MONASH University

Design and Analysis of Numerical Schemes with Characteristic Methods on Generic Grids for Flows in Porous Media

Hanz Martin Cheng

A thesis submitted for the degree of Doctor of Philosophy at
Monash University in 2019
School of Mathematics

Abstract

This thesis focuses on a mathematical model describing the recovery of oil in a process known as miscible displacement, in which a solvent, such as a short-chain hydrocarbon or pressurised carbon-dioxide, is injected into the oil reservoir to reduce the viscosity of the resident oil and push it towards the production wells. The model is an initial-boundary value problem for a nonlinearly-coupled elliptic-parabolic system. The main unknowns are the pressure of the fluid mixture, and the concentration of the injected solvent. Exact solutions of this model are usually inaccessible, especially with data as encountered in applications. The design and convergence analysis of numerical schemes for the model is therefore of particular importance. The purpose of this thesis is to design, test, and analyse numerical schemes for the complete coupled model.

The main contribution of this thesis is the development and the convergence analysis of a family of characteristic-based schemes on generic polygonal meshes. Instead of selecting one particular discretisation of the diffusive terms, we work inside a framework that enables a simultaneous analysis of various such discretisations. Hence, the generic framework of the gradient discretisation method was used for the discretisation of the diffusive terms, and characteristic-based methods for the advective terms. The first part of the thesis gives a short summary of the gradient discretisation method for Neumann boundary conditions. In order to perform characteristic tracking, we need the normal components of the velocity field to be continuous along the edges or faces of the cells, so that its flow is well defined. Also, in order to avoid the creation of artificial sources or sinks, the divergence of the velocity field should be preserved. Thus, we develop a novel method for reconstructing velocity fields which satisfies these properties. Following this, we study the mass conservation properties of characteristic-based methods. Here, we propose a combination of two characteristic-based methods, and we also devise an original post-processing technique, which ensures local and global mass conservation. Numerical tests are performed on a variety of polygonal meshes, producing very similar results regardless of the mesh (as

long as it is not too distorted), showing a certain robustness of the numerical scheme. Upon attempting to mitigate the grid effects for distorted meshes, the simplest solution we found was mesh refinement. Finally, we perform a rigorous convergence analysis for this family of characteristic-based schemes, using only weak regularity assumptions.

List of publications

The following is a list of publications/unpublished manuscripts resulting from this thesis.

1. H.M. Cheng and J. Droniou. Combining the hybrid mimetic mixed method and the Eulerian Lagrangian localised adjoint method for approximating miscible flows in porous media. In *Finite volumes for complex applications VIII–hyperbolic, elliptic and parabolic problems*, volume 200 of *Springer Proc. Math. Stat.*, pages 367-376. Springer, Cham, 2017.

2. H.M. Cheng and J. Droniou. An HMM–ELLAM scheme on generic polygonal meshes for miscible incompressible flows in porous media. *Journal of Petroleum Science and Engineering*, 172:707–723, 2019.

3. H.M. Cheng, J. Droniou, K.-N. Le. Convergence analysis of a family of ELLAM schemes for a fully coupled model of miscible displacement in porous media. *Numerische Mathematik*, 141(2):353–397, 2019.

4. H.M. Cheng, J. Droniou, K.-N. Le. A combined GDM–ELLAM–MMOC scheme for advection dominated PDEs. Submitted.

Declaration

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes two original papers published in peer reviewed journals, one conference proceeding, and one submitted publication. The core theme of the thesis is numerical analysis. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself,

the student, working within the School of Mathematics, Monash University, under the supervision of A/Prof. Jerome Droniou.

The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research. As is standard in mathematics, each author contributes equally to each aspect of the work.

I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author names Nature and % of Co-author's contribution	Co-authors, Monash student Y/N
2,5	Combining the hybrid mimetic mixed method and the Eulerian Lagrangian localised adjoint method for approximating flows in porous media.	Published	50%. Theoretical analysis and numerical tests.	1. Jerome Droniou. 50%. Theoretical analysis and numerical tests.	No
2,5	An HMM–ELLAM scheme on generic polygonal meshes for miscible incompressible flows in porous media.	Published	50%. Theoretical analysis and numerical tests.	1. Jerome Droniou. 50%. Theoretical analysis and numerical tests.	No

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author names Nature and % of Co-author's contribution	Co-authors, Monash student Y/N
4,5	A combined GDM–ELLAM–MMOC scheme for advection dominated PDEs.	Submitted	50%. Theoretical analysis and numerical tests.	1. Jerome Droniou. 20%. 2. Kim Ngan Le. 30%. Theoretical analysis and numerical tests.	No No
6	Convergence analysis of a family of ELLAM schemes for a fully coupled model of miscible displacement in porous media.	Published	50%. Theoretical analysis.	1. Jerome Droniou. 20%. 2. Kim Ngan Le. 30%. Theoretical analysis.	No No

Student signature: Hanz Martin Cheng **Date:** May 31, 2019

I hereby certify that the above declaration correctly reflects the nature and extent of the students and co-authors contributions to this work.

Main Supervisor signature: Jerome Droniou **Date:** May 31, 2019

Acknowledgements

I would want to thank my family, in particular, my parents and my brother, for being very supportive towards my studies and research. I would also want to express my gratitude to my supervisor Jerome, who patiently supervised me throughout the course of my PhD. Through his guidance and supervision, I was able to develop my knowledge and confidence in doing research in applied mathematics and numerical analysis. It was also through his encouragement that I was able to push myself to generate interesting ideas and improvements to some numerical schemes, which later on lead to their publication in journals. I am also very thankful for his advice on which conferences/workshops would be helpful for building up my knowledge and background, and as well as building up networks for my planned career in academia. I would also want to thank one of our colleagues, Ngan, who worked on the characteristic-based schemes together with us. I am also very thankful for some of the tips that she gave me on how to find and apply for postdoctoral positions. I am also very grateful to one of our collaborators, Jerome Bonelle, and the EDF team, who supported my research visit at EDF, France. He has worked with me and helped me understand the details of face-based CDO schemes in Code Saturne, which would be very useful for the planned extension to 3D of the velocity reconstructions. I would also want to thank one of my fellow students, Daniel, for writing the codes which served as the framework/backbone of the implementation of hybrid high order schemes. I am also thankful to Monash University, the Australian Mathematical Society, the Australia and New Zealand Industrial and Applied Mathematics society, and the Australian Research Council for supporting my conference travels and research visits. I would also like to thank my milestone panelists Simon, Paul, and Todd, for asking me questions and for giving me feedback during my milestones, which helped me further understand the model that I am studying. I also thank the referees, Prof. Franck Boyer and Prof. Ian Turner, for the detailed reading of my thesis, and for the comments and suggestions that helped to improve the quality and presentation of my thesis. I would also want to thank John, Linda and Karen for helping me

out with all of the administrative requirements. I would also want to thank my fellow PhD students for making these years fun and enjoyable.

Contents

Chapter 1

Introduction

1.1 The miscible flow model

This thesis focuses on a mathematical model describing the recovery of oil in a process known as miscible displacement, in which a solvent, such as a short-chain hydrocarbon or pressurised carbon-dioxide, is injected into the oil reservoir to reduce the viscosity of the resident oil and push it towards the production wells. One of the models that describes the said process is the miscible flow model, which was first introduced in [67].

Let Ω be a bounded domain in $\mathbb{R}^d$ and $[0, T]$ be a time interval. Denote by $\mathbf{K}(\mathbf{x})$ and $\phi(\mathbf{x})$ the permeability tensor and the porosity of the medium, respectively. Then, neglecting gravity, the miscible flow model is given by:

$$\begin{aligned} \nabla \cdot \mathbf{u} &= q^+ - q^- := q \qquad \text{on } \Omega \times [0, T] \\ \mathbf{u} &= -\frac{\mathbf{K}}{\mu(c)} \nabla p \qquad \text{on } \Omega \times [0, T] \end{aligned} \tag{1.1a}$$

$$\phi \frac{\partial c}{\partial t} + \nabla \cdot (\mathbf{u}c - \mathbf{D}(\mathbf{x}, \mathbf{u}) \nabla c) = q^+ - cq^- := q_c \qquad \text{on } \Omega \times [0, T] \tag{1.1b}$$

with unknowns $p(\boldsymbol{x}, t), \mathbf{u}(\boldsymbol{x}, t)$, and $c(\boldsymbol{x}, t)$ which denote the pressure of the mixture, the Darcy velocity, and the concentration of the injected solvent, respectively. We note that the model (1.1) is derived under the assumption that the fluid and the rock are incompressible. The more generic formulation for the pressure equation (1.1a) is given by $\nabla \cdot \mathbf{u} = q + \phi c_s \dfrac{\partial p}{\partial t}$, where c_s is the total compressibility of the system. The concentration equation (1.1b) is then modified accordingly. In particular, the compressibility c_s is related to $\dfrac{\partial \rho}{\partial p}$ where $\rho(p)$ is the density of the fluid, and the assumption that $c_s = 0$

implies that the density is constant. However, in many cases, c_s is very small and negligible, and hence the assumption on incompressibility is not very restrictive [45]. As a matter of fact, this assumption is also used in some engineering applications [26, 68, 79]. In particular, for petroleum engineering, this may be used in a gas cap drive reservoir or when the reservoir pressure drops below the bubble-point pressure [62, Chapter 7]. The model (1.1) is understood in the following manner: (1.1a) gives us the conservation of mass for the total fluid (mixture of oil and solvents), whereas (1.1b) gives us the conservation of mass of the injected solvents. This captures the physical phenomenon of two or more components (oil and solvents) flowing along a single phase, for which each of the components have different concentration; the derivation and a more detailed interpretation of this model are given in [45, Chapter 2].

The functions q^+ and q^- represent the injection and production wells respectively, and $\mathbf{D}(\mathbf{x}, \mathbf{u})$ denotes the diffusion tensor

$$\begin{aligned} &\mathbf{D}(\mathbf{x}, \mathbf{u}) = \phi(\mathbf{x}) \left[d_m \mathbf{I} + d_l |\mathbf{u}| E(\mathbf{u}) + d_t |\mathbf{u}| \left(\mathbf{I} - E(\mathbf{u}) \right) \right], \\ &\text{with } E(\mathbf{u}) = \left(\frac{u_i u_j}{|\mathbf{u}|^2} \right)_{i,j}. \end{aligned} \tag{1.1c}$$

Here, d_m is the molecular diffusion coefficient, d_l and d_t are the longitudinal and transverse dispersion coefficients respectively, and $E(\mathbf{u})$ is the projection matrix along the direction of $\mathbf{u}$. Also, $\mu(c) = \mu(0)[(1-c) + M^{1/4}c]^{-4}$ is the viscosity of the fluid mixture, where $M = \mu(0)/\mu(1)$ is the mobility ratio of the two fluids. As usually considered in numerical tests, we consider no-flow boundary conditions, and zero initial conditions for the concentration:

$$\mathbf{u} \cdot \mathbf{n} = (\mathbf{D}\nabla c) \cdot \mathbf{n} = 0 \text{ on } \partial\Omega \times [0, T], \qquad c(\cdot, 0) = 0 \text{ in } \Omega. \tag{1.1d}$$

Essentially, the boundary conditions means that nothing flows into or out of the domain, except through the sources and sinks, located at the injection and production wells. The pressure is fixed by imposing a zero average:

$$\int_\Omega p(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} = 0 \qquad \forall t \in [0, T] \tag{1.2}$$

Remark 1.1.1 (Injection concentration and gravity)**.** *The model* (1.1) *assumes an injection concentration of* 1 *(since this is the case in most numerical tests) and neglects the gravity effects. A generic injection concentration* $\widehat{c}$ *could be considered upon the trivial modification* $q^+ \rightsquigarrow \widehat{c}q^+$ *in* (1.1b)*. To include gravity effect, we would have to set* $\mathbf{u} = -\frac{\mathbf{K}}{\mu(c)}(\nabla p - \rho \mathbf{g})$ *, where* ρ *is the density of the fluid. The analysis and numerical schemes we develop thereafter can easily be adapted to both changes.*

Exact solutions of this model are usually inaccessible, especially with data as encountered in applications. The design and convergence analysis of numerical schemes for (1.1) is therefore of particular importance. The purpose of this thesis is to design, test, and analyse numerical schemes for the complete coupled model (1.1).

1.2 Notations for Sobolev spaces

In this section, we present some of the common notations and definitions for Sobolev spaces, which will commonly be used throughout the thesis. We start by defining the L^p spaces.

Definition 1.2.1 (L^p spaces)**.** *Let $1 \leq p \leq \infty$. $f \in L^p(\Omega)$ if and only if $\|f\|_{L^p(\Omega)} < \infty$, where*

$$\|f\|_{L^p(\Omega)} = \left(\int_\Omega |f|^p \right)^{\frac{1}{p}} \text{ if } p < \infty$$
$$\|f\|_{L^\infty(\Omega)} = \inf\{M \text{ such that } |f(x)| \leq M \text{ for a.e. } x \in \Omega\}.$$

The Sobolev spaces $W^{s,p}$ are then defined to be:

Definition 1.2.2 (Sobolev space $W^{s,p}(\Omega)$)**.** *Let s be an integer with $s \geq 0$ and $p \in \mathbb{R}$ such that $1 \leq p \leq \infty$. The Sobolev space $W^{s,p}(\Omega)$ is defined to be*

$$W^{s,p}(\Omega) := \{f \in L^p(\Omega) \text{ such that } D^\alpha f \in L^p(\Omega), |\alpha| \leq s\},$$

where

$$D^\alpha f := \frac{\partial^{|\alpha|} f}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_n} \dots \partial x_n^{\alpha_n}}, \text{ with } \sum_{i=1}^{n} \alpha_i = |\alpha|,$$

and the derivatives are taken in the weak sense.

In particular, when $p = 2$, the space $W^{s,2}(\Omega)$ is a Hilbert space, and is often written $H^s(\Omega) = W^{s,2}(\Omega)$.

Here, we remark that velocity fields in the miscible flow model (1.1) are in practice discontinuous, and hence the space $H^1(\Omega)$ is not the proper space for which the velocity fields are defined. We however require that the normal component of the velocity field to be continuous across the edges, otherwise the flow of the velocity field will not be well defined. We thus introduce, for the velocity field, the $H(\text{div}, \Omega)$ space, defined to be

$$H(\text{div}, \Omega) := \{\mathbf{u} \in L^2(\Omega)^d \text{ such that } \text{div}(\mathbf{u}) \in L^2(\Omega)\}.$$

In particular, functions that belong to the $H(\mathrm{div}, \Omega)$ space have normal components which are continuous across the edges [16, Chapter III.1]. We also recall the definition of Bochner spaces, which deals with space-time functions by treating them as a collection of functions of space, parametrized by time.

Definition 1.2.3 (Bochner space $L^p(0,T;X)$)**.** *Let X denote a separable Banach space, with norm $\|\cdot\|_X$. The space $L^p(0,T;X)$ consists of all measurable functions $u : [0,T] \to X$ with*

$$\|u\|_{L^p(0,T;X)} := \left(\int_0^T \|u(t)\|_X^p \, dt \right)^{\frac{1}{p}} < \infty, \ \textit{for } p \in [1,\infty),$$

and

$$\|u\|_{L^\infty(0,T;X)} := \inf\{M \textit{ such that } \|u(t)\|_X \leq M \textit{ for a.e. } t \in [0,T]\} < \infty.$$

1.3 Review of literature

A number of numerical methods have been used to approximate the solutions of (1.1) and similar models, from finite difference techniques [34, 69], to finite element schemes [33, 48], to discontinuous Galerkin methods [11, 72], to control volume methods with flux limiting [55, 56], to finite volume methods [19, 20]. Closer to the focus of this work are the Modified Method of Characteristics (MMOC) and the Eulerian-Lagrangian Localized Adjoint Method (ELLAM). These methods, designed to deal with the advective terms, have been applied in conjunction with mixed finite elements (for the pressure equation) and conforming finite elements (for the diffusion terms in the concentration equation) in [21, 47, 78]. The combination with FE methods severely restricts the cell geometries that can be managed with such methods. On the contrary, recent finite volume (FV) methods can accommodate very generic mesh geometries, see the review [39] and references therein. Among those, the hybrid mimetic mixed (HMM) method of [41] is a family of numerical schemes for diffusion equations which gathers three separately developed numerical methods: hybrid finite volumes [51], mixed-hybrid mimetic finite differences [15], and mixed finite volume [38]. The HMM has been adapted in [20] to the model (1.1), using an upwind discretisation of the advective term. The drawback of such a discretisation is an additional numerical diffusion, which leads to a widening of the transition layer between the regions $c \approx 1$ and $c \approx 0$.

On the other hand, an overview of studies and analysis involving ELLAM schemes is presented in [74]. Convergence analysis was performed for MFEM–ELLAM schemes (or similar) in [9, 76]. We note here that [9] only considers

the concentration equation (1.1b) (assuming that $\mathbf{u}$ is given), whereas [76] provides error estimates for the complete coupled model (1.1). However, these analyses were carried out under restrictive regularity assumptions on the porosity ϕ and on the solution $(p, \mathbf{u}, c)$ to the model; in particular, the minimal assumptions in [76] are $c \in H^1(0,T;H^2(\Omega)) \cap L^\infty(0,T;W^{2,r}(\Omega))$ (for $r > 2$) and $\mathbf{u} \in W^{1,\infty}(\Omega \times (0,T))$, and [9] supposes that $c, \mathbf{D}\nabla c \in C^1(0,T;H^1(\Omega))$ and $\phi, \mathbf{u} \in W^{1,\infty}(\Omega \times (0,T))$. However, in reservoir modeling, transitions between different rock layers are usually discontinuous; thus, the permeability may vary rapidly over several orders of magnitude, with local variations in the range of 1mD to 10D, where D is the Darcy unit [65]. Due to this discontinuity of $\mathbf{K}$, the solutions to (1.1) cannot expect to satisfy the regularity conditions stated above. Actually, all reported numerical tests [20, 18, 23, 78] seem to have been on cases for which such regularity of the data and/or the solutions does not hold.

More recent developments of ELLAM techniques involve Volume Corrected Characteristic Mixed Methods (VCCMM), which are, in essence, ELLAM schemes with volume adjustment to achieve local mass conservation. Convergence analysis, as well as stability, monotonicity, maximum and minimum principles for these schemes have been studied in [7, 8]. However, these studies only consider a single pure advection model (that is, (1.1b) with $\mathbf{D} = 0$), and assume the regularity $\mathbf{u} \in C^1(\Omega \times (0,T))$, which, as explained above, is not expected in applications. Without accounting for diffusion, the maximum principle is accessible, and the analysis strongly benefits from the resulting L^∞ bounds on the approximate solution. On the contrary, in the presence of anisotropic heterogeneous diffusion $\mathbf{K}$ and $\mathbf{D}(\mathbf{u})$, and on grids as encountered in applications, constructing schemes that satisfy the maximum principle is extremely difficult – to this day, only *nonlinear* schemes are known to preserve the maximum principle in general, and even these do not necessarily have nice coercivity features [39].

As a matter of fact, the convergence analysis of numerical approximations of (1.1) under weak regularity assumptions has recently received an increasing interest; see, e.g., [20, 19] for finite volume methods and [57, 72] for discontinuous Galerkin methods. It therefore seems natural to consider doing such an analysis for characteristic-based discretisation of the advection term.

1.4 Thesis aims

The aim of this thesis is to design and implement characteristic-based schemes on generic polygonal meshes, and to analyse their convergence without as-

suming smoothness on the data or solution, which is not observed in practice. We want to conduct this numerical analysis in a framework that is applicable to various choices of discretisation of the diffusion terms.

1.5 Outline

We start by giving a short summary of the gradient discretisation method (GDM) for Neumann boundary conditions in Chapter 2. Two numerical schemes which fall under this framework will be presented, namely, the hybrid mimetic mixed (HMM) schemes, and the hybrid high order (HHO) schemes. Since our aim is to use characteristic-based schemes, we need to make sure that the reconstructed velocity field satisfies the no-flow boundary conditions, and also the preservation of divergence in (1.1a). The general idea to achieve these properties is to reconstruct $H(\mathrm{div}, \Omega)$ velocity fields from the fluxes obtained from the HMM and HHO schemes, which is discussed in Chapter 3. In particular, we focus on $\mathbb{RT}_k$ elements on simplices, then on quadrilaterals. For $\mathbb{RT}_k$ elements on simplices, computation of internal sub-fluxes are required. Our main contribution in this chapter comes in the form of computing sub-fluxes using the C and A methods in Sections 3.1.3 and 3.1.4, respectively. Moreover, the A method can be extended into 3D, as seen in Section 3.1.5.

We then proceed to give a short summary of two of the characteristic-based numerical schemes: the Eulerian Lagrangian Localized Adjoint Method (ELLAM) and the Modified Method of Characteristics (MMOC) in Chapter 4. Our main contributions in this chapter are:

- the theory which establishes the existence, and some estimates on the flow under minimal regularity assumptions (Section 4.2);
- precise mass balance analysis for the MMOC scheme (Section 4.4);
- combined ELLAM–MMOC scheme for the advection-reaction equation (Section 4.5);
- a new volume adjustment algorithm which ensures that the numerical approximations satisfy local volume conservation (Section 4.6.2).

Having described the numerical schemes for the diffusive and advective components, we then form the combined GDM–ELLAM–MMOC (GEM) scheme for the miscible flow model (1.1) in Chapter 5. The codes used for the numerical implementation of the GEM scheme (with HMM as the choice for the gradient discretisation) can be found in `https://github.`

com/hanzcheng/HMM-GEM-2D. As a point of reference, the CPU runtime presented for all numerical tests were obtained from a Windows desktop with an Intel i7-4790 processor, 3.60Ghz, 8MB cache, and 16GB of RAM. Here, our main contributions are the presentation of the GEM scheme, and adaptations of the local volume adjustment algorithm to the miscible flow model. Some other contributions were the introduction of a modification in the approximation of the trace-forward region around the injection wells for the HMM–ELLAM, so that it may physically be interpreted as the volume injected from the well being transported to the surrounding cells proportionally. We also determined, depending on the time step and how regular/irregular the cell is, the proper amount of points to track along the edges of each cell, in order to have a good initial approximation of the trace-back and trace-forward regions. We started by performing numerical tests using an HMM–ELLAM scheme. These results obtained from HMM–ELLAM were then compared to those from HMM–upwind schemes in order to show the advantages of using characteristic-based schemes. A comparison with MFEM–ELLAM also shows the advantages of the HMM–ELLAM, in terms of having a cheaper computational cost, achieving a good preservation of the physical bounds on c, and the capability to be adapted to more generic meshes. The HMM–GEM, which serves as an improvement over the HMM–ELLAM, is then presented. For the sake of completeness, the numerical results from the HMM–GEM and HMM–ELLAM were compared to those from HMM–MMOC. In general, the numerical results from the HMM–GEM scheme achieves both a good preservation of the physical bounds on c, and of mass conservation. Moreover, the local volume conservation property of HMM–GEM is better than that of HMM–ELLAM, especially on distorted cells. For all these tests, grid effects were present for highly distorted meshes.

Grid effects are then studied in more detail by considering less distorted grids. High order methods are expected to perform better on coarse meshes, and hence the idea was to use a HHO scheme for the gradient discretisation, while maintaining piecewise constant approximations for the concentration c. However, this did not help mitigate the grid effects. On the other hand, it was observed that mesh refinement can reduce grid effects. These lead to the conclusion that retaining the piecewise constant approximations for the advective components of the scheme is the main cause of the grid effects. Due to this, we believe that going for a fully high order approximation of c might be able to mitigate the grid effects on coarse grids; however, this comes with a computational cost that is much more expensive. Hence, at this stage, an efficient way to mitigate the grid effects on coarse distorted grids for characteristic-based schemes is still an open problem.

Finally, in Chapter 6, we present the convergence analysis of numerical

schemes that fall in the GEM framework. The main difference between the analysis presented here and most of those presented in the literature is the fact that the results are established using only weak regularity assumptions on the solution (which are satisfied in practical applications). The convergence is established in detail for the GDM–ELLAM schemes. Minor modifications to the proofs for the GDM–ELLAM are needed in order to establish convergence for the GDM–MMOC schemes. The convergence of schemes in the GEM framework are then obtained by combining the convergence results for the GDM–ELLAM and GDM–MMOC schemes. We note however that this convergence analysis assumes a perfect computation of the tracked regions; future work will address the issue of accounting for approximation in tracked regions, and adjustment strategies, in the convergence analysis.

1.6 Mesh

For our discretisations, "mesh" is to be understood in the simplest intuitive way: a partition of Ω into polygonal (in 2D) or polyhedral (in 3D) sets. Following the notations in [40, Definition 7.2], we denote $\mathcal{T} = (\mathcal{M}, \mathcal{E})$ to be the set of cells K and faces (edges in 2D) σ of our mesh, respectively. This definition covers a large variety of meshes. In particular, the cells are not assumed to be convex (as in Figure 2.1), and the common boundary of two neighbouring cells can include more than one face (as in Figure 1.2, left). For a cell $K \in \mathcal{M}$, $\mathcal{E}_K \subset \mathcal{E}$ denotes the set of faces (edges) of the cell K.

1.6.1 Types of meshes

Our numerical tests will be usually performed on these four types of meshes: Cartesian type meshes, hexahedral meshes (see Fig.1.1), non-conforming meshes, and finally on Kershaw type meshes, as described in [59] (see Fig. 1.2).

Unless otherwise specified, over the domain $\Omega = (0, 1) \times (0, 1)$, Cartesian meshes consist of square cells that have dimension 0.0625×0.0625; hexahedral meshes of hexagonal cells, with diameters ranging from 0.0353 to 0.1297 units; non-conforming meshes are assumed to be locally refined over the top right region, with square cells of dimension 0.0475×0.0475, and square cells of dimension 0.0583×0.0583 on the lower left region, and rectangular cells elsewhere; Kershaw meshes are made up of of quadrilateral cells (most of which are distorted), with cell diameters ranging from 0.0831 to 0.3287 units. These are to be scaled appropriately when dealing with the domain $\Omega = (0, 1000) \times (0, 1000)$.

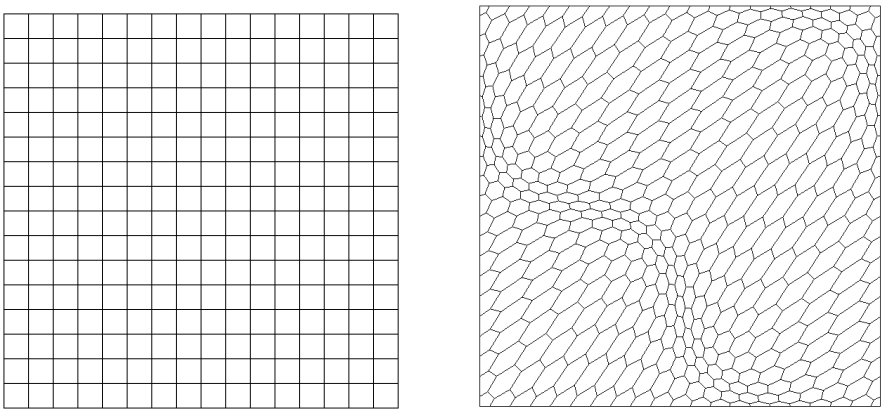


Figure 1.1: Mesh types (left: Cartesian ; right: hexahedral).

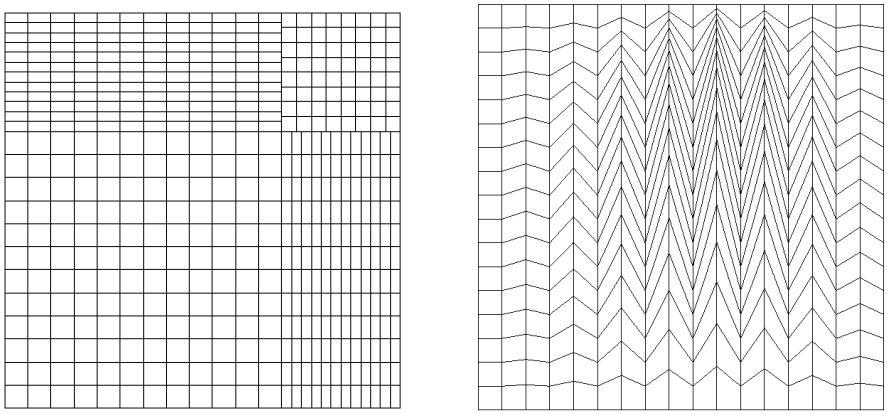


Figure 1.2: Mesh types (left: non-conforming ; right: Kershaw).

Chapter 2

Gradient discretisation method for anisotropic diffusion problems

Consider the anisotropic diffusion problem with Neumann boundary condition

$$\begin{aligned} -\mathrm{div}(\Lambda\nabla p) &= f \text{ on } \Omega \\ \Lambda\nabla p\cdot\mathbf{n} &= g \text{ on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $\mathbf{n}$ is the unit outward normal to $\partial\Omega$. Here, we assume that

- Ω is an open connected subset of $\mathbb{R}^d$ (where d is a positive integer) with a Lipschitz boundary , (2.2a)
- Λ is a measurable function from Ω to the set of $d\times d$ symmetric matrices and there exists $\underline{\lambda},\overline{\lambda}>0$ such that, for a.e. $\boldsymbol{x}\in\Omega$, the eigenvalues of $\Lambda(\boldsymbol{x})$ are in $[\underline{\lambda},\overline{\lambda}]$, (2.2b)
- $f\in L^2(\Omega)$ and $g\in L^2(\partial\Omega)$ with

$$\int_\Omega f+\int_{\partial\Omega} g d\zeta = 0. \tag{2.2c}$$

We note here that the solution of (2.1) is only unique up to an additive constant. For a unique solution, we need to impose an additional condition on p. In particular, we set

$$\int_\Omega p = 0. \tag{2.3}$$

Aside from its applications to flows in porous media (as in the pressure equation (1.1a)), diffusion problems of this type can be applied to image

processing, spread of heat, etc. Aside from the unknown p, the other main quantities of interest are the gradient ∇p and the fluxes along the faces of each cell $\int_\sigma \Lambda \nabla p \cdot \mathbf{n}_{K,\sigma}$. To be specific, it is of particular importance in model (1.1) that the approximations to the fluxes $F_{K,\sigma} \approx -\int_\sigma \Lambda \nabla p \cdot \mathbf{n}_{K,\sigma}$ are accurate, as they are crucial in the reconstruction of the Darcy velocity, as discussed in [23].

In this chapter, gradient schemes with different orders of accuracy, starting with the low order Hybrid Mimetic Mixed (HMM) [41], followed by the Hybrid High Order (HHO) schemes [30], will be presented. The performance of each of the schemes will be measured through the L^2 norm of the error upon comparison with the actual solution.

2.1 Gradient scheme for the diffusion problem

To write the weak formulation for (2.1), we first introduce the space

$$H^1_*(\Omega) := \{u \in H^1(\Omega) : \int_\Omega u = 0\}.$$

We then multiply (2.1) by a test function $\xi \in H^1_*(\Omega)$ and perform integration by parts to obtain the following weak form: Find $p \in H^1_*(\Omega)$ such that

$$\int_\Omega \Lambda \nabla p \cdot \nabla \xi = \int_\Omega f\xi + \int_{\partial\Omega} g\gamma(\xi) d\zeta\,, \quad \forall \xi \in H^1_*(\Omega). \tag{2.4}$$

Here, $\gamma : H^1(\Omega) \to L^2(\partial\Omega)$ is the trace operator, with

$$\gamma(\xi) = \xi_{|\partial\Omega}, \qquad \forall \xi \in H^1(\Omega).$$

Owing to assumption (2.2c), an equivalent form for the weak formulation (2.4) is: Find $p \in H^1(\Omega)$ such that

$$\int_\Omega \Lambda \nabla p \cdot \nabla \xi + \int_\Omega p \int_\Omega \xi = \int_\Omega f\xi + \int_{\partial\Omega} g\gamma(\xi) d\zeta\,, \quad \forall \xi \in H^1(\Omega). \tag{2.5}$$

This can be seen by taking $\xi \equiv 1$ in (2.5), which results to $\int_\Omega p = 0$. The weak formulation (2.4) is then written in its discretised form by replacing the continuous functions and their gradients by their discrete counterparts. This is known as the gradient discretisation method (GDM) [40]. This framework contains many classical methods, including finite elements and finite volumes. The discrete elements are given by what is a called a gradient discretisation

(GD), and the convergence of the resulting schemes (called gradient schemes (GS)) can be established under a few assumptions on the gradient discretisations. We give here a brief presentation of GDs for homogeneous Neumann boundary conditions and the standard properties that ensure the convergence of the corresponding GSs for standard elliptic and parabolic PDEs.

Definition 2.1.1 (GD for homogeneous Neumann boundary conditions)**.** *A gradient discretisation for homogeneous Neumann boundary conditions is* $\mathcal{D} = (X_\mathcal{D}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$*, where*

- $X_\mathcal{D}$ *is a finite-dimensional real space, describing the unknowns of the chosen scheme,*
- *the function reconstruction* $\Pi_\mathcal{D} : X_\mathcal{D} \to L^\infty(\Omega)$ *is linear,*
- *the gradient reconstruction* $\nabla_\mathcal{D} : X_\mathcal{D} \to L^\infty(\Omega)^d$ *is linear.*

The operators $\Pi_\mathcal{D}$ *and* $\nabla_\mathcal{D}$ *must be chosen so that*

$$\|v\|_\mathcal{D} := \left(\|\nabla_\mathcal{D} v\|^2_{L^2(\Omega)} + \left| \int_\Omega \Pi_\mathcal{D} v \right|^2 \right)^{\frac{1}{2}}$$

is a norm on $X_\mathcal{D}$*.*

Considering for example (2.5) with $g = 0$ and replacing the space $H^1(\Omega)$ with $X_\mathcal{D}$, the functions by reconstructions using $\Pi_\mathcal{D}$ and the gradients by reconstructions $\nabla_\mathcal{D}$, we obtain the corresponding gradient scheme: Find $p \in X_\mathcal{D}$ such that

$$\int_\Omega \Lambda \nabla_\mathcal{D} p \cdot \nabla_\mathcal{D} v + \int_\Omega \Pi_\mathcal{D} p \int_\Omega \Pi_\mathcal{D} v = \int_\Omega f \Pi_\mathcal{D} v \,, \quad \forall v \in X_\mathcal{D}. \tag{2.6}$$

Remark 2.1.2. *For any GD for which there is a* $v \in X_\mathcal{D}$ *such that* $\nabla_\mathcal{D} v = 0$ *and* $\Pi_\mathcal{D} v = 1$*,* (2.6) *implies that* $\int_\Omega \Pi_\mathcal{D} p = 0$*. Hence, for these GDs,* (2.6) *can equivalently be written as: Find* $p \in X_\mathcal{D}$ *such that*

$$\int_\Omega \Lambda \nabla_\mathcal{D} p \cdot \nabla_\mathcal{D} v = \int_\Omega f \Pi_\mathcal{D} v \,, \quad \forall v \in X_\mathcal{D},$$
$$\int_\Omega \Pi_\mathcal{D} p = 0.$$

The accuracy of a GD and convergence properties of the resulting GS are measured through three parameters, that respectively correspond to a discrete Poincaré–Wirtinger constant, an interpolation error, and a measure of defect of conformity (error in a discrete Stokes formula):

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}}} \frac{\|\Pi_{\mathcal{D}} v\|_{L^2(\Omega)}}{\|v\|_{\mathcal{D}}}, \tag{2.7a}$$

$$\forall \varphi \in H^1(\Omega)\,,\ S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} (\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)}), \tag{2.7b}$$

$$\forall \boldsymbol{\phi} \in H(\mathrm{div}, \Omega)\,,$$

$$W_{\mathcal{D}}(\boldsymbol{\phi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v \cdot \boldsymbol{\phi} + \Pi_{\mathcal{D}} v \mathrm{div}\boldsymbol{\phi}) \right|, \tag{2.7c}$$

Definition 2.1.3 (Properties of GDs)**.** *A sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ of space gradient discretisations is*

- *coercive if there exists $C_p \in \mathbb{R}_+$ such that $C_{\mathcal{D}_m} \le C_p$ for all $m \in \mathbb{N}$,*
- *GD-consistent if, for all $\varphi \in H^1(\Omega)$, $S_{\mathcal{D}_m}(\varphi) \to 0$ as $m \to \infty$,*
- *limit-conforming if, for all $\boldsymbol{\phi} \in H(\mathrm{div}, \Omega)$, $W_{\mathcal{D}_m}(\boldsymbol{\phi}) \to 0$ as $m \to \infty$,*
- *compact if for any sequence $v_m \in X_{\mathcal{D}_m}$ such that $(\|v_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded, the sequence $(\Pi_{\mathcal{D}_m} v_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(\Omega)$.*

Remark 2.1.4. *The limit-conformity or compactness of a sequence of space GDs implies its coercivity [40, Lemmas 3.8 and 3.9]. The latter is explicitly mentioned since a bound on $C_{\mathcal{D}_m}$ is useful for the analysis.*

Remark 2.1.5. *For non-homogeneous Neumann boundary conditions, a linear trace reconstruction $\mathbb{T}_{\mathcal{D}} : X_{\mathcal{D}} \to L^\infty(\partial\Omega)$ would be needed, and the three parameters associated with the convergence of a gradient scheme are modified as follows:*

$$C_{\mathcal{D}} = \max_{v \in X_{\mathcal{D}}} \left\{ \frac{\|\Pi_{\mathcal{D}} v\|_{L^2(\Omega)}}{\|v\|_{\mathcal{D}}}, \frac{\|\mathbb{T}_{\mathcal{D}} v\|_{L^2(\Omega)}}{\|v\|_{\mathcal{D}}} \right\},$$

$$\forall \varphi \in H^1(\Omega)\,,\ S_{\mathcal{D}}(\varphi) = \min_{v \in X_{\mathcal{D}}} \big(\|\Pi_{\mathcal{D}} v - \varphi\|_{L^2(\Omega)} + \|\mathbb{T}_{\mathcal{D}} v - \gamma\varphi\|_{L^2(\partial\Omega)} + \|\nabla_{\mathcal{D}} v - \nabla\varphi\|_{L^2(\Omega)} \big),$$

$$\forall \boldsymbol{\phi} \in H(\mathrm{div}, \Omega)\,,$$

$$W_{\mathcal{D}}(\boldsymbol{\phi}) = \max_{v \in X_{\mathcal{D}} \setminus \{0\}} \frac{1}{\|v\|_{\mathcal{D}}} \left| \int_\Omega (\nabla_{\mathcal{D}} v(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}) + \Pi_{\mathcal{D}} v(\boldsymbol{x}) \mathrm{div}\boldsymbol{\phi}(\boldsymbol{x}))\, d\boldsymbol{x} \right.$$

$$- \int_{\partial\Omega} \mathbb{T}_{\mathcal{D}} v(\boldsymbol{x}) \gamma_{\mathbf{n}} \boldsymbol{\phi}(\boldsymbol{x}) d\gamma(\boldsymbol{x}) \Big|,$$

where $\gamma_{\mathbf{n}}\boldsymbol{\phi}$ is the normal trace of $\boldsymbol{\phi}$ in $\partial\Omega$. For more details and for other types of boundary conditions, we refer the reader to [40, Chapters 2-3].

2.1.1 Error Estimates and Convergence

Lemma 2.1.6 (Regularity of the limit)**.** *Let $(\mathcal{D}_m)_{m\in\mathbb{N}}$ be a limit-conforming sequence of gradient discretisations in the sense of Definition 2.1.3. Let $p_m \in X_{\mathcal{D}_m}$ be such that $(\|p_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$ is bounded. Then there exists $p \in H^1(\Omega)$ such that, up to a subsequence,*

$$\Pi_{\mathcal{D}_m} p_m \to p \text{ weakly in } L^2(\Omega)$$
$$\nabla_{\mathcal{D}_m} p_m \to \nabla p \text{ weakly in } L^2(\Omega)^d.$$

Proof. Owing to [40, Lemma 3.8], the sequence $(\mathcal{D}_m)_{m\in\mathbb{N}}$ is coercive and hence the sequence $(\Pi_{\mathcal{D}_m} p_m)_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega)$. Therefore, there exists a subsequence of $(\Pi_{\mathcal{D}_m} p_m)_{m\in\mathbb{N}}$, $p \in L^2(\Omega)$, and $\mathbf{v} \in L^2(\Omega)^d$ such that $(\Pi_{\mathcal{D}_m} p_m)_{m\in\mathbb{N}}$ converges to p weakly in $L^2(\Omega)$ and $(\nabla_{\mathcal{D}_m} p_m)_{m\in\mathbb{N}}$ converges to $\mathbf{v}$ weakly in $L^2(\Omega)^d$. Due to the boundedness of $(\|p_m\|_{\mathcal{D}_m})_{m\in\mathbb{N}}$, and the limit conformity of the sequence, we pass through the limit in $W_{\mathcal{D}_m}(\boldsymbol{\phi})$ in order to obtain

$$\forall \boldsymbol{\phi} \in H(\mathrm{div}, \Omega), \quad \int_\Omega (\mathbf{v} \cdot \boldsymbol{\phi} + p \mathrm{div} \boldsymbol{\phi}) = 0.$$

This being true for all $\boldsymbol{\phi} \in (C_c^\infty(\Omega))^d$ implies that $\mathbf{v} = \nabla p$, and hence $p \in H^1(\Omega)$ and $\nabla_{\mathcal{D}_m} p_m \to \nabla p$ weakly in $L^2(\Omega)^d$. ∎

Lemma 2.1.7 (Error Estimate)**.** *Under Assumptions (2.2), let $p \in H^1_*(\Omega)$ be the solution of (2.4). Let $\mathcal{D}$ be a GD for a Neumann problem in the sense of Definition 2.1.3. Then there exists a unique solution $p_{\mathcal{D}} \in X_{\mathcal{D}}$ solution to the GS (2.6), satisfying the following inequalities:*

$$\|\nabla p - \nabla_{\mathcal{D}} p_{\mathcal{D}}\|_{L^2(\Omega)^d} \le \mathrm{Err}_{\mathcal{D}} + S_{\mathcal{D}}(p) \tag{2.8}$$
$$\|p - \Pi_{\mathcal{D}} p_{\mathcal{D}}\|_{L^2(\Omega)} \le C_{\mathcal{D}} \mathrm{Err}_{\mathcal{D}} + S_{\mathcal{D}}(p), \tag{2.9}$$

where

$$\mathrm{Err}_{\mathcal{D}} := \frac{1}{\min(\underline{\lambda}, 1)} \big[W_{\mathcal{D}}(\Lambda \nabla p) + (\overline{\lambda} + |\Omega|^{\frac{1}{2}} C_{\mathcal{D}}) S_{\mathcal{D}}(p) \big],$$

with $C_{\mathcal{D}}$, $S_{\mathcal{D}}$, and $W_{\mathcal{D}}$ defined as in (2.7).

Proof. We start by proving the inequalities (2.8)–(2.9). Let $p_{\mathcal{D}} \in X_{\mathcal{D}}$ be a solution to (2.6). Take $\boldsymbol{\phi} = \Lambda \nabla p$ in the definition (2.7c) of $W_{\mathcal{D}}$. Using the fact that p is a solution to (2.4), we have $\mathrm{div}(\Lambda \nabla p) = -f$, and thus for any $v \in X_{\mathcal{D}}$,

$$\left| \int_\Omega [\nabla_{\mathcal{D}} v \cdot \Lambda \nabla p - \Pi_{\mathcal{D}} v f] \right| \leq \|v\|_{\mathcal{D}} W_{\mathcal{D}}(\Lambda \nabla p).$$

Using the fact that $p_{\mathcal{D}}$ is a solution to (2.6), we substitute the corresponding expression for $\int_\Omega \Pi_{\mathcal{D}} v f$ into the above inequality in order to obtain

$$\left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla p - \nabla_{\mathcal{D}} p_{\mathcal{D}}) - \int_\Omega \Pi_{\mathcal{D}} p_{\mathcal{D}} \int_\Omega \Pi_{\mathcal{D}} v \right| \leq \|v\|_{\mathcal{D}} W_{\mathcal{D}}(\Lambda \nabla p). \quad (2.10)$$

Now, we introduce the element

$$I_{\mathcal{D}}(p) = \underset{w \in X_{\mathcal{D}}}{\mathrm{argmin}} \left(\|\Pi_{\mathcal{D}} w - p\|_{L^2(\Omega)} + \|\nabla_{\mathcal{D}} w - \nabla p\|_{L^2(\Omega)^d} \right) \in X_{\mathcal{D}}$$

and add the expression

$$\left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla p) + \int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}} p \int_\Omega \Pi_{\mathcal{D}} v \right|$$

onto both sides of (2.10). Applying a triangle inequality on the left hand side, we obtain

$$\begin{aligned} &\left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla_{\mathcal{D}} p_{\mathcal{D}}) + \int_\Omega (\Pi_{\mathcal{D}} I_{\mathcal{D}} p - \Pi_{\mathcal{D}} p_{\mathcal{D}}) \int_\Omega \Pi_{\mathcal{D}} v \right| \\ &\leq \|v\|_{\mathcal{D}} W_{\mathcal{D}}(\Lambda \nabla p) + \left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla p) + \int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}} p \int_\Omega \Pi_{\mathcal{D}} v \right|. \end{aligned} \quad (2.11)$$

Next, since $p \in H^1_*(\Omega)$ we write, by triangle inequality, followed by Cauchy-Schwarz,

$$\begin{aligned} &\left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla p) + \int_\Omega \Pi_{\mathcal{D}} I_{\mathcal{D}} p \int_\Omega \Pi_{\mathcal{D}} v \right| \\ &\leq \left| \int_\Omega \Lambda \nabla_{\mathcal{D}} v \cdot (\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla p) \right| + \left| \int_\Omega (\Pi_{\mathcal{D}} I_{\mathcal{D}} p - p) \int_\Omega \Pi_{\mathcal{D}} v \right| \\ &\leq \|\Lambda \nabla_{\mathcal{D}} v\|_{L^2(\Omega)^d} \|\nabla_{\mathcal{D}} I_{\mathcal{D}} p - \nabla p\|_{L^2(\Omega)^d} + |\Omega|^{\frac{1}{2}} \|\Pi_{\mathcal{D}} I_{\mathcal{D}} p - p\|_{L^2(\Omega)} \|\Pi_{\mathcal{D}} v\|_{L^2(\Omega)} \\ &\leq \|v\|_{\mathcal{D}} S_{\mathcal{D}}(p)(\overline{\lambda} + |\Omega|^{\frac{1}{2}} C_{\mathcal{D}}), \end{aligned}$$

where the last inequality was obtained by using assumption (2.2b), and invoking the definitions (2.7a) and (2.7b) of $C_{\mathcal{D}}$ and $S_{\mathcal{D}}$, respectively. Taking $v = I_{\mathcal{D}} p - p_{\mathcal{D}}$, and substituting the above expression into (2.11), we obtain

$$\min(\underline{\lambda}, 1) \|I_{\mathcal{D}} p - p_{\mathcal{D}}\|_{\mathcal{D}} \leq \|v\|_{\mathcal{D}} W_{\mathcal{D}}(\Lambda \nabla p) + \|v\|_{\mathcal{D}} S_{\mathcal{D}}(p)(\overline{\lambda} + |\Omega|^{\frac{1}{2}} C_{\mathcal{D}}),$$

and so

$$\|I_{\mathcal{D}}p - p_{\mathcal{D}}\|_{\mathcal{D}} \leq \mathrm{Err}_{\mathcal{D}}. \tag{2.12}$$

Inequality (2.8) follows by noting that

$$\|I_{\mathcal{D}}p - p_{\mathcal{D}}\|_{\mathcal{D}} \geq \|\nabla_{\mathcal{D}} I_{\mathcal{D}}p - \nabla_{\mathcal{D}} p_{\mathcal{D}}\|_{L^2(\Omega)^d},$$

adding $\|\nabla p - \nabla_{\mathcal{D}} I_{\mathcal{D}}p\|_{L^2(\Omega)^d}$ onto both sides of (2.12), and applying the triangle inequality and the definition of $S_{\mathcal{D}}$. Next, to establish the inequality (2.9), we apply the definition of $C_{\mathcal{D}}$ to obtain

$$\|\Pi_{\mathcal{D}}(I_{\mathcal{D}}p - p_{\mathcal{D}})\|_{L^2(\Omega)} \leq C_{\mathcal{D}} \|I_{\mathcal{D}}p - p_{\mathcal{D}}\|_{\mathcal{D}}.$$

Using (2.12) and adding $\|p - \Pi_{\mathcal{D}} I_{\mathcal{D}}p\|_{L^2(\Omega)}$ to both sides of the inequality, we have

$$\|\Pi_{\mathcal{D}}(I_{\mathcal{D}}p - p_{\mathcal{D}})\|_{L^2(\Omega)} + \|p - \Pi_{\mathcal{D}} I_{\mathcal{D}}p\|_{L^2(\Omega)} \leq C_{\mathcal{D}}\mathrm{Err}_{\mathcal{D}} + \|p - \Pi_{\mathcal{D}} I_{\mathcal{D}}p\|_{L^2(\Omega)},$$

which, by triangle inequality and definition of $S_{\mathcal{D}}$, leads to

$$\|p - \Pi_{\mathcal{D}} p_{\mathcal{D}}\|_{L^2(\Omega)} \leq C_{\mathcal{D}}\mathrm{Err}_{\mathcal{D}} + S_{\mathcal{D}}(p).$$

Finally, we prove that the solution $p_{\mathcal{D}}$ to (2.6) is unique. This is equivalent to showing that the only solution $p_{\mathcal{D}} \in X_{\mathcal{D}}$ to

$$\int_\Omega \Lambda \nabla_{\mathcal{D}} p_{\mathcal{D}} \cdot \nabla_{\mathcal{D}} v + \int_\Omega \Pi_{\mathcal{D}} p_{\mathcal{D}} \int_\Omega \Pi_{\mathcal{D}} v = 0, \forall v \in X_{\mathcal{D}}$$

is $p_{\mathcal{D}} = 0$. Now, note that by taking $f = 0$, the solution p to (2.4) is $p = 0$. The inequality (2.8) will then imply that $\|\nabla_{\mathcal{D}} p_{\mathcal{D}}\|_{L^2(\Omega)^d} = 0$. Taking $v = p_{\mathcal{D}}$ in the above equation will then yield $\int_\Omega \Pi_{\mathcal{D}} p_{\mathcal{D}} = 0$, which implies that $p_{\mathcal{D}} = 0$ by the definition of the norm $\|\cdot\|_{\mathcal{D}}$. ∎

The following result follows directly from Lemmas 2.1.6 and 2.1.7.

Corollary 2.1.8 (Convergence)**.** *Under assumptions* (2.2)*, let* $(\mathcal{D}_m)_{m\in\mathbb{N}}$ *be a sequence of GDs in the sense of Definition 2.1.1, which is GD-consistent and limit-conforming in the sense of Definition 2.1.3. Then, for any* $m \in \mathbb{N}$*, there exists a unique solution* $p_m \in X_{\mathcal{D}_m}$ *to the GS* (2.6) *and if* p *is the solution of* (2.4) *then, as* $m \to \infty$*,* $\Pi_{\mathcal{D}_m} p_m$ *converges to* p *in* $L^2(\Omega)$ *and* $\nabla_{\mathcal{D}_m} p_m$ *converges to* ∇p *in* $L^2(\Omega)^d$*.*

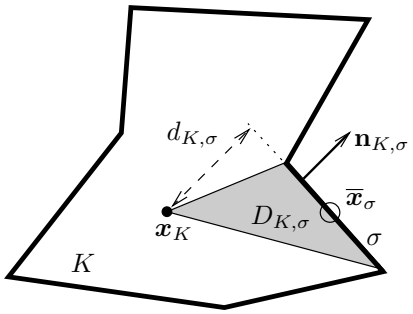
2.2 HMM scheme

For an HMM scheme, the unknowns are given by cell and edge values. The space of unknowns can thus be written as

$$X_{\mathcal{D}} := \{w = ((w_K)_{K\in\mathcal{M}}, (w_\sigma)_{\sigma\in\mathcal{E}_K}) : w_K \in \mathbb{R}, w_\sigma \in \mathbb{R}\}.$$

The reconstructed functions $\Pi_{\mathcal{D}} w$ are piecewise constant on each cell with $(\Pi_{\mathcal{D}} w)|_K = w_K$. A piecewise constant gradient is then defined on a sub-triangulation of cells.

Figure 2.1: Notations in a generic cell in dimension $d = 2$.



Let $\boldsymbol{x}_K \in K$ be a point in cell K such that K is star-shaped with respect to K. That is, $d_{K,\sigma} > 0$ for all $\sigma \in \mathcal{E}_K$, where $d_{K,\sigma}$ is the signed orthogonal distance between $\boldsymbol{x}_K$ and σ (see Figure 2.1). We then set $\overline{\boldsymbol{x}}_\sigma$ to be the centre of mass of σ. Following [41], if $K \in \mathcal{M}$ and $(D_{K,\sigma})_{\sigma\in\mathcal{E}_K}$ is the convex hull of σ and $\boldsymbol{x}_K$ (see Figure 2.1), we set

$$\begin{aligned}
&\forall w \in X_{\mathcal{D}}\,,\ \forall \boldsymbol{x} \in D_{K,\sigma}\,,\\
&\nabla_{\mathcal{D}} w(\mathbf{x}) = \overline{\nabla}_K w + \frac{\sqrt{d}}{d_{K,\sigma}}[w_\sigma - w_K - \overline{\nabla}_K w \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)]\mathbf{n}_{K,\sigma}\,,\\
&\text{where } \overline{\nabla}_K w = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}_K} |\sigma| w_\sigma \mathbf{n}_{K,\sigma}.
\end{aligned} \tag{2.13}$$

Note that $\overline{\nabla}_K w$ is a linearly exact reconstruction of the gradient, that is, if $(w_\sigma)_{\sigma\in\mathcal{E}_K}$ interpolates an affine function A at the edge midpoints, then $\overline{\nabla}_K w = \nabla A$. Owing to the fact that $\sum_{\sigma\in\mathcal{E}_K} |\sigma| w_K \mathbf{n}_{K,\sigma} = 0$ for any constant w_K, we may also write

$$\overline{\nabla}_K w = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}_K} |\sigma|(w_\sigma - w_K)\mathbf{n}_{K,\sigma}. \tag{2.14}$$

In (2.13), the second term can be seen as a stabilisation of the gradient involving a discrete 2nd order Taylor expansion; this term also enforces the coercivity of the discrete bilinear form.

Given $p \in X_{\mathcal{D}}$, the discrete fluxes $(F_{K,\sigma})_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K}$, approximations of $-\int_\sigma \Lambda\nabla p\cdot\mathbf{n}_{K,\sigma}$, are then defined by the relation

$$\begin{aligned}&\forall K\in\mathcal{M}\,,\ \forall v\in X_{\mathcal{D}}\,,\\ &\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}(v_K-v_\sigma)=\int_K \Lambda\nabla_{\mathcal{D}}p(\mathbf{x})\cdot\nabla_{\mathcal{D}}v(\mathbf{x})d\mathbf{x}.\end{aligned}\tag{2.15}$$

As a remark, we note that the fluxes $F_{K,\sigma}$ are uniquely defined. In particular, we can see from (2.13) and (2.14) that $\nabla_{\mathcal{D}}v$ is uniquely determined by the value of $(v_\sigma - v_K)$. Hence, for a given edge $\sigma\in\mathcal{E}_K$, $F_{K,\sigma}$ can be uniquely determined by setting, for example, $v_\sigma - v_K = 1$ and $v_{\sigma'} - v_K = 0$ for all the other edges $\sigma'\in\mathcal{E}_K$.

We now write the gradient scheme (2.6) using the HMM scheme, in its finite volume form. Let $K\in\mathcal{M}$. By taking $v\in X_{\mathcal{D}}$ such that $v_K = 1$ and 0 elsewhere, we obtain the balance of fluxes

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} + |K|\sum_{M\in\mathcal{M}}|M|p_M = \int_K f \quad \text{for all } K\in\mathcal{M}.\tag{2.16a}$$

Now, for $\sigma\in\mathcal{E}$, take $v\in X_{\mathcal{D}}$ such that $v_\sigma = 1$, and 0 for all other components. This gives us the conservativity of fluxes

$$\begin{aligned}F_{K,\sigma}+F_{L,\sigma}&=0 \quad \text{for all edges } \sigma \text{ between two different cells } K \text{ and } L,\\ F_{K,\sigma}&=0 \quad \text{for all edges } \sigma \text{ of } K \text{ lying on } \partial\Omega.\end{aligned}\tag{2.16b}$$

The balance (2.16a) and conservativity (2.16b) of fluxes are fundamental in the formulation of finite volume schemes [39]. With definitions (2.13)–(2.15), system (2.16) provides an approximation $p\in X_{\mathcal{D}}$ of p, as well as approximate fluxes $(F_{K,\sigma})_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K}$.

Remark 2.2.1. *We note here that for the HMM, by selecting $v\in X_{\mathcal{D}}$ such that $v_K = v_\sigma = 1$ for all $\sigma\in\mathcal{E}_K$, we are in the context of Remark 2.1.2. Hence, as an alternative, we may write, for* (2.16a)*, $\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = \int_K f$ for all $K\in\mathcal{M}$, and $\sum_{M\in\mathcal{M}}|M|p_M = 0$.*

2.3 HHO scheme

Following [29], where the HHO scheme, together with other arbitrary order schemes have been presented in the framework of gradient schemes, the

discrete unknowns are written, for $l \geq 0$ a polynomial degree,

$$\underline{U}^l := \big(\bigtimes_{K\in\mathcal{M}} \mathbb{P}^l(K)\big) \times \big(\bigtimes_{\sigma\in\mathcal{E}} \mathbb{P}^l(\sigma)\big).$$

Here, $\mathbb{P}^l$ denotes the polynomial space in d and $d-1$ variables, on K and σ respectively, with degree less than or equal to l. As with the HMM, the unknowns are given by cell and edge values. Unlike the HMM, whose unknowns are constant values on cells and edges, the unknowns for HHO are polynomials on cells and edges. Hence, for $\underline{w} \in \underline{U}^l$, we write

$$\underline{w} = ((w_K)_{K\in\mathcal{M}}, (w_\sigma)_{\sigma\in\mathcal{E}}).$$

For all $\underline{w} \in \underline{U}^l$, we also define w, a broken polynomial field, such that

$$w_{|K} = w_K \qquad \forall K \in \mathcal{M}.$$

The restriction of $\underline{U}^l$ to a cell K is written $\underline{U}^l_K$. Since our approximations are piecewise polynomials, we project generic functions into these polynomial spaces locally through L^2 orthogonal projectors $\Pi^l_Z : L^1(Z) \to \mathbb{P}^l(Z)$, defined by: For all $v \in L^1(Z)$, $\Pi^l_Z v$ is the unique polynomial in $\mathbb{P}^l(Z)$ such that

$$\int_Z (\Pi^l_Z v - v)w = 0 \qquad \forall w \in \mathbb{P}^l(Z). \tag{2.17}$$

Here, Z is either a cell or a face. The space of discrete unknowns and the reconstruction of the function for the HHO scheme is then given by

$$X_\mathcal{D} := \underline{U}^l \quad \text{and} \quad \Pi_\mathcal{D}\underline{w} := w \text{ for all } \underline{w} \in \underline{U}^l.$$

Finally, we describe the gradient reconstruction $\nabla_\mathcal{D}$. We start with a high order reconstruction (maps to a polynomial of degree $l+1$ instead of l) $r^{l+1}_K : \underline{U}^l_K \to \mathbb{P}^{l+1}(K)$ such that for all $\underline{w}_K \in \underline{U}^l_K$, $r^{l+1}_K \underline{w}_K$ satisfies, for all $v \in \mathbb{P}^{l+1}(K)$

$$\int_K \Lambda\nabla r^{l+1}_K \underline{w}_K \cdot \nabla v = \int_K \Lambda\nabla w_K \cdot \nabla v + \sum_{\sigma\in\mathcal{E}_K} \int_\sigma (w_\sigma - w_K)\nabla v \cdot (\Lambda \mathbf{n}_{K,\sigma}). \tag{2.18}$$

Equation (2.18) was inspired by the following integration by parts formula, valid for any $u \in W^{1,1}(K), \boldsymbol{\phi} \in C^\infty(\bar{K})^d$ [29]:

$$\int_K \Lambda\nabla u \cdot \boldsymbol{\phi} = -\int_K u\mathrm{div}(\Lambda\boldsymbol{\phi}) + \sum_{\sigma\in\mathcal{E}_K} \int_\sigma u(\Lambda\boldsymbol{\phi})\cdot\mathbf{n}_{K,\sigma}. \tag{2.19}$$

In particular, setting w_K and w_σ to be the projection of u into $\mathbb{P}^l(K)$ and $\mathbb{P}^l(\sigma)$, respectively, and taking $\nabla r_K^{l+1}\underline{w}_K$ to be the projection of ∇u on $\mathbb{P}^l(K)^d$, (2.18) is obtained by taking $\boldsymbol{\phi} = \nabla v$ in (2.19), where $v \in \mathbb{P}^{l+1}(K)$. Here, the projections are defined as in (2.17). Now, (2.18) defines $r_K^{l+1}\underline{w}_K$ up to an additive constant. To fully determine $r_K^{l+1}\underline{w}_K$, we further impose

$$\int_K r_K^{l+1}\underline{w}_K = \int_K w_K.$$

We now introduce the difference operators $\delta_K^l : \underline{U}_K^l \to \mathbb{P}^l(K)$ for all $K \in \mathcal{M}$ and $\delta_{K,\sigma}^l : \underline{U}_K^l \to \mathbb{P}^l(\sigma)$ for all $\sigma \in \mathcal{E}_K$, defined to be , for $\underline{w}_K \in \underline{U}_K^l$,

$$\delta_K^l\underline{w}_K := \Pi_K^l(r_K^{l+1}\underline{w}_K - w_K) \qquad \delta_{K,\sigma}^l\underline{w}_K := \Pi_\sigma^l(r_K^{l+1}\underline{w}_K - w_\sigma) \quad \forall \sigma \in \mathcal{E}_K. \tag{2.20}$$

As with the HMM, the reconstructed gradient consists of a consistent and limit conforming part, ∇r_K^{l+1} described above, accompanied by a stabilising contribution. Hence, the discrete gradient $\nabla_{\mathcal{D}} : \underline{U}^l \to L^2(\Omega)^d$ is built such that:

$$(\nabla_{\mathcal{D}}\underline{w})_{|K} = \nabla r_K^{l+1}\underline{w}_K + \mathbf{S}_K\underline{w}_K \qquad \forall \underline{w} \in \underline{U}^l, \quad \forall K \in \mathcal{M}. \tag{2.21}$$

In particular, the stabilising contribution $\mathbf{S}_K : \underline{U}_K^l \to L^2(K)^d$ should satisfy the following conditions:

- L^2 stability and boundedness: For all $K \in \mathcal{M}$ and all $\underline{w}_K \in \underline{U}_K^l$, it holds that
$$\|\mathbf{S}_K\underline{w}_K\|_{L^2(K)^d} \simeq |\underline{w}_K|_{2,\partial K}. \tag{2.22}$$
Here, $a \simeq b$ means that there is a real number $C > 0$ independent of h and K, but possibly depending on d and on the other discretisation parameters, such that $Ca \le b \le C^{-1}a$. We also set
$$|\underline{w}_K|_{2,\partial K} := \sum_{\sigma\in\mathcal{E}_K} \frac{1}{|\sigma|} \left\|(\delta_{K,\sigma}^l - \delta_K^l)\underline{w}_K\right\|_{L^2(\sigma)}^2.$$

- Orthogonality: For all $\underline{w}_K \in \underline{U}_K^l$ and all $\boldsymbol{\phi} \in \mathbb{P}^l(K)^d$, it holds that
$$(\mathbf{S}_K\underline{w}_K, \boldsymbol{\phi})_K = 0. \tag{2.23}$$

Under the assumption that Λ is piecewise constant in each cell K, and owing to the orthogonality property (2.23) of the stabilisation term $\mathbf{S}_K$, we have

$$\begin{aligned}\int_K \Lambda\nabla_{\mathcal{D}}v \cdot \nabla_{\mathcal{D}}w &= \int_K \Lambda\nabla r_K^{l+1}\underline{v}_K \cdot \nabla r_K^{l+1}\underline{w}_K + \int_K \Lambda\mathbf{S}_K\underline{v}_K \cdot \mathbf{S}_K\underline{w}_K \\ &:= a_K(\underline{v}_K, \underline{w}_K).\end{aligned}$$

To determine $a_K(\underline{v}_K, \underline{w}_K)$ completely, we are left to define the bilinear form $s_K(\underline{v}_K, \underline{w}_K) = \int_K \Lambda \mathbf{S}_K \underline{v}_K \cdot \mathbf{S}_K \underline{w}_K$, which is defined as

$$s_K(\underline{v}_K, \underline{w}_K) := \sum_{\sigma \in \mathcal{E}_K} \frac{\Lambda_{K,\sigma}}{|\sigma|} \int_\sigma (\delta^l_{K,\sigma} - \delta^l_K)\underline{v}_K (\delta^l_{K,\sigma} - \delta^l_K)\underline{w}_K, \tag{2.24}$$

where $\Lambda_{K,\sigma} = \left\| \mathbf{n}_{K,\sigma} \cdot \Lambda_{|K} \mathbf{n}_{K,\sigma} \right\|_{L^\infty(\sigma)}$. Computing $s_K(\underline{w}_K, \underline{w}_K)$ in (2.24) shows that the stabilisation term indeed satisfies the stability and boundedness property (2.22). An explicit construction of $\mathbf{S}_K$, built from a lifting of face-based differences, can be seen in [29, Section 3.6.3]. We now present the HHO scheme for the diffusion problem (2.1), obtained using $(X_\mathcal{D}, \Pi_\mathcal{D}, \nabla_\mathcal{D})$ above in (2.6): Find $\underline{p} \in \underline{U}^l$ such that

$$\sum_{K \in \mathcal{M}} a_K(\underline{p}_K, \underline{w}_K) = \int_\Omega f \underline{w} \quad \forall \underline{w} \in \underline{U}^l. \tag{2.25}$$

As with the HMM, the HHO falls under Remark 2.1.2, and hence (2.6) also implies that $\int_\Omega \Pi_\mathcal{D} p = 0$. The advantage of (2.25) is that the system has local stencils, and static condensation may be employed to solve the system in a more efficient manner [25].

Remark 2.3.1. *The HHO scheme with $l = 0$ is equivalent to the HMM scheme [30, Proposition 7].*

Owing to [25, Proposition 3.1], we may define the fluxes obtained from the HHO in the following manner.

Definition 2.3.2. *Let $K \in \mathcal{M}$ and $\underline{p} \in \underline{U}^l$ be the solution to* (2.25)*. We define the discrete fluxes $G_{K,\sigma}$ so that for any $\underline{w}_K \in \underline{U}_K$,they satisfy*

$$\sum_{\sigma \in \mathcal{E}_K} \int_\sigma G_{K,\sigma}(w_K - w_\sigma) = \sum_{\sigma \in \mathcal{E}_K} \int_\sigma \left(-\Lambda \nabla r^{l+1}_K \underline{p}_K \cdot \mathbf{n}_{K,\sigma}\right)(w_K - w_\sigma) + s_K(\underline{p}_K, \underline{w}_K).$$

Remark 2.3.3. *The quantity $G_{K,\sigma}$ gives an approximation to $-\Lambda \nabla p \cdot \mathbf{n}_{K,\sigma}$. Here, we can see that $G_{K,\sigma}$ is made up of 2 components, the first component of which is obtained by simply using the solution $\underline{p}$ to the discrete problem, added to a second component, which can be viewed as a stabilisation term.*

Remark 2.3.4. *By writing $F_{K,\sigma} = \int_\sigma G_{K,\sigma} \quad \forall \sigma \in \mathcal{E}_K$, it can be checked that $(F_{K,\sigma})_{\sigma_\in \mathcal{E}_K}$ satisfy the balance and conservation of fluxes* (2.16)*. In particular,* (2.16a) *is satisfied by taking $w_K = 1$ and $w_\sigma = 0$ for all $\sigma \in \mathcal{E}_K$ in Definition 2.3.2, and* (2.16b) *is satisfied by taking $w_\sigma = -1$ for an edge $\sigma \in \mathcal{E}_K$, $w_{\sigma'} = 0$ for all other edges $\sigma' \in \mathcal{E}_K$, and $w_K = 0$ in Definition 2.3.2.*

2.4 Numerical tests

Numerical tests will be performed on the unit square $\Omega = (0,1) \times (0,1)$, with the following types of mesh discretisations: Cartesian meshes, hexahedral meshes, and Kershaw meshes (see Figures 1.1 and 1.2).

The numerical tests will be performed on the following diffusion tensors:

- Test case 1: $\Lambda = \mathbf{I}$, where $\mathbf{I}$ is the identity matrix
- Test case 2: mild anisotropy $\Lambda = \begin{bmatrix} 1.5 & 0.5 \\ 0.5 & 1.5 \end{bmatrix}$,
- Test case 3: strong anisotropy $\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-6} \end{bmatrix}$,

with the prescribed solution $p = \cos(\pi x)\cos(\pi y)$. The first and third test cases with diagonal Λ will give us homogeneous Neumann boundary conditions whereas the second test case will give us non-homogeneous Neumann boundary conditions. Here, we test the accuracy of the approximations to the function and the fluxes, $\|\Pi_{\mathcal{D}} p - p\|_{L^2(\Omega)}$ and $\max_{K \in \mathcal{M},\, \sigma \in \mathcal{E}_K} \frac{1}{|\sigma|} \left| F_{K,\sigma} + \int_\sigma \Lambda \nabla p \cdot \mathbf{n}_{K,\sigma} \right|$ respectively. The discretisations are performed via HMM for $k = 0$ and HHO for $k > 0$. Of course, in order to improve the accuracy of the numerical approximations, one may opt to use mesh refinement. However, in practical applications for the complete coupled model (1.1), we would desire good approximations even on coarse meshes, which is the reason why we consider high order schemes. Although the numerical tests are performed only in dimension $d = 2$, convergence of gradient schemes satisfying the properties in Definition 2.1.3 have been established even for dimension $d = 3$.

Table 2.1: $\|\Pi_{\mathcal{D}} p - p\|_{L^2(\Omega)}$, Test case 1

Mesh \ Order	$k = 0$	$k = 1$	$k = 2$	$k = 3$
Cartesian	1.6094e-03	9.6153e-05	4.451e-06	1.2307e-07
hexahedral	1.7540e-02	4.1209e-05	1.5395e-06	6.0908e-08
Kershaw	7.2798e-03	4.6917e-04	5.4068e-05	1.5072e-06

First, upon comparing Tables 2.1, 2.3 and 2.5, we note that the errors in the reconstruction of the function for the third test case are much larger than those of the first and second test cases. This is expected due to the strong anisotropy, and agrees with the error bound presented in [31, Theorem 3.18]. In particular, we see here the dependence of the error on the square root of

Table 2.2: $\max_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K} \frac{1}{|\sigma|}\left|F_{K,\sigma} + \int_\sigma \Lambda\nabla p\cdot\mathbf{n}_{K,\sigma}\right|$, Test case 1

Mesh \ Order	$k=0$	$k=1$	$k=2$	$k=3$
Cartesian	5.0279e-03	2.1636e-07	8.0998e-12	2.0008e-11
hexahedral	1.1830e-01	2.9876e-03	3.2947e-05	1.2288e-06
Kershaw	2.5388e-01	3.0662e-02	1.8582e-03	4.9996e-05

Table 2.3: $\|\Pi_{\mathcal{D}}p - p\|_{L^2(\Omega)}$, Test case 2

Mesh \ Order	$k=0$	$k=1$	$k=2$	$k=3$
Cartesian	6.7099e-03	6.8497e-04	9.7495e-06	2.4915e-07
hexahedral	3.9076e-02	2.4989e-04	8.8879e-05	3.4669e-05
Kershaw	4.2337e-02	8.1298e-03	1.0283e-03	9.1672e-05

Table 2.4: $\max_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K} \frac{1}{|\sigma|}\left|F_{K,\sigma} + \int_\sigma \Lambda\nabla p\cdot\mathbf{n}_{K,\sigma}\right|$, Test case 2

Mesh \ Order	$k=0$	$k=1$	$k=2$	$k=3$
Cartesian	7.3073e-03	4.4281e-02	1.8217e-04	2.6709e-06
hexahedral	1.4947e-01	5.7095e-02	1.1017e-02	6.0377e-03
Kershaw	3.9801e-01	3.1263e-01	9.4187e-02	6.1661e-03

Table 2.5: $\|\Pi_{\mathcal{D}}p - p\|_{L^2(\Omega)}$, Test case 3

Mesh \ Order	$k=0$	$k=1$	$k=2$	$k=3$
Cartesian	1.6094e-03	4.7203e-04	3.6427e-05	4.4400e-07
hexahedral	6.4631e+02	1.1442	5.2313e-03	2.6065e-05
Kershaw	3.1011e+03	3.0009	2.3103e-01	1.0343e-02

Table 2.6: $\max_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K} \frac{1}{|\sigma|}\left|F_{K,\sigma} + \int_\sigma \Lambda\nabla p\cdot\mathbf{n}_{K,\sigma}\right|$, Test case 3

Mesh \ Order	$k=0$	$k=1$	$k=2$	$k=3$
Cartesian	5.0279e-03	4.4349e-05	2.8073e-05	2.8043e-05
hexahedral	4.3213	8.1994e-03	9.0264e-05	2.4202e-05
Kershaw	6.1570e-01	2.8906e-02	9.1089e-04	3.4813e-05

the ratio between the largest and smallest eigenvalues of Λ, $\sqrt{\frac{\lambda_M}{\lambda_m}}$, which has a value of 10^3 in the third test case. It is also notable that for Cartesian type meshes, the errors in the function reconstruction remain at the same magnitude as the first and second test cases, which seems to indicate that strong anisotropy does not affect the numerical solutions on the Cartesian type meshes.

Now, we look at the errors in the approximation of the fluxes. We see from Tables 2.2, 2.4 and 2.6 that the fluxes computed using HMM for Kershaw type meshes are much worse than those obtained using HMM for Cartesian meshes. It is only for $k \geq 2$ in the first and third test cases, and even $k \geq 3$ in the second test case, that the fluxes from Kershaw type meshes are comparable to those obtained from HMM for Cartesian meshes. This tells us that, if our numerical scheme for the coupled model (1.1) has strong dependence on the accuracy of the fluxes, a high order scheme for the pressure equation (1.1a) is recommended.

Chapter 3

Velocity reconstructions

The HMM and the HHO yield piecewise constant and piecewise polynomial approximations, respectively, of the pressure p. However, for characteristic-based schemes such as the ELLAM and MMOC, we would need to solve a characteristic equation of the following form:

$$\frac{dF_t(\boldsymbol{x})}{dt} = \frac{\mathbf{u}(F_t(\boldsymbol{x}))}{\phi(F_t(\boldsymbol{x}))} \quad \text{for } t \in [-T, T], \qquad F_0(\boldsymbol{x}) = \boldsymbol{x}. \tag{3.1}$$

Two important features of the velocity need to be accounted for: the no-flow boundary conditions $\mathbf{u} \cdot \mathbf{n} = 0$ on $\partial\Omega$, which ensures that the solutions to (3.1) do not exit the computational domain, and the preservation of the divergence in (1.1a), to avoid creating regions with artificial wells or sinks (which lead to non-physical flows).

These are not satisfied when $\mathbf{u}$ is obtained from the piecewise constant and polynomial approximations of the pressure p from the HMM and HHO schemes, respectively. The general idea is to use the fluxes obtained from these schemes to reconstruct elements in $H(\mathrm{div}, \Omega)$. One of the most common types of $H(\mathrm{div}, \Omega)$ elements are the $\mathbb{RT}_k$ finite elements on simplices and on quadrilaterals [13, 16]. Considering a mesh $\mathcal{M}$ which consists of simplices or quadrilaterals, a global interpolant $\mathcal{I}^{\mathrm{glob}}_{\mathbb{RT}} : H(\mathrm{div}, \Omega) \cap \Pi_{K\in\mathcal{M}} H^1(K)^d \to \mathbb{RT}_k$ is defined such that for any $\mathbf{u} \in H(\mathrm{div}, \Omega) \cap \Pi_{K\in\mathcal{M}} H^1(K)^d$, $(\mathcal{I}^{\mathrm{glob}}_{\mathbb{RT}}\mathbf{u})_{|K} = \mathcal{I}_{\mathbb{RT}_k(K)}(\mathbf{u})_{|K}$, where $\mathcal{I}_{\mathbb{RT}_k(K)}$ is a local interpolation operator from the space $H^1(K)^d$ to $\mathbb{RT}_k(K)$. More details about the local interpolation operator $\mathcal{I}_{\mathbb{RT}_k(K)}$ for simplices and quadrilaterals will be given in Sections 3.1 and 3.2, respectively.

3.1 $\mathbb{RT}_k$ elements on simplices

Over a simplex K of dimension d (triangle for $d = 2$, tetrahedron for $d = 3$), an $\mathbb{RT}_k$ finite element is defined to be

$$\mathbb{RT}_k(K) := (\mathbb{P}^k(K))^d + \boldsymbol{x}\mathbb{P}^k(K).$$

This space has dimension $d\binom{k+d}{k} + \binom{k+d-1}{k}$. In particular, given $\mathbf{u} \in H^1(K)^d$, an interpolant $\mathcal{I}_{\mathbb{RT}_k(K)} : H^1(K)^d \to \mathbb{RT}_k(K)$ can be uniquely determined in the following way.

Lemma 3.1.1 (Existence and Uniqueness)**.** *Given $\mathbf{u} \in H^1(K)^d$, there exists a unique $\mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \in \mathbb{RT}_k(K)$ satisfying the equations:*

$$\int_\sigma \mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \cdot \mathbf{n}_{K,\sigma} p_\sigma = \int_\sigma \mathbf{u} \cdot \mathbf{n}_{K,\sigma} p_\sigma \text{ for all } p_\sigma \in \mathbb{P}^k(\sigma) \text{ and } \sigma \in \mathcal{E}_K, \quad (3.2a)$$

and for $k \geq 1$,

$$\int_K \mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \cdot \mathbf{p}_K = \int_K \mathbf{u} \cdot \mathbf{p}_K \text{ for all } \mathbf{p}_K \in (\mathbb{P}^{k-1}(K))^d \ . \quad (3.2b)$$

Proof. Since we are only working locally on cell K, we write $\mathcal{I}_{\mathbb{RT}_k}$ in lieu of $\mathcal{I}_{\mathbb{RT}_k(K)}$ for legibility. We start by showing that the number of equations is equal to the dimension of the space $\mathbb{RT}_k$. We look first at (3.2a), and note that $\mathbb{P}^k(\sigma)$ is the space of $(d-1)$–variable polynomials with degree at most k, which has a dimension of $\binom{k+d-1}{k}$. Moreover, each simplex has $d+1$ faces, and hence (3.2a) gives $(d+1)\binom{k+d-1}{k}$ equations. Now, $\mathbb{P}^{k-1}(K)$ is the space of d–variable polynomials with degree at most $k-1$, which has dimension $\binom{k+d-1}{k-1}$. Since (3.2b) holds for all $\mathbf{p}_K \in (\mathbb{P}^{k-1}(K))^d$, it consists of $d\binom{k+d-1}{k-1}$ equations. The total number of equations is then

$$d\binom{k+d-1}{k-1} + (d+1)\binom{k+d-1}{k} = d\binom{k+d}{k} + \binom{k+d-1}{k},$$

which is the same as the dimension of $\mathbb{RT}_k(K)$, or the number of unknowns. In order to show the existence of $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}$, we can show uniqueness instead. Hence, in the next step, we show that the solution to the system of equations

$$\int_\sigma \mathcal{I}_{\mathbb{RT}_k}\mathbf{u} \cdot \mathbf{n}_{K,\sigma} p_\sigma = 0 \text{ for all } p_\sigma \in \mathbb{P}^k(\sigma) \text{ and } \sigma \in \mathcal{E}_K, \quad (3.3a)$$

and for $k \geq 1$,

$$\int_K \mathcal{I}_{\mathbb{RT}_k}\mathbf{u} \cdot \mathbf{p}_K = 0 \text{ for all } \mathbf{p}_K \in (\mathbb{P}^{k-1}(K))^d \ , \quad (3.3b)$$

is $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u} = 0$. First we note that $\boldsymbol{x}\cdot\mathbf{n}_{K,\sigma}$ is constant on σ, which implies that $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma} \in \mathbb{P}^k(\sigma)$; hence, by (3.3a), we obtain $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma} = 0$ on σ, for all $\sigma \in \mathcal{E}_K$. We then have

$$\int_K (\mathrm{div}\mathcal{I}_{\mathbb{RT}_k}\mathbf{u})^2 = -\int_K \mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\nabla(\mathrm{div}\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}) = 0,$$

since $\nabla(\mathrm{div}\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}) \in (\mathbb{P}^{k-1}(K))^d$. We can then deduce that $\mathrm{div}\mathcal{I}_{\mathbb{RT}_k}\mathbf{u} = 0$ in K. Taking note that $\mathrm{div}(\mathbb{P}^k(K)^d) \subset \mathbb{P}^{k-1}(K)$ and $\mathrm{div}(\boldsymbol{x}\mathbb{P}^k(K)) \subset \mathbb{P}^k(K)$, the fact that $\mathrm{div}\mathcal{I}_{\mathbb{RT}_k}\mathbf{u} = 0$ in K would then imply that $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u} \in \mathbb{P}^k(K)^d$. Now, since $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma}$ is a polynomial of degree k in K that vanishes for all $\boldsymbol{x} \in \sigma$, we may write $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma} = \ell_\sigma q_{k-1}$, where $\ell_\sigma = 0$ on σ and $q_{k-1} \in \mathbb{P}^{k-1}(K)$. We then use (3.3b) to find that

$$\int_K \mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma}p_K = 0 \text{ for all } p_K \in \mathbb{P}^{k-1}(K).$$

Taking $p_K = q_{k-1}$, we obtain

$$\int_K \ell_\sigma q_{k-1}^2 = 0.$$

This implies that $q_{k-1} = 0$ and thus $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}\cdot\mathbf{n}_{K,\sigma} = 0$ on K, for all $\sigma \in \mathcal{E}_K$. Since $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u}$ vanishes in $d+1$ linearly independent directions, we may then conclude $\mathcal{I}_{\mathbb{RT}_k}\mathbf{u} = 0$. ∎

Owing to Lemma 3.1.1, we see that we may use, for the degrees of freedom of $\mathbb{RT}_k$:

- the moments of up to order k of $\mathbf{u}\cdot\mathbf{n}_{K,\sigma}$ on the sides or faces of K;
- the moments of up to order $k-1$ of $\mathbf{u}$ on K.

The degrees of freedom used for $\mathbb{RT}_0$ and $\mathbb{RT}_1$ on triangles are illustrated in Figure 3.1.

In general, the cells K in the mesh are not simplices, and hence a sub-triangulation of cells must be performed (see Figure 3.2 for 2D and Figure 3.4 for 3D).

For the HMM, we start with an approximation for the fluxes $F_{K,\sigma} = \int_\sigma \mathbf{u}\cdot\mathbf{n}_{K,\sigma}$ along each face (edge in 2D) $\sigma \in \mathcal{E}_K$. These will be used to reconstruct the velocity $\mathbf{u}$ via $\mathbb{RT}_0$ finite elements on a sub-triangulation of the cell K, which requires computation of sub-fluxes along interior faces (edges in 2D).

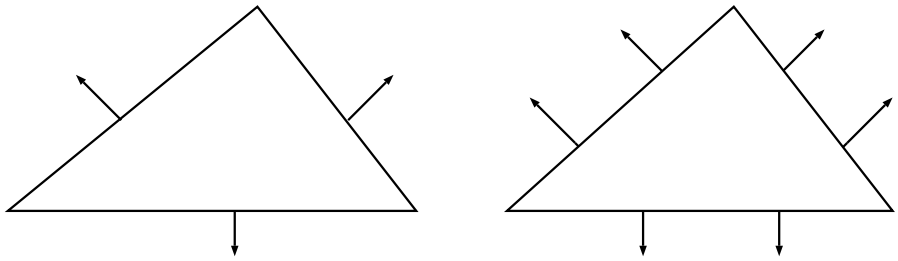


Figure 3.1: DOFs of $\mathbb{RT}_k$ finite elements on triangles (left: $\mathbb{RT}_0$, right: $\mathbb{RT}_1$)

3.1.1 Formulation of the problem

Each cell $K \in \mathcal{M}$ is divided into simplices (triangles in 2D, tetrahedra in 3D), gathered in a set $\mathcal{S}_K$, that share $\boldsymbol{x}_K$ as apex and whose bases are faces or subsets of the faces of K (see Figures 3.2 and 3.4). We then denote by $\mathcal{E}_K^*$ the set of internal faces of K, that is, all of the faces of the simplices $S \in \mathcal{S}_K$, that do not lie on $\sigma \in \mathcal{E}_K$. For every simplex $S \in \mathcal{S}_K$, we denote by σ_S the face of K on which it sits, $\widetilde{\sigma}_S$ the part of σ_S it occupies, and $\mathcal{E}_S^*$ its internal faces (that is, all its faces except σ_S). For every internal face $\sigma^* \in \mathcal{E}_S^*$ that lies on a simplex S, we denote by S' the simplex which shares σ^* with S. We then impose the conservativity of the fluxes

$$\forall \sigma^* \in \mathcal{E}_S^* \,,\; F_{S,\sigma^*} + F_{S',\sigma^*} = 0 \tag{3.4}$$

and the balance (so that the divergence of these fluxes in each simplex is equal to the divergence of the fluxes on K):

$$\forall S \in \mathcal{S}_K \,,\; \sum_{\sigma^* \in \mathcal{E}_S^*} F_{S,\sigma^*} + \frac{|\widetilde{\sigma}_S|}{|\sigma_S|} F_{K,\sigma_S} = \frac{|S|}{|K|} \sum_{\sigma' \in \mathcal{E}_K} F_{K,\sigma'}. \tag{3.5}$$

The second term in the left hand side is the contribution of the external face $\widetilde{\sigma}_S$ of S, on which we assume that the flux is the corresponding proportion of the flux $F_{K,\sigma}$. We note however that the system (3.5) is rank-deficient. This can be seen by taking the sum over all $S \in \mathcal{S}_K$ in (3.5), which, in view of (3.4), leads to the trivial relation

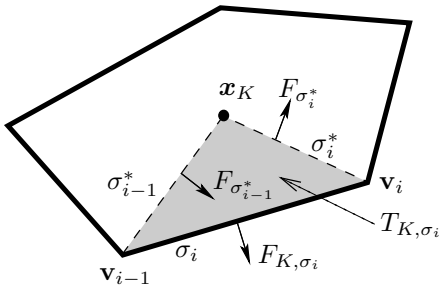
$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}.$$

Techniques to deal with this rank deficiency will be discussed in detail in both 2D and 3D.

3.1.1.1 Implementation in 2D

For the 2 dimensional case, a triangulation of each cell K is performed by choosing a point $\boldsymbol{x}_K$ in the interior of K, and forming a triangle with apex $\boldsymbol{x}_K$ and base σ for each edge $\sigma \in \mathcal{E}_K$. As can be seen in Figure 3.2, for each triangle T_{k,σ_i}, only one of the fluxes (in particular F_{K,σ_i}) is known. An oriented interior flux F_{σ^*} needs to be computed on each internal edge created by this subdivision (as in Figure 3.2). In order to satisfy the conservativity of fluxes (3.4), the idea is to look for one flux F_{σ^*} on each internal edge σ^*. The orientation of the flux F_{σ^*} with respect $T_{K,\sigma}$ is then indicated by $s^{\sigma}_{\sigma^*}$.

Figure 3.2: Triangulation of a generic cell. Here, $s^{\sigma}_{\sigma^*} = +1$ and $s^{\sigma}_{\sigma^{**}} = -1$.



The translation of (3.5) then reads:

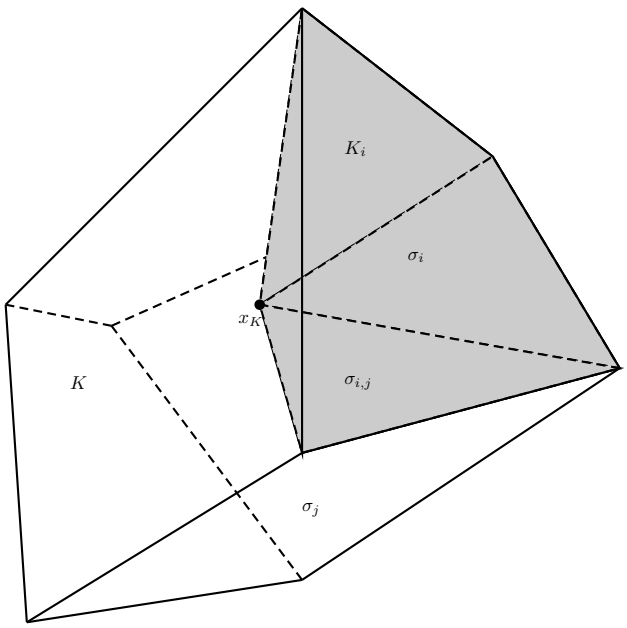
$$\forall \sigma \in \mathcal{E}_K\,, \quad \frac{1}{|T_{K,\sigma}|}\left(\sum_{\sigma^* \in \mathcal{E}^*_{K,\sigma}} s^{\sigma}_{\sigma^*} F_{\sigma^*} + F_{K,\sigma}\right) = \frac{1}{|K|}\sum_{\sigma' \in \mathcal{E}_K} F_{K,\sigma'}, \tag{3.6}$$

where $\mathcal{E}^*_{K,\sigma}$ is the set of edges of the triangle $T_{K,\sigma}$ which are in the interior of K, and $s^{\sigma}_{\sigma^*} = 1$ if F_{σ^*} is oriented outside $T_{K,\sigma}$ and -1 otherwise. Then, $\mathbf{u}$ is the $\mathbb{RT}_0$ function reconstructed from these fluxes on the triangular subdivision. This function belongs to $H(\mathrm{div}, \Omega)$ and by (3.6), this reconstruction is divergence-preserving. Using the notation n_e for the number of edges in cell K, we see that (3.6) gives us n_e equations in n_e unknowns. As seen in Section 3.1.1, the local system of equations is underdetermined. More specifically, its rank is $n_e - 1$. There are several methods to resolve this. Here, we illustrate three methods (see Sections 3.1.2, 3.1.3, and 3.1.4), and compare the numerically reconstructed $\mathbb{RT}_0$ velocities generated by each of the methods by running tests in 2D.

3.1.1.2 Implementation in 3D

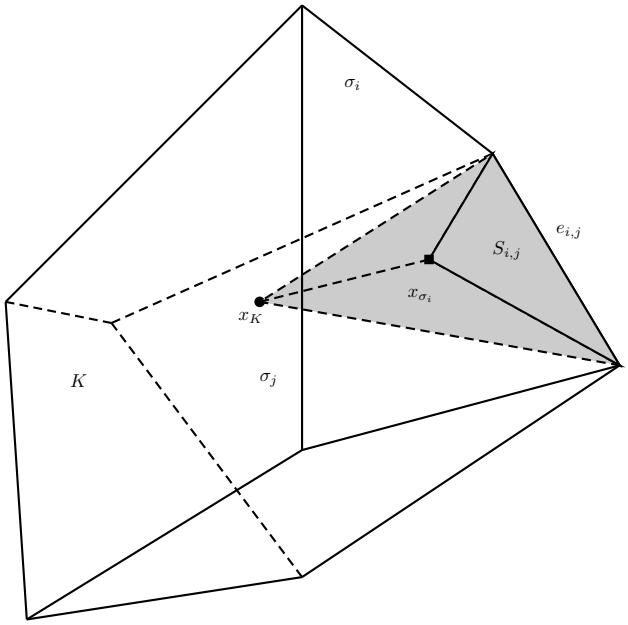
We only describe the process of sub-dividing a cell into simplices for the 3 dimensional case, since the translation of the flux conservativity (3.4) and flux balance (3.5) can be performed in a manner similar to that in the 2 dimensional case. A cell K is subdivided into simplices by first picking a point $\boldsymbol{x}_K$ in the interior of K and forming, for $i = 1, 2, \ldots, n_f$, sub-cells with vertices $\boldsymbol{x}_K, v_{\sigma_{i,k}}$ for $k = 1, 2, \ldots, (n_v)_i$, where n_f and $(n_v)_i$ denote the number of faces of the cell K and the number of vertices in the face σ_i of cell K, respectively. We then denote the sub-cell in cell K associated to face σ_i as K_i (see Figure 3.3). This results to n_f polyhedra, since each face corresponds to one sub-cell. Since we work locally on a cell K, we drop the subscript K for legibility.

Figure 3.3: Division of a polyhedron into sub-cells



A simplex $S_{i,j}$ is then constructed by joining the point $\boldsymbol{x}_K$ to a triangular base formed by joining the edge $e_{i,j}$ being shared by the faces σ_i and σ_j, to a point $\boldsymbol{x}_{\sigma_i}$ on the face σ_i (see Figure 3.4).

Figure 3.4: Triangulation of a polyhedron



3.1.2 Minimal l^2 norm (KR method)

Firstly, we may take the solution to (3.6) with minimal l^2 norm, as in [64]. The velocity field reconstructed from these fluxes will be referred to as *KR velocities* ('KR' since this method of computing internal fluxes is attributed to Y. Kuznetsov and S. Repin). In this case, extension into 3D is quite simple, as the least norm solution of a system of linear equations can easily be computed. However, if for example, we have a constant velocity field $\mathbf{u}$, with fluxes $F_{K,\sigma} = |\sigma|\mathbf{u}\cdot\mathbf{n}_{K,\sigma}$, then the internal fluxes F_{K,σ^*} obtained from the least norm solution might not correspond to the exact value $|\sigma^*|\mathbf{u}\cdot\mathbf{n}_{T_{K,\sigma},\sigma^*}$. An incorrect approximation of these internal fluxes will lead to an incorrect reconstruction of the velocity field. Actually, given the velocity field $\mathbf{u} = (0,1)$, the numerical tests in Section 3.3.1 show that the KR velocity deviates from $\mathbf{u}$, especially over nonstandard meshes.

3.1.3 Consistency condition (C method)

In the two dimensional case, since the system (3.6) is only rank deficient by 1, we may simply remove one of the n_e equations, and replace it with a closing equation so that the local system (3.6) is of full rank. Since only 1

closing equation is needed, we want this equation to involve all of the interior fluxes. To form this closing equation in a consistent way, we assume that our velocity $\mathbf{u}$ is an $\mathbb{RT}_0$ function over the cell K. Note that this ensures that the interior fluxes that we compute will still satisfy (3.6) (i.e. divergence is still preserved locally) since the divergence of an $\mathbb{RT}_0$ function is constant on the entire cell, and thus the same on each of the sub-triangles.

Lemma 3.1.2 (Consistency condition)**.** *Let $\mathbf{v}_1, \mathbf{v}_2, \dots \mathbf{v}_m$ be distinct vertices of cell K. If the velocity $\mathbf{u}$ is an $\mathbb{RT}_0$ function over the cell K, $F_{\sigma_i^*} = \int_{\sigma_i^*} \mathbf{u} \cdot \boldsymbol{n}_{T_{K,\sigma_i},\sigma_i^*}$ and if $\boldsymbol{x}_K$ is a point such that*

$$\sum_{i=1}^{m} \alpha_i \mathbf{v}_i = \boldsymbol{x}_K \quad \textit{with} \quad \sum_{i=1}^{m} \alpha_i = 1, \tag{3.7}$$

then

$$\sum_{i=1}^{m} \alpha_i F_{\sigma_i^*} = 0.$$

Proof. Let σ_i^* be the segment defined by the points $\boldsymbol{x}_K$ and $\mathbf{v}_i$. Since $\mathbf{u}$ is an $\mathbb{RT}_0$ function, $\mathbf{u} = a\boldsymbol{x} + \mathbf{b}$ for some constant a and vector $\mathbf{b}$. Now

$$\begin{aligned} F_{\sigma_i^*} &= \int_{\sigma_i^*} \mathbf{u} \cdot \mathbf{n}_{T_{K,\sigma_i},\sigma_i^*} \\ &= (a\boldsymbol{x}_{\sigma_i^*} + \mathbf{b}) \cdot \mathbf{n}_{T_{K,\sigma_i},\sigma_i^*} |\sigma_i^*| \quad \text{where } \boldsymbol{x}_{\sigma_i^*} \text{ is the center of } \sigma_i^* \\ &= (a\boldsymbol{x}_{\sigma_i^*} + \mathbf{b}) \cdot Rot(\boldsymbol{x}_K - \mathbf{v}_i) \quad \text{where } Rot \text{ represents a clockwise rotation by } \frac{\pi}{2} \end{aligned}$$

Since $(\boldsymbol{x}_{\sigma_i^*} - \boldsymbol{x}_K) \perp (Rot(\boldsymbol{x}_K - \mathbf{v}_i)$, we deduce

$$\begin{aligned} \sum_{i=1}^{m} \alpha_i F_{\sigma_i^*} &= \sum_{i=1}^{m} \alpha_i (a\boldsymbol{x}_{\sigma_i^*} + \mathbf{b}) \cdot Rot(\boldsymbol{x}_K - \mathbf{v}_i) \\ &= \sum_{i=1}^{m} \alpha_i (a(\boldsymbol{x}_{\sigma_i^*} - \boldsymbol{x}_K) + a\boldsymbol{x}_K + \mathbf{b}) \cdot Rot(\boldsymbol{x}_K - \mathbf{v}_i) \\ &= \sum_{i=1}^{m} \alpha_i (a\boldsymbol{x}_K + \mathbf{b}) \cdot Rot(\boldsymbol{x}_K - \mathbf{v}_i) \\ &= (a\boldsymbol{x}_K + \mathbf{b}) Rot\left(\sum_{i=1}^{m} \alpha_i (\boldsymbol{x}_K - \mathbf{v}_i) \right) \\ &= 0, \end{aligned}$$

where the conclusion follows by (3.7). ∎

For an arbitrary choice of the point $\boldsymbol{x}_K$ for a generic polygon K with n_v vertices, we note that we may form n_v triangles by taking 3 consecutive vertices (in counter clockwise order) $\mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2}$ for $i = 1, 2, \ldots, n_v$, where we define $\mathbf{v}_{n_v+1} = \mathbf{v}_1$ and $\mathbf{v}_{n_v+2} = \mathbf{v}_2$. The choice of n_v triangles will then give us n_v equations that relate the vertices $\mathbf{v}_i$ with the point $\boldsymbol{x}_K$. We note that we may even have more relations, since, in general, we may form $\binom{n_v}{3}$ triangles. At this stage, we recall that the system (3.6) is rank-deficient by 1, and hence we only need one closing equation. Needing only one equation, we do not want to create any bias in constructing it. We therefore use all vertices $(\mathbf{v}_1, \ldots, \mathbf{v}_{n_v})$ of K to form this relation. In particular, the n_v triangles with vertices $\mathbf{v}_i, \mathbf{v}_{i+1}, \mathbf{v}_{i+2}$, would be enough to determine a closing equation that involves all of the vertices and interior fluxes. Hence, for each of these triangles, express $\boldsymbol{x}_K$ in terms of barycentric coordinates

$$\begin{aligned}
\alpha_{1,1}\mathbf{v}_1 + \alpha_{1,2}\mathbf{v}_2 + \alpha_{1,3}\mathbf{v}_3 &= \boldsymbol{x}_K \\
\alpha_{2,2}\mathbf{v}_2 + \alpha_{2,3}\mathbf{v}_3 + \alpha_{2,4}\mathbf{v}_4 &= \boldsymbol{x}_K \\
&\vdots \\
\alpha_{n_v-1,1}\mathbf{v}_1 + \alpha_{n_v-1,n_v-1}\mathbf{v}_{n_v-1} + \alpha_{n_v-1,n_v}\mathbf{v}_{n_v} &= \boldsymbol{x}_K \\
\alpha_{n_v,1}\mathbf{v}_1 + \alpha_{n_v,2}\mathbf{v}_2 + \alpha_{n_v,n_v}\mathbf{v}_{n_v} &= \boldsymbol{x}_K,
\end{aligned}$$

where $\sum_{j=i}^{i+2} \alpha_{i,j} = 1$, with

$$\alpha_{n_v-1,n_v+1} = \alpha_{n_v-1,1},\ \alpha_{n_v,n_v+1} = \alpha_{n_v,1},\ \text{and}\ \alpha_{n_v,n_v+2} = \alpha_{n_v,2}.$$

Adding up all the equations, dividing both sides by n_v, and denoting the coefficient of $\mathbf{v}_i$ as α_i, we have $\sum_{i=1}^{n_v} \alpha_i \mathbf{v}_i = \boldsymbol{x}_K$ with $\sum_{i=1}^{n_v} \alpha_i = 1$, so we deduce from Lemma 3.1.2, the consistency condition

$$\sum_{i=1}^{n_v} \alpha_i F_{\sigma_i^*} = 0. \tag{3.8}$$

This holds if the velocity $\mathbf{u}$ is an $\mathbb{RT}_0$ function in K. In our case, we extend this notion and use it more generally by reconstructing velocities from fluxes that satisfy (3.6) and (3.8), which we denote as *C velocities* ('C' for 'consistent').

Remark 3.1.3 (Particular $\boldsymbol{x}_K$ and barycentric combinations)**.** *If $\boldsymbol{x}_K$ is the iso-barycenter of the vertices of K, i.e. $\boldsymbol{x}_K = \frac{1}{n_v}\sum_{i=1}^{n_v} \mathbf{v}_i$, then a consistency relation is simply $\sum_{i=1}^{n_v} F_{\sigma_i^*} = 0$. If $\boldsymbol{x}_K$ is the center of mass of K, then a consistency relation is*

$$\sum_{i=1}^{n_v} \frac{|T_{K,\sigma_{i-1}}| + |T_{K,\sigma_i}|}{2|K|} F_{\sigma_i^*} = 0,$$

where $T_{K,\sigma_{i-1}}$ is the triangle that shares edge σ^*_{i-1} with T_{K,σ_i}.

The system of equations (3.6)–(3.8) has a unique and explicit solution. Indeed, set

$$a_i = \frac{|T_{K,\sigma_i}|}{|K|}\left(\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}\right) - F_{K,\sigma_i} \qquad \text{for } i = 1,\dots,n_v.$$

The system (3.6) is then

$$\begin{aligned} F_{\sigma^*_1} &= F_{\sigma^*_{n_v}} + a_1 \\ F_{\sigma^*_2} &= F_{\sigma^*_1} + a_2 \\ &\vdots \\ F_{\sigma^*_{n_v-1}} &= F_{\sigma^*_{n_v-2}} + a_{n_v-1}. \end{aligned}$$

From these, we easily deduce that

$$F_{\sigma^*_k} = F_{\sigma^*_{n_v}} + \sum_{j=1}^{k} a_j, \quad k = 1,2,\dots,n_v-1. \tag{3.9}$$

By noticing that $\sum_{i=1}^{n_v} a_i = 0$, we see that (3.9) also holds for $k = n_v$. Multiplying (3.9) by α_k, summing over $k = 1,\cdots,n_v$, using the fact that $\sum_{k=1}^{n_v}\alpha_k = 1$ and (3.8), we obtain an explicit expression for $F_{\sigma^*_{n_v}}$, given by

$$F_{\sigma^*_{n_v}} = -\sum_{k=1}^{n_v}\left(\alpha_k \sum_{j=1}^{k} a_j\right). \tag{3.10}$$

Equation (3.10), together with (3.9), give us explicit expressions for $F_{\sigma^*_k}$, $k = 1,2,\dots,n_v$.

These computations show an advantage of this method over the technique consisting in selecting a minimal norm solution of (3.6). Here, we do not need to solve any local system, as we have explicit expressions for the fluxes, as seen in equations (3.9)–(3.10). The weakness of this reconstruction is the fact that it is highly dependent on the fact that we only need one closing equation in 2D, and thus, extension into 3D is non-trivial.

3.1.4 Introducing auxiliary cell-centered unknowns (A method)

The idea here is to provide a setting so that the flux reconstructions may be extended to 3D easily. To do so, we look for internal fluxes that are composed

of a consistent flux coming from a constant velocity in the cell, and an added stabilisation term, similar to a Brezzi-Pitkäranta stabilisation (a discrete inconsistent Laplacian on the submesh). This was actually inspired by the post-processing technique in [12].

We note that if ξ is a constant velocity field, then for all $\sigma \in \mathcal{E}_K$,

$$\int_\sigma \xi \cdot \mathbf{n}_{K,\sigma} = |\sigma| \xi \cdot \mathbf{n}_{K,\sigma}.$$

Definition 3.1.4 (Constant, consistent approximation of a velocity field)**.** *Let ξ be a constant velocity field. We say that $\mathbf{u}_K$ is a constant, consistent approximation of ξ if and only if, given fluxes $F_{K,\sigma} = |\sigma| \xi \cdot \mathbf{n}_{K,\sigma}$, we have that $\mathbf{u}_K = \xi$.*

For example, taking

$$\mathbf{u}_K = \frac{1}{|K|} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K)$$

for a constant velocity field ξ, we have that $F_{K,\sigma} = |\sigma| \xi \cdot \mathbf{n}_{K,\sigma}$. Using the area formula

$$|K| \mathbf{e} = \sum_{\sigma \in \mathcal{E}_K} |\sigma| \mathbf{e} (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) \cdot \mathbf{n}_{K,\sigma},$$

valid for any vector $\mathbf{e}$, we deduce that $\mathbf{u}_K = \xi$ for constant vector fields ξ.

Having cut the cell K into simplices, for each simplex $S \in \mathcal{S}_K$ we then look for internal fluxes of the form

$$\forall \sigma^* \in \mathcal{E}_S^* \,,\; F_{S,\sigma^*} = |\sigma^*| \mathbf{u}_K \cdot \mathbf{n}_{S,\sigma^*} + \frac{|\sigma^*|}{h_{\sigma^*}} (Q_S - Q_{S'}), \tag{3.11}$$

where $\mathbf{u}_K$ is a constant, consistent approximation of a velocity field as in Definition 3.1.4, S' is the simplex on the other side of σ^*, h_{σ^*} is a characteristic distance between S and S' (for example, the distance between their centers of mass), and $(Q_S)_{S \in \mathcal{S}_K}$ are real numbers (if $\mathbf{u}$ is a Darcy velocity $-\Lambda \nabla p$, then these could be considered as potentials inside each simplex).

Substituting (3.11) into (3.5), we obtain the following square system on the unknowns $(Q_S)_{S \in \mathcal{S}_K}$:

$$\forall S \in \mathcal{S}_K \,,\; \sum_{\sigma^* \in \mathcal{E}_S^*} \frac{|\sigma^*|}{h_{\sigma^*}} (Q_S - Q_{S'}) = b_S \tag{3.12}$$

where b_S depends on $\mathbf{u}_K$ and the fluxes around K. We note here that the solution to the system (3.12) is not unique. Indeed, when $b_S = 0$ for all $S \in \mathcal{S}_K$, a set of solutions is given by $Q_S = Q_{S'}$ for all $S \in \mathcal{S}_K$. Actually, we recognise here a (non-consistent) 2-point discretisation of the Laplacian on the submesh $\mathcal{S}_K$, with homogeneous Neumann boundary conditions.

Lemma 3.1.5. *If $(Q_S)_{S\in\mathcal{S}_K}$ is a solution of (3.12), then*

$$\sum_{\sigma^*\in\mathcal{E}_K^*}\frac{|\sigma^*|}{h_{\sigma^*}}(Q_S-Q_{S'})^2=\sum_{S\in\mathcal{S}_K}b_SQ_S.$$

Proof. Upon multiplying equation (3.12) by Q_S, we obtain for all $S\in\mathcal{S}_K$,

$$Q_S\sum_{\sigma^*\in\mathcal{E}_S^*}\frac{|\sigma^*|}{h_{\sigma^*}}(Q_S-Q_{S'})=b_SQ_S.$$

Considering the simplex S' which shares the face σ^* with the simplex S, we have

$$Q_{S'}\sum_{\sigma^*\in\mathcal{E}_S^*}\frac{|\sigma^*|}{h_{\sigma^*}}(Q_{S'}-Q_S)=b_SQ_{S'},$$

and thus

$$-Q_{S'}\sum_{\sigma^*\in\mathcal{E}_S^*}\frac{|\sigma^*|}{h_{\sigma^*}}(Q_S-Q_{S'})=b_SQ_{S'}.$$

Now, we note that each equation in (3.12) involves each internal face $\sigma^*\in\mathcal{E}_K^*$ exactly once, and each face $\sigma^*\in\mathcal{E}_K^*$ is shared by exactly two simplices S and S' in $\mathcal{S}_K$. Hence, upon taking the sum over $S\in\mathcal{S}_K$, and gathering the terms by internal faces $\sigma^*\in\mathcal{E}_K^*$, we obtain

$$\sum_{\sigma^*\in\mathcal{E}_K^*}\frac{|\sigma^*|}{h_{\sigma^*}}(Q_S-Q_{S'})^2=\sum_{S\in\mathcal{S}_K}b_SQ_S.$$

■

In particular, if $(b_S)_{S\in\mathcal{S}_K}=0$ then all $(Q_S)_{S\in\mathcal{S}_K}$ are identical. Hence, the matrix of (3.12) only has the constant vector $\mathbf{1}$ in its kernel, and it is therefore of rank $\sharp\mathcal{S}_K-1$ (rank–deficient by 1). This also shows that, upon choosing one of the Q_S, the system (3.12) has a unique solution.

We now detail some computations in 2D and 3D which show that the solutions to (3.12) are actually easy to compute. In 2D, we may obtain explicit expressions for the solutions to (3.12); whereas in 3D, we would need to perform a 2-step process, the first of which involves solving a local linear system of n_f equations, followed by a second step, which gives explicit expressions for the fluxes.

3.1.4.1 Detailed computations in 2D

Upon ordering the edges σ_i, and thus the corresponding associated triangles T_{K,σ_i} of cell K in counterclockwise order, we denote by σ_i^* the edge shared between T_{K,σ_i} and $T_{K,\sigma_{i+1}}, i = 1, \ldots, n_e$, with the convention that $\sigma_{n_e+1} = \sigma_1$ (see Figure 3.2). We introduce an auxiliary unknown Q_i associated to each of the sub-cells T_{K,σ_i} and write the corresponding 2D equation for (3.11), given by

$$F_{\sigma_i^*} = \bar{F}_{\sigma_i^*} + \frac{|\sigma_i^*|}{h_{i,i+1}}(Q_i - Q_{i+1}) \tag{3.13}$$

where $h_{i,i+1} = \frac{1}{2}(h_i + h_{i+1})$ with h_i and h_{i+1} being the diameters of T_{K,σ_i} and $T_{K,\sigma_{i+1}}$ respectively, and $\bar{F}_{\sigma_i^*} = |\sigma_i^*|\mathbf{u}_K \cdot \mathbf{n}_{K,\sigma_i^*}$, with $\mathbf{n}_{K,\sigma_i^*}$ the outward unit normal along edge σ_i^* of the triangle T_{K,σ_i}, and $\mathbf{u}_K$ being a constant consistent approximation of a velocity field $\mathbf{u}$ in a cell K. The new system in terms of $\mathbf{Q} = (Q_1, \ldots, Q_{n_e})$ is still of rank $n_e - 1$, and the matrix for this system can be viewed as a type of discrete Laplacian, with kernel given by the $n_e \times 1$ vector of all ones $\mathbf{1}$. More specifically, by writing $\beta_i = \dfrac{|\sigma_i^*|}{h_{i,i+1}}$, the system of equations (3.12) can be written as $A\mathbf{Q} = \mathbf{b}$, with unknowns $\mathbf{Q}$, where A is a sparse $n_e \times n_e$ symmetric matrix with entries

$$\begin{aligned} a_{i,i-1} &= -\beta_{i-1}, \\ a_{i,i} &= \beta_i + \beta_{i-1}, \\ a_{i,i+1} &= -\beta_i, \end{aligned}$$

for $i = 1, \ldots, n_e$, where the entries a_{n_e,n_e+1} and $a_{1,0}$ refer to $a_{n_e,1}$ and a_{1,n_e}, respectively, and $\beta_0 = \beta_{n_e}$. Setting $\bar{F}_{\sigma_0^*} = \bar{F}_{\sigma_{n_e}^*}$, the vector $\mathbf{b}$ is composed of entries

$$b_i = \frac{|T_{K,\sigma_i}|}{|K|} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} - F_{K,\sigma_i} - \bar{F}_{\sigma_i^*} + \bar{F}_{\sigma_{i-1}^*}.$$

In matrix form, we may see

$$A = \begin{bmatrix} \beta_1 + \beta_{n_e} & -\beta_1 & 0 & \cdots & -\beta_{n_e} \\ -\beta_1 & \beta_1 + \beta_2 & -\beta_2 & \cdots & 0 \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ 0 & \cdots & -\beta_{n_e-2} & \beta_{n_e-2} + \beta_{n_e-1} & -\beta_{n_e-1} \\ -\beta_{n_e} & 0 & \cdots & -\beta_{n_e-1} & \beta_{n_e-1} + \beta_{n_e} \end{bmatrix},$$

and further note that $A = P^T DP$ where D is the diagonal matrix with $d_{i,i} = \beta_i, i = 1, \ldots, n_e$ and P is the matrix such that for $i = 1, \ldots, n_e$

$$\begin{aligned} p_{i,i} &= 1 \\ p_{i,i+1} &= -1, \end{aligned}$$

where the entry p_{n_e,n_e+1} is equal to $p_{n_e,1}$:

$$P = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \cdots & 0 \\ \mathbf{0} & \ddots & \ddots & \ddots & \mathbf{0} \\ 0 & 0 & \cdots & 1 & -1 \\ -1 & 0 & \cdots & 0 & 1 \end{bmatrix}.$$

Setting one of the Q_i to an arbitrary constant, we will be able to recover a unique set of interior fluxes.

Remark 3.1.6 (Constant velocity fields)**.** *Given a constant velocity field* $\mathbf{u}$*, the fluxes* $F_{K,\sigma} = |\sigma|\mathbf{u} \cdot \mathbf{n}_{K,\sigma}$*. Hence,* $\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = 0$*. Also, we note that* $\bar{F}_{\sigma_i^*}, \bar{F}_{\sigma_{i-1}^*}$*, and* F_{K,σ_i} *are fluxes along the edges of the triangle* T_{K,σ_i}*, which leads to* $-F_{K,\sigma_i} - \bar{F}_{\sigma_i^*} + \bar{F}_{\sigma_{i-1}^*} = 0$*. This then gives* $b_i = 0$ *for all* $i = 1, \ldots, n_e$*, which means that we are solving the system* $A\boldsymbol{Q} = \boldsymbol{0}$*. Owing to Lemma 3.1.5, we deduce that* $Q_i = Q_j$ *for* $i, j = 1, \ldots, n_e$*. Hence, the interior fluxes* $F_{K,\sigma^*} = \bar{F}_{\sigma^*}$ *are exactly what we would have if* $\mathbf{u}$ *is a constant velocity field. Due to this, our reconstruction will be able to recover a constant velocity field* $\mathbf{u}$ *exactly.*

In general, the quantities $q_i = (Q_i - Q_{i+1})$ can uniquely be determined by removing any one of the equations in the system (3.5) and replacing it with a relation involving the q_i's. In our case, we remove the equation corresponding to the n_e-th row of A, and replace it with $\sum_{i=1}^{n_e} q_i = 0$. This will then yield a matrix system $\hat{A}\mathbf{q} = \hat{\mathbf{b}}$, where $\mathbf{q}$ is the $n_e \times 1$ vector with ith entry q_i, $\hat{\mathbf{b}}$ is the $n_e \times 1$ column vector with the first $n_e - 1$ entries identical to $\mathbf{b}$ but with last entry 0, and $\hat{A}$ being the matrix formed by the first $n_e - 1$ rows of $P^T D$, augmented by the $1 \times n_e$ row vector of all ones $\mathbf{1}$. We note now that $\hat{A}$ has full rank. Moreover, in this form, we may explicitly obtain the values

$q_i = (Q_i - Q_{i+1})$. To be specific, we have

$$\begin{aligned}
\beta_1 q_1 &= \beta_{n_e} q_{n_e} + b_1 \\
\beta_2 q_2 &= \beta_{n_e} q_{n_e} + b_1 + b_2 \\
&\vdots \\
\beta_{n_e-1} q_{n_e-1} &= \beta_{n_e} q_{n_e} + \sum_{i=1}^{n_e-1} b_i \\
\sum_{i=1}^{n_e} q_i &= 0.
\end{aligned} \tag{3.14}$$

Upon substituting the first $n_e - 1$ equations into the last equation, we then have

$$\begin{aligned}
q_{n_e} \sum_{i=1}^{n_e} \frac{\beta_{n_e}}{\beta_i} &= -\sum_{i=1}^{n_e-1} \sum_{j=i}^{n_e-1} \frac{b_i}{\beta_j} \\
q_{n_e} &= -\frac{1}{\beta_{n_e}} \frac{\sum_{i=1}^{n_e-1} \sum_{j=i}^{n_e-1} \frac{b_i}{\beta_j}}{\sum_{i=1}^{n_e} \frac{1}{\beta_i}}.
\end{aligned}$$

The values $q_i, i = 1, \ldots, n_e - 1$ can easily be obtained by simply substituting the value q_{n_e} into the equations in (3.14). Aside from this, another advantage of this method is that it can easily be extended to 3D (see Section 3.1.5). Velocities reconstructed from fluxes that satisfy (3.6) and equations (3.13)–(3.14) will be denoted as *A velocities* ('A' for 'auxiliary').

Remark 3.1.7 (Comparison with C-velocities)**.** *Given a constant velocity field* $\mathbf{u}$*, the A velocities will be able to recover* $\mathbf{u}$ *exactly, as discussed in Remark 3.1.6. The same is true for C velocities. However, if the velocity field* $\mathbf{u}$ *is not constant, then, in general, this reconstruction is different from C velocities. This can be seen because the fluxes in* (3.13) *do not necessarily satisfy the final relation* (3.8) *which is used to define C velocities. We can, however, make this reconstruction equivalent to the C velocities by setting the values of the diagonal matrix D to be* $\beta_i = \frac{1}{\alpha_i}$*, where* α_i *is as described in* (3.8).

3.1.5 KR, C, and A velocities in 3D

In this chapter, we explore the option of extending the notions of the KR, C and A velocities into 3D. First, we note that the extension of the KR

velocity into 3D is easy to implement, since we simply have to find the least norm solution to the system (3.5). However, when we look at the case of C velocities, things get a little bit more complicated. In particular, there are 2 simplices corresponding to each edge $e_{i,j}$ of the cell K: namely, $S_{i,j}$ formed by $\boldsymbol{x}_K, \boldsymbol{x}_{\sigma_i}$ and $e_{i,j}$, and $S_{j,i}$ formed by $\boldsymbol{x}_K, \boldsymbol{x}_{\sigma_j}$ and $e_{i,j}$. However, there are 3 internal faces, and hence unknown fluxes corresponding to each edge $e_{i,j}$ of the cell K: namely, the face formed by joining a vertex v_i of $e_{i,j}$ to the point $\boldsymbol{x}_K$ and $\boldsymbol{x}_{\sigma_i}$, the one formed by joining a vertex v_i of $e_{i,j}$ to the point $\boldsymbol{x}_K$ and $\boldsymbol{x}_{\sigma_j}$, and finally the one formed by joining the edge $e_{i,j}$ to the point $\boldsymbol{x}_K$. This leads us to a system of $2n_e$ equations in $3n_e$ unknowns, where n_e is the number of edges of cell K. Compared to the 2D case, which only required to find one closing equation (3.8), we need several additional equations in the 3D case. Hence, the possibility of fully extending the C velocities into 3D still remains an open question. Finally, we discuss the extension of the A velocities into 3D. An option of partially extending the C velocities into 3D by first going through the process needed for an A velocity will also be discussed.

Remark 3.1.8 (A mix of KR and C velocities)**.** *One option for extending the C velocities into 3D would be to find a least norm solution to the system of $2n_e + n_f$ equations given by* (3.5) *and, for $i = 1, \dots, n_f$, an analogue of* (3.8)*, which involves all sub-internal fluxes associated with the sub-cell K_i. However, even with the additional n_f equations, there is no guarantee that this would lead to a reconstruction that recovers constant velocity fields. Moreover, by having to solve a least norm problem, we lose the advantage of the C method in 2D: the availability of an explicit expression for the fluxes.*

We start by partitioning each cell K into n_f sub-cells as in Figure 3.3. Auxiliary cell-centered unknowns (1 for each sub-cell) are then introduced as in Section 3.1.4, but over generic polyhedrons, instead of simplices. This gives us n_f equations in n_f unknowns, which is rank deficient by 1. We note however, that we have a total of n_e interior fluxes. Each interior flux F_{σ_i,σ_j} corresponds to the face $\sigma_{i,j}$ formed by joining the interior point $\boldsymbol{x}_K$ to an edge $e_{i,j}$ of K being shared by the faces σ_i and σ_j of the cell K. We then write

$$F_{\sigma_i,\sigma_j} = \bar{F}_{\sigma_i,\sigma_j} + \frac{|\sigma_{i,j}|}{h_{i,j}}(Q_i - Q_j) \tag{3.15}$$

where $\bar{F}_{\sigma_i,\sigma_j} = |\sigma_{i,j}|\mathbf{u}_K \cdot \mathbf{n}_{K_i,\sigma_{i,j}}$ and $h_{i,j} = \frac{1}{2}(h_i + h_j)$, where h_i and h_j are the diameters of the sub-cells K_i and K_j respectively. For simplicity of notation, write $\beta_{i,j} = \dfrac{|\sigma_{i,j}|}{h_{i,j}}$. We then generalise (3.5) into generic cells (i.e. the average

of the fluxes for each sub-cell is equal to the average of the divergence of the fluxes on K). Denoting by $(n_e)_i$ the number of edges of a face σ_i of cell K, we then have, for preservation of divergence on each sub-cell K_i,

$$\sum_{j=1}^{(n_e)_i} \beta_{i,j}(Q_i - Q_j) = \frac{|K_i|}{|K|} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} - \sum_{j=1}^{(n_e)_i} \bar{F}_{\sigma_i,\sigma_j} - F_{K,\sigma_i}. \tag{3.16}$$

As with the two dimensional case, the values $(Q_i - Q_j)$ can be uniquely determined by fixing one of the values Q_i. Moreover, expressing the system (3.16) in matrix form $A\mathbf{Q} = \mathbf{b}$, we find that $A = P^T D P$ where P is an $n_e \times n_f$ matrix and D is an $n_e \times n_e$ diagonal matrix. Each row of the matrix P corresponds to an edge $e_{i,j}$ of the sub-cell being shared by the faces σ_i and σ_j. Without loss of generality, we may assume that $i < j$, and denote the entries of the k-th row of the matrix P to be $p_{k,i} = 1$, $p_{k,j} = -1$. Correspondingly, the diagonal matrix D has entries $d_{k,k} = \beta_{i,j}$.

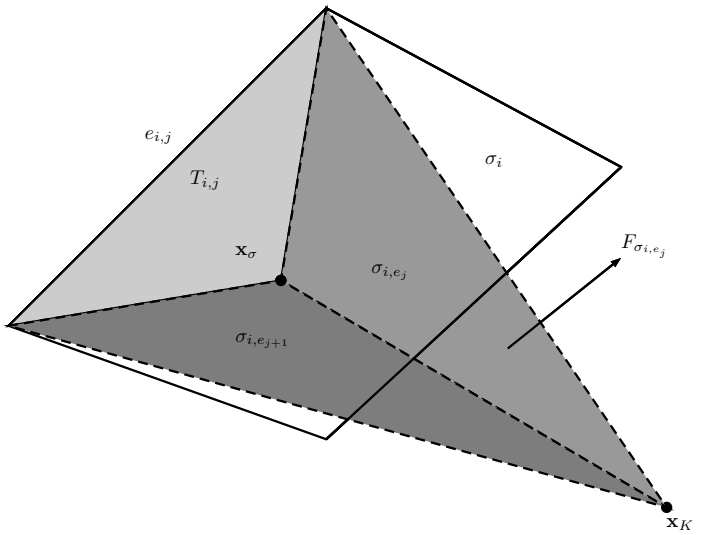
Remark 3.1.9 (Mesh structure and efficient implementation)**.** *In practice, we rarely have to invert n_K matrices of size $n_f \times n_f$, where n_K is the number of cells in the mesh. If two cells K, L have the same topological structure, then the local matrices P_K, P_L (which determines the connectivity between cells) corresponding to cells K and L, respectively, are the same. If a mesh has a lot of cells with the same topological structure (which is common with meshes encountered in applications), then we only need to invert and store a number of matrices equal to the number of cells with different topological structures, which is much smaller than n_K. In particular, if the mesh is made up of identical cells (as in Cartesian meshes), then we only have to store and invert one matrix.*

At this stage, we recall that our aim is to reconstruct $\mathbb{RT}_0$ functions over simplices. Hence, we proceed by breaking each of the sub-cells K_i into simplices. On each sub-cell K_i, we pick a point $\boldsymbol{x}_{\sigma_i}$ on the face σ_i and associate, for each edge e_j ($j = 1, \dots (n_e)_i$) of the face σ_i an interior face σ_{i,e_j} (Note here that since we are only working locally on a sub-cell K_i, the index i has been dropped from e, hence writing e_j instead of $e_{i,j}$). We then form a simplex $S_{i,j}$ with base on the sub-triangle $T_{i,j}$ in σ_i and faces $\sigma_{i,j}, \sigma_{i,e_j}, \sigma_{i,e_{j+1}}$ (see Figure 3.5).

Since (3.16) ensures that each sub-cell K_i preserves the divergence of the entire cell K, we only need each simplex to preserve the divergence of the sub-cell K_i it resides in. Hence, for each edge e_j of the face σ_i, the equation for preservation of divergence is then given by

$$F_{\sigma_i,e_j} + F_{\sigma_i,e_{j+1}} = \frac{|S_{i,j}|}{|K_i|} \sum_{\sigma \in \mathcal{E}_{K_i}} F_{K,\sigma} - F_{\sigma_i,\sigma_j} - \frac{|T_{i,j}|}{|\sigma_i|} F_{K,\sigma_i}, \tag{3.17}$$

Figure 3.5: Triangulation of the sub-cells



where F_{σ_i,e_j} is the interior flux along the interior face σ_{i,e_j}, oriented outward of the simplex $S_{i,j}$. We note here that this consists of $(n_e)_i$ simplices and $(n_e)_i$ interior fluxes for each sub-cell, which corresponds to $(n_e)_i$ equations in $(n_e)_i$ unknowns, but is rank deficient by 1. This system of equations looks exactly the same as those obtained in 2D (i.e. the connectivity is determined through the adjacency of the triangles $T_{i,j}$ and $T_{i,j+1}$, which correspond to the edges e_j and e_{j+1} of the face σ_i, respectively). The fluxes F_{σ_i,e_j} may then be approximated via the C method, or in the same manner as in (3.15). For the latter approach, moving $\bar{F}_{\sigma_{i,e_j}} + \bar{F}_{\sigma_{i,e_{j+1}}}$ to the right hand side of (3.17), and setting $Q_{i,j}$ to be the unknown associated with the simplex $S_{i,j}$, we then have

$$\beta_{\sigma_i,e_j}(Q_{i,j-1} - Q_{i,j}) + \beta_{\sigma_i,e_{j+1}}(Q_{i,j} - Q_{i,j+1}) = b_j,$$

where

$$b_j = \frac{|S_{i,j}|}{|K_i|} \sum_{\sigma \in \mathcal{E}_{K_i}} F_{K,\sigma} - F_{\sigma_i,\sigma_j} - \frac{|T_{i,j}|}{|\sigma_i|} F_{K,\sigma_i} - \bar{F}_{\sigma_{i,e_j}} - \bar{F}_{\sigma_{i,e_{j+1}}}.$$

At this stage, we recognise that writing the system of equations corresponding to (3.17) in matrix form leads to solving essentially the same system of equations as in the 2D case. Hence, expressions for $(Q_{i,j} - Q_{i,j+1})$ may be obtained explicitly. If we will be using the C method, a 3D extension of the closing equation (3.8) can be obtained by finding a barycentric combination of $\boldsymbol{x}_{\sigma_i}$. We note here however that the velocity field obtained by the

C method is a mix of both A and C velocities, since we went through the first step of approximating the fluxes in each of the sub-cells K_i using the A method before finding a closing relation for the simplices.

Remark 3.1.10 ($\mathbb{RT}_k$ finite elements for $k \geq 1$)**.** *Suppose that we want to approximate the velocity field via an $\mathbb{RT}_k$ finite element, where $k \geq 1$. Then higher order moments along the interior faces (edges) and sub-cells would be needed. This can be achieved by solving the diffusion problem* (2.1) *via the HHO scheme locally on each cell K. Access to a cheap reconstruction of the higher order moments similar to those described in Sections 3.1.3–3.1.4 is still an open question, both in 2D and 3D.*

3.2 Mixed finite elements on quadrilaterals

When we are dealing with quadrilateral cells, the use of quadrilateral mixed finite elements holds the advantage of not needing to reconstruct interior fluxes. Moreover, for $k \geq 1$, access to higher order moments are readily provided by HHO schemes upon solving the diffusion problem (2.1). We start by illustrating the $\mathbb{RT}_k$ on rectangular cells. If K is a rectangle, then for $k \geq 0$,

$$\mathbb{RT}_k(K) := \mathcal{Q}_{k+1,k} \times \mathcal{Q}_{k,k+1},$$

where $\mathcal{Q}_{k,m}$ is the space of polynomials of the form $q(x,y) = \sum_{i=0}^{k} \sum_{j=0}^{m} a_{i,j} x^i y^j$. From this, we see that the space $\mathbb{RT}_k(K)$ has $2(k+1)(k+2)$ degrees of freedom. In particular, given $\mathbf{u} \in H^1(K)^2$, an interpolant $\mathcal{I}_{\mathbb{RT}_k(K)} : H^1(K)^2 \to \mathbb{RT}_k(K)$ can be uniquely determined.

Lemma 3.2.1 (Existence and Uniqueness)**.** *Given $\mathbf{u} \in H^1(K)^2$, there exists a unique $\mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \in \mathbb{RT}_k(K)$ satisfying the equations:*

$$\int_\sigma \mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \cdot \mathbf{n}_{K,\sigma} p_\sigma = \int_\sigma \mathbf{u} \cdot \mathbf{n}_{K,\sigma} p_\sigma \text{ for all } p_\sigma \in \mathbb{P}^k(\sigma) \text{ and } \sigma \in \mathcal{E}_K, \quad (3.18a)$$

and for $k \geq 1$,

$$\int_K \mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u} \cdot \boldsymbol{\phi}_K = \int_K \mathbf{u} \cdot \boldsymbol{\phi}_K \text{ for all } \boldsymbol{\phi}_K \in \mathcal{Q}_{k-1,k} \times \mathcal{Q}_{k,k-1}. \quad (3.18b)$$

The proof is very similar to that of Lemma 3.1.1, and will be omitted. Essentially, from Lemma 3.2.1, we see that we may use, for the degrees of freedom of $\mathbb{RT}_k$:

- the moments of up to order k of $\mathbf{u} \cdot \mathbf{n}_{K,\sigma}$ on the sides or faces of K;

- the moments of up to $x^{k-1}y^k$ and $x^k y^{k-1}$ for the first and second arguments of $\mathbf{u}$ on K, respectively.

The degrees of freedom for $\mathbb{RT}_0$ and $\mathbb{RT}_1$ on rectangles are shown in Figure 3.6.

Remark 3.2.2 (Comparison with $\mathbb{RT}_k$ on Simplices)**.** *The dimension of $\mathbb{RT}_k$ on quadrilaterals is larger than the one of $\mathbb{RT}_k$ elements on simplices. This implies that quadrilateral elements need more degrees of freedom, and hence, for a fixed value k, quadrilateral elements should offer more accuracy compared to simplicial elements.*

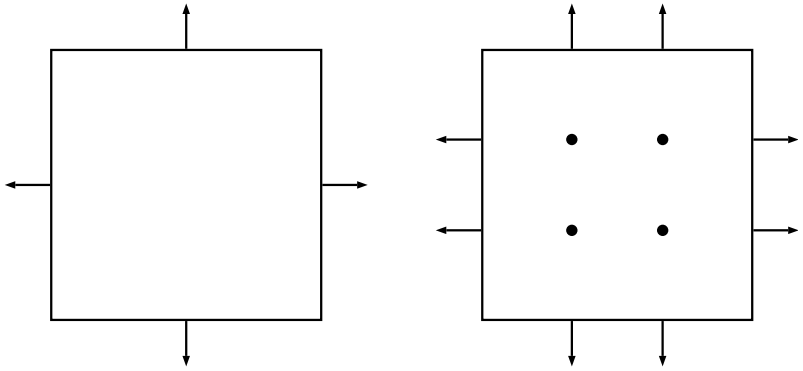


Figure 3.6: DOFs of $\mathbb{RT}_k$ finite elements on rectangles (left: $\mathbb{RT}_0$, right: $\mathbb{RT}_1$)

For generic convex quadrilaterals, the idea is to consider a reference element (usually a square) and perform a Piola transform. Suppose that $F : R \to K$ is the bilinear map from the reference element R to the quadrilateral K. Applying the Piola transform for a function $\mathbf{u}_R : R \to \mathbb{R}$, we recover a function $\mathbf{u}_K : K \to \mathbb{R}$, given by

$$\mathbf{u}_K(F(x)) := JF(x)^{-1} DF(x) \mathbf{u}_R(x), \tag{3.19}$$

where DF is the Jacobian of the bilinear map F, and $JF = |\det DF|$.

3.2.1 Properties of the Piola transform

The following are some important properties of the Piola transform [16, Lemma 1.5]: Suppose that $F : R \to K$ is the bilinear map that sends the reference element R to the quadrilateral K and that $\mathbf{u}_K$ is obtained from $\mathbf{u}_R$

via a Piola transform as in (3.19). Then for all sufficiently smooth $v : R \to \mathbb{R}$ we have that

$$\int_K \mathbf{u}_K \cdot \nabla(v \circ F^{-1}) = \int_R \mathbf{u}_R \cdot \nabla v, \tag{3.20a}$$

$$\int_K (v \circ F^{-1}) \mathrm{div} \mathbf{u}_K = \int_R v \mathrm{div} \mathbf{u}_R, \tag{3.20b}$$

$$\int_{\partial K} (v \circ F^{-1}) \mathbf{u}_K \cdot \mathbf{n}_K = \int_{\partial R} v \mathbf{u}_R \cdot \mathbf{n}_R. \tag{3.20c}$$

In particular, if a function is approximated to be $\mathbf{u}_K$ in a cell K, then the equations (3.20a)–(3.20c) allow us to compute the degrees of freedom inside the reference element R via the equations in (3.18). This determines a unique reconstruction $\mathcal{I}_{\mathbb{RT}_k(R)}\mathbf{u}$. $\mathcal{I}_{\mathbb{RT}_k(K)}\mathbf{u}$ is then obtained by applying the Piola transform to $\mathcal{I}_{\mathbb{RT}_k(R)}\mathbf{u}$.

Remark 3.2.3. *If the quadrilateral K is a parallelogram, then the map $F : R \to K$ is affine, and hence the Jacobian is a constant matrix. This means that if the function $\mathbf{u}_R : R \to \mathbb{R}$ is a polynomial, then so is $\mathbf{u}_K : K \to \mathbb{R}$. However, if K is not a parallelogram, the map F is strictly bilinear, and hence a polynomial function $\mathbf{u}_R : R \to \mathbb{R}$ will be mapped into a rational function $\mathbf{u}_K : K \to \mathbb{R}$. This is especially bad for distorted cells, as the determinant of the Jacobian would either be extremely large or small, leading to an inaccuracy in the approximation of numerical solutions.*

3.2.2 Limitations, possible outlooks and explorations

We recall here that the aim of reconstructing $H(\mathrm{div}, \Omega)$ elements in this thesis is to perform characteristic tracking and solve the characteristic equation (3.1). For $\mathbb{RT}_0$ elements, (3.1) is a system of linear ODEs, and hence, an exact solution is readily available. For $\mathbb{RT}_k$, $k \geq 1$, an exact solution to (3.1) is no longer accessible. The most simple numerical scheme to solve (3.1) would be a first order Euler scheme. However, the higher accuracy gained by using $\mathbb{RT}_k$, $k \geq 1$ will be lost by using a first order Euler scheme. To preserve the accuracy gained by using $\mathbb{RT}_k$, $k \geq 1$, either very small time steps in an Euler scheme, or high order schemes would need to be implemented to solve (3.1). Taking very small time steps in an Euler scheme is expensive; high order schemes are, however, not trivial to implement. In particular, it is difficult to compute the time a point exits a cell K and enters a cell L (as described in Section 4.6) for high order methods.

More recent quadrilateral $H(\mathrm{div}, \Omega)$ elements involve the Arnold-Boffi-Falk (ABF) [10], Arbogast-Correa (AC) [3], and the direct serendipity and

mixed finite elements [4]. All of these try to improve the accuracy that is lost in general quadrilateral elements after performing a Piola transform. The ABF elements still use a Piola transform, but introduces additional degrees of freedom on the reference element, in order to increase the accuracy. On the other hand, the implementation of direct serendipity and mixed finite elements [4] base the reconstructions on the quadrilateral itself, without going through a reference element. These improvements over the $\mathbb{RT}_k$ elements are not explored in this thesis, due to the limitations and difficulties encountered when solving (3.1).

3.3 Numerical tests

3.3.1 Tests in 2D

In this section, we illustrate the advantages of the C and A velocities over the KR velocities of Section 3.1.2. In particular, aside from the cheaper computational costs as outlined in Sections 3.1.3 and 3.1.4, the reconstructed C and A velocities are more precise when compared to KR velocities, especially on skewed meshes. We start by solving (2.1) via the HMM method in order to obtain an approximation of the fluxes $F_{K,\sigma}$ for all $K \in \mathcal{M}, \sigma \in \mathcal{E}_K$. The sub-interior fluxes will then be obtained through the methods described in Sections 3.1.2–3.1.4. These sub-interior fluxes will then be used to construct $\mathbb{RT}_0$ velocities over each sub-cell.

We will consider tests on the domain $\Omega = (0,1) \times (0,1)$, for 3 types of velocity fields:

- a constant velocity field $V = (0,1)$, obtained by solving (2.1) with $\Lambda = I$, $f = 0$ and $g = (0,1) \cdot \mathbf{n}$ on $\partial\Omega$.
- an $\mathbb{RT}_0$ velocity field $V = (x,y)$, obtained by solving (2.1) with $\Lambda = I$, $f = -2$ and $g = (x,y) \cdot \mathbf{n}$ on $\partial\Omega$.
- a generic velocity field obtained from the first numerical test in Section 2.4, i.e. $V = -\Lambda\nabla p = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$.

Tables 3.1, 3.3, and 3.4 present the relative errors obtained between the exact and reconstructed velocities for each of these test cases, respectively, on a variety of mesh geometries. Here, we denote by V_{KR}, V_{C}, and V_{A} the KR, C, and A velocities, respectively, reconstructed through the sub-interior fluxes obtained from the KR, C, and A method, on a triangular sub-mesh of each cell K of the mesh. The accuracy of these velocities are measured

Table 3.1: Relative errors in velocity reconstruction, constant velocity field $V = (0, 1)$.

Mesh	$\frac{\|V - V_{\mathrm{KR}}\|}{\|V\|}$	$\frac{\|V - V_{\mathrm{C}}\|}{\|V\|}$	$\frac{\|V - V_{\mathrm{A}}\|}{\|V\|}$
Cartesian	1.1809e-14	1.1836e-14	1.1837e-14
Hexahedral	3.6428e-02	2.9124e-13	2.9067e-13
Non-conforming	3.4737e-02	8.0722e-14	8.0722e-14
Kershaw	3.0571e-01	5.1485e-14	5.1051e-14

through the norm in $L^2(\Omega)^2$. Hence, in Tables 3.1–3.5 $||\cdot||$ refers to taking the norm in $L^2(\Omega)^2$.

As can be seen in Table 3.1, for square cells (Cartesian mesh), all three methods reconstruct the velocity accurately. However, for cells from hexahedral (Fig. 1.1, right), non-conforming, and Kershaw meshes (Fig. 1.2), KR velocities noticeably deviate from the actual velocity, by more than 30% on distorted cells. On the other hand, as expected, using the auxiliary unknowns (3.13) for the A velocities and the consistency relation (3.8) as a closure equation for the C velocities enable us to recover the velocity V up to machine precision, regardless of the mesh.

Table 3.2: CPU runtime (in seconds) for the reconstruction of a velocity field

Velocity / Mesh	KR	C	A
Cartesian	7.3528	7.3468	7.3498
hexahedral	33.8409	33.5048	33.1472
Kershaw	10.2301	10.1973	9.9306

To give an indication of the computational costs involved, Table 3.2 presents the CPU runtime (in seconds) for the construction of a velocity field: performing a triangulation, computing the interior fluxes, and reconstruction of the $\mathbb{RT}_0$ velocity over the entire mesh. Tables 3.1 and 3.2 show us that the C and A velocities are able to achieve a better accuracy compared to the KR velocities. We note however that there is no observed gain in terms of computational cost. This is due to the fact that we are only solving a very small system of equations (at most 6×6 for the hexahedral meshes) for the flux reconstructions. The measure in CPU runtime is only presented for the first test case, since the same computational time would be needed for the other test cases (only the right hand side of the system changes).

Looking at Tables 3.3–3.4, we observe that on generic grids with dis-

Table 3.3: Relative errors in velocity reconstruction, $\mathbb{RT}_0$ velocity field $V = (x, y)$.

Mesh	$\frac{\|\|V - V_{\mathrm{KR}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{C}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{A}}\|\|}{\|\|V\|\|}$
Cartesian	5.8625e-15	5.7916e-15	5.7918e-15
Hexahedral	2.3477e-02	2.7260e-03	2.7059e-03
Non-conforming	5.0411e-02	3.1049e-04	3.1019e-04
Kershaw	2.5989e-01	4.2460e-03	4.1511e-03

Table 3.4: Relative errors in velocity reconstruction, $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$.

Mesh	$\frac{\|\|V - V_{\mathrm{KR}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{C}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{A}}\|\|}{\|\|V\|\|}$
Cartesian	5.1202e-02	5.1202e-02	5.1202e-02
Hexahedral	5.7948e-02	3.9618e-02	3.9212e-02
Non-conforming	4.9047e-02	4.4937e-02	4.4936e-02
Kershaw	5.4057e-01	4.1091e-01	3.6090e-01

tortion, C and A velocities perform better than KR velocities. We note, however, for Table 3.3, that although the percentage errors for C and A velocities are smaller than 1%, we have been able to establish in Section 3.1.3 that the C velocities should be able to reconstruct an $\mathbb{RT}_0$ velocity up to machine precision. The lack of accuracy of the reconstructed velocities can be explained by the fact that the fluxes obtained from the HMM are not exact, especially on distorted grids (see Chapter 2). It was also demonstrated in Section 2.4 that on distorted meshes with diffusion tensor $\Lambda = I$, solving (2.1) with an HHO scheme with $k = 2$ gives fluxes that have an accuracy comparable to those that come from an HMM on Cartesian type meshes. Hence, we reconstruct our velocities for the $\mathbb{RT}_0$ and generic velocity test case, this time using HHO with $k = 2$ to obtain the approximation of the fluxes $F_{K,\sigma}$ as in Remark 2.3.4.

Now, the results in Table 3.5 illustrate what is expected: recovery of the $\mathbb{RT}_0$ up to machine precision for C velocities, which is much better than what we get for KR velocities. It is also interesting to note here that A velocities were able to recover the $\mathbb{RT}_0$ velocity $V = (x, y)$ up to machine precision. Finally, we look at Table 3.6 for the generic velocity field test case. As can be seen, for the less distorted meshes, the errors are all less than 7%, regardless of the type of reconstruction, with the A and C velocities performing slightly better than the KR velocities, by 1 - 2 %. However, the

Table 3.5: Relative errors in velocity reconstruction, $\mathbb{RT}_0$ velocity field $V = (x, y)$, fluxes from HHO, $k = 2$.

Mesh	$\frac{\|\|V - V_{\mathrm{KR}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{C}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{A}}\|\|}{\|\|V\|\|}$
Cartesian	7.0803e-13	7.0803e-13	7.0803e-13
Hexahedral	2.3399e-02	7.5597e-14	7.5014e-14
Non-conforming	5.0410e-02	3.5458e-13	3.5458e-13
Kershaw	2.5998e-01	7.8583e-12	7.7132e-12

reconstructed velocities on the very distorted Kershaw type meshes are much worse than those from the less distorted meshes, by almost a factor of 10, regardless of whether we use KR, C, or A velocities. On the other hand, it is noticeable that the C and A velocities are better than the KR velocities, by around 10 - 20%.

Table 3.6: Relative errors in velocity reconstruction, $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$, fluxes from HHO, $k = 2$.

Mesh	$\frac{\|\|V - V_{\mathrm{KR}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{C}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{A}}\|\|}{\|\|V\|\|}$
Cartesian	5.6753e-02	5.6753e-02	5.6753e-02
Hexahedral	6.0423e-02	4.3009e-02	4.2625e-02
Non-conforming	5.3510e-02	4.9766e-02	4.9766e-02
Kershaw	6.1089e-01	4.8196e-01	4.2759e-01

This illustrates the fact that a generic velocity field can hardly be approximated with an $\mathbb{RT}_0$ function created through sub-fluxes over distorted grids. We now try to explore, on the Kershaw type meshes, quadratic $\mathbb{RT}_k$ elements, as discussed in Section 3.2. The $\mathbb{RT}_k$ elements will be constructed in two ways: firstly, since the Kershaw mesh is composed of distorted quadrilaterals, we follow the standard technique of reconstructing the $\mathbb{RT}_k$ element on the reference square $[-1, 1] \times [-1, 1]$, and then performing a Piola transform. Secondly, we note that we are solving the diffusion equation (2.1) via an HHO scheme, and not a mixed finite element method. This tells us that the Piola transform is not really necessary, and we can assume that our reconstructed velocity is in $\mathcal{Q}_{k+1,k} \times \mathcal{Q}_{k,k+1}$ for each cell K, with degrees of freedom described as in (3.18) (i.e. for the construction of the $\mathbb{RT}_k$ functions, we treat both regular/irregular quadrilaterals as though they are rectangles). We will denote these velocities by V_{P} (Piola transformed) and V_{RT} (retains the form of the $\mathbb{RT}_k$ functions) respectively.

Table 3.7: Relative errors in velocity reconstruction, Kershaw mesh, rectangular $\mathbb{RT}_k$ elements, $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$.

	$\frac{\|\|V - V_{\mathrm{P}}\|\|}{\|\|V\|\|}$	$\frac{\|\|V - V_{\mathrm{RT}}\|\|}{\|\|V\|\|}$
$k = 0$	3.3574e-01	4.0903e-01
$k = 1$	4.2444	1.1824e-01
$k = 2$	6.3313	3.5294e-01
$k = 3$	103.9067	2.4938e-02

As expected, due to the huge distortion and hence bad approximation of the rational functions, V_{P} gives a bad approximation to the velocity V. On the other hand, V_{RT} gives a better approximation to the velocity. Upon comparison with the triangular $\mathbb{RT}_0$ elements (see Table 3.6), we see that the additional degree of freedom for a quadrilateral $\mathbb{RT}_0$ function only gives a slight improvement in terms of accuracy, by around 2%. Using a high order approximation, the percentage error is reduced to 11.82% for $k = 1$ and 2.49% for $k = 3$. This supports the observation from Table 3.6 that $\mathbb{RT}_0$ functions cannot, in general, give a good approximation of a generic velocity field over distorted meshes. However, we take note of the strange behavior at $k = 2$, which yields a percentage error which is much larger than the error that was obtained when $k = 1$. To further understand where the problem comes from, we assumed that the velocity field $V = (\pi \sin(\pi x)\cos(\pi y), \pi \cos(\pi x)\sin(\pi y))$ is known, so that the moments in (3.18) are calculated exactly. In this case, the relative errors obtained for the rectangular $\mathbb{RT}_1$ and $\mathbb{RT}_2$ velocities are given by 5.2615e-02 and 9.3167e-02, respectively. Also, as expected, when going for $\mathbb{RT}_3$ velocities, the relative error drops down to 1.6999e-02, which is much better than both $\mathbb{RT}_1$ and $\mathbb{RT}_2$. Similar results were obtained by running tests on other types of velocity fields V. Even with exact moments, $\mathbb{RT}_2$ still performs worse than $\mathbb{RT}_1$, which indicates that $\mathbb{RT}_2$ velocities are not suitable for Kershaw type meshes.

Chapter 4

Characteristic-based schemes for advection dominated PDEs

4.1 Introduction

In this chapter, we start by presenting a time-dependent advection-dominated PDE (4.1), and study some numerical schemes for this equation that are based on characteristic methods. These types of PDEs are encountered in many important fields, such as mathematical models in porous medium flow (e.g. reservoir simulation), and fluid dynamics (e.g. Navier–Stokes equations). A short summary, which includes most of the commonly used numerical schemes for advection–diffusion–reaction models, together with their advantages and disadvantages, have been presented in [49].

In particular, our work focuses on two types of numerical schemes based on characteristic methods, namely the Eulerian Lagrangian Localised Adjoint Method (ELLAM) and the Modified Method of Characteristics (MMOC). The advantages of these schemes lie on the fact that they are based on characteristic methods, and thus capture the advective component of the PDE better than upwinding schemes. Moreover, these schemes are not limited by CFL constraints, and hence large time steps can be taken for numerical simulations. These are usually combined with finite difference (FD), finite element (FE) or finite volume (FV) discretisations, in order to provide a complete numerical scheme for advection–diffusion models. To cite a few examples, the FE–MMOC [37], FE–ELLAM [17], and FV–ELLAM [58], have been used to discretise advection-diffusion models. Other variants of the ELLAM, as well as a summary of its properties, have also been presented in [74]. More recent variants of the ELLAM involve the volume corrected characteristics mixed method (VCCMM) [5, 8]. Aside from the global mass conservation property

of ELLAM, these ensure that local volume conservation is achieved. On the other hand, more recent studies of the MMOC involves MMOC with adjusted advection (MMOCAA) [35]. Compared to the MMOC, MMOCAA has better mass conservation properties, which is usually required for an accurate numerical simulation of models that are related to engineering problems. On the other hand, of particular difficulty in the implementation of ELLAM is an accurate evaluation of integrals involving steep back-tracked functions (see Remark 4.3.4). An inaccurate evaluation of these integrals will yield a loss in mass conservation, which leads to severe overshoots or undershoots around these regions. A fix in order to simplify the evaluation of these integrals, which will preserve the mass conservation property, has recently been proposed in [8].

This chapter focuses on the advective–reactive component of the advection–diffusion–reaction model and the characteristic-based schemes used to discretise this equation. We start by presenting some of the assumptions on the data, and under these assumptions, existence of the flow and some estimates useful for the mass balance analysis is then established in Section 4.2. The ELLAM scheme and the MMOC scheme are then presented in Sections 4.3 and 4.4 respectively. In particular, we present a precise analysis of mass balance errors for the MMOC scheme. In order to minimise the mass balance error brought about by MMOC, as in (4.34), and to avoid the high computational costs associated with steep back-tracked functions for ELLAM (see Remark 4.3.4), we then propose in Section 4.5 a combined ELLAM–MMOC scheme. Having achieved global mass balance, a novel, less expensive adjustment yielding local volume conservation is then proposed. The complete coupled scheme then consists of a characteristic component (the combined ELLAM–MMOC), accompanied by a discretisation of the diffusive terms using the Gradient Discretisation Method (GDM) framework [40]. The complete coupled scheme, named GEM (for GDM–ELLAM–MMOC), therefore presents in one form several possible discretisations of the advection–diffusion–reaction model.

4.1.1 Models

Our objective is to design a robust, characteristic-based numerical scheme for a model of miscible displacement in porous media. This model, described in Section 1.1, involves an elliptic equation for the pressure, and an advection–diffusion–reaction equation for the concentration of the invading fluid. For simplicity, we describe the characteristic-based scheme for the concentration equation without explicitly referring to the pressure equation. We therefore

consider the scalar model

$$\begin{cases} \phi\dfrac{\partial c}{\partial t} + \mathrm{div}(\mathbf{u}c - \Lambda\nabla c) = f(c) & \text{on } Q_T := \Omega \times (0,T) \\ \Lambda\nabla c \cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0,T), \\ c(\cdot,0) = c_{\mathrm{ini}} & \text{on } \Omega, \end{cases} \tag{4.1}$$

in which $T > 0$, Ω is an open bounded domain of $\mathbb{R}^d$ ($d \geq 1$), the diffuson tensor Λ and the velocity $\mathbf{u}$ are given, $\mathbf{u} \cdot \mathbf{n} = 0$ on $\partial\Omega$, and $f(c) = f(c, \boldsymbol{x}, t)$ is a function $\mathbb{R} \times Q_T \to \mathbb{R}$. The unknown $c(\boldsymbol{x}, t)$ represents the amount of material (a fraction) present at $(\boldsymbol{x}, t)$. The characteristic method only deals with the advective part of the model, and will therefore be described on the advection–reaction equation (corresponding to $\Lambda \equiv 0$):

$$\begin{cases} \phi\dfrac{\partial c}{\partial t} + \mathrm{div}(\mathbf{u}c) = f(c) & \text{on } Q_T := \Omega \times (0,T) \\ c(\cdot,0) = c_{\mathrm{ini}} & \text{on } \Omega. \end{cases} \tag{4.2}$$

Note that the boundary is non-characteristic due to the assumption $\mathbf{u} \cdot \mathbf{n} = 0$ on $\partial\Omega$, and thus no boundary conditions need to be enforced in (4.2).

4.1.2 Assumptions on the data, and numerical setting

We assume the following properties:

$$\begin{aligned} &c_{\mathrm{ini}} \in L^\infty(\Omega) \\ &f : \mathbb{R} \times Q_T \to \mathbb{R} \text{ is Lipschitz continuous w.r.t. its first variable} \\ &\text{and } f(0,\cdot,\cdot) \in L^\infty(Q_T) \end{aligned} \tag{4.3a}$$

$$\mathbf{u} \in L^\infty(0,T; L^2(\Omega)^d) \text{ and } \mathrm{div}\mathbf{u} \in L^\infty(Q_T). \tag{4.3b}$$

Our objective in this chapter is to describe numerical methods for the complete model (4.1) in a general setting, to ensure that our design and analysis of ELLAM–MMOC schemes applies to various possible spatial discretisations (e.g. finite-element or finite-volume based). To achieve this, we use, in particular, the Gradient Discretisation Method (GDM)[40] for Neumann boundary conditions, introduced in Chapter 2, which fits in the context of this problem. Although most of our work will be done here on the advective–reactive parts of (4.1), we will demonstrate that the GDM also provides all the required tools to describe ELLAM and MMOC schemes. Since the advection–diffusion–reaction model is time dependent, we need to define a few additional terms in the context of the GDM. In particular, we define a space–time GD:

Definition 4.1.1. *A space–time gradient discretisation is* $\mathcal{D}^T = (\mathcal{D}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ *such that* $\mathcal{D}$ *is a space GD in the sense of Definition 2.1.1,* $0 = t^{(0)} < \cdots < t^{(N)} = T$ *are time steps, and* $\mathcal{I}_\mathcal{D} : L^\infty(\Omega) \to X_\mathcal{D}$ *is an operator used to interpolate initial conditions onto the unknowns.*

As an example, for an HMM gradient discretisation (as in Section 2.2), the interpolant is defined in the following manner: $\mathcal{I}_\mathcal{D} : L^\infty(\Omega) \to X_\mathcal{D}$ is such that

$$\forall \phi \in L^\infty(\Omega), \quad \mathcal{I}_\mathcal{D}(\phi) = ((\phi_K)_{K\in\mathcal{M}}, (\phi_\sigma)_{\sigma\in\mathcal{E}_K}), \text{ with}$$
$$\phi_K = \frac{1}{|K|}\int_K \phi(x)dx \text{ and } \phi_\sigma = 0.$$

Remark 4.1.2. *In the GDM, the interpolant* $\mathcal{I}_\mathcal{D}$ *is usually defined on* $L^2(\Omega)$*; in the context of Problem* (1.1)*, the initial condition is always assumed to be bounded and it is therefore natural to only consider interpolants of initial conditions in* $L^\infty(\Omega)$*.*

The properties of coercivity, limit conformity, compactness, and consistency in Definition 2.7 are naturally extended in the following manner:

Definition 4.1.3. *A sequence of space–time gradient discretisations* $(\mathcal{D}_m^T)_{m\in\mathbb{N}}$ *is coercive, limit-conforming or compact if its underlying sequence of space gradient discretisations satisfy the corresponding property. Finally,* $(\mathcal{D}_m^T)_{m\in\mathbb{N}}$ *is GD-consistent if the underlying sequence of spatial GDs is GD-consistent and if*

- *with* $\delta t_m^{(n+\frac{1}{2})} = t_m^{(n+1)} - t_m^{(n)}$, $\max_{n=0,\dots,N_m-1} \delta t_m^{(n+\frac{1}{2})} \to 0$ *as* $m \to \infty$,
- *for all* $\varphi \in L^\infty(\Omega)$, $(\Pi_{\mathcal{D}_m}\mathcal{I}_{\mathcal{D}_m}\varphi)_{m\in\mathbb{N}}$ *is bounded in* $L^\infty(\Omega)$ *and converges to* φ *in* $L^2(\Omega)$ *as* $m \to \infty$*.*

Finally, we assume in the following that $\mathbf{u}$ is approximated on each time interval $(t^{(n)}, t^{(n+1)})$ by a function

$$\mathbf{u}^{(n+1)} \in L^2(\Omega)^d \text{ such that } \mathrm{div}\mathbf{u}^{(n+1)} \in L^\infty(\Omega). \tag{4.4}$$

In the rest of the chapter, the variables are only made explicit in the integrands when there is a risk of confusion. Otherwise we simply write, e.g., $\int_\Omega q$.

4.2 Existence and some estimates on the flow

To simplify the notations in this section, we write $\mathbf{u}^{(n+1)} = \mathbf{V}$. Key to the definition of characteristic-based schemes is the characteristic equation: For $\boldsymbol{x} \in \Omega$, $t \mapsto F_t(\boldsymbol{x})$ solves

$$\frac{dF_t(\boldsymbol{x})}{dt} = \frac{\mathbf{V}(F_t(\boldsymbol{x}))}{\phi(F_t(\boldsymbol{x}))} \quad \text{for } t \in [-T,T], \qquad F_0(\boldsymbol{x}) = \boldsymbol{x}. \tag{4.5}$$

Associated with the flow equation (4.5) is the advection equation

$$\phi \partial_t w + \mathbf{V} \cdot \nabla w = 0. \tag{4.6}$$

A function w is a solution to such an equation if it satisfies, for all $s,t \in [-T,T]$ such that $s-t \in [-T,T]$ and for a.e. $\boldsymbol{x} \in \Omega$, $w(\boldsymbol{x},t) = w(F_{s-t}(\boldsymbol{x}),s)$. We note that F_t depends on n through $\mathbf{V}$, but this dependency is not explicitly indicated when there is no risk of confusion.

Our leading assumption here is: there is a mesh $\mathcal{M}$ (that is, a partition of Ω into polygonal/polyhedral cells) such that

$$\begin{aligned}
&\phi \text{ is piecewise smooth on } \mathcal{M} \text{ and there exists } \phi_*, \phi^* > 0 \text{ such that} \\
&\qquad \phi_* \le \phi \le \phi^*, \\
&\mathbf{V} \in H(\mathrm{div},\Omega) \text{ is piecewise polynomial on } \mathcal{M}, \\
&\text{There is } \Gamma_{\mathrm{div}} \ge 0 \text{ such that } |\mathrm{div}\mathbf{V}| \le \Gamma_{\mathrm{div}} \text{ on } \Omega, \text{ and } \mathbf{V}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega.
\end{aligned} \tag{4.7}$$

Lemma 4.2.1 (The flow is well-defined)**.** *Under Assumption* (4.7)*, there exists a closed set $\mathcal{C} \subset \Omega$ with zero Lebesgue measure such that, for any $\boldsymbol{x} \in \Omega\backslash\mathcal{C}$, there is a unique Lipschitz-continuous map $t \in [-T,T] \mapsto F_t(\boldsymbol{x}) \in \Omega\backslash\mathcal{C}$ that satisfies* (4.5) *(except at an at most countable number of times for the ODE). Moreover, F_t has classical flows properties: for all $t \in [-T,T]$, $F_t : \Omega\backslash\mathcal{C} \to \Omega\backslash\mathcal{C}$ is a locally Lipschitz-continuous homeomorphism (which can thus be used for changes of variables in integrals), and $F_{t+s} = F_t \circ F_s$ for all $s,t \in [-T,T]$ such that $s+t \in [-T,T]$.*

Proof. By smoothness of $\mathbf{V}$ and ϕ in each cell, the flow $t \mapsto F_t(\boldsymbol{x})$ of $\mathbf{V}/\phi$ can clearly be defined until it reaches a cell boundary. Assume that it reaches at a time $t = t_\sigma$ a cell boundary at a point $\boldsymbol{y}$ that is not a vertex or on an edge of the cell (we use here the 3D nomenclature), that is, $\boldsymbol{y}$ is in the relative interior of a face σ. Denote by H_1 and H_2 the two half-spaces on each side of σ, and by $\mathbf{n}_\sigma$ the normal to σ from H_1 to H_2. Since $\mathbf{V} \in H(\mathrm{div},\Omega)$, $\mathbf{V}\cdot\mathbf{n}_\sigma$ is continuous across σ. The function ϕ being positive, it means that

the *sign*, if not the value, of $(\mathbf{V}/\phi)\cdot\mathbf{n}_\sigma$ is continuous across σ. Assuming for example that $(\mathbf{V}/\phi)_{|H_1}(\boldsymbol{y})\cdot\mathbf{n}_\sigma > 0$, then the flow arrives at $\boldsymbol{y}$ from H_1 and, $(\mathbf{V}/\phi)_{|H_2}(\boldsymbol{y})\cdot\mathbf{n}_\sigma$ being also strictly positive, $t\mapsto F_t(\boldsymbol{x})$ can be restarted from $(t_\sigma,\boldsymbol{y})$ by considering $(\mathbf{V}/\phi)_{|H_2}$ (which drives the flow into H_2). Note that the $H(\mathrm{div})$-property of $\mathbf{V}$ is essential here to ensure that the flow can indeed be continued into H_2, and that the values of $\mathbf{V}/\phi$ at $\boldsymbol{y}$ from H_1 and H_2 do not simultaneously drive the flow in the other domain, thus freezing it at $\boldsymbol{y}$.

Following this process, the flow can be continued as long as it does not cross (or starts from) a vertex/edge or, for a face σ, the set $Z_\sigma = \{\boldsymbol{y}\in\sigma\ :\ \mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma = 0\}$. Let $\mathcal{C}$ be the set consisting of all $\boldsymbol{x}\in\Omega$ whose flow arrive (or starts from) at a vertex/edge, or one of the sets Z_σ. The set $\mathcal{C}$ can be obtained by tracking back on $[-T,T]$, following the process above, the vertices, edges or sets Z_σ (until the flow can no longer be constructed, that is, the trace-back process arrives on a vertex, edge or a set $Z_{\sigma'}$). Since each such set is closed, $\mathcal{C}$ is closed. Moreover, vertices and edges have dimension $d-2$ or less, and are therefore tracked back by the flow into sets of zero d-dimensional measure. Consider now a set Z_σ. Since $\mathbf{V}\cdot\mathbf{n}_\sigma$ is a polynomial, either $Z_\sigma=\sigma$ or Z_σ has dimension $d-2$ or less. In the latter case, as for vertices/edges, its trace-back set has zero d-dimensional measure. If $Z_\sigma=\sigma$, then $\mathbf{V}$ is parallel to σ (whatever the side we consider for the values of $\mathbf{V}$) and the trace-back region of Z_σ is contained in $\overline{\sigma}$, which has zero d-dimensional measure. Hence, $\mathcal{C}$ has zero d-dimensional measure. This reasoning also shows that the flow never crosses the boundary of Ω, since $\mathbf{V}\cdot\mathbf{n}=0$ on $\partial\Omega$.

This construction ensures that, for all $\boldsymbol{x}\notin\mathcal{C}$, the flow $t\mapsto F_t(\boldsymbol{x})\in\Omega\backslash\mathcal{C}$ is well-defined on $[-T,T]$, satisfies the ODEs except at a countable number of points (where it intersects faces), is Lipschitz-continuous (since it is globally continuous and Lipschitz inside each cell, with a Lipschitz constant bounded by $\|\mathbf{V}\|_{L^\infty(\Omega)}/\phi_*$), and satisfies the flow property $F_{t+s}=F_t\circ F_s$. To see that it is locally Lipschitz on $\Omega\backslash\mathcal{C}$ with respect to its base point $\boldsymbol{x}$, we simply have to notice that for $\boldsymbol{x}\notin\mathcal{C}$, by construction of $\mathcal{C}$, there is a ball $B(\boldsymbol{x},\theta)$ centered at $\boldsymbol{x}$ such that, for any $\boldsymbol{y}\in B(\boldsymbol{x},\theta)$, the flow $t\mapsto F_t(\boldsymbol{y})$ travels into the same cells and crosses the same faces as $t\mapsto F_t(\boldsymbol{x})$. Since, in each cell, the flow is Lipschitz-continuous w.r.t. its base point with a uniform Lipschitz constant (because $\mathbf{V}$ and ϕ are smooth in each cell, with bounded derivatives), gluing the Lipschitz estimate thanks to the flow property we can check that $\boldsymbol{y}\mapsto F_t(\boldsymbol{y})$ is Lipschitz continuous on $B(\boldsymbol{x},\theta)$. Note that because the open set $\Omega\backslash\mathcal{C}$ can be disconnected, this does not prove a global Lipschitz property of the flow.

The homoeomorphism property follows from the flow property which shows that, on $\Omega\backslash\mathcal{C}$, $F_t\circ F_{-t}=F_0=\mathrm{Id}$. ∎

Let us now establish some relations and estimates on this flow.

Lemma 4.2.2 (Estimates on the flow)**.** *Under Assumptions* (4.7)*, for a.e.* $\boldsymbol{x} \in \Omega$ *and all* $s \in [-T, T]$*, denoting by* JF_t *the Jacobian determinant of* F_t*,*

$$\int_0^s |JF_t(\boldsymbol{x})|(\mathrm{div}\mathbf{V}) \circ F_t(\boldsymbol{x})dt = \phi(F_s(\boldsymbol{x}))|JF_s(\boldsymbol{x})| - \phi(\boldsymbol{x}) \tag{4.8}$$

and

$$|JF_s(\boldsymbol{x})| \leq C_1(s) := \frac{\phi^*}{\phi_*} \exp\left(\frac{\Gamma_{\mathrm{div}}}{\phi_*}|s|\right). \tag{4.9}$$

Moreover, let $w \geq 0$ *be a solution of* (4.6)*. Then, for all* $s, t \in [-T, T]$ *such that* $s - t \in [-T, T]$*,*

$$\int_\Omega \phi(\boldsymbol{x})w(\boldsymbol{x}, t-s)d\boldsymbol{x} \leq \left(1 + \frac{\Gamma_{\mathrm{div}}C_1(T)}{\phi_*}|s|\right) \int_\Omega \phi(\boldsymbol{x})w(\boldsymbol{x}, t)d\boldsymbol{x} \tag{4.10}$$

and

$$\int_\Omega w(\boldsymbol{x}, t-s)d\boldsymbol{x} \leq \frac{C_1(T)}{\phi_*} \int_\Omega \phi(\boldsymbol{x})w(\boldsymbol{x}, t)d\boldsymbol{x}. \tag{4.11}$$

Proof.

Step 1: we establish the following generalised Liouville formula: for any measurable set $A \subset \Omega$,

$$\frac{d}{dt}\int_{F_t(A)} \phi(\boldsymbol{y})d\boldsymbol{y} = \int_{F_t(A)} \mathrm{div}\mathbf{V}(\boldsymbol{y})d\boldsymbol{y}, \tag{4.12}$$

where the time derivative $\dfrac{d}{dt}$ is taken in the sense of distributions (this also shows that the function $t \mapsto \int_{F_t(A)} \phi(\boldsymbol{y})d\boldsymbol{y}$ belongs to $W^{1,1}(-T, T)$).

Let $v_0 \in C_c^\infty(\Omega)$ and set $v(\boldsymbol{x}, t) = v_0(F_{-t}(\boldsymbol{x}))$. Then v is Lipschitz-continuous with respect to t and, by the flow property, $v(\boldsymbol{x}, t) = v(F_{s-t}(\boldsymbol{x}), s)$. Hence,

$$\partial_t v(\boldsymbol{x}, t) = \nabla v(F_{s-t}(\boldsymbol{x}), s) \cdot \frac{d}{dt}(F_{s-t}(\boldsymbol{x})) = -\nabla v(F_{s-t}(\boldsymbol{x}), s) \cdot \frac{\mathbf{V}(F_{s-t}(\boldsymbol{x}))}{\phi(F_{s-t}(\boldsymbol{x}))}.$$

Given the piecewise regularity assumptions on $\mathbf{V}$ and ϕ, for a.e. $\boldsymbol{x} \in \Omega$ we can let $s \to t$ in the above relation to find $\partial_t v(\boldsymbol{x}, t) = -\nabla v(\boldsymbol{x}, t) \cdot \frac{\mathbf{V}(\boldsymbol{x})}{\phi(\boldsymbol{x})}$. Hence, since $\mathbf{V} \in H_{\mathrm{div}}(\Omega)$ with $\mathbf{V} \cdot \mathbf{n} = 0$ on $\partial\Omega$,

$$\frac{d}{dt}\int_\Omega \phi(\boldsymbol{x})v(\boldsymbol{x}, t)d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x})\partial_t v(\boldsymbol{x}, t)d\boldsymbol{x}$$

$$= -\int_\Omega \nabla v(\boldsymbol{x},t)\cdot \mathbf{V}(\boldsymbol{x})d\boldsymbol{x} = \int_\Omega v(\boldsymbol{x},t)\mathrm{div}\mathbf{V}(\boldsymbol{x})d\boldsymbol{x}.$$

Let us now take a sequence $(v_0^{(n)})_{n\in\mathbb{N}}$ in $C_c^\infty(\Omega)$ that converges a.e. on Ω to the characteristic function $\mathbf{1}_A$ of A, and such that $0 \le v_0^{(n)} \le 1$. The relation above yields

$$\frac{d}{dt}\int_\Omega \phi(\boldsymbol{x})v_0^{(n)}(F_{-t}(\boldsymbol{x}))d\boldsymbol{x} = \int_\Omega v_0^{(n)}(F_{-t}(\boldsymbol{x}))\mathrm{div}\mathbf{V}(\boldsymbol{x})d\boldsymbol{x}. \tag{4.13}$$

As $n\to\infty$, the right-hand side converges (by dominated convergence) to

$$\int_\Omega \mathbf{1}_A(F_{-t}(\boldsymbol{x}))\mathrm{div}\mathbf{V}(\boldsymbol{x})d\boldsymbol{x} = \int_{F_t(A)} \mathrm{div}\mathbf{V}(\boldsymbol{x})d\boldsymbol{x}.$$

The sequence of mappings $t\mapsto \int_\Omega \phi(\boldsymbol{x})v_0^{(n)}(F_{-t}(\boldsymbol{x}))d\boldsymbol{x}$ converge pointwise to

$$t\mapsto \int_\Omega \phi(\boldsymbol{x})\mathbf{1}_A(F_{-t}(\boldsymbol{x}))d\boldsymbol{x} = \int_{F_t(A)}\phi(\boldsymbol{x})d\boldsymbol{x},$$

while remaining bounded. Hence, they converge weakly-$*$ in $L^\infty(-T,T)$. We can therefore pass to the distributional limit in (4.13) to see that (4.12) holds.

Step 2: estimates on JF_t.

Set $A = B(\boldsymbol{x},r)$ a ball of center $\boldsymbol{x}$ and radius r contained in Ω. Integrating (4.12) with respect to time from 0 to s and using a change of variables $\boldsymbol{y} = F_{-t}(\boldsymbol{x})$, we obtain

$$\int_{B(\boldsymbol{x},r)}\phi(F_s(\boldsymbol{y}))|JF_s(\boldsymbol{y})|d\boldsymbol{y} - \int_{B(\boldsymbol{x},r)}\phi(\boldsymbol{y})d\boldsymbol{y} = \int_0^s\int_{B(\boldsymbol{x},r)}|JF_t(\boldsymbol{y})|(\mathrm{div}\mathbf{V})\circ F_t(\boldsymbol{y})dtd\boldsymbol{y}.$$

Dividing by the measure of $B(\boldsymbol{x},r)$ and taking the limit as $r\to 0$, we obtain (4.8) for a.e. $\boldsymbol{x}\in\Omega$, due to the piecewise smoothness of $\mathbf{V}$ and ϕ.

Assume to simplify the writing that $s\ge 0$ and use the assumption on $\mathrm{div}\mathbf{V}$ to deduce from (4.8) that $\phi(F_s(\boldsymbol{x}))|JF_s(\boldsymbol{x})| - \phi(\boldsymbol{x}) \le \Gamma_{\mathrm{div}}\int_0^s |JF_t(\boldsymbol{x})|dt$, and thus that

$$|JF_s(\boldsymbol{x})| \le \frac{\phi^*}{\phi_*} + \frac{\Gamma_{\mathrm{div}}}{\phi_*}\int_0^s |JF_t(\boldsymbol{x})|dt.$$

Use then Gronwall's inequality to obtain (4.9).

Step 3: Estimates on w.

We recall that $w(\boldsymbol{x}, t-s) = w(F_s(\boldsymbol{x}), t)$. Hence, a change of variables and (4.8) yield

$$\int_\Omega \phi(\boldsymbol{x}) w(\boldsymbol{x}, t-s) d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x}) w(F_s(\boldsymbol{x}), t) d\boldsymbol{x} = \int_\Omega w(\boldsymbol{y}, t) \phi(F_{-s}(\boldsymbol{y})) |JF_{-s}(\boldsymbol{y})| d\boldsymbol{y}$$
$$= \int_\Omega w(\boldsymbol{y}, t) \left(\phi(\boldsymbol{y}) + \int_0^{-s} |JF_\rho(\boldsymbol{y})| (\mathrm{div}\mathbf{V}) \circ F_\rho(\boldsymbol{y}) d\rho \right) d\boldsymbol{y}.$$

Estimate (4.10) follows by writing, thanks to (4.9), for a.e. $\boldsymbol{y} \in \Omega$,

$$\left| \int_0^{-s} |JF_\rho(\boldsymbol{y})| (\mathrm{div}\mathbf{V}) \circ F_\rho(\boldsymbol{y}) d\rho \right| \leq \Gamma_{\mathrm{div}} C_1(T) |s| \leq \frac{\Gamma_{\mathrm{div}} C_1(T)}{\phi_*} |s| \phi(\boldsymbol{y}).$$

To establish (4.11), we simply write, still using a change of variables,

$$\int_\Omega w(\boldsymbol{x}, t-s) d\boldsymbol{x} = \int_\Omega w(F_s(\boldsymbol{x}), t) d\boldsymbol{x} = \int_\Omega w(\boldsymbol{y}, t) |JF_{-s}(\boldsymbol{y})| d\boldsymbol{y}$$

and we use (4.9) and $\phi \geq \phi_*$ to conclude. ∎

Corollary 4.2.3 (Generalised Liouville formula)**.** *Under Assumptions* (4.7)*, we have*

$$\frac{d}{dt} \int_{F_t(A)} \phi(\boldsymbol{y}) d\boldsymbol{y} = \int_{F_t(A)} \mathrm{div}\mathbf{V}(\boldsymbol{y}) d\boldsymbol{y} \tag{4.14}$$

for any measurable set A.

4.3 ELLAM scheme for the advection–reaction equation

4.3.1 Motivation

For any sufficiently smooth function φ, the product rule yields

$$\varphi \frac{\partial c}{\partial t} = \frac{\partial (c\varphi)}{\partial t} - c \frac{\partial \varphi}{\partial t}.$$

Hence, (4.2) gives, for any time interval $(t^{(n)}, t^{(n+1)})$,

$$\int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \phi(\boldsymbol{x}) \frac{\partial (c\varphi)}{\partial t}(\boldsymbol{x}, t) d\boldsymbol{x} dt$$
$$- \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega c(\boldsymbol{x}, t) \left[\phi(\boldsymbol{x}) \frac{\partial \varphi}{\partial t}(\boldsymbol{x}, t) + \mathbf{u}(\boldsymbol{x}, t) \cdot \nabla \varphi(\boldsymbol{x}, t) \right] d\boldsymbol{x} dt \tag{4.15}$$
$$= \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(c, \boldsymbol{x}, t) \varphi(\boldsymbol{x}, t) d\boldsymbol{x} dt.$$

To simplify the second term on the left hand side of the above equation, the ELLAM requires that test functions φ satisfy

$$\phi\frac{\partial\varphi}{\partial t}+\mathbf{u}\cdot\nabla\varphi=0 \quad \text{ on } \Omega\times(t^{(n)},t^{(n+1)}), \tag{4.16}$$

with $\varphi(\cdot,t^{(n+1)})$ given. The equation (4.15) then leads to the relation

$$\int_\Omega \phi(\boldsymbol{x})(c\varphi)(\boldsymbol{x},t^{(n+1)})d\boldsymbol{x}-\int_\Omega \phi(\boldsymbol{x})(c\varphi)(\boldsymbol{x},t^{(n)})d\boldsymbol{x} \\ =\int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega f(c,\boldsymbol{x},t)\varphi(\boldsymbol{x},t)d\boldsymbol{x}dt.$$

4.3.2 ELLAM scheme

The ELLAM scheme consists in exploiting the motivation above, in the discrete context of the GDM in which trial and test functions are replaced by reconstructions $\Pi_{\mathcal{C}}$ applied to trial and test vectors in $X_{\mathcal{C}}$.

Definition 4.3.1 (ELLAM scheme)**.** *Let $\mathcal{C}^T$ be a space-time gradient discretisation in the sense of Definition 4.1.1. Using a weighted trapezoid rule with weight $\varpi_n\in[0,1]$ for the time-integration of the source term, the ELLAM scheme for* (4.2) *reads as: find $(c^{(n)})_{n=0,\dots,N}\in X_{\mathcal{C}}^{N+1}$ such that $c^{(0)}=\mathcal{I}_{\mathcal{C}}c_{\rm ini}$ and, for all $n=0,\dots,N-1$, $c^{(n+1)}$ satisfies*

$$\int_\Omega \phi\Pi_{\mathcal{C}}c^{(n+1)}\Pi_{\mathcal{C}}z-\int_\Omega\phi\Pi_{\mathcal{C}}c^{(n)}v_z(t^{(n)}) \\ =\varpi_n\delta t^{(n+\frac12)}\int_\Omega f_nv_z(t^{(n)})+(1-\varpi_n)\delta t^{(n+\frac12)}\int_\Omega f_{n+1}\Pi_{\mathcal{C}}z \qquad \forall z\in X_{\mathcal{C}}, \tag{4.17}$$

where v_z is the solution to

$$\phi\partial_tv_z+\mathbf{u}^{(n+1)}\cdot\nabla v_z=0 \quad \text{ on } (t^{(n)},t^{(n+1)}),\ \text{with } v_z(\cdot,t^{(n+1)})=\Pi_{\mathcal{C}}z\ . \tag{4.18}$$

Here and in the rest of the chapter, we let $f_k:=f(\Pi_{\mathcal{C}}c^{(k)},\cdot,t^{(k)})$ (or with a suitable average over $(t^{(k)},t^{(k+1)})$ if f is not continuous in time).

Remark 4.3.2 (About the time integration)**.** *The velocity field $\mathbf{u}$ was approximated by its value at time $t^{(n+1)}$, given by $\mathbf{u}^{(n+1)}$. Other choices for the approximation of $\mathbf{u}$, such as a centred approximation $\frac12(\mathbf{u}^{(n)}+\mathbf{u}^{(n+1)})$, may be made, but we noticed in our tests that this does not noticeably change the numerical solution. A weighted trapezoid rule is applied for time integration in Definition 4.3.1 for the purpose of achieving mass conservation, as discussed in [5]. More details about the choice of ϖ_n and its dependence on time will be discussed in Chapter 5.*

Define the flow $F_t : \Omega \to \Omega$ such that, for a.e. $\boldsymbol{x} \in \Omega$,

$$\frac{dF_t(\boldsymbol{x})}{dt} = \frac{\mathbf{u}^{(n+1)}(F_t(\boldsymbol{x}))}{\phi(F_t(\boldsymbol{x}))} \quad \text{for } t \in [-T, T], \qquad F_0(\boldsymbol{x}) = \boldsymbol{x}. \tag{4.19}$$

Under Assumption (4.7) with $\mathbf{V} = \mathbf{u}^{(n+1)}$, the existence of this flow is proved in Lemma 4.2.1. The solution to (4.18) is then understood in the sense: for $t \in (t^{(n)}, t^{(n+1)}]$ and a.e. $\boldsymbol{x} \in \Omega$, $v_z(\boldsymbol{x}, t) = \Pi_{\mathcal{C}} z(F_{t^{(n+1)}-t}(\boldsymbol{x}))$. In particular,

$$v_z(\cdot, t^{(n)}) = \Pi_{\mathcal{C}} z(F_{\delta t^{(n+\frac{1}{2})}}(\cdot)). \tag{4.20}$$

For any functions f and g, defining the vector functions $f^{(n,\varpi_n)}$ and g_F by

$$\begin{aligned} f^{(n,\varpi_n)}(\boldsymbol{x}) &:= \Big(\varpi_n f_n, (1-\varpi_n) f_{n+1}\Big), \\ g_F(\boldsymbol{x}) &:= \Big(g(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})), g(\boldsymbol{x})\Big), \end{aligned} \tag{4.21}$$

the time-stepping (4.17) can be rewritten in the condensed form

$$\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} \Pi_{\mathcal{C}} z - \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} v_z(t^{(n)}) = \delta t^{(n+\frac{1}{2})} \int_\Omega f^{(n,\varpi_n)} \cdot (\Pi_{\mathcal{C}} z)_F. \tag{4.22}$$

4.3.3 Physical interpretation

We provide a simple physical interpretation of the ELLAM, by supposing that $\Pi_{\mathcal{C}}$ is a piecewise-constant reconstruction on a given mesh $\mathcal{M}$. We also assume that for each cell $K \in \mathcal{M}$, there is $z_K \in X_{\mathcal{C}}$ such that $\Pi_{\mathcal{C}} z_K = \mathbb{1}_K$. Writing $\Pi_{\mathcal{C}} c^{(k)} = \sum_{K \in \mathcal{M}} c_K^{(k)} \mathbb{1}_K$ and taking z_K as a test function, (4.20) and (4.22) give

$$\begin{aligned} \int_K \phi \Pi_{\mathcal{C}} c^{(n+1)} d\boldsymbol{x} = & \int_\Omega \phi \sum_{M \in \mathcal{M}} c_M^{(n)} \mathbb{1}_M(\boldsymbol{x}) \mathbb{1}_K(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) d\boldsymbol{x} \\ & + \delta t^{(n+\frac{1}{2})} \int_\Omega f^{(n,\varpi_n)} \cdot (\mathbb{1}_K)_F d\boldsymbol{x}, \end{aligned}$$

which reduces to

$$|K|_\phi c_K^{(n+1)} = \sum_{M \in \mathcal{M}} |M \cap F_{-\delta t^{(n+\frac{1}{2})}}(K)|_\phi c_M^{(n)} + \delta t^{(n+\frac{1}{2})} \int_\Omega f^{(n,\varpi_n)} \cdot (\mathbb{1}_K)_F d\boldsymbol{x}, \tag{4.23}$$

where $|E|_\phi = \int_E \phi$ is the available porous volume in a set $E \subset \mathbb{R}^d$. The first term on the right hand side of (4.23) tells us that the amount of material

$c_K^{(n+1)}$ present in a particular cell $K \in \mathcal{M}$ at time $t^{(n+1)}$ is obtained by locating where the material in cell K comes from, hence tracing back the cell K to $F_{-\delta t^{(n+1/2)}}(K)$, measuring how much of the material $c_M^{(n)}$ is taken from each $M \in \mathcal{M}$, and transporting this material into the cell K. These are accompanied by the contribution of the source term f in the particular cell K, which is given by the second term. We note here that this second term has a very similar treatment as the first term, i.e. the contribution that comes from f at time $t^{(n)}$ is determined by the trace-back region associated to cell K.

4.3.4 Mass balance properties

One desirable property for numerical schemes is conservation of mass. Essentially, we want a discrete form of the following equation, obtained by integrating (4.2) over Ω and which tells us that the change in c is dictated by the amount of inflow/outflow given by the source term:

$$\int_\Omega \phi(\boldsymbol{x}) c(\boldsymbol{x}, t^{(n+1)}) d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x}) c(\boldsymbol{x}, t^{(n)}) d\boldsymbol{x} + \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega f(c, \boldsymbol{x}, t) d\boldsymbol{x} dt. \tag{4.24}$$

To evaluate the discrete preservation of mass, we need to define a measure of the mass balance error. Following (4.24) and setting $\mathbf{e} := (1, 1)$, the (discrete) mass balance error is defined by

$$e_{\text{mass}} := \left| \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} d\boldsymbol{x} - \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} d\boldsymbol{x} - \delta t^{(n+\frac{1}{2})} \int_\Omega f^{(n,\varpi_n)} \cdot \mathbf{e}\, d\boldsymbol{x} \right|. \tag{4.25}$$

Remark 4.3.3 (Source term in the mass balance error)**.** *A weighted trapezoidal rule was chosen for the source term in e_{mass} since this is the choice we made for our schemes. Other time quadrature rules could be considered, depending on how this source term is discretised in the considered numerical schemes.*

A mass balance-preserving method is one for which $e_{\text{mass}} = 0$. The ELLAM scheme (4.17) satisfies this property. Indeed, taking $z_1 = \sum_{K \in \mathcal{M}} z_K$, which satisfies $\Pi_{\mathcal{C}} z_1 = 1$ over Ω, as a test function in (4.22) gives $e_{\text{mass}} = 0$.

Remark 4.3.4 (Steep back-tracked functions)**.** *The natural physical interpretation of ELLAM, together with its mass conservation property, seem to indicate that the ELLAM scheme should be preferred over other numerical schemes for the advection equation* (4.2)*. However, for Darcy velocities typically encountered in reservoir engineering, the streamlines of the flow F_t*

concentrate around injection wells, and the functions v_z defined by (4.20) *are then extremely steep in these regions. An accurate approximation of the integral of these functions in cells close to the injection well then requires to track a lot of quadrature points, which is very costly [75]. In some instances, even tracking several points along these regions would not give an accurate depiction of the integral. This is one of the main issues with ELLAM implementations. Fixes have been proposed, but they consist is resorting to a different approach, near the injection wells, than the ELLAM process [8]. We aim at designing a numerical scheme that readily behaves well, without having to implement specific fixes in certain regions. The MMOC will be instrumental to that objective.*

4.4 MMOC scheme for the advection–reaction equation

4.4.1 Motivation

We use the product rule to write $\mathrm{div}(\mathbf{u}c) = c\mathrm{div}(\mathbf{u}) + \mathbf{u}\cdot\nabla c$. By treating $\phi\dfrac{\partial c}{\partial t} + \mathbf{u}\cdot\nabla c$ as a directional derivative in space-time, and denoting by τ the associated characteristic direction, we rewrite (4.2) as follows

$$\zeta\frac{\partial c}{\partial \tau} = f - c\,\mathrm{div}(\mathbf{u}), \tag{4.26}$$

where $\zeta = (\phi^2 + |\mathbf{u}|^2)^{\frac{1}{2}}$. The MMOC then approximates $\dfrac{\partial c}{\partial \tau}$ by performing a finite difference along the characteristic direction:

$$\begin{aligned}\left(\zeta\frac{\partial c}{\partial \tau}\right)(\boldsymbol{x},t) &\approx \zeta(\boldsymbol{x})\frac{c(\boldsymbol{x},t^{(n+1)}) - c(\bar{\boldsymbol{x}},t^{(n)})}{((\boldsymbol{x}-\overline{\boldsymbol{x}})^2 + (\delta t^{(n+\frac{1}{2})})^2)^{\frac{1}{2}}}\\ &= \phi(\boldsymbol{x})\frac{c(\boldsymbol{x},t^{(n+1)}) - c(\bar{\boldsymbol{x}},t^{(n)})}{\delta t^{(n+\frac{1}{2})}}.\end{aligned}$$

Here, $\overline{\boldsymbol{x}} := \boldsymbol{x} - \dfrac{\mathbf{u}^{(n+1)}(\boldsymbol{x})}{\phi(\boldsymbol{x})}\delta t^{(n+\frac{1}{2})}$ is a first order finite difference approximation of the solution $F_t(\boldsymbol{x})$ at time $t = -\delta t^{(n+\frac{1}{2})}$ to the flow equation (4.19). A better approximation is given by taking $F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})$ instead of $\bar{\boldsymbol{x}}$:

$$\left(\zeta\frac{\partial c}{\partial \tau}\right)(\boldsymbol{x},t) \approx \phi(\boldsymbol{x})\frac{c(\boldsymbol{x},t^{(n+1)}) - c(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}),t^{(n)})}{\delta t^{(n+\frac{1}{2})}}. \tag{4.27}$$

By integrating (4.26) over the time interval $[t^{(n)}, t^{(n+1)}]$ and using the approximation (4.27) of the characteristic derivative, we obtain

$$\phi(\boldsymbol{x})\Big(c(\boldsymbol{x}, t^{(n+1)}) - c(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}), t^{(n)})\Big) \approx \int_{t^{(n)}}^{t^{(n+1)}} f(c(\boldsymbol{x}, t), \boldsymbol{x}, t) - c(\boldsymbol{x}, t)\mathrm{div}\mathbf{u}^{(n+1)}(\boldsymbol{x})dt. \quad (4.28)$$

4.4.2 MMOC scheme

The MMOC scheme is written, in the GDM setting, by exploiting (4.28).

Definition 4.4.1 (MMOC scheme)**.** *Given a space-time gradient discretisation $\mathcal{C}^T$ and using a weighted trapezoid rule with weight $\varpi_n \in [0,1]$ for the time-integration of the source term, the MMOC scheme for* (4.2) *reads as: find $(c^{(n)})_{n=0,\dots,N} \in X_{\mathcal{C}}^{N+1}$ such that $c^{(0)} = \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}}$ and, for all $n = 0, \dots, N-1$, $c^{(n+1)}$ satisfies*

$$\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} \Pi_{\mathcal{C}} z d\boldsymbol{x} - \int_\Omega \phi(\boldsymbol{x})(\Pi_{\mathcal{C}} c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))\Pi_{\mathcal{C}} z(\boldsymbol{x}) d\boldsymbol{x} = \delta t^{(n+\frac{1}{2})} \int_\Omega \big[\big(f^{(n,\varpi_n)} - (\Pi_{\mathcal{C}} c)^{(n,\varpi_n)}\mathrm{div}\mathbf{u}^{(n+1)}\big)\cdot \mathbf{e}\big]\Pi_{\mathcal{C}} z d\boldsymbol{x} \qquad \forall z \in X_{\mathcal{C}}, \quad (4.29)$$

where we recall that $\mathbf{e} := (1,1)$, and where we have set (by generalising the notation (4.21)*)*

$$(\Pi_{\mathcal{C}} c)^{(n,\varpi_n)}(\boldsymbol{x}) := \big(\varpi_n \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}), (1-\varpi_n)\Pi_{\mathcal{C}} c^{(n+1)}(\boldsymbol{x})\big). \quad (4.30)$$

4.4.3 Physical interpretation

As with the ELLAM, an interpretation of the MMOC will be provided for the simple case wherein we have a piecewise constant approximation of c. Fixing $K \in \mathcal{M}$ and taking in (4.29) the test vector z_K, such that $\Pi_{\mathcal{C}} z_K = \mathbb{1}_K$, we have

$$\int_K \phi \Pi_{\mathcal{C}} c^{(n+1)} d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x}) \sum_{M\in\mathcal{M}} c_M^{(n)} \mathbb{1}_M(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))\mathbb{1}_K(\boldsymbol{x}) d\boldsymbol{x} + \delta t^{(n+\frac{1}{2})} \int_K f^{(n,\varpi_n)}\cdot \mathbf{e} d\boldsymbol{x} - \delta t^{(n+\frac{1}{2})} \int_K \big[\big(\Pi_{\mathcal{C}} c\big)^{(n,\varpi_n)}\mathrm{div}\mathbf{u}^{(n+1)}\big]\cdot \mathbf{e} d\boldsymbol{x},$$

and thus

$$|K|_\phi c_K^{(n+1)} = \sum_{M\in\mathcal{M}} |F_{\delta t^{(n+\frac{1}{2})}}(M)\cap K|_\phi c_M^{(n)} + \delta t^{(n+\frac{1}{2})}\int_K f^{(n,\varpi_n)}\cdot \mathbf{e}d\boldsymbol{x} - \delta t^{(n+\frac{1}{2})}\int_K \left[(c_K)^{(n,\varpi_n)}\mathrm{div}\mathbf{u}^{(n+1)}\right]\cdot \mathbf{e}d\boldsymbol{x}. \tag{4.31}$$

The first term on the right hand side of the equation tells us that the amount of material $c_K^{(n+1)}$ present in a particular cell $K\in\mathcal{M}$ at time $t^{(n+1)}$ is obtained by taking all cells $M\in\mathcal{M}$, advecting material from each of these cells (by computing the trace-forward regions $F_{\delta t^{(n+1/2)}}(M)$), and determining which portion of each cell flows into K. The second term simply represents the change that comes from the source term f. We note that, unlike in the ELLAM, the contribution of the source term f for the MMOC is taken exactly to be from cell K. By itself, this term tells us that, if the source term f is nonconstant over regions close to one another, the MMOC will give either an excess or miss some amount that has flowed into the region K. The third term in (4.31) represents taking away a fraction of the net inflow/outflow in cell K and, in some sense, attempts to balance out the excessive or missing amount resulting from the second term.

Remark 4.4.2 (Comparison of ELLAM and MMOC schemes)**.** *The ELLAM and the MMOC schemes are equivalent in a cell K if the velocity field is divergence free, and the source term f is constant, in the region $F_{[-\delta t^{(n+\frac{1}{2})},0]}(K) = \cup_{t\in[-\delta t^{(n+\frac{1}{2})},0]}F_t(K)$. Physically, the equivalence is expected, as we are now just comparing the first terms of equations* (4.23) *and* (4.31)*, which both compute the amount of substance that has flowed into cell K. This can be done in two ways: Either we first locate the regions from which the substance has come from (ELLAM), or we let the substances in all cells flow, and determine which ones enter the cell K (MMOC). Mathematically, this equivalence can be established by performing a change of variables in $|F_{\delta t^{(n+\frac{1}{2})}}(M)\cap K|_\phi$ and using the property $\phi(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))|JF_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})| = \phi(\boldsymbol{x})$ if the flow occurs in a region where* $\mathrm{div}\mathbf{u}^{(n+1)} = 0$ *(see* (4.8)*).*

4.4.4 Analysis of mass balance error

Consider the MMOC scheme (4.29). By taking the test function $z_1 = \sum_{K\in\mathcal{M}} z_K$, we have $\Pi_{\mathcal{C}} z_1 = 1$ in Ω and we obtain thus the discrete mass

balance equation

$$\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x})(\Pi_{\mathcal{C}} c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))d\boldsymbol{x} + \delta t^{(n+\frac{1}{2})} \int_\Omega \left(f^{(n,\varpi_n)} - (\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\right) \cdot \mathbf{e} d\boldsymbol{x}. \tag{4.32}$$

From this, we see that one of the disadvantages of MMOC schemes over ELLAM schemes is that, in general, MMOC schemes do not preserve mass. The mass balance error e_{mass} for MMOC is estimated by using (4.32) to substitute $\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} d\boldsymbol{x}$ in (4.25). Performing a change of variables, we obtain

$$\begin{aligned} e_{\text{mass}} &= \left| \int_\Omega \phi(\boldsymbol{x})(\Pi_{\mathcal{C}} c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))d\boldsymbol{x} - \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} d\boldsymbol{x} \right. \\ &\quad \left. - \delta t^{(n+\frac{1}{2})} \int_\Omega \left((\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\right) \cdot \mathbf{e} d\boldsymbol{x} \right| \\ &= \left| \int_\Omega \phi(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}) |JF_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})| d\boldsymbol{x} - \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} d\boldsymbol{x} \right. \\ &\quad \left. - \delta t^{(n+\frac{1}{2})} \int_\Omega \left((\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\right) \cdot \mathbf{e} d\boldsymbol{x} \right|. \end{aligned}$$

Using equation (4.8) with $s = \delta t^{(n+\frac{1}{2})}$ in Lemma 4.2.2, we see that

$$\phi(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))|JF_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})| - \phi(\boldsymbol{x}) = \int_0^{\delta t^{(n+\frac{1}{2})}} |JF_t(\boldsymbol{x})| (\mathrm{div}\mathbf{u}^{(n+1)}) \circ F_t(\boldsymbol{x}) dt.$$

Hence,

$$\begin{aligned} e_{\text{mass}} &= \left| \int_\Omega \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}) \int_0^{\delta t^{(n+\frac{1}{2})}} |JF_t(\boldsymbol{x})| (\mathrm{div}\mathbf{u}^{(n+1)}) \circ F_t(\boldsymbol{x}) dt d\boldsymbol{x} \right. \\ &\quad \left. - \delta t^{(n+\frac{1}{2})} \int_\Omega \left((\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\right) \cdot \mathbf{e} d\boldsymbol{x} \right| \\ &= \left| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \Pi_{\mathcal{C}} c^{(n)}(F_{-t}(\boldsymbol{x})) \mathrm{div}\mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \right. \\ &\quad \left. - \delta t^{(n+\frac{1}{2})} \int_\Omega \left((\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\right) \cdot \mathbf{e} d\boldsymbol{x} \right|. \end{aligned} \tag{4.33}$$

By the triangle inequality and recalling the definition (4.30) of $(\Pi_{\mathcal{C}} c)^{(n,\varpi_n)}$, we infer

$$e_{\text{mass}} \leq \varpi_n \left| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \Big(\Pi_{\mathcal{C}} c^{(n)}(F_{-t}(\boldsymbol{x})) - \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}) \Big) \text{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \right|$$
$$+ (1-\varpi_n) \left| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \big(\Pi_{\mathcal{C}} c^{(n)}(F_{-t}(\boldsymbol{x})) - \Pi_{\mathcal{C}} c^{(n+1)}(\boldsymbol{x}) \big) \text{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \right|. \tag{4.34}$$

This estimate shows that the mass balance error e_{mass} is minimal when $\delta t^{(n+\frac{1}{2})}$ tends to 0 (as $F_{-t} \to Id$ as $t \to 0$) or if the approximate amount of substance c in the trace-back of the non divergence-free regions, denoted by U, is almost constant. More precisely,

$$\Pi_{\mathcal{C}} c^{(n)} = \Pi_{\mathcal{C}} c^{(n+1)} = \text{Const} \quad \text{on } F_{[-\delta t^{(n+\frac{1}{2})},0]}(U),$$

with

$$U = \{\boldsymbol{x} \in \Omega : \text{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) \neq 0\}.$$

Remark 4.4.3 (Conservation of mass for the MMOC)**.** *Estimate* (4.34) *shows that if the velocity field is divergence free then the MMOC scheme conserves mass, which is consistent with Remark 4.4.2.*

Remark 4.4.4 (Forward tracking and cost near the injection cells)**.** *Contrary to the ELLAM, the MMOC requires to forward-track test functions (see* (4.31) *in the case of piecewise constant approximations). Hence, in the MMOC, functions whose support is near the injection wells are not backward tracked into the injection cells, which makes them very steep and difficult to integrate (see Remark 4.3.4), but forward tracked far from these cells into non-steep functions that are easier to integrate.*

4.5 A combined ELLAM–MMOC scheme for the advection–reaction equation

Here, we propose a combined ELLAM–MMOC scheme, to benefit from the mass balance property of the ELLAM and mitigate its costly implementation near the injection wells by using the MMOC method, much less expensive in these regions.

We start by applying a pure ELLAM scheme over the first few time steps, until c is almost constant in areas near the non divergence-free regions. After which, we do a hybrid ELLAM–MMOC scheme, where we apply MMOC over

these areas, and ELLAM elsewhere. The interest of such a scheme is twofold. First, the computational cost is reduced compared to a pure ELLAM scheme as we no longer have to compute integrals of steep functions (see Remark 4.3.4). Second, upon using MMOC only in regions where div$\mathbf{u} = 0$ or c is already almost constant, no mass balance error occurs. This combined scheme removes the main disadvantages of both methods.

4.5.1 Presentation of the ELLAM–MMOC scheme

Take α a function of the space variable, write $c = \alpha c + (1-\alpha)c$ and decompose the model (4.2) into

$$\phi\frac{\partial(\alpha c)}{\partial t} + \mathrm{div}((\alpha c)\mathbf{u}) + \phi\frac{\partial((1-\alpha)c)}{\partial t} + \mathrm{div}(((1-\alpha)c)\mathbf{u}) = \alpha f + (1-\alpha)f. \quad (4.35)$$

Discretise this by applying ELLAM (4.17) on the first part αc (and αf) and MMOC (4.29) on the second part $(1-\alpha)c$ (and $(1-\alpha)f$). In this case, we define $\Pi_{\mathcal{C}}(\alpha c)^{(n)}$ as $\alpha\Pi_{\mathcal{C}}c^{(n)}$.

Remark 4.5.1 (Interpretation of the combined ELLAM–MMOC)**.** *An interpretation can be given by considering $c_1 = \alpha c$ and $c_2 = (1-\alpha)c$ as two miscible fluids (that are also miscible in their surroundings) that do not react with each other, and are advected by the velocity* $\mathbf{u}$*. We can consider the combination of these two as one single fluid with concentration c, that is advected at velocity* $\mathbf{u}$ *(one can also consider that c_1 is made of red molecules, c_2 of green molecules, in which case the combination c is yellow; to advect this yellow fluid, one can advect the red molecules with* $\mathbf{u}$ *and the green ones with* $\mathbf{u}$ *too). The presentations* (4.2) *or* (4.35) *correspond to one or the other of these interpretations: do we want to consider both fluids together, or do we treat them separately. For the numerical method, it consists in applying ELLAM on one and MMOC on the other.*

The following definition summarises the combined ELLAM–MMOC scheme.

Definition 4.5.2 (ELLAM–MMOC scheme)**.** *Given a space-time gradient discretisation $\mathcal{C}^T$ and using a weighted trapezoid rule with weight $\varpi_n \in [0,1]$ for the time-integration of the source term, the ELLAM–MMOC scheme for* (4.2) *reads as: find $(c^{(n)})_{n=0,\dots,N} \in X_{\mathcal{C}}^{N+1}$ such that $c^{(0)} = \mathcal{I}_{\mathcal{C}}c_{\mathrm{ini}}$ and, for all*

$n = 0, \ldots, N-1$, $c^{(n+1)}$ *satisfies*

$$\begin{aligned}
&\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} \Pi_{\mathcal{C}} z d\boldsymbol{x} - \int_\Omega \phi(\boldsymbol{x}) \alpha(\boldsymbol{x}) \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}) \Pi_{\mathcal{C}} z(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) d\boldsymbol{x} \\
&- \int_\Omega \phi(\boldsymbol{x}) \left[(1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right](F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \Pi_{\mathcal{C}} z(\boldsymbol{x}) d\boldsymbol{x} \\
&\quad = \delta t^{(n+\frac{1}{2})} \int_\Omega \alpha f^{(n,\varpi_n)} \cdot (\Pi_{\mathcal{C}} z)_F + \delta t^{(n+\frac{1}{2})} \int_\Omega \left[(1-\alpha) f^{(n,\varpi_n)} \cdot \mathbf{e}\right] \Pi_{\mathcal{C}} z \\
&\qquad - \delta t^{(n+\frac{1}{2})} \int_\Omega \left[(1-\alpha) \mathrm{div} \mathbf{u}^{(n+1)} \left(\Pi_{\mathcal{C}} c\right)^{(n,\varpi_n)} \cdot \mathbf{e}\right] \Pi_{\mathcal{C}} z \qquad \forall z \in X_{\mathcal{C}}.
\end{aligned} \tag{4.36}$$

4.5.2 Analysis of mass balance error

Taking $z_1 = \sum_{K \in \mathcal{M}} z_K$ (so that $\Pi_{\mathcal{C}} z_1 = 1$ in Ω) in (4.36) and plugging into (4.25), the mass balance error e_{mass} of the ELLAM–MMOC scheme is estimated as follows:

$$\begin{aligned}
e_{\text{mass}} &= \Bigg| \int_\Omega \phi(\boldsymbol{x}) \left[(1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right](F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) d\boldsymbol{x} \\
&\qquad - \int_\Omega \phi(1-\alpha)\Pi_{\mathcal{C}} c^{(n)} d\boldsymbol{x} - \delta t^{(n+\frac{1}{2})} \int_\Omega (1-\alpha) \mathrm{div} \mathbf{u}^{(n+1)} \left(\Pi_{\mathcal{C}} c\right)^{(n,\varpi_n)} \cdot \mathbf{e}\, d\boldsymbol{x} \Bigg| \\
&= \Bigg| \int_\Omega \phi \left((F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}) \right) \left[(1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right](\boldsymbol{x}) |J F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})| d\boldsymbol{x} \\
&\qquad - \int_\Omega \phi(1-\alpha)\Pi_{\mathcal{C}} c^{(n)} d\boldsymbol{x} - \delta t^{(n+\frac{1}{2})} \int_\Omega (1-\alpha) \mathrm{div} \mathbf{u}^{(n+1)} \left(\Pi_{\mathcal{C}} c\right)^{(n,\varpi_n)} \cdot \mathbf{e}\, d\boldsymbol{x} \Bigg|.
\end{aligned}$$

By using (4.8) and doing a change of variable F_{-t} as in (4.33), we obtain

$$\begin{aligned}
e_{\text{mass}} &= \Bigg| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \left[(1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right](F_{-t}(\boldsymbol{x})) \mathrm{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \\
&\qquad - \delta t^{(n+\frac{1}{2})} \int_\Omega (1-\alpha) \mathrm{div} \mathbf{u}^{(n+1)} \left(\Pi_{\mathcal{C}} c\right)^{(n,\varpi_n)} \cdot \mathbf{e}\, d\boldsymbol{x} \Bigg|. \\
&\leq \varpi_n \Bigg| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \Big[\left((1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right)(F_{-t}(\boldsymbol{x})) \\
&\qquad\qquad - \left((1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right)(\boldsymbol{x}) \Big] \mathrm{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \Bigg| \\
&\quad + (1-\varpi_n) \Bigg| \int_0^{\delta t^{(n+\frac{1}{2})}} \int_\Omega \Big[\left((1-\alpha)\Pi_{\mathcal{C}} c^{(n)}\right)(F_{-t}(\boldsymbol{x})) \\
&\qquad\qquad - \left((1-\alpha)\Pi_{\mathcal{C}} c^{(n+1)}\right)(\boldsymbol{x}) \Big] \mathrm{div} \mathbf{u}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} dt \Bigg|.
\end{aligned}$$

Hence, the mass balance error e_{mass} of the ELLAM–MMOC scheme is minimal when $\delta t^{(n+\frac{1}{2})}$ tends to 0 or, setting $U = \{\boldsymbol{x} \in \Omega \,:\, \text{div}\mathbf{u}^{(n+1)}(\boldsymbol{x}) \neq 0\}$, if

$$\big(1-\alpha\big)\Pi_{\mathcal{C}} c^{(n)} \approx \big(1-\alpha\big)\Pi_{\mathcal{C}} c^{(n+1)} \approx \text{Const} \quad \text{on } F_{[-\delta t^{(n+\frac{1}{2})},0]}(U).$$

Remark 4.5.3 (mass conserving α)**.** *In particular, mass conservation is achieved if*

- $\alpha = 1$ *on* $F_{[-\delta t^{(n+\frac{1}{2})},0]}(U)$ *(that is, pure ELLAM is used on the trace-back of non-divergence free regions), or*
- $\Pi_{\mathcal{C}} c^{(n)} \approx \Pi_{\mathcal{C}} c^{(n+1)} \approx C_1$ *and* $\alpha \approx C_2$*, where each* C_i *is a constant, on*

$$D := \{\boldsymbol{x} \in \Omega \,:\, \alpha(\boldsymbol{x}) \neq 1\} \cap F_{[-\delta t^{(n+\frac{1}{2})},0]}(U)$$

(that is, if MMOC is used –partially or entirely– on a domain D *that is inside the trace-back of non-divergence free regions, then the approximate concentration should almost be constant and stationary on* D*, and* α *should also be constant on* D*).*

4.5.3 Implementation for piecewise constant test functions

As with the ELLAM and MMOC, we consider a piecewise constant approximation for c. Then, considering the ELLAM–MMOC scheme in (4.36), we write $\Pi_{\mathcal{C}} c^{(n)} = \sum_{M\in\mathcal{M}} c_M^{(n)} \mathbb{1}_M$ and find $(c^{(n)})_{n=0,\dots,N} \in X_{\mathcal{C}}^{N+1}$ such that $c^{(0)} = \mathcal{I}_{\mathcal{C}} c_{\text{ini}}$ and, for all $n = 0, \dots, N-1$, $c^{(n+1)}$ satisfies

$$\begin{aligned}
&\int_\Omega \phi c_K^{(n+1)} \mathbb{1}_K d\boldsymbol{x} - \int_\Omega \phi(\boldsymbol{x})\alpha(\boldsymbol{x}) \sum_{M\in\mathcal{M}} c_M^{(n)} \mathbb{1}_M(\boldsymbol{x}) \mathbb{1}_K(F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))d\boldsymbol{x} \\
&- \int_\Omega \phi(\boldsymbol{x})\,(1-\alpha)\,(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \sum_{M\in\mathcal{M}} c_M^{(n)} \mathbb{1}_M(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \mathbb{1}_K(\boldsymbol{x}) d\boldsymbol{x} \\
&\quad = \delta t^{(n+\frac{1}{2})} \int_\Omega \alpha f^{(n,\varpi_n)} \cdot (\mathbb{1}_K)_F + \delta t^{(n+\frac{1}{2})} \int_\Omega \big[(1-\alpha) f^{(n,\varpi_n)} \cdot \mathbf{e}\big] \mathbb{1}_K \\
&\qquad - \delta t^{(n+\frac{1}{2})} \int_\Omega \big[(1-\alpha)\text{div}\mathbf{u}^{(n+1)} \big(\Pi_{\mathcal{C}} c\big)^{(n,\varpi_n)} \cdot \mathbf{e}\big] \mathbb{1}_K \quad \forall K \in \mathcal{M}.
\end{aligned}$$

Assume that α is piecewise constant on $\mathcal{M}$ and only takes the values 0 and 1. Each cell $M \in \mathcal{M}$ can then be classified as $\mathcal{M}_{\text{ELLAM}}$ (corresponding to

$\alpha = 1$) or $\mathcal{M}_{\mathrm{MMOC}}$ (corresponding to $\alpha = 0$). The above relation is then re-written

$$\begin{aligned} & c_K^{(n+1)}|K|_\phi \\ & - \sum_{M\in\mathcal{M}_{\mathrm{ELLAM}}} c_M^{(n)}|M \cap F_{-\delta t^{(n+\frac{1}{2})}}(K)|_\phi - \sum_{M\in\mathcal{M}_{\mathrm{MMOC}}} c_M^{(n)}|F_{\delta t^{(n+\frac{1}{2})}}(M) \cap K|_\phi \\ & \quad = \delta t^{(n+\frac{1}{2})} \int_\Omega \alpha f^{(n,\varpi_n)} \cdot (\mathbb{1}_K)_F + \delta t^{(n+\frac{1}{2})} \int_\Omega \big[(1-\alpha) f^{(n,\varpi_n)} \cdot \mathbf{e}\big] \mathbb{1}_K \\ & \qquad - \delta t^{(n+\frac{1}{2})} \int_\Omega \big[(1-\alpha)\mathrm{div}\mathbf{u}^{(n+1)} \big(\Pi_{\mathcal{C}} c\big)^{(n,\varpi_n)} \cdot \mathbf{e}\big] \mathbb{1}_K \quad \forall K \in \mathcal{M}. \end{aligned} \tag{4.37}$$

If $K \in \mathcal{M}_{\mathrm{ELLAM}}$, then we only need to compute the integral of the first term on the right hand side of (4.37) since the latter terms will be zero. Otherwise, only the second and third terms are computed. These are approximated by taking the average values of f and $\mathrm{div}\mathbf{u}^{(n+1)}$ on the respective cells.

We will demonstrate in Section 5.5.1 that, with a proper choice of α, the combined ELLAM–MMOC scheme can be implemented with an equivalent or cheaper computational cost than ELLAM, with reduced overshoots compared to ELLAM, and does not degrade the mass conservation properties (contrary to MMOC).

4.5.4 Comparison with the MMOCAA

Of particular interest is a comparison with the MMOC scheme with adjusted advection (MMOCAA), first introduced in [35]. The MMOCAA is a modification of MMOC designed to conserve the discrete mass. Simply stated, the modification consists of perturbing the foot $F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})$ of the characteristic by a term of order $O((\Delta t)^2)$. Fixing $\eta \in (0,1)$, set

$$\begin{aligned} \bar{\boldsymbol{x}}^+ &= F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}) + \eta \frac{\mathbf{u}^{(n+1)}(\boldsymbol{x})}{\phi(\boldsymbol{x})} (\delta t^{(n+\frac{1}{2})})^2 \\ \bar{\boldsymbol{x}}^- &= F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}) - \eta \frac{\mathbf{u}^{(n+1)}(\boldsymbol{x})}{\phi(\boldsymbol{x})} (\delta t^{(n+\frac{1}{2})})^2. \end{aligned}$$

For simplicity of notation, we denote the defect of mass balance for the MMOC scheme d_{mass} to be

$$\begin{aligned} d_{\mathrm{mass}} := & \int_\Omega \Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}) - \int_\Omega \phi(\boldsymbol{x}) (\Pi_{\mathcal{C}} c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \\ & - \delta t^{(n+\frac{1}{2})} \int_\Omega \big((\Pi_{\mathcal{C}} c)^{(n,\varpi_n)} \mathrm{div}\mathbf{u}^{(n+1)}\big) \cdot \mathbf{e} d\boldsymbol{x}. \end{aligned}$$

We then define

$$\widehat{\Pi_{\mathcal{C}}c^{(n)}}(\boldsymbol{x}) := \begin{cases} \max(\Pi_{\mathcal{C}}c^{(n)}(\bar{\boldsymbol{x}}^+), \Pi_{\mathcal{C}}c^{(n)}(\bar{\boldsymbol{x}}^-)) & \text{if } d_{\text{mass}} \leq 0 \\ \min(\Pi_{\mathcal{C}}c^{(n)}(\bar{\boldsymbol{x}}^+), \Pi_{\mathcal{C}}c^{(n)}(\bar{\boldsymbol{x}}^-)) & \text{otherwise.} \end{cases}$$

In order to enforce a discrete conservation of mass, the term $\Pi_{\mathcal{C}}c(F_{-\delta t^{(n+\frac{1}{2})}}, t^{(n)})$ in (4.29) is then replaced by

$$\Pi_{\mathcal{C}}c_\gamma(\boldsymbol{x}, t^{(n)}) := \gamma\Pi_{\mathcal{C}}c(F_{-\delta t^{(n+\frac{1}{2})}}, t^{(n)}) + (1-\gamma)\widehat{\Pi_{\mathcal{C}}c^{(n)}}(\boldsymbol{x}),$$

where γ is chosen so that d_{mass}, with $\Pi_{\mathcal{C}}c(F_{-\delta t^{(n+\frac{1}{2})}}, t^{(n)})$ replaced by $\Pi_{\mathcal{C}}c_\gamma(\boldsymbol{x}, t^{(n)})$, is equal to 0. For a more detailed presentation and implementation of the MMOCAA, we refer the reader to [35, 36, 61].

It should be noted that, contrary to the underlying principles of the ELLAM–MMOC scheme, there is no physical justification for using this parameter γ to enforce the mass conservation (that is, $d_{\text{mass}} = 0$). Moreover, in some instances, at the first few time steps of a simulation,

$$\int_\Omega \phi(\boldsymbol{x})(\Pi_{\mathcal{C}}c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))\Pi_{\mathcal{C}}z(\boldsymbol{x})d\boldsymbol{x} = \int_\Omega \phi(\boldsymbol{x})\widehat{\Pi_{\mathcal{C}}c^{(n)}}(\boldsymbol{x})\Pi_{\mathcal{C}}z(\boldsymbol{x})d\boldsymbol{x}, \tag{4.38}$$

and thus mass conservation cannot be achieved for any γ [36]. Also, in order to be able to determine the proper value for γ, one needs to evaluate both

$$\int_\Omega \phi(\boldsymbol{x})(\Pi_{\mathcal{C}}c^{(n)})(F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}))\Pi_{\mathcal{C}}z(\boldsymbol{x})d\boldsymbol{x} \text{ and } \int_\Omega \phi(\boldsymbol{x})\widehat{\Pi_{\mathcal{C}}c^{(n)}}(\boldsymbol{x})\Pi_{\mathcal{C}}z(\boldsymbol{x})d\boldsymbol{x}.$$

Evaluating these integrals is the most expensive part of the scheme since it involves tracking points along the characteristics (as well as implementing a proper quadrature rule). For piecewise constant test functions, this involves taking intersections of polygonal regions. The ELLAM–MMOC method, in most cases, only requires one evaluation of an integral of this type where MMOCAA requires two evaluations. Hence, in general, if N is the number of cells, ELLAM–MMOC requires the computation of only N such integrals whereas MMOCAA requires $2N$ such integrals.

4.6 Details on the implementation

4.6.1 Approximate trace-back region, and tracking points through vertices

To compute the regions of intersection in equations (4.23), (4.31), (4.37), a precise description of the trace-back and trace-forward regions $F_{-\delta t^{(n+\frac{1}{2})}}(K)$

and $F_{-\delta t^{(n+\frac{1}{2})}}(M)$, respectively, is needed. However, in general, we cannot get an exact representation of these regions. Hence, for each cell K, the trace-back region $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ is approximated by a polygonal region $\widetilde{K}$ in the following manner: we select points $(\boldsymbol{x}_i)_{i=1,\ldots,\ell_K}$ along the boundary of K (at least all the vertices and edge midpoints are selected), we solve (4.19) starting from any of these points, thus getting curves $(\widehat{\boldsymbol{x}}_i)_{i=1,\ldots,\ell_K} := F_{[-\delta t^{(n+\frac{1}{2})},0]}(\boldsymbol{x}_i)_{i=1,\ldots,\ell_K}$, and we approximate $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ by the polygon defined by the points $F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}_i)_{i=1,\ldots,\ell_K}$. Figure 4.1 (right) gives an illustration of the approximate trace-back region $\widetilde{K}$ obtained by tracing the vertices, together with the edge midpoints of the cell K.

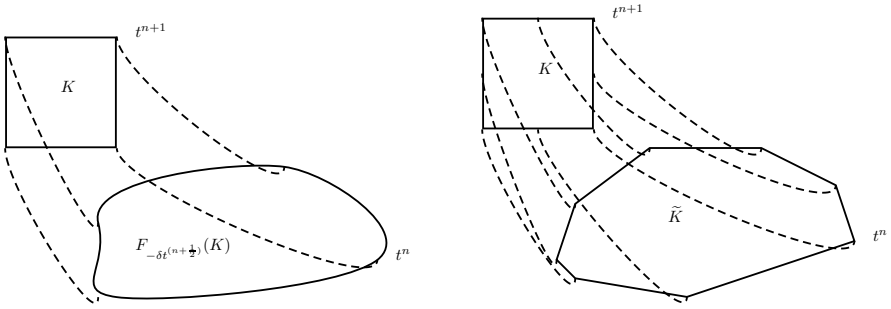


Figure 4.1: Trace-back region $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ (left: exact; right: approximation).

We illustrate here the procedure when the reconstructed velocity $\mathbf{u}^{(n+1)}$ is an $\mathbb{RT}_0$ function on a triangular sub-mesh; similar procedures apply for other types of $H(\mathrm{div},\Omega)$ elements (such as quadrilateral $\mathbb{RT}_k$ elements). Tracking a point through (4.19) is naturally done cell-by-cell, using the value of $\mathbf{u}^{(n+1)}$ in a cell K as long as $\widehat{\boldsymbol{x}}$ stays in K, and then, when $\widehat{\boldsymbol{x}}$ exits K to enter a cell L, continuing the tracking by using the value of $\mathbf{u}^{(n+1)}$ in L. This type of tracking, which determines the location at which a tracked point exits a cell K by choosing the minimal time of flight, was first implemented by Pollock [70] on meshes characterised by orthogonal grid blocks, e.g. Cartesian meshes. Pollock's algorithm was then extended to more generic types of cells in [71], and further improved in [63]. Because the fluxes of $\mathbf{u}^{(n+1)}$ are continuous across the edges, this tracking procedure ensures that a point will never do a U-turn, that is, $-\mathbf{u}^{(n+1)}_{|L}$ (we use $-\mathbf{u}^{(n+1)}$ since we are tracking backwards) will not force $\widehat{\boldsymbol{x}}$ to re-enter K before even entering L (this would in effect freeze $\widehat{\boldsymbol{x}}$ on the interface between K and L).

During this tracking, special care must be taken with points that start or pass through a vertex. An initial position $\boldsymbol{x}$ corresponding to a vertex

is involved with several triangles, and could thus be initially tracked using any of the Darcy velocities in these triangles (see Fig. 4.2). To avoid non-physical initial tracking, we compute the Darcy velocities at the vertex $\boldsymbol{x}$ in each of the triangles involved with it. Picking one of these Darcy velocities at random is not acceptable, since if its opposite vector points outside the corresponding triangle, this means that $\boldsymbol{x}$ would never be tracked back inside this triangle, and that the chosen Darcy velocity is thus not the correct one.

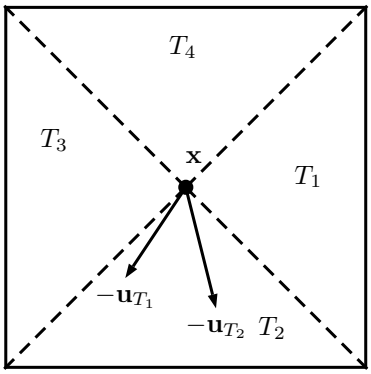


Figure 4.2: Choosing the proper triangle to initialize the tracking.

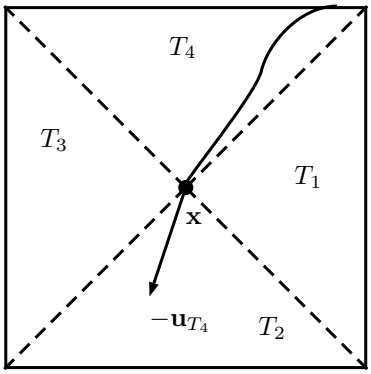


Figure 4.3: Choosing the proper triangle to continue the tracking. Here $\mathbf{u}_{T_4}$ represents the vector that we obtain by computing the Darcy velocity on point $\boldsymbol{x}$ using the reconstructed velocity at T_4.

We therefore loop over the triangles T_1, T_2, etc. until we reach a triangle T_n such that $-\mathbf{u}_{|T_n}^{(n+1)}(\boldsymbol{x})$ points inside T_n. In Figure 4.2, this corresponds to triangle T_2. As a note, regardless of the mesh, our numerical tests suggest that such a triangle always exists. For points that will be tracked into a vertex at some time in $[t^{(n)}, t^{(n+1)})$ (see Fig. 4.3), we consider the negative

of the Darcy velocity in the triangle it came from (in this case, T_4) and determine which triangle it points into (in this case, it points into triangle T_2); the tracking is then continued based on the reconstructed velocity in this latter triangle.

4.6.2 Local volume conservation

In general, the polygonal approximation $\widetilde{K}$ of the tracked region $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ will not be able to preserve its volume, i.e. $|\widetilde{K}| \neq |F_{-\delta t^{(n+\frac{1}{2})}}(K)|$. However, the equality of these volumes is essential in numerical simulations to preserve accuracy. To illustrate this point, consider the simple case of a divergence free velocity field in (4.2), with $\phi = 1, f = 0$ and $c_{\text{ini}} = 1$. In this test case, the exact solution is given by $c(\boldsymbol{x}, t) = 1$. In theory, upon implementing an ELLAM scheme with piecewise constant approximations for the unknown c, we should have the following simplified form of (4.23) at the first time step:

$$|K|c_K^{(1)} = \sum_{M \in \mathcal{M}} |M \cap F_{-\delta t^{(n+\frac{1}{2})}}(K)|(c_{\text{ini}})_M.$$

However, due to the approximation of the trace-back region, we only have

$$|K|c_K^{(1)} = \sum_{M \in \mathcal{M}} |M \cap \widetilde{K}|(c_{\text{ini}})_M = |\widetilde{K}|$$

and thus

$$c_K^{(1)} = \frac{|\widetilde{K}|}{|K|} \neq 1$$

in general. This example shows that an inaccurate approximation of the volume of the tracked cell renders the numerical scheme unable to recover constant solutions. Hence, we need to perform some adjustments on the polygonal region $\widetilde{K}$ in order to yield $|\widetilde{K}| = |F_{-\delta t^{(n+\frac{1}{2})}}(K)|$, which we shall define as the *local volume constraint* for K.

Remark 4.6.1. *Tracking more than one point along each edge of every cell $K \in \mathcal{M}$ gives a better polygonal approximation $\widetilde{K}$ of the trace-back region $F_{-\delta t^{(n+\frac{1}{2})}}(K)$. However, in general, the polygonal approximation still does not satisfy $|\widetilde{K}| = |F_{-\delta t^{(n+\frac{1}{2})}}(K)|$, and hence the local volume constraint is still violated.*

In order to achieve local volume conservation, [6] adjusts the tracked edge midpoints for each of the cells in the mesh. This was shown to work for square

cells, by choosing to adjust the edge midpoints in either a row-wise, column-wise or a staircase-like pattern. An illustration is shown in Figure 4.4 on how the adjustments are made in a column-wise or row-wise manner. Consider, for example, a column-wise adjustment. After tracking the vertices of the square cells and their midpoints, we have a collection of trace-back regions $\widetilde{K}_i$ which correspond to the cells $K_i \in \mathcal{M}$. The adjustments are made in the following order: Beginning from $\widetilde{K}_1$, which corresponds to the bottom left cell (see Figure 4.4), we adjust the midpoint between cells $\widetilde{K}_1$ and $\widetilde{K}_2$ so that the volume of $\widetilde{K}_1$ is correct. We then proceed upwards to the next cell $\widetilde{K}_2$ and adjust its upper midpoint (the one that lies between $\widetilde{K}_2$ and $\widetilde{K}_7$) until we reach the top of the domain (in this case $\widetilde{K}_7$). This top element is not yet adjusted for volume balance. Instead, we move to the next right column and repeat (for $\widetilde{K}_3$ moving to $\widetilde{K}_4$, and $\widetilde{K}_5$ to $\widetilde{K}_6$). Finally, we adjust the right-hand side midpoints of the top row, starting from the left and working right ($\widetilde{K}_7$ to $\widetilde{K}_8$). The final cell $\widetilde{K}_9$ needs no adjustment, due to global volume conservation. The row-wise adjustment is performed in a similar manner. One of the main difficulties in implementing this algorithm is the fact that a row-wise or column-wise adjustment of tracked midpoints may not work. In particular, it was mentioned in [5] that the ordering of the cells (to be adjusted) need to be determined in such a way that the volume errors cancel out each other upon moving through the cells $K \in \mathcal{M}$; otherwise, volume errors may build up in the last few cells. The extension of this algorithm to a generic or unstructured mesh is not trivial (and possibly not doable). Moreover, there is no guarantee that such adjustments will terminate or yield a valid mesh configuration.

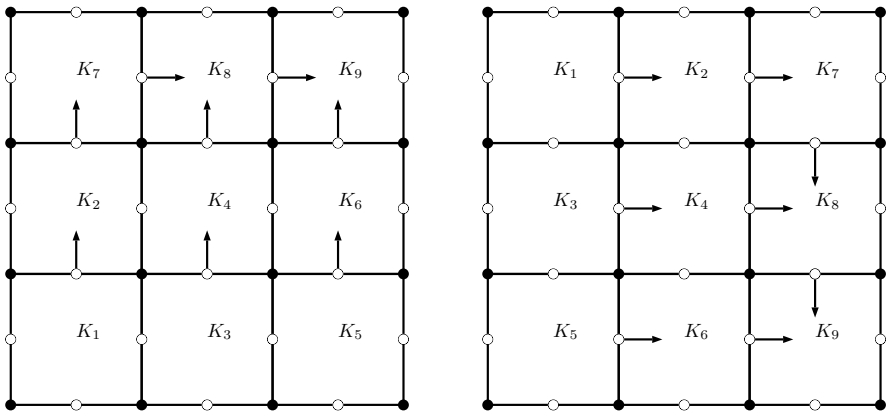


Figure 4.4: Original cells $K \in \mathcal{M}$, shaded points are vertices, hollow points are midpoints (left: column-wise adjustment; right: row-wise adjustment).

A more generic approach is taken in [27], which attains the local volume constraint by solving an optimisation problem. Let $\widetilde{\mathcal{M}}$ be the mesh formed by the polygonal approximations $\widetilde{K_i}$ of the trace-back regions $F_{-\delta t^{(n+\frac{1}{2})}}(K_i)$, for all $K_i \in \mathcal{M}$. We then denote by $\{\widetilde{\mathbf{p}}_i\}_{i=1,\dots,p}$ the set of points in the mesh $\widetilde{K_i}$. Of course, the set $\{\widetilde{\mathbf{p}}_i\}_{i=1,\dots,p}$ consists of points (vertices or edge midpoints) that have been tracked from the cells $K \in \mathcal{M}$. The optimisation problem seeks to find a mesh $\mathcal{M}^*$ consisting of cells K^* which forms a solution to the problem: Find $\mathcal{M}^*$ which minimises $|\mathcal{M}^* - \widetilde{\mathcal{M}}|$, subject to the constraints

- $|K_i^*| = |F_{-\delta t^{(n+\frac{1}{2})}}(K_i)|$ for all $K_i^* \in \mathcal{M}^*$;
- each cell K_i^* of $\mathcal{M}^*$ is valid;
- boundary points in $\mathcal{M}^*$ correspond to boundary points in $\widetilde{\mathcal{M}}$.

Here, $|\mathcal{M}^* - \widetilde{\mathcal{M}}| = \left(\sum_{i=1}^{p} |\mathbf{p}_i^* - \widetilde{\mathbf{p}}_i|^2 \right)^{\frac{1}{2}}$, where $\{\mathbf{p}_i^*\}_{i=1,\dots,p}$ denotes the set of points in the mesh $\mathcal{M}^*$. Hence, we may view $|\mathcal{M}^* - \widetilde{\mathcal{M}}|$ as a measure of how much adjustment is made on the points of the mesh $\widetilde{\mathcal{M}}$ in order to obtain the mesh $\mathcal{M}^*$. Essentially, the optimisation problem tells us that we are seeking to find a minimal adjustment of the points that were back tracked, so that the cells related to the adjusted points still form a mesh, and so that local volume constraint is achieved for each cell in the mesh.

Common to the algorithms in [5, 27] is an explicit expression for the adjusted trace-back regions. However, as can be seen in (4.23), this is not necessary for piecewise constant approximations, standard in reservoir simulations based on finite volume methods. The important aspect is the computation of the quantities $|M \cap \widetilde{K}|$.

We propose an algorithm which adjusts $|M \cap \widetilde{K}|$ for each cell K, in the sense that these adjusted volumes would be something we expect to recover from a mesh obtained by adjusting the tracked points. The proposed algorithm works on any type of cells, but for simplicity of exposure we illustrate it in Figure 4.5 using square cells, and trace-back regions $\widetilde{K}_i$ approximated by tracking the vertices and edge midpoints of K_i. In Figure 4.5, the blue lines denote the trajectory taken by the velocity field, and the red squares form part of the original mesh cells $K \in \mathcal{M}$, whereas the black cells are their trace-back regions: shaded points are the tracked vertices, and the hollow points are the tracked edge midpoints. In practice, after performing the tracking, aside from the final location of these vertices and midpoints, we also store the cell that they finally reside in. The algorithm is implemented cell-wise, starting from the first cell K_1, and proceeds as follows: Consider

a cell K_1 with neighbors K_2, K_3, etc. This leads to a trace-back region $\widetilde{K}_1$ with neighbors $\widetilde{K}_2$, $\widetilde{K}_3$, etc. Suppose that $\widetilde{K}_1$ intersects the residing cells M_1, M_2, M_3 and M_4 (see Figure 4.5, left).

i) We start by measuring the error $e_{K_1} := |\widetilde{K}_1| - |F_{-\delta t^{(n+\frac{1}{2})}}(K_1)|$. The relation $e_{K_1} > 0$ (resp. $e_{K_1} < 0$) means that we overestimate (resp. underestimate) the volume of the trace-back region.

ii) We then compute the magnitude $|\mathbf{u}|$ of $\mathbf{u}$ at the tracked midpoints and also check whether $\mathbf{u}$ points into $\widetilde{K}_1$ or not. If the velocity points towards the same direction for two consecutive midpoints, then we also compute the magnitude of $\mathbf{u}$ for the vertex in between them.

iii) We now illustrate how to adjust the volumes of the regions. If $e_{K_1} < 0$, for example, then it means that the volume $|\widetilde{K}_1|$ should be increased. Based on the velocity field $\mathbf{u}$ in Figure 4.5, this should be done by increasing along the direction of $\widetilde{K}_2$ and $\widetilde{K}_3$. Now, the velocity field along the edge midpoints that are located at $\widetilde{K}_1 \cap M_2$ and $\widetilde{K}_1 \cap M_3$ points outward of $\widetilde{K}_1$ and hence the vertex at $\widetilde{K}_1 \cap M_4$ should also be included. Denote then by $|\mathbf{u}_{1,i}|$ the magnitude of the velocity field evaluated at these tracked points in $\widetilde{K}_1 \cap M_i, i = 2, 3, 4$. We will then adjust each of the volumes by subtracting, to each $|\widetilde{K}_1 \cap M_i|$, the quantity $\frac{|\mathbf{u}_{1,i}|}{\sum_{j=2}^{4} |\mathbf{u}_{1,j}|} e_{K_1}$. This will then make K_1 satisfy the local volume constraint.

iv) To make sure that this quantity really represents something that would have come from a perturbed mesh (see Figure 4.5, right), the quantities $|\widetilde{K}_2 \cap M_2|, |\widetilde{K}_2 \cap M_4|, |\widetilde{K}_3 \cap M_3|$, and $|\widetilde{K}_3 \cap M_4|$ are adjusted accordingly (i.e. $\frac{|\mathbf{u}_{1,2}|}{\sum_{j=2}^{4} |\mathbf{u}_{1,j}|} e_{K_1}$ should be added onto $|\widetilde{K}_2 \cap M_2|$, and the corresponding quantities for the other edges).

We then proceed to adjust the other $\widetilde{K}_i$ in the same manner so that they satisfy their respective local volume constraints.

As a remark, we note that one set of adjustments may not suffice to satisfy the local volume constraints for all cells. For example, if, after the adjustment of $\widetilde{K}_1$, we have that $e_{K_2} > 0$, then we need to decrease the volume of $\widetilde{K}_2$. This can only be done (respecting the direction of the velocity field) by moving along the direction of $\widetilde{K}_1$. This will lead to $e_{K_1} > 0$. Hence, after all the adjustments in the first stage, we need to re-evaluate e_{K_i} and re-adjust the volumes. Of course, from a computational point of view, not all cells would be able to satisfy $e_{K_i} = 0$ exactly, so we terminate our algorithm

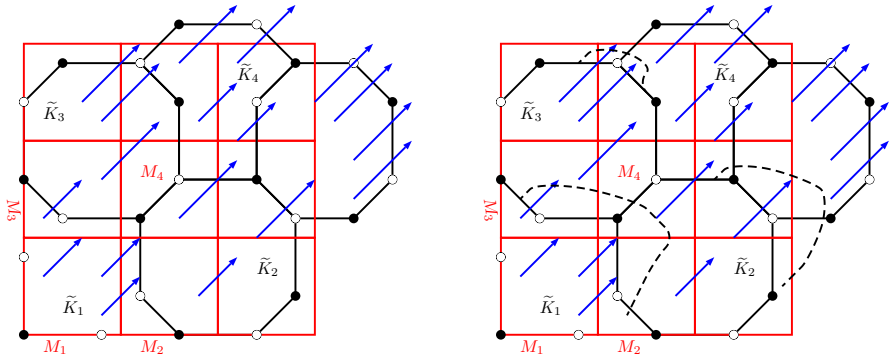


Figure 4.5: Trace-back regions $\widetilde{K}_i$ (left: initial; right: illustration of possible perturbed cells after proposed volume adjustment).

once $|e_{K_i}|$ is below a desired tolerance value for all cells. Another potential issue that may be encountered is when K_1 is a boundary cell (see Figure 4.6). If $\widetilde{K}_1$ lies on the boundary and $e_{K_1} > 0$, then, we need adjustments which will decrease the volume of $\widetilde{K}_1$. Thus, adjusting along the direction of the velocity $\mathbf{u}$ will only worsen the problem, since it will further increase the volume of $\widetilde{K}_1$ and hence the value of e_{K_1}. In this case, the idea is to consider $-\mathbf{u}$ instead. This translates to pushing inwards the points that would have been pushed outwards if $e_{K_1} < 0$. The re-adjustment of the other cells follow accordingly, and convergence is still expected. Making the volume adjustments in the direction of $\mathbf{u}$ (resp. $-\mathbf{u}$) corresponds to saying that we have tracked backward (resp. forward) too much, and hence to fix these, we must track forward (resp. backward) a little bit further.

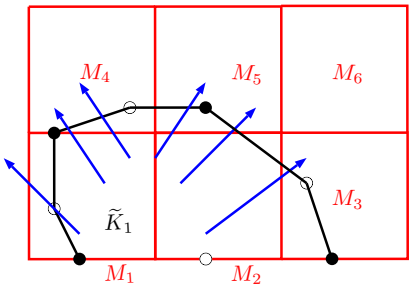


Figure 4.6: Trace-back region $\widetilde{K}_1$ of a cell at the boundary.

As has been mentioned in [23], when dealing with irregular cells, we need to track more than just edge midpoints so that $\widetilde{K}$ is close to $F_{-\delta t^{(n+\frac{1}{2})}}(K)$. Hence, for irregular cells, a slight modification for our algorithm should be made: we may opt to take smaller time steps (to make sure that the errors

e_{K_i} are small to start with, and adjustments can be made in a similar manner as with square cells, by only placing markers on the tracked midpoints and adjusting accordingly), or we may mark more than just the tracked midpoints (in order to have a better idea of the geometry of the trace-back region, and provide a more comprehensive list of the cell volumes to adjust). For our tests in Chapter 5, when such a modification was required, we chose to take smaller time steps.

4.7 Numerical tests

In this section, we perform numerical tests on a Cartesian mesh for the advection-reaction equation (4.2), with a given velocity field $\mathbf{u} = ((1-2y)(x-x^2), -(1-2x)(y-y^2))$, and zero source term (i.e. $f = 0$) on the domain $\Omega = (0,1) \times (0,1)$. Note that $\mathbf{u}$ is a divergence-free velocity field which simulates a rotation with some stretching, and the centre of this rotation is located at $(0.5, 0.5)$ (see Figure 4.7).

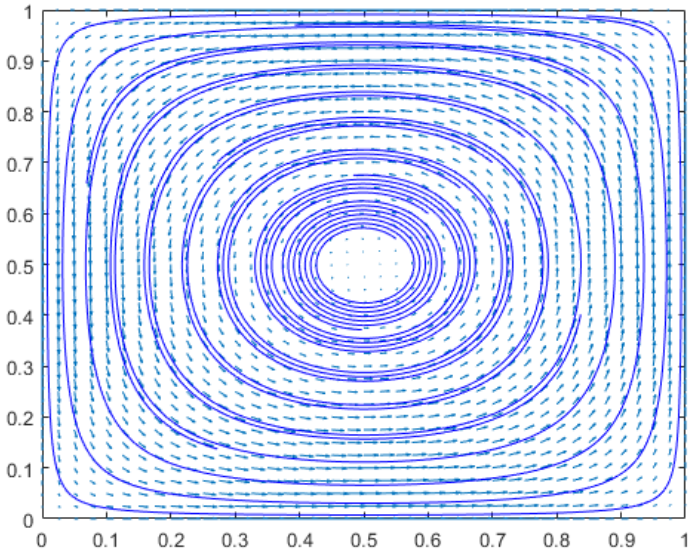


Figure 4.7: Streamlines of the velocity field $\mathbf{u} = ((1-2y)(x-x^2), -(1-2x)(y-y^2))$.

The initial condition is set to be

$$c(\boldsymbol{x}, 0) = \begin{cases} 1 & \text{if } (x-\frac{1}{4})^2 + (y-\frac{3}{4})^2 < \frac{1}{64} \\ 0 & \text{elsewhere} \end{cases}.$$

We seek the concentration at time $t = 8$, i.e. $c(\boldsymbol{x}, 8)$. Essentially, this assumes that we have a substance near the top-left corner of our domain (see Figure 4.8, left), having been rotated, and somehow stretched for $t = 8$ time units. A benchmark solution, obtained by solving (4.2) by the method of characteristics over a very fine grid, with a very small time step $\delta t = 0.001$, is presented in Figure 4.8, right. We start by presenting the concentration

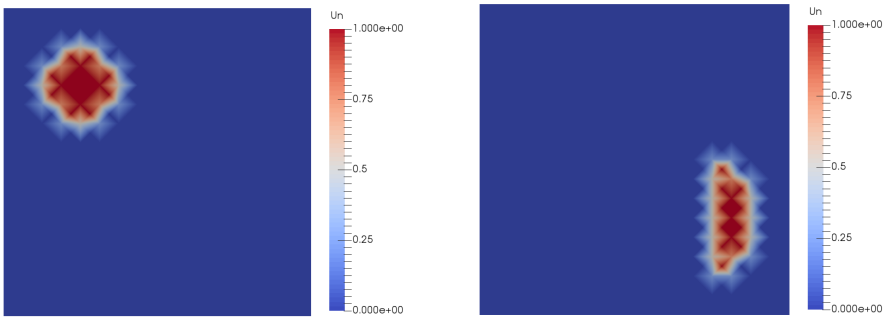


Figure 4.8: $c(\boldsymbol{x}, t)$ (left: initial condition at $t = 0$; right: benchmark solution profile at $t = 8$).

profiles obtained by solving (4.2) using an upwind scheme, with $\delta t = 2$ and $\delta t = 0.5$ (see Figure 4.9). Here, we note that the solution obtained from an upwind scheme exhibits excessive numerical diffusion.

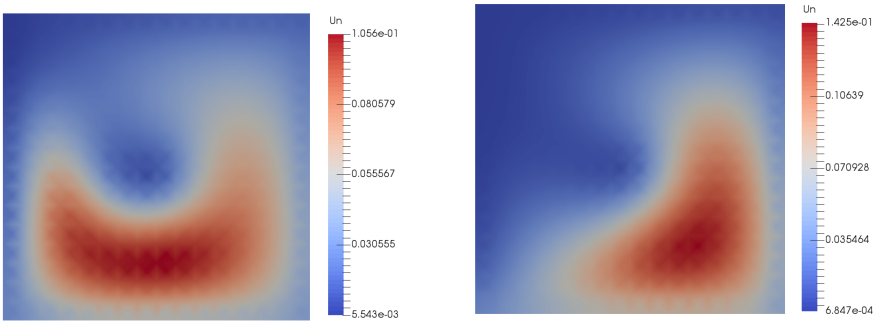


Figure 4.9: Concentration profile $c(\boldsymbol{x}, 8)$ obtained from upwind scheme (left: $\delta t = 2$; right: $\delta t = 0.5$).

We now compare the concentration profiles obtained from an upwind scheme to those obtained from ELLAM and MMOC schemes, with $\delta t = 2$ (see Figure 4.10). Here, a first order explicit Euler scheme is used to solve the characteristic equation (4.19) in order to approximate the trace-back and trace-forward regions, for the ELLAM and MMOC scheme, respectively.

Since $\mathbf{u}$ is a divergence free velocity field, the MMOC is also expected to exhibit global mass balance. Hence, for both ELLAM and MMOC, we use the algorithm described in Section 4.6.2, to perform adjustments in order to achieve local mass balance.

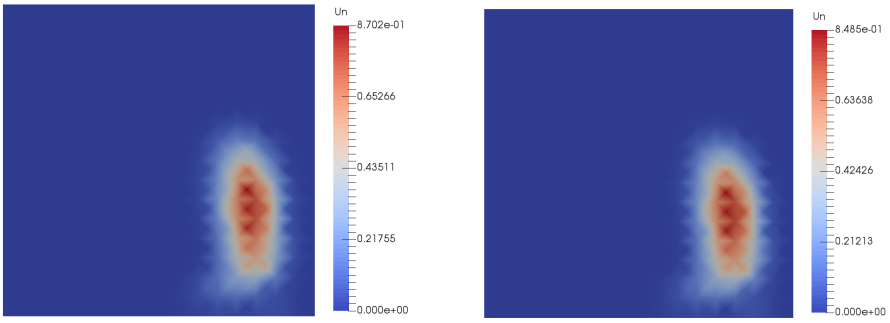


Figure 4.10: Concentration profile $c(\boldsymbol{x}, 8)$ obtained with $\delta t = 2$ (left: ELLAM; right: MMOC).

As can be seen, the concentration profiles obtained for both ELLAM and MMOC with $\delta t = 2$ captures the shape of the exact solution much better than those obtained from the upwind scheme. However, the maximum value of c is only at 0.87 for both schemes, which signals the presence of numerical diffusion. We also note that the concentration profiles for ELLAM and MMOC are quite similar, which is expected (see Remark 4.4.2). We now explore the concentration profiles obtained from ELLAM and MMOC upon taking a smaller time step of $\delta t = 0.5$ (see Figure 4.11).

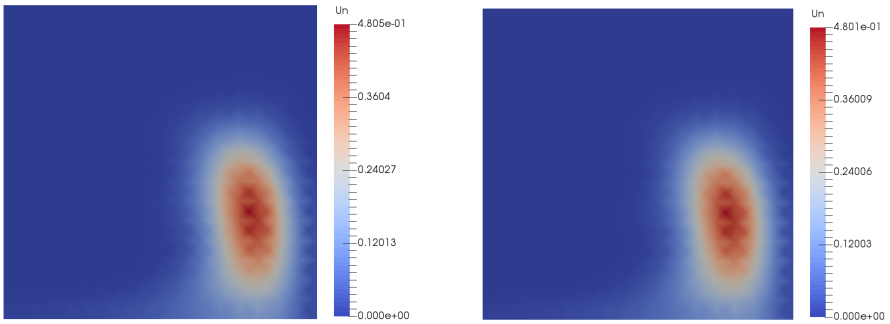


Figure 4.11: Concentration profile $c(\boldsymbol{x}, 8)$ obtained with $\delta t = 0.5$ (left: ELLAM; right: MMOC).

Upon looking at the concentration profiles, we note that although the ELLAM and MMOC still perform better than the upwind scheme, the numerical

diffusion is much worse than those obtained when $\delta t = 2$. In particular, the maximum value of c dropped to around 0.48 for both schemes. Actually, tests on ELLAM schemes with piecewise constant approximations have always been run with large time steps [6, 77]. As a matter of fact, it was indicated in [66, 73] that for a pure advection problem, when using piecewise constant approximations, a reverse CFL condition should be satisfied, i.e. $\delta x \leq |\mathbf{u}|\delta t$, where $|\mathbf{u}|$ measures the magnitude of the velocity field $\mathbf{u}$, δx and δt are the space and time steps, respectively.

Remark 4.7.1 (time-stepping)**.** *The numerical schemes were performed using Euler time steps. Future research may involve studying the implications of using other time-stepping methods, such as the Crank-Nicolson method.*

Table 4.1: CPU runtime (in seconds) for the pure advection problem (4.2) on a Cartesian mesh

Time step \ Scheme	Upwind	ELLAM	MMOC
$\delta t = 0.5$	0.6839	13.0515	14.3441
$\delta t = 2$	2.1480	25.4040	24.3347

In Table 4.1, it can be seen that in terms of computational cost, using an upwind scheme is more than ten times faster than an ELLAM or a MMOC scheme. However, as noted above, this introduces excessive numerical diffusion, and hence we pay a great price in terms of accuracy. Hence, for advection dominated problems, an ELLAM or MMOC scheme would be preferred, since they do not introduce a lot of numerical diffusion. Also, as noted above, for an efficient implementation of ELLAM and MMOC schemes, when we opt to use a coarse mesh, then we should use a relatively large time step. On the other hand, if we opt to use a small time step, then we should use a very fine mesh. This is demonstrated in Figures 4.12 and 4.13, where time steps of $\delta t = 1$ and $\delta t = 0.5$ are used, on meshes which consists of square cells with dimension 0.03125×0.03125 and 0.015625×0.015625 units, respectively. We note that the excessive numerical diffusion that was present upon taking a time step of $\delta t = 0.5$ on the coarse mesh with squares of size 0.0625×0.0625 units in Figure 4.11 is no longer present. In particular, this is due to the fact that the mesh is refined by a factor of 4, which is the same as the factor by which the time step is reduced.

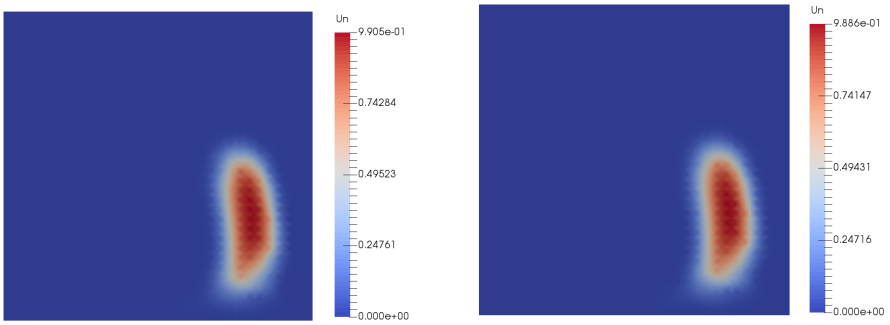


Figure 4.12: Concentration profile $c(\boldsymbol{x}, 8)$ obtained with $\delta t = 1$ on a refined Cartesian mesh with square cells of dimension 0.03125×0.03125 units (left: ELLAM; right: MMOC).

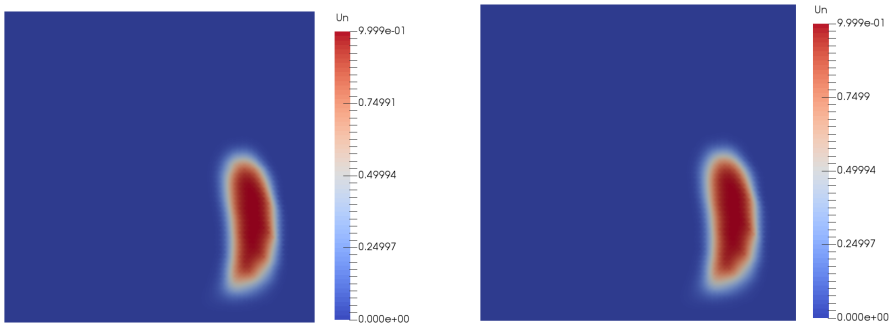


Figure 4.13: Concentration profile $c(\boldsymbol{x}, 8)$ obtained with $\delta t = 0.5$ on a refined Cartesian mesh with square cells of dimension 0.015625×0.015625 units (left: ELLAM; right: MMOC).

Chapter 5

Numerical schemes for the miscible flow model

In this chapter, we explore some numerical schemes for the complete coupled model (1.1). We start by presenting the framework of GDM–characteristic schemes, which covers several numerical schemes, some of which are the HMM–ELLAM, HMM–MMOC, and HMM–GEM. For the numerical tests, we then start by using HMM for the gradient discretisation, coupled by ELLAM for the advective component. This will be followed by an HMM–MMOC scheme and then an HMM-GEM scheme, which has a decent mass conservation property relative to HMM–ELLAM, and improved local volume conservation on hexahedral meshes. These numerical schemes are, however, prone to grid effects on coarse meshes. We start to study these grid effects by creating meshes with less distortion as compared to Kershaw type meshes. In particular, we create two types of meshes: one composed of very thin rectangular elements, and a second one composed of slightly perturbed Kershaw-like elements. Upon refining the mesh, the distortion on the solution is no longer as prominent as those of the coarse mesh. This, together with the fact that velocity fields reconstructed using fluxes which come from a low order scheme are inaccurate (see Section 3.3.1), leads us to consider the use of a high order scheme in space. Hence, we use HHO for the gradient discretisation, coupled with GEM for the advection to try to mitigate the grid effects.

5.1 GDM–characteristic schemes

In this section, we present a generic framework for combining gradient schemes and characteristic-based schemes for diffusive and advective components of

the miscible flow model, respectively. The idea is to implement a time-marching algorithm, wherein gradient discretisations (as described in Chapter 2) are used to approximate the diffusive terms for both (1.1a) and (1.1b). Some examples for which different GDs are applied for each equation in (1.1) are presented in Section 6.3.1.1. Note that the GDs used for the pressure equation do not need to involve the time components (time steps and interpolant of the initial condition). Also, as highlighted in Section 4.5, combining the ELLAM and MMOC for the treatment of the advective terms removes the main disadvantages of each of the schemes. Hence, we propose to use the combined ELLAM–MMOC scheme for the advective component. We will refer to the combination of the GDM with the ELLAM–MMOC scheme for the complete coupled model (1.1) as the GDM–ELLAM–MMOC (GEM) scheme.

The following definition of the GEM scheme is inspired by the construction of the GDM–ELLAM scheme in [23, 24] and by the design of the ELLAM–MMOC scheme for the advection–reaction model (Definition 4.5.2).

Definition 5.1.1 (GEM scheme)**.** *Let $\mathcal{P} = (X_\mathcal{P}, \Pi_\mathcal{P}, \nabla_\mathcal{P})$ be a space GD for the pressure, and $\mathcal{C}^T = (X_\mathcal{C}, \Pi_\mathcal{C}, \nabla_\mathcal{C}, \mathcal{I}_\mathcal{D}, (t^{(n)})_{n=0,\dots,N})$ be a space–time GD for the concentration. Let $\alpha : \Omega \to [0,1]$ be measurable. The GEM scheme for* (1.1) *reads as: find $(p^{(n)})_{n=1,\dots,N} \in X_\mathcal{P}^N$ and $(c^{(n)})_{n=0,\dots,N} \in X_\mathcal{C}^{N+1}$ such that $c^{(0)} = \mathcal{I}_\mathcal{C} c_{\rm ini}$ and, for all $n = 0, \dots, N-1$,*

i) $p^{(n+1)}$ solves

$$\begin{aligned} &\int_\Omega \Pi_\mathcal{P} p^{(n+1)} = 0 \text{ and} \\ &\int_\Omega \frac{\mathbf{K}(\boldsymbol{x})}{\mu(\Pi_\mathcal{C} c^{(n)})} \nabla_\mathcal{P} p^{(n+1)} \cdot \nabla_\mathcal{P} z = \int_\Omega (q_n^+ - q_n^-) \Pi_\mathcal{P} z \,, \qquad \forall z \in X_\mathcal{P} \end{aligned} \tag{5.1}$$

where $q_n^\pm(\cdot) = \frac{1}{\delta t^{(n+\frac{1}{2})}} \int_{t^{(n)}}^{t^{(n+1)}} q^\pm(\cdot, s) ds$ (or, alternatively, $q_n^\pm = q^\pm(t^{(n)})$ if $q^\pm$ are continuous in time).

ii) A Darcy velocity $\mathbf{u}_\mathcal{P}^{(n+1)}$ is reconstructed from $p^{(n+1)}$ (as in Chapter 3, for example) and, to account for the advection term in the concentration equation, the following advection equation is considered; it defines space-time test functions from chosen final values:

$$\phi \partial_t v + \mathbf{u}_\mathcal{P}^{(n+1)} \cdot \nabla v = 0 \quad \text{on } (t^{(n)}, t^{(n+1)}) \,, \text{ with } v(\cdot, t^{(n+1)}) \text{ given.} \tag{5.2}$$

iii) Using a weighted trapezoid rule with weight $\varpi_n \in [0,1]$ for the time-integration of the source term and setting $\mathbf{U}_\mathcal{P}^{(n+1)} = \dfrac{\mathbf{K}(\boldsymbol{x})}{\mu(\Pi_\mathcal{C} c^{(n)})} \nabla_\mathcal{P} p^{(n+1)}$,

$c^{(n+1)}$ *satisfies*

$$\begin{aligned}
&\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n+1)} \Pi_{\mathcal{C}} z d\boldsymbol{x} - \int_\Omega \phi(\boldsymbol{x}) \left[\alpha \Pi_{\mathcal{C}} c^{(n)}\right](\boldsymbol{x}) \Pi_{\mathcal{C}} z (F_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) d\boldsymbol{x} \\
&- \int_\Omega \phi(\boldsymbol{x}) \left[(1-\alpha)(\Pi_{\mathcal{C}} c^{(n)})\right](F_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})) \Pi_{\mathcal{C}} z(\boldsymbol{x}) d\boldsymbol{x} \\
&\qquad + \delta t^{(n+\frac{1}{2})} \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}}^{(n+1)}) \nabla_{\mathcal{C}} c^{(n+1)} \cdot \nabla_{\mathcal{C}} z \\
&\quad = \delta t^{(n+\frac{1}{2})} \int_\Omega \alpha \left[(q^+ - \Pi_{\mathcal{C}} c q^-)\right]^{(n,\varpi_n)} \cdot (\Pi_{\mathcal{C}} z)_F \\
&\qquad + \delta t^{(n+\frac{1}{2})} \int_\Omega \left[(1-\alpha)\big(q^+(1-\Pi_{\mathcal{C}} c)\big)^{(n,\varpi_n)} \cdot \mathbf{e}\right] \Pi_{\mathcal{C}} z, \qquad \forall z \in X_{\mathcal{C}},
\end{aligned} \tag{5.3}$$

where we recall the notations (4.21) *and* (4.30)*, and we set* $q_N^\pm = q_{N-1}^\pm$ *if these quantities are defined by averages on time intervals (as there is no time interval* $(t^{(N)}, t^{(N+1)})$*).*

Remark 5.1.2 (GDM–ELLAM and GDM–MMOC)**.** *Taking* $\alpha \equiv 1$ *everywhere corresponds to the GDM–ELLAM scheme whereas taking* $\alpha \equiv 0$ *everywhere corresponds to the GDM–MMOC scheme.*

Key to an efficient and accurate implementation of the GEM scheme is a proper choice of α so that mass conservation is achieved without having to deal with the steep source terms encountered in ELLAM. Remarks 4.3.4 and 4.5.3 give us an idea of how to define the function α. In the context of the complete coupled model (1.1), the non-divergence free regions are the injection and production cells. Moreover, it is expected that, for an injection cell C_+, $F_{[-\delta t^{(n+\frac{1}{2})},0]}(C_+) \subset C_+$ since the Darcy velocity flows outward the injection well. On the contrary, for a production cell C_-, we have that $C_- \subset F_{[-\delta t^{(n+\frac{1}{2})},0]}(C_-)$. This indicates that for an efficient application of the GEM scheme, the MMOC component should be implemented on regions near the injection cells once the concentration c is almost constant in these regions. This happens after some time T_+ when the injection cells C_+ are almost filled up, i.e. $c \approx 1$ in C_+. Before this, we should implement a pure ELLAM scheme. Hence, we start by defining $\alpha = 1$ over Ω for all n such that $t^{(n)} \leq T_+$, the time where the injection cells are filled up; typically, $T_+ \approx$ 1 to 1.5 years. Note that T_+ can actually be found during the simulation, by checking if the concentration is almost constant in and around the injection cells or not. To be specific, T_+ is determined to be the time $t^{(n)}$ such that $|\Pi_{\mathcal{C}} c^{(n)} - \Pi_{\mathcal{C}} c^{(n+1)}| < \epsilon$ in the cells surrounding the injection well(s) and the

well(s) themselves. For our numerical tests, we take $\epsilon = 10^{-4}$. For n such that $t^{(n)} > T_+$, and assuming for simplicity one injection cell C_+ and one production cell C_-, a possible choice is

$$\alpha(\boldsymbol{x}) = \begin{cases} 1 & \text{if } |\boldsymbol{x} - C_+| \geq |\boldsymbol{x} - C_-| \\ 0 & \text{otherwise.} \end{cases} \tag{5.4}$$

Here, $|\boldsymbol{x} - C_+|$ and $|\boldsymbol{x} - C_-|$ denote the distance between $\boldsymbol{x}$ and the center of the cells C_+ and C_-, respectively. This tells us to use ELLAM for regions far from the injection well, and MMOC otherwise. In case of multiple injection and production wells, the same rule can be applied by taking $\alpha(\boldsymbol{x}) = 1$ if the closest well to $\boldsymbol{x}$ is a production well, and $\alpha(\boldsymbol{x}) = 0$ if the closest well to $\boldsymbol{x}$ is an injection well.

5.1.1 Adaptation of local volume conserving adjustments to the miscible flow model

In this section, we write the algorithm for local volume conservation introduced in Section 4.6.2 in the context of the miscible flow model, for the GDM–ELLAM and GEM scheme. Since the GDM–MMOC does not achieve a global mass balance, it does not make sense to post-process the data and enforce local mass balance for each cell in the mesh. For the miscible flow model (1.1), the velocity field is divergence free in most regions, except for the injection and production wells (see (1.1a)). Hence, for most cells $K \in \mathcal{M}$, by the generalised Liouville's formula (4.14), we have $|F_{-\delta t^{(n+\frac{1}{2})}}(K)|_\phi = |K|_\phi$.

Remark 5.1.3. *The algorithm for local volume conservation proposed in [27] can easily be adapted for the miscible flow model, as long as the constraint on $|F_{-\delta t^{(n+\frac{1}{2})}}(K)|_\phi$ is known. However, for [6], additional steps are required, which involve separating the domain into regions, and performing adjustments along each region in order to achieve global, and then local mass balance. For complete details on how to adapt the algorithm in [6] to the miscible flow model, we refer the reader to [5]. We also note that both these algorithms are more expensive than the volume adjustment algorithm that we propose here.*

5.1.1.1 GDM–ELLAM

For the GDM–ELLAM, all cells are tracked backward. Hence, the cells K for which $|F_{-\delta t^{(n+\frac{1}{2})}}(K)|_\phi$ cannot be determined exactly are: the production cells, and the cells $K \neq C_+$ that track into the injection cells (fully or partially).

The approximation of the trace-back region of the production cells is obtained by tracking more points along their edges (compared to the other cells), so that the error here would be very small. We then formulate an approximate local volume constraint for the cells $K \neq C_+$ that track into the injection cells. We start by giving an approximation of the volume of the trace-back region of an injection cell C_+, which is exact if div$\mathbf{u}$ is constant. We note that our numerical scheme has div$\mathbf{u}$ constant in C_+, since the reconstructed velocity field is in $\mathbb{RT}_0$. Using the generalised Liouville's formula (4.14), the volume of the trace-back region of an injection cell C_+ is approximated to be $|F_{-\delta t^{(n+\frac{1}{2})}}(C_+)|_\phi \approx e^{-\beta}|C_+|_\phi$, where

$$\beta = \frac{\int_{C_+} q(t^{(n+1)})}{\int_{C_+} \phi} \delta t^{(n+\frac{1}{2})}.$$

We then compute an approximate trace-forward region $\widetilde{C}_+$ of C_+ by forming a polygon with vertices and edge points of C_+ tracked forward in time. In practice, compared to the other cells K in the mesh, we track more points along the edges of C_+ in order to obtain a good enough approximation of $\widetilde{C}_+$. In particular, if, on average, n points are tracked along the edges of a cell $K \in \mathcal{M}$, then we found that $4n+1$ is an appropriate number of points to track along the edges of C_+. We then set the approximate local volume constraint for the cells $K \neq C_+$ that track into C_+ to be

$$|\widetilde{K}|_\phi = |K|_\phi - |K \cap \widetilde{C}_+|_\phi + \frac{|K \cap \widetilde{C}_+|_\phi}{|\widetilde{C}_+ \setminus C_+|_\phi}(1 - e^{-\beta})|C_+|_\phi. \tag{5.5}$$

Essentially, (5.5) may be interpreted in the following manner (see Figure 5.1): the region $K \cap \widetilde{C}_+$ (shaded region inside cell K in Figure 5.1) is the part of K that is expected to track back into the injection cell C_+. Hence, since the other part of K (unshaded region inside cell K in Figure 5.1) is tracked back into a divergence free region, its volume $|K|_\phi - |K \cap \widetilde{C}_+|_\phi$ remains unchanged. The changed volume is then approximated as a ratio of the part in C_+ that will be tracked out of itself (i.e. $|C_+ \setminus F_{-\delta t^{(n+\frac{1}{2})}}(C_+)|_\phi = |C_+|_\phi - |F_{-\delta t^{(n+\frac{1}{2})}}(C_+)|_\phi = (1 - e^{-\beta})|C_+|_\phi$).

We can then adapt the algorithm proposed in Section 4.6.2 to the miscible flow model. We note that the local approximation (5.5) is valid only if we have a good approximation of the trace-forward region of the injection cells, which is why we track more points on the injection cells. As an additional fix, since we expect the injection cells to be eventually filled with the injected fluid, we set $c = 1$ on these cells once $c \approx 1$ in C_+. Assuming that

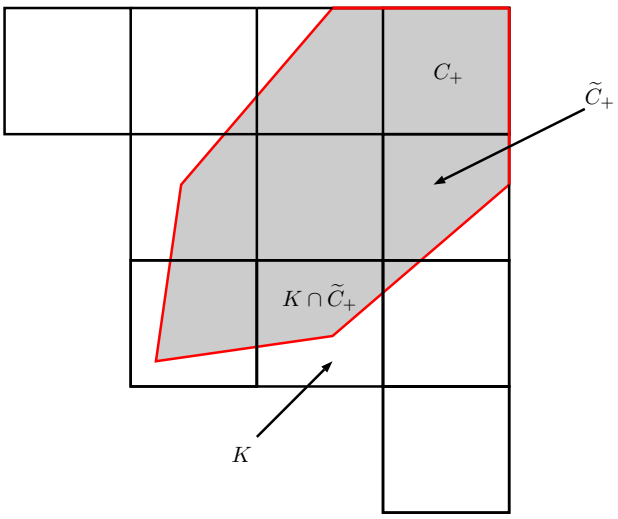


Figure 5.1: Trace-forward region of an injection cell C_+ and the affected residing cells.

(5.5) gives a good enough approximation, the local volume constraint for the production cells should then be satisfied approximately (since we have global mass conservation).

5.1.1.2 GEM

For the GEM scheme, the only regions with unknown volumes are the trace-forward of the injection cells $F_{\delta t^{(n+\frac{1}{2})}}(C_+)$ (for the MMOC component) and the trace-back of the production cells $F_{-\delta t^{(n+\frac{1}{2})}}(C_-)$ (for the ELLAM component). If the initial approximations made by tracking back the vertices and edges points for the two cells C_- and C_+ are good enough, then each of the approximations to the quantities $|F_{-\delta t^{(n+\frac{1}{2})}}(C_-)|_\phi$ and $|F_{\delta t^{(n+\frac{1}{2})}}(C_+)|_\phi$ should be very close to their actual values. Here, it is acceptable to track more points than what we track on average for the other cells $K \in \mathcal{M}$ to ensure that such an accuracy is obtained, since the global impact in terms of cost is minimal. In particular, one clear advantage of the GEM scheme comes from the absence of the approximations (5.5), which tells us that GEM is expected to give a better local mass conservation property compared to the GDM–ELLAM.

As a summary, for the GEM scheme, the following treatments are imposed:

- Since the choice of tracking expands C_- and C_+, good approximations of $|F_{-\delta t^{(n+\frac{1}{2})}}(C_-)|_\phi$ and $|F_{\delta t^{(n+\frac{1}{2})}}(C_+)|_\phi$ are obtained by using polygonal

regions formed by tracking more points (compared to the other cells). This is acceptable since in practice, there are only a few injection and production cells.

- For all other cells in the mesh, we impose the local volume constraint $|F_{\pm\delta t^{(n+\frac{1}{2})}}(K)|_\phi = |K|_\phi$. This constraint is valid as long as our time step is such that none of the MMOC and ELLAM cells track into C_- and C_+, respectively, which is not too restrictive.

5.2 Test data

Unless stated otherwise, the numerical simulations are performed under the following standard data (see, e.g., [78]):

1. $\Omega = (0, 1000) \times (0, 1000)$ ft^2,
2. injection well at $(1000, 1000)$ and production well at $(0, 0)$, both with flow rate of 30ft^2/day,
3. constant porosity $\phi = 0.1$ and constant permeability tensor $\mathbf{K} = 80\mathbf{I}$ mD,
4. oil viscosity $\mu(0) = 1.0$ cp and mobility ratio $M = 41$,
5. $\phi d_m = 0.0$ft^2/day, $\phi d_l = 5.0$ft, and $\phi d_t = 0.5$ft

For the time discretisation, we take a uniform time step of $\delta t^{(n+\frac{1}{2})} = 36$ days for $n = 0, \ldots, N-1$. Since we take a uniform time step, for the tests, we will refer to the time step as Δt in lieu of $\delta t^{(n+\frac{1}{2})}$. These will be simulated on Cartesian type meshes, hexahedral meshes (see Fig.1.1), non-conforming meshes, and finally on Kershaw type meshes (see Fig. 1.2). There are several parameters which may be switched upon performing these tests, such as the type of velocity reconstruction used, number of points to track along the edge of each cell, etc. Hence, to avoid confusion, starting with Figure 5.6, a short summary of which parameters are used for each figure will be presented in Appendix A.

5.3 HMM–ELLAM

Implementing an HMM scheme for diffusion and an ELLAM scheme for advection is essentially taking Definition 5.1.1 with $\alpha \equiv 1$ and the HMM gradient discretisation as described in Section 2.2. In this case, the reconstructed

velocity field is piecewise $\mathbb{RT}_0$ over a sub-triangulation of the mesh, as described in Section 3.1. We now provide full details of the choice of the weight to be used for the weighted trapezoid rule.

5.3.1 Source term and the weighted trapezoid rule

The integral involving the source term in the right hand side of Equation (5.3) should be treated carefully, otherwise the numerical results will feature severe undershoots or overshoots, especially over the regions around the injection well. Note that the left-hand and right-hand quadrature rules correspond to $\varpi_n = 1$ and $\varpi_n = 0$, respectively. To determine the proper weight ϖ_n, we consider an injection cell $K = C_+$. We mainly focus on injection cells since these are the cells which might cause mass conservation to fail. A proper weight that will yield mass conservation has been derived in [5], and is given by

$$\varpi_n = \frac{1}{1 - e^{-\beta}} - \frac{1}{\beta}, \quad \text{where} \quad \beta = \frac{\int_{C_+} q(t^{(n+1)})}{\int_{C_+} \phi} \delta t^{(n+\frac{1}{2})}. \tag{5.6}$$

Hence, for each cell K (injection or not), we use the weighted trapezoid rule with weight ϖ_n as in (5.6), for some C_+ related to K – see below. We treat the computation of the integral over $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ in the right hand side of (5.3) in different manners, depending on whether the cell K is

i) an injection cell,

ii) a cell tracked back into an injection cell (but not an injection cell itself),

iii) or a cell that does not track back into an injection cell.

i) If the cell K is an injection cell C_+, then it tracks back entirely into itself. Hence, over the entire interval $[t^{(n)}, t^{(n+1)}]$, $\nabla \cdot \mathbf{u}^{(n+1)} = \frac{1}{|C_+|} \int_{C_+} q(t^{(n+1)})$ and thus we obtain, through the generalised Liouville's formula (4.14):

$$\int_{F_{-\delta t^{(n+\frac{1}{2})}}(K)} q_{c^{(n+1)}}(t^{(n)}) = e^{-\beta} \int_K q_{c^{(n+1)}}(t^{(n+1)}), \tag{5.7}$$

where β is given by (5.6).

ii) If the cell K is not an injection cell, but is tracked back (at least partially) into an injection cell C_+, then we use a forward tracking algorithm similar to that described in [8]. Denoting by $\widetilde{C}_+$ the approximation to the

trace-forward region of C_+, the integral over the trace-back region of K is then approximated by

$$\int_{F_{-\delta t^{(n+\frac{1}{2})}}(K)} q_{c^{(n+1)}} \approx \frac{|K \cap (\widetilde{C}_+ \setminus C_+)|_\phi}{|\widetilde{C}_+ \setminus C_+|_\phi}(1 - e^{-\beta}) \int_{C_+} q_{c^{(n+1)}}.$$

In physical terms, this means that the volume injected from the well C_+ is transported into each of the cells K proportionally to their occupancy of $\widetilde{C}_+ \setminus C_+$, the trace-forward region of C_+ that they intersect. Note that, on the contrary to [8], only a fraction $(1 - e^{-\beta})$ of $\int_{C_+} q_{c^{(n+1)}}$ is being spread in the cells K around C_+, since a fraction $e^{-\beta}$ of $\int_{C_+} q_{c^{(n+1)}}$ has already been allocated to C_+, as can be seen in (5.7).

iii) Finally, if a cell K does not track back into an injection cell, then the integral $\int_{F_{-\delta t^{(n+\frac{1}{2})}}(K)} q_{c^{(n+1)}}$ will be computed using the (approximate) trace-back regions as described in Section 4.3.3. Actually, in that situation, either:

- K is not a production cell, $F_{-\delta t^{(n+\frac{1}{2})}}(K)$ is disjoint from injection cells and production cells (due to $-\mathbf{u}^{(n+1)}$ pointing outside production cells). Therefore, the integrals involving the source term are equal to zero.

- or K is a production cell, in which case, by nature of the Darcy flow, it is expected that $K \subset F_{-\delta t^{(n+\frac{1}{2})}}(K)$, so both integrals for the source term are equal and the value of ϖ_n is irrelevant.

5.3.2 Effect of the quadrature rule

The following simulations are based on KR velocities (see Section 3.1.2). Figure 5.2 shows the numerical solution for the concentration at $t = 10$ years on a Cartesian mesh using the left and the right hand rule, respectively. These results support the observations made in [22], i.e., that the left and right hand quadrature rules provide severe underestimates and overshoots, respectively, at the injection well, and are thus not good choices.

Figure 5.3 (left) shows the numerical solution for the concentration using the proper weight for the trapezoidal rule, and computation of the integrals as described in Section 5.3.1. This presents a significant improvement from the results obtained through the right and left rule.

The overshoot seen in this figure is at worst around 7%, which is commensurate with (or even less than) overshoots already noticed in other other characteristic methods in the absence of specific tweaks or post-processing [46].

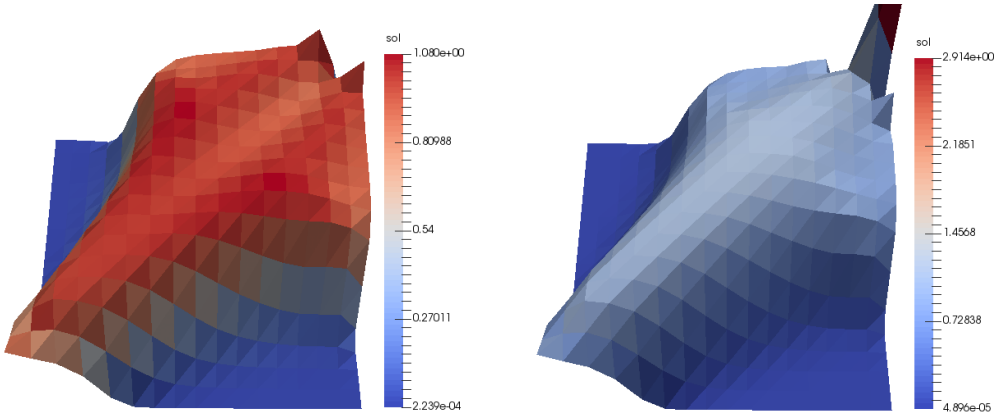


Figure 5.2: Cartesian mesh, $t = 10$ years, KR velocities (left: left rule for source terms; right: right rule for source terms).

5.3.3 Effect of achieving local volume conservation

As has been discussed in Section 4.6.2, local volume conservation is an important property that should be satisfied by numerical schemes. Hence, in this section, we use the technique described in Sections 4.6.2 and 5.1.1 to post-process the volumes of the tracked regions in order to achieve local volume conservation.

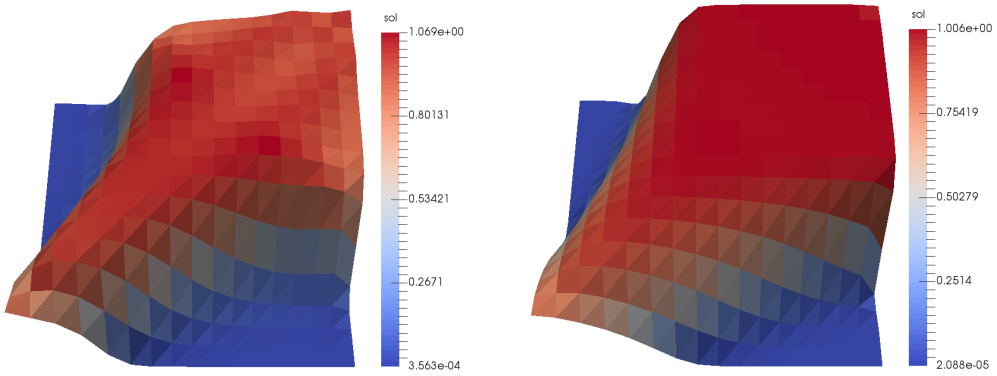


Figure 5.3: Cartesian mesh, weighted trapezoid rule for source terms, KR velocities, $t = 10$ years (left: without local volume conservation; right: with local volume conservation).

Upon imposing local volume conservation, a significant improvement in the concentration profile is observed. In particular, the overshoot of around 7% has been reduced to less than 1% (see Figure 5.3).

5.3.4 Comparison with forward tracking in [8]

We now compare our numerical results to the original algorithm of [8]. Instead of implementing i) and ii) as described in Section 5.3.1, the following process is applied:

- For an injection cell C_+, $c_{C_+}^{(n+1)}$ is fixed at 1 (for all time steps), as this is the concentration of the injected solvent.
- for cells K tracked back (at least partially) into an injection cell C_+, the following approximation is used:

$$\int_{F_{-\delta t^{(n+\frac{1}{2})}}(K)} q_{c^{(n+1)}} \approx \frac{|K \cap (\widetilde{C_+} \setminus C_+)|}{|\widetilde{C_+} \setminus C_+|} \int_{C_+} q_{c^{(n+1)}}, \tag{5.8}$$

  where $\widetilde{C_+}$ is the trace-forward region of C_+.

For some discretisation parameters, this implementation might lead to degraded results due to its physical implications. Setting $c_{C_+}^{(n+1)} = 1$ corresponds to distributing a fraction of $\int_{C_+} q_{c^{(n+1)}}$ into injection cells C_+. In this instance, a good estimate would be given by (5.7). However, computation of the integral for cells K that track back into an injection cell C_+ by (5.8) means that we spread the whole of $\int_{C_+} q_{c^{(n+1)}}$ onto the cells K. This then means that at time level $n+1$, an excessive amount of $e^{-\beta} \int_{C_+} q_{c^{(n+1)}}$ of fluid has been injected in the cells around C_+. If $\delta t^{(n+\frac{1}{2})}$ is large enough, then this is a negligible excess as $e^{-\beta} \approx 0$. However, for moderate to small $\delta t^{(n+\frac{1}{2})}$, the numerical results do not model the physical phenomenon properly. It is important to note that even though characteristic methods aim for computations using large time steps, we should still have an acceptable numerical result even when the time steps are small. We start by presenting in Figure 5.4 a numerical test with a relatively large time step of $\Delta t = 90$ days, which is of the same scale as the time step taken in [8]. As expected, due to the fact that $e^{-\beta} \approx 0$, the concentration profiles obtained from both algorithms are very similar.

Figure 5.5 (right) then shows the numerical solutions obtained at $t = 10$ years upon computing the integrals as in [8], with the moderate time step of $\Delta t = 36$ days. Due to injection of too much fluid, the overshoot at the right of Figure 5.5 (around 4.5%) is larger than the one on the left (less than 1%). This particular feature is even more evident if we take smaller time steps. Due to this, we see that the implementation we propose in Section 5.3.1 is more accurate.

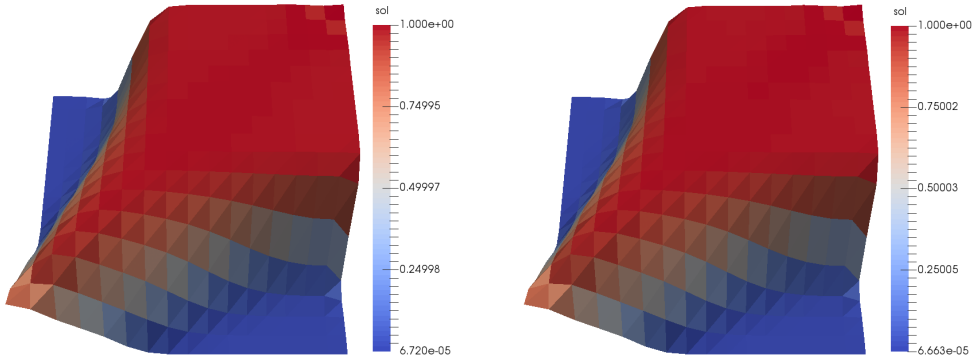


Figure 5.4: Cartesian mesh, weighted trapezoid rule for source terms, KR velocities, $\Delta t = 90$ days, $t = 10$ years (left: trace-forward as in Section 5.3.1; trace-forward as in [8]).

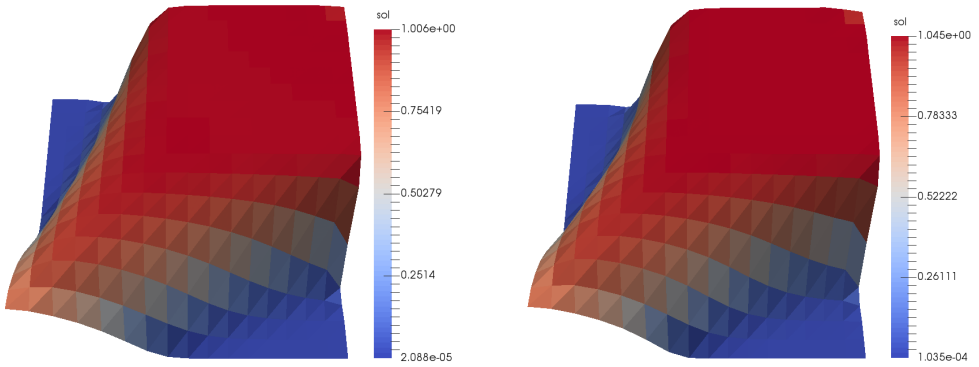


Figure 5.5: Cartesian mesh, weighted trapezoid rule for source terms, KR velocities, $\Delta t = 36$ days, $t = 10$ years (left: trace-forward as in Section 5.3.1; trace-forward as in [8]).

5.3.5 A criterion for choosing the number of points per edge

In general, a polygon formed by tracking only vertices and edge midpoints might not give a good approximation to the trace-back region $F_{-\delta t^{(n+\frac{1}{2})}}(K)$. Of course, with very bad polygonal approximations of the trace-back regions, we do not expect to be able to implement the local volume corrections efficiently. Hence, in this section, we quantify how many points must be tracked along the edge of each cell in order to obtain an acceptable polygonal approximation to $F_{-\delta t^{(n+\frac{1}{2})}}(K)$, which can be used for local volume correction.

The mesh regularity parameter, defined as

$$m_{\mathcal{M}\text{reg}} := \max_{K\in\mathcal{M}} \frac{\operatorname{diam}(K)^2}{|K|},$$

has been used as a criterion for determining the proper number of points to track along the edge of each cell (see Table 5.1). We found in our tests that at least $\lceil \log_2(m_{\mathcal{M}\text{reg}}) \rceil$ points per edge should be tracked in order to obtain a reasonable concentration profile (without local volume adjustments). These results have been established in [22] for KR velocities, and in [23] for C velocities. In this section, we verify that this still holds for A velocities.

Mesh	$m_{\mathcal{M}\text{reg}}$	$\log_2(m_{\mathcal{M}\text{reg}})$	points per edge
Cartesian	2	1	1
Hexahedral	5.4772	2.4534	3
Non-conforming	2.7619	1.4657	2
Kershaw	32.0274	5.0012	6

Table 5.1: Regularity parameter of the meshes and number of points to approximate the trace-back regions.

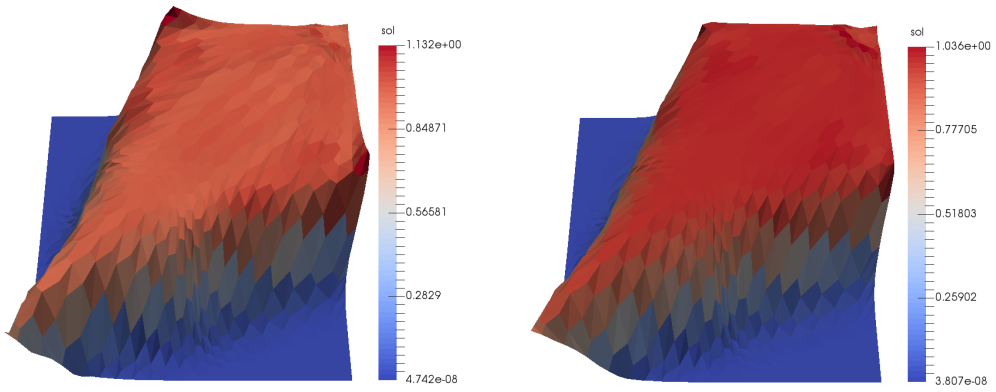


Figure 5.6: Hexahedral mesh, weighted trapezoid rule for source terms, A velocities, $t = 10$ years (left: edge midpoint, right: 3 points per edge).

We start by demonstrating on hexahedral meshes that even for A velocities, tracking only vertices and edge midpoints does not give a good approximation (see Fig. 5.6 left). As expected, taking 3 points per edge, as suggested in Table 5.1, then gives a better result, with an overshoot less than 4% (see Fig. 5.6 right).

This heuristic choice of number of points along each edge is further backed up by the numerical solutions for the non-conforming meshes, and also for the very distorted 'Kershaw' meshes (see Figure 5.7).

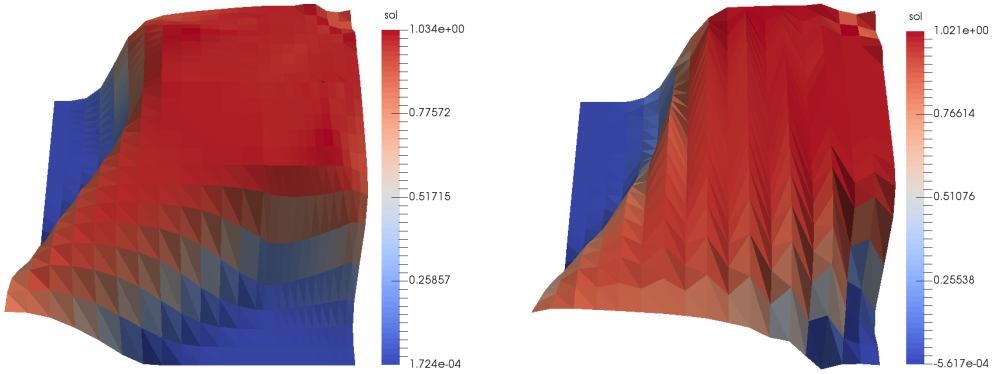


Figure 5.7: $\lceil \log_2(m_{\mathcal{M}\text{reg}}) \rceil$ points per edge, weighted trapezoid rule for source terms, A velocities, $t = 10$ years (left: non-conforming mesh, right: Kershaw mesh).

5.3.5.1 A more efficient implementation

We note however that not all cells are highly "irregular"; thus, tracking a lot of points along each edge for the whole mesh introduces unnecessary numerical cost. If the cell K is an injection or production cell, then we track more than $\lceil \log_2(m_{\mathcal{M}\text{reg}}) \rceil$ points along each edge (see Section 5.1.1). As an improvement, if K is neither an injection nor production cell, we determine the number of points to track along each edge of cell K by measuring instead the cell regularity parameter defined to be

$$m_{K\text{reg}} := \frac{\text{diam}(K)^2}{|K|} \tag{5.9}$$

and track $\lceil \log_2(m_{K\text{reg}}) \rceil$ points along each edge of cell K. By doing so, we reduce the computational cost without degrading much the quality of the numerical solutions. This will be illustrated on the slightly distorted hexahedral meshes and on the very distorted Kershaw mesh (see Figure 5.8).

Remark 5.3.1 (Number of points to track on an edge shared by two cells)**.** *For two neighboring cells K, L sharing an edge $\sigma_{K,L}$, the algorithm prescribed above tells us to track $\lceil \log_2(m_{K\text{reg}}) \rceil$ and $\lceil \log_2(m_{L\text{reg}}) \rceil$ points along $\sigma_{K,L}$ when viewed as a part of cell K and cell L, respectively. This is not practical*

since, for example, 3 equispaced points are totally different from 5 equispaced points along an edge, leading to more points to track. Hence, in the case that $\lceil \log_2(m_{K reg}) \rceil \neq \lceil \log_2(m_{L reg}) \rceil$, *we track* $\max(\lceil \log_2(m_{K reg}) \rceil, \lceil \log_2(m_{L reg}) \rceil)$ *points on the edge* $\sigma_{K,L}$.

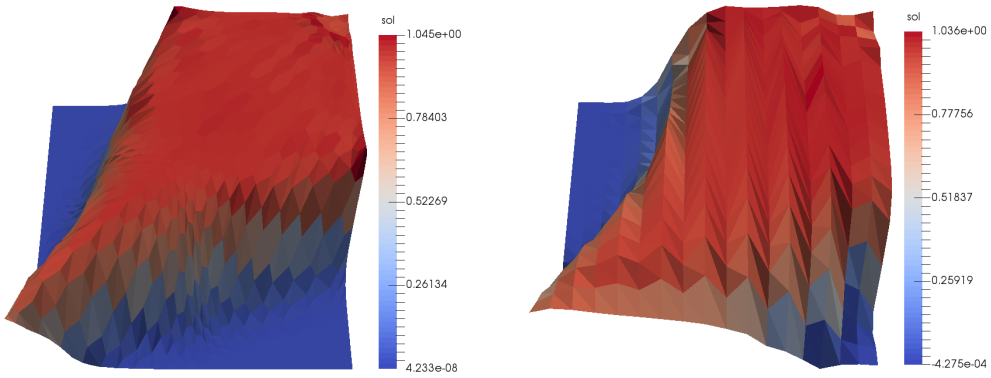


Figure 5.8: $\lceil \log_2(m_{K\text{reg}}) \rceil$ points per edge, weighted trapezoid rule for source terms, A velocities, $t = 10$ years (left: hexahedral mesh, right: Kershaw mesh).

As can be seen in Figures 5.6–5.8, the overshoots of the concentration profiles, obtained from A velocities, tracking $\lceil \log_2(m_{K\text{reg}}) \rceil$ points per edge are all less than 5%. We also note that by reducing the number of points tracked along each edge, the overshoot only increased by around $1 - 2\%$, which is a small price to pay considering the reduced computational cost that comes with it. We note here that the quantity $\lceil \log_2(m_{K\text{reg}}) \rceil$ only depends on the ratio between the area and the diameter of the cell, and not on the time step. Fixing $\lceil \log_2(m_{K\text{reg}}) \rceil$ to track along the edge of each cell would mean that taking a relatively large time step, say $\Delta t = 36$ for small cells would lead to a larger error than taking the same time step of $\Delta t = 36$ for larger cells. This is because small cells will be tracked through several regions, which will then result to a poor polygonal approximation to the trace-back region if we only track $\lceil \log_2(m_{K\text{reg}}) \rceil$ points along each edge; whereas large cells will be tracked through only a few regions, and the trace-back region can be well approximated even when only tracking $\lceil \log_2(m_{K\text{reg}}) \rceil$ points along each edge. The volume correction algorithm requires that the initial errors in approximating the trace-back regions should be small, and using $\lceil \log_2(m_{K\text{reg}}) \rceil$ as a basis, we introduce an improved formula for determining the number of points to be tracked along each edge of a cell K, given by

$$n_K := 2^{\lceil 2\Delta t/\Delta x \rceil} \lceil \log_2(m_{K\text{reg}}) \rceil + 1.$$

Here, we notice the dependence of n_K on $\Delta t/\Delta x$, which measures how large the time step is, relative to the size of the cell. Tracking n_K points along the edge of each cell K will then give a good enough polygonal approximation to the trace-back regions $F_{-\delta t^{(n+\frac{1}{2})}}(K)$, which may then be post-processed in order to achieve local volume conservation. For some meshes, such as hexahedral meshes, when n_K is large, we take a smaller time step in order to reduce n_K (see Figure 5.28 and Table 5.4). Before performing our post-processing, we first compare these results obtained from A velocities, to those obtained by using the KR and C velocities.

5.3.6 Comparison with the other reconstructions of the Darcy velocity

In this section, we compare the concentration profile at $t = 10$ years obtained from KR and C velocities to those obtained when we use A velocities. These will be performed over hexahedral, non-conforming and Kershaw type meshes, by tracking $\lceil \log_2(m_{K\text{reg}}) \rceil$ points along each edge of every cell.

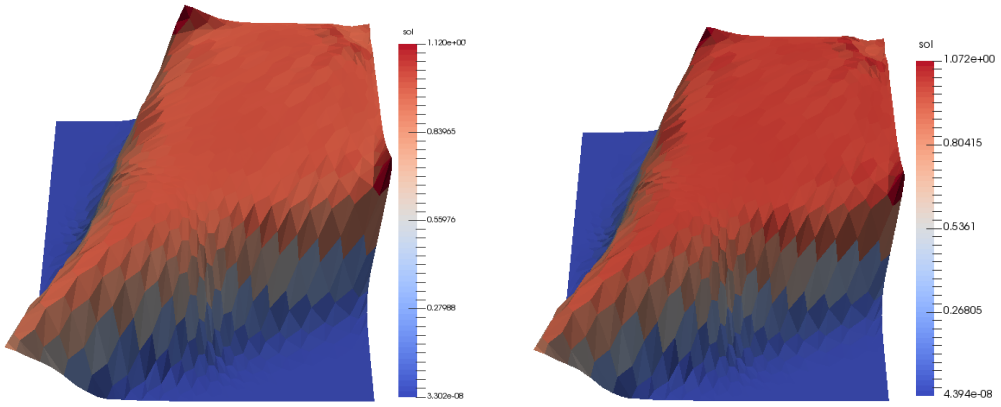


Figure 5.9: Hexahedral mesh, $\lceil \log_2(m_{K\text{reg}}) \rceil$ points per edge, weighted trapezoid rule for source terms, $t = 10$ years (left: KR velocities, right: C velocities).

Upon looking at Figures 5.9–5.11, we see that for all cases, the C velocities perform better than the KR velocities. Moreover, the overshoots in the concentration profiles obtained from KR velocities tend to increase and become much larger as the mesh becomes more distorted (i.e. starting from a small overshoot of around 4.5% on a non-conforming mesh to an overshoot of around 14.5% in a Kershaw mesh). Although the overshoots from C and A velocities also increase as the mesh becomes more distorted, it is not too

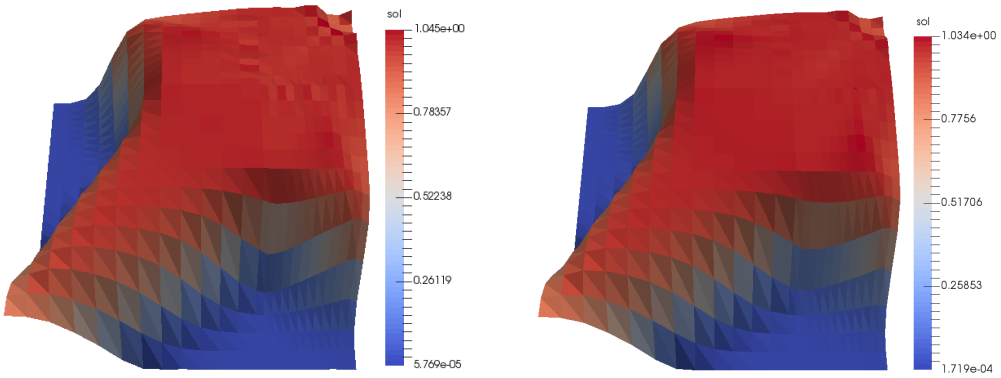


Figure 5.10: Non-conforming mesh, $\lceil \log_2(m_{K\text{reg}}) \rceil$ points per edge, weighted trapezoid rule for source terms, $t = 10$ years (left: KR velocities, right: C velocities).

badly behaved, with an overshoot of at most 7%. It is also notable that for distorted meshes (hexahedral and Kershaw), the overshoots of the concentration profiles obtained from A velocities is slightly less than those from C velocities (around $1 - 2\%$). Actually, these results are expected, and exhibit a similar trend to the results obtained in Section 3.3.1: For regular meshes, the differences between KR, C, and A velocities are minimal; whereas for distorted meshes, C and A velocities perform much better than KR velocities, with A velocities performing slightly better than C velocities. Based on these tests, the most efficient and accurate implementation of an HMM–ELLAM would involve: using A velocities, tracking $\lceil \log_2(m_{K\text{reg}}) \rceil$ points (or n_K points, for meshes with cells of highly varying sizes) along the edge of each cell, and taking a weighted trapezoid rule for source terms. After doing so, we perform a post-processing technique to adjust the volumes so that local volume conservation is achieved. It can also be found that the most efficient and accurate implementation of HMM–MMOC and HMM–GEM also involves these parameters. Henceforth and in the rest of the thesis, an HMM–ELLAM, HMM–MMOC, and HMM–GEM scheme would refer to this most efficient and accurate implementation, unless specified otherwise.

5.3.7 Numerical results from an HMM–ELLAM scheme

In this section, we now perform numerical tests using the "best" implementation of HMM–ELLAM on Cartesian, hexahedral, nonconforming, and Kershaw type meshes.

Upon looking at Figure 5.13 (left) and Figure 5.14, it is noticeable that

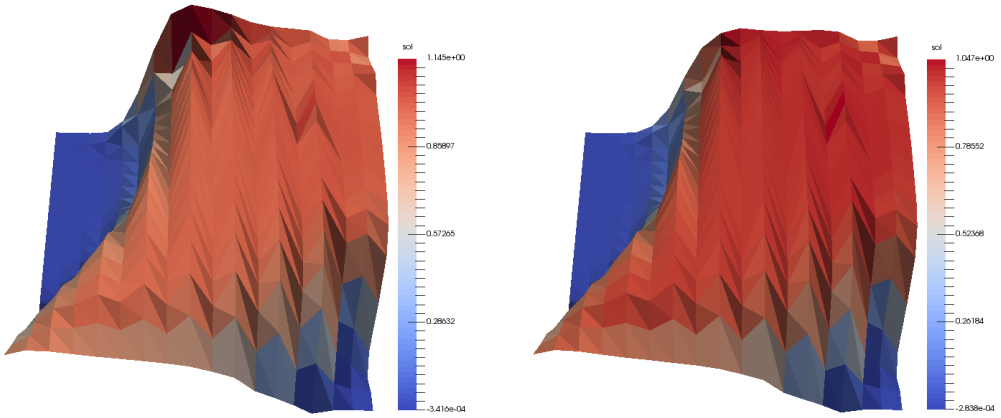


Figure 5.11: Kershaw mesh, $\lceil \log_2(m_{K\text{reg}}) \rceil$ points per edge, weighted trapezoid rule for source terms, $t = 10$ years (left: KR velocities, right: C velocities).

the solution from HMM–ELLAM on hexahedral meshes has a large discrepancy and overshoot near the injection well. This is due to the fact that the algorithm in Section 5.1.1 fails to converge. We note here that such a behavior was not observed in the literature, such as [8], since the tests were run only on Cartesian meshes.

There are two possible explanations for why the algorithm in Section 5.1.1 fails to converge for hexahedral meshes: Firstly, compared to the Cartesian and Kershaw type meshes, the volume of the injection cell, and each of the cells around it (around 700 to 1000 square units), is much smaller than the volume of the other cells in the mesh (on average, 2000 square units). Since these cells are already small to begin with, tracking them backwards will lead to trace-back regions which are much smaller, and hence will be more prone to errors. Taking $\Delta t = 36$ days, the expected volume of the trace-back region of the injection cell is $e^{-\beta}|C_+| \approx 10^{-4}$ square units, which is only around $10^{-8}\%$ of the total volume. This gives us an idea that taking a smaller time step might be able to mitigate these errors; however, even taking a much smaller time step of $\Delta t = 1$ day, there is still no convergence. We also note here that the nonconforming mesh has similar geometric properties: injection cell and cells around it are much smaller than the other cells. However, such a problem was not encountered with the nonconforming meshes. Due to this, we conclude that it is caused by the second possibility, i.e. for hexahedral meshes, (5.5) does not give a good approximation of the local volume constraint for the cells tracked back into the injection cell. This can be seen more clearly upon comparing Figure 5.14 to Figure 5.13, left. Without enforcing

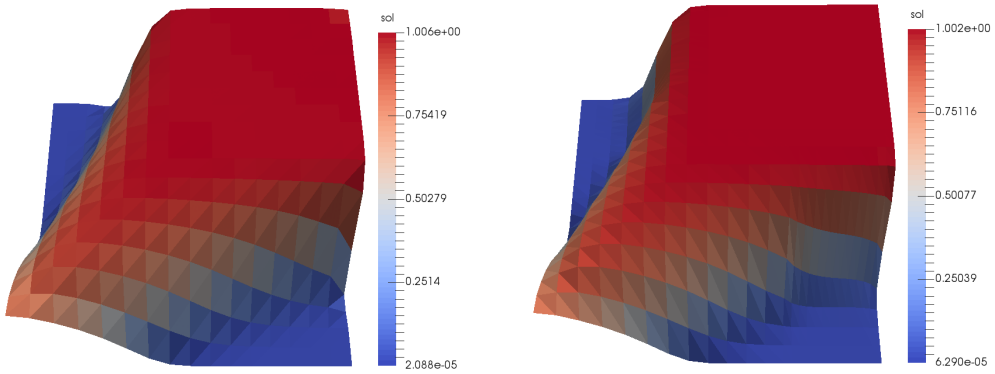


Figure 5.12: Concentration profile at $t = 10$ years, HMM–ELLAM (left: Cartesian mesh, right: nonconforming mesh).

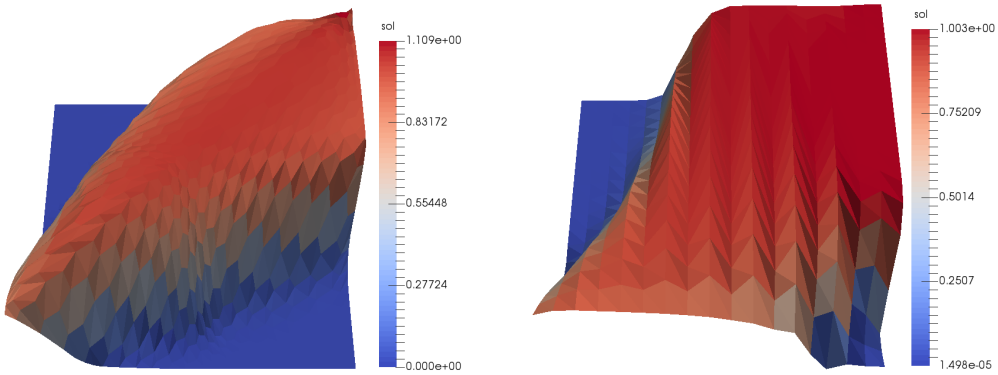


Figure 5.13: Concentration profile at $t = 10$ years, HMM–ELLAM (left: hexahedral mesh, right: Kershaw mesh).

(5.5), the solution seems to behave better. Actually, upon implementation of (5.5), the local volume constraint is satisfied on the cells which are not tracked back either into injection or into production cells. Having eliminated the local mass balance errors from these cells, they accumulate onto the cells near the injection cell. Since (5.5) does not give a good approximation, the accumulated error around this region does not spread and cancel out properly, and hence severely distorts the quality of the numerical solution. We note here that such a failure of the adjustment algorithm was not noticed on Cartesian meshes, whether in the tests conducted above or (with a different approach to the adjustment) in [8]. Our tests on hexahedral meshes demonstrate here the difficulty of designing a robust algorithm to locally adjust the mass balance for ELLAM. As a matter of fact, it has been pointed out in [27] that there is no guarantee that adjustments for ELLAM type schemes,

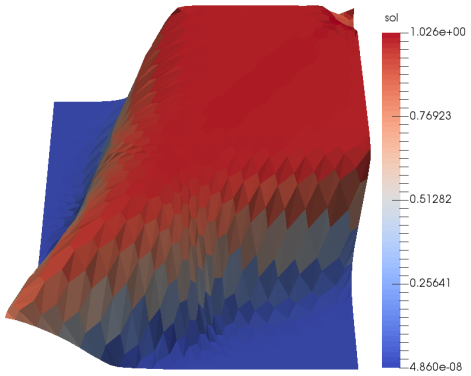


Figure 5.14: concentration profile at $t = 10$ years, hexahedral mesh, HMM–ELLAM, local volume conservation without (5.5).

such as those in [8], will terminate or yield a valid mesh configuration. The same issue could happen to our proposed adjustment applied on the GEM method but, as seen in the tests in Section 5.5.1, this adjustment seems to be more robust.

5.4 Comparison with HMM–upwind and MFEM–ELLAM

In this section, we compare the numerical results obtained from HMM–ELLAM to numerical results obtained from other schemes, such as HMM with upwinding [20] and MFEM–ELLAM [78]. This will be done on two test cases. The first test case will be done under the same test data and parameters considered above. The second test case will be done instead with an inhomogeneous permeability tensor $\mathbf{K} = 20\mathbf{I}$ mD over the region $(200, 400) \times (200, 400) \cup (200, 400) \times (600, 800) \cup (600, 800) \times (200, 400) \cup (600, 800) \times (600, 800)$ and $\mathbf{K} = 80\mathbf{I}$ mD elsewhere (see Figure 5.15), while holding all other test data and parameters to be the same as those of the initial test case. In particular, we note that the four regions around the middle of the domain have lower permeability, and hence it is more difficult for the fluid to flow through these regions.

This comparison is performed on a Cartesian mesh of size 50×50 ft (so that the discontinuities present in the second test case are aligned with the edges of the cells); other meshes could be considered (triangular for MFEM–ELLAM, and any polytopal mesh for HMM–upwind), with similar conclusions. For the second test case, five points are tracked along the edge

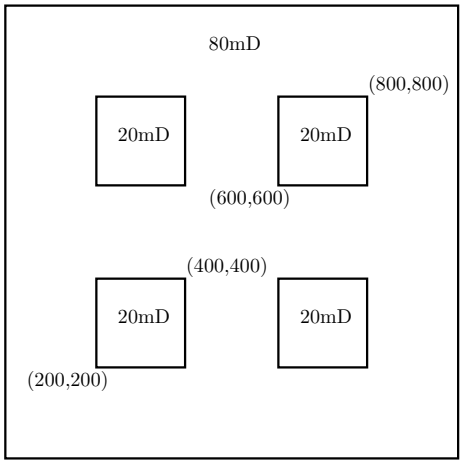


Figure 5.15: Permeability of the medium (test case 2).

of each cell, in order to get a better approximation of the trace-back regions, due to the discontinuities in the permeability tensor. As a point of reference, we present in Figures 5.16 and 5.17 the concentration profile and contour plot at $t = 10$ years for the HMM–ELLAM, for the first and second test cases, respectively.

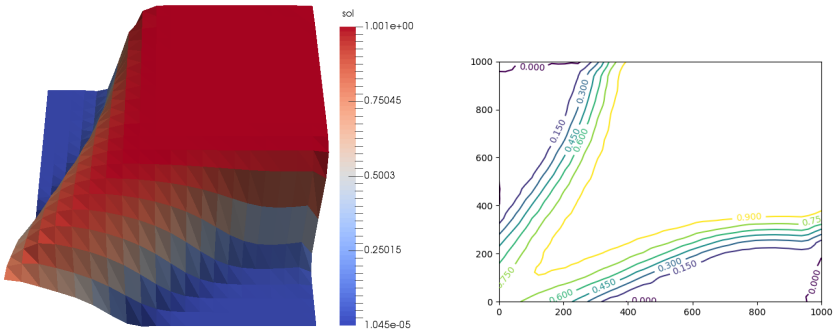


Figure 5.16: Numerical concentration obtained through HMM–ELLAM at $t = 10$ years, homogeneous permeability (left: profile; right: contour plot).

We note that Figure 5.17 exhibits some fingers near the boundary. To be specific, consider the square region $(400, 1000) \times (400, 1000)$. At the top boundary of the domain, the fluid penetrates the region to the left of the line $x = 400$, whereas at the right boundary of the domain, the fluid penetrates the region at the bottom of the line $y = 400$. A similar behavior has also been observed in the numerical tests in [42], and a fix, which modifies the

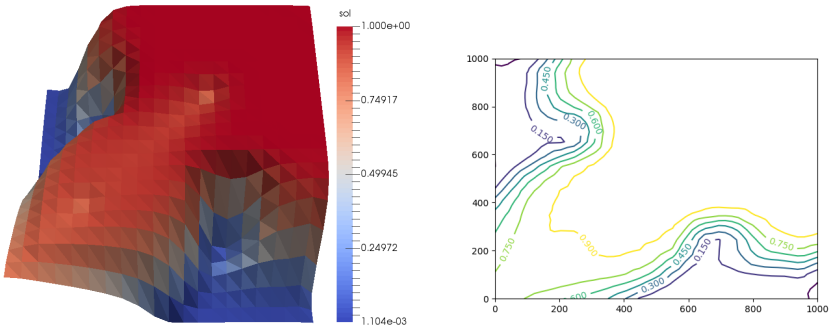


Figure 5.17: Numerical concentration obtained through HMM–ELLAM at $t = 10$ years, inhomogeneous permeability (left: profile; right: contour plot).

diffusion tensor $\mathbf{D}(\boldsymbol{x}, \mathbf{u})$ in (1.1c) by setting

$$(\mathbf{D}(\boldsymbol{x}, \mathbf{u}))_{i,i} = \max((\mathbf{D}(\boldsymbol{x}, \mathbf{u}))_{i,i}, \phi|\mathbf{u}|h),$$

where h is the size of the mesh, has been proposed. This amounts to introducing a vanishing diffusion, which scales with the magnitude of the Darcy velocity, and vanishes with the mesh size in the same way as upstream numerical diffusions. As can be seen in Figure 5.18, upon introducing the vanishing diffusion, the artificial fingers along the boundary have been reduced.

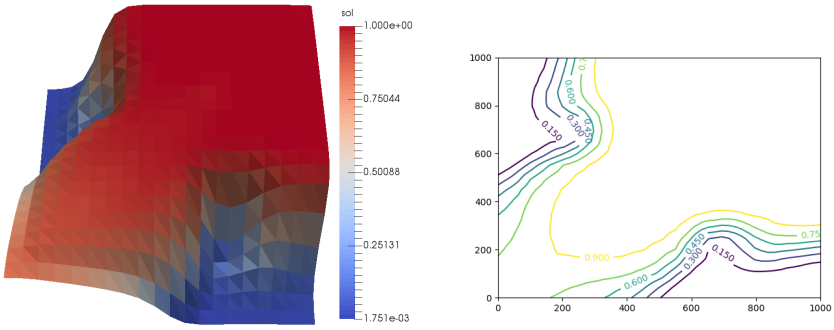


Figure 5.18: Numerical concentration obtained through HMM–ELLAM at $t = 10$ years, inhomogeneous permeability, modified diffusion tensor (left: profile; right: contour plot).

5.4.1 MFEM–ELLAM

As both MFEM–ELLAM and HMM–ELLAM are characteristic methods, tracking is implemented for both schemes for the concentration equation.

Typically, HMM–ELLAM schemes only need to track the vertices, together with 1 point per edge for Cartesian type meshes, unless there are discontinuities in the permeability tensor (as in test case 2), or when the time step is too large relative to the spatial discretisation (which will result to either a degenerate or self intersecting polygon approximating the trace-back region). This can be avoided by reducing the time step or increasing the number of points tracked along each edge. However, MFEM–ELLAM schemes need to track a bare minimum of 3-4 points in each cell to get a correct quadrature rule to integrate the basis functions (and much more than 4 points in case these bases functions become too distorted by the tracking velocity [75]). On the other hand, implementing an HMM–ELLAM scheme requires post-processing in order to achieve mass conservation, as discussed in Section 5.1.1.

Aside from computational cost, a more important thing to consider would be the quality of the numerical solutions. Figures 5.19 and 5.20 give us the numerical solution and contour plot obtained from MFEM–ELLAM at $t = 10$ years for the first and second test cases, respectively. These numerical outputs were obtained by a straight application of the MFEM–ELLAM algorithm as presented in [78], with several hundred of quadrature points per cell around the injection well.

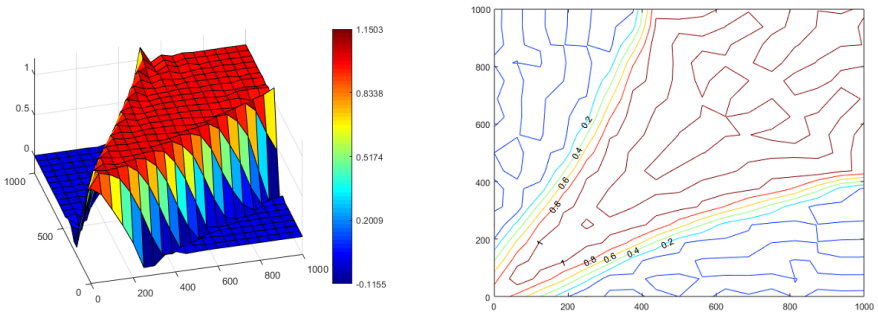


Figure 5.19: Numerical concentration obtained through MFEM–ELLAM at $t = 10$ years, homogeneous permeability (left: profile; right: contour plot).

The shape of the concentration profile and the contour plots obtained from both schemes look quite similar. However, we note that the MFEM–ELLAM has overshoots and undershoots (around 15% for homogeneous permeability and 20% for inhomogeneous permeability), that are typical of characteristic-based methods in the absence of post-processing [46]. On the other hand, the overshoots from the HMM–ELLAM is minimal (less than 1%) for both cases. Note that for the source terms, the MFEM–ELLAM integrates a non-constant function through quadrature rules. The main source of error encountered upon computing these integrals arise due to the presence

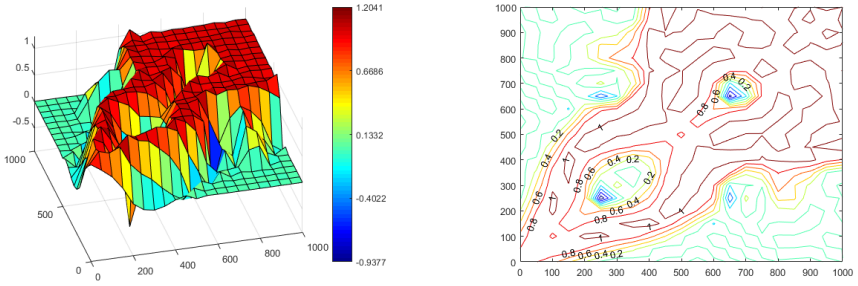


Figure 5.20: Numerical concentration obtained through MFEM–ELLAM at $t = 10$ years, inhomogeneous permeability (left: profile; right: contour plot).

of steep source terms. However, for the HMM–ELLAM, a different treatment of the source terms (see Section 5.3.1) was implemented. We recall that this can be physically interpreted as spreading the injected fluid around the region surrounding the injection well. Moreover, a post-processing on the local volumes which lead to local volume conservation was implemented. Naturally, these resulted to the HMM–ELLAM having a much smaller overshoot than the MFEM-ELLAM. We also note that the MFEM–ELLAM exhibits severe undershoots (up to around 11% for homogeneous permeability near the production cell and 90% for inhomogeneous permeability in the low-permeability regions), which is not present in the HMM–ELLAM. This severe undershoot might be due to the fact that conforming FE methods, used for solving the concentration equation, have unknowns on the vertices that sit at the permeability discontinuities. It has been noticed that, for transport of species in heterogeneous domains, schemes with unknowns at the vertices may lead to unacceptable results on coarse meshes, see [53]. On the other hand, the transition layer (from $c \approx 0$ to $c \approx 1$) is thinner for the MFEM–ELLAM than for the HMM–ELLAM.

5.4.2 HMM–upwind

Over each time step, the HMM–upwind requires, for the concentration, the solution of a linear system which has the same sparsity and number of unknowns as the HMM–ELLAM. Moreover, due to the absence of characteristic tracking and computation of integrals over trace-back regions, the computational cost of HMM–upwind scheme is much cheaper than that of the HMM–ELLAM. Next, we compare the quality of the solutions obtained by looking at Figures 5.21 and 5.22. It is quite notable that the solution remains bounded between 0 and 1 (actually, no undershoot occurs, and the overshoot is less

than 0.01%). However, upwind schemes tend to introduce excess diffusion, and thus the strong viscous fingering effects we expect have been spread out. This can be seen more clearly by looking at the contour plot (Figure 5.21 right). Upon comparison with contour plots obtained for the HMM–ELLAM and MFEM–ELLAM schemes (Figures 5.16 and 5.19, right), we indeed see that the strong viscous fingering expected along the diagonal has been spread out by the upwind scheme. A similar conclusion can be drawn for the inhomogeneous permeability tensors upon comparing Figure 5.22 to Figures 5.17 and 5.20. Upon comparing Figures 5.18, 5.20, and 5.22 we notice that the concentration profile obtained for the HMM–ELLAM have almost completely filled in the regions along the diagonal with low permeability, namely the regions $(200, 400) \times (200, 400)$ and $(600, 800) \times (600, 800)$. In comparison, the concentration profiles obtained from upwinding and MFEM–ELLAM has yet to fill these regions. Upon comparing these to a hybrid high order scheme [2], the concentration profile from the HMM–ELLAM appears to be the one which depicts the sweeped regions most accurately. To further strengthen this argument, another point of comparison, which measures the amount of oil recovered, will be presented.

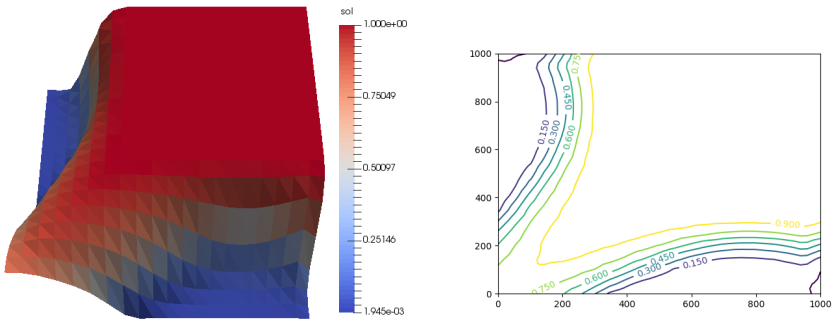


Figure 5.21: Numerical concentration obtained through HMM–upwinding at $t = 10$ years, homogeneous permeability (left: profile; right: contour plot).

5.4.3 Recovered oil

A particular quantity of interest in performing numerical simulations of the model (1.1) is the amount of oil (percentage of domain) recovered at time T, given by $|\Omega|_\phi^{-1} \int_\Omega \phi c(x, T)$. Figure 5.23 shows the percentage of domain recovered at $t = 10$ years, for each of the three numerical schemes on Cartesian meshes. These results were obtained starting with a very coarse mesh consisting of square cells of dimension 200×200, while refining the spatial

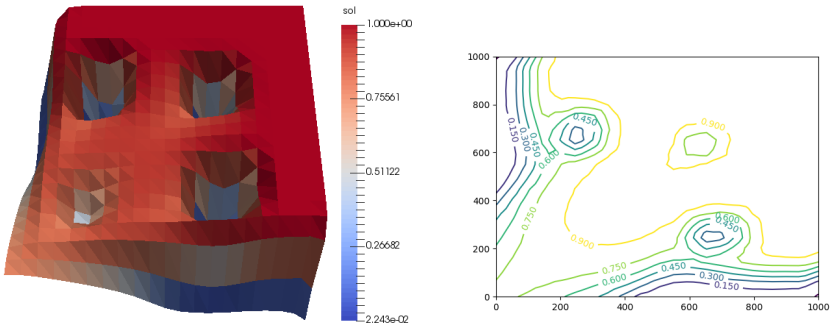


Figure 5.22: Numerical concentration obtained through HMM–upwinding at $t = 10$ years, inhomogeneous permeability (left: profile; right: contour plot).

discretisation by a factor of 2, leading to a final mesh consisting of square cells of dimension 25×25.

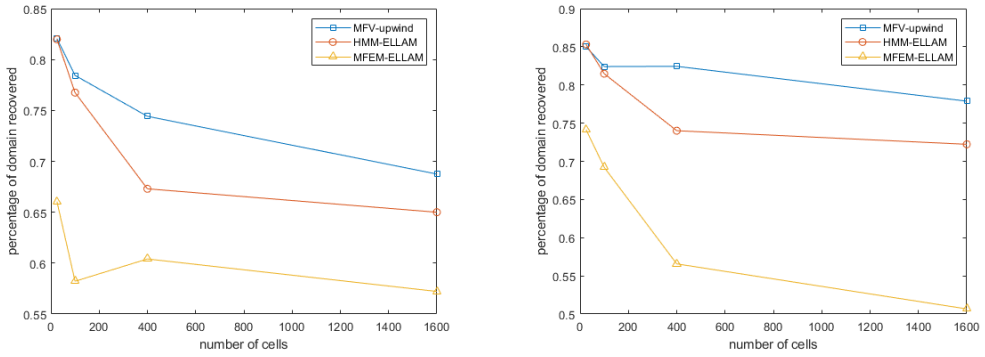


Figure 5.23: Percentage of domain recovered at $t = 10$ years (left: homogeneous permeability; right: inhomogeneous permeability).

It can be seen here that each of the three schemes seem to yield numerical results that converge to different quantities. The HMM–upwind scheme probably overestimates the amount of oil recovered due to the diffusion it introduces, and thus a "wider" region has been sweeped. On the other hand, it is likely that the MFEM–ELLAM underestimates the amount of oil recovered due to the presence of undershoots, as we have noted from Figures 5.19 and 5.20 (left). Thus, we expect the amount of oil recovered to be somewhere between these two values. Here, the solution obtained from HMM–ELLAM converges to such a value, which seems to sit about 65% for the first test case and 73% for the second test case.

As an element of comparison, we consider the results obtained in [2] on

the same model with the Hybrid High-Order (HHO) method. This method is based on polynomials, with an arbitrary chosen degree, in the cells and on the faces, and can theoretically achieve any order of accuracy on any polytopal mesh (at an increased computational cost compared to HMM and ELLAM methods, of course); in practice, though, tests are usually ran up to order 4 or 5; we refer to [30] for the presentation on the pure diffusion equation, and to [28] for advection–diffusion–reaction models. The tests in [2] were performed up to order 4 and seem to indicate that the expected recovery after 10 years is around 65% for the homogeneous test case, and around 75% for the inhomogeneous test case. The low-order, less expensive HMM–ELLAM method seems to provide very similar results, on the contrary to HMM–upwind and MFEM–ELLAM. The latter, in particular, only predicts a 50% recovery for the inhomogeneous test case, which is much lower than all other methods; this is naturally expected because of the presence of severe undershoots. The concentration profiles obtained from the HMM–ELLAM are also similar to the one obtained through the HHO in [2]. Moreover, due to the thinner transition layer present in the HMM–ELLAM, this solution is preferred over the HMM–upwind scheme.

Now, we give an indication of the computational cost that comes with the accuracy offered by the HMM–ELLAM in Table 5.2. Firstly, we note that the HMM–ELLAM is more expensive to implement than the MFEM–ELLAM, due to the post-processing that needs to be performed in order to achieve mass conservation. Also, HMM–upwind performs faster than HMM–ELLAM by more than a factor of ten, which is similar to the observation made for the pure advection test case in Table 4.1. However, it can be checked that even upon performing one or two levels of mesh refinement and taking a smaller time step (which would then lead to the same computational cost as the HMM–ELLAM on a coarse mesh), neither HMM–upwind nor MFEM–ELLAM can reach the level of accuracy achieved by the HMM–ELLAM. Hence, the accuracy gained by the HMM–ELLAM is worth the price paid in terms of computational cost.

Table 5.2: CPU runtime (in seconds) for the miscible flow model (1.1) on a Cartesian mesh

Test case \ Scheme	HMM–upwind	HMM–ELLAM	MFEM–ELLAM
1	27.0348	406.2903	77.3863
2	27.1139	452.2373	80.8103

5.4.4 Strengths and weaknesses of the HMM–ELLAM scheme

Upon comparison with HMM–upwind and MFEM–ELLAM schemes, we got to see that the HMM–ELLAM captures the concentration profile better than upwinding, although as compared to MFEM–ELLAM we have a larger transition layer, as presented in the contour plots in Figures 5.16 to 5.22 (right). As for the amount of oil recovered, it seems that the HMM–ELLAM also performs better than both MFEM–ELLAM and HMM–upwind schemes for both test cases. Hence, overall, an HMM–ELLAM scheme is preferred over MFEM–ELLAM or HMM–upwind schemes. Moreover, the results from HMM–ELLAM were obtained using a time step of $\Delta t = 36$ days, and are quite close to those obtained from the hybrid high order scheme implemented in [2] (which was second order in time, implemented with a time step of $\Delta t = 7.2$ days).

There are still some issues to be resolved for the HMM–ELLAM. Firstly, the concentration profile exhibits some spikes/wiggles near the injection well for hexahedral meshes. This is due to the bad approximation for steep back-tracked functions, as noted in Remark 4.3.4, and also due to bad approximations for the local volume constraints given by (5.5). Secondly, the HMM–ELLAM exhibits severe grid effects for the very distorted Kershaw-type meshes.

5.5 HMM–MMOC and HMM–GEM

In this section, we improve the approximations of the steep back-tracked functions near the injection well. This will be done by exploring the HMM–GEM scheme, starting with the Cartesian mesh, followed by the slightly distorted hexahedral mesh, and then on the locally refined nonconforming mesh, and then finally on the severely distorted Kershaw-type meshes. The HMM–GEM for the miscible flow model is implemented by taking Definition 5.1.1 with α as in (5.4) and the HMM gradient discretisation as described in Section 2.2. Numerical results obtained from the GEM scheme are compared to those obtained from HMM–ELLAM. Both of these schemes employ the post-processing technique outlined in Section 5.1.1 to achieve local mass balance. For completeness, we will also present a comparison with HMM–MMOC, but without the local adjustments. The HMM–MMOC for the miscible flow model is implemented by taking Definition 5.1.1 with $\alpha \equiv 0$ and the HMM gradient discretisation. As with the GEM scheme, for the HMM–MMOC, an ELLAM scheme is first implemented for the first few time steps,

when $t^{(n)} \leq T_+$, after which, a pure MMOC scheme is implemented, i.e. for $t^{(n)} > T_+$, we take $\alpha = 0$ over Ω in (5.3).

5.5.1 Numerical results

We start by presenting the concentration profiles on a Cartesian mesh at $t =$ 10 years obtained through HMM–ELLAM and the HMM–MMOC schemes in Figure 5.24. This is followed by a solution obtained by a HMM–GEM scheme in Figure 5.25. These are accompanied by Table 5.3, which presents some important features, such as the number of points tracked along each edge, overshoots, $e^{(N)}_{\text{mass}}$, and the approximate amount of oil recovered after 10 years, $|\Omega|_\phi^{-1} \int_\Omega \phi \Pi_{\mathcal{C}} c^{(N)}$. Here, $e^{(N)}_{\text{mass}}$ refers to the accumulated mass balance error (percentage) obtained over all the time steps, i.e.

$$e^{(N)}_{\text{mass}} = \frac{\left| \int_\Omega \phi \Pi_{\mathcal{C}} c^{(N)} - \int_\Omega \phi \Pi_{\mathcal{C}} c^{(0)} - \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \int_\Omega (q^+ - \Pi_{\mathcal{C}} c q^-)^{(n,w)} \right|}{\left| \int_\Omega \phi \Pi_{\mathcal{C}} c^{(0)} + \sum_{n=0}^{N-1} \delta t^{(n+\frac{1}{2})} \int_\Omega (q^+ - \Pi_{\mathcal{C}} c q^-)^{(n,w)} \right|}.$$

In practice, we have $e^{(N)}_{\text{mass}} = \max_{n=1,\ldots,N} e^{(n)}_{\text{mass}}$, as the mass balance error accumulates over each time step (in the numerical tests, no compensation is observed).

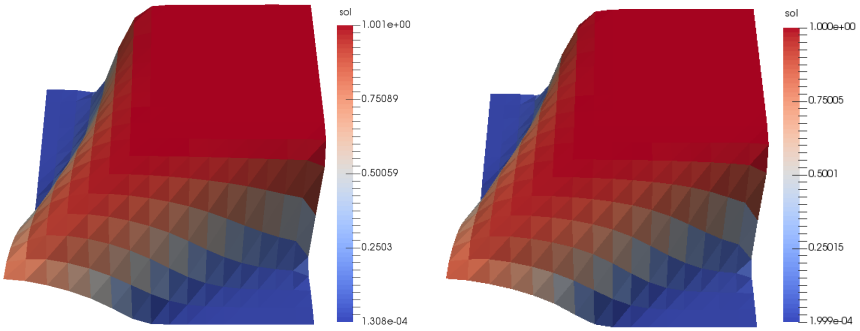


Figure 5.24: concentration profile at $t = 10$ years, Cartesian mesh (left: HMM–ELLAM, right: HMM–MMOC).

Upon comparing the concentration profiles, we see that for all of our numerical schemes, the overshoot is very low, with the maximum overshoot being less than 0.2%. However, it can be noted in Table 5.3 that ELLAM's 0.18% overshoot is much larger, by a factor of almost 20, than those of the

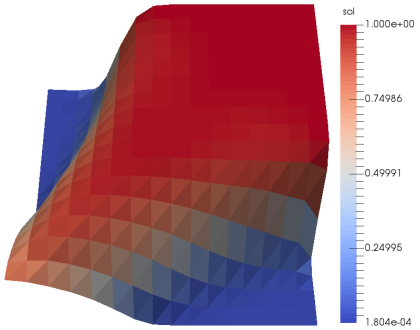


Figure 5.25: concentration profile at $t = 10$ years, Cartesian mesh, HMM–GEM.

MMOC and GEM, even on a very simple square mesh. Aside from the overshoots, there are no noticeable differences between the concentration profiles for the HMM–ELLAM and the HMM–GEM. On the other hand, we note that the HMM–MMOC scheme introduces some artificial diffusion along the diagonal, which slightly smears the expected fingering effect.

Table 5.3: Comparison between HMM–ELLAM, HMM–MMOC and HMM–GEM schemes, Cartesian mesh.

	points per edge	overshoot	$e^{(N)}_{\text{mass}}$	recovery
HMM–ELLAM	1	1.11%	0.19%	70.09%
HMM–ELLAM	3	0.18%	0.21%	69.76%
HMM–MMOC	1	$< 0.01\%$	5.60%	71.97%
HMM–MMOC	3	$< 0.01\%$	2.80%	69.94%
HMM–GEM	1	$< 0.01\%$	2.35%	68.44%
HMM–GEM	3	$< 0.01\%$	0.85%	69.14%

Next, upon comparing the approximate amount of oil recovered after 10 years, the 68.44% to 69.14% obtained for the HMM–GEM scheme is comparable to the amount from the HMM–ELLAM, which ranges from 69.76% to 70.09%. The HMM–MMOC, on the other hand, provides an overestimate of the oil recovered when the tracked cells are approximated only by vertices and edge midpoints, due to the excess diffusion it introduces along the diagonal becoming more prominent. When 3 points are tracked along each edge, the amount of oil recovered for HMM–MMOC is almost the same as that for HMM–ELLAM and HMM–GEM.

Lastly, we compare the mass balance errors. In particular, we focus on the mass balance errors obtained once we track 3 or more points along each

edge. The error obtained from the GEM (0.85%) is much better than the one from MMOC (2.80%), and close to that obtained from ELLAM (0.21%). These results agree with the analysis provided in Sections 4.4.4 and 4.5.2, due to the fact that the MMOC will fail to conserve mass as soon as the fluid starts invading the production well (which translates to $|\Pi_{\mathcal{C}} c^{(n)} - \Pi_{\mathcal{C}} c^{(n+1)}|$ being large on $F_{[-\delta t^{(n+\frac{1}{2})},0]}(C_-)$).

We then compare the numerical results on hexahedral meshes. Unlike the regular square cells for the Cartesian type meshes, the cells for the hexahedral meshes are irregular. As discussed in Section 5.3.5.1, since the hexahedral meshes have cells with highly varying areas, n_K points need to be tracked along each edge of cell K in order to have a good polygonal approximation of the trace-back and trace-forward regions, which will then be post-processed as described in Section 5.1.1. For a complete presentation, three tests were performed for the HMM–ELLAM: the first of which does not involve any adjustment to achieve local mass conservation, followed by an adjustment only on the cells that are not involved with either the injection or production cells (i.e. the algorithm in Section 5.1.1 without (5.5)), and finally an adjustment based on the full algorithm in Section 5.1.1.

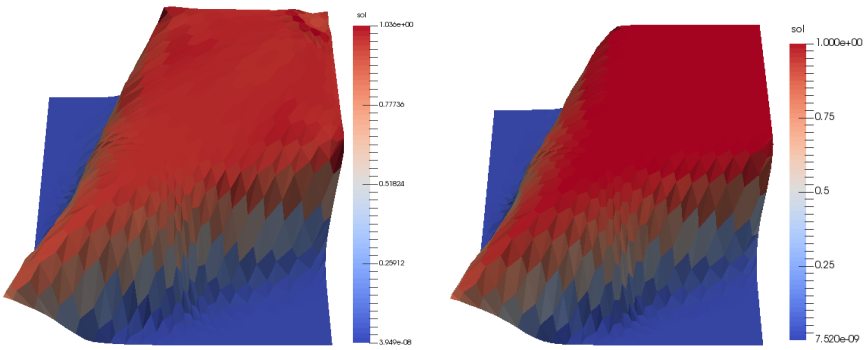


Figure 5.26: concentration profile at $t = 10$ years, hexahedral mesh (left: HMM–ELLAM (no adjustments), right: HMM–MMOC).

As was noted in Section 5.3.7, (5.5) does not give a good approximation, which means that the accumulated error around this region does not spread and cancel out properly, and hence severely distorts the quality of the numerical solution of the HMM–ELLAM. Contrary to the HMM–ELLAM, due to the absence of (5.5), this problem is not as severe with the GEM scheme, and can be solved by taking a slightly smaller time step of $\Delta t = 18$.

With the exception of the case for which ELLAM is adjusted with the local volume constraint (5.5), the amount of oil recovered from all three schemes are comparable, as they are within 2% of each other. Upon comparing the

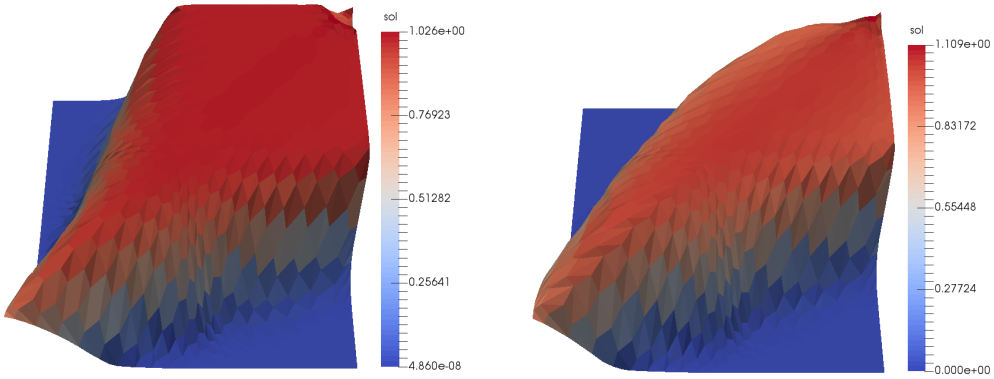


Figure 5.27: Concentration profile at $t = 10$ years, hexahedral mesh, HMM–ELLAM (left: local volume conservation without (5.5) , right: local volume conservation with (5.5)).

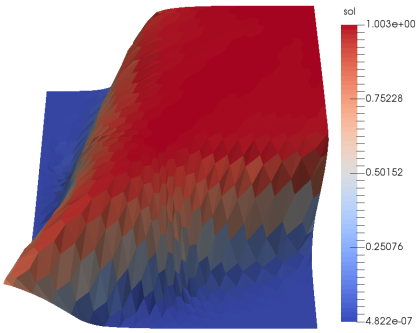


Figure 5.28: concentration profile at $t = 10$ years, hexahedral mesh, GEM scheme.

Table 5.4: Comparison between HMM–ELLAM, HMM–MMOC and HMM–GEM scheme, hexahedral mesh, $\Delta t = 18$ days.

	points per edge	overshoot	e_{mass}	recovery
HMM–ELLAM (no adjustment)	$\lceil \log_2(m_{K\text{reg}}) \rceil$	3.65%	0.62%	62.50%
HMM–ELLAM (adjustment without (5.5))	n_K	2.56%	0.26%	63.92%
HMM–ELLAM (adjustment with (5.5))	n_K	10.90%	0.44%	58.49%
HMM–MMOC	$\lceil \log_2(m_{K\text{reg}}) \rceil$	$< 0.01\%$	1.82%	61.43%
HMM–GEM	n_K	0.34%	0.54%	64.02%

mass balance errors, we note that the HMM–GEM (0.54%) outperforms the HMM–MMOC (1.82%), and is in the same range as the best HMM–ELLAM implementation (0.26%). This example shows that, on non-Cartesian meshes, due to the absence of (5.5), the HMM–GEM is able to provide a better solution, with reduced overshoots and acceptable mass conservation properties, compared to the HMM–ELLAM method. Now, we compare the results on a nonconforming mesh.

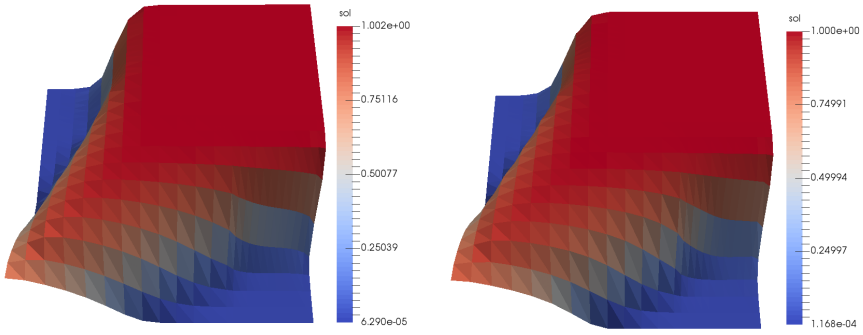


Figure 5.29: concentration profile at $t = 10$ years, nonconforming mesh (left: HMM–ELLAM, right: HMM–MMOC).

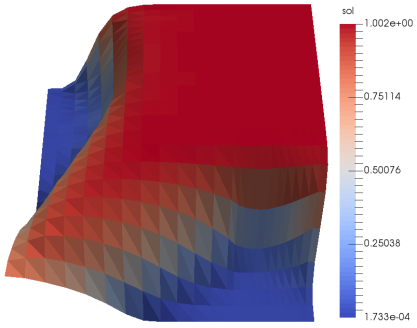


Figure 5.30: concentration profile at $t = 10$ years, nonconforming mesh, HMM–GEM.

As expected, Figures 5.29 to 5.30 show concentration profiles which are quite similar to those obtained from a Cartesian mesh. However, we take note that in Table 5.5, n_K points are tracked along the edge of each cell (as discussed in Section 5.3.5.1), which is much more than the usual $\lceil \log_2(m_{K\text{reg}}) \rceil$. In particular, for the nonconforming mesh, this is explained by the fact that the cells near and around the injection well are very small, and hence the trace-forward regions of these cells should be approximated with more

Table 5.5: Comparison between HMM–ELLAM, HMM–MMOC and HMM–GEM scheme, nonconforming mesh (Note: The $*$ in the third row does not represent an undershoot, i.e. it means that c attains a maximum value of 0.8819 and its minimum value is still positive.).

	points per edge	overshoot	e_{mass}	recovery
HMM–ELLAM	$\lceil \log_2(m_{K\text{reg}}) \rceil$	0.16%	0.49%	69.17%
HMM–MMOC	$\lceil \log_2(m_{K\text{reg}}) \rceil$	$< 0.01\%$	5.22%	70.97%
HMM–GEM	$\lceil \log_2(m_{K\text{reg}}) \rceil$	$-11.81\%^*$	32.27%	53.41%
HMM–GEM	n_K	0.18%	0.28%	69.02%

points. Otherwise, initially very bad approximations of the local volumes will become much worse after local volume adjustments (third row of Table 5.5). We note here that such a bad behavior was not observed for the HMM–MMOC scheme, due to the absence of the local volume adjustments. Actually, if no local volume adjustments were made for the HMM–GEM, tracking $\lceil \log_2(m_{K\text{reg}}) \rceil$ points along the edge of each cell would yield an overshoot of 0.56% which is only slightly worse than HMM–ELLAM, and a global mass conservation of 0.36%, which is slightly better than that of the HMM–ELLAM.

Finally, we compare the numerical results on a much more challenging mesh, the Kershaw mesh.

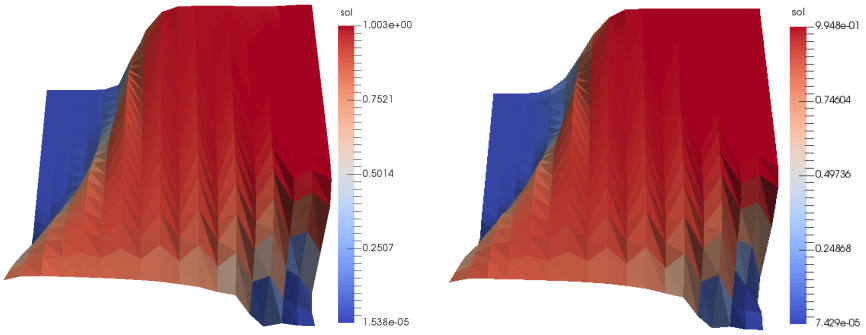


Figure 5.31: concentration profile at $t = 10$ years, Kershaw mesh (left: HMM–ELLAM, right: HMM–MMOC).

As with the Cartesian mesh test case, no significant difference can be observed between the numerical solutions obtained from HMM–ELLAM and HMM–GEM. Also, as expected, the mass balance error for the HMM–MMOC is quite large. Notably, the numerical solution on Kershaw type meshes is skewed towards the lower right corner. This is expected due to the fact that

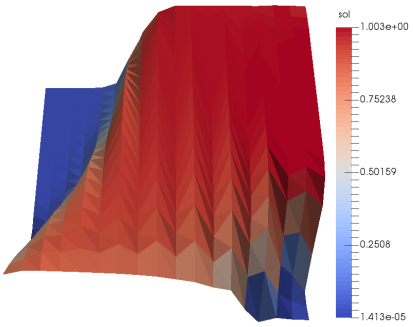


Figure 5.32: concentration profile at $t = 10$ years, Kershaw mesh, HMM–GEM.

Table 5.6: Comparison between HMM–ELLAM, HMM–MMOC and HMM–GEM scheme, Kershaw mesh

	points per edge	overshoot	e_{mass}	recovery
HMM–ELLAM	$\lceil \log_2(m_{K\text{reg}}) \rceil$	0.28%	0.38%	72.63%
HMM–MMOC	$\lceil \log_2(m_{K\text{reg}}) \rceil$	0%	4.28%	73.21%
HMM–GEM	$\lceil \log_2(m_{K\text{reg}}) \rceil$	0.32%	0.13%	72.36%

the numerical fluxes for HMM schemes on this type of mesh are prone to grid effects, as explained in [23].

To summarise the previous tests, the HMM–GEM exhibits a slightly better global mass conservation property than the HMM–ELLAM. This is due to the local volume constraint for the HMM–ELLAM being inexact in the sense that it depended on (5.5) to give a good approximation; whereas for the HMM–GEM scheme, exact local volume constraints were imposed. In particular, on non-Cartesian meshes with small or mildly distorted cells, HMM–GEM can control local volume constraints more easily and more accurately than HMM–ELLAM, as was seen in the test case on hexahedral meshes. Hence, with the choice of α driven by the discussion in Section 4.5.2, GEM achieves both a good preservation of the physical bounds on c, and of mass conservation.

We observe that the solutions on the Cartesian, hexahedral, and non-conforming meshes are very similar, showing a certain robustness of the method with respect to the choice of mesh. However, the solution on the Kershaw mesh is noticeably different, i.e. it is skewed towards the lower right region, which signals the presence of a grid effect. We consider streamlines to help us understand why grid effects are present in Kershaw type meshes.

5.5.2 Streamlines

We plot here the streamlines for the velocities reconstructed using A velocities on Cartesian, hexahedral, nonconforming and Kershaw meshes, which can be seen in Figures 5.33 and 5.34. For the following figures, the particles are assumed to have travelled for 3600 days, which is approximately 10 years.

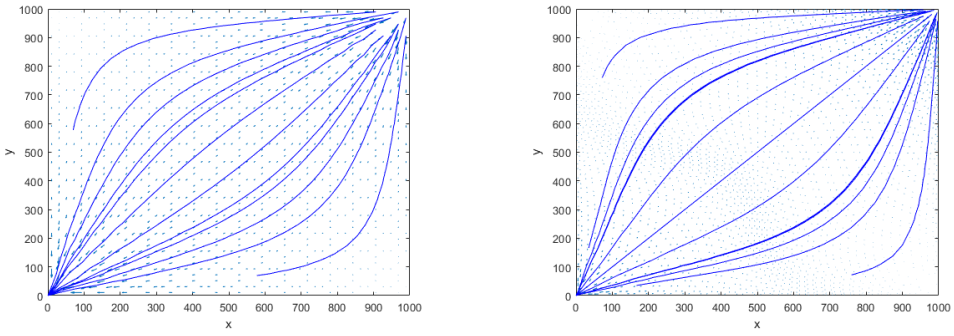


Figure 5.33: Streamlines at 3600 days, A velocities (left: Cartesian mesh; right: Hexahedral mesh).

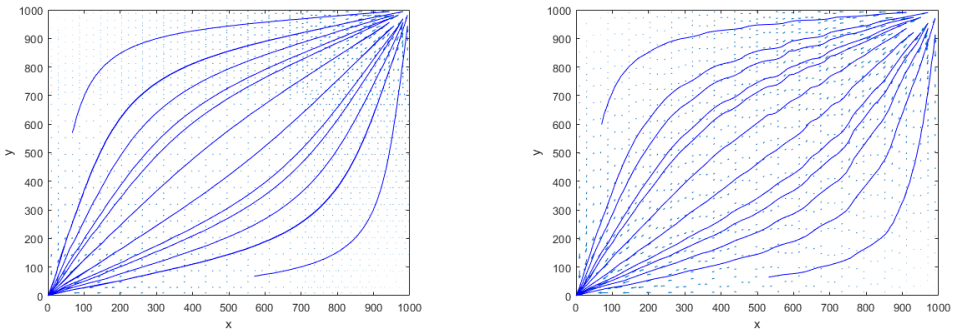


Figure 5.34: Streamlines at 3600 days, A velocities (left: non-conforming mesh; right: Kershaw mesh).

As can be seen in Figures 5.33 to 5.34, the streamlines along all four types of meshes are quite similar. The similarity between the streamlines on Cartesian, hexahedral, and nonconforming meshes explains why the numerical solutions obtained for the concentration on these meshes are very similar to one another. However, at this stage, the streamlines for the Kershaw meshes does not seem to be helpful on indicating why the grid effects are present. Upon careful comparison of the streamlines of hexahedral type

meshes against those of Cartesian or non-conforming meshes, we note that there is a slight difference in how the fluid travels. In particular, we take note that the streamline arising from the rightmost position of the plots ends near position (770, 60) for the hexahedral meshes, and near position (570, 60) for the other 2 meshes. This particular difference can also be seen in the concentration profiles upon comparing Figure 5.28 with Figures 5.25 and 5.30. In particular, the concentration profile obtained on a hexahedral mesh exhibits a sharper fingering effect along the diagonal, as compared to the concentration profile obtained on the other 2 meshes. This phenomenon is caused by using fluxes generated by the low-order HMM method, which are prone to grid effects, as discussed and demonstrated in Section 2.4; hence, in Section 5.6.2, we will be exploring, for the diffusive terms, the usage of high-order methods such as the HHO scheme (as presented in Section 2.3) to improve the quality of the numerical fluxes on distorted meshes. Now, to understand why grid effects are present on Kershaw type meshes, we track the particles in the streamline for a shorter time period of 2520 days, or approximately 7 years.

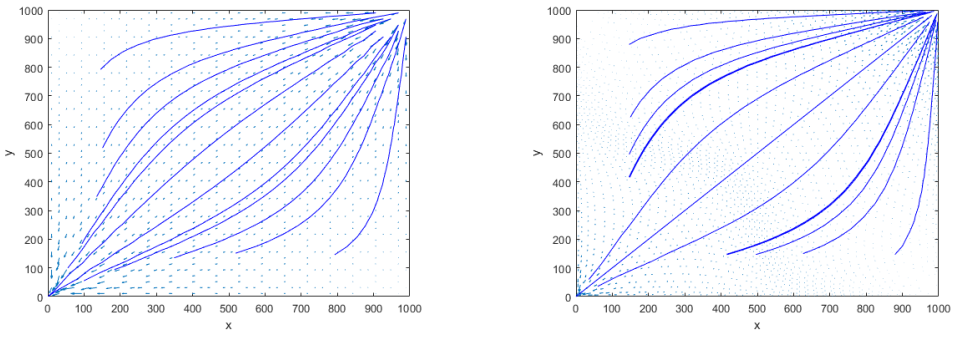


Figure 5.35: Streamlines at 2520 days, A velocities (left: Cartesian mesh; right: Hexahedral mesh).

As can be seen in Figure 5.35, the streamlines resulting from both Cartesian and hexahedral meshes are almost symmetric with respect to the line $y = x$. A similar observation can be made for non-conforming meshes (see Fig. 5.36 left). On the contrary, due to the large distortion of the Kershaw mesh, the advection field on this mesh is such that particles travelling below the line $y = x$ reach the production well $(0, 0)$ faster than those travelling above the line. In particular, upon looking at the plot on the right of Figure 5.36, we focus on the third streamlines from the right and top boundaries, which should be symmetric about the diagonal $x = y$. The streamline below $y = x$ has travelled near the point (275,100), whereas the streamline above

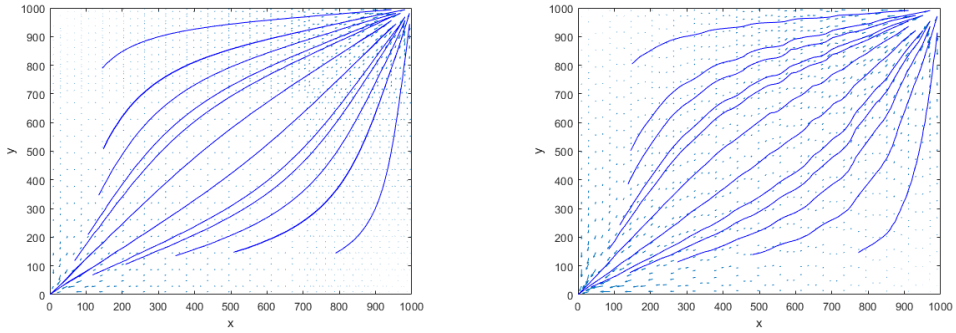


Figure 5.36: Streamlines at 2520 days, A velocities (left: non-conforming mesh; right: Kershaw mesh).

the line $y = x$ has only reached some point near (150,400). Hence, the distorted mesh leads to an advection of the fluid that is skewed towards the lower part of the domain, thus leading to numerical results that vary from those obtained in the other types of meshes. We may also compare these third streamlines to the third streamlines obtained from the other types of meshes. This comparison confirms that, for Kershaw meshes, advection below the line $y = x$ is much faster than expected.

Remark 5.5.1. *Figures 5.33 to 5.36 were obtained from the velocity profile at the first time step; hence, the dependency of the velocity profile on the concentration c due to a high mobility ratio of* $M = 41$ *was not visible. To complete the presentation, we show in Figures 5.37 and 5.38 the velocity profile obtained at the final time step. Also, the particles along the streamline are assumed to have traveled 3600 days, or approximately 10 years. Indeed, upon looking at these figures side by side with Figures 5.24 to 5.32, we see the dependence of the velocity profile on the concentration, i.e., it tends to flow along the region(s) with high concentration.*

Streamlines for KR and C velocities have been plotted in [23], but are not presented here, since similar observations are obtained: streamlines look symmetric for Cartesian, hexahedral and nonconforming meshes, and are skewed towards the lower right corner for Kershaw meshes.

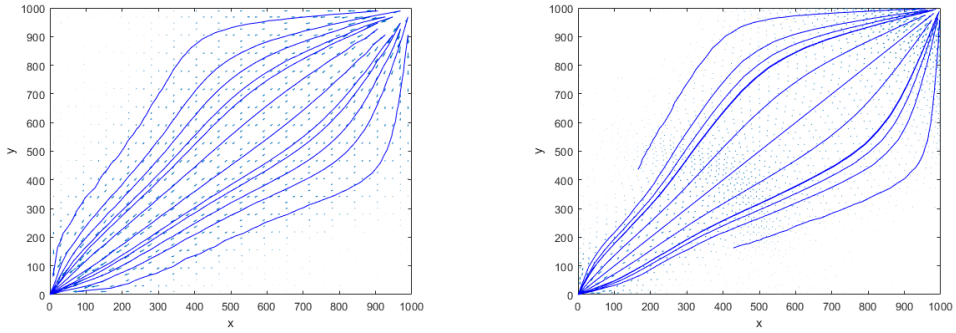


Figure 5.37: Streamlines using velocity profile at final time step, A velocities (left: Cartesian mesh; right: Hexahedral mesh).

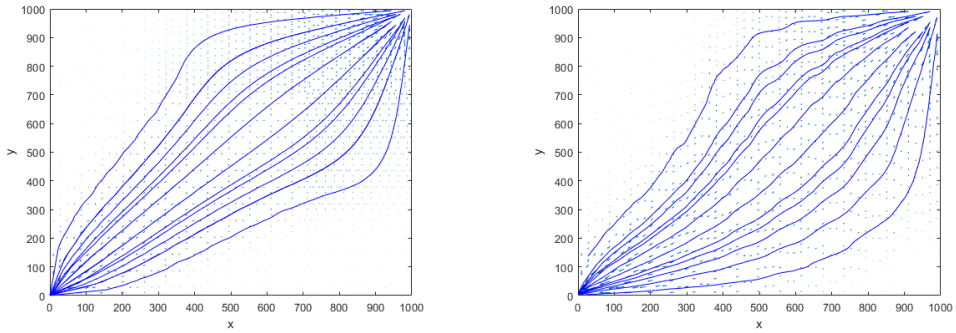


Figure 5.38: Streamlines using velocity profile at final time step, A velocities (left: non-conforming mesh; right: Kershaw mesh).

5.6 Studying the grid effects

5.6.1 Less distorted grids

In this section, we study the sensitivity of numerical solutions with respect to the distortion of a grid in more detail by creating less distorted cells. Consider a square domain $\Omega = (x_L, x_U) \times (y_L, y_U)$, with $y_U - y_L = x_U - x_L$. Without loss of generality, we may translate the domain onto the square $\hat{\Omega} = (0, L) \times (0, L)$, where $L = (x_U - x_L)$. Upon partitioning it into a Cartesian mesh with $N \times N$ square cells, where $N \geq 2$, each cell will have a side length of $h = \frac{L}{N}$. Starting with this Cartesian mesh, we create two sequences of meshes with varying degrees of distortion, where a highly distorted mesh of the first type would consist of very thin rectangular cells, whereas a highly distorted mesh of the second type is similar to a Kershaw type mesh.

Remark 5.6.1. *In this section, mesh refinement refers to starting with a* $2N \times 2N$ *Cartesian mesh, and performing the same adjustments described below for thin rectangular meshes and Kershaw like meshes.*

5.6.1.1 Thin rectangular meshes

We start by defining the thinness factor $\beta \in (0, 1)$, which will give rectangles of width $\beta\ell$, where ℓ is the length of the rectangle. Next, we specify the location and the number of cells to perturb. After the adjustments, since the width has been cut down, we have a mesh for a rectangular region with length L and width $w < L$. Finally, we scale this rectangular region so that we obtain again a mesh for the domain $(0, L) \times (0, L)$. For our case, we will be adjusting the vertical component of the mesh, using $\beta = 0.25$, with thin regions near the top, centre, and bottom of the mesh, respectively (see Figures 5.39–5.40, top). With this choice of β, $m_{\mathcal{M}_{\text{reg}}} = 3.0596$, and hence at most 2 to 3 points need to be tracked along the edge of each cell.

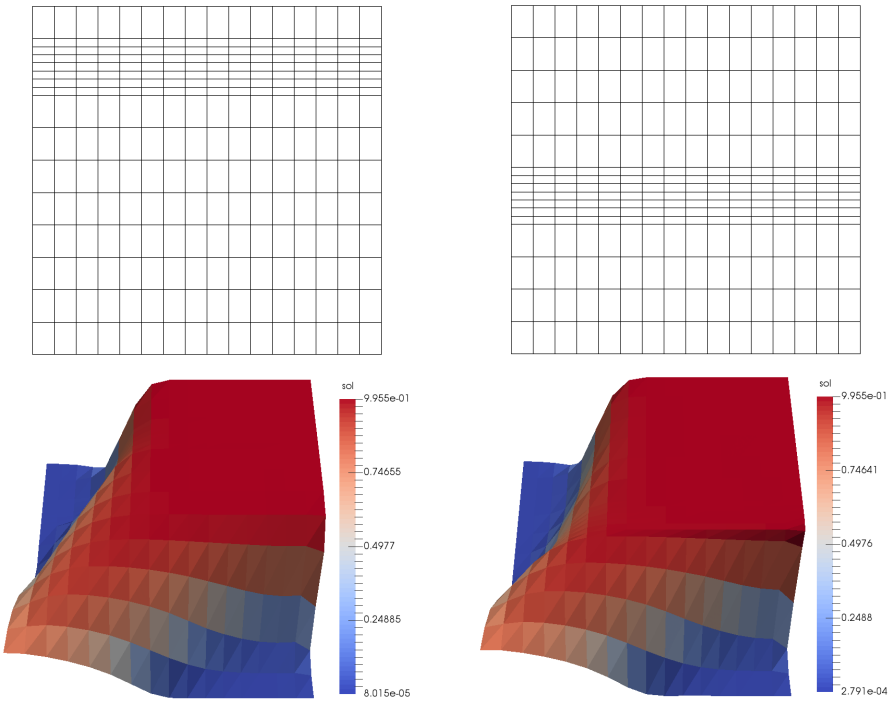


Figure 5.39: top: mesh; bottom: concentration profile at $t = 10$ years, HMM–GEM; left: thin rectangular regions near the top of the mesh ; right: thin rectangular regions near the middle of the mesh.

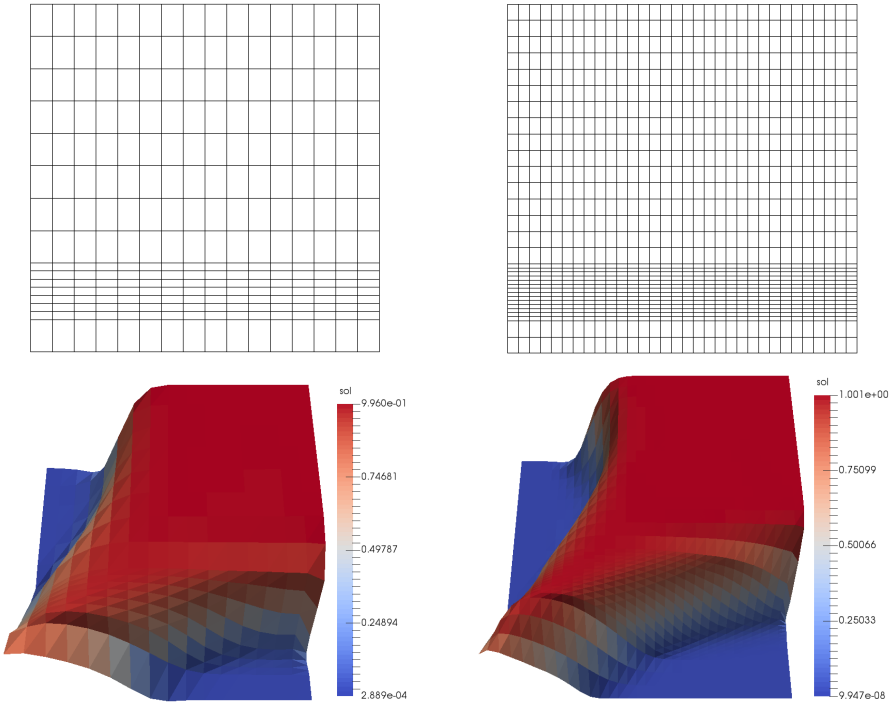


Figure 5.40: top: mesh, thin rectangular regions near the bottom of the mesh; bottom: concentration profile at $t = 10$ years, HMM–GEM; left: without refinement; right: refined mesh.

Figures 5.39 and 5.40, bottom, exhibit the numerical results upon perfoming an HMM–GEM scheme of the test case described in Section 5.2 on the meshes with thin rectangular elements. Here, we note that the effect of the mesh distortion is minimal if it is located near the injection well or around the middle region of the mesh, as seen in Figure 5.39. On the other hand, if the distortion occurs near the production well (see Figure 5.40, left), the sharp fingering effect along the diagonal has slightly been smeared and spread towards the lower right region of the mesh. In particular, the sweeped region, which initially looked like a square at the top right for most test cases (Cartesian, non-conforming, and those of the thin rectangular regions near the top and middle of the mesh), now looks rectangular. Upon performing a mesh refinement, we recover the expected fingering effect along the diagonal, and the skewness of the concentration profile towards the lower right region has been reduced, as observed in Figure 5.40, right. Also, the sweeped top-right region now looks like a square, which agrees with the results from the other test cases. We now explore the effect of locally refining the mesh near and around the very thin cells. In Figure 5.41 left, each of the rect-

angles have been divided into four regions of equal area, whereas in Figure 5.41 right, each of the rectangles have been divided into nine regions of equal area. Of course, cutting the rectangles into equal parts maintains the aspect ratio and hence we still have $m_{\mathcal{M}_{\text{reg}}} = 3.0596$. Following the discussion in Section 5.3.5.1, although the aspect ratio was maintained, this local refinement resulted to having to track n_K, instead of $\lceil \log_2(m_{K\text{reg}}) \rceil$ points, along the edges of each cell (since these cells are now much smaller than the time step) in order to have a good approximation of the trace-back regions.

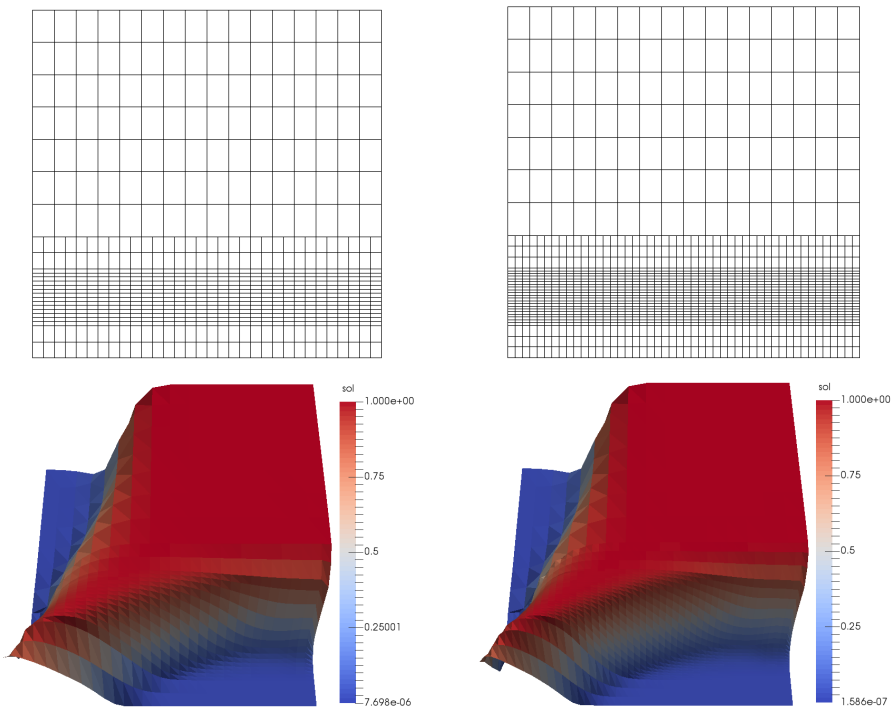


Figure 5.41: top: mesh, thin rectangular regions near the bottom of the mesh; bottom: concentration profile at $t = 10$ years, HMM–GEM; left: local refinement into 2×2 rectangles of equal size; right: local refinement into 3×3 rectangles of equal size.

Here, we note that the local refinement helped improve the resolution around the diagonal (which now features the sharp fingering effect prominently). Only a slight improvement was observed for the sweeped region at the right side of the mesh, since it is still slightly rectangular, as compared to the square shape featured in Figure 5.40, lower right, where the mesh was fully refined.

5.6.1.2 Kershaw-like meshes

To construct these meshes, we start by traversing the cells in the mesh horizontally, adjusting every other vertex by performing a translation in the vertical direction. In particular, we adjust the vertices $\{(x_k, y_j)|x_k = L-(2\lfloor N/2\rfloor -2k-1)h, y_j = jh$ where $k = 1, 2, \ldots \lfloor N/2\rfloor, j = 1, 2, \ldots N-1\}$. As with Kershaw type meshes, we want the distortion to be most pronounced at the central region of the mesh. This is achieved by scaling the adjustment factor with respect to the distance of (x_k, y_j) towards the boundary of our domain. Denoting by (x_C, y_C) the centre of the domain, we generate a family of mesh sequences which has distortions concentrated on the central region $(x_C - \alpha h, x_C + \alpha h) \times (0, L)$. Denote by K_{x_k} and K_{y_j} the distance of x_k and y_j respectively to the closest endpoint of the segment $(0, L)$, i.e. $K_{x_k} = \min(x_k, L - x_k), K_{y_j} = \min(y_j, L - y_j)$. For each k, define $\alpha = \frac{2}{L}K_{x_k}, \beta = \frac{2}{L}K_{y_j}$. The point (x_k, y_j) is then translated to the point

$$(\hat{x}_k, \hat{y}_j) := (x_k, y_j + Ch\alpha\beta) \tag{5.10}$$

for some constant $C \geq 0$. We note here that C will dictate the amount of distortion in the mesh. A small value of C will lead to a mesh that is only slightly distorted, whereas a large value of C will lead to a highly distorted mesh. In order to preserve the cell-edge connectivity of the mesh, we require that the mesh obtained after distortion has all cells being quadrilaterals. We also need to check that we still have an admissible mesh after adjusting the vertices by (5.10). To be specific, we need first to restrict C so that $y_N > \hat{y}_{N-1}$. Hence, we must have

$$\begin{aligned} y_N - \hat{y}_{N-1} &= h - 4\frac{Ch}{L^2}K_{x_k}K_{y_{N-1}} \\ &= h - 4\frac{C}{N^2}K_{x_k} \\ &> 0 \text{ for all } k. \end{aligned}$$

In particular, this should hold for the extreme case when $K_{x_k} = \frac{L}{2}$. This will then imply that $C < \frac{N}{2}$. Now, we need to show for each k that $\hat{y}_{j+1} > \hat{y}_j$ for all j.

$$\begin{aligned} \hat{y}_{j+1} &= y_{j+1} + 4\frac{Ch}{L^2}K_{x_k}K_{y_{j+1}} \\ \hat{y}_{j+1} - \hat{y}_j &= h + 4\frac{Ch}{L^2}K_{x_k}(K_{y_{j+1}} - K_{y_j}). \end{aligned} \tag{5.11}$$

Indeed if $K_{y_s} = y_s$ or $K_{y_s} = L - y_s$ for both $s = j$ and $s = j + 1$, then it is clear that $\hat{y}_{j+1} > \hat{y}_j$. We are left to show that this is true for $K_{y_j} = y_j$ and

$K_{y_{j+1}} = L - y_{j+1}$. This can be deduced by substituting the above expressions into (5.11) and using the fact that $|K_{y_{j+1}} - K_{y_j}| < h$ and $C < \frac{N}{2}$. For our numerical tests, for $N = 16$, we consider $C = 2, 3, 4$ respectively (see Figures 5.42–5.43, top). Here, the regularity factor is $m_{\mathcal{M}_{\text{reg}}} = 8.3169, 14.1989,$ and 22.4050 for $C = 2, 3, 4$ respectively. For the Kershaw-like meshes, the distortion in the numerical solution gets evident starting with the distortion factor of $C = 4$ (see Figure 5.43, lower left). As can be seen in Figure 5.43, lower right, the skewness of the numerical solutions towards the lower right region has been reduced upon performing a mesh refinement. However, for the refined mesh, n_K points needed to be tracked along the edge of each cell in order to have a good approximation of the trace-back region. This agrees with the observations made on the very thin rectangular meshes, and hence indicates that the grid effects might be caused by a lack of accuracy in the approximation on space.

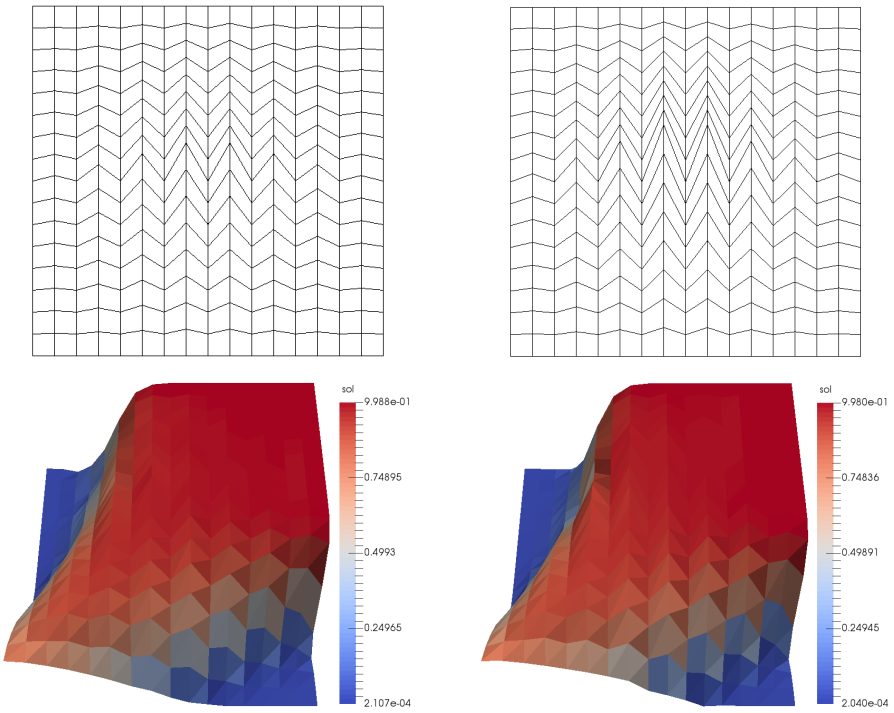


Figure 5.42: top: mesh; bottom: concentration profile at $t = 10$ years, HMM–GEM; left: distortion factor $C = 2$; right: distortion factor $C = 3$.

Remark 5.6.2. *For thin rectangular meshes, the refinement is "exact" in the sense that every rectangle is cut into smaller rectangles, all with the same size. For Kershaw-like meshes, the refinement is inexact. However,*

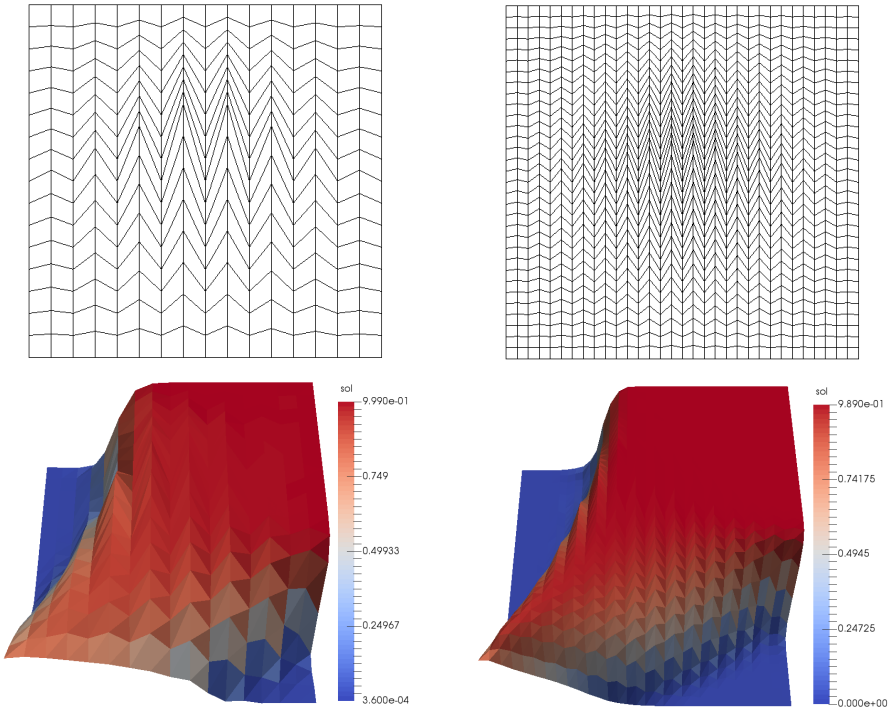


Figure 5.43: top: mesh, distortion factor $C = 4$; bottom: concentration profile at $t = 10$ years, HMM–GEM; left: without refinement; right: refined mesh

the refined mesh shares a similar property to the original mesh, i.e. the mesh regularity factor of the refined mesh is very close to that of the original mesh. For example, with the distortion factor $C = 4$, the refined mesh has a mesh regularity factor of $m_{\mathcal{M}_{\text{reg}}} = 24.2304$, which is close to the value of 22.4050 on the coarse mesh. Hence, the numerical results obtained from this refinement is comparable to the numerical results that could be obtained from an actual refinement.

5.6.2 Using a high order approximation in space

The streamlines presented in Section 5.5.2, and the tests on the $\mathbb{RT}_0$ velocities in Section 3.3.1 hint that a high order approximation should be performed for the pressure equation (1.1a). Also, the numerical results on the thin rectangular meshes and Kershaw-like meshes indicate that a high order approximation in space might be needed for the concentration c in (1.1b). Here, we present a partially high order approximation for the concentration c by performing a splitting technique on the concentration equation (1.1b) [5, 74].

That is, starting with a high order approximation of c, we project onto the space of piecewise constant functions, and take, over one time step, an approximation of the hyperbolic part of the equation

$$\phi\partial_t c + \nabla\cdot(\mathbf{u}c) = q_c,$$

via the combined ELLAM–MMOC scheme. Using the value of c obtained from this scheme as an initial condition, we then restart and approximate, over one time step, the parabolic part

$$\phi\partial_t c - \nabla\cdot(\mathbf{D}(\mathbf{x},\mathbf{u})\nabla c) = 0,$$

by the HHO scheme. We now present in Figure 5.44 the concentration profiles obtained upon implementing this HHO–GEM scheme (which is a GEM scheme with a gradient discretisation taken to be the GD of HHO, combined with a projection when applying the characteristics), with degree $k = 2$ for the HHO, on the thin rectangular meshes (on the lower region) and on the Kershaw-like meshes with a distortion factor of $C = 4$.

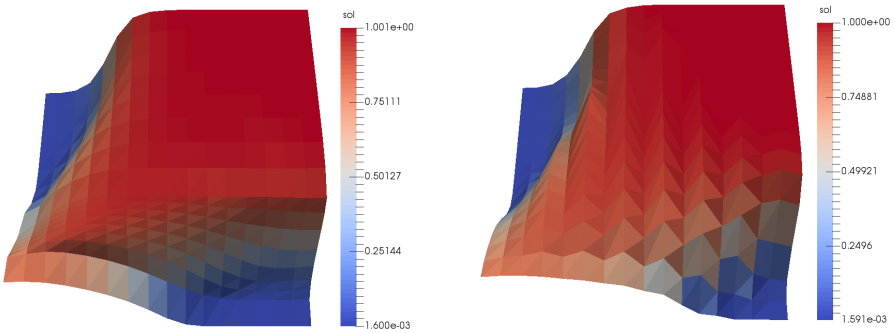


Figure 5.44: Concentration profile at $t = 10$ years, HHO–GEM, triangular $\mathbb{RT}_0$ elements for Darcy velocities (left: thin rectangular mesh ; right: Kershaw-like mesh).

As can be observed, upon comparing Figure 5.44 with Figures 5.40, left and 5.43, left, only a slight improvement is obtained upon performing an HHO–GEM scheme over an HMM–GEM scheme. One possible factor that might have caused this is the fact that $\mathbb{RT}_0$ reconstructions do not give a good approximation to velocity fields on generic meshes, as observed in Section 3.3.1. Hence, we try to further improve the accuracy of our approximation by approximating the velocity field using quadrilateral $\mathbb{RT}_k$ elements instead. We start by looking at the concentration profiles on a thin rectangular mesh obtained by using an HHO–GEM scheme with quadrilateral $\mathbb{RT}_0$ and $\mathbb{RT}_2$ approximations to the velocity field in Figure 5.45.

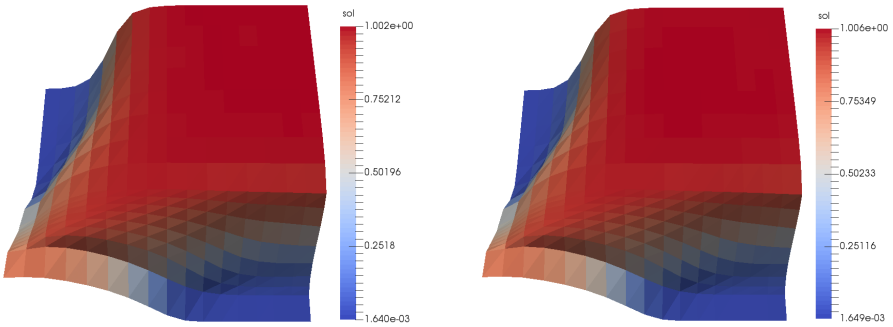


Figure 5.45: Concentration profile at $t = 10$ years, HHO–GEM, rectangular $\mathbb{RT}_k$ elements for Darcy velocities (left: $\mathbb{RT}_0$; right: $\mathbb{RT}_2$).

As expected, upon comparing the Figures 5.44 and 5.45, left, using quadrilateral $\mathbb{RT}_0$ elements do not give much improvement compared to $\mathbb{RT}_0$ elements on simplices. However, even with the more accurate quadrilateral $\mathbb{RT}_2$ elements (Figure 5.45, right), the grid effects are still present. Since this improvement on the approximation of the velocity field does not help mitigate the grid effects for thin rectangular meshes, we do not expect it to mitigate the grid effects in the Kershaw-like mesh, for which the distortion is much worse.

We note that for the concentration profile on Figure 5.45, right, only the hyperbolic part of the concentration equation is approximated using piecewise constants. Due to this, we believe that going for a fully high order approximation of c might be able to mitigate the grid effects on coarse grids; however, this comes with a lot of new problems to deal with. Firstly, the nice features of piecewise constant approximations, for which computing the integrals boil down to finding intersections of polygonal regions, is no longer present. This leads to the problem of how to efficiently perform numerical integration over the intersections of polygonal regions. Secondly, since we now have to find accurate quadrature points over the tracked polygonal regions, the algorithm described in Section 5.1.1 can no longer be used. In particular, an explicit expression for the mesh is needed, hence local volume conservation should be imposed in a manner similar to that of [27]. Thirdly, this comes with a computational cost that is much more expensive and less efficient compared to solving the PDE using the current HMM–GEM on a refined mesh, or using a fully high order Crank Nicholson scheme as in [2]. Hence, at this stage, mitigation of grid effects on coarse distorted grids for characteristic-based schemes is still an open problem.

Chapter 6

Convergence analysis for the GDM–characteristic schemes for the Peaceman model

In this chapter, we state and prove convergence of the GEM, GDM–ELLAM and GDM–MMOC schemes for the miscible flow model (1.1). In particular, we will start by establishing the proofs for the convergence of the GDM–ELLAM scheme. These schemes include, but are not limited to, Mixed Finite Element–ELLAM and Hybrid Mimetic Mixed–ELLAM schemes. A complete convergence analysis is presented on the coupled model, using only weak regularity assumptions on the solution (which are satisfied in practical applications), and not relying on L^∞ bounds (which are impossible to ensure at the discrete level given the anisotropic diffusion tensors and the general grids used in applications). This will be followed by an outline of the changes needed to be done in order to adapt the proof of the GDM–ELLAM scheme to the GDM–MMOC scheme. Finally, the convergence of the GEM scheme can be established by combining the elements of the proof from both GDM–ELLAM and GDM–MMOC schemes.

We start by presenting the weak formulation of the model (1.1). Section 6.1 then presents the main results: existence and uniqueness of the solution to the GEM scheme, and its convergence to the weak solution of (1.1) under weak regularity assumptions. We then focus on the GDM–ELLAM, which corresponds to taking Definition 5.1.1 with $\alpha = 1$ in (5.3). Since ELLAM schemes are based on characteristic methods, we need to solve characteristics along which the solution flows. Existence of the flow and some of its basic properties were established in Section 4.2. In Section 6.2, we present other useful properties of the flow (related to translation estimates). These

properties are not trivial to establish due to the weak regularity assumptions on the reconstructed velocities. Section 6.3 then gives some of the numerical schemes that are covered by the GEM framework, together with proofs on why they satisfy the assumptions (**A1**)–(**A5**), which are required for our convergence analysis.

A priori estimates are then obtained in Section 6.4, which lead us to compactness arguments that will help establish the proof of convergence. We then prove the convergence result for GDM–ELLAM in Section 6.5. The ELLAM discretisation of the advection term makes the energy estimates and the convergence analysis of the corresponding terms rather tricky. The results from Sections 6.2 and 6.4 are instrumental to obtain the major estimates and the proper convergence of the advection term.

We then extend this convergence result to GDM–MMOC schemes, by introducing only slight modifications to the proof. Finally, convergence of the GEM scheme will be established. We note that at the core of our convergence analysis lies some generic compactness results of [40], which are flexible enough to be used even outside a purely GDM framework (as in the GDM–ELLAM, GDM–MMOC, and GEM framework here).

Throughout the chapter we assume the following properties, satisfied by $\mathbf{D}$, $\mathbf{K}$ and μ described in model (1.1).

$$c_{\text{ini}} \in L^\infty(\Omega) \text{ and } q^+, q^- \in L^\infty(\Omega \times (0,T)) \text{ with } |q^+| \le M_{q^+},\ |q^-| \le M_{q^-}. \tag{6.1a}$$

$$\begin{aligned}&\phi \text{ is piecewise smooth on a mesh, and}\\ &\text{there exists } \phi_*, \phi^* > 0 \text{ such that } \phi_* \le \phi \le \phi^* \text{ on } \Omega.\end{aligned} \tag{6.1b}$$

$$\begin{aligned}&A := \mathbf{K}/\mu \text{ is Carathéodory and there exists } \alpha_A \text{ and } \Lambda_A \text{ s.t. for a.e. } \boldsymbol{x} \in \Omega,\\ &\forall (s,\xi) \in \mathbb{R} \times \mathbb{R}^d \ :\ A(\boldsymbol{x},s)\xi \cdot \xi \ge \alpha_A |\xi|^2 \text{ and } |A(\boldsymbol{x},s)| \le \Lambda_A.\end{aligned} \tag{6.1c}$$

$$\begin{aligned}&\mathbf{D} \text{ is Carathéodory and there exists } \alpha_{\mathbf{D}} \text{ and } \Lambda_{\mathbf{D}} \text{ s.t. for a.e. } \boldsymbol{x} \in \Omega,\\ &\forall \xi, \zeta \in \mathbb{R}^d \ :\ \mathbf{D}(\boldsymbol{x},\zeta)\xi \cdot \xi \ge \alpha_{\mathbf{D}}(1+|\zeta|)|\xi|^2 \text{ and } |\mathbf{D}(\boldsymbol{x},\zeta)| \le \Lambda_{\mathbf{D}}(1+|\zeta|).\end{aligned} \tag{6.1d}$$

Here, "Carathéodory" means measurable with respect to $\boldsymbol{x}$ and continuous with respect to the other variables. As mentioned in Section 1.6, "mesh" is to be understood in the simplest intuitive way: a partition of Ω into polygonal (in 2D) or polyhedral (in 3D) sets. Under these assumptions, we consider the following standard notion of weak solution to (1.1) (see, e.g., [54]).

Definition 6.0.3 (Weak solution to the miscible displacement model)**.** *A couple* (p, c) *is a weak solution of* (1.1) *if*

$$
\begin{aligned}
&p \in L^\infty(0,T;H^1(\Omega))\,,\ \int_\Omega p(\boldsymbol{x},t)d\boldsymbol{x} = 0 \text{ for a.e. } t \in (0,T), \text{ and}\\
&\int_0^T \int_\Omega \frac{\mathbf{K}(\boldsymbol{x})}{\mu(c(\boldsymbol{x},t))} \nabla p(\boldsymbol{x},t)\cdot\nabla\psi(\boldsymbol{x},t)d\boldsymbol{x}\mathrm{d}t\\
&\qquad = \int_0^T \int_\Omega (q^+(\boldsymbol{x},t) - q^-(\boldsymbol{x},t))\psi(\boldsymbol{x},t)d\boldsymbol{x}dt\,, \quad \forall \psi \in C^\infty(\overline{\Omega}\times[0,T]),
\end{aligned}
\tag{6.2}
$$

and, setting $\mathbf{u}(\boldsymbol{x},t) = -\frac{\mathbf{K}(\boldsymbol{x})}{\mu(c(\boldsymbol{x},t))}\nabla p(\boldsymbol{x},t)$*,*

$$
\begin{aligned}
&c \in L^2(0,T;H^1(\Omega))\,,\ (1+|\mathbf{u}|)^{1/2}\nabla c \in L^2(\Omega\times(0,T))^d\,,\\
&-\int_\Omega \phi(\boldsymbol{x})c_{\mathrm{ini}}(\boldsymbol{x})\varphi(\boldsymbol{x},0)d\boldsymbol{x} - \int_0^T\int_\Omega \phi(\boldsymbol{x})c(\boldsymbol{x},t)\frac{\partial\varphi}{\partial t}(\boldsymbol{x},t)d\boldsymbol{x}dt\\
&\quad + \int_0^T\int_\Omega \mathbf{D}(\boldsymbol{x},\mathbf{u}(\boldsymbol{x},t))\nabla c(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}dt\\
&\quad - \int_0^T\int_\Omega c(\boldsymbol{x},t)\mathbf{u}(\boldsymbol{x},t)\cdot\nabla\varphi(\boldsymbol{x},t)d\boldsymbol{x}dt + \int_0^T\int_\Omega q^-(\boldsymbol{x},t)c(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)d\boldsymbol{x}dt\\
&\quad = \int_0^T\int_\Omega q^+(\boldsymbol{x},t)\varphi(\boldsymbol{x},t)d\boldsymbol{x}dt\,, \quad \forall\varphi\in C_c^\infty(\overline{\Omega}\times[0,T)).
\end{aligned}
\tag{6.3}
$$

6.1 Convergence results

We use the following notations: If $\mathcal{D}$ is a space GD, $0 = t^{(0)} < \cdots < t^{(N)} = T$ are time steps and $z = (z^{(n)})_{n=0,\ldots,N} \in X_{\mathcal{D}}^{N+1}$, we define the space–time reconstructions $\Pi_{\mathcal{D}}z \in L^\infty(\Omega\times(0,T))$, $\widetilde{\Pi}_{\mathcal{D}}z \in L^\infty(\Omega\times(0,T))$ and $\nabla_{\mathcal{D}}z \in L^\infty(\Omega\times(0,T))^d$ by

$$
\begin{aligned}
&\forall n = 0,\ldots,N-1\,,\ \forall t\in(t^{(n)},t^{(n+1)}]\,,\ \text{for a.e. } \boldsymbol{x}\in\Omega\,,\\
&\Pi_{\mathcal{D}}z(\boldsymbol{x},t) = \Pi_{\mathcal{D}}z^{(n+1)}(\boldsymbol{x})\,,\ \widetilde{\Pi}_{\mathcal{D}}z(\boldsymbol{x},t) = \Pi_{\mathcal{D}}z^{(n)}(\boldsymbol{x})\\
&\text{and } \nabla_{\mathcal{D}}z(\boldsymbol{x},t) = \nabla_{\mathcal{D}}z^{(n+1)}(\boldsymbol{x}).
\end{aligned}
$$

The convergence theorem is then established under the following assumptions.

(**A1**) $(\mathcal{P}_m)_{m\in\mathbb{N}}$ and $(\mathcal{C}_m^T)_{m\in\mathbb{N}}$ are coercive, GD-consistent and limit-conforming sequences of GDs, and $(\mathcal{C}_m^T)_{m\in\mathbb{N}}$ is moreover compact. Denoting by

$0 = t_m^{(0)} < \cdots < t_m^{(N_m)} = T$ the time steps of $\mathcal{C}_m^T$, it is assumed that there exists $M_t \geq 0$ such that, for all $m \in \mathbb{N}$ and $n = 1, \ldots, N-1$, $\delta t_m^{(n+1/2)} \leq M_t \delta t_m^{(n-1/2)}$.

(**A2**) There exists $M_F \geq 0$ such that, for all $m \in \mathbb{N}$, all $z \in X_{\mathcal{C}_m}$, all $n = 0, \ldots, N_m - 1$, and all $s \in [-T, T]$,

$$\|\Pi_{\mathcal{C}_m} z(F_s) - \Pi_{\mathcal{C}_m} z\|_{L^1(\Omega)} \leq M_F |s| \left\| \mathbf{u}_{\mathcal{P}_m}^{(n+1)} \right\|_{L^2(\Omega)} \|\nabla_{\mathcal{C}_m} z\|_{L^2(\Omega)},$$

where F_s is the flow defined by (4.5) with $\mathbf{V} = \mathbf{u}_{\mathcal{P}_m}^{(n+1)}$.

(**A3**) For all $m \in \mathbb{N}$ there is an interpolant $\mathcal{J}_{\mathcal{C}_m} : C^\infty(\overline{\Omega}) \to X_{\mathcal{C}_m}$ such that, for all $\varphi \in C^\infty(\overline{\Omega})$, as $m \to \infty$, $\nabla_{\mathcal{C}_m} \mathcal{J}_{\mathcal{C}_m} \varphi \to \nabla \varphi$ in $L^4(\Omega)^d$ and $\Pi_{\mathcal{C}_m} \mathcal{J}_{\mathcal{C}_m} \varphi \to \varphi$ in $L^\infty(\Omega)$.

(**A4**) There exists $M_{\text{div}} > 0$ such that, for all $m \in \mathbb{N}$ and $n = 0, \ldots, N_m - 1$, $\mathbf{u}_{\mathcal{P}_m}^{(n+1)} \in H_{\text{div}}(\Omega)$ is piecewise polynomial on a mesh, $\mathbf{u}_{\mathcal{P}_m}^{(n+1)} \cdot \mathbf{n} = 0$ on $\partial\Omega$, and $|\text{div} \mathbf{u}_{\mathcal{P}_m}^{(n+1)}| \leq M_{\text{div}}$ on Ω.

(**A5**) If $(p_m, c_m) \in X_{\mathcal{P}_m}^{N_m} \times X_{\mathcal{C}_m}^{N_m+1}$ is a solution to the GDM–ELLAM scheme with $(\mathcal{P}, \mathcal{C}^T) = (\mathcal{P}_m, \mathcal{C}_m^T)$ and $\mathbf{u}_{\mathcal{P}_m} : \Omega \times (0,T) \to \mathbb{R}^d$ is defined by $\mathbf{u}_{\mathcal{P}_m}(\cdot, t) = \mathbf{u}_{\mathcal{P}_m}^{(n+1)}$ for all $t \in (t_m^{(n)}, t_m^{(n+1)})$ and $n = 0, \ldots, N_m - 1$, then, when (**A1**)–(**A4**) hold:

(a) $\left\| \mathbf{u}_{\mathcal{P}_m} \right\|_{L^\infty(0,T;L^2(\Omega))} \leq C_m \|\nabla_{\mathcal{P}_m} p_m\|_{L^\infty(0,T;L^2(\Omega))}$ with $(C_m)_{m \in \mathbb{N}}$ bounded.

(b) if $p \in L^2(0,T; H^1(\Omega))$ and $c \in L^2(\Omega \times (0,T))$ are such that, as $m \to \infty$, $\Pi_{\mathcal{P}_m} p_m \to p$, $\Pi_{\mathcal{C}_m} c_m \to c$ and $\nabla_{\mathcal{P}_m} p_m \to \nabla p$ in $L^2(\Omega \times (0,T))$, then $\mathbf{u}_{\mathcal{P}_m} \to \mathbf{u} = -\frac{\mathbf{K}}{\mu(c)} \nabla p$ weakly in $L^2(\Omega \times (0,T))^d$.

We show in Section 6.3 that various finite element and finite volume methods are given by GDs that satisfy these assumptions.

Theorem 6.1.1 (Convergence of the GEM scheme)**.** *Under Assumptions* (6.1) *and* (**A1**)–(**A5**)*, for any $m \in \mathbb{N}$ there is a unique $(p_m, c_m) \in X_{\mathcal{P}_m}^{N_m} \times X_{\mathcal{C}_m}^{N_m+1}$ solution of the GEM scheme (Definition 5.1.1) with $(\mathcal{P}, \mathcal{C}^T) = (\mathcal{P}_m, \mathcal{C}_m^T)$. Moreover, there is a weak solution (p, c) of* (1.1)*, such that, up to a subsequence as $m \to \infty$,*

- *$\Pi_{\mathcal{P}_m} p_m \to p$ and $\nabla_{\mathcal{P}_m} p_m \to \nabla p$ weakly-$*$ in $L^\infty(0,T; L^2(\Omega))$ and strongly in $L^r(0,T; L^2(\Omega))$ for all $r < \infty$,*

- $\Pi_{\mathcal{C}_m} c_m \to c$ *weakly-*$*$ *in* $L^\infty(0,T;L^2(\Omega))$ *and strongly in* $L^r(0,T;L^2(\Omega))$ *for all* $r < \infty$,
- $\nabla_{\mathcal{C}_m} c_m \to \nabla c$ *weakly in* $L^2(\Omega \times (0,T))^d$.

We will first prove Theorem 6.1.1 for the GDM–ELLAM, that is, taking $\alpha = 1$ in (5.3). After which, a slight modification of the proof will be needed in order to establish the convergence of the GDM–MMOC. The proof of Theorem 6.1.1 will then follow by combining the proofs for GDM–ELLAM and GDM–MMOC.

Remark 6.1.2 (About the assumptions)**.** *Assumption* (**A1**) *is standard in analysis of gradient schemes, except for the assumption on the time steps, which is not very restrictive in practice (it is for example satisfied by uniform time steps, used in most numerical tests on* (1.1)*, see e.g. [20, 78]). Assumption* (**A2**) *is probably the most technical to check for specific methods; we however provide two results (Lemmas 6.2.1 and 6.2.3) which show that it is satisfied for a wide range of conforming or non-conforming methods. Assumption* (**A3**) *is satisfied by all standard interpolants associated with numerical methods for diffusion equations. Assumption* (**A4**) *is natural given the pressure equation* (1.1a) *and the boundedness assumption* (6.1a) *on* q^+ *and* q^-*. Finally, Assumption* (**A5**) *is also rather natural since it is expected that the reconstructed Darcy velocity* $\mathbf{u}_\mathcal{P}$ *is closely related to the reconstructed concentration* $\Pi_\mathcal{C} c$ *and pressure gradient* $\nabla_\mathcal{P} p$.

Remark 6.1.3 (One GD per time step)**.** *In some particular cases, most notably the discretisation via mixed finite elements (see Section 6.3.1.1), the gradient discretisation* $\mathcal{P}$ *changes with each time step. Each equation* (5.1) *is written with a specific gradient discretisation* $\mathcal{P}^{(n+1)}$*. Hence, the choice* $\mathcal{P}$ *of a gradient discretisation for the pressure actually amounts to choosing a family* $\mathcal{P} = (\mathcal{P}^{(i)})_{i=1,\dots,N}$*. Theorem 6.1.1 remains valid provided that the coercivity, GD-consistency and limit-conformity of a sequence* $(\mathcal{P}_m)_{m\in\mathbb{N}} = ((\mathcal{P}_m^{(i)})_{i=1,\dots,N_m})_{m\in\mathbb{N}}$ *of such families of GDs are defined as in Definition 2.1.3 with*

$$C_{\mathcal{P}_m} = \max_{i=1,\dots,N_m} C_{\mathcal{P}_m^{(i)}}, \quad S_{\mathcal{P}_m} = \max_{i=1,\dots,N_m} S_{\mathcal{P}_m^{(i)}} \quad \text{and} \quad W_{\mathcal{P}_m} = \max_{i=1,\dots,N_m} W_{\mathcal{P}_m^{(i)}}.$$

6.2 Properties of the flow

The following lemma is used to prove that conforming discretisations satisfy Assumption (**A2**) (see Section 6.3.1.2), and to establish convergence properties, as the time step tends to 0, of functions transported by the flow (see Lemma 6.2.5).

Lemma 6.2.1 (Translation estimate for Sobolev functions)**.** *Under Assumption* (4.7)*, let F_t be the flow defined by* (4.5)*, and let $r, \alpha \in [1, \infty]$ be such that $\frac{1}{\alpha} = \frac{1}{2} + \frac{1}{r}$. Then, for any $f \in W^{1,r}(\Omega)$ and $s \in [-T, T]$,*

$$\|f(F_s) - f\|_{L^\alpha(\Omega)} \leq \frac{C_1(T)^{1/\alpha}}{\phi_*} |s| \, \|\mathbf{V}\|_{L^2(\Omega)} \, \|\nabla f\|_{L^r(\Omega)},$$

where $C_1(T) = \frac{\phi^*}{\phi_*} \exp(\frac{\Gamma_{\mathrm{div}} T}{\phi_*})$ as in (4.9)*.*

Proof. By density it suffices to prove the estimate for $f \in C^1(\overline{\Omega})$ (in the case $r = \infty$, we first establish it for $r < \infty$ and corresponding α_r, using the density of smooth functions in $W^{1,r}$, and then let $r \to \infty$). For a.e. $\boldsymbol{x} \in \Omega$,

$$\begin{aligned} f(F_s(\boldsymbol{x})) - f(\boldsymbol{x}) = \int_0^s \frac{d}{dt} f(F_t(\boldsymbol{x})) dt &= \int_0^s \nabla f(F_t(\boldsymbol{x})) \cdot \frac{dF_t(\boldsymbol{x})}{dt} dt \\ &= \int_0^s \nabla f(F_t(\boldsymbol{x})) \cdot \frac{\mathbf{V}(F_t(\boldsymbol{x}))}{\phi(F_t(\boldsymbol{x}))} dt. \end{aligned}$$

Take the absolute value, the power α (using Jensen's inequality) and integrate over Ω. Using $\phi \geq \phi_*$ and applying a change of variables $\boldsymbol{y} = F_t(\boldsymbol{x})$ along with (4.9), this leads to

$$\begin{aligned} \int_\Omega |f(F_s(\boldsymbol{x})) - f(\boldsymbol{x})|^\alpha d\boldsymbol{x} &\leq \frac{|s|^{\alpha-1}}{\phi_*^\alpha} \int_\Omega \int_{[0,s]} |\nabla f(F_t(\boldsymbol{x}))|^\alpha |\mathbf{V}(F_t(\boldsymbol{x}))|^\alpha dt d\boldsymbol{x} \\ &\leq \frac{|s|^{\alpha-1}}{\phi_*^\alpha} \int_{[0,s]} \left(\int_\Omega |\nabla f(F_t(\boldsymbol{x}))|^\alpha |\mathbf{V}(F_t(\boldsymbol{x}))|^\alpha d\boldsymbol{x} \right) dt \\ &\leq \frac{C_1(T)|s|^\alpha}{\phi_*^\alpha} \int_\Omega |\nabla f(\boldsymbol{y})|^\alpha |\mathbf{V}(\boldsymbol{y})|^\alpha d\boldsymbol{y}. \end{aligned}$$

The proof is complete by applying Hölder's estimate with exponents r/α and $2/\alpha$, and by taking the power $1/\alpha$ of the resulting inequality. ∎

We now want to establish a similar result but for piecewise-constant functions. This will be useful to establish that discretisations based on piecewise-constant approximations, such as most FV methods, satisfy Assumption (**A2**). Before stating this lemma, we need a preliminary result.

Lemma 6.2.2 (Volume sweeped by a face transported by the flow)**.** *Under Assumption* (4.7)*, let F_t be the flow defined by* (4.5)*. Let σ be a face of the mesh over which $\mathbf{V}$ and ϕ are piecewise smooth. Let $V_t = |F_{[0,t]}(\sigma)|$ be the volume of the region sweeped by σ when transported over $[0, t]$ by the flow, that is, $V_t = |\{F_s(\boldsymbol{y}) \, : \, s \in [0, t] \, , \, \boldsymbol{y} \in \sigma\}|$. Then*

$$\forall t \in [-T, T] \, , \; V_t \leq \frac{C_1(T)}{\phi_*} |t| \int_\sigma |\mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma| ds(\boldsymbol{y}), \tag{6.4}$$

where $C_1(T)$ is given by (4.9) and $\mathbf{n}_\sigma$ is a normal to σ.

Proof. Notice first that since $\mathbf{V} \in H(\mathrm{div}, \Omega)$, the normal components of $\mathbf{V}$ across the faces of the mesh are continuous, and thus $|\mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma|$ is independent of the side of σ chosen to compute $\mathbf{V}$. Without loss of generality, we assume $t \geq 0$.

If the face σ is such that $Z_\sigma := \{\boldsymbol{y} \in \sigma \ : \ \mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma = 0\} = \sigma$, then even though $\sigma \subset \mathcal{C}$ (see Lemma 4.2.1 and its proof), we clearly have $V_t = 0$ since each point on the face is transported inside the face to one of its vertices/edges, which are $(d-2)$-dimensional objects then transported by the flow onto null sets in Ω (whatever side of σ chosen to compute $\mathbf{V}$ and ϕ). Hence, (6.4) holds for such faces.

Let us now assume that $Z_\sigma \neq \sigma$. Then, since $\mathbf{V} \cdot \mathbf{n}_\sigma$ is polynomial, Z_σ is a negligible set in σ for the $(d-1)$-dimensional measure and $F_t(\boldsymbol{y})$ is defined for all $\boldsymbol{y} \in \sigma \backslash Z_\sigma$. Since $F_{[0,t+h]}(\sigma) = F_{[0,t]}(\sigma) \sqcup F_{(t,t+h]}(\sigma)$, the flow property, a change of variables and (4.9) yield

$$\begin{aligned} V_{t+h} - V_t = |F_{(t,t+h]}(\sigma)| &= |F_t(F_{(0,h]}(\sigma))| \\ &= \int_{F_{(0,h]}(\sigma)} |JF_t(\boldsymbol{y})| d\boldsymbol{y} \leq C_1(T) |F_{(0,h]}(\sigma)|. \end{aligned} \tag{6.5}$$

Choose an orthonormal basis of $\mathbb{R}^d$ such that $\sigma \subset \{0\} \times \mathbb{R}^{d-1}$ and $\mathbf{n}_\sigma = (1, 0, \ldots, 0)$, and define $G : \mathbb{R} \times \sigma \to \mathbb{R}^d$ by $G(t, \boldsymbol{y}) = F_t(\boldsymbol{y})$. Using the area formula [44, Theorem 1] we have

$$\begin{aligned} |F_{(0,h]}(\sigma)| &= \int_{\mathbb{R}^d} \mathbf{1}_{G((0,h] \times \sigma))}(\boldsymbol{x}) d\boldsymbol{x} \leq \int_{\mathbb{R}^d} \mathrm{Card}\left[((0,h] \times \sigma) \cap G^{-1}(\{\boldsymbol{x}\})\right] d\boldsymbol{x} \\ &= \int_{(0,h] \times \sigma} |JG(t, \boldsymbol{y})| dt d\mathrm{s}(\boldsymbol{y}) = \int_0^h \left(\int_\sigma |JG(t, \boldsymbol{y})| d\mathrm{s}(\boldsymbol{y}) \right) dt \end{aligned} \tag{6.6}$$

where JG is the Jacobian determinant of G. Given the choice of basis in the range of G,

$$\begin{aligned} JG(t, \boldsymbol{y}) &= \det \begin{bmatrix} \frac{\partial G}{\partial t}(t, \boldsymbol{y}) & \frac{\partial G}{\partial y_1}(t, \boldsymbol{y}) & \cdots & \frac{\partial G}{\partial y_{d-1}}(t, \boldsymbol{y}) \end{bmatrix} \\ &= \det \begin{bmatrix} \frac{dF_t}{dt}(\boldsymbol{y}) & \frac{\partial F_t}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial F_t}{\partial y_{d-1}}(\boldsymbol{y}) \end{bmatrix} \\ &= \det \begin{bmatrix} \frac{\mathbf{V}(F_t(\boldsymbol{y}))}{\phi(F_t(\boldsymbol{y}))} & \frac{\partial F_t}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial F_t}{\partial y_{d-1}}(\boldsymbol{y}) \end{bmatrix}. \end{aligned}$$

For a fixed $\boldsymbol{y} \in \sigma \backslash Z_\sigma$ and for small t the flow $F_t(\boldsymbol{y})$ occurs in a region where $\mathbf{V}$ and ϕ (and thus F_t) are smooth – namely, the side of σ determined by

the sign of $\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma$. Hence, since $F_0=\mathrm{Id}$, denoting by $(\mathbf{V}_1,\dots,\mathbf{V}_d)$ the components of $\mathbf{V}$ in the chosen basis and recalling that $\mathbf{n}_\sigma=(1,0,\dots,0)$,

$$\begin{aligned}\lim_{t\to 0} JG(t,\boldsymbol{y}) &= \det\begin{bmatrix} \frac{\mathbf{V}(\boldsymbol{y})}{\phi(\boldsymbol{y})} & \frac{\partial F_0}{\partial y_1}(\boldsymbol{y}) & \cdots & \frac{\partial F_0}{\partial y_{d-1}}(\boldsymbol{y}) \end{bmatrix}\\ &= \det\begin{bmatrix} \frac{\mathbf{V}_1(\boldsymbol{y})}{\phi(\boldsymbol{y})} & 0 & \cdots & \cdots & 0\\ \vdots & 1 & 0 & \cdots & 0\\ \vdots & 0 & \ddots & \ddots & \vdots\\ \vdots & \vdots & \ddots & \ddots & 0\\ \frac{\mathbf{V}_d(\boldsymbol{y})}{\phi(\boldsymbol{y})} & 0 & \cdots & 0 & 1 \end{bmatrix} = \frac{\mathbf{V}_1(\boldsymbol{y})}{\phi(\boldsymbol{y})} = \frac{\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma}{\phi(\boldsymbol{y})}. \quad (6.7)\end{aligned}$$

Here, the value of ϕ is of course considered on the side of σ into which $F_t(\boldsymbol{y})$ flows for small $t>0$ (as already noticed, the value of $\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma$ does not depend on the considered side). Recalling that (6.7) holds for $\boldsymbol{y}\in\sigma\backslash Z_\sigma$ and that Z_σ has zero $(d-1)$-dimensional measure, the dominated convergence theorem thus shows that

$$\int_\sigma |JG(t,\boldsymbol{y})|\mathrm{d}s(\boldsymbol{y}) \to \int_\sigma \frac{|\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|}{\phi(\boldsymbol{y})}\mathrm{d}s(\boldsymbol{y}) \text{ as } t\to 0.$$

Dividing (6.6) by h, letting $h\to 0$, and plugging the result in (6.5) we infer that

$$\frac{dV_t}{dt} \le \frac{C_1(T)}{\phi_*}\int_\sigma |\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|\mathrm{d}s(\boldsymbol{y}).$$

The mapping $t\mapsto V_t$ is a non-decreasing function, so its derivative in the sense of distributions always exists as a positive measure; the relation above shows that this derivative is actually a bounded function, and thus that $t\mapsto V_t$ is Lipschitz-continuous. Integrating this relation and using $V_0=0$ leads to (6.4). ∎

We can now state a result that mimics Lemma 6.2.1 but for piecewise-constant functions. This result is used in Section 6.3.2.1 to prove that HMM schemes, among others, satisfy (**A2**).

Lemma 6.2.3 (Translation estimate for piecewise-constant functions)**.** *Let $\mathcal{M}$ be a polytopal mesh and $Y_\mathcal{M}$ be the set of piecewise-constant functions on $\mathcal{M}$. Define the discrete H^1-semi norm on $Y_\mathcal{M}$ by*

$$\forall f\in Y_\mathcal{M},\ |f|_\mathcal{M} = \left(\sum_{\sigma\in\mathcal{E}_{\rm int}} |\sigma| d_\sigma \left|\frac{f_K-f_L}{d_\sigma}\right|^2\right)^{1/2}, \quad (6.8)$$

where f_K is the constant value of f on $K\in\mathcal{M}$, $\mathcal{E}_{\rm int}$ is the set of internal faces (that is, $\sigma\in\mathcal{E}$ such that $\sigma\subset\Omega$), K and L are the two cells on each

side of σ, and $d_\sigma = d_{K,\sigma} + d_{L,\sigma}$ (see Figure 2.1). Assume that $(\phi, \mathbf{V})$ satisfy (4.7) on the sub-mesh made of $(D_{K,\sigma})_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K}$ and let k be the maximal polynomial degree of $\mathbf{V}$.

We then define the mesh regularity parameter

$$\varrho_{\mathcal{M}} = \max_{K\in\mathcal{M}} \mathrm{Card}(\mathcal{E}_K) + \max_{K\in\mathcal{M}} \max_{\sigma\in\mathcal{E}_K} \frac{\mathrm{diam}(D_{K,\sigma})}{\mathrm{inrad}(D_{K,\sigma})}, \tag{6.9}$$

where $\mathrm{inrad}(D_{K,\sigma})$ *is the radius of the largest ball included in $D_{K,\sigma}$. If $\varrho \geq \varrho_{\mathcal{M}}$, there exists R depending only on k, d and ϱ such that, for all $s \in [-T, T]$,*

$$\forall f \in Y_{\mathcal{M}}\,,\ \|f(F_s) - f\|_{L^1(\Omega)} \leq R\frac{C_1(T)}{\phi_*}|s|\,\|\mathbf{V}\|_{L^2(\Omega)}\,|f|_{\mathcal{M}}$$

where $C_1(T) = \frac{\phi^*}{\phi_*}\exp(\frac{\Gamma_{\mathrm{div}}T}{\phi_*})$ as in (4.9).

Proof. We start by writing $f(F_s(\boldsymbol{x})) - f(\boldsymbol{x})$ as the sum of the jumps of f along the curve $(F_t(\boldsymbol{x}))_{t\in[0,s]} =: F_{[0,s]}(\boldsymbol{x})$. For $\sigma \in \mathcal{E}_{\mathrm{int}}$, letting $\chi_\sigma(\boldsymbol{x}) = 1$ if $\sigma \cap F_{[0,s]}(\boldsymbol{x}) \neq \emptyset$ and $\chi_\sigma(\boldsymbol{x}) = 0$ otherwise, this leads to

$$|f(F_s(\boldsymbol{x})) - f(\boldsymbol{x})| \leq \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}} \chi_\sigma(\boldsymbol{x})|f_K - f_L|. \tag{6.10}$$

Notice that $\sigma \cap F_{[0,s]}(\boldsymbol{x}) \neq \emptyset$ if and only if $F_{[-s,0]}(\sigma) \cap \{\boldsymbol{x}\} \neq \emptyset$, that is, $\boldsymbol{x}$ belongs to the region sweeped by σ transported by the flow over $[-s, 0]$. Lemma 6.2.2 gives

$$\int_\Omega \chi_\sigma(\boldsymbol{x})d\boldsymbol{x} \leq \frac{C_1(T)}{\phi_*}|s|\int_\sigma |\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|d\mathrm{s}(\boldsymbol{y})$$

where $\mathbf{n}_\sigma$ is a unit normal to σ. Hence, letting $C = \frac{C_1(T)}{\phi_*}$ and using the Cauchy–Schwarz inequality (on the combined sum and integral terms),

$$\begin{aligned}
&\int_\Omega |f(F_s(\boldsymbol{x})) - f(\boldsymbol{x})|d\boldsymbol{x}\\
&\leq C|s|\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\int_\sigma |\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|\,|f_K - f_L|d\mathrm{s}(\boldsymbol{y})\\
&= C|s|\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\int_\sigma \sqrt{d_\sigma}|\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|\,\frac{1}{\sqrt{d_\sigma}}|f_K - f_L|d\mathrm{s}(\boldsymbol{y})\\
&\leq C|s|\left(\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\int_\sigma d_\sigma|\mathbf{V}(\boldsymbol{y})\cdot\mathbf{n}_\sigma|^2 d\mathrm{s}(\boldsymbol{y})\right)^{1/2}\left(\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\int_\sigma \frac{1}{d_\sigma}|f_K - f_L|^2 d\mathrm{s}(\boldsymbol{y})\right)^{1/2}
\end{aligned}$$

$$= C|s| \left(\sum_{\sigma \in \mathcal{E}_{\rm int}} d_\sigma \int_\sigma |\mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma|^2 d\mathrm{s}(\boldsymbol{y}) \right)^{1/2} |f|_{\mathcal{M}} \,. \tag{6.11}$$

Since $\mathbf{V}$ is polynomial on each $D_{K,\sigma}$, we can use the discrete trace inequality of [32, Lemma 1.46] to find R depending only on k, d and ϱ such that

$$\forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{E}_K\,,\ \mathrm{diam}(D_{K,\sigma}) \int_\sigma |\mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma|^2 d\mathrm{s}(\boldsymbol{y}) \le R^2 \int_{D_{K,\sigma}} |\mathbf{V}(\boldsymbol{x})|^2 d\boldsymbol{x}.$$

Noticing that $d_{K,\sigma} \le \mathrm{diam}(D_{K,\sigma})$, we infer

$$d_\sigma \int_\sigma |\mathbf{V}(\boldsymbol{y}) \cdot \mathbf{n}_\sigma|^2 d\mathrm{s}(\boldsymbol{y}) \le R^2 \int_{D_{K,\sigma} \cup D_{L,\sigma}} |\mathbf{V}(\boldsymbol{x})|^2 d\boldsymbol{x}.$$

The proof of the lemma is completed by plugging this estimate into (6.11). ∎

Remark 6.2.4 (Estimate in L^α norm?)**.** *A natural question would be the extension of Lemma 6.2.3 to estimate the L^α norm of $f(F_s) - f$, as in Lemma 6.2.1, by using the discrete $W^{1,r}$-semi norm $|f|_{\mathcal{M},r}$ of f obtained by replacing 2 with r in* (6.8)*. Considering for example the simple case of a constant unit velocity $\mathbf{V} = \mathbf{V}_0$ (and forgetting about boundary conditions for simplification), this would amount to estimating $\|f(\cdot + s\mathbf{V}_0) - f\|_{L^\alpha(\Omega)}$ in terms of $|s|\,|f|_{\mathcal{M},r}$. For meshes admissible for the TPFA finite volume scheme, such an estimate is known with $\alpha = r = 2$ and $|s|$ replaced by $\sqrt{|s|(|s| + \max_{K\in\mathcal{M}} \mathrm{diam}(K))}$ [50, Lemma 3.3]. For general meshes, however, no similar estimate seems to be attainable if $\alpha > 1$.*

The next lemma is instrumental in passing to the limit in the reaction and advection terms of the GEM scheme. Let us first introduce some notations. Given time steps $0 = t^{(0)} < t^{(1)} < \ldots < t^{(N)} = T$ and velocities $\mathbf{V} = (\mathbf{V}^n)_{n=1,\ldots,N}$ that satisfy (4.7), we identify $\mathbf{V}$ with the global velocity $\Omega \times (0,T) \to \mathbb{R}^d$ given by $\mathbf{V}(\cdot,t) = \mathbf{V}^{(n+1)}$ for all $t \in (t^{(n)}, t^{(n+1)}]$ and all $n = 0,\ldots,N-1$. Define $\mathcal{T}_{\mathbf{V}}$ and $\widehat{\mathcal{T}}_{\mathbf{V}}$ as the linear "transport" mappings $L^2(\Omega \times (0,T)) \to L^2(\Omega \times (0,T))$ such that, for $\psi \in L^2(\Omega \times (0,T))$,

$$\begin{aligned} &\text{for a.e. } \boldsymbol{x} \in \Omega, \text{ for all } t \in (t^{(n)}, t^{(n+1)}) \text{ and } n = 0,\ldots,N-1, \\ &\mathcal{T}_{\mathbf{V}}\psi(\boldsymbol{x},t) = \psi\left(F^{(n+1)}_{t^{(n+1)}-t^{(n)}}(\boldsymbol{x}), t\right) \quad \text{and} \quad \widehat{\mathcal{T}}_{\mathbf{V}}\psi(\boldsymbol{x},t) = \psi\left(F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{x}), t\right) \end{aligned} \tag{6.12}$$

where $F^{(n+1)}_t$ is defined by (4.5) for the velocity $\mathbf{V}^{(n+1)}$. The difference between $\mathcal{T}_{\mathbf{V}}$ and $\widehat{\mathcal{T}}_{\mathbf{V}}$ is the time at which this flow is considered.

Lemma 6.2.5 (Convergence of functions transported by the flow)**.** *Let ϕ satisfy* (6.1b) *and, for each $m \in \mathbb{N}$, take $0 = t_m^{(0)} < t_m^{(1)} < \ldots < t_m^{(N_m)} = T$ time steps and $\mathbf{V}_m = (\mathbf{V}_m^n)_{n=1,\ldots,N_m}$ that satisfy* (4.7) *with Γ_{div} not depending on m. Assume that $\delta t_m := \max_{n=0,\ldots,N_m-1}(t_m^{(n+1)} - t_m^{(n)}) \to 0$ as $m \to \infty$ and that $(\mathbf{V}_m)_{m\in\mathbb{N}}$ is bounded in $L^2(\Omega \times (0,T))$. Then $\mathcal{T}_{\mathbf{V}_m}$ and $\widehat{\mathcal{T}}_{\mathbf{V}_m}$ satisfy the following properties.*

1. *There is C not depending on m such that, for $\psi \in L^2(\Omega \times (0,T))$,*

$$\|\mathcal{T}_{\mathbf{V}_m}\psi\|_{L^2(\Omega\times(0,T))} + \left\|\widehat{\mathcal{T}}_{\mathbf{V}_m}\psi\right\|_{L^2(\Omega\times(0,T))} \leq C\,\|\psi\|_{L^2(\Omega\times(0,T))}\,. \tag{6.13}$$

2. *The dual operators $\mathcal{T}^*_{\mathbf{V}_m}$ and $\widehat{\mathcal{T}}^*_{\mathbf{V}_m}$ of $\mathcal{T}_{\mathbf{V}_m}$ and $\mathcal{T}_{\mathbf{V}_m}$ are given by: for $\psi \in L^2(\Omega \times (0,T))$,*

$$\begin{aligned}
\mathcal{T}^*_{\mathbf{V}_m}\psi &= \phi\mathcal{T}_{-\mathbf{V}_m}\left(\frac{\psi}{\phi}\right) + R_m\mathcal{T}_{-\mathbf{V}_m}\psi \\
\widehat{\mathcal{T}}^*_{\mathbf{V}_m}\psi &= \phi\widehat{\mathcal{T}}_{-\mathbf{V}_m}\left(\frac{\psi}{\phi}\right) + \widehat{R}_m\widehat{\mathcal{T}}_{-\mathbf{V}_m}\psi
\end{aligned} \tag{6.14}$$

   where $R_m, \widehat{R}_m \in L^\infty(\Omega \times (0,T))$ and, over each interval $[t^{(n)}, t^{(n+1)}]$, $R_m, \widehat{R}_m$ are bounded by $\delta t^{(n+\frac{1}{2})}\phi_*^{-1}\Gamma_{\mathrm{div}}C_1(T)$.

3. *If $f_m \to f$ strongly (resp. weakly) in $L^2(\Omega \times (0,T))$ as $m \to \infty$, then $\mathcal{T}_{\mathbf{V}_m}f_m \to f$ and $\widehat{\mathcal{T}}_{\mathbf{V}_m}f_m \to f$ strongly (resp. weakly) in $L^2(\Omega\times(0,T))$.*

Proof.

We only prove the results for $\mathcal{T}_{\mathbf{V}_m}$, as the proof for $\widehat{\mathcal{T}}_{\mathbf{V}_m}$ follows by simply replacing $F^{(n+1)}_{t^{(n+1)}-t^{(n)}}(\boldsymbol{y})$ by $F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{y})$. In the first two steps, we drop the index m in $\mathbf{V}_m$ and N_m for simplicity of notation.

Step 1: bound on the norms of $\mathcal{T}_{\mathbf{V}}$ and $\widehat{\mathcal{T}}_{\mathbf{V}}$.

By a change of variables and invoking (4.9), there is C not depending on m, $s \in [-T,T]$ or $n \in \{0,\ldots,N-1\}$ such that, for all $h \in L^2(\Omega)$, $\|h(F_s^{(n+1)}(\cdot))\|_{L^2(\Omega)} \leq C\|h\|_{L^2(\Omega)}$. Estimate (6.13) easily follows from this.

Step 2: description of the dual operator.

A change of variables yields, for any $\varphi, \psi \in L^2(\Omega \times (0,T))$,

$$\begin{aligned}
&\int_{\Omega\times(0,T)} (\mathcal{T}_{\mathbf{V}}\varphi)(\boldsymbol{x},t)\psi(\boldsymbol{x},t)d\boldsymbol{x}dt \\
&\qquad = \sum_{n=0}^{N-1}\int_{t^{(n)}}^{t^{(n+1)}}\int_\Omega \varphi(F^{(n+1)}_{t^{(n+1)}-t^{(n)}}(\boldsymbol{x}),t)\psi(\boldsymbol{x},t)d\boldsymbol{x}dt
\end{aligned}$$

$$= \sum_{n=0}^{N-1} \int_{t^{(n)}}^{t^{(n+1)}} \int_\Omega \varphi(\boldsymbol{y},t)\psi(F^{(n+1)}_{t^{(n)}-t^{(n+1)}}(\boldsymbol{y}),t)|JF^{(n+1)}_{t^{(n)}-t^{(n+1)}}(\boldsymbol{y})|d\boldsymbol{y}dt. \tag{6.15}$$

Relation (4.8) and Estimate (4.9) shows that

$$|JF^{(n+1)}_{t^{(n)}-t^{(n+1)}}(\boldsymbol{y})| = \frac{\phi(\boldsymbol{y})}{\phi(F^{(n+1)}_{t^{(n)}-t^{(n+1)}}(\boldsymbol{y}))} + R(\boldsymbol{y},t^{(n)}) \tag{6.16}$$

with $|R(\boldsymbol{y},t^{(n)})| \leq \delta t^{(n+\frac{1}{2})}\phi_*^{-1}\Gamma_{\text{div}}C_1(T)$. Since $t \mapsto F^{(n+1)}_{t-t^{(n+1)}}(\boldsymbol{y})$ is the flow corresponding to $-\mathbf{V}$, Relations (6.15) and (6.16) then yield (6.14) for $\mathcal{T}^*_{\mathbf{V}}$.

Step 3: proof of the strong convergence.

For simplicity of notation, denote $\|\cdot\|_2 = \|\cdot\|_{L^2(\Omega\times(0,T))}$. Assume that $f_m \to f$ strongly in $L^2(\Omega\times(0,T))$, and let f^ε be a smooth approximation of f such that $\|f-f^\varepsilon\|_2 \leq \varepsilon$. The triangle inequality and (6.13) yield

$$\begin{aligned}\|\mathcal{T}_{\mathbf{V}_m}f_m - f\|_2 \leq\ & \|\mathcal{T}_{\mathbf{V}_m}(f_m-f)\|_2 + \|\mathcal{T}_{\mathbf{V}_m}(f-f^\varepsilon)\|_2 + \|\mathcal{T}_{\mathbf{V}_m}f^\varepsilon - f^\varepsilon\|_2 \\ &+ \|f^\varepsilon - f\|_2 \\ \leq\ & C\|f_m-f\|_2 + (C+1)\varepsilon + \|\mathcal{T}_{\mathbf{V}_m}f^\varepsilon - f^\varepsilon\|_2 .\end{aligned}$$

Invoking Lemma 6.2.1 with $\alpha = 2$, $r=\infty$ and $f^\varepsilon(\cdot,t)$ instead of f gives C' not depending on m or ε such that, if $F^{(n+1)}_{m,t}$ is the flow for the velocity $\mathbf{V}^{(n+1)}_m$,

$$\begin{aligned}\|\mathcal{T}_{\mathbf{V}_m}f^\varepsilon - f^\varepsilon\|_2^2 &= \sum_{n=0}^{N_m-1}\int_{t^{(n)}}^{t^{(n+1)}} \left\|f^\varepsilon(F^{(n+1)}_{m,t^{(n+1)}-t^{(n)}}(\cdot),t) - f^\varepsilon(\cdot,t)\right\|^2_{L^2(\Omega)} dt \\ &\leq C'\delta t_m^2 \sum_{n=0}^{N_m-1}\int_{t^{(n)}}^{t^{(n+1)}} \left\|\mathbf{V}^{(n+1)}_m\right\|^2_{L^2(\Omega)} \|\nabla f^\varepsilon(\cdot,t)\|^2_{L^\infty(\Omega)}\, dt \\ &= C'\delta t_m^2 \|\mathbf{V}_m\|_2^2 \|\nabla f^\varepsilon\|^2_{L^\infty(\Omega\times(0,T))} .\end{aligned}$$

Hence,

$$\|\mathcal{T}_{\mathbf{V}_m}f_m - f\|_2 \leq C\|f_m-f\|_2 + (1+C)\varepsilon + \sqrt{C'}\delta t_m \|\mathbf{V}_m\|_2 \|\nabla f^\varepsilon\|_{L^\infty(\Omega\times(0,T))} .$$

Taking the superior limit as $m\to\infty$ and using the boundedness of $(\mathbf{V}_m)_{m\in\mathbb{N}}$ in $L^2(\Omega\times(0,T))$ thus yields $\limsup_{m\to\infty}\|\mathcal{T}_{\mathbf{V}_m}f_m - f\|_2 \leq (1+C)\varepsilon$. Letting $\varepsilon\to 0$ concludes the proof that $\mathcal{T}_{\mathbf{V}_m}f_m \to f$ strongly in $L^2(\Omega\times(0,T))$.

Step 4: proof of the weak convergence.

Assume that $f_m \to f$ weakly in $L^2(\Omega \times (0,T))$. Then, for all $\psi \in L^2(\Omega \times (0,T))$,

$$\begin{aligned}\int_{\Omega\times(0,T)} (\mathcal{T}_{\mathbf{V}_m} f_m - f)\psi &= \int_{\Omega\times(0,T)} \mathcal{T}_{\mathbf{V}_m}(f_m - f)\psi + \int_{\Omega\times(0,T)} (\mathcal{T}_{\mathbf{V}_m} f - f)\psi \\ &= \int_{\Omega\times(0,T)} (f_m - f)\mathcal{T}^*_{\mathbf{V}_m}\psi + \int_{\Omega\times(0,T)} (\mathcal{T}_{\mathbf{V}_m} f - f)\psi. \end{aligned} \tag{6.17}$$

Since $\psi/\phi \in L^2(\Omega \times (0,T))$, the formula (6.14), the fact that $R_m \to 0$ in $L^\infty(\Omega \times (0,T))$, the estimate (6.13) and the result of Step 3 applied to $-\mathbf{V}_m$ instead of $\mathbf{V}_m$ show that $\mathcal{T}^*_{\mathbf{V}_m}\psi \to \psi$ strongly in $L^2(\Omega \times (0,T))$ as $m \to \infty$. Hence, the first term in the right-hand side of (6.17) tends to 0 since $f_m - f \to 0$ weakly $L^2(\Omega \times (0,T))$. The second term in the right-hand side of (6.17) also converges to 0 since, by Step 3 (applied to $f_m = f$ for all m), $\mathcal{T}_{\mathbf{V}_m} f - f \to 0$ in $L^2(\Omega \times (0,T))$. The proof that $\mathcal{T}_{\mathbf{V}_m} f_m \to f$ weakly in $L^2(\Omega \times (0,T))$ is therefore complete. ∎

6.3 Sample methods covered by the analysis

The ELLAM and MMOC are ways to deal with the advection term in the concentration equation. Various numerical methods can be chosen to discretise the diffusion terms in this equation, as well as in the pressure equation. These methods correspond to selecting specific gradient discretisations $\mathcal{C}$ and $\mathcal{P}$. Here, we detail some of the GDs corresponding to methods used in the literature in conjunction with the ELLAM or MMOC, and we show that they all satisfy the assumptions of Theorem 6.1.1. As a consequence, our convergence result applies to all these methods.

In the following, for simplicity of notations, we drop the index m in the gradient discretisations and we consider Assumptions (**A1**)–(**A5**) 'as the mesh size and time step go to zero' (as opposed to 'as $m \to \infty$').

6.3.1 Conforming/mixed finite-element methods

When discretising the model (1.1) using finite element methods for the diffusion terms and a characteristic method for the advection term, it is natural to use a mixed method for the pressure equation and a conforming method for the concentration equation. The mixed method provides an appropriate Darcy velocity that can be used to build the characteristics. This approach was considered in [76, 78] for ELLAM. We show here that such

a mixed/conforming FE–ELLAM scheme fits into our GEM framework, so that the convergence result of Theorem 6.1.1 applies to the schemes in the aforementioned references. Notice that, contrary to the convergence analysis done for example in [76], our convergence result relies on very weak regularity assumptions on the data and solution, that are usually satisfied in practical applications.

6.3.1.1 Description of the conforming and mixed FE GDs

Any conforming Galerkin approximation fits into the GDM framework. This applies to conforming finite element methods, such as $\mathbb{P}_k$ FE on simplices or $\mathbb{Q}_k$ FE on Cartesian grids. A finite-dimensional subspace V_h of $H^1(\Omega)$ being chosen, define $(X_{\mathcal{C}}, \Pi_{\mathcal{C}}, \nabla_{\mathcal{C}})$ by $X_{\mathcal{C}} = V_h$ and, for $v \in V_h$, $\Pi_{\mathcal{C}} v = v$ and $\nabla_{\mathcal{C}} v = \nabla v$. The interpolant $\mathcal{I}_{\mathcal{C}}$ can be either chosen as the orthogonal projection on V_h, in the case of an abstract space, or as the standard nodal interpolant for specific FE spaces.

We now describe a gradient discretisation $\mathcal{P}$ corresponding to the $\mathbb{RT}_0$ mixed finite element method. The following construction can be extended to higher order $\mathbb{RT}_k$ finite elements [52] or [40, Chapter 10]. A conforming simplicial or Cartesian mesh $\mathcal{M}$ being chosen, define

$$\boldsymbol{V}_{h,0} = \{\boldsymbol{v} \in H(\mathrm{div}, \Omega) \,:\, \boldsymbol{v}_{|K} \in \mathbb{RT}_0(K)\,,\ \forall K \in \mathcal{M}\,,\ \boldsymbol{v}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}, \tag{6.18a}$$

$$W_h = \{z \in L^2(\Omega) \,:\, z_{|K} \text{ constant}\,,\ \forall K \in \mathcal{M}\}, \tag{6.18b}$$

where $\mathbb{RT}_0$ is the lowest order Raviart–Thomas space on the cell K (the description of $\mathbb{RT}_0$ depends if this cell is a simplex or Cartesian cell). After choosing a diffusion tensor $\mathcal{A}$ – that is, a symmetric, uniformly bounded and coercive matrix-valued function $\Omega \to M_d(\mathbb{R})$ – a gradient discretisation $\mathcal{P} = (X_{\mathcal{P}}, \Pi_{\mathcal{P}}, \nabla_{\mathcal{P}})$ is constructed by setting $X_{\mathcal{P}} = W_h$ and, for $z \in W_h$, $\Pi_{\mathcal{P}} z = z$. The reconstructed gradient $\nabla_{\mathcal{P}} z$ is defined as the solution to

$$\begin{aligned} &\mathcal{A}\nabla_{\mathcal{P}} z \in \boldsymbol{V}_{h,0} \text{ and, for all } \boldsymbol{w} \in \boldsymbol{V}_{h,0},\\ &\int_\Omega \boldsymbol{w}(\boldsymbol{x})\cdot\nabla_{\mathcal{P}} z(\boldsymbol{x})d\boldsymbol{x} = -\int_\Omega z(\boldsymbol{x})\mathrm{div}\boldsymbol{w}(\boldsymbol{x})d\boldsymbol{x}. \end{aligned} \tag{6.19}$$

The existence and uniqueness of $\nabla_{\mathcal{P}} z$ follows by applying the Riesz representation theorem in $\boldsymbol{V}_{h,0}$ with the inner product $(\boldsymbol{w}, \boldsymbol{v}) \mapsto \int_\Omega \boldsymbol{w}\cdot\mathcal{A}^{-1}\boldsymbol{v}d\boldsymbol{x}$.

Taking $\mathcal{A}(\boldsymbol{x}) = \frac{\mathbf{K}(\boldsymbol{x})}{\mu(\Pi_{\mathcal{C}} c^{(n)}(\boldsymbol{x}))}$, the scheme (5.1) is exactly an $\mathbb{RT}_0$ mixed finite element discretisation of the pressure equation at the n-th time step. We notice here that $\mathcal{A}$, and thus the gradient discretisation $\mathcal{P}$ built above, changes with each time step; we are therefore in the context of Remark 6.1.3.

6.3.1.2 Assumptions (A1)–(A5)

We show here that all required assumptions for Theorem 6.1.1 are satisfied by sequences of GDs as in Section 6.3.1.1.

Under usual mesh regularity properties, Assumption (**A1**) for $(\mathcal{C}_m^T)_{m\in\mathbb{N}}$ follows from [40, Chapter 8] (note that $W_\mathcal{C} \equiv 0$ and $C_\mathcal{C} \leq C_P$, where C_P is the Poincaré–Wirtinger constant in $H^1(\Omega)$). For the GD $\mathcal{P}$ built on the $\mathbb{RT}_0$ mixed FE, although the matrix $\mathcal{A}$ changes with each time step, it always remains uniformly bounded and coercive; the analysis in [52] or [40, Chapter 10] thus shows that the notions of coercivity, GD-consistency and limit-conformity as in Remark 6.1.3 are verified.

Thanks to (6.1a), the standard Darcy velocity $\mathbf{u}_\mathcal{P}^{(n+1)} = -\frac{\mathbf{K}}{\mu(\Pi_\mathcal{C} c^{(n)})}\nabla_\mathcal{P} p^{(n+1)}$ resulting from the $\mathbb{RT}_0$ discretisation of the pressure equation already satisfies Assumption (**A4**), and is therefore naturally used as the tracking velocity. Assumption (**A5**)a) is trivially satisfied since $|\mathbf{u}_\mathcal{P}| \leq \Lambda_A |\nabla_\mathcal{P} p|$. Moreover, under (**A1**), if $\Pi_\mathcal{C} c \to c$ in $L^2(\Omega\times(0,T))$ as the mesh size and time step go to 0, then $\widetilde{\Pi}_\mathcal{C} c$ also converges to c in the same space (see, e.g., end of Section 6.5.1); thus, if $\nabla_\mathcal{P} p \to \nabla p$ in $L^2(\Omega\times(0,T))$, the assumption (6.1c) on $\mathbf{K}/\mu$ ensures that $\mathbf{u}_\mathcal{P} = -\frac{\mathbf{K}}{\mu(\widetilde{\Pi}_\mathcal{C} c)}\nabla_\mathcal{P} p$ strongly converges in $L^2(\Omega\times(0,T))$ to $\mathbf{u} = -\frac{\mathbf{K}}{\mu(c)}\nabla p$, which proves (**A5**)b).

For $\mathcal{C}$ coming from a conforming finite element method, the standard nodal interpolation $\mathcal{J}_\mathcal{C}$ clearly satisfies (**A3**) (see [14, Theorem 4.4.20]). Finally, Assumption (**A2**) follows from Lemma 6.2.1 applied to $f = \Pi_\mathcal{C} z \in H^1(\Omega)$, $\alpha = 1$ and $r = 2$.

6.3.2 Finite-volume based

A number of finite volume numerical schemes can be embedded in the gradient discretisation method [40]. In particular, one of them is the Hybrid Mimetic Mixed method (HMM), which was presented in Section 2.2. The HMM method was used in [22, 23] to discretise the diffusion terms in both the pressure and concentration equations, together with the ELLAM for the advection term. The analysis carried out here also applies to many other numerical schemes based on piecewise-constant reconstructions, such as the VAG scheme, the MPFA-O FV method, mass-lumped FE methods or nodal Mimetic Finite Differences [40].

6.3.2.1 Assumptions (A1)–(A3)

Under a boundedness assumption on $\varrho_\mathcal{M}$ (defined by (6.9)), the basic properties (**A1**) (with both $\mathcal{C}$ and $\mathcal{P}$ given by an HMM GD as in Section 2.2)

follow from the results in [40, Chapter 13]. The appendix of [1] describes an interpolant $\mathcal{J}_\mathcal{D}$ and shows that it satisfies Assumption (**A3**).

Denoting by $Y_\mathcal{M}$ the space of piecewise constant functions on $\mathcal{M}$, we have $\Pi_\mathcal{D}(X_\mathcal{D}) \subset Y_\mathcal{M}$. Recalling the definition (6.8) of the discrete H^1-semi norm on $Y_\mathcal{M}$, [40, Lemma 13.11 and Remark 7.5] show that $|\Pi_\mathcal{D}\cdot|_\mathcal{M} \le \beta_\mathcal{D} \|\nabla_\mathcal{D}\cdot\|_{L^2(\Omega)}$ with $\beta_\mathcal{D}$ depending only on an upper bound of $\varrho_\mathcal{M}$ (this estimate is not specific to the HMM; it holds for all currently known GDs such that $\Pi_\mathcal{D}(X_\mathcal{D}) \subset Y_\mathcal{M}$). Assumption (**A2**) is therefore a consequence of Lemma 6.2.3, provided that the reconstructed Darcy velocity is piecewise polynomial (which is usually the case – see next section).

6.3.2.2 Reconstructed Darcy velocity and Assumptions (A4)–(A5)

For methods like the HMM that produce piecewise-constant gradients $\nabla_\mathcal{P} p^{(n+1)}$ and/or piecewise-constant concentration $\Pi_\mathcal{C} c^{(n)}$, the natural Darcy velocity $-\frac{\mathbf{K}}{\mu(\Pi_\mathcal{C} c^{(n)})}\nabla_\mathcal{P} p^{(n+1)}$ does not belong to $H(\mathrm{div},\Omega)$. It is therefore not suitable to define the characteristics used in the ELLAM, and another velocity must be reconstructed to be used in (5.2). Finite-volume methods naturally produce numerical fluxes on the mesh faces, that satisfy the balance and conservativity relations (2.15)–(2.16). Such fluxes can be used to reconstruct a Darcy velocity in a Raviart–Thomas space on a sub-mesh of $\mathcal{M}$.

In [22, 23], this idea is applied to the HMM method on the sub-mesh of pyramids $(D_{K,\sigma})_{K\in\mathcal{M},\sigma\in\mathcal{E}_K}$ (see Chapter 3). A velocity $\mathbf{u}_\mathcal{P}^{(n+1)} \in H(\mathrm{div},\Omega)$ is constructed from the pressure unknowns such that its restriction to each diamond $D_{K,\sigma}$ belongs to $\mathbb{RT}_0$ and that, for each cell $K \in \mathcal{M}$,

$$\begin{aligned} &\text{For a.e. } \boldsymbol{x} \in K,\ \mathrm{div}\mathbf{u}_\mathcal{P}^{(n+1)}(\boldsymbol{x}) = \frac{1}{|K|}\sum_{\sigma\in\mathcal{E}_K} \mathcal{F}_{K,\sigma}(p^{(n+1)})\,, \\ &\forall \sigma \in \mathcal{E}_K\,,\ \forall \boldsymbol{y} \in \sigma\,,\ |\sigma| \mathbf{u}_\mathcal{P}^{(n+1)}(\boldsymbol{y})\cdot\mathbf{n}_{K,\sigma} = \mathcal{F}_{K,\sigma}(p^{(n+1)}). \end{aligned} \tag{6.20}$$

Using the flux balance equation (2.16a) with $f = q^+ - q^-$, this reconstruction of $\mathbf{u}_\mathcal{P}^{(n+1)}$ satisfies Assumption (**A4**) with $M_\mathrm{div} = M_{q^+} + M_{q^-}$ (see (6.1a)).

Let us now establish the estimate on $\mathbf{u}_\mathcal{P}$ stated in (**A5**). In the following, $A \lesssim B$ means that $A \le CB$ with C depending only on an upper bound of $\varrho_\mathcal{M}$, and of α_A and Λ_A in (6.1c). Fix $K \in \mathcal{M}$. The relations (6.20) boil down to a linear system for internal fluxes in K – that is, fluxes $\mathcal{F}_\tau$ on $(\partial D_{K,\sigma} \setminus \sigma)_{\sigma\in\mathcal{E}_K}$ – in which the right-hand side is $(\mathcal{F}_{K,\sigma}(p^{(n+1)}))_{\sigma\in\mathcal{E}_K}$. Introducing auxiliary unknowns (as in the A method in Section 3.1.4), augmenting this system with a consistency relation (as in the C method in Section 3.1.3) or fixing the solution to be of minimal ℓ^2 norm (as in the KR method in Section 3.1.2), leads to a linear system $M_K(\mathcal{F}_\tau)_\tau = (\mathcal{F}_{K,\sigma}(p^{(n+1)}))_{\sigma\in\mathcal{E}_K}$ with M_K depending

only on the number of faces of K, not on the geometry of this cell. Hence, $\sum_\tau |\mathcal{F}_\tau|^2 \lesssim \sum_{\sigma\in\mathcal{E}_K} |\mathcal{F}_{K,\sigma}(p^{(n+1)})|^2$. Due to the shape regularity assumption (which implies $|\tau|^{-1} \lesssim \mathrm{diam}(K)/|K|$ for any face τ of any pyramid $D_{K,\sigma}$) and by construction of $\mathbb{RT}_0$ functions, we infer that

$$\begin{aligned}\left\|\mathbf{u}_{\mathcal{P}}^{(n+1)}\right\|_{L^2(D_{K,\sigma})}^2 &\lesssim \sum_{\tau\subset\partial D_{K,\sigma}} \frac{\mathrm{diam}(K)}{|\tau|}|\mathcal{F}_\tau|^2 \\ &\lesssim \frac{\mathrm{diam}(K)^2}{|K|}\sum_{\sigma\in\mathcal{E}_K}|\mathcal{F}_{K,\sigma}(p^{(n+1)})|^2. \end{aligned} \tag{6.21}$$

Fix $\sigma\in\mathcal{E}_K$ and take, in (2.15), $v_\sigma = 1$ and $v_K = v_{\sigma'} = 0$ if $\sigma\neq\sigma'$. The definition (2.13) of $\nabla_{\mathcal{D}}$ easily shows that $|\nabla_{\mathcal{D}} v| \lesssim \mathrm{diam}(K)^{-1}$ and (2.15) therefore yields $\mathrm{diam}(K)\sum_{\sigma\in\mathcal{E}_K}|\mathcal{F}_{K,\sigma}(p^{(n+1)})| \lesssim \int_K |\nabla_{\mathcal{P}} p^{(n+1)}(\boldsymbol{x})|d\boldsymbol{x}$. Hence, by the Cauchy–Schwarz inequality,

$$\frac{\mathrm{diam}(K)^2}{|K|}\sum_{\sigma\in\mathcal{E}_K}|\mathcal{F}_{K,\sigma}(p^{(n+1)})|^2 \lesssim \left\|\nabla_{\mathcal{P}} p^{(n+1)}\right\|_{L^2(K)}^2.$$

Combined with (6.21) this proves (**A5**)a).

Because of this bound, the weak convergence in (**A5**)b) follows if we can show that $\mathbf{u}_{\mathcal{P}}$ converges to $\mathbf{u}$ against any $\boldsymbol{\varphi}\in C_c^\infty(\Omega\times(0,T))^d$. To establish this convergence, we first evaluate $\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}$, where $\mathbf{U}_{\mathcal{P}} = -\frac{\mathbf{K}}{\mu(\widetilde{\Pi}_c c)}\nabla_{\mathcal{P}} p$. Fix $\boldsymbol{\xi}\in\mathbb{R}^d$ and apply the divergence theorem between $\mathbf{u}_{\mathcal{P}}^{(n+1)}\in H(\mathrm{div},K)$ and the affine map $x\mapsto\boldsymbol{\xi}\cdot(\boldsymbol{x}-\boldsymbol{x}_K)$ to write

$$\begin{aligned}\int_K \mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})\cdot\boldsymbol{\xi}d\boldsymbol{x} &= \int_K \mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})\cdot\nabla(\boldsymbol{\xi}\cdot(\boldsymbol{x}-\boldsymbol{x}_K))d\boldsymbol{x} \\ &= \sum_{\sigma\in\mathcal{E}_K}\int_\sigma \mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{y})\cdot\mathbf{n}_{K,\sigma}[\boldsymbol{\xi}\cdot(\boldsymbol{y}-\boldsymbol{x}_K)]ds(\boldsymbol{y}) \\ &\quad - \int_K \mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})[\boldsymbol{\xi}\cdot(\boldsymbol{x}-\boldsymbol{x}_K)]d\boldsymbol{x}.\end{aligned}$$

Using (6.20) and $\frac{1}{|\sigma|}\int_\sigma \boldsymbol{y}ds(\boldsymbol{y}) = \overline{\boldsymbol{x}}_\sigma$, where we recall that $\overline{\boldsymbol{x}}_\sigma$ is the centre of mass of σ, then leads to

$$\begin{aligned}\int_K \mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})\cdot\boldsymbol{\xi}d\boldsymbol{x} &= \sum_{\sigma\in\mathcal{E}_K}\mathcal{F}_{K,\sigma}(p^{(n+1)})\boldsymbol{\xi}\cdot(\overline{\boldsymbol{x}}_\sigma-\boldsymbol{x}_K) \\ &\quad - \int_K \mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})[\boldsymbol{\xi}\cdot(\boldsymbol{x}-\boldsymbol{x}_K)]d\boldsymbol{x}.\end{aligned} \tag{6.22}$$

Apply (2.16a) with $v \in X_{\mathcal{D}}$ the interpolant of the linear mapping $\boldsymbol{x} \mapsto \boldsymbol{\xi} \cdot \boldsymbol{x}$, that is, $v_K = \boldsymbol{\xi} \cdot \boldsymbol{x}_K$ and $v_\sigma = \boldsymbol{\xi} \cdot \overline{\boldsymbol{x}}_\sigma$. The $\mathbb{P}_1$-exactness property of $\nabla_{\mathcal{D}}$ [40, Lemma 13.10] shows that $\nabla_{\mathcal{D}} v = \boldsymbol{\xi}$ and (2.16a) thus gives

$$\sum_{\sigma \in \mathcal{E}_K} \mathcal{F}_{K,\sigma}(p^{(n+1)}) \boldsymbol{\xi} \cdot (\overline{\boldsymbol{x}}_\sigma - \boldsymbol{x}_K) = \int_K \mathbf{U}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x}) \cdot \boldsymbol{\xi} d\boldsymbol{x}.$$

Combining with (6.22) and using the generality of $\boldsymbol{\xi}$ then yields

$$\int_K \mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} - \int_K \mathbf{U}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x}) d\boldsymbol{x} = -\int_K \mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}(\boldsymbol{x})(\boldsymbol{x} - \boldsymbol{x}_K) d\boldsymbol{x}.$$

Denoting by $\mathrm{Pr}_{\mathcal{M}} : L^2(\Omega)^d \to L^2(\Omega)^d$ the orthogonal projection on the piecewise constant functions on $\mathcal{M}$ (that is, $(\mathrm{Pr}_{\mathcal{M}} f)_{|K} = \frac{1}{|K|} \int_K f(\boldsymbol{x}) d\boldsymbol{x}$ for all $K \in \mathcal{M}$), the above relation gives

$$\left\| \mathrm{Pr}_{\mathcal{M}}(\mathbf{u}_{\mathcal{P}}^{(n+1)} - \mathbf{U}_{\mathcal{P}}^{(n+1)}) \right\|_{L^1(\Omega)} \le h_{\mathcal{M}} \left\| \mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)} \right\|_{L^1(\Omega)},$$

where $h_{\mathcal{M}} = \max_{K \in \mathcal{M}} \mathrm{diam}(K)$ is the mesh size. Owing to the boundedness of $\mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}$, this shows that $\mathrm{Pr}_{\mathcal{M}}(\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}) \to 0$ in $L^\infty(0,T; L^1(\Omega))$ as $h_{\mathcal{M}} \to 0$. Take now $\boldsymbol{\varphi} \in C_c^\infty(\Omega \times (0,T))^d$. Using the orthogonality property of $\mathrm{Pr}_{\mathcal{M}}$,

$$\begin{aligned}
&\Big| \int_{\Omega \times (0,T)} (\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}) \cdot \boldsymbol{\varphi} \Big| \\
&\quad \le \Big| \int_{\Omega \times (0,T)} (\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}) \cdot (\boldsymbol{\varphi} - \mathrm{Pr}_{\mathcal{M}} \boldsymbol{\varphi}) \Big| + \Big| \int_{\Omega \times (0,T)} (\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}) \cdot \mathrm{Pr}_{\mathcal{M}} \boldsymbol{\varphi} \Big| \\
&\quad \le \|\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}\|_1 h_{\mathcal{M}} \|D\boldsymbol{\varphi}\|_\infty + \Big| \int_{\Omega \times (0,T)} \mathrm{Pr}_{\mathcal{M}}(\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}) \cdot \boldsymbol{\varphi} \Big| \\
&\quad \le \|\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}}\|_1 h_{\mathcal{M}} \|D\boldsymbol{\varphi}\|_\infty + \|\mathrm{Pr}_{\mathcal{M}}(\mathbf{u}_{\mathcal{P}} - \mathbf{U}_{\mathcal{P}})\|_1 \|\boldsymbol{\varphi}\|_\infty, \qquad (6.23)
\end{aligned}$$

where $\|\cdot\|_r = \|\cdot\|_{L^r(\Omega \times (0,T))}$ and we have used $\|\boldsymbol{\varphi} - \mathrm{Pr}_{\mathcal{M}} \boldsymbol{\varphi}\|_\infty \le h_{\mathcal{M}} \|D\boldsymbol{\varphi}\|_\infty$. The strong convergence of $\Pi_{\mathcal{C}} c$ ensures the strong convergence of $\widetilde{\Pi}_{\mathcal{C}} c$ (see end of Section 6.5.1); hence, the strong convergences assumed in (**A5**) imply that $\mathbf{U}_{\mathcal{P}} \to \mathbf{u} = -\frac{\mathbf{K}}{\mu(c)} \nabla p$ in $L^2(\Omega \times (0,T))^d$. Since the right-hand side of (6.23) tends to 0 as $h_{\mathcal{M}} \to 0$, this concludes the proof that $\mathbf{u}_{\mathcal{P}} \to \mathbf{u}$ weakly in $L^2(\Omega \times (0,T))^d$ as the mesh size and time step tend to 0.

6.4 A Priori Estimates

Throughout this section, $A \lesssim B$ means that $A \le CB$, where C is a constant depending only on the quantities $|\Omega|$, T, ϕ_*, ϕ^*, α_A, $\alpha_{\mathbf{D}}$, Λ_A, $\Lambda_{\mathbf{D}}$, M_{q^-}, M_{q^+},

M_t, M_F, $M_{\rm div}$, $\sup_{m\in\mathbb{N}} C_{\mathcal{P}_m}$, $\sup_{m\in\mathbb{N}} C_{\mathcal{C}_m}$ appearing in Assumptions (6.1) and (**A1**)–(**A5**) ($C_{\mathcal{P}_m}$ and $C_{\mathcal{C}_m}$ are given by (2.7a)). Likewise, in the proofs, C denotes a generic constant that can change from one line to the other, but only depends on the aforementioned parameters.

We also consider that (p_m, c_m) is a solution to the GDM–ELLAM scheme with $(\mathcal{P}, \mathcal{C}^T) = (\mathcal{P}_m, \mathcal{C}_m^T)$ and we drop the index m for legibility. Let $\mathbf{U}_\mathcal{P} = -\frac{\mathbf{K}}{\mu(\widetilde{\Pi}_\mathcal{C} c)} \nabla_\mathcal{P} p$.

Lemma 6.4.1 (Estimates on the pressure)**.** *The following estimate holds:*

$$\|\Pi_\mathcal{P} p\|_{L^\infty(0,T;L^2(\Omega))} + \|\nabla_\mathcal{P} p\|_{L^\infty(0,T;L^2(\Omega))} + \|\mathbf{U}_\mathcal{P}\|_{L^\infty(0,T;L^2(\Omega))} \lesssim 1.$$

Proof. Setting $z = p^{(n+1)}$ in the gradient scheme (5.1), we get:

$$\int_\Omega A(\boldsymbol{x}, \Pi_\mathcal{C} c^{(n)}) \nabla_\mathcal{P} p^{(n+1)} \cdot \nabla_\mathcal{P} p^{(n+1)} = \int_\Omega (q_n^+ - q_n^-) \Pi_\mathcal{P} p^{(n+1)}.$$

Using (6.1c) for the left hand side, followed by Cauchy–Schwarz' inequality

$$\left\|\nabla_\mathcal{P} p^{(n+1)}\right\|_{L^2(\Omega)}^2 \lesssim \left\|q_n^+ - q_n^-\right\|_{L^2(\Omega)} \left\|\Pi_\mathcal{P} p^{(n+1)}\right\|_{L^2(\Omega)} \lesssim \left\|\nabla_\mathcal{P} p^{(n+1)}\right\|_{L^2(\Omega)} \tag{6.24}$$

where we used

$$\left\|\Pi_\mathcal{P} p^{(n+1)}\right\|_{L^2(\Omega)} \lesssim \left\|p^{(n+1)}\right\|_\mathcal{P} = \left\|\nabla_\mathcal{P} p^{(n+1)}\right\|_{L^2(\Omega)} \tag{6.25}$$

since $\int_\Omega \Pi_\mathcal{P} p^{(n+1)} = 0$. Equation (6.24) proves the estimate on $\nabla_\mathcal{P} p$ which gives the bound on $\mathbf{U}_\mathcal{P}$ (owing to (6.1c)) and, using (6.25) once more, provides the estimate on $\Pi_\mathcal{P} p$. ∎

Lemma 6.4.2 (Estimates on the concentration)**.** *The following estimate holds:*

$$\|\Pi_\mathcal{C} c\|_{L^\infty(0,T;L^2(\Omega))} + \left\|(1 + |\mathbf{U}_\mathcal{P}|)^{1/2} \nabla_\mathcal{C} c\right\|_{L^2(0,T;L^2(\Omega))} \lesssim 1 + \|\Pi_\mathcal{C} \mathcal{I}_\mathcal{C} c_{\rm ini}\|_{L^2(\Omega)}.$$

As a consequence, $\|\nabla_\mathcal{C} c\|_{L^2(0,T;L^2(\Omega))} \lesssim 1 + \|\Pi_\mathcal{C} \mathcal{I}_\mathcal{C} c_{\rm ini}\|_{L^2(\Omega)}$.

Proof. Set $Y_n = \left\|\Pi_\mathcal{C} c^{(n)} \sqrt{\phi}\right\|_{L^2(\Omega)}$. The gradient scheme (5.3) with $z = c^{(n+1)}$ yields

$$\begin{aligned} Y_{n+1}^2 &- \int_\Omega \phi \Pi_\mathcal{C} c^{(n)} v(t^{(n)}) + \delta t^{(n+\frac12)} \int_\Omega D(\boldsymbol{x}, \mathbf{U}_\mathcal{P}^{(n+1)}) \nabla_\mathcal{C} c^{(n+1)} \cdot \nabla_\mathcal{C} c^{(n+1)} \\ &+ \varpi_n \delta t^{(n+\frac12)} \int_\Omega \Pi_\mathcal{C} c^{(n)} v(t^{(n)}) q_n^- + (1 - \varpi_n) \delta t^{(n+\frac12)} \int_\Omega (\Pi_\mathcal{C} c^{(n+1)})^2 q_{n+1}^- \\ &= \varpi_n \delta t^{(n+\frac12)} \int_\Omega q_n^+ v(t^{(n)}) + (1 - \varpi_n) \delta t^{(n+\frac12)} \int_\Omega q_{n+1}^+ \Pi_\mathcal{C} c^{(n+1)} =: \Delta. \end{aligned}$$

By Cauchy-Schwarz, recalling that $0 \leq \varpi_n \leq 1$ and that $|q_n^-/\sqrt{\phi}| \leq M_{q^-}/\sqrt{\phi_*}$, and using the coercivity property of the diffusion tensor $\mathbf{D}$,

$$\Delta \geq Y_{n+1}^2 - Y_n \left\| v(t^{(n)})\sqrt{\phi} \right\|_{L^2(\Omega)} + \alpha_{\mathbf{D}} \delta t^{(n+\frac{1}{2})} \left\| (1 + |\mathbf{U}_{\mathcal{P}}^{(n+1)}|)|\nabla_{\mathcal{C}} c^{(n+1)}|^2 \right\|_{L^1(\Omega)} - \frac{M_{q^-}}{\sqrt{\phi_*}} \delta t^{(n+\frac{1}{2})} Y_n \left\| v(t^{(n)}) \right\|_{L^2(\Omega)}.$$

Consider the term $Y_n \left\| v(t^{(n)})\sqrt{\phi} \right\|_{L^2(\Omega)}$ in the right hand side of the inequality. Estimate (4.10) with $w(\boldsymbol{x}, t) = v(\boldsymbol{x}, t)^2$ and $s = \delta t^{(n+\frac{1}{2})}$ (so that $v(t^{(n+1)} - s) = v(t^{(n)})$) followed by Young's inequality gives, for any $\varepsilon > 0$,

$$\begin{aligned} Y_n \left\| v(t^{(n)})\sqrt{\phi} \right\|_{L^2(\Omega)} &\leq Y_n Y_{n+1} \sqrt{1 + C\delta t^{(n+\frac{1}{2})}} \leq Y_n Y_{n+1}(1 + C\delta t^{(n+\frac{1}{2})}) \\ &\leq \frac{1}{2} Y_n^2 + \frac{1}{2} Y_{n+1}^2 + \frac{C^2 \delta t^{(n+\frac{1}{2})}}{2\varepsilon} Y_n^2 + \frac{\delta t^{(n+\frac{1}{2})} \varepsilon}{2} Y_{n+1}^2. \end{aligned} \tag{6.26}$$

Using (4.11),

$$Y_n \left\| v(t^{(n)}) \right\|_{L^2(\Omega)} \leq C Y_n Y_{n+1} \leq \frac{C^2}{2\varepsilon} Y_n^2 + \frac{\varepsilon}{2} Y_{n+1}^2. \tag{6.27}$$

Using (6.26) together with (6.27), we then have

$$\begin{aligned} \Delta \geq Y_{n+1}^2 &- \left(\frac{1}{2} Y_n^2 + \frac{1}{2} Y_{n+1}^2 + \frac{C^2 \delta t^{(n+\frac{1}{2})}}{2\varepsilon} Y_n^2 + \frac{\delta t^{(n+\frac{1}{2})} \varepsilon}{2} Y_{n+1}^2 \right) \\ &+ \alpha_{\mathbf{D}} \delta t^{(n+\frac{1}{2})} \left\| (1 + |\mathbf{U}_{\mathcal{P}}^{(n+1)}|)|\nabla_{\mathcal{C}} c^{(n+1)}|^2 \right\|_{L^1(\Omega)} \\ &- \frac{M_{q^-}}{\sqrt{\phi_*}} \delta t^{(n+\frac{1}{2})} \left(\frac{C^2}{2\varepsilon} Y_n^2 + \frac{\varepsilon}{2} Y_{n+1}^2 \right), \end{aligned}$$

which implies that

$$\begin{aligned} \frac{1}{2} Y_{n+1}^2 - \frac{1}{2} Y_n^2 + \alpha_{\mathbf{D}} \delta t^{(n+\frac{1}{2})} &\left\| (1 + |\mathbf{U}_{\mathcal{P}}^{(n+1)}|)|\nabla_{\mathcal{C}} c^{(n+1)}|^2 \right\|_{L^1(\Omega)} \\ &\lesssim \Delta + \frac{\delta t^{(n+\frac{1}{2})}}{\varepsilon} Y_n^2 + \varepsilon \delta t^{(n+\frac{1}{2})} Y_{n+1}^2. \end{aligned} \tag{6.28}$$

Now, using the boundedness of q^+, Young's inequality, the fact that $\varpi_n \in [0, 1]$ and (4.11) with $w(\boldsymbol{x}, t) = v(\boldsymbol{x}, t)^2$ and $s = \delta t^{(n+\frac{1}{2})}$,

$$\begin{aligned} \Delta &\lesssim \delta t^{(n+\frac{1}{2})} \left(\left\| v(t^{(n)}) \right\|_{L^2(\Omega)} + \left\| \Pi_{\mathcal{C}} c^{(n+1)} \right\|_{L^2(\Omega)} \right) \\ &\lesssim \delta t^{(n+\frac{1}{2})} \left[\frac{1}{\varepsilon} + \varepsilon \left\| v(t^{(n)}) \right\|_{L^2(\Omega)}^2 + \varepsilon Y_{n+1}^2 \right] \lesssim \frac{\delta t^{(n+\frac{1}{2})}}{\varepsilon} + \delta t^{(n+\frac{1}{2})} \varepsilon Y_{n+1}^2. \end{aligned}$$

Combining with (6.28), we find

$$\frac{1}{2}Y_{n+1}^2 - \frac{1}{2}Y_n^2 + \alpha_{\mathbf{D}}\delta t^{(n+\frac{1}{2})}\left\|(1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|)|\nabla_{\mathcal{C}}c^{(n+1)}|^2\right\|_{L^1(\Omega)} \lesssim \frac{\delta t^{(n+\frac{1}{2})}}{\varepsilon} + \frac{\delta t^{(n+\frac{1}{2})}}{\varepsilon}Y_n^2 + \varepsilon\delta t^{(n+\frac{1}{2})}Y_{n+1}^2,$$

which, upon taking a telescoping sum over n, yields

$$\begin{aligned}\frac{1}{2}Y_{n+1}^2 - \frac{1}{2}Y_0^2 + \alpha_{\mathbf{D}}\sum_{k=0}^{n}\delta t^{(k+\frac{1}{2})}\left\|(1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|)|\nabla_{\mathcal{C}}c^{(n+1)}|^2\right\|_{L^1(\Omega)} &\lesssim \frac{1}{\varepsilon}\sum_{k=0}^{n}\delta t^{(k+\frac{1}{2})} + \frac{1}{\varepsilon}\sum_{k=0}^{n}\delta t^{(k+\frac{1}{2})}Y_k^2 + \varepsilon\sum_{k=1}^{n+1}\delta t^{(k-\frac{1}{2})}Y_k^2 \\ &\lesssim \frac{1}{\varepsilon}T + \frac{1}{\varepsilon}\delta t^{(\frac{1}{2})}Y_0^2 + \varepsilon\delta t^{(n+\frac{1}{2})}Y_{n+1}^2 + \left(\frac{1}{\varepsilon}+\varepsilon\right)\sum_{k=1}^{n}(\delta t^{(k+\frac{1}{2})} + \delta t^{(k-\frac{1}{2})})Y_k^2.\end{aligned}$$

Denoting by C the hidden multiplicative constant in the last $\lesssim$ above, choose $\varepsilon = 1/(4CT)$ to absorb the term $\varepsilon\delta t^{(n+\frac{1}{2})}Y_{n+1}^2$ in the left-hand side. Since ε depends only on fixed quantities, we no longer make it explicit and it disappears into the $\lesssim$ symbols. Setting $\delta t^{(-\frac{1}{2})} = 0$ the term $\delta t^{(\frac{1}{2})}Y_0^2$ can be integrated in the last sum and we find

$$Y_{n+1}^2 + \left\|(1+|\mathbf{U}_{\mathcal{P}}|)|\nabla_{\mathcal{C}}c|^2\right\|_{L^1(\Omega\times(0,t^{(n+1)}))} \lesssim 1 + Y_0^2 + \sum_{k=0}^{n}(\delta t^{(k+\frac{1}{2})} + \delta t^{(k-\frac{1}{2})})Y_k^2. \tag{6.29}$$

Dropping for a moment the second term in the left-hand side, and letting C denote the hidden multiplicative constant in $\lesssim$, a discrete Gronwall's inequality [60, Section 5] yields, for any $n = 0, \ldots, N-1$,

$$Y_{n+1}^2 \le C(1+Y_0^2)\exp\Big(\sum_{k=0}^{n}C(\delta t^{(k+\frac{1}{2})} + \delta t^{(k-\frac{1}{2})})\Big) \le C(1+Y_0^2)\exp(2CT).$$

By noticing that $Y_0 \le \sqrt{\phi^*}\left\|\Pi_{\mathcal{C}}c^{(0)}\right\|_{L^2(\Omega)} = \sqrt{\phi^*}\left\|\Pi_{\mathcal{C}}\mathcal{I}_{\mathcal{C}}c_{\mathrm{ini}}\right\|_{L^2(\Omega)}$, this proves the estimate on $\|\Pi_{\mathcal{C}}c\|_{L^\infty(0,T;L^2(\Omega))}$. Plugging this estimate in (6.29) with $n = N-1$ yields the estimate on $\left\|(1+|\mathbf{U}_{\mathcal{P}}|)^{1/2}\nabla_{\mathcal{C}}c\right\|_{L^2(0,T;L^2(\Omega))}$ which, in turn, trivially provides a bound on $\|\nabla_{\mathcal{C}}c\|_{L^2(0,T;L^2(\Omega))}$. ∎

Remark 6.4.3 (Estimate of the advection–reaction terms)**.** *A formal integration-by-parts shows that, if* $\mathbf{u}$ *satisfies* (1.1a)*,*

$$\int_\Omega \mathrm{div}(c\mathbf{u})c + \int_\Omega q^- c^2 = \frac{1}{2}\int_\Omega (q^+ + q^-)c^2 \ge 0.$$

When using c as a test function in the continuous equation, the advection and reaction terms thus combine together to create a non-negative quantity that can simply be discarded from the estimates (which thus hold under very weak assumptions on $q^\pm$). This can be reproduced at the discrete level for upwind discretisations [20, 19]. However, the structure of the ELLAM discretisation does not seem to lend itself to such an easy estimate of the advection–reaction terms, which is why the proof of Lemma 6.4.2 is a bit technical, and requires the boundedness of $q^\pm$ (to bound the Jacobian of the changes of variables – note that we do not require a bound on $\mathbf{u}$ itself, though).

A crucial step in the convergence proof is to establish the strong compactness of $\Pi_{\mathcal{C}} c$. This is done by using a discrete version of the Aubin–Simon theorem. The gradient estimates in Lemma 6.4.2 provides the compactness in space, which must be complemented by some sort of boundedness (in a dual norm) of the discrete time-derivative of c. Establishing this boundedness is the purpose of the following lemma. A dual norm $\|\cdot\|_{\star,\phi,\mathcal{C}}$ is defined on $\Pi_{\mathcal{C}}(X_{\mathcal{C}})$ the following way:

$$\forall w \in \Pi_{\mathcal{C}}(X_{\mathcal{C}})$$
$$\|w\|_{\star,\phi,\mathcal{C}} := \sup\left\{ \int_\Omega \phi w \Pi_{\mathcal{C}} v \,:\, v \in X_{\mathcal{C}}\,,\ \|\nabla_{\mathcal{C}} v\|_{L^4(\Omega)} + \|\Pi_{\mathcal{C}} v\|_{L^\infty(\Omega)} = 1 \right\}.$$

It can easily be checked that this is indeed a norm (if $w \neq 0$, write $w = \Pi_{\mathcal{C}} z$, take $v = z/\mathcal{N}$ where $\mathcal{N} = \|\nabla_{\mathcal{C}} z\|_{L^4(\Omega)} + \|\Pi_{\mathcal{C}} z\|_{L^\infty(\Omega)} > 0$, and notice that $\|w\|_{\star,\phi,\mathcal{C}} \geq \int_\Omega \phi w(\boldsymbol{x}) \Pi_{\mathcal{C}} v(\boldsymbol{x}) d\boldsymbol{x} = \mathcal{N}^{-1} \left\| \sqrt{\phi} w \right\|^2_{L^2(\Omega)}$).

Lemma 6.4.4. *Defining the discrete time derivative of c by*

$$\delta_{\mathcal{C}} c(t) = \frac{\Pi_{\mathcal{C}} c^{(n+1)} - \Pi_{\mathcal{C}} c^{(n)}}{\mathscr{d}t^{(n+\frac{1}{2})}} \text{ for all } t \in (t^{(n)}, t^{(n+1)}) \text{ and all } n = 0, \ldots, N-1,$$

we have

$$\int_0^T \|\delta_{\mathcal{C}} c\|^2_{\star,\phi,\mathcal{C}}\, dt \lesssim 1 + \|\Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}}\|^2_{L^\infty(\Omega)}\,.$$

Proof. Take $z \in X_{\mathcal{C}}$ arbitrary in (5.3). Subtract and add $\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} \Pi_{\mathcal{C}} z$

to get

$$\begin{aligned}
&\int_\Omega \phi(\Pi_{\mathcal{C}} c^{(n+1)} - \Pi_{\mathcal{C}} c^{(n)})\Pi_{\mathcal{C}} z \\
&= -\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)}(\Pi_{\mathcal{C}} z - v(t^{(n)})) - \delta t^{(n+\frac{1}{2})} \int_\Omega D(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}}^{(n+1)}) \nabla_{\mathcal{C}} c^{(n+1)} \cdot \nabla_{\mathcal{C}} z \\
&\quad - \varpi_n \delta t^{(n+\frac{1}{2})} \int_\Omega \Pi_{\mathcal{C}} c^{(n)} v(t^{(n)}) q_n^- - (1-\varpi_n)\delta t^{(n+\frac{1}{2})} \int_\Omega \Pi_{\mathcal{C}} c^{(n+1)} \Pi_{\mathcal{C}} z q_{n+1}^- \\
&\quad + \varpi_n \delta t^{(n+\frac{1}{2})} \int_\Omega q_n^+ v(t^{(n)}) + (1-\varpi_n)\delta t^{(n+\frac{1}{2})} \int_\Omega q_{n+1}^+ \Pi_{\mathcal{C}} z.
\end{aligned}$$

The terms on the right hand side of the equation are referred to as $T_1, T_2, \ldots, T_6$, respectively. For the term T_1, recall that $v(\boldsymbol{x}, t^{(n)}) = \Pi_{\mathcal{C}} z(F_{\delta t^{(n+1/2)}}(\boldsymbol{x}))$. If $n = 0$, noticing that $c^{(0)} = \mathcal{I}_{\mathcal{C}} c_{\text{ini}}$ and applying (**A2**) shows that

$$\begin{aligned}
|T_1| &\lesssim \|\Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\text{ini}}\|_{L^\infty(\Omega)} \|\Pi_{\mathcal{C}} z - \Pi_{\mathcal{C}} z(F_{\delta t^{(1/2)}})\|_{L^1(\Omega)} \\
&\lesssim \delta t^{(\frac{1}{2})} \|\Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\text{ini}}\|_{L^\infty(\Omega)} \left\|\mathbf{u}_{\mathcal{P}}^{(1)}\right\|_{L^2(\Omega)} \|\nabla_{\mathcal{C}} z\|_{L^2(\Omega)}.
\end{aligned} \tag{6.30}$$

If $n \neq 0$, a change of variables yields

$$\begin{aligned}
-T_1 = &\int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} \Pi_{\mathcal{C}} z \\
&- \int_\Omega \phi\big(F_{-\delta t^{(n+1/2)}}(\boldsymbol{x})\big)\, \Pi_{\mathcal{C}} c^{(n)}\big(F_{-\delta t^{(n+1/2)}}(\boldsymbol{x})\big)\, \Pi_{\mathcal{C}} z(\boldsymbol{x})\, \big|JF_{-\delta t^{(n+1/2)}}(\boldsymbol{x})\big|\, d\boldsymbol{x}.
\end{aligned}$$

Applying (4.8) with $s = -\delta t^{(n+\frac{1}{2})}$, we can thus write $-T_1 = T_{11} - T_{12}$ with

$$\begin{aligned}
T_{11} &= \int_\Omega \phi \Pi_{\mathcal{C}} c^{(n)} \Pi_{\mathcal{C}} z - \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}} c^{(n)}\left(F_{-\delta t^{(n+1/2)}}(\boldsymbol{x})\right) \Pi_{\mathcal{C}} z(\boldsymbol{x}) d\boldsymbol{x} \\
T_{12} &= \int_\Omega \Bigg[\Pi_{\mathcal{C}} c^{(n)}\left(F_{-\delta t^{(n+1/2)}}(\boldsymbol{x})\right) \Pi_{\mathcal{C}} z(\boldsymbol{x}) \\
&\qquad \times \int_0^{-\delta t^{(n+\frac{1}{2})}} |JF_t(\boldsymbol{x})| (\text{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}) \circ F_t(\boldsymbol{x}) dt \Bigg] d\boldsymbol{x}.
\end{aligned}$$

Using (**A2**) leads to

$$\begin{aligned}
|T_{11}| &\leq \int_\Omega \left|\phi \Pi_{\mathcal{C}} z \left(\Pi_{\mathcal{C}} c^{(n)} - \Pi_{\mathcal{C}} c^{(n)}(F_{-\delta t^{(n+1/2)}})\right)\right| \\
&\lesssim \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{C}} z\|_{L^\infty(\Omega)} \left\|\mathbf{u}_{\mathcal{P}}^{(n+1)}\right\|_{L^2(\Omega)} \left\|\nabla_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)}.
\end{aligned}$$

The boundedness of $\mathrm{div}\mathbf{u}_{\mathcal{P}}^{(n+1)}$ in (**A4**) and of $|JF_t|$ (see (4.9)) yield, by a change of variables,

$$\begin{aligned}|T_{12}| &\lesssim \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n)}(F_{-\delta t^{(n+1/2)}})\right\|_{L^2(\Omega)} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)} \\ &\lesssim \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)}.\end{aligned}$$

For the term T_2, the property (6.1d) of the diffusion tensor $\mathbf{D}$ and Hölder's inequality with exponents 4, 2 and 4 give

$$\begin{aligned}|T_2| &\lesssim \delta t^{(n+\frac{1}{2})} \int_\Omega \sqrt{1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|}\left(\sqrt{1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|}\,|\nabla_{\mathcal{C}} c^{(n+1)}|\right)|\nabla_{\mathcal{C}} z| \\ &\lesssim \delta t^{(n+\frac{1}{2})} \left\|1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|\right\|_{L^2(\Omega)}^{\frac{1}{2}} \left\|(1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|)^{\frac{1}{2}}|\nabla_{\mathcal{C}} c^{(n+1)}|\right\|_{L^2(\Omega)} \|\nabla_{\mathcal{C}} z\|_{L^4(\Omega)}.\end{aligned}$$

The terms T_3 to T_6 are estimated by using the Cauchy–Schwarz inequality and the fact that $\varpi_n \in [0,1]$:

$$\begin{aligned}|T_3| &\lesssim \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} \left\|v(t^{(n)})\right\|_{L^2(\Omega)}, \\ |T_4| &\lesssim \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n+1)}\right\|_{L^2(\Omega)} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)}, \\ |T_5+T_6| &\lesssim \delta t^{(n+\frac{1}{2})} \left\|v(t^{(n)})\right\|_{L^2(\Omega)} + \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)} \lesssim \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)}\end{aligned}$$

(we have used (4.11) with $w = v^2$ and $s = \delta t^{(n+\frac{1}{2})}$ to obtain $\left\|v(t^{(n)})\right\|_{L^2(\Omega)} \lesssim \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)}$). For $n \neq 0$, combining the estimates from T_1 to T_6 leads to

$$\begin{aligned}&\int_\Omega \phi(\Pi_{\mathcal{C}} c^{(n+1)} - \Pi_{\mathcal{C}} c^{(n)})\Pi_{\mathcal{C}} z \\ &\lesssim \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{C}} z\|_{L^\infty(\Omega)} \left\|\mathbf{u}_{\mathcal{P}}^{(n+1)}\right\|_{L^2(\Omega)} \left\|\nabla_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} \\ &\quad + \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)} \\ &\quad + \delta t^{(n+\frac{1}{2})} \left\|1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|\right\|_{L^2(\Omega)}^{\frac{1}{2}} \left\|(1+|\mathbf{U}_{\mathcal{P}}^{(n+1)}|)^{\frac{1}{2}}|\nabla_{\mathcal{C}} c^{(n+1)}|\right\|_{L^2(\Omega)} \|\nabla_{\mathcal{C}} z\|_{L^4(\Omega)} \\ &\quad + \delta t^{(n+\frac{1}{2})} \left\|\Pi_{\mathcal{C}} c^{(n+1)}\right\|_{L^2(\Omega)} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)} + \delta t^{(n+\frac{1}{2})} \|\Pi_{\mathcal{C}} z\|_{L^2(\Omega)}.\end{aligned} \tag{6.31}$$

Divide both sides by $\delta t^{(n+\frac{1}{2})}$ and take the supremum over all $z \in X_{\mathcal{C}}$ with $\|\nabla_{\mathcal{C}} z\|_{L^4(\Omega)} + \|\Pi_{\mathcal{C}} z\|_{L^\infty(\Omega)} = 1$ to obtain, for all $n = 1, \dots, N-1$ and $t \in (t^{(n)}, t^{(n+1)})$,

$$\|\delta_{\mathcal{C}} c(t)\|_{\star,\phi,\mathcal{C}} \lesssim \left\|\mathbf{u}_{\mathcal{P}}^{(n+1)}\right\|_{L^2(\Omega)} \left\|\nabla_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} + \left\|\Pi_{\mathcal{C}} c^{(n)}\right\|_{L^2(\Omega)} + \left\|\Pi_{\mathcal{C}} c^{(n+1)}\right\|_{L^2(\Omega)}$$

$$+ \left\| 1 + |\mathbf{U}_{\mathcal{P}}^{(n+1)}| \right\|_{L^2(\Omega)}^{\frac{1}{2}} \left\| (1 + |\mathbf{U}_{\mathcal{P}}^{(n+1)}|)^{\frac{1}{2}} \nabla_{\mathcal{C}} c^{(n+1)} \right\|_{L^2(\Omega)} + 1. \quad (6.32)$$

Square this, integrate for $t \in (t^{(n)}, t^{(n+1)})$ and sum over $n = 1, \ldots, N-1$. The assumption on the time steps in (**A1**) ensures that

$$\sum_{n=1}^{N-1} \delta t^{(n+1/2)} \left\| \nabla_{\mathcal{C}} c^{(n)} \right\|_{L^2(\Omega)}^2 \lesssim \sum_{n=1}^{N-1} \delta t^{(n-1/2)} \left\| \nabla_{\mathcal{C}} c^{(n)} \right\|_{L^2(\Omega)}^2$$
$$= \sum_{n=0}^{N-2} \delta t^{(n+1/2)} \left\| \nabla_{\mathcal{C}} c^{(n+1)} \right\|_{L^2(\Omega)}^2 \leq \left\| \nabla_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2$$

(and similarly for the terms involving $\Pi_{\mathcal{C}} c^{(n)}$), so that

$$\int_{t^{(1)}}^{T} \left\| \delta_{\mathcal{C}} c(t) \right\|_{\star,\phi,\mathcal{C}}^2 dt \lesssim \left\| \mathbf{u}_{\mathcal{P}} \right\|_{L^\infty(0,T;L^2(\Omega))}^2 \left\| \nabla_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2 + \left\| \Pi_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2$$
$$+ \left\| 1 + |\mathbf{U}_{\mathcal{P}}| \right\|_{L^\infty(0,T;L^2(\Omega))} \left\| (1 + |\mathbf{U}_{\mathcal{P}}|)^{\frac{1}{2}} \nabla_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2 + 1. \quad (6.33)$$

To estimate $\int_0^{t^{(1)}} \left\| \delta_{\mathcal{C}} c(t) \right\|_{\star,\phi,\mathcal{C}}^2 dt$, we come back to (6.31) with $n = 0$. The first term in the right-hand side of this inequality must be replaced by the right-hand side of (6.30), and thus the first term in (6.32) is replaced by $\delta t^{(1/2)} \left\| \Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}} \right\|_{L^\infty(\Omega)} \left\| \mathbf{u}_{\mathcal{P}}^{(1)} \right\|_{L^2(\Omega)}$. Hence,

$$\int_0^{t^{(1)}} \left\| \delta_{\mathcal{C}} c(t) \right\|_{\star,\phi,\mathcal{C}}^2 dt \lesssim \delta t^{(1/2)} \left\| \Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}} \right\|_{L^\infty(\Omega)}^2 \left\| \mathbf{u}_{\mathcal{P}}^{(1)} \right\|_{L^2(\Omega)}^2$$
$$+ \delta t^{(1/2)} \left\| \Pi_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}} \right\|_{L^2(\Omega)}^2 + \left\| \Pi_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2$$
$$+ \left\| 1 + |\mathbf{U}_{\mathcal{P}}| \right\|_{L^\infty(0,T;L^2(\Omega))} \left\| (1 + |\mathbf{U}_{\mathcal{P}}|)^{\frac{1}{2}} \nabla_{\mathcal{C}} c \right\|_{L^2(\Omega \times (0,T))}^2 + 1. \quad (6.34)$$

The reason for separating the case $n \neq 0$ from the case $n = 0$ is that, for $n = 0$, (6.31) involves $\nabla_{\mathcal{C}} c^{(0)} = \nabla_{\mathcal{C}} \mathcal{I}_{\mathcal{C}} c_{\mathrm{ini}}$ on which no bound has been imposed. The proof is completed by adding together (6.33) and (6.34), and by invoking Assumption (**A5**) and Lemmas 6.4.1 and 6.4.2. ∎

6.5 Proof of the main theorem (GDM–ELLAM)

At each time step, (5.1) and (5.3) are square linear equations on $p^{(n+1)}$ and $c^{(n+1)}$. The estimates of Lemma 6.4.1 and 6.4.2, together with the definition

of the norms in $X_{\mathcal{P}}$ and $X_{\mathcal{C}}$, show that the solutions to these linear systems remain bounded. Hence, the matrices associated with these systems has an empty null space, which ensures the existence and uniqueness of (p, c) solution to the GDM–ELLAM scheme.

We now establish the compactness of $(\Pi_{\mathcal{C}_m} c_m)_{m\in\mathbb{N}}$, which is essential to proving the convergence of the pressure. Once this latter is establish, we conclude the proof by dealing with the convergence of the concentration.

6.5.1 Compactness and initial convergence of $\Pi_{\mathcal{C}_m} c_m$

Theorem 6.5.1. *Under the assumptions and notations of Theorem 6.1.1, the sequence $(\Pi_{\mathcal{C}_m} c_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(0,T;L^2(\Omega))$.*

Proof. The idea is to apply Theorem 6.7.3 in Section 6.7 with $X_m = \Pi_{\mathcal{C}_m}(X_{\mathcal{C}_m})$ equipped with the norm

$$\|u\|_{X_m} = \min\{\|w\|_{\mathcal{C}_m} : w \in X_{\mathcal{C}_m} \text{ s.t. } \Pi_{\mathcal{C}_m} w = u\}$$

and $Y_m = X_m$ with the norm $\|\cdot\|_{Y_m} = \|\cdot\|_{\star,\phi,\mathcal{C}_m}$.

Let us show that $(X_m, Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in $L^2(\Omega)$ (Definition 6.7.2). Item 1 follows by the compactness of $(\mathcal{C}_m)_{m\in\mathbb{N}}$, see Definition 2.1.3. Take now $(u_m)_{m\in\mathbb{N}}$ as prescribed in Item 2 and let u be the limit in $L^2(\Omega)$ of this sequence. Let $\varphi \in C_c^\infty(\Omega)$ and consider the interpolant $\mathcal{J}_{\mathcal{C}_m}$ given by Assumption (**A3**). Then $\|\Pi_{\mathcal{C}_m}\mathcal{J}_{\mathcal{C}_m}\varphi\|_{L^\infty(\Omega)} + \|\nabla_{\mathcal{C}_m}\mathcal{J}_{\mathcal{C}_m}\varphi\|_{L^4(\Omega)} \le C_\varphi$ for some $C_\varphi > 0$ not depending on m, and thus, by definition of $\|\cdot\|_{Y_m} = \|\cdot\|_{\star,\phi,\mathcal{C}_m}$,

$$\left|\int_\Omega \phi u_m \frac{\Pi_{\mathcal{C}_m}\mathcal{J}_{\mathcal{C}_m}\varphi}{C_\varphi}\right| \le \|u_m\|_{Y_m}.$$

Taking the limit as $m \to \infty$, we get $\int_\Omega \phi u\varphi = 0$. Since this is true for all $\varphi \in C_c^\infty(\Omega)$, we deduce that $u = 0$ as required.

We are left to show that the sequence $(f_m)_{m\in\mathbb{N}} = (\Pi_{\mathcal{C}_m} c_m)_{m\in\mathbb{N}}$ satisfies the properties in Theorem 6.7.3. The first property is trivially satisfied by the definition f_m, whereas the second and third one follow from Lemma 6.4.2 and the definition of the norm $\|\cdot\|_{\mathcal{C}_m}$ (Definition 2.1.1). The last property holds due to Lemma 6.4.4.

Thus, we may use Theorem 6.7.3 to conclude that the sequence $(\Pi_{\mathcal{C}_m} c_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(0,T;L^2(\Omega))$. ∎

Theorem 6.5.1 together with Lemma 6.7.1 give $c \in L^2(0,T;H^1(\Omega))$ such that, up to a subsequence as $m \to \infty$, $\Pi_{\mathcal{C}_m} c_m \to c$ strongly in $L^2((0,T)\times\Omega)$ and $\nabla_{\mathcal{C}_m} c_m \to \nabla c$ weakly in $L^2((0,T)\times\Omega)^d$. From here on we always consider

subsequences that satisfy these convergences. Let $\alpha_m : [0,T] \to \mathbb{R}$ be the piecewise affine map that maps each interval $(t_m^{(n)}, t_m^{(n+1)})$ onto $(t_m^{(n-1)}, t_m^{(n)})$, for $n = 1, \dots, N_m - 1$. That is,

$$\alpha_m(t) = t - \left(1 - \frac{\delta t_m^{(n-1/2)}}{\delta t_m^{(n+1/2)}}\right)(t_m - t_m^{(n)}) - (t_m^{(n)} - t_m^{(n-1)}) \text{ for } t \in (t_m^{(n)}, t_m^{(n+1)}).$$

Recalling the definition of $\widetilde{\Pi}_{\mathcal{C}_m} c_m$ at the start of Section 6.4, it holds $\widetilde{\Pi}_{\mathcal{C}_m} c_m = \Pi_{\mathcal{C}_m} c_m(\cdot, \alpha_m(\cdot))$ on $\Omega \times (t^{(1)}, T)$ and $\widetilde{\Pi}_{\mathcal{C}_m} c_m = \Pi_{\mathcal{C}_m} \mathcal{I}_{\mathcal{C}_m} c_{\rm ini}$ on $\Omega \times (0, t^{(1)})$. We have $\alpha_m(t) \to t$ uniformly as $m \to \infty$ and, due to (**A1**), the derivative of the inverse function α_m^{-1} is uniformly bounded. Hence, a triangle inequality, a change of variables in time using α_m^{-1}, and the strong convergence of $(\Pi_{\mathcal{C}_m} c_m)_{m\in\mathbb{N}}$ show that $\widetilde{\Pi}_{\mathcal{C}_m} c_m \to c$ in $L^2(\Omega \times (0,T))$ as $m \to \infty$.

6.5.2 Convergence of the pressure

Step 1: *weak convergences of* $\Pi_{\mathcal{P}_m} p_m$ *and* $\nabla_{\mathcal{P}_m} p_m$ We use Lemmas 6.4.1 and 6.7.1 to obtain $p \in L^\infty(0,T; H^1(\Omega))$ such that, up to a subsequence,

$$\begin{aligned}
&\Pi_{\mathcal{P}_m} p_m \to p \quad \text{weakly-}* \text{ in } L^\infty(0,T; L^2(\Omega)) \text{ and} \\
&\nabla_{\mathcal{P}_m} p_m \to \nabla p \quad \text{weakly-}* \text{ in } L^\infty(0,T; L^2(\Omega)^d).
\end{aligned}$$

The zero-average condition in (5.1) shows that $\int_\Omega \Pi_{\mathcal{P}_m} p_m(\cdot, t) = 0$ for all $t \in (0,T)$. Hence, the weak-$*$ convergence of $\Pi_{\mathcal{P}_m} p_m$ ensures that $\int_\Omega p(\cdot, t) = 0$ for a.e. $t \in (0,T)$ (test the zero-average condition on $\Pi_{\mathcal{P}_m} p_m$ with functions $\rho \in L^\infty(0,T)$ and pass to the limit).

Consider $\psi(\boldsymbol{x}, t) = \Xi(t)\eta(\boldsymbol{x})$ with $\Xi \in C^\infty([0,T])$ and $\eta \in C^\infty(\overline{\Omega})$. Define $\Xi_{\delta t_m}(t) = \Xi(t^{(n+1)})$ on $(t^{(n)}, t^{(n+1)})$ for each n and note that $(\Xi_{\delta t_m})_{m\in\mathbb{N}}$ converges to Ξ uniformly.

By consistency of $(\mathcal{P}_m)_{m\in\mathbb{N}}$, there exists $z_m \in \mathcal{P}_m$ such that $\Pi_{\mathcal{P}_m} z_m \to \eta$ and $\nabla_{\mathcal{P}_m} z_m \to \nabla\eta$ strongly in $L^2(\Omega)$. Recalling that $A = K/\mu$ satisfies (6.1c), [40, Lemma D.9] shows that $A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m)\nabla_{\mathcal{P}_m} z_m \to A(\boldsymbol{x}, c)\nabla\eta$ strongly in $L^2(\Omega \times (0,T))^d$. Apply the second equation of (5.1) to $z = \Xi(t^{(n+1)}) z_m$, multiply by $\delta t_m^{(n+\frac{1}{2})}$, and take the sum over $n = 0, \dots, N_m - 1$. denoting by $q^\pm_{\delta t_m}$ the piecewise-constant-in-time functions equal to $q_n^\pm$ on $(t^{(n)}, t^{(n+1)})$, we obtain

$$\begin{aligned}
\int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m)\nabla_{\mathcal{P}_m} p_m \cdot (\Xi_{\delta t_m} \nabla_{\mathcal{P}_m} z_m) \\
= \int_0^T \int_\Omega (q^+_{\delta t_m} - q^-_{\delta t_m}) \Xi_{\delta t_m} \Pi_{\mathcal{P}_m} z_m. \quad (6.35)
\end{aligned}$$

By symmetry of A, strong convergence of $\widetilde{\Pi}_{\mathcal{C}_m} c_m$ and of $\nabla_{\mathcal{P}_m} z_m$, together with the weak convergence of $\nabla_{\mathcal{P}_m} p_m$, a weak–strong convergence result (see, e.g., [40, Lemma D.8]) shows that the left-hand side of (6.35) converges to $\int_0^T \int_\Omega A(\boldsymbol{x}, c) \nabla p \cdot \Xi \nabla \eta$. Moreover, $q^{\pm}_{\delta t_m} \to q^{\pm}$ in $L^1(0,T; L^2(\Omega))$ and thus the right-hand side of (6.35) converges to $\int_0^T \int_\Omega (q^+ - q^-) \Xi \eta$. This shows that p satisfies the second equation in (6.2) when $\psi = \Xi\eta$. By linear combination, this equation is also satisfied for all tensorial functions and, by a density argument, for all smooth functions. Hence, p satisfies (6.2).

Step 2: *strong convergence of* $\nabla_{\mathcal{P}_m} p_m$ *and* $\mathbf{U}_{\mathcal{P}_m}$ Let $z = p_m^{(n+1)}$ in (5.1), multiply by $\delta t_m^{(n+\frac{1}{2})}$ and take the sum over $n = 0, \dots, N_m - 1$. By weak convergence of $\Pi_{\mathcal{P}_m} p_m$ and since p satisfies (6.2) (which also holds, by density, for $\psi \in L^1(0,T; H^1(\Omega))$),

$$\begin{aligned}
&\lim_{m\to\infty} \int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) \nabla_{\mathcal{P}_m} p_m \cdot \nabla_{\mathcal{P}_m} p_m \\
&= \lim_{m\to\infty} \int_0^T \int_\Omega (q^+_{\delta t_m} - q^-_{\delta t_m}) \Pi_{\mathcal{P}_m} p_m = \int_0^T \int_\Omega (q^+ - q^-) p = \int_0^T \int_\Omega A(\boldsymbol{x}, c) \nabla p \cdot \nabla p.
\end{aligned}$$

This convergence, the weak convergence of $\nabla_{\mathcal{P}_m} p_m$ and the strong convergence of $A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) \nabla p$ show that

$$\begin{aligned}
&\int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) (\nabla_{\mathcal{P}_m} p_m - \nabla p) \cdot (\nabla_{\mathcal{P}_m} p_m - \nabla p) \\
&\quad = \int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) \nabla_{\mathcal{P}_m} p_m \cdot \nabla_{\mathcal{P}_m} p_m \\
&\qquad - \int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) \nabla_{\mathcal{P}_m} p_m \cdot \nabla p \\
&\qquad - \int_0^T \int_\Omega A(\boldsymbol{x}, \widetilde{\Pi}_{\mathcal{C}_m} c_m) \nabla p \cdot (\nabla_{\mathcal{P}_m} p_m - \nabla p) \to 0.
\end{aligned}$$

By coercivity of A (Assumption (6.1c)), we infer that $\nabla_{\mathcal{P}_m} p_m \to \nabla p$ strongly in $L^2(\Omega \times (0,T))^d$. Moreover, since $\nabla_{\mathcal{P}_m} p_m$ is bounded in $L^\infty(0,T; L^2(\Omega))$ (Lemma 6.4.1), this implies that $\nabla_{\mathcal{P}_m} p_m \to \nabla p$ strongly in $L^r(0,T; L^2(\Omega))^d$ for any $r \in (1, \infty)$.

Up to a subsequence $\widetilde{\Pi}_{\mathcal{C}_m} c_m \to c$ a.e. on $\Omega \times (0,T)$. The properties (6.1c) of A and the above convergence of $\nabla_{\mathcal{P}_m} p_m$ show that

$$\mathbf{U}_{\mathcal{P}_m} = -\frac{\mathbf{K}}{\mu(\widetilde{\Pi}_{\mathcal{C}_m} c_m)} \nabla_{\mathcal{P}_m} p_m \to \mathbf{U} = -\frac{\mathbf{K}}{\mu(c)} \nabla p \text{ strongly in } L^r(0,T; L^2(\Omega))^d.$$

Step 3: *strong convergence of* $\Pi_{\mathcal{P}_m} p_m$ Since $p \in L^2(0,T;H^1(\Omega))$, by [40, Lemma 4.10] we can find $P_m \in X_{\mathcal{P}_m}^{N_m+1}$ such that $\Pi_{\mathcal{P}_m} P_m \to p$ and $\nabla_{\mathcal{P}_m} P_m \to \nabla p$ strongly in $L^2(0,T;L^2(\Omega))$. Then, for each $t \in (0,T)$, by definition of the coercivity constant $C_{\mathcal{P}_m}$,

$$\begin{aligned}&\|\Pi_{\mathcal{P}_m}(P_m - p_m)\|_{L^2(\Omega)}^2 \\ &\qquad \le C_{\mathcal{P}_m}^2 \left(\|\nabla_{\mathcal{P}_m}(P_m - p_m)\|_{L^2(\Omega)}^2 + \left| \int_\Omega \Pi_{\mathcal{P}_m}(P_m - p_m) \right|^2 \right).\end{aligned}$$

Integrating from 0 to T and using $\int_\Omega p = \int_\Omega \Pi_{\mathcal{P}_m} p_m = 0$ yields

$$\begin{aligned}&\|\Pi_{\mathcal{P}_m}(P_m - p_m)\|_{L^2(\Omega\times(0,T))}^2 \\ &\qquad \le C_{\mathcal{P}_m}^2 \|\nabla_{\mathcal{P}_m}(P_m - p_m)\|_{L^2(\Omega\times(0,T))^d}^2 + C_{\mathcal{P}_m}^2 \int_0^T \left| \int_\Omega (\Pi_{\mathcal{P}_m} P_m - p) \right|^2.\end{aligned}$$

The first term on the right hand side converges to 0 since both $\nabla_{\mathcal{P}_m} P_m$ and $\nabla_{\mathcal{P}_m} p_m$ converge strongly to ∇p (and $(C_{\mathcal{P}_m})_{m\in\mathbb{N}}$ is bounded by coercivity of $(\mathcal{P}_m)_{m\in\mathbb{N}}$). The second term converges to 0 since $\Pi_{\mathcal{P}_m} P_m$ converges to p strongly. This shows that $\Pi_{\mathcal{P}_m} p_m$ also converges strongly to p in this space, and the convergence in $L^r(0,T;L^2(\Omega))$ follows due to the bound on $\Pi_{\mathcal{P}_m} p_m$ in Lemma 6.4.1.

6.5.3 Convergence of the concentration

The proof of Theorem 6.1.1 is concluded by showing that c satisfies (6.3). It has already been established that $c \in L^2(0,T;H^1(\Omega))$. Lemma 6.4.2 shows that $(1+|\mathbf{U}_{\mathcal{P}_m}|)^{1/2}\nabla_{\mathcal{C}_m} c_m$ is bounded in $L^2(\Omega\times(0,T))^d$ and therefore weakly converges, up to a subsequence, in this space to some $\mathcal{W}$. Since $\mathbf{U}_{\mathcal{P}_m}$ converges strongly in $L^2(\Omega\times(0,T))^d$ and $\nabla_{\mathcal{C}_m} c \to \nabla c$ converges weakly in this space, $(1+|\mathbf{U}_{\mathcal{P}_m}|)^{1/2}\nabla_{\mathcal{C}_m} c_m \to (1+|\mathbf{U}|)^{1/2}\nabla c$ in the sense of distributions. Hence, $(1+|\mathbf{U}|)^{1/2}\nabla c = \mathcal{W} \in L^2(\Omega\times(0,T))^d$. It remains to prove that the equation in (6.3) is satisfied.

Take a test function $\varphi(\boldsymbol{x},t) = \Theta(t)\omega(\boldsymbol{x})$ with $\Theta \in C^\infty([0,T))$ and $\omega \in C^\infty(\overline{\Omega})$. For $m \in \mathbb{N}$ let $\Theta_{\delta t_m} : (0,T) \to \mathbb{R}$ be such that $\Theta_{\delta t_m} = \Theta(t^{(n+1)})$ on $(t^{(n)}, t^{(n+1)}]$ for all $n = 0,\dots,N_m-1$ (for legibility, we drop the index m in the time steps $t_m^{(k)}$). Using Assumption (**A3**), define the interpolant $z_m := \mathcal{J}_{\mathcal{C}_m}\omega$ of ω. Now, consider $z = \Theta(t^{(n+1)})z_m \in X_{\mathcal{C}_m}$ in (5.3), so that $v = v_m^{(n)}$ is given by $v_m^{(n)}(\boldsymbol{x},t^{(n)}) = \Theta(t^{(n+1)})\Pi_{\mathcal{C}_m} z_m(F^{(n+1)}_{t^{(n+1)}-t^{(n)}}(\boldsymbol{x}))$ (here, we make explicit the dependency on the flow $F_t^{(n+1)}$ with respect to the time step n, but not with

respect to m). Sum the resulting equations over $n = 0, \dots, N_m - 1$ and recall the definition (6.12) of $\mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}}$. Letting $q^{\pm}_{\delta t_m}$ (resp. $\hat{q}^{\pm}_{\delta t_m}$, resp. $\varpi_{\delta t_m}$) be the function equal to $q^{\pm}_n$ (resp. $q^{\pm}_{n+1}$, resp. ϖ_n) on $(t^{(n)}, t^{(n+1)})$ for all $n = 0, \dots, N_m - 1$, we obtain

$$\begin{aligned}
&\Big[\sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n+1)} \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m - \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} v_m^{(n)}(t^{(n)}) \Big] \\
&\quad + \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}_m}) \nabla_{\mathcal{C}_m} c_m \cdot \Theta_{\delta t_m}(t) \nabla_{\mathcal{C}_m} z_m \\
&\quad + \int_0^T \int_\Omega \Big[\varpi_{\delta t_m} \widetilde{\Pi}_{\mathcal{C}_m} c_m \mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}} [\Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m] q^-_{\delta t_m} \\
&\qquad\qquad + (1 - \varpi_{\delta t_m}) \Pi_{\mathcal{C}_m} c_m \Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m \hat{q}^-_{\delta t_m} \Big] \\
&\quad = \int_0^T \int_\Omega \left[\varpi_{\delta t_m} \mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}} [\Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m] q^+_{\delta t_m} + (1 - \varpi_{\delta t_m}) \hat{q}^+_{\delta t_m} \Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m \right].
\end{aligned}$$

Let us write $T_1^{(m)} + T_2^{(m)} + T_3^{(m)} = T_4^{(m)}$ this relation.

The limit of $T_2^{(m)}$ is the easiest to establish. Since $\mathbf{U}_{\mathcal{P}_m} \to \mathbf{U}$ strongly in $L^2(\Omega \times (0,T))^d$, the growth assumption (6.1d) on $\mathbf{D}$ ensures that (see, e.g., [43, Lemma A.1])

$$\mathbf{D}(\cdot, \mathbf{U}_{\mathcal{P}_m})^{1/2} \to \mathbf{D}(\cdot, \mathbf{U})^{1/2} \text{ strongly in } L^4(\Omega \times (0,T))^{d \times d}. \tag{6.36}$$

By Lemma 6.4.2 the sequence $\mathbf{D}(\cdot, \mathbf{U}_{\mathcal{P}_m})^{1/2} \nabla_{\mathcal{C}_m} c_m$ is bounded in $L^2(\Omega \times (0,T))^d$. The weak convergence of $\nabla_{\mathcal{C}_m} c_m$ in $L^2(\Omega \times (0,T))^d$ and [43, Lemma A.3] thus show that $\mathbf{D}(\cdot, \mathbf{U}_{\mathcal{P}_m})^{1/2} \nabla_{\mathcal{C}_m} c_m \to \mathbf{D}(\cdot, \mathbf{U})^{1/2} \nabla c$ weakly in $L^2(\Omega \times (0,T))^d$. Using (6.36) and the fact that $\Theta_{\delta t_m} \to \Theta$ uniformly, the strong convergence $\nabla_{\mathcal{C}_m} z_m \to \nabla \omega$ in $L^4(\Omega)^d$ (see (**A3**)) shows that, as $m \to \infty$,

$$\begin{aligned}
T_2^{(m)} &= \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}_m})^{1/2} \nabla_{\mathcal{C}_m} c_m \cdot \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}_m})^{1/2} \Theta_{\delta t_m}(t) \nabla_{\mathcal{C}_m} z_m \\
&\to \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U})^{1/2} \nabla c \cdot \mathbf{D}(\boldsymbol{x}, \mathbf{U})^{1/2} \nabla \varphi = \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}) \nabla c \cdot \nabla \varphi.
\end{aligned} \tag{6.37}$$

We now turn to $T_3^{(m)}$. Let

$$\begin{aligned}
T_{3,\star}^{(m)} = \int_0^T \int_\Omega c \varphi q^-_{\delta t_m} = &\int_0^T \int_\Omega \left[\varpi_{\delta t_m} c \varphi q^-_{\delta t_m} + (1 - \varpi_{\delta t_m}) c \varphi \hat{q}^-_{\delta t_m} \right] \\
&+ \int_0^T \int_\Omega (1 - \varpi_{\delta t_m}) c \varphi (q^-_{\delta t_m} - \hat{q}^-_{\delta t_m}).
\end{aligned}$$

By construction of $q^-_{\delta t_m}$ we easily see that $T^{(m)}_{3,\star} \to \int_0^T\!\!\int_\Omega c\varphi q^-$ as $m \to \infty$. Since $\varpi_{\delta t_m}$ takes its values in $[0,1]$, we can write

$$\begin{aligned}|T^{(m)}_3 - T^{(m)}_{3,\star}| \le M_{q^-} \int_0^T\!\!\int_\Omega \left|\widetilde{\Pi}_{\mathcal{C}_m} c_m \mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}}[\Theta_{\delta t_m}(t)\Pi_{\mathcal{C}_m} z_m] - c\varphi\right| \\ + M_{q^-} \int_0^T\!\!\int_\Omega |\Pi_{\mathcal{C}_m} c_m \Theta_{\delta t_m}(t)\Pi_{\mathcal{C}_m} z_m - c\varphi| \\ + \int_0^T\!\!\int_\Omega |c\varphi|\,|q^-_{\delta t_m} - \hat{q}^-_{\delta t_m}| =: T^{(m)}_{3,1} + T^{(m)}_{3,2} + T^{(m)}_{3,3}\end{aligned}$$

Together with Lemma 6.2.5, the strong convergences in $L^2(\Omega\times(0,T))$ of $\widetilde{\Pi}_{\mathcal{C}_m} c_m$, $\Pi_{\mathcal{C}_m} c_m$ and $\Theta_{\delta t_m}\Pi_{\mathcal{C}_m} z_m$ show that

$$\begin{aligned}\widetilde{\Pi}_{\mathcal{C}_m} c_m \mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}}[\Theta_{\delta t_m}(t)\Pi_{\mathcal{C}_m} z_m] \to c\varphi \text{ in } L^1(\Omega\times(0,T)), \\ \Pi_{\mathcal{C}_m} c_m \Theta_{\delta t_m}(t)\Pi_{\mathcal{C}_m} z_m \to c\varphi \text{ in } L^1(\Omega\times(0,T)).\end{aligned}$$

Hence, $T^{(m)}_{3,1} + T^{(m)}_{3,2} \to 0$. By continuity of the translations in $L^2(\Omega\times(0,T))$, $q^-_{\delta t_m} - \hat{q}^-_{\delta t_m} \to 0$ in this space and since $c\varphi \in L^2(\Omega\times(0,T))$ we deduce that $T^{(m)}_{3,3} \to 0$. This shows that $T^{(m)}_3 - T^{(m)}_{3,\star} \to 0$ as $m\to\infty$ and thus that

$$T^{(m)}_3 \to \int_0^T \int_\Omega c\varphi q^-. \tag{6.38}$$

A similar reasoning yields

$$T^{(m)}_4 \to \int_0^T \int_\Omega q^+\varphi. \tag{6.39}$$

We finally consider $T^{(m)}_1$. Since $\Theta(t^{(N_m)}) = 0$, a change of index in the

first sum of $T_1^{(m)}$ and recalling the definition of $v_m^{(n)}(t^{(n)})$ yield

$$\begin{aligned}
T_1^{(m)} &= \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n)}) \Pi_{\mathcal{C}_m} z_m - \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \Theta(t^{(0)}) \Pi_{\mathcal{C}_m} z_m \\
&\quad - \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m \Big(F^{(n+1)}_{\delta t_m^{(n+1/2)}}(\boldsymbol{x}) \Big) \\
&= \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} (\Theta(t^{(n)}) - \Theta(t^{(n+1)})) \Pi_{\mathcal{C}_m} z_m \\
&\quad - \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \Theta(t^{(0)}) \Pi_{\mathcal{C}_m} z_m \\
&\quad - \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n+1)}) \left(\Pi_{\mathcal{C}_m} z_m \Big(F^{(n+1)}_{\delta t_m^{(n+1/2)}}(\boldsymbol{x}) \Big) - \Pi_{\mathcal{C}_m} z_m \right) \\
&= T_{11}^{(m)} - T_{12}^{(m)} - T_{13}^{(m)}.
\end{aligned}$$

Since $c_m^{(0)} = \mathcal{I}_{\mathcal{C}_m} c_{\mathrm{ini}}$, the consistency of $(\mathcal{C}_m)_{m\in\mathbb{N}}$ (see Definition 2.1.3) ensures that

$$T_{12}^{(m)} \to \int_\Omega \phi c_{\mathrm{ini}} \Theta(0) \omega = \int_\Omega \phi c_{\mathrm{ini}} \varphi(\cdot, 0). \tag{6.40}$$

Since $\Theta(t^{(n)}) - \Theta(t^{(n+1)}) = -\int_{t^{(n)}}^{t^{(n+1)}} \Theta'$ the strong convergences of $\Pi_{\mathcal{C}_m} z_m$ and $\widetilde{\Pi}_{\mathcal{C}_m} c_m$ show that

$$T_{11}^{(m)} = -\int_0^T \int_\Omega \phi \widetilde{\Pi}_{\mathcal{C}_m} c_m \Theta' \Pi_{\mathcal{C}_m} z_m \to -\int_0^T \int_\Omega \phi c \frac{\partial \varphi}{\partial t}. \tag{6.41}$$

It remains to analyse $T_{13}^{(m)}$. Let $\zeta_m = \Pi_{\mathcal{C}_m} z_m - \omega$ and write

$$\Pi_{\mathcal{C}_m} z_m (F^{(n+1)}_{\delta t^{(n+1/2)}}) - \Pi_{\mathcal{C}_m} z_m = \Big(\omega(F^{(n+1)}_{\delta t^{(n+1/2)}}) - \omega \Big) + \zeta_m (F^{(n+1)}_{\delta t^{(n+1/2)}}) - \zeta_m.$$

Letting I be the identity map and $\kappa(t)$ be the piecewise-constant function equal to $\delta t_m^{(n+\frac{1}{2})}$ on $(t^{(n)}, t^{(n+1)})$, this yields

$$\begin{aligned}
T_{13}^{(m)} &= \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n+1)}) \left(\omega \Big(F^{(n+1)}_{\delta t_m^{(n+1/2)}} \Big) - \omega \right) \\
&\quad + \int_{t^{(1)}}^T \int_\Omega \frac{\phi}{\kappa(t)} \widetilde{\Pi}_{\mathcal{C}_m} c_m \left(\mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}} - \mathrm{I} \right) [\Theta_{\delta t_m}(t) \zeta_m] \\
&\quad + \Theta(t^{(1)}) \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \left(\zeta_m (F^{(1)}_{t^{(1)}}) - \zeta_m \right).
\end{aligned}$$

We note that, in the last two terms, the case $n > 0$ is separated from the case $n = 0$, as we do not have any information regarding the boundedness of $\nabla_{\mathcal{C}_m} c_m^{(0)}$ (which would arise in the estimates after invoking (**A2**)). For a.e. $\boldsymbol{x} \in \Omega$, $t \mapsto F_t^{(n+1)}(\boldsymbol{x})$ is Lipschitz-continuous and the chain rule therefore yields

$$\begin{aligned}\omega(F^{(n+1)}_{\delta t_m^{(n+1/2)}}(\boldsymbol{x})) - \omega(\boldsymbol{x}) &= -\int_{t^{(n)}}^{t^{(n+1)}} \partial_t \left[\omega(F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{x}))\right] \\ &= \int_{t^{(n)}}^{t^{(n+1)}} \nabla\omega(F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{x})) \cdot \frac{\mathbf{u}^{(n+1)}_{\mathcal{P}_m}(F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{x}))}{\phi((F^{(n+1)}_{t^{(n+1)}-t}(\boldsymbol{x})))}. \quad (6.42)\end{aligned}$$

The operator $\mathcal{T}_{\mathbf{u}_{\mathcal{P}_m}}$ does not directly act on the time component in $L^2(\Omega \times (0,T))$. Hence, the representation (6.14) of its dual is also valid in $L^2(\Omega \times (t^{(1)},T))$, and space-independent functions can be taken out of these operators. Using this representation, (6.42) and recalling the definition (6.12) of $\widehat{\mathcal{T}}_{\mathbf{u}_{\mathcal{P}_m}}$, we obtain

$$\begin{aligned}T_{13}^{(m)} = &\int_0^T \int_\Omega \phi \widetilde{\Pi}_{\mathcal{C}_m} c_m \widehat{\mathcal{T}}_{\mathbf{u}_{\mathcal{P}_m}} \left[\Theta_{\delta t_m}(t) \nabla\omega \cdot \frac{\mathbf{u}_{\mathcal{P}_m}}{\phi}\right] \\ &+ \int_{t^{(1)}}^T \int_\Omega \frac{\phi}{\kappa(t)} \left(\mathcal{T}_{-\mathbf{u}_{\mathcal{P}_m}} - \mathrm{I}\right)(\widetilde{\Pi}_{\mathcal{C}_m} c_m)\Theta_{\delta t_m}(t)\zeta_m \\ &+ \int_{t^{(1)}}^T \int_\Omega \frac{R_m}{\kappa(t)} \mathcal{T}_{-\mathbf{u}_{\mathcal{P}_m}}(\phi\widetilde{\Pi}_{\mathcal{C}_m} c_m)\Theta_{\delta t_m}(t)\zeta_m \\ &+ \Theta(t^{(1)}) \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \left(\zeta_m(F^{(1)}_{t^{(1)}}) - \zeta_m\right) = T_{131}^{(m)} + \cdots + T_{134}^{(m)}.\end{aligned} \quad (6.43)$$

By weak convergence of $\Theta_{\delta t_m}(t)\nabla\omega \cdot \mathbf{u}_{\mathcal{P}_m}/\phi$ (owing to (b) in (**A5**)) and strong convergence of $\widetilde{\Pi}_{\mathcal{C}_m} c_m$, Lemma 6.2.5 shows that $T_{131}^{(m)} \to \int_0^T \int_\Omega c\mathbf{u} \cdot \Theta\nabla\omega = \int_0^T \int_\Omega c\mathbf{u} \cdot \nabla\varphi$. Using (**A2**) we have, for $n = 1, \ldots, N_m - 1$,

$$\frac{\left\|\Pi_{\mathcal{C}_m} c_m^{(n)}(F^{(n+1)}_{-\delta t_m^{(n+1/2)}}) - \Pi_{\mathcal{C}_m} c_m^{(n)}\right\|_{L^1(\Omega)}}{\delta t_m^{(n+1/2)}} \le M_F \left\|\mathbf{u}_{\mathcal{P}_m}^{(n+1)}\right\|_{L^2(\Omega)} \left\|\nabla_{\mathcal{C}_m} c_m^{(n)}\right\|_{L^2(\Omega)}.$$

Hence, invoking (**A1**),

$$\begin{aligned}|T_{132}^{(m)}| \le \phi^* M_F \|\zeta_m\|_{L^\infty(\Omega)} \|\Theta\|_{L^\infty(0,T)} \\ \times \sum_{n=1}^{N_m-1} \delta t_m^{(n+1/2)} \left\|\mathbf{u}_{\mathcal{P}_m}^{(n+1)}\right\|_{L^2(\Omega)} \left\|\nabla_{\mathcal{C}_m} c_m^{(n)}\right\|_{L^2(\Omega)}\end{aligned}$$

$$\leq \phi^* M_F \|\zeta_m\|_{L^\infty(\Omega)} \|\Theta\|_{L^\infty(0,T)} \left\|\mathbf{u}_{\mathcal{P}_m}\right\|_{L^\infty(0,T;L^2(\Omega))} \times M_t \sum_{n=0}^{N_m-2} \delta t_m^{(n+1/2)} \left\|\nabla_{\mathcal{C}_m} c_m^{(n+1)}\right\|_{L^2(\Omega)}$$
$$\leq \phi^* M_F \|\zeta_m\|_{L^\infty(\Omega)} \|\Theta\|_{L^\infty(0,T)} \left\|\mathbf{u}_{\mathcal{P}_m}\right\|_{L^\infty(0,T;L^2(\Omega))} \|\nabla_{\mathcal{C}_m} c_m\|_{L^1(0,T;L^2(\Omega))}.$$

Using the bounds on $\mathbf{u}_{\mathcal{P}_m}$ and $\nabla_{\mathcal{C}_m} c_m$ given by (a) in (**A5**) and Lemmas 6.4.1 and 6.4.2, and the convergence $\zeta_m = \Pi_{\mathcal{C}_m} \mathcal{J}_{\mathcal{C}_m} \omega - \omega \to 0$ in $L^\infty(\Omega)$ from (**A3**), we infer that $T_{132}^{(m)} \to 0$. The term $T_{133}^{(m)}$ also converges to 0, due to the bound on R_m in Lemma 6.2.5 (which cancels out the term $1/\kappa(t)$), the bound (6.13) and the convergence of ζ_m to 0 in $L^\infty(\Omega)$.

Finally, let us study $T_{134}^{(m)}$. Since $\Pi_{\mathcal{C}_m} c_m^{(0)} = \Pi_{\mathcal{C}_m} \mathcal{I}_{\mathcal{C}_m} c_{\text{ini}}$ is bounded in $L^\infty(\Omega)$ (see Definition 2.1.3), there is C not depending on m such that $|\Theta(t^{(1)}) \Pi_{\mathcal{C}_m} c_m^{(0)}| \leq C$ a.e. on Ω. Split $\zeta_m = \Pi_{\mathcal{C}_m} z_m - \omega$ and write, using (**A2**) on z_m and Lemma 6.2.1 on ω,

$$|T_{134}^{(m)}| \leq C \left(\left\| \Pi_{\mathcal{C}_m} z_m(F_{t^{(1)}}^{(1)}) - \Pi_{\mathcal{C}_m} z_m \right\|_{L^1(\Omega)} + \left\| \omega(F_{t^{(1)}}^{(1)}) - \omega \right\|_{L^1(\Omega)} \right)$$
$$\leq C \left\| \mathbf{u}_{\mathcal{P}_m}^{(1)} \right\|_{L^2(\Omega)} |\delta t_m^{(1)}| \left(M_F \|\nabla_{\mathcal{C}_m} z_m\|_{L^2(\Omega)} + \frac{C_1(T)}{\phi_*} \|\nabla \omega\|_{L^2(\Omega)} \right).$$

The bounds on $\mathbf{u}_{\mathcal{P}}^{(1)}$ (from (a) in (**A5**) and Lemma 6.4.1) and on $\nabla_{\mathcal{C}_m} z_m$ (from (**A3**)) then show that $T_{134}^{(m)} \to 0$.

Hence, $T_{13}^{(m)} \to \int_0^T \int_\Omega c\mathbf{u} \cdot \nabla \varphi$. Together with (6.40) and (6.41), this shows that

$$T_1^{(m)} \to -\int_0^T \int_\Omega \phi c \frac{\partial \varphi}{\partial t} - \int_\Omega \phi c_{\text{ini}} \varphi(\cdot, 0) - \int_0^T \int_\Omega c\mathbf{u} \cdot \nabla \varphi.$$

Gathering this with (6.37), (6.38) and (6.39), we infer that c satisfies the equation in (6.3) whenever $\varphi = \Theta\omega$. By linear combination, this equation is also satisfied for all tensorial functions and, by density argument, for all smooth functions. This concludes the proof that c satisfies (6.3).

6.6 Outline of the proof of the main theorem (GDM–MMOC)

In this section, we now outline the proof for the convergence of the GDM–MMOC scheme. Firstly, we look at the a priori estimates. If (p_m, c_m) is a solution to the GDM–MMOC scheme with $(\mathcal{P}, \mathcal{C}^T) = (\mathcal{P}_m, \mathcal{C}_m^T)$, then the a priori estimates in Lemmas 6.4.2 and 6.4.4 hold true. This can be established

by using arguments that are very similar to those in the proofs of Lemmas 6.4.2 and 6.4.4. For the proof of the main theorem, we only modify some parts of the proof in Section 6.5.3.

We follow the notations of the proof of the GDM–ELLAM scheme in Section 6.5.3. Take a test function $\varphi(\boldsymbol{x}, t) = \Theta(t)\omega(\boldsymbol{x})$ with $\Theta \in C^\infty([0,T))$ and $\omega \in C^\infty(\overline{\Omega})$. For $m \in \mathbb{N}$ let $\Theta_{\delta t_m} : (0,T) \to \mathbb{R}$ be such that $\Theta_{\delta t_m} = \Theta(t^{(n+1)})$ on $(t^{(n)}, t^{(n+1)}]$ for all $n = 0, \ldots, N_m - 1$ (for legibility, we drop the index m in the time steps $t_m^{(k)}$). Using Assumption (**A3**), set $z_m := \mathcal{J}_{\mathcal{C}_m}\omega$ of ω. Use $z = \Theta(t^{(n+1)})z_m \in X_{\mathcal{C}_m}$ in (5.3) and sum the resulting equations over $n = 0, \ldots, N_m - 1$.

$$
\begin{aligned}
&\sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n+1)} \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} \\
&- \sum_{n=0}^{N_m-1} \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)} \left(F^{(n+1)}_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x}) \right) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m(\boldsymbol{x}) d\boldsymbol{x} \\
&+ \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}_m}) \nabla_{\mathcal{C}_m} c_m \cdot \Theta_{\delta t_m}(t) \nabla_{\mathcal{C}_m} z_m \\
&\quad = \int_0^T \int_\Omega \left[\left(q^+(1 - \Pi_{\mathcal{C}_m} c_m) \right)^{(n,w)} \cdot \mathbf{e} \right] \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} dt,
\end{aligned}
\tag{6.44}
$$

Using an argument similar to those for $T_2^{(m)}, T_3^{(m)}$ and $T_4^{(m)}$ in Section 6.5.3, it can be shown that as $m \to \infty$,

$$
\int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}_{\mathcal{P}_m}) \nabla_{\mathcal{C}_m} c_m \cdot \Theta_{\delta t_m}(t) \nabla_{\mathcal{C}_m} z_m \to \int_0^T \int_\Omega \mathbf{D}(\boldsymbol{x}, \mathbf{U}) \nabla c \cdot \nabla \varphi,
$$

and also, the right hand side of (6.44) converges to $\int_0^T \int_\Omega (q^+(1-c))\varphi$.

We now deal with the remaining terms on the left hand side of (6.44), which we refer to as $R_1^{(m)}$. By performing a change of index and noting that

$\Theta(t^{(N_m)}) = 0$, we have

$$
\begin{aligned}
R_1^{(m)} = & \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} - \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \Theta(t^{(0)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} \\
& - \sum_{n=0}^{N_m-1} \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)}(\boldsymbol{x}) \big(F^{(n+1)}_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m(\boldsymbol{x}) d\boldsymbol{x} \\
= & \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \left[\Theta(t^{(n)}) - \Theta(t^{(n+1)})\right] \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} \\
& - \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \Theta(t^{(0)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} + \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x} \\
& - \sum_{n=0}^{N_m-1} \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)} \big(F^{(n+1)}_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m(\boldsymbol{x}) d\boldsymbol{x}.
\end{aligned}
$$

Introducing $\pm \Pi_{\mathcal{C}_m} z_m\big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big)$ in the third summand, we obtain $R_1^{(m)} = R_{11}^{(m)} - R_{12}^{(m)} + R_{13}^{(m)} + R_{14}^{(m)}$ with

$$
R_{11}^{(m)} = \sum_{n=0}^{N_m-1} \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(n)} \left[\Theta(t^{(n)}) - \Theta(t^{(n+1)})\right] \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x},
$$

$$
R_{12}^{(m)} = \int_\Omega \phi \Pi_{\mathcal{C}_m} c_m^{(0)} \Theta(t^{(0)}) \Pi_{\mathcal{C}_m} z_m d\boldsymbol{x},
$$

$$
R_{13}^{(m)} = \sum_{n=0}^{N_m-1} \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)}(\boldsymbol{x}) \Theta(t^{(n+1)}) \left[\Pi_{\mathcal{C}_m} z_m(\boldsymbol{x}) - \Pi_{\mathcal{C}_m} z_m\big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big)\right] d\boldsymbol{x},
$$

and

$$
\begin{aligned}
R_{14}^{(m)} = \sum_{n=0}^{N_m-1} \Bigg[& \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)}(\boldsymbol{x}) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m \big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) d\boldsymbol{x} \\
& - \int_\Omega \phi(\boldsymbol{x}) \Pi_{\mathcal{C}_m} c_m^{(n)} \big(F^{(n+1)}_{-\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m(\boldsymbol{x}) d\boldsymbol{x} \Bigg].
\end{aligned}
$$

Using the same arguments as those of $T_{11}^{(m)}, T_{12}^{(m)}, T_{13}^{(m)}$ and $T_{14}^{(m)}$ in Section 6.5.3, we deduce that as $m \to \infty$,

$$
R_{11}^{(m)} \to - \int_0^T \int_\Omega \phi(\boldsymbol{x}) c(\boldsymbol{x},t) \frac{\partial \varphi}{\partial t}(\boldsymbol{x},t) d\boldsymbol{x} dt\,,\ R_{12}^{(m)} \to \int_\Omega \phi(\boldsymbol{x}) c_{\text{ini}}(\boldsymbol{x}) \varphi(\boldsymbol{x},0) d\boldsymbol{x}
$$

$$
\text{and } R_{13}^{(m)} \to - \int_0^T \int_\Omega c(\boldsymbol{x},t) \mathbf{u}(\boldsymbol{x},t) \cdot \nabla \varphi(\boldsymbol{x},t) d\boldsymbol{x} dt.
$$

We deal with $R_{14}^{(m)}$ by performing a change of variables to obtain, owing to (4.8),

$$\begin{aligned}
R_{14}^{(m)} &= \sum_{n=0}^{N_m-1} \int_\Omega \Pi_{\mathcal{C}_m} c_m^{(n)}(\boldsymbol{x}) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m\big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \\
&\qquad \times \Big[\phi(\boldsymbol{x}) - \phi\big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \big| JF^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big|\Big] \, d\boldsymbol{x} \\
&= -\sum_{n=0}^{N_m-1} \int_\Omega \Pi_{\mathcal{C}_m} c_m^{(n)}(\boldsymbol{x}) \Theta(t^{(n+1)}) \Pi_{\mathcal{C}_m} z_m\big(F^{(n+1)}_{\delta t^{(n+\frac{1}{2})}}(\boldsymbol{x})\big) \\
&\qquad \times \int_0^{\delta t^{(n+\frac{1}{2})}} \big| JF_t^{(n+1)}(\boldsymbol{x})\big| \big(\mathrm{div}\mathbf{u}^{(n+1)}_{\mathcal{P}_m}\big) \circ F_t^{(n+1)}(\boldsymbol{x}) dt d\boldsymbol{x} \\
&= -\int_\Omega \int_0^T \widetilde{\Pi}_{\mathcal{C}_m} c_m \mathcal{T}_m(\Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m) g_m dt d\boldsymbol{x}.
\end{aligned}$$

Setting κ_m to be a piecewise constant function in time defined by $\kappa_m(t) = \delta t^{(n+\frac{1}{2})}$ for all $t \in (t^{(n)}, t^{(n+1)})$ and all $n = 0, \ldots, N_m - 1$, we write

$$\begin{aligned}
g_m(\boldsymbol{x}, t) &:= \frac{\int_0^{\kappa_m(t)} \big| JF_s^{(n+1)}(\boldsymbol{x})\big| \big(\mathrm{div}\mathbf{u}_{\mathcal{P}_m}(\cdot, t)\big) \circ F_s^{(n+1)}(\boldsymbol{x}) ds}{\kappa_m(t)} \\
&= \int_0^1 \big| JF^{(n+1)}_{s\kappa_m(t)}(\boldsymbol{x})\big| \big(\mathrm{div}\mathbf{u}_{\mathcal{P}_m}(\cdot, t)\big) \circ F^{(n+1)}_{s\kappa_m(t)}(\boldsymbol{x}) ds.
\end{aligned}$$

As $m \to \infty$, $\widetilde{\Pi}_{\mathcal{C}_m} c_m \to c$ strongly in $L^2(\Omega \times (0,T))$, and $\mathcal{T}_m(\Theta_{\delta t_m}(t) \Pi_{\mathcal{C}_m} z_m) \to \varphi$ strongly in $L^2(\Omega \times (0,T))$. We then want to establish that $g_m \to \mathrm{div}\mathbf{u}$ in $L^\infty(\Omega \times (0,T))$ weak-$*$.

By Assumptions (**A4**) and (6.1a), $(g_m - \mathrm{div}\mathbf{u})_{m \in \mathbb{N}}$ is bounded in $L^\infty(\Omega \times (0,T))$. Therefore, its weak-$*$ convergence in $L^\infty(\Omega \times (0,T))$ only has to be established against smooth test functions because of the density of $C_c^\infty(\Omega \times (0,T))$ in $L^1(\Omega \times (0,T))$. For any $\zeta \in C_c^\infty(\Omega \times (0,T))$ we have, performing a change of variables, performing an integration by parts, using (4.3b), and noticing that $0 \le \kappa_m(t) \le \max_{n=0,\ldots,N_m-1} \delta t^{(n+\frac{1}{2})}$ for all $t \in (0,T)$,

$$\begin{aligned}
&\Big| \int_0^T \int_\Omega \big(g_m(\boldsymbol{x}, t) - \mathrm{div}\mathbf{u}(\boldsymbol{x}, t)\big) \zeta(\boldsymbol{x}, t) d\boldsymbol{x} dt \Big| \\
&\quad \le \Big| \int_0^T \int_\Omega \big(g_m(\boldsymbol{x}, t) - \mathrm{div}\mathbf{u}_{\mathcal{P}_m}(\boldsymbol{x}, t)\big) \zeta(\boldsymbol{x}, t) d\boldsymbol{x} dt \Big| \\
&\qquad + \Big| \int_0^T \int_\Omega \big(\mathrm{div}\mathbf{u}_{\mathcal{P}_m}(\boldsymbol{x}, t) - \mathrm{div}\mathbf{u}(\boldsymbol{x}, t)\big) \zeta(\boldsymbol{x}, t) d\boldsymbol{x} dt \Big|
\end{aligned}$$

$$
\begin{aligned}
&\le \left| \int_0^T \int_\Omega \mathrm{div}\mathbf{u}_{\mathcal{P}_m}(\boldsymbol{x},t) \int_0^1 \left[\zeta(F_{-s\kappa_m(t)}(\boldsymbol{x}),t) - \zeta(\boldsymbol{x},t)\right] ds d\boldsymbol{x} dt \right| \\
&\quad + \left| \int_0^T \int_\Omega \left(\mathbf{u}_{\mathcal{P}_m}(\boldsymbol{x},t) - \mathbf{u}(\boldsymbol{x},t)\right) \cdot \nabla\zeta(\boldsymbol{x},t) d\boldsymbol{x} dt \right| \\
&\lesssim \|\nabla\zeta\|_{L^\infty(\Omega\times(0,T))} \max_{n=0,\dots,N_m-1} \delta t^{(n+\frac{1}{2})} \\
&\quad + \left| \int_0^T \int_\Omega \left(\mathbf{u}_{\mathcal{P}_m}(\boldsymbol{x},t) - \mathbf{u}(\boldsymbol{x},t)\right) \cdot \nabla\zeta(\boldsymbol{x},t) d\boldsymbol{x} dt \right|.
\end{aligned}
$$

By the consistency of $(\mathcal{C}_m)_{m\in\mathbb{N}}$, and by the weak convergence of $\mathbf{u}_{\mathcal{P}_m}$ to $\mathbf{u}$ in $L^2(\Omega\times(0,T))^d$, the last quantity tends to 0 as $m\to\infty$. This proves that $g_m \to \mathrm{div}\mathbf{u}$ weakly-$*$ in $L^\infty(\Omega\times(0,T))$ as required.

The convergences of $\widetilde{\Pi}_{\mathcal{C}_m} c_m$, $\mathcal{T}_m(\Theta_{\delta t_m}(t)\Pi_{\mathcal{C}_m} z_m)$ and g_m show that

$$
R_{14}^{(m)} \to -\int_\Omega \int_0^T c\varphi \mathrm{div}(\mathbf{u}) \quad \text{as } m\to\infty.
$$

The proof that c is a solution of (6.3) is complete by gathering the convergences of $R_{11}^{(m)}$, $R_{12}^{(m)}$, $R_{13}^{(m)}$ and $R_{14}^{(m)}$ into $R_1^{(m)} = R_{11}^{(m)} - R_{12}^{(m)} + R_{13}^{(m)} + R_{14}^{(m)}$, and by plugging the resulting convergence in (6.44) (in which we recall that $R_1^{(m)}$ is the sum of the first two terms).

The convergence for the GEM scheme is then established by combining the proofs of the GDM–ELLAM and GDM–MMOC.

6.7 Generic compactness results

The following results are particular cases of more general theorems on the GDM that can be found in [40].

Lemma 6.7.1 (Regularity of the limit, space-time problems [40, Lemma 4.8])**.**
Let $p\in(1,\infty)$, and $((\mathcal{D}^T)_m)_{m\in\mathbb{N}}$ be a coercive and limit-conforming sequence of space-time GDs. For each $m\in\mathbb{N}$, take $u_m \in X_{\mathcal{D}_m}^{N_m+1}$ (identified with a piecewise-constant function $[0,T]\to X_{\mathcal{D}_m}$) and assume that $(\|u_m\|_{L^p(0,T;X_{\mathcal{D}_m})})_{m\in\mathbb{N}}$ is bounded. Then there exists $u\in L^p(0,T;H^1(\Omega))$ such that, up to a subsequence as $m\to\infty$, $\Pi_{\mathcal{D}_m}u_m \to u$ and $\nabla_{\mathcal{D}_m}u_m \to \nabla u$ weakly in $L^p(0,T;L^2(\Omega))$. The same property holds with $p=+\infty$, provided that the weak convergences are replaced by weak-$*$ convergences.

Definition 6.7.2 (Compactly–continuously embedded sequence)**.** *Let $(X_m, \|\cdot\|_{X_m})_{m\in\mathbb{N}}$ be a sequence of Banach spaces included in $L^2(\Omega)$, and $(Y_m, \|\cdot\|_{Y_m})_{m\in\mathbb{N}}$ be*

a sequence of Banach spaces. The sequence $(X_m, Y_m)_{m\in\mathbb{N}}$ is compactly–continuously embedded in $L^2(\Omega)$ if:

1. *If $u_m \in X_m$ for all $m \in \mathbb{N}$ and $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ is bounded, then $(u_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(\Omega)$.*
2. *$X_m \subset Y_m$ for all $m \in \mathbb{N}$ and for any sequence $(u_m)_{m\in\mathbb{N}}$ such that*
   (a) *$u_m \in X_m$ for all $m \in \mathbb{N}$ and $(\|u_m\|_{X_m})_{m\in\mathbb{N}}$ is bounded,*
   (b) *$\|u_m\|_{Y_m} \to 0$ as $m \to \infty$,*
   (c) *$(u_m)_{m\in\mathbb{N}}$ converges in $L^2(\Omega)$,*

   it holds that $u_m \to 0$ in $L^2(\Omega)$.

Theorem 6.7.3 (Discrete Aubin–Simon compactness [40, Theorem C.8])**.** *Let $(X_m, Y_m)_{m\in\mathbb{N}}$ be compactly–continuously embedded in $L^2(\Omega)$, $T > 0$ and $(f_m)_{m\in\mathbb{N}}$ be a sequence in $L^2(0,T;L^2(\Omega))$ such that*

- *For all $m \in N$, there exists $N \in \mathbb{N}^*$, $0 = t^{(0)} < \cdots < t^{(N)} = T$ and $(v^{(n)})_{n=0,\dots,N} \in X_m^{N+1}$ such that $f_m(t) = v^{(n+1)}$ for all $n = 0, \dots, N-1$ and a.e. $t \in (t^{(n)}, t^{(n+1)})$, $f_m(t) = v^{(n+1)}$. We then set*

$$\delta_m f_m(t) = \frac{v^{(n+1)} - v^{(n)}}{t^{(n+1)} - t^{(n)}} \text{ for } n = 0, \dots, N-1 \text{ and } t \in (t^{(n)}, t^{(n+1)}).$$

- *The sequence $(f_m)_{m\in\mathbb{N}}$ is bounded in $L^2(0,T;L^2(\Omega))$.*
- *The sequence $(\|f_m\|_{L^2(0,T;X_m)})_{m\in\mathbb{N}}$ is bounded.*
- *The sequence $(\|\delta_m f_m\|_{L^2(0,T;Y_m)})_{m\in\mathbb{N}}$ is bounded.*

Then $(f_m)_{m\in\mathbb{N}}$ is relatively compact in $L^2(0,T;L^2(\Omega))$.

Chapter 7

Conclusion

We developed a family of characteristic-based schemes for a coupled model of miscible fluid flow in porous media, applicable on generic meshes, involved for example in tertiary oil recovery. The diffusive terms were discretised in the generic framework of the gradient discretisation method (GDM), whereas the advective terms were discretised by characteristic-based schemes, such as the Eulerian Lagrangian Localised Adjoint Method (ELLAM) and the Modified Method of Characteristics (MMOC). We started by giving a short summary of the GDM for Neumann boundary conditions in Chapter 2. In particular, two gradient schemes, the HMM and the HHO, were presented. It was noted in Section 2.4 that the fluxes obtained from an HMM scheme are not accurate for highly distorted meshes, and a high order scheme with degree 2 (for isotropic diffusion tensors), or even degree 3 (for anisotropic diffusion tensors) would be needed for accurate approximations of the fluxes.

The normal component of the velocity field needed to be continuous across the edges so that the flow is well defined, and can be used to solve the characteristic equation. Moreover, the divergence of the velocity field needed to be preserved in each cell, in order to avoid the introduction of artificial sources or sinks. We resolved this by the reconstruction of $H(\mathrm{div})$ velocity fields on simplices and on quadrilaterals, as in Chapter 3. In particular, our contribution here came in the design of the C and A methods in Sections 3.1.3 and 3.1.4, respectively. As opposed to the KR velocities in the literature [64], the C and A velocities are cheaper to compute, due to the availability of explicit expressions for the fluxes; moreover, C and A velocities are more accurate, and can recover constant velocity fields exactly, as seen in the tests in Section 3.3.1. We also note here that KR and A velocities can both be extended to 3D, whereas an easy way to fully extend C velocities into 3D is still an open problem. Another avenue of possible exploration would be the extension of the C and A methods, so that they can approximate high order

moments along the interior faces, which can then be used to reconstruct $\mathbb{RT}_k$ velocity fields for $k \geq 1$.

We then presented a summary of two characteristic-based schemes, the ELLAM and the MMOC. The weakness of ELLAM comes in the difficulty of approximating the integrals of steep back-tracked functions, whereas the weakness of MMOC is in the fact that it does not conserve mass. One of our main contributions in this chapter is the proposed combined ELLAM–MMOC scheme (Section 4.5), which mitigates the weakness of both the ELLAM and the MMOC schemes. A detailed discussion on how to implement these characteristic-based schemes is then presented in Section 4.6. The main difficulty of implementing these characteristic-based schemes is the violation of the local volume constraint. We developed here in Section 4.6.2 a novel volume adjustment algorithm, which can be used in conjunction with schemes that have piecewise constant approximations, such as finite volume type schemes. The algorithm we developed is applicable on generic meshes, and does not compute an explicit expression for the final (re-adjusted) mesh. It is thus cheaper to implement than those in the literature [5, 27].

We extended the combined ELLAM-MMOC to the miscible flow model, using the GDM framework for the diffusive terms, thus obtaining the GDM-ELLAM-MMOC (GEM) scheme. Numerical tests were then performed for schemes that fall under the GEM framework. Comparison between the performances of the HMM–ELLAM, MFEM–ELLAM, and HMM–upwind schemes show that in general, the HMM–ELLAM performs better than the other two, namely:

- HMM–upwind has no overshoots and undershoots, while HMM–ELLAM has overshoot $< 1\%$ and no undershoot. MFEM–ELLAM has very high overshoots and undershoots, especially for the test case with inhomogeneous permeability.

- HMM–upwind introduces a lot of numerical diffusion, and hence smears out the expected viscous fingering effect. On the other hand, HMM–ELLAM, being a characteristic-based scheme, captures the fingering effect better than HMM–upwind; however, the transition layer from $c \approx 0$ to $c \approx 1$ is thinner for MFEM–ELLAM.

- HMM–ELLAM gives a better approximation of the amount of oil recovered, compared to both MFEM–ELLAM and HMM–upwind.

Tests were then performed on generic meshes to compare the HMM–ELLAM and the HMM–GEM. In particular, it was observed that a better local volume conservation is achieved for the HMM–GEM scheme compared

to the HMM–ELLAM scheme, especially on meshes with distortion. The grid effects, however, remain persistent for the very distorted Kershaw type meshes. Attempts to mitigate the grid effects were made, by studying thin rectangular meshes and Kershaw-like meshes with less distortion. High order methods are expected to perform better on coarse meshes, and hence the idea was to use a HHO scheme for the gradient discretisation, while maintaining piecewise constant approximations for the concentration c. However, this did not help mitigate the grid effects, which lead to the conclusion that a fully high order scheme might be needed. This is not trivial to implement, and is not covered by the scope of the thesis. Future research may involve finding a way to mitigate grid effects on coarse distorted grids for characteristic-based schemes.

Finally, we analysed the convergence of these GDM characteristic-based schemes for the complete coupled model (1.1). Our analysis applies to a wide range of schemes, given the variety of numerical methods for diffusion problems that fit into the GDM. To cite a few examples, our results apply to MFEM–ELLAM of [78] and to the HMM–ELLAM of [23]. The GEM framework also gives an easy way to construct other characteristic-based schemes, by discretising the diffusion terms using any of the method known to fit into the GDM.

Contrary to previous convergence analysis of schemes involving the ELLAM or MMOC, the analysis here relies neither on L^∞ bounds on the concentration (which, given the anisotropic diffusive terms and generic meshes used in reservoir engineering, would not hold at the discrete level), nor on the smoothness of the data or the solutions (which cannot be established in practical situations, with discontinuous data such as the permeability, porosity, etc.). The convergence is established under minimal regularity assumptions on the data, using energy estimates and discrete compactness techniques. To carry out this analysis, fine properties of the flow of possibly discontinuous Darcy velocities have been established. We note however that this convergence analysis assumes a perfect computation of the tracked regions; future work will address the issue of accounting for approximation in tracked regions, and adjustment strategies, in the convergence analysis.

To summarise, we were able to design and implement characteristic-based schemes on generic polygonal meshes, and also analyse their convergence without assuming smoothness on the data or solution. This was done by the use of gradient schemes, which enables us to work inside a framework that allows a simultaneous analysis for various choices of discretisations for the diffusive terms. In terms of the implementation of characteristic-based schemes, several details needed to be taken care of. Firstly, we had to make sure that the flow of the velocity field is well-defined, which was done by

the construction of an $H(\text{div})$ velocity field. Core to the implementation of characteristic-based schemes is the approximation of the trace-back regions by polygons. It was found that for Cartesian or square meshes, approximating the trace-back regions by a polygon formed by tracking only the vertices and edge midpoints is sufficient. However, for irregular cells, more points need to be tracked along the edges of each cell. Next, to ensure the conservation of mass, an adjustment algorithm was developed. Finally, in order to ensure that mass conservation can be achieved, particular care has to be taken into account for the discretisation of the steep source terms.

At this stage, we would want to bring about some of the challenges that has been encountered, together with some perspectives for future work. Firstly, by using a high order approximation for the diffusive terms and a piecewise constant approximation for the advective terms, we noted that grid effects were prominent for distorted meshes. Future work may look into the mitigation of grid effects by using a high order approximation for the advective terms. Another interesting aspect to consider would be the extension of these GDM characteristic-based schemes into 3D. The main challenge that would be encountered here is the computation of integrals over the trace-back regions. In 2D, this involved taking the intersection between polygons, for which an algorithm is readily available. However, for 3D, taking the intersection between polyhedra is not trivial to implement, especially if there is no assumption on the convexity of the polyhedra. In this thesis, the application of this family of GDM characteristic-based methods is focused on petroleum engineering. Actually, characteristic-based methods are relevant in models that are advection dominated, which appear in many situations. It would therefore be interesting to consider its application onto other areas, such as groundwater flow, nuclear waste storage, computational fluid dynamics (Navier-Stokes), etc.

Appendix A

List of figures and test parameters

In this appendix, we present a table which gives the choice of the parameters used in the numerical tests in Chapter 5. In particular, we take note of the type of velocity reconstruction used, whether a local adjustment has been made for volume conservation or not, and the number of points tracked along the edge of each cell. If different parameters are used for the left and right side of one figure, the left side of the figure will be indicated by (L), whereas the right side will be indicated by (R). By default, for MMOC schemes, no adjustments are made, and hence a distinction between the parameters used for the left and right side will not be made if a figure contains a concentration profile obtained from an ELLAM scheme on the left and a MMOC scheme on the right, such as Figures 5.24, 5.26, 5.29, 5.31. A partial local adjustment (only applicable for ELLAM) means that we do not use (5.5).

Figure	velocity	local adjustments	points per edge
Fig. 5.6	A	none	$\lceil \log_2(m_{\mathcal{M}\text{reg}}) \rceil$
Fig. 5.7	A	none	$\lceil \log_2(m_{\mathcal{M}\text{reg}}) \rceil$
Fig. 5.8	A	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.9 (L)	KR	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.9 (R)	C	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.10 (L)	KR	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.10 (R)	C	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.11 (L)	KR	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.11 (R)	C	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Figs. 5.12–5.13	A	full	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.14	A	partial	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.16	A	full	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.17	A	full	5
Figs. 5.24–5.25	A	full	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.26	A	none	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.27 (L)	A	partial	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Fig. 5.27 (R)	A	full	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Figs. 5.29–5.30	A	full	n_K
Figs. 5.31–5.32	A	full	$\lceil \log_2(m_{K\text{reg}}) \rceil$
Figs. 5.39–5.43	A	full	n_K

Table A.1: Parameters used for the HMM–GEM test cases

Bibliography

[1] Yahya Alnashri and Jérôme Droniou. A gradient discretization method to analyze numerical schemes for nonlinear variational inequalities, application to the seepage problem. *SIAM J. Numer. Anal.*, 56(4):2375–2405, 2018.

[2] Daniel Anderson and Jérôme Droniou. An arbitrary-order scheme on generic meshes for miscible displacements in porous media. *SIAM J. Sci. Comput.*, 40(4):B1020–B1054, 2018.

[3] T. Arbogast and M. Correa. Two families of H(div) mixed finite elements on quadrilaterals of minimal dimension. *SIAM Journal on Numerical Analysis*, 54(6):3332–3356, 2016.

[4] T. Arbogast and Z. Tao. Direct Serendipity and Mixed Finite Elements on Convex Quadrilaterals. *ArXiv e-prints*, September 2018.

[5] Todd Arbogast and Chieh-Sen Huang. A fully mass and volume conserving implementation of a characteristic method for transport problems. *SIAM J. Scientific Computing*, 28:2001–2022, 2006.

[6] Todd Arbogast and Chieh-Sen Huang. A fully conservative Eulerian-Lagrangian method for a convection-diffusion problem in a solenoidal field. *J. Comput. Phys.*, 229(9):3415–3427, 2010.

[7] Todd Arbogast and Wen-Hao Wang. Convergence of a fully conservative volume corrected characteristic method for transport problems. *SIAM J. Numer. Anal.*, 48(3):797–823, 2010.

[8] Todd Arbogast and Wen-Hao Wang. Stability, monotonicity, maximum and minimum principles, and implementation of the volume corrected characteristic method. *SIAM J. Sci. Comput.*, 33(4):1549–1573, 2011.

[9] Todd Arbogast and Mary F. Wheeler. A characteristics-mixed finite element method for advection-dominated transport problems. *SIAM J. Numer. Anal.*, 32(2):404–424, 1995.

[10] Douglas N. Arnold, Daniele Boffi, and Richard S. Falk. Quadrilateral H(div) finite elements. *SIAM Journal on Numerical Analysis*, 42(6):2429–2451, 2005.

[11] S. Bartels, M. Jensen, and R. Müller. Discontinuous Galerkin finite element convergence for incompressible miscible displacement problems of low regularity. *SIAM J. Numer. Anal.*, 47(5):3720–3743, 2009.

[12] Lourenço Beirão da Veiga, Gianmarco Manzini, and Mario Putti. Post processing of solution and flux for the nodal mimetic finite difference method. *Numer. Methods Partial Differential Equations*, 31(1):336–363, 2015.

[13] Daniele Boffi, Franco Brezzi, Leszek F. Demkowicz, Ricardo G. Durán, Richard S. Falk, and Michel Fortin. *Mixed finite elements, compatibility conditions, and applications*, volume 1939 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin; Fondazione C.I.M.E., Florence, 2008. Lectures given at the C.I.M.E. Summer School held in Cetraro, June 26–July 1, 2006, Edited by Boffi and Lucia Gastaldi.

[14] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[15] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15(10):1533–1551, 2005.

[16] Franco Brezzi and Michel Fortin. *Mixed and hybrid finite element methods*, volume 15 of *Springer Series in Computational Mathematics*. Springer-Verlag, New York, 1991.

[17] Michael A. Celia, Thomas F. Russell, Ismael Herrera, and Richard E. Ewing. An Eulerian-Lagrangian localized adjoint method for the advection-diffusion equation. *Advances in Water Resources*, 13(4):187 – 206, 1990.

[18] C. Chainais-Hillairet, S. Krell, and A. Mouton. Study of discrete duality finite volume schemes for the Peaceman model. *SIAM J. Sci. Comput.*, 35(6):A2928–A2952, 2013.

[19] C. Chainais-Hillairet, S. Krell, and A. Mouton. Convergence analysis of a DDFV scheme for a system describing miscible fluid flows in porous

media. *Numer. Methods Partial Differential Equations*, 31(3):723–760, 2015.

[20] Claire Chainais-Hillairet and Jérôme Droniou. Convergence analysis of a mixed finite volume scheme for an elliptic-parabolic system modeling miscible fluid flows in porous media. *SIAM J. Numer. Anal.*, 45(5):2228–2258 (electronic), 2007.

[21] Z. Chen, R.E. Ewing, Q. Jiang, and A.M. Spagnuolo. Error analysis for characteristics-based methods for degenerate parabolic problems. *SIAM J. Numer. Anal.*, 40(4):1491–1515, 2002.

[22] Hanz Martin Cheng and Jérôme Droniou. Combining the hybrid mimetic mixed method and the Eulerian Lagrangian localised adjoint method for approximating miscible flows in porous media. In *Finite volumes for complex applications VIII—hyperbolic, elliptic and parabolic problems*, volume 200 of *Springer Proc. Math. Stat.*, pages 367–376. Springer, Cham, 2017.

[23] Hanz Martin Cheng and Jérôme Droniou. An HMM–ELLAM scheme on generic polygonal meshes for miscible incompressible flows in porous media. *Journal of Petroleum Science and Engineering*, 172:707 – 723, 2019.

[24] Hanz Martin Cheng, Jérôme Droniou, and Kim-Ngan Le. Convergence analysis of a family of ELLAM schemes for a fully coupled model of miscible displacement in porous media. *Numerische Mathematik*, 141(2):353–397, Feb 2019.

[25] Bernardo Cockburn, Daniele A. Di Pietro, and Alexandre Ern. Bridging the hybrid high-order and hybridizable discontinuous galerkin methods. *ESAIM: M2AN*, 50(3):635–650, 2016.

[26] Daniel A. Cogswell and Michael L. Szulczewski. Simulation of incompressible two-phase flow in porous media with large timesteps. *Journal of Computational Physics*, 345:856 – 865, 2017.

[27] Marta D'Elia, Denis Ridzal, Kara J. Peterson, Pavel Bochev, and Mikhail Shashkov. Optimization-based mesh correction with volume and convexity constraints. *Journal of Computational Physics*, 313:455 – 477, 2016.

[28] Daniele A. Di Pietro, Jérôme Droniou, and Alexandre Ern. A discontinuous-skeletal method for advection-diffusion-reaction on general meshes. *SIAM J. Numer. Anal.*, 53(5):2135–2157, 2015.

[29] Daniele A. Di Pietro, Jérôme Droniou, and Gianmarco Manzini. Discontinuous Skeletal Gradient Discretisation methods on polytopal meshes. *J. Comput. Phys.*, 355:397–425, 2018.

[30] Daniele A. Di Pietro, Alexandre Ern, and Simon Lemaire. An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. *Comput. Methods Appl. Math.*, 14(4):461–472, 2014.

[31] Daniele Antonio Di Pietro and Jérôme Droniou. *The Hybrid High-Order Method for Polytopal Meshes*. June 2019. Version 1, https://hal.archives-ouvertes.fr/hal-02151813/file/hho-book.pdf.

[32] Daniele Antonio Di Pietro and Alexandre Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.

[33] J. Douglas, R.E. Ewing, and M.F. Wheeler. The approximation of the pressure by a mixed method in the simulation of miscible displacement. *RAIRO Anal. Numér.*, 17(1):17–33, 1983.

[34] J. Douglas, Jr. Finite difference methods for two-phase incompressible flow in porous media. *SIAM J. Numer. Anal.*, 20(4):681–696, 1983.

[35] Jim Douglas, Frederico Furtado, and Felipe Pereira. On the numerical simulation of waterflooding of heterogeneous petroleum reservoirs. *Computational Geosciences*, 1(2):155–190, Aug 1997.

[36] Jim Douglas, Jr., Chieh-Sen Huang, and Felipe Pereira. The modified method of characteristics with adjusted advection. *Numerische Mathematik*, 83(3):353–369, Sep 1999.

[37] Jim Douglas, Jr. and Thomas F. Russell. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM J. Numer. Anal.*, 19(5):871–885, 1982.

[38] J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105(1):35–71, 2006.

[39] Jérôme Droniou. Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Math. Models Methods Appl. Sci.*, 24(8):1575–1619, 2014.

[40] Jérôme Droniou, Robert Eymard, Thierry Gallouët, Cindy Guichard, and Raphaèle Herbin. *The gradient discretisation method*, volume 82 of *Mathematics & Applications*. Springer, 2018.

[41] Jérôme Droniou, Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Models Methods Appl. Sci.*, 20(2):265–295, 2010.

[42] Jérôme Droniou, Robert Eymard, Alain Prignet, and Kyle S. Talbot. Unified convergence analysis of numerical schemes for a miscible displacement problem. *Foundations of Computational Mathematics*, Mar 2018.

[43] Jérôme Droniou and Kyle S. Talbot. On a miscible displacement model in porous media flow with measure data. *SIAM J. Math. Anal.*, 46(5):3158–3175, 2014.

[44] Lawrence C. Evans and Ronald F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[45] R. Ewing. *The Mathematics of Reservoir Simulation*. Society for Industrial and Applied Mathematics, 1983.

[46] R.E. Ewing, T.F. Russell, and M.F. Wheeler. Simulation of miscible displacement using mixed methods and a modified method of characteristics. In *SPE Reservoir Simulation Symposium, 15-18 November, San Francisco, California*. Society of Petroleum Engineers, 1983.

[47] R.E. Ewing, T.F. Russell, and M.F. Wheeler. Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics. *Comput. Methods Appl. Mech. Engrg.*, 47(1-2):73–92, 1984.

[48] R.E. Ewing and M.F. Wheeler. Galerkin methods for miscible displacement problems in porous media. *SIAM J. Numer. Anal.*, 17(3):351–365, 1980.

[49] Richard E. Ewing and Hong Wang. A summary of numerical methods for time-dependent advection-dominated partial differential equations. *J. Comput. Appl. Math.*, 128(1-2):423–445, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.

[50] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In P. G. Ciarlet and J.-L. Lions, editors, *Techniques of Scientific Computing, Part III*, Handbook of Numerical Analysis, VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[51] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4):1009–1043, 2010.

[52] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. $\mathcal{RT}_k$ mixed finite elements for some nonlinear problems. *Math. Comput. Simulation*, 118:186–197, 2015.

[53] Robert Eymard, Cindy Guichard, Raphaèle Herbin, and Roland Masson. Vertex centred discretization of two-phase darcy flows on general meshes. *ESAIM: Proc.*, 35:59–78, 2012.

[54] Xiaobing Feng. On existence and uniqueness results for a coupled system modeling miscible displacement in porous media. *J. Math. Anal. Appl.*, 194(3):883–910, 1995.

[55] P.A. Forsyth, Y.S. Wu, and K. Pruess. Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media. *Advances in Water Resources*, 18(1):25 – 38, 1995.

[56] Peter A. Forsyth. A control volume finite element approach to napl groundwater contamination. *SIAM J. Scientific Computing*, 12:1029–1057, 1991.

[57] Vivette Girault, Jizhou Li, and Beatrice M. Rivière. Strong convergence of the discontinuous galerkin scheme for the low regularity miscible displacement equations. *Numerical Methods for Partial Differential Equations*, 33(2):489–513, 2017.

[58] Richard W. Healy and Thomas F. Russell. Solution of the advection-dispersion equation in two dimensions by a finite-volume eulerian-lagrangian localized adjoint method. *Advances in Water Resources*, 21(1):11 – 26, 1998.

[59] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.

[60] J.M. Holte. Discrete Gronwall lemma and applications. In *MAA-NCS meeting at the University of North Dakota*, volume 24, pages 1–7, 2009. http://homepages.gac.edu/∼holte/publications/GronwallLemma.pdf.

[61] Chieh-Sen Huang. Convergence analysis of a mass-conserving approximation of immiscible displacement in porous media by mixed finite elements and a modified method of characteristics with adjusted advection. *Computational Geosciences*, 4(2):165–184, 2000.

[62] M. R. Islam, M. E. Hossain, S. H. Moussavizadegan, S. Mustafiz, and J. H. Abou-Kassem. *Advanced Petroleum Reservoir Simulation - Towards Developing Reservoir Emulators (2nd Edition)*. John Wiley & Sons, 2016.

[63] E. Jimenez, K. Sabir, A. Datta-Gupta, and M.J. King. Spatial error and convergence in streamline simulation. *Spe Reservoir Evaluation & Engineering*, 10(3):221–232, 2007.

[64] Yu. Kuznetsov and S. Repin. New mixed finite element method on polygonal and polyhedral meshes. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 18(3), 2003.

[65] Knut-Andreas Lie and Bradley T. Mallison. *Mathematical Models for Oil Reservoir Simulation*, pages 850–856. Springer Berlin Heidelberg, Berlin, Heidelberg, 2015.

[66] Shlomo P. Neuman. An Eulerian-Lagrangian numerical scheme for the dispersion-convection equation using conjugate space-time grids. *J. Comput. Phys.*, 41(2):270–294, 1981.

[67] D. W. Peaceman and H. H. Rachford, Jr. Numerical calculation of multidimensional miscible displacement. *Society of Petroleum Engineers Journal*, 2(4):327–339, 1962.

[68] Donald W. Peaceman. *Fundamentals of Numerical Reservoir Simulation*. Elsevier, 1977.

[69] D.W. Peaceman. Improved treatment of dispersion in numerical calculation of multidimensional miscible displacement. *Soc. Pet. Eng. J.*, 6(3):213–216, 1966.

[70] David W. Pollock. Semianalytical computation of path lines for finite-difference models. *Ground Water*, 26(6):743–750, 1988.

[71] Mathieu Prevost, Michael G. Edwards, and Martin J. Blunt. Streamline tracing on curvilinear structured and unstructured grids. *SPE Journal*, 7(2):139–148, 2002.

[72] B.M. Rivière and N.J. Walkington. Convergence of a discontinuous Galerkin method for the miscible displacement equation under low regularity. *SIAM J. Numer. Anal.*, 49(3):1085–1110, 2011.

[73] Thomas F. Russell. Numerical dispersion in Eulerian-Lagrangian methods. *Computational Methods in Water Resources*, 2:963970, 2002.

[74] Thomas F. Russell and Michael A. Celia. An overview of research on Eulerian–Lagrangian localized adjoint methods (ELLAM). *Advances in Water Resources*, 25(8):1215 – 1231, 2002.

[75] J. Sweeney. Numerical methods for an oil recovery model, 2015. Honours thesis, Monash University.

[76] Hong Wang. An optimal-order error estimate for a family of ELLAM-MFEM approximations to porous medium flow. *SIAM J. Numer. Anal.*, 46(4):2133–2152, 2008.

[77] Hong Wang, Helge K. Dahle, Richard E. Ewing, Magne S. Espedal, Robert C. Sharpley, and Shushuang Man. An ELLAM scheme for advection-diffusion equations in two dimensions. *SIAM J. Sci. Comput.*, 20(6):2160–2194, 1999.

[78] Hong Wang, Dong Liang, Richard E. Ewing, Stephen L. Lyons, and Guan Qin. An approximation to miscible fluid flows in porous media with point sources and sinks by an Eulerian-Lagrangian localized adjoint method and mixed finite element methods. *SIAM J. Sci. Comput.*, 22(2):561–581 (electronic), 2000.

[79] Stephen Whitaker. Diffusion and dispersion in porous media. *AIChE Journal*, 13(3):420–427, 1967.