm 5.3, d_{ij} represents the Euclidean distance between \mathcal{E}_{ij} and \mathbf{x}_0 , $j = 1, \dots, m_{ij}$.

Algorithm 5.2 Algorithm for integrating all monomials up to order p over \mathcal{E}

$\partial\mathcal{E} = \{\mathcal{E}_1, \dots, \mathcal{E}_m\}$ where $\mathcal{E}_i \subset \partial\mathcal{E}$;

$F = \text{FaceIntegrals}(d - 1, \mathcal{E}_1, \dots, \mathcal{E}_m, k_1, \dots, k_d)$; cf. Algorithm 5.3

for $a_1 = 0 : k_1, \dots, a_d = 0 : k_d; k_1 + k_2 + \dots + k_d \leq p$ **do**

$$V(a_1, \dots, a_d) = \frac{1}{d + \sum_{n=1}^d a_n} \sum_{i=1}^m b_i F(a_1, \dots, a_d, i);$$

end for

Algorithm 5.3 Algorithm $F = \text{FaceIntegrals}(N, \mathcal{E}_1, \dots, \mathcal{E}_m, k_1, \dots, k_d)$;

 $F(-1 : k_1, \dots, -1 : k_d, 1 : m) = 0;$
for $i=1:m$ **do**

 choose \mathbf{x}_0 as the first vertex of \mathcal{E}_i ;

 $\partial\mathcal{E}_i = \{\mathcal{E}_{i1}, \dots, \mathcal{E}_{im_i}\}$ where $\mathcal{E}_{ij} \subset \partial\mathcal{E}_i$, $j = 1, \dots, m_i$;

if $N - 1 > 0$ **then**
 $E = \text{FaceIntegrals}(N - 1, \mathcal{E}_{i1}, \dots, \mathcal{E}_{im_i}, k_1, \dots, k_d);$
else if $N-1=0$ ($\mathcal{E}_{ij} = (v_1, \dots, v_d) \in \mathbb{R}^d$ is a point) **then**
 $E(a_1, \dots, a_d, j) = v_1^{a_1} \dots v_d^{a_d} \quad \forall 0 \leq a_n \leq k_n, n = 1, \dots, d, j = 1, \dots, m_i;$
end if
for $a_1 = 0 : k_1, \dots, a_d = 0 : k_d; k_1 + k_2 + \dots + k_d \leq p$ **do**

$$F(a_1, \dots, a_d, i) = \frac{1}{N + \sum_{n=1}^d a_n} \left(\sum_{j=1}^{m_i} d_{ij} E(a_1, \dots, a_d, j) + x_{0,1} k_1 F(a_1 - 1, \dots, a_d, i) \right. \\ \left. + \dots + x_{0,d} k_d F(a_1, \dots, a_d - 1, i) \right);$$

end for
end for

5.3.2 Volume and Interface Integrals over Polytopical Mesh Elements

To fix the ideas, we restrict our discussion to the two-dimensional scalar case, but note that the three-dimensional and vector-/tensor-valued cases follow in a completely analogous manner. Let $\{\phi_i\}_{i=1}^{N_h}$ be a basis for the discrete space Q_h^{DG} defined as in (5.2) whose dimension is N_h . For the construction of the discrete space Q_h^{DG} we can exploit, for example, the approach presented in [67], based on employing polynomial spaces defined over the bounding box of each element, cf. Remark 5.1. More precisely, given an element $E \in \mathcal{T}_h$, we first construct the Cartesian bounding box B_E , such that $\bar{E} \subset \bar{B}_E$ and define a linear map between $\mathbf{F}_E : \hat{B} \rightarrow B_E$ such that

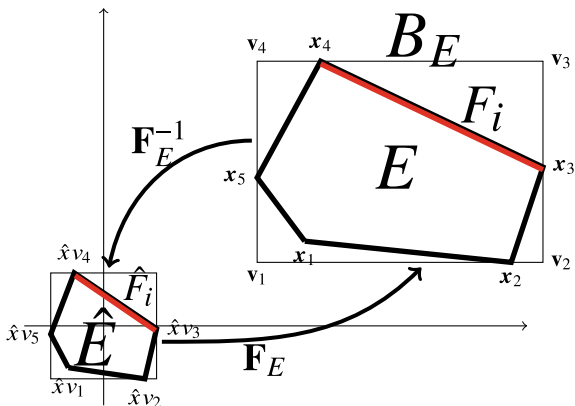
$$\mathbf{F}_E : \hat{\mathbf{x}} \in \hat{B} \mapsto \mathbf{F}_E(\hat{\mathbf{x}}) = \mathbf{J}_E \hat{\mathbf{x}} + \mathbf{t}_E, \quad (5.24)$$

where $\hat{B} = (-1, 1)^d$ and $\mathbf{J}_E \in \mathbb{R}^{d \times d}$ is the (diagonal) Jacobi matrix of the transformation, and $\mathbf{t}_E \in \mathbb{R}^d$ is the translation between the point $\mathbf{0} \in \hat{B}$ and the baricenter of the bounded box B_E , see Fig. 5.5.

We first discuss the application of Algorithm 5.2 for the efficient computation of the local volume integrals over polytopical mesh elements, focusing on the local mass and stiffness volume matrices defined as

$$\mathbf{M}_{i,j}^E = \int_{\Omega} \phi_{i,E} \phi_{j,E}, \quad \mathbf{V}_{i,j}^E = \int_{\Omega} \nabla \phi_{i,E} \cdot \nabla \phi_{j,E}, \quad i, j = 1, \dots, N_{PE}, \quad (5.25)$$

Fig. 5.5 Example of a polygonal element $E \in \mathcal{T}_h$, the relative bounding box B_E , the map \mathbf{F}_E and $\hat{E} = \mathbf{F}_E^{-1}(E)$. Figure taken from [22]



respectively, for all $E \in \mathcal{T}_h$. Here, N_{p_E} is the dimension of the local discrete space, and $\phi_{i,E}$ and $\phi_{j,E}$ are the restriction to E of ϕ_i and ϕ_j , respectively. Employing the transformation \mathbf{F}_E given in (5.24) we have for the mass matrix

$$\mathbf{M}_{i,j}^E = \int_E \phi_{i,E} \phi_{j,E} = \int_{\hat{E}} \hat{\phi}_i \hat{\phi}_j |\mathbf{J}_E|, \quad i, j = 1, \dots, N_{p_E}, \quad (5.26)$$

where $\hat{E} = \mathbf{F}_E^{-1}(E) \subset \hat{B}$, see Fig. 5.5, and the Jacobian of the transformation \mathbf{F}_E is constant and is given by $|\mathbf{J}_E| = (\mathbf{J}_E)_{1,1}(\mathbf{J}_E)_{2,2}$, thanks to the definition of the map (5.24).

In order to employ Algorithm 5.2, we need to identify the coefficients of the homogeneous polynomial expansion for the function $\hat{\phi}_i(\hat{x}, \hat{y})\hat{\phi}_j(\hat{x}, \hat{y})$. We can write, for example, any shape function $\hat{\phi} = \hat{\phi}_i(\hat{x}, \hat{y})$ as the product of one-dimensional Legendre polynomial \mathcal{L}_i , i.e., $\hat{\phi}_i(\hat{x}, \hat{y}) = \mathcal{L}_{i_1}(\hat{x})\mathcal{L}_{i_2}(\hat{y})$, and each Legendre polynomial can be expanded as

$$\mathcal{L}_{i_1}(\hat{x}) = \sum_{m=0}^{i_1} C_{i_1,m} \hat{x}^m, \quad \mathcal{L}_{i_2}(\hat{y}) = \sum_{n=0}^{i_2} C_{i_2,n} \hat{y}^n. \quad (5.27)$$

Therefore, we have

$$\mathbf{M}_{i,j}^E = \int_{\hat{E}} \left(\sum_{m=0}^{i_1} C_{i_1,m} \hat{x}^m \right) \left(\sum_{n=0}^{i_2} C_{i_2,n} \hat{y}^n \right) \left(\sum_{s=0}^{j_1} C_{j_1,s} \hat{x}^s \right) \left(\sum_{r=0}^{j_2} C_{j_2,r} \hat{y}^r \right) |\mathbf{J}_E| \quad (5.28)$$

$$= \int_{\hat{E}} \left(\sum_{k=0}^{i_1+j_1} C_{i_1,j_1,k} \hat{x}^k \right) \left(\sum_{l=0}^{i_2+j_2} C_{i_2,j_2,l} \hat{y}^l \right) |\mathbf{J}_E| \quad (5.29)$$

$$= \sum_{k=0}^{i_1+j_1} \sum_{l=0}^{i_2+j_2} C_{i_1,j_1,k} C_{i_2,j_2,l} |\mathbf{J}_E| \int_{\hat{E}} \hat{x}^k \hat{y}^l, \quad (5.30)$$

where we have defined the compact notation

$$C_{i,j,k} = \sum_{n+m=k} (C_{i,n} C_{j,m}), \quad \text{for } 0 \leq i, j \leq p_E, \quad 0 \leq k \leq i + j, \quad (5.31)$$

and where we stress that the coefficients $C_{i,j,k}$ can be evaluated, once and for all, independently of the polygonal element E .

Concerning the general element of the volume matrix $\mathbf{V}_{i,j}^E$, cf. (5.25), we can proceed as before; indeed, following [22], we obtain

$$\begin{aligned} \mathbf{V}_{i,j}^E &= \sum_{k=0}^{i_1+j_1-2} \sum_{l=0}^{i_2+j_2} C'_{i_1,j_1,k} C_{i_2,j_2,l} (\mathbf{J}_E^{-1})_{1,1}^2 |\mathbf{J}_E| \int_{\hat{E}} \hat{x}^k \hat{y}^l \\ &+ \sum_{k=0}^{i_1+j_1} \sum_{l=0}^{i_2+j_2-2} C_{i_1,j_1,k} C'_{i_2,j_2,l} (\mathbf{J}_E^{-1})_{2,2}^2 |\mathbf{J}_E| \int_{\hat{E}} \hat{x}^k \hat{y}^l, \end{aligned}$$

where $C_{i,j,k}$ is defined in (5.31), and

$$C'_{i,j,k} = \sum_{n+m=k} C'_{i,n} C'_{j,m}, \quad 1 \leq i, j \leq p_E, \quad \text{for } 0 \leq k \leq i + j - 2. \quad (5.32)$$

Here, $C'_{i,n} = (n+1)C_{i,n+1}$, $C'_{j,m} = (m+1)C_{j,m+1}$ are the expansion coefficients of the derivatives of the Legendre polynomials which are again computable independent of the element E , $E \in \mathcal{T}_h$, i.e.,

$$\mathcal{L}'_0(\hat{x}) = 0, \quad \mathcal{L}'_i(\hat{x}) = \sum_{m=0}^{i-1} (m+1)C_{i,m+1} \hat{x}^m = \sum_{m=0}^{i-1} C'_{i,m} \hat{x}^m, \quad \text{for } i > 0. \quad (5.33)$$

We next recall how to compute the key terms that arise in the interface integrals when PolyDG methods are employed for the numerical approximation of second-order partial differential equations. As before, we can transform the integral over a physical face $F \subset \partial E$ to the corresponding integral over the face $\hat{F} = \mathbf{F}_E^{-1}(F) \subset \partial \hat{E}$ on the reference rectangular element \hat{E} . From the definition of the jump and average operators, cf. (5.5), on each face $F \in \mathcal{F}_h^I$ shared by the elements E^+ and E^- we need to assemble contributions of the form

$$\mathbf{S}_{i,j}^{\pm/\mp} = \int_F (\phi_{i,E^\pm} \mathbf{n}^\pm) \cdot (\phi_{j,E^\mp} \mathbf{n}^\mp), \quad i = 1, \dots, N_{p_{E^\pm}}, \quad j = 1, \dots, N_{p_{E^\mp}}, \quad (5.34)$$

$$\mathbf{I}_{i,j}^{\pm/\mp} = \frac{1}{2} \int_F (\nabla \phi_{i,E^\pm} \cdot \mathbf{n}^\pm) \phi_{j,E^\mp}, \quad i = 1, \dots, N_{p_{E^\pm}}, \quad j = 1, \dots, N_{p_{E^\mp}}. \quad (5.35)$$

Analogously, on the boundary face $F \in \mathcal{F}_h^B$ belonging to $E^+ \in \mathcal{T}_h$ we only have to compute

$$\mathbf{S}_{i,j}^{+/+} = \int_F \phi_{i,E^+} \phi_{j,E^+}, \quad \mathbf{I}_{i,j}^{+/+} = \int_F (\nabla \phi_{i,E^+} \cdot \mathbf{n}^+) \phi_{j,E^+}, \quad (5.36)$$

for $i, j = 1, \dots, N_{p_{E^+}}$. We next recall how to efficiently compute terms of the form

$$\begin{aligned} \mathbf{S}_{i,j}^{+/+} &= \int_F (\phi_{i,E^+} \mathbf{n}^+) \cdot (\phi_{j,E^+} \mathbf{n}^+) = \int_F \phi_{i,E^+} \phi_{j,E^+}, \\ \mathbf{S}_{i,j}^{+/-} &= \int_F (\phi_{i,E^+} \mathbf{n}^+) \cdot (\phi_{j,E^-} \mathbf{n}^-) = - \int_F \phi_{i,E^+} \phi_{j,E^-}, \end{aligned} \quad (5.37)$$

and refer to [22] for further details and discussion on the efficient computation of the terms $\mathbf{I}_{i,j}^{\pm/\mp}$. Reasoning as before, we obtain

$$\begin{aligned} \mathbf{S}_{i,j}^{+/+} &= \sum_{k=0}^{i_1+j_1} \sum_{l=0}^{i_2+j_2} C_{i_1,j_1,k} C_{i_2,j_2,l} \mathcal{J}_{F^+} \int_{\hat{F}^+} \hat{\mathbf{x}}^k \hat{\mathbf{y}}^l, \\ \mathbf{S}_{i,j}^{+/-} &= - \sum_{k=0}^{i_1+j_1} \sum_{l=0}^{i_2+j_2} \tilde{X}_{i_1,j_1,k} \tilde{Y}_{i_2,j_2,l} \mathcal{J}_{F^+} \int_{\hat{F}^+} \hat{\mathbf{x}}^k \hat{\mathbf{y}}^l, \end{aligned} \quad (5.38)$$

where \mathcal{J}_{F^+} is defined as $\mathcal{J}_{F^+} = \|\mathbf{J}_{E^+}^{-\top} \hat{\mathbf{n}}_{\hat{F}^+}\| |\mathbf{J}_{E^+}|$ and $\hat{\mathbf{n}}_{\hat{F}^+}$ is the unit outward normal vector to \hat{F}^+ . In (5.38), the coefficients $C_{i,j,k}$ are defined as in (5.31), whereas \tilde{X} and \tilde{Y} are defined as

$$\left. \begin{aligned} \tilde{X}_{i,j,k} &= \sum_{n+m=k} (C_{i,n} \tilde{X}_{j,m}) \\ \tilde{Y}_{i,j,k} &= \sum_{n+m=k} (C_{i,n} \tilde{Y}_{j,m}) \end{aligned} \right\} \text{for } 0 \leq i \leq p_{E^+}, 0 \leq j \leq p_{E^-}, 0 \leq k \leq i+j. \quad (5.39)$$

Here, as before, $C_{i,n}$ are the coefficients of the homogeneous function expansion of the Legendre polynomials in $(-1, 1)$, while $\tilde{X}_{j,m}$ and $\tilde{Y}_{j,m}$ are defined by

$$\left. \begin{aligned} \tilde{X}_{j,m} &= \sum_{r=m}^j C_{j,r} \binom{r}{m} (\tilde{\mathbf{J}}_{1,1})^m (\tilde{\mathbf{t}}_1)^{r-m} \\ \tilde{Y}_{j,m} &= \sum_{r=m}^j C_{j,r} \binom{r}{m} (\tilde{\mathbf{J}}_{2,2})^m (\tilde{\mathbf{t}}_2)^{r-m} \end{aligned} \right\} \text{for } 0 \leq m \leq p_{E^-}, m \leq j \leq p_{E^-}. \quad (5.40)$$

Finally, in the definition above $\tilde{\mathbf{t}}_1$ and $\tilde{\mathbf{t}}_2$ are the two components of the vector $\tilde{\mathbf{t}}$ of the composite map $\tilde{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{F}_{E^-}^{-1}(\mathbf{F}_{E^+}(\hat{\mathbf{x}}))$, cf. Figure 5.6, defined as

$$\tilde{\mathbf{F}}(\hat{\mathbf{x}}) = \mathbf{J}_{E^-}^{-1}(\mathbf{J}_{E^+} \hat{\mathbf{x}} + \mathbf{t}_{E^+}) - \mathbf{J}_{E^-}^{-1} \mathbf{t}_{E^-} = \underbrace{\mathbf{J}_{E^-}^{-1} \mathbf{J}_{E^+}}_{\tilde{\mathbf{j}}} \hat{\mathbf{x}} + \underbrace{\mathbf{J}_{E^-}^{-1}(\mathbf{t}_{E^+} - \mathbf{t}_{E^-})}_{\tilde{\mathbf{t}}}. \quad (5.41)$$

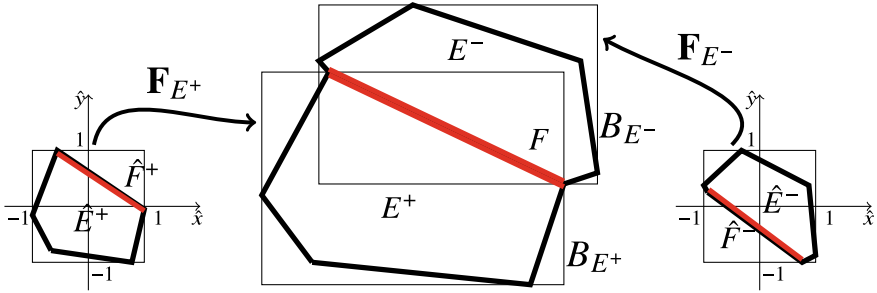


Fig. 5.6 Example of polygonal elements $E^\pm \in \mathcal{T}_h$, together with the bounded boxes B_{E^\pm} , and the local maps $\mathbf{F}_{E^\pm} : \hat{E} \rightarrow E^\pm$ for the common face $F \subset E^\pm$. Figure taken from [22]

We conclude this section by observing that for the computation of the local forcing term

$$\int_E f(\mathbf{x}) \phi_{i,E}(\mathbf{x}) d\mathbf{x}, \quad i = 1, \dots, N_{p_E}, \tag{5.42}$$

the *quadrature free* method allows to exactly evaluate (5.42) when f is a polynomial function. If f is a general function, an explicit polynomial approximation of f is required.

5.3.3 Numerical Results

The aim of this section is to present some numerical computations to assess the practical performance of the *quadrature free* algorithm.

5.3.3.1 Example 1: Integration of Bivariate Polynomials over a Given Polygon

We first present some numerical results in order to test the performance of the method proposed in Algorithm 5.1 for the integration of bivariate polynomials over a given polygon $\mathcal{P} \subset \mathbb{R}^2$ based on employing the recursive algorithm described in Sect. 5.3.1, i.e., $\int_{\mathcal{P}} x^k y^l = I(2, \mathcal{P}, k, l)$.

For the sake of comparison, we also present the analogous computations carried out based on employing the sub-tessellation technique. In this case, the domain of integration \mathcal{P} is firstly decomposed into triangles; then on each sub-triangle we employ a tensor product Gauss quadrature rule consisting of \mathcal{N}^2 nodes and weights, which is defined based on application of the Duffy transformation. In order to guarantee the exact integration of $x^k y^l$, we select $\mathcal{N} = \lceil \frac{k+l}{2} \rceil + 1$. In order to compare both

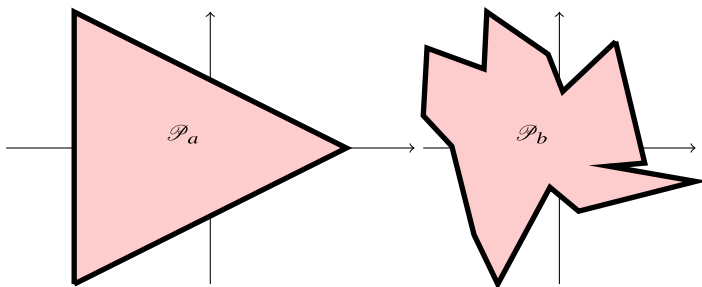


Fig. 5.7 Example 1 (Sect. 5.3.3.1). Domains of integration \mathcal{P} : triangle (\mathcal{P}_a , left) and an irregular polygon with 15 faces (\mathcal{P}_b , right)

Table 5.2 Example 1 (Sect. 5.3.3.1). Vertex coordinates of polygons \mathcal{P}_a and \mathcal{P}_b of Fig. 5.7

	vertex	(x, y)-coordinates
\mathcal{P}_a	1	(-1.0000000000000000, -1.0000000000000000)
	2	(1.0000000000000000, 0.0000000000000000)
	3	(-1.0000000000000000, 1.0000000000000000)
\mathcal{P}_b	1	(0.413058522141662, 0.781696234443715)
	2	(0.024879797655533, 0.415324992429711)
	3	(-0.082799691823524, 0.688810136531751)
	4	(-0.533191422779328, 1.0000000000000000)
	5	(-0.553573605852999, 0.580958514816226)
	6	(-0.972432940212767, 0.734117068746903)
	7	(-1.0000000000000000, 0.238078507228890)
	8	(-0.789986179147920, 0.012425068086110)
	9	(-0.627452906935866, -0.636532897516109)
	10	(-0.452662174765764, -1.0000000000000000)
	11	(-0.069106265580153, -0.289054989277619)
	12	(0.141448047807069, -0.464417038155806)
	13	(1.0000000000000000, -0.245698820584615)
	14	(0.363704451489016, -0.134079689960635)
	15	(0.627086024018283, -0.110940423607648)

approaches, we integrate bivariate polynomials of different degrees on the triangle and the irregular polygon depicted in Fig. 5.7; see Table 5.2 for the list of vertex coordinates for both domains. In Figs. 5.8 and 5.9 we compare the CPU time (in seconds) taken to evaluate the underlying integral (on \mathcal{P}_a and \mathcal{P}_b , respectively) up to machine precision, using the *quadrature free* algorithm and the sub-tessellation method. We remark that the times for the *quadrature free* algorithm include the computation of b_i , \mathbf{n}_i , and d_{ij} , $j = 1, \dots, m_i$, $i = 1, \dots, m$. The times for sub-tessellation method include the one-time computation of the \mathcal{N}^2 nodes and weights on the reference triangle, the time required for sub-tessellation, as well as the time needed for numer-

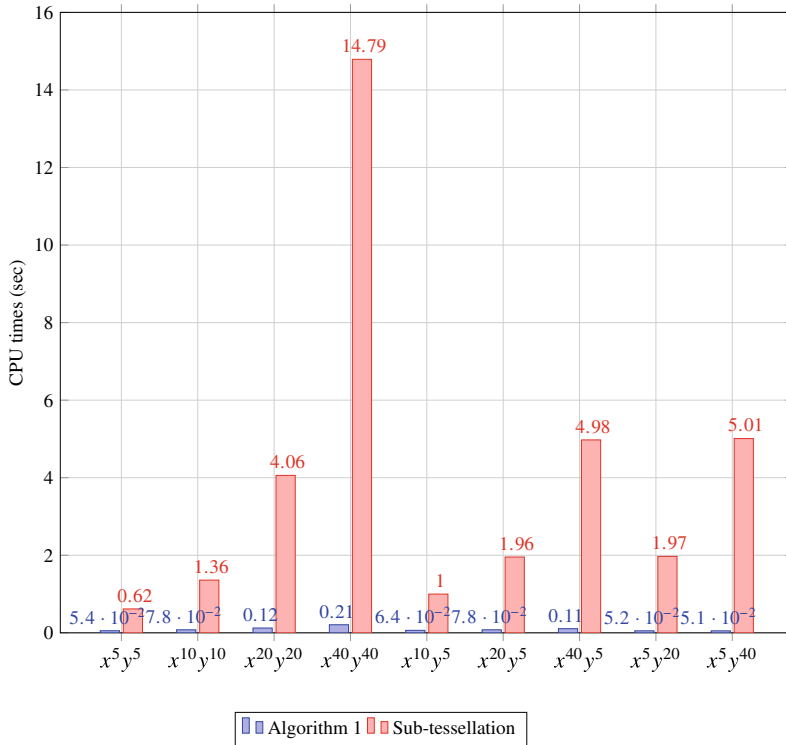


Fig. 5.8 Example 1 (Sect. 5.3.3.1). CPU times as a function of the integrand. Integration domain \mathcal{P}_a of Fig. 5.7

ical integration on each sub-triangle. From the results reported in Figs. 5.8 and 5.9 it is clear that the *quadrature free* algorithm outperforms sub-tessellation; indeed, for both domains of integration, we observe an improvement in the CPU-time required to evaluate the underlying integral of between one- to two-orders of magnitude when the former approach is employed. Moreover, even when the integration domain consists of a triangle (\mathcal{P}_a), the *quadrature free* algorithm still outperforms classical quadrature rules, even though in this case no sub-tessellation is undertaken.

5.3.3.2 Example 2: Computation of the PolyDG Stiffness and Mass Matrices in Three-Dimensions

We now compare the performance of the *quadrature free* algorithm and the sub-tessellation method when employed to assemble the stiffness and mass matrices for the PolyDG approximation of a second-order elliptic diffusion-reaction problem in three-dimensions. Here, the polyhedral grids have been obtained by agglomeration starting from a partition Ω consisting of hexahedral elements. The agglomeration is performed based on employing the *METIS* library for graph partitioning, cf., for

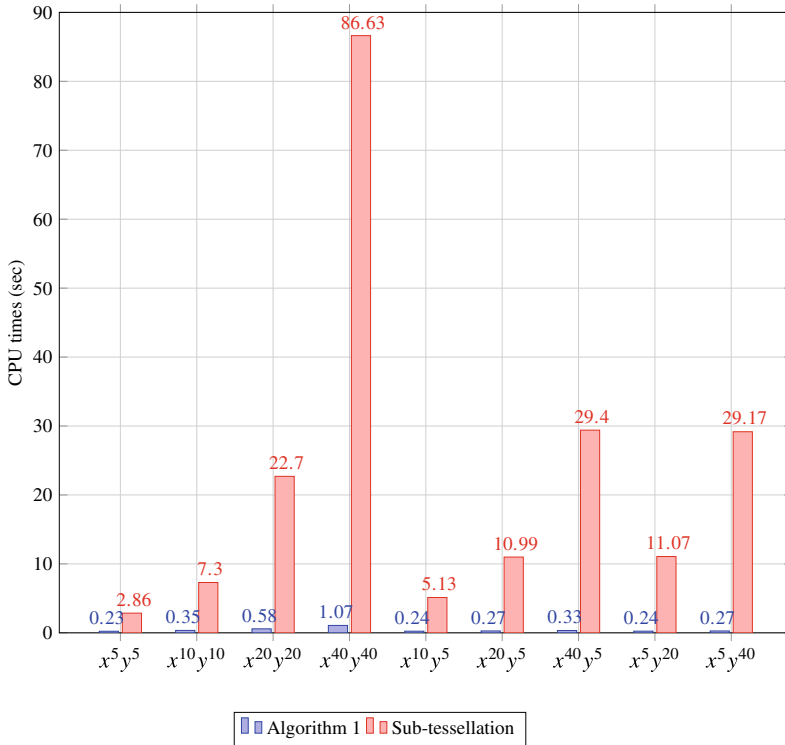


Fig. 5.9 Example 1 (Sect. 5.3.3.1). CPU times as a function of the integrand. Integration domain \mathcal{P}_b of Fig. 5.7

example, [108, 109] so that each polyhedral element is typically non-convex. In Fig. 5.10 we show three typical examples of polyhedral elements generated from the agglomeration process. We now compare the CPU time required by the *quadrature free* method with the quadrature integration/sub-tessellation approach to assemble the volume and mass matrices, denoted by \mathbf{V} and \mathbf{M} , respectively, as well for the computation of the interface matrices \mathbf{S} and \mathbf{I} ; cf. Sect. 5.3.2. We point out that, to assemble the volume and mass matrices based on employing the sub-tessellation algorithm, we exploit the fact that the polyhedral mesh is obtained by agglomeration of hexahedral elements, so that the sub-tessellation into hexahedra of each polyhedral mesh element is already given. In Fig. 5.11 (left) we report the CPU time needed for the computation of the volume matrices \mathbf{V} and \mathbf{M} , for a set of agglomerated polyhedral grids where we fix the polynomial approximation degree $p \in \{1, 2, 3, 4, 5\}$ and we vary the number of elements $N_e \in \{5, 40, 320, 2560, 20480\}$; in all cases the agglomerated elements are formed from approximately 10 (fine) hexahedral elements. The analogous results obtained based on computing the interface matrices \mathbf{S} and \mathbf{I} (right) are shown in Fig. 5.11 (right); furthermore, these timings are compared on a log-log plot in Fig. 5.12. From the computations shown in Figs. 5.11 and 5.12, we clearly observe that the quadrature free method substantially outperforms the



Fig. 5.10 Example 2 (Sect. 5.3.3.2). Typical agglomerated element shapes

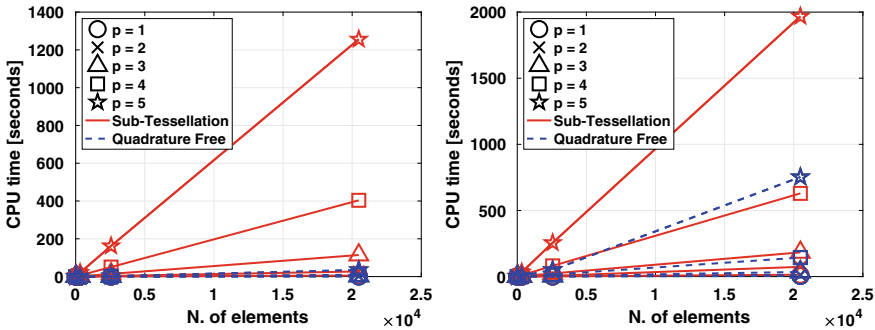


Fig. 5.11 Example 2 (Sect. 5.3.3.2). Comparison of the CPU time needed to assemble the volume matrices \mathbf{M} and \mathbf{V} (left) and the interface matrices \mathbf{S} and \mathbf{I} (right) for a three-dimensional problem by using the proposed quadrature free method and the classical sub-tessellation method. Each line is obtained by fixing the polynomial approximation degree $p \in \{1, 2, 3, 4, 5\}$ and measuring the CPU time by varying the number of elements (N_e) of the underlying mesh

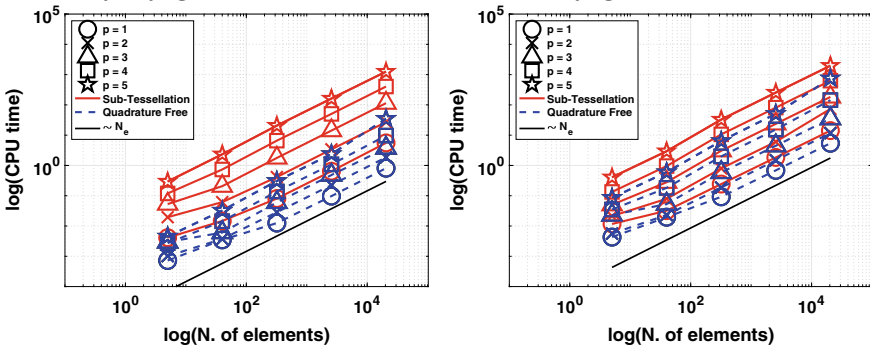


Fig. 5.12 Example 2 (Sect. 5.3.3.2). Comparison of the CPU time needed to assemble the volume matrices \mathbf{M} and \mathbf{V} (left) and the interface matrices \mathbf{S} and \mathbf{I} (right) for a three-dimensional problem by using the proposed quadrature free method and the classical sub-tessellation method. Each line is obtained by fixing the polynomial approximation degree $p \in \{1, 2, 3, 4, 5\}$ and measuring the CPU time by varying the number of elements (N_e) of the underlying mesh

sub-tessellation quadrature approach, both for the computation of the volume and the face integrals. We refer to [22] for additional numerical computations, where the issue of computational complexity is also addressed.

5.4 PolyDG Methods for Seismic Wave Propagation

In this section we present an overview of high-order PolyDG methods for the approximate solution of wave propagation problems modeled by the elastodynamics equations on computational meshes consisting of polytopic elements. In particular, we discuss the model problem, analyze the well-posedness of the semidiscrete formulation and derive an *hp*-version *a priori* error bound. The theoretical estimates are then validated through two-dimensional numerical computations carried out on both benchmark, as well as real test cases. The dispersion analysis, in two-dimensions, is not reported here, for the sake of brevity, and can be found in [26], where it has been shown that polygonal meshes behave similarly to classical simplicial/quadrilateral grids in terms of dispersion errors. For the sake of brevity, we focus here on the elastodynamics equation; more sophisticated model problems can be successfully treated as well, for example, see [10, 11, 28], respectively, for elasto-acoustic coupling and non-linear sound waves phenomena.

5.4.1 Model Problem and Its PolyDG Semidiscretization

We consider an elastic body occupying an open, bounded polyhedral domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, and denote by \mathbf{n} the outward normal unit vector to its boundary. The boundary $\partial\Omega$ is assumed to be composed of two disjoint portions $\Gamma_D \neq \emptyset$ and Γ_N , i.e., $\Gamma_D \cap \Gamma_N = \emptyset$. For a final observation time $T > 0$, let $\mathbf{u} : \Omega \times [0, T] \rightarrow \mathbb{R}^d$ be the displacement vector. The equations of the initial/boundary-value problem of (linear) elastodynamics are given by

$$\left\{ \begin{array}{ll} \rho \ddot{\mathbf{u}} - \nabla \cdot \boldsymbol{\sigma} = \mathbf{f}, & \text{in } \Omega \times (0, T], \\ \mathbf{u} = \mathbf{0}, & \text{on } \Gamma_D \times (0, T], \\ \boldsymbol{\sigma} \mathbf{n} = \mathbf{0}, & \text{on } \Gamma_N \times (0, T], \\ \mathbf{u} = \mathbf{u}_0, & \text{in } \Omega \times \{0\}, \\ \dot{\mathbf{u}} = \mathbf{u}_1, & \text{in } \Omega \times \{0\}. \end{array} \right. \quad (5.43)$$

Here, $\mathbf{f} \in L^2((0, T]; L^2(\Omega))$ is the (given) external load and $\mathbf{u}_0 \in \mathbf{H}_{0, \Gamma_D}^1(\Omega)$ and $\mathbf{u}_1 \in L^2(\Omega)$ are (given) initial data, where $\mathbf{H}_{0, \Gamma_D}^1(\Omega)$ denotes the space of vector-valued functions in $\mathbf{H}^1(\Omega)$ whose trace vanishes on Γ_D . Finally, $\rho \in L^\infty(\Omega)$ is the medium density. As constitutive law for the stress tensor $\boldsymbol{\sigma} : \Omega \times [0, T] \rightarrow \mathbb{S}$, \mathbb{S} being the space of $d \times d$ symmetric real-valued matrices, $d = 2, 3$, we assume the generalized Hooke's law, i.e., $\boldsymbol{\sigma}(\mathbf{u}) = \mathcal{D}\boldsymbol{\varepsilon}(\mathbf{u})$, where the fourth order stiffness tensor $\mathcal{D} : \mathbb{S} \rightarrow \mathbb{S}$ is defined as $\mathcal{D}\boldsymbol{\tau} = 2\mu\boldsymbol{\tau} + \lambda\text{tr}(\boldsymbol{\tau})\mathbf{I}$ for any $\boldsymbol{\tau} \in \mathbb{S}$, and $\boldsymbol{\varepsilon}(\mathbf{u})$ is the symmetric gradient of \mathbf{u} , i.e., $\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}(\nabla\mathbf{u} + \nabla\mathbf{u}^\top)$. Here, \mathbf{I} is the identity tensor, $\text{tr}(\cdot)$ represents the trace operator, and $\lambda, \mu \in L^\infty(\Omega)$ are the first and the second Lamé parameters, respectively. We assume that \mathcal{D} is symmetric, positive definite and uniformly bounded over Ω . We recall that the compressional (P) and shear

(S) wave velocities can be obtained through the relations $c_P = \sqrt{(\lambda + 2\mu)/\rho}$ and $c_S = \sqrt{\mu/\rho}$, respectively. The weak formulation of problem (5.43) reads as follows: for all $t \in (0, T]$ find $\mathbf{u} = \mathbf{u}(t) \in \mathbf{H}_{0,\Gamma_D}^1(\Omega)$ such that:

$$\begin{cases} \int_{\Omega} \rho \ddot{\mathbf{u}} \cdot \mathbf{v} + \int_{\Omega} \mathcal{D}\boldsymbol{\varepsilon}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathbf{H}_{0,\Gamma_D}^1(\Omega), \\ \mathbf{u}(\cdot, 0) = \mathbf{u}_0, \quad \dot{\mathbf{u}}(\cdot, 0) = \mathbf{u}_1. \end{cases} \quad (5.44)$$

Problem (5.44) is well posed and its unique solution $\mathbf{u} \in C((0, T]; \mathbf{H}_{0,\Gamma_D}^1(\Omega)) \cap C^1((0, T]; \mathbf{L}^2(\Omega))$, see [130, Theorem 8–3.1].

Based on employing the notation of Sect. 5.2, we introduce the PolyDG semidiscretization of problem (5.44): for all $t \in (0, T]$, find $\mathbf{u}_h = \mathbf{u}_h(t) \in \mathbf{W}_h^{DG}$ such that

$$\int_{\Omega} \rho \ddot{\mathbf{u}}_h \cdot \mathbf{v} + \mathcal{B}(\mathbf{u}_h, \mathbf{v}) = \int_{\Omega} \mathbf{f} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathbf{W}_h^{DG}, \quad (5.45)$$

supplemented with the initial conditions $\mathbf{u}_h(0) = \mathbf{u}_h^0$ and $\dot{\mathbf{u}}_h(0) = \mathbf{u}_h^1$, where $\mathbf{u}_h^0, \mathbf{u}_h^1 \in \mathbf{W}_h^{DG}$ are suitable approximations of \mathbf{u}_0 and \mathbf{u}_1 , respectively. Here, we also assume that \mathcal{D} and ρ are element-wise constant over the mesh \mathcal{T}_h . The bilinear form $\mathcal{B}(\cdot, \cdot) : \mathbf{W}_h^{DG} \times \mathbf{W}_h^{DG} \rightarrow \mathbb{R}$ is defined as

$$\begin{aligned} \mathcal{B}(\mathbf{u}, \mathbf{v}) &= \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \boldsymbol{\varepsilon}(\mathbf{v}) + \int_{\Omega} \boldsymbol{\sigma}(\mathbf{u}) : \mathcal{R}(\llbracket \llbracket \mathbf{v} \rrbracket \rrbracket) + \int_{\Omega} \mathcal{R}(\llbracket \llbracket \mathbf{u} \rrbracket \rrbracket) : \boldsymbol{\sigma}(\mathbf{v}) + \\ &\int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \eta \llbracket \llbracket \mathbf{u} \rrbracket \rrbracket : \llbracket \llbracket \mathbf{v} \rrbracket \rrbracket \end{aligned} \quad (5.46)$$

for all $\mathbf{u}, \mathbf{v} \in \mathbf{W}_h^{DG}$. Here, $\mathcal{R}(\cdot) : \mathbf{L}^1(\mathcal{F}_h^I \cup \mathcal{F}_h^D) \rightarrow \mathcal{W}_h^{DG}$ is the *lifting operator* of the traces of $d \times d$ symmetric tensors defined as

$$\int_{\Omega} \mathcal{R}(\llbracket \llbracket \mathbf{w} \rrbracket \rrbracket) : \boldsymbol{\sigma}(\mathbf{v}) = - \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \llbracket \llbracket \mathbf{w} \rrbracket \rrbracket : \{\boldsymbol{\sigma}(\mathbf{v})\} \quad \forall \mathbf{v} \in \mathbf{W}_h^{DG}. \quad (5.47)$$

The penalization function $\eta : \mathcal{F}_h \rightarrow \mathbb{R}_+$ in (5.46) is defined face-wise as

$$\eta = \sigma_0 \overline{\mathcal{D}}_E \begin{cases} \max_{E \in \{E_1, E_2\}} \left(\frac{\rho_E^2}{h_E} \right), & F \in \mathcal{F}_h^I, F \subset \partial E_1 \cap \partial E_2, \\ \frac{\rho_E^2}{h_E}, & F \in \mathcal{F}_h^D, F \subset \partial E \cap \Gamma_D. \end{cases} \quad (5.48)$$

where $\overline{\mathcal{D}}_E = |(\mathcal{D}|_E)^{1/2}|_2^2$ for any $E \in \mathcal{T}_h$ (here $|\cdot|_2$ is the operator norm induced by the l_2 -norm on \mathbb{R}^n , where n denotes the dimension of the space of symmetric second-order tensors, i.e., $n = 3$ if $d = 2$, $n = 6$ if $d = 3$), and σ_0 is a positive parameter at our disposal.

5.4.2 Well-Posedness, Stability and Error Analysis of the Semidiscrete Formulation

In this section we prove stability and error estimates for the PolyDG semidiscretization defined in (5.45). To this end, we define the space $\tilde{\mathbf{W}}_h^{DG} = \mathbf{W}_h^{DG} \oplus \mathbf{H}_{0,\Gamma_D}^1(\Omega)$ endowed with the following DG norm

$$\|\mathbf{v}\|_{DG}^2 = \left\| \mathcal{D}^{\frac{1}{2}} \boldsymbol{\varepsilon}(\mathbf{v}) \right\|_{L^2(\Omega)}^2 + \left\| \eta^{\frac{1}{2}} [[[\mathbf{v}]]] \right\|_{L^2(\mathcal{F}_h^I \cup \mathcal{F}_h^D)}^2 \quad \forall \mathbf{v} \in \tilde{\mathbf{W}}_h^{DG}; \quad (5.49)$$

here, we have used the compact notation $\|\cdot\|_{L^2(\mathcal{F}_h^I \cup \mathcal{F}_h^D)}^2 = \sum_{F \in \mathcal{F}_h^I \cup \mathcal{F}_h^D} \|\cdot\|_{L^2(F)}^2$. From the preliminary results of Sect. 5.2 we immediately have the following estimates; we refer to [26] for more details.

Lemma 5.3 *Assume that \mathcal{T}_h satisfies Assumption 5.1. Then, for any $\mathbf{w} \in \mathbf{W}_h^{DG}$ we have that*

$$\|\eta^{-1/2}\{\mathbf{w}\}\|_{L^2(\mathcal{F}_h^I \cup \mathcal{F}_h^D)}^2 \lesssim \frac{1}{\sigma_0} \|\mathbf{w}\|_{L^2(\Omega)}, \quad (5.50)$$

where the hidden constant is independent of p_E , $|E|$, and \mathbf{w} , and where σ_0 is the constant appearing in the definition of the penalty function, cf. (5.48).

Lemma 5.4 *Assume that \mathcal{T}_h satisfies Assumption 5.1. For any $\mathbf{v} \in \tilde{\mathbf{W}}_h^{DG}$ we have that*

$$\|\mathcal{R}([[[\mathbf{v}]]])\|_{L^2(\Omega)}^2 \lesssim \frac{1}{\sigma_0} \|\eta^{\frac{1}{2}} [[[\mathbf{v}]]] \|_{L^2(\mathcal{F}_h^I \cup \mathcal{F}_h^D)}^2,$$

where σ_0 is the constant appearing in the definition of the penalty function, cf. (5.48).

Proof The proof follows by observing that if $\mathbf{v} \in \mathbf{H}_{0,\Gamma_D}^1(\Omega)$, then $[[[\mathbf{v}]]] = \mathbf{0}$ and the estimate is trivial. If $\mathbf{v} \in \mathbf{W}_h^{DG}$, by using the definition of the lifting operator (5.47) together with Lemma 5.3 the result follows immediately.

Based on employing the above results and standard DG arguments, the well-posedness of the PolyDG formulation (5.45) can be established.

Lemma 5.5 *Assume that \mathcal{T}_h satisfies Assumption 5.1, and that the constant σ_0 appearing in the definition (5.48) of the penalization function is chosen sufficiently large. Then,*

$$\mathcal{B}(\mathbf{v}, \mathbf{v}) \gtrsim \|\mathbf{v}\|_{DG}^2, \quad \mathcal{B}(\mathbf{v}, \mathbf{w}) \lesssim \|\mathbf{v}\|_{DG} \|\mathbf{w}\|_{DG} \quad \forall \mathbf{v}, \mathbf{w} \in \tilde{\mathbf{W}}_h^{DG}.$$

We next provide a stability result of the semidiscrete PolyDG formulation (5.45) in the following energy norm

$$\|\mathbf{u}_h(t)\|_E^2 = \|\rho^{\frac{1}{2}} \dot{\mathbf{u}}_h(t)\|_{L^2(\Omega)}^2 + \|\mathbf{u}_h(t)\|_{DG}^2 \quad \forall t \in (0, T]. \quad (5.51)$$

Proposition 5.2 *Let $\mathbf{f} \in L^2((0, T]; L^2(\Omega))$ and $\mathbf{u}_h \in C^2((0, T]; \mathbf{W}_h^{DG})$ be the approximate solution of (5.45) obtained with the stability constant σ_0 defined in (5.48) chosen sufficiently large. Then,*

$$\|\mathbf{u}_h(t)\|_E \lesssim \|\mathbf{u}_h^0\|_E + \int_0^t \|\mathbf{f}(\tau)\|_{L^2(\Omega)}, \quad 0 < t \leq T. \quad (5.52)$$

Proof Selecting $\mathbf{v} = \dot{\mathbf{u}}_h \in \mathbf{W}_h^{DG}$ in (5.45), integrating in time between 0 and t , employing Lemma 5.3 together with the arithmetic-geometric inequality, and choosing σ_0 large enough, we get

$$\|\mathbf{u}_h\|_E^2 + 2 \int_{\Omega} \mathcal{R}([\![\mathbf{u}_h]\!]]) : \boldsymbol{\sigma}(\mathbf{u}_h) \gtrsim \|\mathbf{u}_h\|_E^2.$$

Moreover, from Lemma 5.4 it also follows that

$$2 \left| \int_{\Omega} \mathcal{R}([\![\mathbf{u}_h^0]\!]]) : \boldsymbol{\sigma}(\mathbf{u}_h^0) \right| \lesssim \frac{1}{\sqrt{\sigma_0}} \|\eta^{\frac{1}{2}} [[[\mathbf{u}_h^0]]] \|_{L^2(\mathcal{F}_h^I \cup \mathcal{F}_h^D)} \|\boldsymbol{\sigma}(\mathbf{u}_h^0)\|_{L^2(\Omega)} \lesssim \frac{1}{\sqrt{\sigma_0}} \|\mathbf{u}_h^0\|_E^2.$$

Therefore, substituting these inequalities, and applying the Cauchy-Schwarz inequality yields

$$\|\mathbf{u}_h\|_E^2 \lesssim \|\mathbf{u}_h^0\|_E^2 + 2 \int_0^t \|\mathbf{u}_h\|_E \|\mathbf{f}\|_{L^2(\Omega)}.$$

The statement of the proposition now follows by employing Gronwall's lemma [129].

Before providing hp -version error bounds, we observe that formulation (5.45) is not strongly-consistent, due to the presence of the lifting operator. It is easy to see that the error $\mathbf{u} - \mathbf{u}_h$ satisfies the following error equation

$$\int_{\Omega} \rho (\ddot{\mathbf{u}} - \ddot{\mathbf{u}}_h) \cdot \mathbf{v}_h + \mathcal{B}(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) + \mathcal{R}_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{W}_h^{DG}, \quad (5.53)$$

where the residual $\mathcal{R}_h(\cdot, \cdot) : \widetilde{\mathbf{W}}_h^{DG} \times \mathbf{W}_h^{DG} \rightarrow \mathbb{R}$ is defined by

$$\mathcal{R}_h(\mathbf{w}, \mathbf{v}_h) = - \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \{\boldsymbol{\sigma}(\mathbf{w})\} : [[[\mathbf{v}_h]]] - \int_{\Omega} \boldsymbol{\sigma}(\mathbf{w}) : \mathcal{R}([\![\mathbf{v}_h]\!]]) ,$$

for all $\mathbf{w} \in \widetilde{\mathbf{W}}_h^{DG}$ and for all $\mathbf{v}_h \in \mathbf{W}_h^{DG}$, and where we have used also that $\mathcal{R}_h(\mathbf{w}_h, \mathbf{v}_h) = 0$ whenever $\mathbf{w}_h \in \mathbf{W}_h^{DG}$, cf. (5.47).

In order to derive *a priori* error bounds for the semidiscrete scheme, we assume that Assumption 5.2 is satisfied; we define, component-wise, the extension opera-

tors $\mathcal{E} : \mathcal{H}^s(\Omega) \rightarrow \mathcal{H}^s(\mathbb{R}^{d \times d})$, $s \in \mathbb{N}_0$, as in Sect. 5.2.4, cf. also [136]; we employ the tensorial and vectorial counterpart of the approximation estimates outlined in Sect. 5.2.4, cf. also [26, 63], to obtain the following bound

$$\|\mathbf{u} - \mathbf{\Pi u}\|_E^2 \lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{2^{(r_E-3/2)} p_E} \left(\|\mathcal{E}\mathbf{u}\|_{\mathbf{H}^{r_E}(T_E)}^2 + \frac{h_E^2}{p_E^3} \|\mathcal{E}\dot{\mathbf{u}}\|_{\mathbf{H}^{r_E}(T_E)}^2 \right), \quad (5.54)$$

where $s_E = \min(p_E + 1, r_E)$. The hidden constant depends on the material parameters and on the shape-regularity of T_E , but is independent of q , h_E , p_E and the number of faces per element. Moreover, the global interpolant $\mathbf{\Pi}$ is defined elementwise as $\mathbf{\Pi u}|_E = \mathbf{\Pi}_E^{p_E} \mathbf{u}$ for any $E \in \mathcal{T}_h$, where $\mathbf{\Pi}_E^{p_E}$ is vector-valued counterpart of the interpolant defined in Lemma 5.2.

The last ingredient we need is the following bound on the residual; we refer to [26] for the proof.

Lemma 5.6 *For any $\mathbf{w} \in \tilde{\mathbf{W}}_h^{DG}$ and $\mathbf{v}_h \in \mathbf{W}_h^{DG}$, the following bound holds*

$$|\mathcal{R}_h(\mathbf{w}, \mathbf{v}_h)| \lesssim \left(\sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{2^{(r_E-3/2)} p_E} \|\mathcal{E}\sigma(\mathbf{w})\|_{\mathcal{H}^{r_E}(T_E)}^2 \right)^{1/2} \|\mathbf{v}_h\|_{DG}, \quad (5.55)$$

where $s_E = \min(p_E + 1, r_E)$ for all $E \in \mathcal{T}_h$. The hidden constant depends on the material parameters and the shape-regularity of T_E , but is independent of q , h_E , p_E , and the number of element faces.

We can now state the main result of this section.

Theorem 5.1 *Let Assumptions 5.1 and 5.2 be satisfied. Moreover, assume that the analytical solution \mathbf{u} of (5.43) is sufficiently regular. For any time $t \in [0, T]$, let $\mathbf{u}_h \in \mathbf{W}_h^{DG}$ be the PolyDG solution of problem (5.45) obtained with a penalty parameter σ_0 appearing in (5.48) sufficiently large. Then, for any time $t \in (0, T]$ the following bound holds*

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h\|_E^2 &\lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{2^{(r_E-3/2)} p_E} \left(\|\mathcal{E}\mathbf{u}\|_{\mathbf{H}^{r_E}(T_E)}^2 + \frac{h_E^2}{p_E^3} \|\mathcal{E}\dot{\mathbf{u}}\|_{\mathbf{H}^{r_E}(T_E)}^2 + \|\mathcal{E}\sigma(\mathbf{u})\|_{\mathcal{H}^{r_E}(T_E)}^2 \right) \\ &\quad + \frac{h_E^{2(s_E-1)}}{2^{(r_E-3/2)} p_E} \int_0^t \left(\|\mathcal{E}\dot{\mathbf{u}}\|_{\mathbf{H}^{r_E}(T_E)}^2 + \frac{h_E^2}{p_E^3} \|\mathcal{E}\ddot{\mathbf{u}}\|_{\mathbf{H}^{r_E}(T_E)}^2 + \|\mathcal{E}\sigma(\dot{\mathbf{u}})\|_{\mathcal{H}^{r_E}(T_E)}^2 \right), \end{aligned} \quad (5.56)$$

with $s_E = \min(p_E + 1, m_k)$ for all $E \in \mathcal{T}_h$. The hidden constants depends on the material parameters and the shape-regularity of T_E , but is independent of q , h_E , p_E and the number of element faces.

Proof We recall the main steps of the proof and refer to [26] for more details. Let $\mathbf{\Pi}$ be defined as (5.54). We write the error as $\mathbf{u} - \mathbf{u}_h = \mathbf{e}_h - \mathbf{e}_I$ with $\mathbf{e}_h = \mathbf{u}_h - \mathbf{\Pi u}$ and $\mathbf{e}_I = \mathbf{u} - \mathbf{\Pi u}$, and rewrite the error equation (5.53) for $\mathbf{v}_h = \dot{\mathbf{e}}_h$, to obtain

$$\int_{\Omega} \rho \ddot{\mathbf{e}}_h \cdot \dot{\mathbf{e}}_h + \mathcal{B}(\mathbf{e}_h, \dot{\mathbf{e}}_h) = \int_{\Omega} \rho \ddot{\mathbf{e}}_I \cdot \dot{\mathbf{e}}_h + \mathcal{B}(\mathbf{e}_I, \dot{\mathbf{e}}_h) + \mathcal{R}_h(\mathbf{e}_I, \dot{\mathbf{e}}_h),$$

where we have also used that $\mathcal{R}_h(\mathbf{e}_h, \dot{\mathbf{e}}_h) = 0$ since $\mathbf{e}_h, \dot{\mathbf{e}}_h \in \mathbf{W}_h^{DG}$. Using the definition of the energy norm (5.51), integrating in time between 0 and t , and exploiting that $\mathbf{e}_h(0) = \mathbf{0}$, and reasoning as in the proof of Proposition 5.2 yields

$$\|\mathbf{e}_h\|_{\mathbb{E}}^2 + 2 \int_{\Omega} \mathcal{R}([\![\mathbf{e}_h]\!]]) : \boldsymbol{\sigma}(\mathbf{e}_h) \gtrsim \|\mathbf{e}_h\|_{\mathbb{E}}^2,$$

provided the penalty parameter is chosen sufficiently large. Therefore, we get

$$\begin{aligned} \|\mathbf{e}_h\|_{\mathbb{E}}^2 &\lesssim \int_0^t \int_{\Omega} \rho \ddot{\mathbf{e}}_I \cdot \dot{\mathbf{e}}_h + \int_0^t \mathcal{B}(\mathbf{e}_I, \dot{\mathbf{e}}_h) + \int_0^t \mathcal{R}_h(\mathbf{e}_I, \dot{\mathbf{e}}_h) \\ &= \int_0^t \int_{\Omega} \rho \ddot{\mathbf{e}}_I \cdot \dot{\mathbf{e}}_h + \mathcal{B}(\mathbf{e}_I, \mathbf{e}_h) - \int_0^t \mathcal{B}(\dot{\mathbf{e}}_I, \mathbf{e}_h) - \mathcal{R}_h(\mathbf{e}_I, \mathbf{e}_h) + \int_0^t \mathcal{R}_h(\dot{\mathbf{e}}_I, \mathbf{e}_h), \end{aligned}$$

where in the second step we have used integration by parts for the second and third term on the right hand side together with $\mathbf{e}_h(0) = \mathbf{0}$. Employing Jensen and Cauchy-Schwarz inequalities for first term on the right hand side, the fact that $\mathcal{R}_h(\mathbf{e}_I, \mathbf{e}_h) = \mathcal{R}_h(\mathbf{u}, \mathbf{e}_h)$, Lemma 5.5, the definition of the energy norm (5.51), and Lemma 5.6, we get

$$\|\mathbf{e}_h\|_{\mathbb{E}}^2 \lesssim \|\mathbf{e}_I\|_{\mathbb{E}} \|\mathbf{e}_h\|_{\mathbb{E}} + \int_0^t \|\dot{\mathbf{e}}_I\|_{\mathbb{E}} \|\mathbf{e}_h\|_{\mathbb{E}} + \mathcal{I}(\mathbf{u}) \|\mathbf{e}_h\|_{\mathbb{E}} + \int_0^t \mathcal{I}(\dot{\mathbf{u}}) \|\mathbf{e}_h\|_{\mathbb{E}},$$

where

$$\mathcal{I}(\mathbf{u}) = \left(\sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{p_E^{2(m_E-3/2)}} \|\boldsymbol{\mathcal{E}}\boldsymbol{\sigma}(\mathbf{u})\|_{\mathcal{H}^{m_E}(T_E)}^2 \right)^{1/2},$$

cf. Lemma 5.6. Applying the arithmetic-geometric inequality with $\delta > 0$ we have

$$(1 - \delta) \|\mathbf{e}_h\|_{\mathbb{E}}^2 \lesssim \frac{1}{\delta} (\|\mathbf{e}_I\|_{\mathbb{E}}^2 + \mathcal{I}^2(\mathbf{u})) + \int_0^t (\|\dot{\mathbf{e}}_I\|_{\mathbb{E}} + \mathcal{I}(\dot{\mathbf{u}})) \|\mathbf{e}_h\|_{\mathbb{E}}.$$

Choosing δ small enough and applying Gronwall's lemma [129] we get

$$\|\mathbf{e}_h\|_{\mathbb{E}}^2 \lesssim \|\mathbf{e}_I\|_{\mathbb{E}}^2 + \mathcal{I}^2(\mathbf{u}) + \int_0^t (\|\dot{\mathbf{e}}_I\|_{\mathbb{E}}^2 + \mathcal{I}^2(\dot{\mathbf{u}})).$$

The proof is completed by employing (5.54) and the definition of $\mathcal{I}(\mathbf{u})$. \square

5.4.3 Numerical Results

Before presenting some numerical experiments, we first discuss the algebraic formulation of the semidiscrete formulation and the time integration of the corresponding system of second-order ordinary differential equations. We suppose that Ω is partitioned into N_{el} disjoint polytopic elements E_r , $r = 1, \dots, N_{el}$, and denote by $n_{pE} = \dim(\mathbb{P}_{pE})$, and set $N_{dof} = \sum_{r=1}^{N_{el}} n_{pE}$ to be the dimension of each component of a function in \mathbf{W}_h^{DG} . We introduce a basis $\{\Phi_i^1, \dots, \Phi_i^d\}_{i=1}^{N_{dof}}$, $d = 2, 3$, for the finite element space \mathbf{W}_h^{DG} . By expressing $\mathbf{u}_h \in \mathbf{W}_h^{DG}$ as a linear combination of the basis functions, i.e.,

$$\mathbf{u}_h(\mathbf{x}, t) = \sum_{s=1}^d \sum_{j=1}^{N_{dof}} \Phi_j^s(\mathbf{x}) U_j^s(t),$$

and writing Eq. (5.45) for any test function $\Phi_i^s(\mathbf{x}) \in \mathbf{W}_h^{DG}$, $s = 1, \dots, d$, we obtain the following system of second order differential equations

$$M\ddot{\mathbf{U}}(t) + B\mathbf{U}(t) = \mathbf{F}(t) \quad \forall t \in (0, T), \quad (5.57)$$

for the displacements $\mathbf{U}(t) = (\mathbf{U}^1(t), \dots, \mathbf{U}^d(t))^T$. Here, $\mathbf{F} = (\mathbf{F}^1(t), \dots, \mathbf{F}^d(t))^T$ represents the external applied load, M and B are the (symmetric and positive definite) mass and stiffness matrices, respectively. To integrate the system (5.57) in time we consider the explicit, second-order accurate, and conditionally stable leap-frog scheme: we subdivide the interval $(0, T]$ into N_T equal subintervals of size $\Delta t = T/N_T$ and at every time level $t_n = n\Delta t$ we solve the system

$$M\mathbf{U}(t_{n+1}) = [2M - \Delta t^2 B]\mathbf{U}(t_n) - M\mathbf{U}(t_{n-1}) + \Delta t^2 \mathbf{F}(t_n), \quad \text{for } n = 1, \dots, N_T, \quad (5.58)$$

with

$$M\mathbf{U}(t_1) = [M - \frac{\Delta t^2}{2} B]\mathbf{U}(t_0) - \Delta t M\dot{\mathbf{U}}(t_0) + \frac{\Delta t^2}{2} \mathbf{F}(t_0), \quad (5.59)$$

supplemented with the initial conditions. We recall that to ensure stability, the explicit time integration leap-frog scheme must satisfy the usual Courant–Friedrichs–Levy (CFL) condition that imposes a restriction on Δt of the form

$$\Delta t \leq C_{\text{CFL}}(c_P, \sigma_0) \frac{h}{p^2},$$

where h is the maximum mesh size and p is the polynomial approximation degree (supposed to be uniform here, for the sake of simplicity). The constant C_{CFL} depends

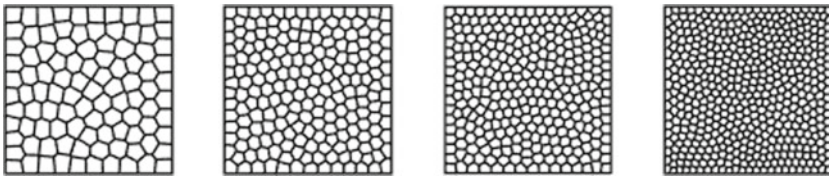


Fig. 5.13 Example 1 (Sect. 5.4.3.1). Mesh configurations considered with increasing number of polygonal elements: $N_{el} = 100, 200, 300, 500$

on the compressional wave velocity $c_P = \sqrt{(\lambda + 2\mu)/\rho}$ and on the stability parameter σ_0 , cf. (5.48), and can be estimated as in [26], cf., also, [29].

5.4.3.1 Example 1: Smooth Problem with a Known Analytical Solution

We first consider the wave propagation problem in $\Omega = (0, 1)^2$, where $\Gamma_N = (0, 1) \times \{1\}$, $\Gamma_D = \partial\Omega \setminus \Gamma_N$, $\lambda = \mu = \rho = 1$ and boundary conditions, initial conditions and the forcing term \mathbf{f} are selected so that the analytical solution of (5.43) is given by

$$\mathbf{u}(\mathbf{x}, t) = \cos(\sqrt{2}\pi t) \begin{bmatrix} -\sin(\pi x) \sin(\pi y)^2 \\ \cos(\pi x) \sin(\pi y)^2 \end{bmatrix}. \quad (5.60)$$

For the proceeding computations we set the final time $T = 0.6$ and time step $\Delta t = 10^{-5}$. Firstly, we consider the convergence of the PolyDG method with p -refinement. To this end, in Fig. 5.14 (left) we plot $\|\mathbf{u}(T) - \mathbf{u}_h(T)\|_E$ versus the polynomial degree $p_E = p$, for all $E \in \mathcal{T}_h$, on a fixed polyhedral mesh \mathcal{T}_h consisting of 300 elements; cf. Figure 5.13. Here, on a semi-logarithmic scale, we observe that the convergence line is approximately straight, thereby indicating exponential convergence of the PolyDG method as p is uniformly enriched. Secondly, we consider the h -convergence of the PolyDG approximation computed on the sequence of meshes depicted in Fig. 5.13. In Fig. 5.14 (right), we observe that $\|\mathbf{u}(T) - \mathbf{u}_h(T)\|_E$ behaves like $\mathcal{O}(h^p)$ as h tends to zero, for each fixed p ; this is in agreement with the *a priori* error bound stated in Theorem 5.1.

5.4.3.2 Example 2: Elastic Wave Propagation in a Heterogeneous Medium

For an application of the presented PolyDG method, we study the elastic wave propagation in the computational domain $\Omega = (0, 38.4) \text{ km} \times (0, 10) \text{ km}$ representing an idealized bidimensional Earth's cross section, see Fig. 5.15. The bottom and the lateral boundaries are set far enough from the point source (white dot in Fig. 5.15) in order to prevent any reflections from the boundaries of the waves of interest. At the top of the model a free-surface boundary condition is imposed, i.e., $\sigma \mathbf{n} = \mathbf{0}$, whereas

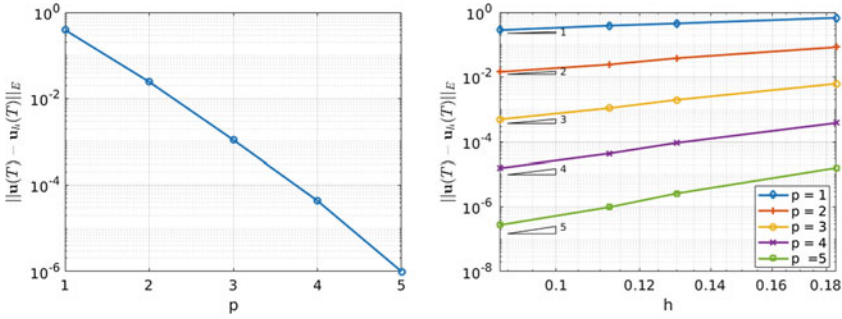


Fig. 5.14 Example 1 (Sect. 5.4.3.1). Computed error $\|\mathbf{u}(T) - \mathbf{u}_h(T)\|_E$ versus the polynomial degree p , fixing $N_{el} = 300$ (left) and versus the mesh size $h = 1/N_{el}$, $N_{el} = 100, 200, 300, 500$ (right) fixing $p = 2, 3, 4, 5$. Results are obtained choosing as observation time $T = 0.6$ with $\Delta t = 10^{-5}$

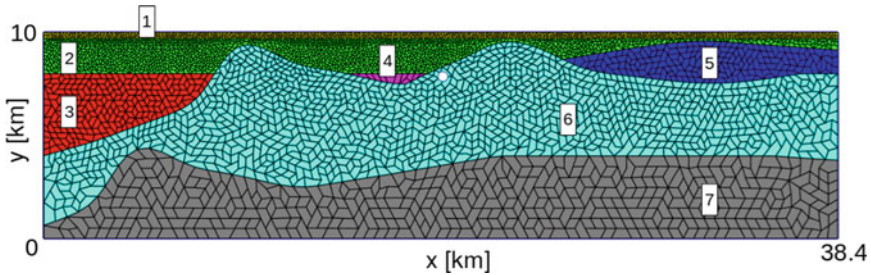


Fig. 5.15 Example 2 (Sect. 5.4.3.2). Unstructured polygonal grid. The mesh spacing varies from $h_E \approx 160$ m for material 1 to $h_E \approx 1500$ m for material 7; cf. Table 5.3. The source location $\mathbf{x}_s = (19.4, 7.8)$ km is indicated by a *white circle*

homogeneous Dirichlet conditions are set in the remaining part of the boundary. We simulate a point source load of the form

$$\mathbf{f}(\mathbf{x}, t) = \left(0, A e^{-10^{-4} \|\mathbf{x} - \mathbf{x}_s\|^2} (1 - 2\pi^2 f_0^2 (t - t_0)^2) e^{-\pi^2 f_0^2 (t - t_0)^2}\right),$$

with $A = 10^3$ N, $f_0 = 2$ Hz and $t_0 = 2$ s applied at the point $\mathbf{x}_s = (19.4, 7.8)$ km. We assign constant material properties within each region as described in Table 5.3. The computational domain is discretized using an unstructured grid consisting of 4870 (agglomerated) polygonal elements, with a mesh size varying from $h_E \approx 160$ m for material 1 to $h_E \approx 1500$ m for material 7; cf. Table 5.3. The grid spacing is chosen small enough not only to describe with sufficient precision the physical profile of the submerged topography, but also to guarantee that over the whole domain there is at least 5 points per wavelength with polynomial degree equal to 4 to keep numerical dispersion and dissipation errors sufficiently small, i.e., of order of machine precision, see [26]. In Fig. 5.16 we report a set of snapshots of the displacement magnitude

Table 5.3 Example 2 (Sect. 5.4.3.2). Material properties used for the computational domain in Fig. 5.15

Material	ρ [kg/m^3]	c_p [m/s]	c_s [m/s]
1	1800	1321	294
2	1800	2024	450
3	2050	1920	600
4	2050	1920	650
5	2050	2000	650
6	2400	3030	1515
7	2450	3200	1600

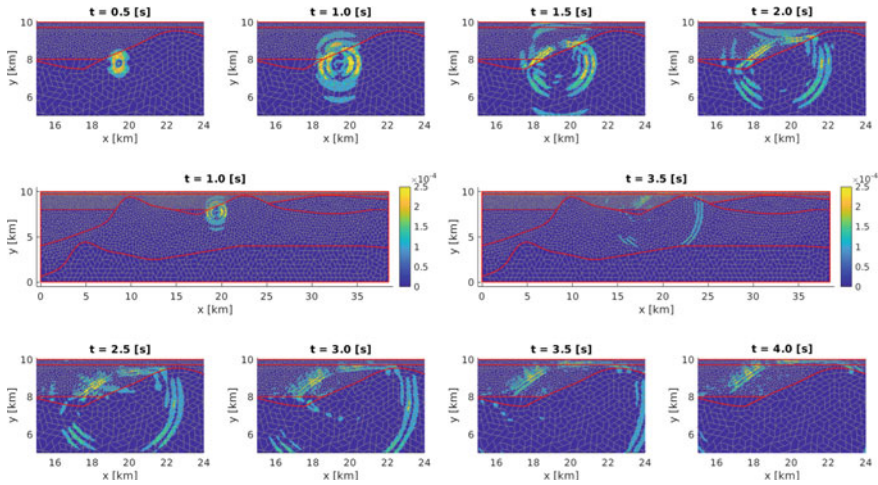


Fig. 5.16 Example 2 (Sect. 5.4.3.2). Snapshots of the computed displacement magnitude $|\mathbf{u}| = \sqrt{\mathbf{u}_1^2 + \mathbf{u}_2^2}$ at different time $t = 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4$ s. Due to the material heterogeneities, high oscillations and perturbations of the wave front can be observed. Waves moving leftwards with respect to the point source location are clearly visible. The displacement magnitude is measured in meters

$|\mathbf{u}| = \sqrt{\mathbf{u}_1^2 + \mathbf{u}_2^2}$ computed with the proposed method (with $\sigma_0 = 10$ and polynomial degree equal to 4) coupled with the leap-frog scheme, fixing the final observation time $T = 5$ s and time step $\Delta t = 10^{-4}$ s. The discontinuities between the mechanical properties of the materials produce oscillations and perturbations on the wave front. In particular, due to the stratigraphy of the model, the energy is focussed towards the left of the domain, reaches the surface of the model and (most of it) remains trapped within the first layer. All these complex and relevant phenomena are well captured by the proposed PolyDG method, see Fig. 5.16.

5.5 PolyDG Methods for Flow in Fractured Porous Media

The aim of this section is to present an overview of the results presented in [16], where a unified formulation and analysis of PolyDG approximations of flows in fractured porous media is provided for all primal-primal, primal-mixed, mixed-primal and mixed-mixed formulations. More precisely, a primal-primal setting consists of having the pressure as only unknown for both the bulk and fracture problems. When dealing with the approximation of Darcy's flow, one may also resort to a mixed-mixed approach, where the flow is described through an additional unknown representing the (averaged) velocity of the fluid in both the bulk and the fracture. This variable, often referred to as Darcy's velocity, is of primary interest in many engineering applications [57, 119], so that the mixed setting is often preferred to the primal one, which may only return the velocity after post-processing the computed pressure, thus entailing a potential loss of accuracy. On the other hand, the primal-primal approach is easier to solve, featuring a smaller number of degrees of freedom. For this reason, our aim is to design a unified setting where, according to the desired approximation properties of the model, one may resort to either a primal or mixed approximation for the problem in the bulk, as well as to a primal or mixed approximation for the problem in the fracture. In particular, for the primal discretizations we employ the Symmetric Interior Penalty discontinuous Galerkin method [32, 148], whereas for the mixed discretizations we employ the local DG (LDG) method of [79], both in their generalization to polytopic grids [13, 63, 64, 66, 67]. Our main reference for the design of such a setting is the work by Arnold et al. [33], where a *unified* analysis of all DG methods present in the literature is undertaken. This framework is based on the flux-formulation, where the so-called numerical fluxes are introduced on elemental interfaces as approximations of the analytical solution. Different choices of the numerical fluxes affect the stability and the accuracy of the underlying PolyDG method and provide conservation properties of desired quantities such as, for example, mass, momentum, and energy [66]. In the particular context of flow in fractured porous media, we also show that the coupling conditions between bulk and fracture problems may be imposed through a suitable definition of the numerical fluxes on the fracture faces. Such an abstract setting allows us to analyze theoretically, in a unified manner, all the possible combinations of primal-primal (PP), mixed-primal (MP), primal-mixed (PM) and mixed-mixed (MM) formulations for the bulk and fracture problems, respectively.

The rest of the section is organized as follows. In Sect. 5.5.1 we introduce the model problem; the discretization based on employing PolyDG methods is presented, in the unified setting of [33], in Sect. 5.5.2. In Sect. 5.5.3, we recall the main theoretical results, namely well-posedness and stability, and present *a priori* error bounds. Illustrative numerical tests are presented in Sect. 5.5.4 to confirm the theoretical bounds. Moreover, we assess the capability of the method in handling more complicated geometries, presenting some test cases featuring networks of partially immersed fractures.

5.5.1 Model Problem

To describe the flow, which we assume to be single-phase flow, we adopt the mathematical model of [118]. This model was first introduced in [2, 3] for fractures with large permeability and is here generalised to handle also the low permeable case. An extension to two-phase flows can be found in [99, 107]. To keep the presentation as simple as possible, we assume that the porous medium is cut by a single, non immersed fracture. We refer to [4] for the extension of the model to totally immersed fractures. Finally, in order to handle networks of intersecting fractures, some physical conditions need to be added to describe the behavior of the flow at the intersection points/lines. A possible choice is to impose pressure continuity and balance of fluxes as in [56, 96]. Other, more general conditions, where the angle between fractures is taken into account and jumps of the pressure across the intersection are allowed, may be found, for example, in [95, 133].

In the following we assume that the porous matrix is represented by the open, bounded, and polygonal/polyhedral domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ and the fracture is described by the $(d - 1)$ -dimensional C^∞ manifold (with no curvature) $\Gamma \subset \mathbb{R}^{d-1}$, $d = 2, 3$. Since we are assuming that Γ is not immersed, it separates Ω into the two connected disjoint subdomains Ω_1 and Ω_2 . We decompose the boundary of Ω into two disjoint subsets $\partial\Omega_D$ and $\partial\Omega_N$, *i.e.*, $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, with $\partial\Omega_D \cap \partial\Omega_N = \emptyset$, and we define $\partial\Omega_{D,i} = \partial\Omega_D \cap \partial\Omega_i$ and $\partial\Omega_{N,i} = \partial\Omega_N \cap \partial\Omega_i$, for $i = 1, 2$. For the fracture domain we set $\partial\Gamma = \Gamma \cap \partial\Omega$ with $\partial\Gamma = \partial\Gamma_D \cup \partial\Gamma_N$. Finally, we denote by \mathbf{n}_Γ the normal unit vector on Γ with a fixed orientation from Ω_1 to Ω_2 . Our model considers Darcy's flow in its mixed form for the problem both in the bulk and the fracture. More precisely, in addition to the Darcy's pressure, we take into account an auxiliary vector-valued variable, called Darcy's velocity. This quantity is of primary interest in many engineering applications, such as oil recovery and groundwater pollution modeling. Indeed, in these cases, in order to be effective, the simulation of the phenomenon requires very accurate approximation of the velocities of the involved fluids. The coupled bulk-fracture model problem in mixed form is given by:

$$\mathbf{u}_i = \mathbf{v}_i \nabla p_i \quad \text{in } \Omega_i, \quad (5.61a)$$

$$-\nabla \cdot \mathbf{u}_i = f_i \quad \text{in } \Omega_i, \quad (5.61b)$$

$$p_i = 0 \quad \text{on } \partial\Omega_{D,i}, \quad (5.61c)$$

$$\mathbf{u}_i \cdot \mathbf{n}_i = 0 \quad \text{on } \partial\Omega_{N,i} \quad (5.61d)$$

$$\mathbf{u}_\Gamma = \mathbf{v}_\Gamma^\tau \ell_\Gamma \nabla_\tau p_\Gamma \quad \text{in } \Gamma, \quad (5.61e)$$

$$-\nabla_\tau \cdot \mathbf{u}_\Gamma = \ell_\Gamma f_\Gamma - \llbracket \mathbf{u} \rrbracket \quad \text{in } \Gamma, \quad (5.61f)$$

$$p_\Gamma = 0 \quad \text{on } \partial\Gamma_D, \quad (5.61g)$$

$$\mathbf{u}_\Gamma \cdot \boldsymbol{\tau} = 0 \quad \text{on } \partial\Gamma_N, \quad (5.61h)$$

$$-\{\mathbf{u}\} \cdot \mathbf{n}_\Gamma = \beta_\Gamma \llbracket p \rrbracket \cdot \mathbf{n}_\Gamma \quad \text{on } \Gamma, \quad (5.61i)$$

$$-\llbracket \mathbf{u} \rrbracket = \alpha_\Gamma (\{p\} - p_\Gamma) \quad \text{on } \Gamma. \quad (5.61j)$$

In the bulk, in each domain Ω_i , $i = 1, 2$, the motion of an incompressible fluid with pressure p_i and velocity \mathbf{u}_i is described by (5.61a)–(5.61b), supplemented by the boundary conditions (5.61c)–(5.61d). Moreover, $f_i \in L^2(\Omega_i)$ represents a source term, and $\mathbf{v}_i = \mathbf{v}_i(x) \in \mathbb{R}^{d \times d}$ is the bulk permeability tensor, which we assume to be symmetric, positive definite, uniformly bounded from below and above and with entries that are bounded, piecewise continuous real-valued functions. Denoting by p_Γ and \mathbf{u}_Γ the fracture pressure and velocity, respectively, on the manifold Γ representing the fracture, we formulate a reduced version of Darcy’s law in the tangential direction, cf. equations (5.61e)–(5.61f), and assume that the fracture permeability tensor \mathbf{v}_Γ , has a block-diagonal structure when written in its normal and tangential components and that $\mathbf{v}_\Gamma^\tau \in \mathbb{R}^{(d-1) \times (d-1)}$ is positive definite and uniformly bounded. Moreover, \mathbf{v}_Γ satisfies the same regularity assumptions as those satisfied by the bulk permeability \mathbf{v} . In (5.61e)–(5.61f)–(5.61g)–(5.61h), $f_\Gamma \in L^2(\Gamma)$, $\boldsymbol{\tau}$ is the vector in the tangent plane of Γ normal to $\partial\Gamma$ and ∇_τ and $\nabla_\tau \cdot$ denote the tangential gradient and divergence operators, respectively. Finally, we close the model providing the interface conditions (5.61i)–(5.61j) where $\beta_\Gamma = \frac{1}{2\eta_\Gamma}$, $\alpha_\Gamma = \frac{2}{\eta_\Gamma(2\xi-1)}$ and $\eta_\Gamma = \frac{\ell_\Gamma}{\mathbf{v}_\Gamma^\tau}$, $\ell_\Gamma > 0$ being the fracture thickness. Finally, in the definition of α_Γ , the closure parameter $\xi > 1/2$ is related to the pressure profile across the fracture aperture. We refer to [118] for a rigorous derivation of the model. An analogous model has been used in Chap. 8, where it has been solved numerically with a mixed virtual element method. Other type of schemes for fracture porous media may be found in Chaps. 3 and 4 of this volume.

To introduce the weak formulation, we first introduce the bulk pressure and velocity spaces:

$$M^b = L^2(\Omega), \quad \mathbf{V}^b = \{\mathbf{v} \in H_{div}(\Omega) : [[\mathbf{v}]]|_\Gamma \in L^2(\Gamma), \{\mathbf{v}\}|_\Gamma \in [L^2(\Gamma)]^d, \mathbf{v} \cdot \mathbf{n}|_{\partial\Omega_N} = 0\}.$$

Similarly, for the fracture pressure and velocity we define the spaces

$$M^\Gamma = L^2(\Gamma), \quad \mathbf{V}^\Gamma = \{\mathbf{v}_\Gamma \in H_{div,\tau}(\Gamma) : \mathbf{v}_\Gamma \cdot \boldsymbol{\tau}|_{\partial\Gamma} = 0\}.$$

We equip the spaces \mathbf{V}^b and \mathbf{V}^Γ with the norms

$$\begin{aligned} \|\mathbf{v}\|_{\mathbf{V}^b}^2 &= \|\mathbf{v}\|_{L^2(\Omega)}^2 + \|\nabla \cdot \mathbf{v}\|_{L^2(\Omega)}^2 + \|[[\mathbf{v}]]\|_{L^2(\Gamma)}^2 + \|\{\mathbf{v}\}\|_{L^2(\Gamma)}^2, \\ \|\mathbf{v}_\Gamma\|_{\mathbf{V}^\Gamma}^2 &= \|\mathbf{v}_\Gamma\|_{L^2(\Gamma)}^2 + \|\nabla_\tau \cdot \mathbf{v}_\Gamma\|_{L^2(\Gamma)}^2, \end{aligned}$$

respectively. Finally, we define the global spaces for the pressure and the velocity as $M = M^b \times M^\Gamma$ and $\mathbf{W} = \mathbf{V}^b \times \mathbf{V}^\Gamma$, respectively, equipped with the canonical norms for product spaces. We can now formulate problem (5.61) in weak form as follows: find $(\mathbf{u}, \mathbf{u}_\Gamma) \in \mathbf{W}$ and $(p, p_\Gamma) \in M$ such that

$$\begin{aligned} A((\mathbf{u}, \mathbf{u}_\Gamma), (\mathbf{v}, \mathbf{v}_\Gamma)) + B((\mathbf{v}, \mathbf{v}_\Gamma), (p, p_\Gamma)) &= \mathbf{0}, \\ -B((\mathbf{u}, \mathbf{u}_\Gamma), (q, q_\Gamma)) &= F^p(q, q_\Gamma) \end{aligned} \tag{5.62}$$

for all $(\mathbf{v}, \mathbf{v}_\Gamma) \in \mathbf{W}$ and $(q, q_\Gamma) \in M$, where the bilinear forms $A(\cdot, \cdot) : \mathbf{W} \times \mathbf{W} \rightarrow \mathbb{R}$ and $B(\cdot, \cdot) : \mathbf{W} \times M \rightarrow \mathbb{R}$ are defined as

$$\begin{aligned} A((\mathbf{u}, \mathbf{u}_\Gamma), (\mathbf{v}, \mathbf{v}_\Gamma)) &= a(\mathbf{u}, \mathbf{v}) + a_\Gamma(\mathbf{u}_\Gamma, \mathbf{v}_\Gamma), \\ B((\mathbf{v}, \mathbf{v}_\Gamma), (q, q_\Gamma)) &= b(\mathbf{v}, q) + b_\Gamma(\mathbf{v}_\Gamma, q_\Gamma) + d(\mathbf{v}, q_\Gamma), \end{aligned}$$

respectively, with

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &= \int_{\Omega} \mathbf{v}^{-1} \mathbf{u} \cdot \mathbf{v} + \int_{\Gamma} \frac{1}{\alpha_\Gamma} \llbracket \mathbf{u} \rrbracket \llbracket \mathbf{v} \rrbracket + \int_{\Gamma} \frac{1}{\beta_\Gamma} \{\mathbf{u}\} \cdot \{\mathbf{v}\}, \\ a_\Gamma(\mathbf{u}_\Gamma, \mathbf{v}_\Gamma) &= \int_{\Gamma} (\mathbf{v}_\Gamma^t \ell_\Gamma)^{-1} \mathbf{u}_\Gamma \cdot \mathbf{v}_\Gamma, \end{aligned}$$

and

$$b(\mathbf{v}, q) = \int_{\Omega} \nabla \cdot \mathbf{v} q, \quad b_\Gamma(\mathbf{v}_\Gamma, q_\Gamma) = \int_{\Gamma} \nabla_\tau \cdot \mathbf{v}_\Gamma q_\Gamma, \quad d(\mathbf{v}, q_\Gamma) = - \int_{\Gamma} \llbracket \mathbf{v} \rrbracket q_\Gamma. \quad (5.63)$$

Finally the linear operator $F^p(\cdot) : M \rightarrow \mathbb{R}$ is defined as $F^p(q, q_\Gamma) = \int_{\Omega} f q + \int_{\Gamma} \ell_\Gamma f_\Gamma q_\Gamma$.

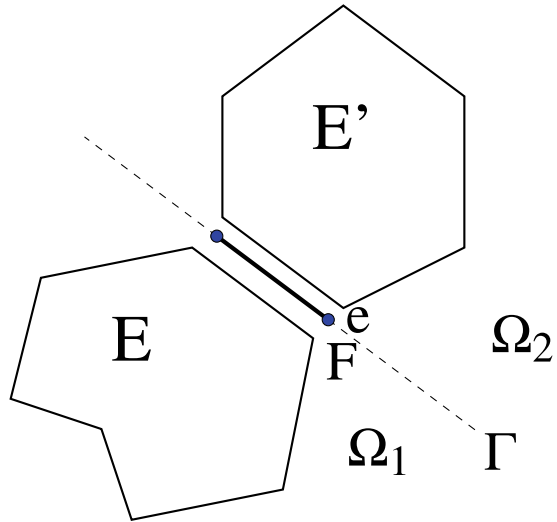
We next recall the following well-posedness result: we refer to [16] for the proof. Note that the existence and uniqueness of the problem can be proven only under the condition that the parameter $\xi > 1/2$.

Theorem 5.2 *Suppose that $\xi > 1/2$, then problem (5.62) admits a unique solution.*

5.5.2 PolyDG Discretization of Flow in Fractured Porous Media: A Unified Approach

In this section we present, in a unified setting, a family of discrete formulations for the coupled bulk-fracture problem (5.62). In particular, the problem in the bulk and the one in the fracture can be either discretized in their mixed or primal form. We then derive four formulations that embrace all the possible combinations of primal-primal, mixed-primal, primal-mixed and mixed-mixed discretizations. The primal discretizations will be based on the Symmetric Interior Penalty DG method (SIPDG) [32, 148], while the mixed approach will exploit the Local Discontinuous Galerkin method (LDG) [69, 79, 128], including their extension to polytopic grids [13, 63, 64, 66, 67]. The derivation follows the approach of [33] based on the introduction of the *numerical fluxes*, which approximate the trace of the solutions on the boundary of each mesh element. In particular, the imposition of the coupling conditions (5.61i)–(5.61j) will be achieved through a proper definition of the numerical fluxes on the faces belonging to the fracture.

Fig. 5.17 Example of two neighboring elements of a polygonal bulk mesh aligned with the fracture and of the induced subdivision



We consider a sequence of meshes \mathcal{T}_h that are aligned with the fracture Γ and we denote, as in Sect. 5.2, by \mathcal{F}_h the set of all the faces of the mesh \mathcal{T}_h , that we can decompose as $\mathcal{F}_h = \mathcal{F}_h^I \cup \mathcal{F}_h^B \cup \Gamma_h$, where now \mathcal{F}_h^I is the set of interior faces not belonging to the fracture, \mathcal{F}_h^B is the set of faces lying on the boundary of the domain $\partial\Omega$ (which can be further decomposed into $\mathcal{F}_h^B = \mathcal{F}_h^D \cup \mathcal{F}_h^N$) and Γ_h is the set of fracture faces. In particular, the induced subdivision of the fracture Γ_h consists of the faces of the elements of \mathcal{T}_h that share part of their boundary with the fracture, so that, according to the definition of \mathcal{F}_h given in Sect. 5.2.1, Γ_h is made up of line segments when $d = 2$ and of triangles when $d = 3$. In the latter case, the triangles are not necessarily shape-regular and they may present hanging nodes, due to the fact that the sub-triangulations of each elemental interface is chosen independently from the others. For this reason, we here extend the concept of *interface* introduced in Sect. 5.2.1 also to the $(d - 2)$ -dimensional facets of elements in Γ_h , defined again as intersection of boundaries of two neighbouring elements. When $d = 2$, the interfaces reduce to points (see Fig. 5.17), while when $d = 3$ they consists of line segments. Moreover, since we aim at employing PolyDG methods also for the discretization of the problem in the fracture, we denote by $\mathcal{E}_{\Gamma,h}$ the set of all the interfaces (that we will also call edges) of the elements in Γ_h , and we write, accordingly to the previous notation, $\mathcal{E}_{\Gamma,h} = \mathcal{E}_{\Gamma,h}^I \cup \mathcal{E}_{\Gamma,h}^B$, with $\mathcal{E}_{\Gamma,h}^B = \mathcal{E}_{\Gamma,h}^D \cup \mathcal{E}_{\Gamma,h}^N$.

For the forthcoming stability and error analysis, we assume that both the bulk and fracture sequence of meshes are *polytopic-regular*, according to Assumption 5.1 and that the covering satisfies Assumption 5.2. Moreover, we suppose that the permeability tensors \mathbf{v} and \mathbf{v}_Γ are piecewise *constant* on mesh elements, i.e., $\mathbf{v}|_E \in [\mathbb{P}_0(E)]^{d \times d}$ for all $E \in \mathcal{T}_h$, and $\mathbf{v}_\Gamma|_F \in [\mathbb{P}_0(F)]^{(d-1) \times (d-1)}$ for all $F \in \Gamma_h$.

First, to each element $E \in \mathcal{T}_h$ and $F \in \Gamma_h$ we associate the integers $p_E \geq 1$ and $p_F \geq 1$, and introduce the finite-dimensional spaces:

$$\begin{aligned}
Q_h^b &= \{q \in L^2(\Omega) : q|_E \in \mathbb{P}_{p_E}(E) \forall E \in \mathcal{T}_h\}, \\
\mathbf{W}_h^b &= \{\mathbf{v} \in [L^2(\Omega)]^d : \mathbf{v}|_E \in [\mathbb{P}_{p_E}(E)]^d \forall E \in \mathcal{T}_h\}, \\
Q_h^\Gamma &= \{q_\Gamma \in L^2(\Gamma) : q_\Gamma|_F \in \mathbb{P}_{p_F}(F) \forall F \in \Gamma_h\}, \\
\mathbf{W}_h^\Gamma &= \{\mathbf{v}_\Gamma \in [L^2(\Gamma)]^{d-1} : \mathbf{v}_\Gamma|_F \in [\mathbb{P}_{p_F}(F)]^{d-1} \forall F \in \Gamma_h\}.
\end{aligned}$$

We remark that the polynomial degrees in the bulk and fracture discrete spaces are chosen *independently* of each other.

We next focus on equations (5.61a)–(5.61b) in the bulk and equations (5.61e)–(5.61f) in the fracture. We proceed as in [33], and multiply equations (5.61a)–(5.61b) by (sufficiently smooth) vector-valued and scalar-valued test functions, respectively, integrate by parts over an element $E \in \mathcal{T}_h$, and sum over all elements. Analogously, we multiply equations (5.61e)–(5.61f) by (sufficiently smooth) test functions, integrate by parts over an element $F \in \Gamma_h$ and sum over all the elements in Γ_h . We then discretize, use identity (5.7), and integrate by parts again the first equation in the bulk and the first equation in the fracture, to get the following general discrete formulation: find $p_h \in Q_h^b$, $\mathbf{u}_h \in \mathbf{W}_h^b$, $p_{\Gamma,h} \in Q_h^\Gamma$, and $\mathbf{u}_{\Gamma,h} \in \mathbf{W}_h^\Gamma$ such that

$$\int_{\mathcal{T}_h} \mathbf{v}^{-1} \mathbf{u}_h \cdot \mathbf{v} = \int_{\mathcal{T}_h} \nabla p_h \cdot \mathbf{v} + \int_{\mathcal{F}_h^I \cup \Gamma_h} \{\hat{p} - p_h\} \llbracket \mathbf{v} \rrbracket + \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^B \cup \Gamma_h} \llbracket \hat{p} - p_h \rrbracket \cdot \{\mathbf{v}\}, \quad (5.64)$$

$$\int_{\mathcal{T}_h} \mathbf{u}_h \cdot \nabla q - \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^B \cup \Gamma_h} \{\hat{\mathbf{u}}\} \cdot \llbracket q \rrbracket - \int_{\mathcal{F}_h^I \cup \Gamma_h} \llbracket \hat{\mathbf{u}} \rrbracket \{q\} = \int_{\mathcal{T}_h} f q, \quad (5.65)$$

$$\begin{aligned}
\int_{\Gamma_h} (\mathbf{v}_\Gamma^\tau \ell_\Gamma)^{-1} \mathbf{u}_{\Gamma,h} \cdot \mathbf{v}_\Gamma &= \int_{\Gamma_h} \nabla p_{\Gamma,h} \cdot \mathbf{v}_\Gamma + \int_{\mathcal{E}_{\Gamma,h}^I} \{\hat{p}_\Gamma - \hat{p}_{\Gamma,h}\} \llbracket \mathbf{v}_\Gamma \rrbracket + \\
&\int_{\mathcal{E}_{\Gamma,h}} \llbracket \hat{p}_\Gamma - \hat{p}_{\Gamma,h} \rrbracket \cdot \{\mathbf{v}_\Gamma\}, \quad (5.66)
\end{aligned}$$

$$\int_{\Gamma_h} \mathbf{u}_{\Gamma,h} \cdot \nabla q_\Gamma - \int_{\mathcal{E}_{\Gamma,h}} \{\hat{\mathbf{u}}_\Gamma\} \cdot \llbracket q_\Gamma \rrbracket - \int_{\mathcal{E}_{\Gamma,h}^I} \llbracket \hat{\mathbf{u}}_\Gamma \rrbracket \{q_\Gamma\} = \int_{\Gamma_h} \ell_\Gamma f_\Gamma q_\Gamma - \int_{\Gamma_h} \llbracket \hat{\mathbf{u}} \rrbracket q_\Gamma \quad (5.67)$$

for all $q \in Q_h^b$, $\mathbf{v} \in \mathbf{W}_h^b$, $q_\Gamma \in Q_h^\Gamma$ and $\mathbf{v}_\Gamma \in \mathbf{W}_h^\Gamma$. We point out that, in order to simplify the notation, we have dropped the subscript τ from the *tangent* gradient and divergence operators. Here, in the spirit of [33], the *numerical fluxes*

$$\hat{p} = (\hat{p}_E)_{E \in \mathcal{T}_h}, \quad \hat{\mathbf{u}} = (\hat{\mathbf{u}}_E)_{E \in \mathcal{T}_h}, \quad \hat{p}_\Gamma = (\hat{p}_{\Gamma,F})_{F \in \Gamma_h}, \quad \hat{\mathbf{u}}_\Gamma = (\hat{\mathbf{u}}_{\Gamma,F})_{F \in \Gamma_h},$$

are approximations to the analytical solutions \mathbf{u} and p , respectively, on the boundary of E and to p_Γ and \mathbf{u}_Γ , respectively, on the boundary of the fracture face F . The numerical fluxes \hat{p} , $\hat{\mathbf{u}}$, \hat{p}_Γ , $\hat{\mathbf{u}}_\Gamma$ must be interpreted as linear functionals taking values in the spaces $\Pi_{E \in \mathcal{T}_h} L^2(\partial E)$, $[\Pi_{E \in \mathcal{T}_h} L^2(\partial E)]^d$, $\Pi_{F \in \Gamma_h} L^2(\partial F)$, $[\Pi_{F \in \Gamma_h} L^2(\partial F)]^d$, respectively. By suitably choosing the numerical fluxes, we can obtain all the possible combinations of primal-primal, mixed-primal, primal-mixed and mixed-mixed

Table 5.4 Primal forms for the DG discretizations of the bulk-fracture problems

Method	Primal bilinear form	Reference equations
Primal-Primal (PP)	$\mathcal{A}_b^P(p, q) + \mathcal{A}_\Gamma^P(p_\Gamma, q_\Gamma) + \mathcal{C}((p, q), (p_\Gamma, q_\Gamma))$	(5.74), (5.75), (5.76)
Mixed-Primal (MP)	$\mathcal{A}_b^M(p, q) + \mathcal{A}_\Gamma^P(p_\Gamma, q_\Gamma) + \mathcal{C}((p, q), (p_\Gamma, q_\Gamma))$	(5.75), (5.76), (5.86)
Primal-Mixed (PM)	$\mathcal{A}_b^P(p, q) + \mathcal{A}_\Gamma^M(p_\Gamma, q_\Gamma) + \mathcal{C}((p, q), (p_\Gamma, q_\Gamma))$	(5.74), (5.76), (5.97)
Mixed-Mixed (MM)	$\mathcal{A}_b^M(p, q) + \mathcal{A}_\Gamma^M(p_\Gamma, q_\Gamma) + \mathcal{C}((p, q), (p_\Gamma, q_\Gamma))$	(5.76), (5.86), (5.97)

formulations for the bulk and fracture, respectively. In Table 5.4 we summarize the bilinear forms for all formulations, whose precise definition will be given in the forthcoming sections.

5.5.2.1 Primal-Primal Formulation

To obtain the primal-primal formulation, based on the *symmetric interior penalty Discontinuous Galerkin* (SIPDG) method, we choose the numerical fluxes $\hat{p} = \hat{p}(p_h)$, $\hat{\mathbf{u}} = \hat{\mathbf{u}}(p_h, p_{\Gamma,h})$, $\hat{p}_\Gamma = \hat{p}_\Gamma(p_{\Gamma,h})$, and $\hat{\mathbf{u}}_\Gamma = \hat{\mathbf{u}}_\Gamma(p_{\Gamma,h})$ as follows

$$\hat{p} = \begin{cases} \{p_h\} & \text{on } \mathcal{F}_h^I \\ 0 & \text{on } \mathcal{F}_h^D \\ p_h & \text{on } \mathcal{F}_h^N \\ p_h & \text{on } \Gamma_h \end{cases} \quad \hat{\mathbf{u}} = \begin{cases} \{\mathbf{v} \nabla p_h\} - \sigma_F \llbracket p_h \rrbracket & \text{on } \mathcal{F}_h^I \\ \mathbf{v} \nabla p_h - \sigma_F p_h \mathbf{n}_F & \text{on } \mathcal{F}_h^D \\ 0 & \text{on } \mathcal{F}_h^N \\ -[\alpha_\Gamma(\{p_h\} - p_{\Gamma,h}) \frac{\mathbf{n}_F}{2} + \beta_\Gamma \llbracket p_h \rrbracket] & \text{on } \Gamma_h \end{cases} \quad (5.68)$$

$$\hat{p}_\Gamma = \begin{cases} \{p_{\Gamma,h}\} & \text{on } \mathcal{E}_{\Gamma,h}^I \\ 0 & \text{on } \mathcal{E}_{\Gamma,h}^D \\ p_{\Gamma,h} & \text{on } \mathcal{E}_{\Gamma,h}^N \end{cases} \quad \hat{\mathbf{u}}_\Gamma = \begin{cases} \{\mathbf{v}_\Gamma^\tau \ell_\Gamma \nabla p_{\Gamma,h}\} - \sigma_e \llbracket p_{\Gamma,h} \rrbracket & \text{on } \mathcal{E}_{\Gamma,h}^I \\ \mathbf{v}_\Gamma^\tau \ell_\Gamma \nabla p_{\Gamma,h} - \sigma_e p_{\Gamma,h} \mathbf{n}_e & \text{on } \mathcal{E}_{\Gamma,h}^D \\ 0 & \text{on } \mathcal{E}_{\Gamma,h}^N \end{cases} \quad (5.69)$$

Here, we have introduced the discontinuity penalization parameters σ and $\sigma_\Gamma \in L^\infty(e^I \cup e^D)$. In particular, they are non-negative bounded functions and their precise definitions will be given in Definition 5.5 below. Moreover, we have used the notation $\sigma_F = \sigma|_F$, for $F \in \mathcal{F}_h^I \cup \mathcal{F}_h^D$ and $\sigma_e = \sigma_\Gamma|_e$ for $e \in e^I \cup e^D$. Note also that, with this choice, the numerical flux \hat{p} is double valued on Γ_h and single valued on $\mathcal{F}_h^I \cup \mathcal{F}_h^B$. By using the above definitions, and after eliminating the velocities \mathbf{u}_h and $\mathbf{u}_{\Gamma,h}$ in an elementwise manner as in [33], based on the fact that $\nabla Q_h \subseteq \mathbf{W}_h$, $\nabla Q_h^\Gamma \subseteq \mathbf{W}_h^\Gamma$ and employing the *lifting operators*

$$\mathcal{L}_b^{SIP} : [L^1(\mathcal{F}_h^I \cup \mathcal{F}_h^D)]^d \rightarrow \mathbf{W}_h^b, \quad \int_{\Omega} \mathcal{L}_b^{SIP}(\boldsymbol{\xi}) \cdot \mathbf{v} = - \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \{\mathbf{v}\} \cdot \boldsymbol{\xi}, \quad (5.70)$$

$$\mathcal{L}_{\Gamma}^{SIP} : [L^1(\mathcal{E}_{\Gamma,h}^I \cup e^D)]^{d-1} \rightarrow \mathbf{W}_{\Gamma}^{\Gamma}, \quad \int_{\Gamma} \mathcal{L}_{\Gamma}^{SIP}(\boldsymbol{\xi}_{\Gamma}) \cdot \mathbf{v}_{\Gamma} = - \int_{\mathcal{E}_{\Gamma,h}^I \cup e^D} \{\mathbf{v}_{\Gamma}\} \cdot \boldsymbol{\xi}_{\Gamma}, \quad (5.71)$$

for all $\mathbf{v} \in \mathbf{W}_h^b$ and $\mathbf{v}_{\Gamma} \in \mathbf{W}_{\Gamma}^{\Gamma}$, respectively, we obtain the following discrete formulation: find $(p_h, p_h^{\Gamma}) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}$ such that

$$\mathcal{A}_h^{PP}((p_h, p_h^{\Gamma}), (q, q_{\Gamma})) = \mathcal{L}_h^{PP}(q, q_{\Gamma}) \quad \forall (q, q_{\Gamma}) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}, \quad (5.72)$$

where the superscript PP stands for *primal-primal* and $\mathcal{L}_h^{PP} : \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma} \rightarrow \mathbb{R}$ is defined as $\mathcal{L}_h^{PP}(q, q_{\Gamma}) = \mathcal{L}_b^p(q) + \mathcal{L}_{\Gamma}^p(q_{\Gamma})$ and $\mathcal{A}_h^{PP} : (\mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}) \times (\mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}) \rightarrow \mathbb{R}$ is given by

$$\mathcal{A}_h^{PP}((p_h, p_h^{\Gamma}), (q, q_{\Gamma})) = \mathcal{A}_b^p(p_h, q) + \mathcal{A}_{\Gamma}^p(p_{\Gamma,h}, q_{\Gamma}) + C((p_h, p_{\Gamma,h}), (q, q_{\Gamma})) \quad (5.73)$$

with

$$\begin{aligned} \mathcal{A}_b^p(p_h, q) &= \int_{\mathcal{T}_h} \mathbf{v} \nabla p_h \cdot \nabla q + \int_{\mathcal{T}_h} \mathbf{v} \mathcal{L}_b^{SIP}(\llbracket p_h \rrbracket) \cdot \nabla q \\ &+ \int_{\mathcal{T}_h} \mathbf{v} \mathcal{L}_b^{SIP}(\llbracket q \rrbracket) \cdot \nabla p_h + \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \sigma_F \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket, \end{aligned} \quad (5.74)$$

$$\begin{aligned} \mathcal{A}_{\Gamma}^p(p_{\Gamma,h}, q_{\Gamma}) &= \int_{\Gamma_h} \mathbf{v}_{\Gamma}^{\tau} \ell_{\Gamma} \nabla p_{\Gamma,h} \cdot \nabla q_{\Gamma} + \int_{\Gamma_h} \mathbf{v}_{\Gamma}^{\tau} \ell_{\Gamma} \mathcal{L}_{\Gamma}^p(\llbracket p_{\Gamma,h} \rrbracket) \cdot \nabla q_{\Gamma} \\ &+ \int_{\Gamma_h} \mathbf{v}_{\Gamma}^{\tau} \ell_{\Gamma} \mathcal{L}_{\Gamma}^{SIP}(\llbracket q_{\Gamma} \rrbracket) \cdot \nabla p_{\Gamma,h} + \int_{\mathcal{E}_{\Gamma,h}^I \cup \mathcal{E}_{\Gamma,h}^D} \sigma_e \llbracket p_{\Gamma,h} \rrbracket \cdot \llbracket q_{\Gamma} \rrbracket, \end{aligned} \quad (5.75)$$

$$C((p_h, p_{\Gamma,h}), (q, q_{\Gamma})) = \int_{\Gamma_h} \beta_{\Gamma} \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket + \int_{\Gamma_h} \alpha_{\Gamma} (\{p_h\} - p_{\Gamma,h}) (\{q\} - q_{\Gamma,h}), \quad (5.76)$$

and

$$\mathcal{L}_b^p(q) = \int_{\mathcal{T}_h} f q, \quad \mathcal{L}_\Gamma^p(q_\Gamma) = \int_{\Gamma_h} \ell_\Gamma f_\Gamma q_\Gamma. \quad (5.77)$$

5.5.2.2 Mixed-Primal Formulation

We next address the choice of the numerical fluxes that leads to a mixed-primal formulation. Here, the mixed formulation will be based on the use of the LDG method [69, 79, 127, 128]. To this end, we define the numerical fluxes $\hat{p} = \hat{p}(p_h)$ and $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\mathbf{u}_h, p_h, p_{\Gamma,h})$ for the bulk as

$$\hat{p} = \begin{cases} \{p_h\} + \mathbf{b} \cdot \llbracket p_h \rrbracket & \text{on } \mathcal{F}_h^I, \\ 0 & \text{on } \mathcal{F}_h^D, \\ p_h & \text{on } \mathcal{F}_h^N, \\ p_h & \text{on } \Gamma_h, \end{cases} \quad \hat{\mathbf{u}} = \begin{cases} \{\mathbf{u}_h\} - \mathbf{b} \llbracket \mathbf{u}_h \rrbracket - \sigma_F \llbracket p_h \rrbracket & \text{on } \mathcal{F}_h^I, \\ \mathbf{u}_h - \sigma_F p_h \mathbf{n}_F & \text{on } \mathcal{F}_h^D, \\ 0 & \text{on } \mathcal{F}_h^N, \\ -[\alpha_\Gamma(\{p_h\} - p_{\Gamma,h}) \frac{\mathbf{n}_F}{2} + \beta_\Gamma \llbracket p_h \rrbracket] & \text{on } \Gamma_h, \end{cases} \quad (5.78)$$

whereas for the numerical fluxes in the fracture we adopt the same definition as in (5.68). Here, $\mathbf{b} \in [L^\infty(\mathcal{F}_h^I)]^d$ is a (possibly null) facewise constant vector-valued function such that $\|\mathbf{b}\|_{\infty, \mathcal{F}_h^I} \lesssim 1$. With this definition of the numerical fluxes, we obtain the following discrete mixed problem: find $((p_h, \mathbf{u}_h), p_\Gamma^\Gamma) \in Q_h^b \times \mathbf{W}_h^b \times Q_\Gamma^\Gamma$ such that

$$\mathcal{M}_b(\mathbf{u}_h, \mathbf{v}) + \mathcal{B}_b(p_h, \mathbf{v}) = \mathbf{0} \quad \forall \mathbf{v} \in \mathbf{W}_h^b, \quad (5.79)$$

$$-\mathcal{B}_b(q, \mathbf{u}_h) + \mathcal{S}_b(p_h, q) + C_1(p_h, q, p_{\Gamma,h}) = \mathcal{L}_b^p(q) \quad \forall q \in Q_h^b, \quad (5.80)$$

$$\mathcal{A}_\Gamma^p(p_{\Gamma,h}, q_\Gamma) + C_2(p_h, p_{\Gamma,h}, q_\Gamma) = \mathcal{L}_\Gamma^p(q_\Gamma) \quad \forall q_\Gamma \in Q_\Gamma^\Gamma, \quad (5.81)$$

where

$$\begin{aligned} \mathcal{M}_b(\mathbf{u}_h, \mathbf{v}) &= \int_{\mathcal{T}_h} \mathbf{v}^{-1} \mathbf{u}_h \cdot \mathbf{v}, \\ \mathcal{B}_b(p_h, \mathbf{v}) &= - \int_{\mathcal{T}_h} \nabla p_h \cdot \mathbf{v} + \int_{\mathcal{F}_h^I} \llbracket p_h \rrbracket \cdot (\{\mathbf{v}\} - \mathbf{b} \llbracket \mathbf{v} \rrbracket) + \int_{\mathcal{F}_h^D} p_h \mathbf{v} \cdot \mathbf{n}_F, \\ \mathcal{S}_b(p_h, q) &= \int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \sigma_F \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket, \\ C_1(p_h, q, p_{\Gamma,h}) &= \int_{\Gamma_h} \beta_\Gamma \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket + \int_{\Gamma_h} \alpha_\Gamma (\{p_h\} - p_{\Gamma,h}) \{q\}, \\ C_2(p_h, p_{\Gamma,h}, q_\Gamma) &= \int_{\Gamma_h} \alpha_\Gamma (p_{\Gamma,h} - \{p_h\}) q_\Gamma, \end{aligned}$$

and $\mathcal{A}_\Gamma^p(\cdot, \cdot)$ and $\mathcal{L}_\Gamma^p(\cdot)$ are defined as in (5.75) and (5.77), respectively. Also note that we have $C((p_h, p_{\Gamma,h}), (q, q_\Gamma)) = C_1(p_h, q, p_{\Gamma,h}) + C_2(p_h, p_{\Gamma,h}, q_\Gamma)$. For the purpose of the analysis, the bulk velocity \mathbf{u}_h can be eliminated elementwise by introducing the *lifting operator*, $\mathcal{L}_b^{LDG} : [L^1(\mathcal{F}_h^I \cup \mathcal{F}_h^D)]^d \rightarrow \mathbf{W}_h^b$, defined by

$$\int_{\mathcal{T}_h} \mathcal{L}_b^{LDG}(\boldsymbol{\xi}) \cdot \mathbf{v} = - \int_{\mathcal{F}_h^I} (\{\mathbf{v}\} - \mathbf{b}[\![\mathbf{v}]\!]) \cdot \boldsymbol{\xi} - \int_{\mathcal{F}_h^D} \boldsymbol{\xi} \cdot \mathbf{v} \quad \forall \mathbf{v} \in \mathbf{W}_h^b \quad (5.82)$$

to obtain the following discrete formulation: find $(p_h, p_h^\Gamma) \in Q_h^b \times Q_h^\Gamma$ such that

$$\mathcal{A}_h^{MP}((p_h, p_h^\Gamma), (q, q_\Gamma)) = \mathcal{L}_h^{MP}(q, q_\Gamma) \quad \forall (q, q_\Gamma) \in Q_h^b \times Q_h^\Gamma, \quad (5.83)$$

where the superscript *MP* stands for *mixed-primal* and $\mathcal{A}_h^{MP} : (Q_h^b \times Q_h^\Gamma) \times (Q_h^b \times Q_h^\Gamma) \rightarrow \mathbb{R}$ is defined as

$$\mathcal{A}_h^{MP}((p_h, p_h^\Gamma), (q, q_\Gamma)) = \mathcal{A}_b^M(p_h, q) + \mathcal{A}_\Gamma^P(p_{\Gamma,h}, q_\Gamma) + C((p_h, p_{\Gamma,h}), (q, q_\Gamma)). \quad (5.84)$$

Here, $\mathcal{L}_h^{MP} : Q_h^b \times Q_h^\Gamma \rightarrow \mathbb{R}$ is given by

$$\mathcal{L}_h^{MP}(q, q_\Gamma) = \mathcal{L}_b^M(q) + \mathcal{L}_\Gamma^P(q_\Gamma), \quad (5.85)$$

with

$$\begin{aligned} \mathcal{A}_b^M(p_h, q) &= \int_{\mathcal{T}_h} \mathbf{v}(\nabla p_h + \mathcal{L}_b^{LDG}(\llbracket p_h \rrbracket)) \cdot (\nabla q + \mathcal{L}_b^{LDG}(\llbracket q \rrbracket)) + \\ &\int_{\mathcal{F}_h^I \cup \mathcal{F}_h^D} \sigma_F \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket + \int_{\Gamma_h} \beta_\Gamma \llbracket p_h \rrbracket \cdot \llbracket q \rrbracket + \int_{\Gamma_h} \alpha_\Gamma (\{p_h\} - p_\Gamma) \{q\}, \end{aligned} \quad (5.86)$$

$$\mathcal{L}_b^M(q) = \int_{\mathcal{T}_h} f q. \quad (5.87)$$

5.5.2.3 Primal-Mixed Formulation

We next address the choice of the numerical fluxes that lead to a primal-mixed formulation, i.e. we approximate the problem in the bulk using the SIPDG method, and the problem in the fracture in mixed form, employing the LDG method. In the bulk we define the numerical fluxes \hat{p} and $\hat{\mathbf{u}}$ as in (5.68), whereas in the fracture we define the numerical fluxes $\hat{p}_\Gamma = \hat{p}_\Gamma(p_{\Gamma,h})$ and $\hat{\mathbf{u}}_\Gamma = \hat{\mathbf{u}}_\Gamma(\mathbf{u}_{\Gamma,h}, p_{\Gamma,h})$ as follows

$$\begin{aligned} \hat{p}_\Gamma &= \begin{cases} \{p_{\Gamma,h}\} + \mathbf{b}_\Gamma \cdot \llbracket p_{\Gamma,h} \rrbracket & \text{on } \mathcal{E}_{\Gamma,h}^I, \\ 0 & \text{on } \mathcal{E}_{\Gamma,h}^D, \\ p_{\Gamma,h} & \text{on } \mathcal{E}_{\Gamma,h}^N, \end{cases} \\ \hat{\mathbf{u}}_\Gamma &= \begin{cases} \{\mathbf{u}_{\Gamma,h}\} - \mathbf{b}_\Gamma \llbracket \mathbf{u}_{\Gamma,h} \rrbracket - \sigma_e \llbracket p_{\Gamma,h} \rrbracket & \text{on } \mathcal{E}_{\Gamma,h}^I, \\ \mathbf{u}_{\Gamma,h} - \sigma_e (p_{\Gamma,h} \mathbf{n}_e - g_\Gamma \mathbf{n}_e) & \text{on } \mathcal{E}_{\Gamma,h}^D, \\ 0 & \text{on } \mathcal{E}_{\Gamma,h}^N. \end{cases} \end{aligned} \quad (5.88)$$

Here, $\mathbf{b}_\Gamma \in [L^\infty(\mathbf{e}^I)]^{d-1}$ is a vector-valued function that is constant on each edge and it is chosen such that $\|\mathbf{b}_\Gamma\|_{\infty, \mathbf{e}^I} \lesssim 1$. This choice leads to the following primal-mixed problem: find $(p_h, (p_h^\Gamma, \mathbf{u}_{\Gamma,h})) \in Q_h^b \times Q_h^\Gamma \times \mathbf{W}_h^\Gamma$ such that

$$\mathcal{A}_b^p(p_h, q) + C_1((p_h, q), p_{\Gamma,h}) = \mathcal{L}_b^p(q) \quad \forall q \in Q_h^b, \quad (5.89)$$

$$\mathcal{M}_\Gamma(\mathbf{u}_{\Gamma,h}, \mathbf{v}_\Gamma) + \mathcal{B}_\Gamma(p_{\Gamma,h}, \mathbf{v}_\Gamma) = 0 \quad \forall \mathbf{v}_\Gamma \in \mathbf{W}_h^\Gamma, \quad (5.90)$$

$$-\mathcal{B}_\Gamma(q_\Gamma, \mathbf{u}_{\Gamma,h}) + \mathcal{S}_\Gamma(p_{\Gamma,h}, q_\Gamma) + C_2(p_h, (p_{\Gamma,h}, q_\Gamma)) = \mathcal{L}_\Gamma^p(q_\Gamma) \quad \forall q_\Gamma \in Q_h^\Gamma, \quad (5.91)$$

where

$$\begin{aligned} \mathcal{M}_\Gamma(\mathbf{u}_{\Gamma,h}, \mathbf{v}_\Gamma) &= \int_{\Gamma_h} (\mathbf{v}_\Gamma^\top \ell_\Gamma)^{-1} \mathbf{u}_{\Gamma,h} \cdot \mathbf{v}_\Gamma, \\ \mathcal{B}_\Gamma(p_{\Gamma,h}, \mathbf{v}_\Gamma) &= - \int_{\Gamma_h} \mathbf{v}_\Gamma \cdot \nabla p_{\Gamma,h} + \int_{\mathcal{E}_{h,\Gamma}^I} \llbracket p_{\Gamma,h} \rrbracket \cdot (\{\mathbf{v}_\Gamma\} - \mathbf{b}_\Gamma \llbracket \mathbf{v}_\Gamma \rrbracket) + \int_{\mathcal{E}_{h,\Gamma}^D} p_{\Gamma,h} \mathbf{v}_\Gamma \cdot \mathbf{n}_e, \\ \mathcal{S}_b(p_{\Gamma,h}, q_\Gamma) &= \int_{\mathcal{E}_{\Gamma,h}} \sigma_e \llbracket p_{\Gamma,h} \rrbracket \cdot \llbracket q_\Gamma \rrbracket, \end{aligned}$$

and $\mathcal{A}_b^p(p_h, q)$ and $\mathcal{L}_b^p(q)$ are defined as in (5.74) and (5.77), respectively. The variable $\mathbf{u}_{\Gamma,h}$ can be eliminated element-wise based on employing the lifting operator, $\mathcal{L}_\Gamma^{LDG} : [L^1(\mathcal{E}_h^I \cup \mathcal{E}_h^D)]^d \rightarrow \mathbf{W}_h^\Gamma$, defined by

$$\int_{\Gamma_h} \mathcal{L}_\Gamma^{LDG}(\boldsymbol{\xi}_\Gamma) \cdot \mathbf{v}_\Gamma = - \int_{\mathcal{E}_{\Gamma,h}^I} (\{\mathbf{v}_\Gamma\} - \mathbf{b}_\Gamma \llbracket \mathbf{v}_\Gamma \rrbracket) \cdot \boldsymbol{\xi}_\Gamma - \int_{\mathcal{E}_{\Gamma,h}^D} \boldsymbol{\xi}_\Gamma \cdot \mathbf{v}_\Gamma \quad \forall \mathbf{v}_\Gamma \in \mathbf{W}_h^\Gamma, \quad (5.92)$$

to obtain the following primal formulation: find $(p_h, p_h^\Gamma) \in Q_h^b \times Q_h^\Gamma$ such that

$$\mathcal{A}_h^{PM}((p_h, p_h^\Gamma), (q, q_\Gamma)) = \mathcal{L}_h^{PM}(q, q_\Gamma) \quad \forall (q, q_\Gamma) \in Q_h^b \times Q_h^\Gamma, \quad (5.93)$$

where the superscript *PM* stands for *primal-mixed* and $\mathcal{A}_h^{PM} : (Q_h^b \times Q_h^\Gamma) \times (Q_h^b \times Q_h^\Gamma) \rightarrow \mathbb{R}$ is defined as

$$\mathcal{A}_h^{PM}((p_h, p_h^\Gamma), (q, q_\Gamma)) = \mathcal{A}_b^p(p_h, q) + \mathcal{A}_\Gamma^M(p_{\Gamma,h}, q_\Gamma) + C((p_h, p_{\Gamma,h}), (q, q_\Gamma)). \quad (5.94)$$

Here, $\mathcal{L}_h^{PM} : Q_h^b \times Q_h^\Gamma \rightarrow \mathbb{R}$ is given by

$$\mathcal{L}_h^{PM}(q, q_\Gamma) = \mathcal{L}_b^p(q) + \mathcal{L}_\Gamma^M(q_\Gamma), \quad (5.95)$$

with

$$\mathcal{A}_\Gamma^M(p_{\Gamma,h}, q_\Gamma) = \int_{\Gamma_h} \mathbf{v}_\Gamma^\top \ell_\Gamma (\nabla p_{\Gamma,h} + \mathcal{L}_\Gamma^{LDG}(\llbracket p_{\Gamma,h} \rrbracket)) \cdot (\nabla q_\Gamma + \mathcal{L}_\Gamma^{LDG}(\llbracket q_\Gamma \rrbracket)) \quad (5.96)$$

$$+ \int_{\mathcal{E}_{\Gamma,h}^f \cup \mathcal{E}_{\Gamma,h}^p} \sigma_e \llbracket p_{\Gamma,h} \rrbracket \cdot \llbracket q_{\Gamma} \rrbracket, \quad (5.97)$$

$$\hat{\mathcal{L}}_{\Gamma}^M(q_{\Gamma}) = \int_{\Gamma_h} \ell_{\Gamma} f_{\Gamma} q_{\Gamma}. \quad (5.98)$$

5.5.2.4 Mixed-Mixed Formulation

Finally, if we approximate both the problem in the bulk and in the fracture with the LDG method by choosing the bulk numerical fluxes $\hat{p} = \hat{p}(p_h)$ and $\hat{\mathbf{u}} = \hat{\mathbf{u}}(\mathbf{u}_h, p_h, p_{\Gamma,h})$ as in (5.78) and the fracture numerical fluxes $\hat{p}_{\Gamma} = \hat{p}_{\Gamma}(p_{\Gamma,h})$ and $\hat{\mathbf{u}}_{\Gamma} = \hat{\mathbf{u}}_{\Gamma}(\mathbf{u}_{\Gamma,h}, p_{\Gamma,h})$ as in (5.88), we obtain the following mixed-mixed formulation: find $(p_h, p_{\Gamma,h}) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}$ and $(\mathbf{u}_h, \mathbf{u}_{\Gamma,h}) \in \mathbf{W}_h^b \times \mathbf{W}_h^{\Gamma}$ such that

$$\mathcal{M}_b(\mathbf{u}_h, \mathbf{v}) + \mathcal{B}_b(p_h, \mathbf{v}) = F_b(\mathbf{v}) \quad \forall \mathbf{v} \in \mathbf{W}_h^b, \quad (5.99)$$

$$-\mathcal{B}_b(q, \mathbf{u}_h) + \mathcal{S}_b(p_h, q) + \mathcal{C}_1(p_h, q, p_{\Gamma,h}) = G_b(q) \quad \forall q \in \mathcal{Q}_h^b, \quad (5.100)$$

$$\mathcal{M}_{\Gamma}(\mathbf{u}_{\Gamma,h}, \mathbf{v}_{\Gamma}) + \mathcal{B}_{\Gamma}(p_{\Gamma,h}, \mathbf{v}_{\Gamma}) = F_{\Gamma}(\mathbf{v}_{\Gamma}) \quad \forall \mathbf{v}_{\Gamma} \in \mathbf{W}_{\Gamma}^{\Gamma}, \quad (5.101)$$

$$-\mathcal{B}_{\Gamma}(q_{\Gamma}, \mathbf{u}_{\Gamma,h}) + \mathcal{S}_{\Gamma}(p_{\Gamma,h}, q_{\Gamma}) + \mathcal{C}_2(p_h, (p_{\Gamma,h}, q_{\Gamma})) = G_{\Gamma}(q_{\Gamma}) \quad \forall q_{\Gamma} \in \mathcal{Q}_h^{\Gamma}. \quad (5.102)$$

Again, based on employing the definition of the lifting operators (5.82) and (5.92), the bulk and fracture velocities can be eliminated, to yield the following equivalent formulation: find $(p_h, p_{\Gamma,h}) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}$ such that

$$\mathcal{A}_h^{MM}((p_h, p_{\Gamma,h}^{\Gamma}), (q, q_{\Gamma})) = \mathcal{L}_h^{MM}(q, q_{\Gamma}) \quad \forall (q, q_{\Gamma}) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}, \quad (5.103)$$

where the superscript MM stands for *mixed-mixed* and $\mathcal{A}_h^{MM} : (\mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}) \times (\mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma}) \rightarrow \mathbb{R}$ is defined as

$$\mathcal{A}_h^{MM}((p_h, p_{\Gamma,h}^{\Gamma}), (q, q_{\Gamma})) = \mathcal{A}_b^M(p_h, q) + \mathcal{A}_{\Gamma}^M(p_{\Gamma,h}, q_{\Gamma}) + \mathcal{C}((p_h, p_{\Gamma,h}), (q, q_{\Gamma})), \quad (5.104)$$

and $\mathcal{L}_h^{MM} : \mathcal{Q}_h^b \times \mathcal{Q}_h^{\Gamma} \rightarrow \mathbb{R}$ is given by

$$\mathcal{L}_h^{MM}(q, q_{\Gamma}) = \mathcal{L}_b^M(q) + \mathcal{L}_{\Gamma}^M(q_{\Gamma}). \quad (5.105)$$

5.5.3 Well-Posedness and Error Estimates

In this section, we recall the main results that ensure that the primal-primal (PP) (5.72), mixed-primal (MP) (5.83), primal-mixed (PM) (5.93) and mixed-mixed (MM)

(5.103) formulations are well-posed. We recall that, for the analysis, we assume the permeability tensors \mathbf{v} and \mathbf{v}_Γ^τ to be piecewise constant and that we employ the following notation $\bar{\mathbf{v}}_E = |\sqrt{\mathbf{v}}|_E|_2^2$ and $\bar{\mathbf{v}}_F^\tau = |\sqrt{\mathbf{v}_\Gamma^\tau}|_F|_2^2$, where $|\cdot|_2$ denotes the l_2 -norm. First, we give an appropriate definition of the discontinuity penalization parameters, so that we can work in a polytopic framework. Taking as a reference [13, 63, 64, 66, 67], we give the following two definitions for the bulk and fracture penalty functions.

Definition 5.5 The penalization parameter $\sigma : \mathcal{F}_h \setminus \Gamma_h \rightarrow \mathbb{R}^+$ for the bulk problem is defined facewise as

$$\sigma(\mathbf{x}) = \sigma_0 \begin{cases} \max_{E \in \{E^+, E^-\}} \frac{\bar{\mathbf{v}}_E p_E^2}{h_E} & \text{if } \mathbf{x} \subset F \in \mathcal{F}_h^I, \bar{F} = \partial \bar{E}^+ \cap \partial \bar{E}^-, \\ \frac{\bar{\mathbf{v}}_E p_E^2}{h_E} & \text{if } \mathbf{x} \subset F \in \mathcal{F}_h^D, \bar{F} = \partial \bar{E} \cap \partial \bar{\Omega}, \end{cases} \quad (5.106)$$

with $\sigma_0 > 0$ independent of p_E , $|E|$, and $|F|$. Analogously, the penalization parameter $\sigma_\Gamma : e \rightarrow \mathbb{R}^+$ for the fracture problem is defined edgewise as

$$\sigma_\Gamma(\mathbf{x}) = \sigma_{0,\Gamma} \begin{cases} \max_{F \in \{F^+, F^-\}} \frac{\bar{\mathbf{v}}_F^\tau p_F^2}{h_F} & \text{if } \mathbf{x} \subset e \in e^I, \bar{e} = \partial \bar{F}^+ \cap \partial \bar{F}^-, \\ \frac{\bar{\mathbf{v}}_F^\tau p_F^2}{h_F}, & \text{if } \mathbf{x} \subset e \in e^D, \bar{e} = \partial \bar{F} \cap \partial \bar{\Omega}, \end{cases} \quad (5.107)$$

with $\sigma_{0,\Gamma} > 0$ independent of p_F , $|F|$, and $|e|$.

Writing $\tilde{Q}^b = \{q = (q_1, q_2) \in H^1(\Omega_1) \times H^1(\Omega_2)\} \cap H^2(\mathcal{T}_h)$ and $\tilde{Q}^\Gamma = H^1(\Gamma) \cap H^2(\Gamma_h)$, we introduce the spaces $Q^b(h) = Q_h^b + \tilde{Q}^b$ and $Q^\Gamma(h) = Q_h^\Gamma + \tilde{Q}^\Gamma$ endowed with the *energy* norm

$$\| (q, q_\Gamma) \|_{\mathcal{C}}^2 = \|q\|_{b,DG}^2 + \|q_\Gamma\|_{\Gamma,DG}^2 + \|(q, q_\Gamma)\|_{\mathcal{C}}^2, \quad (5.108)$$

where

$$\begin{aligned} \|q\|_{b,DG}^2 &= \|\mathbf{v}^{1/2} \nabla q\|_{0,\mathcal{T}_h}^2 + \|\sigma_F^{1/2} \llbracket q \rrbracket\|_{0,\mathcal{F}_h^I \cup \mathcal{F}_h^D}^2, \\ \|q_\Gamma\|_{\Gamma,DG}^2 &= \|(\mathbf{v}_\Gamma^\tau \ell_\Gamma)^{1/2} \nabla q_\Gamma\|_{0,\Gamma_h}^2 + \|\sigma_e^{1/2} \llbracket q_\Gamma \rrbracket\|_{0,e^I \cup e^D}^2, \\ \|(q, q_\Gamma)\|_{\mathcal{C}}^2 &= \|\beta_\Gamma^{1/2} \llbracket q \rrbracket\|_{0,\Gamma_h}^2 + \|\alpha_\Gamma^{1/2} (\{q\} - q_\Gamma)\|_{0,\Gamma_h}^2. \end{aligned}$$

We remark that all the bilinear forms $\mathcal{A}_h^{**}(\cdot, \cdot)$, $** \in \{PP, MP, MM, PM\}$, defined in Sect. 5.5.2 are also well-defined on the extended space $Q^b(h) \times Q^\Gamma(h)$. We now recall the following result, and refer to [16] for the proof.

Lemma 5.7 *The following bounds hold*

$$\mathcal{A}_b^p(q, q) \gtrsim \|q\|_{b,DG}^2 \quad \forall q \in Q_h^b, \quad (5.109)$$

$$\mathcal{A}_b^p(p, q) \lesssim \|p\|_{b,DG} \|q\|_{b,DG} \quad \forall p, q \in Q^b(h), \quad (5.110)$$

$$\mathcal{A}_\Gamma^p(q_\Gamma, q_\Gamma) \gtrsim \|q_\Gamma\|_{\Gamma,DG}^2 \quad \forall q_\Gamma \in Q_h^\Gamma, \quad (5.111)$$

$$\mathcal{A}_\Gamma^p(p_\Gamma, q_\Gamma) \lesssim \|p_\Gamma\|_{\Gamma, DG} \|q_\Gamma\|_{\Gamma, DG} \quad \forall p_\Gamma, q_\Gamma \in Q^\Gamma(h), \quad (5.112)$$

$$\mathcal{A}_b^M(q, q) \gtrsim \|q\|_{b, DG}^2 \quad \forall q \in Q_h^b, \quad (5.113)$$

$$\mathcal{A}_b^M(p, q) \lesssim \|p\|_{b, DG} \|q\|_{b, DG} \quad \forall p, q \in Q^b(h), \quad (5.114)$$

$$\mathcal{A}_\Gamma^M(q_\Gamma, q_\Gamma) \gtrsim \|q_\Gamma\|_{\Gamma, DG}^2 \quad \forall q_\Gamma \in Q_\Gamma^h, \quad (5.115)$$

$$\mathcal{A}_\Gamma^M(p_\Gamma, q_\Gamma) \lesssim \|p_\Gamma\|_{\Gamma, DG} \|q_\Gamma\|_{\Gamma, DG} \quad \forall p_\Gamma, q_\Gamma \in Q^\Gamma(h). \quad (5.116)$$

The first and third estimates hold provided that σ_0 and $\sigma_{0,\Gamma}$ are chosen sufficiently large.

Employing Lemma 5.7, we can easily prove the well-posedness of all of our discrete problems, as stated in the following proposition.

Proposition 5.3 *Let the penalization parameters σ for the problem in the bulk and in the fracture be defined as in (5.106) and (5.107), respectively, and suppose that for the primal formulations σ_0 and $\sigma_{0,\Gamma}$ are chosen sufficiently large. Then, all the formulations (5.72), (5.83), (5.93) and (5.103) are well-posed.*

Next we prove error bounds in the discrete energy norm (5.108). To this end, for each subdomain Ω_i , $i = 1, 2$, we denote by \mathcal{E}_i the classical continuous extension operator (cf. [136]) $\mathcal{E}_i : H^s(\Omega_i) \rightarrow H^s(\mathbb{R}^d)$, for $s \in \mathbb{N}_0$. Similarly, we denote by \mathcal{E}_Γ the continuous extension operator $\mathcal{E}_\Gamma : H^s(\Gamma) \rightarrow H^s(\mathbb{R}^{d-1})$, for $s \in \mathbb{N}_0$. We then make the following regularity assumptions for the analytical solution (p, p_Γ) of problem (5.62).

Assumption 5.3 Let $\mathcal{T}_\# = \{T_E\}$ and $\mathcal{F}_\# = \{T_F\}$ denote the associated coverings of Ω and Γ , respectively, cf. Definition 5.2. We assume that the analytical solution (p, p_Γ) is such that:

- A1 For every $E \in \mathcal{T}_h$, if $E \subset \Omega_i$, we have $\mathcal{E}_i p_i|_{T_E} \in H^{r_E}(T_E)$, where $r_E \geq 1 + d/2$ and $T_E \in \mathcal{T}_\#$, with $E \subset T_E$;
- A2 For every $F \in \mathcal{T}_h$, we have $\mathcal{E}_\Gamma p_\Gamma|_{T_F} \in H^{r_F}(T_F)$, where $r_F \geq 1 + (d-1)/2$ and $T_F \in \mathcal{F}_\#$, with $F \subset T_F$. \square

Assumption 5.4 We assume that the normal components of the exact fluxes $\mathbf{v}\nabla p$ and $\ell_\Gamma \mathbf{v}_\Gamma^\tau \nabla p_\Gamma$ are continuous across mesh interfaces, that is $[[\mathbf{v}\nabla p]] = 0$ on \mathcal{F}_h^I and $[[\ell_\Gamma \mathbf{v}_\Gamma^\tau \nabla p_\Gamma]] = 0$ on e^I . \square

From Proposition 5.3 and Strang's second lemma, the following abstract error bound follows directly.

Lemma 5.8 *Let the hypotheses of Proposition 5.3 be satisfied. Then, for all the discrete formulations presented in Sect. 5.5.2, the following abstract error bound holds*

$$\begin{aligned} |||(p, p_\Gamma) - (p_h, p_{\Gamma,h})||| &\lesssim \inf_{(q, q_\Gamma) \in Q_h^b \times Q_\Gamma^h} |||(p, p_\Gamma) - (q, q_\Gamma)||| \\ &+ \sup_{(w, w_\Gamma) \in Q_h^b \times Q_\Gamma^h} \frac{|\mathcal{R}_h^{**}((p, p_\Gamma), (w, w_\Gamma))|}{|||(w, w_\Gamma)|||}, \end{aligned} \quad (5.117)$$

where the residual \mathcal{R}_h^{**} is defined as

$$\mathcal{R}_h^{**}((p, p_\Gamma), (w, w_\Gamma)) = \mathcal{A}_h^{**}((p, p_\Gamma), (w, w_\Gamma)) - \mathcal{L}_h^{**}(w, w_\Gamma),$$

with $** \in \{PP, MP, MM, PM\}$.

We now recall the following result that provides a bound on the residuals stemming from formulations (5.72), (5.83), (5.93) and (5.103).

Lemma 5.9 [16, Lemma 5.6, Lemma 5.7] *Let (p, p_Γ) be the analytical solution of problem (5.62) satisfying the regularity Assumptions 5.3 and 5.4. Then, for every $w \in Q^b(h)$ and $w_\Gamma \in Q^\Gamma(h)$, we have that*

$$|\mathcal{R}_b^P(p, w)|^2 \lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{p_E^{2(r_E-1)}} \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 \left[\bar{\mathbf{v}}_E^2 \right] \cdot \|w\|_{b,DG}^2, \quad (5.118)$$

$$|\mathcal{R}_\Gamma^P(p_\Gamma, w_\Gamma)|^2 \lesssim \sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{p_F^{2(r_F-1)}} \|\mathcal{E} p_\Gamma\|_{H^{r_F}(T_F)}^2 \left[(\bar{\mathbf{v}}_F^\tau \ell_\Gamma)^2 \right] \cdot \|w_\Gamma\|_{\Gamma,DG}^2, \quad (5.119)$$

$$|\mathcal{R}_b^M(p, w)|^2 \lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{p_E^{2(r_E-1)}} \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 \left[\bar{\mathbf{v}}_E^2 \right] \cdot \|w\|_{b,DG}^2, \quad (5.120)$$

$$|\mathcal{R}_\Gamma^M(p_\Gamma, w_\Gamma)|^2 \lesssim \sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{p_F^{2(r_F-1)}} \|\mathcal{E} p_\Gamma\|_{H^{r_F}(T_F)}^2 \left[\bar{\mathbf{v}}_F^\tau \ell_\Gamma \right]^2 \cdot \|w_\Gamma\|_{\Gamma,DG}^2. \quad (5.121)$$

The above bounds, together with the observation that, for all the cases, the residual can always be split into two contributions: one involving the approximation of the problem in the bulk and one involving the approximation of the problem in the fracture, i.e.,

$$\mathcal{R}_h^{**}((p, p_\Gamma), (w, w_\Gamma)) = \mathcal{R}_b^*(p, w) + \mathcal{R}_\Gamma^*(p_\Gamma, w_\Gamma), \quad (5.122)$$

are the key ingredients required to derive main result of this section.

Theorem 5.3 *Let $\mathcal{T}_\# = \{T_E\}$ and $\mathcal{F}_\# = \{T_F\}$ denote the associated coverings of Ω and Γ , respectively, consisting of shape-regular simplices as in Definition 5.2, satisfying Assumption 5.2. Let (p, p_Γ) be the solution of problem (5.62) and $(p_h, p_{\Gamma,h}) \in Q_h^b \times Q_h^\Gamma$ be its approximation obtained with the method PP, MP, MM or PM, with the penalization parameters given by (5.106) and (5.107) and σ_0 and $\sigma_{0,\Gamma}$ sufficiently large for the primal formulations. Moreover, suppose that the analytical solution (p, p_Γ) satisfies the regularity Assumptions 5.3 and 5.4. Then, the following error bound holds*

$$\begin{aligned} |||(p, p_\Gamma) - (p_h, p_{\Gamma,h})|||^2 &\lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{p_E} G_E^*(\bar{\mathbf{v}}_E) \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 + \\ &\sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{p_F} G_F^*(\bar{\mathbf{v}}_F^\tau) \|\mathcal{E}_\Gamma p_\Gamma\|_{H^{r_F}(T_F)}^2, \end{aligned}$$

where $\mathcal{E} p$ is to be interpreted as $\mathcal{E}_1 p_1$ when $E \subset \Omega_1$ or as $\mathcal{E}_2 p_2$ when $E \subset \Omega_2$. Here, $s_E = \min(p_E + 1, r_E)$, $s_F = \min(p_F + 1, r_F)$, and the constants satisfy

$$G_E^P(\bar{\mathbf{v}}_E) \lesssim \bar{\mathbf{v}}_E \quad G_F^P(\bar{\mathbf{v}}_F^\tau) \lesssim \bar{\mathbf{v}}_F^\tau, \quad G_E^M(\bar{\mathbf{v}}_E) \lesssim \bar{\mathbf{v}}_E \quad G_F^M(\bar{\mathbf{v}}_F^\tau) \lesssim \bar{\mathbf{v}}_F^\tau \ell_\Gamma.$$

Proof From Lemma 5.8 we deduce that the error satisfies the following abstract bound

$$\begin{aligned} |||(p, p_\Gamma) - (p_h, p_{\Gamma,h})||| &\lesssim \underbrace{\inf_{(q, q_\Gamma) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^\Gamma} |||(p, p_\Gamma) - (q, q_\Gamma)|||}_I + \\ &\underbrace{\sup_{(w, w_\Gamma) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^\Gamma} \frac{|\mathcal{R}_h((p, p_\Gamma), (w, w_\Gamma))|}{|||(w, w_\Gamma)|||}}_{II}. \end{aligned} \quad (5.123)$$

For the term I , exploiting the approximation results stated in Lemma 5.2, we obtain

$$I \lesssim \sum_{E \in \mathcal{T}_h} \bar{\mathbf{v}}_E \frac{h_E^{2(s_E-1)}}{p_E} \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 + \sum_{F \in \Gamma_h} \bar{\mathbf{v}}_F^\tau \ell_\Gamma \frac{h_F^{2(s_F-1)}}{p_F} \|\mathcal{E}_\Gamma p_\Gamma\|_{H^{r_F}(T_F)}^2. \quad (5.124)$$

The statement of the theorem follows from (5.124), together with the bound on Term II deriving from what observed in (5.122) and Lemma 5.9.

If the hypotheses of Theorem 5.3 hold, we can also derive error estimates for the velocities \mathbf{u} and \mathbf{u}_Γ for the mixed-primal, primal-mixed, and mixed-mixed formulations. More precisely, if $(\mathbf{u}, \mathbf{u}_\Gamma) \in \mathbf{W}$ and $(p, p_\Gamma) \in M$ is the solution of problem (5.62), then, if $((p_h, \mathbf{u}_h), p_{\Gamma,h}) \in \mathcal{Q}_h^b \times \mathbf{W}_h^b \times \mathcal{Q}_h^\Gamma$ is the approximation obtained with the mixed-primal method (5.80), we have that

$$\|\mathbf{u} - \mathbf{u}_h\|_{0, \mathcal{T}_h}^2 \lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{p_E} G_E^M \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 + \sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{p_F} G_F^P \|\mathcal{E}_\Gamma p_\Gamma\|_{H^{r_F}(T_F)}^2.$$

Analogously, if $(p_h, (p_{\Gamma,h}, \mathbf{u}_{\Gamma,h})) \in \mathcal{Q}_h^b \times \mathcal{Q}_h^\Gamma \times \mathbf{W}_h^\Gamma$ is the approximation computed with the primal-mixed method (5.90), we deduce that

$$\|\mathbf{u}_\Gamma - \mathbf{u}_{\Gamma,h}\|_{0,\Gamma_h}^2 \lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{P_E^{2(r_E-1)}} G_E^P \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 + \sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{P_F^{2(r_F-1)}} G_F^M \|\mathcal{E}_\Gamma p_\Gamma\|_{H^{r_F}(T_F)}^2.$$

Finally, if $((p_h, \mathbf{u}_h), (p_{\Gamma,h}, \mathbf{u}_{\Gamma,h})) \in Q_h^b \times \mathbf{W}_h^b \times Q_h^\Gamma \times \mathbf{W}_h^\Gamma$ is the approximation obtained with the mixed-mixed method (5.100), then the following bound holds

$$\begin{aligned} \|\mathbf{u} - \mathbf{u}_h\|_{0,\mathcal{T}_h}^2 + \|\mathbf{u}_\Gamma - \mathbf{u}_{\Gamma,h}\|_{0,\Gamma_h}^2 &\lesssim \sum_{E \in \mathcal{T}_h} \frac{h_E^{2(s_E-1)}}{P_E^{2(r_E-1)}} G_E^M \|\mathcal{E} p\|_{H^{r_E}(T_E)}^2 \\ &+ \sum_{F \in \Gamma_h} \frac{h_F^{2(s_F-1)}}{P_F^{2(r_F-1)}} G_F^M \|\mathcal{E}_\Gamma p_\Gamma\|_{H^{r_F}(T_F)}^2. \end{aligned}$$

Here, the constants G_E^M , G_F^P , G_E^P and G_F^M are defined as in Theorem 5.3. We refer to [16] for further details.

5.5.4 Numerical Results

In this section we present three sets of two-dimensional numerical experiments employing the paradigmatic *primal-primal* and *mixed-primal* settings. With the first set of experiments we aim to validate the theoretical convergence results of Sect. 5.5.3, by considering a test case with known analytical solution. With the second and third sets of experiments, we assess the capability of the method for handling more complicated geometries, namely networks of partially immersed fractures and networks of *intersecting* fractures. All the numerical tests have been implemented in MATLAB[®] and the polygonal meshes conforming to the fractures have been obtained by suitably modifying the code `PolyMesher` [142].

5.5.4.1 Example 1: Problem with a Known Analytical Solution

We consider the domain $\Omega = (0, 1)^2$ and the fracture $\Gamma = \{(x, y) \in \Omega : x = 0.5\}$. Following [14, 73], we select the analytical solution in the bulk and the fracture as follows

$$p = \begin{cases} \sin(4x) \cos(\pi y) & \text{if } x < 0.5, \\ \cos(4x) \cos(\pi y) & \text{if } x > 0.5, \end{cases} \quad p_\Gamma = \xi [\cos(2) + \sin(2)] \cos(\pi y),$$

so that they satisfy the coupling conditions (5.61i)–(5.61j) with $\mathbf{v} = \mathbf{I}$, provided that $\beta_\Gamma = 2$, that is $\mathbf{v}_\Gamma^n / \ell_\Gamma = 4$. In particular, here we choose the tangential and normal components of the permeability tensor in the fracture as $\mathbf{v}_\Gamma^t = 10^2$ and $\mathbf{v}_\Gamma^n = 4 \cdot 10^{-2}$, respectively, and the fracture thickness $\ell_\Gamma = 10^{-2}$. Moreover, in the experiments we set $\xi = \frac{3}{4}$. We impose Dirichlet boundary conditions on the whole $\partial\Omega$ and also on

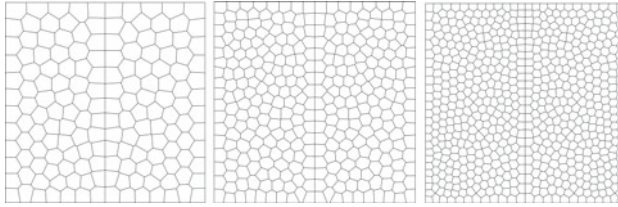


Fig. 5.18 Example 1 (Sect. 5.5.4.1). Three refinements of the polygonal mesh grid conforming to the fracture

$\partial\Gamma$. Finally the source terms are chosen accordingly as

$$f = \begin{cases} \sin(4x) \cos(\pi y)(16 + \pi^2) & \text{if } x < 0.5, \\ \cos(4x) \cos(\pi y)(16 + \pi^2) & \text{if } x > 0.5, \end{cases}$$

$$f_\Gamma = \cos(\pi y)[\cos(2) + \sin(2)](\xi \mathbf{v}_\Gamma^\tau \pi^2 + \frac{4}{\ell_\Gamma}).$$

In Fig. 5.18, we show three levels of refinement of the polygonal mesh conforming to the fracture employed in the computations. In order to test the h -convergence properties of our methods, thus validating the error estimate for the *energy* norm stated in Theorem 5.3, we compute the quantity $\|p - p_h\|_{1, \mathcal{T}_h} + \|p_\Gamma - p_{\Gamma, h}\|_{1, \Gamma_h}$. The plots in Fig. 5.19 show the computed errors as a function of the inverse of the mesh size h (loglog scale), together with the expected convergence rates. In particular, Fig. 5.19a shows the results obtained with the primal-primal approximation, while Fig. 5.19b shows the analogous results for the mixed-primal method. Each plot consists of four lines: every line shows the behaviour of the energy norm of the error for a different polynomial degree in the bulk (we consider uniform polynomial degrees $p_E = 1, 2, 3, 4$ for all $E \in \mathcal{T}_h$). For the fracture problem we always choose a uniform quadratic polynomial degree, i.e., $k_F = 2$ for all $F \in \Gamma_h$. For both the (PP) and (MP) method the theoretical convergence rates are clearly obtained, coinciding with $\min(p_E, p_F)$. In particular, the convergence rate is equal to 1 in the linear case, i.e., when $p_E = 1$ for all $E \in \mathcal{T}_h$, and it is equal to 2 in all the other cases, since the approximation of the fracture problem is always quadratic. Note also that the (PP) and (MP) methods achieve the same level of accuracy.

5.5.4.2 Example 2: Immersed Fractures

We now investigate the capability of our discretization methods to deal with *immersed* fractures. To this end, we take as a reference [4], where the mathematical model [118] is extended to fully immersed fractures. In particular, we supplement Eqs. (5.61) with a condition prescribing the behaviour of the fluid at the fracture tips immersed in the porous medium. As in [4], we impose that $\mathbf{v}_\Gamma^\tau \nabla_\tau p_\Gamma \cdot \boldsymbol{\tau} = 0$ on $\partial\Gamma \setminus \partial\Omega$, i.e., that the mass transfer across the immersed tips can be neglected.

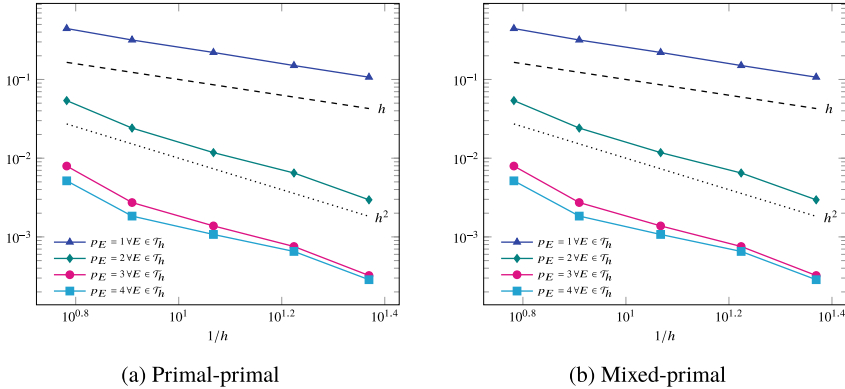
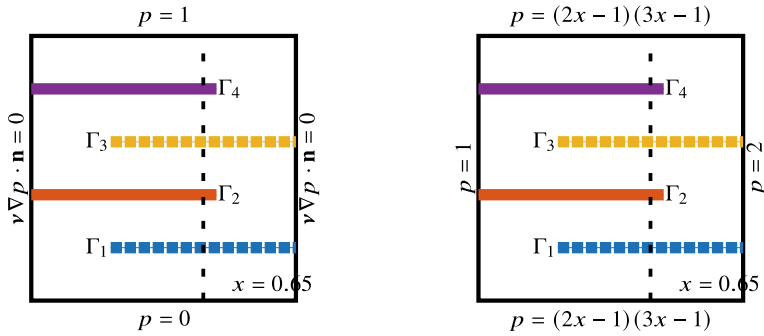


Fig. 5.19 Example 1 (Sect.5.5.4.1). Computed errors as a function of $1/h$ (loglog scale) and expected convergence rates for uniform bulk polynomial degrees $p_E = 1, 2, 3, 4$ for all $E \in \mathcal{T}_h$ and fixed uniform fracture polynomial degree $p_F = 2$ for all $F \in \Gamma_h$. Case Primal-Primal on the left and Mixed-Primal on the right

We employ again the paradigmatic primal-primal and mixed-primal approximation schemes to reproduce some numerical experiments already proposed in [4]. We consider the computational domain $\Omega = (0, 1)^2$ cut by four partially immersed fractures, namely $\Gamma_1 = \{(x, y) \in (0, 1)^2 : x \geq 0.3, y = 0.2\}$, $\Gamma_2 = \{(x, y) \in (0, 1)^2 : x \leq 0.7, y = 0.4\}$, $\Gamma_3 = \{(x, y) \in (0, 1)^2 : x \geq 0.3, y = 0.6\}$, $\Gamma_4 = \{(x, y) \in (0, 1)^2 : x \leq 0.7, y = 0.8\}$. The fractures Γ_2 and Γ_4 are impermeable ($\mathbf{v}_\Gamma^\tau = \mathbf{v}_\Gamma^n = 10^{-2}$), while Γ_1 and Γ_3 are partially permeable ($\mathbf{v}_\Gamma^n = 10^{-2}$, $\mathbf{v}_\Gamma^\tau \in \{100, 1\}$). With the aim of investigating the dependence of the flow on the physical properties of the fractures, we consider two different configurations (A and B), by varying the value of the permeability \mathbf{v}_Γ^τ on the partially permeable fractures Γ_1 and Γ_3 and the boundary conditions as illustrated in Fig. 5.20. At the extremities of the fractures that are non-immersed, i.e., $\partial\Gamma \cap \partial\Omega$, we impose boundary conditions that are consistent with those imposed on $\partial\Omega$ at that point. In both cases we consider an isotropic bulk permeability tensor, i.e., $\mathbf{v} = \mathbf{I}$ and we assume that all the fractures have aperture $\ell_\Gamma = 0.01$. Moreover, we take the forcing terms $f = f_\Gamma = 0$, so that the flow is only generated by the boundary conditions. Finally, we choose the parameter $\xi = 0.55$. Our results have been obtained with Cartesian grids aligned with the fractures, consisting of 26243 elements; this is approximately the same as in [4]. We remark that each immersed fracture tips coincides with a mesh vertex (in the case when the fracture ends at an edge of an element, the tip is considered as an additional vertex for the quadrilateral, which then becomes a pentagon). For both the (PP) and (MP) approximations we choose uniform *linear* polynomial degrees for both the bulk and fracture problems. In Fig. 5.21 we show the results obtained with the (PP) and (MP) methods for configuration A; in Fig. 5.22 we show analogous results for the configuration B. In particular, in both figures, we report the pressure field in the bulk with the streamlines of the velocity (left), the value of the bulk pressure along the line



(a) Configuration A: $v_T^r = 100$ on Γ_1, Γ_3

(b) Configuration B: $v_T^r = 1$ on Γ_1, Γ_3

Fig. 5.20 Example 2 (Sect. 5.5.4.2). Immersed fractures: configurations and boundary condition for the test cases A and B

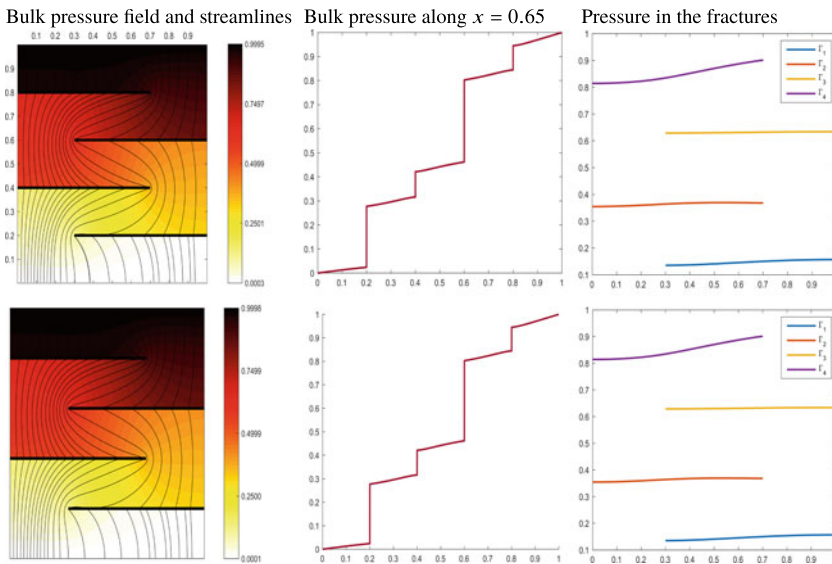


Fig. 5.21 Example 2 (Sect. 5.5.4.2). Immersed fractures; configuration A, primal-primal approximation (top) and mixed-primal approximation (bottom)

$x = 0.65$ (middle) and the pressure field inside the four fractures (right). The top line of each figure encloses the results obtained with the (PP) approximation, while the bottom line presents those obtained with the (MP) method. For both the (PP) and (MP) schemes, our results are in perfect agreement with those obtained in [4], thus showing that our approximation schemes can be easily extended to the treatment of more complex situations. Moreover, for this example, we observe that the (PP) and (MP) methods deliver the same level of accuracy.

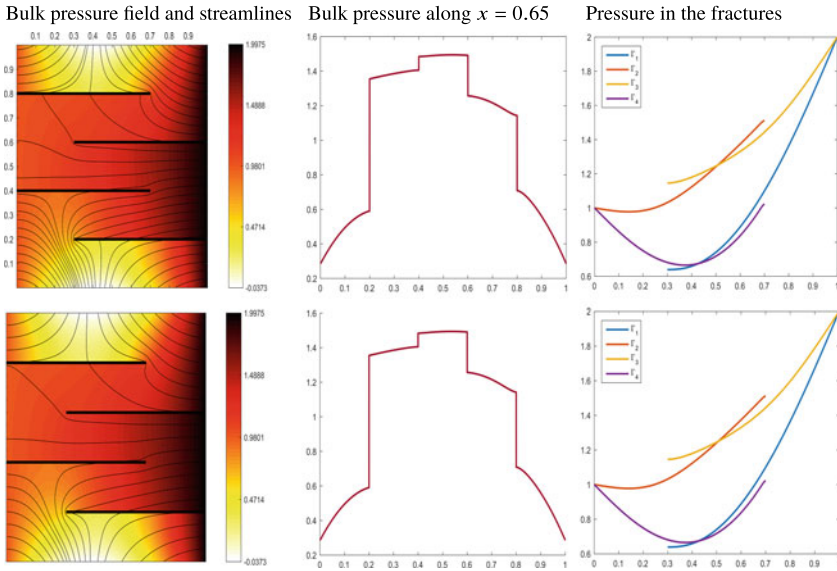


Fig. 5.22 Example 2 (Sect. 5.5.4.2). Immersed fractures; configuration B, primal-primal approximation (top) and mixed-primal approximation (bottom)

5.5.4.3 Example 3: Network of Intersecting Fractures

We conclude with a test case, already presented in [14], that aims at investigating the capability of our method for dealing with a network of *intersecting* fractures, which is also totally immersed in the bulk domain. In order to proceed, we need to complement our mathematical model (5.61) with some conditions at the intersection points, prescribing the behaviour of the fluid. In particular, we impose pressure continuity and flux conservation, as in [49, 56, 96]. At the immersed tips we impose the no flux condition $\mathbf{v}_\Gamma^t \nabla_\tau p_\Gamma \cdot \boldsymbol{\tau} = 0$ as above. We also mention that this numerical experiment was first presented in [18] employing the mimetic finite difference method. Here, we employ a suitable modification of the *primal-primal* scheme, which is able to handle intersecting fractures by virtue of an appropriate definition of jump and average operators at the intersection points. We refer to [15] for a detailed analysis of this scheme.

In the numerical simulations, we consider the bulk domain $\Omega = (0, 1)^2$ and the network made of 10 intersecting fractures that is shown in Fig. 5.23a.

We impose homogeneous Dirichlet boundary conditions on the whole $\partial\Omega$ and define the source terms in the bulk and in the fracture as

$$f(x, y) = \begin{cases} 10 & \text{if } (x - 0.1)^2 + (y - 0.1)^2 \leq 0.04, \\ -10 & \text{if } (x - 0.9)^2 + (y - 0.9)^2 \leq 0.04, \end{cases} \quad f_\Gamma = 0,$$

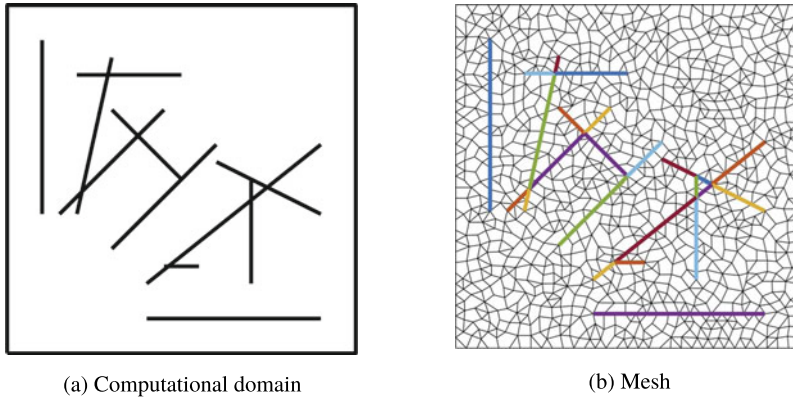


Fig. 5.23 Example 3 (Sect. 5.5.4.3). Network of intersecting fractures: computational domain (left) and a sample of the polygonal mesh employed for the computations (right)

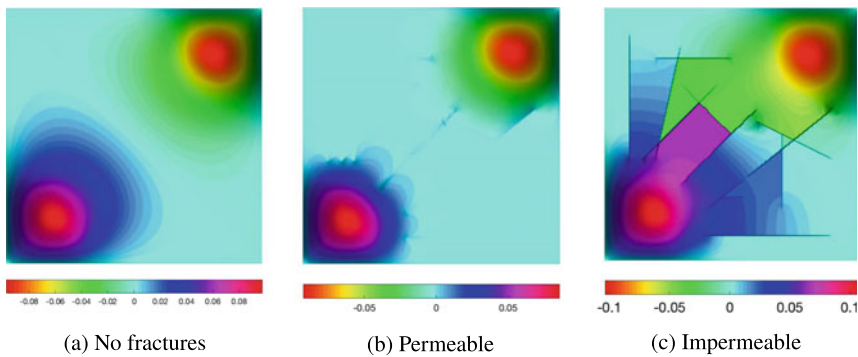


Fig. 5.24 Example 3 (Sect. 5.5.4.3). Network of intersecting fractures: discrete pressure in the bulk for the three test cases, no fractures (left), permeable network $\nu_{\Gamma}^{\tau} = \nu_{\Gamma}^n = 1000$ (middle), impermeable network $\nu_{\Gamma}^{\tau} = \nu_{\Gamma}^n = 0.001$ (right)

respectively. We note that the source term in the bulk is defined so that a source is present in the lower left corner of the domain and a sink in its top right corner. We assume that the porous medium in the bulk is isotropic and homogeneous, i.e., $\mathbf{v} = \text{Id}$. With the aim of testing the behaviour of the bulk pressure depending on the permeability properties of the fracture network, we consider three different configurations:

1. **No fractures** are present in the porous medium;
2. **Permeable** network: all the fractures have high permeability properties with $\nu_{\Gamma}^{\tau} = \nu_{\Gamma}^n = 1000$ and constant thickness $\ell_{\Gamma} = 0.01$;
3. **Impermeable** network: all the fractures have blocking properties with $\nu_{\Gamma}^{\tau} = \nu_{\Gamma}^n = 0.001$ and constant thickness $\ell_{\Gamma} = 0.01$.

In Fig. 5.23b we show a detail of the polygonal mesh conforming to the fracture network that we employed for the simulations. The discrete pressures for the problem in the bulk, obtained with the primal-primal approximation, in the three cases outlined above, are presented in Fig. 5.24. In particular, one may observe that, when the network is permeable, the bulk pressure is only marginally affected by the presence of the fractures, so that it reaches maximum and minimum values that are only slightly lower than those of the non-fractured case (see the comparison between Fig. 5.24a and Fig. 5.24b). In contrast, in the impermeable case, jumps of the bulk pressure across the fractures are clearly observed, cf., Fig. 5.24c. Finally, we note that our results are in good agreement with those obtained in [96].

5.6 Conclusions

In this work we have provided a comprehensive review of the current development of PolyDG methods for geophysical applications, addressing as paradigmatic applications the numerical modeling of seismic wave propagation and fracture reservoir simulations. After having recalled the theoretical background of the analysis of PolyDG methods (cf. Section 5.2), in Sect. 5.3 we discussed the issue of efficiently implementing DG methods on polytopic meshes, addressing in detail the issue of numerical quadrature and recalling the main results contained in [22], where a new *quadrature free* algorithm for the numerical evaluation of the integrals required to assemble the mass and stiffness matrices has been proposed. More precisely, a cubature method, which does not require the definition of a set of nodes and weights on the domain of integration, and allows for the exact integration of polynomial functions based on evaluating the integrand only at the vertices of the polytopic integration domain, is presented and tested in both two- and three-dimensions. This approach shows a remarkable gain in terms of CPU time with respect to classical quadrature rules, while maintaining the same degree of accuracy. In Sect. 5.4 we presented PolyDG methods for the approximate solution of the elastodynamics equations on computational meshes consisting of polytopic elements. We analysed the well-posedness of the numerical formulation and proved *hp*-version *a priori* error estimates for the semi-discrete scheme. The fully discrete method is then obtained based on employing the leap-frog scheme for the time discretization. To test the numerical performance and fully exploit the flexibility in the process of mesh design offered by polytopic elements numerical experiments have been presented. Section 5.5 focused on the problem of modeling the flow in a fractured porous medium. For ease of presentation and analysis we have assumed the medium to be cut by a single non-immersed fracture and have reviewed the unified development and analysis of PolyDG methods for this class of problems. These error bounds have been validated through numerical tests. Moreover, we have demonstrated that our approach can be extended to handle networks of partially immersed fractures and networks of *intersecting* fractures, cf. [15]. To conclude we mention that the current developments of PolyDG methods, not discussed here for the sake of brevity, include the exploitation of agglomeration-

based algorithms to design multilevel and multigrid methods for the efficient iterative solution of the (linear) system of equations stemming from the PolyDG discretization. Indeed, multigrid/multilevel solvers require the definition of a succession of coarse grids, based on the original ‘fine’ grid. The process of defining the coarser grids involves what is called agglomeration, i.e., the combination of several nodes or control volumes or coefficients from the original grid. In this context, the flexibility offered by polytopic grids can be fully exploited. Some pioneering works on the analysis of agglomeration-based multigrid/multilevel solvers and preconditioners can be found in [21, 23, 30]; cf. also the classical approach based on a sequence of simplicial/quadrilateral meshes [31].

References

1. J. Aghili, D.A. Di Pietro, B. Ruffini, An *hp*-hybrid high-order method for variable diffusion on general meshes. *Comput. Methods Appl. Math.* **17**(3), 359–376 (2017)
2. C. Alboin, J. Jaffré, J.E. Roberts, C. Serres, Modeling fractures as interfaces for flow and transport in porous media, in *Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment* (South Hadley, MA, 2001), vol. 295 of *Contemp. Math.*, pp. 13–24. Amer. Math. Soc. (Providence, RI, 2002)
3. C. Alboin, J. Jaffré, J.E. Roberts, X. Wang, C. Serres, Domain decomposition for some transmission problems in flow in porous media, in *Numerical Treatment of Multiphase Flows in Porous Media*, vol. 552 of *Lecture Notes in Phys.*, (Springer, Berlin, 2000), pp. 22–34
4. P. Angot, F. Boyer, F. Hubert, Asymptotic and numerical modelling of flows in fractured porous media. *M2AN Math. Model. Numer. Anal.* **43**(2), 239–275 (2009)
5. P. Antonietti, M. Verani, C. Vergara, S. Zonca, Numerical solution of fluid-structure interaction problems by means of a high order Discontinuous Galerkin method on polygonal grids. *Finite Elem. Anal. Des.* **159**, 1–14 (2019)
6. P.F. Antonietti, B. Ayuso de Dios, I. Mazzieri, A. Quarteroni, Stability analysis of discontinuous Galerkin approximations to the elastodynamics problem. *J. Sci. Comput.* **68**, 143–170 (2016)
7. P.F. Antonietti, L. Beirão da Veiga, N. Bigoni, M. Verani, Mimetic finite differences for nonlinear and control problems. *Math. Models Methods Appl. Sci.* **24**(8), 1457–1493 (2014)
8. P.F. Antonietti, L. Beirão da Veiga, D. Mora, M. Verani, A stream virtual element formulation of the Stokes problem on polygonal meshes. *SIAM J. Numer. Anal.* **52**(1), 386–404 (2014)
9. P.F. Antonietti, L. Beirão da Veiga, S. Scacchi, M. Verani, A C^1 virtual element method for the Cahn-Hilliard equation with polygonal meshes. *SIAM J. Numer. Anal.* **54**(1), 34–56 (2016)
10. P.F. Antonietti, F. Bonaldi, I. Mazzieri, Simulation of three-dimensional elastoacoustic wave propagation based on a discontinuous Galerkin spectral element method. *Internat. J. Numer. Methods Engrg.* **121**(10), 2206–2226 (2020)
11. P.F. Antonietti, F. Bonaldi, I. Mazzieri, A high-order discontinuous Galerkin approach to the elasto-acoustic problem. *Comput. Methods Appl. Mech. Engrg.* **358**(29), 112634 (2020)
12. P.F. Antonietti, F. Brezzi, L.D. Marini, Bubble stabilization of discontinuous Galerkin methods. *Comput. Methods Appl. Mech. Engrg.* **198**(21–26), 1651–1659 (2009)
13. P.F. Antonietti, A. Cangiani, J. Collis, Z. Dong, E.H. Georgoulis, S. Giani, P. Houston, Review of Discontinuous Galerkin finite element methods for partial differential equations on complicated domains. *Lect. Notes Comput. Sci. Eng.* **114**, 281–310 (2015)
14. P.F. Antonietti, C. Facciola, A. Russo, M. Verani, Discontinuous Galerkin approximation of flows in fractured porous media on polytopic grids. *SIAM J. Sci. Comput.* **41**(1), A109–A138 (2019)

15. P.F. Antonietti, C. Facciola, M. Verani, Polytopic discontinuous Galerkin approximation of flows in porous media with networks of fractures. MOX Report 8/2020. Submitted (2020)
16. P.F. Antonietti, C. Facciola, M. Verani, Unified formulation for polytopic discontinuous Galerkin approximation of flows in fractured porous media. *Math. Eng.* **2**(1), 340–385 (2020)
17. P.F. Antonietti, A. Ferroni, I. Mazziere, R. Paolucci, A. Quarteroni, C. Smerzini, M. Stupazzini, Numerical modeling of seismic waves by discontinuous spectral element methods, *43-ème Congrès National d'Analyse Numérique, CANUM2016, volume 61 of ESAIM Proc, Surveys (EDP Sci, Les Ulis, 2018)*, pp. 1–37
18. P.F. Antonietti, L. Formaggia, A. Scotti, M. Verani, N. Verzotti, Mimetic finite difference approximation of flows in fractured porous media. *M2AN Math. Model. Numer. Anal.* **50**(3), 809–832 (2016)
19. P.F. Antonietti, S. Giani, P. Houston, *hp*-version composite discontinuous Galerkin methods for elliptic problems on complicated domains. *SIAM J. Sci. Comput.* **35**(3), A1417–A1439 (2013)
20. P.F. Antonietti, S. Giani, P. Houston, Domain decomposition preconditioners for discontinuous Galerkin methods for elliptic problems on complicated domains. *J. Sci. Comput.* **60**(1), 203–227 (2014)
21. P.F. Antonietti, P. Houston, X. Hu, M. Sarti, M. Verani, Multigrid algorithms for *hp*-version interior penalty discontinuous Galerkin methods on polygonal and polyhedral meshes. *Calcolo* **54**(4), 1169–1198 (2017)
22. P.F. Antonietti, P. Houston, G. Pennesi, Fast numerical integration on polytopic meshes with applications to discontinuous Galerkin finite element methods. *J. Sci. Comput.* **77**, 1339–1370 (2018)
23. P.F. Antonietti, P. Houston, G. Pennesi, E. Süli, An agglomeration-based massively parallel non-overlapping additive Schwarz preconditioner for high-order discontinuous Galerkin methods on polytopic grids. *Math. Comp.* (2020). <https://doi.org/10.1090/mcom/3510>
24. P.F. Antonietti, G. Manzini, M. Verani, The fully nonconforming Virtual Element Method for biharmonic problems. *Math. Models Methods Appl. Sci.* **28**(02), 387–407. *M3AS Math (Models Methods Appl. Sci. 2018)*
25. P.F. Antonietti, C. Marcati, I. Mazziere, A. Quarteroni, High order discontinuous Galerkin methods on simplicial elements for the elastodynamics equation. *Numer. Algorithms* **71**(1), 181–206 (2016)
26. P.F. Antonietti, I. Mazziere, High-order Discontinuous Galerkin methods for the elastodynamics equation on polygonal and polyhedral meshes. *Comput. Methods Appl. Mech. Engrg.* **342**, 414–437 (2018)
27. P.F. Antonietti, I. Mazziere, N. Dal Santo, A. Quarteroni, A high-order discontinuous Galerkin approximation to ordinary differential equations with applications to elastodynamics. *IMA J. Numer. Anal.* **38**(4), 1709–1734 (2018)
28. P.F. Antonietti, I. Mazziere, M. Muhr, V. Nikolić, B. Wohlmuth, A high-order discontinuous Galerkin method for nonlinear sound waves. *J. Comput. Phys* **415**, (2020)
29. P.F. Antonietti, I. Mazziere, A. Quarteroni, F. Rapetti, Non-conforming high order approximations of the elastodynamics equation. *Comput. Methods Appl. Mech. Engrg.* **209**, 212–238 (2012)
30. P.F. Antonietti, G. Pennesi, V-cycle multigrid algorithms for discontinuous Galerkin methods on non-nested polytopic meshes. *J. Sci. Comput.* **78**(1), 625–652 (2019)
31. P.F. Antonietti, M. Sarti, M. Verani, Multigrid algorithms for *hp*-discontinuous Galerkin discretizations of elliptic problems. *SIAM J. Numer. Anal.* **53**(1), 598–618 (2015)
32. D.N. Arnold, An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.* **19**(4), 742–760 (1982)
33. D.N. Arnold, F. Brezzi, B. Cockburn, L.D. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.* **39**(5), 1749–1779 (2001/02)
34. D.N. Arnold, F. Brezzi, R.S. Falk, L.D. Marini, Locking-free Reissner-Mindlin elements without reduced integration. *Comput. Methods Appl. Mech. Engrg.* **196**(37–40), 3660–3671 (2007)

35. B. Ayuso de Dios, K. Lipnikov, G. Manzini, The nonconforming virtual element method. *ESAIM Math. Model. Numer. Anal.* **50**(3), 879–904 (2016)
36. F. Bassi, L. Botti, A. Colombo, Agglomeration-based physical frame dG discretizations: an attempt to be mesh free. *Math. Models Methods Appl. Sci.* **24**(8), 1495–1539 (2014)
37. F. Bassi, L. Botti, A. Colombo, D.A. Di Pietro, P. Tesini, On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *J. Comput. Phys.* **231**(1), 45–65 (2012)
38. F. Bassi, L. Botti, A. Colombo, S. Rebay, Agglomeration based discontinuous Galerkin discretization of the Euler and Navier-Stokes equations. *Comput. Fluids* **61**, 77–85 (2012)
39. L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D. Marini, A. Russo, Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.* **23**(1), 199–214 (2013)
40. L. Beirão da Veiga, F. Brezzi, F. Dassi, L.D. Marini, A. Russo, A family of three-dimensional virtual elements with applications to magnetostatics. *SIAM J. Numer. Anal.* **56**(5), 2940–2962 (2018)
41. L. Beirão da Veiga, F. Dassi, A. Russo, High-order virtual element method on polyhedral meshes. *Comput. Math. Appl.* **74**(5), 1110–1122 (2017)
42. L. Beirão da Veiga, K. Lipnikov, G. Manzini, *The Mimetic Finite Difference method for elliptic problems*, vol. 11 (Springer, Cham, 2014)
43. L. Beirão da Veiga, D. Mora, G. Rivera, Virtual elements for a shear-deflection formulation of Reissner-Mindlin plates. *Math. Comp.* **88**(315), 149–178 (2019)
44. L. Beirão da Veiga, D. Mora, G. Vacca, The Stokes complex for virtual elements with application to Navier-Stokes flows. *J. Sci. Comput.* **81**(2), 990–1018 (2019)
45. L. Beirão da Veiga, A. Russo, G. Vacca, The virtual element method with curved edges. *ESAIM Math. Model. Numer. Anal.* **53**(2), 375–404 (2019)
46. L. Beirão da Veiga, F. Brezzi, L. Marini, A. Russo, Virtual element method for general second-order elliptic problems on polygonal meshes. *Math. Models Methods Appl. Sci.* **26**(04), 729–750 (2016)
47. M.F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, S. Scialò, A hybrid mortar virtual element method for discrete fracture network simulations. *J. Comput. Phys.* **306**, 148–166 (2016)
48. M.F. Benedetto, S. Berrone, S. Pieraccini, S. Scialò, The virtual element method for discrete fracture network simulations. *Comput. Methods Appl. Mech. Engrg.* **280**, 135–156 (2014)
49. M.F. Benedetto, S. Berrone, S. Scialò, A globally conforming method for solving flow in discrete fracture networks using the virtual element method. *Finite Elem. Anal. Des.* **109**, 23–36 (2016)
50. A. Bermúdez, P. Gamallo, L. Hervella-Nieto, R. Rodríguez, Finite element analysis of pressure formulation of the elastoacoustic problem. *Numer. Math.* **95**, 29–51 (2003)
51. A. Bermúdez, L. Hervella-Nieto, R. Rodríguez, Finite element computation of three-dimensional elastoacoustic vibrations. *J. Sound Vib.* **219**, 279–306 (1999)
52. J. Bielak, O. Ghattas, E. Kim, Parallel octree-based finite element method for large-scale earthquake ground motion simulation. *CMES–Comput. Model. Eng. Sci.* **10**, 99–112 (2005)
53. L. Botti, D.A. Di Pietro, Assessment of hybrid high-order methods on curved meshes and comparison with discontinuous Galerkin methods. *J. Comput. Phys.* **370**, 58–84 (2018)
54. L. Botti, D.A. Di Pietro, J. Droniou, A hybrid high-order method for the incompressible Navier-Stokes equations based on Temam’s device. *J. Comput. Phys.* **376**, 786–816 (2019)
55. M. Botti, D.A. Di Pietro, P. Sochala, A hybrid high-order method for nonlinear elasticity. *SIAM J. Numer. Anal.* **55**(6), 2687–2717 (2017)
56. K. Brenner, J. Hennicker, R. Masson, P. Samier, Gradient discretization of hybrid-dimensional Darcy flow in fractured porous media with discontinuous pressures at matrix-fracture interfaces. *IMA J. Numer. Anal.* **37**(3), 1551–1585 (2016)
57. F. Brezzi, T.J. Hughes, L.D. Marini, A. Masud, Mixed discontinuous Galerkin methods for Darcy flow. *J. Sci. Comput.* **22**(1–3), 119–145 (2005)
58. F. Brezzi, K. Lipnikov, M. Shashkov, Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.* **43**(5), 1872–1896 (2005) (electronic)

59. F. Brezzi, K. Lipnikov, 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)
60. F. Brezzi, K. Lipnikov, V. Simoncini, A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.* **15**(10), 1533–1551 (2005)
61. E. Burman, P. Hansbo, M.G. Larson, K. Larsson, Cut finite elements for convection in fractured domains. *Comput. Fluids* **179**, 726–734 (2019)
62. A. Cangiani, Z. Dong, E. Georgoulis, *hp*-version discontinuous Galerkin methods on essentially arbitrarily-shaped elements. Submitted for publication (2019)
63. A. Cangiani, Z. Dong, E.H. Georgoulis, *hp*-version space-time discontinuous Galerkin methods for parabolic problems on prismatic meshes. *SIAM J. Sci. Comput.* **39**(4), A1251–A1279 (2017)
64. A. Cangiani, Z. Dong, E.H. Georgoulis, P. Houston, *hp*-version discontinuous Galerkin methods for advection-diffusion-reaction problems on polytopic meshes. *ESAIM Math. Model. Numer. Anal.* **50**(3), 699–725 (2016)
65. A. Cangiani, Z. Dong, E.H. Georgoulis, P. Houston, *hp-Version Discontinuous Galerkin Methods on Polygonal and Polyhedral Meshes* (SpringerBriefs in Mathematics, 2017)
66. A. Cangiani, Z. Dong, E.H. Georgoulis, P. Houston, *hp-Version Discontinuous Galerkin Methods on Polytopic Meshes* (SpringerBriefs in Mathematics, Springer International Publishing, 2017)
67. A. Cangiani, E.H. Georgoulis, P. Houston, *hp*-version discontinuous Galerkin methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.* **24**(10), 2009–2041 (2014)
68. A. Cangiani, G. Manzini, O.J. Sutton, Conforming and nonconforming virtual element methods for elliptic problems. *IMA J. Numer. Anal.* **37**(3), 1317–1354 (2017)
69. P. Castillo, B. Cockburn, I. Perugia, D. Schötzau, An a priori error analysis of the local discontinuous Galerkin method for elliptic problems. *SIAM J. Numer. Anal.* **38**(5), 1676–1706 (2000)
70. E. Chaljub, Y. Capdeville, J.P. Vilotte, Solving elastodynamics in a fluid-solid heterogeneous sphere: a parallel spectral element approximation on non-conforming grids. *J. Comput. Phys.* **187**, 457–491 (2003)
71. F. Chave, D.A. Di Pietro, L. Formaggia, A hybrid high-order method for Darcy flows in fractured porous media. *SIAM J. Sci. Comput.* **40**(2), A1063–A1094 (2018)
72. F. Chave, D.A. Di Pietro, F. Marche, F. Pigeonneau, A hybrid high-order method for the Cahn-Hilliard problem in mixed form. *SIAM J. Numer. Anal.* **54**(3), 1873–1898 (2016)
73. F.A. Chave, D. Di Pietro, L. Formaggia, A hybrid high-order method for Darcy flows in fractured porous media. *SIAM J. Sci. Comput.* **40**(2), A1063–A1094 (2018)
74. E.B. Chin, J.B. Lasserre, N. Sukumar, Numerical integration of homogeneous functions on convex and nonconvex polygons and polyhedra. *Comput. Mech.* **56**(6), 967–981 (2015)
75. B. Cockburn, B. Dond, J. Guzmán, A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems. *Math. Comp.* **77**(264), 1887–1916 (2008)
76. B. Cockburn, J. Gopalakrishnan, R. Lazarov, Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.* **47**(2), 1319–1365 (2009)
77. B. Cockburn, J. Gopalakrishnan, F.-J. Sayas, A projection-based error analysis of HDG methods. *Math. Comp.* **79**(271), 1351–1367 (2010)
78. B. Cockburn, J. Guzmán, H. Wang, Superconvergent discontinuous Galerkin methods for second-order elliptic problems. *Math. Comp.* **78**(265), 1–24 (2009)
79. B. Cockburn, C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems. *SIAM J. Numer. Anal.* **35**(6), 2440–2463 (1998)
80. G. Cohen, X. Ferrieres, S. Pernet, A spatial high-order hexahedral discontinuous Galerkin method to solve maxwell’s equations in time domain. *J. Comput. Phys.* **217**(2), 340–363 (2006)

81. C. D'Angelo, A. Scotti, A mixed finite element method for Darcy flow in fractured porous media with non-matching grids. *ESAIM Math. Model. Numer. Anal.* **46**(02), 465–489 (2012)
82. J.D. De Basabe, M.K. Sen, A comparison of finite-difference and spectral-element methods for elastic wave propagation in media with a fluid-solid interface. *Geophys. J. Int.* **200**, 278–298 (2015)
83. J.D. De Basabe, M.K. Sen, M.F. Wheeler, The interior penalty discontinuous Galerkin method for elastic wave propagation: grid dispersion. *Geophys. J. Int.* **175**(1), 83–93 (2008)
84. D.A. Di Pietro, J. Droniou, A hybrid high-order method for Leray-Lions elliptic equations on general meshes. *Math. Comp.* **86**(307), 2159–2191 (2017)
85. D.A. Di Pietro, J. Droniou, $W^{s,p}$ -approximation properties of elliptic projectors on polynomial spaces, with application to the error analysis of a hybrid high-order discretisation of Leray-Lions problems. *Math. Models Methods Appl. Sci.* **27**(5), 879–908 (2017)
86. D.A. Di Pietro, A. Ern, Hybrid high-order methods for variable-diffusion problems on general meshes. *C. R. Math. Acad. Sci. Paris* **353**(1), 31–34 (2015)
87. D.A. Di Pietro, S. Krell, A hybrid high-order method for the steady incompressible Navier-Stokes problem. *J. Sci. Comput.* **74**(3), 1677–1705 (2018)
88. J. Dolbow, N. Moes, T. Belytschko, An extended finite element method for modeling crack growth with frictional contact. *Comput. Methods Appl. Mech. Engrg.* **190**(51–52), 6825–6846 (2001)
89. V. Dolean, H. Fol, S. Lanteri, R. Perrussel, Solution of the time-harmonic maxwell equations using discontinuous Galerkin methods. *J. Comput. Appl. Math.* **218**(2), 435–445 (2008)
90. J. Droniou, R. Eymard, R. Herbin, Gradient schemes: generic tools for the numerical analysis of diffusion equations. *ESAIM Math. Model. Numer. Anal.* **50**(3), 749–781 (2016)
91. M. Dumbser, M. Käser, An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes – II. The three-dimensional isotropic case. *Geophys. J. Int.* **167**(1), 319–336 (2006)
92. A. Ferroni, P. Antonietti, I. Mazzieri, A. Quarteroni, Dispersion-dissipation analysis of 3-d continuous and discontinuous spectral element methods for the elastodynamics equation. *Geophys. J. Int.* **211**(3), 1554–1574 (2017)
93. A. Fichtner, H. Igel, H.-P. Bunge, B.L.N. Kennett, Simulation and inversion of seismic wave propagation on continental scales based on a spectral-element method. *JNAIAM J. Numer. Anal. Ind. Appl. Math.* **4**(1–2), 11–22 (2009)
94. B. Flemisch, A. Fumagalli, A. Scotti, A review of the XFEM-based approximation of flow in fractured porous media, in *Advances in Discretization Methods* (Springer, 2016), pp. 47–76
95. L. Formaggia, A. Fumagalli, A. Scotti, P. Ruffo, A reduced model for Darcy's problem in networks of fractures. *ESAIM: Math. Model. Numer. Anal.* **48**(4), 1089–1116 (2014)
96. L. Formaggia, A. Scotti, F. Sottocasa, Analysis of a mimetic finite difference approximation of flows in fractured porous media. *ESAIM Math. Model. Numer. Anal.* **52**(2), 595–630 (2018)
97. T.-P. Fries, T. Belytschko, The extended/generalized finite element method: an overview of the method and its applications. *Internat. J. Numer. Methods Engrg.* **84**(3), 253–304 (2010)
98. N. Frih, J.E. Roberts, A. Saada, Modeling fractures as interfaces: a model for Forchheimer fractures. *Comput. Geosci.* **12**(1), 91–104 (2008)
99. A. Fumagalli, A. Scotti, A numerical method for two-phase flow in fractured porous media with non-matching grids. *Adv. Water Resour.* **62**(Part C), 454–464 (2013)
100. P. Galvez, J.-P. Ampuero, L.A. Dalguer, S.N. Somala, T. Nissen-Meyer, Dynamic earthquake rupture modelled with an unstructured 3-D spectral element method applied to the 2011 M9 Tohoku earthquake. *Geophys. J. Int.* **198**(2), 1222–1240 (2014)
101. A. Gerstenberger, A.W. Wall, An extended finite element method/Lagrange multiplier based approach for fluid-structure interaction. *Comput. Methods Appl. Mech. Engrg.* **197**(19–20), 1699–1714 (2008)
102. F.X. Giraldo, T. Warburton, A high-order triangular discontinuous Galerkin oceanic shallow water model. *Internat. J. Numer. Methods Fluids* **7**, 899–925 (2008)
103. W. Hackbusch, S.A. Sauter, Composite finite elements for problems containing small geometric details. Part II: implementation and numerical results. *Comput. Visual Sci.* **1**(4), 15–25 (1997)

104. W. Hackbusch, S.A. Sauter, Composite finite elements for the approximation of PDEs on domains with complicated micro-structures. *Numer. Math.* **75**(4), 447–472 (1997)
105. D.J. Holdych, D.R. Noble, R.B. Secor, Quadrature rules for triangular and tetrahedral elements with generalized functions. *Internat. J. Numer. Methods Engrg.* **73**(9), 1310–1327 (2015)
106. J. Hyman, M. Shashkov, S. Steinberg, The numerical solution of diffusion problems in strongly heterogeneous non-isotropic materials. *J. Comput. Phys.* **132**(1), 130–148 (1997)
107. J. Jaffré, M. Mnejjja, J. Roberts, A discrete fracture model for two-phase flow with matrix-fracture interaction. *Procedia Comput. Sci.* **4**, 967–973 (2011)
108. G. Karypis, V. Kumar, A fast and high quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.* **20**(1), 359–392 (1998)
109. G. Karypis, V. Kumar, Metis: unstructured graph partitioning and sparse matrix ordering system, version 4.0 (2009). <http://www.cs.umn.edu/~metis>
110. M. Käser, M. Dumbser, An arbitrary high-order discontinuous Galerkin method for elastic waves on unstructured meshes - I: the two-dimensional isotropic case with external source terms. *Geophys. J. Int.* **166**(2), 855–877 (2006)
111. M. Käser, M. Dumbser, A highly accurate discontinuous Galerkin method for complex interfaces between solids and moving fluids. *Geophysics* **73**, T23–T35 (2008)
112. D. Komatitsch, J. Tromp, Spectral-element simulations of global seismic wave propagation-I. Validation. *Geophys. J. Int.* **149**(2), 390–412 (2002)
113. D. Komatitsch, J. Tromp, Spectral-element simulations of global seismic wave propagation-II. Three-dimensional models, oceans, rotation and self-gravitation. *Geophys. J. Int.* **150**(1), 303–318 (2002)
114. J.B. Lasserre, Integration on a convex polytope. *Proc. Amer. Math. Soc.* **126**(8), 2433–2441 (1998)
115. C.-J. Li, P. Lambertu, C. Dagnino, Numerical integration over polygons using an eight-node quadrilateral spline finite element. *J. Comput. Appl. Math.* **233**(2), 279–292 (2009)
116. J.N. Lyness, G. Monegato, Quadrature rules for regions having regular hexagonal symmetry. *SIAM J. Numer. Anal.* **14**(2), 283–295 (1977)
117. J. Ma, V. Rokhlin, S. Wandzura, Generalized Gaussian quadrature of systems of arbitrary functions. *SIAM J. Numer. Anal.* **33**(3), 971–996 (1996)
118. V. Martin, J. Jaffré, J.E. Roberts, Modeling fractures and barriers as interfaces for flow in porous media. *SIAM J. Sci. Comput.* **26**(5), 1667–1691 (2005)
119. A. Masud, T.J. Hughes, A stabilized mixed finite element method for Darcy flow. *Comput. Methods Appl. Mech. Engrg.* **191**(39–40), 4341–4370 (2002)
120. I. Mazzieri, M. Stupazzini, R. Guidotti, C. Smerzini, SPEED: spectral elements in elastodynamics with discontinuous Galerkin: a non-conforming approach for 3D multi-scale problems. *Internat. J. Numer. Methods Engrg.* **95**(12), 991–1010 (2013)
121. E.D. Mercerat, N. Glinsky, A nodal high-order discontinuous Galerkin method for elastic wave propagation in arbitrary heterogeneous media. *Geophys. J. Int.* **201**(2), 1101–1118 (2015)
122. N. Moes, J. Dolbow, T. Belytschko, A finite element method for crack growth without remeshing. *Int. J. Numer. Meth. Eng.* **46**(1), 131–150 (1999)
123. S.E. Mousavi, N. Sukumar, Numerical integration of polynomials and discontinuous functions on irregular convex polygons and polyhedrons. *Comput. Mech.* **47**(5), 535–554 (2011)
124. S.E. Mousavi, H. Xiao, N. Sukumar, Generalized Gaussian quadrature rules on arbitrary polygons. *Internat. J. Numer. Methods Engrg.* **82**(1), 99–113 (2010)
125. W.A. Mulder, E. Zhebel, S. Minisini, Time-stepping stability of continuous and discontinuous finite-element methods for 3-D wave propagation. *Geophys. J. Int.* **196**(2), 1123–1133 (2014)
126. S. Natarajan, S. Bordas, D.R. Mahapatra, Numerical integration over arbitrary polygonal domains based on Schwarz-Christoffel conformal mapping. *Internat. J. Numer. Methods Engrg.* **80**(1), 103–134 (2009)
127. I. Perugia, D. Schötzau, An *hp*-analysis of the local discontinuous Galerkin method for diffusion problems. *J. Sci. Comput.* **17**(1), 561–571 (2002)
128. I. Perugia, D. Schötzau, The *hp*-local discontinuous Galerkin method for low-frequency time-harmonic maxwell equations. *Math. Comp.* **72**(243), 1179–1214 (2003)

129. A. Quarteroni, *Numerical Models for Differential Problems*, vol. 2, (Springer Science & Business Media, 2014)
130. P.-A. Raviart, J.-M. Thomas, *Introduction à l'analyse numérique des équations aux dérivées partielles* (Masson, 1983)
131. B. Rivière, S. Shaw, M.F. Wheeler, J.R. Whiteman, Discontinuous Galerkin finite element methods for linear elasticity and quasistatic linear viscoelasticity. *Numer. Math.* **95**(2), 347–376 (2003)
132. B. Rivière, M.F. Wheeler, Discontinuous finite element methods for acoustic and elastic wave problems. *Contemp. Math.* **329**, 271–282 (2003)
133. N. Schwenck, B. Flemisch, R. Helmig, B.I. Wohlmuth, Dimensionally reduced flow models in fractured porous media: crossings and boundaries. *Comput. Geosci.* **19**(6), 1219–1230 (2015)
134. C.P. Simon, L.E. Blume, *Mathematics for Economists* (W. W. Norton and Company, New York, 1996)
135. A. Sommariva, M. Vianello, Product Gauss cubature over polygons based on Green's integration formula. *BIT* **47**(2), 441–453 (2007)
136. E. Stein, *Singular Integrals and Differentiability Properties of Functions* (Princeton University Press, Princeton, N.J., 1970)
137. A.H. Stroud, D. Secrest, Gaussian quadrature formulas. *ZAMM Z. Angew. Math. Mech.* **47**(2), 138–139 (1967)
138. Y. Sudhakar, W.A. Wall, Quadrature schemes for arbitrary convex/concave volumes and integration of weak form in enriched partition of unity methods. *Comput. Methods Appl. Mech. Engrg.* **258**(1), 39–54 (2013)
139. N. Sukumar, N. Moes, T. Belytschko, Extended finite element method for three-dimensional crack modelling. *Internat. J. Numer. Methods Engrg.* **48**(11), 1549–1570 (2000)
140. N. Sukumar, A. Tabarraei, Conforming polygonal finite elements. *Internat. J. Numer. Methods Engrg.* **61**(12), 2045–2066 (2004)
141. A. Tabarraei, N. Sukumar, Extended finite element method on polygonal and quadtree meshes. *Comput. Methods Appl. Mech. Engrg.* **197**(5), 425–438 (2008)
142. C. Talischi, G.H. Paulino, A. Pereira, I.F. Menezes, Polymesher: a general-purpose mesh generator for polygonal elements written in matlab. *Struct. Multidiscip. Optim.* **45**(3), 309–328 (2012)
143. M.E. Taylor, *Partial Differential Equations: Basic Theory* (Springer, New York, 1996)
144. S. Terrana, J.P. Vilotte, L. Guillot, A spectral hybridizable discontinuous Galerkin method for elastic-acoustic wave propagation. *Geophys. J. Int.* **213**, 574–602 (2018)
145. G. Ventura, On the elimination of quadrature subcells for discontinuous functions in the extended finite-element method. *Internat. J. Numer. Methods Engrg.* **66**(5), 761–795 (2006)
146. G. Ventura, E. Benvenuti, Equivalent polynomials for quadrature in Heaviside function enriched elements. *Internat. J. Numer. Methods Engrg.* **102**(3–4), 688–710 (2015)
147. T. Warburton, J.S. Hesthaven, On the constants in hp -finite element trace inverse inequalities. *Comput. Methods Appl. Mech. Engrg.* **192**(25), 2765–2773 (2003)
148. M.F. Wheeler, An elliptic collocation-finite element method with interior penalties. *SIAM J. Numer. Anal.* **15**(1), 152–161 (1978)
149. L.C. Wilcox, G. Stadler, C. Burstedde, O. Ghattas, A high-order discontinuous Galerkin method for wave propagation through coupled elastic-acoustic media. *J. Comput. Phys.* **229**, 9373–9396 (2010)
150. N. Yarvin, V. Rokhlin, Generalized Gaussian quadratures and singular value decompositions of integral operators. *SIAM J. Sci. Comput.* **20**(2), 669–718 (1998)
151. A.M. Yogitha, K.T. Shivaram, Numerical integration of arbitrary functions over a convex and non convex polygonal domain by eight noded linear quadrilateral finite element method. *Aust. J. Basic Appl. Sci.* **10**(16), 104–110 (2016)

Chapter 6

A Hybrid High-Order Method for Multiple-Network Poroelasticity



Lorenzo Botti, Michele Botti, and Daniele A. Di Pietro

Abstract We develop Hybrid High-Order methods for multiple-network poroelasticity, modelling seepage through deformable fissured porous media. The proposed methods are designed to support general polygonal and polyhedral elements. This is a crucial feature in geological modelling, where the need for general elements arises, e.g., due to the presence of fracture and faults, to the onset of degenerate elements to account for compaction or erosion, or when nonconforming mesh adaptation is performed. We use as a starting point a mixed weak formulation where an additional total pressure variable is added, that ensures the fulfilment of a discrete inf-sup condition. A complete theoretical analysis is performed, and the results are demonstrated on a panel of numerical tests.

Keywords Hybrid High-Order methods · Discontinuous Galerkin methods · Polytopal methods · Multi-network poroelasticity · Barenblatt-Biot equations

L. Botti

Department of Engineering and Applied Sciences, University of Bergamo, Bergamo, Italy
e-mail: lorenzo.botti@unibg.it

M. Botti

MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica,
Politecnico di Milano, Milan, Italy
e-mail: michele.botti@polimi.it

D. A. Di Pietro (✉)

IMAG, University of Montpellier, CNRS, Montpellier, France
e-mail: daniele.di-pietro@umontpellier.fr

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2021

D. A. Di Pietro et al. (eds.), *Polyhedral Methods in Geosciences*,

SEMA SIMAI Springer Series 27,

https://doi.org/10.1007/978-3-030-69363-3_6

6.1 Introduction

In this work, we develop and analyse Hybrid High-Order (HHO) methods for the multiple-network poroelastic problem.

In the standard quasi-static poroelasticity theory [18], the medium is modelled as a continuous superposition of solid and fluid phases. The corresponding set of equations, named after Biot in recognition of his pioneering contributions [7, 8], result from the balances of force and mass. Specifically, mechanical equilibrium is assumed, with the total stress tensor decomposed into one contribution due to the strain of the porous matrix and one due to the pore pressure; see [32]. A standard description of the flow, on the other hand, is obtained combining the mass balance with the Darcy law. This simplified description can fail to capture physically relevant phenomena in fissured media. A modification of the Darcy model accounting for the simultaneous presence of pore and fissure networks was originally proposed by Barenblatt et al. in [4] for the rigid case. Plugging this description into the Biot model gives rise to the so-called Barenblatt–Biot equations. These ideas can be naturally extended to M porous networks, finding applications, e.g., in the modelling of the interactions between biological fluids and tissue; see, e.g. [33]. A different extension of the Biot model is considered in Chap. 4, where thermal effects are incorporated into a single network model.

In the context of computational geosciences, the use of discretisation methods that support general polytopal meshes and, possibly, high-order has been recently advocated by several authors; see, e.g., [2, 3, 6, 15–17, 27, 31] and references therein. The support of polyhedral meshes enables, e.g., a seamless treatment of degenerate elements which may arise due to erosion or compaction in corner-point descriptions of petroleum basins, of non-matching interfaces across fractures or faults, and of non-conforming mesh refinement or agglomeration [5]. High-order methods, on the other hand, typically lead to a better usage of computational resources than low-order methods whenever the solution exhibits sufficient (local) regularity or mesh adaptation is available.

Our focus is here on a specific family of polytopal discretisations, HHO methods. Originally introduced in [23] in the context of linear elasticity, HHO methods rely on two key ingredients: *local reconstructions* obtained by solving small, embarrassingly parallel problems inside each element and *stabilisation terms* that penalise, inside each element, residuals designed so as to preserve optimal approximation properties. A general and up-to-date overview of HHO methods can be found in the recent monograph [22]. Hybrid High-Order methods are linked to the hybridized version of the Mixed Virtual Element methods considered in Chaps. 7 and 8; see [1, 24] and also [22, Sects. 5.4 and 5.5]. Concerning their application to poroelasticity, we can cite, in particular: the HHO-Discontinuous Galerkin method for the Biot problem proposed and analysed in [9], based in turn on the methods of [23] for the mechanics and [25] for the flow; its extension to nonlinear elastic laws proposed in [14], where the mechanical term is discretised according to [13]; its application to the treatment of stochastic coefficients considered in [12] in conjunction with Polynomial Chaos

techniques. An abstract analysis framework covering general schemes for the linear Biot problem in fully discrete formulation (cf. [20]) has been recently proposed in [10] including, in particular, a variation of the method of [9] where also the flow equation is discretised in the HHO spirit. Other applications of HHO methods to problems in geosciences include flows in fractured porous media [16, 17] and miscible fluid flows in porous media [2].

The method proposed in the present work uses as a starting point a mixed formulation inspired by [30], where an additional total pressure variable is introduced that accounts for the pore and mechanical pressures. Given an integer polynomial degree $k \geq 0$, the discretisation of the mechanical term in the equilibrium equation follows [13] if $k \geq 1$ and [12] if $k = 0$. This choice induces a natural discretisation for the total pressure in the space of broken polynomials of total degree $\leq k$, which ensures inf-sup stability. As it has been done in [10], we consider two different discretisations of the Darcy term in the mass balance equations (enforcing mass conservation in each pore network). The first scheme is based on the HHO method of [26], so the discrete unknowns for the pore pressures are located both at elements and faces. The second scheme is obtained by using the Discontinuous Galerkin (DG) method of [25]. In both cases, the linear exchange terms as well as the porosity are discretised using element unknowns only. The resulting methods have several appealing features: they support general polytopal meshes and high-order; they can be applied to an arbitrary number $M \geq 1$ of pore networks; they are well-behaved for quasi-incompressible porous matrices; they deliver an L^2 -error estimate for the total pressure robust in the entire range of geophysical parameters.

From the practical standpoint, a relevant difference between the two schemes is that the HHO-HHO version can benefit from static condensation, leading to linear systems where the only globally coupled unknowns are displacement and pore pressure at faces, and global pressures at elements. On typical meshes, this results in fewer unknowns compared to the HHO-DG scheme and better computational efficiency, particularly in three space dimensions; see, e.g., the numerical tests on meshes with planar faces in [11]. On the other hand, the HHO-DG scheme may be easier to implement, as it does not require the introduction of pore pressures at faces, nor the computation of local pore pressure reconstructions or static condensation. From the theoretical point of view, the analysis of the HHO-DG scheme requires elliptic regularity (in Theorem 6.2, the convexity of the domain is assumed) to achieve optimal orders of convergence. As pointed out in [10], this is not the case for the HHO-HHO scheme. In this paper, we focus on the HHO-DG scheme for the numerical tests of Sect. 6.5, and postpone a comparison with the HHO-HHO scheme to a future work.

The rest of this paper is organised as follows. In Sect. 6.2 we establish the continuous setting and state the multiple-network poroelasticity problem in weak formulation. Section 6.3 describes the discrete setting and contains the statement of the discrete problem. The analysis of the method is carried out in Sect. 6.4 focusing, for the sake of simplicity, on the HHO-HHO variant. The pivotal result is here an a priori estimate for an abstract problem whose purpose is twofold: when applied to

the HHO scheme, it yields its well-posedness; when applied to the error equations, it establishes a basic error estimate. Finally, Sect. 6.5 contains a thorough numerical validation of the method.

6.2 Continuous Setting

In what follows, given an open bounded set $X \subset \mathbb{R}^d$, we denote by $(\cdot, \cdot)_X$ the usual scalar product of $L^2(X; \mathbb{R})$, $L^2(X; \mathbb{R}^d)$, or $L^2(X; \mathbb{R}^{d \times d})$, according to the context. When $X = \Omega$, the subscript is omitted. Given a vector space V and two real numbers $a < b$, we additionally denote by $C^0([a, b]; V)$ the spaces of continuous V -valued functions of time on $[a, b]$ and by $H^m(a, b; V)$ the space of V -valued functions that are square-integrable along with their derivatives up to the m -th on (a, b) , equipped with the usual norms.

We consider the evolution over a finite time $t_F > 0$ of a porous medium which, in its reference configuration, occupies a fixed region of space $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, and hosts $M \geq 1$ pore networks. For the sake of simplicity, we assume that Ω is a polygon or a polyhedron, so that it can be covered exactly by a spatial mesh made of polygonal or polyhedral elements. Denote by $\mu > 0$ and $\lambda \geq 0$ the Lamé parameters of the matrix and, for any $i \in \llbracket 1, M \rrbracket$, by $C_i \geq 0$, $\alpha_i \in (0, 1]$, and $K_i > 0$, respectively, the constrained specific storage, Biot–Willis, and permeability coefficients of each network. We additionally denote by $\mathbf{f} \in H^1(0, t_F; L^2(\Omega; \mathbb{R}^d))$ a volumetric force and, for any $i \in \llbracket 1, M \rrbracket$, by $g_i \in C^0([0, t_F]; L^2(\Omega; \mathbb{R}))$ a source term for the i th pore network. The above physical parameters and forcing terms will be collectively referred to as the *problem data*.

Let $\mathbf{U} := H_0^1(\Omega; \mathbb{R}^d)$, $P_0 := \{q \in L^2(\Omega; \mathbb{R}) : \int_\Omega q = 0\}$, and, for all $i \in \llbracket 1, M \rrbracket$, $P_i := H_0^1(\Omega; \mathbb{R})$. We also set, for the sake of brevity, $\boldsymbol{\alpha} := (1, \alpha_1, \dots, \alpha_M) \in \mathbb{R}^{M+1}$ and, denoting by p_0 the total pressure field and, for any $i \in \llbracket 1, M \rrbracket$, by p_i the pressure field in the i th porous network, $\mathbf{p} := (p_0, p_1, \dots, p_M)$. We consider a weak formulation inspired by (but not coincident with) the one considered in [30]: Find the displacement $\mathbf{u} \in C^0([0, t_F]; \mathbf{U})$, the total pressure $p_0 \in H^1(0, t_F; P_0)$ and, for all $i \in \llbracket 1, M \rrbracket$, the i th pore network pressure $p_i \in C^0([0, t_F]; P_i) \cap H^1(0, t_F; L^2(\Omega; \mathbb{R}))$ such that it holds, for almost every $t \in (0, t_F]$, all $\mathbf{v} \in \mathbf{U}$, all $q_0 \in P_0$, and all $q_i \in P_i$, $i \in \llbracket 1, M \rrbracket$,

$$2\mu a(\mathbf{u}(t), \mathbf{v}) + b(\mathbf{v}, p_0(t)) = (\mathbf{f}(t), \mathbf{v}) \quad (6.1a)$$

$$b(\mathbf{u}(t), q_0) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{p}, q_0) = 0, \quad (6.1b)$$

$$(d_i \psi_i(\mathbf{p}(t)), q_i) + (S_i(\mathbf{p}(t)), q_i) + K_i c(p_i, q_i) = (g_i(t), q_i) \quad \forall i \in \llbracket 1, M \rrbracket, \quad (6.1c)$$

where we have set, for all $i \in \llbracket 1, M \rrbracket$ and all $\mathbf{q} \in \mathbb{R}^{M+1}$,

$$\psi_i(\mathbf{q}) := C_i q_i + \alpha_i \lambda^{-1} \boldsymbol{\alpha} \cdot \mathbf{q}, \quad (6.2)$$

and we have introduced the bilinear forms $a : \mathbf{U} \times \mathbf{U} \rightarrow \mathbb{R}$, $b : \mathbf{U} \times P_0 \rightarrow \mathbb{R}$, and $c : H^1(\Omega; \mathbb{R}) \times H^1(\Omega; \mathbb{R}) \rightarrow \mathbb{R}$ such that, for all $\mathbf{w}, \mathbf{v} \in \mathbf{U}$, all $q_0 \in P_0$, and all $r, q \in H^1(\Omega; \mathbb{R})$,

$$a(\mathbf{w}, \mathbf{v}) := (\nabla_s \mathbf{w}, \nabla_s \mathbf{v}), \quad b(\mathbf{v}, q_0) := (\nabla \cdot \mathbf{v}, q_0), \quad c(r, q) := (\nabla r, \nabla q). \quad (6.3)$$

In the expression of the bilinear form a , ∇_s denotes the symmetric part of the gradient applied to vector fields. In (6.1b), the exchange term is expressed by the function $S_i : \mathbb{R}^{M+1} \rightarrow \mathbb{R}$ such that, for any $\mathbf{q} \in \mathbb{R}^{M+1}$,

$$S_i(\mathbf{q}) := \sum_{j=1}^M \xi_{i \leftarrow j} (q_i - q_j),$$

where $\{\xi_{i \leftarrow j} : i, j \in \llbracket 1, M \rrbracket\}$ is a family of nonnegative real numbers such that $\xi_{i \leftarrow j} = \xi_{j \leftarrow i}$ for all $i, j \in \llbracket 1, M \rrbracket$. We assume that the initial pressures $p_i^0 \in P_i$, $i \in \llbracket 0, M \rrbracket$, are given, so that an initial equilibrium displacement $\mathbf{u}^0 \in \mathbf{U}$ can be computed from (6.1a).

6.3 Discrete Setting

6.3.1 Space and Time Meshes

We consider spatial meshes corresponding to couples $\mathcal{M}_h := (\mathcal{T}_h, \mathcal{F}_h)$, where \mathcal{T}_h is a finite collection of polyhedral elements such that $h := \max_{T \in \mathcal{T}_h} h_T > 0$ with h_T denoting the diameter of T , while \mathcal{F}_h is a finite collection of planar faces. It is assumed henceforth that the mesh \mathcal{M}_h matches the geometrical requirements detailed in [22, Definition 1.4]. This covers, essentially, any reasonable partition of Ω into polyhedral sets, not necessarily convex.

For every mesh element $T \in \mathcal{T}_h$, we denote by \mathcal{F}_T the subset of \mathcal{F}_h containing the faces that lie on the boundary ∂T of T . For any mesh element $T \in \mathcal{T}_h$ and each face $F \in \mathcal{F}_T$, \mathbf{n}_{TF} is the constant unit vector normal to F pointing out of T . Boundary faces lying on $\partial\Omega$ and internal faces contained in Ω are collected in the sets \mathcal{F}_h^b and \mathcal{F}_h^i , respectively. For any $F \in \mathcal{F}_h^i$, we denote by T_1 and T_2 the elements of \mathcal{T}_h such that $F \subset \partial T_1 \cap \partial T_2$. The numbering of T_1 and T_2 is arbitrary but fixed once and for all, and we set $\mathbf{n}_F := \mathbf{n}_{T_1 F}$.

Our focus being on the h -convergence analysis, we consider a sequence of refined polygonal or polyhedral meshes that is regular in the sense of [22, Definition 1.9]. This implies, in particular, that the diameter h_T of a mesh element $T \in \mathcal{T}_h$ is comparable to the diameter h_F of each face $F \in \mathcal{F}_T$ uniformly in h , and that the number of faces in \mathcal{F}_T is bounded above by an integer N_∂ independent of h ; see [22, Lemma 1.12]. In order to have the stability of the bilinear form discretising the mechanical term

when discrete unknowns are polynomials of degree $k \geq 1$, we will further assume that every element $T \in \mathcal{T}_h$ is star-shaped with respect to every point of a ball of diameter uniformly comparable to h_T . This assumption ensures, in particular, that uniform local Korn inequalities hold inside each element; cf. the Appendix of [11] and also [22, Chap. 7].

The time mesh is obtained subdividing $[0, t_F]$ into $N \in \mathbb{N}^*$ uniform subintervals. We introduce the timestep $\tau := t_F/N$ and the discrete times $t^n := n\tau$, $n \in \llbracket 0, N \rrbracket$.

For all $n \in \llbracket 1, N \rrbracket$ and all $\varphi \in C^0([0, t_F]; V)$ we let, for the sake of brevity,

$$\varphi^n := \varphi(t^n)$$

and define the discrete backward time derivative operator $\delta_t^n : C^0([0, t_F]; V) \rightarrow V$ at time n as

$$\delta_t^n \varphi := \frac{\varphi^n - \varphi^{n-1}}{\tau}. \quad (6.4)$$

Denoting by $(\cdot, \cdot)_V$ an inner product in V with associated norm $\|\cdot\|_V$, and letting $\varphi \in H^1(0, t_F; V)$, it holds

$$\sum_{n=1}^N \tau \|\delta_t^n \varphi\|_V^2 \leq \|\varphi\|_{H^1(0, t_F; V)}^2. \quad (6.5)$$

6.3.2 Local and Broken Spaces and Projectors

Let a polynomial degree $l \geq 0$ be fixed. For all $X \in \mathcal{T}_h \cup \mathcal{F}_h$, denote by $\mathbb{P}^l(X; \mathbb{R})$ the space spanned by the restriction to X of d -variate polynomials of total degree $\leq l$, and let $\pi_X^l : L^1(X; \mathbb{R}) \rightarrow \mathbb{P}^l(X; \mathbb{R})$ be the corresponding L^2 -orthogonal projector such that, for any $v \in L^1(X; \mathbb{R})$,

$$(\pi_X^l v - v, w)_X = 0 \quad \forall w \in \mathbb{P}^l(X; \mathbb{R}).$$

Denoting by $m \geq 1$ an integer, the vector version $\pi_X^l : L^1(X; \mathbb{R}^m) \rightarrow \mathbb{P}^l(X; \mathbb{R}^m)$, is obtained applying π_X^l component-wise. We will also need, in what follows, the space of $d \times d$ symmetric matrix-valued fields with polynomial entries, denoted by $\mathbb{P}^l(T; \mathbb{R}_{\text{sym}}^{d \times d})$.

At the global level, we introduce the broken polynomial space

$$\mathbb{P}^l(\mathcal{T}_h; \mathbb{R}) := \{v \in L^1(\Omega; \mathbb{R}) : v|_T \in \mathbb{P}^l(T; \mathbb{R}) \quad \forall T \in \mathcal{T}_h\},$$

the corresponding vector version $\mathbb{P}^l(\mathcal{T}_h; \mathbb{R}^d)$, and the space $\mathbb{P}^l(\mathcal{T}_h; \mathbb{R}_{\text{sym}}^{d \times d})$ of $d \times d$ symmetric matrix-valued fields with broken polynomial entries. The L^2 -orthogonal projector on $\mathbb{P}^l(\mathcal{T}_h; \mathbb{R})$ is $\pi_h^l : L^1(\Omega; \mathbb{R}) \rightarrow \mathbb{P}^l(\mathcal{T}_h; \mathbb{R})$ such that, for all $v \in L^1(\Omega; \mathbb{R})$,

$$(\pi_h^l v)|_T = \pi_T^l v|_T \quad \forall T \in \mathcal{T}_h. \quad (6.6)$$

Broken polynomial spaces constitute special instances of the broken Sobolev spaces $H^m(\mathcal{T}_h; \mathbb{R}) := \{v \in L^2(\Omega; \mathbb{R}) : v|_T \in H^m(T; \mathbb{R}) \quad \forall T \in \mathcal{T}_h\}$, which will be used, along with their vector-valued counterparts, to express the regularity requirements on the exact solution in the error estimate of Theorems 6.1 and 6.2. For any function $v \in H^1(\mathcal{T}_h; \mathbb{R})$ we define, for all $F \in \mathcal{F}_h^i$, the jump operator such that

$$[v]_F := v|_{T_1} - v|_{T_2},$$

where we remind the reader that T_1 and T_2 are the mesh elements that share F as a face, taken in an arbitrary but fixed order. On boundary faces, the jump operator simply returns the trace of its argument on $\partial\Omega$.

6.3.3 Discrete Spaces and Reconstructions

To formulate the discrete problem, we need scalar and vector HHO spaces. From this point on, we let an integer $k \geq 0$ be fixed, corresponding to the polynomial degrees of the discrete unknowns.

6.3.3.1 Scalar HHO Space and Pressure Reconstruction

The scalar HHO space, that will be used to discretise network pressures in the HHO-HHO scheme (6.23), is

$$\underline{Q}_h^k := \left\{ \underline{q}_h = ((q_T)_{T \in \mathcal{T}_h}, (q_F)_{F \in \mathcal{F}_h}) : \right. \\ \left. q_T \in \mathbb{P}^k(T; \mathbb{R}) \text{ for all } T \in \mathcal{T}_h \text{ and } q_F \in \mathbb{P}^k(F; \mathbb{R}) \text{ for all } F \in \mathcal{F}_h \right\}.$$

The interpolator $\underline{I}_h^k : H^1(\Omega; \mathbb{R}) \rightarrow \underline{Q}_h^k$ is defined setting, for all $q \in H^1(\Omega; \mathbb{R})$,

$$\underline{I}_h^k q := ((\pi_T^k q|_T)_{T \in \mathcal{T}_h}, (\pi_F^k q|_F)_{F \in \mathcal{F}_h}).$$

For all $\underline{q}_h \in \underline{Q}_h^k$, we define the broken polynomial function $q_h \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ obtained patching element unknowns, that is,

$$(q_h)|_T := q_T \quad \forall T \in \mathcal{T}_h.$$

For any element $T \in \mathcal{T}_h$, we denote by \underline{Q}_T^k the restriction of \underline{Q}_h^k to T , and we introduce the pressure reconstruction $\mathbf{r}_T^{k+1} : \underline{Q}_T^k \rightarrow \mathbb{P}^{k+1}(T; \mathbb{R})$ such that, for all $\underline{q}_T \in \underline{Q}_T^k$,

$$(\nabla \mathbf{r}_T^{k+1} \underline{q}_T, \nabla \mathbf{w})_T = -(q_T, \Delta \mathbf{w})_T + \sum_{F \in \mathcal{F}_T} (q_F, \nabla \mathbf{w} \cdot \mathbf{n}_{TF})_F \quad \forall \mathbf{w} \in \mathbb{P}^{k+1}(T; \mathbb{R}),$$

$$\int_T \mathbf{r}_T^{k+1} \underline{q}_T = \int_T q_T.$$

The global pressure reconstruction operator $\mathbf{r}_h^{k+1} : \underline{Q}_h^k \rightarrow \mathbb{P}^{k+1}(\mathcal{T}_h; \mathbb{R})$ is obtained patching the local ones: For all $\underline{q}_h \in \underline{Q}_h^k$,

$$(\mathbf{r}_h^{k+1} \underline{q}_h)|_T := \mathbf{r}_T^{k+1} \underline{q}_T \quad \forall T \in \mathcal{T}_h.$$

6.3.3.2 Vector HHO Space, Strain, and Displacement Reconstructions

The vector HHO space, that will be used to discretise the displacement, is

$$\underline{V}_h^k := \left\{ \underline{\mathbf{v}}_h = ((\mathbf{v}_T)_{T \in \mathcal{T}_h}, (\mathbf{v}_F)_{F \in \mathcal{F}_h}) : \right. \\ \left. \mathbf{v}_T \in \mathbb{P}^k(T; \mathbb{R}^d) \text{ for all } T \in \mathcal{T}_h \text{ and } \mathbf{v}_F \in \mathbb{P}^k(F; \mathbb{R}^d) \text{ for all } F \in \mathcal{F}_h \right\}.$$

For all $\underline{\mathbf{v}}_h \in \underline{V}_h^k$, we let $\mathbf{v}_h \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}^d)$ be such that

$$(\mathbf{v}_h)|_T := \mathbf{v}_T \quad \forall T \in \mathcal{T}_h.$$

The interpolator $\underline{\mathbf{I}}_h^k : H^1(\Omega; \mathbb{R}^d) \rightarrow \underline{V}_h^k$ is such that, for any $\mathbf{v} \in H^1(\Omega; \mathbb{R}^d)$,

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

For any element $T \in \mathcal{T}_h$, we denote by $\underline{\mathbf{V}}_T^k$ the restriction of \underline{V}_h^k to T and we introduce the strain reconstruction $\mathbf{E}_T^k : \underline{V}_T^k \rightarrow \mathbb{P}^k(T; \mathbb{R}_{\text{sym}}^{d \times d})$ such that, for all $\underline{\mathbf{v}}_T \in \underline{V}_T^k$,

$$(\mathbf{E}_T^k \underline{\mathbf{v}}_T, \boldsymbol{\tau})_T = -(\mathbf{v}_T, \nabla \cdot \boldsymbol{\tau})_T + \sum_{F \in \mathcal{F}_T} (\mathbf{v}_F, \boldsymbol{\tau} \mathbf{n}_{TF})_F \quad \forall \boldsymbol{\tau} \in \mathbb{P}^k(T; \mathbb{R}_{\text{sym}}^{d \times d}).$$

For any $\underline{\mathbf{v}}_T \in \underline{V}_T^k$, we reconstruct from $\mathbf{E}_T^k \underline{\mathbf{v}}_T$ a high-order displacement $\mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T \in \mathbb{P}^{k+1}(T; \mathbb{R}^d)$ enforcing the following conditions:

$$(\nabla_s \mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T - \mathbf{E}_T^k \underline{\mathbf{v}}_T, \nabla_s \mathbf{w})_T = 0 \quad \forall \mathbf{w} \in \mathbb{P}^{k+1}(T; \mathbb{R}^d),$$

$$\int_T \mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T = \int_T \mathbf{v}_T, \text{ and } \int_T \nabla_{\text{ss}} \mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T = \frac{1}{2} \sum_{F \in \mathcal{F}_T} \int_F (\mathbf{v}_F \otimes \mathbf{n}_{TF} - \mathbf{n}_{TF} \otimes \mathbf{v}_F),$$

where ∇_{ss} denotes the skew-symmetric part of the gradient applied to vector fields. The global strain and displacement reconstructions $\mathbf{E}_h^k : \underline{V}_h^k \rightarrow \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}_{\text{sym}}^{d \times d})$ and

$\mathbf{r}_h^{k+1} : \underline{\mathbf{V}}_h^k \rightarrow \mathbb{P}^{k+1}(\mathcal{T}_h; \mathbb{R}^d)$ are obtained setting, for all $\mathbf{v}_h \in \underline{\mathbf{V}}_h^k$,

$$(\mathbf{E}_h^k \mathbf{v}_h)|_T := \mathbf{E}_T^k \mathbf{v}_T \text{ and } (\mathbf{r}_h^{k+1} \mathbf{v}_h)|_T := \mathbf{r}_T^{k+1} \mathbf{v}_T \text{ for all } T \in \mathcal{T}_h.$$

We also define a global divergence reconstruction $\mathbf{D}_h^k : \underline{\mathbf{V}}_h^k \rightarrow \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ as the trace of \mathbf{E}_h^k , that is, for all $\mathbf{v}_h \in \underline{\mathbf{V}}_h^k$,

$$\mathbf{D}_h^k \mathbf{v}_h := \text{tr}(\mathbf{E}_h^k \mathbf{v}_h).$$

6.3.3.3 Displacement and Pressure Spaces

The discrete spaces for the displacement including the strongly enforced homogeneous boundary conditions and for the total pressure including the zero-average condition are, respectively:

$$\underline{\mathbf{U}}_h^k := \{ \mathbf{v}_h \in \underline{\mathbf{V}}_h^k : \mathbf{v}_F = \mathbf{0} \text{ for all } F \in \mathcal{F}_h^b \} \text{ and } P_{h,0}^k := \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}) \cap P_0,$$

with P_0 defined in Sect. 6.2. When using the HHO method for the discretisation of the flow equations, for any $i \in \llbracket 1, M \rrbracket$, the space for the i th network pressure is

$$\underline{P}_{h,i}^k := \underline{Q}_{h,D}^k \text{ with } \underline{Q}_{h,D}^k := \left\{ \underline{q}_h \in \underline{Q}_h^k : q_F = 0 \text{ for all } F \in \mathcal{F}_h^b \right\},$$

while, when using the DG method, we use instead

$$P_{h,i}^k := \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}).$$

6.3.4 Discrete Bilinear Forms

We discuss in this section the approximation of the continuous bilinear forms defined in (6.3). In order to alleviate the exposition, from this point on we use the abridged notation $a \lesssim b$ for the inequality $a \leq Cb$ with real number $C > 0$ independent of the meshsize, the time step and, for local inequalities, on the mesh element or face. Further dependencies of the hidden constant will be specified when appropriate.

6.3.4.1 Mechanical Term

The discrete counterpart of the continuous bilinear form a is $\mathbf{a}_h : \underline{\mathbf{V}}_h^k \times \underline{\mathbf{V}}_h^k \rightarrow \mathbb{R}$ such that, for all $\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h \in \underline{\mathbf{V}}_h^k$,

$$a_h(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) := \begin{cases} (\mathbf{E}_h^k \underline{\mathbf{w}}_h, \mathbf{E}_h^k \underline{\mathbf{v}}_h) + s_{a,h}(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) & \text{if } k \geq 1, \\ (\mathbf{E}_h^0 \underline{\mathbf{w}}_h, \mathbf{E}_h^0 \underline{\mathbf{v}}_h) + s_{a,h}(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) + j_h(\mathbf{r}_h^1 \underline{\mathbf{w}}_h, \mathbf{r}_h^1 \underline{\mathbf{v}}_h) & \text{if } k = 0, \end{cases}$$

with stabilising bilinear form $s_{a,h} : \mathbf{V}_h^k \times \mathbf{V}_h^k \rightarrow \mathbb{R}$ and jump penalisation bilinear form $j_h : H^1(\mathcal{T}_h; \mathbb{R}^d) \times H^1(\mathcal{T}_h; \mathbb{R}^d) \rightarrow \mathbb{R}$ such that

$$s_{a,h}(\underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h) := \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} h_F^{-1} (\delta_{TF}^k \underline{\mathbf{w}}_T, \delta_{TF}^k \underline{\mathbf{v}}_T)_F \quad \forall \underline{\mathbf{w}}_h, \underline{\mathbf{v}}_h \in \mathbf{V}_h^k,$$

$$j_h(\underline{\mathbf{w}}, \underline{\mathbf{v}}) := \sum_{F \in \mathcal{F}_h} h_F^{-1} ([\underline{\mathbf{w}}]_F, [\underline{\mathbf{v}}]_F)_F \quad \forall \underline{\mathbf{w}}, \underline{\mathbf{v}} \in H^1(\mathcal{T}_h; \mathbb{R}^d),$$

where, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, $\delta_{TF}^k \underline{\mathbf{v}}_T := \boldsymbol{\pi}_F^k(\mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T - \mathbf{v}_F) - \boldsymbol{\pi}_T^k(\mathbf{r}_T^{k+1} \underline{\mathbf{v}}_T - \mathbf{v}_T)$. A discussion on the case $k = 0$, including a justification of the term involving the bilinear form j_h , can be found in [12]; see also [22, Sect. 7.6].

Following [22, Chap. 7], the bilinear form a_h defines an inner product on $\underline{\mathbf{U}}_h^k$, and we denote by $\|\cdot\|_{a,h}$ the induced norm. The corresponding dual norm $\|\cdot\|_{a,h,*}$ is defined such that, for any linear form $\ell_h : \underline{\mathbf{U}}_h^k \rightarrow \mathbb{R}$,

$$\|\ell_h\|_{a,h,*} := \sup_{\underline{\mathbf{v}}_h \in \underline{\mathbf{U}}_h^k \setminus \{\mathbf{0}\}} \frac{\ell_h(\underline{\mathbf{v}}_h)}{\|\underline{\mathbf{v}}_h\|_{a,h}}. \quad (6.7)$$

The following consistency property holds: For all $\underline{\mathbf{w}} \in \mathcal{U} \cap H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)$,

$$\|\mathcal{E}_{a,h}(\underline{\mathbf{w}}; \cdot)\|_{a,h,*} \lesssim h^{k+1} |\underline{\mathbf{w}}|_{H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)}, \quad (6.8)$$

where the hidden constant is independent of both h and $\underline{\mathbf{w}}$ and the consistency error linear form $\mathcal{E}_{a,h}(\underline{\mathbf{w}}; \cdot) : \underline{\mathbf{U}}_h^k \rightarrow \mathbb{R}$ is such that, for all $\underline{\mathbf{v}}_h \in \underline{\mathbf{U}}_h^k$,

$$\mathcal{E}_{a,h}(\underline{\mathbf{w}}; \underline{\mathbf{v}}_h) := -(\nabla \cdot \nabla_s \underline{\mathbf{w}}, \underline{\mathbf{v}}_h) - a_h(\underline{\mathbf{I}}_h^k \underline{\mathbf{w}}, \underline{\mathbf{v}}_h). \quad (6.9)$$

We additionally have the following discrete Korn–Poincaré inequality:

$$\|\underline{\mathbf{v}}_h\|_{L^2(\Omega; \mathbb{R}^d)} \leq C_K \|\underline{\mathbf{v}}_h\|_{a,h} \quad \forall \underline{\mathbf{v}}_h \in \underline{\mathbf{U}}_h^k, \quad (6.10)$$

where the real number $C_K > 0$ is independent of h , but possibly depends on Ω , d , k , and the mesh regularity parameter. In the case $k \geq 1$, this inequality results from [22, Eq. (7.75) with $2\mu = 1$ and $\lambda = 0$ together with Remark 7.26] whereas, in the case $k = 0$, it is a consequence of [22, Eq. (7.109) with $\lambda = 0$ and Remark 7.26].

6.3.4.2 Pressure–Displacement Coupling

The coupling between the total pressure and the displacement is realised by means of the bilinear form $\mathbf{b}_h : \underline{\mathbf{V}}_h^k \times \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ such that, for all $(\mathbf{v}_h, q_h) \in \underline{\mathbf{V}}_h^k \times \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$,

$$\mathbf{b}_h(\mathbf{v}_h, q_h) := (\mathbf{D}_h^k \underline{\mathbf{v}}_h, q_h).$$

The following inf-sup condition holds: There is a real number $\beta > 0$ independent of h , but possibly depending on Ω , d , k , and the mesh regularity parameter, such that

$$\beta \|q_h\|_{L^2(\Omega; \mathbb{R})} \leq \|\mathbf{b}_h(\cdot, q_h)\|_{a,h,*} \quad \forall q_h \in P_{h,0}^k. \quad (6.11)$$

Moreover, we have the following consistency properties: For all $\mathbf{v} \in \mathbf{U}$,

$$\mathbf{b}_h(\underline{\mathbf{I}}_h^k \mathbf{v}, q_h) = b(\mathbf{v}, q_h) \quad \forall q_h \in P_{h,0}^k \quad (6.12)$$

and, for all $q \in H^1(\Omega; \mathbb{R}) \cap H^{k+1}(\mathcal{T}_h; \mathbb{R})$,

$$\|\mathcal{E}_{b,h}(q; \cdot)\|_{a,h,*} \lesssim h^{k+1} |q|_{H^{k+1}(\mathcal{T}_h; \mathbb{R})}, \quad (6.13)$$

where the hidden constant is independent of both h and q and the consistency error linear form $\mathcal{E}_{b,h}(q; \cdot) : \underline{\mathbf{U}}_h^k \rightarrow \mathbb{R}$ is such that, for all $\mathbf{v}_h \in \underline{\mathbf{U}}_h^k$,

$$\mathcal{E}_{b,h}(q; \mathbf{v}_h) := -(\nabla q, \mathbf{v}_h) - \mathbf{b}_h(\mathbf{v}_h, \pi_h^k q). \quad (6.14)$$

6.3.4.3 HHO Discretisation of the Darcy Term

Denote by ∇_h the broken gradient acting element-wise. The Darcy bilinear form c is approximated by $c_h^{\text{hho}} : \underline{\mathcal{Q}}_h^k \times \underline{\mathcal{Q}}_h^k \rightarrow \mathbb{R}$ such that, for all $\underline{\mathbf{r}}_h, \underline{\mathbf{q}}_h \in \underline{\mathcal{Q}}_h^k$,

$$c_h^{\text{hho}}(\underline{\mathbf{r}}_h, \underline{\mathbf{q}}_h) := (\nabla_h \mathbf{r}_h^{k+1}, \nabla_h \mathbf{r}_h^{k+1} \underline{\mathbf{q}}_h) + s_{c,h}(\underline{\mathbf{r}}_h, \underline{\mathbf{q}}_h),$$

with stabilising bilinear form

$$s_{c,h}(\underline{\mathbf{r}}_h, \underline{\mathbf{q}}_h) := \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_T} h_F^{-1} (\delta_{TF}^k \mathbf{r}_T, \delta_{TF}^k \mathbf{q}_T)_F,$$

where, for all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, $\delta_{TF}^k \mathbf{r}_T := \pi_F^k(\mathbf{r}_T^{k+1} \underline{\mathbf{q}}_T - \mathbf{q}_F) - \pi_T^k(\mathbf{r}_T^{k+1} \underline{\mathbf{q}}_T - \mathbf{q}_T)$. The bilinear form c_h^{hho} defines an inner product on $\underline{\mathcal{Q}}_{h,D}^k$ as a consequence of [22, Eq. (2.41) and Corollary 2.16], and we denote by $\|\cdot\|_{c,h,\text{hho}}$ the induced norm. The corresponding dual norm is such that, for any linear form $\ell_h : \underline{\mathcal{Q}}_{h,D}^k \rightarrow \mathbb{R}$,

$$\|\ell_h\|_{c,h,*} := \sup_{\underline{q}_h \in \underline{Q}_{h,D}^k \setminus \{0\}} \frac{\ell_h(\underline{q}_h)}{\|\underline{q}_h\|_{c,h,\text{hho}}}. \tag{6.15}$$

It follows from [22, Eq. (2.42)] that, for all $r \in H_0^1(\Omega; \mathbb{R}) \cap H^{k+2}(\mathcal{T}_h; \mathbb{R})$ such that $\Delta r \in L^2(\Omega; \mathbb{R})$,

$$\|\mathcal{E}_{c,h}^{\text{hho}}(r; \cdot)\|_{c,h,*} \lesssim h^{k+1} |r|_{H^{k+2}(\mathcal{T}_h; \mathbb{R})}, \tag{6.16}$$

where the hidden constant is independent of both h and r , and the consistency error linear form $\mathcal{E}_{c,h}^{\text{hho}}(r; \cdot) : \underline{Q}_{h,D}^k \rightarrow \mathbb{R}$ is such that, for all $\underline{q}_h \in \underline{Q}_{h,D}^k$,

$$\mathcal{E}_{c,h}^{\text{hho}}(r; \underline{q}_h) := -(\Delta r, q_h) - c_h^{\text{hho}}(\underline{I}_h^k r, \underline{q}_h). \tag{6.17}$$

The following discrete Poincaré inequality results combining [22, Lemma 2.15 and Eq. (2.41)]: For all $\underline{q}_h \in \underline{Q}_{h,D}^k$,

$$\|q_h\|_{L^2(\Omega; \mathbb{R})} \leq C_P \|\underline{q}_h\|_{c,h,\text{hho}}, \tag{6.18}$$

with real number $C_P > 0$ independent of h and \underline{q}_h , but possibly depending on Ω , d , k , and the mesh regularity parameter.

6.3.4.4 DG Discretisation of the Darcy Term

For the DG approximation of the Darcy operator we need to assume $k \geq 1$ to have consistency. Let the normal trace average operator be defined such that, for all $\psi \in H^1(\mathcal{T}_h; \mathbb{R}^d)$ and all $F \in \mathcal{F}_h^i$ shared by the mesh elements T_1 and T_2 ,

$$\{\psi \cdot \mathbf{n}\}_F := \frac{1}{2} (\psi|_{T_1} + \psi|_{T_2})|_F \cdot \mathbf{n}_F.$$

The DG method hinges on the bilinear form $c_h^{\text{dg}} : \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}) \times \mathbb{P}^k(\mathcal{T}_h; \mathbb{R}) \rightarrow \mathbb{R}$ such that, for all $r_h, q_h \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$,

$$\begin{aligned} c_h^{\text{dg}}(r_h, q_h) &:= (\nabla_h r_h, \nabla_h q_h) + \sum_{F \in \mathcal{F}_h} \frac{\eta}{h_F} ([r_h]_F, [q_h]_F)_F \\ &\quad - \sum_{F \in \mathcal{F}_h} [([r_h]_F, \{\nabla_h q_h \cdot \mathbf{n}\}_F)_F + (\{\nabla_h r_h \cdot \mathbf{n}\}_F, [q_h]_F)_F], \end{aligned} \tag{6.19}$$

where the stabilisation parameter $\eta > 0$ is chosen large enough to ensure coercivity with respect to the norm $\|\cdot\|_{c,h,\text{dg}}$ defined such that, for all $q_h \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$,

$$\|q_h\|_{c,h,\text{dg}} := \left(\|\nabla_h q_h\|_{L^2(\Omega; \mathbb{R}^d)}^2 + \sum_{F \in \mathcal{F}_h} h_F^{-1} \|[q_h]_F\|_{L^2(F)}^2 \right)^{\frac{1}{2}}.$$

Let $r \in H_0^1(\Omega, \mathbb{R})$ be such that $\Delta r \in L^2(\Omega, \mathbb{R})$, and consider the elliptic projection problem that consists in finding $r_h \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ such that

$$c_h^{\text{dg}}(r_h, q_h) = -(\Delta r, q_h)_{L^2(\Omega)} \quad \forall q_h \in \mathbb{P}^k(\mathcal{T}_h, \mathbb{R}). \quad (6.20)$$

It is inferred from [21, Appendix A] that, if Ω is convex and $r \in H^{m+1}(\mathcal{T}_h, \mathbb{R})$ for some $m \in \{0, \dots, k\}$, it holds

$$\|r_h - r\|_{L^2(\Omega)} + h\|r_h - r\|_{c,h,\text{dg}} \lesssim h^{m+1}|r|_{H^{m+1}(\mathcal{T}_h)}, \quad (6.21)$$

with hidden constant independent of h and r .

6.3.5 Discrete Problems

Assume the initial pressures given, and denote by $\mathbf{u}^0 \in \mathbf{U}$ the corresponding initial equilibrium displacement. Enforce the initial condition by setting

$$\underline{\mathbf{u}}_h^0 := \underline{\mathbf{I}}_h^k \mathbf{u}^0, \quad p_{h,i}^0 := \pi_h^k p_i^0 \quad \forall i \in \llbracket 0, M \rrbracket. \quad (6.22)$$

The discrete problem with HHO discretisation of the Darcy term (HHO-HHO scheme) reads:

Problem 6.1 (HHO-HHO scheme) *For $n = 1, \dots, N$, find $\underline{\mathbf{u}}_h^n \in \underline{\mathbf{U}}_h^k$, $p_{h,0}^n \in P_{h,0}^k$ and, for all $i \in \llbracket 1, M \rrbracket$, $\underline{p}_{h,i}^n \in \underline{P}_{h,i}^k$ such that, for all $\underline{\mathbf{v}}_h \in \underline{\mathbf{U}}_h^k$, all $q_{h,0} \in P_{h,0}^k$, and all $\underline{q}_{h,i} \in \underline{P}_{h,i}^k$, $i \in \llbracket 1, M \rrbracket$,*

$$2\mu \mathbf{a}_h(\underline{\mathbf{u}}_h^n, \underline{\mathbf{v}}_h) + \mathbf{b}_h(\underline{\mathbf{v}}_h, p_{h,0}^n) = (\mathbf{f}^n, \underline{\mathbf{v}}_h), \quad (6.23a)$$

$$\mathbf{b}_h(\underline{\mathbf{u}}_h^n, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{p}_h^n, q_{h,0}) = 0, \quad (6.23b)$$

$$(\delta_t^n \psi_i(\mathbf{p}_h), q_{h,i}) + (S_i(\mathbf{p}_h^n), q_{h,i}) + K_i c_h^{\text{hho}}(\underline{p}_{h,i}^n, \underline{q}_{h,i}) = (g_i^n, q_{h,i}) \quad \forall i \in \llbracket 1, M \rrbracket, \quad (6.23c)$$

where we have set, for any $n \in \llbracket 0, N \rrbracket$, $\mathbf{p}_h^n := (p_{h,0}^n, p_{h,1}^n, \dots, p_{h,M}^n)$ and we remind the reader that ψ_i is defined by (6.2).

The problem resulting from the DG approximation of the flow operator (HHO-DG scheme) reads:

Problem 6.2 (HHO-DG scheme) *For $n = 1, \dots, N$, find $\underline{\mathbf{u}}_h^n \in \underline{\mathbf{U}}_h^k$ and $p_{h,0}^n \in P_{h,0}^k$ such that (6.23a)–(6.23b) hold for all $\underline{\mathbf{v}}_h \in \underline{\mathbf{U}}_h^k$ and all $q_{h,0} \in P_{h,0}^k$, respectively, and,*

for all $i \in \llbracket 1, M \rrbracket$, $p_{h,i}^n \in P_{h,i}^k$ such that, for all $q_{h,i} \in P_{h,i}^k$, $i \in \llbracket 1, M \rrbracket$,

$$(\delta_i^n \psi_i(\mathbf{p}_h), q_{h,i}) + (S_i(\mathbf{p}_h^n), q_{h,i}) + K_i c_h^{\text{dg}}(p_{h,i}^n, q_{h,i}) = (g_i^n, q_{h,i}) \quad \forall i \in \llbracket 1, M \rrbracket. \quad (6.24)$$

6.4 Convergence Analysis

We carry out a convergence analysis for the methods formulated in Sect. 6.3.5. For the sake of conciseness, the focus is on the HHO-HHO scheme (6.23). The modifications needed to adapt the results to the HHO-DG scheme are discussed in Sect. 6.4.4. A unified analysis covering both HHO-HHO and HHO-DG methods for the single-network Biot problem can be found in [10].

6.4.1 An Abstract A Priori Estimate

We derive an a priori estimate for an auxiliary problem analogous to (6.23), but with modified right-hand side. Applied to the discrete problem (6.23), this estimate can be used to infer its well-posedness. Applied to the error equations (6.50) below, it gives a basic error estimate.

Problem 6.3 (HHO-HHO scheme with abstract right-hand side) *Let the families of linear forms $(\ell_1^n : \underline{U}_h^k \rightarrow \mathbb{R})_{n \in \llbracket 0, N \rrbracket}$, and, for all $i \in \llbracket 1, M \rrbracket$, $(\ell_{2,i}^n : P_{h,i}^k \rightarrow \mathbb{R})_{n \in \llbracket 1, N \rrbracket}$, be given. Assume $\underline{\mathbf{w}}_h^0 \in \underline{U}_h^k$, $r_{h,0}^0 \in P_{h,0}^k$, and, for all $i \in \llbracket 1, M \rrbracket$, $\underline{r}_{h,i}^0 \in P_{h,i}^k$ also given. For $n = 1, \dots, N$, $\underline{\mathbf{w}}_h^n \in \underline{U}_h^k$, $r_{h,0}^n \in P_{h,0}^k$ and, for all $i \in \llbracket 1, M \rrbracket$, $\underline{r}_{h,i}^n \in P_{h,i}^k$ are such that, for all $\mathbf{v}_h \in \underline{U}_h^k$, all $q_h \in P_{h,0}^k$, and all $\underline{q}_{h,i} \in P_{h,i}^k$, $i \in \llbracket 1, M \rrbracket$,*

$$2\mu a_h(\underline{\mathbf{w}}_h^n, \mathbf{v}_h) + \mathbf{b}_h(\mathbf{v}_h, r_{h,0}^n) = \ell_1^n(\mathbf{v}_h), \quad (6.25a)$$

$$\mathbf{b}_h(\underline{\mathbf{w}}_h^n, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{r}_h^n, q_{h,0}) = 0, \quad (6.25b)$$

$$(\delta_i^n \psi_i(\mathbf{r}_h), q_{h,i}) + (S_i(\mathbf{r}_h^n), q_{h,i}) + K_i c_h^{\text{hho}}(\underline{r}_{h,i}^n, \underline{q}_{h,i}) = \ell_{2,i}^n(\underline{q}_{h,i}) \quad \forall i \in \llbracket 1, M \rrbracket, \quad (6.25c)$$

where, for any $n \in \llbracket 0, N \rrbracket$, $\mathbf{r}_h^n := (r_{h,0}^n, r_{h,1}^n, \dots, r_{h,M}^n)$.

Applying discrete time derivation to (6.25b) we obtain, for all $n \in \llbracket 1, N \rrbracket$,

$$\mathbf{b}_h(\delta_t^n \underline{\mathbf{w}}_h, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \delta_t^n \mathbf{r}_h, q_{h,0}) = 0 \quad \forall q_{h,0} \in P_{h,0}^k. \quad (6.26)$$

Lemma 6.1 (Abstract a priori estimate) *Assuming τ small enough (with threshold independent of h), the solution to (6.25) satisfies the following a priori estimate:*

$$\begin{aligned} \max_{n \in \llbracket 1, N \rrbracket} & \left(\mu \|\underline{\mathbf{w}}_h^n\|_{\mathbf{a},h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^n\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|r_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\ & + \sum_{n=1}^N \tau \|r_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|r_{h,i}^n\|_{\mathbf{c},h,\text{hho}}^2 \leq \exp\left(\frac{t_F}{1-\tau}\right) (\mathcal{N}_\ell + \mathcal{N}_0), \end{aligned} \quad (6.27)$$

where we have introduced the exchange norm

$$\|\mathbf{r}_h^n\|_{\xi}^2 := \sum_{i=1}^M \sum_{j=1}^M \|\xi_{i \leftarrow j}(r_{h,i}^n - r_{h,j}^n)\|_{L^2(\Omega; \mathbb{R})}^2$$

and we have set

$$\mathcal{N}_\ell := \frac{1}{2\mu} \max_{n \in \llbracket 1, N \rrbracket} \|\ell_1^n\|_{\mathbf{a},h,*}^2 + \frac{1}{\mu} \sum_{n=1}^N \tau \|\delta_t^n \ell_1\|_{\mathbf{a},h,*}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} \|\ell_{2,i}^n\|_{\mathbf{c},h,*}^2, \quad (6.28a)$$

$$\mathcal{N}_0 := 2 \|\ell_1^0\|_{\mathbf{a},h,*} \|\underline{\mathbf{w}}_h^0\|_{\mathbf{a},h} + 2\mu \|\underline{\mathbf{w}}_h^0\|_{\mathbf{a},h}^2 + \frac{1}{\lambda} \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^0\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|r_{h,i}^0\|_{L^2(\Omega; \mathbb{R})}^2. \quad (6.28b)$$

Moreover, it holds

$$\frac{\beta^2}{\mu} \max_{n \in \llbracket 1, N \rrbracket} \|r_{h,0}^n\|_{L^2(\Omega; \mathbb{R})}^2 \leq \frac{2}{\mu} \max_{n \in \llbracket 1, N \rrbracket} \|\ell_1^n\|_{\mathbf{a},h,*}^2 + 4\beta^2 \exp\left(\frac{t_F}{1-\tau}\right) (\mathcal{N}_\ell + \mathcal{N}_0). \quad (6.29)$$

Proof We start by deriving a basic energy estimate and then, leveraging the discrete inf-sup condition (6.11), deduce from the latter the estimate on the total pressure.

(i) *Basic energy estimate.* Let $\mathbf{N} \in \llbracket 1, N \rrbracket$ and $n \in \llbracket 1, \mathbf{N} \rrbracket$. Taking $\underline{\mathbf{v}}_h = \delta_t^n \underline{\mathbf{w}}_h$ in (6.25a), $q_{h,0} = -r_{h,0}^n$ in (6.26), and, for all $i \in \llbracket 1, M \rrbracket$, $\underline{q}_{h,i} = r_{h,i}^n$ in (6.25c), and summing the resulting equations we obtain, after expanding $\delta_t^n \psi_i(\mathbf{r}_h)$ according to its definition,

$$\begin{aligned} 2\mu \mathbf{a}_h(\underline{\mathbf{w}}_h^n, \delta_t^n \underline{\mathbf{w}}_h^n) + \lambda^{-1} (\boldsymbol{\alpha} \cdot \delta_t^n \mathbf{r}_h^n, \boldsymbol{\alpha} \cdot \mathbf{r}_h^n) + \sum_{i=1}^M C_i (\delta_t^n r_{h,i}, r_{h,i}^n) \\ + \sum_{i=1}^M (S_i(r_h^n), r_{h,i}^n) + \sum_{i=1}^M K_i \text{c}^{\text{hho}}(\underline{r}_{h,i}^n, r_{h,i}^n) = \ell_1^n(\delta_t^n \underline{\mathbf{w}}_h) + \sum_{i=1}^M \ell_{2,i}(r_{h,i}^n). \end{aligned} \quad (6.30)$$

Denote by $\mathcal{L}^n = \mathcal{L}_1^n + \dots + \mathcal{L}_5^n$ and $\mathcal{R}^n = \mathcal{R}_1^n + \mathcal{R}_2^n$, respectively, the left- and right-hand side of the above expression, and set $\mathcal{L} := \sum_{n=1}^N \tau \mathcal{L}^n$ and, for $i \in \{1, 2\}$,

$$\mathcal{R}_i := \sum_{n=1}^N \tau \mathcal{R}_i^n.$$

(i.A) Lower bound for \mathcal{L} . Recalling the definition (6.4) of the discrete time derivative and using multiple times the formula

$$x(x-y) = \frac{1}{2} (x^2 + (x-y)^2 - y^2) \quad (6.31)$$

with $x = \bullet^n$ and $y = \bullet^{n-1}$, we can write for the first three terms in \mathcal{L}^n

$$\begin{aligned} \mathcal{L}_1^n &= \frac{\mu}{\tau} (\|\mathbf{w}_h^n\|_{a,h}^2 + \|\mathbf{w}_h^n - \mathbf{w}_h^{n-1}\|_{a,h}^2 - \|\mathbf{w}_h^{n-1}\|_{a,h}^2), \\ \mathcal{L}_2^n &= \frac{1}{2\lambda\tau} \left(\|\boldsymbol{\alpha} \cdot \mathbf{r}_h^n\|_{L^2(\Omega; \mathbb{R})}^2 + \|\boldsymbol{\alpha} \cdot (\mathbf{r}_h^n - \mathbf{r}_h^{n-1})\|_{L^2(\Omega; \mathbb{R})}^2 - \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^{n-1}\|_{L^2(\Omega; \mathbb{R})}^2 \right), \\ \mathcal{L}_3^n &= \sum_{i=1}^M \frac{C_i}{2\tau} \left(\|r_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 + \|r_{h,i}^n - r_{h,i}^{n-1}\|_{L^2(\Omega; \mathbb{R})}^2 - \|r_{h,i}^{n-1}\|_{L^2(\Omega; \mathbb{R})}^2 \right). \end{aligned} \quad (6.32)$$

For the fourth term, using again (6.31) this time with $x = r_{h,i}^n$ and $y = r_{h,j}^{n-1}$ along with $\xi_{i \leftarrow j} = \xi_{j \leftarrow i}$, we get

$$\begin{aligned} \mathcal{L}_4^n &= \sum_{i=1}^M \sum_{j=1}^M (\xi_{i \leftarrow j} (r_{h,i}^n - r_{h,j}^{n-1}), r_{h,i}^n) \\ &= \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \left(\|\xi_{i \leftarrow j}^{\frac{1}{2}} r_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 + \|\xi_{i \leftarrow j}^{\frac{1}{2}} (r_{h,i}^n - r_{h,j}^{n-1})\|_{L^2(\Omega; \mathbb{R})}^2 - \|\xi_{j \leftarrow i}^{\frac{1}{2}} r_{h,j}^{n-1}\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\ &= \frac{1}{2} \sum_{i=1}^M \sum_{j=1}^M \|\xi_{i \leftarrow j}^{\frac{1}{2}} (r_{h,i}^n - r_{h,j}^{n-1})\|_{L^2(\Omega; \mathbb{R})}^2 = \frac{1}{2} \|\mathbf{r}_h^n\|_{\xi}^2. \end{aligned} \quad (6.33)$$

Multiplying (6.30) by τ , summing over $n \in \llbracket 1, N \rrbracket$, using (6.32) and (6.33), and telescoping out the appropriate summands, we get

$$\begin{aligned} \mu \|\mathbf{w}_h^N\|_{a,h}^2 + \frac{1}{2\lambda} \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^N\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M \frac{C_i}{2} \|r_{h,i}^N\|_{L^2(\Omega; \mathbb{R})}^2 + \frac{1}{2} \sum_{n=1}^N \tau \|r_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|L_{h,i}^n\|_{c,h,hho}^2 \\ \leq \mathcal{R} + \mu \|\mathbf{w}_h^0\|_{a,h}^2 + \frac{1}{2\lambda} \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^0\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M \frac{C_i}{2} \|r_{h,i}^0\|_{L^2(\Omega; \mathbb{R})}^2. \end{aligned} \quad (6.34)$$

(i.B) Upper bound for \mathcal{R} . A discrete integration by parts in time gives for the first term

$$\begin{aligned}
\mathcal{R}_1 &= \ell_1^N(\underline{\mathbf{w}}_h^N) - \ell_1^0(\underline{\mathbf{w}}_h^0) - \sum_{i=1}^N \tau (\delta_t^n \ell_1)(\underline{\mathbf{w}}_h^{n-1}) \\
&\leq \|\ell_1^N\|_{a,h,*} \|\underline{\mathbf{w}}_h^N\|_{a,h} + \|\ell_1^0\|_{a,h,*} \|\underline{\mathbf{w}}_h^0\|_{a,h} + \sum_{n=1}^N \tau \mu^{-\frac{1}{2}} \|\delta_t^n \ell_1\|_{a,h,*} \mu^{\frac{1}{2}} \|\underline{\mathbf{w}}_h^{n-1}\|_{a,h} \\
&\leq \frac{1}{4\mu} \|\ell_1^N\|_{a,h,*}^2 + \frac{\mu}{2} \|\underline{\mathbf{w}}_h^N\|_{a,h}^2 + \|\ell_1^0\|_{a,h,*} \|\underline{\mathbf{w}}_h^0\|_{a,h} \\
&\quad + \frac{1}{2\mu} \sum_{n=1}^N \tau \|\delta_t^n \ell_1\|_{a,h,*}^2 + \frac{\mu}{2} \sum_{n=0}^N \tau \|\underline{\mathbf{w}}_h^n\|_{a,h}^2,
\end{aligned} \tag{6.35}$$

where we have used multiple times the definition of dual norm (6.7) to pass to the second line and we have concluded invoking the standard and generalised Young inequalities and rearranging.

Moving to the second term, we use the definition (6.15) of the dual norm and the Young inequality to write, for all $i \in \llbracket 1, M \rrbracket$,

$$\begin{aligned}
\sum_{n=1}^N \tau \ell_{2,i}^n(\underline{\mathcal{L}}_{h,i}^n) &\leq \sum_{n=1}^N \tau K_i^{-\frac{1}{2}} \|\ell_{2,i}^n\|_{c,h,*} K_i^{\frac{1}{2}} \|\underline{\mathcal{L}}_{h,i}^n\|_{c,h,\text{hho}} \\
&\leq \frac{1}{2} \sum_{n=1}^N \tau K_i^{-1} \|\ell_{2,i}^n\|_{c,h,*}^2 + \frac{1}{2} \sum_{n=1}^N \tau K_i \|\underline{\mathcal{L}}_{h,i}^n\|_{c,h,\text{hho}}^2.
\end{aligned}$$

Hence, summing over $i \in \llbracket 1, M \rrbracket$,

$$\mathcal{R}_2 \leq \frac{1}{2} \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} \|\ell_{2,i}^n\|_{c,h,*}^2 + \frac{1}{2} \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|\underline{\mathcal{L}}_{h,i}^n\|_{c,h,\text{hho}}^2. \tag{6.36}$$

Gathering (6.35) and (6.36) and rearranging, we arrive at

$$\begin{aligned}
\mathcal{R} &\leq \frac{\mu}{2} \|\underline{\mathbf{w}}_h^N\|_{a,h}^2 + \frac{1}{2} \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|\underline{\mathcal{L}}_{h,i}^n\|_{c,h,\text{hho}}^2 + \frac{\mu}{2} \sum_{n=0}^N \tau \|\underline{\mathbf{w}}_h^n\|_{a,h}^2 \\
&\quad + \frac{1}{4\mu} \|\ell_1^N\|_{a,h,*}^2 + \frac{1}{2\mu} \sum_{n=1}^N \tau \|\delta_t^n \ell_1\|_{a,h,*}^2 + \frac{1}{2} \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} \|\ell_{2,i}^n\|_{c,h,*}^2 \\
&\quad + \|\ell_1^0\|_{a,h,*} \|\underline{\mathbf{w}}_h^0\|_{a,h}.
\end{aligned} \tag{6.37}$$

(i.C) Basic estimate. Combining (6.34) and (6.37) and multiplying by 2, we arrive at

$$\begin{aligned}
& \mu \|\underline{\mathbf{w}}_h^N\|_{\mathbf{a},h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \mathbf{r}_h^N\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|r_{h,i}^N\|_{L^2(\Omega; \mathbb{R})}^2 \\
& + \sum_{n=1}^N \tau \|r_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|r_{h,i}^n\|_{\mathbf{c},h,\text{hho}}^2 \leq \mu \sum_{n=0}^N \tau \|\underline{\mathbf{w}}_h^n\|_{\mathbf{a},h}^2 + \mathcal{N}_\ell + \mathcal{N}_0.
\end{aligned} \tag{6.38}$$

The estimate (6.27) follows from the discrete Gronwall inequality of [28, Lemma 5.1].

(ii) *Estimate on the total pressure.* For all $n \in \llbracket 1, N \rrbracket$, using the inf-sup stability (6.11) of the pressure-displacement coupling, we can write

$$\begin{aligned}
\beta \|r_{h,0}^n\|_{L^2(\Omega; \mathbb{R})} & \leq \sup_{\mathbf{v}_h \in \underline{\mathbf{U}}_h^n \setminus \{\mathbf{0}\}} \frac{\mathbf{b}_h(\mathbf{v}_h, r_{h,0}^n)}{\|\mathbf{v}_h\|_{\mathbf{a},h}} \\
& \leq \sup_{\mathbf{v}_h \in \underline{\mathbf{U}}_h^n \setminus \{\mathbf{0}\}} \frac{\ell_1^n(\mathbf{v}_h) - 2\mu \mathbf{a}_h(\underline{\mathbf{w}}_h^n, \mathbf{v}_h)}{\|\mathbf{v}_h\|_{\mathbf{a},h}} \\
& \leq \|\ell_1^n\|_{\mathbf{a},h,*} + 2\mu \|\underline{\mathbf{w}}_h^n\|_{\mathbf{a},h},
\end{aligned} \tag{6.39}$$

where we have used (6.25a) in the second line and we have concluded using the definition (6.7) of dual norm for the first term and a Cauchy–Schwarz inequality on the symmetric positive definite bilinear form \mathbf{a}_h for the second. Squaring, dividing both sides by μ , passing to the maximum over $n \in \llbracket 1, N \rrbracket$, and using (6.27) to estimate the second term in the right-hand side, (6.41) follows.

6.4.2 A Priori Estimate for the HHO-HHO Scheme

The following lemma contains an a priori estimate on the discrete solution, from which the well posedness of problem (6.23) can be inferred.

Lemma 6.2 (A priori estimate on the discrete solution) *Assuming τ small enough, any solution $(\underline{\mathbf{u}}_h^n, p_{h,0}^n, (p_{h,i})_{1 \leq i \leq M})_{1 \leq n \leq N}$ to the discrete problem (6.23) satisfies the following a priori bound:*

$$\begin{aligned}
& \max_{n \in \llbracket 1, N \rrbracket} \left(\mu \|\underline{\mathbf{u}}_h^n\|_{\mathbf{a},h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \mathbf{p}\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|p_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\
& + \sum_{n=1}^N \tau \|r_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|p_{h,i}^n\|_{\mathbf{c},h,\text{hho}}^2 \leq \exp\left(\frac{t_F}{1-\tau}\right) (\mathcal{A} + \mathcal{B}),
\end{aligned} \tag{6.40}$$

where

$$\begin{aligned}
\mathcal{A} &:= \frac{C_K^2}{2\mu} \|\mathbf{f}\|_{C^0([0,t_F];L^2(\Omega;\mathbb{R}^d))}^2 + \frac{1}{\mu} \|\mathbf{f}\|_{H^1(0,t_F;L^2(\Omega;\mathbb{R}^d))}^2 \\
&\quad + C_{\text{PTF}} \sum_{i=1}^M \frac{1}{K_i} \|\mathbf{g}_i\|_{C^0([0,t_F];L^2(\Omega;\mathbb{R}))}^2 \\
\mathcal{B} &:= 2C_K \|\mathbf{f}^0\|_{L^2(\Omega;\mathbb{R}^d)} \|\underline{\mathbf{u}}_h^0\|_{\mathbf{a},h} + 2\mu \|\underline{\mathbf{u}}_h^0\|_{\mathbf{a},h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \mathbf{p}_h^0\|_{L^2(\Omega;\mathbb{R})}^2 \\
&\quad + \sum_{i=1}^M C_i \|p_{h,i}^0\|_{L^2(\Omega;\mathbb{R})}^2.
\end{aligned}$$

Moreover, it holds

$$\frac{\beta^2}{\mu} \max_{n \in \llbracket 1, N \rrbracket} \|p_{h,0}^n\|_{L^2(\Omega;\mathbb{R})}^2 \leq \frac{2C_K^2}{\mu} \|\mathbf{f}\|_{C^0([0,t_F];L^2(\Omega;\mathbb{R}^d))}^2 + 4\beta^2 \exp\left(\frac{t_F}{1-\tau}\right) (\mathcal{A} + \mathcal{B}). \quad (6.41)$$

Proof We apply Lemma 6.1 with $\ell_1^n = (\underline{\mathbf{U}}_h^k \ni \mathbf{v}_h \mapsto (\mathbf{f}, \mathbf{v}_h) \in \mathbb{R})$ for all $n \in \llbracket 0, N \rrbracket$ and $\ell_2^n = (\underline{\mathbf{P}}_{h,i}^n \ni \underline{q}_{h,i} \mapsto (g_i, q_{h,i}) \in \mathbb{R})$ for all $n \in \llbracket 1, N \rrbracket$ and all $i \in \llbracket 1, M \rrbracket$, and show that

$$\mathcal{N}_\ell \leq \mathcal{A} \text{ and } \mathcal{N}_0 \leq \mathcal{B}. \quad (6.42)$$

Let us prove the first bound in (6.42). Denote by $\mathcal{N}_{\ell,i}$, $i \in \llbracket 1, 3 \rrbracket$, the terms in the right-hand side of (6.28a). We start by noticing that, for all $n \in \llbracket 0, N \rrbracket$,

$$\begin{aligned}
\|\ell_1^n\|_{\mathbf{a},h,*} &= \sup_{\mathbf{v}_h \in \underline{\mathbf{U}}_h^k \setminus \{\mathbf{0}\}} \frac{\ell_1^n(\mathbf{v}_h)}{\|\mathbf{v}_h\|_{\mathbf{a},h}} \\
&= \sup_{\mathbf{v}_h \in \underline{\mathbf{U}}_h^k \setminus \{\mathbf{0}\}} \frac{\|\mathbf{f}^n\|_{L^2(\Omega;\mathbb{R}^d)} \|\mathbf{v}_h\|_{L^2(\Omega;\mathbb{R}^d)}}{\|\mathbf{v}_h\|_{\mathbf{a},h}} \\
&= \sup_{\mathbf{v}_h \in \underline{\mathbf{U}}_h^k \setminus \{\mathbf{0}\}} \frac{C_K \|\mathbf{f}^n\|_{L^2(\Omega;\mathbb{R}^d)} \|\mathbf{v}_h\|_{\mathbf{a},h}}{\|\mathbf{v}_h\|_{\mathbf{a},h}} \leq C_K \|\mathbf{f}^n\|_{L^2(\Omega;\mathbb{R}^d)},
\end{aligned} \quad (6.43)$$

where we have used the definition (6.7) of the dual norm in the first line, a Cauchy–Schwarz inequality to pass to the the second line, and the discrete Korn inequality (6.10) to pass to the third line. As a consequence,

$$\mathcal{N}_{\ell,1} \leq \frac{C_K^2}{2\mu} \max_{n \in \llbracket 1, N \rrbracket} \|\mathbf{f}^n\|_{L^2(\Omega;\mathbb{R}^d)}^2 = \frac{C_K^2}{2\mu} \|\mathbf{f}\|_{C^0([0,t_F];L^2(\Omega;\mathbb{R}^d))}^2. \quad (6.44)$$

Proceeding similarly for the second term and invoking the boundedness (6.5) of the discrete time derivative with $V = L^2(\Omega; \mathbb{R}^d)$ and $\varphi = \mathbf{f}$, we get

$$\mathcal{N}_{\ell,2} \leq \frac{C_K^2}{2\mu} \sum_{n=1}^n \tau \|\delta_t^n \mathbf{f}\|_{L^2(\Omega; \mathbb{R}^d)}^2 \leq \frac{C_K^2}{2\mu} \|\mathbf{f}\|_{H^1(0,t_F; L^2(\Omega; \mathbb{R}^d))}^2. \tag{6.45}$$

To bound the third term, we observe that, using the definition (6.15) of the dual norm and the Poincaré inequality in a similar manner as above, it holds, for all $n \in \llbracket 1, N \rrbracket$ and all $i \in \llbracket 1, M \rrbracket$, $\|\ell_{2,i}^n\|_{c,h,*} \leq K_i^{-1} C_P \|g_i^n\|_{L^2(\Omega; \mathbb{R})}$, hence

$$\begin{aligned} \mathcal{N}_{\ell,3} &\leq C_P \sum_{i=1}^M \frac{1}{K_i} \sum_{n=1}^N \tau \|g_i^n\|_{L^2(\Omega; \mathbb{R})}^2 \\ &\leq C_{PTF} \sum_{i=1}^M \frac{1}{K_i} \max_{n \in \llbracket 1, N \rrbracket} \|g_i^n\|_{L^2(\Omega; \mathbb{R})}^2 = C_{PTF} \sum_{i=1}^M \frac{1}{K_i} \|g_i\|_{C^0([0,t_F]; L^2(\Omega; \mathbb{R}))}^2. \end{aligned} \tag{6.46}$$

Gathering (6.44)–(6.46), the first bound in (6.30) follows. The second bound in (6.30) is an immediate after invoking (6.43) with $n = 0$. This concludes the proof.

6.4.3 Error Estimate for the HHO-HHO Scheme

Following the general ideas of [20], we estimate the error such that, for all $n \in \llbracket 0, N \rrbracket$,

$$\underline{\mathbf{e}}_h^n := \underline{\mathbf{u}}_h^n - \hat{\underline{\mathbf{u}}}_h^n, \quad \epsilon_{h,0}^n := p_{h,0}^n - \hat{p}_{h,0}^n, \quad \epsilon_{h,i}^n := \underline{p}_{h,i}^n - \hat{\underline{p}}_{h,i}^n \quad \forall i \in \llbracket 1, M \rrbracket, \tag{6.47}$$

where the interpolate of the continuous solution is obtained setting, for all $n \in \llbracket 0, N \rrbracket$,

$$\hat{\underline{\mathbf{u}}}_h^n := \underline{\mathbf{I}}_h^k \mathbf{u}^n, \quad \hat{p}_{h,0}^n := \pi_h^k p_0^n, \quad \hat{\underline{p}}_{h,i}^n := \underline{\mathbf{I}}_h^k p_i^n \quad \forall i \in \llbracket 1, M \rrbracket. \tag{6.48}$$

The starting point for the error analysis is the following proposition, which establishes that the errors solve the auxiliary problem (6.25) for a suitable choice of the right-hand sides ℓ_1 and $\ell_{2,i}$, $i \in \llbracket 1, M \rrbracket$.

Proposition 6.1 (Error equations) *We have that*

$$\underline{\mathbf{e}}_h^0 = \mathbf{0}, \quad \epsilon_{h,0}^0 = 0, \quad \epsilon_{h,i}^0 = \underline{0} \quad \forall i \in \llbracket 1, M \rrbracket. \tag{6.49}$$

Additionally, for $n = 1, \dots, N$, it holds, for all $\mathbf{v}_h \in \underline{\mathbf{U}}_h^k$, all $q_{h,0} \in P_{h,0}^k$,

$$2\mu \mathbf{a}_h(\underline{\mathbf{e}}_h^n, \mathbf{v}_h) + \mathbf{b}_h(\mathbf{v}_h, \epsilon_{h,0}^n) = 2\mu \mathcal{E}_{a,h}(\mathbf{u}^n; \mathbf{v}_h) + \mathcal{E}_{b,h}(p_0^n; \mathbf{v}_h), \tag{6.50a}$$

$$\mathbf{b}_h(\underline{\mathbf{e}}_h^n, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h^n, q_{h,0}) = 0, \tag{6.50b}$$

and, for all $i \in \llbracket 1, M \rrbracket$ and all $q_{h,i} \in \underline{P}_{h,i}^k$,

$$\begin{aligned}
& (\delta_t^n \psi_i(\boldsymbol{\epsilon}_h), q_{h,i}) + (S_i(\boldsymbol{\epsilon}_h^n), q_{h,i}) + K_i c_h^{\text{hho}}(\underline{\boldsymbol{\epsilon}}_{h,i}^n, \underline{q}_{h,i}) \\
& = (d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p}), q_{h,i}) + \mathcal{E}_{c,h}^{\text{hho}}(p_i^n; \underline{q}_{h,i}), \tag{6.50c}
\end{aligned}$$

where we have set, for all $n \in \llbracket 0, N \rrbracket$, $\boldsymbol{\epsilon}_h^n := (\epsilon_{h,0}^n, \epsilon_{h,1}^n, \dots, \epsilon_{h,M}^n)$ and, given a function of time φ smooth enough, we have introduced the abridged notation $d_t^n \varphi := d_t \varphi(t^n)$.

Proof Equation (6.49) is an immediate consequence of the definition (6.47) of the errors along with the discrete initial condition (6.22).

Let now $n \in \llbracket 1, N \rrbracket$. To prove (6.50a), it suffices to subtract from both sides of (6.23a) the quantity $2\mu a_h(\hat{\underline{\mathbf{u}}}_h^n, \mathbf{v}_h) + b_h(\mathbf{v}_h, \hat{p}_{h,0}^n)$, observe that $\mathbf{f}^n = -2\mu \nabla \cdot (\nabla_s \mathbf{u}^n) - \nabla p_0^n$ almost everywhere in Ω , and recall the definitions (6.9) and (6.14) of the consistency error linear forms associated with a_h and b_h .

Moving to (6.50b), we observe that, for all $q_{h,0} \in P_{h,0}^k$,

$$\begin{aligned}
b_h(\hat{\underline{\mathbf{u}}}_h^n, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \hat{\mathbf{p}}_h^n, q_{h,0}) &= b_h(\mathbf{I}_h^k \mathbf{u}^n, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \boldsymbol{\pi}_h^k p^n, q_{h,0}) \\
&= b(\mathbf{u}, q_{h,0}) - \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{p}^n, q_{h,0}) = 0, \tag{6.51}
\end{aligned}$$

where, to pass to the second line, we have used the consistency property (6.12) of b_h together with the definition (6.6) of the global L^2 -orthogonal projector and $q_{h,0} \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ to remove it from the second term, while the conclusion follows from (6.1a) after observing that $P_{h,0}^k \subset P_0$. The error equation (6.50b) then follows subtracting (6.51) from (6.23b) and using the linearity of the bilinear forms in the left-hand side.

Finally, to prove (6.50c) for a given $i \in \llbracket 1, M \rrbracket$ and $\underline{q}_{h,i} \in P_{h,i}^k$, we subtract from both sides the quantity $(\delta_t^n \psi_i(\hat{\mathbf{p}}_h), q_{h,i}) + (S_i(\hat{\mathbf{p}}_h^n), q_{h,i}) + K_i c_h^{\text{hho}}(\hat{\underline{p}}_{h,i}^n, \underline{q}_{h,i})$ and observe that

$$\begin{aligned}
(g_i^n, q_{h,i}) &= (d_t^n \psi_i(\mathbf{p}), q_{h,i}) + (S_i(\mathbf{p}^n), q_{h,i}) - (K_i \Delta p_i^n, q_{h,i}) \\
&= (d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p}), q_{h,i}) + \mathcal{E}_{c,h}^{\text{hho}}(p_i^n; \underline{q}_{h,i}) \\
&\quad + (\delta_t^n \psi_i(\hat{\mathbf{p}}_h), q_{h,i}) + (S_i(\hat{\mathbf{p}}_h^n), q_{h,i}) + K_i c_h^{\text{hho}}(\hat{\underline{p}}_{h,i}^n, \underline{q}_{h,i}),
\end{aligned}$$

where, to pass to the second line, we have added and subtracted $(\delta_t^n \psi_i(\hat{\mathbf{p}}_h), q_{h,i}) + c_h^{\text{hho}}(\hat{\underline{p}}_{h,i}^n, \underline{q}_{h,i})$, used the fact that $q_{h,i} \in \mathbb{P}^k(\mathcal{T}_h; \mathbb{R})$ along with the linearity of ψ and the definition (6.6) of the global L^2 -orthogonal projector to write $(\delta_t^n \psi_i(\hat{\mathbf{p}}_h), q_{h,i}) = (\delta_t^n \psi_i(\mathbf{p}), q_{h,i})$, and recalled the definition (6.17) of the consistency error associated with the bilinear form c_h^{hho} .

Theorem 6.1 (Error estimate for the HHO-HHO scheme) *Assume the additional regularity*

$$\begin{aligned} \mathbf{u} &\in H^1(0, t_F; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)), \\ p_0 &\in H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R})), \\ \forall i \in \llbracket 1, M \rrbracket, \quad p_i &\in C^0([0, t_F]; H^{k+2}(\mathcal{T}_h; \mathbb{R})), \\ \forall i \in \llbracket 1, M \rrbracket, \quad \psi_i(\mathbf{p}) &\in H^2(0, t_F; L^2(\Omega; \mathbb{R})). \end{aligned}$$

Then, for a time step τ small enough (with threshold independent of h), it holds that

$$\begin{aligned} \max_{n \in \llbracket 1, N \rrbracket} &\left(\mu \|\boldsymbol{\xi}_h^n\|_{\mathbf{a},h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h^n\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|\epsilon_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 + \frac{\beta^2}{\mu} \|\epsilon_{h,0}^n\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\ &+ \sum_{n=1}^N \tau \|\epsilon_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|\underline{\epsilon}_{h,i}^n\|_{\mathbf{c},h,\text{hho}}^2 \lesssim h^{2(k+1)} \mathcal{A} + \tau^2 \mathcal{B}, \end{aligned} \quad (6.52)$$

where the hidden constant is independent of h , τ , of the problem data, of \mathbf{u} , and of p_i , $i \in \llbracket 0, M \rrbracket$, but possibly depends on Ω , t_F , the mesh regularity parameter, and k , and we have set

$$\begin{aligned} \mathcal{A} &:= \|\mathbf{u}\|_{H^1(0, t_F; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d))}^2 + \mu^{-1} \|p_0\|_{H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R}^d))}^2 \\ &\quad + \sum_{i=1}^M K_i^{-1} \|p_i\|_{C^0([0, t_F]; H^{k+2}(\mathcal{T}_h; \mathbb{R}))}^2, \\ \mathcal{B} &:= \sum_{i=1}^M K_i^{-1} \|\psi_i(\mathbf{p})\|_{H^2(0, t_F; L^2(\Omega; \mathbb{R}))}^2. \end{aligned}$$

Proof For the sake of brevity, denote by $\mathcal{E}_{h\tau}$ the left-hand side of (6.52). Applying Lemma 6.1 with, for all $n \in \llbracket 1, N \rrbracket$,

$$\begin{aligned} \ell_1^n &= 2\mu \mathcal{E}_{\mathbf{a},h}(\mathbf{u}^n; \cdot) + \mathcal{E}_{\mathbf{b},h}(p_0^n; \cdot), \\ \ell_{2,i}^n &= (\mathbf{d}_i^n \psi_i(\mathbf{p}) - \delta_i^n \psi_i(\mathbf{p}), \cdot) + \mathcal{E}_{\mathbf{c},h}^{\text{hho}}(p_i; \cdot) \quad \forall i \in \llbracket 1, M \rrbracket, \end{aligned}$$

using multiple times the triangle inequality, and rearranging the terms, we arrive at

$$\begin{aligned} \mathcal{E}_{h\tau} &\lesssim \mu^{-1} \max_{n \in \llbracket 1, N \rrbracket} \|2\mu \mathcal{E}_{\mathbf{a},h}(\mathbf{u}^n; \cdot) + \mathcal{E}_{\mathbf{b},h}(p_0^n; \cdot)\|_{\mathbf{a},h,*}^2 \\ &\quad + \mu^{-1} \sum_{n=1}^N \tau \|\delta_i^n (2\mu \mathcal{E}_{\mathbf{a},h}(\mathbf{u}; \cdot) + \mathcal{E}_{\mathbf{b},h}(p_0; \cdot))\|_{\mathbf{a},h,*}^2 \\ &\quad + \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} \|\mathcal{E}_{\mathbf{c},h}^{\text{hho}}(p_i^n; \cdot)\|_{\mathbf{c},h,*}^2 \end{aligned}$$

$$+ \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} \|(\mathfrak{d}_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p}), \cdot)\|_{\mathfrak{c},h,*}^2 =: \mathfrak{T}_1 + \dots + \mathfrak{T}_4. \quad (6.53)$$

We proceed to bound the terms in the right-hand side of the above expression. For the first term, we write

$$\begin{aligned} \mathfrak{T}_1 &\lesssim \mu^{-1} \left(\max_{n \in \llbracket 1, N \rrbracket} \|2\mu \mathcal{E}_{a,h}(\mathbf{u}^n; \cdot)\|_{\mathfrak{a},h,*}^2 + \max_{n \in \llbracket 1, N \rrbracket} \|\mathcal{E}_{b,h}(p_0^n; \cdot)\|_{\mathfrak{a},h,*}^2 \right) \\ &\lesssim h^{2(k+1)} \mu^{-1} \max_{n \in \llbracket 1, N \rrbracket} \left(2\mu |\mathbf{u}^n|_{H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)}^2 + |p_0^n|_{H^{k+1}(\mathcal{T}_h; \mathbb{R})}^2 \right) \\ &\leq h^{2(k+1)} \left(2\|\mathbf{u}\|_{C^0([0, t_F]; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d))}^2 + \mu^{-1} \|p_0\|_{C^0([0, t_F]; H^{k+1}(\mathcal{T}_h; \mathbb{R}))}^2 \right) \\ &\lesssim h^{2(k+1)} \mathcal{A}, \end{aligned} \quad (6.54)$$

where, to pass to the second line, we have used the consistency properties (6.8) of \mathfrak{a}_h and (6.13) of \mathfrak{b}_h , while the conclusion follows from the embedding $H^1(0, t_F; V) \hookrightarrow C^0([0, t_F]; V)$ valid in dimension 1.

For the second term, we write

$$\begin{aligned} \mathfrak{T}_2 &\lesssim \mu^{-1} \sum_{n=1}^N \tau \left(\|2\mu \mathcal{E}_{a,h}(\delta_t^n \mathbf{u}; \cdot)\|_{\mathfrak{a},h,*}^2 + \|\mathcal{E}_{b,h}(\delta_t^n p_0; \cdot)\|_{\mathfrak{a},h,*}^2 \right) \\ &\lesssim h^{2(k+1)} \mu^{-1} \sum_{n=1}^N \tau \left(2\mu |\delta_t^n \mathbf{u}|_{H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)}^2 + |\delta_t^n p_0|_{H^{k+1}(\mathcal{T}_h; \mathbb{R})}^2 \right) \\ &\lesssim h^{2(k+1)} \left(\|\mathbf{u}\|_{H^1(0, t_F; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d))}^2 + \mu^{-1} \|p_0\|_{H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R}))}^2 \right) \\ &\lesssim h^{2(k+1)} \mathcal{A}, \end{aligned} \quad (6.55)$$

where, in the first line, we have used the fact that $\delta_t^n (2\mu \mathcal{E}_{a,h}(\mathbf{u}; \cdot) + \mathcal{E}_{b,h}(p_0; \cdot)) = 2\mu \mathcal{E}_{a,h}(\delta_t^n \mathbf{u}; \cdot) + \mathcal{E}_{b,h}(\delta_t^n p_0; \cdot)$ followed by a triangle inequality, we have invoked the consistency (6.8) of \mathfrak{a}_h and (6.13) of \mathfrak{b}_h to pass to the second line, and the boundedness (6.5) of the backward time derivative operator to pass to the third line.

For the third term, the consistency properties (6.16) of $\mathfrak{c}_h^{\text{ho}}$ readily give

$$\begin{aligned} \mathfrak{T}_3 &\leq h^{2(k+1)} \sum_{i=1}^M \sum_{n=1}^N \tau K_i^{-1} |p_i^n|_{H^{k+2}(\mathcal{T}_h; \mathbb{R})}^2 \\ &\lesssim h^{2(k+1)} t_F \sum_{i=1}^M K_i^{-1} \|p_i\|_{C^0([0, t_F]; H^{k+2}(\mathcal{T}_h; \mathbb{R}))}^2 \lesssim h^{2(k+1)} \mathcal{A}. \end{aligned} \quad (6.56)$$

Let us now move to the fourth term. For the sake of conciseness, we let, for all $i \in \llbracket 1, M \rrbracket$, $\psi_i := \psi_i(\mathbf{p})$, regarded as an element $H^1(0, t_F; L^2(\Omega; \mathbb{R}))$, and we

conventionally denote $\psi(\mathbf{x}, t) := \psi(t)(\mathbf{x})$ for all $t \in [0, t_F]$ and almost every $\mathbf{x} \in \Omega$. Let $i \in \llbracket 1, M \rrbracket$. It holds, for all $n \in \llbracket 1, N \rrbracket$,

$$\begin{aligned} d_t^n \psi_i - \delta_t^n \psi_i &= d_t^n \psi_i - \frac{1}{\tau} \int_{t^{n-1}}^{t^n} d_t \psi_i(t) \, dt \\ &= d_t^n \psi_i - \frac{1}{\tau} \int_{t^{n-1}}^{t^n} \left(d_t^n \psi_i - \int_t^{t^n} d_t^2 \psi_i(s) \, ds \right) \, dt \\ &= \frac{1}{\tau} \int_{t^{n-1}}^{t^n} \int_t^{t^n} d_t^2 \psi_i(s) \, ds \, dt \leq \int_{t^{n-1}}^{t^n} |d_t^2 \psi_i(t)| \, dt. \end{aligned}$$

Combining this result with the Jensen inequality, we infer

$$\begin{aligned} \|d_t^n \psi_i - \delta_t^n \psi_i\|_{L^2(\Omega; \mathbb{R})}^2 &\leq \int_{\Omega} \left| \int_{t^{n-1}}^{t^n} |d_t^2 \psi_i(\mathbf{x}, t)| \, dt \right|^2 \, d\mathbf{x} \\ &\leq \tau \int_{t^{n-1}}^{t^n} \|d_t^2 \psi_i(t)\|_{L^2(\Omega; \mathbb{R})}^2 \, dt \\ &\leq \tau \|\psi_i\|_{H^2(t^{n-1}, t^n; L^2(\Omega; \mathbb{R}))}^2. \end{aligned} \tag{6.57}$$

We next write, for all $n \in \llbracket 1, N \rrbracket$, all $i \in \llbracket 1, M \rrbracket$, and all $\underline{q}_{h,i} \in \underline{P}_{h,i}^k$,

$$\begin{aligned} |(d_t^n \psi_i - \delta_t^n \psi_i, q_{h,i})| &\leq \|d_t^n \psi_i - \delta_t^n \psi_i\|_{L^2(\Omega; \mathbb{R})} \|q_{h,i}\|_{L^2(\Omega; \mathbb{R})} \\ &\leq \tau^{\frac{1}{2}} \|\psi_i\|_{H^2(t^{n-1}, t^n; L^2(\Omega; \mathbb{R}))} \|q_{h,i}\|_{L^2(\Omega; \mathbb{R})} \\ &\lesssim \tau^{\frac{1}{2}} \|\psi_i\|_{H^2(t^{n-1}, t^n; L^2(\Omega; \mathbb{R}))} \|\underline{q}_{h,i}\|_{c,h,hho}, \end{aligned}$$

where we have used a Cauchy–Schwarz inequality in the first line, the bound (6.57) in the second line, and a discrete global Poincaré inequality in HHO spaces (resulting from a combination of [19, Proposition 5.4] and [26, Lemma 4]) to conclude. Using the above estimate in conjunction with the definition (6.15) of the dual norm, we have that

$$\|(d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p}), \cdot)\|_{c,h,*}^2 \lesssim \tau \|\psi_i(\mathbf{p})\|_{H^2(t^{n-1}, t^n; L^2(\Omega; \mathbb{R}))}^2.$$

Using this bound, we obtain

$$\begin{aligned} \mathfrak{T}_4 &\lesssim \sum_{i=1}^M \sum_{n=1}^N \tau^2 K_i^{-1} \|\psi_i(\mathbf{p})\|_{H^2(t^{n-1}, t^n; L^2(\Omega; \mathbb{R}))}^2 \\ &= \tau^2 \sum_{i=1}^N K_i^{-1} \|\psi_i(\mathbf{p})\|_{H^2(0, t_F; L^2(\Omega; \mathbb{R}))}^2 = \tau^2 \mathcal{B}. \end{aligned} \tag{6.58}$$

Plugging (6.54)–(6.58) into (6.53) yields (6.52).

6.4.4 Error Estimate for the HHO-DG Scheme

The proof of the error estimate for the HHO-DG scheme follows by adapting the arguments used in Theorem 6.1 to a different choice of the interpolates of the continuous pressures in (6.48). For all $n \in \llbracket 0, N \rrbracket$ and all $i \in \llbracket 1, M \rrbracket$, we set

$$\epsilon_{h,i}^n := P_{h,i}^n - \hat{P}_{h,i}^n,$$

where $\hat{P}_{h,i}^0 := \pi_h^k p_i^0$ and, for $n \geq 1$, $\hat{P}_{h,i}^n$ is the solution of problem (6.20) with $r = p_i^n$.

Theorem 6.2 (Error estimate for the HHO-DG scheme) *Assume $k \geq 1$, Ω convex, and the additional regularity*

$$\begin{aligned} \mathbf{u} &\in H^1(0, t_F; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d)), \\ p_0 &\in H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R})), \\ \psi_0(\mathbf{p}) &\in H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R})) \\ \forall i \in \llbracket 1, M \rrbracket, \quad S_i(\mathbf{p}) &\in C^0(\llbracket 0, t_F \rrbracket; H^{k+1}(\mathcal{T}_h; \mathbb{R})), \\ \forall i \in \llbracket 1, M \rrbracket, \quad \psi_i(\mathbf{p}) &\in H^2(0, t_F; L^2(\Omega; \mathbb{R})) \cap H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R})), \end{aligned}$$

with $\psi_0(\mathbf{p}) := \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{p} - p_0)$. Then, for a time step τ small enough (with threshold independent of h), it holds that

$$\begin{aligned} \max_{n \in \llbracket 1, N \rrbracket} &\left(\mu \|\underline{\mathbf{e}}_h^n\|_{a,h}^2 + \lambda^{-1} \|\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h^n\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|\epsilon_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 + \frac{\beta^2}{\mu} \|\epsilon_{h,0}^n\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\ &+ \sum_{n=1}^N \tau \|\boldsymbol{\epsilon}_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|\epsilon_{h,i}^n\|_{c,h,dg}^2 \lesssim h^{2(k+1)} \mathcal{A}^{dg} + \tau^2 \mathcal{B}^{dg}, \end{aligned} \quad (6.59)$$

where the hidden constant is independent of h , τ , of the problem data, of \mathbf{u} , and of p_i , $i \in \llbracket 0, M \rrbracket$, but possibly depends on Ω , t_F , k , and we have set

$$\begin{aligned} \mathcal{A}^{dg} &:= \|\mathbf{u}\|_{H^1(0, t_F; H^{k+2}(\mathcal{T}_h; \mathbb{R}^d))}^2 + \mu^{-1} \|p_0\|_{H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R}^d))}^2 \\ &+ \sum_{i=0}^M \lambda \alpha_i^{-2} \|\psi_i(\mathbf{p})\|_{H^1(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R}))}^2 + \sum_{i=1}^M \lambda \alpha_i^{-2} \|S_i(\mathbf{p})\|_{L^2(0, t_F; H^{k+1}(\mathcal{T}_h; \mathbb{R}))}^2, \\ \mathcal{B}^{dg} &:= \sum_{i=1}^M \lambda \alpha_i^{-2} \|\psi_i(\mathbf{p})\|_{H^2(0, t_F; L^2(\Omega; \mathbb{R}))}^2. \end{aligned}$$

Proof Proceeding as in the proof of Proposition 6.1 and recalling the definition of the elliptic projection in (6.20), it is readily inferred that

$$\underline{\mathbf{e}}_h^0 = \mathbf{0}, \quad \epsilon_{h,i}^0 = 0, \quad \forall i \in \llbracket 0, M \rrbracket \quad (6.60a)$$

and, for $n \in \llbracket 1, N \rrbracket$, it holds, for all $\mathbf{v}_h \in \underline{\mathbf{U}}_h^k$, all $q_{h,0} \in P_{h,0}^k$,

$$2\mu \mathbf{a}_h(\underline{\mathbf{e}}_h^n, \mathbf{v}_h) + \mathbf{b}_h(\mathbf{v}_h, \epsilon_{h,0}^n) = 2\mu \mathcal{E}_{a,h}(\mathbf{u}^n; \mathbf{v}_h) + \mathcal{E}_{b,h}(p_0^n; \mathbf{v}_h), \quad (6.60b)$$

$$\mathbf{b}_h(\delta_t^n \underline{\mathbf{e}}_h, q_{h,0}) - \lambda^{-1}(\delta_t^n(\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h), q_{h,0}) = -(\delta_t^n(\psi_0(\mathbf{p} - \hat{\mathbf{p}}_h), q_{h,0})), \quad (6.60c)$$

and, for all $i \in \llbracket 1, M \rrbracket$ and $q_{h,i} \in P_{h,i}^k$,

$$\begin{aligned} & (\delta_t^n \psi_i(\boldsymbol{\epsilon}_h), q_{h,i}) + (S_i(\boldsymbol{\epsilon}_h^n), q_{h,i}) + K_i \mathbf{c}_h^{\text{dg}}(\boldsymbol{\epsilon}_{h,i}^n, q_{h,i}) \\ &= (S_i(\mathbf{p}^n - \hat{\mathbf{p}}_h^n), q_{h,i}) + (d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p}), q_{h,i}) + (\delta_t^n \psi_i(\mathbf{p} - \hat{\mathbf{p}}_h), q_{h,i}), \end{aligned} \quad (6.60d)$$

where, in (6.60c), we have applied discrete time derivation and introduced the linear function ψ_0 defined such that, for all $\mathbf{q} \in \mathbb{R}^{M+1}$, $\psi_0(\mathbf{q}) := \lambda^{-1}(\boldsymbol{\alpha} \cdot \mathbf{q} - q_0)$. Then, following the first two step of the proof of Lemma 6.1 we obtain an estimate similar to (6.34), namely, for an arbitrary $\mathbf{N} \in \llbracket 1, N \rrbracket$ it holds

$$\begin{aligned} & \mu \|\underline{\mathbf{e}}_h^{\mathbf{N}}\|_{a,h}^2 + \frac{\|\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h^{\mathbf{N}}\|_{L^2(\Omega; \mathbb{R})}^2}{2\lambda} + \sum_{i=1}^M \frac{C_i}{2} \|\epsilon_{h,i}^{\mathbf{N}}\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{n=1}^{\mathbf{N}} \frac{\tau}{2} \|\epsilon_h^n\|_{\xi}^2 + \sum_{i=1}^M \sum_{n=1}^{\mathbf{N}} \tau K_i \|\epsilon_{h,i}^n\|_{c,h,\text{dg}}^2 \\ & \leq \sum_{n=1}^{\mathbf{N}} \tau (2\mu \mathcal{E}_{a,h}(\mathbf{u}^n; \delta_t^n \underline{\mathbf{e}}_h) + \mathcal{E}_{b,h}(p_0^n; \delta_t^n \underline{\mathbf{e}}_h)) + \sum_{i=0}^M \sum_{n=1}^{\mathbf{N}} \tau (\mathcal{E}_{i,h}^n(\mathbf{p}), \epsilon_{h,i}^n), \end{aligned} \quad (6.61)$$

with $\mathcal{E}_{0,h}^n(\mathbf{p}) := \delta_t^n \psi_0(\mathbf{p} - \hat{\mathbf{p}}_h)$ and, for all $i \in \llbracket 1, M \rrbracket$,

$$\mathcal{E}_{i,h}^n(\mathbf{p}) := (d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p})) + S_i(\mathbf{p}^n - \hat{\mathbf{p}}_h^n) + \delta_t^n \psi_i(\mathbf{p} - \hat{\mathbf{p}}_h).$$

The first term in the right-hand side of (6.61) can be bounded as in (6.35). We bound the second term by using the Cauchy–Schwarz and Young inequality to write

$$\sum_{i=0}^M \sum_{n=1}^{\mathbf{N}} \tau (\mathcal{E}_{i,h}^n(\mathbf{p}), \epsilon_{h,i}^n) \leq \sum_{i=0}^M \sum_{n=1}^{\mathbf{N}} \frac{\tau \lambda}{2\alpha_i^2} \|\mathcal{E}_{i,h}^n(\mathbf{p})\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{n=1}^{\mathbf{N}} \frac{\tau}{2\lambda} \|\boldsymbol{\alpha} \cdot \boldsymbol{\epsilon}_h^n\|_{L^2(\Omega; \mathbb{R})}^2.$$

Therefore, proceeding as in steps (i.C) and (ii) of Lemma 6.1, yields

$$\begin{aligned}
& \max_{n \in \llbracket 1, N \rrbracket} \left(\mu \|e_h^n\|_{a,h}^2 + \lambda^{-1} \|\alpha \cdot \epsilon_h^n\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M C_i \|\epsilon_{h,i}^n\|_{L^2(\Omega; \mathbb{R})}^2 + \frac{\beta^2}{\mu} \|\epsilon_{h,0}^n\|_{L^2(\Omega; \mathbb{R})}^2 \right) \\
& + \sum_{n=1}^N \tau \|\epsilon_h^n\|_{\xi}^2 + 2 \sum_{i=1}^M \sum_{n=1}^N \tau K_i \|\epsilon_{h,i}^n\|_{c,h,\text{dg}}^2 \lesssim \exp\left(\frac{t_F}{1-\tau}\right) (\mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3^{\text{dg}} + \mathfrak{T}_4^{\text{dg}}),
\end{aligned} \tag{6.62}$$

where

$$\begin{aligned}
\mathfrak{T}_3^{\text{dg}} &:= \sum_{n=1}^N \tau \left(\sum_{i=0}^M \lambda \alpha_i^{-2} \|\delta_t^n \psi_i(\mathbf{p} - \hat{\mathbf{p}}_h)\|_{L^2(\Omega; \mathbb{R})}^2 + \sum_{i=1}^M \lambda \alpha_i^{-2} \|S_i(\mathbf{p}^n - \hat{\mathbf{p}}_h^n)\|_{L^2(\Omega; \mathbb{R})}^2 \right), \\
\mathfrak{T}_4^{\text{dg}} &:= \sum_{i=1}^M \sum_{n=1}^N \tau \lambda \alpha_i^{-2} \|d_t^n \psi_i(\mathbf{p}) - \delta_t^n \psi_i(\mathbf{p})\|_{L^2(\Omega; \mathbb{R})}^2,
\end{aligned}$$

and the terms \mathfrak{T}_1 and \mathfrak{T}_2 are defined in (6.53) and bounded in (6.54) and (6.55), respectively. The term $\mathfrak{T}_4^{\text{dg}}$ can be bounded using (6.57) and (6.58) to obtain $\mathfrak{T}_4^{\text{dg}} \lesssim \tau^2 \mathcal{B}^{\text{dg}}$. Hence, it only remains to bound $\mathfrak{T}_3^{\text{dg}}$. Owing to the linearity of the backward time derivative δ_t^n and the functions ψ_i and S_i for all $i \in \llbracket 1, M \rrbracket$, the approximation property (6.21) of the elliptic projection, and the boundedness property (6.5), we infer

$$\begin{aligned}
\mathfrak{T}_3^{\text{dg}} &\lesssim h^{2(k+1)} \sum_{n=1}^N \tau \left(\sum_{i=0}^M \lambda \alpha_i^{-2} \|\delta_t^n \psi_i(\mathbf{p})\|_{H^{k+1}(\mathcal{T}_h; \mathbb{R})}^2 + \sum_{i=1}^M \lambda \alpha_i^{-2} \|S_i(\mathbf{p}^n)\|_{H^{k+1}(\mathcal{T}_h; \mathbb{R})}^2 \right) \\
&\lesssim h^{2(k+1)} \mathcal{A}^{\text{dg}}.
\end{aligned}$$

Combining the previous bounds with (6.62) leads to the conclusion.

Table 6.1 Model parameters

Parameter	Unit	Set <i>i</i>	Set <i>ii</i>	Set <i>iii</i>	Set <i>iv</i>
μ	MPa	4.2	4.2	4.2	4.2
λ	MPa	2.4	$2.4 \cdot 10^5$	2.4	2.4
α_1	–	0.95	0.95	0.95	0.95
α_2	–	0.12	0.12	0.12	0.12
C_1	MPa^{-1}	0.054	0.054	0.0	0.054
C_2	MPa^{-1}	0.014	0.014	0.0	0.014
K_1	$\text{m}^2 \text{MPa}^{-1} \text{s}^{-1}$	$6.18 \cdot 10^{-6}$	$6.18 \cdot 10^{-6}$	$6.18 \cdot 10^{-6}$	10^{-12}
K_2	$\text{m}^2 \text{MPa}^{-1} \text{s}^{-1}$	$2.72 \cdot 10^{-5}$	$2.72 \cdot 10^{-5}$	$2.72 \cdot 10^{-5}$	10^{-11}
$\xi_{1 \leftarrow 2}$	$\text{MPa}^{-1} \text{s}^{-1}$	0.01	0.01	0.01	0.01

Table 6.2 Convergence rates for the HHO-DG discretisation with polynomial degree $k = 1$ based on manufactured solutions of the Barenblatt–Biot problem, see text for details

Set	$\ \underline{e}_{h\tau}\ _{\infty,1}$	EOC	$\ \epsilon_{0,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{1,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{2,h\tau}\ _{\infty,0}$	EOC
<i>i</i>	2.39e−01	–	5.60e−01	–	4.78e−01	–	2.48e−01	–
	6.23e−02	1.94	1.11e−01	2.24	9.31e−02	2.36	4.80e−02	2.37
	1.51e−02	2.05	2.28e−02	2.28	1.88e−02	2.31	1.01e−02	2.24
	3.73e−03	2.01	4.92e−03	2.21	3.83e−03	2.29	2.52e−03	2.01
	9.39e−04	1.99	1.08e−03	2.19	7.55e−04	2.34	6.28e−04	2.00
<i>ii</i>	2.43e−01	–	8.25e−01	–	1.43e−01	–	1.32e−01	–
	6.26e−02	1.95	1.55e−01	2.41	3.76e−02	1.92	3.86e−02	1.77
	1.51e−02	2.05	3.09e−02	2.33	9.16e−03	2.04	9.52e−03	2.02
	3.73e−03	2.02	6.84e−03	2.18	2.34e−03	1.97	2.49e−03	1.93
	9.35e−04	2.00	1.71e−03	2.00	6.04e−04	1.95	6.27e−04	1.99
<i>iii</i>	2.39e−01	–	5.67e−01	–	4.79e−01	–	3.08e−01	–
	6.23e−02	1.94	1.14e−01	2.31	9.43e−02	2.34	6.48e−02	2.25
	1.51e−02	2.05	2.40e−02	2.24	1.97e−02	2.26	1.40e−02	2.21
	3.73e−03	2.01	5.50e−03	2.13	4.45e−03	2.15	3.27e−03	2.10
	9.35e−04	2.00	1.38e−03	1.99	1.12e−03	1.99	8.19e−04	2.00
<i>iv</i>	2.42e−01	–	8.00e−01	–	7.78e−01	–	4.14e−01	–
	6.25e−02	1.95	1.46e−01	2.46	1.41e−01	2.47	6.28e−02	2.72
	1.51e−02	2.05	2.79e−02	2.39	2.62e−02	2.43	1.11e−02	2.50
	3.73e−03	2.01	5.58e−03	2.32	4.88e−03	2.42	2.61e−03	2.09
	9.39e−04	1.99	1.12e−03	2.31	8.43e−04	2.53	6.40e−04	2.03

6.5 Numerical Tests

In this section, we present some numerical examples to illustrate the theoretical results. In order to confirm the convergence rates predicted in Theorem 6.2, we rely on a manufactured smooth solution of a two-network poroelasticity problem (i.e. the Barenblatt–Biot problem) on the unit square domain $\Omega = (0, 1)^2$ and time interval $[0, t_F = 1)$. The exact displacement \mathbf{u} and exact pressures p_1 and p_2 are given by,

$$\mathbf{u}(\mathbf{x}, t) = \sin(\pi t) \begin{pmatrix} -\cos(\pi x_1) \cos(\pi x_2) \\ \sin(\pi x_1) \sin(\pi x_2) \end{pmatrix},$$

$$p_1(\mathbf{x}, t) = \pi \sin(\pi t) [\sin(\pi x_1) \cos(\pi x_2) + \cos(\pi x_1) \sin(\pi x_2)],$$

$$p_2(\mathbf{x}, t) = \pi \sin(\pi t) [\sin(\pi x_1) \cos(\pi x_2) - \cos(\pi x_1) \sin(\pi x_2)].$$

The total pressure p_0 , volumetric load \mathbf{f} , and source terms g_1 and g_2 are inferred from the exact solution. In order to assess the robustness with respect to the model coefficients, we consider the four sets of parameters depicted in Table 6.1. The first set of model parameters is taken from [29]. The second, third, and fourth sets are meant

Table 6.3 Convergence rates for the HHO-DG discretisation with polynomial degree $k = 2$ based on manufactured solutions of the Barenblatt–Biot problem, see text for details

Set	$\ \underline{e}_{h\tau}\ _{\infty,1}$	EOC	$\ \epsilon_{0,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{1,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{2,h\tau}\ _{\infty,0}$	EOC
<i>i</i>	3.29e−02	–	8.38e−02	–	7.16e−02	–	3.31e−02	–
	4.05e−03	3.02	7.36e−03	3.51	6.15e−03	3.54	2.65e−03	3.64
	5.40e−04	2.91	8.04e−04	3.19	6.15e−04	3.32	3.48e−04	2.93
	6.93e−05	2.96	8.58e−05	3.23	5.70e−05	3.43	4.55e−05	2.93
	8.68e−06	3.00	9.43e−06	3.19	5.68e−06	3.33	5.68e−06	3.00
<i>ii</i>	3.36e−02	–	1.22e−01	–	1.69e−02	–	2.16e−02	–
	4.05e−03	3.05	9.88e−03	3.63	2.33e−03	2.86	2.46e−03	3.13
	5.37e−04	2.91	1.17e−03	3.08	3.21e−04	2.86	3.47e−04	2.83
	6.82e−05	2.98	1.46e−04	3.00	4.20e−05	2.94	4.56e−05	2.93
	8.52e−06	3.00	1.81e−05	3.01	5.52e−06	2.93	5.69e−06	3.00
<i>iii</i>	3.29e−02	–	8.61e−02	–	7.23e−02	–	4.77e−02	–
	4.04e−03	3.02	7.84e−03	3.46	6.56e−03	3.46	4.39e−03	3.44
	5.38e−04	2.91	9.51e−04	3.04	7.90e−04	3.05	5.39e−04	3.02
	6.83e−05	2.98	1.20e−04	2.99	9.90e−05	3.00	6.81e−05	2.99
	8.54e−06	3.00	1.49e−05	3.01	1.23e−05	3.01	8.45e−06	3.01
<i>iv</i>	3.35e−02	–	1.14e−01	–	1.12e−01	–	4.67e−02	–
	4.05e−03	3.05	8.78e−03	3.71	8.36e−03	3.75	2.90e−03	4.01
	5.40e−04	2.91	8.94e−04	3.30	7.69e−04	3.44	3.61e−04	3.01
	6.93e−05	2.96	8.97e−05	3.32	6.56e−05	3.55	4.59e−05	2.98
	8.68e−06	3.00	9.45e−06	3.25	5.80e−06	3.50	5.69e−06	3.01

to check the robustness of the method in the nearly incompressible case (i.e. large values of λ), in the vanishing storage coefficients case, and in the small permeabilities case, respectively. We remark that the value of μ and λ considered in the second test corresponds to a Poisson ratio $\nu = 0.49999$.

We consider the HHO method described in Sect. 6.3 with DG discretisation of the Darcy term for polynomial degrees $k \in \{1, 2, 3\}$ over a trapezoidal elements mesh sequence $(\mathcal{T}_h)_j$ with 2^{2+2j} elements, for $j \in \llbracket 1, 5 \rrbracket$. The time discretisation is based on Backward Differentiation Formulas (BDF) of order $(k + 1)$ with a fixed time step $\tau = 10^{-3}$. The boundary conditions are inferred from the exact solution. On the bottom edge $\{\mathbf{x} \in \partial\Omega : x_2 = 0\}$, we enforce Dirichlet conditions for the displacement and Neumann conditions for both the network pressures p_1 and p_2 . On the rest of the domain boundary we set Neumann conditions for the displacement and Dirichlet for the two pressures. Initial conditions are specified by means of L^2 -projections over mesh elements according to (6.22). Initialisation is performed at several time points ($t_i = -\tau i$, $i = 1, \dots, k + 1$), in agreement with the BDF order.

In Tables 6.2, 6.3 and 6.4 we report the convergence rates for the four set of model parameters indicated in Table 6.1. We use the following shorthand notations for the error measures:

Table 6.4 Convergence rates for the HHO-DG discretisation with polynomial degree $k = 3$ based on manufactured solutions of the Barenblatt–Biot problem, see text for details

Set	$\ \mathbf{e}_{h\tau}\ _{\infty,1}$	EOC	$\ \epsilon_{0,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{1,h\tau}\ _{\infty,0}$	EOC	$\ \epsilon_{2,h\tau}\ _{\infty,0}$	EOC
<i>i</i>	3.30e−03	–	8.57e−03	–	7.41e−03	–	2.77e−03	–
	2.42e−04	3.77	5.34e−04	4.00	4.48e−04	4.05	1.66e−04	4.06
	1.42e−05	4.09	2.64e−05	4.34	2.03e−05	4.46	9.44e−06	4.14
	9.26e−07	3.94	1.41e−06	4.23	8.87e−07	4.52	6.29e−07	3.91
	5.79e−08	4.00	7.49e−08	4.24	3.89e−08	4.51	3.89e−08	4.02
<i>ii</i>	3.36e−03	–	1.19e−02	–	1.94e−03	–	1.83e−03	–
	2.43e−04	3.79	7.14e−04	4.06	1.42e−04	3.77	1.57e−04	3.54
	1.42e−05	4.10	3.83e−05	4.22	8.91e−06	4.00	9.39e−06	4.07
	9.14e−07	3.96	2.37e−06	4.01	5.94e−07	3.91	6.28e−07	3.90
	5.66e−08	4.01	1.45e−07	4.03	3.83e−08	3.96	3.89e−08	4.01
<i>iii</i>	3.31e−03	–	8.94e−03	–	7.62e−03	–	4.80e−03	–
	2.42e−04	3.77	5.78e−04	3.95	4.88e−04	3.97	3.17e−04	3.92
	1.42e−05	4.10	3.15e−05	4.20	2.65e−05	4.20	1.73e−05	4.19
	9.14e−07	3.95	1.99e−06	3.99	1.67e−06	3.99	1.10e−06	3.98
	5.67e−08	4.01	1.22e−07	4.02	1.03e−07	4.02	6.78e−08	4.02
<i>iv</i>	3.34e−03	–	1.09e−02	–	1.08e−02	–	3.25e−03	–
	2.42e−04	3.78	6.23e−04	4.13	5.95e−04	4.18	1.78e−04	4.19
	1.42e−05	4.09	2.91e−05	4.42	2.53e−05	4.56	9.62e−06	4.21
	9.27e−07	3.94	1.45e−06	4.33	9.93e−07	4.67	6.31e−07	3.93
	5.79e−08	4.00	7.47e−08	4.28	3.94e−08	4.66	3.89e−08	4.02

$$\|\mathbf{e}_{h\tau}\|_{\infty,1} := \max_{n \in \llbracket 1, N \rrbracket} \|\mathbf{u}_h^n - \mathbf{I}_h^k \mathbf{u}^n\|_{a,h},$$

$$\|\epsilon_{i,h\tau}\|_{\infty,0} := \max_{n \in \llbracket 1, N \rrbracket} \|p_{i,h}^n - \pi_h^k p_i^n\|_{L^2(\Omega; \mathbb{R})}, \quad \forall i \in \llbracket 0, 2 \rrbracket.$$

Each error measure is accompanied by the corresponding estimated order of convergence (EOC). The observed convergence rates are in agreement with the error estimate of Theorem 6.2. We remark that the performance is not affected by the different choices of the model parameters. Hence, the method is robust in all the limit cases of vanishing storage, nearly incompressible, and poorly permeable media.

Acknowledgements M. Botti acknowledges funding from the European Commission through the H2020-MSCA-IF-EF project PDGeoFF, Polyhedral Discretisation Methods for Geomechanical Simulation of Faults and Fractures in Poroelastic Media (Grant no. 896616).

References

1. J. Aghili, S. Boyaval, D.A. Di Pietro, Hybridization of mixed high-order methods on general meshes and application to the Stokes equations. *Comput. Methods Appl. Math.* **15**(2), 111–134 (2015)
2. D. Anderson, J. Droniou, An arbitrary order scheme on generic meshes for miscible displacements in porous media. *SIAM J. Sci. Comput.* **40**(4), B1020–B1054 (2018)
3. P.F. Antonietti, C. Facciola, A. Russo, M. Verani, Discontinuous Galerkin approximation of flows in fractured porous media on polytopic grids. *SIAM J. Sci. Comput.* **41**(1), A109–A138 (2019)
4. G.I. Barenblatt, Iu.P. Zheltov, I.N. Kochina, Basic concepts in the theory of seepage of homogeneous liquids in fissured rocks. *J. Appl. Math. Mech.* **24**, 1286–1303 (1960)
5. F. Bassi, L. Botti, A. Colombo, D.A. Di Pietro, P. Tesini, On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *J. Comput. Phys.* **231**(1), 45–65 (2012)
6. S. Berrone, A. Borio, C. Fidelibus, S. Pieraccini, S. Scialò, F. Vicini, Advanced computation of steady-state fluid flow in discrete fracture-matrix models: FEM-BEM and VEM-VEM fracture-block coupling. *GEM Int. J. Geomath.* **9**(2), 377–399 (2018)
7. M.A. Biot, General theory of three dimensional consolidation. *J. Appl. Phys.* **12**(2), 155–164 (1941)
8. M.A. Biot, Theory of elasticity and consolidation for a porous anisotropic solid. *J. Appl. Phys.* **26**(2), 182–185 (1955)
9. D. Boffi, M. Botti, D.A. Di Pietro, A nonconforming high-order method for the Biot problem on general meshes. *SIAM J. Sci. Comput.* **38**(3), A1508–A1537 (2016)
10. L. Botti, M. Botti, D.A. Di Pietro, An abstract analysis framework for monolithic discretisations of poroelasticity with application to Hybrid High-Order methods (2020). Published online. <https://dx.doi.org/10.1016/j.camwa.2020.06.004>
11. L. Botti, D.A. Di Pietro, J. Droniou, A Hybrid High-Order discretisation of the Brinkman problem robust in the Darcy and Stokes limits. *Comput. Methods Appl. Mech. Eng.* **341**, 278–310 (2018)
12. M. Botti, D.A. Di Pietro, A. Guglielmana, A low-order nonconforming method for linear elasticity on general meshes. *Comput. Methods Appl. Mech. Eng.* **354**, 96–118 (2019)
13. M. Botti, D.A. Di Pietro, P. Sochala, A Hybrid High-Order method for nonlinear elasticity. *SIAM J. Numer. Anal.* **55**(6), 2687–2717 (2017)
14. M. Botti, D.A. Di Pietro, P. Sochala, A Hybrid High-Order discretisation method for nonlinear poroelasticity. *Comput. Methods Appl. Math.* **20**(2), 227–249 (2020)
15. K. Brenner, M. Groza, C. Guichard, R. Masson, Vertex approximate gradient scheme for hybrid dimensional two-phase Darcy flows in fractured porous media. *ESAIM Math. Model. Numer. Anal.* **49**(2), 303–330 (2015)
16. F. Chave, D.A. Di Pietro, L. Formaggia, A Hybrid High-Order method for Darcy flows in fractured porous media. *SIAM J. Sci. Comput.* **40**(2), A1063–A1094 (2018)
17. F. Chave, D.A. Di Pietro, L. Formaggia, A Hybrid High-Order method for passive transport in fractured porous media. *Int. J. Geomath.* **10**(12) (2019)
18. O. Coussy, *Poromechanics* (Wiley, 2004)
19. D.A. Di Pietro, J. Droniou, A Hybrid High-Order method for Leray-Lions elliptic equations on general meshes. *Math. Comput.* **86**(307), 2159–2191 (2017)
20. D.A. Di Pietro, J. Droniou, A third Strang lemma for schemes in fully discrete formulation. *Calcolo* **55**(40) (2018)
21. D.A. Di Pietro, J. Droniou, A third Strang lemma for schemes in fully discrete formulation, 4 2018. Preprint arXiv 1804.09484. <https://arxiv.org/abs/1804.09484>. Contains an additional Appendix with respect to the published paper
22. D.A. Di Pietro, J. Droniou, *The Hybrid High-Order Method for Polytopal Meshes*. Number 19 in Modeling, Simulation and Application (Springer International Publishing, 2020). <https://dx.doi.org/10.1007/978-3-030-37203-3>

23. D.A. Di Pietro, A. Ern, A hybrid high-order locking-free method for linear elasticity on general meshes. *Comput. Methods Appl. Mech. Eng.* **283**, 1–21 (2015)
24. D.A. Di Pietro, A. Ern, Arbitrary-order mixed methods for heterogeneous anisotropic diffusion on general meshes. *IMA J. Numer. Anal.* **37**(1), 40–63 (2017)
25. D.A. Di Pietro, A. Ern, J.-L. Guermond, Discontinuous Galerkin methods for anisotropic semidefinite diffusion with advection. *SIAM J. Numer. Anal.* **46**(2), 805–831 (2008)
26. D.A. Di Pietro, A. Ern, S. Lemaire, An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. *Comput. Meth. Appl. Math.* **14**(4), 461–472 (2014)
27. R. Eymard, T. Gallouët, C. Guichard, R. Herbin, R. Masson, TP or not TP, that is the question. *Comput. Geosci.* **18**(3–4), 285–296 (2014)
28. J.G. Heywood, R. Rannacher, Finite-element approximation of the nonstationary Navier-Stokes problem. IV. Error analysis for second-order time discretization. *SIAM J. Numer. Anal.* **27**(2), 353–384 (1990)
29. A.E. Kolesov, P.N. Vabishchevich, Splitting schemes with respect to physical processes for double-porosity poroelasticity problems (2016)
30. J.J. Lee, E. Piersanti, K.-A. Mardal, M.E. Rognes, A mixed finite element method for nearly incompressible multiple-network poroelasticity. *SIAM J. Sci. Comput.* **41**(2), A722–A747 (2019)
31. J.M. Nordbotten, Stable cell-centered finite volume discretization for Biot equations. *SIAM J. Numer. Anal.* **54**(2), 942–968 (2016)
32. K. Terzaghi, *Theoretical Soil Mechanics* (Wiley, New York, 1943)
33. B. Tully, Y. Ventikos, Cerebral water transport using multiple-network poroelastic theory: application to normal pressure hydrocephalus. *J. Fluid Mech.* **667**, 188–215 (2011)

Chapter 7

The Mixed Virtual Element Method for the Richards Equation



Dibyendu Adak, Gianmarco Manzini, and Sundararajan Natarajan

Abstract The time-dependent Richards equation can be reformulated as a nonlinear, possibly degenerate, parabolic problem in mixed form by applying the Kirchhoff transformation. A preliminary time integration yields the variational formulation. A numerical treatment of this problem using polygonal and polyhedral meshes is, then, feasible by applying the mixed virtual element method. In this setting, we study a semi-discrete and a fully-discrete virtual element approximation. The theoretical analysis shows that our virtual element formulations are well-posed and convergent, and optimal convergence rates for the approximation errors can be proved. Such theoretical results are confirmed and the accuracy is assessed by investigating the behavior of the method on a set of representative numerical experiments.

Keywords Richard equation · Mixed virtual element method · Polygonal mesh · Low-order approximation · Convergence analysis

7.1 Introduction

The mathematical model of the water flow in an unsaturated soil under the effect of gravity and the action of capillarity is based on the Richards equation. This equation was proposed for the first time by the English mathematician and physicist Lewis F. Richardson in his book *Weather prediction by numerical process* published in 1922 (for the new edition see Ref. [77]). Nonetheless, credits were later attributed to Lorenzo A. Richards for the work in his Ph.D. thesis *Capillary conduction of liquids through porous mediums* published in 1933 [76].

D. Adak · S. Natarajan
Department of Mechanical Engineering, Indian Institute of Technology Madras, Chennai, India
e-mail: snatarajan@iitm.ac.in; sundararajan.natarajan@gmail.com

G. Manzini (✉)
Istituto di Matematica Applicata e Tecnologie Informatiche “E. Magenes”, via Ferrata 1,
27100 Pavia, Italy
e-mail: marco.manzini@imati.cnr.it

Due to the complexity of the nonlinear phenomena that are taken into account by the Richards equation, only a few simplified cases offer an analytical solution. Therefore, the numerical approach is the only option that is really available in the majority of the situations found in practice, which normally involve very different initial and boundary conditions and a wide range of soils. A significant amount of work to improve effectiveness of numerical simulation in term of robustness and reliability has been carried out in the last three decades, but the computer resolution of the Richards equation still challenges the numerical modelers. A review of the major advancements in the development of numerical methods for solving the Richards equation is beyond the scope of this chapter, but can be found in the recently published paper of Ref. [55].

In short, among the major difficulties of the numerical approximation of the Richards equation, we find that such equation can change type being elliptic in the fully saturated flow regime and parabolic in the partially saturated flow regime, and its solutions can be characterized by an important lack of regularity. Theoretical properties of the solution of degenerate nonlinear parabolic problems were studied in [4, 70, 80]. To overcome the issue of the poor regularity of the solutions, a nonlinear mixed formulation discretized by mixed finite elements was proposed in Ref. [69]. It was, then, further extended in [8], where a nonlinear mixed finite element method is proposed for a degenerate parabolic equation arising in flows in porous media. Other discretizations that provide good approximations to the solution of the Richards equation are given by relaxation schemes [58], multiscale mixed/mimetic methods on corner-point grids [1], mixed transform finite element methods for solving the nonlinear equation for flow in variably saturated porous media [10] finite volumes [54, 64], mixed finite element discretizations [81], mixed finite element discretization on non-matching multiblock grids [83], discontinuous Galerkin finite element methods [59] also with adaptivity [60]. In case of implicit schemes, iterative methods are considered for solving the resulting nonlinear equations (see, e.g., [34, 50, 56, 74]). Convergence results for implicit discretization schemes are found in [73, 75, 78].

In this paper, we consider the mixed finite element formulation that was originally proposed in [78] in the new framework of mixed virtual element method (mixed VEM) recently introduced in [16]. Such discretization can also be interpreted as a generalization of the low-order Raviart-Thomas mixed finite element method for simplexes to more general polytopal meshes in the two-dimensional (2D) and three-dimensional (3D) setting. In Chap. 8, the Mixed VEM is applied to the numerical treatment of single-phase flows in underground media.

Despite its recentness, the virtual element method has been proved to be successful in many different domains of numerical analysis of partial differential equations (PDEs). A brief historical overview and some background material is presented in the next subsection.

7.1.1 Background Material on the VEM

The VEM was originally developed as a variational reformulation of the *nodal* mimetic finite difference (MFD) method [20, 27, 41, 65] for solving diffusion problems on unstructured polygonal meshes. A survey on the MFD method can be found in the review paper [63] and the research monograph [21]. The VEM inherits the flexibility of the MFD method with respect to the admissible meshes and this feature is well reflected in the many significant applications that have been developed so far, see, for example, [5, 7, 12–19, 22–26, 28–33, 35, 37–39, 44, 45, 51, 52, 67, 68, 71, 72, 79, 82]. Moreover, the connection between the VEM and the finite elements on polygonal/polyhedral meshes is thoroughly investigated in [48, 53, 66], between VEM and discontinuous skeletal gradient discretizations in [53], and between the VEM and the BEM-based FEM method in [47]. The VEM was originally formulated in [11] as a conforming FEM for the Poisson problem. It was later extended to convection-reaction-diffusion problems with variable coefficients in [3, 18]. Meanwhile, the nonconforming formulation for diffusion problems was proposed in [9] as the finite element reformulation of [62] and later extended to general elliptic problems [36, 49], Stokes problem [46], eigenvalue problems [57], and the biharmonic equation [6, 84]. Mixed VEM for elliptic problems were introduced in [42] in a BDM-like setting and further developed in [16] in an RT-like setting. The connection with De Rham diagrams and Nedelec elements with application to electromagnetism has been explored in [15].

7.1.2 Structure of the Paper

The outline of the paper is as follows. In Sect. 7.2, we introduce the Richards equation, which we reformulate in mixed form after the Kirchhoff transformation and a preliminary time integration that yields the variational formulation. In Sect. 7.3, we discuss the virtual element approximation of the resulting nonlinear parabolic problem and formulate the semi-discrete and fully-discrete formulations. In Sect. 7.4, we investigate the convergence of both formulations, and prove optimal convergence rates. In Sect. 7.5, we assess the accuracy of the virtual element approximation by investigating the behavior of the method in solving two representative benchmark problems. In Sect. 7.6, we offer our final conclusions.

7.1.3 Notation and a Few Technical Definitions

We use the standard definition and notation of Sobolev spaces, norms and seminorms, cf. [2]. Let k be a nonnegative integer number. The Sobolev space $H^k(\omega)$ consists of all square integrable functions with all square integrable weak derivatives up to order

k that are defined on the open, bounded, connected subset ω of \mathbb{R}^d , $d = 2, 3$. As usual, if $k = 0$, we prefer the notation $L^2(\omega)$. Norm and seminorm in $H^k(\omega)$ are denoted by $\|\cdot\|_{k,\omega}$ and $|\cdot|_{k,\omega}$, respectively, and $(\cdot, \cdot)_\omega$ denote the L^2 -inner product. We omit the subscript ω in the L^2 -inner product notation when ω is the whole computational domain Ω . In a few situations, for the sake of clarity, we may prefer to use the integral notation of the inner product.

We denote the linear space of polynomials of degree up to ℓ defined on ω by $\mathbb{P}_\ell(\omega)$, with the useful conventional notation that $\mathbb{P}_{-1}(\omega) = \{0\}$. Space $\mathbb{P}_\ell(\omega)$ is the span of the finite set of *scaled monomials of degree up to ℓ* , that are given by

$$\mathcal{M}_\ell(\omega) = \left\{ \left(\frac{\mathbf{x} - \mathbf{x}_\omega}{h_\omega} \right)^\alpha \text{ with } |\alpha| \leq \ell \right\},$$

where

- \mathbf{x}_ω denotes the center of gravity of ω and h_ω its characteristic length, as, for instance, the edge length or the cell diameter for $d = 2$ and 3 , respectively;
- $\alpha = (\alpha_1, \dots, \alpha_d)$ is the d -dimensional multi-index of nonnegative integers α_i with degree $|\alpha| = \alpha_1 + \dots + \alpha_d \leq \ell$ and such that $\mathbf{x}^\alpha = x_1^{\alpha_1} \dots x_d^{\alpha_d}$ for any $\mathbf{x} \in \mathbb{R}^d$ and $\partial^{|\alpha|}/\partial \mathbf{x}^\alpha = \partial^{|\alpha|}/\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}$.

We will also use the set of *scaled monomials of degree exactly equal to ℓ* , denoted by $\mathcal{M}_\ell^*(\omega)$ and obtained by setting $|\alpha| = \ell$ in the definition above. The dimension of $\mathbb{P}_\ell(\omega)$ equals N_ℓ , the cardinality of the basis set $\mathcal{M}_\ell(\omega)$. For $d = 2$, it holds $N_\ell = (\ell + 1)(\ell + 2)/2$; for $d = 3$, it holds $N_\ell = (\ell + 1)(\ell + 2)(\ell + 3)/6$.

To ease the exposition, we assume that Ω is a polytopal domain, i.e., a polygon for $d = 2$ and a polyhedron for $d = 3$. We consider a family of mesh decompositions denoted by $\mathcal{T} = \{\Omega_h\}_h$, where each mesh Ω_h is a set of non-overlapping, bounded elements \mathbf{P} such that $\overline{\Omega} = \cup_{\mathbf{P} \in \Omega_h} \overline{\mathbf{P}}$. The subindex h that labels each mesh Ω_h is the supremum of the diameters $h_{\mathbf{P}} = \sup_{\mathbf{x}, \mathbf{y} \in \mathbf{P}} |\mathbf{x} - \mathbf{y}|$ of the elements of Ω_h . For $d = 2$, each element \mathbf{P} has a non-intersecting *polygonal* boundary $\partial \mathbf{P}$ formed by $N_{\mathbf{P}}^{\mathcal{E}}$ straight edges e connecting $N_{\mathbf{P}}^{\mathcal{V}}$ vertices; note that $N_{\mathbf{P}}^{\mathcal{V}} = N_{\mathbf{P}}^{\mathcal{E}}$. In this case, the sequence of vertices forming $\partial \mathbf{P}$ is oriented in the counter-clockwise order and the vertex coordinates are denoted by $\mathbf{x}_i := (x_i, y_i)$, $i = 1, 2, \dots, N^{\mathcal{V}}$. For $d = 3$, each element \mathbf{P} has a non-intersecting *polyhedral* boundary $\partial \mathbf{P}$ formed by $N_{\mathbf{P}}^{\mathcal{F}}$ planar faces f connecting $N_{\mathbf{P}}^{\mathcal{V}}$ vertices with coordinates $\mathbf{x}_i := (x_i, y_i, z_i)$, $i = 1, 2, \dots, N^{\mathcal{V}}$. We denote the measure of \mathbf{P} by $|\mathbf{P}|$ and its barycenter (center of gravity) by $\mathbf{x}_{\mathbf{P}} := (x_{\mathbf{P}}, y_{\mathbf{P}})$ when $d = 2$ or $\mathbf{x}_{\mathbf{P}} := (x_{\mathbf{P}}, y_{\mathbf{P}}, z_{\mathbf{P}})$ when $d = 3$. Hereafter, we will refer to a 2D edge and a 3D face as the “mesh face” (or, simply, the “face”), and we will try to be dimension independent if not otherwise specified. We denote the unit normal vector to mesh face $f \in \partial \mathbf{P}$ by $\mathbf{n}_{\mathbf{P},f}$ and assume that these vectors are pointing out of \mathbf{P} . Moreover, we assume that the orientation of the mesh faces in every mesh is fixed *once and for all*, so that we can unambiguously introduce \mathbf{n}_f , the unit normal vector to face f . The orientation of this vector is independent of the elements \mathbf{P} to which f may belong, and may differ from $\mathbf{n}_{\mathbf{P},f}$ only by the multiplicative factor -1 .

7.2 The Mixed Variational Formulation of the Richards Equation

We consider the Richards equation

$$\frac{\partial \theta(\psi)}{\partial t} - \operatorname{div} (k_{\text{rel}}(\theta(\psi)) \nabla(\psi + z)) = f, \quad (7.1)$$

defined for $(\mathbf{x}, t) \in \Omega \times \mathbb{R}^+$ where $\Omega \subset \mathbb{R}^d$ for $d = 2, 3$ is the computational domain, the scalar variable $\psi(\mathbf{x}, t)$ is the pressure head, $\theta(\psi)$ is the saturation curve, $k_{\text{rel}}(\psi)$ is the relative permeability curve, z is the vertical coordinate oriented against the gravity direction and parallel to $\hat{\mathbf{z}} = (0, 1)^T$ for $d = 2$, $\hat{\mathbf{z}} = (0, 0, 1)^T$ for $d = 3$, and $f(\mathbf{x}, t)$ is the right-hand side source term. The Richards equation models the flow of water in sub-surface soils, and is generally non-linear and degenerate due to the non-linear dependence of the saturation and relative permeability curves on the pressure head. A source of degeneracy lies in the non-linear dependence of the relative permeability on ψ , which can become zero. A rather common way to cope with this kind of non-linearities consists in reformulating the partial differential equation by the Kirchhoff transformation [8, 54, 69, 75, 78, 81, 83], which introduces the alternative unknown p to be used instead of the pressure head variable ψ . Furthermore, Eq. (7.1) can change type in $\Omega \subset \mathbb{R}^d$, being elliptic in the fully saturated flow regime, i.e., when $\theta(\psi)$ is constant in time, and parabolic in the partially saturated flow regime. To deal with such situations, after the Kirchhoff transformation we consider the mixed form of the Richards equation where the mathematical model is split in the flux equation and the mass conservation equation.

To solve the Richards equation numerically, we consider the mixed formulation based upon the Kirchhoff transformation, which is given by:

$$\mathcal{K} : \mathbb{R} \rightarrow \mathbb{R} \quad \text{such that} \quad \psi \rightarrow \mathcal{K}(\psi) = \int_0^\psi k_{\text{rel}}(\theta(s)) ds.$$

As $k_{\text{rel}} \circ \theta(\cdot)$ is invertible, we rewrite Eq. (7.1) in terms of $p = \mathcal{K}(\psi)$ by noting that $\nabla p = \mathcal{K}'(\psi) \nabla \psi$ and setting $\beta(p) = \theta(\psi) = (\theta \circ \mathcal{K}^{-1})(p)$ and $\gamma(p) = \mathcal{K}'(\psi) = k_{\text{rel}}(\theta(\psi))$.

We recall that Ω is a polytopal domain (i.e., a polyhedron for $d = 3$; a polygon for $d = 2$) with Lipschitz continuous boundary $\Gamma = \bar{\Gamma}_D \cup \bar{\Gamma}_N$, where Γ_D and Γ_N are disjoint subsets of the domain boundary. We use such partition of Γ to set Dirichlet-type boundary conditions on Γ_D and Neumann-type boundary conditions on Γ_N . We also assume that $|\Gamma_D| \neq 0$, where $|\Gamma_D|$ is the $(d - 1)$ -Lebesgue measure of Γ_D . Let $J = (0, T] \subset \mathbb{R}^+$ be a finite time interval with T being the final time. The model problem associated with Eq. (7.1), which we shall approximate numerically, is given by

$$\frac{\partial \beta(p)}{\partial t} - \operatorname{div}((\nabla p + \gamma(p)\hat{\mathbf{z}})) = f \quad \text{in } J \times \Omega, \tag{7.2}$$

$$p = g_D \quad \text{on } J \times \Gamma_D, \tag{7.3}$$

$$(\nabla p + \gamma(p)\hat{\mathbf{z}}) \cdot \mathbf{n} = 0 \quad \text{on } J \times \Gamma_N, \tag{7.4}$$

$$p = p^0 \quad \text{at } t = 0 \text{ in } \Omega. \tag{7.5}$$

Well-posedness and solvability in weak sense of elliptic-parabolic problems like (7.2)–(7.5) have been investigated by many authors. In this work, we mainly refer to [4] for the definition of weak solution and the basic assumptions on the equation coefficients, that we list below:

(A1) $\beta(\cdot)$ is a C^1 , non-decreasing, and Lipschitz continuous function; thus,

$$|\beta(\xi) - \beta(\eta)| \leq C|\xi - \eta|, \quad \forall \xi, \eta \in \mathbb{R};$$

(A2) $\gamma(\cdot)$ is a continuous and bounded function satisfying

$$|\gamma(\xi) - \gamma(\eta)|^2 \leq C(\beta(\xi) - \beta(\eta))(\xi - \eta), \quad \forall \xi, \eta \in \mathbb{R};$$

(A3) $\beta(p^0)$ is essentially bounded in Ω and the initial state p^0 is taken in $L^2(\Omega)$;

(A4) $g_D \in L^2(J; H^1(\Gamma_D)) \cap L^\infty(J; L^\infty(\Gamma_D))$;

(A5) $f \in L^2(J; L^2(\Omega))$.

Assumption **(A2)** implies that

$$\|\gamma(\xi) - \gamma(\eta)\|_{0,\Omega}^2 \leq C(\beta(\xi) - \beta(\eta))(\xi - \eta) \tag{7.6}$$

whenever $\xi, \eta : \Omega \rightarrow \mathbb{R}$ are integrable functions. Moreover, under Assumptions **(A1)–(A5)**, existence and uniqueness of a solution $p \in L^2(J; H^1(\Omega))$ in weak sense is proved in [4] and the following regularity results hold:

$$\beta(p) \in L^\infty(J; L^1(\Omega)), \quad \partial \beta(p) / \partial t \in L^2(J; H^{-1}(\Omega)),$$

and

$$\mathbf{u} = -(\nabla p + \gamma(p)\hat{\mathbf{z}}) \in L^2(J; (L^2(\Omega))^d).$$

Since $\partial \beta(p) / \partial t \in L^2(J; H^{-1}(\Omega))$, a variational formulation of problem (7.2)–(7.5) would require test functions in $H^1(\Omega)$, which is a quite restrictive condition. An alternative approach involving more regular terms in the equations and, therefore, less regular test functions is possible by a preliminary time integration [69]. In fact, it holds that $\beta(p) \in L^2(J; H^1(\Omega))$ for $p \in L^2(J; H^1(\Omega))$ because $\beta(p)$ is a Lipschitz continuous function from **(A1)**. Since $\partial \beta(p) / \partial t \in L^2(J; H^{-1}(\Omega))$ we obtain that

$\beta(p) \in C^0(J; L^2(\Omega))$, see [61, Chap. I]. Thus, we integrate Eq. (7.2) in time and, for (almost) every $t \in J$, we obtain

$$\beta(p) + \operatorname{div} \int_0^t \mathbf{u}(s) \, ds = \beta(p^0) + \int_0^t f(s) \, ds,$$

where $\mathbf{u} = -(\nabla p + \gamma(p)\hat{\mathbf{z}})$ and p^0 is the initial solution state introduced in (A3). It was also proved that [8]:

$$\int_0^t \mathbf{u}(s) \, ds \in H^1(J; (L^2(\Omega))^d) \cap L^2(J; (H^1(\Omega))^d).$$

The *mixed variational formulation* of problem (7.2)–(7.5) is given by: For all $t \in J$, find $(\mathbf{u}(t, \cdot), p(t, \cdot)) \in H(\operatorname{div}; \Omega) \times L^2(\Omega)$ such that

$$(\beta(p), q) + \left(\operatorname{div} \int_0^t \mathbf{u}(s) \, ds, q \right) = (\beta(p^0), q) + \left(\int_0^t f(s) \, ds, q \right) \quad \forall q \in L^2(\Omega), \quad (7.7)$$

$$(\mathbf{u}, \mathbf{v}) - (p, \operatorname{div}(\mathbf{v})) + (\gamma(p)\hat{\mathbf{z}}, \mathbf{v}) = \langle g_D, \mathbf{n} \cdot \mathbf{v} \rangle \quad \forall \mathbf{v} \in H(\operatorname{div}; \Omega), \quad (7.8)$$

where the continuous bilinear form that provides the Dirichlet boundary condition in the right-hand side of (7.8) is given by

$$\langle g_D, \mathbf{n} \cdot \mathbf{v} \rangle = - \int_{\Gamma_D} g_D \mathbf{n} \cdot \mathbf{v} \, ds,$$

and \mathbf{n} is the unit vector orthogonal to Γ_D and pointing out of Ω .

7.3 The Mixed Virtual Element Method for the Richards Equation

In this section, we introduce the mixed virtual element method and consider its application to the numerical resolution of the Richards equation. To this end, we first present the basic assumptions on the mesh families that are admissible in the refinement process. Then, we define the local and global mixed virtual element space for the low-order approximation of the flux vector fields and the piecewise constant approximation of the hydraulic head (pressure) field. We equip the virtual element space with a stabilized virtual element inner product, which is used in the semi-discrete approximation of the Richards equation in mixed form. The virtual inner product possesses the three properties of stability, continuity and consistency, which are crucial to prove the convergence of the method. By integrating on the

time subintervals of a suitable partition of the time domain, we eventually derive the fully-discrete virtual element formulation.

Mesh regularity assumptions In order to use the interpolation and projection error estimates from the theory of polynomial approximation of functions in Sobolev spaces, we need a few regularity assumptions on the family of mesh decompositions $\mathcal{T} = \{\Omega_h\}_h$. We state the mesh regularity assumptions for the 2D and 3D case as follows.

Assumption (*Mesh regularity*)

- For $d = 2$, there exists a positive constant ϱ independent of h such that for every polygonal element P it holds that
 - (M1) P is star-shaped with respect to a disk with radius $\geq \varrho h_P$;
 - (M2) for every edge $e \in \partial P$ it holds that $h_e \geq \varrho h_P$.
- For $d = 3$, there exists a positive constant ϱ independent of h such that for every polyhedral element P and every mesh face $f \in \partial P$ it holds that
 - (M1) P is star-shaped with respect to a ball with radius $\geq \varrho h_P$ and every f is star-shaped with respect to a disk with radius $\geq \varrho h_f$;
 - (M2) for every edge $e \in \partial f$ and every face f it holds that $h_e \geq \varrho h_f \geq \varrho^2 h_P$. \square

Remark 7.1 The star-shapedness property (M1) implies that the elements and the mesh faces are *simply connected* subsets of \mathbb{R}^d and \mathbb{R}^{d-1} , respectively. The scaling property (M2) implies that the numbers of edges and faces in the elemental boundaries is uniformly bounded over the whole mesh family \mathcal{T} .

These mesh assumptions are quite general and, as observed from the very first publication on the VEM, see, for example, [11], allow us a great flexibility in the geometric shape of the elements of each mesh used in the numerical formulation. For example, non-convex elements or elements with hanging nodes are admissible. As already mentioned in Sect. 7.1.3, we retain a few additional but absolutely reasonable restrictions, e.g., elemental boundaries are given by portions of straight lines for $d = 2$ and mesh faces are planar for $d = 3$. We also avoid elements with intersecting boundaries, elements with “holes”, and elements totally surrounding other elements. It is worth mentioning, however, that examples of calculations using meshes with such kind of “exotic”-shaped elements have already been presented to the VEM community to challenge the robustness of the method [71].

Virtual element space The low regularity of the exact solution motivates us to consider the low-order mixed virtual element approximation. To this end, we define the global virtual element space for the vector fields as:

$$\mathbf{V}_h := \left\{ \mathbf{v}_h \in H(\text{div}; \Omega) : \mathbf{v}_h|_P \in \mathbf{V}_h(P) \quad \forall P \in \Omega_h \right\}, \tag{7.9}$$

where $\mathbf{V}_h(P)$ is the local (elemental) virtual element space. According to [16], we define it as

$$\mathbf{V}_h(\mathbf{P}) := \left\{ \mathbf{v}_h \in H(\operatorname{div}; \mathbf{P}) : \mathbf{v}_h \cdot \mathbf{n}_{\mathbf{P},f} \in \mathbb{P}_0(f) \forall f \in \partial \mathbf{P}, \operatorname{div} \mathbf{v}_h \in \mathbb{P}_0(\mathbf{P}), \operatorname{rot} \mathbf{v}_h = 0 \right\} \tag{7.10}$$

for a two-dimensional polygon. A similar definition holds when \mathbf{P} is a three-dimensional polyhedron, which can be given by using the condition $\operatorname{curl} \mathbf{v}_h = 0$ instead of $\operatorname{rot} \mathbf{v}_h = 0$ in definition (7.10), see Ref. [15]. The degrees of freedom that uniquely characterize the virtual element vector-valued fields \mathbf{v}_h in \mathbf{V}_h are the average on each mesh face of the normal component of \mathbf{v}_h :

- **(D1)** $\frac{1}{|f|} \int_f \mathbf{v}_h \cdot \mathbf{n}_f ds$ for all faces f .

This choice of degrees of freedom perfectly matches the degrees of freedom at $k = 0$ considered in the mixed VEM formulation of Refs. [15, 16], where a proof of unisolvence can be found. The restriction of **(D1)** to the boundary faces of the polytopal element \mathbf{P} are the degrees of freedom of the virtual element vector-valued fields in $\mathbf{V}_h(\mathbf{P})$, and can be proved to be unisolvent for such functions. It is worth noting that $\mathbf{V}_h(\mathbf{P})$ is indeed the local Raviart-Thomas space $\operatorname{RT}_0(\mathbf{P})$ when element \mathbf{P} is a d -simplex (triangle for $d = 2$, tetrahedron for $d = 3$), so, in this sense, the space $\mathbf{V}_h(\mathbf{P})$ is a generalization of the lowest-order Raviart-Thomas space to a polytopal cell. The degrees of freedom in **(D1)** make it possible to compute the divergence of every vector field in \mathbf{V}_h , which by definition (7.10) is a piecewise constant scalar function on mesh Ω_h . From the Gauss-Green theorem and since $\mathbf{v}_h \cdot \mathbf{n}_{\mathbf{P},f}$ is constant on every elemental face f , we find that

$$\operatorname{div} \mathbf{v}_h|_{\mathbf{P}} = \frac{1}{|\mathbf{P}|} \sum_{f \in \partial \mathbf{P}} |f| \mathbf{v}_h \cdot \mathbf{n}_{\mathbf{P},f},$$

where we recall that $\mathbf{n}_{\mathbf{P},f}$ is the unit normal vector to f pointing out of \mathbf{P} . This choice of degrees of freedom is also consistent with the fact that any vector-valued field \mathbf{v}_h of \mathbf{V}_h belongs to $H(\operatorname{div}; \Omega)$ and that its normal components must be continuous across the mesh faces. As the degree of freedom associated with each mesh face is unique, this condition is automatically satisfied by the mixed virtual element discretization.

Using the degrees of freedom of \mathbf{v}_h , we can compute the elemental orthogonal projection operator onto the constant vector fields

$$\int_{\mathbf{P}} (\mathbf{v}_h - \Pi_0^0(\mathbf{v}_h)) \cdot \mathbf{q} dV = 0 \quad \forall \mathbf{q} \in (\mathbb{P}_0(\mathbf{P}))^2. \tag{7.11}$$

A global projection operator, which we still denote as Π_0^0 with some abuse of notation, can be defined by setting $(\Pi_0^0 \mathbf{v}_h)|_{\mathbf{P}} = \Pi_0^0(\mathbf{v}_h|_{\mathbf{P}})$, where the Π_0^0 on the right must be intended as the local operator defined by (7.11). We also denote the interpolation of a vector field $\mathbf{v} \in H(\operatorname{div}; \Omega)$ by \mathbf{v}_I , which is the vector field in \mathbf{V}_h with the same degrees of freedom of \mathbf{v} . From the standard polynomial approximation theory for

Sobolev spaces [40], we know that

$$\|\mathbf{v} - \mathbf{v}_I\|_{0,\Omega} + \|\mathbf{v} - \Pi_0^0(\mathbf{v})\|_{0,\Omega} \leq Ch^s |\mathbf{v}|_{s,\Omega} \quad 0 < s \leq 1. \quad (7.12)$$

Inequality (7.12) will be useful in the convergence analysis of Sect. 7.4.

Virtual element inner product in \mathbf{V}_h . We recall that

$$\begin{aligned} (\mathbf{v}_h, \mathbf{w}_h) &= \int_{\Omega} \mathbf{v}_h(\mathbf{x}) \cdot \mathbf{w}_h(\mathbf{x}) dV = \sum_{\mathbf{P} \in \Omega_h} \int_{\mathbf{P}} \mathbf{v}_h(\mathbf{x}) \cdot \mathbf{w}_h(\mathbf{x}) dV \\ &= \sum_{\mathbf{P} \in \Omega_h} (\mathbf{v}_h, \mathbf{w}_h)_{\mathbf{P}} \end{aligned} \quad (7.13)$$

is the usual L^2 inner product between vector fields.

On the virtual element space \mathbf{V}_h , we consider the bilinear form defined by

$$\begin{aligned} (\mathbf{u}_h, \mathbf{v}_h)_{\mathbf{V}_h} &= \sum_{\mathbf{P} \in \Omega_h} (\mathbf{u}_h, \mathbf{v}_h)_{\mathbf{V}_h(\mathbf{P})} \\ &= \sum_{\mathbf{P} \in \Omega_h} \left(\int_{\mathbf{P}} \Pi_0^0 \mathbf{u}_h \cdot \Pi_0^0 \mathbf{v}_h dV + S_{\mathbf{P}} \left((1 - \Pi_0^0) \mathbf{u}_h, (1 - \Pi_0^0) \mathbf{v}_h \right) \right). \end{aligned} \quad (7.14)$$

The last summation argument on the right, viz. $S_{\mathbf{P}}(\cdot, \cdot)$, is the stabilization term, and its properties are such that the bilinear form (7.14) is indeed an inner product on \mathbf{V}_h . Any symmetric and coercive bilinear form that scales like the inner product $(\cdot, \cdot)_{\mathbf{P}}$ can be used as the stabilization of $(\cdot, \cdot)_{\mathbf{V}_h(\mathbf{P})}$. Formally, we assume that there exist two real positive constants c_* and c^* that are independent of h (and \mathbf{P}), and such that

$$c_* (\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{P}} \leq S_{\mathbf{P}}(\mathbf{v}_h, \mathbf{v}_h) \leq c^* (\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{P}} \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \quad (7.15)$$

This assumption implies that $S_{\mathbf{P}}(\cdot, \cdot)$ is ‘‘spectrally equivalent’’ to the inner product in $L^2(\mathbf{P})$, and, hence, is an inner product. It is clear at this point that relation (7.14) defines an inner product on \mathbf{V}_h and the induced norm is given by:

$$\|\mathbf{v}_h\|_{\mathbf{V}_h}^2 = (\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{V}_h} = \sum_{\mathbf{P} \in \Omega_h} \left(\int_{\mathbf{P}} |\Pi_0^0 \mathbf{v}_h|^2 dV + S_{\mathbf{P}} \left((1 - \Pi_0^0) \mathbf{v}_h, (1 - \Pi_0^0) \mathbf{v}_h \right) \right)$$

for every $\mathbf{v}_h \in \mathbf{V}_h$. Note, indeed, that $\|\mathbf{v}_h\|_{\mathbf{V}_h} = 0$ implies, for any \mathbf{P} , that $\|\Pi_0^0 \mathbf{v}_h\|_{0,\mathbf{P}} = 0$ and, using (7.15), that $\|(1 - \Pi_0^0) \mathbf{v}_h\|_{0,\mathbf{P}} = 0$. Therefore, $\Pi_0^0 \mathbf{v}_h = 0$ and $(1 - \Pi_0^0) \mathbf{v}_h = 0$, and the decomposition $\mathbf{v}_h = \Pi_0^0 \mathbf{v}_h + (1 - \Pi_0^0) \mathbf{v}_h$ implies that $\mathbf{v}_h = 0$. The other properties of an inner product follows from the definition of $(\cdot, \cdot)_{\mathbf{V}_h(\mathbf{P})}$.

The local bilinear form $(\cdot, \cdot)_{\mathbf{V}_h}$ has three important properties, that follows from its definition: stability, continuity, and consistency.

- **Stability:** there exist two constants α_* , $\alpha^* > 0$ independent of h such that

$$\alpha_*(\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{P}} \leq (\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{V}_h(\mathbf{P})} \leq \alpha^*(\mathbf{v}_h, \mathbf{v}_h)_{\mathbf{P}} \quad (7.16)$$

for all $\mathbf{v}_h \in \mathbf{V}_h$ and every element \mathbf{P} . Adding the contribution of all the elements yields the equivalence between the two norms $\|\cdot\|_{0,\Omega}$ and $|||\cdot|||_{\mathbf{V}_h}$:

$$(\alpha_*)^{\frac{1}{2}} \|\mathbf{v}_h\|_{0,\Omega} \leq |||\mathbf{v}_h|||_{\mathbf{V}_h} \leq (\alpha^*)^{\frac{1}{2}} \|\mathbf{v}_h\|_{0,\Omega}. \quad (7.17)$$

- **Continuity:** using the same constant α^* introduced above, which, we recall, is independent of h , it holds that

$$(\mathbf{v}_h, \mathbf{w}_h)_{\mathbf{V}_h} \leq \alpha^* \|\mathbf{v}_h\|_{0,\Omega} \|\mathbf{w}_h\|_{0,\Omega} \quad (7.18)$$

for all $\mathbf{v}_h, \mathbf{w}_h \in \mathbf{V}_h$. This property is an obvious consequence of the Cauchy-Schwarz inequality $(\mathbf{v}_h, \mathbf{w}_h)_{\mathbf{V}_h} \leq |||\mathbf{v}_h|||_{\mathbf{V}_h} |||\mathbf{w}_h|||_{\mathbf{V}_h}$ and the stability established in (7.17), which gives us the bounds $|||\mathbf{v}_h|||_{\mathbf{V}_h} \leq (\alpha^*)^{1/2} \|\mathbf{v}_h\|_{0,\Omega}$ and $|||\mathbf{w}_h|||_{\mathbf{V}_h} \leq (\alpha^*)^{1/2} \|\mathbf{w}_h\|_{0,\Omega}$.

- **Consistency:** for every element $\mathbf{P} \in \Omega_h$ and every constant vector-valued field \mathbf{q} defined on \mathbf{P} , it holds that

$$(\mathbf{v}_h, \mathbf{q})_{\mathbf{V}_h(\mathbf{P})} = (\mathbf{v}_h, \mathbf{q})_{\mathbf{P}} \quad (7.19)$$

for all $\mathbf{v}_h \in \mathbf{V}_h(\mathbf{P})$.

We will use this properties systematically in the convergence analysis of Sect. 7.4.

Approximation of the scalar variable We approximate the scalar variable p by a scalar function that is piecewise constant on the elements of mesh Ω_h . For the sake of exposition, we denote the space of these functions by $Q_h = \mathbb{P}_0(\Omega_h)$. Such a space is clearly a subspace of $L^2(\Omega)$, and its formal definition is:

$$Q_h := \{q_h \in L^2(\Omega) : q_h|_{\mathbf{P}} \in \mathbb{P}_0(\mathbf{P}) \quad \forall \mathbf{P} \in \Omega_h\}. \quad (7.20)$$

Every scalar function q_h in Q_h is uniquely identified by the set of constant values associated with the mesh elements, i.e., $q_h = (q_{\mathbf{P}})_{\mathbf{P} \in \Omega_h}$, where

- **(D2):** $q_{\mathbf{P}} = \frac{1}{|\mathbf{P}|} \int_{\mathbf{P}} q_h dV$.

We introduce the local orthogonal projection operators $\mathcal{P}_h^{\mathbf{P}} : L^2(\mathbf{P}) \rightarrow \mathbb{P}_0(\mathbf{P})$, which allows us to associate a square integrable function q with the set of its cell averages

on the cells of the current mesh Ω_h , and a global orthogonal projection operator $\mathcal{P}_h : L^2(\Omega) \rightarrow \mathcal{Q}_h$, such that $(\mathcal{P}_h q)|_{\mathbf{P}} = \mathcal{P}_h^{\mathbf{P}} q = (1/|\mathbf{P}|) \int_{\mathbf{P}} q \, dV$ for every $q \in L^2(\Omega)$ and $\mathbf{P} \in \Omega_h$. To ease the exposition, we will use the subscript p_I as an alternative notation for $\mathcal{P}_h p$, and, with some abuse of notation, for $\mathcal{P}_h^{\mathbf{P}} p$ without specifying the element \mathbf{P} .

Standard results from the theory of the polynomial approximation in Sobolev spaces state that:

$$\|p - p_I\|_{0,\mathbf{P}} \leq Ch_{\mathbf{P}}^s |p|_{s,\mathbf{P}} \quad 0 < s \leq 1,$$

see, again, Ref. [40].

Semi-discrete virtual element formulation The semi-discrete virtual element formulation corresponding to the semi-discrete mixed variational formulation (7.7)–(7.8) is given by: For all $t \in J$, find $(\mathbf{u}_h(t), p_h(t)) \in \mathbf{V}_h \times \mathcal{Q}_h$ such that

$$\begin{aligned} (\beta(p_h), q_h) + \left(\operatorname{div} \int_0^t \mathbf{u}_h(s) \, ds, q_h \right) &= (\beta(p_h^0), q_h) + \left(\int_0^t f(s) \, ds, q_h \right) \\ &\quad \forall q_h \in \mathcal{Q}_h, \end{aligned} \tag{7.21}$$

$$(\mathbf{u}_h, \mathbf{v}_h)_{\mathbf{V}_h} - (p_h, \operatorname{div} \mathbf{v}_h) + \sum_{\mathbf{P} \in \Omega_h} \gamma(p_{\mathbf{P}}) (\hat{\mathbf{z}}_I, \mathbf{v}_h)_{\mathbf{P}} = \langle \bar{g}_D, \mathbf{v}_h \rangle_h \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{7.22}$$

In (7.21), $\beta(p_h)$ is the discrete scalar field $\{\beta(p_{\mathbf{P}})\} \in \mathcal{Q}_h$. Moreover, we require that the initial solution $p_h^0 := p_h(0)$ satisfies

$$(\beta(p^0) - \beta(p_h^0), q_h) = 0 \quad \forall q_h \in \mathcal{Q}_h, \tag{7.23}$$

where we recall that p^0 is the initial solution state introduced in (A3). Condition (7.23) may be accomplished by taking

$$p_{h|\mathbf{P}}^0 = \beta^{-1} \left(\frac{1}{|\mathbf{P}|} \int_{\mathbf{P}} \beta(p^0) \, dV \right) \quad \forall \mathbf{P} \in \Omega_h.$$

The Dirichlet boundary condition on the right-hand side of (7.8) is numerically approximated in (7.22) by the bilinear form

$$\langle \bar{g}_D, \mathbf{v}_h \rangle_h = - \sum_{\mathbf{f} \subset \Gamma_D} |\mathbf{f}| \mathbf{v}_{\mathbf{P},\mathbf{f}} \bar{g}_{D,\mathbf{f}} \quad \text{where} \quad \bar{g}_{D,\mathbf{f}} = \frac{1}{|\mathbf{f}|} \int_{\mathbf{f}} g_D \, ds.$$

Since $\mathbf{n} \cdot \mathbf{v}_h|_{\mathbf{f}}$ is a constant on each \mathbf{f} due to definition of the degrees of freedom (D1), it holds that $\langle \bar{g}_D, \mathbf{v}_h \rangle_h = \langle g_D, \mathbf{n} \cdot \mathbf{v}_h \rangle$.

Fully-discrete virtual element formulation The fully-discrete scheme is obtained on a partition of the time interval J formed by N_T non-overlapping sub-intervals $[t^{n-1}, t^n]$ of size $\Delta t^n = t^n - t^{n-1}$ for $n = 1, \dots, N_T$, where $t^0 = 0$ and $t^{N_T} = T$ and such that $T = \sum_{n=1}^{N_T} \Delta t^n$. We denote the maximum time step by Δt and take the reasonable assumption that there exists a constant C providing a uniform bound from below for each time step Δt^n , i.e. $C \Delta t \leq \Delta t^n$. For any vector or scalar time dependent field $\eta(t, \cdot)$ defined on $J \times \Omega$ we use the notation $\eta^n = \eta(t^n, \cdot)$, and we denote the average of $\eta(t, \cdot)$ over the time interval $[t^{n-1}, t^n]$ by

$$\langle \eta \rangle^n = \frac{1}{\Delta t^n} \int_{t^{n-1}}^{t^n} \eta(t, \cdot) dt \quad \text{so that} \quad \int_0^{t^n} \eta(t, \cdot) dt = \sum_{k=1}^n \Delta t^k \langle \eta \rangle^k.$$

We formulate the fully-discrete virtual element scheme by numerically approximating the time integral of \mathbf{u}_h in (7.21) as follows

$$\int_0^{t^n} \mathbf{u}_h(s) ds \approx \sum_{k=0}^n \Delta t^k \mathbf{u}_h^k. \quad (7.24)$$

This expression can be interpreted as a first-order accurate approximation of the interpolant of the time integral of \mathbf{u} in (7.7):

$$\int_0^{t^n} \mathbf{u}_I(s) ds = \sum_{k=1}^n \Delta t^k \langle \mathbf{u}_I \rangle^k \quad \text{where} \quad \langle \mathbf{u}_I \rangle^k = \frac{1}{\Delta t^k} \int_{t^{k-1}}^{t^k} \mathbf{u}_I(s) ds.$$

We obtain a fully-discrete numerical approximation to (7.7)–(7.8), which is based on a time-stepping scheme equivalent to the forward Euler method for ordinary differential equations. Thus, the fully-discrete virtual element formulation is given by: *For every $n = 1, \dots, N_T$, find $(\mathbf{u}_h^n, p_h^n) \in \mathbf{V}_h \times \mathcal{Q}_h$ such that*

$$\begin{aligned} (\beta(p_h^n), q_h) + \left(\operatorname{div} \sum_{k=0}^n \Delta t^k \mathbf{u}_h^k, q_h \right) &= (\beta(p_h^0), q_h) + \left(\sum_{k=0}^n \Delta t^k \langle f \rangle^k, q_h \right) \\ &\quad \forall q_h \in \mathcal{Q}_h, \end{aligned} \quad (7.25)$$

$$\begin{aligned} (\mathbf{u}_h^n, \mathbf{v}_h)_{\mathbf{V}_h} - (p_h^n, \operatorname{div} \mathbf{v}_h) + \sum_{P \in \Omega_h} \gamma(p_P^n) (\hat{\mathbf{z}}_I, \mathbf{v}_h)_P &= (\bar{g}_D^n, \mathbf{v}_h)_{\mathbf{V}_h} \\ &\quad \forall \mathbf{v}_h \in \mathbf{V}_h \end{aligned} \quad (7.26)$$

and

$$\langle \bar{g}_D^n, \mathbf{v}_h \rangle_h = - \sum_{f \in \Gamma_D} |f| \mathbf{v}_h|_f \bar{g}_{D,f}^n. \quad (7.27)$$

In (7.27), we denote the face average value of g_D at the mesh face f and time instant t^n by $\bar{g}_{D,f}^n$. Since $\mathbf{n} \cdot \mathbf{v}_h|_f$ is a constant on each f due to definition of the degrees of freedom (D1), it holds that $\langle \bar{g}_D^n, \mathbf{v}_h \rangle_h = \langle g_D^n, \mathbf{n} \cdot \mathbf{v}_h \rangle$, where g_D^n is the Dirichlet boundary data taken at time t^n .

7.4 Convergence Analysis

The main results of this section are stated in Theorem 7.1 for the convergence of the semi-discrete scheme, and in Theorem 7.2 for the convergence of the fully discrete scheme. In both cases, optimal convergence rates with respect to the order of the approximation are derived.

Theorem 7.1 (Convergence of the semi-discrete scheme) *Let $(\mathbf{u}(t, \cdot), p(t, \cdot))$ for $t \in (0, T]$ be the flux and scalar solution of the mixed variational formulation (7.7)–(7.8) under Assumptions (A1)–(A5) and $(\mathbf{u}_I(t, \cdot), p_I(t, \cdot))$ their interpolants in $\mathbf{V}_h \times Q_h$. Let $(\mathbf{u}_h(t, \cdot), p_h(t, \cdot)) \in \mathbf{V}_h \times Q_h$ for $t \in (0, T]$ be the flux and scalar solution of the semi-discrete virtual element formulation (7.21)–(7.22) under the mesh regularity assumptions (M1)–(M2). Then, for every time T there exists a constant $C(T) = \mathcal{O}(T \exp(T))$ independent of h such that*

$$\left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2 + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) dt \right\|_{0,\Omega}^2 \leq C(T)h^2. \tag{7.28}$$

Theorem 7.2 (Convergence of the fully-discrete scheme) *Let $(\mathbf{u}(t, \cdot), p(t, \cdot))$ for $t \in (0, T]$ be the flux and scalar solution of the mixed variational formulation (7.7)–(7.8) under Assumptions (A1)–(A5) and $(\mathbf{u}_I(t, \cdot), p_I(t, \cdot))$ their interpolants in $\mathbf{V}_h \times Q_h$. Let $(\mathbf{u}_h(t, \cdot), p_h(t, \cdot)) \in \mathbf{V}_h \times Q_h$ for $t \in (0, T]$ be the flux and scalar solution of the fully-discrete virtual element formulation (7.25)–(7.26) under mesh regularity assumptions (M1)–(M2). We assume that $\partial \mathbf{u} / \partial t \in L^2(J, (L^2(\Omega))^d)$, and $\partial p / \partial t \in L^2(J, L^2(\Omega))$. Then, for every time T there exists a constant $C(T) = \mathcal{O}(T \exp(T))$ independent of h such that*

$$\left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - \langle p_I \rangle^n) \right\|_{0,\Omega}^2 + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \leq C(T)(\Delta t^2 + h^2). \tag{7.29}$$

The proofs of both theorems require some preliminary technical results that are stated and proved in the next subsections. More precisely, in Sect. 7.4.1 we prove the inf-sup condition and derive the error equations for both the semi-discrete and the fully-discrete VEM. In Sect. 7.4.2 we transform the error equations, introduce some estimates of the approximation error and, at the end of the section, we prove Theorem 7.1. Similar arguments are applied in Sect. 7.4.3 and fully discrete analogs

of the results of Sect. 7.4.2 are derived. Theorem 7.2 is proved at the end of the section.

7.4.1 Inf-sup Condition and Error Equations

The first lemma of this section provides the inf-sup condition for the pair of discrete spaces $\mathbf{V}_h - Q_h$.

Lemma 7.1 (Inf-sup condition) *Let us assume that (A1)–(A5) and (M1)–(M2) hold, and, further, that Ω is a convex domain with Lipschitz continuous boundary. Then, for every scalar function $q_h \in Q_h$ there exists a vector function $\tilde{\mathbf{v}}_q \in \mathbf{V}_h$ such that*

$$\operatorname{div} \tilde{\mathbf{v}}_{q_h} = q_h \quad \text{and} \quad \|\tilde{\mathbf{v}}_q\|_{0,\Omega} \leq C \|\operatorname{div} \tilde{\mathbf{v}}_q\|_{0,\Omega} = C \|q_h\|_{0,\Omega},$$

for some real positive constant C independent of h , q and \mathbf{v}_q .

Proof Let $q \in Q_h$. Due to the discrete inf-sup condition shown in [43], there exists a vector $\mathbf{v}_q \in H(\operatorname{div}; \Omega) \cap (L^{2+\epsilon}(\Omega))^2$, $\epsilon > 0$, such that

$$\operatorname{div} \mathbf{v}_q = q_h \quad \text{and} \quad \|\mathbf{v}_q\|_{\mathbf{V}_h} \leq C \|q_h\|_{0,\Omega}, \quad (7.30)$$

where the constant C is independent of h . The lemma follows by taking $\tilde{\mathbf{v}}_q = (\mathbf{v}_q)_I \in \mathbf{V}_h$, and noting that in each element P we have

$$(\operatorname{div} \tilde{\mathbf{v}}_q)_{|P} = (\operatorname{div} (\mathbf{v}_q)_I)_{|P} = \mathcal{P}_h(\operatorname{div} (\mathbf{v}_q))_{|P} = \mathcal{P}_h(q_h)_{|P} = q_h|_P = q_P,$$

and

$$\|\tilde{\mathbf{v}}_q\|_{0,\Omega} = \|(\mathbf{v}_q)_I\|_{0,\Omega} \leq \|\mathbf{v}_q\|_{0,\Omega} \leq \alpha^* \|\mathbf{v}_q\|_{\mathbf{V}_h} \leq C \|q_h\|_{0,\Omega}.$$

In the last chain of inequalities the final constant C absorbs α^* and the constant of the inequality in (7.30). \square

The error equations for the semi-discrete and fully-discrete formulations are derived in the following lemma.

Lemma 7.2 (Error equations)

(i) **Semi-discrete error equations.** *Let (\mathbf{u}, p) be the solution of (7.7)–(7.8) and (\mathbf{u}_h, p_h) be the solution of (7.21)–(7.22). Then, it holds:*

$$(\beta(p_h) - \beta(p), q_h) + \left(\operatorname{div} \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds, q_h \right) = 0 \quad \forall q_h \in Q_h, \quad (7.31)$$

$$(\mathbf{u}_h, \mathbf{v}_h)_{\mathbf{V}_h} - (\mathbf{u}, \mathbf{v}_h) - (p_h - p_I, \operatorname{div} \mathbf{v}_h) + ((\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}}, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \quad (7.32)$$

(ii) **Fully-discrete error equations.** Let (\mathbf{u}, p) be the solution of (7.7)–(7.8) and (\mathbf{u}_h, p_h) be the solution of (7.25)–(7.26). Then, it holds:

$$(\beta(p_h^n) - \beta(p^n), q_h) + \left(\operatorname{div} \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k), q_h \right) = 0 \quad \forall q_h \in Q_h, \tag{7.33}$$

$$(\mathbf{u}_h^n, \mathbf{v}_h)_{\mathbf{v}_h} - (\mathbf{u}^n, \mathbf{v}_h) - (p_h^n - p_I^n, \operatorname{div} \mathbf{v}_h) + ((\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \tag{7.34}$$

Proof Assertion (i) follows on taking the difference between (7.7) and (7.21) and using (7.23), and between (7.8) and (7.22), and noting that $(\operatorname{div} \mathbf{u}(s), q_h) = (\operatorname{div} \mathbf{u}_I(s), q_h)$ for every $s \in (0, t]$ and $\langle \bar{g}_D, \mathbf{v}_h \rangle_h = \langle g_D, \mathbf{n} \cdot \mathbf{v}_h \rangle$.

Assertion (ii) follows on taking the difference between (7.7) (with $t = t^n$) and (7.25) and using (7.23), and between (7.8) and (7.26) (with $t = t^n$), and noting that $(\operatorname{div} \langle \mathbf{u} \rangle^k, q_h) = (\operatorname{div} \langle \mathbf{u}_I \rangle^k, q_h)$ at any time instant t^k with $1 \leq k \leq n$ and $\langle \bar{g}_D^n, \mathbf{v}_h \rangle_h = \langle g_D^n, \mathbf{n} \cdot \mathbf{v}_h \rangle$. \square

7.4.2 Convergence of the Semi-discrete Approximation

To prove the convergence of the semi-discrete approximation (7.21)–(7.22), we need to estimate the time-integral between 0 and T of the two errors $(p_h - p_I)$ and $(\mathbf{u}_h - \mathbf{u}_I)$. These two errors measure the distance between the virtual element solution pair (p_h, \mathbf{v}_h) and the pair (p_I, \mathbf{u}_I) interpolating the exact solution fields (p, \mathbf{u}) . To derive the estimate, we need two preliminary results that are stated and proved in Lemmas 7.3–7.4, and using error equations (7.31)–(7.32). All the constants that appear in these mathematical developments are independent of the mesh size parameter h but may depend on the final time T . When this occurs, we denote this dependence by $C(T)$. In the next lemmas, we use the symbol \mathbf{u}_π to denote the piecewise constant vector-valued field given by taking the elemental averages of \mathbf{u} , which is the exact flux solution field evaluated at time t .

Lemma 7.3 *Under Assumptions (A1)–(A5) and (M1)–(M2), there exists a positive constant C , which is independent of h and T , such that*

$$\begin{aligned} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2 &\leq C \left(T \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt \right. \\ &\quad + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) dt \right\|_{0,\Omega}^2 + \left\| \int_0^T (\mathbf{u} - \mathbf{u}_I) dt \right\|_{0,\Omega}^2 \\ &\quad \left. + \left\| \int_0^T (\mathbf{u} - \mathbf{u}_\pi) dt \right\|_{0,\Omega}^2 \right), \end{aligned} \tag{7.35}$$

where \mathbf{u}_π denotes the piecewise constant vector-valued field given by taking the averages of \mathbf{u} on elements $P \in \Omega_h$.

Proof First, we integrate error equation (7.32) in time between $t = 0$ and T :

$$\begin{aligned} & \int_0^T \left((\mathbf{u}_h, \mathbf{v}_h)_{\mathbf{V}_h} - (\mathbf{u}, \mathbf{v}_h) \right) dt - \left(\int_0^T (p_h - p_I) dt, \operatorname{div} \mathbf{v}_h \right) \\ & + \left(\int_0^T ((\gamma(p_h) - \gamma(p))\hat{\mathbf{z}}) dt, \mathbf{v}_h \right) = 0. \end{aligned} \quad (7.36)$$

In view of Lemma 7.1, there exists a vector $\tilde{\mathbf{v}}_h \in \mathbf{V}_h$ such that

$$\operatorname{div} \tilde{\mathbf{v}}_h = \int_0^T (p_h - p_I) dt \text{ and } \|\tilde{\mathbf{v}}_h\|_{0,\Omega} \leq C \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}. \quad (7.37)$$

Constant C is independent of h and $q_h = \int_0^T (p_h - p_I) dt$, cf. Lemma 7.1, and, thus, it is independent of T . We set $\mathbf{v}_h = \tilde{\mathbf{v}}_h$ and use the expression of $\operatorname{div} \tilde{\mathbf{v}}_h$ in (7.36); then, we reorder the terms of the equation to obtain:

$$\begin{aligned} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2 &= \int_0^T \left((\mathbf{u}_h, \tilde{\mathbf{v}}_h)_{\mathbf{V}_h} - (\mathbf{u}, \tilde{\mathbf{v}}_h) \right) dt \\ &+ \left(\int_0^T (\gamma(p_h) - \gamma(p))\hat{\mathbf{z}} dt, \tilde{\mathbf{v}}_h \right). \end{aligned} \quad (7.38)$$

We use the consistency condition from (7.19) to transform the above relation as follows:

$$\begin{aligned} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2 &= \int_0^T \left((\mathbf{u}_h - \mathbf{u}_\pi, \tilde{\mathbf{v}}_h)_{\mathbf{V}_h} - (\mathbf{u} - \mathbf{u}_\pi, \tilde{\mathbf{v}}_h) \right) dt \\ &+ \left(\int_0^T (\gamma(p_h) - \gamma(p))\hat{\mathbf{z}} dt, \tilde{\mathbf{v}}_h \right) = (I) + (II) + (III). \end{aligned} \quad (7.39)$$

We estimate separately the three terms in the right-hand side of (7.39) by using the Cauchy-Schwarz and the Young inequality for some non-negative coefficient ϵ , the inequality in (7.37), and, for the first term, the stability condition on the \mathbf{V}_h -inner product:

$$\begin{aligned} (I) &= \left| \int_0^T (\mathbf{u}_h - \mathbf{u}_\pi, \tilde{\mathbf{v}}_h)_{\mathbf{V}_h} dt \right| \leq \frac{\alpha^*}{2\epsilon} \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_\pi) dt \right\|_{0,\Omega}^2 + \frac{\epsilon\alpha^*}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\ &\leq \frac{\alpha^*}{2\epsilon} \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_\pi) dt \right\|_{0,\Omega}^2 + \frac{\epsilon\alpha^*C}{2} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2. \end{aligned} \quad (7.40)$$

$$\begin{aligned} (II) &= \left| \int_0^T (\mathbf{u} - \mathbf{u}_\pi, \tilde{\mathbf{v}}_h) dt \right| \leq \frac{1}{2\epsilon} \left\| \int_0^T (\mathbf{u} - \mathbf{u}_\pi) dt \right\|_{0,\Omega}^2 + \frac{\epsilon}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\ &\leq \frac{1}{2\epsilon} \left\| \int_0^T (\mathbf{u} - \mathbf{u}_\pi) dt \right\|_{0,\Omega}^2 + \frac{\epsilon C}{2} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2. \end{aligned} \quad (7.41)$$

$$\begin{aligned}
 (III) &= \left| \left(\int_0^T (\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}} \, dt, \tilde{\mathbf{v}}_h \right) \right| \\
 &\leq \frac{1}{2\epsilon} \left\| \int_0^T (\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}} \, dt \right\|_{0,\Omega}^2 + \frac{\epsilon}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\
 &\leq \frac{1}{2\epsilon} \left\| \int_0^T (\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}} \, dt \right\|_{0,\Omega}^2 + \frac{\epsilon C}{2} \left\| \int_0^T (p_h - p_I) \, dt \right\|_{0,\Omega}^2.
 \end{aligned} \tag{7.42}$$

By choosing a suitable value of ϵ , we absorb the pressure term, i.e., the term that depends on $p_h - p_I$, in the right-hand side of (7.40), (7.41), and (7.42) within the left-hand side of (7.39) and we obtain the inequality:

$$\begin{aligned}
 \left\| \int_0^T (p_h - p_I) \, dt \right\|_{0,\Omega}^2 &\leq C \left(\left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_\pi) \, dt \right\|_{0,\Omega}^2 + \left\| \int_0^T (\mathbf{u} - \mathbf{u}_\pi) \, dt \right\|_{0,\Omega}^2 \right. \\
 &\quad \left. + \left\| \int_0^T (\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}} \, dt \right\|_{0,\Omega}^2 \right).
 \end{aligned} \tag{7.43}$$

We transform the first term of the right-hand side of (7.43) by adding and subtracting \mathbf{u}_I and \mathbf{u} to the integral argument and applying the triangular inequality

$$\begin{aligned}
 \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_\pi) \, dt \right\|_{0,\Omega}^2 &\leq \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) \, dt \right\|_{0,\Omega}^2 + \left\| \int_0^T (\mathbf{u}_I - \mathbf{u}) \, dt \right\|_{0,\Omega}^2 \\
 &\quad + \left\| \int_0^T (\mathbf{u} - \mathbf{u}_\pi) \, dt \right\|_{0,\Omega}^2
 \end{aligned} \tag{7.44}$$

We transform the last term of the right-hand side of (7.43) by applying Jensen’s inequality and (7.6), which follows from assumption (A2), and we obtain:

$$\begin{aligned}
 \left\| \int_0^T (\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}} \, dt \right\|_{0,\Omega}^2 &\leq T \int_0^T \|\gamma(p_h) - \gamma(p)\|_{0,\Omega}^2 \, dt \\
 &\leq CT \int_0^T (\beta(p_h) - \beta(p), p_h - p) \, dt,
 \end{aligned} \tag{7.45}$$

where constant C is independent of h and T . The assertion of the lemma follows on using inequalities (7.44) and (7.45) in (7.43) to obtain (7.35). \square

Lemma 7.4 *Under Assumptions (A1)–(A5) and (M1)–(M2), for every time $T > 0$ there exists a positive constant $C(T) = \mathcal{O}(\exp(T))$, which is independent of h , such that*

$$\begin{aligned}
 &\int_0^T (\beta(p_h) - \beta(p), p_h - p) \, dt + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) \, dt \right\|_{0,\Omega}^2 \\
 &\leq C(T) \left(\int_0^T \|p - p_I\|_{0,\Omega}^2 \, dt + \int_0^T \|\mathbf{u} - \mathbf{u}_I\|_{0,\Omega}^2 \, dt + \int_0^T \|\mathbf{u} - \mathbf{u}_\pi\|_{0,\Omega}^2 \, dt \right),
 \end{aligned} \tag{7.46}$$

where \mathbf{u}_π denotes the piecewise constant vector-valued field given by taking the averages of \mathbf{u} on elements $P \in \Omega_h$.

Proof Adding error equation (7.31) with $q_h = p_h - p_I$ and error equation (7.32) with $\mathbf{v}_h = \int_0^t (\mathbf{u}_h(s) - \mathbf{u}_I(s)) ds$ yields

$$\begin{aligned} & (\beta(p_h) - \beta(p), p_h - p_I) + \left(\mathbf{u}_h, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right)_{\mathbf{V}_h} - \left(\mathbf{u}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) \\ & + \left((\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) = 0. \end{aligned} \quad (7.47)$$

We subtract \mathbf{u}_π by exploiting consistency property (7.19) and we obtain

$$\begin{aligned} & (\beta(p_h) - \beta(p), p_h - p_I) + \left(\mathbf{u}_h - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right)_{\mathbf{V}_h} \\ & - \left(\mathbf{u} - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) + \left((\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) = 0, \end{aligned} \quad (7.48)$$

and, then, we add and subtract \mathbf{u}_I and the scalar solution field p in (7.48) and we find that

$$\begin{aligned} & (\beta(p_h) - \beta(p), p_h - p) + \left(\mathbf{u}_h - \mathbf{u}_I, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right)_{\mathbf{V}_h} = (\beta(p_h) - \beta(p), p_I - p) \\ & + \left(\mathbf{u} - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) - \left(\mathbf{u}_I - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right)_{\mathbf{V}_h} \\ & + \left((\gamma(p) - \gamma(p_h)) \hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right). \end{aligned} \quad (7.49)$$

The second term in the left-hand side of (7.49) can be further transformed by noting that

$$\left(\mathbf{u}_h - \mathbf{u}_I, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right)_{\mathbf{V}_h} = \frac{d}{dt} \frac{1}{2} \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{\mathbf{V}_h}^2.$$

Then, we remove the time derivative by integrating in time from $t = 0$ to the final time T , thus obtaining

$$\begin{aligned}
 & \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt + \frac{1}{2} \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{\mathbf{V}_h}^2 \\
 &= \int_0^T (\beta(p_h) - \beta(p), p_I - p) dt + \int_0^T (\mathbf{u} - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds) dt \\
 &\quad - \int_0^T (\mathbf{u}_I - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds)_{\mathbf{V}_h} dt \\
 &\quad + \int_0^T ((\gamma(p_h) - \gamma(p))\hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds) dt. \tag{7.50}
 \end{aligned}$$

The global equivalence relation (7.17) implies that

$$\begin{aligned}
 & \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt + \frac{1}{2} \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 \\
 & \leq \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt + \frac{1}{2\alpha_*} \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{\mathbf{V}_h}^2. \tag{7.51}
 \end{aligned}$$

We develop the first term of the right-hand side of (7.50) by applying the Young inequality for some positive coefficient ϵ that will be determined in the following and we use Assumption (A1) to obtain:

$$\begin{aligned}
 & \left| \int_0^T (\beta(p_h) - \beta(p), p_I - p) dt \right| \leq \int_0^T |(\beta(p_h) - \beta(p), p_I - p)| dt \\
 & \leq \frac{1}{2\epsilon} \int_0^T \|p_I - p\|_{0,\Omega}^2 dt + \frac{\epsilon}{2} \int_0^T \|\beta(p_h) - \beta(p)\|_{0,\Omega}^2 dt \\
 & \leq \frac{1}{2\epsilon} \int_0^T \|p_I - p\|_{0,\Omega}^2 dt + \frac{C\epsilon}{2} \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt. \tag{7.52}
 \end{aligned}$$

We apply the Young inequality to the second and third term of the right-hand side of (7.50), and the stability condition of the inner product in \mathbf{V}_h to obtain the two inequalities

$$\begin{aligned}
 & \left| \int_0^T (\mathbf{u} - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds) dt \right| \leq \frac{1}{2\epsilon} \int_0^T \|(\mathbf{u} - \mathbf{u}_\pi)\|_{0,\Omega}^2 dt \\
 & \quad + \frac{\epsilon}{2} \int_0^T \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 dt; \tag{7.53}
 \end{aligned}$$

$$\begin{aligned}
 & \left| \int_0^T (\mathbf{u}_I - \mathbf{u}_\pi, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds)_{\mathbf{V}_h} dt \right| \leq \frac{\alpha^*}{2\epsilon} \int_0^T \|(\mathbf{u}_I - \mathbf{u}_\pi)\|_{0,\Omega}^2 dt \\
 & \quad + \frac{\alpha^*\epsilon}{2} \int_0^T \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 dt. \tag{7.54}
 \end{aligned}$$

Finally, we develop the fourth term of the right-hand side of (7.50) by applying the Young inequality and Assumption (A2) to obtain:

$$\begin{aligned}
& \left| \int_0^T \left((\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) dt \right| \\
& \leq \int_0^T \left| \left((\gamma(p_h) - \gamma(p)) \hat{\mathbf{z}}, \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right) \right| dt \\
& \leq \frac{1}{2\epsilon} \int_0^T \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 dt + \frac{\epsilon}{2} \int_0^T \|\gamma(p_h) - \gamma(p)\|_{0,\Omega}^2 ds \\
& \leq \frac{1}{2\epsilon} \int_0^T \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 dt + \frac{C\epsilon}{2} \int_0^T (\beta(p_h) - \beta(p), p_h - p) ds.
\end{aligned} \tag{7.55}$$

We apply inequalities (7.52), (7.53), (7.54) and (7.55) to (7.50). By taking a suitable value of ϵ , the second term of the right-hand side of (7.52) and (7.55) may be absorbed by the left-hand side of (7.50). Combining this with (7.51) and collecting all constants in C yields:

$$\begin{aligned}
& \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 \\
& \leq C \left(\int_0^T \|\mathbf{u}_I - \mathbf{u}_\pi\|_{0,\Omega}^2 dt + \int_0^T \|\mathbf{u} - \mathbf{u}_\pi\|_{0,\Omega}^2 dt \right. \\
& \quad \left. + \int_0^T \|p_I - p\|_{0,\Omega}^2 dt + \int_0^T \left\| \int_0^t (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 dt \right). \tag{7.56}
\end{aligned}$$

We eliminate the last term in the right-hand side of (7.56) by applying the *Gronwall inequality*, which also determines the dependence $C(T) = \mathcal{O}(\exp(T))$ on the value of the final time T , to obtain:

$$\begin{aligned}
& \int_0^T (\beta(p_h) - \beta(p), p_h - p) dt + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) ds \right\|_{0,\Omega}^2 \\
& \leq C(T) \left(\int_0^T \|\mathbf{u}_I - \mathbf{u}_\pi\|_{0,\Omega}^2 dt + \int_0^T \|p_I - p\|_{0,\Omega}^2 dt \right). \tag{7.57}
\end{aligned}$$

The lemma follows by adding and subtracting \mathbf{u} in the first term in the right-hand side and using the triangular inequality.

Proof of Theorem 7.1 We use inequality (7.35) from Lemma 7.3 to bound the first term of the left-hand side of (7.28). Then, we use inequality (7.46) from Lemma (7.4). We obtain

$$\begin{aligned} \left\| \int_0^T (p_h - p_I) dt \right\|_{0,\Omega}^2 + \left\| \int_0^T (\mathbf{u}_h - \mathbf{u}_I) dt \right\|_{0,\Omega}^2 \leq C(T) & \left(\int_0^T \|p - p_I\|_{0,\Omega}^2 dt \right. \\ & \left. + \int_0^T \|\mathbf{u} - \mathbf{u}_I\|_{0,\Omega}^2 dt + \int_0^T \|\mathbf{u} - \mathbf{u}_\pi\|_{0,\Omega}^2 dt \right), \end{aligned} \tag{7.58}$$

where $C(T) = \mathcal{O}(T \exp(T))$. Finally, we estimate the spatial convergence rate by using the interpolation bounds (7.12) for $s = 1$.

7.4.3 Convergence of the Fully-Discrete Approximation

To prove the convergence of the fully-discrete approximation (7.25)–(7.26), we need to estimate the time-integral between 0 and T of the two errors $(p_h - p_I)$ and $(\mathbf{u}_h - \mathbf{u}_I)$. These two errors measure the distance between the virtual element solution pair (p_h, \mathbf{v}_h) and the pair (p_I, \mathbf{u}_I) interpolating the exact solution fields (p, \mathbf{u}) . To derive the estimate, we need two preliminary results that are stated and proved in Lemmas 7.5 and 7.7 from the error equations (7.33)–(7.34). It is worth noting that these lemmas are the discrete analog of the Lemmas 7.3, and 7.4 of the semi-discrete formulation. In the following analysis, all the constants that appear in the inequality chains are independent of the mesh size parameter h but may depend on the final time T . When this occurs, we denote this dependence by $C(T)$. In the proof of the next lemmas, we use the symbol \mathbf{u}_π^n to denote the piecewise constant vector field given by taking the elemental averages of \mathbf{u}^n , which is the exact flux field evaluated at time t^n .

Lemma 7.5 *Under Assumptions (A1)–(A5) and (M1)–(M2), there exists a positive constant C , which is independent of h and T , such that*

$$\begin{aligned} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2 \leq C & \left(T \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) \right. \\ & + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\ & \left. + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 \right), \end{aligned} \tag{7.59}$$

where \mathbf{u}_π^n denotes the piecewise constant vector-valued field given by taking the averages of \mathbf{u} on elements $P \in \Omega_h$ at time t^n .

Proof We use a fully discrete analog of the argument that we used to prove Lemma 7.5. First, we multiply the error equation (7.34) by Δt^n and sum for $n = 1, \dots, N_T$ to obtain:

$$\begin{aligned} & \left(\sum_{n=1}^{N_T} \Delta t^n \mathbf{u}_h^n, \mathbf{v}_h \right)_{\mathbf{V}_h} - \left(\sum_{n=1}^{N_T} \Delta t^n \mathbf{u}^n, \mathbf{v}_h \right) - \left(\sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n), \operatorname{div} \mathbf{v}_h \right) \\ & + \left(\sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \mathbf{v}_h \right) = 0. \end{aligned} \quad (7.60)$$

In view of Lemma 7.1, there exists a vector $\tilde{\mathbf{v}}_h \in \mathbf{V}_h$ such that

$$\operatorname{div} \tilde{\mathbf{v}}_h = \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \quad \text{and} \quad \|\tilde{\mathbf{v}}_h\|_{0,\Omega} \leq C \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}. \quad (7.61)$$

Constant C is independent of h and $q_h = \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n)$, cf. Lemma 7.1, and, thus, it is independent of T . We set $\mathbf{v}_h = \tilde{\mathbf{v}}_h$ and use the expression of $\operatorname{div} \tilde{\mathbf{v}}_h$ in (7.60); then, we reorder the terms of the equation to obtain:

$$\begin{aligned} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2 &= \left(\sum_{n=1}^{N_T} \Delta t^n \mathbf{u}_h^n, \tilde{\mathbf{v}}_h \right)_{\mathbf{V}_h} - \left(\sum_{n=1}^{N_T} \Delta t^n \mathbf{u}^n, \tilde{\mathbf{v}}_h \right) \\ &+ \left(\sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \tilde{\mathbf{v}}_h \right). \end{aligned} \quad (7.62)$$

We use the consistency condition from (7.19) to transform the above relation as follows:

$$\begin{aligned} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2 &= \left(\sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n), \tilde{\mathbf{v}}_h \right)_{\mathbf{V}_h} - \left(\sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n), \tilde{\mathbf{v}}_h \right) \\ &+ \left(\sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \tilde{\mathbf{v}}_h \right) = (I) + (II) + (III). \end{aligned} \quad (7.63)$$

We estimate separately the three terms in the right-hand side of (7.63) by using the Cauchy-Schwarz and the Young inequality for some non-negative coefficient ϵ , the inequality in (7.61), and, for the first term, the stability condition on the \mathbf{V}_h -inner product:

$$\begin{aligned} (I) &= \left| \left(\sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n), \tilde{\mathbf{v}}_h \right)_{\mathbf{V}_h} \right| \leq \frac{\alpha^*}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 + \frac{\epsilon \alpha^*}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\ &\leq \frac{\alpha^*}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 + \frac{\epsilon \alpha^* C}{2} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2. \end{aligned} \quad (7.64)$$

$$\begin{aligned}
(II) &= \left| \left(\sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n), \tilde{\mathbf{v}}_h \right) \right| \leq \frac{1}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 + \frac{\epsilon}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\
&\leq \frac{1}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 + \frac{\epsilon C}{2} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.65}$$

$$\begin{aligned}
(III) &= \left| \left(\sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \tilde{\mathbf{v}}_h \right) \right| \\
&\leq \frac{1}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}} \right\|_{0,\Omega}^2 + \frac{\epsilon}{2} \|\tilde{\mathbf{v}}_h\|_{0,\Omega}^2 \\
&\leq \frac{1}{2\epsilon} \left\| \sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}} \right\|_{0,\Omega}^2 + \frac{\epsilon C}{2} \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.66}$$

By choosing a suitable value of ϵ we absorb the pressure term, i.e., the term that depends on $p_h^n - p_I^n$, in the right-hand side of (7.64), (7.65) and (7.66) within the left-hand side of (7.63) and we obtain the inequality:

$$\begin{aligned}
\left\| \sum_{n=1}^{N_T} \Delta t^n (p^n - p_I^n) \right\|_{0,\Omega}^2 &\leq C \left(\left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 \right. \\
&\quad \left. + \left\| \sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}} \right\|_{0,\Omega}^2 \right).
\end{aligned} \tag{7.67}$$

We transform the first term of the right-hand side of (7.67) by adding and subtracting $(\mathbf{u}_I)^n$ and \mathbf{u}^n to the summation argument and applying the triangular inequality

$$\begin{aligned}
\left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2 &\leq \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - (\mathbf{u}_I)^n) \right\|_{0,\Omega}^2 + \left\| \sum_{n=1}^{N_T} \Delta t^n ((\mathbf{u}_I)^n - \mathbf{u}^n) \right\|_{0,\Omega}^2 \\
&\quad + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}^n - \mathbf{u}_\pi^n) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.68}$$

We transform the last term of the right-hand side of (7.67) by using Jensen's inequality and inequality (7.6), cf. Assumption (A2), and noting that $\Delta t^n \leq T$; hence,

$$\begin{aligned}
& \left\| \sum_{n=1}^{N_T} \Delta t^n (\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}} \right\|_{0,\Omega}^2 \leq \sum_{n=1}^{N_T} (\Delta t^n)^2 \|\gamma(p_h^n) - \gamma(p^n)\|_{0,\Omega}^2 \\
& \leq C \sum_{n=1}^{N_T} (\Delta t^n)^2 (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) \\
& \leq CT \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n), \tag{7.69}
\end{aligned}$$

where constant C is independent of h and T . The assertion of the lemma follows on using inequalities (7.68) and (7.69) in inequality (7.67) to obtain (7.59). \square

As we have already noted in the opening comments of this section, next Lemma 7.7 is a fully discrete analog of Lemma 7.4. In its proof, all time integrals must be substituted by summations over the time index n . To simplify the proof, we find it convenient to introduce the special summation symbol:

$$\mathcal{S}_h^n = \begin{cases} 0 & \text{for } n = 0, \\ \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) & \text{for } n \geq 1. \end{cases} \tag{7.70}$$

The properties listed in Lemma 7.6 will be used in the proof of Lemma 7.7.

Lemma 7.6 *For $n \geq 1$, there hold:*

$$(i) \quad \mathcal{S}_h^n - \mathcal{S}_h^{n-1} = \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n); \tag{7.71}$$

$$(ii) \quad 2 (\mathcal{S}_h^n - \mathcal{S}_h^{n-1}, \mathcal{S}_h^n)_{\mathbf{V}_h} = \|\mathcal{S}_h^n\|_{\mathbf{V}_h}^2 - \|\mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2 + \|\mathcal{S}_h^n - \mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2; \tag{7.72}$$

$$(iii) \quad 2 \sum_{n=1}^{N_T} (\mathcal{S}_h^n - \mathcal{S}_h^{n-1}, \mathcal{S}_h^n)_{\mathbf{V}_h} = \|\mathcal{S}_h^{N_T}\|_{\mathbf{V}_h}^2 + \sum_{n=1}^{N_T} \|\mathcal{S}_h^n - \mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2. \tag{7.73}$$

Proof Item (i) is an immediate consequence of definition (7.70). Item (ii) follows from this obvious equality:

$$\begin{aligned}
2 (\mathcal{S}_h^n - \mathcal{S}_h^{n-1}, \mathcal{S}_h^n)_{\mathbf{V}_h} &= (\mathcal{S}_h^n, \mathcal{S}_h^n)_{\mathbf{V}_h} - (\mathcal{S}_h^{n-1}, \mathcal{S}_h^{n-1})_{\mathbf{V}_h} \\
&\quad + (\mathcal{S}_h^n, \mathcal{S}_h^n)_{\mathbf{V}_h} + (\mathcal{S}_h^{n-1}, \mathcal{S}_h^{n-1})_{\mathbf{V}_h} - 2 (\mathcal{S}_h^{n-1}, \mathcal{S}_h^n)_{\mathbf{V}_h} \tag{7.74}
\end{aligned}$$

Item (iii) follows from item (ii) by summing (7.72) from $n = 1$ to $n = N_T$ to obtain

$$2 \sum_{n=1}^{N_T} (\mathcal{S}_h^n - \mathcal{S}_h^{n-1}, \mathcal{S}_h^n)_{\mathbf{V}_h} = \sum_{n=1}^{N_T} (\|\mathcal{S}_h^n\|_{\mathbf{V}_h}^2 - \|\mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2) + \sum_{n=1}^{N_T} \|\mathcal{S}_h^n - \mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2$$

and noting that the first term on the right-hand side is a telescopic sum. \square

Lemma 7.7 *Under Assumptions (A1)–(A5) and (M1)–(M2), for every time $T > 0$ there exists a positive constant $C(T) = \mathcal{O}(\exp(T))$ such that*

$$\begin{aligned} & \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\ & \leq C(T) \left(\sum_{n=1}^{N_T} \Delta t^n \|p_h^n - p^n\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n - \langle \mathbf{u}_I \rangle^n\|_{0,\Omega}^2 \right. \\ & \quad \left. + \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n - \mathbf{u}_\pi^n\|_{0,\Omega}^2 \right), \end{aligned} \tag{7.75}$$

where constant $C(T)$ is independent of the mesh size parameter h , and \mathbf{u}_π^n denotes the piecewise constant vector-valued field given by taking the averages of \mathbf{u} on elements $P \in \Omega_h$ at time t^n .

Proof We use (7.70) with $n = N_T$ to reformulate the left-hand side of (7.75) as follows:

$$\begin{aligned} & \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\ & = \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \|\mathcal{S}_h^{N_T}\|_{0,\Omega}^2 \\ & \leq \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + (\alpha_*)^{-1} \|\mathcal{S}_h^{N_T}\|_{\mathbf{V}_h}^2. \end{aligned} \tag{7.76}$$

We add $(\alpha_*)^{-1} \sum_{n=1}^{N_T} \|\mathcal{S}_h^n - \mathcal{S}_h^{n-1}\|_{\mathbf{V}_h}^2$ to the right-hand side of (7.76) and, then, we substitute (7.73) to obtain the inequality:

$$\begin{aligned} & \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\ & \leq 2C \sum_{n=1}^{N_T} \Delta t^n \left((\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left(\frac{\mathcal{S}_h^n - \mathcal{S}_h^{n-1}}{\Delta t^n}, \mathcal{S}_h^n \right)_{\mathbf{V}_h} \right), \end{aligned} \tag{7.77}$$

where constant C depends on α_* but is independent of h . Next, we derive an expression for the summation argument in the right-hand side of (7.77), which depends on the interpolation errors $p^n - p_I^n$ and $\mathbf{u}^n - \langle \mathbf{u}_I \rangle^n$.

Adding error equation (7.33) with $q_h = p_h^n - p_I^n$ and error equation (7.34) with $\mathbf{v}_h = \mathcal{S}_h^n$ yields

$$\begin{aligned} & (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + (\mathbf{u}_h^n, \mathcal{S}_h^n)_{\mathbf{V}_h} - (\mathbf{u}^n, \mathcal{S}_h^n) \\ & + ((\gamma(p_h^n) - \gamma(p^n))\hat{\mathbf{z}}, \mathcal{S}_h^n) = 0. \end{aligned} \quad (7.78)$$

We subtract \mathbf{u}_π by exploiting the consistency property (7.19) and we obtain

$$\begin{aligned} & (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + (\mathbf{u}_h^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)_{\mathbf{V}_h} \\ & - (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n) + ((\gamma(p_h^n) - \gamma(p^n))\hat{\mathbf{z}}, \mathcal{S}_h^n) = 0, \end{aligned} \quad (7.79)$$

and, then, we add and subtract $\langle \mathbf{u}_I \rangle^n$ and the scalar solution field p^n

$$\begin{aligned} & (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n, \mathcal{S}_h^n)_{\mathbf{V}_h} = (\beta(p_h^n) - \beta(p^n), p_I^n - p^n) \\ & + (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n) - (\langle \mathbf{u}_I \rangle^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)_{\mathbf{V}_h} + ((\gamma(p^n) - \gamma(p_h^n))\hat{\mathbf{z}}, \mathcal{S}_h^n). \end{aligned} \quad (7.80)$$

Using (7.71) and adding and subtracting \mathbf{u}^n in the third term of the right-hand side yield:

$$\begin{aligned} & (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left(\frac{\mathcal{S}_h^n - \mathcal{S}_h^{n-1}}{\Delta t^n}, \mathcal{S}_h^n \right)_{\mathbf{V}_h} \\ & = (\beta(p_h^n) - \beta(p^n), p_I^n - p^n) \\ & + (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n) - (\langle \mathbf{u}_I \rangle^n - \mathbf{u}^n, \mathcal{S}_h^n)_{\mathbf{V}_h} - (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)_{\mathbf{V}_h} \\ & + ((\gamma(p_h^n) - \gamma(p^n))\hat{\mathbf{z}}, \mathcal{S}_h^n). \end{aligned} \quad (7.81)$$

Finally, we substitute (7.81) in the right-hand side of (7.77) to obtain the inequality:

$$\begin{aligned} & \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\ & \leq 2C \sum_{n=1}^{N_T} \Delta t^n \left((\beta(p_h^n) - \beta(p^n), p_I^n - p^n) + (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n) \right. \\ & \quad \left. - (\langle \mathbf{u}_I \rangle^n - \mathbf{u}^n, \mathcal{S}_h^n)_{\mathbf{V}_h} - (\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)_{\mathbf{V}_h} \right. \\ & \quad \left. + ((\gamma(p_h^n) - \gamma(p^n))\hat{\mathbf{z}}, \mathcal{S}_h^n) \right). \end{aligned} \quad (7.82)$$

To estimate the first term of the right-hand side of (7.82), we apply the Young inequality for some positive coefficient ϵ that will be determined in the following, use Assumption (A1), and obtain:

$$\begin{aligned}
& \sum_{n=1}^{N_T} \Delta t^n |(\beta(p_h^n) - \beta(p^n), p_I^n - p^n)| \\
& \leq \frac{1}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \|p_I^n - p^n\|_{0,\Omega}^2 + \frac{\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n \|\beta(p_h^n) - \beta(p^n)\|_{0,\Omega}^2 \\
& \leq \frac{1}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \|p_I^n - p^n\|_{0,\Omega}^2 + \frac{C\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n).
\end{aligned} \tag{7.83}$$

To estimate the second, third and fourth terms of the right-hand side of (7.82) we apply the Young inequality and definition (7.70), and the consistency condition for the inner product in \mathbf{V}_h , and we obtain:

$$\begin{aligned}
\sum_{n=1}^{N_T} \Delta t^n |(\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)| & \leq \frac{1}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \|(\mathbf{u}^n - \mathbf{u}_\pi^n)\|_{0,\Omega}^2 \\
& \quad + \frac{\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.84}$$

$$\begin{aligned}
\sum_{n=1}^{N_T} \Delta t^n |(\langle \mathbf{u}_I \rangle^n - \mathbf{u}^n, \mathcal{S}_h^n)_{\mathbf{V}_h}| & \leq \frac{\alpha^*}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \|(\langle \mathbf{u}_I \rangle^n - \mathbf{u}^n)\|_{0,\Omega}^2 \\
& \quad + \frac{\alpha^*\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.85}$$

$$\begin{aligned}
\sum_{n=1}^{N_T} \Delta t^n |(\mathbf{u}^n - \mathbf{u}_\pi^n, \mathcal{S}_h^n)_{\mathbf{V}_h}| & \leq \frac{\alpha^*}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \|(\mathbf{u}^n - \mathbf{u}_\pi^n)\|_{0,\Omega}^2 \\
& \quad + \frac{\alpha^*\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2.
\end{aligned} \tag{7.86}$$

To estimate the last term of the right-hand side of (7.82), we apply the Cauchy-Schwarz and Young inequalities with the same coefficient ϵ , Assumption **(A2)**, and definition (7.70), and we obtain:

$$\begin{aligned}
& \left| \sum_{n=1}^{N_T} \Delta t^n \left((\gamma(p_h^n) - \gamma(p^n)) \hat{\mathbf{z}}, \mathcal{S}_h^n \right) \right| \leq \sum_{n=1}^{N_T} \Delta t^n \left\| \gamma(p_h^n) - \gamma(p^n) \right\|_{0,\Omega} \left\| \mathcal{S}_h^n \right\|_{0,\Omega} \\
& \leq \frac{1}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2 + \frac{\epsilon}{2} \sum_{n=1}^{N_T} \Delta t^n \left\| \gamma(p_h^n) - \gamma(p^n) \right\|_{0,\Omega}^2 \\
& \leq \frac{1}{2\epsilon} \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=1}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2 + \frac{\epsilon C}{2} \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n). \quad (7.87)
\end{aligned}$$

We apply inequalities (7.83)–(7.87) to (7.82). By taking a suitable value of ϵ , the second term of the right-hand side of (7.83) and (7.87) is absorbed by the left-hand side of (7.82), thus giving

$$\begin{aligned}
& \sum_{n=1}^{N_T} \Delta t^n (\beta(p_h^n) - \beta(p^n), p_h^n - p^n) + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\
& \leq C \left(\sum_{n=1}^{N_T} \Delta t^n \|p_h^n - p^n\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \|(\mathbf{u}^n - \mathbf{u}_\pi)\|_{0,\Omega}^2 \right. \\
& \quad \left. + \sum_{n=1}^{N_T} \Delta t^n \|(\langle \mathbf{u}_I \rangle^n - \mathbf{u}^n)\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \left\| \sum_{k=0}^n \Delta t^k (\mathbf{u}_h^k - \langle \mathbf{u}_I \rangle^k) \right\|_{0,\Omega}^2 \right). \quad (7.88)
\end{aligned}$$

Finally, we eliminate the last term in the right-hand side of (7.88) by applying the discrete Gronwall inequality, and we find the assertion of the lemma with a constant $C(T) = \mathcal{O}(\exp(T))$, which is independent of h . \square

Proof of Theorem 7.2 Adding and subtracting p_I^n to the first term on the left-hand side of (7.29) and using the triangular inequality yield:

$$\begin{aligned}
\left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - \langle p_I \rangle^n) \right\|_{0,\Omega}^2 & \leq 2 \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - p_I^n) \right\|_{0,\Omega}^2 \\
& \quad + 2 \left\| \sum_{n=1}^{N_T} \Delta t^n (p_I^n - \langle p_I \rangle^n) \right\|_{0,\Omega}^2. \quad (7.89)
\end{aligned}$$

The first term on the right-hand side is controlled by using inequality (7.59) from Lemma 7.5. Then, we use inequality (7.75) from Lemma 7.7 and the Jensen inequality. We obtain

$$\begin{aligned}
& \left\| \sum_{n=1}^{N_T} \Delta t^n (p_h^n - \langle p_I \rangle^n) \right\|_{0,\Omega}^2 + \left\| \sum_{n=1}^{N_T} \Delta t^n (\mathbf{u}_h^n - \langle \mathbf{u}_I \rangle^n) \right\|_{0,\Omega}^2 \\
& \leq C(T) \left(\sum_{n=1}^{N_T} \Delta t^n \|p_I^n - p^n\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \|p_I^n - \langle p_I \rangle^n\|_{0,\Omega}^2 \right. \\
& \quad \left. + \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n - \langle \mathbf{u}_I \rangle^n\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n - \mathbf{u}_\pi^n\|_{0,\Omega}^2 \right) \\
& = C(T) (\mathbb{T}_1 + \mathbb{T}_2 + \mathbb{T}_3 + \mathbb{T}_4), \tag{7.90}
\end{aligned}$$

where $C(T) = \mathcal{O}(T \exp(T))$. We estimate the error terms \mathbb{T}_i , $i = 1, 4$ in the right-hand side of (7.90) as follows [78], by noting that $p_h = \mathcal{P}_h p_h$, $p_I = \mathcal{P}_h p$, $\langle p_I \rangle^n = \mathcal{P}_h \langle p \rangle^n$, $\Delta t^n < T$ and the projection operator \mathcal{P}_h for $h \rightarrow 0$ is (uniformly) bounded. We find that:

$$\begin{aligned}
\mathbb{T}_1 & \leq Ch^2 \sum_{n=1}^{N_T} \Delta t^n \|p^n\|_{1,\Omega}^2 \leq Ch^2 \left(\int_0^T \|p\|_{1,\Omega}^2 dt + \mathcal{O}(\Delta t) \right), \\
\mathbb{T}_2 & \leq \sum_{n=1}^{N_T} \Delta t^n \|\mathcal{P}_h(p^n - \langle p \rangle^n)\|_{0,\Omega}^2 \leq C \Delta t^2 \int_0^T \left\| \frac{\partial p}{\partial t} \right\|_{0,\Omega}^2 dt, \\
\mathbb{T}_3 & \leq \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n - \mathbf{u}_I^n\|_{0,\Omega}^2 + \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}_I^n - \langle \mathbf{u}_I \rangle^n\|_{0,\Omega}^2 \\
& \leq C \left(h^2 \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n\|_{1,\Omega}^2 + \Delta t^2 \int_0^T \left\| \frac{\partial \mathbf{u}}{\partial t} \right\|_{0,\Omega}^2 dt \right) \\
& \leq C \left(h^2 \left(\int_0^T \|\mathbf{u}\|_{1,\Omega}^2 dt + \mathcal{O}(\Delta t) \right) + \Delta t^2 \int_0^T \left\| \frac{\partial \mathbf{u}}{\partial t} \right\|_{0,\Omega}^2 dt \right), \\
\mathbb{T}_4 & \leq Ch^2 \sum_{n=1}^{N_T} \Delta t^n \|\mathbf{u}^n\|_{1,\Omega}^2 \leq Ch^2 \left(\int_0^T \|\mathbf{u}\|_{1,\Omega}^2 dt + \mathcal{O}(\Delta t) \right).
\end{aligned}$$

and the theorem follows by using these estimates in inequality (7.90).

7.5 Numerical Experiments

In this section, we study the convergence properties of the proposed mixed VEM numerically. To perform this study, we consider four different types of meshes, viz.,

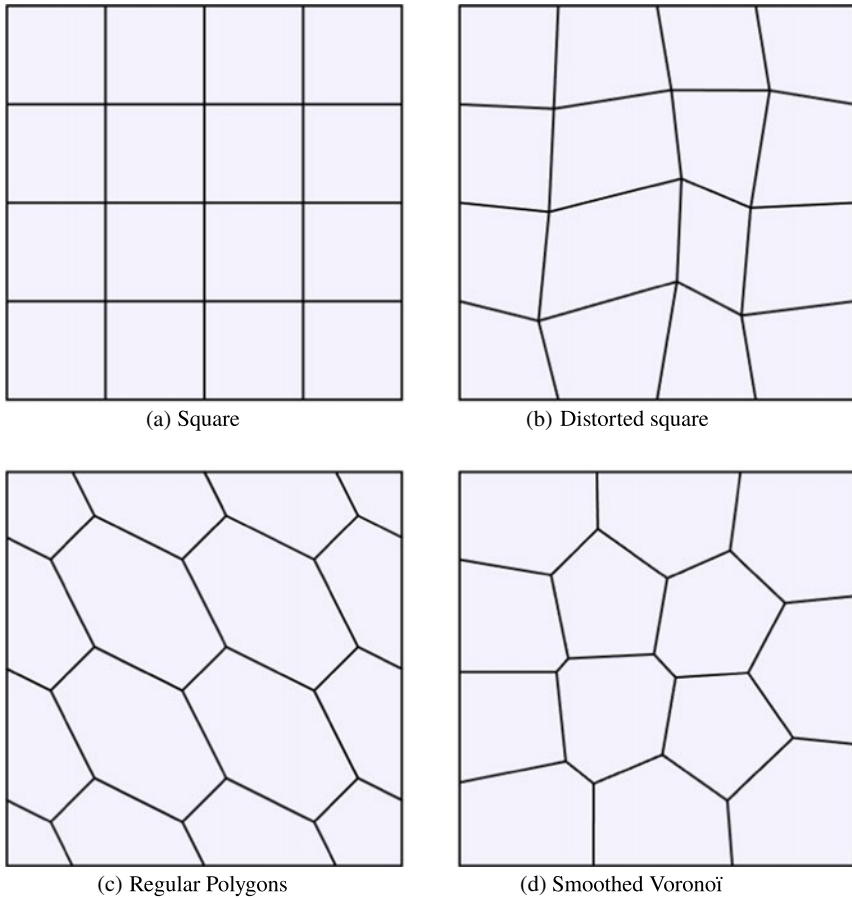


Fig. 7.1 A schematic representation of different discretizations employed in this study

structured square, distorted square, regular polygonal and Voronoi meshes. Figure 7.1 shows a schematic representation of the meshes employed in this study.

Before studying the Richards equation, the mixed VEM developed herein is validated for the simplest model problem:

$$-\operatorname{div}(\nabla p) = f \quad \text{in } \Omega, \tag{7.91}$$

$$p = g_D \quad \text{on } \Gamma, \tag{7.92}$$

where Ω is the unit square, and f and Dirichlet boundary conditions g_D are chosen on the whole boundary Γ so that the exact solution is $p = \sin(\pi x) \sin(\pi y)$. We discretize the domain with the four mesh families shown in Fig. 7.1. Figure 7.2 shows the convergence of the error for the pressure approximation measured in the

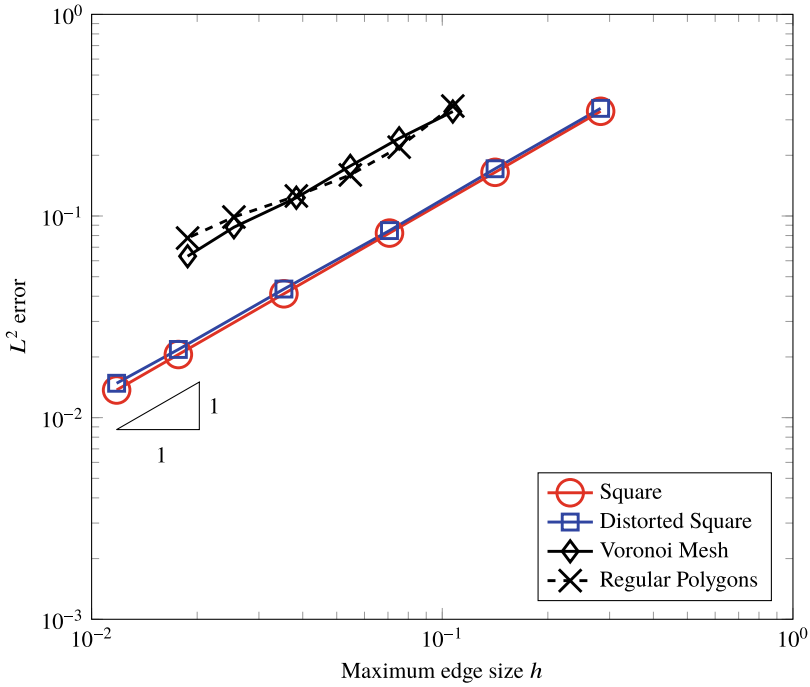


Fig. 7.2 Convergence of the relative pressure error in the L^2 norm with respect to mesh size h for different mesh discretizations

L^2 norm versus the mesh size parameter h . The optimal convergence rate for the approximation of p , which must be proportional to $\mathcal{O}(h)$, is clearly observed.

To investigate the performance of the mixed VEM in treating numerically the Richards equation, we consider the manufactured solution test case proposed in [74], with exact solution given by

$$p(x, y, t) = \begin{cases} \frac{-2(e^s - 1)}{e^s + 1} & \text{for } s \geq 0, \\ -s & \text{for } s < 0, \end{cases}$$

where $s = x - y - t$. The forcing term f and Dirichlet boundary conditions g_D on the whole boundary Γ are enforced accordingly. We assume that $\gamma = 0$ and take

$$\beta(p) = \begin{cases} \frac{\pi^2 - p^2}{2} & \text{for } p \geq 0, \\ \frac{\pi^2}{2} & \text{for } p < 0 \end{cases}.$$

The virtual element discretization is solved over the unit square Ω and from $t = 0$ up to the final time $T = 1$.

We consider the nonlinear implicit scheme given by taking the difference of both Eqs. (7.25) and (7.26) at two consecutive time instants t^{n-1} and t^n . To solve the resulting nonlinear system for the degrees of freedom of the virtual element approximations \mathbf{u}_h and p_h , we apply the Picard linearization procedure discussed in [74]:

$$\kappa(p_h^{n,i}, q_h) + (\operatorname{div} \mathbf{u}_h^n, q_h) + \left(\frac{\beta(p_h^{n,i-1}) - \beta(p_h^{n-1})}{\Delta t^n}, q_h \right) = \left(\langle f \rangle^n, q_h \right) + \kappa(p_h^{n,i-1}, q_h) \quad (7.93)$$

$$\left(\mathbf{u}_h^{n,i}, \mathbf{v}_h \right)_{\mathbf{V}_h} - \left(p_h^{n,i}, \operatorname{div} \mathbf{v}_h \right) = \left(\bar{g}^n, \mathbf{v}_h \right)_h \quad \forall \mathbf{v}_h \in \mathbf{V}_h \quad (7.94)$$

where i and $i - 1$ are two consecutive inner iteration steps and κ is a suitable constant scaling factor (in all our calculations we set $\kappa = 2$). A Newton-type linearization is an alternative and effective choice for the solution of the implicit nonlinear problem provided by the VEM. However, as this problem has only a mild non-linearity, we preferred staying on the Picard's iterative method. The computations are performed over the four mesh families shown in Fig. 7.1. The initial mesh size of all the mesh families is taken such that $h \approx 0.35$; the time increment is $\Delta T = 0.01$. Except for the Voronoi meshes, the run parameters h and ΔT are approximately halved at each refinement step until $\Delta T = 0.000625$. In the case of Voronoi meshes, the tolerance for the stopping criteria is set as 1×10^{-10} .

Figure 7.4 shows the plots of the L^2 relative errors on the pressure approximation with respect to h when the various meshes are refined. We can infer from these plots that our numerical approach yields a rate of convergence of about 0.94, which is, thus, very close to the theoretical estimate of 1 from Theorem 7.2.

Next, we study the convergence of the relative error in the pressure approximation in L^2 norm for two cases: (a) Case A: constant Δt , varying h and (b) Case B: constant h , varying Δt , when the domain is discretized with square and Voronoi meshes. For Case A, Δt is assumed to be 0.00125 and for Case B, h is considered to be ≈ 0.03143 . Figure 7.3 show the convergence of the relative error in the L^2 norm for Case A and B, respectively. It can be seen that in both cases, the rate of convergence is close to 1, as predicted by theoretical estimate (Fig. 7.4).

Remark 7.2 Finally, we note that in these practical calculations, instead of (7.93)–(7.94) we could use the alternative equation for $n > 1$

$$\left(\frac{\beta(p_h^n) - \beta(p_h^{n-1})}{\Delta t^n}, q_h \right) + (\operatorname{div} \mathbf{u}_h^n, q_h) = \left(\langle f \rangle^n, q_h \right) \quad \forall q_h \in Q_h. \quad (7.95)$$

Indeed, Eqs. (7.95) and (7.25) are equivalent as the former is readily derived by subtracting the latter equation taken at the time steps n and $n - 1$. Picard (or Newton) linearization can be also applied for the numerical resolution.

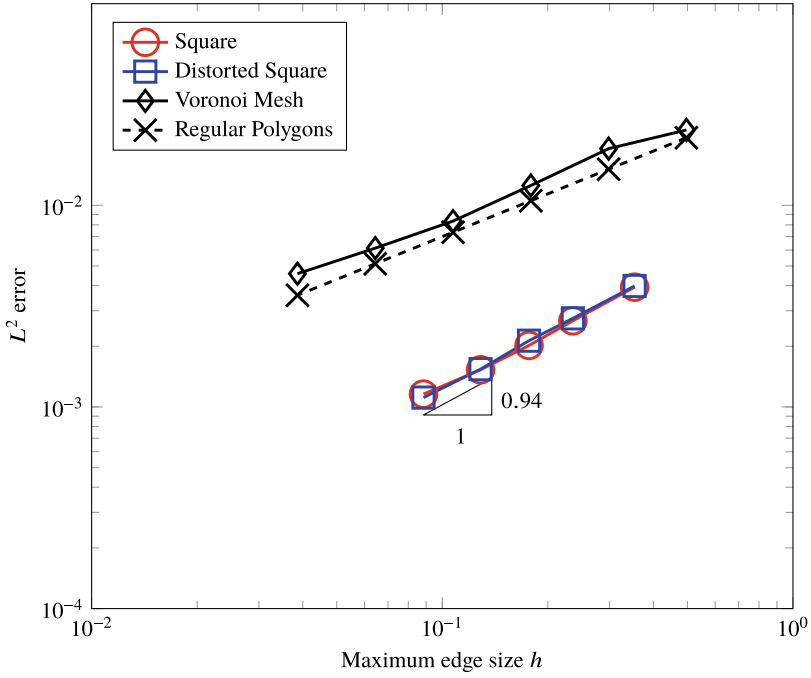


Fig. 7.3 Richards equation: convergence of the relative pressure error in the L^2 norm with respect to mesh size h for different mesh discretizations. The mesh size h and the time increment Δt are approximately halved at each refinement step

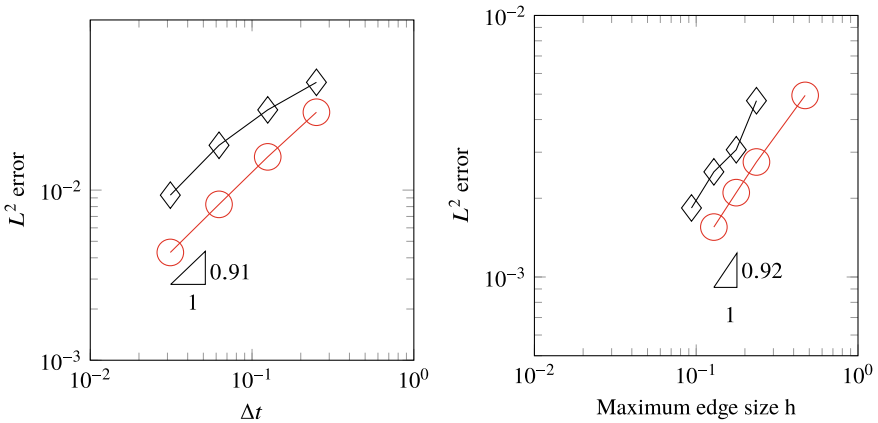


Fig. 7.4 Convergence of the relative pressure error in the L^2 norm: **a** with respect to time increment, Δt for constant mesh size h and **b** with respect to the mesh size for a constant time increment Δt . Results are shown for the square mesh family (circles) and the Voronoi mesh family (diamonds)

7.6 Conclusions

We applied the mixed virtual element method to the numerical treatment of the Richards equation in weak form for the computer modeling of flows in soils in partially to fully saturated regimes. We obtain such a variational formulation by applying the Kirchoff transformation and a preliminary integration in time. Then, we approximated the resulting nonlinear parabolic problem in mixed form by the mixed virtual element method in space on unstructured polytopal meshes to obtain the semi-discrete and fully-discrete mixed VEM. We theoretically proved that both formulations are convergent and derive an estimate of the convergence rates. Finally, we studied the behaviour and the accuracy of the mixed virtual element discretization and assessed its flexibility with respect to the geometric shape of the mesh elements. To this end, we considered a steady-state and a time-dependent benchmark problem that we solved on a set of polygonal meshes. The numerical results are in perfect agreement with the theoretical expectations from the convergence analysis.

Acknowledgements GM has been partially supported by the ERC Project CHANGE, which has received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (grant agreement No. 694515).

References

1. J.E. Aarnes, S. Krogstad, K.-A. Lie, Multiscale mixed/mimetic methods on corner-point grids. *Comput. Geosci.* **12**(3), 297–315 (2007)
2. R.A. Adams, J.J.F. Fournier, *Sobolev Spaces*. Pure and Applied Mathematics, 2 edn. (Academic Press, 2003)
3. B. Ahmad, A. Alsaedi, F. Brezzi, L.D. Marini, A. Russo, Equivalent projectors for virtual element methods. *Comput. Math. Appl.* **66**, 376–391 (2013)
4. H.W. Alt, S. Luckhaus, Quasilinear elliptic-parabolic differential equations. *Math. Z.* **183**(3), 311–341 (1983)
5. P.F. Antonietti, L. Beirão da Veiga, D. Mora, M. Verani, A stream virtual element formulation of the stokes problem on polygonal meshes. *SIAM J. Numer. Anal.* **52**(1), 386–404 (2014)
6. P.F. Antonietti, G. Manzini, M. Verani, The fully nonconforming virtual element method for biharmonic problems. *M3AS Math. Models Methods Appl. Sci.* **28**(2) (2018)
7. P.F. Antonietti, G. Manzini, M. Verani, The conforming virtual element method for polyharmonic problems. *Comput. Math. Appl.* (2019). Published online: 4 October 2019
8. T. Arbogast, M.F. Wheeler, N.-Y. Zhang, A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.* **33**(4), 1669–1687 (1996)
9. B. Ayuso de Dios, K. Lipnikov, G. Manzini, The non-conforming virtual element method. *ESAIM: Math Model. Numer. Anal.* **50**(3), 879–904 (2016)
10. R.G. Baca, J.N. Chung, D.J. Mulla, Mixed transform finite element method for solving the non-linear equation for flow in variably saturated porous media. *Internat. J. Numer. Methods Fluids* **24**(5), 441–455 (1997)
11. L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L.D. Marini, A. Russo, Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.* **23**, 119–214 (2013)

12. L. Beirão da Veiga, F. Brezzi, L.D. Marini, Virtual elements for linear elasticity problems. *SIAM J. Numer. Anal.* **51**(2), 794–812 (2013)
13. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, The Hitchhiker’s guide to the virtual element method. *Math. Models Methods Appl. Sci.* **24**(8), 1541–1573 (2014)
14. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, Virtual element methods for general second order elliptic problems on polygonal meshes. *Math. Models Methods Appl. Sci.* **26**(04), 729–750 (2015)
15. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, $H(\text{div})$ and $H(\text{curl})$ -conforming VEM. *Numer. Math.* **133**(2), 303–332 (2016)
16. L. Beirão da Veiga, F. Brezzi, L. D. Marini, A. Russo, Mixed virtual element methods for general second order elliptic problems on polygonal meshes. *ESAIM: Math. Model. Numer. Anal.* **50**(3), 727–747 (2016)
17. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, Serendipity nodal vem spaces. *Comput. Fluids* **141**, 2–12 (2016)
18. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, Virtual element methods for general second order elliptic problems on polygonal meshes. *Math. Models Methods Appl. Sci.* **26**(4), 729–750 (2016)
19. L. Beirão da Veiga, A. Chernov, L. Mascotto, A. Russo, Basic principles of hp virtual elements on quasiuniform meshes. *Math. Models Methods Appl. Sci.* **26**(8), 1567–1598 (2016)
20. L. Beirão da Veiga, K. Lipnikov, G. Manzini, Arbitrary order nodal mimetic discretizations of elliptic problems on polygonal meshes. *SIAM J. Numer. Anal.* **49**(5), 1737–1760 (2011)
21. L. Beirão da Veiga, K. Lipnikov, G. Manzini, *The Mimetic Finite Difference Method*, in *MS&A. Modeling, Simulations and Applications*, vol. 11, 1 edn. (Springer, 2014)
22. L. Beirão da Veiga, C. Lovadina, D. Mora, A virtual element method for elastic and inelastic problems on polytope meshes. *Comput. Methods Appl. Mech. Eng.* **295**, 327–346 (2015)
23. L. Beirão da Veiga, C. Lovadina, G. Vacca, Divergence free virtual elements for the Stokes problem on polygonal meshes. *ESAIM: Math. Model. Numer. Anal.* **51**(2), 509–535 (2017)
24. L. Beirão da Veiga, G. Manzini, A virtual element method with arbitrary regularity. *IMA J. Numer. Anal.* **34**(2), 782–799 (2014). <https://doi.org/10.1093/imanum/drt018> (first published online 2013)
25. L. Beirão da Veiga, G. Manzini, Residual a posteriori error estimation for the virtual element method for elliptic problems. *ESAIM: Math. Model. Numer. Anal.* **49**, 577–599 (2015)
26. L. Beirão da Veiga, G. Manzini, L. Mascotto, A posteriori error estimation and adaptivity in hp virtual elements. *Numer. Math.* **143**, 139–175 (2019)
27. L. Beirão da Veiga, G. Manzini, M. Putti, Post-processing of solution and flux for the nodal mimetic finite difference method. *Numer. Methods Partial Differ. Equ.* **31**(1), 336–363 (2015)
28. M.F. Benedetto, S. Berrone, A. Borio, The virtual element method for underground flow simulations in fractured media, in *Advances in Discretization Methods*. SEMA SIMAI Springer Series, vol. 12 (Springer International Publishing, Switzerland, 2016), pp. 167–186
29. M.F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, S. Scialò, A hybrid mortar virtual element method for discrete fracture network simulations. *J. Comput. Phys.* **306**, 148–166 (2016)
30. M.F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, S. Scialò, The virtual element method for discrete fracture network flow and transport simulations, in *ECCOMAS Congress 2016—Proceedings of the 7th European Congress on Computational Methods in Applied Sciences and Engineering*, vol. 2 (2016), pp. 2953–2970
31. M.F. Benedetto, S. Berrone, S. Pieraccini, S. Scialò, The virtual element method for discrete fracture network simulations. *Comput. Methods Appl. Mech. Eng.* **280**, 135–156 (2014)
32. M.F. Benedetto, S. Berrone, S. Scialò, A globally conforming method for solving flow in discrete fracture networks using the virtual element method. *Finite Elem. Anal. Des.* **109**, 23–36 (2016)
33. E. Benvenuti, A. Chiozzi, G. Manzini, N. Sukumar, Extended virtual element method for the Laplace problem with singularities and discontinuities. *Comput. Methods Appl. Mech. Eng.* **356**, 571–597 (2019)

34. L. Bergamaschi, M. Putti, Mixed finite elements and newton-type linearizations for the solution of richards' equation. *Int. J. Numer. Methods Eng.* **45**(8), 1025–1046 (1999)
35. S. Berrone, A. Borio, Orthogonal polynomials in badly shaped polygonal elements for the virtual element method. *Finite Elem. Anal. Des.* **129**, 14–31 (2017)
36. S. Berrone, A. Borio, Manzini, SUPG stabilization for the nonconforming virtual element method for advection-diffusion-reaction equations. *Comput. Methods Appl. Mech. Eng.* **340**, 500–529 (2018)
37. S. Berrone, A. Borio, S. Scialò, A posteriori error estimate for a PDE-constrained optimization formulation for the flow in DFNs. *SIAM J. Numer. Anal.* **54**(1), 242–261 (2016)
38. S. Berrone, S. Pieraccini, S. Scialò, Towards effective flow simulations in realistic discrete fracture networks. *J. Comput. Phys.* **310**, 181–201 (2016)
39. S. Berrone, S. Pieraccini, S. Scialò, F. Vicini, A parallel solver for large scale DFN flow simulations. *SIAM J. Sci. Comput.* **37**(3), C285–C306 (2015)
40. S.C. Brenner, L.R. Scott, *The Mathematical Theory of Finite Element Methods, Texts in Applied Mathematics*, vol. 15 (Springer, New York, 1994)
41. F. Brezzi, A. Buffa, K. Lipnikov, Mimetic finite differences for elliptic problems. *M2AN Math. Model. Numer. Anal.* **43**, 277–295 (2009)
42. F. Brezzi, R.S. Falk, L.D. Marini, Basic principles of mixed virtual element methods. *ESAIM Math. Model. Numer. Anal.* **48**(4), 1227–1240 (2014)
43. F. Brezzi, K. Lipnikov, M. Shashkov, Convergence of mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Num. Anal.* **43**, 1872–1896 (2005)
44. F. Brezzi, L.D. Marini, Virtual element methods for plate bending problems. *Comput. Methods Appl. Mech. Eng.* **253**, 455–462 (2013)
45. A. Cangiani, E.H. Georgoulis, T. Pryer, O.J. Sutton, A posteriori error estimates for the virtual element method. *Numer. Math.* 1–37 (2017)
46. A. Cangiani, V. Gyrya, G. Manzini, The non-conforming virtual element method for the Stokes equations. *SIAM J. Numer. Anal.* **54**(6), 3411–3435 (2016)
47. A. Cangiani, V. Gyrya, G. Manzini, O. Sutton, Chapter 14: Virtual element methods for elliptic problems on polygonal meshes, in *Generalized Barycentric Coordinates in Computer Graphics and Computational Mechanics*, K. Hormann, N. Sukumar, eds. (CRC Press, Taylor & Francis Group, 2017), pp. 1–20
48. A. Cangiani, G. Manzini, A. Russo, N. Sukumar, Hourglass stabilization of the virtual element method. *Int. J. Numer. Methods Eng.* **102**(3–4), 404–436 (2015)
49. A. Cangiani, G. Manzini, O. Sutton. Conforming and nonconforming virtual element methods for elliptic problems. *IMA J. Numer. Anal.* **37**, 1317–1354 (2017) (online August 2016)
50. M.A. Celia, E.T. Bouloutas, R.L. Zarba, General mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.* **26**, 1483–1496 (1990)
51. O. Certik, F. Gardini, G. Manzini, L. Mascotto, G. Vacca, The p- and hp-versions of the virtual element method for elliptic eigenvalue problems. *Comput. Math. Appl.* (2019). Published online: 31 October 2019
52. O. Certik, F. Gardini, G. Manzini, G. Vacca, The virtual element method for eigenvalue problems with potential terms on polytopical meshes. *Appl. Math.* **63**(3), 333–365 (2018)
53. D.A. Di Pietro, J. Droniou, G. Manzini, Discontinuous skeletal gradient discretisation methods on polytopal meshes. *J. Comput. Phys.* **355**, 397–425 (2018)
54. R. Eymard, M. Gutnic, D. Hilhorst, The finite volume method for Richards equation. *Comput. Geosci.* **3**(3–4), 259–294 (2000), 1999
55. M.W. Farthing, F.L. Ogden, Numerical solution of Richards' equation: a review of advances and challenges. *Soil Sci. Soc. Am. J* 1257–1269 (2017)
56. C. Fassino, G. Manzini, Fast-secant algorithms for the non-linear Richards' equation. *Commun. Numer. Methods Eng.* **14**(10), 921–930 (1998)
57. F. Gardini, G. Manzini, G. Vacca, The nonconforming virtual element method for eigenvalue problems. *ESAIM: Math. Model. Numer. Anal.* **53**, 749–774 (2019). Accepted for publication: 29 November 2018. <https://doi.org/10.1051/m2an/2018074>

58. W. Jäger, J. Kav cur, Solution of doubly nonlinear and degenerate parabolic problems by relaxation schemes. *ESAIM: Math. Model. Numer. Anal.* **29**(5), 605–627 (1995)
59. H. Li, M.W. Farthing, C.N. Dawson, C.T. Miller, Local discontinuous Galerkin approximations to Richards' equation. *Adv. Water Resour.* **30**(3), 555–575 (2007)
60. H. Li, M.W. Farthing, C.T. Miller, Adaptive local discontinuous Galerkin approximation to Richards' equation. *Adv. Water Resour.* **30**(9), 1883–1901 (2007)
61. J.-L. Lions, E. Magenes, *Problèmes aux limites non homogènes et applications*, vol. 2. Travaux et Recherches Mathématiques, No. 18 (Dunod, Paris, 1968)
62. K. Lipnikov, G. Manzini, A high-order mimetic method for unstructured polyhedral meshes. *J. Comput. Phys.* **272**, 360–385 (2014)
63. K. Lipnikov, G. Manzini, M. Shashkov, Mimetic finite difference method. *J. Comput. Phys.* **257**(Part B), 1163–1227 (2014) Review paper
64. G. Manzini, S. Ferraris, Mass-conservative finite-volumes on unstructured grids for the Richards' equation. *Adv. Water Resour.* **27**, 1199–1215 (2004)
65. G. Manzini, K. Lipnikov, J.D. Moulton, M. Shashkov, Convergence analysis of the mimetic finite difference method for elliptic problems with staggered discretizations of diffusion coefficients. *SIAM J. Numer. Anal.* **55**(6), 2956–2981 (2017)
66. G. Manzini, A. Russo, N. Sukumar, New perspectives on polygonal and polyhedral finite element methods. *Math. Models Methods Appl. Sci.* **24**(8), 1621–1663 (2014)
67. D. Mora, G. Rivera, R. Rodríguez, A virtual element method for the Steklov eigenvalue problem. *Math. Models Methods Appl. Sci.* **25**(08), 1421–1445 (2015)
68. S. Natarajan, P.A. Bordas, E.T. Ooi, Virtual and smoothed finite elements: a connection and its application to polygonal/polyhedral finite element methods. *Int. J. Numer. Methods Eng.* **104**(13), 1173–1199 (2015)
69. R.H. Nochetto, C. Verdi, Approximation of degenerate parabolic problems using numerical integration. *SIAM J. Numer. Anal.* **25**(4), 784–814 (1988)
70. F. Otto, L^1 -contraction and uniqueness for quasilinear elliptic-parabolic equations. *C. R. Acad. Sci. Paris Sér. I Math.* **321**(8), 1005–1010 (1995)
71. G.H. Paulino, A.L. Gain, Bridging art and engineering using Escher-based virtual elements. *Struct. Multidiscip. Optim.* **51**(4), 867–883 (2015)
72. I. Perugia, P. Pietra, A. Russo, A plane wave virtual element method for the Helmholtz problem. *ESAIM: Math. Model. Numer. Anal.* **50**(3), 783–808 (2016)
73. I.S. Pop, Error estimates for a time discretization method for the Richards' equation. *Comput. Geosci.* **6**(2), 141–160 (2002)
74. I.S. Pop, F. Radu, P. Knabner, Mixed finite elements for the Richards' equation: linearization procedure. *J. Comput. Appl. Math.* **168**, 365–373 (2004)
75. F. Radu, I.S. Pop, P. Knabner, Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. *SIAM J. Numer. Anal.*, **42**(4), 1452–1478 (electronic) (2004)
76. L.A. Richards, Capillary conduction of liquids through porous mediums. *Physics (IEEE)* **1**, 318 (1931)
77. L.F. Richardson, P. Lynch, *Weather Prediction by Numerical Process*, 2nd edn. (Cambridge University Press, Cambridge Mathematical Library, 2007)
78. E. Schneid, P. Knabner, F. Radu, A priori error estimates for a mixed finite element discretization of the Richards' equation. *Numer. Math.* **98**(2), 353–370 (2004)
79. G. Vacca, L. Beirão da Veiga, Virtual element methods for parabolic problems on polygonal meshes. *Numer. Methods Partial Differ. Equ. Int. J.* **31**(6), 2110–2134 (2015)
80. C.J. van Duijn, L.A. Peletier, Nonstationary filtration in partially saturated porous media. *Arch. Ration. Mech. Anal.* **78**, 173–198 (1982)
81. C.S. Woodward, C.N. Dawson, Analysis of expanded mixed finite element methods for a nonlinear parabolic equation modeling flow into variably saturated porous media. *SIAM J. Numer. Anal.*, **37**(3), 701–724 (electronic) (2000)
82. P. Wriggers, W.T. Rust, B.D. Reddy, A virtual element method for contact. *Comput. Mech.* **58**(6), 1039–1050 (2016)

83. I. Yotov, A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arising in porous media flow. *East-West J. Numer. Math.* **5**(3), 211–230 (1997)
84. J. Zhao, S. Chen, B. Zhang, The nonconforming virtual element method for plate bending problems. *Math. Models Methods Appl. Sci.* **26**(9), 1671–1687 (2016)

Chapter 8

Performances of the Mixed Virtual Element Method on Complex Grids for Underground Flow



Alessio Fumagalli, Anna Scotti , and Luca Formaggia 

Abstract The numerical simulation of physical processes in the underground frequently entails challenges related to the geometry and/or data. The former are mainly due to the shape of sedimentary layers and the presence of fractures and faults, while the latter are connected to the properties of the rock matrix which might vary abruptly in space. The development of approximation schemes has recently focused on the overcoming of such difficulties with the objective of obtaining numerical schemes with good approximation properties. In this work we carry out a numerical study on the performance of the Mixed Virtual Element Method (MVEM) for the solution of a single-phase flow model in fractured porous media. This method is able to handle grid cells of polytopal type and treat hybrid dimensional problems. It has been proven to be robust with respect to the variation of the permeability field and of the shape of the elements. Our numerical experiments focus on two test cases that cover several of the aforementioned critical aspects.

Keywords Virtual element method · Fracture flow · Grid generation · Mixed-dimensional problems · spe10 benchmark

8.1 Introduction

The numerical simulation of subsurface flows is of paramount importance in many environmental and energy related applications such as the management of groundwater resources, geothermal energy production, subsurface storage of carbon dioxide.

A. Fumagalli · A. Scotti (✉) · L. Formaggia
MOX-Laboratory for Modeling and Scientific Computing, Dipartimento di Matematica,
Politecnico di Milano, via Bonardi 9, 20133 Milan, Italy
e-mail: anna.scotti@polimi.it

A. Fumagalli
e-mail: alessio.fumagalli@polimi.it

L. Formaggia
e-mail: luca.formaggia@polimi.it

The physical processes are usually modeled, under suitable assumptions, by Darcy's law and its generalization to multiphase flow.

In spite of the simplicity of the Darcy model, the simulation of subsurface flow is often a numerical challenge due to the strong heterogeneity of the coefficients, porosity and permeability of the porous medium, and to the geometrical complexity of the domains of interest. At the spatial scale of reservoirs, or sedimentary basins, the porous medium has a layered structure due to the deposition and erosion of sediments, and tectonic stresses can create, over millions or years, deformations, folds, faults and fractures. In realistic cases the construction of a computational grid that honours the geometry of layers and a large number of fractures is not only a difficult task, but can also give poor results in terms of quality, creating, for instance, very small or badly shaped elements in the vicinity of the interfaces.

In the framework of Finite Volume and Finite Elements methods one possibility is to consider formulations that allow for coarse/agglomerated and regular grids cut by the interfaces in arbitrary ways. The Embedded Discrete Fracture Model, for instance, [43, 47, 53], can represent permeable fracture that cut the background grid by adding additional transmissibility in the matrix resulting from the Finite Volumes discretization; on the other hand the eXtended Finite Element Method can be used to generalize a classical FEM discretization allowing for discontinuities inside an element of the grid, see for example [23, 26, 31, 32] for the application of this technique to Darcy's problem.

A promising alternative consists in the use of numerical methods that are robust in the presence of more general grids, in particular polygonal/polyhedral grids, and that impose mild restriction on element shape: this is the case for the Virtual Element Method (VEM), introduced in [6, 7, 18] and successfully applied now to a variety of problems, including elliptic problems in mixed form which is the case of the Darcy model considered in this work. See also [9, 10, 38, 40, 41]. By avoiding the explicit construction of basis function VEM can indeed handle very general grids, which might be useful in the aforementioned cases where the heterogeneity of the medium and the presence of internal interfaces pose constraints to grid generation. In the context of porous media simulations, mixed methods, i.e. methods that consider both velocity and pressure as unknowns of the problem, are of particular interest since they provide a good approximation of pressure as well as an accurate (and conservative) velocity field. For these reasons, we focus our attention on the Mixed Virtual Element Method (MVEM).

MVEM may be considered to belong to the general family of "Discontinuous Skeletal Methods" described in [14]. Its formulation falls in the finite element method framework, where however shape functions are defined only implicitly by their properties, and degrees of freedom are obtained by suitable projection operators that enable to compute the approximate bilinear forms. The latter include a computable stabilization term necessary to recover well posedness. Low-order MVEM gives rise to an algebraic problem akin to that produced by Mimetic Finite Differences. A link among Mimetic Finite Differences and Hybrid Finite Volumes may be found also in [28].

The aim of this work is to consider practical grid generation strategies to deal with such complex geometries and to test the performance of the MVEM method on the different types of grid proposed. In particular, we want to investigate the impact of grid type and element shape on properties of the linear system such as sparsity and condition number, and eventually compare the errors. To this aim we will consider two test cases from the literature, in particular two layers from the well-known 10th SPE Comparative Solution Project (SPE10) dataset, described in [22], characterized by a complex permeability field, and a test case for fractured media taken from [30]. We focus our attention on grid generation strategies that can be applicable in realistic cases: if it is certainly true that MVEM can handle general polytopal grid the construction of such grids is often a difficult task. For this reason, in addition to classical Delaunay triangular grids we consider the case of Voronoi grids, rectangular Cartesian grids cut by fractures, and grids generated by agglomeration. This latter strategy can be applied as a post-processing to all other grid types with the advantage of reducing the number of unknowns. For the numerical implementation of the test cases we have used the publicly available library PorePy [46].

The paper is structured as follows: in Sect. 8.2 we present the mathematical model, i.e. the single phase Darcy model in the presence of fractures approximated as codimension 1 interfaces. Section 8.3 is devoted to the weak formulation of the problem just introduced. Section 8.4 introduces the numerical discretization by the Virtual Element method, while in Sect. 8.5 we describe the grid generation strategies used in the paper. Section 8.6 presents the numerical tests, and Sect. 8.7 is devoted to conclusions.

8.2 Governing Equations

We now introduce the mathematical models considered in this work. The realistic modeling of subsurface flows requires a complex set of non-linear equations and constitutive laws, however one of the key ingredient (upon a suitable linearisation) is the single-phase flow model for a porous media, like the one already in Chap. 1, Sect. 1.2.1, based on Darcy's law and mass conservation. We are here studying this model, keeping in mind that it might be seen as a part of a more complex model. In addition, it is of our interest to consider also fractures in the porous media, and this calls for a more sophisticated approach.

As already mentioned, we set our study in a saturated porous medium represented by the domain $\Omega \subset \mathbb{R}^2$. The boundary of Ω , indicated with $\partial\Omega$, is supposed regular enough (e.g. Lipschitz continuous). The boundary is divided into two disjoint parts $\partial_u\Omega$ and $\partial_p\Omega$ such that $\partial_u\Omega \cap \partial_p\Omega = \emptyset$ and $\overline{\partial_u\Omega} \cup \overline{\partial_p\Omega} = \overline{\partial\Omega}$. These portions of the boundary will be used to define boundary conditions.

8.2.1 Single-Phase Flow in the Bulk Domain

We briefly recall the mathematical model of single-phase flow in porous media, referring to classical results in literature, see [4], for details. We are interested in the computation of the vector field Darcy velocity \mathbf{u} and scalar field pressure p , which are solutions of the following problem

$$\begin{aligned} \mathbf{u} + K \nabla p &= \mathbf{0} && \text{in } \Omega, \\ \nabla \cdot \mathbf{u} &= f && \\ \mathbf{u} \cdot \mathbf{n}_\partial &= \bar{u} && \text{on } \partial_u \Omega, \\ p &= \bar{p} && \text{on } \partial_p \Omega. \end{aligned} \tag{8.1}$$

The parameter K is the 2×2 permeability tensor, which is symmetric and positive definite. For simplicity, the dynamic viscosity of the fluid is included into K . The source or sink term is named f . Finally, \mathbf{n}_∂ is the outward unit normal on $\partial\Omega$, \bar{u} and \bar{p} given boundary data.

We recall that the permeability tensor, for real applications, may vary several order of magnitude from region to region (i.e., grid cells) and can be discontinuous.

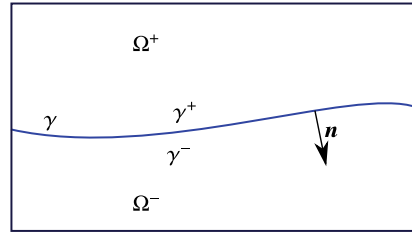
8.2.2 Fracture Flow

We are interested in the simulation of single-phase flow in porous media in the presence of fractures. For simplicity we start with a single fracture. The model we are considering is the result of a model reduction procedure that approximates the fracture as a lower dimensional object and derives new equations and coupling conditions for the Darcy velocity and pressure both in the fracture and surrounding porous medium. More details on this subject can be found in the following, not exhaustive, list of works [2, 3, 12, 15, 17, 20, 23, 32, 33, 41, 48, 51, 58, 60], as well as Chaps. 3, 4 and 5 of this Book.

In the following, the fracture is indicated with γ , and quantities related to the porous media and the fracture are indicated with the subscript Ω and γ , respectively. The fracture is described by a planar surface with normal vector denoted by \mathbf{n} , which also defines a positive and negative side of γ , indicated as γ^+ and γ^- , see Fig. 8.1 as an example. Given a field u in $\Omega \setminus \gamma$ we indicate its trace on γ^+ and γ^- as tru_+ and tru_- , respectively.

The fracture is characterized by an aperture ϵ_γ which, in the reduced model where the fracture has co-dimension one, is only a model parameter. Finally, if the fracture touches the boundary we can apply natural or essential given boundary conditions; we denote as $\partial_p \gamma$ and $\partial_u \gamma$ the portions of $\partial\gamma$ where pressure and velocity are imposed. We assume that $\partial_p \overset{\circ}{\gamma} \cap \partial_u \overset{\circ}{\gamma} = \emptyset$ as well as $\partial_p \gamma \cup \partial_u \gamma = \partial\gamma$. If a fracture tip does

Fig. 8.1 Hybrid-dimensional representation of a fracture immersed in a porous media



not touch the physical boundary a no-flow condition is imposed, so in this case we assume that the immersed tip belongs to $\partial_u \gamma$ with an homogeneous condition.

We recall the system of equations that will be used in the sequel. In the bulk porous medium $\Omega \setminus \gamma$ the problem is governed by the classic Darcy’s equations already presented in (8.1), which we rewrite using the subscript Ω to identify quantities in $\Omega \setminus \gamma$

$$\begin{aligned}
 \mathbf{u}_\Omega + K_\Omega \nabla p_\Omega &= \mathbf{0} && \text{in } \Omega \setminus \gamma, \\
 \nabla \cdot \mathbf{u}_\Omega &= f_\Omega && \\
 \mathbf{u}_\Omega \cdot \mathbf{n}_\partial &= \bar{u}_\Omega && \text{on } \partial_u \Omega \setminus \partial \gamma, \\
 p_\Omega &= \bar{p}_\Omega && \text{on } \partial_p \Omega \setminus \partial \gamma.
 \end{aligned}
 \tag{8.2a}$$

We assume that also the flow in the fracture is governed by Darcy’s law, however the differential operators operate now on the tangent space. Yet, for the sake of simplicity, with an abuse of notation we use the same symbols to denote them. The system of equations in the fracture is then given by

$$\begin{aligned}
 \epsilon_\gamma^{-1} \mathbf{u}_\gamma + K_\gamma \nabla p_\gamma &= \mathbf{0} && \text{in } \gamma, \\
 \nabla \cdot \mathbf{u}_\gamma - \text{tr} \mathbf{u}_+ \cdot \mathbf{n} + \text{tr} \mathbf{u}_- \cdot \mathbf{n} &= f_\gamma && \\
 \mathbf{u}_\gamma \cdot \mathbf{n}_\partial &= \bar{u}_\gamma && \text{on } \partial_u \gamma, \\
 p_\gamma &= \bar{p}_\gamma && \text{on } \partial_p \gamma.
 \end{aligned}
 \tag{8.2b}$$

Here, \bar{u}_γ and \bar{p}_γ are given boundary data, and we recall that possible fracture tips are in $\partial_u \gamma$ with $\bar{u}_\gamma = 0$. The parameter K_γ is the tangential effective permeability in γ . In the 2D setting, where the reduced fracture model is one-dimensional, K_γ is a positive quantity. In the 3D setting, it may be in general a rank-2 symmetric and positive tensor. We may note in the equation representing the conservation of mass the presence of an additional term that describes the flux exchange with the surrounding porous media. To close the problem we need to complete the coupling between fracture and bulk, and we consider the following Robin-type condition on both sides of γ

$$\epsilon_\gamma \text{tr} \mathbf{u}_\pm \cdot \mathbf{n} \pm \kappa_\gamma (p_\gamma - \text{tr} p_\pm) = 0 \quad \text{on } \gamma^\pm,
 \tag{8.2c}$$

with $\kappa_\gamma > 0$ being the normal effective permeability. Problem (8.2) consists of the system of equations that describe the Darcy velocity and pressure in both the fracture and surrounding porous medium. An analysis may be found, for instance, in [33] or [40].

The case of $N > 1$ non-intersecting fractures the problem is analogous to the one just described where $\gamma = \cup_{i=1}^N \gamma_i$. However, if two or more fractures intersect we need to introduce new conditions to describe the flux interchange between connected fractures. At each intersection ι we denote with I_ι the set of intersecting fractures and we consider the following conditions on ι ,

$$\begin{cases} \epsilon_\iota \alpha_j \operatorname{tr} \mathbf{u}_j \cdot \mathbf{t}_j + \kappa_\iota (p_\iota - \operatorname{tr} p_j) = 0 & \forall \gamma_j \in I_\iota \\ \sum_{\gamma_j \in I_\iota} \alpha_j \operatorname{tr} \mathbf{u}_j \cdot \mathbf{t}_j = 0 \end{cases} \quad \text{on } \iota, \quad (8.3)$$

where ϵ_ι is the measure of the intersection, p_ι is the pressure at the intersection, κ_ι is the permeability at the intersection and $\alpha \in \{-1, 1\}$ depends on the orientation chosen for the normal \mathbf{t}_j to $\partial\gamma_j$ at the intersection. Note that \mathbf{t}_j is indeed on the tangent plane of γ_j . System (8.3) can be simplified by noting that it implies that p_ι is equal to the average of the p_j .

8.3 Weak Formulation

The numerical scheme that we will present in Sect. 8.4 is based on the weak formulation of problem (8.1) and (8.2). Therefore, we will present in the following the functional setting and the weak form we have used as basis for the numerical discretization. We indicate with $L^2(A)$ the Lebesgue space of square integrable functions on A , while $H_{\operatorname{div}}(A)$ is the space of square integrable vector functions whose distributional divergence is in $L^2(A)$. They are Hilbert spaces with standard norms and inner products. In particular, we denote with $(\cdot, \cdot)_A$ the $L^2(A)$ -scalar product. Moreover, given a functional space V and its dual V' we use $\langle a, b \rangle$, with $a \in V$ and $b \in V'$ to denote the duality pairing between the given functional spaces.

8.3.1 Single-Phase Bulk Flow Without Fractures

If fractures are not present, the setting is rather standard. For simplicity, we assume homogeneous essential boundary conditions $\bar{u}_\Omega = 0$, otherwise a lifting technique can be used to recover the original problem. We introduce the following functional spaces for vector and scalar field, respectively,

$$V(\Omega) = \{\mathbf{v} \in H_{\operatorname{div}}(\Omega) : \operatorname{tr} \mathbf{v} \cdot \mathbf{n}_\partial = 0 \text{ on } \partial_u \Omega\} \quad \text{and} \quad Q(\Omega) = L^2(\Omega). \quad (8.4)$$

Here tr is the normal trace operator $\text{tr} : H_{\text{div}}(\Omega) \rightarrow H^{-\frac{1}{2}}(\partial_u \Omega)$, which is linear and bounded, see [13].

We can now introduce the following bilinear forms and functionals

$$\begin{aligned} a_\Omega : V(\Omega) \times V(\Omega) &\rightarrow \mathbb{R} : a_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) = (H_\Omega \mathbf{u}_\Omega, \mathbf{v}_\Omega)_\Omega \\ b_\Omega : V(\Omega) \times Q(\Omega) &\rightarrow \mathbb{R} : b_\Omega(\mathbf{v}_\Omega, p_\Omega) = -(\nabla \cdot \mathbf{v}_\Omega, q_\Omega)_\Omega \\ G_\Omega : V(\Omega) &\rightarrow \mathbb{R} : G_\Omega(\mathbf{v}_\Omega) = -(\text{tr} \mathbf{v}_\Omega \cdot \mathbf{n}_\partial, \bar{p}_\Omega) \\ F_\Omega : Q(\Omega) &\rightarrow \mathbb{R} : F_\Omega(q_\Omega) = -(f_\Omega, q_\Omega)_\Omega \end{aligned}$$

where $H_\Omega = K_\Omega^{-1}$. We assume that $K_\Omega \in [L^\infty(\Omega)]^{2 \times 2}$, with $\underline{\alpha} \|\mathbf{y}\|^2 \leq \mathbf{y}^T K_\Omega \mathbf{y} \leq \bar{\alpha} \|\mathbf{y}\|^2$, a.e. in Ω , where $\mathbf{y} \in \mathbb{R}^2$ and $0 < \underline{\alpha} \leq \bar{\alpha}$.

Furthermore, we take $\bar{p}_\Omega \in H_{00}^{\frac{1}{2}}(\partial_p \Omega)$, and $f_\Omega \in L^2(\Omega)$. Let us note that $a_\Omega : V(\Omega) \times V(\Omega) \rightarrow \mathbb{R}$ is continuous, coercive and symmetric, being K_Ω symmetric.

We can now state the weak formulation of our problem: find $(\mathbf{u}_\Omega, p_\Omega) \in V(\Omega) \times Q(\Omega)$ such that

$$\begin{aligned} a_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) + b_\Omega(\mathbf{v}_\Omega, p_\Omega) &= G_\Omega(\mathbf{v}_\Omega) \quad \forall \mathbf{v}_\Omega \in V(\Omega) \\ b_\Omega(\mathbf{u}_\Omega, q_\Omega) &= F_\Omega(q_\Omega) \quad \forall q_\Omega \in Q(\Omega) \end{aligned} \quad (8.5)$$

The previous problem is well posed, provided $|\partial_p \Omega| > 0$. See, for example, [13] for a proof.

8.3.2 Fracture Flow

We extend now the weak formulation for problem (8.2), with the simplifying assumption that only one fracture is considered. Its extension to multiple fractures is straightforward, see for example [15, 33]. Also in this case we assume homogeneous essential boundary conditions, otherwise a lifting technique can be used.

We need to introduce the space $H_{\text{div}}(\Omega \setminus \gamma)$ as the space of vector function in $L^2(\Omega \setminus \gamma)$ (which may be identified by $L^2(\Omega)$ since γ has zero measure) whose distributional divergence is in $L^2(\Sigma)$ for all measurable $\Sigma \subset (\Omega \setminus \gamma)$. We need also to impose some extra regularity on the trace on γ^\pm , due to the Robin-type condition (8.2c). The reader may refer to [13, 36, 48] for a more detailed discussion on this matter. In particular, we require that, for a $\mathbf{v}_\Omega \in H_{\text{div}}(\Omega \setminus \gamma)$, $\text{tr} \mathbf{v}_+ \cdot \mathbf{n} \in L^2(\gamma)$ and $\text{tr} \mathbf{v}_- \cdot \mathbf{n} \in L^2(\gamma)$, where tr here indicates the trace of \mathbf{v} on the two sides of the fracture. This space is equipped with the inner product

$$(\mathbf{u}, \mathbf{v})_{H_{\text{div}}(\Omega \setminus \gamma)} = (\mathbf{u}, \mathbf{v})_\Omega + (\nabla \cdot \mathbf{u}, \nabla \cdot \mathbf{v})_\Omega + (\text{tr} \mathbf{u}_+ \cdot \mathbf{n}, \text{tr} \mathbf{v}_+ \cdot \mathbf{n})_\gamma + (\text{tr} \mathbf{u}_- \cdot \mathbf{n}, \text{tr} \mathbf{v}_- \cdot \mathbf{n})_\gamma,$$

and induced norm. The new space for vector fields in the bulk is given by

$$\hat{V}(\Omega) = \{\mathbf{v}_\Omega \in \mathbf{H}_{\text{div}}(\Omega \setminus \gamma) : \text{tr} \mathbf{v}_\Omega \cdot \mathbf{n}_\partial = 0 \text{ on } \partial_u \Omega\}.$$

The functional spaces for vector and scalar fields defined on the fracture are

$$V(\gamma) = \{\mathbf{v}_\gamma \in \mathbf{H}_{\text{div}}(\gamma) : \text{tr} \mathbf{v}_\gamma \cdot \mathbf{n}_\partial = 0\} \quad \text{and} \quad Q(\gamma) = L^2(\gamma),$$

where in this case the trace operator in $V(\gamma)$ is given by $\text{tr} : \mathbf{H}_{\text{div}}(\gamma) \rightarrow H^{-\frac{1}{2}}(\partial_u \gamma)$. Note that in the case of 2D problems like the ones treated in this work, $V(\gamma)$ is in fact a subspace of $H^1(\gamma)$ and the trace reduces to the value at the boundary point.

We introduce now the bilinear forms and functional for the weak formulation of problem (8.2). First, we modify the bilinear form a_Ω by taking into account the coupling terms from (8.2c) as

$$\hat{a}_\Omega : \hat{V}(\Omega) \times \hat{V}(\Omega) \rightarrow \mathbb{R} : \quad \hat{a}_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) = a_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega)_\Omega + \sum_{* \in \{+, -\}} (\eta_\gamma \text{tr} \mathbf{u}_* \cdot \mathbf{n}, \text{tr} \mathbf{u}_* \cdot \mathbf{n})_\gamma$$

where $\eta_\gamma = \epsilon_\gamma \kappa_\gamma^{-1}$ and we have assumed that $\eta_\gamma \in L^\infty(\gamma)$. Second, the bilinear forms associated with the fracture are given by

$$\begin{aligned} a_\gamma : V(\gamma) \times V(\gamma) &\rightarrow \mathbb{R} : \quad a_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) = (H_\gamma \mathbf{u}_\gamma, \mathbf{v}_\gamma)_\gamma \\ b_\gamma : V(\gamma) \times Q(\gamma) &\rightarrow \mathbb{R} : \quad b_\gamma(\mathbf{v}_\gamma, p_\gamma) = -(\nabla \cdot \mathbf{v}_\gamma, p_\gamma)_\gamma \\ G_\gamma : V(\gamma) &\rightarrow \mathbb{R} : \quad G_\gamma(\mathbf{v}_\gamma) = -\langle \text{tr} \mathbf{v}_\gamma \cdot \mathbf{n}_\partial, \bar{p}_\gamma \rangle \\ F_\gamma : V(\gamma) \times \mathbb{R} &\rightarrow \mathbb{R} : \quad F_\gamma(q_\gamma) = -(f_\gamma, q_\gamma)_\gamma \end{aligned}$$

where we have $H_\gamma^{-1} = \epsilon_\gamma K_\gamma$ and we have assumed that $H_\gamma \in L^\infty(\gamma)$, $\bar{p}_\gamma \in H^{\frac{1}{2}}(\partial_p \gamma)$, and $f_\gamma \in L^2(\gamma)$. Third, we introduce the bilinear forms responsible for the flux exchange between the fracture and the bulk medium

$$\begin{aligned} c^\pm : \hat{V}(\Omega) \times Q(\gamma) &\rightarrow \mathbb{R} : \quad c^\pm(\mathbf{u}_\Omega, q_\gamma) = \pm(\text{tr} \mathbf{u}_\pm \cdot \mathbf{n}, q_\gamma)_\gamma \\ c : \hat{V}(\Omega) \times Q(\gamma) &\rightarrow \mathbb{R} : \quad c(\mathbf{u}_\Omega, q_\gamma) = \sum_{* \in \{+, -\}} c^*(\mathbf{u}_\Omega, q_\gamma). \end{aligned}$$

Finally, we can write the weak formulation for problem (8.2): find $(\mathbf{u}_\Omega, p_\Omega, \mathbf{u}_\gamma, p_\gamma) \in \hat{V}(\Omega) \times Q(\Omega) \times V(\gamma) \times Q(\gamma)$ such that

$$\begin{aligned} \hat{a}_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) + b_\Omega(\mathbf{v}_\Omega, p_\Omega) + c(\mathbf{v}_\Omega, p_\gamma) &= G_\Omega(\mathbf{v}_\Omega) & \forall \mathbf{v}_\Omega \in \hat{V}(\Omega) \\ b_\Omega(\mathbf{u}_\Omega, q_\Omega) &= F_\Omega(q_\Omega) & \forall q_\Omega \in Q(\Omega) \\ a_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) + b_\gamma(\mathbf{v}_\gamma, p_\gamma) &= G_\gamma(\mathbf{v}_\gamma) & \forall \mathbf{v}_\gamma \in V(\gamma) \\ b_\gamma(\mathbf{u}_\gamma, q_\gamma) + c(\mathbf{u}_\Omega, q_\gamma) &= F_\gamma(q_\gamma) & \forall q_\gamma \in Q(\gamma) \end{aligned} \quad (8.6)$$

The reader can refer to [23, 26, 32] for proofs of the well posedness of the problem, provided suitable boundary conditions.

8.4 Numerical Approximation by MVEM

The challenges in terms of heterogeneity of physical data and complexity of the geometry due to the presence of fractures influence the choice of the numerical scheme. A possible choice is the mixed finite element method, see [13, 56, 57]. However, this class of methods, capable of providing accurate results for pressure and velocity fields, even in the presence of high heterogeneities, requires grids made either of simplexes (triangles or tetrahedra) or quad/hexahedra. This can be inefficient for the problem at hand, where instead methods able to operate on grids formed by arbitrary polytopes are rather appealing. For this reason finite volume schemes, see [27] for a review, are very much used in practice. However, they normally treat the primal formulation and require good quality grids to obtain an accurate solution and a good reconstruction of the velocity field. Indeed, it is known that convergence of the method is guaranteed only if the grid has special properties.

Therefore, we focus here our attention on the low-order Mixed Virtual Element Method, a numerical schemes that operates on polytopal grids and that has shown to be rather robust with respect to irregularities in the data and in the computational grid. We consider first the case of porous medium without fractures, focusing on problems with highly heterogeneous permeability, and then the case of a fractured porous medium, using the model described in Sect. 8.4.2. A different application of the Mixed VEM for the numerical treatment of the Richards equations can be found in Chap. 7.

The actual implementation in PorePy adopts a flux mortar technique that allows non-conforming coupling between inter-dimensional grids. We do not exploit the possibility of having grids non-conforming to the fractures in this work, nevertheless in Sect. 8.4.2 we will describe the mortar approach more in detail.

8.4.1 Bulk Flow Without Fractures

In this part we present the MVEM discretization of problem (8.5). A key point of the virtual method is to use an implicit definition of suitable basis functions, and obtain computable discrete local matrices by manipulating the different terms in the weak formulation appropriately. In this work we consider only the low order case, yet the method can be extended to higher order formulations.

We indicate the computational grid, approximation of Ω , as $\mathcal{T}(\Omega)$. We assume that Ω has polygonal boundary, so that $\mathcal{T}(\Omega)$ covers Ω exactly. The set of faces of $\mathcal{T}(\Omega)$ is denoted as $\mathcal{E}(\Omega)$, with the distinction between the internal and boundary faces indicated by $\mathcal{E}(\overset{\circ}{\Omega})$ and $\mathcal{E}(\partial\Omega)$, respectively. We also specify the edges on a specific portion of the boundary of Ω as $\mathcal{E}(\partial_u\Omega)$ and $\mathcal{E}(\partial_p\Omega)$. We clearly have $\mathcal{E}(\overset{\circ}{\Omega}) \cup \mathcal{E}(\partial\Omega) = \mathcal{E}(\Omega)$ as well as $\mathcal{E}(\overset{\circ}{\Omega}) \cap \mathcal{E}(\partial\Omega) = \emptyset$. In the sequel, we generally indicate as $C \in \mathcal{T}(\Omega)$ a grid cell and $e \in \mathcal{E}(\Omega)$ a face between cells. Element C can be a generic polygon (polyhedra in the 3D case).

We introduce the finite dimensional subspaces, approximation of the continuous spaces given in (8.4), as

$$V_h(\Omega) = \{v_\Omega \in V(\Omega) : \nabla \cdot v_\Omega|_C \in \mathbb{P}_0(C) \text{ and } \nabla \times v_\Omega|_C = \mathbf{0}, \forall C \in \mathcal{T}(\Omega), \\ \text{tr} v_\Omega \cdot \mathbf{n}_e \in \mathbb{P}_0(e), \forall e \in \mathcal{E}(\Omega)\},$$

with $\mathbb{P}_0(X)$ being the space of constant polynomials on X , while tr and \mathbf{n}_e the trace and the normal associated to edge e . For the scalar field we set

$$Q_h(\Omega) = \{q_\Omega \in Q(\Omega) : q_\Omega|_C \in \mathbb{P}_0(C) \forall C \in \mathcal{T}(\Omega)\}.$$

Clearly, $V_h(\Omega) \subset V(\Omega)$ and $Q_h(\Omega) \subset Q(\Omega)$. The degrees of freedom associated with $V_h(\Omega)$ and $Q_h(\Omega)$ are one scalar value for each face and one scalar value for each cell, respectively. More precisely, if we generically indicate with dof_i the functional associated with the i -th degree of freedom, we have, for a $v_\Omega \in V_h(\Omega)$ and a $q_\Omega \in Q_h(\Omega)$

$$\text{dof}_i v_\Omega = \text{tr} v_\Omega \cdot \mathbf{n}_{e_i} \quad \text{and} \quad \text{dof}_i q_\Omega = q_\Omega|_{C_i},$$

where e_i and C_i are the i -th edge and cell, respectively, and tr now indicates the trace associated to the edge e_i .

Moreover, we can observe that in case of triangular grids $V_h(\Omega)$ coincides with $\mathbb{RT}_0(\Omega)$, so the former can be viewed as a generalization of the well known Raviart-Thomas finite elements.

By performing exact integration, the numerical approximation of the bilinear form b_Ω and of the functionals G_Ω , F_Ω are computable with the given definition of the discrete spaces. However, for the term a_Ω we need further manipulations to obtain a computable expression. To this purpose, we define a suitable subspace of $V_h(\Omega)$, defined as

$$\mathcal{V}_h(\Omega) = \{v_\Omega \in V_h(\Omega) : v_\Omega|_C = K_C \nabla v_C \text{ for a } v_C \in \mathbb{P}_1(C) \forall C \in \mathcal{T}(\Omega)\},$$

where K_C is a suitable constant approximation of $K_\Omega|_C$, and we define a projection operator $\Pi_\Omega : V_h(\Omega) \rightarrow \mathcal{V}_h(\Omega)$ so that for a $v \in V_h(\Omega)$ we have

$$a_\Omega(v - \Pi_\Omega v, w) = 0, \quad \forall w \in \mathcal{V}_h(\Omega).$$

We now set $T_\Omega = I - \Pi_\Omega$, where $T_\Omega : V_h(\Omega) \rightarrow \mathcal{V}_h^\perp(\Omega)$ and the orthogonality condition is governed by the bilinear form a_Ω , which, being symmetric, continuous and coercive, defines an inner product. Indeed, from the definition of Π_Ω we have $a_\Omega(T_\Omega v_\Omega, \Pi_\Omega w_\Omega) = 0$ for all $v_\Omega, w_\Omega \in V_h(\Omega)$. Considering this fact, we have the following decomposition

$$a_\Omega(u_\Omega, v_\Omega) = a_\Omega((\Pi_\Omega + T_\Omega)u_\Omega, (\Pi_\Omega + T_\Omega)v_\Omega) = a_\Omega(\Pi_\Omega u_\Omega, \Pi_\Omega v_\Omega) + a_\Omega(T_\Omega u_\Omega, T_\Omega v_\Omega).$$

Now, thanks to the definition of $\mathcal{V}_h(\Omega)$ the first term is computable in terms of the degrees of freedom, see for instance [39], but not the second one. However, since it gives the contribution of a_Ω only on $\mathcal{V}_h^\perp(\Omega)$, it can be approximated with a suitable stabilizing bilinear form $s : V_h(\Omega) \times V_h(\Omega) \rightarrow \mathbb{R}$, i.e.

$$a_\Omega(T_\Omega \mathbf{u}_\Omega, T_\Omega \mathbf{v}_\Omega) \approx s_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega).$$

For more details about s_Ω refer to the works [5, 7, 18, 25, 40, 41]. The form s_Ω must satisfy the following equivalence condition:

$$\exists v_*, v^* \in \mathbb{R}^+ : v_* a_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) \leq s_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) \leq v^* a_\Omega(\mathbf{u}_\Omega, \mathbf{v}_\Omega) \quad \forall \mathbf{u}_\Omega, \mathbf{v}_\Omega \in V_h(\Omega).$$

To illustrate our choice of s_Ω , let us denote with $\boldsymbol{\varphi}$ a generic element of the basis of $V_h(\Omega)$. The stabilization term, in our case, can be computed as

$$s_\Omega(\boldsymbol{\varphi}_\theta, \boldsymbol{\varphi}_\chi) = \sum_{C \in \mathcal{T}(\Omega)} \|H_\Omega\|_{L^\infty(C)} \sum_{i=1}^{N_{dof}(C)} \text{dof}_i(T_\Omega \boldsymbol{\varphi}_\theta) \text{dof}_i(T_\Omega \boldsymbol{\varphi}_\chi), \quad (8.7)$$

where $N_{dof}(C)$ is the total number of degrees of freedom for the vector field for the cell C and dof_i gives the value of the argument at the i th-dof. The K_Ω norm is a scaling factor in order to consider also strong oscillations of physical parameters. With the definition of the stabilization term now all the terms are computable and the global system can be assembled. For more details on the actual computation of the local matrices refer to [6, 40].

8.4.2 Fracture Flow

We introduce now the numerical scheme used for the approximation of problem (8.6). We consider the notations and terms for the porous media from the previous section. In fact, the derivation of the discrete setting for the porous media is similar to what already presented. We focus now on the fracture discretization as well as on the coupling term with the surrounding porous media.

In particular, for the implementation we have chosen PorePy [46], that considers an additional interface γ^\pm between the fracture and the porous media along with a flux mortar technique to couple domains of different dimensions, allowing also non-conforming grids between the domains. However, to avoid additional complexity we consider only conforming grids so that the mortar variable behaves as a Lagrange multiplier λ_h . The latter is the normal flux exchange from the higher to lower dimensional domain. See Fig. 8.2 as an example. Geometrically (i) the interface between the porous media and the fracture, (ii) the fracture, and (iii) the two interfaces coincide but they are represented by different objects with suitable operators for their coupling. In the case of conforming discretizations these operators simply

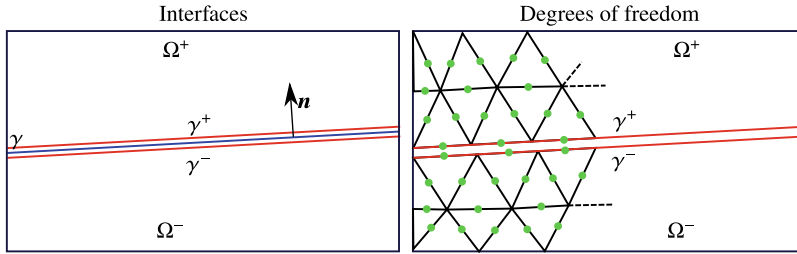


Fig. 8.2 On the left, the hybrid-dimensional representation of a fracture immersed in a porous media with the two interfaces γ^+ and γ^- , in red. On the right, the representation of the degrees of freedom for vector fields

map the corresponding degrees of freedom, however in the case of non-conforming discretizations projection operators should be considered.

As done before, we consider the special case of a single fracture, being its generalization straightforward. First of all, the velocity degrees of freedom for the rock matrix in the proximity of the fracture are doubled as Fig. 8.2 shows. We can thus represent $\text{tr} \mathbf{u}_\pm \cdot \mathbf{n} = \lambda_h^\pm$ for both sides \pm of the fracture itself. The term \hat{a}_Ω involves the actual integration of the basis functions for each grid cells, which is not possible since they are not, in general, analytically known.

Many of the following steps are similar to what already done for the bulk porous media. We introduce a tessellation of γ into non-overlapping cells (segments in this case), the grid is indicated with $\mathcal{T}(\gamma)$ and the set of faces (edges) as $\mathcal{E}(\gamma)$. Also in this case, we divide the internal faces and the external faces $\mathcal{E}(\overset{\circ}{\gamma})$ and $\mathcal{E}(\partial\gamma)$. Moreover, the latter can also be divided into subset depending on the boundary conditions $\mathcal{E}(\partial_u\gamma)$ and $\mathcal{E}(\partial_p\gamma)$. Clearly, we have $\mathcal{E}(\gamma) = \mathcal{E}(\overset{\circ}{\gamma}) \cup \mathcal{E}(\partial\gamma)$ as well as $\mathcal{E}(\partial_u\gamma) \cup \mathcal{E}(\partial_p\gamma) = \mathcal{E}(\partial\gamma)$. We introduce the functional spaces for the variables defined on the fracture, for the vector fields we have

$$V_h(\gamma) = \{ \mathbf{v}_\gamma \in V(\gamma) : \nabla \cdot \mathbf{v}_\gamma|_C \in \mathbb{R}, \nabla \times \mathbf{v}_\gamma|_C = \mathbf{0} \forall C \in \mathcal{T}(\gamma), \text{tr} \mathbf{v}_\gamma \cdot \mathbf{n}_e \in \mathbb{R} \forall e \in \mathcal{E}(\gamma) \},$$

while for the scalar fields we consider the discrete space

$$Q_h(\gamma) = \{ q_\gamma \in Q(\gamma) : q_\gamma|_C \in \mathbb{R} \forall C \in \mathcal{T}(\gamma) \}.$$

By keeping the same approach as before, we assume exact integration so that the numerical approximation of the bilinear form b_γ as well as functionals G_γ and F_γ are computable with the given definition of the discrete spaces. The term a_γ is not directly computable, we thus introduce the subspace of $V_h(\gamma)$ as

$$\mathcal{V}_h(\gamma) = \{ \mathbf{v}_\gamma \in V_h(\gamma) : \mathbf{v}_\gamma|_C = K_\gamma|_C \nabla v_C \text{ for a } v_C \in \mathbb{P}_1(C) \forall C \in \mathcal{T}(\gamma) \}.$$

We introduce the projection operator Π_γ from $V_h(\gamma) \rightarrow \mathcal{V}_h(\gamma)$ such that for a $\mathbf{v} \in V_h(\gamma)$ we have $a_\gamma(\mathbf{v} - \Pi_\gamma \mathbf{v}, \mathbf{w}) = 0$ for all $\mathbf{w} \in \mathcal{V}_h$. By introducing the operator $T_\gamma = I - \Pi_\gamma$, we have the decomposition

$$a_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) = a_\gamma((\Pi_\gamma + T_\gamma)\mathbf{u}_\gamma, (\Pi_\gamma + T_\gamma)\mathbf{v}_\gamma) = a_\gamma(\Pi_\gamma \mathbf{u}_\gamma, \Pi_\gamma \mathbf{v}_\gamma) + a_\gamma(T_\gamma \mathbf{u}_\gamma, T_\gamma \mathbf{v}_\gamma).$$

By the definition of $\mathcal{V}_h(\gamma)$ the first term is now computable, while the second term, which is not computable, is replaced by the stabilization term

$$a_\gamma(T_\gamma \mathbf{u}_\gamma, T_\gamma \mathbf{v}_\gamma) \approx s_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma)$$

with the request that s_γ scales as a_γ , meaning that

$$\exists u_*, v^* \in \mathbb{R} : \quad u_* s_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) \leq a_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) \leq v^* s_\gamma(\mathbf{u}_\gamma, \mathbf{v}_\gamma) \quad \forall \mathbf{u}_\gamma, \mathbf{v}_\gamma \in V_h(\gamma).$$

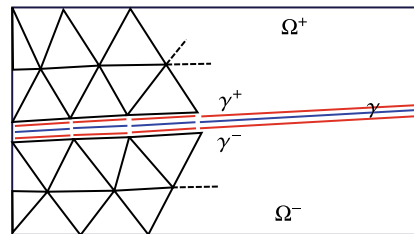
Denoting an element of the basis of $V_h(\gamma)$ as ϕ , the actual construction of s_γ is given by the formula

$$s_\gamma(\phi_\theta, \phi_\chi) = \sum_{C \in \mathcal{T}(\gamma)} h \|K_\gamma^{-1}\|_{L^\infty(C)} \sum_{i=1}^{N_{dof}(C)} \text{dof}_i(T_\gamma \phi_\theta) \text{dof}_i(T_\gamma \phi_\chi),$$

with h the diameter of the current cell C . With the previous choices all the terms are computable and the fracture problem can be assembled. For more details see [5, 7, 18, 40, 41].

To couple the bulk and fracture flow, a Lagrange multiplier λ_h^\pm is used to represent the flux exchange between the fracture and the surrounding porous media. We assume conforming grids, meaning that the fracture grid is conforming with the interface grid as well as the faces of the porous media are conforming with the interface grid. See Fig. 8.3 as an example. For space compatibility, we assume the Lagrange multiplier be a piece-wise constant polynomial. The interface condition (8.2c) is directly computable with the degrees of freedom introduced providing a suitable projection of the pressure p_Ω at the fracture interface. Our choice is to consider the same value of the pressure at neighbouring cells, however other approaches can be used, see for example [57].

Fig. 8.3 Representation of the conforming computational grids for the porous media, the fracture, and the two interfaces



8.5 Grid Generation

The generation of grids for realistic fractured porous media geometries is a challenging task, whose complete automatic solution is still an open problem, particularly for 3D configurations. We here give a brief overview of some techniques that have been proposed, with no pretence of being exhaustive.

8.5.1 *Constrained Delaunay*

The generation of a grid of simplexes (triangles in 2D, tetrahedra in 3D) conformal to a fracture network may be performed in principle by employing a constrained Delaunay algorithm. It is an extension of the well known Delaunay algorithm to the case where the mesh has to honour internal constraints (or describe a non-convex domain). Usually it starts from a representation of the domain and in 3D it first generates constrained Delaunay triangulation on the fracture and boundary geometry, then new nodes are added in the domain to generate a final grid that satisfies a relaxed Delaunay criterion to honour the internal interfaces. The description of the constrained Delaunay procedure may be found, for instance, in [21]. Another general reference for mesh generation procedures is [34]. However, in practical situations several issues may arise. The presence of fractures intersecting with small angles, for instance, may produce an excessive refinement near the intersections in order to maintain the Delaunay property. In 3D there is the additional issue of the possible generation of extremely badly distorted elements, often called slivers, whose automatic removal is problematic, when not impossible, under the constraint of conformity with complex internal surfaces.

Several techniques have been proposed to ameliorate the procedure. For instance in [49, 50] the authors present a procedure that modifies the fracture network trying to maintain its characteristics of connectivity and effective permeability, while eliminating geometrical situations where that may impair the effectiveness of a Delaunay triangulation. In the second reference, a special decision strategy (called “Gabriel criterion”) is used to select a part of the fracture network to which triangulation can be constrained without leading to an excessive degradation in cells quality, or excessively fine grids. The procedure has proved rather effective on moderately complex network in 2D, while the results for 3D configurations seem less convincing.

We mention for completeness that an alternative procedure for generating simplicial grids is the one based on the front advancing technique (maybe coupled with the Delaunay procedure). It is implemented in several software tools, see for instance [35, 59]. However, its use in the context of fractures media is at the moment very limited, probably because of the lack of results of the termination of the procedure, contrary to the Delaunay algorithm where one can prove that, under mild conditions, the generation terminates in a finite number of steps. Moreover it has a much higher computational cost. The interested reader may consult the cited references.

In our case PorePy considers the software Gmsh [44] for the generation of the Delaunay bidimensional grids. The grid size in the configuration file is specifically tuned to obtain high quality triangles. Indeed, we consider distances between fractures, between a fracture and the domain boundary, and length of fracture branches. With these precautions, we usually obtain quality grids that are suitable for numerical studies.

8.5.2 Grids Cut by the Fracture Network

An alternative possibility to create a grid conforming to fractures or, in general, planar interfaces, consists in cutting a regular Cartesian or simplex mesh, as shown in Fig. 8.4 for the case of a Cartesian mesh. The resulting grid will be formed by polytopal elements in the vicinity of the fractures. The main issue in this procedure is the possible generation of badly shaped or very small elements as a consequence of the cut. Another technical problem is the necessity of having efficient techniques for computing intersections and constructing the polytope. To this respect, one may adopt the tools available in specialized libraries like CGAL [19], or developed by the RING Consortium [54]. Clearly, the adoption of this technique calls for numerical schemes able to operate on general polytopal elements. This method, when applied to Cartesian grids, has the advantage of maintaining a structured grid away from the fracture network, where the sparsity of the linear system may ease its numerical solution, but it does not allow local refinements (unless by using hanging nodes, which increase computational complexity). In general it is a valid alternative to a direct triangulation provided the numerical scheme be robust with respect to the presence of small or high aspect ratio elements.

We outline a possible algorithm for the case of a Cartesian background grid, adopted in this work. We start by creating a Cartesian mesh of rectangular elements and compute the intersections among the edges of the grid and the segments describing the fractures. This step is rather straightforward for Cartesian grids. The intersection points can be easily sorted according to a parametric coordinate to create the mesh of each fracture. Then, each cell cut by one or more fractures is split into two, three or four polygonal sub cells as follows: (i) for each point, the signed dis-

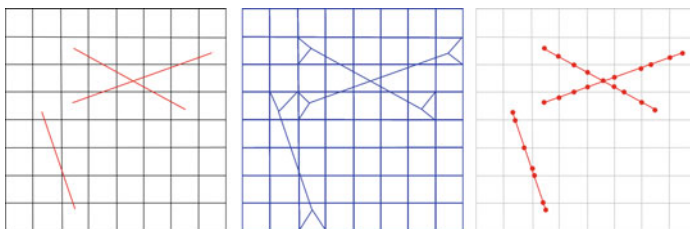


Fig. 8.4 Creation of a polygonal mesh from a regular Cartesian grid

tance from the fracture is computed, and *(ii)* points on the same side of the fracture are grouped, and sorted in counter-clockwise direction.

To avoid non-convex cells the cells containing the fracture tips are split in three by connecting the tip with the nodes of the edge that is crossed by the prolongation of the fracture. However, in principle it would be possible to consider a single cell with two coincident faces.

8.5.3 Agglomeration

Polytopal grids can be generated by agglomerating simplicial elements produced, for instance, by a constrained Delaunay procedure. For example, in [16], tetrahedra are agglomerated (and nodes moved) to try to produce hexahedral elements in large part of the domain, with a twofold objective: on the one hand the reduction of the total number of degrees of freedom and consequent reduction of computational complexity, on the other hand, the generation of a grid more suitable for finite volume schemes based on two-point flux approximation (TPFA).

In a more general setting, agglomeration may join together elements whose value of physical parameters are similar, with the final objective of reducing computational cost, as well as eliminating excessively small elements. The numerical method, however, should be able to operate properly on the possible irregularly shaped and non-convex elements generated by the procedure. The technique is clearly a post-processing one, since it requires to have a mesh to start with. Its basic implementation is however rather simple and is similar to that used in some multigrid solvers, like in [45].

In our case, PorePy has the capability to agglomerate cells based on two different criteria: *(i)* by volume, meaning that cells with small volumes are grouped with neighbouring cells. This procedure continues until the new created cells have volumes that are comparable with an average volume in the grid. This procedure can be effective in presence of uniform physical data in different part of the computational domain and in particular in presence of fracture networks. In the case of highly variable data, e.g. permeability, the previous procedure may not be effective since cells with very different properties may be merged together. For this reason PorePy implements another strategy, *(ii)* based on the agglomeration in the algebraic multigrid method. It adopts a measure of the strength of connections between DOFs to select the cells to be joined, based on a two-point flux approximation discretization, for more details see [40, 61]. Examples of these strategies are given in [38, 40–42, 51].

Remark 1 The agglomeration procedure is even more effective when a time dependent problem is solved, like linear and non-linear transport of a tracer or the heat equation. Other strategies might be more appropriate to optimize the grid for a specific physical process.

8.5.4 Voronoi

Voronoi grids are of particular interest for methods such as Finite Volumes with TPFA, since they guarantee that the line connecting the centroids of neighbouring cells is always orthogonal to the shared face. Under this assumption the two point approximation of the flux is consistent if the permeability tensor is diagonal. However, producing Voronoi diagrams that honour the internal interfaces represented by the fracture is not an easy task, particularly for complex 3D configurations. An attempt in that direction has been performed in [11, 55].

In this work, limited to 2D cases, we generate Voronoi diagrams that honour the geometry of the fractures and the boundaries of the domain by first creating a Cartesian grid (see Sect. 8.5.2) and positioning a seed at the centre of cells not cut by the fractures. Then, we start from the discretization of the fractures induced by the intersection with the background grid, and for each fracture cell we position two seeds on opposite sides of the fracture at a small distance δ as shown in Fig. 8.5. This will create a Voronoi cell with a face exactly on the fracture. The same technique is used to obtain boundary faces in the desired position. Close to each fracture tip x_T we position four seeds in $x_T \pm \delta_1 \mathbf{n} \pm \delta_2 \mathbf{t}$ where \mathbf{n} and \mathbf{t} are the normal and the tangent unit vectors to the fracture and $\delta_{1,2}$ are user defined distances. This ensures that the fracture is honoured up to the tip and has the correct length. Similar strategies are applied at fracture intersections. The position of the seeds and faces close to the intersections is also shown in Fig. 8.5. Note that with this strategy the Voronoi cells far from fractures are rather regular, since they reflect the underlying Cartesian grid.

An advantage of Voronoi grids is that faces are planar and cells are convex by construction. However, an important drawback is that the number of faces per cell can be quite large. Moreover, as pointed out before, the construction of a constrained grid in general realistic configurations is an open problem.

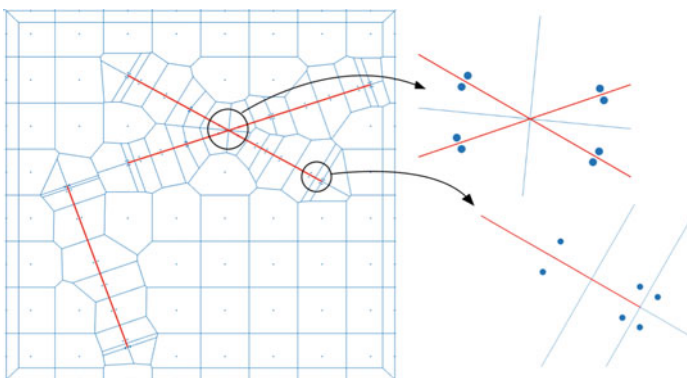


Fig. 8.5 On the left, graphical representation of Voronoi grid with fractures. On the right, details on the construction for fracture intersection and fracture tip

8.6 Numerical Results

In this section, we present two test cases to show the performances and the potentiality of the previously introduced algorithms. In particular, in the first test case we have a setting where the permeability experiences a high variation between neighbouring cells. In the second test case a network of fractures is considered with different types of intersections: in this case the challenge is more related to the geometrical complexity to create the computational grid. In both test cases, agglomerating procedures are used to reduce the computational cost of the simulations.

8.6.1 Heterogeneous Porous Medium: Layers from SPE10

The aim of this test case is to validate the effectiveness of the MVEM scheme in presence of highly heterogeneous permeability. We consider two distinct layers of the SPE10 [22] benchmark problem, in particular layer 4 and 35 (by starting the numeration from 1), from now on denoted as L4 and L35, respectively. The main difference between them is that the latter has distinctive channels of high permeability which are not present in layer 4. The permeability is assumed to be scalar in each cell, and each layer is composed by a computational grid of 60×220 . Figure 8.6 on the left shows the permeability fields for both layers. Note that in both cases permeability spans about six order of magnitude.

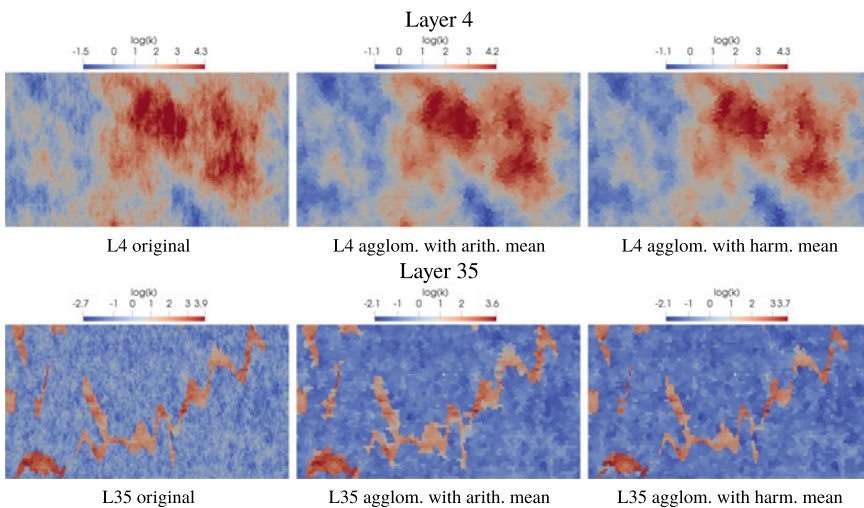


Fig. 8.6 Permeability field for the test case of Sect. 8.6.1 for layer 4 on the top and 35 on the bottom. On the left the reference values, on the centre and right the values obtained after the clustering with arithmetic and harmonic mean, respectively. The values are given in \log_{10}

Table 8.1 Average, minimum and maximum value of cell area and number of faces per cell for the six grids employed for test case Sect. 8.6.1

	Aspect ratio			Cell area			n_{faces}		
	Average	Min	Max	Average	Min	Max	Average	Min	Max
L4	2.37	1.50	4.37	108	37.2	242	12.2	6	20
L35	2.37	1.13	5.83	111	37.2	297	12.2	6	22

To lighten the computational effort, we apply an agglomerating procedure to group cells and obtain a smaller problem. Starting from square cells the algorithm creates cells by considering the procedure in Sect. 8.5.3 and, for each agglomerated cell, the associated permeability will be computed in two different ways: as the arithmetic and harmonic average. The former is more suited for flow parallel to layers of different permeability, while for orthogonal flow the harmonic average gives more realistic results. For a more detailed discussion see [52]. We consider both approaches, see Fig. 8.6 on centre and right, which represents the agglomerated permeability of both layers by considering the arithmetic and harmonic means. For layer 4 the figures look similar, while for layer 35 the channels for the agglomerated problem with harmonic mean are narrower than the original ones and than those obtained in the agglomerated grid with arithmetic mean.

In Table 8.1, we summarize the geometric properties of the grids obtained by means of cells clustering for the two layers. We can observe that the area of the cells and the average number of faces per cell is similar in the two cases, however, in layer 35 we have slightly more elongated elements on average, reflecting the channelized permeability field. The aspect ratio is estimated using the area of the cells, the maximum distance between points and is rescaled so that square cells (or equilateral triangles, see Sect. 8.6.2) have aspect ratio 1.

We impose a pressure gradient from left to right with synthetic values 1 and 0, respectively. The other boundaries are sealed with homogeneous Neumann conditions.

To compare the accuracy of the proposed clustering techniques, we compute the errors in the pressure with respect to the problem on the original grid solved with a two-point flux approximation scheme [1, 37], which, in this case since the grid is K -orthogonal, is consistent and converges quadratically to the exact solution, thus can be considered as a valid reference. We name this solution “reference” and we indicate the pressure as p_{ref} . The error is computed as

$$err = \frac{\|\Pi_{\text{ref}} p - p_{\text{ref}}\|_{L^2(\Omega)}}{\|p_{\text{ref}}\|_{L^2(\Omega)}}$$

where Π_{ref} is the piecewise constant projection operator that maps from the current grid to the reference one. Due to the clustering procedure its construction is rather straightforward, since the cells of the original mesh are nested in the agglomerated one. We can notice that the errors obtained for the layer 4 with both averaging pro-

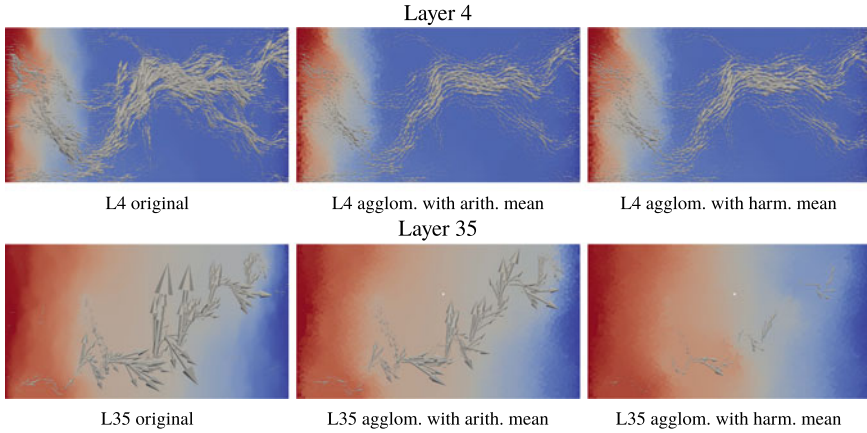


Fig. 8.7 Pressure and Darcy velocity fields for the test case of Sect. 8.6.1 for layer 4 on the top and 35 on the bottom. On the left the reference solution, on the centre and right the values obtained after the clustering with arithmetic and harmonic mean, respectively. The arrows are scaled by the same value in each layer and the pressure ranges from 0 to 1, blue to red respectively

cedure are comparable and around 4%, which can be acceptable in most of real applications. In the case of layer 35 the situation is more involved, in fact the arithmetic mean gives an error of approximately 3.5% while the harmonic mean of 13%. We can explain this big discrepancy by noticing that, when a channel of high permeability is composed by few cells in its normal direction, during the agglomeration procedure it is possible that some of these cells are grouped with the surrounding lower permeability cells. The harmonic mean will bring the permeability value of the agglomerated cell closer to the lower value than the higher, dramatically changing the connectivity properties of the obtained permeability field. This can be noticed in the permeability field reported in Fig. 8.6, suggesting that harmonic averaging can be unsuited for parallel flow in strongly channelled domains.

Figure 8.7 shows the pressure fields for both layers and for the two approaches. On top of the pressure fields the Darcy velocity is also represented with grey arrows. We notice that for layer 4 pressures and velocities look very similar, while for layer 35 the pressure field and velocity of the agglomerated problem with harmonic mean look quite different compared with the reference solutions as well as that obtained with the agglomeration strategy that uses the arithmetic mean.

To improve the effectiveness of this approach, a local numerical upscaling technique could be considered to compute a more representative value of the permeability for grouped cells. However, in this case we might expect a higher computational cost. See [29] for a more detailed presentation of upscaling techniques.

To conclude this test case, let us now analyse the properties of the system matrix to verify what is the impact of element size and shape in the different cases. Note that the problem is in mixed form and our analysis considers the entire saddle point matrix. Since the number of unknowns is not exactly the same after grid agglomeration we

Table 8.2 Matrix properties for test case Sect. 8.6.1

	N_{DOF}	N_{cells}	N_{faces}	\bar{n}	$K(A)$
L4 (mean K)	16345	2269	14076	22.17	8.29e+06
L4 (harmonic K)	16345	2269	14076	22.17	8.44e+06
L35 (mean K)	16010	2210	13800	22.53	8.29e+06
L35 (harmonic K)	16010	2210	13800	22.53	8.39e+06

describe matrix sparsity by means of the average number of non-zero entries per row \bar{n} , computed as

$$\bar{n} = \frac{n_z}{N_{\text{DOF}}},$$

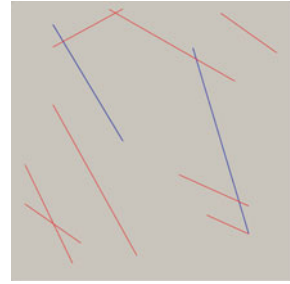
where n_z is the number of non-zero entries and N_{DOF} is the number of unknowns. Moreover we will compare the values of condition number $K(A)$ estimated by the method `condEst` provided by `Matlab`[®]. In Table 8.2 we consider the two layers, L4 and L35, and by “mean K”, “harmonic K” we identify the averaging of permeability in the agglomerated cells, the arithmetic and harmonic mean respectively. This choice has no impact on the matrix size or sparsity but may result in different condition numbers. We can observe that the four matrices are very similar in terms of size, sparsity and condition number, and that the large number of faces per element reflects in the average number of entries per row. It can be also observed that mesh agglomeration is slightly more effective in layer L35 due to its channelized permeability distribution.

8.6.2 Fracture Network

This test case considers the Benchmark 3 of the study [30] presented in Sect. 4.3. Our objective is to study the impact of the grid on the solution quality provided by the MVEM. The domain contains a fracture network made of 10 fractures and 6 intersections, one of which is of L -shape. For the detailed fracture geometry, we refer to the aforementioned work. See Fig. 8.8 for a representation of the problem geometry.

We consider three types of grids: Delaunay, Cartesian cut, and Voronoi. Since the fracture network may create small cells, on top of these three grids an agglomeration algorithm is used to agglomerate cells of small volume. These cells are merged with neighbouring cells, trying to obtain a more uniform cell size in the grid. The Delaunay grid is created by the software `Gmsh` [44], tuned to provide high quality elements in proximity of small fracture branches or almost intersecting fractures. The six different grids we are considering are reported in Fig. 8.9 along with the number of cells associated to the rock matrix and fractures.

Fig. 8.8 Geometry of the domain for the benchmark used in Sect. 8.6.2



We see that some of the agglomerated elements have internal cuts, in particular for Delaunay agglomerated grid in Fig. 8.9, and for all the clustered grids we have cells that are not shape regular and in some cases not even star-shaped. For classical finite elements or finite volumes we might expect low quality results.

Another result of the agglomeration is a reduction of the number of very small or very stretched cells. In Fig. 8.10 we can observe histograms of an estimate of the cells aspect ratio for the different grids. We can see that for the Cartesian cut grid and the Voronoi grid the maximum aspect ratio decreases remarkably with the agglomeration, while in the case of a Delaunay grid we have the opposite effect. As

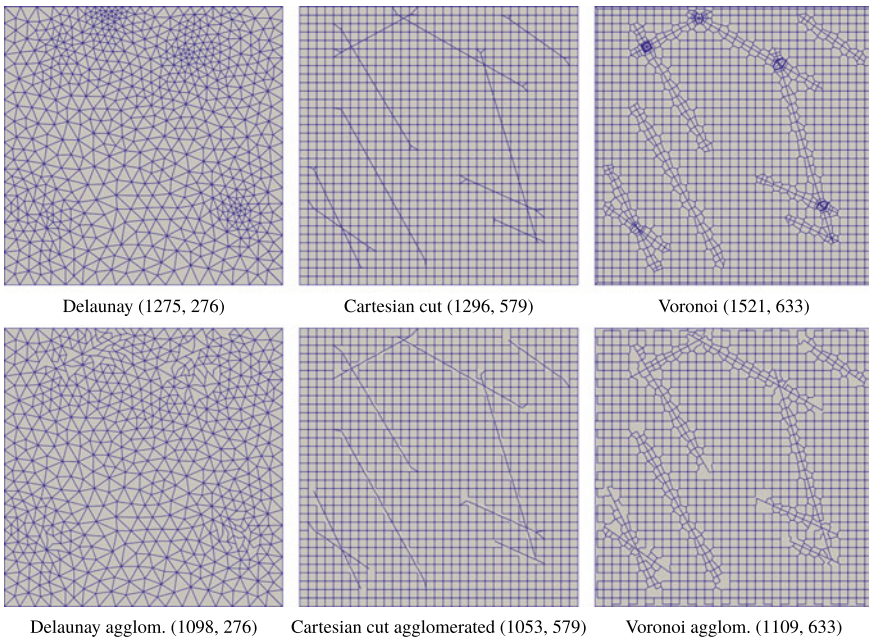


Fig. 8.9 Benchmark 3 of Sect. 8.6.2: Fracture network on top left, on the others the grids for different approaches. In the brackets the number of cells (bulk, fracture)

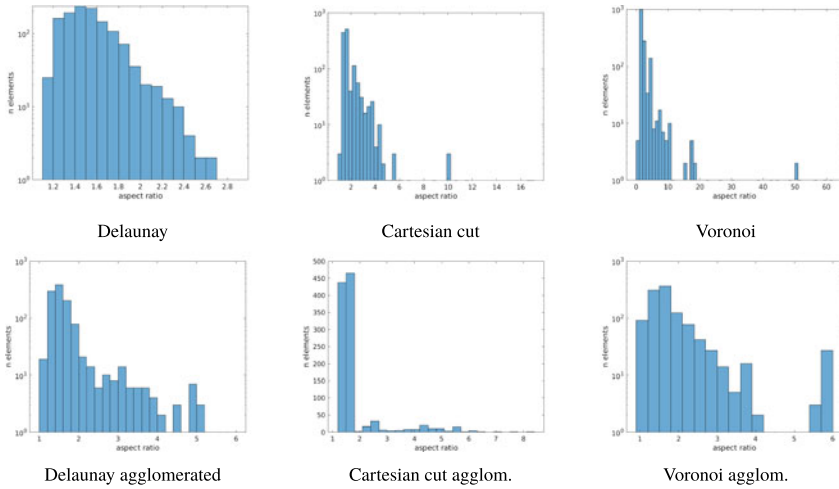


Fig. 8.10 Histograms of the cells aspect ratio for the different types of grid in test case Sect. 8.6.2

Table 8.3 Average, minimum and maximum value of cell area and number of faces per cell for the six grids employed for test case Sect. 8.6.2

	Cell area			n_{faces}		
	Average	Min	Max	Average	Min	Max
Delaunay	7.8431e-04	8.4186e-05	2.1020e-03	3	3	3
Delaunay agglom.	9.1075e-04	3.9631e-04	2.1767e-03	3.1557	3	8
Cut	7.7160e-04	8.4664e-08	9.1833e-04	3.9769	3	6
Cut agglom.	9.4967e-04	3.9945e-04	2.2589e-03	4.4311	3	10
Voronoi	6.5746e-04	4.6260e-07	1.2686e-03	4.4694	3	14
Voronoi agglom.	9.0171e-04	3.3000e-04	3.4502e-03	5.1109	4	16

we will show later high anisotropy can result in a less effective stabilization for the MVEM matrix. Moreover, in Table 8.3 we show that cells agglomeration leads to an increase of the mean and minimum cell areas, but also to an increase of the number of faces per cell.

Referring to the colour code given in Fig. 8.8, we set the aperture $\epsilon = 10^{-4}$ for all the fractures and the permeability is set to $k_{\gamma} = \kappa_{\gamma} = 10^4$ for all the fractures depicted in red and $k_{\gamma} = \kappa_{\gamma} = 10^{-4}$ for the ones in blue. The former behave as high flow channels while the latter as low permeable barriers. The rock matrix is characterized by a unit scalar permeability. In [30] two sets of boundary conditions were considered, left-to-right and bottom-to-top. In our case we choose the former, meaning that we set a value of pressure equal to 4 on the left side of Ω and to 1 on the right side of Ω . The other two boundaries are considered as impervious.

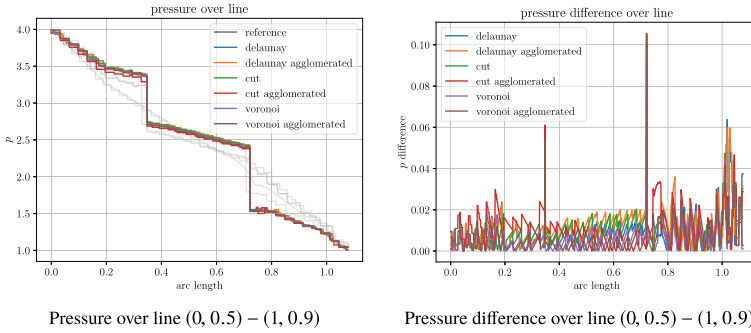


Fig. 8.11 On the left, pressure over line for the test case of Sect. 8.6.2. The grey solutions are the one reported in [30]. Most of the plots overlap with the reference solution, in black. On the right, the difference over the same line between a solution and the reference one

In Fig. 8.11 (left) we report the plot of pressure over the line (0, 0.5) – (1, 0.9), by using the grids shown in Fig. 8.9. In light grey we present the results obtained in the benchmark [30] and in black the reference solution. The latter has been calculated with mimetic finite difference, on a very refined grid that represents fractures as the same dimension of the porous media. We clearly see that all the proposed methods overlap with the reference solution showing high accuracy even on such coarse grids. In particular, results do not deteriorate with the agglomerating procedure. Moreover, comparing with the results obtained in [30] the ones given by the MVEM are, generally, of higher quality.

In Fig. 8.11 (right), we show the pressure difference between the reference solution and the ones obtained with the considered grids, over the reference solution itself. The errors are quite small except for the two peaks in correspondence of the pressure jump in the picture at the left of the same Figure. The reason can be associated to the sampling procedure used in the extraction of these data.

Finally, as done in [30] we compute the errors in the rock matrix between the reference and the computed solution. We consider the following formula

$$err_m^2 = \frac{1}{|\Omega|(\Delta p_{ref})^2} \sum_{f=K_m \cap K_{ref,m}} |f| (p_m|_{K_m} - p_{ref}|_{K_{ref,m}})^2, \quad (8.8)$$

where $p_m|_{K_m}$ is the pressure of the m -method at cell K_m , p_{ref} is the reference pressure at cell $K_{ref,m}$, and Δp_{ref} is the maximum variation of the pressure on all the domain. These errors are reported in Table 8.4. All the errors are quite small and comparable with those reported in [30]. When the agglomeration procedure is adopted, the errors slightly increase due to the smaller number of cells except for the Cartesian cut case where the error doubles, remaining nevertheless acceptable.

Let us now analyse the properties of the system matrix to verify what is the impact of element size and shape in the different cases. We remind that the grids have been generated with comparable resolution to obtain similar numbers of degrees

Table 8.4 Pressure error between the reference solution and the compute with the MVEM by using formula 8.8

	Original	Agglomerated
Delaunay	0.013008	0.014267
Cartesian cut	0.012865	0.025827
Voronoi	0.0085291	0.010037

Table 8.5 Matrix properties for test case Sect. 8.6.2

	N_{DOF}	N_{cells}	N_{faces}	\bar{n}	$K(A)$
Delaunay	3741	1373	2162	5.15	4.82e+10
Delaunay agglom.	3384	1196	1982	5.51	3.85e+10
Cut	4961	1495	1296	6.00	4.23e+10
Cut agglom.	4474	1252	2814	7.42	3.67e+10
Voronoi	6095	1738	3913	7.33	4.10e+10
Voronoi agglom.	5118	1326	3348	9.32	3.21e+10

of freedom, however, the number of unknowns is not exactly the same. Results are summarized in Table 8.5. From the point of view of the degrees of freedom the Voronoi grid is the most demanding because, for a given space resolution it generates very small cells close to the intersections and tips, however, it is also the one that benefits the most from agglomeration. The conditioning is of the same order of magnitude for all grids, and improves with agglomeration. In particular the best result is obtained for the agglomerated Voronoi grid despite the large number of faces per element that results from clustering of general polygons and reflects in the slightly larger number of non-zero entries per row.

We can also observe that, even if the sparsity of the matrices is similar in all cases, the pattern can change significantly. In Fig. 8.12 we compare the matrix structure corresponding to a Delaunay grid and a Cartesian cut one: the underlying structure of the Cartesian grid has a visible impact on the sparsity pattern. A similar structure is observed for the case of the Voronoi grid since, away from the fracture network, the seeds are positioned to obtain a Cartesian grid. Let T_α denote the time required to solve 1000 times the system arising from the discretization on a mesh α with the “\” method from Matlab®, and let $\tilde{T}_\alpha = \frac{T_\alpha}{(N_{DOF}^\alpha)^3}$ be the time normalized against the third power of the system size. The corresponding values, reported in Table 8.6, seem to indicate that, for the same sparsity, a faster solution is obtained with a more compact pattern. Solution strategies for this kind of problem can be found in [24].

We can also compare the performances of an iterative solver on the same matrices. Given the small size of the problem and the fact that the preconditioner we adopt is not *ad hoc* for the problem it is not fair to compare the computational time of a direct

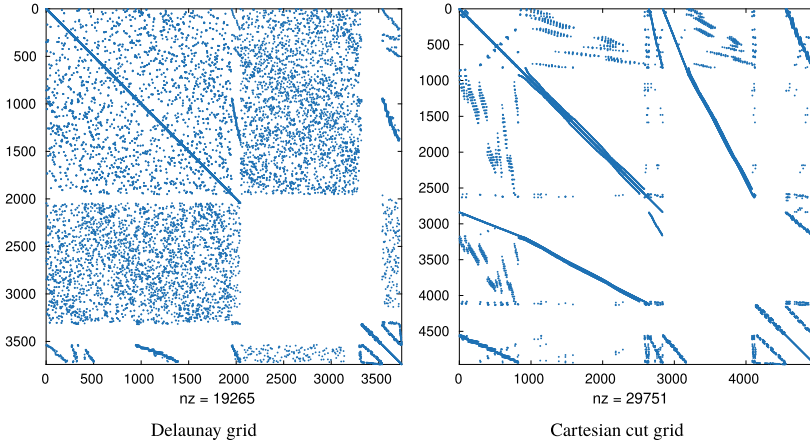


Fig. 8.12 The sparsity patterns for a Delaunay grid (left) and the Cartesian cut grid (right)

Table 8.6 Normalized time for the solution of the linear systems corresponding to the different grids

	Delaunay	Delaunay agglom.	Cut	Cut agglom.	Voronoi	Voronoi agglom.
\tilde{T}_α	2.630e-10	3.410e-10	1.889-e10	2.262e-10	1.394e-10	2.246e-10

and iterative solver, but we can highlight the differences in number of iterations for the different grids. Since the system matrix can be rearranged as

$$A = \begin{bmatrix} M & \hat{B}^T \\ -B & C \end{bmatrix}$$

we employed the following block preconditioner

$$P = \begin{bmatrix} M & 0 \\ 0 & -\tilde{S} \end{bmatrix}$$

where \tilde{S} is approximated using the lumped version of M , called \tilde{M} , i.e. $\tilde{S} = -C - B\tilde{M}^{-1}\hat{B}^T$, and applied GMRES with a tolerance on the normalized residual of 10^{-6} . Results are summarized in Table 8.7. The number of iterations reflects the differences in condition number; note that the chosen preconditioner reduces conditioning of approximately 4 orders of magnitude in all cases except for the case of the agglomerated cut (and, to a lesser extent, Voronoi) grid where it is slightly less effective.

Finally, we study the effect of element shape on the MVEM stabilization term. We define element-wise an index

Table 8.7 Number of GMRES iterations for the solution of the linear systems corresponding to the different grids

	Delaunay	Delaunay agglom.	Cut	Cut agglom.	Voronoi	Voronoi agglom.
N_{it}	27	30	24	75	34	47
$K(P^{-1}A)$	8.72e+06	8.64e+06	7.92e+06	3.12e+07	8.95e+06	8.97e+06

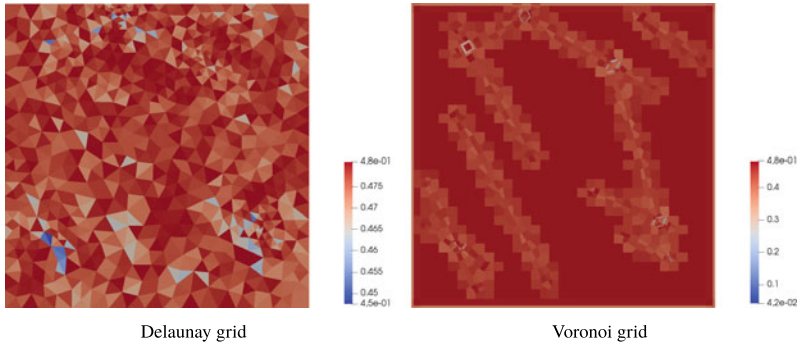


Fig. 8.13 On the left, κ_i on the Delaunay grid, on the right the same index on the Voronoi grid before clustering

$$\kappa_i = \frac{\|S_i\|}{\|S_i\| + \|A_i\|}$$

where S_i and A_i are the local stability and consistency contributions to the matrix arising from the discretization of the bilinear form a_Ω on the i -th element.

As shown in Fig. 8.13 in the case of a Delaunay grid the norm of the stabilization term in each local matrix is comparable to the norm of the consistency term, i.e. $\kappa_i \simeq 0.5$ everywhere. In the Voronoi grid instead we have elements with extremely high aspect ratios (up to 60), or, in other words, we have small edges compared to the typical mesh size. In this latter case the norm of the stability term is one order of magnitude smaller in elements with very small edges. A discussion of the stability bounds for grids in the case of small edges can be found in [7, 8] for the primal formulation of elliptic problems.

8.7 Conclusion

In this work we have presented and discussed the performances of the Mixed Virtual Finite Element Method applied to underground problems. One of its main advantages is the possibility to handle, in a natural way, grid cells of any shape becoming suitable for its usage in problems with complex geometries, such as subsurface flows. A

second strong point is the ability of the scheme to handle, in a robust way, strong variations of the permeability matrix which is again a common aspect for underground processes. Finally, the numerical scheme is also locally mass conservative making it very suitable in the coupling of other physical processes, like transport problems. We have tested the capabilities of the scheme with respect to two test cases that are known in literature and stress the two aforementioned critical points: heterogeneity and geometrical complexity. A first remark is that the mixed virtual element method gives high quality results also for challenging grids and physical data, making it a promising and interesting scheme for industrial applications. Moreover we performed some comparisons of the system matrices arising from the discretization of the problem on different types of grids: Delaunay, Voronoi, Cartesian grids cut by fractures. We observed similar condition numbers and sparsity, but a better sparsity patterns for grids obtained from the modification of structured ones. We also applied agglomeration by means of permeability based and volume based clustering: besides reducing the computational cost this technique allowed us to eliminate small cells and, in some cases, cells with very large aspect ratios where the MVEM stabilization term employed in this work does not scale correctly. Future research may focus on the choice of the most effective stabilization term formulation for the grid type, as well as to the generalization of this work to the three dimensional case, including the discussion of corner point grids, which are widely used in subsurface flows but pose many challenges due to the presence of non-planar faces and non-convex elements.

Acknowledgements We acknowledge the PorePy development team: Eirik Keilegavlen, Runar Berge, Michele Starnoni, Ivar Stefansson, Jhabriel Varela, Inga Berre.

References

1. I. Aavatsmark, Interpretation of a two-point flux stencil for skew parallelogram grids. *Comput. Geosci.* **11**(3), 199–206 (2007)
2. C. Alboin, J. Jaffré, J.E. Roberts, C. Serres, Modeling fractures as interfaces for flow and transport in porous media, in *Fluid Flow and Transport in Porous Media: Mathematical and Numerical Treatment* (South Hadley, MA, 2001), vol. 295 of Contemporary Mathematics, pp. 13–24. American Mathematical Society (Providence, RI, 2002)
3. P.F. Antonietti, L. Formaggia, A. Scotti, M. Verani, N. Verzotti, Mimetic finite difference approximation of flows in fractured porous media. *ESAIM: M2AN* **50**(3), 809–832 (2016)
4. J. Bear, *Dynamics of Fluids in Porous Media* (American Elsevier, 1972)
5. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, H(div) and H(curl)-conforming VEM. *Numerische Mathematik* **133**(2), 303–332 (2014)
6. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, The hitchhiker’s guide to the virtual element method. *Math. Model. Methods Appl. Sci.* **24**(08), 1541–1573 (2014)
7. L. Beirão da Veiga, F. Brezzi, L.D. Marini, A. Russo, Mixed virtual element methods for general second order elliptic problems on polygonal meshes. *ESAIM: M2AN* **50**(3), 727–747 (2016)
8. L. Beirão da Veiga, C. Lovadina, A. Russo, Stability analysis for the virtual element method. *Math. Model. Methods Appl. Sci.* **27**(13), 2557–2594 (2017)
9. M.F. Benedetto, S. Berrone, A. Borio, S. Pieraccini, S. Scialò, A hybrid mortar virtual element method for discrete fracture network simulations. *J. Comput. Phys.* **306**, 148–166 (2016)

10. M.F. Benedetto, S. Berrone, S. Pieraccini, S. Scialò, The virtual element method for discrete fracture network simulations. *Comput. Methods Appl. Mech. Eng.* **280**, 135–156 (2014)
11. R.L. Berge, Ø.S. Klemetsdal, K.-A. Lie, Unstructured Voronoi grids conforming to lower dimensional objects. *Comput. Geosci.* **23**(1), 169–188 (2019)
12. I. Berre, F. Doster, E. Keilegavlen, Flow in fractured porous media: a review of conceptual models and discretization approaches. *Transp. Porous Media* **130**(1), 215–236 (2019)
13. D. Boffi, F. Brezzi, M. Fortin, *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics (Springer, Berlin Heidelberg, 2013)
14. D. Boffi, D.A. Di Pietro, Unified formulation and analysis of mixed and primal discontinuous skeletal methods on polytopal meshes. *ESAIM: Math. Model. Numer. Anal.* **52**(1), 1–28 (2018)
15. W.M. Boon, J.M. Nordbotten, I. Yotov, Robust discretization of flow in fractured porous media. *SIAM J. Numer. Anal.* **56**(4), 2203–2233 (2018)
16. A. Botella, B. Lévy, G. Caumon, Indirect unstructured hex-dominant mesh generation using tetrahedra recombination. *Comput. Geosci.* **20**(3), 437–451 (2015)
17. K. Brenner, J. Hennicker, R. Masson, P. Samier, Gradient discretization of hybrid-dimensional darcy flow in fractured porous media with discontinuous pressures at matrix-fracture interfaces. *IMA J. Numer. Anal.* (2016)
18. F. Brezzi, R.S. Falk, D.L. Marini, Basic principles of mixed virtual element methods. *ESAIM: M2AN* **48**(4), 1227–1240 (2014)
19. The Computational Geometry Algorithms Library. <http://www.cgal.org>
20. F. Chave, D.A. Di Pietro, L. Formaggia, A hybrid high-order method for Darcy flows in fractured porous media. *SIAM J. Sci. Comput.* **40**(2), A1063–A1094 (2018)
21. S.-W. Cheng, T.K. Dey, J. Shewchuk, *Delaunay Mesh Generation* (Chapman and Hall/CRC, 2012)
22. M.A. Christie, M.J. Blunt, *SPE-66599-MS, Chapter Tenth SPE Comparative Solution Project: A Comparison of Upscaling Techniques* (Society of Petroleum Engineers, Houston, Texas, 2001), p. 13
23. C. D’Angelo, A. Scotti, A mixed finite element method for Darcy flow in fractured porous media with non-matching grids. *Math. Model. Numer. Anal.* **46**(02), 465–489 (2012)
24. F. Dassi, S. Scacchi, Parallel solvers for virtual element discretizations of elliptic equations in mixed form. *Comput. Math. Appl.* (2019)
25. F. Dassi, G. Vacca, Bricks for the mixed high-order virtual element method: projectors and differential operators. *Appl. Numer. Math.* (2019)
26. M. Del Pra, A. Fumagalli, A. Scotti, Well posedness of fully coupled fracture/bulk Darcy flow with XFEM. *SIAM J. Numer. Anal.* **55**(2), 785–811 (2017)
27. J. Droniou, Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Model. Methods Appl. Sci.* **24**(08), 1575–1619 (2014)
28. J. Droniou, R. Eymard, T. Gallouët, R. Herbin, A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Model. Methods Appl. Sci.* **20**(02), 265–295 (2010)
29. L.J. Durlofsky, Upscaling of geocellular models for reservoir flow simulation: a review of recent progress, in *7th International Forum on Reservoir Simulation Bühl/Baden-Baden, Germany*, pp. 23–27 (2003)
30. B. Flemisch, I. Berre, W. Boon, A. Fumagalli, N. Schwenck, A. Scotti, I. Stefansson, A. Tatomir, Benchmarks for single-phase flow in fractured porous media. *Adv. Water Resour.* **111**, 239–258 (2018)
31. B. Flemisch, A. Fumagalli, A. Scotti, A review of the XFEM-based approximation of flow in fractured porous media, chapter advances in discretization methods, in *Advances in Discretization Methods: Discontinuities, Virtual Elements, Fictitious Domain Methods*, vol. 12, SEMA SIMAI Springer Series, ed. by G. Ventura, E. Benvenuti (Springer International Publishing, Cham, 2016), pp. 47–76
32. L. Formaggia, A. Fumagalli, A. Scotti, P. Ruffo, A reduced model for Darcy’s problem in networks of fractures. *ESAIM: Math. Model. Numer. Anal.* **48**(7), 1089–1116 (2014)

33. L. Formaggia, A. Scotti, F. Sottocasa, Analysis of a mimetic finite difference approximation of flows in fractured porous media. *ESAIM: M2AN* **52**(2), 595–630 (2018)
34. P. Frey, P.L. George, *Mesh Generation: Application to Finite Elements* (Wiley, 2013)
35. P.J. Frey, H. Borouchaki, P.-L. George, 3D Delaunay mesh generation coupled with an advancing-front approach. *Comput. Methods Appl. Mech. Eng.* **157**(1–2), 115–131 (1998)
36. N. Frih, V. Martin, J.E. Roberts, A. Saâda, Modeling fractures as interfaces with nonmatching grids. *Comput. Geosci.* **16**(4), 1043–1060 (2012)
37. H.A. Friis, M.G. Edwards, J. Mykkeltveit, Symmetric positive definite flux-continuous full-tensor finite-volume schemes on unstructured cell-centered triangular grids. *SIAM J. Sci. Comput.* **31**(2), 1192–1220 (2009)
38. A. Fumagalli, Dual virtual element method in presence of an inclusion. *Appl. Math. Lett.* **86**, 22–29 (2018)
39. A. Fumagalli, E. Keilegavlen, Dual virtual element method for discrete fractures networks. Technical report, [arXiv:1610.02905](https://arxiv.org/abs/1610.02905) [math.NA] (2017)
40. A. Fumagalli, E. Keilegavlen, Dual virtual element method for discrete fractures networks. *SIAM J. Sci. Comput.* **40**(1), B228–B258 (2018)
41. A. Fumagalli, E. Keilegavlen, Dual virtual element methods for discrete fracture matrix models. *Oil Gas Sci. Technol.—Revue d’IFP Energies Nouvelles* **74**(41), 1–17 (2019)
42. A. Fumagalli, E. Keilegavlen, S. Scialò, Conforming, non-conforming and non-matching discretization couplings in discrete fracture network simulations. *J. Comput. Phys.* **376**, 694–712 (2019)
43. A. Fumagalli, L. Pasquale, S. Zonca, S. Micheletti, An upscaling procedure for fractured reservoirs with embedded grids. *Water Resour. Res.* **52**(8), 6506–6525 (2016)
44. C. Geuzaine, J.-F. Remacle, Gmsh: a 3D finite element mesh generator with built-in pre- and post-processing facilities. *Int. J. Numer. Methods Eng.* **79**(11), 1309–1331 (2009)
45. J.E. Jones, P.S. Vassilevski, AMGE based on element agglomeration. *SIAM J. Sci. Comput.* **23**(1), 109–133 (2001)
46. E. Keilegavlen, R. Berge, A. Fumagalli, M. Staronni, I. Stefansson, J. Varela, I. Berre, Porepy: an open-source software for simulation of multiphysics processes in fractured porous media. Technical report, [arXiv:1908.09869](https://arxiv.org/abs/1908.09869) [math.NA] (2019)
47. L. Li, S.H. Lee, Efficient field-scale simulation of black oil in a naturally fractured reservoir through discrete fracture networks and homogenized media. *SPE Reserv. Eval. Eng.* **11**, 750–758 (2008)
48. V. Martin, J. Jaffré, J.E. Roberts, Modeling fractures and barriers as interfaces for flow in porous media. *SIAM J. Sci. Comput.* **26**(5), 1667–1691 (2005)
49. H. Mustapha, A Gabriel-Delaunay triangulation of 2D complex fractured media for multiphase flow simulations. *Comput. Geosci.* **18**(6), 989–1008 (2014)
50. H. Mustapha, K. Mustapha, A new approach to simulating flow in discrete fracture networks with an optimized mesh. *SIAM J. Sci. Comput.* **29**(4), 1439–1459 (2007)
51. J.M. Nordbotten, W. Boon, A. Fumagalli, E. Keilegavlen, Unified approach to discretization of flow in fractured porous media. *Comput. Geosci.* (2018)
52. J.M. Nordbotten, M.A. Celia, *Geological Storage of CO₂: modeling approaches for large-scale simulation* (Wiley, 2011)
53. P. Panfili, A. Cominelli, Simulation of miscible gas injection in a fractured carbonate reservoir using an embedded discrete fracture model, in *Abu Dhabi International Petroleum Exhibition and Conference, 10–13 November* (Society of Petroleum Engineers, Abu Dhabi, UAE, 2014)
54. J. Pellerin, A. Botella, F. Bonneau, A. Mazuyer, B. Chauvin, B. Lévy, G. Caumon, RINGMesh: a programming library for developing mesh-based geomodeling applications. *Comput. Geosci.* **104**, 93–100 (2017)
55. J. Pellerin, B. Lévy, G. Caumon, Toward mixed-element meshing based on restricted Voronoi diagrams. *Procedia Eng.* **82**, 279–290 (2014)
56. P.-A. Raviart, J.-M. Thomas, A mixed finite element method for second order elliptic problems. *Lect. Notes Math.* **606**, 292–315 (1977)

57. J.E. Roberts, J.-M. Thomas, Mixed and hybrid methods, in *Handbook of Numerical Analysis*, vol. II, Handb. Numer. Anal., II, (North-Holland, Amsterdam, 1991), pp. 523–639
58. T.H. Sandve, I. Berre, J.M. Nordbotten, An efficient multi-point flux approximation method for discrete fracture-matrix simulations. *J. Comput. Phys.* **231**(9), 3784–3800 (2012)
59. J. Schöberl, NETGEN an advancing front 2D/3D-mesh generator based on abstract rules. *Comput. Vis. Sci.* **1**(1), 41–52 (1997)
60. I. Stefansson, I. Berre, E. Keilegavlen, Finite-volume discretisations for flow in fractured porous media. *Transp. Porous Media* **124**(2), 439–462 (2018)
61. U. Trottenberg, C.W. Oosterlee, A. Schüller. *Multigrid* (Elsevier Academic Press, 2001)

Index

B

Barenblatt–Pattle solution, 58
Biot equations, 139

C

Constitutive laws, 125
Convergence, 147

D

Darcy, 141
Darcy's law, 135
Darcy velocity, 302
Discrete inequalities
 Korn–Poincaré in HHO spaces, 236
 Poincaré in HHO spaces, 238
 trace, 171

E

Error equations, 246

F

Finite volume methods, 123
Fourier's law, 141
Fracture, 302
 Darcy model, 154, 197, 304
 network, 319

G

Global reconstructions
 displacement, 235

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Switzerland AG 2021
D. A. Di Pietro et al. (eds.), *Polyhedral Methods in Geosciences*,
SEMA SIMAI Springer Series 27,
<https://doi.org/10.1007/978-3-030-69363-3>

divergence, 235
pressure, 234
strain, 234

Gradient discretization, 47
 gradient scheme, 48

Gradient discretization method
 coercivity, 31
 compactness, 32
 consistency, 32
 limit conformity, 32

H

HHO spaces
 scalar, 233
 vector, 234
Hooke's law, 136

I

Inf-sup condition, 237

L

Lifting operator, 202, 204, 206
 tensor traces, 187
Local discontinuous Galerkin method, 204
Local reconstructions
 displacement, 234
 divergence, 267
 pressure, 233
 strain, 234

M

Mass-lumping operator, 8

- Material strain, 136
 - Mesh family
 - agglomerated, 314
 - constrained Delaunay, 312
 - finite volume, 126
 - grid-cut, 313
 - polytopic-regular, 168
 - regular, 231, 266
 - Voronoi, 315
 - Mixed virtual elements, 260, 307
 - error equations, 272
 - inf-sup condition, 273
 - Model
 - Barenblatt–Biot equation, 254
 - degenerate parabolic, 38
 - weak solution, 39
 - multiple-network poroelasticity, 230
 - porous medium equation, 28, 38
 - Stefan equation, 7, 25, 38
 - Monotonicity, 144
 - Multi-Point Flux Approximation (MPFA), 135
 - convergence, 145
 - robustness, 148
 - MPSA, 137
 - MPxA methods, 145
 - Multi-point finite volumes, 119
 - Multi-point flux approximation, *see* MPFA
- N**
- Non-conforming finite element method, 3, 261
 - Non-conforming space, 6
 - Non-linear iterative schemes, 67
- P**
- Penalty function, 130
 - PolyDG method, 161, 166
 - discrete spaces, 169
 - elastodynamics, 186
 - error estimates, 190
 - stability, 188
 - fractured porous media, 196
 - unified formulation, 199
 - well-posedness, 208
 - inverse trace estimate, 169
 - Polynomial approximation, 172
 - Polytopal mesh, 5, 166, 231
 - PorePy, 307
 - Poroelastic material, 139
 - Projectors
 - elliptic, 239
 - L^2 -orthogonal, 232, 267
- Q**
- Quadrature free integration, 175
 - volume and interface integrals, 177
- R**
- Richards equation, 263
 - full discrete formulation, 271
 - convergence, 272, 287
 - semi-discrete formulation, 270
 - convergence, 272, 279
- S**
- Seismic wave propagation, 161, 193
 - Single-phase flow, 302
 - SIPDG method, 202
 - SPE10 benchmark, 316
 - Stabilization parameter, 187, 238
 - Static condensation, 23, 229
 - Strong symmetry, 137
- T**
- Thermo-poroelasticity, 141
 - Trace operators
 - average, 170, 238
 - jump, 170, 233
- W**
- Weak symmetry, 136, 138