# Finite volume/finite element schemes for compressible two-phase flows inheterogeneous and anisotropic porous media

El Houssaine Quenjel

# THESE DE DOCTORAT DE

L'ÉCOLE CENTRALE DE NANTES
COMUE UNIVERSITE BRETAGNE LOIRE
ET L'UNIVERSITE MOULAY ISMAIL MEKNES

ECOLE DOCTORALE N° 601
*Mathématiques et Sciences et Technologies de l'Information et de la Communication*
Spécialité : *Mathématiques et leurs Interactions*

Par

**EL Houssaine QUENJEL**

## « Volumes finis/Eléments finis pour des écoulements diphasiques compressibles en milieux poreux hétérogènes et anisotropes »

**Thèse présentée et soutenue publiquement au Maroc, le 15 décembre 2018**
**Unité de recherche : Laboratoire de Mathématiques Jean Leray (LMJL)**

**Rapporteurs avant soutenance** :

| | |
|---|---|
| Roland MASSON | Professeur des universités, Université de Nice Sophia Antipolis |
| Mohammed AFIF | Professeur, Université Cadi Ayyad |
| Driss SGHIR | Professeur, Université Moulay Ismail |

**Composition du Jury :**

| | | |
|---|---|---|
| Président : | EL Hassan EL KINANI | Professeur, Université Moulay Ismail |
| Rapporteurs : | Roland MASSON | Professeur des universités, Université de Nice Sophia Antipolis |
| | Mohammed AFIF | Professeur, Université Cadi Ayyad |
| | Driss SGHIR | Professeur, Université Moulay Ismail |
| Examinateurs : | Boris ANDREIANOV | Professeur des universités, Université de Tours |
| | Laurent DI MENZA | Professeur des universités, Université de Reims Champagne-Ardenne |
| Dir. de thèse : | Mazen SAAD | Professeur des universités, Ecole Centrale de Nantes |
| Co-dir. de thèse : | Mustapha GHILANI | Professeur, ENSAM-Meknès, Université Moulay Ismail |

# Acknowledgment

First of all, I would like to express my best thanks to my supervisors Mustapha GHILANI, the Professor at college of engineering (ENSAM) in Meknès and Mazen SAAD, the Professor at Centrale Nantes, for their gentleness, ready assistance, pertinent comments, encouragement and advices. Throughout this work, I have really learned from their scientific qualities many skills in applied mathematics, especially in the numerical analysis and programing. Without them, I would have never achieved what I reached today.

Special thanks are addressed to the Professors: Roland MASSON, Mohammed AFIF and Driss SGHIR. I deeply acknowledge them for accepting to be members of the referees of this thesis. Particular acknowledgments go to : Boris ANDREIANOV, Laurent DI MENZA and Hassan EL KINANI who kindly accepted to be members of the defense committee.

The last chapter of this thesis has been carried out in collaboration with the Professor Marianne BESSEMOULIN-CHATARD at University of Nantes who I think sincerely for her valuable comments.

I am thankful to Miriam ANCEL for her gentleness and modesty to revise a part of this report and give me some important and valuable comments in English. I thank all the mathematics' faculty for their contribution in my education, their encouragement and help. Also, I wish to thank my friends and colleagues in Morocco and France for the unforgotten memories and the best moments we have lived and shared all together.

At last, not at least, I am gratefully indebted to my parents Mohammed and Moulati, for being always there, their continuous support, unconditional love and encouragement. My deep and sincere acknowledgments are devoted to the rest of my family. They all have my respect and appreciation.

This work is partially supported by: Ministère de l'Enseignement Supérieur, de la Recherche Scientifique et de la Formation des Cadres du Maroc, CNRST and l'Institue Français au Maroc, which I appreciate very much.

*To my parents*

# Abstract

The objective of this thesis is the development and the analysis of robust and consistent numerical schemes for the approximation of compressible two-phase flow models in anisotropic and heterogeneous porous media. A particular emphasis is set on the anisotropy together with the geometric complexity of the medium. The mathematical problem is given in a system of two degenerate and coupled parabolic equations whose main variables are the nonwetting saturation and the global pressure. In view of the difficulties manifested in the considered system, its cornerstone equations are approximated with two different classes of the finite volume family.

The first class consists of combining finite elements and finite volumes. Based on standard assumptions on the space discretization and on the permeability tensor, a rigorous convergence analysis of the scheme is carried out thanks to classical arguments. To dispense with the underlined assumptions on the anisotropy ratio and on the mesh, the model has to be first formulated in the factional flux formulation. Moreover, the diffusive term is discretized by a Godunov-like scheme while the convective fluxes are approximated using an upwind technique. The resulting scheme preserves the physical ranges of the computed solution and satisfies the coercivity property. Hence, the convergence investigation holds. Numerical results show a satisfactory qualitative behavior of the scheme even if the medium of interest is anisotropic.

The second class allows to consider more general meshes and tensors. It is about a new positive nonlinear discrete duality finite volume method. The main point is to approximate a part of the fluxes using a nonstandard technique. The application of this idea to a nonlinear diffusion equation yields surprising results. Indeed, not only is the discrete maximum property fulfilled but also the convergence of the scheme is established. Practically, the proposed method shows great promises since it provides a positivity-preserving and convergent scheme with an optimal convergence rate.

**Mots clés :** porous media, two-phase flow, compressible, immiscible, finite volumes, finite elements, positive, DDFV monotone.

# Résumé

Cette thèse est centrée autour du développement et de l'analyse des schémas volumes finis robustes afin d'approcher les solutions du modèle diphasique compressible en milieux poreux hétérogènes et anisotropes. Le modèle à deux phases compressibles comprend deux équations paraboliques dégénérées et couplées dont les variables principales sont la saturation du gaz et la pression globale. Ce système est discrétisé à l'aide de deux méthodes différentes (CVFE et DDFV) qui font partie de la famille des volumes finis.

La première classe à laquelle on s'intéresse consiste à combiner la méthode des volumes finis et celle des éléments finis. Dans un premier temps, on considère un schéma volume finis upwind pour la partie convective et un schéma de type éléments finis conformes pour la diffusion capillaire. Sous l'hypothèse que les coefficients de transmissibilités sont positifs, on montre que la saturation vérifie le principe du maximum et on établit des estimations d'énergies permettant de démontrer la convergence du schéma. Dans un second temps, on a mis en place un schéma positif qui corrige le précédent. Ce schéma est basé sur une approximation des flux diffusifs par le schéma de Godunov. L'avantage est d'établir la bornitude des solutions approchées ainsi que les estimations uniformes sur les gradients discrets sans aucune contrainte ni sur le maillage ni sur la perméabilité. En utilisant des arguments classiques de compacité, on prouve rigoureusement la converge du schéma. Chaque schéma est validé par des simulations numériques qui montrent bien le comportement attendu d'une telle solution.

Concernant la deuxième classe, on s'intéressera tout d'abord à la construction et à l'étude d'un nouveau schéma de type DDFV (Discrete Duality Finite Volume) pour une équation de diffusion non linéaire dégénérée. Cette méthode permet d' avantage de prendre en compte des maillages très généraux et des perméabilités quelconques. L'idée clé de cette discrétisation est d'approcher les flux dans la direction normale par un schéma centré et d'utiliser un schéma décentré dans la direction tangentielle. Par conséquent, on démontre que la solution approchée respecte les bornes physiques et on établit aussi des estimations d'énergie. La convergence du schéma est également établie. Des résultats numériques confirment bien ceux de la théorie. Ils exhibent en outre que la méthode est presque d'ordre deux.

**Mots clés :** milieux poreux, diphasique compressible, immiscible, volumes finis, éléments finis, positif, DDFV monotone.

# Contents

# Introduction

## 0.1 Contexte général et objectifs de la thèse

Ce mémoire de thèse est consacré à l'analyse numérique des schémas volumes finis positifs pour des écoulements diphasiques compressibles en milieux poreux hétérogènes et anisotropes.

La modélisation des écoulements en milieux poreux joue un rôle important dans la compréhension de nombreux phénomènes issus de la physique, de l'hydrogéologie, de l'environnement ... La mise en équations de tels phénomènes permet de mieux les comprendre et espérer prévoir le comportement de leurs variables potentielles.

La modélisation des problèmes physiques fait souvent intervenir des phénomènes de convection et de diffusion. Celle ci conduit à des systèmes d'équations aux dérivées partielles de type hyperboliques ou/et elliptiques non linéaires. La résolution ou l'analyse mathématique (existence, unicité et stabilité), des solutions du système obtenu n'est pas toujours évidente. Elle est même délicate à cause de la non linéarité, la dégénérescence et au couplage des équations. A cet effet l'approximation de ces systèmes reste la seule possibilité pour se faire une idée sur les profils de leurs solutions. Cependant cette démarche génère d'autres difficultés spécifiques à l'approximation tels que la consistance, la stabilité numérique et la convergence. Tout cela constituera le cœur de notre préoccupation dans ce travail.

Le développement et l'analyse des méthodes numériques pour l'approximation des solutions des systèmes en question a connu une rapide expansion dans les dernières décennies. Plusieurs méthodes ont été utilisées pour leurs approximations : différences finies, éléments finis, éléments finis mixtes, volumes finis, ... . Dans ce travail nous utiliserons des schémas numériques qui combinent les volumes finis et les éléments finis (CVFE) d'une part et la méthode des DDFV (discrete duality finite volume) d'autre part.

Le problème mathématique auquel nous nous intéressons au cours de ce mémoire de thèse est constitué de deux équations paraboliques non linéaires dégénérées et fortement couplées. Outre la non linéarité, l'anisotropie du milieu présente une sérieuse difficulté. Celle ci rend difficile la satisfaction des solutions approchées des bornes physiques. A tire d'exemple, la saturation approchée doit être comprise entre 0 et 1. Ce problème porte le nom du principe du maximum.

Outre le traitement des géométries complexes, la conservation des flux est parmi les avantages qui motive l'usage des méthodes volumes finis en général, CVFE et DDFV en particulier. Dans notre cas, l'ensemble des équations qui régissent les écoulements diphasiques résultent essentiellement de

la loi de conservation de la masse. Ainsi, la conservation locale de la masse est une propriété tout à fait naturelle qui doit être satisfaite à travers des interfaces des sous-domaines adjacents.

L'objectif primordial de ce travail est de concevoir des schémas volumes finis de type CVFE et DDFV qui soient capables de simuler le modèle diphasique compressible et anisotrope. L'analyse mathématique de chaque schéma est détaillée avec des validations numériques dans des chapitres spécifiques. Les résumés des méthodes CVFE et DDFV sont présentés dans la section suivante.

## 0.2 Synthèse du manuscrit

Ce manuscrit est divisé en quatre chapitres. Le premier chapitre présente une brève motivation de la thèse. Nous rappelons les concepts de base des écoulements en milieux poreux, notamment la saturation, les perméabilités, l'anisotropie, la pression capillaire, la pression globale et les différentes formulations des équations qui régissent des écoulements compressibles dans un milieu poreux. Ce survol a comme but de mieux comprendre l'objectif des travaux réalisés tout au long de cette thèse.

Le deuxième chapitre est consacré à l'analyse numérique d'un schéma volumes finis/éléments finis pour un système d'équations aux dérivées partielles paraboliques dégénérées. Ceci suppose la positivité des coefficients de transmissivité comme hypothèse principale. En revanche, quand le milieu est fortement anisotrope cette hypothèse n'est pas forcement vérifiée. Pour s'affranchir de cette contrainte, nous proposons dans le troisième chapitre un schéma qui assure la positivité des saturations malgré l'anisotropie. L'idée clé de cette approche consiste à approximer le terme de diffusion comme s'il était de type hyperbolique.

Dans le quatrième chapitre et afin d'utiliser des maillages très généraux et des perméabilités quelconques, nous construisons un schéma DDFV pour un problème de diffusion non linéaire. L'apport majeur de ce nouveau schéma est qu'il vérifie le principe du maximum discret ainsi que des estimations de type énergie au même temps.

Dans ce qui suit nous présenterons un survol descriptif de tous les chapitres du mémoire de thèse.

### 0.2.1 Chapitre 2 : Analyse numérique d'un schéma volumes finis de type "vertex-centered" pour un modèle eau-gaz en milieux poreux

Ce chapitre est centré sur l'analyse numérique d'un schéma volumes finis dit "vertex-centered", de type CVFE (Control Volume Finite Element) approchant un système d'équations de convection-diffusion qui modélise des écoulements immiscibles eau-gaz en milieux poreux. La phase gazeuse est considérée compressible tandis que celle de l'eau est incompressible. Ce système est obtenu par la loi de conservation de masse pour chaque phase et où la vitesse de l'écoulement est donnée par la loi de Darcy-Muskat généralisée.

#### Modèle mathématique

Le problème est posé sur le cylindre $Q_{\mathfrak{T}} = \Omega \times (0, \mathfrak{T}))$ où $\Omega$ est un ouvert connexe, borné et polyédrique de $R^d$ $(d = 2, 3)$. $\mathfrak{T}$ est un réel strictement positif qui représente le temps physique.

Après certaines transformations, le système auquel on s'intéresse s'écrit sous la forme :

$$\begin{aligned}\partial_t(\phi\rho(p)s) - \operatorname{div}\Big(\Lambda\rho(p)M_g(s)\nabla p\Big) - \operatorname{div}\Big(\Lambda\rho(p)\nabla\xi(s)\Big)\\ + \operatorname{div}\Big(\Lambda\rho^2(p)M_g(s)\vec{\mathbf{g}}\Big) + \rho(p)sq^P = 0,\end{aligned} \tag{0.2.1}$$

$$\begin{aligned}\partial_t(\phi s) + \operatorname{div}\Big(\Lambda M_w(s)\nabla p\Big) - \operatorname{div}\Big(\Lambda\nabla\xi(s)\Big)\\ - \operatorname{div}\Big(\Lambda M_w(s)\vec{\mathbf{g}}\Big) + sq^P = q^P - q^I.\end{aligned} \tag{0.2.2}$$

Les variables principales sont la saturation du gaz $s$ et la pression globale $p$. On précise que la saturation de l'eau est bien évidement égale à $1-s$. Les autres coefficients sont : $\phi$ la porosité du milieu, $\rho$ la densité de la phase gazeuse, $\Lambda$ désigne la matrice de la perméabilité absolue, $M_g$ (resp. $M_w$) la mobilité du gaz (resp. de l'eau) et $\xi$ est une primitive qui s'annule en 0 du terme

$$\gamma(s) = \frac{M_w(s)M_g(s)}{M(s)}p_c'(s) \geq 0, \quad \xi(s) = \int_0^s \gamma(u)\, du.$$

où $p_c(s)$ est la pression capillaire, qui est ici une fonction de la saturation du gaz. De plus, le vecteur $\vec{\mathbf{g}}$ est l'accélération du pesanteur et $q^P, q^I$ sont des fonctions sources. Le système (0.2.1)-(0.2.2) est complété par la donnée de conditions initiales

$$p(x,0) = p^0(x) \text{ et } s(x,0) = s^0(x) \quad \text{pour tout } x \in \Omega, \tag{0.2.3}$$

et la donnée des conditions aux limites

$$\begin{cases} p = 0,\ s = 0 & \text{sur} \quad \Gamma_D \times (0,\mathfrak{T}) \\ \Big(M_g(s)\Lambda\nabla p + \Lambda\nabla\xi(s) - \rho(p)M_g(s)\Lambda\vec{\mathbf{g}}\Big)\cdot\mathbf{n} = 0 & \text{sur} \quad \Gamma_N \times (0,\mathfrak{T}) \\ \Big(M_w(s)\Lambda\nabla p - \Lambda\nabla\xi(s) - \rho_w M_w(s)\Lambda\vec{\mathbf{g}}\Big)\cdot\mathbf{n} = 0 & \text{sur} \quad \Gamma_N \times (0,\mathfrak{T}) \end{cases}. \tag{0.2.4}$$

où $\{\Gamma_N, \Gamma_D\}$ est une partition du bord $\partial\Omega$ dont la mesure superficielle de $\Gamma_D$ est strictement positive. Le vecteur $\mathbf{n}$ désigne la normale sortante de $\partial\Omega$.

Nous supposons que les fonctions du problème (0.2.1)-(0.2.4) vérifient les hypothèses suivantes.

($\mathrm{H}_0$) La pression globale initiale $p^0$ est dans $L^2(\Omega)$ et la saturation du gaz initiale $s_0$ est dans $L^\infty(\Omega)$ avec $0 \leq s_0(x) \leq 1$ p.p. dans $\Omega$.

($\mathrm{H}_1$) La porosité $\phi$ est une fonction de $L^\infty$ avec $\phi_0 \leq \phi(x) \leq \phi_1$ p.p. dans $\Omega$ où $\phi_0$ et $\phi_1$ sont deux constantes strictement positives.

($\mathrm{H}_2$) La mobilité du gaz $M_g$, (resp. de l'eau $M_w$) est une fonction croissante (resp. décroissante) de $[0,1]$ dans $\mathbb{R}^+$. Elle est prolongeable par 0 sur l'intervalle $]-\infty, 0]$ (resp. $[1, +\infty[$). De plus, il existe une constante strictement positive telle que

$$m_0 \leq M_g(s) + M_w(s), \quad \forall s \in [0,1].$$

($\mathrm{H}_3$) La perméabilité absolue est donnée par une matrice symétrique définie positive dont les coefficients appartiennent à $L^\infty(\Omega)$. De plus, elle est uniformément elliptique i.e. ils existent $\underline{\Lambda}, \overline{\Lambda} > 0$ telles qu'on ait :

$$\underline{\Lambda}\,|\zeta|^2 \leq \Lambda(x)\zeta\cdot\zeta \leq \overline{\Lambda}\,|\zeta|^2, \text{ pour tout } \zeta \in \mathbb{R}^d \text{ et p.p. dans } \Omega.$$

(H$_4$) La fonction $\gamma$ est continue sur $[0,1]$ et

$$\begin{cases}\gamma(s) > 0 & \text{pour } 0 < s < 1 \\ \gamma(0) = \gamma(1) = 0\end{cases}.$$

On suppose également une continuité hölderienne d'ordre $\theta \in (0,1]$ sur la fonction $\xi^{-1}$. Autrement-dit, il existe une constante positive $C$ telle que :

$$\text{pour tous } a, b \in [0, \xi(1)], |\xi^{-1}(a) - \xi^{-1}(b)| \leq C|a-b|^{\theta}.$$

(H$_5$) Le terme d'injection $q^I$ et de production $q^P$ sont deux fonctions positives de $L^2(Q_{\mathfrak{T}})$.

(H$_6$) La densité du gaz est différentiablement continue sur $\mathbb{R}$ et uniformément bornée i.e. $\rho_0 \leq \rho(p_g) \leq \rho_1$ où $\rho_0$ et $\rho_1$ sont deux constantes positives.

Le cadre fonctionnel dans lequel on cherche les solutions au sens faible du problème continu n'est autre que l'espace de Sobolev classique

$$H^1_{\Gamma_D}(\Omega) = \{v \in H^1(\Omega) \ / \ v = 0 \text{ on } \Gamma_D\}.$$

C'est un espace de Hilbert muni de la norme

$$||v||_{H^1_{\Gamma_D}(\Omega)} = ||\nabla v||_{(L^2(\Omega))^d}$$

**Definition 0.2.1.** *(Solutions faibles) Sous les hypothèses* $(H_0)$*–*$(H_6)$*, un couple de fonctions mesurables* $(p,s)$ *est dit solution faible du problème (0.2.1)-(0.2.4) si les conditions suivantes sont satisfaites :*

$$\begin{aligned} 0 \leq s \leq 1 \ p.p. \ \ &dans \ \ Q_{\mathfrak{T}}, \\ \xi(s) \ \ &\in L^2(0,\mathfrak{T}; H^1_{\Gamma_D}(\Omega)), \\ p \ \ &\in L^2(0,\mathfrak{T}; H^1_{\Gamma_D}(\Omega)), \end{aligned}$$

*et pour tout* $\varphi, \psi \in \mathcal{C}^{\infty}_c(\Omega \times [0,\mathfrak{T}))$*, on a*

$$\begin{aligned} &-\int_{Q_{\mathfrak{T}}} \phi\rho(p)s \ \partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_{\Omega} \phi \ \rho(p^0)s^0 \ \varphi(x,0)\,\mathrm{d}x \\ &\quad + \int_{Q_{\mathfrak{T}}} \rho(p)M_g(s)\Lambda\nabla p \cdot \nabla\varphi \,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} \rho(p)\Lambda\nabla\xi(s)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t \\ &\quad - \int_{Q_{\mathfrak{T}}} \rho^2(p)M_g(s)\Lambda \ \vec{\boldsymbol{g}}\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} \rho(p)sq^P \ \varphi\,\mathrm{d}x\,\mathrm{d}t = 0, \end{aligned} \tag{0.2.5}$$

$$\begin{aligned} &-\int_{Q_{\mathfrak{T}}} \phi s \ \partial_t\psi\,\mathrm{d}x\,\mathrm{d}t - \int_{\Omega} \phi(x)s^0 \ \psi(x,0)\,\mathrm{d}x - \int_{Q_{\mathfrak{T}}} M_w(s)\Lambda\nabla p\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{Q_{\mathfrak{T}}} \Lambda\nabla\xi(s)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} M_w(s)\Lambda \ \vec{\boldsymbol{g}}\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{Q_{\mathfrak{T}}} sq^P\psi\,\mathrm{d}x\,\mathrm{d}t = \int_{Q_{\mathfrak{T}}} (q^P - q^I)\psi\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{0.2.6}$$

L'étude théorique du système précédent, notamment l'existence de solutions faibles, est réalisée dans l'article [84]. La question d'unicité reste encore ouverte.

## Le maillage CVFE

Nous entamerons cette partie par la description du maillage utilisé et nous continuerons par l'approximation des termes importants pour obtenir en fin le schéma final.

Nous nous plaçons dans le cas de la dimension deux en espace. En principe, la méthode CVFE est définie sur un maillage dual qui est construit à partir d'un maillage initial. Ce dernier est une triangulation $\mathcal{T}$ conforme au sens des éléments finis. La discrétisation $\mathcal{T}$ est alors un recouvrement du domaine $\Omega$ c'est-à-dire $\overline{\Omega} = \bigcup_{T\in\mathcal{T}} T$. Étant donné un triangle $T$, on dénote $x_T$ son barycentre, $h_T$ son diamètre, $\varrho_T$ le rayon de la plus grande boule inscrite dans $T$ et $|T|$ son aire. L'ensemble des arêtes de $T$ est noté $\mathcal{E}_T$. La famille de tous les sommets de la triangulation est dénotée $\mathcal{V}$. Pour chaque sommet $K \in \mathcal{V}$, on définit $\mathcal{K}_T$ l'ensemble des triangles qui partagent le sommet $K$. On lui associe également une et une seule maille duale, appelée volume de contrôle $\omega_K$. Elle est construite en connectant le barycentre de chaque triangle de $\mathcal{K}_T$ aux milieux des arêtes ayant $K$ comme extrémité. Le centre de $\omega_K$ est le sommet $x_K$ et sa surface $|\omega_K|$. Pour deux volumes de contrôles adjacents $K$ et $L$ qui s'intersectent dans un triangle $T$, on définit le segment $\sigma_{KL}^T = \partial K \cap \partial L \cap T$ où $|\sigma_{KL}^T|$ est sa longueur et $n_{\sigma_{KL}^T}$ la normale dirigée de $K$ vers $L$. Une illustration de ces deux maillages est présentée sur la figure (1) ci-dessous.

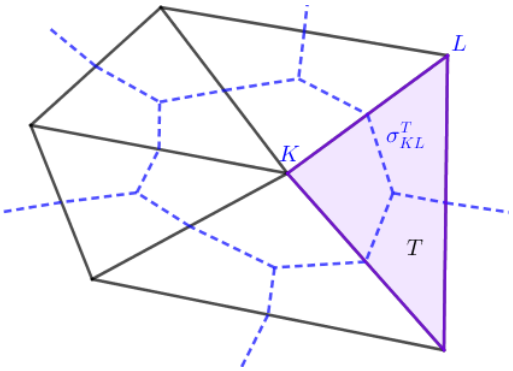



Figure 1: Exemple de maillages initial, dual et volume de contrôle

On note $h := \max_{T\in\mathcal{T}} h_T$ le pas du maillage triangulaire et $\theta_{\mathcal{T}} := \max_{T\in\mathcal{T}} \frac{h_T}{\varrho_T}$. Pour toute suite de maillages $(\theta_{\mathcal{T}_m})_m$ on suppose que la suite $(\theta_{\mathcal{T}_m})_m$ est uniformément majorée. Cette hypothèse constitue une condition nécessaire de la convergence de la méthode des éléments finis. Elle est appelée condition de régularité de la suite des maillages.

La discrétisation de l'intervalle $(0, \mathfrak{T})$ est donnée par une suite croissante de réels $(t^n)_{n=0,\dots,N}$, telle que :

$$t^0 = 0 \ < t^1 \ < \cdots < \ t^{N-1} \ < \ t^N = \mathfrak{T}.$$

Afin de simplifier l'exposé cette subdivision est considérée uniforme et le pas de temps est noté $\delta t$.

Soit $\{u_K^n\}_{\{K\in\mathcal{V},\ n=0,\dots,N\}}$ une famille de nombre réels. Dans le contexte du schéma CVFE, on précise deux reconstructions de la solution :

(i) Une solution volume fini $\tilde{u}_{h,\delta t}$ construite sur le maillage dual. Elle est constante par maille

et définie presque partout sur le sous-domaine $\bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T})$ par

$$\tilde{u}_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0\ \chi_{\mathring{\omega}_K}(x),\ \forall x \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K,$$

$$\tilde{u}_{h,\delta t}(x,t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} u_K^{n+1}\ \chi_{\mathring{\omega}_K\times(t^n,t^{n+1}]}(x,t),\ \forall (x,t) \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T}).$$

L 'ensemble de toutes ces fonctions est désigné par $W_{h,\delta t}$.

(ii) Une solution élément fini $\mathbb{P}_1$ en espace et constante par morceaux en temps

$$u_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0\varphi_K(x),\quad \forall x \in \Omega,$$

$$u_{h,\delta t}(x,t) = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} u_K^{n+1}\varphi_K(x)\ \chi_{(t^n,t^{n+1}]}(t),\quad \forall\ (x,t) \in \Omega\times(0,\mathfrak{T}).$$

L'ensemble de toute ces fonctions forment l'espace $X_{h,\delta t}$. Si $F(u)$ est une fonction non linéaire, on notera son interpolation au sens des volumes (resp. éléments) finis par $F(\tilde{u}_{h,\delta t})$ (resp. $F(u_{h,\delta t})$).

## Esquisse de la discrétisation du schéma CVFE pour le modèle diphasique

Pour rendre notre étude plus claire on néglige la gravité car elle ne pose aucune difficulté majeure. Étant donné un entier $n$ et un volume de contrôle dual $\omega_K$. On suit la démarche volumes finis. A cet effet, on intègre sur la maille $(t^n,t^{n+1}]\times\omega_K$ l'équation de la phase gazeuse, on applique la formule de Green. On obtient alors

$$\int_{t^n}^{t^{n+1}}\int_{\omega_K}\phi(x)\partial_t\Big(\rho(p)s\Big)\ \mathrm{d}x - \underbrace{\sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_\sigma \rho(p)M_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t}_{\text{terme convectif}}$$

$$\underbrace{-\sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_\sigma \rho(p)\ \Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t}_{\text{terme diffusif capillaire}}$$

$$+\int_{t^n}^{t^{n+1}}\int_{\omega_K}\rho(p)\ s\ q^P\,\mathrm{d}x\,\mathrm{d}t = 0,$$

où $\mathcal{E}_K$ est l'ensemble des arêtes duales et $\mathbf{n}_{\sigma K}$ la normale à l'interface $\sigma$ sortante de $\omega_K$.

Le terme évolutif est approximé à l'aide du schéma d'Euler implicite

$$\int_{t^n}^{t^{n+1}}\int_{\omega_K}\phi(x)\partial_t\Big(\rho(p)s\Big)\ \mathrm{d}x\,\mathrm{d}t \approx |\omega_K|\phi_K\Big(\rho(p_K^{n+1})\ s_K^{n+1} - \rho(p_K^n)\ s_K^n\Big).$$

Pour qu'on puisse obtenir un schéma stable, le flux convectif est discrétisé en utilisant un schéma décentré. Un tel schéma est essentiellement décrit par la donnée d'un flux numérique $G_g$. Ainsi, l'approximation de la partie convective devient

$$-\int_\sigma \rho(p)M_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \approx \rho_{KL}^{n+1}\ \Lambda_{KL}^T\ G_g(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p),$$

où le coefficient $\rho_{KL}^{n+1}$ est la moyenne de la densité sur l'interface $\sigma_{KL}^T$

$$\rho_{KL}^{n+1} := \begin{cases} \dfrac{1}{p_K^{n+1} - p_L^{n+1}} \displaystyle\int_{p_L^{n+1}}^{p_K^{n+1}} \rho(z)\,\mathrm{d}z, & \text{si } \; p_L^{n+1} \neq p_K^{n+1} \\ \rho(p_K^{n+1}), & \text{sinon} \end{cases}, \tag{0.2.7}$$

et $\Lambda_{KL}^T$ représente le coefficient de transmissivité à travers l'interface $\sigma_{KL}^T$ des deux volumes de contrôles $K$ et $L$ dans le triangle $T$

$$\begin{cases} \Lambda_{KL}^T := -\displaystyle\int_T \Lambda(x)\nabla\varphi_K \cdot \nabla\varphi_L\,\mathrm{d}x = \Lambda_{LK}^T, & \text{pour } \; K \neq L \\ \Lambda_{KK}^T := \displaystyle\sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T = \int_T \Lambda(x)\nabla\varphi_K \cdot \nabla\varphi_K\,\mathrm{d}x & \text{sinon} \end{cases}. \tag{0.2.8}$$

Le signe de ce coefficient est très important pour l'analyse du schéma. A cet effet, nous allons dans ce chapitre supposer que toutes les transmissivités sont positives. Cette hypothèse n'est pas forcement vérifiée pour n'importe quel maillage et tenseur. Par contre elle l'est dans des cas particuliers. A titre d'exemple, si tous les angles des triangles sont inférieurs à $\pi/2$ et si le tenseur de perméabilité se réduit à un scalaire alors toutes les transmissivités sont positives.

La fonction $G_\alpha$ (pour $\alpha = g, w$) est un flux numérique, qui prend trois arguments et vérifie les propriétés suivantes :

($C_1$) $G_\alpha(\cdot, b; c)$ est croissante $\forall b, c \in \mathbb{R}$ et $G_\alpha(a, \cdot; c)$ est décroissante $\forall a, c \in \mathbb{R}$,

($C_2$) $G_\alpha(a, b; c) = -G_\alpha(b, a; -c)$ $\forall a, b, c \in \mathbb{R}$,

($C_3$) $G_g(a, a; c) = M_g(a)(-c)$, et $G_w(a, a; c) = M_w(a)c$, $\forall a, c \in \mathbb{R}$. En outre, il existe une constante $C$ telle que :

$$\forall\, a, b, c \in \mathbb{R} \quad |G_\alpha(a, b; c)| \leq C(|a| + |b|)|c|,$$

($C_4$) Il existe $\beta > 0$ tel que

$$\forall\, a, b, c \in \mathbb{R} \quad (G_w(a, b; c) - G_g(a, b; c))c \geq \beta|c|^2.$$

($C_5$) Il existe un module de continuité $\eta : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ tel que l'inégalité suivante ait lieu

$$\forall\, a, b, c, a', b' \in \mathbb{R} \quad |G_\alpha(a, b; c) - G_\alpha(a', b'; c)| \leq \; \eta\,(|a - a'| + |b - b'|)|c|.$$

Un exemple classique de choix de la fonction $G_\alpha$ qui répond aux hypothèses ($C_1$)-($C_5$) est celui d'Engquist–Osher. Pour le construire, il suffit de décomposer la mobilité $M_\alpha$ en sa partie croissante $M_{\alpha\,\uparrow}$ et sa partie décroissante $M_{\alpha\,\downarrow}$. On écrit alors

$$G_\alpha(a, b; c) = c^+\Big(M_{\alpha\,\uparrow}(a) + M_{\alpha\,\downarrow}(b)\Big) - c^-\Big(M_{\alpha\,\uparrow}(b) + M_{\alpha\,\downarrow}(a)\Big),$$

où $c^+ = \max(c, 0)$ et $c^- = -\min(c, 0)$. Mais, la mobilité du gaz est croissante tandis que celle de l'eau est décroissante selon l'hypothèse ($H_2$). Par conséquent, la formule précédente se réduit à

$$G_g(a, b; c) = -M_g(b)c^+ + M_g(a)c^-, \quad G_w(a, b; c) = M_w(b)c^+ - M_w(a)c^-.$$

On exploite à nouveau l'hypothèse $(H_2)$ pour voir que les propriétés $(C_1)$-$(C_5)$ sont bien satisfaites.

Maintenant, on va décrire l'approximation du terme capillaire. Comme il s'agit d'un terme de diffusion, ce dernier est naturellement approché par un schéma centré pour des raisons encore une fois de stabilité.

$$-\int_\sigma \rho(p)\ \Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \approx -\rho_{KL}^{n+1}\ \Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1}).$$

Finalement, on approxime le terme source par sa moyenne sur le volume de contrôle $K$,

$$\int_{t^n}^{t^{n+1}}\int_{\omega_K}\rho(p)\ s\ q^P\,\mathrm{d}x\,\mathrm{d}t = \delta t\ \omega_K\ \rho(p_K^{n+1})\ s_K^{n+1}\ q_{P,K}^{n+1}. \tag{0.2.9}$$

**Le schéma numérique**

On assemble toutes les approximations qu'on vient de proposer afin d'acquérir le schéma CVFE pour le système (0.2.1)-(0.2.4)

$$p_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} p_0(x)\,\mathrm{d}x, \qquad \forall K\in\mathcal{V}, \tag{0.2.10}$$

$$s_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} s_0(x)\,\mathrm{d}x, \qquad \forall K\in\mathcal{V}. \tag{0.2.11}$$

$$\begin{aligned}\phi_K\Big(\rho(p_K^{n+1})\ s_K^{n+1}-\rho(p_K^n)\ s_K^n\Big) &+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\rho_{KL}^{n+1}\ \Lambda_{KL}^T G_g(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)\\ &- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\rho_{KL}^{n+1}\ \Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1})\\ &+ \delta t\ \rho(p_K^{n+1})\ s_K^{n+1}\ q_{P,K}^{n+1}=0,\end{aligned} \tag{0.2.12}$$

$$\begin{aligned}\phi_K\Big(s_K^{n+1}-s_K^n\Big) &+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T\ G_w(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)\\ &- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1})\\ &+ \delta t\ (s_K^{n+1}-1)\ q_{P,K}^{n+1} = -\delta t q_{I,K}^{n+1},\quad \forall n=0,\dots,N-1,\quad \forall K\in\mathcal{V}, x_K\notin\Gamma_D.\end{aligned} \tag{0.2.13}$$

**Résultats obtenus**

.

Sous la positivité des transmissivités on obtient les résultats suivants.

**Lemma 0.2.1.** *(Principe du maximum discret) Soit $n=0,\dots,N-1$. Si $(p_K^{n+1},s_K^{n+1})_{K\in\mathcal{V}}$ est une solution du schéma (0.2.10)-(0.2.13) alors la saturation*

$$0\le s_K^{n+1}\le 1 \quad \textit{pour tout volume de contrôle } \omega_K.$$

L'idée de la preuve est standard. Pour établir que $s_K^{n+1} \geq 0$, on considère tout d'abord un volume de contrôle $\omega_K$ tel que $s_K^{n+1} = \min_{L\in\mathcal{V}}(u_L^{n+1})$, puis on multiplie l'équation discrète du gaz (0.2.12) par $(s_K^{n+1})^-$. La monotonie et la consistance du flux numérique $G_g$ renforce la monotonie de la partie convective. En outre, la monotonie du terme dispersif découle de la positivité des transmissivités. Finalement, on procède de la même façon pour démonter que $s_K^{n+1} \leq 1$. Il suffit justement de choisir un $\omega_K$ avec $s_K^{n+1} = \max_{L\in\mathcal{V}}(u_L^{n+1})$, multiplier l'équation (0.2.13) par $(s_K^{n+1}-1)^+$ et suivre le même raisonnement.

**Proposition 0.2.1.** *(Estimations a priori) Soit $n = 0, \ldots, N-1$. Si le couple $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$ est une solution du système (0.2.10)-(0.2.13) alors ils existent deux constantes $C_p$ et $C_\xi$ dépendantes de $\Omega, \mathfrak{T}, p_0, s_0, m_0, q^P, q^I, \underline{\Lambda}, \overline{\Lambda}$ telles que :*

$$\sum_{n=0}^{N-1} \delta t \left\| p_h^{n+1} \right\|_{X_h}^2 \leq C_p, \tag{0.2.14}$$

*et*

$$\sum_{n=0}^{N-1} \delta t \left\| \xi(s_h^{n+1}) \right\|_{X_h}^2 \leq C_\xi. \tag{0.2.15}$$

On présente en premier la preuve de l'inégalité (0.2.14). Pour ceci, l'équation du gaz est multipliée par $p_K^{n+1}|\omega_K|$ et celle de l'eau par $g(p_K^{n+1})|\omega_K|$ où $g\prime(p) = -\rho(p)$. On ajoute les équations résultantes et on somme sur tous les sommets $K \in \mathcal{V}$ et $n = 0, \ldots, N-1$. On fait appel à la conservation du schéma pour pouvoir utiliser la formule d'intégration par parties. On introduit le choix crucial du coefficient $\rho_{KL}^{n+1}$ qui permet de découpler la corrélation de $p$ et $\xi$. Grâce à la positivité des coefficients $\Lambda_{KL}^T$ on prouve une sorte de coercivité sur la pression globale.

Le traitement du terme source repose sur la sous-linéarité de $g$ et l'inégalité de Poincaré discrète.

En combinant toutes ces estimations et en utilisant l'inégalité de Young on arrive à majorer le gradient discret de $p$.

Pour la démonstration de la deuxième inégalité, on multiplie l'équation discrète de l'eau par $\xi(s_K^{n+1})$, on somme sur tous les sommets $K \in \mathcal{V}$ et $n = 0, \ldots, N-1$. Pour conclure, on se sert des techniques similaires qu'auparavant ainsi que l'estimation d'énergie sur la pression globale.

**Lemma 0.2.2.** *(Existence d'une solution) Le système d'équations (0.2.3)-(0.2.4) sous les hypothèses $(H_0)$-$(H_6)$ admet un couple de solution $(s_h, p_h)$.*

Le principe du maximum et les estimations uniformes sur les gradients permettent d'établir le résultat d'existence grâce au théorème de monotonie.

**Theorem 0.2.1.** *Sous les hypothèses $(H_0)$-$(H_6)$, soit $(\mathcal{T}_m)_m$ une suite régulière de maillages de $\Omega$. On suppose en plus que les coefficients de transmissivités sont tous positifs. Soit $(p_{h,\delta t}, s_{h,\delta t})$ une suite de solutions du schéma numérique (0.2.12)-(0.2.13). Quand $(h, \delta t)$ tend vers $(0,0)$ cette suite admet une sous-suite convergente vers une solution $(s,p)$ faible du problème (0.2.1)-(0.2.4) au sens de la Définition 0.2.1.*

Après avoir établi des estimations sur les translatés en espace et en temps sur la suite $\xi(s_{h,\delta t})$, on peut extraire, selon le fameux théorème de Kolmogorov, une sous-suite encore notée $(p_{h,\delta t}, s_{h,\delta t})$ telle qu'on ait la convergence forte du terme évolutif (en temps) pour chaque phase. Par contre on

dispose seulement de la convergence faible à la fois de la pression globale et des gradients discrets. Le passage à la limite suit des idées classiques, mais à un moment donné on doit prendre en compte le fait que la suite $\rho(p_{h,\delta t})\xi(s_{h,\delta t})$ converge fortement. Cette remarque permet davantage de remplacer l'absence de la convergence forte de $p_{h,\delta t}$.

A la fin de ce chapitre nous présentons les résultats de simulations numériques sur un maillage triangulaire dont tous les angles sont aigus avec une perméabilité scalaire afin d'assurer la positivité des coefficients de transmissivité. Le système algébrique non linéaire issu du schéma numérique est résolu par la méthode de Newton-Raphson.

Le premier test est pris d'un benchmark. On voit que la méthode est d'ordre un, ce qui est tout à fait naturel pour les schémas décentrés. Dans le deuxième test, nous nous intéressons à simuler la récupération secondaire du gaz. Nous retenons de cette application que l'approche CVFE manifeste un comportement acceptable et semblable à la réalité physique dans le cas d'un milieu poreux isotrope. En particulier, quand le terme capillaire est négligeable i.e. la pression du gaz est identique à celle de l'eau, la méthode permet alors de capter le cas purement hyperbolique du système diphasique sans aucune oscillation.

### 0.2.2 Chapitre 3 : Un schéma volumes finis/éléments finis positif pour un modèle diphasique compressible dégénéré en milieux poreux anisotropes

Dans ce chapitre nous allons proposer un schéma beaucoup plus général que celui du chapitre 3. En effet, nous avons vu que l'étude du schéma CVFE est essentiellement basée sur la positivité de transmissivités. Cette hypothèse est très restrictive et elle ne tolère pas de maillages généraux, en particulier quand certains angles sont obtus. En outre que la nature du maillage, l'anisotropie du milieu poreux peut rendre l'hypothèse en question inutile. Par conséquent, on pourra perdre le principe du maximum qui est une propriété primordiale qu'on souhaite établir.

Pour faire en sorte que ce problème de positivité soit surmonté nous allons corriger le schéma du chapitre précédent. La première idée est de reformuler le système continu d'une façon équivalente comme ci-après

$$\partial_t(\phi\rho(p)s) - \mathrm{div}\left(\rho(p)M(s)f_g(s)\Lambda\nabla p\right) - \mathrm{div}\left(\rho(p)\gamma(s)\Lambda\nabla s\right) + \rho(p)sq^P = 0, \tag{0.2.16}$$

$$\partial_t(\phi s) + \mathrm{div}\left(M(s)f_w(s)\Lambda\nabla p\right) - \mathrm{div}\left(\gamma(s)\Lambda\nabla s\right) + sq^P = q^P - q^I, \tag{0.2.17}$$

où $M$ n'est autre que la mobilité totale $M = M_g + M_w$ qui est bien évidement une fonction minorée uniformément loin de 0. La fonction $f_\alpha := \dfrac{M_\alpha}{M}$ représente le flux fractionnaire de la phase $\alpha$. Ainsi, sa monotonie est induite par celles des mobilités.

Nous allons garder les mêmes hypothèses sur le modèle mathématique ainsi que les notations du maillage et celles des fonctions discrètes. Cependant, on n'impose pas de restrictions majeures sur les mailles. Cela veut dire que les angles des triangles peuvent être obtus comme ils peuvent être aigus. De plus, la perméabilité du milieu peut être une matrice pleine dont le ratio d'anisotropie peut considérablement varier. Il en résulte que les coefficients de transmissivité sont désormais dans $\mathbb{R}$ et non dans $\mathbb{R}^+$.

**Le schéma CVFE positif pour le modèle diphasique compressible**

Comme dans le chapitre précédent nous allons présenter l'approximation de l'équation du gaz à l'aide du schéma CVFE positif. On fait de même pour l'équation de la phase eau.

Soient $\omega_K$ une maille duale et $n = 0, \cdots, N-1$. On intègre l'équation (0.2.16) sur $(t^n, t^{n+1}) \times \omega_K$ et on applique le formule de Gauss-Green

$$\begin{aligned}\int_{t^n}^{t^{n+1}} \int_{\omega_K} \phi(x)\partial_t\Big(\rho(p)s\Big)\,\mathrm{d}x - \sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_\sigma \rho(p)M(s)f_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t \\ - \sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_\sigma \rho(p)\,\Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t \\ + \int_{t^n}^{t^{n+1}}\int_{\omega_K}\rho(p)sq^P\,\mathrm{d}x\,\mathrm{d}t = 0. \qquad (0.2.18)\end{aligned}$$

Ici, on approche le terme en temps par le schéma d'Euler implicite

$$\int_{t^n}^{t^{n+1}} \int_{\omega_K} \phi(x)\partial_t\Big(\rho(p)s\Big)\,\mathrm{d}x\,\mathrm{d}t \approx |\omega_K|\phi_K\Big(\rho(p_K^{n+1})\,s_K^{n+1} - \rho(p_K^n)\,s_K^n\Big).$$

Puis, le terme de convection est discrétisé par un schéma décentré i.e.

$$-\int_\sigma \rho(p)M_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \approx \rho_{KL}^{n+1}M_T^{n+1}G_g\Big(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\delta_{KL}^{n+1}p\Big). \qquad (0.2.19)$$

$\rho_{KL}^{n+1}$ est donnée par l'expression (0.2.20).

$$\frac{1}{\rho_{KL}^{n+1}} := \begin{cases} \dfrac{1}{p_K^{n+1}-p_L^{n+1}}\displaystyle\int_{p_L^{n+1}}^{p_K^{n+1}}\frac{1}{\rho(z)}\,\mathrm{d}z, & \text{si } \ p_L^{n+1}\neq p_K^{n+1} \\ \dfrac{1}{\rho(p_K^{n+1})}, & \text{sinon} \end{cases} \qquad (0.2.20)$$

La quantité $M_T^{n+1}$ désigne l'approximation de la mobilité totale sur le triangle $T$. Elle est exprimée par un schéma centré

$$M_T^{n+1} = \frac{1}{\#\mathcal{V}_T}\Big(\sum_{K\in\mathcal{V}_T} M(s_K^{n+1})\Big).$$

On rappelle également que $\Lambda_{KL}^T \in \mathbb{R}$ s'écrit

$$\begin{cases} \Lambda_{KL}^T := -\displaystyle\int_T \Lambda(x)\nabla\varphi_K\cdot\nabla\varphi_L\,\mathrm{d}x = \Lambda_{LK}^T, \quad \text{pour} \quad K\neq L \\ \Lambda_{KK}^T := \displaystyle\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T \quad \text{sinon} \end{cases}. \qquad (0.2.21)$$

La fonction $G_g$ signifie un flux numérique à trois arguments. Son expression peut être définie à partir de celle de $G_w$. Soit alors $g_w(a,b)$ un flux monotone quelconque de $f_w$ dans le sens suivant :

($g_1$) $g_w$ est une fonction croissante par rapport à la première variable et elle est décroissante par rapport à la deuxième variable.

($g_2$) $g_w$ est consistent i.e. $g_w(a,a) = f_w(a), \quad \forall a \in R.$

($g_3$) $g_w$ est lipschitzienne par rapport à $a$ et $b$.

On définit maintenant

$$G_w(a,b,c) = g_w(a,b)c^+ - g_w(b,a)c^-, \tag{0.2.22}$$
$$G_g(a,b,c) = G_w(a,b,c) - c. \tag{0.2.23}$$

Ce dernier choix de $G_g$ est remarquable. Certes, on s'en servira pour contrôler surtout le gradient discret de la pression globale. On voit que les flux numériques $G_g$ et $G_w$ sont fortement liés au choix de la fonction $g_w$. La construction de $g_w$ est classique, il suffit de prendre en considération la technique décentré pour en déduire que

$$g_w(a,b) = f_w(b), \quad \forall a,b \in \mathbb{R}.$$

D'après l'hypothèse sur le flux fractionnaire $f_w$ il vient que $g_w$ vérifie les propriétés ($g_1$)-($g_3$).

Ensuite, la discrétisation du terme de diffusion est complètement différente de celle que nous avons exposé au Chapitre 1. En fait, on traite la partie elliptique comme si elle était hyperbolique. Elle est alors approximée par

$$\int_\sigma \rho(p)\Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \approx \rho_{KL}^{n+1}\gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1} - s_K^{n+1}), \tag{0.2.24}$$

où $\gamma_{KL}^{n+1}$ est défini par le schéma de Goudnov

$$\gamma_{KL}^{n+1} := \begin{cases} \max\limits_{s\in I_{KL}^{n+1}} \gamma(s) & \text{si } \Lambda_{KL}^T \geq 0 \\ \min\limits_{s\in I_{KL}^{n+1}} \gamma(s) & \text{sion} \end{cases}, \tag{0.2.25}$$

et

$$I_{KL}^{n+1} := \left[\min(s_K^{n+1}, s_L^{n+1}), \max(s_K^{n+1}, s_L^{n+1})\right].$$

Enfin, on approche le terme source par sa moyenne.

Pour résumer, le schéma CVFE positif pour le modèle diphasique consiste à trouver $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$, pour $n = 0, \cdots, N-1$ qui résolvent le système discret suivant

$$p_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} p^0(x)\,\mathrm{d}x, \qquad \forall K \in \mathcal{V}, \tag{0.2.26}$$
$$s_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} s^0(x)\,\mathrm{d}x, \qquad \forall K \in \mathcal{V}. \tag{0.2.27}$$

$$\begin{aligned} \phi_K\Big(\rho(p_K^{n+1})\,s_K^{n+1} - \rho(p_K^n)\,s_K^n\Big) &+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\,M_T^{n+1}G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\,\delta_{KL}^{n+1}p) \\ &- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\,\gamma_{KL}^{n+1}\,\Lambda_{KL}^T(s_L^{n+1} - s_K^{n+1}) \\ &+ \delta t\,\rho(p_K^{n+1})\,s_K^{n+1}\,q_{P,K}^{n+1} = 0, \end{aligned} \tag{0.2.28}$$

$$\phi_K\left(s_K^{n+1}-s_K^n\right)+\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} M_T^{n+1}\,G_w(s_K^{n+1},s_L^{n+1};\Lambda_{KL}^T\,\delta_{KL}^{n+1}p)$$
$$-\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\gamma_{KL}^{n+1}\,\Lambda_{KL}^T(s_L^{n+1}-s_K^{n+1})$$
$$+\delta t\,(s_K^{n+1}-1)\,q_{P,K}^{n+1}=-\delta t q_{I,K}^{n+1},\quad \forall n=0,\dots,N-1,\quad \forall K\in\mathcal{V},x_K\notin\Gamma_D. \tag{0.2.29}$$

Du point de vue numérique, l'avantage est que la méthode n'incorpore pas la fonction de Kirchoff $\xi(s)$. Ceci implique un gain significatif en terme de performance dans l'implémentation du schéma car $\xi$ est une primitive d'une fonction non linéaire. En effet, on considère le schéma du Chapitre 1. Dans la mise en œuvre de son terme elliptique, on est amené à approximer une intégrale pour chaque valeur de saturation. Quand le maillage devient de plus en plus fin, le solveur demande beaucoup plus de temps à cause du calcul des intégrales. En revanche, dans le cas du schéma positif, on n'a pas ce problème.

Du point de vue analyse, l'avantage est que les résultats que nous allons présenter dans la suite, à savoir la bornitude de la saturation, les estimations d'énergies, sont indépendants de la triangulation choisie ainsi que du tenseur de perméabilité considéré.

**Analyse du schéma CVFE positif**

Au cours de l'analyse de convergence du schéma CVFE positif (0.2.26)-(0.2.29) on fait largement appel à la théorie des éléments finis que celle des volumes finis. Pour en savoir plus, nous allons utiliser par exemple le lemme suivant dans plusieurs endroits de preuves des résultats ci-après.

**Lemma 0.2.3.** *On considère la fonction* $\psi_{\mathcal{T}}=\sum_{K\in\mathcal{V}}\psi_K\varphi_K$, *où* $(\varphi_K)_{K\in\mathcal{V}}$ *est la famille de fonctions de base* $\mathbb{P}_1$. *Alors, il existe* $C_0=C_0(\Lambda,\theta_{\mathcal{T}})$ *tel que*

$$\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\left|\Lambda_{KL}^T\right|(\psi_K-\psi_L)^2\le C_0\int_\Omega\Lambda\nabla\psi_{\mathcal{T}}\cdot\nabla\psi_{\mathcal{T}}\,\mathrm{d}x. \tag{0.2.30}$$

**Lemma 0.2.4.** *(Principe du maximum discret) Soit* $n=0,\dots,N-1$. *On suppose que le système* (0.2.26)-(0.2.29) *possède une solution du schéma alors la saturation du gaz est bornée i.e.*

$$0\le s_K^{n+1}\le 1,\quad \forall K\in\mathcal{V}.$$

La preuve de cet énoncé est légèrement différente de celle de la section précédente. En effet, comme certaines transmissivités pourraient être de signes quelconques, le terme diffusif du schéma peut engendrer des saturations négatives. L'introduction du schéma de Godunov permet de corriger cette insuffisance et renforce sa monotonie en utilisant la dégénérescence de la fonction capillaire $\gamma$. Dans le terme convectif discret, le coefficient $\Lambda_{KL}^T$ est mis à l'intérieur de l'expression du flux numérique. Le schéma décentré prend en compte son signe ce qui permet de renforcer la monotonie du terme hyperbolique.

**Proposition 0.2.2.** *(Estimations d'énergie) Soit* $n=0,\dots,N-1$. *Soit* $(p_K^{n+1},s_K^{n+1})_{K\in\mathcal{V}}$ *une solution éventuelle du système* (0.2.26)-(0.2.29) *alors ils existent deux constantes* $C_p'$ *et* $C_\xi'$ *dépendantes de* $\Omega,\mathfrak{T},p^0,s^0,m_0,q^P,q^I,\underline{\Lambda},\overline{\Lambda},\theta_{\mathcal{T}}$ *telles que :*

$$\sum_{n=0}^{N-1}\delta t\left\|p_h^{n+1}\right\|_{X_h}^2\le C_p', \tag{0.2.31}$$

*et*

$$\sum_{n=0}^{N-1} \delta t \left\| \xi(s_h^{n+1}) \right\|_{X_h}^2 \leq C'_\xi. \tag{0.2.32}$$

La preuve de l'estimation à priori sur la pression globale est préformée comme suite. On multiplie l'équation du gaz par la fonction test $|\omega_K|\, g(p_K^{n+1})$ et celle de l'eau par $-|\omega_K|\, p_K^{n+1}$ où $g'(p) = 1/\rho(p)$. On effectue l'addition des deux équations et on somme sur tous les sommets $K$ du maillage et $n = 0, \ldots, N-1$. Après avoir introduit la formule de l'intégration par partie discrète, chaque terme est majoré ou minoré proprement. Aussi, on utilise Lemme (0.2.3) et le principe du maximum pour conclure. Idem pour la deuxième inégalité.

**Theorem 0.2.2.** *Supposons* $(H_0)$*-*$(H_6)$ *et soit* $(\mathcal{T}_{h,\delta t})_{h,\delta t}$ *une suite de maillages triangulaires du domaine* $\Omega$ *dont la régularité est uniformément bornée. On considère* $(p_{h,\delta t}, s_{h,\delta t})$ *une suite de solutions du schéma* (0.2.26)*-*(0.2.29)*. Étant donné que* $(h, \delta t)$ *tend vers* $(0,0)$ *cette suite est convergente à une sous-suite près vers une solution* $(p, s)$ *faible du problème continu (0.2.1)-(0.2.4) au sens de la Définition (0.2.1).*

On rappelle que la gravité est négligeable pour que l'énoncé soit cohérent avec la définition de la solution faible. La démonstration du théorème de convergence est grandement basée sur des résultats de compacité. On établit alors des estimations uniformes sur les translatés en espace et en temps sur la fonction $\xi(s_{h,\delta t})$. On applique à nouveau le théorème de Kolmogorov pour montrer la convergence forte de la suite $(s_{h,\delta t})$, à une sous-suite près. Le passage à la limite découle également de la convergence forte des termes évolutifs discrets. Toutes ces démarches sont détaillées dans ce chapitre 3.

Dans la fin du chapitre nous présentons des simulations numériques en dimension deux. Elles mettent en évidence l'effet de l'anisotropie et son influence sur le déplacement de l'eau. On vérifie en outre que la saturation ne dépasse pas les bornes physiques.

### 0.2.3 Chapitre 4 : Convergence d'un schéma DDFV monotone pour les équations parabolique non linéaires dégénérées

Ce chapitre a comme objectif de développer et d'analyser une méthode de volumes finis de type "Discrete Duality Finite Volume" (DDFV) pour une équation de diffusion non linéaire. Les points forts de cette discrétisation proposée sont la possibilité d'utiliser des maillages et des perméabilités quelconques. Elle est en plus inconditionnellement coercitive.

Dans ce chapitre on met l'accent notamment sur le principe du maximum qui n'est pas connu en général pour les schéma DDFV existants dans la littérature. La convergence du schéma est rigoureusement prouvée à l'aide des outils classiques. La mise en œuvre montre qu'elle est d'ordre deux indépendamment du maillage et du tenseur. Cela rend notre approche originale et novatrice dans le contexte des méthodes DDFV.

### Équation de diffusion non linéaire

Soit $\Omega$ un ouvert borné connexe polygonal de $\mathbb{R}^d$ et $\mathfrak{T}$ un réel positif. Le problème auquel on s'intéresse est :

$$\begin{cases} \partial_t u - \nabla \cdot \Big( f(u) \Lambda \nabla u \Big) = 0 & \text{dans} \quad Q_{\mathfrak{T}} := \Omega \times (0, \mathfrak{T}) \\ u = 0 & \text{sur} \quad \partial\Omega \times (0, \mathfrak{T}), \\ u(\cdot, 0) = u^0 & \text{dans} \quad \Omega \end{cases} \tag{0.2.33}$$

où $f$ est une fonction positive, $\Lambda$ une matrice carrée d'ordre $d$ et $u^0$ la donnée initiale. La fonction $u$ désigne l'inconnue principale du problème. Du point de vu physique, cette quantité décrit, par exemple, une concentration, une saturation ou une température. L'étude théorique (existence et unicité de la solution) de ce problème modèle est classique. Nous nous intéressons alors à approximer les solutions de ce modèle. A cet effet, nous précisons tout d'abord les hypothèses nécessaires pour l'investigation du schéma.

($\mathrm{A}_1$) La donnée initiale $u^0$ est dans $L^\infty(\Omega)$ avec $0 \leq u \leq 1$.

($\mathrm{A}_2$) La fonction $f$ appartint à $\mathcal{C}^0([0,1], \mathbb{R})$ telle que :

$$\begin{cases} f(u) > 0, & \text{si } u \in (0,1), \\ f(u) = 0, & \text{si } u \in \mathbb{R} \backslash (0,1). \end{cases}$$

On notera par $F$ (resp. $\xi$) la primitive de $f$ (resp. $v := \sqrt{f}$) qui s'annule en 0. En conséquent, $F, \xi$ sont lipschitziennes. Par ailleurs, on suppose que la fonction $v$ est absolument continue.

($\mathrm{A}_3$) Le tenseur $\Lambda : \Omega \to \mathcal{S}_d(\mathbb{R})$, où $\mathcal{S}_d(\mathbb{R})$ est l'espace des matrices symétriques d'ordre $d$, est supposé dans $L^\infty(\Omega)^{d \times d}$. Il vérifie en plus la condition d'ellipticité uniforme :

$$\underline{\Lambda} \, |\zeta|^2 \leq \Lambda(x) \zeta \cdot \zeta \leq \overline{\Lambda} \, |\zeta|^2 , \text{ pour tout } \zeta \in \mathbb{R}^d \text{ and p.p. } x \in \Omega,$$

où $\underline{\Lambda}, \overline{\Lambda}$ sont des constantes positives.

D'après la continué absolue de la fonction $v$, on est en droit de donner un sens au schéma Engquist-Osher qu'on définira plus loin.

Le cadre fonctionnel dans lequel la solution vit est l'espace classique de Sobolev

$$H_0^1(\Omega) = \{ v \in H^1(\Omega) \ / \ v = 0 \text{ sur } \partial\Omega \}.$$

On rappelle que c'est un espace de Hilbert muni de la norme

$$||v||_{H_0^1(\Omega)} = ||\nabla v||_{L^2(\Omega)^d}.$$

On définit maintenant la notion de la solution faible de l'équation de diffusion non linéaire.

**Definition 0.2.2.** *Une fonction mesurable $u : Q_{\mathfrak{T}} \longrightarrow [0,1]$ est dite solution faible du problème (0.2.33) si*

$$\xi(u) \in L^2(0, \mathfrak{T}; H_0^1(\Omega)),$$
$$- \int_{Q_{\mathfrak{T}}} u \, \partial_t \varphi \, \mathrm{d}x \, \mathrm{d}t + \int_{Q_{\mathfrak{T}}} \Lambda \nabla F(u) \cdot \nabla \varphi \, \mathrm{d}x \, \mathrm{d}t - \int_\Omega u^0 \ \varphi(\cdot, 0) \, \mathrm{d}x = 0, \quad \forall \varphi \in \mathcal{C}_c^\infty(\Omega \times [0, \mathfrak{T})).$$

**Maillages DDFV et notations**

Nous décrivons les divers maillages utilisés dans le cadre des méthodes DDFV et nous précisons aussi les notations adoptées dans ce chapitre.

Un maillage DDFV $\mathcal{T}$ est défini par trois maillages différents : primal, dual et diamant, notés par $\overline{\mathfrak{M}} = \mathfrak{M} \cup \partial\mathfrak{M}$, $\overline{\mathfrak{M}^*} = \mathfrak{M}^* \cup \partial\mathfrak{M}^*$ et $\mathfrak{D}$ respectivement. Le maillage primal intérieur $\mathfrak{M}$ est la donnée d'une famille finie de polygones $(K)_{K\in\mathfrak{M}}$ qui couvrent le domaine $\Omega$ i.e. $\bigcup_{K\in\mathfrak{M}} \overline{K} = \overline{\Omega}$. Les sous-ensembles $K$ sont souvent appelés des volumes de contrôle et ils ne sont pas forcement convexes. Le maillage primal du bord $\partial\mathfrak{M}$ est constitué des arêtes de $\overline{\mathfrak{M}}$ qui se trouvent sur le bord $\partial\Omega$. Ces arêtes sont considérées comme des mailles dégénérées. Ensuite, à chaque volume de contrôle $K$ on lui associe un et un seul centre $x_K$. Par exemple $x_K$ pourrait être choisi comme le centre de masse de $K$. On note $\mathcal{V}$ l'ensemble de tous ces centres ainsi que $\mathcal{V}^*$ l'ensemble de tous les sommets de $\mathfrak{M}$.

La construction du maillage dual $\overline{\mathfrak{M}^*}$ est basée sur celle du maillage primal $\overline{\mathfrak{M}}$, en particulier les sommets et les centres. Pour chaque somment $x_{K^*} \in \mathcal{V}^*$ on lui associe un unique polygone $K^*$ dit volume de contrôle dual dont les sommets sont les centres des mailles primales qui partagent le même point $x_{K^*}$. Si le centre $x_{K^*} \in \mathcal{V}^* \cap \partial\Omega$ alors il devient sommet de $K^*$ en plus des centres des volumes de contrôle ayant $x_{K^*}$ comme sommet. Les arrêtes des volumes de contrôle du maillage dual sont constitués par les segments constitués par les centres du maillage primal.

On désigne par $\mathcal{E}$ (resp. $\mathcal{E}^*$ ) l'ensemble des arêtes primales (resp. duales). Deux volumes de contrôle sont dits voisins s'ils partagent au moins une arête. Autrement-dit, si $K, L \in \overline{\mathfrak{M}}$ (resp. $K^*, L^* \in \overline{\mathfrak{M}^*}$) sont adjacents alors il existe au moins une arête $\sigma \in \mathcal{E}$ (resp. $\sigma^* \in \mathcal{E}^*$) telle que $\sigma = \overline{K} \cap \overline{L}$ (resp. $\sigma^* = \overline{K^*} \cap \overline{L^*}$).

Le maillage diamant $\mathfrak{D} = (\mathcal{D}_{\sigma,\sigma^*})_{(\sigma,\sigma^*)\in\mathcal{E}\times\mathcal{E}^*}$ est un recouvrement de $\Omega$ construit à partir des arêtes primales. Ainsi, pour $\sigma \in \mathcal{E}$ on lui associe un unique diamant $\mathcal{D} := \mathcal{D}_{\sigma,\sigma^*}$ qui est un polygone obtenu en joignant les extrémités de $\sigma$ et les centres des volumes de contrôle qui la partagent (dans sens horaire par exemple). Si $\sigma$ est incluse dans $\partial\Omega$ le diamant correspondant n'est autre qu'un triangle (Figure 2).

On pose $\mathcal{T} = (\mathfrak{M}, \overline{\mathfrak{M}^*})$. Pour $M \in \mathcal{T}$ on désigne donc par $m_M$ l'aire de $M$, $\mathcal{E}_M$ l'ensemble des arêtes de $M$, $\mathcal{D}_M$ l'ensemble constitué de diamants $\mathcal{D}_{\sigma,\sigma^*}$ tel que la mesure de Lebesgue $m(\mathcal{D}_{\sigma,\sigma^*} \cap \mathfrak{M}) > 0$ et $d_M$ le diamètre de $M$. Pour chaque diamant $\mathcal{D} \in \mathfrak{D}$ de sommets $(x_K, x_{K^*}, x_L, x_{L^*})$, on définit son centre $x_\mathcal{D}$ par l'intersection de ces principales diagonales, sa mesure par $m_\mathcal{D}$ et son diamètre par $d_\mathcal{D}$. La longueur de l'arête $e \in \mathcal{E} \cup \mathcal{E}^*$ est notée par $m_e$. La notation $\mathbf{n}_{\sigma K}$ (resp. $\mathbf{n}_{\sigma^* K^*}$) signifie la normale à $\sigma$ (resp. $\sigma^*$) sortante de $K$ (resp. $K^*$). D'une manière similaire, $\boldsymbol{\tau}_{\boldsymbol{K,L}}$ (resp. $\boldsymbol{\tau}_{\boldsymbol{K^*,L^*}}$) est la tangente à $\sigma$ (resp. $\sigma^*$) dirigée de $K$ (resp. $K^*$) à $L$ (resp. $L^*$). On considère $\alpha_\mathcal{D}$ l'angle entre les deux vecteurs $\boldsymbol{\tau}_{\boldsymbol{K,L}}$ et $\boldsymbol{\tau}_{\boldsymbol{K^*,L^*}}$.

Le pas du maillage $\mathcal{T}$ est défini par $h_\mathfrak{D} = \max\{d_\mathcal{D}, \mathcal{D} \in \mathfrak{D}\}$. On détermine en outre le nombre réel $\alpha_\mathcal{T} \in ]0, \frac{\pi}{2}]$ tel que

$$\sin(\alpha_\mathcal{T}) := \min_{\mathcal{D}\in\mathfrak{D}} |\sin(\alpha_\mathcal{D})| .$$

On considère $\rho_K$ (resp. $\rho_{K^*}$) le rayon de la plus grande boule incluse dans $K$ (resp. $K^*$) dont le

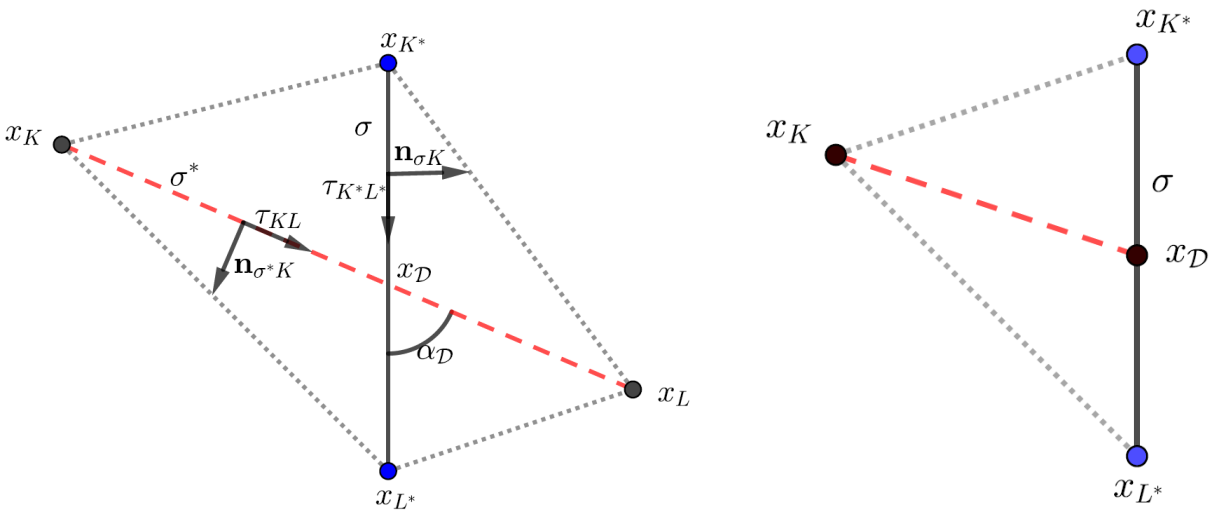



Figure 2: Exemple des diamants intérieur (gauche) et extérieur (droite).

centre est $x_K$ (resp. $x_{K^*}$). On définit la régularité du maillage $\mathrm{reg}(\mathcal{T})$ par

$$\mathrm{reg}(\mathcal{T}) = \max\left(\frac{1}{\sin(\alpha_{\mathcal{T}})}, \max_{\mathcal{D}\in\mathfrak{D}} \frac{h_{\mathfrak{D}}}{\sqrt{m_{\mathcal{D}}}}, \max_{K\in\mathfrak{M}} \frac{d_K}{\sqrt{m_K}}, \max_{K^*\in\mathfrak{M}^*} \frac{d_{K^*}}{\sqrt{m_{K^*}}}, \max_{K\in\mathfrak{M}}\Big(\frac{d_K}{\rho_K} + \frac{\rho_K}{d_K}\Big), \max_{K^*\in\overline{\mathfrak{M}^*}}\Big(\frac{d_{K^*}}{\rho_{K^*}} + \frac{\rho_{K^*}}{d_{K^*}}\Big)\right).$$

Cette quantité doit être bornée uniformément pour pouvoir démonter le théorème de convergence.

On désigne par $\mathbb{R}^{\#\mathcal{T}}$ l'espace des vecteurs $u_{\mathcal{T}}$ ayant la forme :

$$u_{\mathcal{T}} = \Big((u_K)_{K\in\mathfrak{M}}, (u_{K^*})_{K^*\in\overline{\mathfrak{M}^*}}\Big).$$

C'est un espace de Hilbert muni du produit scalaire

$$[\![u_{\mathcal{T}}, v_{\mathcal{T}}]\!]_{\mathcal{T}} = \frac{1}{2}\Big(\sum_{K\in\mathfrak{M}} m_K u_K v_K + \sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*} u_{K^*} v_{K^*}\Big), \quad \forall u_{\mathcal{T}}, v_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}.$$

Pour chaque $u_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}$ on construit deux fonctions discrètes $u_{\mathfrak{M}}$ et $u_{\overline{\mathfrak{M}^*}}$ telles que

$$u_{\mathfrak{M}} = \sum_{K\in\mathfrak{M}} u_K \mathbf{1}_K, \quad u_{\overline{\mathfrak{M}^*}} = \sum_{K^*\in\overline{\mathfrak{M}^*}} u_{K^*}\mathbf{1}_{K^*}.$$

On notera également $X_{\mathcal{T}}$ l'espace des fonctions $u_h = \frac{1}{2}\Big(u_{\mathfrak{M}} + u_{\overline{\mathfrak{M}^*}}\Big)$. Pour prendre en compte le temps on définit la reconstruction suivante : pour $n = 0, \cdots, N-1$ et $t \in (t^n, t^{n+1}]$

$$u_{h,\delta t}(x,t) = u_h(x)^{n+1}, \quad \text{où } u_h \in X_{\mathcal{T}}.$$

Soit $X_{\mathcal{T},\delta t} \subset L^1(Q_{\mathfrak{T}})$ l'ensemble de toutes ces fonctions discrètes. Par analogie, on donne les définitions des fonctions $u_{\mathfrak{M}_h,\delta t}, u_{\overline{\mathfrak{M}^*}_h,\delta t}$ qui sont constantes par morceaux.

La formule du gradient discret est parmi l'un des avantages d'utilisation de la méthode DDFV. Cet opérateur est défini sur le maillage diamant $\mathfrak{D}$. Son expression s'écrit sur chaque diamant $\mathcal{D} \in \mathfrak{D}$ dans la base $(\mathbf{n}_{\sigma K}, \mathbf{n}_{\sigma^* K^*})$ par

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{\sin(\alpha_{\mathcal{D}})}\Big(\frac{u_L - u_K}{m_{\sigma^*}}\mathbf{n}_{\sigma K} + \frac{u_{L^*} - u_{K^*}}{m_\sigma}\mathbf{n}_{\sigma^* K^*}\Big).$$

Afin de rendre notre approche lisible et plus compacte on introduit les notations suivantes

$$\begin{aligned}
&a_{KL} := \frac{1}{\sin(\alpha_{\mathcal{D}})}\frac{m_\sigma}{m_{\sigma^*}}\,\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma K} > 0, \quad \eta^{\mathcal{D}}_{\sigma\sigma^*} := \frac{1}{\sin(\alpha_{\mathcal{D}})}\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} \in \mathbb{R},\\
&g_M := g(u_M), \quad \forall M \in \{K, L, K^*, L^*\} \text{ and } g \in \{F, \xi\},\\
&\delta_{LK}u := u_L - u_K, \quad \delta_{L^*K^*}u := u_{L^*} - u_{K^*},
\end{aligned}$$

où $\Lambda^{\mathcal{D}}$ signifie la moyenne

$$\Lambda^{\mathcal{D}} = \frac{1}{m_{\mathcal{D}}}\int_{\mathcal{D}} \Lambda(x)\,\mathrm{d}x.$$

Lorsque les paramètres de la discrétisation tendent vers 0, les suites $(u_{\mathfrak{M}_h,\delta t})$ et $\underline{u_{\mathfrak{M}_h,\delta t}}$ ne convergent pas forcement vers la même limite. Ceci mène à des problèmes dans le passage à la limite dans des termes non linéaires. Pour remédier cette difficulté on est obligé de pénaliser le schéma numérique par la fonction $\mathcal{P}^{\mathcal{T}}$ qui est définie de $\mathbb{R}^{\#\mathcal{T}}$ dans $\mathbb{R}^{\#\mathcal{T}}$, pour tout $u_{\mathcal{T}}$, par

$$\mathcal{P}^{\mathcal{T}} u_{\mathcal{T}} = \Big(\mathcal{P}^{\mathfrak{M}} u_{\mathcal{T}}, \mathcal{P}^{\mathfrak{M}^*} u_{\mathcal{T}}, \mathcal{P}^{\partial\mathfrak{M}^*} u_{\mathcal{T}}\Big),$$

où $\mathcal{P}^{\mathfrak{M}} u_{\mathcal{T}} = (\mathcal{P}_K u_{\mathcal{T}})_{K\in\mathfrak{M}}$, $\mathcal{P}^{\mathfrak{M}^*} u_{\mathcal{T}} = (\mathcal{P}_{K^*} u_{\mathcal{T}})_{K^*\in\mathfrak{M}^*}$ , $\mathcal{P}^{\partial\mathfrak{M}^*} u_{\mathcal{T}} = (\mathcal{P}_{K^*} u_{\mathcal{T}})_{K^*\in\partial\mathfrak{M}^*}$. Chaque composante s'écrit explicitement comme suite :

$$\mathcal{P}_K u_{\mathcal{T}} = \frac{1}{m_K}\frac{1}{h^{\varepsilon}_{\mathfrak{D}}}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K\cap K^*}\Big(F(u_K) - F(u_{K^*})\Big), \quad \forall\, K \in \mathfrak{M}, \tag{0.2.34}$$

$$\mathcal{P}_{K^*} u_{\mathcal{T}} = \frac{1}{m_{K^*}}\frac{1}{h^{\varepsilon}_{\mathfrak{D}}}\sum_{K\in\mathfrak{M}} m_{K\cap K^*}\Big(F(u_{K^*}) - F(u_K)\Big), \quad \forall\, K^* \in \overline{\mathfrak{M}^*}. \tag{0.2.35}$$

L'exposant $\varepsilon$ est dans l'intervalle $(0, 2)$. On signale que le choix de ce terme de pénalisation n'est pas optimale.

### Le schéma DDFV monotone pour l'équation de diffusion non linéaire

On présente brièvement ici la discrétisation de l'équation (0.2.33) sur le maillage primal et quant au maillage dual la démarche est analogue. Soient $n \in \{0, \dots, N-1\}$ et $K$ un volume de contrôle primal. On intègre sur $K\times]t^n, t^{n+1}]$, on utilise la formule de Gauss-Green et un schéma implicite en temps. On obtient donc

$$\int_{t^n}^{t^{n+1}}\int_K \partial_t u \;\mathrm{d}x\,\mathrm{d}t - \sum_{\sigma\in\mathcal{E}_K}\int_{t^n}^{t^{n+1}}\int_\sigma f(u)\,\Lambda\nabla u\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t = 0. \tag{0.2.36}$$

La dérivée temporelle est approchée grâce au schéma d'Euler

$$\int_{t^n}^{t^{n+1}}\int_K \partial_t u \;\mathrm{d}x\,\mathrm{d}t \approx\; m_K\Big(u_K^{n+1} - u_K^n\Big), \tag{0.2.37}$$

où $u_K^m$ est la moyenne de $u(.,t^m)$ sur $K$ pour $m = n, n+1$. Puis, le terme de diffusion non linéaire est approximée par un schéma centré et un schéma décentré au même temps

$$-\int_{t^n}^{t^{n+1}}\int_\sigma f(u)\Lambda\nabla u\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t \approx \delta t\,\Big(a_{KL}(F(u_K^{n+1}) - F(u_L^{n+1})) + v_{KL}^{n+1}\eta^{\mathcal{D}}_{\sigma\sigma^*}(\xi(u_{K^*}^{n+1}) - \xi(u_{L^*}^{n+1}))\Big),$$

où $F$ (resp. $\xi$) est la transformation de Kirchoff (resp. semi-Kirchoff) et $v_{KL}^{n+1}$ est une approximation "upstream" de $v$ sur l'arête primale $\sigma$. Elle est fournie par le schéma d' Engquist-Osher :

$$v_{KL}^{n+1} = \begin{cases} v_\downarrow(u_L^{n+1}) + v_\uparrow(u_K^{n+1}) & \text{si } \eta_{\sigma\sigma^*}^{\mathcal{D}} \, (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}) \geq 0 \\ v_\downarrow(u_K^{n+1}) + v_\uparrow(u_L^{n+1}) & \text{sinon} \end{cases}. \tag{0.2.38}$$

Les fonctions $v_\downarrow, v_\uparrow$ sont calculées à partir des formules suivantes :

$$v_\uparrow(u) := \int_0^u \Big(v'(s)\Big)^+ \, \mathrm{d}s, \quad v_\downarrow(u) := -\int_0^u \Big(v'(s)\Big)^- \, \mathrm{d}s,$$

où $x^+ = \max(x, 0)$, $x^- = \max(-x, 0)$ pour tout $x \in \mathbb{R}$. A la lumière de l'hypothèse $(A_2)$, les intégrales définissants $v_\uparrow, v_\downarrow$ existent.

On remarque que la quantité $v_{KL}^{n+1} \eta_{\sigma\sigma^*}^{\mathcal{D}} (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})$ peut être exprimée autrement à l'aide d'un flux numérique $G$ tel que

$$G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{L^*K^*}^{n+1} \xi(u)) = v_{KL}^{n+1} \eta_{\sigma\sigma^*}^{\mathcal{D}} (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}).$$

On rappelle qu'un flux numérique $G$ est une fonction à trois arguments $(a, b, c) \in \mathbb{R}^3$ qui répond aux axiomes suivants :

$$\begin{cases} (G_1) & G(\cdot, b, c) \text{ est croissante et continue pour tout } b, c \in \mathbb{R}, \\ & \text{et } G(a, \cdot, c) \text{ est décroissante et continue pour tout } a, c \in \mathbb{R}; \\ (G_2) & G(a, b, c) = -G(a, b, -c) \text{ pour tout } a, b, c \in \mathbb{R}; \\ (G_3) & G(a, a, c) = v(a)c \text{ pour tout } a, c \in \mathbb{R}. \end{cases} \tag{0.2.39}$$

Finalement, notre schéma s'écrit en deux parties. La première (resp. deuxième) partie correspond à la discrétisation du problème (0.2.33) sur le maillage primal (resp. dual).

$$u_M^0 = \frac{1}{m_M} \int_M u^0(x) \, \mathrm{d}x, \quad \forall M \in \mathcal{T}, \tag{0.2.40}$$

$$\begin{aligned} &\frac{m_K}{\delta t}\Big(u_K^{n+1} - u_K^n\Big) \\ &\qquad + \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_K} \Big(a_{KL} \, (F_K^{n+1} - F_L^{n+1}) + G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{L^*K^*}^{n+1} \xi(u))\Big) \\ &\qquad\qquad + \gamma \, \mathcal{P}_K u_{\mathcal{T}}^{n+1} = 0, \qquad \forall K \in \mathfrak{M}, \quad n \geq 0, \end{aligned} \tag{0.2.41}$$

$$\begin{aligned} &\frac{m_{K^*}}{\delta t}\Big(u_{K^*}^{n+1} - u_{K^*}^n\Big) \\ &\qquad + \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_{K^*}} \Big(a_{K^*L^*} \, (F_{K^*}^{n+1} - F_{L^*}^{n+1}) + G(u_{K^*}^{n+1}, u_{L^*}^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{LK}^{n+1} \xi(u))\Big) \\ &\qquad\qquad + \gamma \, \mathcal{P}_{K^*} u_{\mathcal{T}}^{n+1} = 0, \qquad \forall K^* \in \mathfrak{M}^*, \quad n \geq 0. \end{aligned} \tag{0.2.42}$$

Le paramètre $\gamma > 0$ représente un coefficient de stabilisation du schéma. On remarque que les mailles duales de $\partial\mathfrak{M}^*$ n'interviennent pas dans le système (0.2.40)-(0.2.42) à cause de la condition aux limites de Dirichlet.

**Résultats principaux**

On va énoncer les résultats obtenus dans ce chapitre. On insiste en particulier sur le principe du maximum discret et les estimations à priori qui représentent majoritairement les points clés pour établir le théorème de convergence du schéma.

**Lemma 0.2.5.** *(La borne $L^\infty$) Soient $0 \leq n \leq N-1$ et $(u_h^{n+1})$ un vecteur de $\mathbb{R}^{\#\mathcal{T}}$ tels que le schéma DDFV (0.2.40)-(0.2.42) soit vérifié. Alors, $u_{\mathfrak{M}}^{n+1}, u_{\overline{\mathfrak{M}^*}}^{n+1}$ sont dans l'intervalle $[0,1]$.*

On démontre tout d'abord par récurrence sur $n$ que $u_{\mathfrak{M}}^{n+1} \in [0,1]$ et on fait de même pour $u_{\overline{\mathfrak{M}^*}}^{n+1}$. A cet effet, on choisit un $K \in \mathfrak{M}$ tel que $u_K^{n+1} = \min_{L\in\mathfrak{M}} u_L^{n+1}$. On multiplie ensuite l'équation (0.2.41) par $-(u_K^{n+1})^-$. On obtient alors une équation dont la partie correspondante au terme de diffusion est positive. Ceci est dû à la monotonie de $F$, le choix de du coefficient $v_{KL}^{n+1}$ et à la dégénérescence de la fonction $v$ en dehors de zéro. Puis, il est facile de voir que la contribution du terme de pénalisation est toujours positive. On utilise enfin l'hypothèse de récurrence pour en déduire que $u_K^{n+1} \geq 0$. Pour prouver que $u_K^{n+1} \geq 1$, on multiplie (0.2.41) par $\left(u_K^{n+1}-1\right)^+$ et on se sert des mêmes arguments afin de conclure.

**Proposition 0.2.3.** *(Estimations d'énergie) Si $u_h^{n+1}$, où $n = 0,\dots,N-1$, est une solution du schéma DDFV (0.2.40)-(0.2.42) alors il existe $C$ indépendamment de $h_{\mathfrak{D}}$ et $\delta t$ tel que*

$$\sum_{n=0}^{N-1} \delta t \left\|\nabla^{\mathfrak{D}} \xi_h^{n+1}\right\|_2^2 + \frac{\gamma}{h_{\mathfrak{D}}^{\varepsilon}} \sum_{n=0}^{N-1} \delta t \left\|\xi(u_{\mathfrak{M}}^{n+1}) - \xi(u_{\overline{\mathfrak{M}^*}}^{n+1})\right\|_{L^2(\Omega)}^2 \leq C, \tag{0.2.43}$$

*et*

$$\sum_{n=0}^{N-1} \delta t \left\|\nabla^{\mathfrak{D}} F_h^{n+1}\right\|_2^2 \leq C. \tag{0.2.44}$$

Pour la démonstration de ce résultat, on multiplie d'abord (0.2.41) par $u_K^{n+1}$, on somme sur $K \in \mathfrak{M}$ et $n = 0,\dots,N-1$. De la même façon, on multiplie maintenant (0.2.42) par $u_{K^*}^{n+1}$, on somme sur $K^* \in \mathfrak{M}^* \cup \partial\mathfrak{M}^*$ et $n = 0,\dots,N-1$. On additionne ensuite les relations résultantes. On effectue des intégrations par parties et le choix du flux numérique permet d'achever l'estimation (0.2.43). La deuxième inégalité (0.2.44) découle automatiquement de (0.2.43).

**Theorem 0.2.3.** *Soient $(\mathcal{T}_h)$ une suite de maillages DDFV telle que $h_{\mathfrak{D}}, \delta t$ tendent vers 0 et $reg(\mathcal{T}_h)$ soit bornée. Alors les convergences ci-dessous sont satisfaites à une sous-suite près.*

$$u_{h,\delta t}, u_{\mathfrak{M}_h,\delta t}, u_{\overline{\mathfrak{M}^*}_h,\delta t} \longrightarrow u \qquad p.p.\ dans\ Q_{\mathfrak{T}}, \tag{0.2.45}$$

$$\nabla^{\mathfrak{D}} F_{h,\delta t} \longrightarrow \nabla F(u) \qquad faiblement\ dans\ L^2(Q_{\mathfrak{T}})^2. \tag{0.2.46}$$

*Ensuite, $0 \leq u \leq 1$ p.p. dans $Q_{\mathfrak{T}}$. Enfin, la fonction $u$ est l'unique solution faible du problème (0.2.33) au sens de la Définition 0.2.2.*

La preuve de ce théorème est effectuée en plusieurs étapes. On établie en premier des estimations sur les translatés en temps et en espace sur les deux suites $\xi(u_{\mathfrak{M}_h,\delta t})$ et $\xi(u_{\overline{\mathfrak{M}^*}_h,\delta t})$. Deuxièmement on s'assure que $u_{\mathfrak{M}_h,\delta t}$, $u_{\overline{\mathfrak{M}^*}_h,\delta t}$ et $u_{h,\delta t}$ convergent vers la même limite grâce à l'introduction du

terme de la pénalisation et de la convergence faible du gradient (0.2.46). Finalement on passe à la limite en utilisant des intégrations par parties, les estimations d'énergies, (0.2.45)-(0.2.46) et en exploitant l'avantage de la pénalisation.

Pour la validation numérique, nous nous intéressons à évaluer et à étudier l'erreur de convergence numériquement du schéma DDFV proposé pour l'équation de diffusion non linéaire dans des cas tests particuliers. On considère pour cela des maillages généraux et des perméabilités dont le ratio d'anisotropie est relativement important. On vérifie donc le principe de maximum discret qui est le point le plus important dans notre étude pour ce type de schéma. Il en résulte que la méthode est sensiblement d'ordre deux malgré la déformation du maillage et l'anisotropie du domaine.

## Liste de publications

Au cours de cette thèse nous avons produit trois publications scientifiques : un article publié dans une revue internationale :

- M. Ghilani, E. Quenjel and M. Saad. *Positive control volume finite element scheme for a degenerate compressible two phase flow in anisotropic porous media*, **Computational Geosciences**, Volume 22, pages : 1–25, 2018.

et deux articles soumis avec révision

- M. Ghilani, E. Quenjel and M. Saad. *Numerical analysis of a vertex-centered finite volume scheme for a gas-water porous media flow model.*
- E. Quenjel, M. Ghilani, M. Saad and M. Bessemoulin-Chatard. *Convergence of a positive nonlinear DDFV scheme for degenerate parabolic equations.*

# Chapter 1

# A review of modeling flows in porous media and state of the art

In this chapter we begin with some real-world applications that motivate the present thesis. We next overview the fundamental concepts related to porous media flows. We also survey some standard mathematical models accounted for the immiscible two-phase displacements in porous media. Finally, we indicate several relevant works dealing with such systems from both theoretical and numerical points of view.

## 1.1 Motivation

In the last decades, special attention has been paid to the two-phase flows in porous media. These kinds of processes arise from a wide range of disciplines such as hydrology, nuclear wastes management, medicine and petroleum engineering. Indeed, a large variety of different experiments in the fields of applications are overpriced or strictly prohibited to be carried out in reality for safety concerns. This has led to their representation thanks to the physical and the mathematical models. Unfortunately, exact solutions to such a model are inaccessible because of several factors related to the used physical data and the nature of the involved system itself. Then, understanding such a system allows to gain insight into the process in question. Hence, the numerical approximation brings along a great contribution and offers an attractive alternative to grasp the underlined models and therefore the studied phenomenon. Popular situations where the two-phase flow model occurs are thereafter highlighted.

First, the groundwater accounts for a major source, about 50 %, of the drinking-water supply for many countries in the world. In addition to irrigation, it is used for countless industrial processes [22, 123]. This natural resource is located in aquifers whose depth depends strongly on geological factors and the kind of climate. Because of intense human activities together with the current industry, the groundwater is subject to inevitable contamination issued from fertilizers, pesticides, storage tanks, landfills, etc (Fig. 1.1). The polluted water is then unsuitable and its use may lead to serious health problems. As a result, preserving the quality of this precious resource and removing the pollutants is of a great advantage for societies. Due to the scientific and the technological progress, some potential cleanup strategies and techniques have been developed and installed in order to remedy the affected zones. However, removal and remediation operations are often very expensive. Therefore, performing efficient numerical simulations of the two-phase flow model is commonly recommended for twofold : it first allows to predict the migration of the contaminants in the underground and minimize the duration together with the costly tasks of the cleaning.

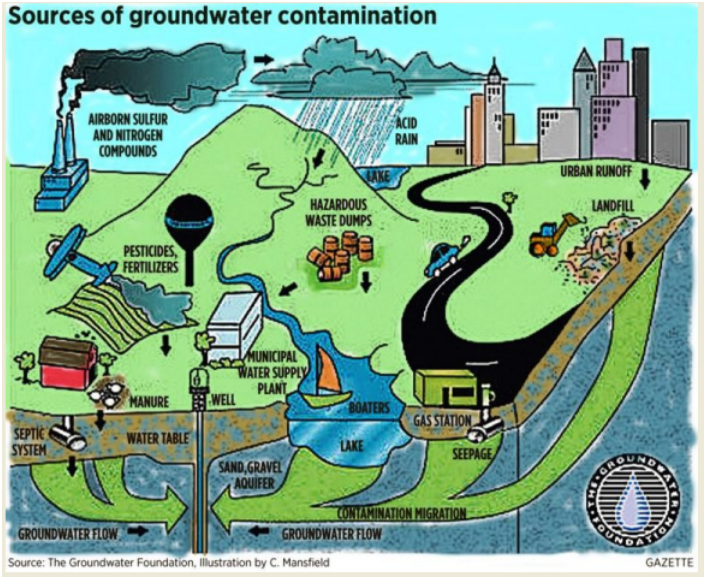



Figure 1.1: Main origins of groundwater contamination.

Research is otherwise active for developing new economical and accurate methods so as to address these issues.

Next, as a secondary source of energy, electricity is of a vital importance. It is generated from several sources namely: coal, natural gas, wind, hydropower, nuclear fission reactions, etc. It is worth mentioning that nuclear power plants provide a significant amount of energy for some developed countries like France, UK, Germany, and US. The latter type of power produces radioactive wastes at every step of the nuclear fuel cycle. Nuclear wastes are accumulated in controlled and safe repositories, but they are still considered as the most hazardous matter to which a person is exposed. Management of such a substance has represented a challenging task that the engaged

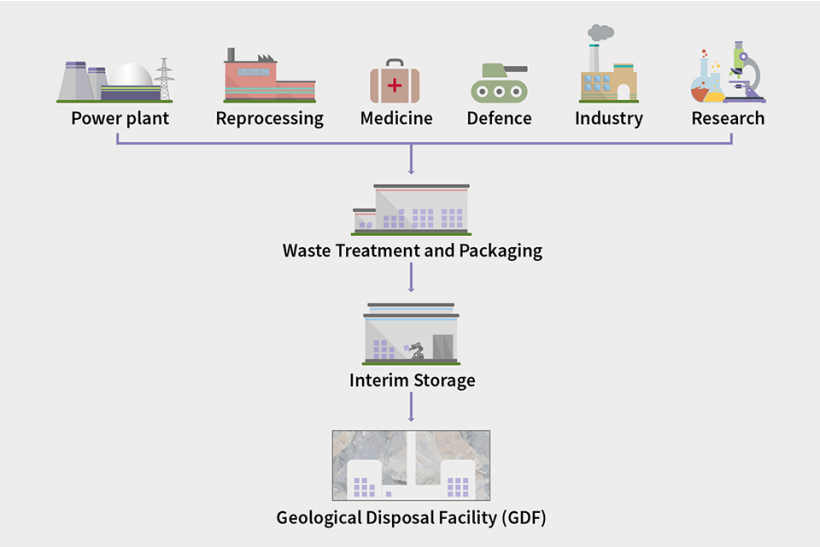



Figure 1.2: Different steps of nuclear waste management.

societies have confronted for many years [87]. The main reasons behind this resides in an exorbitant

cost of the waste management and the fact that the radioactive decay may take at least one million years. Consequently, a safe disposal of nuclear wastes has been a major and long-term goal of the current research. There exit already some options to get rid of these wastes. For instance, one solution consists in burying them through geological disposal facilities. Another possibility is to inject and store them in deep geological formations. Real applications of the latter is not actually permitted yet, but pragmatic installations will probably take place in the coming few years. In any case, the study of the two-phase flow model in porous media provides information on the ability of such an option to ensure the safety of the environment.

Finally, in petroleum engineering many techniques have been developed for the recuperation of the oil [47]. For instance, one can consider the enhanced recovery technique. Typically, this situation consists of placing two wells within the field under consideration. One of them is referred to as the injection well whereas the other one stands for the production well. A liquid such as water with or without chemical substances is then injected in the concerned zone with a high pressure in order to guarantee the migration of the hydrocarbons toward the production area. The cycle of the recovery is illustrated in Fig 1.3.

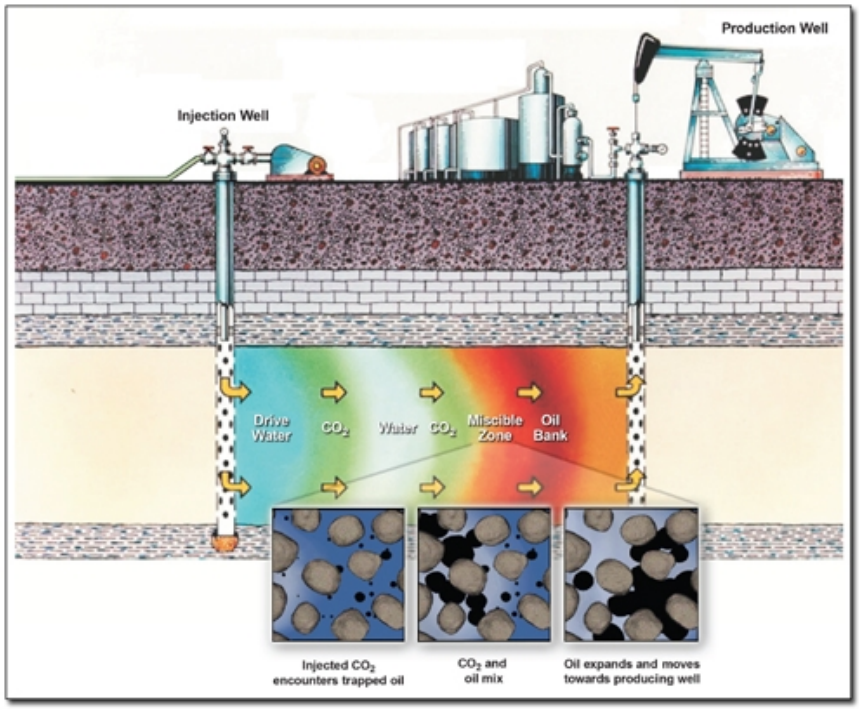



Figure 1.3: Enhanced oil recovery.

To model the underlined situation one may resorts to the diphasic model in porous media. This latter can give important ideas related to the motion of hydrocarbons in the field. It also allows to predict and optimize the production rates of oil.

## 1.2 Basic porous media concepts

This section is devoted to setting up the basic ingredients that are necessary to define the governed equations for the diphasic flows in a porous medium. Namely, this requires the properties of the porous medium and the characteristics of the involved phases.

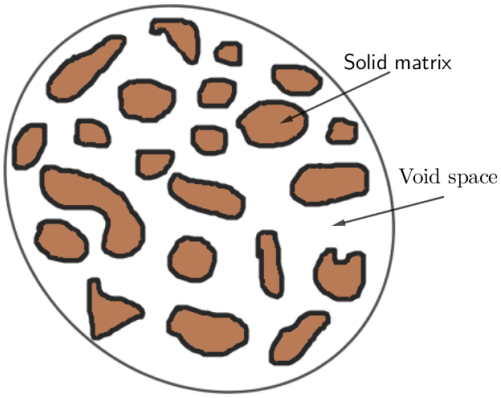



Figure 1.4: A simple example of a porous medium.

### 1.2.1 The porous medium

According to Cory [55] a porous medium is a kind of a matter made of the solid matrix and void (non-solid) space usually called pore space (see Fig. 1.4). This empty part can be occupied with one or more fluids e.g. water, oil and/or gas. Typical examples of porous media are soil, sand, bread, and lungs. In fact, not all physical object can belong to the framework of porous media. In the mentioned reference the author stresses that every porous medium must fulfill the following restrictions

- ($M_1$) the empty space of the porous medium is connected via free paths,
- ($M_2$) the smallest dimension of the pore space must be large enough compared to the mean-free path of the fluid molecules,
- ($M_3$) dimensions of the pore space must be small enough so that the fluid flow is largely controlled by interfacial forces, that is, only adhesive and cohesive forces govern the flow whenever interfaces occur between two fluids.

The first point ensures that the fluid particles may move in the whole void region of the medium. The second restriction allows to consider the continuum approach of the porous medium so that one can apply fluid mechanics laws. This continuum approach consists of considering the porous medium as an union of averaging volumes. The third assumption deprives many matters such as the network of pipes from the preceding definition of a porous medium.

A phase is referred to as a fluid like liquid or gas. We talk about a single-phase flow when the whole void space of the medium is filled by only one fluid e.g. water or air. Two phases are said to be immiscible if they cannot be mixed up such as oil and water. In a multiphase flow system, the pore space is completely occupied by more than two immiscible fluids.

### 1.2.2 Porosity and representative elementary volume

Based on the continuum approach, one can derive the equations of flows on the macroscopic level. In this case, the porous medium can be viewed as a system made of averaging volumes called representative elementary volumes (REV). The choice and size of the latter is strongly depending on the porosity. The porosity is a spacial function that allows to describe the distribution of the

void space within the continuum. Formally, the porosity denoted by $\phi$ is defined to be the fraction of the volume of the pores to the total volume of the REV

$$\phi = \frac{\text{volume of the void in REV}}{\text{total volume of REV}}.$$

Note that the porosity is a dimensionless quantity and it is comprised between 0 and 1. To determine the porosity of a porous medium e.g. a rock we resort to experimental results. For more information, we refer the reader to Cory's book [55].

### 1.2.3 Saturation

The saturation of a phase $\alpha$ represents the portion of this fluid within the representative elementary volume i.e.

$$s_\alpha = \frac{\text{volume of the fluid } \alpha \text{ in REV}}{\text{volume of the void in REV}}.$$

It is a time-space dependent function and there holds

$$\sum_\alpha s_\alpha(x,t) = 1, \quad 0 \leq s_\alpha(x,t) \leq 1.$$

Combining this definition of the saturation and that of the porosity we deduce that the volume of the phase $\alpha$ within the medium is given by $\phi s_\alpha$.

### 1.2.4 Capillary pressure law

When two immiscible fluids flow simultaneously in a porous medium, fluid-fluid and fluid-solid interactions occur at the separation interface. To illustrate this fact, we consider a vertical capillary tube that is partially immersed in a beaker of water as depicted in Fig 1.5. After reaching its maximum level, water forms a curved surface with an angle $\theta < \pi/2$. This angle characterizes the water-gas flow, according to which the water will be called the wetting phase while the gas will be referred to as the non-wetting phase. This definition of the wettability can be extended to any two-phase system. In the sequel, the index $w$ refers to the wetting-phase whereas $g$ stands for the non-wetting phase.

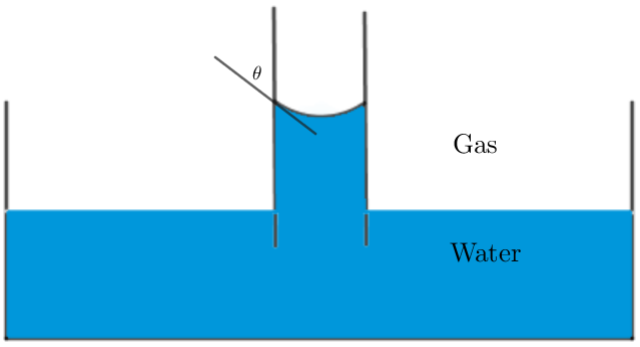



Figure 1.5: Water-gas interface in a capillary tube.

The origin of the curved surface between water and gas is due to the capillary forces. On the microscopic level, the fluid molecules are attracted to the solid by the adhesive forces whereas water

molecules are attracted to that of gas (and vice versa) by the cohesive forces. At the contact surface these forces are not balanced. Therefore, they give rise to a jump in terms of pressures between the water and gas which yields the capillary pressure law.

By definition, the capillary pressure denoted by $p_c$ is the difference between the pressure of the non-wetting phase (e.g. gas) $p_g$ and the pressure of the wetting phase (e.g. water) $p_w$.

$$p_c(s_w) = p_g - p_w. \tag{1.2.1}$$

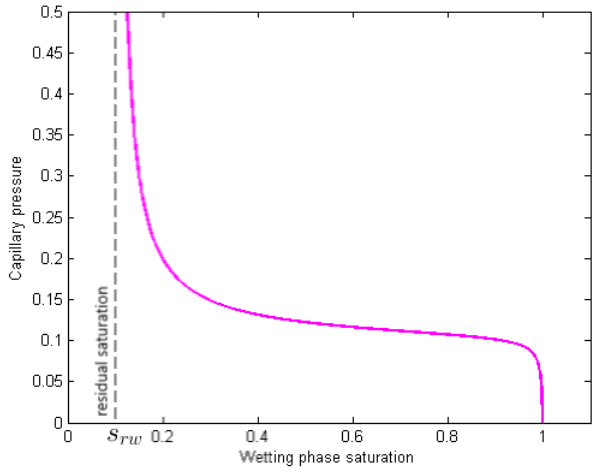



Figure 1.6: Typical shape of the capillary function $p_c(s_w)$.

Moreover, it is assumed to be dependent only on the wetting phase saturation $s_w$. It is also a positive and nonincreasing function with respect to $s_w$. Although it is often determined empirically or experimentally, the capillary pressure can be expressed in analytical formulas for particular porous media problems. For instance, the most common examples for applications to the air-water system are the models of Van Genuchten and Brooks–Corey. Their expressions of the capillary pressure are written in terms of the effective and the residual saturation.

The residual saturation $s_{rw}$ of the wetting phase (e.g. water) refers to the minimum amount of the wetting fluid which remains within the pores after the drainage process. When $s_w$ comes close to $s_{rw}$ the flow of water becomes very slow. At the same time, the capillary pressure increases rapidly and approaches a vertical asymptote at the point $s_w = s_{rw}$, see Fig. 1.6 . On the other hand, it is possible to obtain the residual saturation of the non-wetting phase. Then, the effective or the renormalized saturation of the wetting phase $s_{ew}$ and that of the non-wetting phase $s_{eg}$ are respectively defined to be

$$s_{ew} = \frac{s_w - s_{rw}}{1 - s_{rg}}, \quad s_{eg} = \frac{s_g - s_{rg}}{1 - s_{rw}}. \tag{1.2.2}$$

Thanks to the above relationships we obtain

$$s_{ew} + s_{eg} = 1 \quad \text{and} \quad s_{ew}, s_{eg} \in [0, 1].$$

Therefore, in the case of the water-gas flow system, the Van Genuchten [116] proposition for the capillary pressure function reads

$$p_c(s_w) = \frac{1}{\lambda}\left(s_{ew}^{(q-1/q)} - 1\right)^{1/q}, \tag{1.2.3}$$

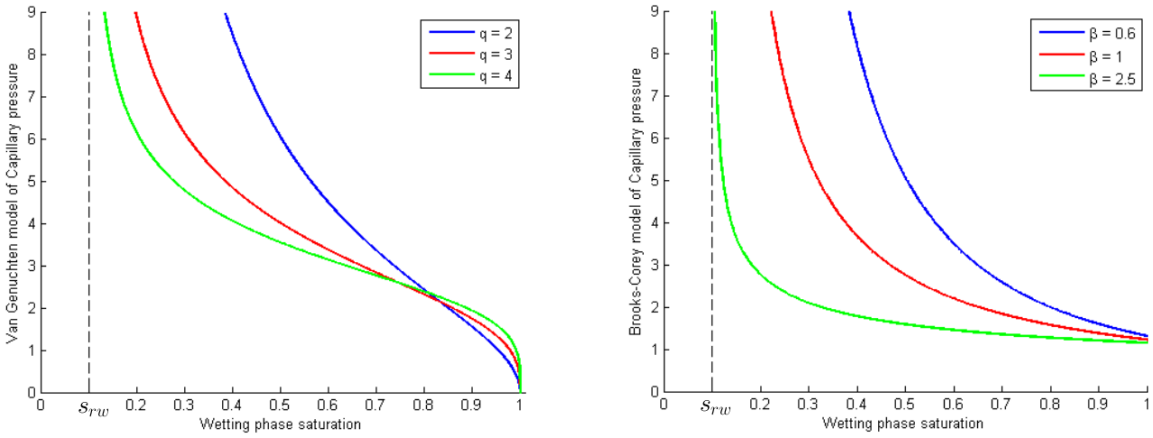



Figure 1.7: The Van Genuchten (left) and Brooks–Corey (right) capillary pressure functions.

with $\lambda$ is a real number and $q$ is typically an integer between 2 and 5. In the left part of Fig. 1.7, we plot the function (1.2.3) for three values of $q$ and a fixed value of $\lambda = 1/3$.

Additionally, the Brooks–Corey [31] model incorporates the concept of the entry pressure $p_e$ of the porous medium. This parameter corresponds to the value of the saturation for which the gradient of capillary pressure becomes much bigger. It must be imposed so that the non-wetting phase can penetrate the medium. Returning back to Fig. 1.6, we observe that $p_c$ increases rapidly around $s_w = 1$. Graphically, one sees that $p_e$ is approximately 0.1. So, the Brooks–Corey capillary pressure model is given by

$$p_c(s_w) = p_e s_{ew}^{-\frac{1}{\beta}}, \tag{1.2.4}$$

where $p_e$ is the entry pressure of the medium and $\beta$ is a real parameter that depends on the pore size distribution. Typical values of $\beta$ are in $[0.2, 0.3]$. In the left part of Fig. 1.7 we display the behavior of Brooks–Corey's capillary pressure functions.

### 1.2.5 Heterogeneity and anisotropy of a porous medium

Heterogeneity of a porous medium is closely related to its properties and provides information on the variation of some parameters, based on the averaging approach, with respect to the spacial variables. Otherwise, if the considered parameter is independent of the location, we then talk about homogeneity. For instance, a medium is called heterogeneous with respect to the porosity if the size of the pores varies spatially. This means that the medium might be composed of large, small and tiny pores that are depending on the position. If the pores are identical therefore the porous medium is homogeneous.

Anisotropy of a porous medium designates the dependency of tensorial quantities, e.g. intrinsic permeability (see its definition below), on directions at a given position. On the other hand, if the underlined quantity has the same value in any direction, the medium is said to be isotropic.

In Fig. 1.8 we reveal the concepts of heterogeneity and the anisotropy of such a porous medium. The top-left sub-figure shows that the void paths are similar and uniformly distributed, hence the medium in question is homogeneous and isotropic. Next, in the top-right sub-figure, we observe that the empty space depends on the position whereas it is independent on the direction. As a consequence, the fluid may flow rapidly through the upper side of the medium compared to the lower one. In this case, we talk about the heterogeneous and isotropic structure of the medium. Now, in the bottom-left sub-figure, the fluid flow in the $y$-direction can be more resistive than that in the

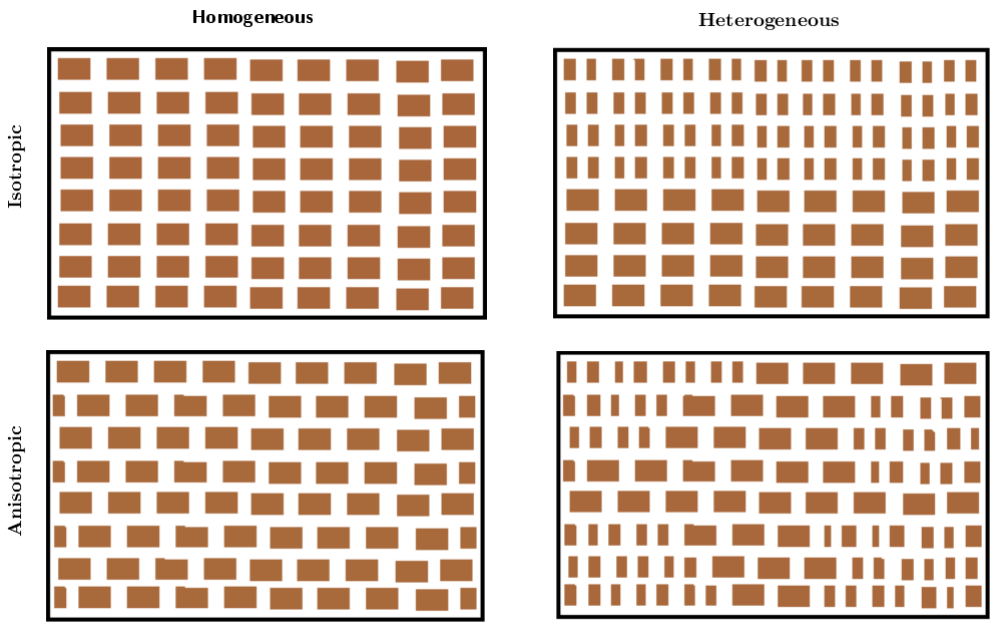



Figure 1.8: Illustration of the heterogeneity and the anisotropy of a porous medium with respect to the porosity and the permeability.

$x$-direction even if the size of the pores does not vary spatially. Then, the medium is homogeneous but anisotropic. Finally, the medium depicted in the last sub-figure possesses a different porosity in several parts as well as the permeability. Hence, it is heterogeneous and anisotropic.

### 1.2.6 Absolute and relative permeabilities

#### Absolute permeability

The absolute (or intrinsic) permeability denoted $\Lambda$ is a peculiar property of the porous medium (e.g. rock). It measures the ability of the rock to permeate any fluid which occupies the whole void space within the rock. Hence, the absolute permeability of the rock depends strongly on the geometric characteristics of the pores. It also remains the same in the medium despite of the fluid nature (gas, oil or water).

From a mathematical point of view, the permeability is represented by a tensor whose physical dimension is the $m^2$. For anisotropic media in the three-dimensional setting, this tensor reads

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} & \Lambda_{xz} \\ \Lambda_{yx} & \Lambda_{yy} & \Lambda_{yz} \\ \Lambda_{zx} & \Lambda_{zy} & \Lambda_{zz} \end{pmatrix}. \tag{1.2.5}$$

Generally, this matrix is assumed to be positive-definite [101]. In the case of an isotropic medium, the underlined matrix reduces to a scalar function.

#### Relative permeability

The relative permeability $K_{r\alpha}$ is a dimensionless quantity that models the ability of the phase $\alpha$ to pass through the porous medium in the presence of other fluids. This parameter allows to compare which fluids can flow together. It further verifies $0 \leq K_{r\alpha} \leq 1$. Note that in case of a single-phase

flow one has $K_{r\alpha} \equiv 1$. In addition, the analytical formula of $K_{r\alpha}$ is assumed to be a function which depends only on the saturation $s_\alpha$ [47]. The most famous relative permeability functions are that of Van Genuchten and Brooks–Corey which are often utilized in the two-phase system (e.g. water and gas). As we have seen for the capillary pressure functions, the Van Genuchten relative permeability functions can be expressed in terms of the effective saturations as follows

$$K_{rw}(s_w) = (s_{ew})^{\lambda}\left(1-\left(1-(s_{ew})^{\frac{q}{q-1}}\right)^{\frac{q-1}{q}}\right)^2, \tag{1.2.6}$$

$$K_{rg}(s_g) = (s_{eg})^{\gamma}\left(1-\left(1-s_{eg}\right)^{\frac{q}{q-1}}\right)^{\frac{2(q-1)}{q}}. \tag{1.2.7}$$

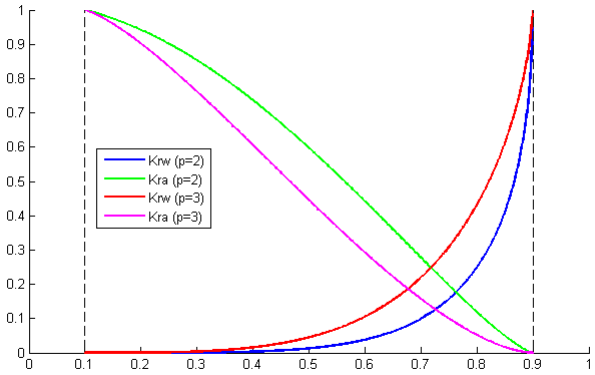



Figure 1.9: Van Genuchten relative permeabilities for residual saturations $s_{rw} = s_{rg} = 0.1$.

The parameters $\lambda$ and $\gamma$ are respectively set to $\lambda = 1/2$ and $\gamma = 1/3$. As before, $q$ is a positive integer chosen between 2 and 5. In Fig. 1.9, we plot the functions (1.2.6)-(1.2.7) for two values of $q$ ($q = 2, 3$) and with the same residual saturations $s_{rw} = s_{rg} = 0.1$. It shows that $K_{rw}$ increases slowly for lower values of saturation of the wetting phase e.g. water. When $s_w$ approaches its maximum value $K_{rw}$ grows rapidly, then the medium is almost saturated with water and the quantity of non-wetting phase e.g. gas becomes very small within the pores. Concerning $K_{rg}$, it behaves in the opposite situation of $K_{rw}$. It is known that the relative permeabilities can present hysteresis. However, the influence of the latter can be neglected, see [55] for more details.

### 1.2.7 Fluid's density and viscosity

#### Fluid's density

As a basic definition, fluid's density denoted $\rho$ is the ratio of fluid's mass to an infinitesimal elementary volume. Therefore it is a positive quantity whose unit is $Kg/m^3$. In particular, if the mass has the same value in any REV then $\rho$ is constant and the flow is said to be incompressible. For example, in practical applications to the water-air flow system, the water is viewed as an incompressible phase compared to the air phase since the compressibility of the latter is much bigger than that of water.

Yet, incompressibility can not be achieved in reality. Indeed, the fluid flow depends strongly on several parameters namely temperature $T$, pressure $p$ and the mass fraction $m_f$ of the chemical

species constituting the fluid. For the sake of simplicity, we assume that the flow is isothermal and the fluid is composed from identical molecules. Under these conditions, the state equation of the density is only written in terms of the fluid pressure as follows [22]

$$\rho = \rho(p).$$

Differentiating this relation yields

$$\frac{\mathrm{d}\rho}{\mathrm{d}p} = \beta_f \rho(p), \tag{1.2.8}$$

where the factor $\beta_f = \dfrac{1}{\rho}\dfrac{\mathrm{d}\rho}{\mathrm{d}p}$ stands for the compressibility of the fluid under study. Notice that in the case that $\beta_f = 0$ then the fluid of interest is incompressible. This coefficient is sometimes considered constant on certain ranges of pressure. In this case (1.2.8) implies an explicit formula of the density

$$\rho(p) = \rho^{\mathrm{ref}} \exp\Big(\beta_f (p - p^{\mathrm{ref}})\Big). \tag{1.2.9}$$

The factor $\rho^{\mathrm{ref}}$ indicates the density at the reference pressure $p^{\mathrm{ref}}$. Thanks to Taylor's series expansion we get

$$\rho(p) = \rho^{\mathrm{ref}} \left(1 + \beta_f (p - p^{\mathrm{ref}}) + \frac{\beta_f^2}{2}(p - p^{\mathrm{ref}})^2 + \dots\right).$$

When the function $p$ takes values around $p^{\mathrm{ref}}$, then $\rho(p)$ can be approximated by the first terms of the previous series

$$\rho(p) \approx \rho^{\mathrm{ref}} \Big(1 + \beta_f (p - p^{\mathrm{ref}})\Big).$$

In this case, the flow is referred to as slightly compressible.

### Fluid's viscosity

The dynamic viscosity $\mu$ of a phase measures the resistance of the fluid when it is subjected to the shear stress. By stress we mean the cause that incites the deformation. Informally, the viscosity allows to determine the thickness of the fluid. For example, honey is more viscous than water and water is less viscous than oil. Moreover, it is known that the viscosity of water becomes small with temperature while that of gas increases with temperature. Likewise, it depends on the pressure and the interaction between molecules of the fluid. Hence, its equation of state has the following form [22]

$$\mu = \mu(p, \mathfrak{T}, \dots). \tag{1.2.10}$$

On the other hand, let us denote by $\tau$ the shear stress and $\mathcal{D}$ the shear rate. These two quantities are linked via the Newton law

$$\tau = \mu \mathcal{D}.$$

The fluid is said to be Newtonian if the viscosity is independent of the shear rate and varies only with respect to the pressure and the temperature. Otherwise, it is called non-Newtonian.

The SI unit of dynamic viscosity is the Pascal second, Pa s. Throughout this thesis, we will assume that the dynamic viscosity is constant.

## 1.3 Mathematical formulations of flows in porous media

This section targets to survey some mathematical models describing the flow and transport in porous media. To this end, we restrain our exposition to the single-phase flow and immiscible two-phase flow models under the isothermal condition. Moreover, all the considered fluids are Newtonian. The of model equations under study are derived from the mass conservation law and Darcy law.

To fix the ideas, we assume that the porous medium $\Omega$ is assimilated to a bounded connected open subset of $\mathbb{R}^d$ with a Lipschitz boundary. In practice, we are only interested in two or three dimensions in space e.g. $d = 2, 3$. The real number $T$ will stand for the physical time. We denote $|\cdot|_{\mathbb{R}^d}$ or simply $|\cdot|$ the euclidean norm in $\mathbb{R}^d$.

### 1.3.1 Mass conservation principle

Assuming that the medium is completely filled with a phase (its saturation equals to one), the principle of the mass conservation of this phase states that the rate change of the total mass of the fluid within a volume $K \subseteq \Omega$ is balanced by the mass flux across the boundary of $K$ and the contribution of sources or sinks within $K$. This statement is known as the integral form of the mass conservation equation written as [109]

$$\frac{\partial}{\partial t}\int_K \phi\rho \,\mathrm{d}K = \int_{\partial V} \rho\mathbf{V}\cdot\mathbf{n}_K \,\mathrm{d}\partial K + \int_K \rho F \,\mathrm{d}K, \tag{1.3.1}$$

where $\mathbf{n}_K$ is the unit outward normal vector to the boundary $\partial K$ and $\mathrm{d}\partial K$ is an appropriate superficial measure upon $\partial K$. Thanks to Gauss–Ostrogradski's formula, the balance equation (1.3.1) becomes

$$\frac{\partial \phi\rho}{\partial t} + \operatorname{div}\rho\mathbf{V} = \rho F, \tag{1.3.2}$$

where we specify the constitutive parameters in the following list.

- $\phi(x)$ Porosity of the porous medium given in $[m^3]$. It is a spacial-dependent function.
- $\rho(x,t)$ Density of the fluid given in $[kg/m^3]$. The density of a fluid is only connected to its pressure $p$. If it is constant then the flow is incompressible. Otherwise it is compressible.
- $\mathbf{V}(x,t)$ Velocity of the fluid in $[m/s]$.
- $F(x,t)$ Source/sink term with dimension $[s^{-1}]$.

In the case that more than two fluids are present in the medium, we actually take into account the saturation $s_\alpha$ of the $\alpha$-phase. Hence, the mass continuity for each phase reads

$$\frac{\partial \phi\rho_\alpha s_\alpha}{\partial t} + \operatorname{div}\rho_\alpha\mathbf{V}_\alpha = \rho_\alpha F_\alpha. \tag{1.3.3}$$

As we are tacitly interested in Darcian flows, the velocity $\mathbf{V}$ given in (1.3.2) (resp. $\mathbf{V}_\alpha$) is expressed according to the Darcy (resp. Darcy–Muskat) law that we detail in the next subsections.

### 1.3.2 Darcy's law

In 1856 H. Darcy [58] observed experimentally that the flow occurs when a difference in pressure is maintained. In the one-dimensional case, he established a linear relationship between fluid's velocity

and the gradient of pressure. Later on this relationship has been extended to the multi-dimensional case as

$$\mathbf{V} = -\frac{1}{\mu}\Lambda\Big(\nabla p - \rho\vec{\mathbf{g}}\Big). \tag{1.3.4}$$

The constitutive inputs are detailed below.

- $p(x,t)$ Fluid's pressure in $[Pa] = [N/m^2]$.
- $\vec{\mathbf{g}}$ Gravitational acceleration in $[m/s^2]$. It is a vector pointing toward the opposite direction of the $z$-coordinate (which points upward). Then, one may set $g = (0, 0, -9.81)^T$.
- $\Lambda(x)$ Symmetric permeability tensor whose dimension is $[m^2]$.
- $\mu(x,t)$ Dynamic viscosity of the fluid is expressed in $[Pa.s]$. It is assumed to be a constant throughout this thesis.

Using the averaging volume method under some convenient assumptions, Darcy's formula can be rigorously obtained from the momentum conservation of the Navier–Stokes equation [119].

It is worth indicating that Darcy's law remains valid only for extremely small velocities [17]. In the case of a flow with a relatively high velocity, then its speed is linked to the pressure head via a nonlinear relationship which is due to Forchheimer [80]. Considering flows with the latter kind of velocity is beyond the scope of this thesis.

### 1.3.3 Darcy–Muskat's Law

When two fluids share the pore space, as seen in the preceding section, each fluid resists to the motion of the other one. This effect is obviously modeled by the relative permeabilities. Therefore, the flow velocity expression for each phase take naturally into account this new parameter. In fact, it is represented by the extended Darcy law known under the name of Darcy–Muskat's law [104] and given by the formula

$$\mathbf{V}_\alpha = -\frac{K_{r\alpha}}{\mu_\alpha}\Lambda(\nabla p_\alpha - \rho_\alpha\vec{\mathbf{g}}), \tag{1.3.5}$$

where, each phase has now its own characteristics :

- $K_{r\alpha}(s_\alpha)$ relative permeability of the phase $\alpha$ (dimensionless quantity),
- $\mu_\alpha(x,t)$ dynamic viscosity of the phase $\alpha$,
- $p_\alpha(x,t)$ pressure of the phase $\alpha$,
- $\rho_\alpha(x,t)$ density of the $\alpha$-phase.

In light of (1.3.5), the velocity of each fluid becomes very slow when $s_\alpha$ comes closer to $s_{r\alpha}$, see Fig. 1.9. In addition, $\mathbf{V}_\alpha$ vanishes when $s_\alpha = s_{r\alpha}$. This is referred to as the degeneracy issue. In fact, the absence phase in some parts of the medium can lead to serious problems in both the theoretical and numerical investigations of the model.

### 1.3.4 Single-phase flow

The simplest model of flows in porous media is the single-phase flow. It is described by the mass conservation equation where the velocity is expressed thanks to Dacry's law. This situation, of course, occurs when the whole medium is saturated by a single fluid with only one component. To acquire its mathematical formulation, one substitutes the relationship (1.3.4) into the equation (1.3.2). Therefore one obtains

$$\frac{\partial \phi \rho(p)}{\partial t} - \operatorname{div} \rho(p) \frac{1}{\mu} \Lambda \Big( \nabla p - \rho(p) \vec{\mathbf{g}} \Big) = \rho(p) F, \quad \text{in } Q_{\mathfrak{T}} := \Omega \times (0, \mathfrak{T}). \tag{1.3.6}$$

The primary unknown here is the pressure $p$. To close the above equation, boundary and initial conditions must be specified. The initial datum $p(\cdot, t = 0) = p^0$ gives the state of the solution at $t = 0$ whereas boundary conditions show in advance the behavior of the solution on some or all parts of the boundary. The most common kinds of boundary conditions that we will take into consideration are that of Dirichlet and Neumann. Let us then split the boundary $\partial\Omega$ into two disjoint parts $\partial\Omega = \Gamma_D \cup \Gamma_N$ with $|\Gamma_D| > 0$. Thereby, we consider

$$p(x,t) = p^D(x,t) \quad \text{on } \Gamma_D \times (0, \mathfrak{T}) \quad \text{and} \quad \rho(p) \mathbf{V} \cdot \mathbf{n} = \eta(x,t) \quad \text{on } \Gamma_N \times (0, \mathfrak{T}). \tag{1.3.7}$$

Without loss of generality, the given functions $p^D, \eta$ can be set to zero. Notice that the equation (1.3.6) is of a parabolic type when $\rho$ is not the constant function. Therefore, the compressibility of the fluid maintains the parabolic nature of the single-phase flow model. Otherwise, it is of an elliptic type.

Next, one might wonder whether (1.3.6) admits analytical expressions of solutions (when they exist) for given data. The answer is positive for some particular situations. For instance, let us consider a homogeneous and an isotropic medium with no gravity effects. We further assume that we have no source term and the density is linear with respect to the pressure. Then (1.3.6) reduces to

$$\frac{\partial p}{\partial t} - \operatorname{div} \nabla p^{\gamma} = 0, \quad \text{with } \gamma = 2. \tag{1.3.8}$$

This identity is said to be the porous medium equation in the literature [117] whose unique classical solution is

$$p(x,t) = \frac{|x|^2}{1-t}, \quad \forall t < 1. \tag{1.3.9}$$

We point out that one can always build suitable exact solutions to (1.3.6). To this end, one only needs to find an obvious physical expression for $p$, computes the left hand side of (1.3.6) using the chosen function and sets the result to the source term. For general physical data and assumptions, the step consists of seeking or studying the existence and/or uniqueness of exact or classical solutions is often skipped due to the nonlinearity kind of the posed problem. Nonetheless, weak solutions might exist implicitly when they are understood in the sense of distributions. For further information about this topic in relation with (1.3.6), we refer to [82].

### 1.3.5 Immiscible two-phase flow

We here highlight a couple formulations to equations modeling the two-phase flow. In contrast to the single-phase flow model, the diphasic one includes some peculiar quantities such as the saturation, relative permeabilities and capillary pressure.

The displacement process is then governed by the mass conservation equation together with Darcy-Muskat's relationship for each phase [22, 47, 95]. For $\alpha \in \{w, g\}$, plugging (1.3.5) into (1.3.3) yields

$$\frac{\partial \phi \rho_\alpha(p_\alpha) s_\alpha}{\partial t} - \operatorname{div} \rho_\alpha(p_\alpha) M_\alpha(s_\alpha) \Lambda \Big( \nabla p_\alpha - \rho_\alpha(p_\alpha) \vec{\mathbf{g}} \Big) = \rho_\alpha(p_\alpha) F_\alpha, \quad \text{in } Q_{\mathfrak{T}}, \tag{1.3.10}$$

where the function $M_\alpha = K_{r\alpha}/\mu_\alpha$ is appended to designate the $\alpha$-phase mobility. We hereafter stress that the source term $F_\alpha$, that will be specified later on, is an affine function in terms of

the $\alpha$-phase saturation. In addition to (1.3.10), the medium being completely occupied by the two fluids entails

$$s_w + s_g = 1. \tag{1.3.11}$$

This identity together with (1.3.10) for each phase lead to a system of two equations and three unknowns, $p_g, p_w$, and $s_w$. Another condition is required to relate these variables. A relevant relation involving the three functions at the same time is provided by the capillary pressure law

$$p(s_w) = p_g - p_w. \tag{1.3.12}$$

Additionally, boundary conditions and initial data must be prescribed. This depends on the choice of the primary variables. Several possibilities exist in order to determine the unknowns. For instance, one selects the two pressures, $p_g$ and $p_w$ and deduces the saturation from (1.3.11)-(1.3.12). Another option consists of taking a fluid pressure (either $p_g$ or $p_w$) and saturation (either $s_g$ or $s_w$) as the main variables.

### Pressure–pressure formulation

In this first formulation, we consider $p_g$ and $p_w$ as the main unknowns. Moreover, the capillary pressure function is assumed to be invertible so that one can compute the saturation. Thus, one gets $s_w = p_c^{-1}(p_g - p_w)$ and $s_g = 1 - p_c^{-1}(p_g - p_w)$. Inserting these equations into (1.3.10) for each phase one finds

$$-\phi \frac{\partial \rho_g p_c^{-1}}{\partial t} - \operatorname{div} \rho_g M_g \Lambda \Big( \nabla p_g - \rho_g \vec{\mathbf{g}} \Big) = \rho_g F_g, \tag{1.3.13}$$

$$\phi \frac{\partial \rho_w p_c^{-1}}{\partial t} - \operatorname{div} \rho_w M_w \Lambda \Big( \nabla p_w - \rho_w \vec{\mathbf{g}} \Big) = \rho_w F_w, \tag{1.3.14}$$

subject to mixed boundary conditions

$$p_g(x,t) = p_g^D \quad \text{on } \Gamma_D \times (0,\mathfrak{T}), \qquad \rho_g \mathbf{V}_g \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0,\mathfrak{T}), \tag{1.3.15}$$

$$p_w(x,t) = p_w^D \quad \text{on } \Gamma_D \times (0,\mathfrak{T}), \qquad \rho_w \mathbf{V}_w \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0,\mathfrak{T}), \tag{1.3.16}$$

together with initial pressures

$$p_w(x,0) = p_w^0(x) \quad \text{and} \quad p_g(x,0) = p_g^0(x) \quad \text{in } \Omega. \tag{1.3.17}$$

Practically, this approach can be achieved for particular cases where the inverse of the capillary pressure has a good behavior especially near $s_w = 0$. When the $p_c$ exhibits singularities near some points, one may resorts to its approximation thanks to a regularization technique. At the discrete level, pertinent approximations of the capillary pressure might, however, give rise to serious issues in the nonlinear solver for very small values of $p_c'$.

Generally, the very weak point of this formulation consists of excluding the hyperbolic occurrence when capillary effects are neglected i.e. $p_c \equiv 0$. This motivates the following alternatives.

### Phase pressure–saturation formulation

First, let us look at the formulation incorporating the nonwetting-phase saturation $s_g$ and the wetting-phase pressure $p_w$ as the primary unknowns while $s_g$ and $p_g$ are substituted by

$$s_w = 1 - s_g, \quad \text{and} \quad p_g = p_w + p_c(1 - s_g).$$

Consequently, (1.3.10) can be recast in the form:

$$\phi\frac{\partial \rho_g s_g}{\partial t} - \operatorname{div} \rho_g M_g(s_g)\Lambda\Big(\nabla p_w + \nabla p_c - \rho_g \vec{\mathbf{g}}\Big) = \rho_g F_g, \tag{1.3.18}$$

$$-\phi\frac{\partial \rho_w s_g}{\partial t} - \operatorname{div} \rho_w M_w(1-s_g)\Lambda\Big(\nabla p_w - \rho_w \vec{\mathbf{g}}\Big) = \rho_w F_w. \tag{1.3.19}$$

For the sake of simplicity, we henceforth consider homogeneous boundary conditions and initial data as follows

$$p_w(x,t) = 0 \quad \text{on } \Gamma_D \times (0,\mathfrak{T}), \qquad \rho_g \mathbf{V}_g \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0,\mathfrak{T}), \tag{1.3.20}$$

$$s_g(x,t) = 0 \quad \text{on } \Gamma_D \times (0,\mathfrak{T}), \qquad \rho_w \mathbf{V}_w \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0,\mathfrak{T}), \tag{1.3.21}$$

$$s_g(x,0) = s_g^0(x) \quad \text{and} \quad p_w(x,0) = p_w^0(x) \quad \text{in } \Omega. \tag{1.3.22}$$

In case of a compressible flow, the system (1.3.18)-(1.3.22) contains at least one degenerate parabolic equation. In addition, it is nonlinear and strongly coupled. Otherwise, in the incompressible case, it changes the type with respect to the presence and the absence of the nonwetting phase. To see this, let us reformulate previous system in the following simple form. Notice that $\rho_\alpha$ is constant. Replacing (1.3.19) with the sum of (1.3.18) and (1.3.19) implies

$$\phi\frac{\partial s_g}{\partial t} - \operatorname{div} M_g(s_g)\Lambda\Big(\nabla p_w + \nabla p_c - \rho_g \vec{\mathbf{g}}\Big) = F_g, \tag{1.3.23}$$

$$- \operatorname{div} M(s_g)\Lambda\Big(\nabla p_w + f_g \nabla p_c - \mathbf{G}\Big) = F_g + F_w, \tag{1.3.24}$$

where the function $M$ is termed the total mobility. It is bounded away from zero. Moreover, the factor $f_g$ denotes the fractional flow of the phase $g$. The term $G$ stands for the modified gravity vector. They are respectively defined as :

$$M = M_g + M_w, \quad f_w = \frac{M_g}{M}, \quad \mathbf{G} = \frac{\rho_g M_g + \rho_w M_w}{M}\vec{\mathbf{g}}.$$

Particularly, we neglect the gravity and source contributions. Consequently, (1.3.24) is of type elliptic with regard to $p_w$. Hence, the initial condition on $p_w$ is not required. The saturation equation is purely hyperbolic with respect to $s_g$ whenever the term including $p_c$ is neglected. In this case, (1.3.23) is reduced to the famous Buckley–Leverett equation [33]. Otherwise, the underlined equation turns out to be a nonlinear degenerate parabolic equation since $M_g(s_g = 0) = 0$. From a numerical perspective, the nonlinear diffusion coefficient $M_g p_c'$ has to be controlled near $s_g = 0$.

Secondly, one can also choose $p_g$ and $s_w$ as the primary unknowns for solving the diphasic mathematic model. This time, a careful attention should be paid to the variation of $M_w p_c'$ near the singularity of $p_c'$.

Finally, the formulations we discussed so far are degenerate, nonlinear and strongly coupled. This can give rise to severe problems in the analysis of model's equations at the continuous setting as well as at the discrete one. This has led to the development of artificial unknowns in order to overcome some of these major difficulties.

### Global pressure alternative

The concept of the global pressure has been originally introduced by G. Chavent et al. [47]. Its basic idea consists of expressing the phase pressures in terms of a unique intermediary pressure with additional perturbations. These corrections exist, are well-defined and are assumed to depend solely on the nonwetting phase saturation. In the sequel, we denote $s = s_g$.

The global pressure is defined as

$$p = p_g + \widetilde{p}_g(s) = p_w + \widetilde{p}_w(s), \tag{1.3.25}$$

so that one has

$$\nabla p = \nabla p_g - f_w(s)\nabla p_c = \nabla p_w + f_g(s)\nabla p_c. \tag{1.3.26}$$

The artificial pressures $\widetilde{p}_g$ and $\widetilde{p}_w$ satisfying (1.3.25)-(1.3.26) are respectively written under the following form

$$\widetilde{p}_g(s) = -\int_0^s f_w(u)p_c'\,\mathrm{d}u, \quad \widetilde{p}_w(s) = \int_0^s f_g(u)p_c'\,\mathrm{d}u. \tag{1.3.27}$$

As a consequence of the preceding formulas and the fact that $f_g + f_w = 1$, one checks that

$$p_g - p_w = \widetilde{p}_w(s) - \widetilde{p}_g(s) = p_c(s).$$

In case of incompressible flows, substituting $p_w = p - \widetilde{p}_w(s)$ into the system (1.3.18)-(1.3.19) provides

$$\phi\frac{\partial s}{\partial t} - \operatorname{div} M_g(s)\Lambda\Big(\nabla p - \rho_g\vec{\mathbf{g}}\Big) = F_g, \tag{1.3.28}$$

$$-\operatorname{div} M(s)\Lambda\Big(\nabla p - \mathbf{G}\Big) = F_g + F_w, \tag{1.3.29}$$

where the main unknowns are now the saturation $s$ and the global pressure $p$. Here the equations are more easily to study theoretically and numerically. One actually sees that the unknowns are decoupled and less degenerate compared to (1.3.18)-(1.3.19) since the difficulty coming from the terms involving $p_c$ is tackled.

On the other hand, when one of fluids' compressibility is included, the global pressure serves, furthermore, to establish a link between its gradient and the gradients of both phases. According to (1.3.26) one finds

$$|\nabla p|^2 + f_g f_w\,|\nabla p_c|^2 = f_g\,|\nabla p_g|^2 + f_w\,|\nabla p_w|^2\,. \tag{1.3.30}$$

This is a great gain since the right hand side of this relationship shows that if one of the phases is absent somewhere within the medium, we automatically lose the gradient of its pressure. Nevertheless, the first term of the left hand side of (1.3.30) is independent of the degeneracy of each phase while the second one is expressed only in terms of the saturation. This is the main strength of the global pressure formulation. The equality (1.3.30) has played a fundamental role to establish the a priori estimates, regardless the degeneracy issue, which are the crucial ingredient to prove the existence of weak solutions to compressible flows [5, 95]. The passage from the phases pressures to the global pressure was a key point in the work [44]. Analogous discrete inequalities to (1.3.30) were also the cornerstone for the convergence analysis of the numerical schemes proposed in [72, 113].

In our study, we will rather focus on flows with small capillary effects. Hence, the density varies a little bit with respect to the capillary pressure function. This amounts to suppose that the density of the gas depends only on the global pressure as pointed out in [47]

$$\rho_\alpha(p_\alpha) = \rho(p).$$

Under this assumption, the governing equations of the compressible two-phase flow model reduce to

$$\frac{\partial \phi \rho_\alpha(p) s_\alpha}{\partial t} - \operatorname{div} \rho_\alpha(p) M_\alpha(s_\alpha) \Lambda \Big( \nabla p + f_{\tau(\alpha)} \nabla p_c - \rho_\alpha(p) \vec{\mathbf{g}} \Big) = \rho_\alpha(p) F_\alpha, \quad \text{in } Q_{\mathfrak{T}}, \tag{1.3.31}$$

with

$$s_w + s_g = 1,$$

where the permutation $\tau$ reads

$$\tau(\alpha) = \begin{cases} w, & \text{if } \alpha = g \\ g, & \text{if } \alpha = w \end{cases}.$$

Moreover, the flow velocity of the phase $\alpha$ is reformulated so that

$$\mathbf{V}_\alpha = -\Lambda \Big( M_\alpha(s_\alpha) \nabla p + \gamma(s) \nabla s_\alpha - M_\alpha \rho_\alpha(p) \vec{\mathbf{g}} \Big).$$

The nonlinear diffusion coefficient $\gamma(s)$ is

$$\gamma(s) = \frac{M_g M_w}{M} p_c' \geq 0.$$

Note upon this that the total velocity in the presence of gravity becomes

$$\mathbf{V}_w + \mathbf{V}_g = -M \Lambda \Big( \nabla p - \mathbf{G}(p) \Big), \quad \text{and,} \quad \mathbf{G}(p) = \frac{\rho_g(p) M_g + \rho_w(p) M_w}{M} \vec{\mathbf{g}}.$$

As usual, the above system should be supplied with boundary conditions and initial data

$$p(x,t) = 0 \quad \text{on } \Gamma_D \times (0, \mathfrak{T}), \qquad \rho_g \mathbf{V}_g \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0, \mathfrak{T}), \tag{1.3.32}$$

$$s(x,t) = 0 \quad \text{on } \Gamma_D \times (0, \mathfrak{T}), \qquad \rho_w \mathbf{V}_w \cdot \mathbf{n} = 0 \quad \text{on } \Gamma_N \times (0, \mathfrak{T}), \tag{1.3.33}$$

$$s(x,0) = s^0(x) \quad \text{and} \quad p(x,0) = p^0(x) \quad \text{in } \Omega. \tag{1.3.34}$$

Under some standard assumptions, existence results for problems like (1.3.31)-(1.3.34) were studied in [32, 83, 84]. Numerical analysis for such a system using the finite volume approximation was done in [23].

## 1.4 Sate of the art

The study and approximation theory of the governing equations of the flow and transport in porous media have become of an increasing interest for the understanding of numerous physical phenomena. This mechanism allows to rigorously understand, predict and optimize the behavior of the phenomenon under consideration.

From a mathematical point of view, most of these models are formulated in a set of partial differential equations together with initial and/or boundary conditions. Throughout this thesis, the envisaged problems include diffusion and convection effects. They then incorporate elliptic and hyperbolic characters. It is well known that the convection-dominated case can produce sharp front and chocks. This means that the solution can vary rapidly within a very small region of the domain. From a numerical perspective, the fact of capturing the localization of shocks is not an obvious task. Design and convergence of efficient numerical approaches for these systems is then of a practical importance.

Once the problem is posed, it is of an advantage to examine its qualitative behavior, namely the existence, uniqueness and properties of the solutions such as smoothness, asymptotic behavior and so on. As a matter of fact, this depends strongly on the structure of the problem at hand and on the kind of solutions sought e.g. classical, weak, entropic, renormalized solutions, etc. Moreover, when the inputs of the system of interest are irregular or present some discontinuities, the nature of the solution can be changed. In practical models of flow and transport in porous media, the equations are extremely complicated to solve analytically. The most encountered difficulties may arise from the physical assumptions on the data, the nonlinearity of the system under study and/or the strong coupling of the constitutive variables.

Numerous theoretical analyses, with various assumptions on the data, of porous media flow models have been reported in the literature of the past few decades. A particular emphasis has been set on the two-phase flow systems. The existence and uniqueness investigation of immiscible incompressible flows has been conducted in plenty contributions, for instance we cite [15, 47, 48, 49, 75, 74, 82, 98, 120] and the references are therein. The miscible incompressible case is also a subject of interest which is treated in [7, 53, 77, 78]. A recent few works are devoted to analyzing the mathematical models for compressible and immiscible displacements in porous media. Assuming that the fluid densities are depending only on the global pressure, introduced in [47], the existence of weak solutions have been studied in [83, 84, 85]. These latter results have been generalized in [95, 96] without any major restriction on the densities. *Amaziane et al.* developed a new global pressure formulation composed of a nonlinear parabolic equation for the global pressure equation coupled to a nonlinear diffusion–convection equation for saturation [5, 6]. Recently, in a one-dimensional case, *Saad* [115] showed an existence result for a slightly compressible and immiscible two-phase flow problem where the density follows an exponential law with a small compressibility factor.

On the other hand, it has become a tendency to introduce the scientific computing as a realistic way to unveil the posed problem. Thanks to appropriate and modern numerical methods, it is possible to design approximated solutions for porous media flow type problems. Popular choices of these methods are: finite differences, (mixed) finite elements, finite volumes, discontinuous Galerkin schemes and gradient scheme methods. As discussed above, we point out that the choice of the discresattization has to take into account the character of the mathematical model in question. In addition, the chosen meshes have to be adequate and adopted to the complex geometry of the porous medium (one can bear in mind too distorted geological layers as a reference for instance). Accordingly, the proposed numerical scheme must work on these kinds of meshes so that good results can be achieved. Given a numerical method (in the framework of the present thesis), it is desirable that it satisfies some relevant and prerequisite properties summarized in:

($i$) stability,

($ii$) preserving the physical ranges of certain unknowns,

($iii$) convergence of the discrete unknowns towards their continuous corresponding ones.

By stability we mean the coercivity of the scheme. This property amounts to establish energy estimates using appropriate discrete norms. The stability property allows also to guarantee the existence of solutions to the equations of the numerical scheme. The second property ensures that the discrete solutions are acceptable and meaningful from a physical perspective. These points are crucial to conduct the analysis convergence of the proposed discretizations. The latter is essentially based on compactness results consisting of space and time translates estimations together with Kolmogorov's criterion [30, 69]. The ultimate goal of the convergence part is to claim that the

sequence of approximate solutions becomes closer to the solution of the problem of interest as the mesh size decreases. A fourth property of importance that can be investigated is the a posteriori error analysis. This step consists of providing an error estimate of the computed solution to the solution of the continuous mathematical model. It additionally requires intricate techniques and tools to prove these kinds of error estimates. The investigation of such an estimate is beyond the scope of the present dissertation.

Uncountable numerical schemes for the approximation of solutions to the two-phase flow models have been extensively implemented and studied in the recent decades. However, there are only a few works dealing with the convergence analysis of such a scheme, especially in the case of compressible flows. We thereafter review the main works related to this context. To begin with, traditional finite difference methods are described in the books [17, 108]. Finite element and mixed finite element schemes have been greatly of importance to deal with general meshes and constraints on some physical data, see for instance [16, 52, 106, 110, 121]. In these references, the convergence of schemes is carried out based on rigorous error estimates. They can be used whenever the governing equations are of diffusion-dominated type. When the convection phenomenon is more important than the diffusion, these schemes may lead to unphysical oscillations. This is due to their weakness to capture the hyperbolic character of the model. For this reason, finite volume discretizations [61, 69] have been developed. They are naturally adopted for conservation laws thanks to their local conservativity of the fluxes. The pioneer finite volume scheme is the two-point flux approximation (TPFA). It enjoys a standard stencil which ensures a simple implementation of the method as well as satisfying the monotonicity property. A convergence result of this approach for incompressible and immiscible flows is provided in [103]. Such a convergence analysis was performed in [12, 72]. Using the feature of the global pressure, a TPFA scheme has been analyzed in [23] for a degenerate compressible system. Similar ideas are extended to the case where the density depends on its own pressure in the paper [113]. Although it is of great advantage, the TPFA methodology requires an orthogonality condition on the mesh and a scalar permeability tensor. To dispense with these restrictions, reliable schemes combining finite elements and finite volumes are addressed in [1, 45, 73, 92, 114] for degenerate parabolic equations arising in the diphasic models. Moreover, the study of a finite element and finite volume discretization owing to a posteriori error analysis can be found for instance in [43, 118]. The gradient schemes method, which includes various discretizations, has been devoted to the incompressible flows in [68]. For more information on this topic, the reader can see the monograph [62].

Our goal in this thesis is to develop and analyze finite volume schemes satisfying the aforementioned properties $(i)$–$(iii)$ for a degenerate compressible two-phase flow system in anisotropic porous media on almost general meshes. The convergence of all the proposed numerical schemes is based on the a priori analysis and classical compactness estimates. Then, we propose two different methods of the finite volume family. The first approach is referred to as the control volume finite element (CVFE) discretization which belongs to the vertex-centered finite volume (VCFV) schemes. It has been studied in several works [34, 35, 36, 40, 42, 45, 67, 92]. The strength of this discretization lies in its ability to deal with general meshes and offers a good approximation of the diffusion counterpart. However, it sometimes suffers from an excessive numerical diffusion. This situation may lead to inaccurate results and therefore underpredict the phenomenon modeled by the considered equations. On the other hand, the method involves finite element meshes (triangles in 2D) which makes it somewhat incapable to treat cases where the meshes coming from physical applications are predefined and too distorted. This has led us to the investigation of other reliable finite volumes schemes. So, the second approach is centered on the schemes of DDFV (Discrete Duality Finite Volume) type. Its analysis in two dimensions has been reported in [11, 39, 46, 59, 60] for various kinds of problems. It was also extended to the dimension three in [8, 9, 56, 97]. The

advantage of the DDFV framework consists in providing a consistent reconstruction of the gradient, which is in duality with the discrete divergence operator thanks to the Stokes-like formula. Furthermore, the method is stable and accurate of second order even if the mesh is too distorted and the ratio of anisotropy is important.

The remainder of the manuscript is structured as follows : in Chapter 2 we start off a classical VCFV discretization for a compressible two-phase flow system composed of degenerate parabolic equations. The convection part is approximated according to the upstream technique including a crucial choice of the mean value of the density on the interfaces. The diffusion term is approximated with a centered scheme. Assuming that the transmissibilities are nonnegative, the discrete saturation remains bounded between 0 and 1 and uniform estimates on the discrete gradients are established. By virtue of space and times translates estimations, the convergence of the scheme is shown. At the end of the chapter we present numerical simulations to illustrate the displacement of water through the domain of computation.

In chapter 3 we generalize the method studied in the previous chapter into a positive CVFE scheme in order to get rid of the sign of stiffness coefficients. This essentially allows us to take the anisotropy into account and utilize general triangular meshes. So, the idea is to write the system in an equivalent form known as the fractional flow formulation. The core of the discretization relies on the treatment of the diffusion contribution. This term is approximated thanks to the upstream approach with respect to the sign of the transmissibilities. The convection fluxes are also approximated in the same spirit as in the preceding chapter. Then, the discrete maximum principle on the gas saturation holds. The coercivity-like property is proven and the control of the discrete gradients is derived. In addition to these main ingredients, compactness estimates are shown. Therefore the convergence of the scheme towards a weak solution of the continuous problem is investigated. Numerical simulations are exhibited and aim to illustrate the impact of the anisotropy on the flow of water in the porous medium.

Chapter 4 is devoted to the construction and analysis of a positive DDFV scheme for unsteady degenerate diffusion equations in two dimensional space. A particular attention is paid to this problem since it turns out to be somehow the cornerstone for the study of the diphasic model encountered in the last chapters and many others models used in hydrology, biology and medicine. Th idea of the presented method is to approximate the fluxes thanks to monotone schemes which ensure the unconditional coercivity of the DDFV approach. Accordingly, one obtains a discrete maximum principle on the solution and establishes easily an a priori estimate. Then, the numerical scheme converges up to a penalization term which is not needed in practice. Numerical results confirms that the method is positive and accurate with optimal convergence rates as expected.

# Chapter 2

# Numerical analysis of a vertex-centered finite volume scheme for a gas-water porous media flow model

This chapter is concerned with the numerical study of a vertex-centered finite volume scheme for a coupled system modeling the simultaneous displacement of gas and water in porous media. This approach requires two kinds of meshes, a primal mesh and barycentric dual mesh. We will then use a $\mathbb{P}_1$-finite element method on the first one while we perform a finite volume discretization on the second one. The analysis of the numerical scheme leans on the non-negativity of the stiffness coefficients and on a classical regularity of the mesh. Some numerical simulations are given in two space dimensions to illustrate the proper behavior of the proposed scheme.

## 2.1 Introduction

As advertised in the previous chapter, the process of modeling flows in porous media takes a privileged place in solving some real-world problems. In this chapter, we are particularly interested in a simplified compressible two-phase flow model with two main variables. It is a system composed of two coupled and degenerate parabolic equations. Many efforts have been put into studying the existence, uniqueness and properties of their solutions. From a physical point of view, illustrating the behavior of these solutions is necessary so that one can grasp the phenomenon involving the considered model. To this end, designing efficient and accurate numerical methods is still a suitable compromise to discover the secret behind the mathematical model.

Plenty of numerical methods wit diverse assumptions on data have been addressed for solving the equations of the two-phase flow model. First, finite difference schemes have been investigated in [17, 108]. This method is generally avoided when system's inputs are not smooth enough and the domain of study is not structured. Due to their ability to deal with complex geometric forms, finite element methods have also been the subject of several works [3, 47, 51]. They are efficient and more accurate for diffusions type problems, but they may produce oscillations in case of an important advection. Being cheap and reliable, finite volume schemes have received a huge attention in the last decades [69, 61]. Basically, they are constructed from a balance equation together with a proper approximation of the fluxes. They are often preferred and used to discretize partial differential equations resulting from conservation laws, and including high dominated convection

terms [1, 12, 21, 72, 103]. Furthermore, they naturally enjoy the local mass conservation property and they can guarantee the discrete maximum principle which are two fundamental materials to analyze such a finite volume scheme. Indeed, the conservation of the fluxes across the interfaces of the control volumes allows to establish some essential arguments for the convergence of the scheme. In addition, the maximum principle gives information about the physical admissibility of discrete solutions. For instance, the saturation is between 0 and 1 by its nature, then any proposed approximation should persevere these ranges, otherwise the obtained solution would not be physically accepted.

The simplest finite volume method is the famous two-point flux approximation (TPFA). It consists of approximating the fluxes by using only the values of the solution at the centers of the two control volumes sharing the same interface. The convergence analysis of TPFA schemes for compressible/incompressible flows has been carried out in a few works [12, 23, 103, 113]. Nevertheless, this approach stipulates an isotropic permeability tensor and an orthogonality constraint on the mesh. These conditions are too restrictive compared to the physical data that already exist in practical applications. To relax the impact of this issue, some schemes have been developed in [1, 45, 92, 114]. The main point of these contributions lies in combining the features of the finite element method, providing a simple discretization of the gradient, and the locally conservativity characteristic of the finite volume approximation. More generally, a new mathematical framework known as the gradient schemes method, including a large variety of discretizations, has been studied and analyzed for incompressible two-phase flows in [62, 68].

The ultimate goal of this chapter is the convergence study of a nonlinear vertex-centered finite volume scheme (VCFV), based on a $\mathbb{P}_1$-finite element approach, in order to approximate the mathematical model of the diphasic flow including the compressibility of the nonwetting phase. We then broaden the ideas presented and developed in [1, 45, 92] to a coupled system made of two degenerate parabolic equations, which are derived from the mass conservation equation for each phase together with the generalized Darcy law. The convergence of the numerical scheme relies on classical compactness arguments.

Before we go further, let us sketch out, without exclusivity, the dating of the used VCFV method. The idea of this method was first introduced in [19] in order to deal with convection-diffusion problems with a high Peclet number. It was also discussed in [34, 35, 36], where the authors analyzed and applied it to elliptic problems. Later on, the convergence analysis of such a scheme for a linear system consisting of a hyperbolic equation and an elliptic one was established in [67]. We also mention the works of Feistauer et al. [76] that developed and analyzed a VCFV scheme for a boundary-value problem incorporating a nonlinear conservation law with a diffusion term. In addition, some variants of the VCFV methodology have been proposed in [50, 91] to discretize the two-phase flow model while no convergence proof is provided. In [45, 92], the convergence analysis of a VCFV scheme has been established for a system involving degenerate convection-diffusion-reaction equations.

The remainder of this chapter is articulated as follows. Section 2.2 presents the mathematical formulation of the compressible two-phase flow in porous media together with mixed boundary conditions and initial data. Next, in Section 2.3 we define the used meshes namely primal and dual meshes and we introduce the discrete functionals spaces. Section 2.5 is devoted to sketching out the VCFV discretization and to how we derive the expected scheme. Section 2.6 is dedicated to the discrete maximum principle and the a priori estimates on the discrete gradients. In Section 2.7, the existence of a discrete solution to the combined scheme is shown. Section 2.8 is concerned with the space and time translates estimations. Finally, in Section 2.9 we concatenate the overall properties of these sections to prove the convergence of the discrete solutions sequence towards a weak solution to the continuous problem, which is the main result of this chapter.

## 2.2 Model's equations

In this section, we briefly give the mathematical formulation for a compressible two-phase flow model in heterogeneous and anisotropic porous media. It is derived from the generalized Darcy law together with the mass conservation equation for each phase. The two considered phases are: gas as a nonwetting phase and water as a wetting-phase. We restrict ourselves to the case where the first fluid is compressible and the second one is incompressible.

Let $\Omega$ be a bounded open subset of $\mathbb{R}^d$, $d \in \{2,3\}$, and $\mathfrak{T}$ a fixed positive real number. We denote $Q_{\mathfrak{T}} = \Omega \times (0,\mathfrak{T})$. Following [83], the governing equations for the compressible flow are :

$$\phi(x)\partial_t(\rho_\alpha(p_\alpha)s_\alpha) + \mathrm{div}(\rho_\alpha(p_\alpha)V_\alpha) + \rho_\alpha(p_\alpha)s_\alpha q^P = \rho_\alpha(p_\alpha)s_\alpha^I q^I, \quad (\alpha = g,w) \text{ in } Q_{\mathfrak{T}}, \tag{2.2.1}$$

where $\phi$ is the porosity of the medium, $u_\alpha$ is the saturation of the $\alpha$-phase, $\rho_\alpha$ is the density of the phase $\alpha$, $q^p$ is a production term, $q^I$ is an injection term, and $s_\alpha^I$ is the saturation of the injected fluid. $V_\alpha$ is the velocity of the $\alpha$-phase given by the generalized Darcy law (i.e. see[18, 22])

$$V_\alpha = -\frac{K_{r\alpha}(s_\alpha)}{\mu_\alpha}\Lambda\big(\nabla p_\alpha - \rho_\alpha(p_\alpha)\vec{\mathbf{g}}\big), \quad \alpha = g,w, \tag{2.2.2}$$

where $\Lambda$ is the absolute permeability of the porous medium, $K_{r\alpha}$ is the relative permeability of the $\alpha$-phase, $\mu_\alpha$ is the viscosity of the phase $\alpha$, which is constant in our study, $p_\alpha$ the pressure of the phase $\alpha$ and $\vec{\mathbf{g}}$ is the gravitational acceleration. We assume that the whole porous medium is occupied with the two fluids, meaning that the following identity is fulfilled

$$s_w + s_g = 1. \tag{2.2.3}$$

In a capillary tube, the contact between the two fluids generates a curvature because of the difference between their corresponding pressures. This jump stands for the capillary effects. The physical function encoding this difference calls the capillary pressure, denoted by $p_c$, and it is assumed to be only in terms of the nonwetting phase saturation

$$p_c(s_g) = p_g - p_w.$$

We hereafter denote by $s$ the gas saturation instead of $s_g$. According to laboratory experiments, see for instance [22], it has been exhibited that the function $s \longrightarrow p_c(s)$ is nondecreasing, $\big(\frac{dp_c(s)}{ds} \geq 0$, for any $s \in [0,1]\big)$. Furthermore, when the gas fluid is completely disappeared this function degenerates, thus one gets $p_c(s=0) = 0$.

From a theoretical point of view, one cannot control the energy of the above system since the relative permeabilities degenerate whenever the saturation vanishes or is equal to 1. At the discrete level, this degeneracy does not also allow to control the discrete gradients of both : the gas and water pressures. This issue has been underlined in several studies [23, 103]. In order to tackle this inconvenience, we make use of the global pressure formulation, which has been originally invented by *Chavent* et al. in [47]. This formulation consists of introducing an intermediary pressure so that the impact of the degeneracy and the strong coupling of the unknowns can be alleviated.

Hereafter $p$ will stand for the global pressure. We then recall that $p$ is defined via the following relationship

$$M(s)\nabla p = M_w(s)\nabla p_w + M_g(s)\nabla p_g, \tag{2.2.4}$$

where $M_\alpha$ designates the mobility of the $\alpha$-phase and $M$ is the total mobility. These quantities are explicitly defined by

$$M_\alpha = \frac{K_{r\alpha}}{\mu_\alpha}, \quad M(s) = M_w(s) + M_g(s).$$

On the other hand, the global pressure $p$ can be viewed as a modification of the gas or water pressure i.e.

$$p = p_g + \overline{p}(s) = p_w + \tilde{p}(s), \tag{2.2.5}$$

where we have set the artificial pressures $\overline{p}, \tilde{p}$, to

$$\overline{p}(s) = -\int_0^s \frac{M_w(u)}{M(u)} p_c'(u)\,\mathrm{d}u \quad \text{and} \quad \tilde{p}(s) = \int_0^s \frac{M_g(u)}{M(u)} p_c'(u)\,\mathrm{d}u. \tag{2.2.6}$$

Now substituting (2.2.5) into (2.2.1) gives rise to a new nonnegative function denoted $\gamma$ whose expression is :

$$\gamma(s) = \frac{M_w(s)M_g(s)}{M(s)} p_c'(s) \geq 0.$$

Let us next perform the Kirchoff transformation $\xi$ of the function $\gamma$. It is simply given by a primitive of $\gamma$ on the interval $[0,1]$ :

$$\begin{aligned} \xi(s) &= \int_0^s \gamma(u)\,\mathrm{d}u = \int_0^s \frac{M_w(u)M_g(u)}{M(u)} p_c'(u)\,\mathrm{d}u, \\ &= -\int_0^s M_g(u)\overline{p}'(u)\,\mathrm{d}u = \int_0^s M_w(u)\tilde{p}'(u)\,\mathrm{d}u. \end{aligned}$$

Thanks to the aforementioned relations, one can express main terms of the velocities given in (2.2.2) with the aid of the global pressure $p$ and the function $\xi$. Then

$$M_w(s)\nabla p_w = M_w(s)\nabla p + \nabla\xi(s), \tag{2.2.7}$$

$$M_g(s)\nabla p_g = M_g(s)\nabla p - \nabla\xi(s). \tag{2.2.8}$$

Consequently

$$M_g(s)\left|\nabla p_g\right|^2 + M_w(s)\left|\nabla p_w\right|^2 = M(s)\left|\nabla p\right|^2 + \frac{M_g M_w}{M}\left|\nabla p_c(s)\right|^2.$$

We point out that the degeneracy of the mobilities $M_g$ and $M_w$ makes it impossible to get a hand on the energy of the system of interest. Nonetheless, the previous identity states that this energy can be estimated by controlling the gradient of the global pressure and the gradient of capillary term $\xi$.

The convergence analysis of the numerical scheme that we will propose later on amounts to impose $s_g^I = 0$. We can also assume that the gas density varies slowly with respect to the capillary pressure (see [47] for more details). In the sequel, we consider $\rho_g \approx \rho(p)$. Now, Plugging (2.2.7)-(2.2.8) into the system (2.2.1)-(2.2.2), we derive the global pressure formulation for the compressible two-phase flow

$$\begin{aligned} &\partial_t(\phi\rho(p)s) - \operatorname{div}\Lambda\rho(p)M_g(s)\nabla p - \operatorname{div}\Lambda\rho(p)\nabla\xi(s) \\ &\qquad + \operatorname{div}\Lambda\rho^2(p)M_g(s)\vec{\mathbf{g}} + \rho(p)sq^P = 0, \end{aligned} \tag{2.2.9}$$

$$\begin{aligned} &\partial_t(\phi s) + \operatorname{div}\Lambda M_w(s)\nabla p - \operatorname{div}\Lambda\nabla\xi(s) \\ &\qquad - \operatorname{div}\Lambda M_w(s)\vec{\mathbf{g}} + sq^P = q^P - q^I. \end{aligned} \tag{2.2.10}$$

where the main unknowns are, from now on, the global pressure $p$ and the gas saturation $s$. Mixed boundary conditions and initial conditions are added to close the system (2.2.9) -( 2.2.10). Then,

the boundary $\partial\Omega$ of $\Omega$ is divided into two parts $\Gamma_D$ and $\Gamma_N$ with $|\Gamma_D| > 0$. On $\Gamma_D$, we prescribe a Dirichlet condition and on $\Gamma_N$ we have a Neumann condition as follows

$$\begin{cases} p(x,t) = 0,\ s(x,t) = 0 & \text{on} \quad \Gamma_D \times (0,\mathfrak{T}) \\ V_w.\mathbf{n} = V_g.\mathbf{n} = 0 & \text{on} \quad \Gamma_N \times (0,\mathfrak{T}) \end{cases}, \tag{2.2.11}$$

where $\mathbf{n}$ is the outward normal vector to $\Gamma_N$. Furthermore, the initial conditions read

$$p(x,0) = p^0(x) \quad \text{in } \Omega, \tag{2.2.12}$$

$$s(x,0) = s^0(x) \quad \text{in } \Omega. \tag{2.2.13}$$

Let us now list the essential assumptions on the physical data and coefficients. They are classical for the study of the two-phase flow model.

($\mathrm{H}_0$) The initial global pressure $p^0$ is in $L^2(\Omega)$ and the initial gas saturation $s^0$ belongs to $L^\infty(\Omega)$ with $0 \leq s^0(x) \leq 1$ a.e. $x \in \Omega$.

($\mathrm{H}_1$) The porosity $\phi$ is a $L^\infty$-function and there exist two positive constants $\phi_0$ and $\phi_1$ such that $\phi_0 \leq \phi(x) \leq \phi_1$ a.e. $x \in \Omega$.

($\mathrm{H}_2$) The gas (resp. water) mobility $M_g$, (resp. $M_w$) is a nondecreasing (resp. nonincreasing) continuous function from $[0,1]$ to $\mathbb{R}$ with $M_g(s) = 0$ for every $s \in ]-\infty, 0]$ and $M_w(s) = 0$ for every $s \in [1, +\infty[)$. Moreover, there exists a positive constant $m_0$ such that

$$m_0 \leq M_g(s) + M_w(s), \quad \forall s \in [0,1].$$

($\mathrm{H}_3$) The absolute permeability $\Lambda$ is a map from $\Omega$ to $\mathcal{S}_d(\mathbb{R})$, where $\mathcal{S}_d(\mathbb{R})$ is the space of $d$-square symmetric matrices. It is also is assumed to be in $L^\infty(\Omega)^{d\times d}$. Furthermore, $\Lambda$ verifies the ellipticity condition i.e. there exist positive constants $\underline{\Lambda}$ and $\overline{\Lambda}$ such that

$$\underline{\Lambda}|\zeta|^2 \leq \Lambda(x)\zeta\cdot\zeta \leq \overline{\Lambda}|\zeta|^2, \text{ for all } \zeta \in \mathbb{R}^d \text{ and a.e. } x \in \Omega.$$

($\mathrm{H}_4$) The function $\gamma$ belongs to $\mathcal{C}^0(\mathbb{R}, \mathbb{R}^+)$ with

$$\begin{cases} \gamma(s) > 0 & \text{for } 0 < s < 1 \\ \gamma(0) = \gamma(1) = 0 & \text{otherwise} \end{cases}.$$

we also assume that $\xi^{-1}$ is a $\theta$-Hölder function on $[0, \xi(1)]$ with $\theta \in (0,1]$. This means that there exists a positive constant $C$ such that for all $a, b \in [0,\xi(1)]$, $|\xi^{-1}(a) - \xi^{-1}(b)| \leq C|a-b|^\theta$.

($\mathrm{H}_5$) The injection term $q^I$ and the production one $q^P$ are $L^2$-functions with $0 \leq q^P(x,t), q^I(x,t)$ a.e. $(x,t) \in Q_{\mathfrak{T}}$.

($\mathrm{H}_6$) The density $\rho \in \mathcal{C}^1(\mathbb{R},\mathbb{R})$ is strictly increasing and uniformly bounded : $\rho_0 \leq \rho(p_g) \leq \rho_1$ for some positive constants $\rho_0, \rho_1$.

We define the natural space where weak solutions are sought

$$H^1_{\Gamma_D}(\Omega) = \{v \in H^1(\Omega) \ / \ v = 0 \text{ on } \Gamma_D\},$$

which is a Hilbert space endowing with the norm

$$||v||_{H^1_{\Gamma_D}(\Omega)} = ||\nabla v||_{(L^2(\Omega))^d}$$

In the rest of this chapter, we assume that the hypotheses $(H_0)$-$(H_6)$ are fulfilled. Now, we are in a position to give the definition of weak solutions.

**Definition 2.2.1.** *(Weak solutions) A pair of measurable functions $(p, s)$ is said to be a weak solution to the problem (2.2.9)-(2.2.12) provided*

$$\begin{aligned} 0 \leq s \leq 1 \ &a.e.\ in\ Q_{\mathfrak{T}}, \\ \xi(s) \ &\in L^2(0,\mathfrak{T}; H^1_{\Gamma_D}(\Omega)), \\ p \ &\in L^2(0,\mathfrak{T}; H^1_{\Gamma_D}(\Omega)), \end{aligned}$$

*and for every $\varphi, \psi \in \mathcal{C}_c^\infty(\Omega \times [0,\mathfrak{T}))$, one has*

$$\begin{aligned} &-\int_{Q_{\mathfrak{T}}} \phi\rho(p)s\partial_t\varphi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\rho(p^0)s^0\varphi(x,0)\,\mathrm{d}x \\ &\quad + \int_{Q_{\mathfrak{T}}} \rho(p)M_g(s)\Lambda\nabla p\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} \rho(p)\Lambda\nabla\xi(s)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t \\ &\quad - \int_{Q_{\mathfrak{T}}} \rho^2(p)M_g(s)\Lambda\vec{\boldsymbol{g}}\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} \rho(p)sq^P\varphi\,\mathrm{d}x\,\mathrm{d}t = 0, \end{aligned} \tag{2.2.14}$$

$$\begin{aligned} &-\int_{Q_{\mathfrak{T}}} \phi s\partial_t\psi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi(x)s^0\psi(x,0)\,\mathrm{d}x - \int_{Q_{\mathfrak{T}}} M_w(s)\Lambda\nabla p\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{Q_{\mathfrak{T}}} \Lambda\nabla\xi(s)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} \rho_w M_w(s)\Lambda\vec{\boldsymbol{g}}\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &+ \int_{Q_{\mathfrak{T}}} sq^P\psi\,\mathrm{d}x\,\mathrm{d}t = \int_{Q_{\mathfrak{T}}} (q^P - q^I)\psi\,\mathrm{d}x\,\mathrm{d}t. \end{aligned} \tag{2.2.15}$$

For the existence of a weak solution to the problem (2.2.14)-(2.2.15), we refer to this work [84].

## 2.3 Meshes and basic notations

In this section, we set up the main discrete tools and notations that are necessary to discretize the considered model. To this purpose, we will define two kinds of meshes of the domain $\Omega$ ; a primal mesh, which is a triangulation if $d = 2$ or a tetrahedralization if $d = 3$, and a barycentric dual mesh which is constructed from the primal discretization. For the sake of simplicity, we will restrict our attention to the case where $d = 2$. We further take into account polygonal connected domains.

A primal mesh $\mathcal{T}$ is a conforming triangulation of $\Omega$ in the sense of the finite element method; that is, the intersection of two triangles is either an edge, a vertex or the empty set and $\overline{\Omega} = \cup_{T\in\mathcal{T}}\overline{T}$. The set of vertices of $\mathcal{T}$ (resp. $T \in \mathcal{T}$) is denoted by $\mathcal{V}$ (resp. $\mathcal{V}_T$). We designate by $\mathcal{E}$ (resp. $\mathcal{E}_T$) the set of all edges of $\mathcal{T}$ (resp. $T$). For a triangle $T \in \mathcal{T}$, we define $x_T$ as its barycenter, $h_T = \mathrm{diam}(T)$ its diameter, and $|T|$ its Lebesgue measure. Let $\varrho_T$ be the diameter of the largest ball inscribed within the triangle $T$. The size and the regularity of the triangulation $\mathcal{T}$ are respectively denoted by $h_\mathcal{T}$ and $\theta_\mathcal{T}$. They are defined by

$$h_\mathcal{T} := \max_{T\in\mathcal{T}}(h_T), \quad \theta_\mathcal{T} := \max_{T\in\mathcal{T}} \frac{h_T}{\varrho_T}.$$

The construction of the dual barycentric mesh involves the vertices of the primal mesh, the centers of the edges and the barycenters of the triangles. For each vertex $K \in \mathcal{V}$ we associate a unique control volume, denoted $\omega_K$, of the dual mesh. We also denote by $\mathcal{V}_D$ the set of these dual sub-domains, then $\Omega = \cup_{K\in\mathcal{V}_D}\omega_K$. Each dual cell is obtained by connecting (in the positive sense

for instance) the barycenter of each triangle whose vertex is $K$ with the midpoint of the edges having $K$ as an end point. For two vertices $K, L \in \mathcal{V}_T$, $\sigma^T_{KL}$ denotes the dual interface contained in $T$ and intersects with the segment $[KL]$ whose extremities are $K$ and $L$. By $|\sigma^T_{KL}|$, we mean the length of the interface $\sigma^T_{KL}$ and by $n^T_{\sigma_{KL}}$ the unit normal vector to $\sigma^T_{KL}$ pointing from $K$ to $L$. Next, for $K \in \mathcal{V}$, $|\omega_K|$ is the $d$ dimensional Lebesgue measure of $\omega_K$. We additionally designate by $\mathcal{K}_T$ the set of all triangles sharing the vertex $K$.

We now assume that the primal mesh is regular in the sense that there exists a positive constant $\theta_0$ such that for any sequence of discretizations $\{\mathcal{T}_m\}_{m\in\mathbb{N}}$, we have

$$\theta_{\mathcal{T}_m} \leq \theta_0. \tag{2.3.1}$$

This inequality is well known as Ciarlet's condition in the finite element literature [54], it prevents the degeneracy of the triangulation. In other words, for any refinement of the mesh, the smallest angle of the triangles is bounded far away from 0.

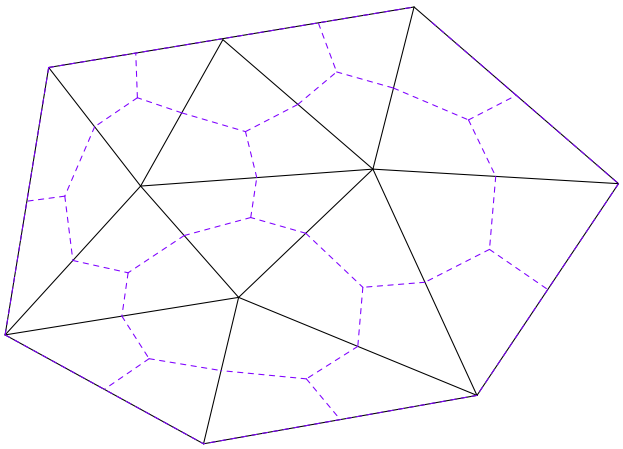


Figure 2.1: Visualization of the 2D primal and dual meshes.

We moreover consider a time discretization of the interval $(0, \mathfrak{T})$. It is given by a strictly increasing sequence of real numbers $(t^n)$, for $n = 0, \ldots, N$, such that :

$$t^0 = 0 \ < t^1 \ < \cdots < \ t^{N-1} \ < \ t^N = \mathfrak{T}.$$

The size of the time cell is denoted by $\delta t^n = t^{n+1} - t^n$, for $n = 0, \ldots, N-1$ and $\delta t = \max\limits_{n=0,\ldots,N-1} \delta t^n$ stands for the size of this discretization. To avoid heavy notations, one can assume that this subdivision is uniform, i.e. $\delta t^n = \delta t$ is constant, for every $n \in \{0, \ldots, N-1\}$.

## 2.4 Approximation spaces and discrete functions

We now describe the approximation spaces where the discrete solutions will be lied in. On one hand we consider a finite volume space usually called trial space, denoted by $W_h$, made of piecewise constant functions on the dual mesh

$$W_h = \{w_h(x) = \sum_{K\in\mathcal{V}} w_K \chi_{\omega_K}(x) \ / \ w_K \in \mathbb{R} \ \ \forall K \in \mathcal{V}\} \subset L^2(\Omega),$$

where $\chi$ is the characteristic function of $\omega_K$, which is equal to 1 on $\omega_K$ and 0 otherwise. On the other hand we define two finite dimensional spaces, denoted receptively by $X_h$, $X_h^0$ and composed of linear piecewise functions

$$X_h = \{\varphi \in \mathcal{C}^0(\overline{\Omega}), \varphi_{|T} \in \mathbb{P}_1, \forall T \in \mathcal{T}\} \subset H^1(\Omega),$$
$$X_h^0 = \{\varphi \in X_h, \phi(x_K) = 0, \forall K \in \mathcal{V},\ K \in \Gamma_D\} \subset H^1_{\Gamma_D}(\Omega).$$

The space $X_h$ has a canonical basis which is comprised of shape functions $(\varphi_K)_{K\in\mathcal{V}}$ with $\varphi_K(x_S) = \delta_{KL}$, where $\delta_{KL}$ is the Kronecker symbol. We recall that, for every $K, L \in \mathcal{V}$, one has

$$\delta_{KL} = \begin{cases} 1 & \text{if } K = L \\ 0 & \text{if } K \neq L \end{cases}.$$

For every $u_h \in X_h$, the function $u_h$ writes

$$u_h(x) = \sum_{K\in\mathcal{V}} u_K \varphi_K(x),$$

hence its gradient is defined as

$$\nabla u_h(x) = \sum_{K\in\mathcal{V}} u_K \nabla\varphi_K(x).$$

One notices that

$$\sum_{K\in\mathcal{V}} \varphi_K = 1, \sum_{K\in\mathcal{V}} \nabla\varphi_K = 0 \quad \text{and} \quad \nabla\varphi_{K_{|T}} = -\frac{|\sigma_K^T|}{2\,|T|} n_{\sigma_K^T},$$

with $\sigma_K^T$ is the edge of the triangle $T$ located in front of the vertex $K$ and $n_{\sigma_K^T}$ is the outward normal to the same interface (see Fig. 2.2). Moreover, the space $X_h$ is equipped by the following semi-norm

$$||u_h||^2_{X_h} := \int_\Omega |\nabla u_h|^2 \,\mathrm{d}x, \quad \forall u_h \in X_h.$$

This latter turns out to be a norm on $X_h^0$ thanks to the Poincaré inequality that will be defined below.

For every $n \in \{0, \ldots, N\}$ and $K \in \mathcal{V}$ we consider $u_K^n$ as an approximation of $u(x_K, t^n)$. Thus, the discrete unknowns will be denoted by $\{u_K^n\}_{\{K\in\mathcal{V},\ n=0,\ldots,N\}}$.

**Definition 2.4.1.** *(Discrete functions)*
*Consider discrete values $\{u_K^n\}_{\{K\in\mathcal{V},\ n=0,\ldots,N\}}$. We define two approximate solutions as follows:*

*(i) A finite volume solution $\tilde{u}_{h,\delta t}$ is piecewise constant and defined almost everywhere in $\bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0, \mathfrak{T})$ with*

$$\tilde{u}_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0 \chi_{\mathring{\omega}_K}(x),\ \forall x \in \bigcup_{\omega_K\in\mathcal{V}} \mathring{\omega}_K,$$
$$\tilde{u}_{h,\delta t}(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} u_K^{n+1} \chi_{\mathring{\omega}_K\times(t^n,t^{n+1}]}(x,t),\ \forall (x,t) \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T}).$$

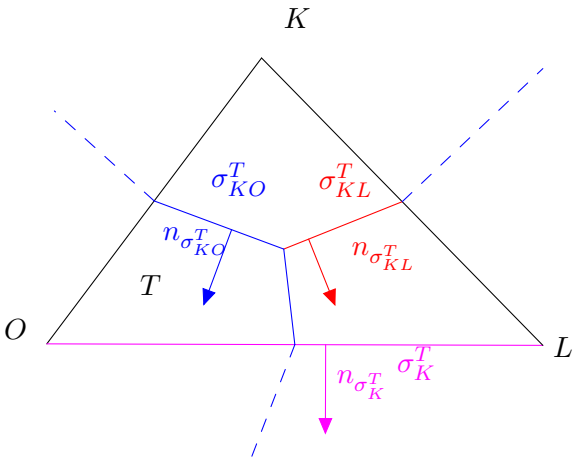



Figure 2.2: Illustration of dual interfaces and their unit normal vectors.

*(ii) A finite element solution $u_{h,\delta t}$ is a continuous function in space, which is $\mathbb{P}_1$ per triangles, and piecewise constant in time, such that :*

$$u_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0 \varphi_K(x), \quad \forall x \in \Omega,$$
$$u_{h,\delta t}(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} u_K^{n+1} \varphi_K(x) \chi_{(t^n,t^{n+1}]}(t), \quad \forall\ (x,t) \in \Omega \times (0,\mathfrak{T}).$$

To discretize nonlinear functions, we make an interpolation approximation. Let $F$ be a nonlinear function, we mean by $F(\tilde{u}_{h,dt})$ the finite volume reconstruction, which is defined almost everywhere, and by $F(u_{h,dt})$ the finite element reconstruction i.e.:

$$F(\tilde{u}_{h,\delta t})(x,0) = \sum_{K\in\mathcal{V}} F(u_K^0)\chi_{\mathring{\omega}_K}(x), \ \forall x \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K,$$
$$F(\tilde{u}_{h,\delta t})(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} F(u_K^{n+1})\chi_{\mathring{\omega}_K\times(t^n,t^{n+1}]}(x,t), \ \forall (x,t) \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T}),$$
$$F(u_{h,\delta t})(x,0) = \sum_{K\in\mathcal{V}} F(u_K^0)\varphi_K(x), \quad \forall x \in \Omega,$$
$$F(u_{h,\delta t})(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} F(u_K^{n+1})\varphi_K(x)\chi_{(t^n,t^{n+1}]}(t), \quad \forall\ (x,t) \in \Omega \times (0,\mathfrak{T}).$$

## 2.5 Numerical scheme for the diphasic flow in porous media

Stability and convergence of the scheme are two mandatory ingredients to ensure the validity of finite volume method. To this purpose, we consider careful approximations of the fluxes, across the interfaces of the dual cells, to guarantee these requirements. First, an implicit Euler scheme

in time is performed. The mobilities are approximated with the aid of an upstream scheme with respect to the sign of the discrete gradient of the global pressure. A mean value of the gas density is introduced so that the effect of compressibility on the analysis of the scheme can be removed. This particular choice is also fundamental to decouple the dependency of the variables. As it is need for diffusion processes, a centered approximation is used for the discretization of the dissipative term. In what follows, we sketch out how to get the scheme of the first equation (2.2.9) and in a similar way we write that of the second one (2.2.10).

We stress that there is no loss of generality in assuming the flow with no gravity, i.e. $\vec{\mathbf{g}} \equiv 0$ as we will point out below. Hence, the term including the gravity has been dropped. Fix a time superscript $n = 0, \cdots, N-1$ and $\omega_K \in \mathcal{V}_\mathcal{D}$ a dual control volume. As it is known for conservation laws, especially for our model, the proposed finite volume approach is essentially based on the balance equation. This standard step consists of integrating the gas equation (2.2.9) on the time-space cell $(t^n, t^{n+1}] \times \omega_K$ and applying the Green-Gauss formula. This gives

$$\begin{aligned}
&\int_{t^n}^{t^{n+1}} \int_{\omega_K} \phi(x) \partial_t \Big(\rho(p)s\Big) \,\mathrm{d}x \\
&\underbrace{- \sum_{T \in \mathcal{K}_T} \sum_{\sigma \in \mathcal{E}_K \cap T} \int_{t^n}^{t^{n+1}} \int_\sigma \rho(p) M_g(s) \Lambda \nabla p \cdot \mathbf{n}_{\sigma K} \,\mathrm{d}\sigma \,\mathrm{d}t}_{\text{convective term}} \\
&\underbrace{- \sum_{T \in \mathcal{K}_T} \sum_{\sigma \in \mathcal{E}_K \cap T} \int_{t^n}^{t^{n+1}} \int_\sigma \rho(p) \Lambda \nabla \xi(s) \cdot \mathbf{n}_{\sigma K} \,\mathrm{d}\sigma \,\mathrm{d}t}_{\text{capillary term}} \\
&+ \int_{t^n}^{t^{n+1}} \int_{\omega_K} \rho(p) s q^P \,\mathrm{d}x \,\mathrm{d}t = 0,
\end{aligned} \tag{2.5.1}$$

where $\mathcal{E}_K$ stands for the set of the edges of the dual control volume associated to $K$, $\mathbf{n}_{\sigma K}$ denotes the unit normal vector to $\sigma$ pointing outward to $\omega_K$ and $\mathrm{d}\sigma$ is the $d-1$ dimensional Lebesgue measure on $\sigma$. Next, the evolution term is approximated using a forward Euler scheme as follows

$$\begin{aligned}
&\int_{t^n}^{t^{n+1}} \int_{\omega_K} \phi(x) \partial_t \Big(\rho(p)s\Big) \,\mathrm{d}x \,\mathrm{d}t \\
&\approx \int_{\omega_K} \phi(x) \Big(\rho(p(x,t^{n+1}))s(x,t^{n+1}) - \rho(p(x,t^n))s(x,t^n)\Big) \,\mathrm{d}x, \\
&\approx \int_{\omega_K} \phi(x) \Big(\rho(\tilde{p}_{h,\delta t}(x,t^{n+1}))\tilde{s}_{h,\delta t}(x,t^{n+1}) - \rho(\tilde{p}_{h,\delta t}(x,t^n))\tilde{s}_{h,\delta t}(x,t^n)\Big) \,\mathrm{d}x, \\
&= |\omega_K| \phi_K \Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big),
\end{aligned} \tag{2.5.2}$$

where $\phi_K$ is the mean value of the porosity function $\phi$ over $\omega_K$. We would like to point out that we do not take into account the contribution of $\sigma \subset \Gamma_N$ due to the homogeneous Neumann condition specified in (2.2.11). Let us now look at the discretization of the elliptic term. To this end, we have extended the ideas presented in [1, 92] where a VCFV scheme has been investigated for degenerate parabolic equations. As a consequence we consider the following approximation which seems to be

natural

$$
\begin{aligned}
&-\sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_{\sigma}\rho(p)\Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\\
&\qquad\approx-\delta t\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\backslash\{K\}}\rho_{KL}^{n+1}\Lambda_{KL}^{T}(\xi_L^{n+1}-\xi_K^{n+1}),
\end{aligned}
\tag{2.5.3}
$$

where the coefficients $\Lambda_{KL}^T$ and $\rho_{KL}^{n+1}$ are respectively given by

$$
\begin{cases}
\Lambda_{KL}^{T}:=-\displaystyle\int_T\Lambda(x)\nabla\varphi_K\cdot\nabla\varphi_L\,\mathrm{d}x=\Lambda_{LK}^{T}, \quad \text{for} \quad K\neq L\\[2ex]
\Lambda_{KK}^{T}:=\displaystyle\sum_{L\in\mathcal{V}_T\backslash\{K\}}\Lambda_{KL}^{T}=\int_T\Lambda(x)\nabla\varphi_K\cdot\nabla\varphi_K\,\mathrm{d}x
\end{cases},
\tag{2.5.4}
$$

and

$$
\rho_{KL}^{n+1}:=\begin{cases}
\dfrac{1}{p_K^{n+1}-p_L^{n+1}}\displaystyle\int_{p_L^{n+1}}^{p_K^{n+1}}\rho(z)\,\mathrm{d}z, & \text{if} \quad p_L^{n+1}\neq p_K^{n+1}\\[2ex]
\rho(p_K^{n+1}), & \text{otherwise}
\end{cases}.
\tag{2.5.5}
$$

This expression of the density on the interface has been proposed in [23] to manage the issue related to the compressibility of the gas. It also allows to tackle the strong coupling of the system.

In case of a dominated-convection flow the upwind approximation of the hyperbolic term produces no oscillations contrary to centered schemes. We thus follow this fashion to approximate the second integral of the right hand side of (2.5.1). Therefore

$$
\begin{aligned}
&-\sum_{T\in\mathcal{K}_T}\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_{\sigma}\rho(p)M_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t\\
&\approx-\delta t\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\backslash\{K\}}\rho_{KL}^{n+1}M_{g_{KL}}^{n+1}\Lambda_{KL}^{T}(p_L^{n+1}-p_K^{n+1}),
\end{aligned}
\tag{2.5.6}
$$

where $M_{g_{KL}}^{n+1}$ is explicitly written by the formula :

$$
M_{g_{KL}}^{n+1}:=\begin{cases}
M_g(s_L^{n+1}) & \text{if } p_L^{n+1}-p_K^{n+1}\geq 0\\[2ex]
M_g(s_K^{n+1}) & \text{otherwise}
\end{cases}.
$$

We can extend the approximation (2.5.6) to more general expressions by the use of the numerical flux function $G_g$ whose entries are $s_K^{n+1}, s_L^{n+1}$ and $\delta_{KL}^{n+1}p:=p_L^{n+1}-p_K^{n+1}$. Now, for $\alpha=g,w$, the function $G_\alpha$ of three arguments $a,b,c\in\mathbb{R}$ is said to be a numerical flux if it satisfies the following items

$(C_1)$ $G_\alpha(\cdot,b,c)$ is nondecreasing for all $b,c\in\mathbb{R}$ and $G_\alpha(a,\cdot,c)$ is nonincreasing for all $a,c\in\mathbb{R}$;

$(C_2)$ $G_\alpha(a,b,c)=-G_\alpha(b,a,-c)$ for all $a,b,c\in\mathbb{R}$;

($C_3$) $G_g(a,a,c) = -M_g(a)c$, and $G_w(a,a,c) = M_w(a)c$ for all $a, c \in \mathbb{R}$, and there exists a positive constant $C$ such that

$$\forall a, b, c \in \mathbb{R} \quad |G_\alpha(a,b,c)| \leq C(|a| + |b|)|c|; \tag{2.5.7}$$

($C_4$) there exists a constant $m_0$ such that

$$\forall a, b, c \in \mathbb{R} \quad (G_w(a,b,c) - G_g(a,b,c))c \geq m_0|c|^2; \tag{2.5.8}$$

($C_5$) there exists a modulus of continuity $\eta : \mathbb{R}^+ \longrightarrow \mathbb{R}^+$ such that

$$\forall a, b, c, a', b' \in \mathbb{R} \quad |G_\alpha(a,b,c) - G_\alpha(a',b',c)| \leq \eta(|a-a'| + |b-b'|)|c|. \tag{2.5.9}$$

**Remark 2.5.1.** *In order to obtain the numerical flux $G_\alpha$ employed in (2.5.6), we consider the nondecreasing part $M_{\alpha\uparrow}$ and the nonincreasing part $M_{\alpha\downarrow}$ of the mobility function $M_\alpha$. As a result*

$$G_\alpha(a,b;c) = c^+\Big(M_{\alpha\uparrow}(a) + M_{\alpha\downarrow}(b)\Big) - c^-\Big(M_{\alpha\uparrow}(b) + M_{\alpha\downarrow}(a)\Big),$$

*where $c^+ = \max(c,0)$ and $c^- = -\min(c,0)$. Let us check that $G_\alpha$ is well-defined. We know that $M_g$ is a nondecreasing function whereas $M_w$ is a nonincreasing function. We then get*

$$\begin{aligned} G_g(a,b;c) &= -M_g(b)c^+ + M_g(a)c^-, \\ G_w(a,b;c) &= M_w(b)c^+ - M_w(a)c^-. \end{aligned}$$

*As a consequence, the properties ($C_1$)-($C_3$) and ($C_5$) hold. To verify the condition ($C_3$), one computes*

$$\begin{aligned} \Big(G_w(a,b;c) - G_g(a,b;c)\Big)c &= (M_g(b) + M_w(b))c^{+^2} + (M_g(a) + M_w(a))c^{-^2} \\ &\geq m_0 c^2. \end{aligned} \tag{2.5.10}$$

*We indicate that this inequality will be of a great importance to establish the energy estimates on the discrete gradient of the global pressure $p$.*

Therefore (2.5.6) becomes

$$\begin{aligned} &- \sum_{\sigma \in \mathcal{E}_K \cap T} \int_\sigma \rho(p) M_g(s) \Lambda \nabla p \cdot \mathbf{n}_{\sigma K} \, \mathrm{d}\sigma \\ &\qquad \approx \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} \rho_{KL}^{n+1} \Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p), \end{aligned} \tag{2.5.11}$$

where

$$G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) := -M_g(s_L^{n+1})(\delta_{KL}^{n+1} p)^+ + M_g(s_K^{n+1})(\delta_{KL}^{n+1} p)^-.$$

To summarize, the numerical scheme reads

$$p_K^0 = \frac{1}{|\omega_K|} \int_{\omega_K} p^0(x) \, \mathrm{d}x, \qquad \forall K \in \mathcal{V}, \tag{2.5.12}$$

$$s_K^0 = \frac{1}{|\omega_K|} \int_{\omega_K} s^0(x) \, \mathrm{d}x, \qquad \forall K \in \mathcal{V}. \tag{2.5.13}$$

$$\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1}-\rho(p_K^n)s_K^n\Big)+\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\rho_{KL}^{n+1}\Lambda_{KL}^T G_g(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)$$
$$-\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\rho_{KL}^{n+1}\Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1})$$
$$+\delta t\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1}=0, \tag{2.5.14}$$

$$\phi_K\Big(s_K^{n+1}-s_K^n\Big)+\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T G_w(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)$$
$$-\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1})$$
$$+\delta t(s_K^{n+1}-1)q_{P,K}^{n+1}=-\delta t q_{I,K}^{n+1},\quad \forall n=0,\dots,N-1,\quad \forall K\in\mathcal{V}, x_K\notin\Gamma_D. \tag{2.5.15}$$

We indicate that every solution to the numerical scheme is known on $\Gamma_D$ according to (2.2.11). Therefore, the equations corresponding to the control volumes whose centers are located at the boundary $\Gamma_D$ do not contribute in the above system.

The coefficient $\Lambda_{KL}^T$ is referred to as the transmissibility between two neighbor control volumes $\omega_K$ and $\omega_L$. As we are interested in the monotony property of the numerical scheme, the sign of $\Lambda_{KL}^T$ is of a huge importance. Precisely, in case that all of these transmissibilities are nonnegative the discrete gas saturation stays in the physical ranges of its initial state as we will see below, otherwise it may exceed these ranges. For instance, if $\Lambda=\lambda I$, where $I$ is the identity matrix, and all of the angles of the triangles are less than $\pi/2$, one has $\Lambda_{KL}^T\geq 0$ for every $\sigma_{KL}^T$. During this chapter, we will assume that:

($H_7$) all the coefficients $\Lambda_{KL}^T$ are nonnegative.

In the next chapter, we will deal with the general case. As a matter of fact, the problem necessitates a reformulation of the convective term as well as the construction of a Godunov-like scheme instead of a centered one in order to correct the diffusive counterpart. This strategy will ensure the physical admissibility of the computed saturation and the satisfiability of the numerical scheme.

**Remark 2.5.2.** *Taking into account gravitational effects ($\vec{g}\neq 0$), a new term denoted by $F_{gK}$ is added to the first equation (2.5.14) of the scheme. This term is the approximation of the integral $\int_{\partial K}\rho_g^2(p)M_g(s)\Lambda\vec{g}\cdot\mathbf{n}\,\mathrm{d}\sigma$. Using the upwind scheme, it is given by*

$$F_{gK}=\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}|\sigma_{KL}^T|\left(\rho_{KL}^{n+1}\right)^2\Big(M_g(s_K^{n+1})Z_{KL}^T-M_g(s_L^{n+1})Z_{LK}^T\Big),$$

*where $Z_{KL}^T=\left(\Lambda\vec{g}\cdot n_{KL}^T\right)^+=\left(\Lambda\vec{g}\cdot n_{LK}^T\right)^-$. In the same way, we add the following expression, denoted $F_{wK}$, to the equation (2.5.15)*

$$F_{wK}=\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}|\sigma_{KL}^T|\,\rho_w\Big(M_w(s_L^{n+1})Z_{KL}^T-M_w(s_K^{n+1})Z_{LK}^T\Big).$$

*Thanks to the monotonicity of the mobilities, the functions $F_{gK}$ and $F_{wK}$ are nondecreasing with respect to $s_K^{n+1}$ and nonincreasing with respect to $s_L^{n+1}$. In addition, they form numerical fluxes which are consistent and conservative. As a consequence, the convergence analysis remains valid.*

## 2.6 Maximum principle and energy estimates

In this section, we prove the maximum principle on the approximate gas saturation and we uniformly estimate the discrete gradient of the global pressure $p$ and the discrete gradient of the Kirchoff function $\xi$. We admit the existence of such solutions to the numerical scheme. The existence result will be the object of the next section.

**Lemma 2.6.1.** *For a time superscript $n = 0, \dots, N-1$. Let $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$ be a solution to the combined scheme (2.5.12)-(2.5.15). Then, the computed saturation $(s_K^{n+1})_{K\in\mathcal{V}}$ belongs to the interval $[0,1]$.*

*Proof.* The proof is conducted by induction on $n$. For $n = 0$, the lemma is a direct consequence of Assumption $(H_1)$. We now assume that the sequence $(s_K^k)_{K\in\mathcal{V}} \subset [0,1]$ for $k \leq n$ and we prove the validity of the claim for $k = n+1$. So, let us consider a vertex $K$ such that $s_K^{n+1} = \min\{s_L^{n+1}\}_{L\in\mathcal{V}}$. Multiplying the gas equation of the numerical scheme (2.5.14) by $-(s_K^{n+1})^- = \min(-s_K^{n+1}, 0)$ reads

$$
\begin{aligned}
&- \phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big)(s_K^{n+1})^- \\
&- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p)(s_K^{n+1})^- \\
&+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T(\xi_L^{n+1} - \xi_K^{n+1})(s_K^{n+1})^- \\
&- \delta t\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1}(s_K^{n+1})^- = 0.
\end{aligned} \tag{2.6.1}
$$

We aim here to establish that the last three terms of the left hand side of (2.6.1) are nonnegative. Obviously, one has $s_L^{n+1} \geq s_K^{n+1}$. Using the fact that the numerical flux $G_g$ is a nonincreasing function with respect to $s_L^{n+1}$ (item $(C_1)$) together with its consistency property (item $(C_3)$), one claims

$$
\begin{aligned}
G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p)(s_K^{n+1})^- &\leq G_g(s_K^{n+1}, s_K^{n+1}; \delta_{KL}^{n+1}p)(s_K^{n+1})^- \\
&= -M_g(s_K^{n+1})\delta_{KL}^{n+1}p(s_K^{n+1})^- = 0.
\end{aligned}
$$

The last identity is satisfied since $M_g$ is extended by zero for $s \leq 0$. Consequently, the second term in the left hand side of the equation (2.6.1) is nonnegative. In addition, the Kirchoff transform $\xi$ is a nondecreasing function and the coefficients $\Lambda_{KL}^T$ are nonnegative. Thereby

$$
\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T\rho_{KL}^{n+1}(\xi_L^{n+1} - \xi_K^{n+1})(s_K^{n+1})^- \geq 0. \tag{2.6.2}
$$

We observe that $s_K^{n+1}(s_K^{n+1})^- = -\left|(s_K^{n+1})^-\right|^2$. Thanks to the induction assumption on $s_K^n$ one gets

$$
\begin{aligned}
&-\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big)(s_K^{n+1})^- \\
&\qquad = \phi_K\Big(\rho(p_K^{n+1})\left|(s_K^{n+1})^-\right|^2 + \rho(p_K^n)s_K^n(s_K^{n+1})^-\Big) \leq 0.
\end{aligned}
$$

As a result $(s_K^{n+1})^- = 0$, which entails that $s_K^{n+1} \geq 0$.

To show that $s_K^{n+1} \leq 1$ for every $n = 0, \dots, N-1$ and $K \in \mathcal{V}$, we continue by induction, but we employ this time the water equation. Let $\omega_K$ be then a dual control volume and $s_K^{n+1}$ the maximum

of the finite family $\{s_L^{n+1}\}_{L\in\mathcal{V}}$. We want to prove that $s_K^{n+1} \le 1$. To this end, we multiply the second equation (2.5.15) of the numerical scheme by $(s_K^{n+1}-1)^+$. Whence

$$\begin{aligned}
&\phi_K\Big(s_K^{n+1}-s_K^n\Big)(s_K^{n+1}-1)^+\\
&+\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T\ G_w(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)(s_K^{n+1}-1)^+\\
&-\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T\ (\xi_L^{n+1}-\xi_K^{n+1})(s_K^{n+1}-1)^+\\
&+\delta t(s_K^{n+1}-1)q_{P,K}^{n+1}(s_K^{n+1}-1)^+ = -\delta t q_{I,K}^{n+1}(s_K^{n+1}-1)^+. \qquad (2.6.3)
\end{aligned}$$

Once more, the function $G_w$ is nonincreasing with respect to the second variable and is consistent. Thus

$$\begin{aligned}
G_w(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)(s_K^{n+1}-1)^+ &\ge G_w(s_K^{n+1},s_K^{n+1};\delta_{KL}^{n+1}p)(s_K^{n+1}-1)^+\\
&= M_w(s_K^{n+1})\delta_{KL}^{n+1}p(s_K^{n+1}-1)^+ = 0,
\end{aligned}$$

since the water mobility degenerates $M_w(s)=0$ for $s\ge 1$. By the nonnegativity of the transmissibilities and the monotonicity of the function $\xi$ we deduce

$$\begin{aligned}
&\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T(\xi_L^{n+1}-\xi_K^{n+1})(s_K^{n+1}-1)^+\\
&\qquad=\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}(\xi_L^{n+1}-\xi_K^{n+1})(s_K^{n+1}-1)^+\le 0.
\end{aligned}$$

One can see in a straightforward way that $\delta t(s_K^{n+1}-1)q_{P,K}^{n+1}(s_K^{n+1}-1)^+ = \delta t\left|(s_K^{n+1}-1)^+\right|^2 q_{P,K}^{n+1}$ and that the right hand side of (2.6.3) is nonpositive. Hence

$$\phi_K\Big(s_K^{n+1}-s_K^n\Big)(s_K^{n+1}-1)^+\le 0.$$

According to this inequality, the induction assumption and the equality

$$\Big(s_K^{n+1}-1\Big) = (s_K^{n+1}-1)^+ - (s_K^{n+1}-1)^-,$$

we demonstrate that $(s_K^{n+1}-1)^+=0$. Finally, we find that

$$s_L^{n+1}\le s_K^{n+1}\le 1,\ \forall\ \ n=0,\dots,N-1,\ \text{ and }\ \forall L\in\mathcal{V}.$$

This concludes the proof. □

**Lemma 2.6.2.** *(Integration by parts) For every $u_h, v_h \in X_h$, there holds*

$$\int_\Omega \Lambda\nabla u_h\cdot\nabla v_h\,\mathrm{d}x = \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T(u_K-u_L)(v_K-v_L). \qquad (2.6.4)$$

*In particular, if $u_h = v_h$, one has*

$$\int_\Omega \Lambda\nabla u_h\cdot\nabla u_h\,\mathrm{d}x = \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T(u_K-u_L)^2. \qquad (2.6.5)$$

*Proof.* We develop the right hand side of (2.6.4)

$$\int_\Omega \Lambda \nabla u_h \cdot \nabla v_h \, \mathrm{d}x = \sum_{T \in \mathcal{T}} \int_T \Lambda \nabla u_h \cdot \nabla v_h \, \mathrm{d}x.$$

Now in each triangle, one has

$$\int_T \Lambda \nabla u_h \cdot \nabla v_h \, \mathrm{d}x = \Big( \sum_{K \in \mathcal{V}_T} u_K \int_T \Lambda \nabla \varphi_{K_{|T}} \, \mathrm{d}x \Big) \cdot \Big( \sum_{K \in \mathcal{V}_T} v_K \nabla \varphi_{L_{|T}} \Big).$$

Thanks to (2.5.4) and $\sum_{K \in \mathcal{V}_T} \nabla \varphi_{K_{|T}} = 0$ we get (to be modified)

$$\begin{aligned}
\int_T \Lambda \nabla u_h \cdot \nabla v_h \, \mathrm{d}x &= \Big( \sum_{K \in \mathcal{V}_T} u_K \int_T \Lambda \nabla \varphi_{K_{|T}} \, \mathrm{d}x \Big) \cdot \Big( \sum_{L \in \mathcal{V}_T} v_L \nabla \varphi_{L_{|T}} \Big) \\
&= \sum_{K \in \mathcal{V}_T} \Big( u_K v_K \int_T \Lambda(x) \nabla \varphi_{K_{|T}} \cdot \nabla \varphi_{K_{|T}} \, \mathrm{d}x - \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda^T_{KL} u_K v_L \Big) \\
&= \sum_{K \in \mathcal{V}_T} \Big( \Lambda^T_{KK} u_K v_K - \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda^T_{KL} u_K v_L \Big) \\
&= \sum_{K \in \mathcal{V}_T} \Big( \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda^T_{KL} u_K v_K - \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda^T_{KL} u_K v_L \Big) \\
&= \sum_{K \in \mathcal{V}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda^T_{KL} (v_K - v_L) u_K \\
&= \sum_{\sigma^T_{KL} \in \mathcal{E}_T} \Lambda^T_{KL} (v_K - v_L)(u_K - u_L).
\end{aligned}$$

This establishes the required relationship. □

As its continuous version, the discrete Poincaré-Inequality is a practical tool for the study of coercive problems. It states that the solution is dominated by its derivatives up to a constant. Its proof can found in [64].

**Lemma 2.6.3.** *(Poincaré's Inequality) There exists $C_{Poin}$ depending only on $\Omega$ such that*

$$\|u_h\|_{L^2(\Omega)} \leq C_{Poin} \|u_h\|_{X_h}, \quad \forall u_h \in X^0_h. \tag{2.6.6}$$

Let $u_\mathcal{T} \in X_h$ and consider the piecewise constant functions $\overline{u}_\mathcal{T}, \underline{u}_\mathcal{T} : \Omega \longrightarrow \mathbb{R}$ defined by

$$\begin{aligned}
\overline{u}_\mathcal{T}(x) = \overline{u}_T &= \sup_{x \in T} u_\mathcal{T}(x), \quad \text{if } x \in T \in \mathcal{T}, \\
\underline{u}_\mathcal{T}(x) = \underline{u}_T &= \inf_{x \in T} u_\mathcal{T}(x), \quad \text{if } x \in T \in \mathcal{T}.
\end{aligned}$$

We recall the following properties whose proofs are stemmed from the finite element literature [28, 64].

**Lemma 2.6.4.** *There holds*

$$\int_\Omega |\overline{u}_\mathcal{T}(x) - \underline{u}_\mathcal{T}(x)| \, \mathrm{d}x \leq \frac{27}{2} h \int_\Omega |\nabla u_\mathcal{T}(x)| \, \mathrm{d}x.$$

**Lemma 2.6.5.** *For $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, let $u_\mathcal{T}$ and $u_\mathfrak{M}$ be respectively the piecewise linear and the piecewise constant reconstructions. Then*

$$\int_T |u_\mathcal{T}(x) - u_\mathfrak{M}(x)|^2 \,\mathrm{d}x \le ch^2 ||\nabla u_\mathcal{T}||^2_{L^2(\Omega)^d},$$

*where $c$ is an absolute constant.*

We hereafter denote $(C_i)_{i\in I}$ a finite collection of constants depending only on the data described in the list of assumptions $(H_1)$–$(H_6)$ and on the mesh regularity. We next determine some uniform estimates on the discrete gradient of the global pressure and on the function $\xi(s)$. This control of the gradients will allow us to ensure the existence of a solution of the scheme and to establish some compactness arguments.

**Proposition 2.6.1.** *For every time level $n = 0, \dots, N-1$ we consider $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$ a solution to the nonlinear system (2.5.12)-(2.5.15). Then, there exist two constants $C_p$ and $C_\xi$ depending only on $\Omega, \mathfrak{T}, \phi_1, \rho_0, \rho_1, , p^0, s^0, m_0, q^P, q^I, \underline{\Lambda}, \overline{\Lambda}$ such that*

$$\sum_{n=0}^{N-1} \delta t \left\|p_h^{n+1}\right\|^2_{X_h} \le C_p, \tag{2.6.7}$$

*and*

$$\sum_{n=0}^{N-1} \delta t \left\|\xi(s_h^{n+1})\right\|^2_{X_h} \le C_\xi. \tag{2.6.8}$$

*Proof.* Let us begin with the proof of the first inequality. To this purpose, we define the function $\mathcal{H}$ to be $\mathcal{H}(p) = g(p) + \rho(p)p$ with $g(p)' = -\rho(p)$ and let $\mathcal{B}$ be a primitive of the Kirchoff function $\xi$. Note that $\mathcal{H}(0) = 0$ and $\mathcal{H}(p) \ge 0$ for all $p \in \mathbb{R}$. We next multiply the first equation (2.5.14) and the second equation (2.5.15) of the numerical scheme by $|\omega_K|\, p_K^{n+1}$ and $|\omega_K|\, g(p_K^{n+1})$, respectively. Adding them together and summing over $K$ and $n$ gives

$$S_1 + S_2 + S_3 + S_4 = 0.$$

Each term of this identity reads

$$S_1 = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K| \phi_K \Big( (\rho(p_K^{n+1}) s_K^{n+1} - \rho(p_K^n) s_K^n) p_K^{n+1} + (s_K^{n+1} - s_K^n) g(p_K^{n+1}) \Big),$$

$$S_2 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Big( \rho_{KL}^{n+1} \Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) p_K^{n+1} + \Lambda_{KL}^T G_w(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) g(p_K^{n+1}) \Big),$$

$$S_3 = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T \Big( \rho_{KL}^{n+1} (\xi_L^{n+1} - \xi_K^{n+1}) p_K^{n+1} + (\xi_L^{n+1} - \xi_K^{n+1}) g(p_K^{n+1}) \Big),$$

$$S_4 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K| \Big( \rho(p_K^{n+1}) s_K^{n+1} q_{P,K}^{n+1} \Big) p_K^{n+1} + \Big( (s_K^{n+1} - 1) q_{P,K}^{n+1} + q_{I,K}^{n+1} \Big) g(p_K^{n+1}).$$

As it classical for the discrete time derivative, we should write or underestimate $S_1$ with some quantities incorporating only the initial and the final states of $p$ and $s$. To this end, we follow the same approach given in [23]. We first establish

$$\Big(\rho(p)s - \rho(p^*)s^*\Big)p + \Big(s - s^*\Big)\Big(\mathcal{H}(p) - \rho(p)p\Big) \geq \mathcal{H}(p)s - \mathcal{H}(p^*)s^*, \tag{2.6.9}$$

for all $s, s^* \geq 0$ and $p, p^* \in \mathbb{R}$. Developing the right hand side of the preceding inequality yields

$$\begin{aligned}
&\Big(\rho(p)s - \rho(p^*)s^*\Big)p + \Big(s - s^*\Big)\Big(\mathcal{H}(p) - \rho(p)p\Big) \\
&\qquad = s\mathcal{H}(p) - s^*\Big(g(p) + \rho(p^*)p\Big) \\
&\qquad = s\mathcal{H}(p) - s^*\mathcal{H}(p^*) + s^*\Big(\mathcal{H}(p^*) - g(p) - \rho(p^*)p\Big).
\end{aligned}$$

What is left is to show that

$$\mathcal{H}(p^*) - g(p) - \rho(p^*)p \geq 0.$$

We observe that

$$\begin{aligned}
\mathcal{H}(p^*) - g(p) - \rho(p^*)p &= g(p^*) + \rho(p^*)p^* - g(p) - \rho(p^*)p \\
&= g(p^*) - g(p) + \rho(p^*)(p^* - p) \\
&= g(p^*) - g(p) - g'(p^*)(p^* - p).
\end{aligned}$$

The concavity of the function $g$, since $g'' = -\rho' \leq 0$, entails

$$g(p) \leq g(p^*) + g'(p^*)(p - p^*).$$

Hence, (2.6.9) is proved. We make use of this fundamental inequality to deduce the following lower bound

$$\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(s_K^N\mathcal{H}(p_K^N) - s_K^0\mathcal{H}(p_K^0)\Big) \leq S_1.$$

By virtue of the discrete maximum principle, one gets

$$|S_1| \leq \sum_{K\in\mathcal{V}} |\omega_K|\phi_K s_K^0\mathcal{H}(p_K^0).$$

Thanks to the conservativity property of the numerical fluxes, we can integrate by parts. We thus obtain

$$\begin{aligned}
S_2 = -\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T\Big(&\rho_{KL}^{n+1}G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p)(p_L^{n+1} - p_K^{n+1}) + \\
&G_w(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p)(g(p_L^{n+1}) - g(p_K^{n+1}))\Big).
\end{aligned}$$

By the definition of the coefficient $\rho_{KL}^{n+1}$ given in (2.5.5), we find

$$\rho_{KL}^{n+1}\Big(p_L^{n+1} - p_K^{n+1}\Big) = -\Big(g(p_L^{n+1}) - g(p_K^{n+1})\Big). \tag{2.6.10}$$

Consequently

$$S_2 = \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \rho_{KL}^{n+1} \Big( G_w(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) - G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) \Big) (p_L^{n+1} - p_K^{n+1}).$$

Bearing in mind the nonnegativity of $\Lambda_{KL}^T$, we deduce from (2.5.10) that

$$m_0\rho_0 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (p_L^{n+1} - p_K^{n+1})^2 \le S_2.$$

In light of Lemma 2.6.4 and the coercivity of $\Lambda$ we claim

$$\rho_0 m_0 \underline{\Lambda} \sum_{n=0}^{N-1} \delta t \left\| p_h^{n+1} \right\|_{X_h}^2 \le S_2.$$

Next, integrating $S_3$ by parts and a using repeatedly (2.6.10) leads to

$$\begin{aligned} S_3 &= -\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (\xi_L^{n+1} - \xi_K^{n+1}) \Big( \rho_{KL}^{n+1} (p_K^{n+1} - p_L^{n+1}) + (g(p_K^{n+1}) - g(p_L^{n+1})) \Big) \\ &= 0. \end{aligned}$$

Now, the sub-linearity of the function $g$ i.e. $|g(p)| \le C_g |p|$, the discrete maximum principle, the Cauchy-Schwarz inequality and the Poincaré inequality imply

$$\begin{aligned} |S_4| &\le (\rho_1 + C_g) \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K| \Big( q_{P,K}^{n+1} + q_{I,K}^{n+1} \Big) p_K^{n+1} \\ &\le (\rho_1 + C_g) \sum_{n=0}^{N-1} \delta t \left\| q_{P,h}^{n+1} + q_{I,h}^{n+1} \right\|_{L^2(\Omega)} \left\| p_h^{n+1} \right\|_{L^2(\Omega)} \\ &\le C_1 \Big( \sum_{n=0}^{N-1} \delta t \left\| p_h^{n+1} \right\|_{X_h}^2 \Big)^{1/2}. \end{aligned}$$

Using the elementary inequality $ab \le \dfrac{a^2}{2} + \dfrac{b^2}{2}$, we get

$$|S_4| \le C_2 + \frac{\rho_0 m_0 \underline{\Lambda}}{2} \sum_{n=0}^{N-1} \delta t \left\| p_h^{n+1} \right\|_{X_h}^2.$$

The proof of the first inequality is concluded by taking

$$C_p = \frac{2}{\rho_0 m_0 \underline{\Lambda}} \Big( C_2 + \left\| \widetilde{\phi}_h \widetilde{s_h}(\cdot, 0) \mathcal{H}(\widetilde{p_h}(\cdot, 0)) \right\|_{L^1(\Omega)} \Big).$$

Let us now bound the discrete gradient of the Kirchoff function $\xi(s)$. So, multiplying the equation (2.5.15) by $\xi(s_K^{n+1})$ , summing up on all $K \in \mathcal{V}$ and $n = 0, \cdots, N-1$ gives

$$T_1 + T_2 + T_3 + T_4 = 0,$$

where

$$
\begin{aligned}
T_1 &= \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K| \phi_K (s_K^{n+1} - s_K^n) \xi(s_K^{n+1}),\\
T_2 &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T G_w(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) \xi(s_K^{n+1}),\\
T_3 &= -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T (\xi_L^{n+1} - \xi_K^{n+1}) \xi(s_K^{n+1}),\\
T_4 &= \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K| \Big( (s_K^{n+1} - 1) q_{P,K}^{n+1} + q_{I,K}^{n+1} \Big) \xi(s_K^{n+1}).
\end{aligned}
$$

The treatment of $T_1$ keeps the same spirit as that of $S_1$. Let $\mathcal{B}$ be a function such that $\mathcal{B}'(s) = \xi(s)$, for every $s \in [0,1]$. One can see in a straightforward way that

$$
\begin{aligned}
\mathcal{B}(b) - \mathcal{B}(a) &= \int_a^b \xi(s)\,\mathrm{d}s\\
&= \xi(b)(b-a) - \overbrace{\int_a^b \gamma(s)(s-a)\,\mathrm{d}s}^{\geq 0}\\
&\leq \xi(b)(b-a), \quad \forall a, b \in [0,1].
\end{aligned}
$$

Therefore

$$
\sum_{K\in\mathcal{V}} |\omega_K| \phi_K \Big( \mathcal{B}(s_K^N) - \mathcal{B}(s_K^0) \Big) \leq T_1.
$$

Utilizing once more the integration by parts property, Cauchy-Schwarz inequality and estimate (2.6.7) we obtain

$$
\begin{aligned}
|T_2| &\leq \|M_w\|_\infty \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \left| p_L^{n+1} - p_K^{n+1} \right| \left| \xi(s_L^{n+1}) - \xi(s_K^{n+1}) \right|\\
&\leq \|M_w\|_\infty \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \left| p_L^{n+1} - p_K^{n+1} \right| \left| \xi(s_L^{n+1}) - \xi(s_K^{n+1}) \right|\\
&\leq \sqrt{C_p}\, \|M_w\|_\infty \Big( \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \left| \xi(s_L^{n+1}) - \xi(s_K^{n+1}) \right|^2 \Big)^{1/2}\\
&\leq \frac{C_p \|M_w\|_\infty^2 \overline{\Lambda}}{2\underline{\Lambda}} + \frac{\underline{\Lambda}}{2} \sum_{n=0}^{N-1} \delta t \left\| \xi(s_h^{n+1}) \right\|_{X_h}^2.
\end{aligned}
$$

Integrate again by parts and use the coercivity of the tensor $\Lambda$ to infer

$$
\begin{aligned}
T_3 &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \Big( \xi(s_L^{n+1}) - \xi(s_K^{n+1}) \Big)^2\\
&\geq \underline{\Lambda} \sum_{n=0}^{N-1} \delta t \left\| \xi(s_h^{n+1}) \right\|_{X_h}^2.
\end{aligned}
$$

Since function $\xi$ is bounded then

$$\begin{aligned} |T_4| &\leq \|\xi\|_\infty \left\|q^P + q^I\right\|_{L^1(Q_\mathfrak{T})} \\ &\leq \sqrt{\mathfrak{T}\,|\Omega|}\,\|\xi\|_\infty \left\|q^P + q^I\right\|_{L^2(Q_\mathfrak{T})}. \end{aligned}$$

In conclusion, one gets

$$C_\xi = \frac{2}{\underline{\Lambda}}\Big(\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\mathcal{B}(s_K^0) + \frac{C_p\,\|M_w\|_\infty^2\,\overline{\Lambda}}{2\underline{\Lambda}} + \sqrt{\mathfrak{T}\,|\Omega|}\,\|\xi\|_\infty \left\|q^P + q^I\right\|_{L^2(Q_\mathfrak{T})}\Big).$$

Hence, the proof of Proposition 2.6.1 is concluded. □

**Remark 2.6.1.** *In the course of the above proof, we record once again that the nonnegativity of the transmissibilities is a key role to derive the a priori estimate on the global pressure. Now if some of them are negative, one can not control the gradients. This issue is addressed in the next chapter. We are obliged to consider the fractional flow formulation of the convective term to save these estimates.*

## 2.7 Existence result

Based on the uniform estimates of the previous section, we show now that the numerical scheme possesses a solution in the next lemma. This essentially relies on the following fundamental argument, that can be found in [65]. The latter result provides a sufficient condition so that a nonlinear specified vector field can admit a zero.

**Lemma 2.7.1.** *Let $\mathcal{A}$ be a finite dimensional space endowed with an inner product $(\cdot,\cdot)_\mathcal{A}$ and its associated norm $||\cdot||_\mathcal{A}$. Let $\mathcal{P}$ be a continuous mapping from $\mathcal{A}$ into itself satisfying*

$$(\mathcal{P}(x), x)_\mathcal{A} > 0, \quad \textit{for} \quad ||x||_\mathcal{A} = r > 0.$$

*Then there exists $x^* \in \mathcal{A}$ with $||x^*||_\mathcal{A} < r$ such that*

$$\mathcal{P}(x^*) = 0.$$

The following proposition states that the numerical scheme admits a solution at each time iteration.

**Proposition 2.7.1.** *(Existence)*
*For $n = 0, \ldots, N-1$, there exists a solution $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$ to the coupled scheme (2.5.12)-(2.5.15).*

*Proof.* To apply Lemma 2.7.1 we should specify the space $\mathcal{A}$, its inner product and the functional $\mathcal{P}$. For the sake of clarity, we prefer to adopt the following notations

$$\begin{aligned} q &:= \text{Card}\{K \in \mathcal{V}/x_K \notin \Gamma_D\}, \\ s &:= \{s_K^{n+1}\}_{K\in\mathbb{R}^q}, \\ p &:= \{p_K^{n+1}\}_{K\in\mathbb{R}^q}. \end{aligned}$$

Hence, we set $\mathcal{A} = \mathbb{R}^q \times \mathbb{R}^q$. It is equipped with its usual scalar product. The definition of the functional $\mathcal{P}$ is not evident and it amounts to construct some adequate functions. To this end, we first define the mapping $\Phi : \mathbb{R}^q \times \mathbb{R}^q \longrightarrow \mathbb{R}^q \times \mathbb{R}^q$, such that

$$\Phi(p,s) = \Big(\{\Phi_{1,K}\}_{K\in\mathcal{V}}, \{\Phi_{2,K}\}_{K\in\mathcal{V}}\Big),$$

where the first (resp. second) component corresponds to the gas (resp. water) equations of the numerical scheme as follows

$$\begin{aligned}
\Phi_{1,K} &= \phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big) \\
&+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p) \\
&- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T(\xi_L^{n+1} - \xi_K^{n+1}) + \delta t\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1}, \\
\Phi_{2,K} &= \phi_K\Big(s_K^{n+1} - s_K^n\Big) + \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T G_w(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p) \\
&- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T(\xi_L^{n+1} - \xi_K^{n+1}) + \delta t(s_K^{n+1} - 1)q_{P,K}^{n+1} + \delta t q_{I,K}^{n+1}.
\end{aligned}$$

Thanks to the assumptions on the data, one can see in a straightforward way that $\Phi$ is well-defined and continuous. To make use of the energy estimates proof, we need to define $\mathcal{F} : \mathbb{R}^q \times \mathbb{R}^q \longrightarrow \mathbb{R}^q \times \mathbb{R}^q$, such that

$$\mathcal{F}(p,s) = (p,v),$$

with $v = \{g(p_K^{n+1}) + \xi(s_K^{n+1})\}_{K\in\mathcal{V}}$. Notice that $\mathcal{F}$ exists and is continuous. As a consequence $\mathcal{F}$ is a homeomorphism. Indeed, the expression of $\mathcal{F}^{-1}$ is:

$$\mathcal{F}^{-1}(p,v) = \Big(u, \xi^{-1}(v - g(p))\Big).$$

Whence, one sees that $\mathcal{F}$ owns similar properties of $\xi$. It is now sufficient to consider the continuous mapping $\mathcal{P}$ defined as

$$\mathcal{P}(p,v) = \Phi \circ \mathcal{F}^{-1}(p,v) = \Phi(p,s).$$

The existence statement will be proved once we establish the inequality below

$$\Big(\mathcal{P}(p,v), (p,v)\Big)_{\mathbb{R}^{2q}} > 0, \quad \text{for} \quad ||(p,v)||_{\mathbb{R}^{2q}} = r, \tag{2.7.1}$$

for some sufficiently large $r$. Reproducing the proof of Proposition 2.6.1 we compute

$$\begin{aligned}
\Big(\mathcal{P}(p,v), (p,v)\Big)_{\mathbb{R}^{2q}} &\geq \frac{1}{\delta t}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(s_K^{n+1}\mathcal{H}(p_K^{n+1}) - s_K^n\mathcal{H}(p_K^n)\Big) \\
&+ \frac{1}{\delta t}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(\mathcal{B}(s_K^{n+1}) - \mathcal{B}(s_K^n)\Big) \\
&+ \frac{\rho_0 m_0 \underline{\Lambda}}{2}||p_h^{n+1}||_{X_h}^2 + \frac{\underline{\Lambda}}{2}||\xi(s_h^{n+1})||_{X_h}^2 - C_p' - C_\xi'.
\end{aligned}$$

For some positive constants $C'_p, C'_\xi$. Consequently

$$\Big(\mathcal{P}(u,v),(u,v)\Big)_{\mathbb{R}^{2q}} \geq -\frac{1}{\delta t}\sum_{K\in\mathcal{V}}|\omega_K|\phi_K\Big(s_K^n\mathcal{H}(p_K^n)+\mathcal{B}(s_K^n)\Big) + \min\Big(\frac{\rho_0 m_0\underline{\Lambda}}{2},\frac{\underline{\Lambda}}{2}\Big)\Big(||p_h^{n+1}||^2_{X_h}+||\xi(s_h^{n+1})||^2_{X_h}\Big)-C'_p-C'_\xi. \tag{2.7.2}$$

Additionally, the usual norm $\|\cdot\|_{\mathbb{R}^{2q}}$ is equivalent to the norm $\|\cdot\|_{\mathcal{V}}$ given by

$$\|u\|_{\mathcal{V}}=\sum_{K\in\mathcal{V}}\omega_K\,|u_K|^2\,.$$

Hence, there exists $C_{\mathcal{V}}>0$ such that $\|u\|_{\mathbb{R}^{2q}}\leq C_{\mathcal{V}}\,\|u\|_{\mathcal{V}}$. In view of Lemma 2.6.5, the Poincaré inequality and the Lipschitz continuity of the function $g$, there exists a positive constant $L$ which is independent of the discretization parameters ($h$ and $\delta t$ ) such that

$$\begin{aligned}\|(p,v)\|^2_{\mathbb{R}^{2q}} &= \Big\|\Big(\{p_K^{n+1}\}_{K\in\mathcal{V}},\{g(p_K^{n+1})+\xi(s_K^{n+1})\}_{K\in\mathcal{V}}\Big)\Big\|^2_{\mathbb{R}^{2q}}\\ &\leq C_{\mathcal{V}}\Big\|\Big(\{p_K^{n+1}\}_{K\in\mathcal{V}},\{g(p_K^{n+1})+\xi(s_K^{n+1})\}_{K\in\mathcal{V}}\Big)\Big\|^2_{\mathcal{V}}\\ &\leq L\Big(\,\|\xi(s_h^{n+1})\|^2_{X_h}+\|p_h^{n+1}\|^2_{X_h}\,\Big).\end{aligned} \tag{2.7.3}$$

Therefore, the last inequality ensures that (2.7.1) is fulfilled for a large enough $r=\|\xi(s_h^{n+1})\|^2_{X_h}+\|p_h^{n+1}\|^2_{X_h}$. The proof is concluded. □

## 2.8 Space and time translates

In this section we aim to establish compactness properties consisting of space and time translates estimations on the mass of gas sequence $\mathbb{U}_{h,\delta t}=\tilde{\phi}_h\rho(p_{h,\delta t})s_{h,\delta t}$ and on the mass of water sequence $\mathbb{V}_{h,\delta t}=\tilde{\phi}_h s_{h,\delta t}$. To do that, we require the following claim. This result affirms that the difference between the finite volume and the finite element reconstructions tend to zero as the size of the mesh goes to zero.

**Lemma 2.8.1.** *Let us denote* $\tilde{\mathbb{U}}_{h,\delta t}=\tilde{\phi}_h\rho(\tilde{p}_{h,\delta t})\tilde{s}_{h,\delta t}$. *Then*

$$\Big\|\mathbb{U}_{h,\delta t}-\tilde{\mathbb{U}}_{h,\delta t}\Big\|_{L^1(Q_{\mathfrak{T}})}\longrightarrow 0 \ \text{ as } \ h\longrightarrow 0.$$

*Proof.* The functions $\phi_h,\rho(p_{h,\delta t})$ and $s_{h,\delta t}$ are bounded. As a consequence

$$\begin{aligned}\Big\|\mathbb{U}_{h,\delta t}-\tilde{\mathbb{U}}_{h,\delta t}\Big\|_{L^1(Q_{\mathfrak{T}})} &= \int_{Q_{\mathfrak{T}}}|\mathbb{U}_{h,\delta t}-\tilde{\mathbb{U}}_{h,\delta t}|\,\mathrm{d}x\,\mathrm{d}t,\\ &\leq D_1+D_2,\end{aligned}$$

where $D_1$ and $D_2$ read

$$\begin{aligned}D_1&=\phi_1\rho_1\int_{Q_{\mathfrak{T}}}|s_{h,\delta t}-\tilde{s}_{h,\delta t}|\,\mathrm{d}x\,\mathrm{d}t,\\ D_2&=\phi_1\int_{Q_{\mathfrak{T}}}|\rho(p_{h,\delta t})-\rho(\tilde{p}_{h,\delta t})|\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

In light of the $\theta$-Hölder continuity of the function $\xi^{-1}$ we write

$$D_1 \leq L_\xi \int_{Q_\mathfrak{T}} |\xi(s_{h,\delta t}) - \xi(\tilde{s}_{h,\delta t})|^\theta \,\mathrm{d}x\,\mathrm{d}t.$$

Next, the application of Hölder's inequality with $\theta \in (0,1]$ leads to

$$D_1 \leq C_3 \Big( \int_{Q_\mathfrak{T}} |\xi(s_{h,\delta t}) - \xi(\tilde{s}_{h,\delta t})| \,\mathrm{d}x\,\mathrm{d}t \Big)^\theta =: C_3 (D_1')^\theta,$$

where

$$D_1' = \int_{Q_\mathfrak{T}} |\xi(s_{h,\delta t}) - \xi(\tilde{s}_{h,\delta t})| \,\mathrm{d}x\,\mathrm{d}t.$$

First, we observe that $\xi_K^{n+1} = \xi(s_{h,\delta t}(x_K, t))$. We develop the expression of $D_1'$ as follows

$$\begin{aligned}
D_1' &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{K\in\mathcal{V}_T} \int_{\omega_K\cap T} |\xi(s_{h,\delta t}) - \xi(\tilde{s}_{h,\delta t})| \,\mathrm{d}x \\
&= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{K\in\mathcal{V}_T} \int_{\omega_K\cap T} |\xi(s_{h,\delta t}(x,t)) - \xi(s_{h,\delta t}(x_K,t))| \,\mathrm{d}x \\
&= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{K\in\mathcal{V}_T} \int_{\omega_K\cap T} \big|\nabla\xi(s_{h,\delta t})_{|T} \cdot (x - x_K)\big| \,\mathrm{d}x \\
&\leq \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{K\in\mathcal{V}_T} \mathrm{diam(T)} |\omega_K \cap T| \,\big|\nabla\xi(s_{h,\delta t})_{|T}\big| \\
&\leq h \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} |T| \,\big|\nabla\xi(s_{h,\delta t})_{|T}\big| \\
&\leq (\mathfrak{T}\,|\Omega|)^{\frac{1}{2}} h \Big( \int_0^\mathfrak{T} \|\nabla\xi(s_{h,\delta t})\|^2_{L^2(\Omega)^2} \,\mathrm{d}t \Big)^{\frac{1}{2}} \\
&\leq C_4 h,
\end{aligned}$$

where we used the Cauchy-Schwarz inequality together with the a priori estimate (2.6.8). Thereby

$$D_1 \leq C_5 h^\theta \to 0 \text{ as } h \to 0.$$

We next recall that the derivative of the density $\rho'$ is bounded, then we estimate

$$D_2 \leq \big\|\rho'\big\|_\infty \int_{Q_\mathfrak{T}} |p_{h,\delta t} - \tilde{p}_{h,\delta t}| \,\mathrm{d}x\,\mathrm{d}t.$$

Similarly one estimates $D_2$

$$D_2 \leq C_6 h \to 0 \text{ as } h \to 0.$$

We finally deduce that the difference between $\mathbb{U}_{h,\delta t}$ and $\tilde{\mathbb{U}}_{h,\delta t}$ converges to zero in $L^1(Q_\mathfrak{T})$ as $h$ tends to zero. This completes the proof. □

**Lemma 2.8.2.** *(Space Translates) Let $(p_{h,\delta t}, s_{h,\delta t})$ be a solution to the system (2.5.12)-(2.5.15). Then the following inequality holds*

$$\int_0^{\mathfrak{T}} \int_{\Omega'} \left| \tilde{\mathbb{U}}_{h,\delta t}(x+y,t) - \tilde{\mathbb{U}}_{h,\delta t}(x,t) \right| \mathrm{d}x\, \mathrm{d}t \leq \beta(|y|), \tag{2.8.1}$$

*for every $y \in \mathbb{R}^d$,where $\Omega' = \{x \in \Omega, \quad [x, x+y] \subset \Omega\}$ and $\beta(|y|) \longrightarrow 0$ as $|y|$ goes to zero.*

*Proof.* By the definition of $\mathbb{U}_{h,\delta t}$, one has

$$\begin{aligned}
&\int_{Q'_{\mathfrak{T}}} \left| \tilde{\mathbb{U}}_{h,\delta t}(x+y,t) - \tilde{\mathbb{U}}_{h,\delta t}(x,t) \right| \mathrm{d}x\, \mathrm{d}t, \\
&= \int_{Q'_{\mathfrak{T}}} \left| \Big( \tilde{\phi_h} \rho(\tilde{p}_{h,\delta t}) \tilde{s}_{h,\delta t} \Big)(x+y,t) - \Big( \tilde{\phi_h} \rho(\tilde{p}_{h,\delta t}) \tilde{s}_{h,\delta t} \Big)(x,t) \right| \mathrm{d}x\, \mathrm{d}t, \\
&\leq W_1 + W_2 + W_3,
\end{aligned}$$

where $W_1, W_2$ and $W_3$ are given by

$$W_1 = \phi_1 \rho_1 \int_{Q'_{\mathfrak{T}}} |\tilde{s}_{h,\delta t}(x+y,t) - \tilde{s}_{h,\delta t}(x,t)| \, \mathrm{d}x\, \mathrm{d}t, \tag{2.8.2}$$

$$W_2 = \phi_1 \int_{Q'_{\mathfrak{T}}} |\rho(\tilde{p}_{h,\delta t}(x+y,t)) - \rho(\tilde{p}_{h,\delta t}(x,t))| \, \mathrm{d}x\, \mathrm{d}t. \tag{2.8.3}$$

$$W_3 = \rho_1 \int_{Q'_{\mathfrak{T}}} \left| \tilde{\phi_h}(x+y) - \tilde{\phi_h}(x) \right| \mathrm{d}x\, \mathrm{d}t. \tag{2.8.4}$$

In order to overestimate $W_1$, we introduce once more the $\theta$-Hölder continuity of $\xi^{-1}$. So, one has

$$W_1 \leq C_7 \int_{Q'_{\mathfrak{T}}} |\xi(\tilde{s}_{h,\delta t}(x+y,t)) - \xi(\tilde{s}_{h,\delta t}(x,t))|^{\theta} \, \mathrm{d}x\, \mathrm{d}t.$$

The Hölder inequality allows us to write

$$W_1 \leq C_8 \Big( \int_{Q'_{\mathfrak{T}}} |\xi(\tilde{s}_{h,\delta t}(x+y,t)) - \xi(\tilde{s}_{h,\delta t}(x,t))| \, \mathrm{d}x\, \mathrm{d}t \Big)^{\theta}.$$

As in the same spirit of [69], we define the function $\chi_{\sigma^T_{KL}}(x)$ for each $\sigma^T_{KL}$ by

$$\chi_{\sigma^K_{MS}}(x) = \begin{cases} 1, & \text{if the line segment } [x, x+y] \text{ intersects with } \sigma^T_{KL}, \\ 0, & \text{else.} \end{cases}$$

for $y \in \mathbb{R}, x \in \Omega'$ and $K, L \in \mathcal{V}_{\mathcal{T}}$. It is known that $\int_{\Omega'} \chi_{\sigma_{MS}^K}(x)\,\mathrm{d}x \leq C_\sigma |\sigma_{KL}^T||y|$. Thereby

$$
\begin{aligned}
W_1 &\leq C_9\Big(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\left|\xi(s_L^{n+1})-\xi(s_K^{n+1})\right|\int_{\Omega'}\chi_{\sigma_{MS}^K}(x)\,\mathrm{d}x\Big)^\theta,\\
&\leq C_{10}|y|^\theta\Big(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\left|\sigma_{KL}^T\right|\left|\xi(s_L^{n+1})-\xi(s_K^{n+1})\right|\Big)^\theta,\\
&\leq C_{11}|y|^\theta\Big(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}|T|^{\frac{1}{2}}\left|\xi(s_L^{n+1})-\xi(s_K^{n+1})\right|\Big)^\theta,\\
&\leq C_{12}|y|^\theta\Big(\sum_{n=0}^{N-1}\delta t \sum_{T\in\mathcal{T}}|T|\left|\nabla\xi(s_{h,\delta t})_{|T}\right|^2\Big)^{\theta/2},\\
&\leq C_{13}|y|^\theta\Big(\int_0^{\mathfrak{T}}\|\nabla\xi(s_{h,\delta t})\|_{L^2(\Omega)^2}^2\,\mathrm{d}t\Big)^{\theta/2},\\
&\leq C_{14}|y|^\theta,
\end{aligned}
$$

where we have mainly used the regularity of the mesh, within the triangle $\omega_K \cap T$, and the Cauchy-Schwarz inequality. Analogous arguments are employed to prove

$$W_2 \leq C_{15}|y|. \tag{2.8.5}$$

It is easy to see from Assumption $(\mathrm{H}_1)$ on the porosity that the space translates are strongly convergent which leads to

$$W_3 \to 0 \text{ as } |y| \to 0.$$

Finally, this inequality together with the previous one establish the required property (2.8.1). □

The time translates on $\tilde{\mathbb{U}}_{h,\delta t}$ are claimed in the following lemma.

**Lemma 2.8.3.** *(Time translates)*
*Let $(p_{h,\delta t}, s_{h,\delta t})$ be a solution to the algebraic system (2.5.12)-(2.5.15). There exists a modulus of continuity $\omega$ that is independent of $h$ and $\delta t$ such that*

$$\int_{\Omega\times(0,\mathfrak{T}-\tau)}\left|\tilde{\mathbb{U}}_{h,\delta t}(x,t+\tau)-\tilde{\mathbb{U}}_{h,\delta t}(x,t)\right|^2\mathrm{d}x\,\mathrm{d}t \leq \omega(\tau), \tag{2.8.6}$$

*for all $\tau \in (0,\mathfrak{T})$. Further, $\omega(\tau) \longrightarrow 0$ as $\tau \longrightarrow 0$.*

*Proof.* The proof mimics similar ideas as in [69] and later in [23]. □

## 2.9 Convergence of the numerical scheme

The scope of this section is to establish the strong convergence of the saturation, the weak convergence of the global pressure and the weak convergence of the gradients. To this purpose, we have to concatenate all the aforementioned properties together with the famous Riesz-Frechet-Kolmogorov compactness criterion.

**Proposition 2.9.1.** *Under Assumptions $(H_1)$-$(H_7)$, let $(p_{h,\delta t}, s_{h,\delta t})$ be a sequence of solutions to the numerical scheme (2.5.12)-(2.5.15). As $(h, \delta t) \longrightarrow (0,0)$, the following convergences hold up to a subsequence*

$$\tilde{\mathbb{U}}_{h,\delta t} \text{ and } \mathbb{U}_{h,\delta t} \longrightarrow U \qquad \text{strongly in } L^r(Q_{\mathfrak{T}}), r \geq 1, \quad \text{and a.e in } Q_{\mathfrak{T}},, \tag{2.9.1}$$

$$\tilde{s}_{h,\delta t} \text{ and } s_{h,\delta t} \longrightarrow s \qquad \text{a.e. in } Q_{\mathfrak{T}}, \tag{2.9.2}$$

$$p_{h,\delta t} \rightharpoonup p \qquad \text{weakly in } L^2(Q_{\mathfrak{T}}), \tag{2.9.3}$$

$$\nabla p_{h,\delta t} \rightharpoonup \nabla p \qquad \text{weakly in } L^2(Q_{\mathfrak{T}})^d, \tag{2.9.4}$$

$$\nabla \xi(s_{h,\delta t}) \rightharpoonup \nabla \xi(s) \qquad \text{weakly in in } L^2(Q_{\mathfrak{T}})^d. \tag{2.9.5}$$

*Moreover, $\xi(s)$ and $p$ are in $L^2(0, \mathfrak{T}; H^1_{\Gamma_D}(\Omega))$ with*

$$0 \leq s \leq 1, \quad U = \phi\rho(p)s \text{ a.e. in } Q_{\mathfrak{T}}. \tag{2.9.6}$$

*Finally, for all functions $\Gamma$ and $\kappa \in \mathcal{C}^0_b(\mathbb{R})$, with $\kappa(0) = 0$, we have*

$$\Gamma(p_{h,\delta t})\kappa(s_{h,\delta t}) \longrightarrow \Gamma(p)\kappa(s) \text{ a.e. in } Q_{\mathfrak{T}} \tag{2.9.7}$$

*Proof.* It follows from the space and the time translates lemmas that the sequence $\tilde{\mathbb{U}}_{h,\delta t}$ is relatively compact in $L^1(Q_{\mathfrak{T}})$ thanks to Kolmogorov's compactness theorem [30, 69]. This yields the strong convergence of an unlabeled subsequence of $\tilde{\mathbb{U}}_{h,\delta t}$ :

$$\tilde{\mathbb{U}}_{h,\delta t} \longrightarrow U \text{ in } L^1(Q_{\mathfrak{T}}) \text{ and a.e. in } \quad Q_{\mathfrak{T}},$$

and in virtue of Lemma 2.8.1, this subsequence $\mathbb{U}_{h,\delta t}$ converges to the same limit $U$. Also, it is bounded and consequently the strong convergence occurs in $L^r(Q_{\mathfrak{T}})$, with $r \geq 1$, which establishes (2.9.1).

Same steps are followed, as for $\tilde{\mathbb{U}}_{h,\delta t}$, to check the space and the time translates on the function $\tilde{\phi_h}\tilde{s}_{h,\delta t}$. We apply once again Kolmogorov's theorem to ensure the convergence almost everywhere of a subsequence, still denoted, $(\tilde{\phi_h}\tilde{s}_{h,\delta t})$. Hence

$$\tilde{\phi_h}\tilde{s}_{h,\delta t} \longrightarrow \phi s \quad \text{a.e. in} \quad Q_{\mathfrak{T}}, \tag{2.9.8}$$

and consequently,

$$s_{h,\delta t}, \quad \tilde{s}_{h,\delta t} \quad \longrightarrow s \quad \text{a.e. in} \quad Q_{\mathfrak{T}}. \tag{2.9.9}$$

Form Proposition 2.6.1, the sequence $(\nabla p_{h,\delta t})$ is bounded in $L^2(Q_{\mathfrak{T}})^d$. Moreover, the Poincaré inequality shows that the sequence $(p_{h,\delta t})$ is also bounded in $L^2(Q_{\mathfrak{T}})$. Hence there exists a function $p \in L^2(0, \mathfrak{T}; H^1_{\Gamma_D}(\Omega))$ such that the following convergences hold up to a subsequence

$$p_{h,\delta t} \rightharpoonup p \qquad \text{weakly in } L^2(Q_{\mathfrak{T}}), \tag{2.9.10}$$

$$\nabla p_{h,\delta t} \rightharpoonup \nabla p \qquad \text{weakly in } \Big(L^2(Q_{\mathfrak{T}})\Big)^d. \tag{2.9.11}$$

Similarly, the estimate (2.6.8) confirms that $(\xi(s_{h,\delta t}))$ is bounded in $L^2(Q_{\mathfrak{T}})$. Thus there exist two functions $\xi^* \in L^2(Q_{\mathfrak{T}})$ and $\zeta \in L^2(Q_{\mathfrak{T}})^d$ such that

$$\xi(s_{h,\delta t}) \rightharpoonup \xi^* \qquad \text{weakly in } L^2(Q_{\mathfrak{T}}), \tag{2.9.12}$$

$$\nabla \xi(s_{h,\delta t}) \rightharpoonup \zeta \qquad \text{weakly in } \Big(L^2(Q_{\mathfrak{T}})\Big)^d. \tag{2.9.13}$$

In view of (2.9.9) we can pasages to the limit thanks to the continuity of $\xi$ :

$$\xi(s_{h,\delta t}) \longrightarrow \xi(s) \quad \text{a.e. in} \quad Q_{\mathfrak{T}}. \tag{2.9.14}$$

The uniqueness of the limit implies

$$\xi^* = \xi(s) \text{ a.e. in } Q_{\mathfrak{T}}.$$

One deduces now that

$$\zeta = \nabla \xi(s),$$

and in the meantime one shows that $\xi(s) \in L^2(0, \mathfrak{T}; H^1_{\Gamma_D}(\Omega))$. We next introduce the fact that $\rho$ is (strictly) increasing to see that

$$\int_{Q_{\mathfrak{T}}} \Big(\tilde{\phi}_h \rho(p_{h,\delta t}) s_{h,\delta t} - \tilde{\phi}_h \rho(\varphi) s_{h,\delta t}\Big)\Big(p_{h,\delta t} - \varphi\Big)\,\mathrm{d}x\,\mathrm{d}t \geq 0, \quad \forall \quad \varphi \in L^2(Q_{\mathfrak{T}}).$$

The convergences (2.9.1) and (2.9.8) allow us to conclude that

$$\int_{Q_{\mathfrak{T}}} \Big(U - \phi\rho(\varphi)s\Big)(p - \varphi)\,\mathrm{d}x\,\mathrm{d}t \geq 0, \quad \forall \quad \varphi \in L^2(Q_{\mathfrak{T}}).$$

We now take $\varphi = p + \varepsilon w$ where $\varepsilon \in ]0, 1]$ and $w \in L^2(Q_{\mathfrak{T}})$. As a consequence

$$\int_{Q_{\mathfrak{T}}} \Big(U - \phi\rho(p + \varepsilon w)s\Big)(\varepsilon w)\,\mathrm{d}x\,\mathrm{d}t \geq 0, \quad \forall \quad \varepsilon \in ]0, 1], \quad \forall\, w \in L^2(Q_{\mathfrak{T}}).$$

Dividing each side by $\varepsilon$, letting $\varepsilon$ go to zero, substituting $w$ by $-w$ leads to

$$\int_{Q_{\mathfrak{T}}} \Big(U - \phi\rho(p)s\Big)w\,\mathrm{d}x\,\mathrm{d}t = 0, \quad \forall\, w \in L^2(Q_{\mathfrak{T}}).$$

In the absence of a strong convergence on the global pressure, we will use the strong convergence of the mass of the gas phase, especially to show (2.9.7). On one hand, if $s_{h,\delta t} \longrightarrow 0$ a.e., then $\Gamma(p_{h,\delta t})\kappa(s_{h,\delta t}) \longrightarrow 0 = \Gamma(p)\kappa(s)$ a.e. (since $\kappa(0) = 0$ and $\Gamma(p)$ is bounded). On the other hand, when $s_{h,\delta t} \longrightarrow s \neq 0$, in light of (2.9.1) we have $\Gamma(p_{h,\delta t}) \longrightarrow \Gamma(p)$ almost everywhere in $Q_{\mathfrak{T}}$. Then, $\Gamma(p_{h,\delta t})\kappa(s_{h,\delta t}) \longrightarrow \Gamma(p)\kappa(s)$ almost everywhere in $Q_{\mathfrak{T}}$, since the functions $\Gamma$, $\kappa$ are continuous. This establishes (2.9.7). □

We finally claim that any sequence of solutions converges towards a weak solution to the considered mathematical model.

**Theorem 2.9.1.** *(Passage to the limit) Under Assumptions $(H_1)$-$(H_7)$, the limit function $(p, s)$ of Proposition 2.9.1 is a weak solution to the problem (2.2.9)-(2.2.12) in the sense of Definition 2.2.1.*

*Proof.* We detail the proof of the first equation of the numerical scheme. Likewise, the proof of the second one is obtained. To this purpose, let $\psi \in \mathcal{C}_c^\infty(\Omega \times [0, \mathfrak{T}))$ and denote $\psi_K^{n+1} = \psi(x_K, t^{n+1})$ for all $K \in \mathcal{V}$ and $n \in \{0, \ldots, N\}$. Multiply the equation (2.5.14) by $\delta t \psi_K^{n+1}$ and sum up on $K \in \mathcal{V}$ and $n \in \{0, \ldots, N\}$ to find

$$\mathcal{G}_1^h + \mathcal{G}_2^h + \mathcal{G}_3^h + \mathcal{G}_4^h + \mathcal{G}_5^h = 0,$$

where

$$\mathcal{G}_1^h = \sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big)\psi_K^{n+1},$$

$$\mathcal{G}_2^h = -\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{V}}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\backslash\{K\}} \Lambda_{KL}^T\rho_{KL}^{n+1}(\xi_L^{n+1} - \xi_K^{n+1})\psi_K^{n+1},$$

$$\mathcal{G}_3^h = \sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{V}}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\backslash\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1}p)\psi_K^{n+1},$$

$$\mathcal{G}_4^h = \sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{V}} |\omega_K|\rho(p_K^{n+1})s_K^{n+1}q_K^{P,n+1}\psi_K^{n+1}.$$

To start off, we treat the convergence of the evolution term $\mathcal{G}_1^h$. Using the discrete integration by parts in time and bearing in mind that $\psi_K^N = \psi(x_K, \mathfrak{T}) = 0$, one gets

$$\begin{aligned}
\mathcal{G}_1^h &= -\sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^{n+1})s_K^{n+1}\Big(\psi_K^{n+1} - \psi_K^n\Big) - \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^0)s_K^0\psi_K^0 \\
&= -\sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}}\int_{t^n}^{t^{n+1}}\int_{\omega_K} \phi_K\rho(p_K^{n+1})s_K^{n+1}\partial_t\psi(x_K,t)\,\mathrm{d}x\,\mathrm{d}t - \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^0)s_K^0\psi_K^0 \\
&= -\int_{Q_{\mathfrak{T}}} \widetilde{\mathbb{U}}_{h,\delta t}\partial_t\widetilde{\psi}_h\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\rho(p^0)s^0\widetilde{\psi}_h(x,0)\,\mathrm{d}x.
\end{aligned}$$

with $\widetilde{\psi}_h(x,t) = \psi(x_K,t)$ for all $x\in K$, $K\in\mathcal{V}$ and $t\in[0,\mathfrak{T})$. Due to the smoothness of the test function, the sequence $\{\widetilde{\psi}_h\}$ (resp. $\{\partial_t\widetilde{\psi}_h\}$) is uniformly convergent towards $\psi$ (resp. $\partial_t\psi$). We recall that $\{\widetilde{\mathbb{U}}_{h,\delta t}\}$ converges strongly to $\phi\rho(p)s$. Owing to Lebesgue's Dominated Convergence Theorem (LDCT) we deduce

$$\lim_{h,\delta t\to 0}\mathcal{G}_1^h = -\int_{Q_{\mathfrak{T}}}\phi\rho(p)s\partial_t\psi\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega\phi\rho(p^0)s^0\psi(x,0)\,\mathrm{d}x.$$

Next, let us study the convergence of the discrete elliptic operator. In other words, let us establish

$$\lim_{h\to 0}\mathcal{G}_2^h = \int_{Q_{\mathfrak{T}}}\rho(p)\Lambda\nabla\xi(s)\cdot\nabla\ \psi\,\mathrm{d}x\,\mathrm{d}t. \tag{2.9.15}$$

The presence of the density in the diffusion term makes it hard to pass to the limit in the latter since we have only a weak convergence on the global pressure. To tackle this issue, we need to introduce the strong convergence result on the mass of gas (2.9.7) as we are going to show below. We firstly reorder the expression $\mathcal{G}_2^h$ by triangles and dual edges to infer

$$\begin{aligned}
\mathcal{G}_2^h &= -\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{V}}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\backslash\{K\}} \Lambda_{KL}^T\rho_{KL}^{n+1}\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)\psi_K^{n+1} \\
&= \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T\rho_{KL}^{n+1}\Big(\xi_L^{n+1} - \xi_K^{n+1}\Big)\Big(\psi_L^{n+1} - \psi_K^{n+1}\Big).
\end{aligned}$$

Now we see that the coefficient $\rho_{KL}^{n+1}$ does not allow us to the above expression under an integral form. As pointed out in [23], it is wiser to approach $\mathcal{G}_2^h$ with another term whose limit is (2.9.15). Such a proposition consists of taking

$$\mathcal{G}_2^{h,*} = \int_{Q_\mathfrak{T}} \rho(p_{h,\delta t})\Lambda\nabla\xi(s_{h,\delta t}) \cdot \nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t.$$

This integral can be rewritten as follows

$$\mathcal{G}_2^{h,*} = \int_{Q_\mathfrak{T}} \Lambda\nabla(\rho(p_{h,\delta t})\xi(s_{h,\delta t})) \cdot \nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\mathfrak{T}} \xi(s_{h,\delta t})\Lambda\nabla\rho(p_{h,\delta t}) \cdot \nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t.$$

We know that $\{\rho(p_{h,\delta t})\}$ and $\{\xi(s_{h,\delta t})\}$ are two bounded sequences and their gradients are so. In addition, $\{\rho(p_{h,\delta t})\xi(s_{h,\delta t})\}$ converges strongly to $\rho(p)\xi(s)$ thanks to (2.9.7). Moreover, similar arguments of the proof of Proposition 2.9.1 are utilized to get

$$\nabla(\rho(p_{h,\delta t})\xi(s_{h,\delta t})) \rightharpoonup \nabla(\rho(p)\xi(s)), \text{ weakly in } L^2(Q_\mathfrak{T})^d.$$

Furthermore, there exists $\rho^* \in L^2(Q_\mathfrak{T})$ such that

$$\rho(p_{h,\delta t}) \rightharpoonup \rho^\star, \text{ weakly in } L^2(Q_\mathfrak{T}),$$

and

$$\nabla\rho(p_{h,\delta t}) \rightharpoonup \nabla\rho^\star, \text{ weakly in } L^2(Q_\mathfrak{T})^d.$$

According to the strong convergence in $L^2(Q_\mathfrak{T})^d$ of the sequences $\{\nabla\psi_{h,\delta t}\}$, $\{\xi(s_{h,\delta t})\nabla\psi_{h,\delta t}\}$ we claim

$$I := \lim_{h,\delta t\to 0} \mathcal{G}_2^{h,*} = \int_{Q_\mathfrak{T}} \Lambda\nabla(\rho(p)\xi(s)) \cdot \nabla\psi\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_\mathfrak{T}} \xi(s)\Lambda\nabla\rho^\star \cdot \nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Extending the first integral in $I$ gives

$$\mathcal{G}_2 = \int_{Q_\mathfrak{T}} \rho(p)\Lambda\nabla\xi(s) \cdot \nabla\psi\,\mathrm{d}x\,\mathrm{d}t + \int_{Q_\mathfrak{T}} \xi(s)(\nabla\rho(p) - \nabla\rho^\star) \cdot \Lambda\nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Applying the integration by parts to the second integral in $I$ leads to

$$\begin{aligned}\int_{Q_\mathfrak{T}} \xi(s)(\nabla\rho(p) - \nabla\rho^\star) \cdot \Lambda\nabla\psi\,\mathrm{d}x\,\mathrm{d}t = &- \int_{Q_\mathfrak{T}} (\rho(p) - \rho^\star)\gamma(s)\nabla s \cdot \Lambda\nabla\psi\,\mathrm{d}x\,\mathrm{d}t \\ &- \int_{Q_\mathfrak{T}} (\rho(p) - \rho^\star)\xi(s)\,\mathrm{div}(\Lambda\nabla\psi)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Now one can check that $\rho(p)\gamma(s) = \rho^\star\gamma(s)$ and $\rho(p)\xi(s) = \rho^\star\xi(s)$ almost everywhere in $Q_\mathfrak{T}$ with the aid of (2.9.7). Therefore, the last two integrals of the preceding equality vanish. Finally, this shows that

$$\lim_{h,\delta t\to 0} \mathcal{G}_2^{h,*} = \int_{Q_\mathfrak{T}} \rho(p)\Lambda\nabla\xi(s) \cdot \nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

What is left is to prove that

$$\lim_{h,\delta t\to 0} |\mathcal{G}_2^h - \mathcal{G}_2^{h,*}| = 0. \tag{2.9.16}$$

To this end, we need to define the following piecewise functions $\overline{u}_{h,\delta t}, \underline{u}_{h,\delta t}$ with $u \in \{p, s\}$.

$$\begin{aligned}&\overline{u}_T^{n+1} := \sup_{x\in T} u_h^{n+1}(x), \qquad \underline{u}_T^{n+1} := \inf_{x\in T} u_h^{n+1}(x),\\ &\overline{u}_{h,\delta t_{|T\times(t^n,t^{n+1}]}} := \overline{u}_T^{n+1}, \quad \underline{u}_{h,\delta t|T\times(t^n,t^{n+1}]} := \underline{u}_T^{n+1}.\end{aligned}$$

Now, setting

$$\mathcal{D}_2^h = \int_{Q_\mathfrak{T}} \rho(\underline{p}_{h,\delta t}) \Lambda \nabla \xi(s_{h,\delta t}) \cdot \nabla \psi_{h,\delta t} \, \mathrm{d}x \, \mathrm{d}t,$$

to deduce

$$\begin{aligned}\left|\mathcal{G}_2^h - \mathcal{G}_2^{h,*}\right| &\leq \left|\mathcal{G}_2^h - \mathcal{D}_2^h\right| + \left|\mathcal{D}_2^h - \mathcal{G}_2^{h,*}\right| \\ &\leq 2 \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \left|\rho(\overline{p}_T^{n+1}) - \rho(\underline{p}_T^{n+1})\right| \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \Lambda_{KL}^T \left|\xi_L^{n+1} - \xi_K^{n+1}\right| \left|\psi_L^{n+1} - \psi_K^{n+1}\right|.\end{aligned}$$

In virtue of the Cauchy-Schwarz inequality and Lemma 2.6.4 we discover

$$\begin{aligned}\left|\mathcal{G}_2^h - \mathcal{G}_2^{h,*}\right| &\leq C_{16} \left\|\nabla \psi\right\|_\infty \left\|\nabla \xi(s_{h,\delta t})\right\|_{L^2(Q_\mathfrak{T})^d} \Big( \sum_{n=0}^{N-1} \delta t \int_\Omega \left|\overline{p}_h^{n+1} - \underline{p}_h^{n+1}\right|^2 \mathrm{d}x \Big)^{1/2} \\ &\leq C_{17} h \quad \longrightarrow 0, \text{ as } h, \delta t \longrightarrow 0.\end{aligned}$$

Let us move on to deal with the convective term $\mathcal{G}_3^h$. We thus rewrite $\mathcal{G}_3^h$ by edges

$$\mathcal{G}_3^h = -\sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \rho_{KL}^{n+1} \Lambda_{KL}^T G_g(s_K^{n+1}, s_L^{n+1}; \delta_{KL}^{n+1} p) \Big( \psi_L^{n+1} - \psi_K^{n+1} \Big).$$

We additionally define

$$\mathcal{D}_3^h = \int_{Q_\mathfrak{T}} \rho(p_{h,\delta t}) M_g(s_{h,\delta t}) \Lambda \nabla p_{h,\delta t} \cdot \nabla \psi_{h,\delta t} \, \mathrm{d}x \, \mathrm{d}t.$$

Using repeatedly (2.9.7), the sequence $\{\rho(p_{h,\delta t}) M_g(s_{h,\delta t}) \nabla \psi_{h,\delta t}\}$ converges strongly to $\{\rho(p) M_g(s) \nabla \psi\}$ in $L^2(Q_\mathfrak{T})^d$. Since $\{\nabla p_{h,\delta t}\}$ converges weakly to $\nabla p$ in $L^2(Q_\mathfrak{T})^d$ then

$$\lim_{h, \delta t \to 0} \mathcal{D}_3^h = \int_{Q_\mathfrak{T}} \rho(p) M_g(s) \Lambda \nabla p \cdot \nabla \psi \, \mathrm{d}x \, \mathrm{d}t.$$

Define now

$$\mathcal{D}_3^{h,1} = \int_{Q_\mathfrak{T}} \rho(\underline{p}_{h,\delta t}) M_g(s_{h,\delta t}) \Lambda \nabla p_{h,\delta t} \cdot \nabla \psi_{h,\delta t} \, \mathrm{d}x \, \mathrm{d}t.$$

We show that $\left|\mathcal{D}_3^h - \mathcal{D}_3^{h,1}\right| \to 0$ as $h, \delta t \longrightarrow 0$. To do that, let us seek an upper bound of this quantity

$$\begin{aligned}\left|\mathcal{D}_3^h - \mathcal{D}_3^{h,1}\right| &\leq C_{18} \left\|\nabla \psi\right\|_\infty \left\|\nabla p_{h,\delta t}\right\|_{L^2(Q_\mathfrak{T})^d} \Big( \sum_{n=0}^{N-1} \delta t \int_\Omega \left|\overline{p}_h^{n+1} - \underline{p}_h^{n+1}\right|^2 \Big)^{1/2} \\ &\leq C_{19} h \quad \longrightarrow 0, \text{ as } h, \delta t \longrightarrow 0.\end{aligned}$$

Let us finally define $\mathcal{G}_3^{h,*}$

$$\mathcal{G}_3^{h,*} = \int_{Q_\mathfrak{T}} \rho(\underline{p}_{h,\delta t}) M_g(\underline{s}_{h,\delta t}) \Lambda \nabla p_{h,\delta t} \cdot \nabla \psi_{h,\delta t} \, \mathrm{d}x \, \mathrm{d}t.$$

Moreover, we prove that

$$\left|\mathcal{D}_3^{h,1}-\mathcal{G}_3^{h,*}\right|\to 0,\ \text{ as } h,\delta t\longrightarrow 0. \tag{2.9.17}$$

Indeed, the Cauchy-Schwarz inequality, the $\theta$-Hölder continuity of the function $\xi^{-1}$, the Hölder inequality with $\theta\in(0,1]$, Proposition 2.6.1 and Lemma 2.6.4 entail

$$\begin{aligned}\left|\mathcal{D}_3^{h,1}-\mathcal{G}_3^{h,*}\right| &\leq C_{20}\Big(\sum_{n=0}^{N-1}\delta t\int_\Omega\left|\overline{s}_{h,\delta t}-\underline{s}_{h,\delta t}\right|^2\Big)^{1/2}\\ &\leq C_{21}\Big(\sum_{n=0}^{N-1}\delta t\int_\Omega\left|\xi(\overline{s}_{h,\delta t})-\xi(\underline{s}_{h,\delta t})\right|^2\Big)^{\theta/2},\quad\longrightarrow 0,\ \text{ as } h,\delta t\longrightarrow 0.\end{aligned}$$

From the aforementioned expressions, we should now check that the sequence $\{\mathcal{G}_3^h-\mathcal{G}_3^{h,*}\}$ tends to zeros as $h,\delta t$ go to zero. In view of the consistency property and the item $(C_3)$ of the numerical flux $G_g$, we infer

$$\begin{aligned}&\left|\rho_{KL}^{n+1}G_g(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)-\Big(-\rho_{K,0}^{n+1}M_g(s_{K,0}^{n+1})\delta_{KL}^{n+1}p\Big)\right|\left|\delta_{KL}^{n+1}\psi\right|\\ &=\left|\rho_{KL}^{n+1}G_g(s_K^{n+1},s_L^{n+1};\delta_{KL}^{n+1}p)-\rho_{K,0}^{n+1}G_g\Big(s_{K,0}^{n+1},s_{K,0}^{n+1};\delta_{KL}^{n+1}p\Big)\right|\left|\delta_{KL}^{n+1}\psi\right|\\ &\leq C_{22}\Big(\eta(\left|s_K^{n+1}-s_{K,0}^{n+1}\right|)+\left|\rho_{KL}^{n+1}-\rho_{K,0}^{n+1}\right|\Big)\left|\delta_{KL}^{n+1}p\right|\left|\delta_{KL}^{n+1}\psi\right|\\ &\leq C_{23}\Big(\eta(\left|s_K^{n+1}-s_{K,0}^{n+1}\right|)+\left|\rho_{KL}^{n+1}-\rho_{K,0}^{n+1}\right|\Big)\Big(\left|\delta_{KL}^{n+1}p\right|^2+\left|\delta_{KL}^{n+1}\psi\right|^2\Big),\end{aligned}$$

where $\eta(\cdot)$ is the modulus of continuity of the numerical flux $G_g$ defined in (2.5.9). Consequently

$$\begin{aligned}\left|\mathcal{G}_3^h-\mathcal{G}_3^{h,*}\right| &\leq C_{22}\sum_{T\in\mathcal{T}}\Big(\eta(\left|\overline{s}_T^{n+1}-\underline{s}_T^{n+1}\right|)+\left|\overline{p}_T^{n+1}-\underline{p}_T^{n+1}\right|\Big)\\ &\qquad\times\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Big(\Lambda_{KL}^T\left|\delta_{KL}^{n+1}p\right|^2+\Lambda_{KL}^T\left|\delta_{KL}^{n+1}\psi\right|^2\Big)\\ &\leq C_{24}\int_{Q_{\mathfrak{T}}}\Big(\eta(\left|\overline{s}_{h,\delta t}-\underline{s}_{h,\delta t}\right|)+\left|\overline{p}_{h,\delta t}-\underline{p}_{h,\delta t}\right|\Big)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Proceeding similarly as in the proof of (2.9.17) we conclude

$$\lim_{h,\delta t\to 0}\left|\mathcal{G}_3^h-\mathcal{G}_3^{h,*}\right|=0. \tag{2.9.18}$$

Finally, one gets

$$\lim_{h,\delta t\to 0}\mathcal{G}_3^h=\int_{Q_{\mathfrak{T}}}\rho(p)M_g(s)\Lambda\nabla p\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

The last limit results from the almost everywhere convergence (2.9.7) and LDCT

$$\lim_{h,\delta t\to 0}\mathcal{G}_4^h=\lim_{h,\delta t\to 0}\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathcal{V}}|\omega_K|\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1}\psi_K^{n+1}=\int_{Q_{\mathfrak{T}}}\rho(p)sq^P\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Whence, the proof of Theorem 2.9.1 is concluded. □

## 2.10 Numerical experiments

In this section we give two numerical tests in the two dimensional space in order to illustrate the behavior and stability of the presented discretization. We point out that the implemented scheme (2.5.12)-(2.5.15) yields a nonlinear algebraic system. Appropriate linearization schemes [100, 111, 112] can be proposed for solving the resulting system namely the Picard iteration technique, the fixed point approach, the L-scheme strategy and the Newton method. For instance, the work [110] is proposing and analyzing a linear iterative scheme for solving a related problem, involving Hölder continuous nonlinearities. Using an improved Newton method, the authors suggested in [27] a variable switching technique in the case of the Richards equation to overcome the issues linked to the fully saturated or fully unsaturated regimes. In our case the underlined system is solved thanks to Newton-Raphson's method. Note that in our system the evolution terms are not in fact degenerate since the conservative variables $(s, \rho(p)s)$ are connected to the variables $(p, \xi(s))$ by a diffeomorphism (see Lemma 4.3 in [84] for more details) and consequently due to the monotonicity of the function $\xi$, the variables $(p, s)$ are uniquely defined. Therefore, it is not necessary to switch the variable with respect to the fully saturated regime. It is worth mentioning that at every time iteration indexed by $n$, the solver is initialized by $(p^n, s^n)$ where the stopping criterion is fixed to $10^{-10}$. It also includes the computation of a Jacobian matrix. In order to avoid the singularity of this matrix, the algorithm necessitates a slight restriction on the time step $\delta t < h$ even if the numerical scheme is unconditionally stable. In the both tests below, we observe that the Newton process requires a few iterations, between three and ten, to converge.

### 2.10.1 Test case 1

Being inspired by the benchmark test [106], this first academic example aims to evaluate the error between the computed solution and the analytical solution to the following model problem

$$\begin{cases} \partial_t u - \operatorname{div}\Big(f(u)\nabla p + \varepsilon \nabla u\Big) = F & \text{in} \quad (0,1)^2 \times (0, t_f) \\ \operatorname{div}\Big(M(u)\nabla p\Big) = 0 & \text{in} \quad (0,1)^2 \times (0, t_f) \end{cases}, \tag{2.10.1}$$

with $f(u) = u/(0.5-0.2u), \varepsilon = 0.01, M(u) = 1/(0.5-0.2u)$ and $t_f = 0.05$. For $F = 2\varepsilon\dfrac{\pi^2}{16}\sin(\dfrac{\pi}{4}(x+y+2t))$, the functions satisfying the above system read

$$u((x,y),t) = \sin(\frac{\pi}{4}(x+y+2t)), \quad p((x,y),t) = \frac{0.2}{\pi/4}\cos(\frac{\pi}{4}(x+y+2t)) + 0.5(x+y). \tag{2.10.2}$$

We supplement the equations of (2.10.1) by Dirichlet boundary conditions and initial conditions which correspond to (2.10.2). We consider a sequence of triangular meshes [88] where all the angles are acute (see Fig 2.3) to discretize the domain $\Omega$. We here indicate that the time step is divided by 4 since the mesh size is divided by 2.

I In Table 2.1 we display the errors in $L^2(Q_T)$ norm and the convergence rates computed on successively refined meshes for the saturation and the pressure at final time $t_f = 0.2$. Although it is accurate of second order in case of linear problems, we can observe that the VCFV method converges of first order with regard to the pressure and the saturation. This is classical and it is due mainly to the upstream technique used in the numerical scheme. Compared to [106] the mixed finite element–finite volume scheme converges also of first order towards the exact solution.

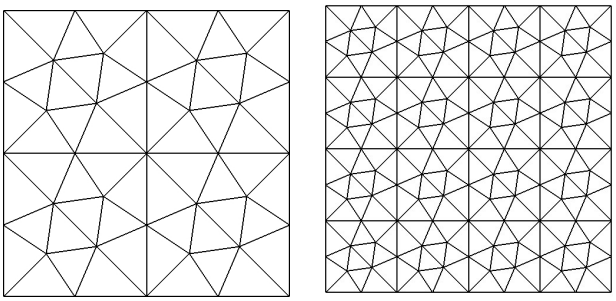


Figure 2.3: Example of meshes used in the test case 1 with $h = 0.250$, 0.125 from left to right.

| h | $\|\|u - u_{h,\delta t}\|\|_{L^2(Q_T)}$ | Rate | $\|\|p - p_{h,\delta t}\|\|_{L^2(Q_T)}$ | Rate | $\min u_{h,\delta t}$ |
|---|---|---|---|---|---|
| 0.250 | 0.530 E-02 | - | 0.182 E-03 | - | 0 |
| 0.125 | 0.321 E-02 | 0.723 | 0.771 E-04 | 1.239 | 0 |
| 0.063 | 0.173 E-02 | 0.903 | 0.345 E-04 | 1.173 | 0 |
| 0.031 | 0.877 E-03 | 0.957 | 0.164 E-04 | 1.053 | 0 |
| 0.016 | 0.437 E-03 | 1.055 | 0.807 E-05 | 1.075 | 0 |

Table 2.1: Error study of the scheme for the saturation and the pressure.

### 2.10.2 Test case 2

In the second example we consider the test case treated in [23]. The domain of computation is $\Omega = (0,1)^2$. The porosity is set to $\phi = 0.206$. The relative permeabilities and the capillary pressure are respectively defined by: $K_{rg}(s) = s^2, K_{rw}(s) = (1-s)^2, p_c(s) = P_{max}\, s$, where $P_{max} = 1.013 \times 10^5 Pa$. The viscosities of the two fluids read: $\mu_w = 10^{-3} Pa.s$, $\mu_g = 9 \times 10^{-5} Pa.s$. The gas density is an affine function: $\rho(p) = \rho_r(1 + c_r(p - p_r))$ with $\rho_r = 400 Kg.m^{-3}$, $c_r = 10^{-6} Pa$, $p_r = 1.013 \times 10^5 Pa$. The absolute permeability is given by $\Lambda = 0.15 \times 10^{-10}[m^2]$. The initial gas saturation and gas pressure are: $s_g(x,0) = 0.9$, $p_g(x,0) = 1.013 \times 10^5 Pa$. We point out that the initial global pressure is obtained by the relationship (2.2.5).

We inject water in the left zone ($x = 1$, $0.8 \leq y \leq 1$) of the medium with a saturation $s_w^l = 0.9$ and a constant pressure $P_g^l = 4.026 \times 10^5 Pa$. The right zone ($x = 1{,}0 \leq y \leq 0.2$) is left to be in contact with the air. Hence we impose $P_g^r = 1.013 \times 10^5 Pa$ and consider a free flow meaning that $\nabla \xi(s) \cdot \mathbf{n} = 0$. The remainder of the boundary is impermeable. We further take no source terms; that is $q^P = q^I = 0$. The final time is fixed to $t_f = 50s$. The figures below illustrate the motion of water saturation in the absence of capillary effects, which means that $\xi = 0$, Fig 2.5 and in the presence of them ($\xi \neq 0$) Fig 2.6 for different times $T = 4s, 20s, 50s$. In both cases we observe that the discrete saturation remains in the interval $[0.1, 1]$ as we have shown in Lemma 2.6.1. Moreover we observe a remarkable displacement of a sharp front between the two fluids in the first test, toward the right zone where the pressure is lower, while a smooth one is recorded in the second test which is due to the diffusive nature of the capillary term as expected. This tests shows the robustness and the ability of the proposed approach to capture the shocks in the pure hyperbolic case of the studied model.

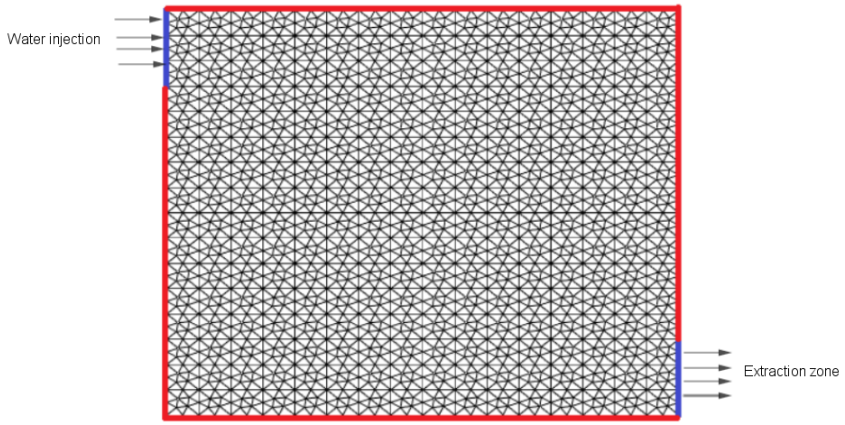



Figure 2.4: Primal mesh with 3584 triangles and 1857 vertices

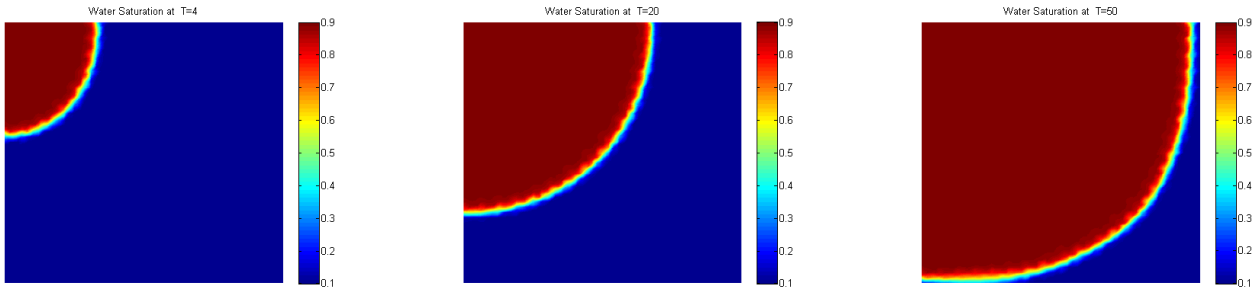



Figure 2.5: Evolution of water saturation with $P_{max} = 0Pa$ at $t = 4s, 20s, 50s$.

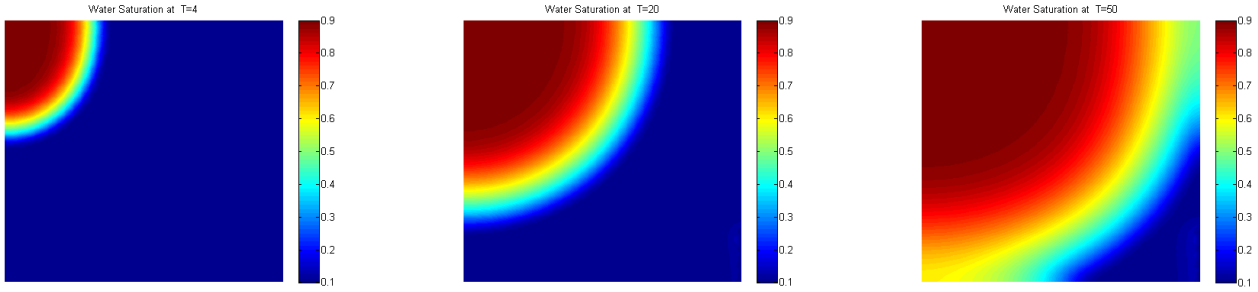



Figure 2.6: Evolution of water saturation with $P_{max} = 1.013 \times 10^5 Pa$ at $t = 4s, 20s, 50s$.

# Chapter 3

# Positive control volume finite element scheme for a degenerate compressible two-phase flow in anisotropic porous media

In this chapter we are concerned with the convergence analysis of a positive control volume finite element scheme (CVFE) for a degenerate compressible two-phase flow model in anisotropic porous media. For this, we consider the global pressure saturation formulation. We next use an implicit Euler scheme in time and a CVFE discretization in space. This approach rests on a particular choice of the mean value of the gas density on the interfaces, a centered scheme of the total mobility and the upwind approximation of fractional fluxes according to the total velocity. Thus, the maximum principle is fulfilled without any constraint on the stiffness coefficients. Moreover uniform estimates on the discrete gradient of the global pressure and the dissipative term are derived. As the mesh size is sent to zero, we establish that the sequence of approximate solutions converges to a weak solution of the continuous problem. Numerical tests are presented in two dimensions to exhibit the behavior of the gas pressure and the water saturation through the medium.

## 3.1 Introduction

We are interested in the two phase flow model in porous media. Its applicability is of a great prominence in engineering. More precisely, it occurs widely in oil recovery where, in general, the phases are a gas and liquid. It can be applied in hydrology and many other fields. The mathematical formulation of the two-phase flow model comprises a coupled nonlinear system of partial differential equations with degenerate coefficients. Then, seeking analytical solutions is usually avoided due to the complexity of the system. As a result, we resort to suitable numerical methods in order to approximate the solutions of interest. Such a method should preserve some properties, which are resumed in robustness and consistency.

Various contributions, with different hypotheses on the data, have been proposed for the discretization of the two-phase flow model. Beginning by finite difference approximation, we refer to the works [17, 108]. This stipulates high regularity on the data and structured domains, which excludes a large part of physical problems. So, finite volume methods have appeared and known

a huge interest in the last decades [61, 69] since they are robust and cheap in view of the computational cost. In addition, they are often used to discretize equations including high dominated convection terms [1, 12, 21, 72, 103, 114]. Concerning compressible flows, we refer to [4, 23, 113] for the convergence analysis of such finite volume schemes requiring both an isotropic permeability tensor and an orthogonality condition on the mesh. The scheme proposed in [113, 114] consists of a finite volume method on a specific mesh together with a phase-by-phase upstream scheme. The authors showed that the proposed scheme satisfies the maximum principle for the saturations, and obtained discrete energy estimates on the pressures under the assumption that the transmissibility coefficients are nonnegative. Practically, this condition is very restrictive. It is satisfied for a scalar permeability and for particular meshes. For instance in case of a triangulation, the angles of the triangles must be acute. Generally, to deal with the anisotropic case, some attempts have been investigated. In these studies, the feature of a finite element scheme and that of the finite volume method are combined. The first one provides a simple discretization of the diffusion counterpart while the second one preserves the locally conservative property of the numerical fluxes [73, 114]. More generally, the so-called gradient schemes method, which includes a large variety of discretizations, has been developed for the incompressible flows in [62, 68]. Nevertheless, this class of schemes fails the preserve the physical ranges of the approximate solution, which is an important property when it comes to deal with positive quantities such as the saturation and concentration.

The main point of this chapter focuses on the numerical analysis of a positive control volume finite element scheme for the approximation of a compressible two-phase flow model. This approach has been applied to a degenerate parabolic equation in [40] in which, the elliptic term is treated as a hyperbolic one so that they could prove the maximum principle and derive an a priori estimate on the discrete gradient. This methodology has been successfully extended to a system consisting of two parabolic equations in [42]. Being inspired by these works, we will propose a nonlinear scheme that will allow us to handle the issue due to the anisotropy of the medium. To get the desired discretization, an implicit Euler scheme in time and a CVFE discretization in space are considered. The convective fluxes are approximated with the aid of an upwind scheme, the total mobility is discretized with a centered scheme and the diffusive term is discretized using a Godunov-like scheme. For more details about the dating and the analysis of the CVFE method for several partial differential equations we refer the reader to this non-exhaustive list [19, 34, 35, 36, 67, 76].

The layout of this chapter is given as follows. In Section 3.2 we state the mathematical formulation of the compressible two-phase flow in porous media, which is derived from the generalized Darcy law and the mass conservation law. Section 3.3 is devoted to defining the primal mesh, the dual mesh and to describing the discrete solution space and the discrete trial space. Section 3.4 is devoted to sketching out the CVFE discretization and how we get the expected scheme. Next, we survey some useful properties in Section 3.5. Moreover, Section 3.6 is dedicated to establishing the maximum principle and a priori estimates on the discrete gradients. In Section 3.7, the existence of a discrete solution to the combined scheme is proved. In Section 3.8, the space and time translates estimations are established. Section 3.9 is concerned with the convergence of a discrete solution towards a weak solution to the continuous problem, which is the main result of the present chapter. Finally, in Section 3.10 some numerical tests are presented to display the flow of water and gas through the medium with different rates of anisotropy.

## 3.2 Presentation of the problem

The mathematical formulation of the compressible two-phase flow model is obtained by substituting the generalized Darcy law into the mass conservation equation for each phase. In addition, the considered phases are: gas, which is compressible and water, which is incompressible. We emphasize that the studied medium is anisotropic and heterogeneous.

To begin with, we consider a porous medium $\Omega$ as a bounded polygonal open of $\mathbb{R}^d$ ($d \geq 1$) and let $\mathfrak{T}$ be a fixed positive real number. We denote $Q_{\mathfrak{T}} = \Omega \times (0, \mathfrak{T})$. According to [83] the governing equations of the compressible flow are given in $Q_{\mathfrak{T}}$ by

$$\phi(x)\partial_t(\rho_\alpha(p_\alpha)s_\alpha) + \operatorname{div}(\rho_\alpha(p_\alpha)V_\alpha) + \rho_\alpha(p_\alpha)s_\alpha q^P = \rho_\alpha(p_\alpha)s_\alpha^I q^I, \quad (\alpha = g, w) \tag{3.2.1}$$

where $\phi$ is the porosity of the medium $\Omega$, $s_\alpha$ is the saturation of the $\alpha$-phase, $\rho_\alpha$ is the density of the phase $\alpha$, $q^P$ is a production term, $q^I$ an injection term, and $s_\alpha^I$ is the saturation of the injected fluid. Moreover, $V_\alpha$ is the velocity of the $\alpha$-phase, which obeys the Darcy-Muskat law [18, 22]

$$V_\alpha = -\frac{K_{r\alpha}(s_\alpha)}{\mu_\alpha}\Lambda\big(\nabla p_\alpha - \rho_\alpha(p_\alpha)\vec{\mathbf{g}}\big), \quad \alpha = g, w, \tag{3.2.2}$$

where $\Lambda$ is the absolute permeability of the porous medium, $K_{r\alpha}$ is the relative permeability of the $\alpha$-phase, $\mu_\alpha$ is the viscosity of the phase $\alpha$, which is considered to be constant, $p_\alpha$ the pressure of the phase $\alpha$ and $\vec{\mathbf{g}}$ is a gravitational term. We assume that the two phases occupy the whole medium, which can be interpreted by the following identity

$$s_w + s_g = 1. \tag{3.2.3}$$

In a capillary tube, the contact between the two fluids incites a curvature, which is due to the difference of their corresponding pressures. This jump represents the capillary pressure law, denoted by $p_c$, and it is assumed to be only in terms of the nonwetting phase saturation. Owing to (3.2.3) we write

$$p_c(s_g) = p_g - p_w. \tag{3.2.4}$$

Physically, the capillary pressure function $p_c := p_c(s_g)$ is nondecreasing, $\big(\dfrac{dp_c(s_g)}{ds_g} > 0$, for any $s_g \in [0, 1]\big)$ [22]. In addition, it degenerates whenever the gas fluid disappears i.e. $p_c(s_g = 0) = 0$. In the sequel, $s = s_g$ will stand for the gas saturation and $s_w = 1 - s$.

In studying the problem (3.2.1)-(3.2.4), the main difficulties are caused by the degeneracy and the strong coupling of the system. To be more precise, the evolution and dissipative terms of each phase vanish whenever the corresponding saturation is equal to zero. As a consequence, we possess no control on the gradients of pressures at the discrete setting. In order to overcome this issue, we need to reformulate this system otherwise with the help of the global pressure feature. This alternative idea has been introduced in [47]. We recall that the global pressure, denoted by $p$, is defined such that the following relationship holds

$$M(s)\nabla p = M_w(s)\nabla p_w + M_g(s)\nabla p_g, \tag{3.2.5}$$

where $M_\alpha$ represents the mobility of the $\alpha$-phase and $M$ is the total mobility. These quantities are defined by

$$M_\alpha = \frac{K_{r\alpha}}{\mu_\alpha}, \quad M(s) = M_w(s) + M_g(s). \tag{3.2.6}$$

Then, the global pressure $p$ can be written in an explicit formula as

$$p = p_g + \overline{p}(s) = p_w + \tilde{p}(s), \tag{3.2.7}$$

with

$$\overline{p}(s) = -\int_0^s \frac{M_w(u)}{M(u)} p_c'(u)\,\mathrm{d}u \quad \text{and} \quad \tilde{p}(s) = \int_0^s \frac{M_g(u)}{M(u)} p_c'(u)\,\mathrm{d}u, \tag{3.2.8}$$

are artificial pressures. We note that the global pressure formulation includes the following function referred to as a capillary term

$$\gamma(s) = \frac{M_w(s)M_g(s)}{M(s)} p_c'(s) \geq 0. \tag{3.2.9}$$

Now, we define $\xi$ as a primitive of the function $\gamma$, which is known under the name of Kirchoff transform

$$\xi(s) = \int_0^s \gamma(u)\,\mathrm{d}u.$$

In case of a regular function $\gamma$, we obtain

$$\nabla\xi(s) = \frac{M_w(s)M_g(s)}{M(s)} \nabla p_c(s).$$

It follows from the definitions of the global pressure and the Kirchoff transform $\xi$ that

$$M_g(s)\nabla p_g = M_g(s)\nabla p + \nabla\xi(s), \tag{3.2.10}$$

$$M_w(s)\nabla p_w = M_w(s)\nabla p - \nabla\xi(s). \tag{3.2.11}$$

Hence, the relations (3.2.10) and (3.2.11) show the strong dependency of these "new" variables on the old ones. At the continuous level, to estimate the gradient of the pressures $p_g$ and $p_w$ we only need to control the gradient of the global pressure $p$ and that of the function $\xi$.

We stress that the water phase is incompressible, meaning $\rho_w$ is constant while the gas density is merely depending on the global pressure, i.e. $\rho_g(p_g) = \rho(p)$, we refer to [47, 83] for more details. We furthermore consider that $s^I = 0$, meaning that no injection of gas is taken into account.

Substituting the previous relationships into the system (3.2.1)-(3.2.2) leads to the global pressure formulation

$$\begin{aligned} &\partial_t(\phi\rho(p)s) - \mathrm{div}\left(\rho(p)M_g(s)\Lambda\nabla p\right) - \mathrm{div}\left(\rho(p)\Lambda\nabla\xi(s)\right) \\ &\quad + \mathrm{div}\left(\rho^2(p)M_g(s)\vec{\mathbf{g}}\right) + \rho(p)sq^P = 0, \end{aligned} \tag{3.2.12}$$

$$\begin{aligned} &\partial_t(\phi s) + \mathrm{div}\left(M_w(s)\Lambda\nabla p\right) - \mathrm{div}\left(\Lambda\nabla\xi(s)\right) \\ &\quad - \mathrm{div}\left(\rho_w M_w(s)\Lambda\vec{\mathbf{g}}\right) + sq^P = q^P - q^I, \end{aligned} \tag{3.2.13}$$

where, henceforth the main unknowns are the global pressure $p$ and the gas saturation $s$. For numerical analysis reasons, we would rather consider this system otherwise. Precisely, the present form of the system yields no energy estimates, especially for the global pressure. So the idea is to

take into account the nondegeneracy of the total mobility and the fraction flow formulation [26]. This formulation reads

$$\partial_t(\phi\rho(p)s) - \operatorname{div}\Big(\rho(p)M(s)f_g(s)\Lambda\nabla p\Big) - \operatorname{div}\Big(\rho(p)\Lambda\nabla\xi(s)\Big) + \operatorname{div}\Big(\rho^2(p)M_g(s)\vec{\mathbf{g}}\Big) + \rho(p)sq^P = 0 \tag{3.2.14}$$

$$\partial_t(\phi s) + \operatorname{div}\Big(M(s)f_w(s)\Lambda\nabla p\Big) - \operatorname{div}\Big(\Lambda\nabla\xi(s)\Big) - \operatorname{div}\Big(\rho_w M_w(s)\Lambda\vec{\mathbf{g}}\Big) + sq^P = q^P - q^I, \tag{3.2.15}$$

where $f_\alpha$ is the fractional flow of the $\alpha$-phase defined by

$$f_\alpha(s) = \frac{M_\alpha(s)}{M(s)}, \quad \alpha = g, w.$$

We further add to the system (3.2.14)-(3.2.15) some mixed boundary conditions of Dirichlet-Neumann type and initial conditions. The boundary $\partial\Omega$ of $\Omega$ comprises two parts $\Gamma_D$ and $\Gamma_N$ whose measures are positive. On $\Gamma_D$, we impose a homogeneous Dirichlet condition and on $\Gamma_N$ we consider a homogeneous Neumann condition as follows

$$\begin{cases} s(x,t) = 0, & \text{on} \quad \Gamma_D \times (0,\mathfrak{T}) \\ p(x,t) = 0, & \text{on} \quad \Gamma_D \times (0,\mathfrak{T}) \\ V_w \cdot \mathbf{n} = V_g \cdot \mathbf{n} = 0 & \text{on} \quad \Gamma_N \times (0,\mathfrak{T}), \end{cases} \tag{3.2.16}$$

where $\mathbf{n}$ is the outward unit normal vector to $\Gamma_N$. Besides, the initial conditions are given by

$$p(x,0) = p^0(x) \quad \text{in } \Omega, \tag{3.2.17}$$

$$s(x,0) = s^0(x) \quad \text{in } \Omega. \tag{3.2.18}$$

Following we list the main assumptions on the physical data.

($\mathrm{H}_1$) The porosity $\phi$ is a $L^\infty(\Omega)$ function such that there exist two positive constants $\phi_0$ and $\phi_1$: $\phi_0 \le \phi(x) \le \phi_1$ a.e. $x \in \Omega$.

($\mathrm{H}_2$) The gas (resp. water) mobility $M_g$ (resp. $M_w$) is a nondecreasing (resp. nonincreasing) Lipschitz continuous function from $[0,1]$ to $\mathbb{R}$ with $M_g(s) = 0$ (resp. $M_w(s) = 0$) for every $s \in ]-\infty, 0]$ (resp. $[1, +\infty[$). Moreover, there exists a positive constant $m_0$ such that, for every $s \in [0,1]$ :

$$0 < m_0 \le M(s) = M_g(s) + M_w(s). \tag{3.2.19}$$

Consequently, the fractional flows verify the same properties as the mobilities. In addition, $f_g(s) + f_w(s) = 1$.

($\mathrm{H}_3$) The absolute permeability $\Lambda$ is a map from $\Omega$ to $\mathcal{S}_d(\mathbb{R})$, where $\mathcal{S}_d(\mathbb{R})$ is the space of $d$-square symmetric matrices. It is also is assumed to be in $L^\infty(\Omega)^{d\times d}$. Furthermore, $\Lambda$ verifies the following inequality

$$\underline{\Lambda}|z|^2 \le \Lambda(x)z \cdot z \le \overline{\Lambda}\,|z|^2, \text{ for all } \ z \in \mathbb{R}^d \text{ and a.e. } \ x \in \Omega$$

for some positive constants $\underline{\Lambda}$ and $\overline{\Lambda}$.

(H$_4$) The function $\gamma$ belongs to $\mathcal{C}^0([0,1],\mathbb{R}^+)$ with

$$\begin{cases} 0<\gamma(s)<1, & \text{for } 0<s<1,\\ \gamma(0)=\gamma(1)=0. \end{cases}$$

We furthermore assume that $\xi^{-1}$ is a $\theta$-Hölder function with $\theta\in(0,1]$ on $[0,\xi(1)]$, which means that there exists a positive constant $L_\xi$ such that for every $a,b\in[0,\xi(1)]$, we have $|\xi^{-1}(a)-\xi^{-1}(b)|\le L_\xi|a-b|^\theta$. This inequality will play a fundamental role in the analysis of the nonlinear CVFE scheme.

(H$_5$) The functions $q^I$ and $q^P$ are in $L^2(Q_\mathfrak{T})$ such that $q^P(x,t),q^I(x,t)\ge 0$ a.e. $(x,t)\in Q_\mathfrak{T}$.

(H$_6$) The density $\rho$ belongs to $\mathcal{C}^1(\mathbb{R},\mathbb{R})$, is strictly increasing, and there exist two constants $\rho_0,\rho_1$ such that $0<\rho_0\le\rho(p)\le\rho_1$.

We next define the natural space where weak solutions are sought

$$H^1_{\Gamma_D}(\Omega)=\{u\in H^1(\Omega)\ /\ u=0 \text{ on } \Gamma_D\}.$$

$H^1_{\Gamma_D}(\Omega)$ is a Hilbert space endowed with the norm

$$||u||_{H^1_{\Gamma_D}(\Omega)}=||\nabla u||_{(L^2(\Omega))^d}.$$

We next give the definition of weak solutions to the continuous problem (3.2.14)-(3.2.18). In the rest of this chapter, we assume that the hypothesis $(H_1)$-$(H_6)$ are fulfilled.

**Definition 3.2.1.** *(Weak solution) Let $p^0$ be a $L^2(\Omega)$-function and $s^0$ be a $L^\infty(\Omega)$-function verifying $0\le s^0(x)\le 1$ a.e. $x\in\Omega$. Then, $(p,s)$ is a weak solution to the problem (3.2.14)-(3.2.18) provided*

$$\begin{aligned} 0\le s(x,t)\le 1 \ \text{ a.e. } (x,t)\in Q_\mathfrak{T},\\ \xi(s)\ \in L^2(0,\mathfrak{T};H^1_{\Gamma_D}(\Omega)),\\ p\ \in L^2(0,\mathfrak{T};H^1_{\Gamma_D}(\Omega)), \end{aligned}$$

*and such that for every $\varphi,\psi\in\mathcal{C}^\infty_c(\Omega\times[0,\mathfrak{T}))$, one has*

$$\begin{aligned} &-\int_{Q_\mathfrak{T}}\phi\rho(p)s\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t-\int_\Omega\phi(x)\rho(p^0)s^0\varphi(x,0)\,\mathrm{d}x\\ &+\int_{Q_\mathfrak{T}}\rho(p)M(s)f_g(s)\Lambda\nabla p\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t+\int_{Q_\mathfrak{T}}\rho(p)\Lambda\nabla\xi(s)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t\\ &-\int_{Q_\mathfrak{T}}\Lambda\rho^2(p)M_g(s)\Lambda\vec{\boldsymbol{g}}\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t+\int_{Q_\mathfrak{T}}\rho(p)sq^P\varphi\,\mathrm{d}x\,\mathrm{d}t=0, \end{aligned}\tag{3.2.20}$$

$$\begin{aligned} &-\int_{Q_\mathfrak{T}}\phi s\partial_t\psi\,\mathrm{d}x\,\mathrm{d}t-\int_\Omega\phi(x)s^0\psi(x,0)\,\mathrm{d}x-\int_{Q_\mathfrak{T}}M(s)f_w(s)\Lambda\nabla p\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t\\ &+\int_{Q_\mathfrak{T}}\Lambda\nabla\xi(s)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t+\int_{Q_\mathfrak{T}}\rho_wM_w(s)\Lambda\vec{\boldsymbol{g}}\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t\\ &+\int_{Q_\mathfrak{T}}sq^P\psi\,\mathrm{d}x\,\mathrm{d}t=\int_{Q_\mathfrak{T}}(q^P-q^I)\psi\,\mathrm{d}x\,\mathrm{d}t. \end{aligned}\tag{3.2.21}$$

For the existence of a weak solution to the problem (3.2.20)-(3.2.21), we refer to this paper [84].

## 3.3 CVFE Mesh and discrete functions

In this section, we present two different types of meshes, a primal mesh and a dual barycentric mesh. We also give a discretization of the time interval. In addition, we define the discrete spaces and functions. To streamline the presentation, we restrict ourselves to the two space dimensions case.

A primal mesh $\mathcal{T}$ is a conforming triangulation, of $\Omega$ in the sense of the finite element method; that is, the intersection of two triangles is either an edge, a vertex or the empty set. The set of vertices of $\mathcal{T}$ (resp. $T \in \mathcal{T}$) is denoted by $\mathcal{V}$ (resp. $\mathcal{V}_T$). We designate by $\mathcal{E}$ (resp. $\mathcal{E}_T$) the set of all edges of $\mathcal{T}$ (resp. $T$). For a triangle $T \in \mathcal{T}$, we define $x_T$ as the barycenter, $h_T = \text{diam}(T)$ the diameter, and $|T|$ the Lebesgue measure of $T$. Let $\varrho_T$ be the diameter of the largest ball inscribed in $T$. The size and regularity of the triangulation $\mathcal{T}$ are respectively denoted by $h$ and $\theta_{\mathcal{T}}$. They are defined to be

$$h := \max_{T\in\mathcal{T}}(h_T), \quad \theta_{\mathcal{T}} := \max_{T\in\mathcal{T}} \frac{h_T}{\varrho_T}.$$

A dual or a barycentric mesh is constructed in the following way. For each vertex $K \in \mathcal{V}$ we associate a unique control volume, denoted $\omega_K$, of the dual mesh. We also denote by $\mathcal{V}_D$ the set of these dual control volumes, then $\Omega = \cup_{K\in\mathcal{V}_D}\omega_K$. Each dual cell $\omega_K$ is obtained by connecting the barycenter of every triangle whose vertex is $K$ with the midpoint of the edges having $K$ as an endpoint. For two vertices $K, L \in \mathcal{V}_T$, we denote by $\sigma^T_{KL}$ the dual interface contained in $T$ and intersects with the segment $[KL]$ whose extremities are $K$ and $L$. By $|\sigma^T_{KL}|$, we mean the length of the interface $\sigma^T_{KL}$ and by $n^T_{\sigma_{KL}}$ the unit normal vector to $\sigma^T_{KL}$ pointing from $K$ to $L$. Next, for $K \in \mathcal{V}$, $|\omega_K|$ is the $d$ dimensional Lebesgue measure of $\omega_K$. We additionally designate by $\mathcal{K}_T$ the set of all triangles sharing the vertex $K$.

We assume that the primal mesh is regular in the sense that there exists a constant $c_0$ such that for any sequence of discretizations $\{\mathcal{T}_m\}_{m\in\mathbb{N}}$ we have

$$\theta_{\mathcal{T}_m} \leq c_0. \tag{3.3.1}$$

**Remark 3.3.1.** *It is worth noticing that the above discretizations of $\Omega$ are still valid and can be obtained in a similar way in case of three dimensions. Indeed, one should perform a tetrahedral mesh with slight changes in the terminology where for instance the triangles are substituted by tetrahedra. Hence, edges and their midpoints are respectively replaced by faces and their barycenters. Also, this 3D partition of $\Omega$ verifies the shape-regularity (3.3.1) condition according to [64].*

A time discretization of the interval $(0, \mathfrak{T})$ is given by a strictly increasing sequence of real numbers $(t^n)_{n=0,\cdots,N}$ with

$$t^0 = 0 \ < t^1 \ < \cdots < \ t^{N-1} \ < \ t^N = \mathfrak{T}.$$

We designate by $\delta t^n = t^{n+1} - t^n$, for $n = 0, \ldots, N-1$ and $\delta t = \max\limits_{n=0,\ldots,N} \delta t^n$. Without loss of generality, we can assume that the time step is uniform.

We now present the approximation spaces, where the discrete unknowns lie in. We also describe the construction of the discrete functions. To do that, let $X_h$ be a finite dimensional space of piecewise linear functions on the primal mesh and $W_h$ the space of piecewise constant functions on the dual mesh. One thus has

$$X_h = \{\varphi \in \mathcal{C}^0(\overline{\Omega}), \varphi_{|T} \in \mathbb{P}_1, \forall T \in \mathcal{T}\} \subset H^1(\Omega). \tag{3.3.2}$$

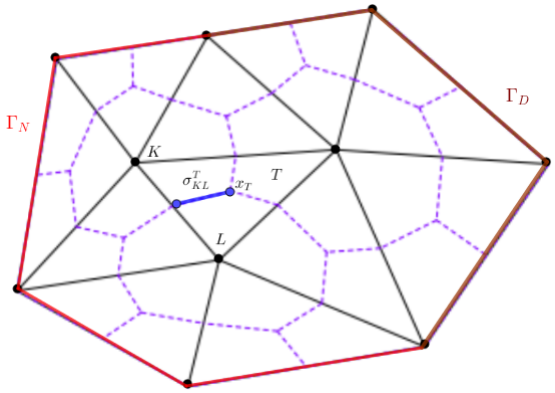



Figure 3.1: Illustration of the primal and the dual meshes.

Let us consider

$$X_h^0 = \{\varphi \in X_h, \varphi(x_K) = 0, \forall K \in \mathcal{V}, x_K \in \Gamma_D\}. \tag{3.3.3}$$

Assuming that the extremities of the Dirichlet boundary $\Gamma_D$ belong to $\mathcal{V}$ as depicted in Fig. 3.1, one gets directly the inclusion $X_h^0 \subset H^1_{\Gamma_D}(\Omega)$. The space $X_h$ possesses a canonical basis of shape functions $(\varphi_K)_{K\in\mathcal{V}}$ with $\varphi_K(x_L) = \delta_{KL}$, where $\delta_{KL}$ is the Kronecker symbol. Furthermore, it is endowed by the following semi-norm

$$||u_h||^2_{X_h} := \int_\Omega |\nabla u_h|^2 \,\mathrm{d}x, \quad \forall u_h \in X_h,$$

which turns out to be a norm on $X_h^0$. Moreover, we recall that

$$\sum_{K\in\mathcal{V}} \varphi_K = 1, \quad \sum_{K\in\mathcal{V}} \nabla\varphi_K = 0 \quad \text{and} \quad \nabla\varphi_{K_{|T}} = -\frac{|\sigma_K^T|}{2\,|T|} n_{\sigma_K^T},$$

where $\sigma_K^T$ is the opposite edge of the vertex $K$ contained in $T$ and $n_{\sigma_K^T}$ is the outward normal to this edge.

For $n \in \{0, \dots, N\}$ and $K \in \mathcal{V}$ we take $u_K^n$ an approximation of $u(x_K, t^n)$. Thus, the discrete unknowns will be denoted by $\{u_K^n\}_{\{K\in\mathcal{V},\ n=0,\dots,N\}}$.

**Definition 3.3.1.** *(Discrete functions)*
*Consider discrete unknowns $\{u_K^n\}_{\{K\in\mathcal{V},\ n=0,\dots,N\}}$. We define two approximate solutions as follows:*

*(i) A finite volume solution $\tilde{u}_{h,\delta t}$ is a piecewise constant function defined almost everywhere in $\bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0, \mathfrak{T})$ with*

$$\tilde{u}_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0 \chi_{\mathring{\omega}_K}(x), \ \forall x \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K,$$

$$\tilde{u}_{h,\delta t}(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} u_K^{n+1} \chi_{\mathring{\omega}_K\times(t^n,t^{n+1}]}(x,t), \ \forall (x,t) \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T}).$$

*(ii) A finite element solution $u_{h,\delta t}$ is a continuous function in space, which is $\mathbb{P}_1$ per triangle, and piecewise constant in time, such that :*

$$u_{h,\delta t}(x,0) = \sum_{K\in\mathcal{V}} u_K^0 \varphi_K(x), \quad \forall x \in \Omega,$$
$$u_{h,\delta t}(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} u_K^{n+1} \varphi_K(x) \chi_{(t^n,t^{n+1}]}(t), \quad \forall\ (x,t) \in \Omega \times (0,\mathfrak{T}).$$

To discretize nonlinear functions, we utilize an interpolation approximation. So, let $F$ be a nonlinear function, we mean by $F(\tilde{u}_{h,\delta t})$ the finite volume reconstruction defined almost everywhere, and by $F(u_{h,\delta t})$ the finite element reconstruction i.e.:

$$F(\tilde{u}_{h,\delta t})(x,0) = \sum_{K\in\mathcal{V}} F(u_K^0)\chi_{\mathring{\omega}_K}(x), \ \forall x \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K,$$
$$F(\tilde{u}_{h,\delta t})(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} F(u_K^{n+1})\chi_{\mathring{\omega}_K\times(t^n,t^{n+1}]}(x,t), \ \forall (x,t) \in \bigcup_{K\in\mathcal{V}} \mathring{\omega}_K \times (0,\mathfrak{T}),$$
$$F(u_{h,\delta t})(x,0) = \sum_{K\in\mathcal{V}} F(u_K^0)\varphi_K(x), \quad \forall x \in \Omega,$$
$$F(u_{h,\delta t})(x,t) = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} F(u_K^{n+1})\varphi_K(x)\chi_{(t^n,t^{n+1}]}(t), \quad \forall\ (x,t) \in \Omega \times (0,\mathfrak{T}).$$

## 3.4 The nonlinear CVFE scheme

In the proposed numerical scheme, we basically carry out a finite volume discretization where the discrete gradient is approximated using a $\mathbb{P}_1$-finite element approximation. In what follows, we sketch out how we obtain the discretization of the gas equation (3.2.14) and in an analogous way we get that of the water equation (3.2.15).

Without loss of generality, we neglect the gravity effects; that is $\vec{\mathbf{g}} \equiv 0$. Then, integrating (3.2.14) on the time-space cell $(t^n, t^{n+1}] \times \omega_K$, for all $n = 0, \ldots, N-1$ and $K \in \mathcal{V}$, and applying the Green-Gauss formula yields

$$\begin{aligned}
&\int_{\omega_K} \phi(x)\Big(\rho(p(x,t^{n+1}))s(x,t^{n+1}) - \rho(p(x,t^n))s(x,t^n)\Big)\,\mathrm{d}x \\
&- \sum_{T\in\mathcal{K}_T} \sum_{\sigma\in\mathcal{E}_K\cap T} \int_{t^n}^{t^{n+1}} \int_{\sigma} \rho(p)M(s)f_g(s)\Lambda\nabla p\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t \\
&- \sum_{T\in\mathcal{K}_T} \sum_{\sigma\in\mathcal{E}_K\cap T} \int_{t^n}^{t^{n+1}} \int_{\sigma} \rho(p)\Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t \\
&+ \int_{t^n}^{t^{n+1}} \int_{\omega_K} \rho(p)sq^P\,\mathrm{d}x\,\mathrm{d}t = 0,
\end{aligned} \tag{3.4.1}$$

where $\mathcal{E}_K$ stands for the set of all edges of the dual control volume associated to $K$ and $\mathbf{n}_{\sigma K}$ denotes the outward unit normal vector to $\sigma$ and $\mathrm{d}\sigma$ is the $d-1$ dimensional Lebesgue measure on $\sigma$. Next,

the evolution term is approximated by Euler's scheme

$$\begin{aligned}
&\int_{\omega_K} \phi(x)\Big(\rho(p(x,t^{n+1}))s(x,t^{n+1}) - \rho(p(x,t^n))s(x,t^n)\Big)\,\mathrm{d}x \\
&\approx \int_{\omega_K} \phi(x)\Big(\rho(\tilde{p}_{h,\delta t}(x,t^{n+1}))\tilde{s}_{h,\delta t}(x,t^{n+1}) - \rho(\tilde{p}_{h,\delta t}(x,t^n))\tilde{s}_{h,\delta t}(x,t^n)\Big)\,\mathrm{d}x, \\
&= |\omega_K|\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big),
\end{aligned} \tag{3.4.2}$$

where $\phi_K$ is the mean value of the porosity function $\phi$ over $\omega_K$. Let us now focus on the discretization of the elliptic term. In the same spirit of [42, 40], this term is approximated as follows

$$\sum_{\sigma\in\mathcal{E}_K\cap T}\int_{t^n}^{t^{n+1}}\int_\sigma \rho(p)\Lambda\nabla\xi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \approx \delta t \sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1}-s_K^{n+1}), \tag{3.4.3}$$

where $\rho_{KL}^{n+1}$ , $\gamma_{KL}^{n+1}$ and $\Lambda_{KL}^T$ are respectively given by

$$\frac{1}{\rho_{KL}^{n+1}} := \begin{cases} \dfrac{1}{p_K^{n+1}-p_L^{n+1}}\displaystyle\int_{p_L^{n+1}}^{p_K^{n+1}} \frac{1}{\rho(z)}\,\mathrm{d}z, & \text{if } \ p_L^{n+1}\neq p_K^{n+1} \\ \dfrac{1}{\rho(p_K^{n+1})}, & \text{otherwise} \end{cases}, \tag{3.4.4}$$

$$\gamma_{KL}^{n+1} := \begin{cases} \max\limits_{s\in I_{KL}^{n+1}}\gamma(s) & \text{if } \Lambda_{KL}^T\geq 0 \\ \min\limits_{s\in I_{KL}^{n+1}}\gamma(s) & \text{otherwise} \end{cases}, \tag{3.4.5}$$

with

$$I_{KL}^{n+1} := \left[\min(s_K^{n+1},s_L^{n+1}),\max(s_K^{n+1},s_L^{n+1})\right],$$

and

$$\begin{cases} \Lambda_{KL}^T := -\displaystyle\int_T \Lambda(x)\nabla\varphi_K\cdot\nabla\varphi_L\,\mathrm{d}x = \Lambda_{LK}^T, \quad \text{for } \ K\neq L, \\ \Lambda_{KK}^T := \displaystyle\sum_{L\in\mathcal{V}_T\setminus\{K\}}\Lambda_{KL}^T. \end{cases} \tag{3.4.6}$$

We point out that the prominence of the choice of $\rho_{KL}^{n+1}$ in (3.4.4) is exhibited in the following identity

$$(p_K^{n+1}-p_L^{n+1}) = \rho_{KL}^{n+1}\Big(g(p_K^{n+1})-g(p_L^{n+1})\Big), \quad \text{with } \ g(p)=\int_0^p\frac{1}{\rho(z)}\,\mathrm{d}z. \tag{3.4.7}$$

One also notices that $g$ is a concave function because $\rho$ is an increasing function. Moreover, the Godunov scheme in (3.4.5) is inspired from [42] and [40] which has been applied to degenerate parabolic equations.

Concerning the convective term, we utilize an upstream value of the fractional flow function $f_g$ on the interface $\sigma_{KL}^T$ with respect to the sign of $\Lambda_{KL}^T(p_L^{n+1}-p_K^{n+1})$. We further use a centered approximation for the total mobility on each triangle $T$ whose vertex is $K$. Consequently, we get

$$\begin{aligned}
&-\sum_{\sigma\in\mathcal{E}_K\cap T}\int_\sigma\left[\Lambda\rho(p)M(s)f_g(s)\nabla p\right]\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}\sigma \\
&\qquad\approx \sum_{L\in\mathcal{V}_T\setminus\{K\}}\rho_{KL}^{n+1}M_T^{n+1}G_g\Big(s_K^{n+1},s_L^{n+1};\Lambda_{KL}^T\delta_{KL}^{n+1}p\Big),
\end{aligned} \tag{3.4.8}$$

where, we hereafter denote $\delta_{KL}^{n+1}p = p_L^{n+1} - p_K^{n+1}$, $G_g$ is a numerical convection flux function. Moreover, $M_T^{n+1}$ is the approximate value of the total mobility

$$M_T^{n+1} = \frac{1}{\#\mathcal{V}_T}\Big(\sum_{K\in\mathcal{V}_T} M(s_K^{n+1})\Big). \tag{3.4.9}$$

The numerical convection flux functions $\{G_\alpha\}_{\alpha=g,w}$, whose arguments are $(a,b,c) \in \mathbb{R}^3$, are defined in the following way. Let $g_w(a,b)$ be any monotone numerical flux for $f_w$, that is:

$(C_1)$ $g_w(a,b)$ is nondecreasing with respect to $a$ and nonincreasing with respect to $b$,

$(C_2)$ $g_w(a,a) = f_w(a)$,

$(C_3)$ $g_w$ is Lipschitz continuous with respect to $a$ and $b$,

then one defines

$$G_w(a,b,c) = g_w(a,b)c^+ - g_w(b,a)c^-, \tag{3.4.10}$$
$$G_g(a,b,c) = G_w(a,b,c) - c, \tag{3.4.11}$$

where $c^+ = \max(c,0)$ and $c^- = -\min(c,0)$. This definition of $G_g$ is required to the coupled nonlinear system and plays a major role to obtain an estimate on the discrete gradient of the global pressure.

**Remark 3.4.1.** *In our context, one possibility to construct the numerical flux $g_w$ is to consider the nondecreasing part $f_{w\uparrow}$ and the nonincreasing part $f_{w\downarrow}$ of the fractional flow $f_w$ such that*

$$g_w(a,b) = f_{w\uparrow}(a) + f_{w\downarrow}(b).$$

*We know that $f_w$ is a nonincreasing function. Then, one gets*

$$g_w(a,b) = f_w(b). \tag{3.4.12}$$

**Lemma 3.4.1.** *According to assumption $(H_2)$ together with (3.4.12), the numerical flux function $g_w$ verifies the properties $(C_1)$-$(C_3)$.*

Finally, the source terms are approximated using the mean values of the functions $\rho(p), s$, $q^P$ and $q^I$.

Gathering the approximations (3.4.2), (3.4.3), (3.4.8) leads to the control volume finite element scheme for the gas equation (3.2.14). In a similar way, we obtain the discretization of the water equation (3.2.15). Then the final scheme reads

$$p_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} p^0(x)\,\mathrm{d}x, \qquad \forall K\in\mathcal{V}, \tag{3.4.13}$$
$$s_K^0 = \frac{1}{|\omega_K|}\int_{\omega_K} s^0(x)\,\mathrm{d}x, \qquad \forall K\in\mathcal{V}. \tag{3.4.14}$$

$$\begin{aligned}\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big) &+ \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}M_T^{n+1}G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\,\delta_{KL}^{n+1}p)\\ &- \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1} - s_K^{n+1})\\ &+ \delta t\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1} = 0,\end{aligned} \tag{3.4.15}$$

$$\phi_K\Big(s_K^{n+1}-s_K^n\Big)+\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}M_T^{n+1}G_w(s_K^{n+1},s_L^{n+1};\Lambda_{KL}^T\,\delta_{KL}^{n+1}p)$$
$$-\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}\gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1}-s_K^{n+1})$$
$$+\delta t(s_K^{n+1}-1)q_{P,K}^{n+1}=-\delta t q_{I,K}^{n+1},\quad \forall n=0,\dots,N-1,\quad \forall K\in\mathcal{V},x_K\notin\Gamma_D. \tag{3.4.16}$$

**Remark 3.4.2.** *Taking into account gravitational effects ($\vec{\boldsymbol{g}}\neq 0$), a new term denoted by $F_{gK}$ would be added to the first equation (3.4.15) of the scheme. This term is the approximation of the integral $\displaystyle\int_{\partial K}\rho_g^2(p)M_g(s)\Lambda\vec{\boldsymbol{g}}\cdot\mathbf{n}\,\mathrm{d}\sigma$. Using the upwind scheme, it is given by*

$$F_{gK}=\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}|\sigma_{KL}^T|\Big(\rho_{KL}^{n+1}\Big)^2\Big(M_g(s_K^{n+1})Z_{KL}^T-M_g(s_L^{n+1})Z_{LK}^T\Big),$$

*where $Z_{KL}^T=\Big(\Lambda\vec{\boldsymbol{g}}\cdot n_{KL}^T\Big)^+=\Big(\Lambda\vec{\boldsymbol{g}}\cdot n_{LK}^T\Big)^-$. In the same way, we add the following expression, denoted $F_{wK}$, to the equation (3.4.16)*

$$F_{wK}=\frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}}|\sigma_{KL}^T|\rho_w\Big(M_w(s_L^{n+1})Z_{KL}^T-M_w(s_K^{n+1})Z_{LK}^T\Big).$$

*Thanks to the monotonicity of the mobilities, the functions $F_{gK}$ and $F_{wK}$ are nondecreasing with respect to $s_K^{n+1}$ and nonincreasing with respect to $s_L^{n+1}$. In addition, they form numerical fluxes which are consistent and conservative. As a consequence, the convergence analysis remains valid.*

## 3.5 Preliminary properties

Throughout we will need these essential properties many times. Their proofs can be found in [40, 42].

**Lemma 3.5.1.** *Let $\psi_{\mathcal{T}}=\sum_{K\in\mathcal{V}}\psi_K\varphi_K\in X_h$, then there exists a constant $C_0=C_0(\Lambda,\theta_{\mathcal{T}})$ such that*

$$\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}|\Lambda_{KL}^T|\,(\psi_K-\psi_L)^2\leq C_0\int_\Omega\Lambda\nabla\psi_{\mathcal{T}}\cdot\nabla\psi_{\mathcal{T}}\,\mathrm{d}x. \tag{3.5.1}$$

**Lemma 3.5.2.** *(Integration by parts) For every $u_h,v_h\in X_h$, there holds*

$$\int_\Omega\Lambda\nabla u_h\cdot\nabla v_h\,\mathrm{d}x=\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T(u_K-u_L)(v_K-v_L). \tag{3.5.2}$$

Let $u_{\mathcal{T}}\in X_h$ and consider the piecewise constant functions $\overline{u}_{\mathcal{T}},\underline{u}_{\mathcal{T}}:\Omega\longrightarrow\mathbb{R}$ defined by

$$\overline{u}_{\mathcal{T}}(x)=\overline{u}_T=\sup_{x\in T}u_{\mathcal{T}}(x),\qquad\text{if } x\in T\in\mathcal{T},$$
$$\underline{u}_{\mathcal{T}}(x)=\underline{u}_T=\inf_{x\in T}u_{\mathcal{T}}(x),\qquad\text{if } x\in T\in\mathcal{T}.$$

**Lemma 3.5.3.** *There exists an absolute constant $c > 0$ such that*

$$\int_\Omega |\overline{u}_\mathcal{T}(x) - \underline{u}_\mathcal{T}(x)|^2 \,\mathrm{d}x \le ch^2 \int_\Omega |\nabla u_\mathcal{T}(x)|^2 \,\mathrm{d}x,$$

*where* $c = \frac{243}{2\pi^2}$.

**Remark 3.5.1.** *The previous lemma holds also in* $L^1(\Omega)$*:*

$$\int_\Omega |\overline{u}_\mathcal{T}(x) - \underline{u}_\mathcal{T}(x)| \,\mathrm{d}x \le \frac{27}{2} h \int_\Omega |\nabla u_\mathcal{T}(x)| \,\mathrm{d}x.$$

**Lemma 3.5.4.** *For* $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$*, let* $u_\mathcal{T}$ *and* $u_\mathfrak{M}$ *be respectively the piecewise linear and the piecewise constant reconstructions. Then*

$$\int_T |u_\mathcal{T}(x) - u_\mathfrak{M}(x)|^2 \,\mathrm{d}x \le ch^2 ||\nabla u_\mathcal{T}||^2_{L^2(\Omega)^d},$$

*where* $c$ *is the same constant as in Lemma 3.5.3.*

## 3.6 Maximum principle and energy estimates

Our goal in this section is to prove the nonnegativity of the approximate saturation and control the gradient of the global pressure $p$ and that of $\xi(s)$. The importance of these estimates will be illustrated below, when we show the convergence of the discrete solutions.

**Lemma 3.6.1.** *(Maximum principle)*
*For* $n = 0, \dots, N-1$*, let* $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$ *be a solution to the combined scheme* (3.4.13)-(3.4.16)*. If* $(s_K^0)_{K\in\mathcal{V}}$ *is in* $[0,1]$ *then* $(\tilde{s}_{h,\delta t})$ *remains also in the interval* $[0,1]$*.*

*Proof.* The claim is performed by induction on $n$. The property is indeed trivial for $n = 0$. We now assume that the sequence $(s_K^k)_{K\in\mathcal{V}} \subset [0,1]$ for $k \le n$ and we prove that the proposition is true for $k = n+1$. For this, let us consider $K \in \mathcal{V}$ such that $s_K^{n+1} = \min\{s_L^{n+1}\}_{L\in\mathcal{V}}$. Multiplying (3.4.15) by $-(s_K^{n+1})^-$ gives

$$\begin{aligned}
&- \phi_K \Big(\rho(p_K^{n+1}) s_K^{n+1} - \rho(p_K^n) s_K^n\Big)(s_K^{n+1})^- \\
&- \frac{\delta t}{|\omega_K|} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1} M_T^{n+1} G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \, \delta_{KL}^{n+1} p)(s_K^{n+1})^- \\
&+ \frac{\delta t}{|\omega_K|} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1} \gamma_{KL}^{n+1} \Lambda_{KL}^T (s_L^{n+1} - s_K^{n+1})(s_K^{n+1})^- \\
&- \delta t \rho(p_K^{n+1}) s_K^{n+1} q_{P,K}^{n+1} (s_K^{n+1})^- = 0.
\end{aligned} \tag{3.6.1}$$

Notice that $s_L^{n+1} \ge s_K^{n+1}$, $G_g$ is a nonincreasing function with respect to the $s_L^{n+1}$ and $G_g$ is consistent i.e. $G_g(a,a,c) = -f_g(a)c$. Thus

$$\begin{aligned}
G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \, \delta_{KL}^{n+1} p)(s_K^{n+1})^- &\le G_g(s_K^{n+1}, s_K^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p)(s_K^{n+1})^- \\
&= -f_g(s_K^{n+1}) \Lambda_{KL}^T \delta_{KL}^{n+1} p (s_K^{n+1})^- = 0,
\end{aligned}$$

where we have used the fact that $f_g$ is extended by zero whenever $s \leq 0$. Hence, the second term in the left hand side of the equation (3.6.1) is nonnegative. Next, thanks to the definition of $\gamma_{KL}^{n+1}$ (3.4.5), and to the fact that $\gamma(s) = 0$ for any $s \leq 0$, we deduce that

$$\gamma_{KL}^{n+1}(s_K^{n+1})^- = 0, \quad \text{if} \quad \Lambda_{KL}^T \leq 0.$$

Indeed, if $s_K^{n+1} \geq 0$ then $(s_K^{n+1})^- = 0$ which gives $\gamma_{KL}^{n+1}(s_K^{n+1})^- = 0$. Conversely, if $s_K^{n+1} < 0$ then we get $\gamma_{KL}^{n+1} = \min_{s \in I_{KL}^{n+1}} \gamma(s) = 0$ since $\Lambda_{KL}^T \leq 0$. Consequently

$$\begin{aligned}
&\frac{\delta t}{|\omega_K|} \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda_{KL}^T \rho_{KL}^{n+1} \gamma_{KL}^{n+1} (s_L^{n+1} - s_K^{n+1})(s_K^{n+1})^- \\
&= \frac{\delta t}{|\omega_K|} \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} (\Lambda_{KL})^+ \rho_{KL}^{n+1} \gamma_{KL}^{n+1} (s_L^{n+1} - s_K^{n+1})(s_K^{n+1})^- \geq 0. \qquad (3.6.2)
\end{aligned}$$

We observe that the source term is nonnegative. Due to the induction assumption on $s_K^n$ we find

$$\begin{aligned}
&-\phi_K \Big( \rho(p_K^{n+1}) s_K^{n+1} - \rho(p_K^n) s_K^n \Big) (s_K^{n+1})^- \\
&\qquad = \phi_K \Big( \rho(p_K^{n+1}) ((s_K^{n+1})^-)^2 + \rho(p_K^n) s_K^n (s_K^{n+1})^- \Big) \leq 0. \qquad (3.6.3)
\end{aligned}$$

We then infer that $(s_K^{n+1})^- = 0$, which implies that $s_K^{n+1} \geq 0$.

In order to prove that $s_K^{n+1} \leq 1$ for every $n = 0, \ldots, N-1$ and $K \in \mathcal{V}$, we similarly argue by induction as above. So, let $\omega_K$ be a dual control volume such that $s_K^{n+1} = \max\{s_L^{n+1}\}_{L \in \mathcal{V}}$ and let us check that $s_K^{n+1} \leq 1$. To do that, we multiply the equation (3.4.16) by $(s_K^{n+1} - 1)^+$

$$\begin{aligned}
&\phi_K \Big( s_K^{n+1} - s_K^n \Big) (s_K^{n+1} - 1)^+ \\
&+ \frac{\delta t}{|\omega_K|} \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} M_T^{n+1} G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \, \delta_{KL}^{n+1} p)(s_K^{n+1} - 1)^+ \\
&- \frac{\delta t}{|\omega_K|} \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda_{KL}^T \gamma_{KL}^{n+1} (s_L^{n+1} - s_K^{n+1})(s_K^{n+1} - 1)^+ \\
&+ \delta t (s_K^{n+1} - 1) q_{P,K}^{n+1} (s_K^{n+1} - 1)^+ = -\delta t q_{I,K}^{n+1} (s_K^{n+1} - 1)^+. \qquad (3.6.4)
\end{aligned}$$

We know that $G_w$ is nonincreasing with respect to the second variable and that is consistent. Thus we have

$$\begin{aligned}
G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \, \delta_{KL}^{n+1} p)(s_K^{n+1} - 1)^+ &\geq G_w(s_K^{n+1}, s_K^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p)(s_K^{n+1} - 1)^+ \\
&= f_w(s_K^{n+1}) \Lambda_{KL}^T \delta_{KL}^{n+1} p (s_K^{n+1} - 1)^+ = 0,
\end{aligned}$$

since the fractional flow $f_w$ is extended by 0 for $s \geq 1$. Next, according to the definition of (3.4.5) and the fact that $\gamma$ is extended by zero whenever $s \geq 1$, we write

$$\gamma_{KL}^{n+1}(s_K^{n+1} - 1)^+ = 0, \quad \text{if} \quad \Lambda_{KL}^T \leq 0. \qquad (3.6.5)$$

This yields

$$\begin{aligned}
&\sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} \Lambda_{KL}^T \gamma_{KL}^{n+1} (s_L^{n+1} - s_K^{n+1})(s_K^{n+1} - 1)^+ \\
&\qquad = \sum_{T \in \mathcal{K}_T} \sum_{L \in \mathcal{V}_T \setminus \{K\}} (\Lambda_{KL})^+ \gamma_{KL}^{n+1} (s_L^{n+1} - s_K^{n+1})(s_K^{n+1} - 1)^+ \leq 0.
\end{aligned}$$

One notices that $\delta t(s_K^{n+1}-1)q_{P,K}^{n+1}(s_K^{n+1}-1)^+ = \delta t\Big((s_K^{n+1}-1)^+\Big)^2 q_{P,K}^{n+1}$ and that the right hand side of (3.6.4) is nonpositive. As a consequence

$$\phi_K\Big(s_K^{n+1}-s_K^n\Big)(s_K^{n+1}-1)^+ \leq 0.$$

Combining this inequality with

$$\Big(s_K^{n+1}-1\Big) = (s_K^{n+1}-1)^+ - (s_K^{n+1}-1)^-,$$

we find that $(s_K^{n+1}-1)^+ = 0$. As a result

$$s_L^{n+1} \leq s_K^{n+1} \leq 1, \ \forall \ \ n = 0, \ldots, N-1, \ \text{and} \ \ \forall L \in \mathcal{V}.$$

This concludes the proof. □

In the sequel, we introduce $(C_i)_i = 1, \cdots$ as a family of values depending only on the data specified in Hypotheses $(H_1)-(H_6)$ and they are independent of the mesh and the time steps. Our concern now is to overestimate the discrete gradient of the global pressure $p$ and that of $\xi$. We also consider the fact that $0 \leq s^0(x) \leq 0$ a.e. $x \in \Omega$.

**Proposition 3.6.1.** *(A priori estimates)*
*Under hypotheses $(H_1)$-$(H_6)$ and the regularity assumption on the mesh (3.3.1), we consider $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$, for each $n = 0, \ldots, N-1$, a solution to the combined scheme (3.4.13)-(3.4.16). Then, there exist two constants $C_p$ and $C_\xi$ depending only on $\Omega$, $T$, $p^0$, $s^0$, $m_0$, $q^P$, $q^I$, $\Lambda$, $\theta_\mathcal{T}$ such that*

$$\begin{aligned}&\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(s_K^N\mathcal{H}(p_K^N) - s_K^0\mathcal{H}(p_K^0)\Big)\\ &+ \frac{m_0}{2}\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T(p_K^{n+1}-p_L^{n+1})^2 \leq C_p,\end{aligned} \tag{3.6.6}$$

*and*

$$\begin{aligned}&\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(B(s_K^N) - B(s_K^0)\Big)\\ &+ \frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{\sigma_{KL}^T\in\mathcal{E}_T}\Lambda_{KL}^T(\xi(s_K^{n+1})-\xi(s_L^{n+1}))^2 \leq C_\xi,\end{aligned} \tag{3.6.7}$$

*where $\mathcal{H}(p) = \rho(p)g(p) - p$ with $g'(p) = \dfrac{1}{\rho(p)}$, and $B'(s) = \xi(s)$.*

*Proof.* We respectively multiply the gas equation (3.4.15) and the water equation (3.4.16) of the combined scheme by $|\omega_K|\, g(p_K^{n+1})$, $-|\omega_K|\, p_K^{n+1}$. Adding them together and summing over $K$ and $n$, leads to

$$A_1 + A_2 + A_3 + A_4 = 0,$$

where

$$A_1 = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big((\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n)g(p_K^{n+1}) - (s_K^{n+1} - s_K^n)p_K^{n+1}\Big),$$

$$A_2 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Big(\rho_{KL}^{n+1} M_T^{n+1} G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p)g(p_K^{n+1}) - M_T^{n+1} G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p)p_K^{n+1}\Big),$$

$$A_3 = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T \Big(\rho_{KL}^{n+1}\gamma_{KL}^{n+1}(s_L^{n+1} - s_K^{n+1})g(p_K^{n+1}) - \gamma_{KL}^{n+1}(s_L^{n+1} - s_K^{n+1})p_K^{n+1}\Big),$$

$$A_4 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K|\Big(\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1}\Big)g(p_K^{n+1}) - \Big((s_K^{n+1} - 1)q_{P,K}^{n+1} + q_{I,K}^{n+1}\Big)p_K^{n+1}.$$

Using the fact that $g$ is concave, we obtain, see [23, 114] for more details, that

$$(\rho(p)s - \rho(p^*)s^*)g(p) - (s - s^*)p \geq \mathcal{H}(p)s - \mathcal{H}(p^*)s^*, \forall s, s^* \in [0,1]. \tag{3.6.8}$$

It follows from this inequality that $A_1$ can be underestimated with a telescopic series. Consequently

$$\sum_{K\in\mathcal{V}} |\omega_K|\Big(\mathcal{H}(p_K^N)s_K^N - \mathcal{H}(p_K^0)s_K^0\Big) \leq A_1. \tag{3.6.9}$$

Now we are interested in seeking a lower bound of $A_2$. First of all, we rearrange the summation by edges, we consider the relationship (3.4.7) and we use the inequality (3.4.11) Therefore

$$\begin{aligned}
A_2 &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} M_T^{n+1}\Big(\rho_{KL}^{n+1} G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p)(g(p_K^{n+1}) - g(p_L^{n+1})) - \\
&\qquad G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p)(p_K^{n+1} - p_L^{n+1})\Big), \\
&= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} M_T^{n+1}\Big(G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p) \\
&\qquad - G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\, \delta_{KL}^{n+1} p)\Big)(p_L^{n+1} - p_K^{n+1}), \\
&= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} M_T^{n+1} \underbrace{\sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (p_L^{n+1} - p_K^{n+1})^2}_{>=0}.
\end{aligned}$$

As a consequence of (3.2.19)

$$m_0 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (p_L^{n+1} - p_K^{n+1})^2 \leq A_2. \tag{3.6.10}$$

Using similar arguments for $A_3$, we can easily check that

$$A_3 = \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \gamma_{KL}^{n+1} \Big(s_K^{n+1} - s_L^{n+1}\Big)\Big(\rho_{KL}^{n+1}\Big(g(p_K^{n+1}) - g(p_L^{n+1})\Big) - (p_K^{n+1} - p_L^{n+1})\Big).$$

It follows from the expression of the coefficient $\rho_{KL}^{n+1}$ defined in (3.4.7) that

$$A_3 = 0. \tag{3.6.11}$$

Owing to the fact that $g$ is sub-linear, i.e. $|g(p)| \leq C_g|p|$, and that $\rho$ is bounded, we deduce

$$|A_4| \leq C_1 \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K|(q_{P,K}^{n+1} + q_{I,K}^{n+1})|p_K^{n+1}|.$$

The Cauchy-Schwarz inequality entails

$$\begin{aligned} |A_4| &\leq C_1\Big(\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K||q_{P,K}^{n+1} + q_{I,K}^{n+1}|^2\Big)^{\frac{1}{2}}\Big(\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K||p_K^{n+1}|^2\Big)^{\frac{1}{2}}, \\ &\leq C_1||q^P + q^I||_{L^2(Q_{\mathfrak{T}})}\Big(\sum_{n=0}^{N-1} \delta t||\tilde{p}_h^{n+1}||_{L^2(\Omega)}^2\Big)^{\frac{1}{2}}. \end{aligned} \tag{3.6.12}$$

An application of the Poincaré inequality [29] yields

$$|A_4| \leq C_2\Big(\sum_{n=0}^{N-1} \delta t||p_h^{n+1}||_{X_h}^2\Big)^{\frac{1}{2}},$$

where $C_2$ is also depending on $||q^P + q^I||_{L^2(Q_{\mathfrak{T}})}$. We now combine the Young inequality ($ab \leq \epsilon a^2 + \frac{b^2}{4\epsilon}$), with $\epsilon = \dfrac{\underline{\Lambda} m_0}{2}$, and the ellipticity of the tensor $\Lambda$ to obtain

$$\begin{aligned} |A_4| &\leq C_3 + \frac{\underline{\Lambda} m_0}{2}\Big(\sum_{n=0}^{N-1} \delta t||\nabla p_h^{n+1}||_{L^2(\Omega)^2}^2\Big), \\ &\leq C_3 + \frac{m_0}{2}\Big(\int_{Q_{\mathfrak{T}}} \Lambda\nabla p_{h,\delta t}\cdot\nabla p_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t\Big). \end{aligned} \tag{3.6.13}$$

Finally, the discrete integration by parts formula (3.5.2) leads to

$$|A_4| \leq C_3 + \frac{m_0}{2}\Big(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (p_L^{n+1} - p_K^{n+1})^2\Big). \tag{3.6.14}$$

Thanks to the relations (3.6.9)-(3.6.11) and (3.6.14), we achieve the proof of the first estimation (3.6.6).

Let us now turn our attention to overestimate the discrete gradient of the capillary term. For this, we routinely multiply the equation (3.4.16) by $\xi(s_K^{n+1})$ and we sum on all $K \in \mathcal{V}$ and $n = 0, \cdots, N-1$. Therefore

$$D_1 + D_2 + D_3 + D_4 = 0,$$

where

$$D_1 = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K|\phi_K(s_K^{n+1} - s_K^n)\xi(s_K^{n+1}),$$

$$D_2 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} M_T^{n+1} G_w\Big(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p\Big)\xi(s_K^{n+1}),$$

$$D_3 = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T \gamma_{KL}^{n+1}(s_L^{n+1} - s_K^{n+1})\xi(s_K^{n+1}),$$

$$D_4 = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K|\Big((s_K^{n+1} - 1)q_{P,K}^{n+1} + q_{I,K}^{n+1}\Big)\xi(s_K^{n+1}).$$

Consider $B$ a primitive of the function $\xi$, i.e, $B'(s) = \xi(s)$, for every $s \in [0,1]$. Observe that

$$B(b) - B(a) = \int_a^b \xi(s)\,\mathrm{d}s = \xi(b)(b-a) - \underbrace{\int_a^b \gamma(s)(s-a)\,\mathrm{d}s}_{\geq 0}.$$

Thereby

$$(a-b)\xi(a) \geq B(a) - B(b), \quad \forall a, b \in [0,1].$$

This inequality gives

$$D_1 = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K|\phi_K(s_K^{n+1} - s_K^n)\xi(s_K^{n+1}) \geq \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(B(s_K^N) - B(s_K^0)\Big). \tag{3.6.15}$$

Reorganizing the expression of $D_2$ by edges, we get

$$D_2 = -\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} M_T^{n+1} G_w\Big(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p\Big)(\xi(s_L^{n+1}) - \xi(s_K^{n+1})).$$

An application of the Young inequality ($ab \leq \dfrac{\epsilon a^2}{2} + \dfrac{b^2}{2\epsilon}$), with $\epsilon = C_0$ (this constant figures in Lemma 3.5.1), yields

$$\begin{aligned}
|D_2| &\leq C_5 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} |\Lambda_{KL}^T||p_K^{n+1} - p_L^{n+1}||\xi(s_L^{n+1}) - \xi(s_K^{n+1})|, \\
&\leq C_6 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} |\Lambda_{KL}^T|\Big(p_K^{n+1} - p_L^{n+1}\Big)^2 \\
&+ \frac{1}{C_0}\frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} |\Lambda_{KL}^T|\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)^2.
\end{aligned}$$

According to Lemma 3.5.1 and relation (3.5.2), we get

$$|D_2| \leq C_6 \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (p_K^{n+1} - p_L^{n+1})^2$$
$$+ \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (\xi(s_K^{n+1}) - \xi(s_L^{n+1}))^2.$$

In virtue of the estimate (3.6.6), we obtain

$$|D_2| \leq C_7 + \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T (\xi(s_K^{n+1}) - \xi(s_L^{n+1}))^2. \tag{3.6.16}$$

Similarly, we reorganize the summation $D_3$ by interfaces. We thereby discover

$$D_3 = \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \gamma_{KL}^{n+1} \Big(s_L^{n+1} - s_K^n\Big)\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big).$$

The regularity of $\xi$ ensures the existence of $s^* \in I_{KL}^{n+1} = [\min(s_K^{n+1}, s_L^{n+1}), \max(s_K^{n+1}, s_L^{n+1})]$ such that

$$\xi(s_L^{n+1}) - \xi(s_K^{n+1}) = \gamma(s^*)(s_L^{n+1} - s_K^{n+1}).$$

Now if $\Lambda_{KL}^T \geq 0$, we get $\Lambda_{KL}^T \gamma(s^*) \leq \Lambda_{KL}^T \gamma_{KL}^{n+1}$ since $\gamma_{KL}^{n+1}$ is the maximum of $\gamma$ on $I_{KL}^{n+1}$. Otherwise, $\Lambda_{KL}^T \leq 0$, $\Lambda_{KL}^T \gamma(s^*) \leq \Lambda_{KL}^T \gamma_{KL}^{n+1}$, since the minimum of $\gamma$ is $\gamma_{KL}^{n+1}$. In both cases, we have $\Lambda_{KL}^T \gamma(s^*) \leq \Lambda_{KL}^T \gamma_{KL}^{n+1}$. Next, $\xi$ is a nondecreasing function, which yields the nonnegativity of the term $\Big(s_L^{n+1} - s_K^n\Big)\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)$. Thus

$$D_3 \geq \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \gamma(s^*) \Big(s_L^{n+1} - s_K^n\Big)\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big),$$
$$= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)^2. \tag{3.6.17}$$

The term $D_4$ can be treated as $A_4$. As a result, we check in a straightforward way that

$$|D_4| \leq C_8 + \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)^2.$$

In conclusion, we get

$$\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T \Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)^2 \leq C_9. \tag{3.6.18}$$

Hence, the proof of Proposition 3.6.1 is complete. □

## 3.7 Existence of discrete solutions

Here we claim that the combined finite volume finite element scheme possesses a solution. This is essentially based on the following fundamental lemma, that can be found in [65]. This lemma provides a sufficient condition so that a vector field can admit a zero.

**Lemma 3.7.1.** *Let $\mathcal{A}$ be a finite dimensional space with inner product $(\cdot,\cdot)$ and norm $||\cdot||$, and let $\mathcal{P}$ be a continuous mapping from $\mathcal{A}$ into itself satisfying*

$$(\mathcal{P}(x), x) > 0 \quad \textit{for} \quad ||x|| = r > 0.$$

*Then there exists $x^* \in \mathcal{A}$ with $||x^*|| < r$ such that*

$$\mathcal{P}(x^*) = 0.$$

We are now in position to state and prove the existence result.

**Proposition 3.7.1.** *(Existence)*
*Under hypotheses $(H_1)$-$(H_6)$ and the regularity assumption on the mesh (3.3.1), there exists at least one solution $(p_K^{n+1}, s_K^{n+1})_{K\in\mathcal{V}}$, for $n = 0, \dots, N$, to the coupled scheme (3.4.13)-(3.4.16)*

*Proof.* For the sake of clarity, we denote

$$\begin{aligned} q &:= \text{Card}\{K \in \mathcal{V}/x_K \notin \Gamma_D\},\\ s &:= \{s_K^{n+1}\}_{K\in\mathbb{R}^q},\\ p &:= \{p_K^{n+1}\}_{K\in\mathbb{R}^q}. \end{aligned}$$

We define the mapping $\Phi : \mathbb{R}^q \times \mathbb{R}^q \longrightarrow \mathbb{R}^q \times \mathbb{R}^q$, such that

$$\Phi(p,s) = \Big(\{\Phi_{1,K}\}_{K\in\mathcal{V}}, \{\Phi_{2,K}\}_{K\in\mathcal{V}}\Big),$$

where

$$\begin{aligned} \Phi_{1,K} &= \phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big)\\ &\quad + \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1} M_T^{n+1} G_g(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p)\\ &\quad - \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1} - s_K^{n+1}) + \delta t\rho(p_K^{n+1})s_K^{n+1}q_{P,K}^{n+1},\\ \Phi_{2,K} &= \phi_K\Big(s_K^{n+1} - s_K^n\Big) + \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} M_T^{n+1} G_w(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p)\\ &\quad - \frac{\delta t}{|\omega_K|}\sum_{T\in\mathcal{K}_T}\sum_{L\in\mathcal{V}_T\setminus\{K\}} \gamma_{KL}^{n+1}\Lambda_{KL}^T(s_L^{n+1} - s_K^{n+1}) + \delta t(s_K^{n+1} - 1)q_{P,K}^{n+1} + \delta t q_{I,K}^{n+1}. \end{aligned}$$

It follows from the assumptions on the data that $\Phi$ is well-defined and continuous. We now define the following homeomorphism $\mathcal{F} : \mathbb{R}^q \times \mathbb{R}^q \longrightarrow \mathbb{R}^q \times \mathbb{R}^q$, such that

$$\mathcal{F}(p,s) = (u,v),$$

where, $u = \{g(p_K^{n+1})\}_{K\in\mathcal{V}}$ and $v = \{-p_K^{n+1} + \xi(s_K^{n+1})\}_{K\in\mathcal{V}}$. We next consider the continuous mapping $\mathcal{P}$ as follows

$$\mathcal{P}(u,v) = \Phi \circ \mathcal{F}^{-1}(u,v) = \Phi(p,s).$$

It remains to check that

$$\Big(\mathcal{P}(u,v),(u,v)\Big) > 0, \quad \text{for} \quad ||(u,v)||_{\mathbb{R}^{2q}} = r, \tag{3.7.1}$$

for some sufficiently large $r$. Being inspired by the calculus of the energy estimates proof, we find

$$\begin{aligned}\Big(\mathcal{P}(u,v),(u,v)\Big) \geq & \frac{1}{\delta t}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(s_K^{n+1}\mathcal{H}(s_K^{n+1}) - s_K^n\mathcal{H}(s_K^n)\Big)\\ & + \frac{1}{\delta t}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(B(s_K^{n+1}) - B(s_K^n)\Big)\\ & + \frac{m_0\underline{\Lambda}}{2}||p_h^{n+1}||_{X_h}^2 + \frac{\underline{\Lambda}}{2}||\xi(s_h^{n+1})||_{X_h}^2 - C_p' - C_\xi',\end{aligned}$$

for some positive constants $C_p', C_\xi'$. Consequently

$$\begin{aligned}\Big(\mathcal{P}(u,v),(u,v)\Big) \geq & -\frac{1}{\delta t}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(s_K^n\mathcal{H}(s_K^n) + B(s_K^n)\Big)\\ & + \min\Big(\frac{m_0\underline{\Lambda}}{2}, \frac{\underline{\Lambda}}{2}\Big)\Big(||p_h^{n+1}||_{X_h}^2 + ||\xi(s_h^{n+1})||_{X_h}^2\Big) - C_p' - C_\xi'.\end{aligned} \tag{3.7.2}$$

In view of Lemma 3.5.4, the Poincaré inequality and the Lipschitz continuity of the function $g$, there exists a positive constant $L$ such that

$$\begin{aligned}\Big\|\Big(u,v\Big)\Big\|_{\mathbb{R}^{2q}}^2 &= \Big\|\Big(\{g(p_K^{n+1})\}_{K\in\mathcal{V}}, \{-p_K^{n+1} + \xi(s_K^{n+1})\}_{K\in\mathcal{V}}\Big)\Big\|_{\mathbb{R}^{2q}}^2,\\ &\leq L\Big(\big\|\xi(s_h^{n+1})\big\|_{X_h}^2 + \big\|p_h^{n+1}\big\|_{X_h}^2\Big).\end{aligned} \tag{3.7.3}$$

Therefore, the last inequality implies that (3.7.1) is fulfilled if $r$ is large enough. □

## 3.8 Space and time translates

In this section we aim to establish some compactness results, consisting of space and time translates on the gas mass sequence $\tilde{\phi}_h\rho(p_{h,\delta t})s_{h,\delta t}$. To do that, we require the following lemma. This result affirms that the difference between the finite volume and the finite element reconstruction of the underlined sequence tends to zero whenever the size of the mesh goes to zero.

**Lemma 3.8.1.** *The hypotheses $(H_1)$-$(H_6)$ and the regularity assumption on the mesh (3.3.1) are assumed to be fulfilled. Denote $\mathbb{U}_{h,\delta t} = \tilde{\phi}_h\rho(p_{h,\delta t})s_{h,\delta t}$ and $\tilde{\mathbb{U}}_{h,\delta t} = \tilde{\phi}_h\rho(\tilde{p}_{h,\delta t})\tilde{s}_{h,\delta t}$. Then*

$$\Big\|\mathbb{U}_{h,\delta t} - \tilde{\mathbb{U}}_{h,\delta t}\Big\|_{L^1(Q_{\mathfrak{T}})} \longrightarrow 0 \text{ as } h \longrightarrow 0.$$

*Proof.* We simply write

$$\begin{aligned}\left\|\mathbb{U}_{h,\delta t}-\tilde{\mathbb{U}}_{h,\delta t}\right\|_{L^1(Q_{\mathfrak{T}})} &= \int_{Q_{\mathfrak{T}}} |\mathbb{U}_{h,\delta t}-\tilde{\mathbb{U}}_{h,\delta t}|\,\mathrm{d}x\,\mathrm{d}t,\\ &= \int_{Q_{\mathfrak{T}}} |\tilde{\phi}_h\rho(p_{h,\delta t})s_{h,\delta t}-\tilde{\phi}_h\rho(\tilde{p}_{h,\delta t})\tilde{s}_{h,\delta t}|\,\mathrm{d}x\,\mathrm{d}t,\\ &\leq E_1+E_2,\end{aligned}$$

where $E_1$ and $E_2$ read

$$\begin{aligned}E_1 &= \phi_1\rho_1\int_{Q_{\mathfrak{T}}} |s_{h,\delta t}-\tilde{s}_{h,\delta t}|\,\mathrm{d}x\,\mathrm{d}t,\\ E_2 &= \phi_1\int_{Q_{\mathfrak{T}}} |\rho(p_{h,\delta t})-\rho(\tilde{p}_{h,\delta t})|\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Using the fact that $\xi^{-1}$ is a $\theta$-Hölder function, we infer

$$E_1 \leq \phi_1 L_\xi \int_{Q_{\mathfrak{T}}} |\xi(s_{h,\delta t})-\xi(\tilde{s}_{h,\delta t})|^\theta\,\mathrm{d}x\,\mathrm{d}t.$$

Now Hölder's inequality with $\theta\in(0,1]$ implies

$$E_1 \leq C\Big(\int_{Q_{\mathfrak{T}}} |\xi(s_{h,\delta t})-\xi(\tilde{s}_{h,\delta t})|\,\mathrm{d}x\,\mathrm{d}t\Big)^\theta =: C(E_1')^\theta,$$

where

$$E_1' = \int_{Q_{\mathfrak{T}}} |\xi(s_{h,\delta t})-\xi(\tilde{s}_{h,\delta t})|\,\mathrm{d}x\,\mathrm{d}t.$$

This expression of $E_1'$ can be developed as follows

$$\begin{aligned}E_1' &= \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{K\in\mathcal{V}_T}\int_{\omega_K\cap T} |\xi(s_{h,\delta t})-\xi(\tilde{s}_{h,\delta t})|\,\mathrm{d}x,\\ &= \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{K\in\mathcal{V}_T}\int_{\omega_K\cap T} |\xi(s_{h,\delta t}(x,t))-\xi(s_{h,\delta t}(x_K,t))|\,\mathrm{d}x,\\ &= \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{K\in\mathcal{V}_T}\int_{\omega_K\cap T} \left|\nabla\xi(s_{h,\delta t})_{|T}\cdot(x-x_K)\right|\mathrm{d}x,\\ &\leq \sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}\sum_{K\in\mathcal{V}_T} \mathrm{diam(T)}|\omega_K\cap T|\left|\nabla\xi(s_{h,\delta t})_{|T}\right|,\\ &\leq h\sum_{n=0}^{N-1}\delta t\sum_{T\in\mathcal{T}}|T|\left|\nabla\xi(s_{h,\delta t})_{|T}\right| \leq (\mathfrak{T}\,|\Omega|)^{\frac{1}{2}}h\Big(\int_0^{\mathfrak{T}}\|\nabla\xi(s_{h,\delta t})\|^2_{L^2(\Omega)^d}\,\mathrm{d}t\Big)^{\frac{1}{2}} \leq C_{10}h.\end{aligned}$$

where we have applied the Cauchy-Schwarz inequality together with (3.6.7). As a result

$$E_1 \leq C_{11}h^\theta \to 0 \text{ as } h\to 0.$$

The function $\rho'$ is bounded, then we estimate

$$E_2 \leq \phi_1 \left\|\rho'\right\|_\infty \int_{Q_\mathfrak{T}} |p_{h,\delta t} - \tilde{p}_{h,\delta t}| \, \mathrm{d}x \, \mathrm{d}t.$$

The same conclusion can be drawn for $E_2$

$$E_2 \leq C_{12} h \to 0 \text{ as } h \to 0.$$

We deduce that the difference between $\mathbb{U}_{h,\delta t}$ and $\tilde{\mathbb{U}}_{h,\delta t}$ tends to zero as $h$ goes to zero. This ends the proof. □

We now give the space translates result on $\tilde{\mathbb{U}}_{h,\delta t}$.

**Lemma 3.8.2.** *(Space Translates)*
*Under the hypotheses $(H_1)$-$(H_6)$ and the regularity assumption on the mesh (3.3.1), let $(p_{h,\delta t}, s_{h,\delta t})$ be a solution to (3.4.13)-(3.4.16). Then the following inequality holds*

$$\int_0^\mathfrak{T} \int_{\Omega'} \left| \tilde{\mathbb{U}}_{h,\delta t}(x+y,t) - \tilde{\mathbb{U}}_{h,\delta t}(x,t) \right| \mathrm{d}x \, \mathrm{d}t \leq \beta(|y|), \tag{3.8.1}$$

*for every $y \in \mathbb{R}^d$, where $\Omega' = \{x \in \Omega, \quad [x, x+y] \subset \Omega\}$ and $\beta(|y|) \longrightarrow 0$ as $|y|$ goes to zero.*

*Proof.* In view of the expression of $\mathbb{U}_{h,\delta t}$ we have

$$\begin{aligned}
&\int_{Q'_\mathfrak{T}} \left| \tilde{\mathbb{U}}_{h,\delta t}(x+y,t) - \tilde{\mathbb{U}}_{h,\delta t}(x,t) \right| \mathrm{d}x \, \mathrm{d}t, \\
&= \int_{Q'_\mathfrak{T}} \left| \left(\tilde{\phi_h} \rho(\tilde{p}_{h,\delta t}) \tilde{s}_{h,\delta t}\right)(x+y,t) - \left(\tilde{\phi_h} \rho(\tilde{p}_{h,\delta t}) \tilde{s}_{h,\delta t}\right)(x,t) \right| \mathrm{d}x \, \mathrm{d}t, \\
&\leq R_1 + R_2 + R_3,
\end{aligned}$$

where $R_1, R_2$ and $R_3$ are given by

$$R_1 = \phi_1 \rho_1 \int_{Q'_\mathfrak{T}} |\tilde{s}_{h,\delta t}(x+y,t) - \tilde{s}_{h,\delta t}(x,t)| \, \mathrm{d}x \, \mathrm{d}t, \tag{3.8.2}$$

$$R_2 = \phi_1 \int_{Q'_\mathfrak{T}} |\rho(\tilde{p}_{h,\delta t}(x+y,t)) - \rho(\tilde{p}_{h,\delta t}(x,t))| \, \mathrm{d}x \, \mathrm{d}t. \tag{3.8.3}$$

$$R_3 = \rho_1 \int_{Q'_\mathfrak{T}} \left| \tilde{\phi_h}(x+y) - \tilde{\phi_h}(x) \right| \mathrm{d}x \, \mathrm{d}t. \tag{3.8.4}$$

In order to estimate $R_1$, we introduce once more the $\theta$-Hölder continuity of $\xi^{-1}$. So, one has

$$R_1 \leq C_{13} \int_{Q'_\mathfrak{T}} |\xi(\tilde{s}_{h,\delta t}(x+y,t)) - \xi(\tilde{s}_{h,\delta t}(x,t))|^\theta \, \mathrm{d}x \, \mathrm{d}t.$$

The Hölder inequality allows us to write

$$R_1 \leq C_{14} \Big( \int_{Q'_\mathfrak{T}} |\xi(\tilde{s}_{h,\delta t}(x+y,t)) - \xi(\tilde{s}_{h,\delta t}(x,t))| \, \mathrm{d}x \, \mathrm{d}t \Big)^\theta.$$

As in the same spirit of [69], we define the function $\chi_{\sigma_{KL}^T}(x)$ for each $\sigma_{KL}^T$ by

$$\chi_{\sigma_{MS}^K}(x) = \begin{cases} 1, & \text{if the line segment } [x, x+y] \text{ intersects } \sigma_{KL}^T, \\ 0, & \text{else.} \end{cases}$$

for $y \in \mathbb{R}, x \in \Omega'$ and $K, L \in \mathcal{V}_{\mathcal{T}}$. It is known that $\int_{\Omega'} \chi_{\sigma_{MS}^K}(x)\,\mathrm{d}x \leq C_\sigma |\sigma_{KL}^T||y|$. Thereby

$$\begin{aligned}
R_1 &\leq C_{14}\Big(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\xi(s_L^{n+1}) - \xi(s_K^{n+1})\right| \int_{\Omega'} \chi_{\sigma_{MS}^K}(x)\,\mathrm{d}x\Big)^\theta, \\
&\leq C_{15}|y|^\theta \Big(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\sigma_{KL}^T\right| \left|\xi(s_L^{n+1}) - \xi(s_K^{n+1})\right| \Big)^\theta, \\
&\leq C_{16}|y|^\theta \Big(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} |T|^{\frac{1}{2}} \left|\xi(s_L^{n+1}) - \xi(s_K^{n+1})\right| \Big)^\theta, \\
&\leq C_{17}|y|^\theta \Big(\sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} |T| \left|\nabla \xi(s_{h,\delta t})_{|T}\right|^2 \Big)^{\theta/2}, \\
&\leq C_{18}|y|^\theta \Big(\int_0^{\mathfrak{T}} \|\nabla \xi(s_{h,\delta t})\|_{L^2(\Omega)^d}^2\,\mathrm{d}t\Big)^{\theta/2} \leq C_{19}|y|^\theta,
\end{aligned}$$

where we have mainly used the regularity of the mesh, within the triangle $\omega_K \cap T$, and the Cauchy-Schwarz inequality. Analogous arguments are employed to prove

$$R_2 \leq C|y||h|. \tag{3.8.5}$$

It is easy to see from the assumption $(\mathrm{H}_1)$ on the porosity that the space translates are strongly convergent in $L^1(\Omega)$ which leads to

$$R_3 \to 0 \text{ as } |y| \to 0.$$

This inequality together with the previous one establish the required property (3.8.1). □

The following lemma asserts the time translates on $\tilde{\mathbb{U}}_{h,\delta t}$ .

**Lemma 3.8.3.** *(Time translates)*
*Under the hypotheses $(H_1)$-$(H_6)$ and the regularity assumption on the mesh (3.3.1), let $(p_{h,\delta t}, s_{h,\delta t})$ be a solution to the algebraic system (3.4.13)-(3.4.16). There exists a modulus of continuity $\omega$ that does not depend on $h$ nor on $\delta t$ such that*

$$\int_{\Omega\times(0,\mathfrak{T}-\tau)} \left|\tilde{\mathbb{U}}_{h,\delta t}(x,t+\tau) - \tilde{\mathbb{U}}_{h,\delta t}(x,t)\right|^2 \mathrm{d}x\,\mathrm{d}t \leq \omega(\tau), \tag{3.8.6}$$

*for all $\tau \in (0,\mathfrak{T})$. Moreover $\omega(\tau) \longrightarrow 0$ as $\tau \longrightarrow 0$.*

*Proof.* The proof follows analogous ideas as provided in [23, 69]. □

## 3.9 Convergence of the control volume finite element scheme

We are now in a position to state and prove the main theorem of this chapter, which asserts the convergence of any sequence of discrete solutions to the nonlinear CVFE scheme towards a weak solution of the continuous problem. This result is essentially based on the energy estimates, and the Kolmogorov compactness theorem.

**Proposition 3.9.1.** *Let $(\mathcal{T}_h)_h$ be a family of meshes of $\Omega$ satisfying the regularity assumption (3.3.1) with $h = size(\mathcal{T}_h) \to 0$ . Under assumptions $(H_1)$-$(H_6)$, let $(p_{h,\delta t}, s_{h,\delta t})$ be a sequence of solutions to the numerical scheme (3.4.13)-(3.4.16). Then, there exists a subsequence of $p_{h,\delta t}, s_{h,\delta t}, \tilde{p}_{h,\delta t}$ and $\tilde{s}_{h,\delta t}$ satisfying the following convergences*

$$\tilde{\mathbb{U}}_{h,\delta t} \text{ and } \mathbb{U}_{h,\delta t} \longrightarrow U \qquad \text{strongly in } L^r(Q_{\mathfrak{T}}), r \geq 1, \quad \text{and a.e in } Q_{\mathfrak{T}}, \tag{3.9.1}$$

$$\tilde{s}_{h,\delta t} \text{ and } s_{h,\delta t} \longrightarrow s \qquad \text{a.e. in } Q_{\mathfrak{T}}, \tag{3.9.2}$$

$$\tilde{p}_{h,\delta t}, p_{h,\delta t} \rightharpoonup p \qquad \text{weakly in } L^2(Q_{\mathfrak{T}}), \tag{3.9.3}$$

$$\nabla p_{h,\delta t} \rightharpoonup \nabla p \qquad \text{weakly in } L^2(Q_{\mathfrak{T}})^d, \tag{3.9.4}$$

$$\nabla \xi(s_{h,\delta t}) \rightharpoonup \nabla \xi(s) \qquad \text{weakly in in } L^2(Q_{\mathfrak{T}})^d. \tag{3.9.5}$$

*Moreover, $\xi(s)$ and $p$ are in $L^2(0, \mathfrak{T}; H^1_{\Gamma_D}(\Omega))$ with*

$$0 \leq s \leq 1 \text{ a.e. in } Q_{\mathfrak{T}}, \tag{3.9.6}$$

$$U = \phi \rho(p) s \text{ a.e. in } Q_{\mathfrak{T}}. \tag{3.9.7}$$

*Finally, for all functions $\Gamma$ and $\kappa \in \mathcal{C}^0_b(\mathbb{R})$, with $\kappa(0) = 0$, we have*

$$\Gamma(p_{h,\delta t})\kappa(s_{h,\delta t}) \longrightarrow \Gamma(p)\kappa(s) \text{ a.e. in } Q_{\mathfrak{T}} \tag{3.9.8}$$

*Proof.* The proof is similar to that of Proposition 2.9.1. □

Let us now demonstrate the main result of this chapter, which attests that any limit of the sequence of solutions is a weak solution of the continuous problem.

**Theorem 3.9.1.** *(Passage to the limit)*
*Under the assumptions of Proposition 3.9.1, the limit function $(p, s)$ given in (3.9.2) and (3.9.3) is a weak solution of the problem (3.2.14)-(3.2.18) in the sense of Definition 3.2.1.*

*Proof.* For the ease of readability, some expressions and quantities exhibit only the index $h$ whereas they depend on both $\delta t$ and $h$. We detail the proof in the case of the gas equation and that of the water equation mimics the same steps. To this purpose, let $\psi \in \mathcal{C}^\infty_c(\Omega \times [0, \mathfrak{T}))$. Multiply the equation (3.4.15) by $\delta t \psi_K^{n+1} := \delta t \psi(x_K, t^{n+1})$ for all $K \in \mathcal{V}$ and $n \in \{0, \ldots, N\}$, sum over $K$ and $n$. Then

$$\mathcal{W}_1^h + \mathcal{W}_2^h + \mathcal{W}_3^h + \mathcal{W}_4^h + \mathcal{W}_5^h = 0,$$

where

$$\mathcal{W}_1^h = \sum_{n=0}^{N-1} \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\Big(\rho(p_K^{n+1})s_K^{n+1} - \rho(p_K^n)s_K^n\Big)\psi_K^{n+1},$$

$$\mathcal{W}_2^h = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}\Lambda_{KL}^T\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)\psi_K^{n+1},$$

$$\mathcal{W}_3^h = -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T\rho_{KL}^{n+1}\Big(\gamma_{KL}^{n+1}(s_L^{n+1} - s_K^{n+1}) - (\xi(s_L^{n+1}) - \xi(s_K^{n+1}))\Big)\psi_K^{n+1},$$

$$\mathcal{W}_4^h = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \rho_{KL}^{n+1}M_T^{n+1}G_g\Big(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T\delta_{KL}^{n+1}p\Big)\psi_K^{n+1},$$

$$\mathcal{W}_5^h = \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K|\rho(p_K^{n+1})s_K^{n+1}q_K^{P,n+1}\psi_K^{n+1}.$$

We first rearrange the summation $\mathcal{W}_1^h$ taking into account $\psi_K^N = \psi(x_K, \mathfrak{T}) = 0$

$$\begin{aligned}\mathcal{W}_1^h &= -\sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^{n+1})s_K^{n+1}\Big(\psi_K^{n+1} - \psi_K^n\Big) - \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^0)s_K^0\psi_K^0,\\ &= -\sum_{n=0}^{N-1}\sum_{K\in\mathcal{V}} \int_{t^n}^{t^{n+1}} \int_{\omega_K} \phi_K\rho(p_K^{n+1})s_K^{n+1}\partial_t\psi(x_K,t)\,\mathrm{d}x\,\mathrm{d}t - \sum_{K\in\mathcal{V}} |\omega_K|\phi_K\rho(p_K^0)s_K^0\psi_K^0.\end{aligned}$$

As in [23] we show in a straightforward way that

$$\lim_{h,\delta t\to 0} \mathcal{W}_1^h = -\int_{Q_{\mathfrak{T}}} \phi\rho(p)s\partial_t\psi(x,t)\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \phi\rho(p^0)s^0\ \psi(x,0)\,\mathrm{d}x.$$

We next demonstrate the following limit

$$\lim_{h,\delta t\to 0} \mathcal{W}_2^h = \int_{Q_{\mathfrak{T}}} \rho(p)\Lambda\nabla\xi(s)\cdot\nabla\ \psi\,\mathrm{d}x\,\mathrm{d}t.$$

To do this, we integrate by parts $\mathcal{W}_2^h$

$$\begin{aligned}\mathcal{W}_2^h &= -\sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} \sum_{T\in\mathcal{K}_T} \sum_{L\in\mathcal{V}_T\setminus\{K\}} \Lambda_{KL}^T\rho_{KL}^{n+1}\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)\psi_K^{n+1},\\ &= \sum_{n=0}^{N-1} \delta t \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \Lambda_{KL}^T\rho_{KL}^{n+1}\Big(\xi(s_L^{n+1}) - \xi(s_K^{n+1})\Big)\Big(\psi_L^{n+1} - \psi_K^{n+1}\Big).\end{aligned}$$

Now consider

$$\mathcal{W}_2^{h,*} = \int_{Q_{\mathfrak{T}}} \rho(p_{h,\delta t})\Lambda\nabla\xi(s_{h,\delta t})\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t, \tag{3.9.9}$$

and let us show that this expression converges to the desired limit. To start off, remark that

$$\mathcal{W}_2^{h,*} = \int_{Q_{\mathfrak{T}}} \Lambda\nabla(\rho(p_{h,\delta t})\xi(s_{h,\delta t}))\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t - \int_{Q_{\mathfrak{T}}} \xi(s_{h,\delta t})\Lambda\nabla\rho(p_{h,\delta t})\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t.$$

Using the fact that the functions $\rho(p_{h,\delta t})$ and $\xi(s_{h,\delta t})$, and their gradients are bounded. We deduce that

$$\nabla(\rho(p_{h,\delta t})\xi(s_{h,\delta t})) \rightharpoonup \nabla(\rho(p)\xi(s)), \text{ weakly in } L^2(Q_{\mathfrak{T}})^d.$$

In addition, there exists $\rho^* \in L^2(Q_{\mathfrak{T}})$ such that

$$\rho(p_{h,\delta t}) \rightharpoonup \rho^\star, \text{ weakly in } L^2(Q_{\mathfrak{T}}),$$

and

$$\nabla\rho(p_{h,\delta t}) \rightharpoonup \nabla\rho^\star, \text{ weakly in } L^2(Q_{\mathfrak{T}})^d.$$

Moreover, it follows from the strong convergence in $L^2(Q_{\mathfrak{T}})^d$ of the sequences $(\nabla\psi_{h,\delta t})$, $(\xi(s_{h,\delta t})\nabla\psi_{h,\delta t})$, when $h, \delta t \to 0$, that

$$\mathcal{W}_2^{h,*} \longrightarrow \mathcal{W}_2 = \int_{Q_{\mathfrak{T}}} \Lambda\nabla(\rho(p)\xi(s)) \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t - \int_{Q_{\mathfrak{T}}} \xi(s)\Lambda\nabla\rho^\star \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t.$$

Expanding the first integral in $\mathcal{W}_2$ gives

$$\mathcal{W}_2 = \int_{Q_{\mathfrak{T}}} \rho(p)\Lambda\nabla\xi(s) \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} (\xi(s)\nabla\rho(p) - \xi(s)\nabla\rho^\star) \cdot \Lambda\nabla\psi \,\mathrm{d}x\,\mathrm{d}t.$$

Finally, integrate once more by parts the second integral in $\mathcal{W}_2$ to obtain

$$\begin{aligned}\int_{Q_{\mathfrak{T}}} \xi(s)(\nabla\rho(p) - \nabla\rho^\star) \cdot \Lambda\nabla\psi \,\mathrm{d}x\,\mathrm{d}t = &- \int_{Q_{\mathfrak{T}}} (\rho(p) - \rho^\star)\gamma(s)\nabla s \cdot \Lambda\nabla\psi \,\mathrm{d}x\,\mathrm{d}t \\ &- \int_{Q_{\mathfrak{T}}} (\rho(p) - \rho^\star)\xi(s)\,\mathrm{div}(\Lambda\nabla\psi) \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

The last two integrals vanish since $\rho(p)\gamma(s) = \rho^\star\gamma(s)$ and $\rho(p)\xi(s) = \rho^\star\xi(s)$ almost everywhere in $Q_{\mathfrak{T}}$. Consequently

$$\lim_{h,\delta t\to 0} \mathcal{W}_2^{h,*} = \mathcal{W}_2 = \int_{Q_{\mathfrak{T}}} \rho(p)\Lambda\nabla\xi(s) \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t.$$

What is left is to show that

$$\lim_{h,\delta t\to 0} |\mathcal{W}_2^h - \mathcal{W}_2^{h,*}| = 0. \tag{3.9.10}$$

To this end, we need to introduce the functions $\overline{p}_{h,\delta t}, \underline{p}_{h,\delta t}$

$$\overline{p}_T^{n+1} := \sup_{x\in T} p_h^{n+1}(x), \qquad \underline{p}_T^{n+1} := \inf_{x\in T} p_h^{n+1}(x) \tag{3.9.11}$$

$$\overline{p}_{h,\delta t_{|T\times(t^n,t^{n+1}]}} := \overline{p}_T^{n+1}, \quad \underline{p}_{h,\delta t|T\times(t^n,t^{n+1}]} := \underline{p}_T^{n+1}. \tag{3.9.12}$$

We define

$$\mathcal{V}_2^h = \int_{Q_{\mathfrak{T}}} \rho(\underline{p}_{h,\delta t})\Lambda\nabla\xi(s_{h,\delta t}) \cdot \nabla\psi_{h,\delta t} \,\mathrm{d}x\,\mathrm{d}t.$$

One observes that

$$\begin{aligned}\left|\mathcal{W}_2^h - \mathcal{W}_2^{h,*}\right| &\le \left|\mathcal{W}_2^h - \mathcal{V}_2^h\right| + \left|\mathcal{V}_2^h - \mathcal{W}_2^{h,*}\right|, \\ &\le 4\int_{Q_{\mathfrak{T}}} \left|\rho(\overline{p}_{h,\delta t}) - \rho(\underline{p}_{h,\delta t})\right| |\Lambda\nabla\xi(s_{h,\delta t}) \cdot \nabla\psi_{h,\delta t}| \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

In view of the Cauchy-Schwarz inequality and Lemma 3.5.3 we find

$$\begin{aligned}\left|\mathcal{W}_2^h - \mathcal{W}_2^{h,*}\right| &\leq \int_{Q_\mathfrak{T}} \left|\rho(\overline{p}_{h,\delta t}) - \rho(\underline{p}_{h,\delta t})\right| |\Lambda \nabla \xi(s_{h,\delta t}) \cdot \nabla \psi_{h,\delta t}| \,\mathrm{d}x\,\mathrm{d}t,\\ &\leq \overline{\Lambda} \left\|\rho'\right\|_\infty \|\nabla \psi\|_\infty \left\|\nabla \xi(s_{h,\delta t})\right\|_{L^2(Q_\mathfrak{T})^d} \Big( \sum_{n=0}^{N-1} \delta t \int_\Omega \left|\overline{p}_h^{n+1} - \underline{p}_h^{n+1}\right|^2 \mathrm{d}x\Big)^{1/2},\\ &\longrightarrow 0, \text{ as } h, \delta t \longrightarrow 0.\end{aligned}$$

Let us now establish that

$$\lim_{h,\delta t \to 0} \mathcal{W}_3^h = 0.$$

For this, let us define the coefficient $\overline{\gamma}_{KL}^{n+1}$

$$\overline{\gamma}_{KL}^{n+1} := \begin{cases} \dfrac{\xi(s_K^{n+1}) - \xi(s_L^{n+1})}{s_K^{n+1} - s_L^{n+1}}, & \text{if } s_K^{n+1} \neq s_L^{n+1} \\ \\ \gamma(s_K^{n+1}), & \text{if } s_K^{n+1} = s_L^{n+1} \end{cases}. \tag{3.9.13}$$

As a consequence $\mathcal{W}_3^h$ becomes

$$\mathcal{W}_3^h = \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \rho_{KL}^{n+1} \Lambda_{KL}^T \Big(\gamma_{KL}^{n+1} - \overline{\gamma}_{KL}^{n+1}\Big)\Big(s_K^{n+1} - s_L^{n+1}\Big)(\psi_K^{n+1} - \psi_L^{n+1}).$$

Using repeatedly the Cauchy-Schwarz inequality yields

$$\left|\mathcal{W}_3^h\right| \leq \rho_1 \left| \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\Lambda_{KL}^T\right| \left(s_K^{n+1} - s_L^{n+1}\right)^2 \right|^{\frac{1}{2}} \times \mathcal{X}_h^{\frac{1}{2}},$$

where

$$\mathcal{X}_h = \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\Lambda_{KL}^T\right| \left(\gamma_{KL}^{n+1} - \overline{\gamma}_{KL}^{n+1}\right)^2 (\psi_K^{n+1} - \psi_L^{n+1})^2. \tag{3.9.14}$$

We next introduce the fact that $\xi^{-1}$ is a $\theta$-Hölder, which yields

$$\left|s_K^{n+1} - s_L^{n+1}\right| \leq L_\xi \left|\xi(s_K^{n+1}) - \xi(s_L^{n+1})\right|^\theta.$$

According to this inequality together with (3.5.1), (3.5.2) and (3.6.7), there exists a positive constant $C$ so that

$$\left| \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\Lambda_{KL}^T\right| \left(s_K^{n+1} - s_L^{n+1}\right)^2 \right|^{\frac{1}{2}} \leq C.$$

Now the function $\gamma \circ \xi^{-1}$ is uniformly continuous on the compact $[0, \xi(1)]$. This ensures the existence of a modulus of continuity of this function, denoted by $\eta$ such that

$$\left|\gamma_{KL}^{n+1} - \overline{\gamma}_{KL}^{n+1}\right| \leq \eta\Big(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\Big), \quad \forall \sigma_{KL}^T,$$

where, for every $T \in \mathcal{T}_h$, we consider

$$\overline{\xi}_T^{n+1} = \xi(\overline{s}_T^{n+1}), \quad \underline{\xi}_T^{n+1} = \xi(\underline{s}_T^{n+1}),$$

and, for all $(x,t) \in T \times (t^n, t^{n+1})$, we define

$$\overline{s}_T^{n+1} := \sup_{x \in T} s_h^{n+1}(x), \qquad \underline{s}_T^{n+1} := \inf_{x \in T} s_h^{n+1}(x). \tag{3.9.15}$$

Consequently, the term $\mathcal{X}_h$ given in (3.9.14) satisfies

$$0 \leq \mathcal{X}_h \leq \mathcal{Y}_h,$$

with $\mathcal{Y}_h$ as written under the following form

$$\mathcal{Y}_h = \sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \left| \eta\Big(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\Big) \right|^2 \sum_{\sigma_{KL}^T \in \mathcal{E}_T} \left|\Lambda_{KL}^T\right| (\psi_K^{n+1} - \psi_L^{n+1})^2.$$

In view of Lemma 3.5.1 and the regularity of the function $\psi$, we claim that

$$0 \leq \mathcal{Y}_h \leq C \left| \eta\Big(\overline{\xi}_{h,dt} - \underline{\xi}_{h,dt}\Big) \right|^2,$$

where $C$ is a positive constant, which is independent of $h$ and $\delta t$. So, to conclude the proof of $\lim_{h \to 0} \mathcal{Y}_h = 0$, we require $\lim_{h \to 0} \Big(\overline{\xi}_{h,dt} - \underline{\xi}_{h,dt}\Big) = 0$ a.e. in $Q_{\mathfrak{T}}$. Indeed, we consider a generalization of Lemma 3.5.3 to get

$$\int_{Q_{\mathfrak{T}}} \left| \overline{\xi}_{h,dt} - \underline{\xi}_{h,dt} \right| \mathrm{d}x\, \mathrm{d}t \leq Ch \Big( \int_{Q_{\mathfrak{T}}} |\nabla \xi(s_{h,\delta t})|^2 \, \mathrm{d}x\, \mathrm{d}t \Big)^{\frac{1}{2}}.$$

Thereby, up to a subsequence, there holds

$$\lim_{h \to 0} \Big(\overline{\xi}_{h,dt} - \underline{\xi}_{h,dt}\Big) = 0 \quad \text{a.e. in } Q_{\mathfrak{T}}.$$

By the continuity of $\xi^{-1}$, we deduce

$$\lim_{h \to 0} \Big(\overline{s}_{h,\delta t} - \underline{s}_{h,\delta t}\Big) = 0 \quad \text{a.e. in } Q_{\mathfrak{T}}. \tag{3.9.16}$$

Consequently

$$\lim_{h,\delta t \to 0} \left| \mathcal{W}_3^h \right| = \lim_{h,\delta t \to 0} \mathcal{Y}_h = \lim_{h,\delta t \to 0} \mathcal{X}_h = 0.$$

Let us next study the convergence of the convective term $\mathcal{W}_4^h$. To this purpose, let us write $\mathcal{W}_4^h$ by edges

$$\mathcal{W}_4^h = -\sum_{n=0}^{N-1} \delta t \sum_{T \in \mathcal{T}} \sum_{\sigma_{KL}^T \in \mathcal{E}_T} M_T^{n+1} \rho_{KL}^{n+1} G_g\Big(s_K^{n+1}, s_L^{n+1}; \Lambda_{KL}^T \delta_{KL}^{n+1} p\Big) \Big(\psi_L^{n+1} - \psi_K^{n+1}\Big).$$

We additionally define

$$\mathcal{V}_4^h = \int_{Q_{\mathfrak{T}}} \rho(p_{h,\delta t}) M(s_{h,\delta t}) f_g(s_{h,\delta t}) \Lambda \nabla p_{h,\delta t} \cdot \nabla \psi_{h,\delta t} \, \mathrm{d}x\, \mathrm{d}t.$$

Thanks to (3.9.8) and the smoothness of the test function, the sequence $\Big(\rho(p_{h,\delta t})M(s_{h,\delta t})f_g(s_{h,\delta t})\nabla\psi_{h,\delta t}\Big)$ converges strongly to $\Big(\rho(p)M(s)f_g(s)\nabla\psi\Big)$ in $L^2(Q_{\mathfrak{T}})^d$. The sequence $(\nabla p_{h,\delta t})$ converges weakly to $\nabla p$ in $L^2(Q_{\mathfrak{T}})^d$. Then one gets

$$\lim_{h,\delta t\to 0}\mathcal{V}_4^h=\int_{Q_{\mathfrak{T}}}\rho(p)M(s)f_g(s)\Lambda\nabla p\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Define now

$$\mathcal{V}_4^{h,1}=\int_{Q_{\mathfrak{T}}}\rho(\underline{p}_{h,\delta t})M(s_{h,\delta t})f_g(s_{h,\delta t})\Lambda\nabla p_{h,\delta t}\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t,$$

where $\underline{p}_{h,\delta t}$ is given in (3.9.12). We show that $\mathcal{V}_4^h-\mathcal{V}_4^{h,1}\to 0$.

$$\begin{aligned}\Big|\mathcal{V}_4^h-\mathcal{V}_4^{h,1}\Big|&\le C\,\|M\|_\infty\,\overline{\Lambda}\,\big\|\rho'\big\|_\infty\,\|\nabla\psi\|_\infty\,\|\nabla p_{h,\delta t}\|_{L^2(Q_{\mathfrak{T}})^d}\Big(\sum_{n=0}^{N-1}\delta t\int_\Omega\Big|\overline{p}_h^{n+1}-\underline{p}_h^{n+1}\Big|^2\,\mathrm{d}x\Big)^{1/2},\\&\le C'h\longrightarrow 0,\ \text{ as } h,\delta t\longrightarrow 0.\end{aligned}$$

We continue in this fashion to define $\mathcal{W}_4^{h,*}$

$$\mathcal{W}_4^{h,*}=\int_{Q_{\mathfrak{T}}}\rho(\underline{p}_{h,\delta t})M(\underline{s}_{h,\delta t})f_g(\underline{s}_{h,\delta t})\Lambda\nabla p_{h,\delta t}\cdot\nabla\psi_{h,\delta t}\,\mathrm{d}x\,\mathrm{d}t,$$

where $\underline{s}_{h,\delta t}$ is defined in (3.9.15). Moreover, we show that

$$\mathcal{V}_4^{h,1}-\mathcal{W}_4^{h,*}\to 0. \tag{3.9.17}$$

Using the Cauchy-Schwarz inequality and the strong convergence (3.9.16), we have

$$\Big|\mathcal{V}_4^{h,1}-\mathcal{W}_4^{h,*}\Big|\le C''\Big(\sum_{n=0}^{N-1}\delta t\int_\Omega\big|\overline{s}_{h,\delta t}-\underline{s}_{h,\delta t}\big|^2\,\mathrm{d}x\Big)^{1/2},\longrightarrow 0,\ \text{ as } h,\delta t\longrightarrow 0.$$

It remains to establish that the sequence $(\mathcal{W}_4^h-\mathcal{W}_4^{h,*})$ goes to zero as $h,\delta t$ tend to zero. To this end, we use the fact that the gas fractional flow, the total mobility and the density are bounded functions together with the consistency and the Lipschitz continuity of the numerical flux $G_g$. To be more precise, we compute

$$\begin{aligned}&\Big|\rho_{KL}^{n+1}M_T^{n+1}G_g\Big(s_K^{n+1},s_L^{n+1};\Lambda_{KL}^T\delta_{KL}^{n+1}p\Big)-\Big(-\rho_{K,0}^{n+1}M(s_{K,0}^{n+1})f_g(s_{K,0}^{n+1})\Lambda_{KL}^T\delta_{KL}^{n+1}p\Big)\Big|\,\big|\delta_{KL}^{n+1}\psi\big|,\\&=\Big|\rho_{KL}^{n+1}M_T^{n+1}G_g\Big(s_K^{n+1},s_L^{n+1};\Lambda_{KL}^T\delta_{KL}^{n+1}p\Big)-\rho_{K,0}^{n+1}M(s_{K,0}^{n+1})G_g\Big(s_{K,0}^{n+1},s_{K,0}^{n+1};\Lambda_{KL}^T\delta_{KL}^{n+1}p\Big)\Big|\,\big|\delta_{KL}^{n+1}\psi\big|,\\&\le C\Big(\eta(\Big|s_K^{n+1}-s_{K,0}^{n+1}\Big|)+\Big|\rho_{KL}^{n+1}-\rho_{K,0}^{n+1}\Big|+\Big|M_T^{n+1}-M(s_{K,0}^{n+1})\Big|\Big)\,\big|\Lambda_{KL}^T\big|\,\big|\delta_{KL}^{n+1}p\big|\,\big|\delta_{KL}^{n+1}\psi\big|,\\&\le C\Big(\eta(\Big|s_K^{n+1}-s_{K,0}^{n+1}\Big|)+\Big|\rho_{KL}^{n+1}-\rho_{K,0}^{n+1}\Big|+\Big|M_T^{n+1}-M(s_{K,0}^{n+1})\Big|\Big)\\&\qquad\times\Big(\big|\Lambda_{KL}^T\big|\,\big|\delta_{KL}^{n+1}p\big|^2+\big|\Lambda_{KL}^T\big|\,\big|\delta_{KL}^{n+1}\psi\big|^2\Big),\end{aligned}$$

where $\eta(\cdot)$ is a modulus of continuity. The last inequality and Lemma 3.5.1 affirm that

$$\begin{aligned}\left|\mathcal{W}_4^h - \mathcal{W}_4^{h,*}\right| \leq & C \sum_{T\in\mathcal{T}} \left(\eta(\left|\overline{s}_T^{n+1} - \underline{s}_T^{n+1}\right|) + \left|\overline{\rho}_T^{n+1} - \underline{\rho}_T^{n+1}\right| + \left|M(\overline{s}_T^{n+1}) - M(\underline{s}_T^{n+1})\right|\right) \\ & \times \sum_{\sigma_{KL}^T\in\mathcal{E}_T} \left(\left|\Lambda_{KL}^T\right| \left|\delta_{KL}^{n+1} p\right|^2 + \left|\Lambda_{KL}^T\right| \left|\delta_{KL}^{n+1}\psi\right|^2\right), \\ & \leq C \int_{Q_{\mathfrak{T}}} \left(\eta(\left|\overline{s}_{h,\delta t} - \underline{s}_{h,\delta t}\right|) + \left|\rho(\overline{p}_{h,\delta t}) - \rho(\underline{p}_{h,\delta t})\right| + \left|M(\overline{s}_{h,\delta t}) - M(\underline{s}_{h,\delta t})\right|\right) \mathrm{d}x\,\mathrm{d}t,\end{aligned}$$

As a consequence of the convergence (3.9.16) and Lebesgue's dominated convergence theorem, it follows that the first and the third integrals on the right hand side go to zero as $h, \delta t$ tend to zero. Using again that the derivative of the density is bounded, Lemma 3.5.3 and the uniform estimate on the global pressure (3.6.6), the second integral on the right hand side goes to zero too. We hence obtain

$$\lim_{h,\delta t\to 0} \left|\mathcal{W}_4^h - \mathcal{W}_4^{h,*}\right| = 0. \tag{3.9.18}$$

Therefore

$$\lim_{h,\delta t\to 0} \mathcal{W}_4^h = \int_{Q_{\mathfrak{T}}} \rho(p) M(s) f_g(s) \Lambda \nabla p \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t.$$

Finally, in order to pass to the limit in $\mathcal{W}_5^h$, we make use of the result (3.9.8) and Lebesgue's dominated convergence theorem to attest that

$$\lim_{h,\delta t\to 0} \mathcal{W}_5^h = \lim_{h,\delta t\to 0} \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathcal{V}} |\omega_K| \rho(p_K^{n+1}) s_K^{n+1} q_{P,K}^{n+1} \psi_K^{n+1} = \int_{Q_{\mathfrak{T}}} \rho(p) s q^P \psi \,\mathrm{d}x\,\mathrm{d}t,$$

as required. □

## 3.10 Numerical experiments

Here we provide some numerical tests in two space dimensions so that we can show the robustness and the stability of the proposed numerical scheme. More precisely, we are interested in the secondary recovery of gas by injecting water. In addition, we consider an anisotropic permeability tensor to illustrate its impact on the displacement of the fluids.

The domain of our study is $\Omega = [0,1]^2$, then the length and the width of the medium are $L_x = Ly = 1m$. Next we perform a primal mesh, which is a triangulation in the sense of the finite element discretization, and a barycentric dual mesh constructed as described in Section 3.3. This mesh consists of 3584 elements and 1857 vertices as depicted in Fig. 3.2. We emphasize that the triangle angles are acute, which allows us to take into account the isotropic case, where the stiffness coefficients are positive. Nevertheless, whenever the permeability tensor is not the identity matrix, this property is no longer valid. Without loss of generality, other triangulations can be suggested.

For these simulations, we require some physical data. For this, we consider the test case of the work [23] where the authors implemented a two-point flux approximation scheme and considered an isotropic tensor. The porosity is then set to $\phi = 0.206$. We recall that $s = s_g$. The relative permeabilities and the capillary pressure are respectively given by: $K_{rg} = s^2, K_{rw} = (1-s)^2, p_c(s) =$

$P_{max}s$, with $P_{max} = 1.013 \times 10^5 Pa$. The viscosities of the two phases are: $\mu_w = 10^{-3} Pa.s, \mu_g = 9 \times 10^{-5} Pa.s$. The gas density is chosen as follows: $\rho(p) = \rho_r(1+c_r(p-p_r))$ with $\rho_r = 400 Kg.m^{-3}$, $c_r = 10^{-6} Pa$, $p_r = 1.013 \times 10^5 Pa$. We pick out the absolute permeability as

$$\Lambda = 0.15 \times 10^{-10} \begin{bmatrix} 1 & 0 \\ 0 & \lambda \end{bmatrix} [m^2],$$

where $\lambda$ is a parameter in $[0, 1]$. Besides, we present three case tests with $\lambda \in \{1, 0.1, 0.001\}$.

The gas saturation and gas pressure are initialized as follows: $s_g(x, 0) = 0.9$, $p_g(x, 0) = 1,013 \times 10^5 Pa$. Next, water is injected on the left zone ($x = 0$, $0.8 \leq y \leq 1$) of the medium with a constant saturation $s_w^l = 0.9$, meaning that $s_g^l = 0.1$ (see Fig. 3.2), and with a maintaining pressure $P_g^l = 4.6732 \times 10^5 Pa$. The extraction zone ($x = 1$,$0 \leq y \leq 0.2$) is in contact with the air. Therefore, in this region, the pressure is $P_g^r = 1,013 \times 10^5 Pa$ and a free flow of the fluids is considered. What remains of the boundary is impermeable. We furthermore have no source terms; that is $q^P = q^I = 0$.

The implemented CVFE scheme provides a nonlinear algebraic system. In order to solve it, we apply the Newton-Raphson method. Moreover, we take $\varepsilon = 10^{-10}$ as a stopping criterion. The final time is set to $t_f = 40s$ for all the tests. The time step is chosen to be $\delta t = 0.05$ for $\lambda = 1, 0.1$ and $\delta t = 0.005$ for $\lambda = 0.001$. We present four numerical tests. The three first ones are devoted to investigating the influence of the anisotropy on the compressible flow within the domain. The last one compares the difference between the compressible and incompressible flows.

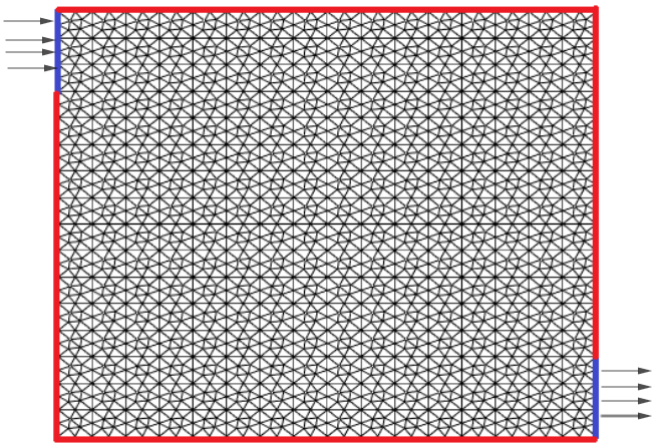
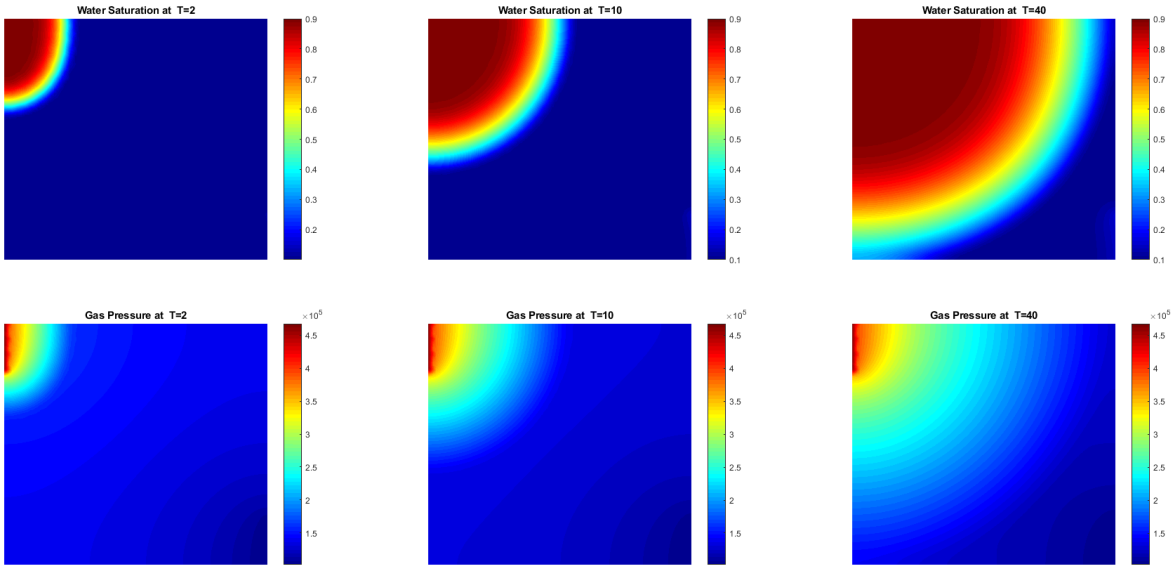


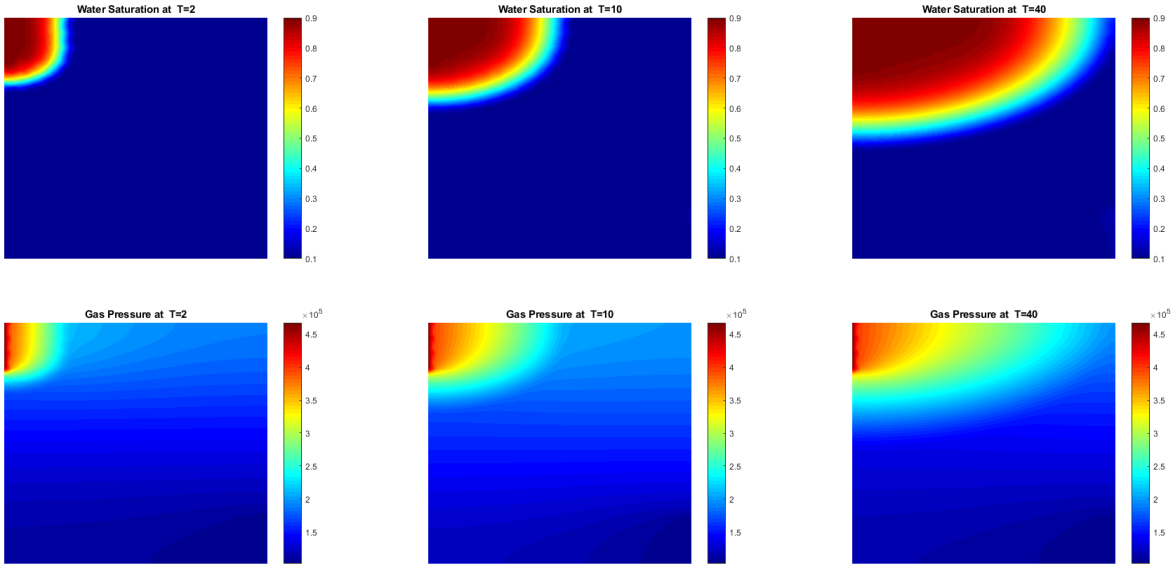
Figure 3.2: Primal mesh with 3584 triangles and 1857 vertices.

### 3.10.1 First test $\lambda = 1$




In the first test, we illustrate the behavior of water saturation (top) and the gas pressure (bottom) through an isotropic medium for different times $t_f = 2s, 10s, 40s$. We then recall that the transmissibility coefficients are nonnegative. We see that the discrete saturation remains in the interval $[0, 1]$ as we have established in Lemma 3.6.1. On one hand we observe a remarkable displacement of a front between the two fluids toward the right zone where the pressure is lower. On the other hand, we notice important diffusive effects on all these figures, which are due to the capillary term.

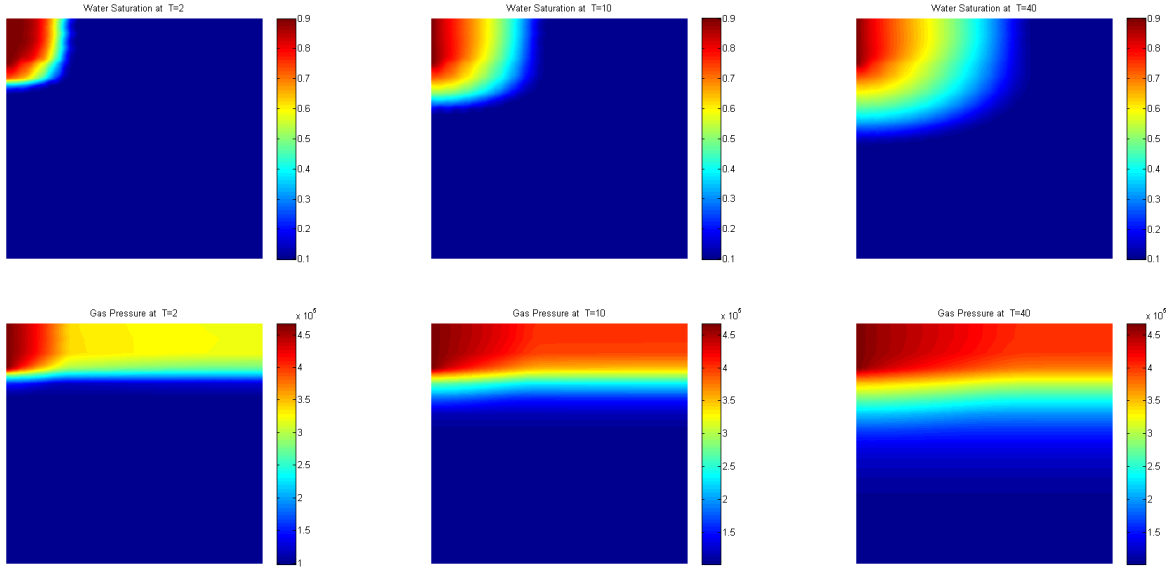
### 3.10.2 Second test $\lambda = 0.1$




In the second test, we consider a weak anisotropy with $\lambda = 0.1$. We then show the influence of this anisotropy on the flow of water through the medium. Contrary to the first test, some stiffness coefficients are nonpositive. However, the physical ranges of the computed saturation are respected

as claimed in Lemma 3.6.1. In addition, we record an important flow of the water from left to right and this is natural since the permeability is much bigger in this direction.

### 3.10.3 Third test $\lambda = 0.001$




In the third simulation, the anisotropy ratio is too large compared to the previous tests. Then some of the transmissibility coefficients are necessarily nonpositive. As noticed before, the water pushes the gas in the $x$-direction. The displacement of the two fluids is very slow since the pores are too tiny in the $y$-direction. We also observe small undershoots on the saturation, which may be caused by the effect of anisotropy together with the Newton solver.

### 3.10.4 Fourth test: comparison between compressible and incompressible flows

In this test we compare the incompressible flow with various compressible flows in the absence of the capillary effects. The capillary pressure is neglected in order to illustrate only the impact of the compressibility of the gas. We finally display in Fig. 3.3–3.5 the evolution of water saturation and gas pressure at three points of the medium $\Omega$. We here consider an identical permeability i.e. $\lambda = 1$ and $c_r \in \{0; 5 \times 10^{-6}; 5 \times 10^{-5}; 5 \times 10^{-4}\}[Pa]$. Even if the flow is slightly compressible, we remark that the velocity of the water through the domain is relatively slow. In the incompressible case the flow is independent of the initial pressure whereas it plays a major role for the compressible flow. As we observe in Fig. 3.3–3.5, there is a significant difference in terms of pressures in the first stage of the evolution.

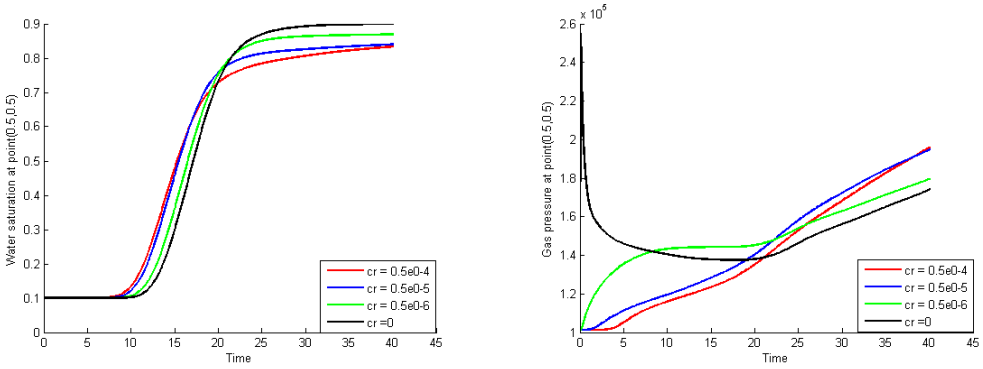



Figure 3.3: Evolution of water saturation (left) and gas pressure (right) at point (0.5,0.5).

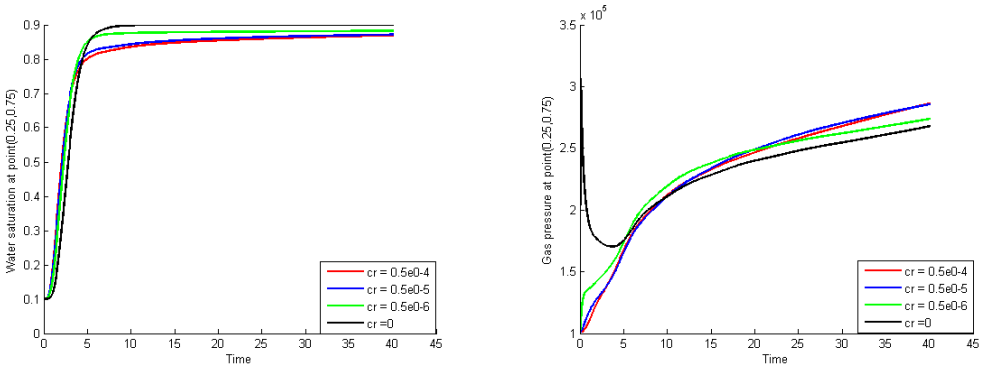



Figure 3.4: Evolution of water saturation (left) and gas pressure (right) at point (0.25,0.75).

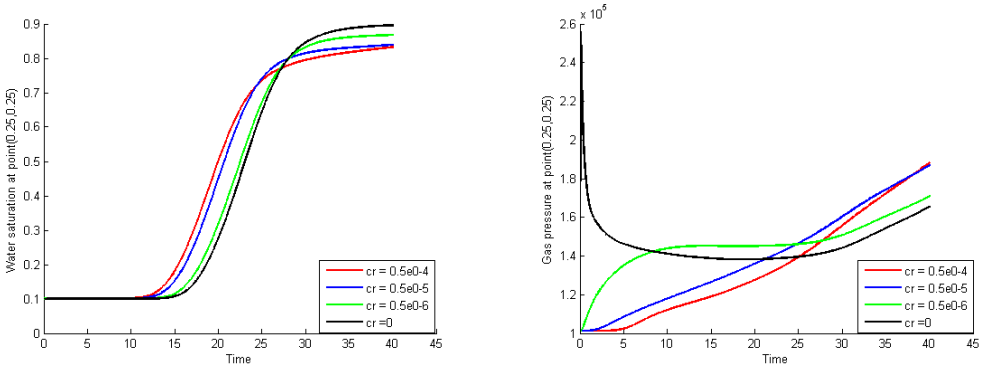



Figure 3.5: Evolution of water saturation (left) and gas pressure (right) at point (0.25,0.25).

# Chapter 4

# Convergence of a monotone nonlinear DDFV scheme for degenerate parabolic equations

In this chapter, we carry out the convergence analysis of a monotone DDFV method for approximating solutions of degenerate parabolic equations. The basic idea rests upon different approximations of the fluxes on the same interface of the control volume. Precisely, the approximated flux is split into two terms corresponding to the primal and dual normal components. Then the first term is discretized using a centered scheme whereas the second one is approximated in a non evident way by an upstream scheme. The novelty of our approach is twofold: on the one hand we prove that the resulting scheme preserves the positivity and on the other hand we establish energy estimates. Some numerical tests are then presented and they show that the scheme in question turns out to be robust and efficient with an accuracy of second order.

## 4.1 Problem statement

Nonlinear degenerate parabolic equations are the main core to study some complex problems arising, for instance, from petroleum engineering, hydrology and biology. Hence, seeking analytical or approximate solutions of these equations is of an immense advantage. Throughout this chapter, we will be interested in approximating, thanks to a new finite volume scheme, the solution to the academic problem:

$$\begin{cases} \partial_t u - \nabla \cdot \Big( f(u) \Lambda \nabla u \Big) = 0 & \text{in} \quad Q_{\mathfrak{T}} := \Omega \times (0, \mathfrak{T}) \\ u = 0 & \text{on} \quad \partial\Omega \times (0, \mathfrak{T}), \\ u(\cdot, 0) = u^0 & \text{in} \quad \Omega \end{cases} \tag{4.1.1}$$

where $\Omega$ is a bounded polygonal open of $\mathbb{R}^d$, $\mathfrak{T}$ a fixed positive number, $\partial\Omega$ the boundary of $\Omega$, $\Lambda$ a given $d$-square matrix (tensor) and $f$ a given nonnegative function. In the context of porous media flows, the function $f$ is usually called the mobility while the tensor $\Lambda$ stands for the permeability. More precisely, the problem (4.1.1) describes the infiltration of a single fluid through a porous medium with no gravity effects [47]. It is derived from the Darcy law together with the mass conservation equation. On the other hand, this problem is known under the name of the porous medium equation [117] whenever $f(u) = u^m$, for some nonnegative real number $m$. In view of the theoretical study, the elliptic term of (4.1.1) can be formulated otherwise by introducing the

so-called Kirchhoff transformation $F$. With some general assumptions on $F$, this formulation is sometimes said to be the simplified Stefan problem [66], which is used to model free boundary value problems. Even if this function seems to have no physical interpretation, it will play a remarkable role to carry out the analysis of the scheme we consider here. It is then defined by

$$F(u) = \int_0^u f(s)\,\mathrm{d}s, \qquad \forall u \in \mathbb{R}. \tag{4.1.2}$$

In this discretization, we will also introduce the semi-Kirchhoff transform denoted by $\xi$ and defined as

$$\xi(u) = \int_0^u \sqrt{f(s)}\,\mathrm{d}s, \qquad \forall u \in \mathbb{R}. \tag{4.1.3}$$

Different approximations, with various assumptions on the data, have been conducted to discretize problems involving nonlinear diffusion equations of type (4.1.1). For an upstream finite difference method, we cite the work [94]. Concerning finite volume schemes, we refer to this battery of contributions [10, 14, 20, 23, 25, 61, 71, 70, 79, 99]. Plenty of these discretizations stipulate restrictive constraints, especially an orthogonality condition on the mesh in the sense of Eymard *et al.* [69], which excludes a large variety of interesting meshes. For example, in Hydrology, most geological layers are quite deformed thus the meshes used to discretize the field are somehow distorted. In this case, the orthogonality condition can not be satisfied for most of the edges. In addition, in the presence of anisotropic media, we may encounter the same difficulties. Yet, some works have combined finite volume and finite element methods [1, 20, 63, 73, 93, 105]. Carrying out the analysis of these schemes, the authors required a positivity assumption on the stiffness coefficients that does not hold for all sort of meshes. To overcome this issue, positive schemes with their convergence studies have been proposed in [40, 42]. More generally, a gradient scheme [62, 66] has been suggested to discretize the Stefan problem, which is an equivalent formulation of (4.1.1) using the Kirchoff transform. The Gradient schemes framewok encompasses a lot of popular discretizations, but it may produce undershoots and overshoots in general. There is no hope of proving such bounds without further assumptions.

In this chapter, we are concerned with the Discrete Duality Finite Volume (DDFV) method for the discretization of the problem (4.1.1). This method belongs to the gradient schemes family and is viewed as a particular class of the finite volume methods. It has been first introduced for the Laplace equation in [89, 90]. It has been also proved to be equivalent to a nonconforming finite element approach in [60]. In two dimensions, the convergence analysis of the DDFV scheme is carried out later on for many types of partial differential equations of second order in several works [11, 39, 46, 59, 60]. Such results have been extended to 3D in [8, 9, 56, 97]. The strength of this discretization consists of producing a consistent discrete whole gradient on almost general grids and any tensor. This is of a great importance since most of the meshes coming from physics are somehow distorted. On the other hand, the reconstruction gradient operator verifies the discrete Stokes formula, which is a powerful tool to analyze such a scheme. Moreover, the DDFV method is unconditionally coercive, which ensures the stability of the scheme.

Practically (cf. FVCA5 benchmark) [88] the DDFV schemes fail to satisfy an explicit discrete maximum principle. This property is crucial whenever we deal with positive physical quantities by their nature like saturation and concentration. As to be more precise, let us consider the DDFV discretization of the linear diffusion equation $-\Delta u = f$ with Dirichlet boundary conditions. Formally, it yields a stiffness matrix which is not monotone in the case of non admissible meshes in the sense of Eymard *et al.* [69]. By choosing an appropriate positive source term, we can

acquire a solution with some negative values. In general, the monotonicity of the DDFV scheme has been a drawback for the method since it has appeared. However, in the work of [37] the authors were able to design a monotone nonlinear DDFV scheme for the diffusion equation. It basically rests upon the DDFV idea together with a nonlinear monotone two-point finite volume method as investigated in [86, 102, 122]. Unfortunately, there is no convergence proof of the numerical schemes proposed in [37, 86, 102, 122] since they suffer from the lack of coercivity as pointed out in [38, 61]. Recently, in [39] the authors have employed a nonlinear technique to establish the nonnegativity of the approximate solution in the case of a linear drift equation enclosed with Neumann boundary conditions. Then, the contribution of our appraoch is to propose a new scheme that fulfills the physical ranges of the discrete solution even on almost general meshes and for possibly anisotropic tensors. Given an interface of a control volume (primal or dual), the key point of our approach consists of approximating the flux across this interface with a TPFA (Two-Point Flux Approximation) scheme with respect to the unit normal to the same interface and an upwind scheme with respect to the corresponding dual interface. This technique is not standard in the framework of DDFV methods, and it gathers the main ingredients to conduct the convergence analysis. From a practical perspective our scheme yields surprising results with optimal convergence rates.

We have chosen to introduce the proposed scheme for degenerate diffusive equations involving homogeneous Dirichlet boundary conditions. The only reason behind the choice of the model problem is the ease readability of our scheme. This approach can be easily extended to more general boundary conditions as done in [13, 60] as well as to models including convective and source terms [39, 46, 57]. Indeed, the convective term does not provide any supplementary difficulties, since it can be approximated using adequate upstream approaches in order to ensure the discrete maximum principle and get the main elements for the convergence analysis.

The remainder of this chapter is structured as follows. In Section 4.2, we give the DDFV discretization, some related notations and definitions of discrete operators. In Section 4.3, we sketch out how to derive the proposed DDFV scheme. In Section 4.4, we prove that this scheme preserves the physical ranges of the approximate solution and we derive some energy estimates on the discrete gradients. In Section 4.5, we establish that the nonlinear algebraic system has a solution using a monotony criterion. In Section 4.6, we state some compactness properties and we apply Kolmogorov's theorem to ensure the existence of a convergent subsequence of a family of discrete solutions. In Section 4.7, we demonstrate that this subsequence tends towards the weak solution of the continuous problem. In Section 4.8, we exhibit some numerical results to show the efficiency and robustness of our scheme.

Let us now formulate the main assumptions on the data.

($A_1$) The initial condition $u^0$ is assumed to be in $L^\infty(\Omega)$ with $0 \leq u \leq 1$.

($A_2$) The function $f$ belongs to $\mathcal{C}^0([0,1],\mathbb{R})$ with

$$\begin{cases} f(u) > 0, & \text{for all } u \in (0,1), \\ f(u) = 0, & \text{for all } u \in \mathbb{R}\backslash(0,1). \end{cases}$$

As a consequence, $F$ and $\xi$ are Lipschitz continuous nondecreasing functions. We also assume that $v := \sqrt{f}$ is absolutely continuous. This latter regularity on $v$ is required so that the Engquist-Osher scheme, to be presented later (Section 4.3), can be defined.

($A_3$) The tensor $\Lambda : \Omega \to \mathcal{S}_d(\mathbb{R})$, where $\mathcal{S}_d(\mathbb{R})$ is the space of $d$-square symmetric matrices, is assumed to be in $L^\infty(\Omega)^{d\times d}$ and verifies the uniform ellipticity condition

$$\underline{\Lambda}\,|\zeta|^2 \leq \Lambda(x)\zeta\cdot\zeta \leq \overline{\Lambda}\;|\zeta|^2\,,\ \text{for all}\ \ \zeta\in\mathbb{R}^d \text{ and a.e.}\ \ x\in\Omega,$$

for some positive constants $\underline{\Lambda}$ and $\overline{\Lambda}$.

We next define the natural space $L^2(0,\mathfrak{T};H_0^1(\Omega))$ where the solution to the problem (4.1.1) will be sought

$$H_0^1(\Omega)=\{v\in H^1(\Omega)\ /\ v=0 \text{ on } \partial\Omega\}.$$

Moreover, $H_0^1(\Omega)$ is a Hilbert space endowed with the norm

$$||v||_{H_0^1(\Omega)}=||\nabla v||_{(L^2(\Omega))^d}.$$

This leads us to the definition of the weak solution.

**Definition 4.1.1.** *(Weak solution) A measurable function $u : Q_{\mathfrak{T}} \longrightarrow [0,1]$ is called a weak solution of the problem (4.1.1) provided*

$$\xi(u)\in L^2(0,\mathfrak{T};H_0^1(\Omega)),$$
$$-\int_{Q_{\mathfrak{T}}} u\,\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t+\int_{Q_{\mathfrak{T}}}\Lambda\nabla F(u)\cdot\nabla\varphi\,\mathrm{d}x\,\mathrm{d}t-\int_\Omega u^0\ \varphi(\cdot,0)\,\mathrm{d}x=0,\quad\forall\varphi\in\mathcal{C}_c^\infty(\Omega\times[0,\mathfrak{T})).$$

The existence of a weak solution to the problem (4.1.1) has been investigated in [2]. The uniqueness proof is already addressed in [81].

## 4.2 DDFV discretization

For the simplicity of the exposition, we follow most of the notations given in the works [11, 46]. From now on, we focus only on the two dimensions (in space) case.

### 4.2.1 Meshes and notations

A DDFV discretization requires three kinds of meshes, a primal mesh, dual mesh and diamond mesh. The primal mesh is denoted by $\overline{\mathfrak{M}}=\mathfrak{M}\cup\partial\mathfrak{M}$, where $\mathfrak{M}$ is a partition of $\Omega$ with polygonal open disjoint subsets usually called control volumes and $\partial\mathfrak{M}$ is the set of boundary edges viewed as degenerate control volumes. These primal grids are not necessarily convex. For every $K\in\overline{\mathfrak{M}}$, the center of gravity of $K$ is denoted by $x_K$. We further define $\mathcal{V}$ as the family of these centers.

We designate by $\mathcal{V}^*$ the set of all the vertices of the mesh $\mathfrak{M}$. It is composed of inner vertices $\mathcal{V}^*_{int}$ and boundary ones $\mathcal{V}^*_{ext}$. For each $x_{K^*}\in\mathcal{V}^*_{int}$ (resp. $x_{K^*}\in\mathcal{V}^*_{ext}$), we associate a unique dual control volume $K^*$ which is a polygon whose vertices are given by the set $\{x_K\in\mathcal{V}/x_{K^*}\in\overline{K},\ K\in\mathfrak{M}\}$ (resp.$\{x_{K^*}\}\cup\{x_K\in\mathcal{V}/x_{K^*}\in\overline{K},\ K\in\partial\mathfrak{M}\}$ ). With these dual sub-domains, we construct the dual mesh denoted by $\overline{\mathfrak{M}^*}=\mathfrak{M}^*\cup\partial\mathfrak{M}^*$ (see Fig. 4.1).

By $\mathcal{E}$ (resp. $\mathcal{E}^*$) we mean the set of all the edges of $\overline{\mathfrak{M}}$ (resp. $\overline{\mathfrak{M}^*}$). Two cells are said to be neighbors if they share at least one edge. To be more precise, for every couple of neighboring primal (resp. dual) control volumes $K$ and $L$ (resp. $K^*$ and $L^*$ ), there exists $\sigma\in\mathcal{E}$ (resp. $\sigma^*\in\mathcal{E}^*$) such that $\sigma=\overline{K}\cap\overline{L}$ (resp. $\sigma^*=\overline{K^*}\cap\overline{L^*}$).

The diamond mesh $\mathfrak{D} = (\mathcal{D}_{\sigma,\sigma^*})_{(\sigma,\sigma^*)\in\mathcal{E}\times\mathcal{E}^*}$ is also a partition of $\Omega$ by diamond cells. For every primal edge $\sigma$ with $\sigma \not\subseteq \partial\overline{\Omega}$, the subset $\mathcal{D}_{\sigma,\sigma^*}$ is a quadrilateral constructed by connecting the endpoints of $\sigma$ and $\sigma^*$. In the case where $\sigma \in \mathcal{E} \cap \partial\overline{\Omega}$, this quadrilateral $\mathcal{D}_{\sigma,\sigma^*}$ is nothing more than a triangle as depicted in Fig. 4.2.

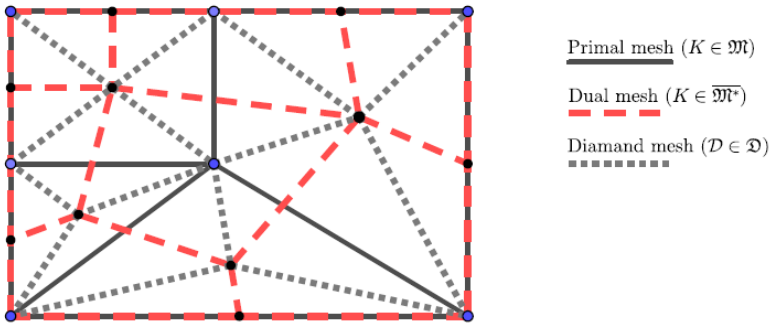



Figure 4.1: Illustration of the DDFV meshes.

The DDFV mesh is then given by the union of $\mathcal{T} = (\mathfrak{M}, \overline{\mathfrak{M}^*})$ and $\mathfrak{D}$. For every $M \in \mathcal{T}$ (primal or dual cell), the notation $m_M$ represents the measure of $M$, $\mathcal{E}_M$ contains all the edges of $M$, $\mathcal{D}_M$ is made of all the diamonds $\mathcal{D}_{\sigma,\sigma^*}$ such that $m(\mathcal{D}_{\sigma,\sigma^*} \cap M) > 0$, and $d_M$ refers to the diameter of $M$. For each $\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}$, the vertices of $\mathcal{D}_{\sigma,\sigma^*}$ are the extremities of both $\sigma$ and $\sigma^*$ i.e. $(x_K, x_{K^*}, x_L, x_{L^*})$. The center $x_\mathcal{D}$ of $\mathcal{D}_{\sigma,\sigma^*} =: \mathcal{D}$ is defined as the intersection of its main diagonals. $m_\mathcal{D}$ stands for the measure of $\mathcal{D}$, $d_\mathcal{D}$ its diameter, and $\alpha_\mathcal{D}$ is the angle between $(x_K, x_L)$ and $(x_{K^*}, x_{L^*})$. For every edge $e \in \mathcal{E} \cup \mathcal{E}^*$, we define $m_e$ as its measure. By $\mathbf{n}_{\sigma K}$ (resp. $\mathbf{n}_{\sigma^* K^*}$) we mean the unit normal to $\sigma$ (resp. $\sigma^*$) outwards $K$ (resp. $K^*$). Similarly, $\boldsymbol{\tau}_{K,L}$ (resp. $\boldsymbol{\tau}_{K^*,L^*}$) is the unit tangent vector to $\sigma$ (resp. $\sigma^*$) oriented from $K$ (resp. $K^*$) to $L$ (resp. $L^*$).

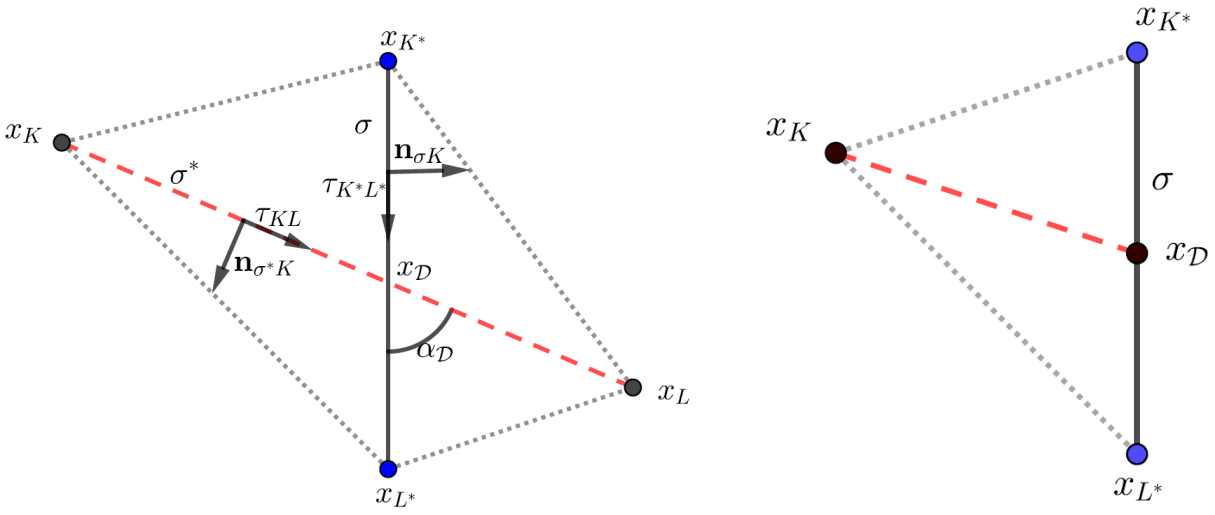



Figure 4.2: Interior (left) and boundary (right) diamond cells.

Now, we define the regularity of the DDFV mesh that determines how flat the diamond cells are. It also provides information about the difference between the size of a primal (resp. dual) control

volume and the size of a diamond cell whenever their intersection is nonempty. This regularity must be controlled for any sequence of meshes in order to perform the convergence analysis of the scheme. Let us denote $h_{\mathfrak{D}}$ the largest diameter of the diamond cells, $\alpha_{\mathcal{T}}$ the unique real number in $]0, \frac{\pi}{2}]$ such that

$$\sin(\alpha_{\mathcal{T}}) := \min_{\mathcal{D}\in\mathfrak{D}} |\sin(\alpha_{\mathcal{D}})|,$$

and $\rho_K$ (resp. $\rho_{K^*}$) the radius of the biggest inscribed ball in $K$ (resp. $K^*$) whose center is $x_K$ (resp. $x_{K^*}$). Then, the regularity of the mesh is defined by

$$\mathrm{reg}(\mathcal{T}) = \max\left(\frac{1}{\sin(\alpha_{\mathcal{T}})}, \max_{\mathcal{D}\in\mathfrak{D}} \frac{h_{\mathfrak{D}}}{\sqrt{m_{\mathcal{D}}}}, \max_{K\in\mathfrak{M}} \frac{d_K}{\sqrt{m_K}}, \max_{K^*\in\mathfrak{M}^*} \frac{d_{K^*}}{\sqrt{m_{K^*}}}, \right.$$
$$\left. \max_{K\in\mathfrak{M}} \Big(\frac{d_K}{\rho_K} + \frac{\rho_K}{d_K}\Big), \max_{K^*\in\overline{\mathfrak{M}^*}} \Big(\frac{d_{K^*}}{\rho_{K^*}} + \frac{\rho_{K^*}}{d_{K^*}}\Big)\right)$$

It follows from this relation that there exists a positive constant $C$ depending only on $\mathrm{reg}(\mathcal{T})$ such that

$$m_\sigma m_{\sigma^*} \le C m_K, \quad m_{\sigma^*}^2 \le C m_{\mathcal{D}}, \quad m_\sigma^2 \le C m_{\mathcal{D}}, \quad m_\sigma m_{\sigma^*} \le C m_{\mathcal{D}}.$$

A time discretization of the interval $(0, \mathfrak{T})$ is given by an increasing sequence of real numbers $(t^n)_{n=0,\dots,N}$ such that

$$t^0 = 0 < t^1 < \cdots < t^N = \mathfrak{T}.$$

For every $n \in \{0, \dots, N-1\}$, we denote $\delta t^n = t^{n+1} - t^n$ and we define $\delta t = \max_{0\le n\le N-1} \delta t^n$. To avoid heavy notations, we assume that the time step $\delta t^n$ is uniform. Then $\delta t = \delta t^n$, for all $n \in \{0, \dots, N-1\}$.

### 4.2.2 Discrete operators

We now survey the discrete version of the unknowns and operators that will allow us to define the nonlinear DDFV discretization for the problem (4.1.1). To begin with, let us specify the structure of the space $\mathbb{R}^{\#\mathcal{T}}$. Any vector $u_{\mathcal{T}}$ of this space is written under the form

$$u_{\mathcal{T}} = \Big((u_K)_{K\in\mathfrak{M}}, (u_{K^*})_{K^*\in\overline{\mathfrak{M}^*}}\Big).$$

Next, $\mathbb{R}^{\#\mathcal{T}}$ is endowed by following scalar product

$$[\![u_{\mathcal{T}}, v_{\mathcal{T}}]\!]_{\mathcal{T}} = \frac{1}{2}\Big(\sum_{K\in\mathfrak{M}} m_K u_K v_K + \sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*} u_{K^*} v_{K^*}\Big), \qquad \forall u_{\mathcal{T}}, v_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}.$$

Additionally, the set $(\mathbb{R}^2)^{\#\mathfrak{D}}$ represents the space of vector fields of the form $\zeta_{\mathfrak{D}} = (\zeta_{\mathcal{D}})_{\mathcal{D}\in\mathfrak{D}}$ whose components are constant on the diamond cells. This space is endowed by the inner product $\Big(\cdot,\cdot\Big)_{\mathfrak{D},\Lambda}$ defined as

$$\Big(\zeta_{\mathfrak{D}}, \varphi_{\mathfrak{D}}\Big)_{\mathfrak{D},\Lambda} = \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \zeta_{\mathcal{D}} \cdot \Lambda^{\mathcal{D}} \varphi_{\mathcal{D}}, \quad \forall \zeta_{\mathfrak{D}}, \varphi_{\mathfrak{D}} \in (\mathbb{R}^2)^{\#\mathfrak{D}},$$

where

$$\Lambda^{\mathcal{D}} = \frac{1}{m_{\mathcal{D}}} \int_{\mathcal{D}} \Lambda(x)\,\mathrm{d}x, \quad \forall \mathcal{D} \in \mathfrak{D}.$$

**Discrete gradient**

In the framework of the DDFV method, the discrete gradient operator denoted $\nabla^{\mathfrak{D}}$ is a linear mapping from $\mathbb{R}^{\#\mathcal{T}}$ to $(\mathbb{R}^2)^{\#\mathfrak{D}}$. It is defined for every $u_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}$ by

$$\nabla^{\mathfrak{D}} u_{\mathcal{T}} = \Big(\nabla^{\mathcal{D}} u_{\mathcal{T}}\Big)_{\mathcal{D}\in\mathfrak{D}},$$

where the quantity $\nabla^{\mathcal{D}} u_{\mathcal{T}}$ is referred to as the restriction of the approximate gradient on the diamond cell $\mathcal{D} \in \mathfrak{D}$. On the one hand, for $\mathcal{D} \in \mathfrak{D}$ with $\overline{\mathcal{D}} \cap \overline{\partial\Omega} \cap \mathcal{E} = \emptyset$, such a restriction is defined so that one can get

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} \cdot \boldsymbol{\tau}_{\boldsymbol{K},\boldsymbol{L}} = \frac{u_L - u_K}{m_{\sigma^*}}, \qquad \nabla^{\mathcal{D}} u_{\mathcal{T}} \cdot \boldsymbol{\tau}_{\boldsymbol{K^*},\boldsymbol{L^*}} = \frac{u_{L^*} - u_{K^*}}{m_\sigma},$$

or equivalently,

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{\sin(\alpha_{\mathcal{D}})}\Big(\frac{u_L - u_K}{m_{\sigma^*}}\mathbf{n}_{\sigma K} + \frac{u_{L^*} - u_{K^*}}{m_\sigma}\mathbf{n}_{\sigma^* K^*}\Big).$$

On the other hand, our model problem is complemented with Dirichlet boundary conditions. This latter states that the solution is known on $\partial\Omega$. Consequently, for every $\mathcal{D} \in \mathfrak{D}$ with $\overline{\mathcal{D}} \cap \overline{\partial\Omega} \subset \mathcal{E}$ (see Fig. 4.2), one has

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{\sin(\alpha_{\mathcal{D}})}\Big(\frac{u_{|\partial\Omega}(x_{\mathcal{D}}) - u_K}{m_{\sigma^*}}\mathbf{n}_{\sigma K} + \frac{u_{|\partial\Omega}(x_{L^*}) - u_{|\partial\Omega}(x_{K^*})}{m_\sigma}\mathbf{n}_{\sigma^* K^*}\Big).$$

Notice that the two components of the discrete gradient are reproduced so that one can ensure a consistent approximation of the continuous gradient. This of course requires supplementary unknowns introduced on the dual cells.

For given $u_{\mathcal{T}}, v_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}}$ and $\mathcal{D} \in \mathfrak{D}$, we define $\delta^{\mathcal{D}} u_{\mathcal{T}} = \begin{bmatrix} u_K - u_L \\ u_{K^*} - u_{L^*} \end{bmatrix}$. Then, one sets

$$\Big(\nabla^{\mathfrak{D}} u_{\mathcal{T}}, \nabla^{\mathfrak{D}} v_{\mathcal{T}}\Big)_{\mathfrak{D},\Lambda} = \sum_{\mathcal{D}\in\mathfrak{D}} \delta^{\mathcal{D}} u_{\mathcal{T}} \cdot \mathbb{A}^{\mathcal{D},\Lambda} \delta^{\mathcal{D}} v_{\mathcal{T}}, \tag{4.2.1}$$

where the local matrix $\mathbb{A}^{\mathcal{D},\Lambda}$ reads

$$\mathbb{A}^{\mathcal{D},\Lambda} = \frac{1}{4m_{\mathcal{D}}}\begin{bmatrix} m_\sigma^2 \Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma K} & m_\sigma m_{\sigma^*}\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} \\ m_\sigma m_{\sigma^*}\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} & m_{\sigma^*}^2\,\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma^* K^*}\cdot\mathbf{n}_{\sigma^* K^*} \end{bmatrix}, \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{4.2.2}$$

One also defines

$$\mathbb{A}^{\mathcal{D}} = \frac{1}{4m_{\mathcal{D}}}\begin{bmatrix} m_\sigma^2 & m_\sigma m_{\sigma^*}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} \\ m_\sigma m_{\sigma^*}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} & m_{\sigma^*}^2 \end{bmatrix}, \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{4.2.3}$$

These matrices are positive-definite as given in Lemma A.0.2. Therefore, the bracket $(\cdot,\cdot)_{\mathfrak{D},\Lambda}$ is indeed an inner product on $(\mathbb{R}^2)^{\#\mathfrak{D}}$.

In order to make a conspicuous scheme later, we will denote

$$\begin{aligned} &a_{KL} := \frac{1}{\sin(\alpha_{\mathcal{D}})}\frac{m_\sigma}{m_{\sigma^*}}\,\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma K} > 0, \quad \eta^{\mathcal{D}}_{\sigma\sigma^*} := \frac{1}{\sin(\alpha_{\mathcal{D}})}\Lambda^{\mathcal{D}}\mathbf{n}_{\sigma K}\cdot\mathbf{n}_{\sigma^* K^*} \in \mathbb{R} \\ &g_M := g(u_M), \quad \forall M \in \{K, L, K^*, L^*\} \text{ and } g \in \{F, \xi\} \\ &\delta_{LK} u := u_L - u_K, \quad \delta_{L^*K^*} u := u_{L^*} - u_{K^*}. \end{aligned}$$

**Discrete divergence**

The discrete divergence has been originally introduced in [60] so as to reproduce a discrete counterpart of Green's formula. It is defined by a mapping from $(\mathbb{R}^2)^{\#\mathfrak{D}}$ to $\mathbb{R}^{\#\mathcal{T}}$ as follows:

$$\mathrm{div}^{\mathcal{T}}\Psi_{\mathfrak{D}} = \Big(\mathrm{div}^{\mathfrak{M}}\Psi_{\mathfrak{D}}, \mathrm{div}^{\mathfrak{M}^*}\Psi_{\mathfrak{D}}, \mathrm{div}^{\partial\mathfrak{M}^*}\Psi_{\mathfrak{D}}\Big), \quad \forall \Psi_{\mathfrak{D}} = (\Psi_{\mathcal{D}})_{\mathcal{D}\in\mathfrak{D}} \in (\mathbb{R}^2)^{\#\mathfrak{D}},$$

with $\mathrm{div}^{\mathfrak{M}}\Psi_{\mathfrak{D}} = (\mathrm{div}_K\Psi_{\mathfrak{D}})_{K\in\mathfrak{M}}$, $\mathrm{div}^{\mathfrak{M}^*}\Psi_{\mathfrak{D}} = (\mathrm{div}_{K^*}\Psi_{\mathfrak{D}})_{K^*\in\mathfrak{M}^*}$ and $\mathrm{div}^{\partial\mathfrak{M}^*}\Psi_{\mathfrak{D}} = (\mathrm{div}_{K^*}\Psi_{\mathfrak{D}})_{K^*\in\partial\mathfrak{M}^*}$. Each component is explicitly given by

$$\mathrm{div}_K\Psi_{\mathfrak{D}} = \frac{1}{m_K}\sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K} m_\sigma \Psi_{\mathcal{D}}\cdot\mathbf{n}_{\sigma K}, \qquad \forall K\in\mathfrak{M},$$

$$\mathrm{div}_{K^*}\Psi_{\mathfrak{D}} = \frac{1}{m_{K^*}}\sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_{K^*}} m_{\sigma^*} \Psi_{\mathcal{D}}\cdot\mathbf{n}_{\sigma^* K^*}, \qquad \forall K^*\in\mathfrak{M}^*,$$

$$\mathrm{div}_{K^*}\Psi_{\mathfrak{D}} = \frac{1}{m_{K^*}}\Big(\sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_{K^*}} m_{\sigma^*} \Psi_{\mathcal{D}}\cdot\mathbf{n}_{\sigma^* K^*} + \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_{K^*}\cap\partial\Omega} \frac{m_{\sigma^*}}{2}\Psi_{\mathcal{D}}\cdot\mathbf{n}_{\sigma^* K^*}\Big), \qquad \forall K^*\in\partial\mathfrak{M}^*.$$

### 4.2.3 Approximation spaces

This subsection is devoted to describing the discrete spaces together with some related notations. First, a DDFV mesh is composed of three different partitions. Let us therefore define the discrete functions on these meshes.

(i) We will denote $u_{\mathfrak{M}}$ (resp. $u_{\overline{\mathfrak{M}^*}}$) the first (resp. second) reconstruction on the primal (resp. dual) mesh, which is a piecewise constant function defined as

$$u_{\mathfrak{M}} = \sum_{K\in\mathfrak{M}} u_K \mathbf{1}_K, \quad u_{\overline{\mathfrak{M}^*}} = \sum_{K^*\in\overline{\mathfrak{M}^*}} u_{K^*}\mathbf{1}_{K^*}.$$

where $\mathbf{1}_K$ is the characteristic function of $K$. We then define the discrete function $u_h$ of $L^1(\Omega)$ as follows:

$$u_h = \frac{1}{2}\Big(u_{\mathfrak{M}} + u_{\overline{\mathfrak{M}^*}}\Big).$$

We henceforth denote $X_{\mathcal{T}}$ the set of all these functions $u_h$.

(ii) The third reconstruction $u_{\mathfrak{D}}$ concerns the diamond mesh. It is about piecewise constant functions of the form $u_{\mathfrak{D}} := \sum_{\mathcal{D}\in\mathfrak{D}} u_{\mathcal{D}}\mathbf{1}_{\mathcal{D}}$ for a given vector $(u_{\mathcal{D}})_{\mathcal{D}\in\mathfrak{D}}$. The set of all these functions will be denoted by $X_{\mathcal{D}}$.

As a consequence, the approximation spaces read:

$$X_{\mathcal{T},\delta t} = \Big\{u_{h,\delta t}\in L^1(Q_{\mathfrak{T}}) :\ u_{h,\delta t}(x,t) = u_h^{n+1}(x)\ /\ u_h\in X_{\mathcal{T}}, \forall t\in(t^n,t^{n+1}], \forall n = 0,\cdots,N-1\Big\}$$
$$X_{\mathfrak{D},\delta t} = \Big\{u_{\mathfrak{D},\delta t}\in L^1(Q_{\mathfrak{T}}) :\ u_{\mathfrak{D},\delta t}(x,t) = u_{\mathfrak{D}}^{n+1}(x)\ /\ u_{\mathcal{D}}\in X_{\mathcal{D}}, \forall t\in(t^n,t^{n+1}], \forall n = 0,\cdots,N-1\Big\}.$$

For each $u_{h,\delta t}\in X_{\mathcal{T},\delta t}$, its gradient $\nabla^{\mathfrak{D}}u_{h,\delta t}\in X_{\mathfrak{D},\delta t}\times X_{\mathfrak{D},\delta t}$ is written by

$$\nabla^{\mathfrak{D}}u_{h,\delta t}(x,t) = \nabla^{\mathfrak{D}}u_h^{n+1}(x) := \nabla^{\mathfrak{D}}u_{\mathcal{T}}^{n+1}(x), \qquad \forall t\in(t^n,t^{n+1}],\ \forall n = 0,\cdots,N-1.$$

As for $u_{h,\delta t} \in X_{\mathcal{T},\delta t}$, we take

$$u_{\mathfrak{M},\delta t}(x,t) = u_{\mathfrak{M}}^{n+1}(x), \quad u_{\overline{\mathfrak{M}^*},\delta t}(x,t) = u_{\overline{\mathfrak{M}^*}}^{n+1}(x) \quad \forall t \in (t^n, t^{n+1}],\ \forall n = 0, \cdots, N-1.$$

Let us now consider a nonlinear function $F : \mathbb{R} \longrightarrow \mathbb{R}$. We will denote by $F_{h,\delta t}$ the mean value of $F(u_{\mathfrak{M},\delta t})$ and $F(u_{\overline{\mathfrak{M}^*},\delta t})$:

$$F_{h,\delta t} = \frac{1}{2}\Big(F(u_{\mathfrak{M},\delta t}) + F(u_{\overline{\mathfrak{M}^*},\delta t})\Big).$$

We next equip the finite dimensional space $X_{\mathcal{T}}$ with the norm $|\cdot|_{p,\mathcal{T}}$. For every $u_h \in X_{\mathcal{T}}$, we define

$$|u_h|_{p,\mathcal{T}} = \begin{cases} \Big(\frac{1}{2}\sum\limits_{K\in\mathfrak{M}} m_K |u_K|^p + \frac{1}{2}\sum\limits_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*}|u_{K^*}|^p\Big)^{1/p} & \text{if} \quad 1 \le p < +\infty \\ \max\Big(\max\limits_{K\in\mathfrak{M}}|u_K|, \max\limits_{K^*\in\overline{\mathfrak{M}^*}}|u_{K^*}|\Big) & \text{if} \quad p = +\infty \end{cases}.$$

This leads us to consider the discrete Sobolev norm as

$$\|u_h\|_{1,p,\mathcal{T}} = \begin{cases} \Big(|u_h|_{p,\mathcal{T}}^p + \big\|\nabla^{\mathfrak{D}} u_h\big\|_p^p\Big)^{1/p} & \text{if} \quad 1 \le p < +\infty \\ |u_h|_{\infty,\mathcal{T}} + \big\|\nabla^{\mathfrak{D}} u_h\big\|_\infty & \text{if} \quad p = +\infty \end{cases},$$

where the norm of the discrete gradient is

$$\Big\|\nabla^{\mathfrak{D}} u_h\Big\|_p^p = \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\nabla^{\mathcal{D}} u_h\right|^p, \quad \forall 1 \le p < +\infty, \text{ and } \Big\|\nabla^{\mathfrak{D}} u_h\Big\|_\infty = \max_{\mathcal{D}\in\mathfrak{D}} \left|\nabla^{\mathcal{D}} u_h\right|.$$

Observe that

$$\Big\|\nabla^{\mathfrak{D}} u_h\Big\|_2^2 = \sum_{\mathcal{D}\in\mathfrak{D}} \delta^{\mathcal{D}} u_{\mathcal{T}} \cdot \mathbb{A}^{\mathcal{D}} \delta^{\mathcal{D}} u_{\mathcal{T}}.$$

Finally we can also give the discrete counterpart of the $L^q(0,\mathfrak{T}; W^{1,p}(\Omega))$-norm

$$\|u_{h,\delta t}\|_{q;1,p,\mathcal{T}} = \begin{cases} \Big(\sum\limits_{n=1}^{N} \delta t \, \|u_h^n\|_{1,p,\mathcal{T}}^q\Big)^{1/q} & \text{if} \quad 1 \le p,q < +\infty \\ \max\limits_{n=1,\cdots,N} \|u_h^n\|_{1,\infty,\mathcal{T}} & \text{if} \quad p = q = +\infty \end{cases}.$$

## 4.3 Numerical scheme

Belonging to the family of finite volume methods, the DDFV scheme is basically obtained by integrating the first equation of (4.1.1) over $M\times]t^n, t^{n+1}]$, where $M$ is a primal or an internal dual cell. Performing Green's formula yields the balance equation. Then the resulting fluxes are approximated by introducing the definition of the discrete gradient and that of the numerical flux function.

For the convenience of the reader, we briefly look at the discretization of (4.1.1) on the primal mesh and it is deduced similarly in the case of the dual mesh. So, let $n \in \{0, \ldots, N-1\}$ and $K$ be a primal control volume. Then, one gets

$$\int_{t^n}^{t^{n+1}} \int_K \partial_t u \ \mathrm{d}x\,\mathrm{d}t - \sum_{\sigma\in\mathcal{E}_K} \int_{t^n}^{t^{n+1}} \int_\sigma f(u)\ \Lambda\nabla u \cdot \mathbf{n}_{\sigma K}\,\mathrm{d}\sigma\,\mathrm{d}t = 0. \tag{4.3.1}$$

The evolution term is approximated thanks to the Euler scheme

$$\int_{t^n}^{t^{n+1}} \int_K \partial_t u \ \mathrm{d}x \, \mathrm{d}t \approx \ m_K \Big( u_K^{n+1} - u_K^n \Big), \tag{4.3.2}$$

where $u_K^m$ is the mean value of $u(.,t^m)$ over $K$ for $m = n, n+1$. Concerning the diffusion part, it is discretized as follows

$$- \int_{t^n}^{t^{n+1}} \int_\sigma f(u) \Lambda \nabla u \cdot \mathbf{n}_{\sigma K} \, \mathrm{d}\sigma \, \mathrm{d}t \approx \delta t \, \Big( a_{KL} (F(u_K^{n+1}) - F(u_L^{n+1})) + v_{KL}^{n+1} \eta_{\sigma\sigma^*}^{\mathcal{D}} (\xi(u_{K^*}^{n+1}) - \xi(u_{L^*}^{n+1})) \Big)$$

where $F$ (resp. $\xi$) is the Kirchoff (resp. semi-Kirchoff) function and $v_{KL}^{n+1}$ is an upstream approximation of $v(u) := \sqrt{f(u)}$ on the primal edge $\sigma$. We next provide a central formula concerning the constructions of $v_{KL}^{n+1}$. This consists of considering the Engquist-Osher scheme [107], which reads

$$v_{KL}^{n+1} = \begin{cases} v_\downarrow(u_L^{n+1}) + v_\uparrow(u_K^{n+1}) & \text{if } \eta_{\sigma\sigma^*}^{\mathcal{D}} \ (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}) \geq 0 \\ v_\downarrow(u_K^{n+1}) + v_\uparrow(u_L^{n+1}) & \text{else} \ , \end{cases} \tag{4.3.3}$$

where the functions $v_\downarrow, v_\uparrow$ are given by

$$v_\uparrow(u) := \int_0^u \Big( v'(s) \Big)^+ \mathrm{d}s, \quad v_\downarrow(u) := - \int_0^u \Big( v'(s) \Big)^- \mathrm{d}s,$$

and $x^+ = \max(x, 0)$, $x^- = \max(-x, 0)$ for all $x \in \mathbb{R}$. This convention will be adopted hereafter. In light of hypothesis $(A_2)$, the functions $v_\uparrow, v_\downarrow$ exist.

We wish to emphasize that one can rewrite the quantity $v_{KL}^{n+1} \eta_{\sigma\sigma^*}^{\mathcal{D}} (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})$ thanks to a numerical flux function $G$ as follows

$$G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{L^*K^*}^{n+1} \xi(u)) = v_{KL}^{n+1} \eta_{\sigma\sigma^*}^{\mathcal{D}} (\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}).$$

We recall that a function $G$ of arguments $(a, b, c) \in \mathbb{R}^3$ is said to be a numerical flux if the assertions below are satisfied:

$$\begin{cases} (H_1) & G(\cdot, b, c) \text{ is nondecreasing and continuous for all } b, c \in \mathbb{R}, \\ & \quad \text{and } G(a, \cdot, c) \text{ is nonincreasing and continuous for all } a, c \in \mathbb{R}; \\ (H_2) & G(a, b, c) = -G(a, b, -c) \text{ for all } a, b, c \in \mathbb{R}; \\ (H_3) & G(a, a, c) = v(a)c \text{ for all } a, c \in \mathbb{R}. \end{cases} \tag{4.3.4}$$

As stressed in [11, 46], we require a penalization operator, which is crucial to pass to the limit in the scheme. This penalty term permits to check that the approximate solution on the primal mesh and the dual mesh tend to the same limit. It will be also a key point in our study for the convergence of the diffusive term. To this purpose, let $\varepsilon \in ]0, 2[$ and $u_\mathcal{T} \in \mathbb{R}^{\#\mathcal{T}}$. The penalization $\mathcal{P}^\mathcal{T}$ is a map from $\mathbb{R}^{\#\mathcal{T}}$ to $\mathbb{R}^{\#\mathcal{T}}$ defined, for all $u_\mathcal{T}$, by

$$\mathcal{P}^\mathcal{T} u_\mathcal{T} = \Big( \mathcal{P}^\mathfrak{M} u_\mathcal{T}, \mathcal{P}^{\mathfrak{M}^*} u_\mathcal{T}, \mathcal{P}^{\partial\mathfrak{M}^*} u_\mathcal{T} \Big),$$

where $\mathcal{P}^\mathfrak{M} u_\mathcal{T} = (\mathcal{P}_K u_\mathcal{T})_{K \in \mathfrak{M}}$, $\mathcal{P}^{\mathfrak{M}^*} u_\mathcal{T} = (\mathcal{P}_{K^*} u_\mathcal{T})_{K^* \in \mathfrak{M}^*}$ , $\mathcal{P}^{\partial\mathfrak{M}^*} u_\mathcal{T} = (\mathcal{P}_{K^*} u_\mathcal{T})_{K^* \in \partial\mathfrak{M}^*}$ such that

$$\mathcal{P}_K u_\mathcal{T} = \frac{1}{m_K} \frac{1}{h_\mathfrak{D}^\varepsilon} \sum_{K^* \in \overline{\mathfrak{M}^*}} m_{K \cap K^*} \Big( F(u_K) - F(u_{K^*}) \Big), \quad \forall \ K \in \mathfrak{M}, \tag{4.3.5}$$

$$\mathcal{P}_{K^*} u_\mathcal{T} = \frac{1}{m_{K^*}} \frac{1}{h_\mathfrak{D}^\varepsilon} \sum_{K \in \mathfrak{M}} m_{K \cap K^*} \Big( F(u_{K^*}) - F(u_K) \Big), \quad \forall \ K^* \in \overline{\mathfrak{M}^*}. \tag{4.3.6}$$

Owing to the homogeneous Dirichlet boundary condition, one sets $\mathcal{P}_{K^*} u_{\mathcal{T}} = 0 \quad \forall\, K^* \in \partial\mathfrak{M}^*$. Based on the elementary inequality

$$(F(a) - F(b))(a-b) \geq (\xi(a) - \xi(b))^2, \qquad \forall a, b \in \mathbb{R}, \tag{4.3.7}$$

one can check that

$$\begin{aligned}
[\![\mathcal{P} u_{\mathcal{T}}, u_{\mathcal{T}}]\!]_{\mathcal{T}} &= \frac{1}{2} \frac{1}{h_{\mathfrak{D}}^{\varepsilon}} \sum_{K^* \in \overline{\mathfrak{M}^*}} \sum_{K \in \mathfrak{M}} m_{K \cap K^*} \, (F(u_K) - F(u_{K^*}))(u_K - u_{K^*}) \\
&\geq \frac{1}{2} \frac{1}{h_{\mathfrak{D}}^{\varepsilon}} \left\| \xi(u_{\mathfrak{M}}) - \xi(u_{\overline{\mathfrak{M}^*}}) \right\|_{L^2(\Omega)}^2 .
\end{aligned} \tag{4.3.8}$$

Thanks to the DDFV discretization, an approximate solution for the problem (4.1.1) is defined as a function $u_{h,\delta t} \in X_{\mathcal{T},\delta t}$ satisfying the set of equations:

$$u_M^0 = \frac{1}{m_M} \int_M u^0(x) \, \mathrm{d}x, \qquad \forall M \in \mathcal{T}, \tag{4.3.9}$$

$$\begin{aligned}
&\frac{m_K}{\delta t} \left( u_K^{n+1} - u_K^n \right) \\
&\qquad + \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_K} \left( a_{KL} \, (F_K^{n+1} - F_L^{n+1}) + G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{L^*K^*}^{n+1} \xi(u)) \right) \\
&\qquad\qquad + \gamma \, \mathcal{P}_K u_{\mathcal{T}}^{n+1} = 0, \qquad \forall K \in \mathfrak{M}, \quad n \geq 0,
\end{aligned} \tag{4.3.10}$$

$$\begin{aligned}
&\frac{m_{K^*}}{\delta t} \left( u_{K^*}^{n+1} - u_{K^*}^n \right) \\
&\qquad + \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathcal{D}_{K^*}} \left( a_{K^*L^*} \, (F_{K^*}^{n+1} - F_{L^*}^{n+1}) + G(u_{K^*}^{n+1}, u_{L^*}^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}} \delta_{LK}^{n+1} \xi(u)) \right) \\
&\qquad\qquad + \gamma \, \mathcal{P}_{K^*} u_{\mathcal{T}}^{n+1} = 0, \qquad \forall K^* \in \mathfrak{M}^*, \quad n \geq 0.
\end{aligned} \tag{4.3.11}$$

The coefficient $\gamma$ is a positive parameter. Let us next check that $G$ is well-defined. This is the object of the following result.

**Lemma 4.3.1.** *The numerical flux function $G$ is well-defined, meaning that assertions $(H_1), (H_2)$ and $(H_3)$ of (4.3.4) are fulfilled.*

*Proof.* Observe that items $(H_1), (H_3)$ of (4.3.4) are direct consequences of the expression of $v_{KL}^{n+1}$ given in (4.3.3) and the assumption $(A_2)$. It remains to check that the assertion $(H_2)$ holds. To this end, we first point out that the discrete gradient on a fixed diamond, which we recall below, is uniquely defined

$$\nabla^{\mathcal{D}} u_{\mathcal{T}} = \frac{1}{\sin(\alpha_{\mathcal{D}})} \left( \frac{u_L - u_K}{m_{\sigma^*}} \mathbf{n}_{\sigma K} + \frac{u_{L^*} - u_{K^*}}{m_\sigma} \mathbf{n}_{\sigma^* K^*} \right).$$

In other words, we associate to the primal interface $\sigma = K|L$ a unique dual interface $\sigma = K^*|L^*$. Now if we permute $K, L$ then $K^*, L^*$ are automatically permuted, but the coefficient $\eta_{\sigma\sigma^*}^{\mathcal{D}}$ keeps the same sign. In particular, this asserts that $\eta_{\sigma\sigma^*}^{\mathcal{D}} = \eta_{\sigma^*\sigma}^{\mathcal{D}}$. Accordingly

$$\eta_{\sigma\sigma^*}^{\mathcal{D}} \left( \xi_{K^*}^{n+1} - \xi_{L^*}^{n+1} \right) = - \eta_{\sigma\sigma^*}^{\mathcal{D}} \left( \xi_{L^*}^{n+1} - \xi_{K^*}^{n+1} \right).$$

According to this identity and the definition of $v_{KL}^{n+1}$ introduced in (4.3.3), one finds

$$v_{KL}^{n+1} = v_{LK}^{n+1}.$$

Hence

$$G\Big(u_K^{n+1}, u_L^{n+1}, \eta_{\sigma\sigma^*}^{\mathcal{D}}(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})\Big) = -\,G\Big(u_L^{n+1}, u_K^{n+1}, \eta_{\sigma\sigma^*}^{\mathcal{D}}(\xi_{L^*}^{n+1} - \xi_{K^*}^{n+1})\Big).$$

□

**Remark 4.3.1.** *In the case where $\Lambda = Id$, the coefficient $\eta_{\sigma\sigma^*}^{\mathcal{D}}$ measures the flatting of the diamond cells. In particular, if $\eta_{\sigma\sigma^*}^{\mathcal{D}} \equiv 0$ for all $\mathcal{D}$, meaning that the mesh is orthogonal [69], the above discretization reduces to the pioneer TPFA (Two-Point Flux Approximation) scheme for the problem (4.1.1) on the primal mesh and on the dual mesh separately. Its convergence analysis can be found in [71].*

**Remark 4.3.2.** *Let us fix the penalty coefficient to $\gamma = 0$. According to Lemma 4.3.1, the above numerical scheme is locally conservative i.e. there exists a unique discrete flux $J_{\mathfrak{D}}^{n+1}$ such that the following relationship holds*

$$[\![u_{\mathcal{T}}^{n+1} - u_{\mathcal{T}}^{n} + \delta t\, \mathrm{div}^{\mathcal{T}}\, J_{\mathfrak{D}}^{n+1}, \psi_{\mathcal{T}}]\!]_{\mathcal{T}} = 0, \quad \forall \psi_{\mathcal{T}} \in \mathbb{R}^{\#\mathcal{T}} \text{ and } n \geq 0. \tag{4.3.12}$$

*Indeed, the function $J_{\mathfrak{D}}^{n+1} = (J_{\mathcal{D}}^{n+1})_{\mathcal{D}\in\mathfrak{D}}$ is defined via its two projections with respect to the primal and dual units normals. In other words, it is sufficient to set*

$$J_{\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K} = \frac{1}{m_\sigma}\Big(a_{KL}\,(F_K^{n+1} - F_L^{n+1}) + G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}}\delta_{L^*K^*}^{n+1}\xi(u))\Big),$$
$$J_{\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma^*K^*} = \frac{1}{m_{\sigma^*}}\Big(a_{K^*L^*}\,(F_{K^*}^{n+1} - F_{L^*}^{n+1}) + G^*(u_{K^*}^{n+1}, u_{L^*}^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}}\delta_{LK}^{n+1}\xi(u))\Big).$$

*As a consequence, $J_{\mathcal{D}}^{n+1}$ is expressed in a unique way thanks to the crucial identity [11]*

$$\sin(\alpha_{\mathcal{D}})J_{\mathcal{D}}^{n+1} = (J_{\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma K})\boldsymbol{\tau}_{\boldsymbol{K,L}} + (J_{\mathcal{D}}^{n+1} \cdot \mathbf{n}_{\sigma^* K^*})\boldsymbol{\tau}_{\boldsymbol{K^*,L^*}}.$$

*Finally, (4.3.12) stems from the definition of the discrete divergence given above and that of the scheme.*

## 4.4 $L^\infty$ bounds and a priori estimates

In this section, we show that any solution to the equations of the proposed scheme verifies a $L^\infty$ bound. In addition, some a priori estimates are derived on the discrete gradient of the Kirchoff function. These materials are of importance to prove the convergence.

### 4.4.1 Boundedness of discrete solutions

**Lemma 4.4.1.** *For each fixed integer $0 \leq n \leq N-1$, let $(u_{\mathcal{T}}^{n+1})$ be a vector of $\mathbb{R}^{\#\mathcal{T}}$ such that the DDFV scheme (4.3.9)-(4.3.11) holds. Then, $u_{\mathfrak{M}}^{n+1}, u_{\mathfrak{M}^*}^{n+1}$ belong to $[0,1]$.*

*Proof.* The proof is carried out by induction on $n$. Fix $n \in \{0, \cdots, N-1\}$. Let us assume that the claim is true for $u_{\mathfrak{M}}^{n}, u_{\mathfrak{M}^*}^{n}$ and check that it is so for $u_{\mathfrak{M}}^{n+1}, u_{\mathfrak{M}^*}^{n+1}$. To this purpose, we perform the proof in two steps .

*Step 1* : We consider $u_K^{n+1} = \min_{L\in\mathfrak{M}}(u_L^{n+1})$. Multiplying (4.3.10) by $-(u_K^{n+1})^-$ yields

$$-\frac{m_K}{\delta t}\Big(u_K^{n+1} - u_K^{n}\Big)(u_K^{n+1})^- - \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K}\Big(a_{KL}\,(F_K^{n+1} - F_L^{n+1}) + G(u_K^{n+1}, u_L^{n+1}; \eta_{\sigma\sigma^*}^{\mathcal{D}}\delta_{L^*K^*}^{n+1}\xi(u))\Big)(u_K^{n+1})^-$$
$$- \gamma\, \mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1})^- = 0.$$

Since $F$ is a nondecreasing function, we obtain $a_{KL}\,(F_K^{n+1}-F_L^{n+1})\le 0$. Furthermore

$$G(u_K^{n+1},u_L^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u))(u_K^{n+1})^-\le 0.$$

Indeed, if $0\le u_K^{n+1}$ then $(u_K^{n+1})^-=0$. Otherwise, we use the fact that the numerical flux function is nonincreasing with respect to the second argument and that it is consistent

$$\begin{aligned}G(u_K^{n+1},u_L^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u))(u_K^{n+1})^- &\le G\Big(u_K^{n+1},u_K^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{K^*L^*}\xi(u)\Big)(u_K^{n+1})^-\\ &= v(u_K^{n+1})\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u)(u_K^{n+1})^-=0.\end{aligned}$$

The previous equality holds thanks to the degeneracy of the function $v$ on $]-\infty,0]$. Let us next demonstrate that

$$-\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1})^-\ge 0. \tag{4.4.1}$$

It follows from the definition of the penalization term highlighted in (4.3.5) that

$$\begin{aligned}-\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1})^- &= \frac{1}{m_K}\frac{1}{h^{\varepsilon}_{\mathfrak{D}}}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K\cap K^*}\Big(-F(u_K^{n+1})(u_K^{n+1})^- + F(u_{K^*}^{n+1})(u_K^{n+1})^-\Big)\\ &= \frac{1}{m_K}\frac{1}{h^{\varepsilon}_{\mathfrak{D}}}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K\cap K^*}F(u_{K^*}^{n+1})(u_K^{n+1})^-,\end{aligned}$$

where $F(u_K^{n+1})(u_K^{n+1})^-=0$. Since $F(u_{K^*}^{n+1})\ge 0$, regardless the sign of $u_{K^*}^{n+1}$, inequality (4.4.1) holds. Whence

$$-\Big(u_K^{n+1}-u_K^n\Big)(u_K^{n+1})^- = \big|(u_K^{n+1})^-\big|^2+(u_K^{n+1})^-u_K^n\le 0,$$

which implies, using the induction assumption, that $(u_K^{n+1})^-=0$. Hence, $u_K^{n+1}\ge 0$.

*Step 2* : We here switch the role of the control volume $K$ and take now $u_K^{n+1}=\max_{L\in\mathfrak{M}}(u_L^{n+1})$. Multiplying (4.3.10) by $(u_K^{n+1}-1)^+$ gives

$$\begin{aligned}&\frac{m_K}{\delta t}\Big(u_K^{n+1}-u_K^n\Big)(u_K^{n+1}-1)^+\\ &\qquad+\sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K}\Big(a_{KL}\,(F_K^{n+1}-F_L^{n+1})+G(u_K^{n+1},u_L^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u))\Big)(u_K^{n+1}-1)^+\\ &\qquad\qquad+\gamma\,\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1}-1)^+=0.\end{aligned}$$

It is now evident that $a_{KL}\,(F_K^{n+1}-F_L^{n+1})(u_K^{n+1}-1)^+\ge 0$. Next, let us establish

$$G(u_K^{n+1},u_L^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u))(u_K^{n+1}-1)^+\ge 0.$$

So, if $u_K^{n+1}\le 1$ then $(u_K^{n+1}-1)^+=0$. Otherwise, $u_K^{n+1}\ge 1$, we utilize once again the consistency of $G$ and the fact that it is decreasing with respect to the second variable. Therefore

$$\begin{aligned}G(u_K^{n+1},u_L^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u))(u_K^{n+1}-1)^+ &\ge G\Big(u_K^{n+1},u_K^{n+1};\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{K^*L^*}\xi(u)\Big)(u_K^{n+1}-1)^+\\ &= v(u_K^{n+1})\eta^{\mathcal{D}}_{\sigma\sigma^*}\delta^{n+1}_{L^*K^*}\xi(u)(u_K^{n+1}-1)^+=0.\end{aligned}$$

Let us show that $\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1}-1)^+\ge 0$. We first observe that

$$\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1}-1)^+=\frac{1}{m_K}\frac{1}{h^{\varepsilon}_{\mathfrak{D}}}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K\cap K^*}\Big(F(u_K^{n+1})-F(1)+F(1)-F(u_{K^*}^{n+1})\Big)(u_K^{n+1}-1)^+.$$

On the one hand, $(F(u_K^{n+1}) - F(1))(u_K^{n+1} - 1)^+ = 0$ for every $u_K^{n+1} \geq 0$. On the other hand, $F(1) - F(u_{K^*}^{n+1}) \geq 0$ for all $u_{K^*}^{n+1} \in \mathbb{R}$. Thus, $\mathcal{P}_K u_{\mathcal{T}}^{n+1}(u_K^{n+1} - 1)^+ \geq 0$. Utilizing now the identity

$$\Big(u_K^{n+1} - u_K^n\Big)(u_K^{n+1} - 1)^+ = (u_K^{n+1} - 1)^{+^2} + (u_K^{n+1} - 1)^+(1 - u_K^n),$$

we deduce that $(u_K^{n+1} - 1)^+ = 0$, which yields $u_K^{n+1} \leq 1$.

Similarly, we mimic the same steps so that we prove the property in the case of the dual mesh. Hence, the proof of the Lemma is concluded. □

**Remark 4.4.1.** *The degeneracy of the function $v$ and the flux splitting scheme (4.3.3) enforce the boundedness of the discrete solution. Also, this particular approach ensures the coercivity of the discrete elliptic operator. One can notice that the Godunov scheme [69] does not fulfill this latter property.*

In the sequel, we will denote by $C$ different constants in various occurrences, which depend only on the physical data together with the regularity of the mesh and are independent of the discretization parameters $\delta t$, $h_{\mathfrak{D}}$.

### 4.4.2 Estimates on the discrete gradients

We first recall the following remarkable formula.

**Lemma 4.4.2.** *(Discrete integration by parts) Let $\mathcal{M}$ be a primal or dual mesh of the domain $\Omega$. For every $K \in \mathcal{M}$, we denote by $N(K)$ the set of neighbors of $K$. Let $A_{KL}$, $K \in \mathcal{M}$ and $L \in N(K)$ be a real value with $A_{KL} = -A_{LK}$, and let $\varphi$ be a piecewise constant function on the cells of $\mathcal{M}$. Then*

$$\sum_{K\in\mathcal{M}} \sum_{L\in N(K)} A_{KL}\varphi_K = -\frac{1}{2} \sum_{K\in\mathcal{M}} \sum_{L\in N(K)} A_{KL}(\varphi_L - \varphi_K).$$

*Particularly, if $A_{KL} = T_{KL}(c_L - c_K)$, with $T_{KL} = T_{LK}$, one infers*

$$\sum_{K\in\mathcal{M}} \sum_{L\in N(K)} T_{KL}(c_L - c_K)\varphi_K = -\frac{1}{2} \sum_{K\in\mathcal{M}} \sum_{L\in N(K)} T_{KL}(c_L - c_K)(\varphi_L - \varphi_K).$$

**Proof.** *The proof of this lemma is omitted since it is similar to that given in [23].*

We next refer to [24, 69] for the proof of the following fundamental inequality.

**Lemma 4.4.3.** *(The discrete Poincaré inequality) Consider $\mathcal{T}$ a mesh of $\Omega$. Then there exists a constant $C_p$, only depending on the diameter of $\Omega$, such that for every $w_h \in X_{\mathcal{T}}$ one has*

$$|w_h|_{2,\mathcal{T}}^2 \leq \frac{1}{2} \|w_{\mathfrak{M}}\|_{L^2(\Omega)}^2 + \frac{1}{2} \left\|w_{\overline{\mathfrak{M}^*}}\right\|_{L^2(\Omega)}^2 \leq C_p \left\|\nabla^{\mathfrak{D}} w_h\right\|_2^2.$$

**Proposition 4.4.1.** *(The discrete gradient estimate) Let $(u_{\mathcal{T}}^n)$ be in $\mathbb{R}^{\#\mathcal{T}}$, for $n = 0, \ldots, N$, such that the DDFV scheme (4.3.9)-(4.3.11) holds. Then*

$$\sum_{n=0}^{N-1} \delta t \left\|\nabla^{\mathfrak{D}} \xi_h^{n+1}\right\|_2^2 + \frac{\gamma}{h_{\mathfrak{D}}^{\varepsilon}} \sum_{n=0}^{N-1} \delta t \left\|\xi(u_{\mathfrak{M}}^{n+1}) - \xi(u_{\overline{\mathfrak{M}^*}}^{n+1})\right\|_{L^2(\Omega)}^2 \leq C, \tag{4.4.2}$$

*for some appropriate positive constant $C$.*

*Proof.* We multiply the first (resp. second) equation of the DDFV scheme (4.3.10)-(4.3.11) by $u_K^{n+1}$ (resp. $u_{K^*}^{n+1}$) and sum up over all the primal (resp. dual) cells and the integers $n$. Adding together the resulting equations leads to

$$T_1 + T_2 + T_3 = 0,$$

where we have set

$$\begin{aligned}
T_1 =& \sum_{n=0}^{N-1} \sum_{K\in\mathfrak{M}} m_K(u_K^{n+1} - u_K^n)u_K^{n+1} + \sum_{n=0}^{N-1} \sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*}(u_{K^*}^{n+1} - u_{K^*}^n)u_{K^*}^{n+1},\\
T_2 =& \sum_{n=0}^{N-1} \delta t \sum_{K\in\mathfrak{M}} \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K} \Big(a_{KL}\,(F_K^{n+1} - F_L^{n+1}) \;+\; v_{KL}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})\Big)u_K^{n+1}\\
&+ \sum_{n=0}^{N-1} \delta t \sum_{K^*\in\overline{\mathfrak{M}^*}} \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_{K^*}} \Big(\,a_{K^*L^*}\,(F_{K^*}^{n+1} - F_{L^*}^{n+1}) \;+\; v_{K^*L^*}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_K^{n+1} - \xi_L^{n+1})\Big)u_{K^*}^{n+1},\\
T_3 =& 2\sum_{n=0}^{N-1} \delta t\,\gamma\,[\![\mathcal{P}u_{\mathcal{T}}^{n+1}, u_{\mathcal{T}}^{n+1}]\!]_{\mathcal{T}}.
\end{aligned}$$

First of all, observe that

$$x(x-y) \geq \frac{1}{2}(x^2 - y^2), \quad \forall x, y \in \mathbb{R}.$$

According to the above inequality, one can underestimate $T_1$

$$\frac{1}{2}\sum_{K\in\mathfrak{M}} m_K\Big((u_K^N)^2 - (u_K^0)^2\Big) + \frac{1}{2}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*}\Big((u_{K^*}^N)^2 - (u_{K^*}^0)^2\Big) \leq T_1. \tag{4.4.3}$$

Let us now turn our attention to the term $T_2$. To this end, we perform a discrete integration by parts as given in Lemma 4.4.2 to obtain

$$T_2 = T_{21} + T_{22},$$

with

$$\begin{aligned}
T_{21} =& \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} \Big(a_{KL}\,(F_K^{n+1} - F_L^{n+1})(u_K^{n+1} - u_L^{n+1}) + a_{K^*L^*}\,(F_{K^*}^{n+1} - F_{L^*}^{n+1})(u_{K^*}^{n+1} - u_{L^*}^{n+1})\Big),\\
T_{22} =& \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} \Big(v_{KL}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})(u_K^{n+1} - u_L^{n+1}) + v_{K^*L^*}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_K^{n+1} - \xi_L^{n+1})(u_{K^*}^{n+1} - u_{L^*}^{n+1})\Big).
\end{aligned}$$

The practical inequality (4.3.7) implies that

$$T_{21} \geq \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} a_{KL}\,(\xi_K^{n+1} - \xi_L^{n+1})^2 + \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} a_{K^*L^*}\,(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})^2.$$

Thanks to the monotonicity of the functions $v_\uparrow, v_\downarrow$ and the definition of $v_{KL}^{n+1}$, we find

$$v_{KL}^{n+1}\Big(u_K^{n+1} - u_L^{n+1}\Big)\eta_{\sigma\sigma^*}^{\mathcal{D}}\Big(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}\Big) \geq \eta_{\sigma\sigma^*}^{\mathcal{D}}\Big(\xi_K^{n+1} - \xi_L^{n+1}\Big)\Big(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}\Big).$$

Similarly

$$v_{K^*L^*}^{n+1}\Big(u_{K^*}^{n+1}-u_{L^*}^{n+1}\Big)\eta_{\sigma\sigma^*}^{\mathcal{D}}\Big(\xi_K^{n+1}-\xi_L^{n+1}\Big)\geq\eta_{\sigma\sigma^*}^{\mathcal{D}}\Big(\xi_K^{n+1}-\xi_L^{n+1}\Big)\Big(\xi_{K^*}^{n+1}-\xi_{L^*}^{n+1}\Big).$$

As a result we get

$$T_{22}\geq 2\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}}\eta_{\sigma\sigma^*}^{\mathcal{D}}\Big(\xi_K^{n+1}-\xi_L^{n+1}\Big)\Big(\xi_{K^*}^{n+1}-\xi_{L^*}^{n+1}\Big).$$

We deduce that

$$T_2\geq\sum_{n=0}^{N-1}\delta t\left(\nabla^{\mathfrak{D}}\xi_h^{n+1},\nabla^{\mathfrak{D}}\xi_h^{n+1}\right)_{\mathfrak{D},\Lambda}. \tag{4.4.4}$$

In view of the relationship (4.2.1) and Lemma A.0.1 we assert

$$T_2\geq C\sum_{n=0}^{N-1}\delta t\left\|\nabla^{\mathcal{D}}\xi_h^{n+1}\right\|_2^2,$$

for some constant $C>0$. Next, owing to (4.3.8), we write

$$T_3\geq\frac{\gamma}{h_{\mathfrak{D}}^{\varepsilon}}\sum_{n=0}^{N-1}\delta t\left\|\xi(u_{\mathfrak{M}}^{n+1})-\xi(u_{\mathfrak{M}^*}^{n+1})\right\|_{L^2(\Omega)}^2. \tag{4.4.5}$$

Combining(4.4.3)-(4.4.5), the energy estimate (4.4.2) follows as required. □

**Corollary 4.4.1.** *From the previous proposition, one gets*

$$\sum_{n=0}^{N-1}\delta t\left\|\nabla^{\mathfrak{D}}F_h^{n+1}\right\|_2^2\leq C,$$

*Proof.* This result is a direct consequence of Lemma A.0.1 together with inequality (4.4.2). It is sufficient to observe that

$$F(a)-F(b)=v(x_0)\Big(\xi(a)-\xi(b)\Big),$$

for some $x_0\in[\min(a,b),\max(a,b)]$ and notice that the function $v$ is bounded. □

## 4.5 Existence of discrete solutions

In this section, we prove that the nonlinear algebraic system, which comes from the DDFV scheme, admits a solution. To this end, we will need the following fundamental lemma, that can be found in [65]. This result ensures the existence of at least one zero of some specific vector fields.

**Lemma 4.5.1.** *Let $\mathcal{A}$ be a finite dimensional Hilbert space with inner product $(\cdot,\cdot)$ and norm $||\cdot||$, and let $\mathcal{L}$ be a continuous mapping from $\mathcal{A}$ into itself which verifies*

$$(\mathcal{L}(x),x)\ >0,\quad \textit{for}\ \ ||x||=r>0.$$

*Then, there exists $x^*\in\mathcal{A}$ with $||x^*||<r$ such that*

$$\mathcal{L}(x^*)=0.$$

We now state the existence result in the proposition below.

**Proposition 4.5.1.** *The DDFV scheme (4.3.9)-(4.3.11) has at least one solution $u_{\mathcal{T}}^{n+1}$ for every $n = 0, \dots, N-1$.*

*Proof.* We proceed by induction on $n$. We then assume that $u_{\mathcal{T}}^{n}$ is given and prove the existence of $u_{\mathcal{T}}^{n+1}$ satisfying the numerical scheme (4.3.10)-(4.3.11). To this purpose, we define the mapping $\mathcal{L} : \mathbb{R}^{\#\mathcal{T}} \longrightarrow \mathbb{R}^{\#\mathcal{T}}$ that associates for each $u_{\mathcal{T}}^{n+1}$ the vector :

$$\mathcal{L}(u_{\mathcal{T}}^{n+1}) \;=\; \Big(\mathcal{L}_M\Big)_{M\in\mathcal{T}},$$

where

$$\begin{aligned}
\mathcal{L}_K \;&=\; \frac{m_K}{\delta t}(u_K^{n+1} - u_K^n) \\
&\quad + \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K} \Big(a_{KL}\,(F_K^{n+1} - F_L^{n+1}) + \; v_{KL}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1})\Big) + \gamma\,\mathcal{P}_K u_{\mathcal{T}}^{n+1}, \;\text{ if }\; M = K \in \mathfrak{M}, \\
\mathcal{L}_{K^*} &= \frac{m_{K^*}}{\delta t}(u_{K^*}^{n+1} - u_{K^*}^n) \\
&\quad + \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_{K^*}} \Big(a_{K^*L^*}\,(F_{K^*}^{n+1} - F_{L^*}^{n+1}) + \; v_{KL}^{n+1}\,\eta_{\sigma\sigma^*}^{\mathcal{D}}\,(\xi_K^{n+1} - \xi_L^{n+1})\Big) + \gamma\,\mathcal{P}_{K^*} u_{\mathcal{T}}^{n+1}, \;\text{ if }\; M = K^* \in \mathfrak{M}^*, \\
\mathcal{L}_{K^*} &= 0, \text{if }\; M = K^* \in \partial\mathfrak{M}^*.
\end{aligned}$$

The functional $\mathcal{L}$ is well-defined and continuous. It remains to demonstrate that

$$\Big(\mathcal{L}(u_{\mathcal{T}}^{n+1}), u_{\mathcal{T}}^{n+1}\Big) \;> 0, \quad \text{for} \;\; ||u_{\mathcal{T}}^{n+1}||_{\mathbb{R}^{\#\mathcal{T}}} = r, \tag{4.5.1}$$

for some sufficiently large $r$. It follows from the calculation of the previous section, Lemma 4.4.1 and the Poincaré inequality given in Lemma 4.4.3 that

$$\begin{aligned}
\Big(\mathcal{L}(u_{\mathcal{T}}^{n+1}), u_{\mathcal{T}}^{n+1}\Big) &\geq \frac{1}{\delta t}\sum_{K\in\mathfrak{M}} m_K\,\Big((u_K^{n+1})^2 - (u_K^n)^2\Big) + \frac{1}{\delta t}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*}\,\Big((u_{K^*}^{n+1})^2 - (u_{K^*}^n)^2\Big) \\
&\quad + C\,\big\|\nabla^{\mathcal{D}}\xi_h^{n+1}\big\|_2^2 \\
&\geq C'\,\big|u_h^{n+1}\big|_{2,\mathcal{T}}^2 - \frac{2\,|\Omega|}{\delta t},
\end{aligned}$$

for some constants $C, C' > 0$. Thanks to the equivalence of the usual norms $||\cdot||_{\mathbb{R}^{\#\mathcal{T}}}, |\cdot|_{2,\mathcal{T}}$ on the finite dimensional space $\mathbb{R}^{\#\mathcal{T}}$, inequality (4.5.1) is fulfilled provided a large $r$. We therefore obtain the existence of at least one solution to the DDFV scheme (4.3.9)-(4.3.11). □

## 4.6 Convergence

We first give some standard compactness properties. Their proofs follow similar arguments as, for instance, in [11, 69].

**Lemma 4.6.1.** *(Space Translates)*
*Let $u_{h,\delta t}$ be a discrete solution to the DDFV scheme (4.3.10)-(4.3.11). Then*

$$\int_0^{\mathfrak{T}}\int_{\Omega'} |\xi_{h,\delta t}(x+y,t) - \xi_{h,\delta t}(x,t)|\,\mathrm{d}x\,\mathrm{d}t \leq\; \omega(|y|), \quad \textit{for every} \;\; y \in \mathbb{R}^2, \tag{4.6.1}$$

*where $\Omega' = \{x \in \Omega \;/\; x + y \in \Omega\}$ and $\omega$ is a modulus of continuity independent of $\delta t, h_{\mathfrak{D}}$, verifying $\omega(|y|) \longrightarrow 0$ as $|y| \longrightarrow 0$.*

*Proof.* The proof of this claim is made in the case of the primal mesh and it is similar for the dual mesh, i.e. it is sufficient to prove that

$$\int_0^{\mathfrak{T}} \int_{\Omega'} |\xi(u_{\mathfrak{M},\delta t}(x+y,t)) - \xi(u_{\mathfrak{M},\delta t}(x,t))| \,\mathrm{d}x \,\mathrm{d}t \leq \omega(|y|), \quad \text{for every} \quad y \in \mathbb{R}^2. \tag{4.6.2}$$

Now, for every $x, y \in \mathbb{R}^2$ and $\sigma = K|L$, we define the characteristic function $\chi_\sigma$ as

$$\chi_\sigma(x,y) = \begin{cases} 1, & \text{if } [x, x+y] \cap \sigma \neq \emptyset, \\ 0, & \text{else.} \end{cases}$$

We know that $\displaystyle\int_{\Omega'} \chi_\sigma(x,y) \,\mathrm{d}x \leq m_\sigma \,|y|$ (see [69] for more details). As a consequence, since we have

$$|\xi(u_{\mathfrak{M},\delta t}(x+y,t)) - \xi(u_{\mathfrak{M},\delta t}(x,t))| \;\leq\; \sum_{\sigma=K|L} \chi_\sigma(x,y) \left|\xi_L^{n+1} - \xi_K^{n+1}\right|,$$

this gives

$$\begin{aligned} \int_{\Omega'} |\xi(u_{\mathfrak{M},\delta t}(x+y,t)) - \xi(u_{\mathfrak{M},\delta t}(x,t))| \,\mathrm{d}x \;&\leq\; |y| \sum_{\sigma=K|L} m_\sigma \left|\xi_L^{n+1} - \xi_K^{n+1}\right| \\ &\leq\; C\ |y| \sum_{\mathcal{D}_{\sigma,\sigma^*} \in \mathfrak{D}} m_{\mathcal{D}} \left| \frac{\xi_L^{n+1} - \xi_K^{n+1}}{m_{\sigma^*}} \right| \end{aligned}$$

for some appropriate $C$ depending on the regularity of the mesh. On the other hand, we have

$$\left| \frac{\xi_L^{n+1} - \xi_K^{n+1}}{m_{\sigma^*}} \right| \;\leq\; \left|\nabla^{\mathcal{D}} \xi_h^{n+1}\right|.$$

As a result of the Cauchy-Schwarz inequality and the energy estimate (4.4.2), one infers

$$\int_0^{\mathfrak{T}} \int_{\Omega'} |\xi(u_{\mathfrak{M},\delta t}(x+y,t)) - \xi(u_{\mathfrak{M},\delta t}(x,t))| \,\mathrm{d}x \,\mathrm{d}t \leq C\,|y|\,.$$

This completes the proof. □

**Lemma 4.6.2.** *(Time translates)*
*Let $u_{h,\delta t}$ be a solution to the DDFV scheme (4.3.9)-(4.3.11). Then there exists a constant $C$ that does not depend on $h_{\mathfrak{D}}$ nor on $\delta t$ such that*

$$\begin{aligned} &\int_0^{\mathfrak{T}-\tau} \int_\Omega \left|\xi(u_{\mathfrak{M},\delta t}(x,t+\tau)) - \xi(u_{\mathfrak{M},\delta t}(x,t))\right|^2 \mathrm{d}x \,\mathrm{d}t \\ &+ \int_0^{\mathfrak{T}-\tau} \int_\Omega \left|\xi(u_{\overline{\mathfrak{M}^*},\delta t}(x,t+\tau)) - \xi(u_{\overline{\mathfrak{M}^*},\delta t}(x,t))\right|^2 \mathrm{d}x \,\mathrm{d}t \leq C\,(\tau + \delta t), \end{aligned} \tag{4.6.3}$$

*for all $\tau \in (0, \mathfrak{T})$.*

*Proof.* The proof follows the main steps of [69, Lemma 4.6]. We will provide the proof of the first integral in (4.6.3) and that of the second one is proved in a similar way. To begin with, let $\tau \in (0, \mathfrak{T})$ and $t \in (0, \mathfrak{T} - \tau)$. We set

$$A = \int_0^{\mathfrak{T}-\tau} \int_\Omega \left| \xi(u_{\mathfrak{M},\delta t}(x, t+\tau)) - \xi(u_{\mathfrak{M},\delta t}(x,t)) \right|^2 \mathrm{d}x.$$

We next define $n_0(t) \in \{0, \cdots, N-1\}$ such that $t^{n_0(t)} < t \leq t^{n_0(t)} + 1$ and $n_1(t) \in \{0, \cdots, N-1\}$ such that $t^{n_1(t)} < t + \tau \leq t^{n_1(t)} + 1$. One then can rewrite $A$ as follows

$$\begin{aligned} A &= \int_0^{\mathfrak{T}-\tau} \sum_{K\in\mathfrak{M}} m_K \left| \xi(u_K^{n_1(t)}) - \xi(u_K^{n_0(t)}) \right|^2 \\ &\leq C \int_0^{\mathfrak{T}-\tau} \sum_{K\in\mathfrak{M}} \left( \left( \xi(u_K^{n_1(t)}) - \xi(u_K^{n_0(t)}) \right) \times \sum_{t\leq n\delta t<t+\tau} m_K \left( u_K^{n+1} - u_K^n \right) \right) \mathrm{d}t, \end{aligned}$$

for some constant $C > 0$ depending only on $\xi$. In light of the definition of the DDFV scheme, one gets

$$\begin{aligned} A \leq L \int_0^{\mathfrak{T}-\tau} &\sum_{K\in\mathfrak{M}} \left( \xi(u_K^{n_1(t)}) - \xi(u_K^{n_0(t)}) \right) \\ \times \sum_{t\leq n\delta t<t+\tau} \delta t \Bigg( &- \sum_{\mathcal{D}_{\sigma,\sigma^*}\in\mathcal{D}_K} \left( a_{KL} \left( F_K^{n+1} - F_L^{n+1} \right) + v_{KL}^{n+1} \, \eta_{\sigma\sigma^*}^{\mathcal{D}} \left( \xi_{K^*}^{n+1} - \xi_{L^*}^{n+1} \right) \right) + \gamma \, \mathcal{P}_K u_{\mathcal{T}}^{n+1} \Bigg) \mathrm{d}t. \end{aligned}$$

Applying the integration by parts and the first mean value theorem ensures the existence of a positive constant $C$ that depends only on the regularity of the mesh, $\xi$, $\|\xi'\|_\infty$ and on $\overline{\Lambda}$ with

$$\begin{aligned} A \leq C \int_0^{\mathfrak{T}-\tau} &\sum_{t\leq n\delta t<t+\tau} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} \Big( \left| \xi_K^{n+1} - \xi_L^{n+1} \right| \left| \xi(u_K^{n_1(t)}) - \xi(u_L^{n_1(t)}) \right| \\ &+ \left| \xi_K^{n+1} - \xi_L^{n+1} \right| \left| \xi(u_L^{n_0(t)}) - \xi(u_K^{n_0(t)}) \right| \\ &+ \left| \xi_{K^*}^{n+1} - \xi_{L^*}^{n+1} \right| \left| \xi(u_K^{n_1(t)}) - \xi(u_L^{n_1(t)}) \right| \\ &+ \left| \xi_{L^*}^{n+1} - \xi_{K^*}^{n+1} \right| \left| \xi(u_L^{n_0(t)}) - \xi(u_K^{n_0(t)}) \right| \Big) \mathrm{d}t \\ - \gamma \int_0^{\mathfrak{T}-\tau} &\sum_{t\leq n\delta t<t+\tau} \delta t \sum_{K\in\mathfrak{M}} \left( \xi(u_K^{n_1(t)}) - \xi(u_K^{n_0(t)}) \right) \mathcal{P}_K u_{\mathcal{T}}^{n+1} \, \mathrm{d}t. \end{aligned}$$

Let us now introduce the characteristic function $\beta$ which is defined by (see [69])

$$\beta(n,t) = \begin{cases} 1 & \text{if } t \leq (n+1)\delta t < t + \tau \\ 0 & \text{otherwise} \end{cases},$$

Using the elementary inequality $ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$ in the previous estimate leads to

$$A \leq \frac{C}{2} \Big( E_1 + E_2 + E_3 + E_4 \Big) + E_5$$

where we have obtained

$$E_1 = \sum_{n=0}^{N-1} \delta t \int_0^{\mathfrak{T}-\tau} \beta(n,t) \sum_{\mathcal{D}\in\mathfrak{D}} \left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \mathrm{d}t,$$

$$E_2 = \sum_{n=0}^{N-1} \delta t \int_0^{\mathfrak{T}-\tau} \beta(n,t) \sum_{\mathcal{D}\in\mathfrak{D}} \left(\xi(u_K^{n_1(t)}) - \xi(u_L^{n_1(t)})\right)^2 \mathrm{d}t,$$

$$E_3 = \sum_{n=0}^{N-1} \delta t \int_0^{\mathfrak{T}-\tau} \beta(n,t) \sum_{\mathcal{D}\in\mathfrak{D}} \left(\xi(u_K^{n_0(t)}) - \xi(u_L^{n_0(t)})\right)^2 \mathrm{d}t,$$

$$E_4 = \sum_{n=0}^{N-1} \delta t \int_0^{\mathfrak{T}-\tau} \beta(n,t) \sum_{\mathcal{D}\in\mathfrak{D}} \left(\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}\right)^2 \mathrm{d}t,$$

$$E_5 = -\gamma \sum_{n=0}^{N-1} \delta t \int_0^{\mathfrak{T}-\tau} \beta(n,t) \sum_{K\in\mathfrak{M}} \left(\xi(u_K^{n_1(t)}) - \xi(u_K^{n_0(t)})\right) \mathcal{P}_K u_{\mathcal{T}}^{n+1} \,\mathrm{d}t.$$

On the other hand, observe that

$$\sum_{\mathcal{D}\in\mathfrak{D}} \left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \le C_1 \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\frac{\xi_K^{n+1} - \xi_L^{n+1}}{m_{\sigma^*}}\right|^2 \le C_1 \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\nabla^{\mathcal{D}} \xi_h^{n+1}\right|^2,$$

where $C_1 > 0$ depends only on the regularity of the mesh. By virtue of the energy estimate (4.4.2) and since $\displaystyle\int_0^{\mathfrak{T}-\tau} \beta(n,t) \le \tau$, there exists an appropriate constant $C$ such that $E_1 \le C\tau$ and $E_4 \le C\tau$. Next, following [69, Lemma 4.6], one claims that

$$E_2 \le C_1 \sum_{m=0}^{N-1} \int_{t^m}^{t^{m+1}} \delta t \sum_{n=0}^{N-1} \beta(n,t) \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\nabla^{\mathcal{D}} \xi_h^m\right|^2 \mathrm{d}t \le C\tau.$$

In an analogous way, one has $E_3 \le C\tau$. Finally, to treat $E_5$, we use the maximum principle and the estimate (4.4.2) to get that $E_5 \le C\tau$. Hence, the proof of the lemma is concluded. □

We now claim a weak convergence of the discrete gradient and a strong convergence of $u_{h,\delta t}$.

**Proposition 4.6.1.** *Let $(\mathcal{T}_h)_h$ be a sequence of DDFV meshes such that $h_{\mathfrak{D}}, \delta t$ tend to zero and reg$(\mathcal{T}_h)$ is bounded. Then, the following convergences hold up to a subsequence:*

$$u_{h,\delta t}, u_{\mathfrak{M}_h,\delta t}, u_{\overline{\mathfrak{M}^*}_h,\delta t} \longrightarrow u \qquad \text{a.e. in } Q_{\mathfrak{T}}, \tag{4.6.4}$$

$$\nabla^{\mathfrak{D}} F_{h,\delta t} \longrightarrow \nabla F(u) \qquad \text{weakly} \quad \text{in } L^2(Q_{\mathfrak{T}})^2. \tag{4.6.5}$$

*Moreover*

$$0 \le u \le 1 \qquad \text{a.e. in } Q_{\mathfrak{T}}. \tag{4.6.6}$$

*Proof.* Thanks to Kolmogorov's compactness theorem [30], the sequences $\xi(u_{\mathfrak{M}_h,\delta t}), \xi(u_{\overline{\mathfrak{M}^*}_h,\delta t})$ are relatively compact in $L^1(Q_{\mathfrak{T}})$. This ensures the existence of unlabeled subsequences of $\xi(u_{\mathfrak{M}_h,\delta t}), \xi(u_{\overline{\mathfrak{M}^*}_h,\delta t})$ converging almost everywhere :

$$\xi(u_{\mathfrak{M}_h,\delta t}) \longrightarrow \xi_1 \ \text{ a.e. in } Q_{\mathfrak{T}}, \quad \text{and} \quad \xi(u_{\overline{\mathfrak{M}^*}_h,\delta t}) \longrightarrow \xi_2 \ \text{ a.e. in } Q_{\mathfrak{T}}.$$

Since $\xi^{-1}$ is continuous, we deduce that

$$u_{\mathfrak{M}_h,\delta t} \longrightarrow u_1 := \xi^{-1}(\xi_1) \ \text{ a.e. in } Q_{\mathfrak{T}}, \ \text{ and } \ u_{\overline{\mathfrak{M}^*}_h,\delta t} \longrightarrow u_2 := \xi^{-1}(\xi_2) \ \text{ a.e. in } Q_{\mathfrak{T}}.$$

In light of Proposition 4.4.1, we assert

$$\left\| \xi(u_{\mathfrak{M},\delta t}) - \xi(u_{\overline{\mathfrak{M}^*},\delta t}) \right\|^2_{L^2(Q_{\mathfrak{T}})} \leq C\, h^{\varepsilon}_{\mathfrak{D}}. \tag{4.6.7}$$

Thus, up to unlabeled subsequence, we get

$$\xi(u_{\mathfrak{M},\delta t}) - \xi(u_{\overline{\mathfrak{M}^*},\delta t}) \longrightarrow 0, \ \text{ a.e. in } Q_{\mathfrak{T}}.$$

Therefore

$$u_{\mathfrak{M},\delta t} - u_{\overline{\mathfrak{M}^*},\delta t} \longrightarrow 0, \ \text{ a.e. in } Q_{\mathfrak{T}}.$$

We then verify that $u_1 = u_2 := u$. Consequently

$$u_{h,\delta t} \longrightarrow u \ \text{ a.e. in } Q_{\mathfrak{T}}, \ \text{ and } \ \xi_{h,\delta t} \longrightarrow \xi(u) \ \text{ a.e. in } Q_{\mathfrak{T}}.$$

Thanks to the $L^\infty$ bound given in Lemma 4.4.1, we deduce from Lebesgue's dominated convergence theorem that

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \|u_{h,\delta t} - u\|_{L^2(Q_{\mathfrak{T}})} = 0.$$

Thereby

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \|F_{h,\delta t} - F(u)\|_{L^2(Q_{\mathfrak{T}})} = 0.$$

Next, thanks to Corollary 4.4.1, the sequence $(\nabla^{\mathfrak{D}} F_{h,\delta t})$ is bounded in $(L^2(Q_{\mathfrak{T}}))^d$. Let us establish that

$$\nabla^{\mathfrak{D}} F_{h,\delta t} \longrightarrow \nabla F(u) \quad \text{weakly in } (L^2(Q_{\mathfrak{T}}))^d.$$

We first show that $\nabla F(u) = G$ is the sense of distribution. To do this, let $\varphi \in \left(\mathcal{C}^\infty(\overline{\Omega}\times[0,\mathfrak{T}])\right)^2$. Due to the weak convergence of $(\nabla^{\mathfrak{D}} F_{h,\delta t})$ and the strong one of $(F_{h,\delta t})$, one can pass to the limit in

$$\begin{aligned} I_{\mathcal{T}_h,\delta t} &:= \int_{Q_{\mathfrak{T}}} \nabla^{\mathfrak{D}} F_{h,\delta t} \cdot \varphi \,\mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} F_{h,\delta t} \operatorname{div} \varphi \,\mathrm{d}x\,\mathrm{d}t \\ &\longrightarrow \int_{Q_{\mathfrak{T}}} G \cdot \varphi \ \mathrm{d}x\,\mathrm{d}t + \int_{Q_{\mathfrak{T}}} F(u) \operatorname{div} \ \varphi \,\mathrm{d}x\,\mathrm{d}t. \end{aligned}$$

The definition of the discrete gradient allows us to write

$$\int_{Q_{\mathfrak{T}}} \nabla^{\mathfrak{D}} F_{h,\delta t} \cdot \varphi \,\mathrm{d}x\,\mathrm{d}t = \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \nabla^{\mathcal{D}} F_{h,\delta t} \cdot \varphi^{n+1}_{\mathcal{D}}, \tag{4.6.8}$$

where $\varphi^{n+1}_{\mathcal{D}} = \dfrac{1}{\delta t\, m_{\mathcal{D}}} \int_{t^n}^{t^{n+1}} \int_{\mathcal{D}} \varphi \,\mathrm{d}x\,\mathrm{d}t$. For every diamond $\mathcal{D} = \mathcal{D}_{\sigma,\sigma^*}$, we introduce $\varphi^{n+1}_{\sigma}, \varphi^{n+1}_{\sigma^*}$ and $\tilde{\varphi}^{n+1}_{\mathcal{D}}$ as follows

$$\begin{aligned} &\varphi^{n+1}_{\sigma} = \frac{1}{\delta t\, m_\sigma} \int_{t^n}^{t^{n+1}} \int_{\sigma} \varphi(s,t) \,\mathrm{d}s\,\mathrm{d}t, \ \varphi^{n+1}_{\sigma^*} = \frac{1}{\delta t\, m_{\sigma^*}} \int_{t^n}^{t^{n+1}} \int_{\sigma^*} \varphi(s,t) \,\mathrm{d}s\,\mathrm{d}t, \\ &\tilde{\varphi}^{n+1}_{\mathcal{D}} \cdot \mathbf{n}_{\sigma K} = \varphi^{n+1}_{\sigma} \cdot \mathbf{n}_{\sigma K}, \quad \tilde{\varphi}^{n+1}_{\mathcal{D}} \cdot \mathbf{n}_{\sigma^* K^*} = \varphi^{n+1}_{\sigma^*} \cdot \mathbf{n}_{\sigma^* K^*}. \end{aligned}$$

Note that $\tilde{\varphi}_{\mathcal{D}}^{n+1}$ is uniquely defined. Thanks to the smoothness of $\varphi$, we derive the estimate

$$\begin{aligned}\left|\varphi_{\mathcal{D}}^{n+1}-\tilde{\varphi}_{\mathcal{D}}^{n+1}\right| &\leq \frac{1}{\sin(\alpha_{\mathcal{T}})}\Big(\left|\varphi_{\mathcal{D}}^{n+1}-\varphi_{\sigma}^{n+1}\right|+\left|\varphi_{\mathcal{D}}^{n+1}-\varphi_{\sigma^*}^{n+1}\right|\Big)\\ &\leq 2\ \mathrm{reg}(\mathcal{T}_h)\ h_{\mathcal{D}}\ \|\nabla\varphi\|_{L^\infty}\,. \end{aligned} \tag{4.6.9}$$

Now, the expression (4.6.8) becomes

$$\begin{aligned}\int_{Q_{\mathfrak{T}}}\nabla^{\mathfrak{D}}F_{h,\delta t}\cdot\varphi\,\mathrm{d}x\,\mathrm{d}t &= \sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}}m_{\mathcal{D}}\nabla^{\mathcal{D}}F_h^{n+1}\cdot\tilde{\varphi}_{\mathcal{D}}^{n+1}\\ &\quad+\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}}m_{\mathcal{D}}\nabla^{\mathcal{D}}F_h^{n+1}\cdot(\varphi_{\mathcal{D}}^{n+1}-\tilde{\varphi}_{\mathcal{D}}^{n+1})\\ &=: A_{\mathcal{T}_h,\delta t}+B_{\mathcal{T}_h,\delta t}.\end{aligned}$$

In addition, inequality (4.6.9) and the energy estimate (4.4.2) lead to

$$\lim_{h_{\mathfrak{D}},\delta t\to 0}B_{\mathcal{T}_h,\delta t}=0.$$

We next return to the definition of the discrete gradient. It implies

$$\begin{aligned}A_{\mathcal{T}_h,\delta t} &= \frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}}m_\sigma m_{\sigma^*}\left(\frac{F_L^{n+1}-F_K^{n+1}}{m_{\sigma^*}}\mathbf{n}_{\sigma K}+\frac{F_{L^*}^{n+1}-F_{K^*}^{n+1}}{m_\sigma}\mathbf{n}_{\sigma^*K^*},\tilde{\varphi}_{\mathcal{D}}^{n+1}\right)\\ &= -\frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{K\in\mathfrak{M}}F_K^{n+1}\sum_{\sigma\in\mathcal{E}_K}m_\sigma\left(\tilde{\varphi}_{\mathcal{D}}^{n+1},\mathbf{n}_{\sigma K}\right)\\ &\quad-\frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{K^*\in\overline{\mathfrak{M}^*}}F_{K^*}^{n+1}\sum_{\sigma^*\in\mathcal{E}_{K^*}}m_{\sigma^*}\left(\tilde{\varphi}_{\mathcal{D}}^{n+1},\mathbf{n}_{\sigma^*K^*}\right),\end{aligned}$$

where we used discrete integration by parts (4.4.2). By virtue of the expression of $\tilde{\varphi}_{\mathcal{D}}^{n+1}$, one gets

$$\begin{aligned}A_{\mathcal{T}_h,\delta t} &= -\frac{1}{2}\sum_{n=0}^{N-1}\sum_{K\in\mathfrak{M}}F_K^{n+1}\sum_{\sigma\in\mathcal{E}_K}\int_{t^n}^{t^{n+1}}\int_\sigma\varphi(s)\cdot\mathbf{n}_{\sigma K}\,\mathrm{d}s\,\mathrm{d}t\\ &\quad-\frac{1}{2}\sum_{n=0}^{N-1}\sum_{K^*\in\overline{\mathfrak{M}^*}}F_{K^*}^{n+1}\sum_{\sigma^*\in\mathcal{E}_{K^*}}\int_{t^n}^{t^{n+1}}\int_{\sigma^*}\varphi(s)\cdot\mathbf{n}_{\sigma^*K^*}\,\mathrm{d}s\,\mathrm{d}t.\end{aligned}$$

Stokes formula entails

$$\begin{aligned}A_{\mathcal{T}_h,\delta t} &= -\frac{1}{2}\sum_{n=0}^{N-1}\sum_{K\in\mathfrak{M}}F_K^{n+1}\int_{t^n}^{t^{n+1}}\int_K\mathrm{div}\,\varphi\,\mathrm{d}x\,\mathrm{d}t\\ &\quad-\frac{1}{2}\sum_{n=0}^{N-1}\sum_{K^*\in\overline{\mathfrak{M}^*}}F_{K^*}^{n+1}\int_{t^n}^{t^{n+1}}\int_{K^*}\mathrm{div}\,\varphi\,\mathrm{d}x\,\mathrm{d}t\\ &= -\int_{Q_{\mathfrak{T}}}F_{h,\delta t}\,\mathrm{div}\,\varphi\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

As a consequence

$$\lim_{h_{\mathfrak{D}},\delta t\to 0}I_{\mathcal{T}_h,\delta t}=0.$$

Thereby we proved that $F(u)\in L^2(0,\mathfrak{T};H_0^1(\Omega))$ and $\nabla F(u)=G$. This finishes up the proof. □

## 4.7 Passage to the limit

In this section we prove that any limit of the approximate solution sequence converges towards the weak solution of the main problem.

**Theorem 4.7.1.** *Under hypotheses $(A_1)$–$(A_3)$ and assuming a uniform boundedness of the mesh regularity, the limit function $u$ of Proposition 4.6.1 is the weak solution to the problem (4.1.1) in the sense of Definition 4.1.1.*

*Proof.* Let $\psi \in \mathcal{C}_c^\infty(\Omega \times [0,\mathfrak{T}))$, we denote by $\psi_K^{n+1} = \psi(x_K, t^{n+1})$ and $\psi_{K^*}^{n+1} = \psi(x_{K^*}, t^{n+1})$. We multiply the equations (4.3.10), (4.3.11) by $\frac{1}{2}\delta t \psi_K^{n+1}$, $\frac{1}{2}\delta t\ \psi_{K^*}^{n+1}$ respectively, sum over $K$, $K^*$ and $n$. Next, one performs an integration by parts, adds and substracts $\sum_{n=0}^{N-1} \delta t \left(\nabla^{\mathfrak{D}} F_h^{n+1}, \nabla^{\mathfrak{D}} \psi_h^{n+1}\right)_{\mathfrak{D},\Lambda}$ to get

$$\mathcal{S}^1_{\mathcal{T}_h,\delta t} + \mathcal{S}^2_{\mathcal{T}_h,\delta t} + \mathcal{S}^3_{\mathcal{T}_h,\delta t} + \mathcal{S}^4_{\mathcal{T}_h,\delta t} = 0,$$

where

$$\begin{aligned}
\mathcal{S}^1_{\mathcal{T}_h,\delta t} &= \sum_{n=0}^{N-1} [\![ u^{n+1}_{\mathcal{T}_h} - u^n_{\mathcal{T}_h}, \psi^{n+1}_{\mathcal{T}_h} ]\!]_{\mathcal{T}_h},\\
\mathcal{S}^2_{\mathcal{T}_h,\delta t} &= \sum_{n=0}^{N-1} \delta t \left(\nabla^{\mathfrak{D}} F_h^{n+1}, \nabla^{\mathfrak{D}} \psi_h^{n+1}\right)_{\mathfrak{D},\Lambda} = \int_{Q_{\mathfrak{T}}} \Lambda \nabla^{\mathfrak{D}} F_h^{n+1} \cdot \nabla^{\mathfrak{D}} \psi_h^{n+1} \,\mathrm{d}x\,\mathrm{d}t,\\
\mathcal{S}^3_{\mathcal{T}_h,\delta t} &= \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}_h} \eta^{\mathcal{D}}_{\sigma\sigma^*} \left[ v^{n+1}_{KL} \left(\xi^{n+1}_{K^*} - \xi^{n+1}_{L^*}\right) - \left(F^{n+1}_{K^*} - F^{n+1}_{L^*}\right)\right] \left(\psi^{n+1}_K - \psi^{n+1}_L\right)\\
&\quad + \frac{1}{2} \sum_{n=0}^{N-1} \delta t \sum_{\mathcal{D}\in\mathfrak{D}_h} \eta^{\mathcal{D}}_{\sigma\sigma^*} \left[ v^{n+1}_{K^*L^*} \left(\xi^{n+1}_{K} - \xi^{n+1}_{L}\right) - \left(F^{n+1}_{K} - F^{n+1}_{L}\right)\right] \left(\psi^{n+1}_{K^*} - \psi^{n+1}_{L^*}\right),\\
\mathcal{S}^4_{\mathcal{T}_h,\delta t} &= \gamma \sum_{n=0}^{N-1} \delta t\ [\![ \mathcal{P} u^{n+1}_{\mathcal{T}_h}, \psi^{n+1}_{\mathcal{T}_h} ]\!]_{\mathcal{T}_h}.
\end{aligned}$$

Let us start off by establishing

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \mathcal{S}^1_{\mathcal{T}_h,\delta t} = -\int_\Omega u^0\ \psi(\cdot,0)\,\mathrm{d}x - \int_{Q_{\mathfrak{T}}} u\ \partial_t\psi \,\mathrm{d}x\,\mathrm{d}t.$$

Using a summation by parts in time and the fact that $\psi_K^N = \psi_{K^*}^N = 0$, yields

$$\begin{aligned}
\mathcal{S}^1_{\mathcal{T}_h,\delta t} &= -[\![ u^0_{\mathcal{T}_h}, \psi_{\mathcal{T}_h}(\cdot,0) ]\!]_{\mathcal{T}_h} - \sum_{n=0}^{N-1} [\![ u^{n+1}_{\mathcal{T}_h}, \psi^{n+1}_{\mathcal{T}_h} - \psi^n_{\mathcal{T}_h} ]\!]_{\mathcal{T}_h}\\
&=: \mathcal{S}^{1,1}_{\mathcal{T}_h,\delta t} + \mathcal{S}^{1,2}_{\mathcal{T}_h,\delta t}.
\end{aligned}$$

Thanks to the strong convergence of $(\psi_{\mathcal{T}_h}(\cdot,0))$, one obtains

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \mathcal{S}^{1,1}_{\mathcal{T}_h,\delta t} = -\int_\Omega u^0 \psi(\cdot,0)\,\mathrm{d}x.$$

Expanding the term $\mathcal{S}^{1,2}_{\mathcal{T}_h,\delta t}$ entails

$$\begin{aligned}\mathcal{S}^{1,2}_{\mathcal{T}_h,\delta t} &= -\sum_{n=0}^{N-1} [\![ u^{n+1}_{\mathcal{T}_h} - \psi^n_{\mathcal{T}_h} ]\!]_{\mathcal{T}_h} \\ &= -\frac{1}{2}\sum_{n=0}^{N-1}\sum_{K\in\mathfrak{M}} m_K \int_{t^n}^{t^{n+1}} u_K^{n+1}\partial_t\psi(x_K,t)\,\mathrm{d}x\,\mathrm{d}t - \frac{1}{2}\sum_{n=0}^{N-1}\sum_{K^*\in\overline{\mathfrak{M}^*}} m_{K^*}\int_{t^n}^{t^{n+1}} u_{K^*}^{n+1}\partial_t\psi(x_{K^*},t)\,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Bearing in mind that $(\partial_t\psi(x_K,\cdot))_{K\in\mathfrak{M}}$ and $(\partial_t\psi(x_{K^*},\cdot))_{K^*\in\overline{\mathfrak{M}^*}}$ converge uniformly towards $\partial_t\psi$, we apply the Lebesgue dominated convergence theorem to find

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \mathcal{S}^{1,2}_{\mathcal{T}_h,\delta t} = -\int_{Q_{\mathfrak{T}}} u\,\partial_t\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Let us next prove the convergence of the diffusion part. To do so, we recall that the sequence $(\nabla^{\mathfrak{D}} F_h^{n+1})$ converges weakly towards $\nabla F(u)$ whereas $(\Lambda\nabla^{\mathfrak{D}}\psi_h^{n+1})$ converges uniformly towards $\Lambda\nabla\psi$. Thereby

$$\lim_{h_{\mathfrak{D}},\delta t\to 0} \mathcal{S}^{2}_{\mathcal{T}_h,\delta t} = \int_{Q_{\mathfrak{T}}} \Lambda\nabla F(u)\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t.$$

Let us turn our attention to the convergence of $\mathcal{S}^{3}_{\mathcal{T}_h,\delta t}$. This term can be split up into two parts as follows

$$\mathcal{S}^{3}_{\mathcal{T}_h,\delta t} = \mathcal{S}^{3,1}_{\mathcal{T}_h,\delta t} + \mathcal{S}^{3,2}_{\mathcal{T}_h,\delta t},$$

where we have set

$$\begin{aligned}\mathcal{S}^{3,1}_{\mathcal{T}_h,\delta t} &= \frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h}\eta^{\mathcal{D}}_{\sigma\sigma^*}\left[v^{n+1}_{KL}\Big(\xi^{n+1}_{K^*}-\xi^{n+1}_{L^*}\Big)-\Big(F^{n+1}_{K^*}-F^{n+1}_{L^*}\Big)\right]\Big(\psi^{n+1}_K-\psi^{n+1}_L\Big),\\ \mathcal{S}^{3,2}_{\mathcal{T}_h,\delta t} &= \frac{1}{2}\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h}\eta^{\mathcal{D}}_{\sigma\sigma^*}\left[v^{n+1}_{K^*L^*}\Big(\xi^{n+1}_{K}-\xi^{n+1}_{L}\Big)-\Big(F^{n+1}_{K}-F^{n+1}_{L}\Big)\right]\Big(\psi^{n+1}_{K^*}-\psi^{n+1}_{L^*}\Big).\end{aligned}$$

Next, the first mean value theorem guarantees the existence of a constant

$$u_{K^*L^*}\in[\min(u^{n+1}_{K^*},u^{n+1}_{L^*}),\max(u^{n+1}_{K^*},u^{n+1}_{L^*})]$$

satisfying

$$F^{n+1}_{K^*}-F^{n+1}_{L^*} = v(u_{K^*L^*})\Big(\xi^{n+1}_{K^*}-\xi^{n+1}_{L^*}\Big).$$

Thus, using assumption $(A_3)$ on the tensor $\Lambda$ and the regularity of the mesh, we get

$$\left|\mathcal{S}^{3,1}_{\mathcal{T}_h,\delta t}\right| \le C\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h} m_{\mathcal{D}}\left|v^{n+1}_{KL}-v(u_{K^*L^*})\right|\left|\nabla^{\mathcal{D}}\xi_{h,\delta t}\right|\left|\nabla^{\mathcal{D}}\psi_{h,\delta t}\right|.$$

for some constant $C>0$. We set

$$\begin{aligned}\overline{\xi}^{n+1}_{\mathcal{D}} &:= \max_{M\in\mathcal{V}_{\mathcal{D}}}\{\xi(u^{n+1}_M)\}, \qquad \underline{\xi}^{n+1}_{\mathcal{D}} := \min_{M\in\mathcal{V}_{\mathcal{D}}}\{\xi(u^{n+1}_M)\}\\ \overline{\xi}_{\mathcal{T}_h,\delta t|\mathcal{D}\times(t^n,t^{n+1}]} &:= \overline{\xi}^{n+1}_{\mathcal{D}}, \qquad \underline{\xi}_{\mathcal{T}_h,\delta t|\mathcal{D}\times(t^n,t^{n+1}]} := \underline{\xi}^{n+1}_{\mathcal{D}},\end{aligned}$$

where $\mathcal{V}_{\mathcal{D}}$ stands for the set of vertices of the diamond $\mathcal{D}$. The function $\xi$ is increasing and continuous on $[0,1]$ then its inverse is continuous on the compact $[0,\xi(1)]$. Therefore, there exists a modulus of continuity $\omega$ of $v\circ\xi^{-1}$, which is continuous and bounded on the same interval with $\omega(0)=0$. Using this latter fact and the Cauchy-Schwarz inequality yields

$$\begin{aligned}
\left|\mathcal{S}^{3,1}_{\mathcal{T}_h,\delta t}\right| &\leq C\,\|\nabla\psi\|_\infty \sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h} m_{\mathcal{D}}\,\omega\Big(\overline{\xi}_{\mathcal{D}}^{n+1}-\underline{\xi}_{\mathcal{D}}^{n+1}\Big)\ \left|\nabla^{\mathcal{D}}\xi_{h,\delta t}\right| \\
&\leq C\left(\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h} m_{\mathcal{D}}\,\omega\Big(\overline{\xi}_{\mathcal{D}}^{n+1}-\underline{\xi}_{\mathcal{D}}^{n+1}\Big)^2\right)^{1/2}\times\left(\sum_{n=0}^{N-1}\delta t\sum_{\mathcal{D}\in\mathfrak{D}_h} m_{\mathcal{D}}\ \left|\nabla^{\mathcal{D}}\xi_{h,\delta t}\right|^2\right)^{1/2} \\
&\leq C\left(\int_{Q_{\mathfrak{T}}}\omega\Big(\overline{\xi}_{\mathcal{T}_h,\delta t}-\underline{\xi}_{\mathcal{T}_h,\delta t}\Big)^2\right)^{1/2}\times\left(\sum_{n=0}^{N-1}\delta t\left\|\nabla^{\mathfrak{D}}\xi_h^{n+1}\right\|_2^2\right)^{1/2},
\end{aligned}$$

for some positive constant $C$. In view of Lemma A.0.2 together with (4.4.2), we deduce that

$$\lim_{h_{\mathfrak{D}},\delta t\to 0}\mathcal{S}^{3,1}_{\mathcal{T}_h,\delta t}=0.$$

Similarly, we establish that

$$\lim_{h_{\mathfrak{D}},\delta t\to 0}\mathcal{S}^{3,2}_{\mathcal{T}_h,\delta t}=0.$$

Finally, let us demonstrate that

$$\lim_{h_{\mathfrak{D}},\delta t\to 0}\mathcal{S}^{4}_{\mathcal{T}_h,\delta t}=0.$$

Owing to the definition of the penalization term we explore

$$\begin{aligned}
\sum_{n=0}^{N-1}\delta t\left|[\![\mathcal{P}u_{\mathcal{T}_h}^{n+1},\psi_{\mathcal{T}_h}^{n+1}]\!]_{\mathcal{T}_h}\right| &= \left|\frac{1}{2}\frac{1}{h_{\mathfrak{D}}^{\varepsilon}}\sum_{n=0}^{N-1}\delta t\sum_{K^*\in\overline{\mathfrak{M}^*}}\sum_{K\in\mathfrak{M}} m_{K\cap K^*}\Big(F(u_K^{n+1})-F(u_{K^*}^{n+1})\Big)\Big(\psi_K^{n+1}-\psi_{K^*}^{n+1}\Big)\right| \\
&\leq \frac{1}{2}\frac{\|v\|_\infty}{h_{\mathfrak{D}}^{\varepsilon}}\sum_{n=0}^{N-1}\delta t\sum_{K^*\in\overline{\mathfrak{M}^*}}\sum_{K\in\mathfrak{M}} m_{K\cap K^*}\left|\xi_K^{n+1}-\xi_{K^*}^{n+1}\right|\left|\psi_K^{n+1}-\psi_{K^*}^{n+1}\right| \\
&\leq \frac{1}{2}\frac{\|v\|_\infty}{h_{\mathfrak{D}}^{\varepsilon}}\left\|\xi_{\mathfrak{M}_h,\delta t}-\xi_{\overline{\mathfrak{M}^*}_h,\delta t}\right\|_{L^2(Q_{\mathfrak{T}})}\left\|\psi_{\mathfrak{M}_h,\delta t}-\psi_{\overline{\mathfrak{M}^*}_h,\delta t}\right\|_{L^2(Q_{\mathfrak{T}})}.
\end{aligned}$$

On the other hand, the regularity of the function $\psi$ ensures the existence of a constant $C$ depending only on the regularity of the mesh such that (see [46] for deep details)

$$\left\|\psi_{\mathfrak{M}_h,\delta t}-\psi_{\overline{\mathfrak{M}^*}_h,\delta t}\right\|_{L^2(Q_{\mathfrak{T}})}\leq Ch_{\mathfrak{D}}\,\|\psi\|_{W^{1,\infty}(\Omega)}.$$

Utilizing the energy estimate (4.4.2) and the fact that $\varepsilon<2$ we obtain

$$\left|\mathcal{S}^{4}_{\mathcal{T}_h,\delta t}\right|\leq Ch_{\mathfrak{D}}^{1-\varepsilon/2}\longrightarrow 0,\quad h_{\mathfrak{D}},\delta t\longrightarrow 0.$$

This ends the proof of the theorem. □

## 4.8 Numerical results

In this section, we present some numerical tests so that we can show the efficiency and the stability of the proposed DDFV scheme. As highlighted in the introduction of this chapter, this method will allow us to take into account almost general meshes and any tensor. We also stress that boundary conditions of Dirichlet type are prescribed. It is sufficient to take the trace of a given exact solution on the boundary. This particularity provides analytical solutions of the continuous problem and enables us to compare them with the discrete ones.

To begin with, let us consider the unit square $\Omega = [0,1]^2$ as the domain of our study. Next, the primal meshes are given by a sequence of distorted quadrangulation, refined Kershaw and triangular meshes of $\Omega$. The first family is denoted by $M1$ while the second one is denoted by $M2$. These kinds of meshes are taken from the FVCA5 benchmark [88]. Their corresponding dual meshes are constructed as described in Section 4.2.

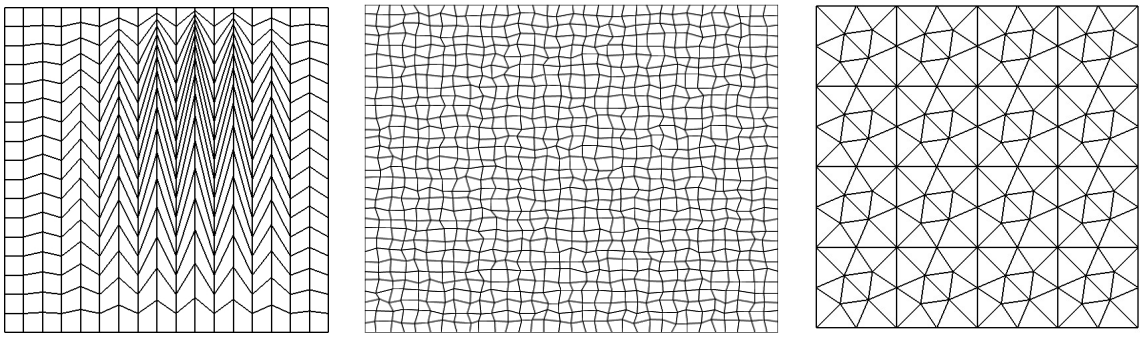


Figure 4.3: From left to right, Kershaw quadrangle and triangular meshes.

Furthermore, the mobility function is chosen as follows

$$f(u) = u^m(1-u)^m, \quad \forall u \in [0,1] \quad \text{and} \quad m \in \{1,2\}.$$

Notice that this function presents some degeneracy in $u = 0$ and $u = 1$. Additionally, we require the computation of the functions $v_\uparrow(u)$ and $v_\downarrow(u)$ in order to calculate the numerical flux. In our study, the function $v$ admits a unique global maximum $\overline{u} = 1/2$. Hence, one gets in a straightforward way that

$$v_\uparrow(u) = v\Big(\min\{u, \frac{1}{2}\}\Big), \quad \text{and } v_\downarrow(u) = v\Big(\max\{u, \frac{1}{2}\}\Big) - v(\frac{1}{2}), \quad \text{for all } u \in (0,1)^2.$$

We also focus on the case of anisotropic media to verify the validity of our discretization. To this end, we select a diagonal tensor $\Lambda$ :

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & 0 \\ 0 & \Lambda_{yy} \end{pmatrix}.$$

The DDFV scheme is formulated in a nonlinear algebraic system, which is solved thanks to Newton's method with a given tolerance $\varepsilon = 1.e^{-10}$. We underline that the numerical scheme (4.3.10)-(4.3.11) is fully implicit in time, unconditionally stable and convergent. Yet, we require the time step to be proportional to the square of the mesh size as mentioned in [39] to assess

numerical error estimates.

As we are interested in the accuracy of the scheme, we are going to evaluate the error of the proposed discretization. In all the tests, we denote by *ERL2* the difference between the analytical solution and the numerical one in $L^\infty(0,T;L^2(\Omega))$-norm. Moreover, we study the error between the gradients of the semi Kirchoff functions in $L^2(\Omega\times(0,T))^2$, which is denoted by *ERGL2*. The convergence rate will be designated by *Rate*. More precisely

$$ERL2 = ||u_{\text{ex}} - u_{h,\delta t}||_{L^\infty(0,T;L^2(\Omega))}, \qquad ERGL2 = ||\nabla\xi(u_{\text{ex}}) - \nabla\xi(u_{h,\delta t})||_{L^2(\Omega\times(0,T))^2}.$$

$$Rate = \frac{\log\left(Err^{i+1}/Err^i\right)}{\log\left(h_{\mathfrak{D}}^{i+1}/h_{\mathfrak{D}}^i\right)}, \quad Err = ERL2, ERGL2,$$

where $i$ refers to the index of the space discretization $\mathcal{T}_i$ for $i = 1,\cdots,5$. In all the tables below $u_{\min}$ (resp. $u_{\max}$) stands for the minimum (resp. maximum) of the computed solution.

### 4.8.1 Test 1

In this test, we investigate the numerical convergence of the DDFV scheme (4.3.9)-(4.3.11) using the exact solution:

$$u_{\text{ex}}(x,t) = 80x_1^2(1-x_1)^2 \times t, \quad \forall x = (x_1,x_2) \in \Omega, \quad t \in (0,T). \tag{4.8.1}$$

Substituting this expression in the main problem (4.1.1) yields a nonnegative source term. One notices that this solution degenerates at the line $\{x_1 = 0\}$ and at $\{x_1 = 1\}$. The mobility function $f(u) = u^2(1-u)^2$ is considered. Here, the final time is fixed to $T = 0.15$.

| | | $\gamma = 0$ | | $\gamma = 0.5$ | |
|---|---|---|---|---|---|
| $h$ | # Unknowns | $\|\|u_{\mathfrak{M},\delta t} - u_{\overline{\mathfrak{M}^*},\delta t}\|\|$ | Rate | $\|\|u_{\mathfrak{M},\delta t} - u_{\overline{\mathfrak{M}^*},\delta t}\|\|$ | Rate |
| 0.3420 | 41 | 0.111 E-01 | - | 0.110 E-01 | - |
| 0.1740 | 145 | 0.575 E-02 | 0.974 | 0.574 E-02 | 0.973 |
| 0.0920 | 545 | 0.296 E-02 | 1.034 | 0.295 E-02 | 1.034 |
| 0.0470 | 2113 | 0.146 E-02 | 1.059 | 0.146 E-02 | 1.059 |
| 0.0195 | 8321 | 0.705 E-03 | 0.823 | 0.705 E-03 | 0.823 |

Table 4.1: The norm $||u_{\mathfrak{M},\delta t} - u_{\overline{\mathfrak{M}^*},\delta t}||_{L^2(Q_{\mathfrak{T}})}$ with and without penalization term for $n = m = 2$.

First, we have seen that the penalization term has played a crucial role to establish that the two reconstructions of the solution on the primal and dual meshes converge to the same limit. Second, this fact holds numerically without the penalty term. To see this, we compute the difference in $L^2(\Omega\times(0,T))$ norm between the approximate solution on the primal mesh and that on the dual mesh. For this, we consider two values of the stabilization parameter $\gamma = 0$ and $\gamma = 0.5$ with a fixed $\varepsilon = 1$. As shown in Table 4.1, the presence or the absence of the penalization term does not influence the convergence of the sequence $||u_{\mathfrak{M},\delta t} - u_{\overline{\mathfrak{M}^*},\delta t}||_{L^2(\Omega\times(0,T))}$. One can as well check that the convergence rate is almost one.

Since the penalty term turns out to be useless numerically then we set the parameter $\gamma$ to zero in the sequel. Let us now return back to the accuracy assessment of the scheme using the exact

solution (4.8.1). In Table 4.2 we list the obtained results with an isotropic tensor $\Lambda_{xx} = \Lambda_{yy} = 1$. We can observe that the convergence rate of the solution is almost of second order for both kinds of meshes. We thus reach the well known order of DDFV schemes for linear problems [39, 59, 90]. Despite of being of order between 1 and 2 for linear problems, the convergence rate of the discrete gradient may be deteriorated with respect to the nonlinearity, the anisotropy and/or the discretization error. For instance we refer to [11] where the authors have found an accuracy of order 0.4 for an anisotropic Laplace equation. Here, for our nonlinear problem, we observe that the convergence rate of the gradient is close to 1 in the case of the mesh family $M1$ whereas it is close to 2 for the Kershaw meshes. We also verify that the computed solution preserves a maximum principle property. Table 4.3 gives the errors in the anisotropic case where the tensor entries are $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 0.01$. It demonstrates that the numerical solution is always nonnegative with convergence rates which are slightly similar to the isotropic case.

| M1 | | | | | | |
|---|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.3420 | 0.127 E-01 | - | 0.324 E-01 | - | 0 | 0.703 |
| 0.1740 | 0.629 E-02 | 1.048 | 0.218 E-01 | 0.590 | 0 | 0.747 |
| 0.0920 | 0.216 E-02 | 1.669 | 0.134 E-02 | 0.755 | 0 | 0.748 |
| 0.0470 | 0.665 E-03 | 1.766 | 0.799 E-02 | 0.781 | 0 | 0.750 |
| 0.0195 | 0.126 E-03 | 1.880 | 0.345 E-02 | 0.947 | 0 | 0.750 |
| M2 | | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.2710 | 0.135 E-02 | - | 0.996 E-01 | - | 0 | 0.703 |
| 0.1355 | 0.369 E-03 | 1.870 | 0.265 E-01 | 1.910 | 0 | 0.744 |
| 0.0903 | 0.168 E-03 | 1.934 | 0.119 E-01 | 1.975 | 0 | 0.747 |
| 0.0677 | 0.959 E-04 | 1.954 | 0.671 E-02 | 1.990 | 0 | 0.749 |
| 0.0542 | 0.619 E-04 | 1.964 | 0.430 E-02 | 1.995 | 0 | 0.750 |

Table 4.2: Numerical convergence with isotropic tensor and $n = m = 2$.

| M1 | | | | | | |
|---|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.342 | 0.130 E-01 | - | 0.327 E-01 | - | 0 | 0.736 |
| 0.174 | 0.649 E-02 | 1.030 | 0.221 E-01 | 0.583 | 0 | 0.748 |
| 0.092 | 0.244 E-02 | 1.529 | 0.142 E-01 | 0.695 | 0 | 0.749 |
| 0.047 | 0.898 E-03 | 1.499 | 0.909 E-02 | 0.666 | 0 | 0.751 |
| 0.0195 | 0.180 E-03 | 1.812 | 0.416 E-02 | 0.883 | 0 | 0.750 |
| M2 | | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.2710 | 0.146 E-02 | - | 0.997 E-01 | - | 0 | 0.739 |
| 0.1355 | 0.402 E-03 | 1.856 | 0.267 E-01 | 1.903 | 0 | 0.746 |
| 0.0903 | 0.184 E-03 | 1.928 | 0.120 E-01 | 1.923 | 0 | 0.748 |
| 0.0677 | 0.105 E-03 | 1.948 | 0.683 E-02 | 1.967 | 0 | 0.749 |
| 0.0542 | 0.680 E-04 | 1.957 | 0.441 E-02 | 1.963 | 0 | 0.750 |

Table 4.3: Numerical convergence with anisotropic tensor and $n = m = 2$.

### 4.8.2 Test 2

We now test the accuracy and the stability of our scheme thanks to the analytical solution

$$u_{\text{ex}}(x,t) = 6x_1^2 \times t, \quad \forall x = (x_1, x_2) \in \Omega, \quad t \in (0,T),$$

where the mobility function is chosen to be $f(u) = u(1-u)$. Note that this function is not a perfect square with $f(0) = f(1) = 0$. This solution fulfills the continuous problem (4.1.1) with a corresponding source term, which is also nonnegative. It vanishes at the line $\{x_1 = 0\}$. The final time is taken as $T = 0.15$. Tables 4.4 and 4.5 present the numerical convergence of the scheme including the isotropic tensor, and anisotropic one (with $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 0.001$) respectively. On the mesh family $M1$, the first table shows that the numerical scheme is accurate of almost second order whereas the second one exhibits an accuracy of order 1.5 which might be explained by the impact of anisotropy. Notwithstanding the distortion of the mesh family $M_2$, we get a super-convergence for the solution and the gradient of its semi Kirchoff transform. In both cases we have not recorded any undershoots nor overshoots.

| M1 | | | | | | |
|---|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{\min}$ | $u_{\max}$ |
| 0.342 | 0.104 E-01 | - | 0.367 E-01 | - | 0 | 0.840 |
| 0.174 | 0.425 E-02 | 1.335 | 0.242 E-01 | 0.622 | 0 | 0.895 |
| 0.092 | 0.132 E-02 | 1.821 | 0.138 E-01 | 0.878 | 0 | 0.897 |
| 0.047 | 0.365 E-03 | 1.933 | 0.696 E-02 | 1.026 | 0 | 0.900 |
| 0.0195 | 0.114 E-03 | 1.312 | 0.385 E-02 | 0.667 | 0 | 0.900 |
| M2 | | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{\min}$ | $u_{\max}$ |
| 0.2710 | 0.124 E-02 | - | 0.102 E-00 | - | 0 | 0.882 |
| 0.1355 | 0.365 E-03 | 1.767 | 0.357 E-01 | 1.519 | 0 | 0.892 |
| 0.0903 | 0.173 E-03 | 1.849 | 0.212 E-01 | 1.286 | 0 | 0.897 |
| 0.0677 | 0.100 E-03 | 1.890 | 0.151 E-01 | 1.171 | 0 | 0.898 |
| 0.0542 | 0.654 E-04 | 1.914 | 0.118 E-01 | 1.111 | 0 | 0.900 |

Table 4.4: Numerical convergence with isotropic tensor and $n = m = 1$.

### 4.8.3 Test 3

This test concerns the porous medium equation. First, we compare our scheme with the following two dimensional exact solution [41] to the main problem (4.1.1)

$$u_{\text{ex}}(x,t) = \frac{\lambda_1(x_1 - 0.5)^2 + \lambda_2(x_2 - 0.5)^2}{1-t}, \quad \forall x = (x_1, x_2) \in \Omega, t \in (0,T),$$

with $\lambda_1 = \frac{1}{16\Lambda_{xx}}$ and $\lambda_2 = \frac{1}{16\Lambda_{yy}}$. The mobility function is $f(u) = 2u$. Note that this choice does not match with the assumption $(A_2)$. We then record the numerical convergence results in Table 4.6 and Table 4.7 with a final time set to $T = 0.2$. On the first mesh sequence, one can check that the method is accurate of second order even in the presence of anisotropy ($\Lambda_{xx} = 0.1$ and $\Lambda_{yy} = 10$). Analogous results have been observed in [41] for the same problem using a VAG (Vertex Approximate Gradient) scheme. In contrast, the super-convergence is lost in the isotropic

| M1 | | | | | | |
|---|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.342 | 0.116 E-01 | - | 0.380 E-01 | - | 0 | 0.840 |
| 0.174 | 0.506 E-02 | 1.245 | 0.263 E-01 | 0.547 | 0 | 0.895 |
| 0.092 | 0.199 E-02 | 1.459 | 0.173 E-01 | 0.659 | 0 | 0.897 |
| 0.047 | 0.754 E-03 | 1.453 | 0.108 E-01 | 0.703 | 0 | 0.900 |
| 0.0195 | 0.207 E-03 | 1.459 | 0.580 E-02 | 0.701 | 0 | 0.900 |
| M2 | | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ | $u_{max}$ |
| 0.2710 | 0.210 E-02 | - | 0.118 E-00 | - | 0 | 0.881 |
| 0.1355 | 0.672 E-03 | 1.646 | 0.533 E-01 | 1.148 | 0 | 0.893 |
| 0.0903 | 0.326 E-03 | 1.783 | 0.359 E-01 | 0.974 | 0 | 0.897 |
| 0.0677 | 0.193 E-03 | 1.833 | 0.272 E-01 | 0.961 | 0 | 0.898 |
| 0.0542 | 0.127 E-03 | 1.852 | 0.220 E-01 | 0.966 | 0 | 0.900 |

Table 4.5: Numerical convergence with anisotropic tensor and $n = m = 1$.

case for the second family of meshes. This is due to the severe distortion of the mesh in the $x_2$-direction. As expected, the second order is recovered in the anisotropic case since the contribution of the term in $x_1$ is less important. In any case, one can see that the method preserves the positivity.

| M1 | | | | | |
|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ |
| 0.342 | 0.426 E-03 | - | 0.945 E-02 | - | 0.206 E-03 |
| 0.174 | 0.260 E-03 | 0.733 | 0.773 E-02 | 0.299 | 0.243 E-04 |
| 0.092 | 0.789 E-04 | 1.860 | 0.462 E-02 | 0.803 | 0.304 E-05 |
| 0.047 | 0.213 E-04 | 1.965 | 0.252 E-02 | 0.909 | 0.612 E-06 |
| 0.0195 | 0.450 E-05 | 1.755 | 0.115 E-02 | 0.882 | 0.234 E-06 |
| M2 | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ |
| 0.2710 | 0.410 E-03 | - | 0.679 E-01 | - | 0.537 E-05 |
| 0.1355 | 0.250 E-03 | 0.713 | 0.481 E-01 | 0.495 | 0.108 E-04 |
| 0.0903 | 0.186 E-03 | 0.732 | 0.367 E-01 | 0.667 | 0.554 E-05 |
| 0.0677 | 0.149 E-03 | 0.764 | 0.294 E-01 | 0.773 | 0.739 E-06 |
| 0.0542 | 0.125 E-03 | 0.802 | 0.244 E-01 | 0.831 | 0.507 E-08 |

Table 4.6: Numerical convergence of the scheme with $\Lambda_{xx} = \Lambda_{yy} = 1$.

Finally, we provide an example which exhibits a low space regularity due to the degenerate nature of the considered problem. This test has been also treated in [41] using the VAG discretization. It is about the one dimensional weak solution

$$u_{\text{ex}}(x,t) = \max(2\Lambda_{xx}t - x_1, 0) \quad \forall x = (x_1, x_2), \in \Omega, t \in (0,T),$$

to the porous medium equation (4.1.1) (we recall $f(u) = 2u$) complemented with the Dirichlet boundary condition corresponding to this exact solution. In this test-case, we consider a sequence of refined triangulations of $\Omega$ as primal meshes. We take $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 10$. The final time is

| M1 | | | | | |
|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ |
| 0.342 | 0.340 E-01 | - | 0.103 E-00 | - | 0.734 E-03 |
| 0.174 | 0.123 E-01 | 1.516 | 0.621 E-01 | 0.752 | 0.131 E-03 |
| 0.092 | 0.336 E-02 | 2.022 | 0.335 E-01 | 0.963 | 0.178 E-04 |
| 0.047 | 0.847 E-03 | 2.068 | 0.168 E-01 | 1.040 | 0.449 E-05 |
| 0.0195 | 0.222 E-03 | 1.509 | 0.884 E-02 | 0.728 | 0.234 E-06 |
| M2 | | | | | |
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ |
| 0.2710 | 0.286 E-02 | - | 0.251 E-00 | - | 0.130 E-05 |
| 0.1355 | 0.713 E-03 | 2.004 | 0.119 E-00 | 0.987 | 0.108 E-05 |
| 0.0903 | 0.317 E-03 | 2.000 | 0.794 E-01 | 0.991 | 0.563 E-06 |
| 0.0677 | 0.179 E-03 | 1.991 | 0.598 E-01 | 0.988 | 0.739 E-07 |
| 0.0542 | 0.115 E-03 | 1.990 | 0.479 E-01 | 0.990 | 0.172 E-08 |

Table 4.7: Numerical convergence of the scheme with $\Lambda_{xx} = 0.1$ and $\Lambda_{yy} = 10$.

$T = 0.25$. The obtained results are given in Table 4.8. As expected, it is shown that the discrete solution is nonnegative. It additionally converges with an order strictly less than 2 because of the anisotropy and its low regularity. This phenomenon has been also indicated in [41].

| Triangular meshes | | | | | |
|---|---|---|---|---|---|
| h | ERL2 | Rate | ERGL2 | Rate | $u_{min}$ |
| 0.250 | 0.176 E-01 | - | 0.165 E-00 | - | 0 |
| 0.125 | 0.106 E-01 | 0.728 | 0.971 E-01 | 0.761 | 0 |
| 0.063 | 0.583 E-02 | 0.865 | 0.612 E-01 | 0.667 | 0 |
| 0.031 | 0.324 E-02 | 0.850 | 0.386 E-01 | 0.663 | 0 |
| 0.017 | 0.177 E-02 | 0.875 | 0.242 E-01 | 0.674 | 0 |

Table 4.8: Numerical convergence of the scheme with $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 10$.

# Conclusion & Perspectives

In conclusion, in this thesis we have first studied a couple of finite volume schemes of CVFE type for solving the governing equations of the two-phase flow model in anisotropic porous media. Secondly, we developed a positive discrete duality finite volume scheme for the approximation of degenerate parabolic equations on almost 2D general meshes. The basic idea for the analysis of both methods is to approximate the fluxes properly thanks to monotone schemes bearing in mind two main points. The first one consists in preserving the natural bounds of the solutions which makes the latters meaningful from a physical point of view. The second point is to derive some a priori estimates on the discrete gradients which is not an easy task especially when the coercivity property is lost. This has led to make use of the upstream techniques so that this central property can be recovered. In addition to these two fundamental elements, establishing classical compactness arguments has played an essential role to carry out the convergence analysis of the numerical scheme. Numerical experiments exhibited the ability of the schemes belonging to the CVFE family to efficiently simulate the displacement of water through the porous medium in question. They also gave interesting evidences on how the fluid moves in anisotropic media whose ratio of anisotropy is important. On the other hand, the implementation of the proposed DDFV scheme on too distorted meshes and anisotropic tensors showed spurring results. Indeed, the method turned out to be accurate of almost second order even if the scheme is based on upwinding techniques which lead generally to an accuracy of first order.

As an outlook of this thesis, we can always envisage the compressible two-phase flow model in porous media with no major restrictions on the physical data. We may particularly consider the case where the density depends on its own pressure. This case presents a challenging task since the coercivity property does not hold due to the non-positivity of the stiffnesses coefficients. We then need to design more reliable finite volume schemes ensuring the aforementioned two points in order to address this issue. This will be the object of future works based on the DDFV framework.

# Appendix A

# Technical lemmas

Let $\mathcal{D}$ be a fixed diamond cell. We define the following $2 \times 2$ matrices

$$\mathbb{A}^{\mathcal{D}} = \frac{1}{4m_{\mathcal{D}}} \begin{bmatrix} m_\sigma^2 & m_\sigma m_{\sigma^*} \\ m_\sigma m_{\sigma^*} & m_{\sigma^*}^2 \end{bmatrix} =: \begin{bmatrix} \mathbb{A}_\sigma^{\mathcal{D}} & \mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D}} \\ \mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D}} & \mathbb{A}_{\sigma^*}^{\mathcal{D}} \end{bmatrix}, \tag{A.0.1}$$

$$\begin{aligned} \mathbb{A}^{\mathcal{D},\Lambda} &= \frac{1}{4m_{\mathcal{D}}} \begin{bmatrix} m_\sigma^2 \Lambda \mathbf{n}_{\sigma K} \cdot \mathbf{n}_{\sigma K} & m_\sigma m_{\sigma^*} \Lambda \mathbf{n}_{\sigma K} \cdot \mathbf{n}_{\sigma^* K^*} \\ m_\sigma m_{\sigma^*} \Lambda \mathbf{n}_{\sigma K} \cdot \mathbf{n}_{\sigma^* K^*} & m_{\sigma^*}^2 \, \Lambda \mathbf{n}_{\sigma^* K^*} \cdot \mathbf{n}_{\sigma^* K^*} \end{bmatrix} \\ &=: \begin{bmatrix} \mathbb{A}_\sigma^{\mathcal{D}} & \mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D},\Lambda} \\ \mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D},\Lambda} & \mathbb{A}_{\sigma^*}^{\mathcal{D},\Lambda} \end{bmatrix}, \end{aligned} \tag{A.0.2}$$

and

$$\mathbb{B}^{\mathcal{D},\Lambda} = \begin{bmatrix} \left|\mathbb{A}_\sigma^{\mathcal{D}}\right| + \left|\mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D},\Lambda}\right| & 0 \\ 0 & \left|\mathbb{A}_{\sigma,\sigma^*}^{\mathcal{D},\Lambda}\right| + \left|\mathbb{A}_{\sigma^*}^{\mathcal{D},\Lambda}\right| \end{bmatrix}, \quad \forall \mathcal{D} \in \mathfrak{D}. \tag{A.0.3}$$

The following lemma claims a crucial property of the matrix $\mathbb{A}^{\mathcal{D},\Lambda}$. In particular, it states that $\mathbb{A}^{\mathcal{D},\Lambda}$ is positive definite.

**Lemma A.0.1.** *[39] There exist some positive constants $\lambda_0$ and $\lambda_1$ depending only on the mesh regularity and on $\underline{\Lambda}$, $\overline{\Lambda}$ satisfying*

$$\mathbb{A}^{\mathcal{D},\Lambda} x \cdot x \leq \mathbb{B}^{\mathcal{D},\Lambda} x \cdot x \leq \lambda_1 \mathbb{A}^{\mathcal{D},\Lambda} x \cdot x, \quad \forall x \in \mathbb{R}^2, \tag{A.0.4}$$

$$\lambda_0 \mathbb{A}^{\mathcal{D}} x \cdot x \leq \mathbb{A}^{\mathcal{D},\Lambda} x \cdot x, \quad \forall x \in \mathbb{R}^2. \tag{A.0.5}$$

*Proof.* For the sake of completeness we reproduce the same proof as given in [39]. Let $x = (x_1, x_2)$ be a fixed vector of $\mathbb{R}^2$. Thus, for every $\mathcal{D} \in \mathfrak{D}$, we have

$$\mathbb{A}^{\mathcal{D},\Lambda} x \cdot x \leq \mathbb{B}^{\mathcal{D},\Lambda} x \cdot x \leq \left\|\mathbb{A}^{\mathcal{D},\Lambda}\right\|_1 |x|^2 .$$

where $\|.\|_p$ is the usual $p$-norm matrix, $p = 1, 2$. The equivalence of norms on $\mathbb{R}^2 \times \mathbb{R}^2$, ensures the existence of a coefficient $L \geq 1$ such that

$$\left\|\mathbb{A}^{\mathcal{D},\Lambda}\right\|_1 |x|^2 \leq L \left\|\mathbb{A}^{\mathcal{D},\Lambda}\right\|_2 |x|^2 \leq L \ \mathrm{Cond}_2(\mathbb{A}^{\mathcal{D},\Lambda}) \mathbb{A}^{\mathcal{D},\Lambda} x \cdot x, \tag{A.0.6}$$

where $\mathrm{Cond}_2$ stands for the condition number with respect to the 2-norm. In addition, this number can be overestimated as follows

$$\mathrm{Cond}_2(\mathbb{A}^{\mathcal{D},\Lambda}) \leq \mathrm{Cond}_2(\Lambda^{\mathcal{D}})\left(\mathcal{Q}_{\mathcal{D}}^2 + \sqrt{\mathcal{Q}_{\mathcal{D}}^2 - \frac{1}{\mathrm{Cond}_2(\Lambda^{\mathcal{D}})}}\right)^2 \leq 4\mathrm{reg}(\mathcal{T})^2 \times \frac{\overline{\Lambda}}{\underline{\Lambda}}.$$

where we have set

$$\mathcal{Q}_{\mathcal{D}} = \frac{1}{2\sin(\alpha_{\mathcal{D}})}\Big(\frac{m_\sigma}{m_{\sigma^*}} + \frac{m_{\sigma^*}}{m_\sigma}\Big) \geq 1.$$

This proves the inequality

$$\mathbb{A}^{\mathcal{D},\Lambda}x \cdot x \leq \mathbb{B}^{\mathcal{D},\Lambda}x \cdot x \leq \lambda_1 \mathbb{A}^{\mathcal{D},\Lambda}x \cdot x,$$

with $\lambda_1 = 4L\,\mathrm{reg}(\mathcal{T})^2 \times \Big(\overline{\Lambda}/\underline{\Lambda}\Big)$. Using the elementary inequality

$$\big|\mathbb{A}^{\mathcal{D}}_{\sigma,\sigma^*}\big|\, x_1 x_2 \leq \frac{1}{2}\Big(\mathbb{A}^{\mathcal{D}}_{\sigma} x_1^2 + \mathbb{A}^{\mathcal{D}}_{\sigma^*} x_2^2\Big),$$

one deduces that

$$\mathbb{A}^{\mathcal{D}}x \cdot x \leq 2\Big(\mathbb{A}^{\mathcal{D}}_{\sigma} x_1^2 + \mathbb{A}^{\mathcal{D}}_{\sigma^*} x_2^2\Big).$$

Now the ellipticity of the tensor $\Lambda$ implies

$$\mathbb{A}^{\mathcal{D}}x \cdot x \leq \frac{2}{\underline{\Lambda}}\Big(\mathbb{A}^{\mathcal{D}}_{\sigma} x_1^2 + \mathbb{A}^{\mathcal{D},\Lambda}_{\sigma^*} x_2^2\Big) \leq \frac{2}{\underline{\Lambda}}\mathbb{B}^{\mathcal{D},\Lambda}x \cdot x.$$

Thanks to inequality (A.0.4), one gets

$$\mathbb{A}^{\mathcal{D}}x \cdot x \leq \frac{2\lambda_1}{\underline{\Lambda}}\mathbb{A}^{\mathcal{D},\Lambda}x \cdot x.$$

Hence, the second inequality follows by setting $\lambda_0 = \dfrac{\underline{\Lambda}}{2\lambda_1}$. □

**Lemma A.0.2.** *Consider the following piecewise constant functions*

$$\overline{\xi}_{\mathcal{D}}^{n+1} := \max_{M\in\mathcal{V}_{\mathcal{D}}}\{\xi(u_M^{n+1})\}, \qquad \underline{\xi}_{\mathcal{D}}^{n+1} := \min_{M\in\mathcal{V}_{\mathcal{D}}}\{\xi(u_M^{n+1})\},$$
$$\overline{\xi}_{\mathcal{T}_h,\delta t|\mathcal{D}\times(t^n,t^{n+1}]} := \overline{\xi}_{\mathcal{D}}^{n+1}, \qquad \underline{\xi}_{\mathcal{T}_h,\delta t|\mathcal{D}\times(t^n,t^{n+1}]} := \underline{\xi}_{\mathcal{D}}^{n+1},$$

*where we denote* $\mathcal{V}_{\mathcal{D}} = \{K, L, K^*, L^*\}$. *Then*

$$\lim_{h_{\mathcal{D}},\delta t\to 0}\Big\|\overline{\xi}_{\mathcal{T}_h,\delta t} - \underline{\xi}_{\mathcal{T}_h,\delta t}\Big\|_{L^2(Q_T)} = 0. \tag{A.0.7}$$

*Proof.* We first observe that

$$\Big|\overline{\xi}_{\mathcal{D}}^{n+1} - \underline{\xi}_{\mathcal{D}}^{n+1}\Big|^2 \leq |\xi(u_K) - \xi(u_L)|^2 + |\xi(u_{K^*}) - \xi(u_{L^*})|^2 + |\xi(u_K) - \xi(u_{K^*})|^2 + |\xi(u_L) - \xi(u_{L^*})|^2.$$

This implies

$$
\begin{aligned}
&\sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\overline{\xi}_{\mathcal{D}}^{n+1} - \underline{\xi}_{\mathcal{D}}^{n+1}\right|^2 \\
&\leq \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \Big( \left|\xi_K^{n+1} - \xi_L^{n+1}\right|^2 + \left|\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}\right|^2 + \left|\xi_K^{n+1} - \xi_{K^*}^{n+1}\right|^2 + \left|\xi_L^{n+1} - \xi_{L^*}^{n+1}\right|^2 \Big) \\
&\leq h_{\mathfrak{D}}^2 \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \Big( \left|\frac{\xi_K^{n+1} - \xi_L^{n+1}}{m_{\sigma^*}}\right|^2 + \left|\frac{\xi_{K^*}^{n+1} - \xi_{L^*}^{n+1}}{m_{\sigma}}\right|^2 \Big) \\
&\qquad + \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\xi_K^{n+1} - \xi_{K^*}^{n+1}\right|^2 + \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\xi_L^{n+1} - \xi_{L^*}^{n+1}\right|^2 \\
&\leq h_{\mathfrak{D}}^2 \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \Big( \left|\nabla^{\mathcal{D}} \xi_h^{n+1} \cdot \boldsymbol{\tau}_{\boldsymbol{K},\boldsymbol{L}}\right|^2 + \left|\nabla^{\mathcal{D}} \xi_h^{n+1} \cdot \boldsymbol{\tau}_{\boldsymbol{K^*},\boldsymbol{L^*}}\right|^2 \Big) \\
&\qquad + \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\xi_K^{n+1} - \xi_{K^*}^{n+1}\right|^2 + \sum_{\mathcal{D}\in\mathfrak{D}} m_{\mathcal{D}} \left|\xi_L^{n+1} - \xi_{L^*}^{n+1}\right|^2 .
\end{aligned}
$$

Due to the estimates (4.4.2) and (4.6.7) one concludes that

$$
\lim_{h_{\mathfrak{D}},\delta t\to 0} \left\| \overline{\xi}_{\mathcal{T}_h,\delta t} - \underline{\xi}_{\mathcal{T}_h,\delta t} \right\|_{L^2(Q_T)} = 0.
$$

□

# Bibliography


[1] Afif, M. and Amaziane, B. (2002). Convergence of finite volume schemes for a degenerate convection–diffusion equation arising in flow in porous media. *Computer Methods in Applied Mechanics and Engineering*, 191(46):5265–5286.

[2] Alt, H. W. and Luckhaus, S. (1983). Quasilinear elliptic-parabolic differential equations. *Mathematische Zeitschrift*, 183(3):311–341.

[3] Amaziane, B. and El Ossmani, M. (2008). Convergence analysis of an approximation to miscible fluid flows in porous media by combining mixed finite element and finite volume methods. *Numerical Methods for Partial Differential Equations*, 24(3):799–832.

[4] Amaziane, B., Jurak, M., and Keko, A. Ž. (2010). Modeling and numerical simulations of immiscible compressible two-phase flow in porous media by the concept of global pressure. *Transport in Porous Media*, 84(1):133–152.

[5] Amaziane, B., Jurak, M., and Keko, A. Ž. (2011). An existence result for a coupled system modeling a fully equivalent global pressure formulation for immiscible compressible two-phase flow in porous media. *Journal of Differential Equations*, 250(3):1685–1718.

[6] Amaziane, B., Jurak, M., and Vrbaški, A. (2012). Existence for a global pressure formulation of water-gas flow in porous media. *Electronic Journal of Differential Equations*, 2012.

[7] Amirat, Y., Hamdache, K., and Ziani, A. (1996). Mathematical analysis for compressible miscible displacement models in porous media. *Mathematical Models and Methods in Applied Sciences*, 6(06):729–747.

[8] Andreianov, B., Bendahmane, M., and Hubert, F. (2013a). On 3D DDFV discretization of gradient and divergence operators: discrete functional analysis tools and applications to degenerate parabolic problems. *Computational Methods in Applied Mathematics*, 13(4):369–410.

[9] Andreianov, B., Bendahmane, M., Hubert, F., and Krell, S. (2012). On 3D DDFV discretization of gradient and divergence operators. I. Meshing, operators and discrete duality. *IMA Journal of Numerical Analysis*, 32(4):1574–1603.

[10] Andreianov, B., Bendahmane, M., and Saad, M. (2011). Finite volume methods for degenerate chemotaxis model. *Journal of Computational and Applied Mathematics*, 235(14):4015–4031.

[11] Andreianov, B., Boyer, F., and Hubert, F. (2007). Discrete duality finite volume schemes for Leray-Lions- type elliptic problems on general 2D meshes. *Numerical Methods for Partial Differential Equations*, 23(1):145–195.

[12] Andreianov, B., Eymard, R., Ghilani, M., and Marhraoui, N. (2013b). Finite volume approximation of degenerate two-phase flow model with unlimited air mobility. *Numerical Methods for Partial Differential Equations*, 29(2):441–474.

[13] Andreianov, B. A., Gutnic, M., and Wittbold, P. (2004). Convergence of finite volume approximations for a nonlinear elliptic-parabolic problem: A "continuous" approach. *SIAM Journal on Numerical Analysis*, 42(1):228–251.

[14] Angelini, O., Brenner, K., and Hilhorst, D. (2013). A finite volume method on general meshes for a degenerate parabolic convection–reaction–diffusion equation. *Numerische Mathematik*, 123(2):219–257.

[15] Arbogast, T. (1992). The existence of weak solutions to single porosity and simple dual-porosity models of two-phase incompressible flow. *Nonlinear Anal*, 19(11):1009–1031.

[16] Arbogast, T. and Wheeler, M. F. (1996). A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM Journal on Numerical Analysis*, 33(4):1669–1687.

[17] Aziz, K. and Settari, A. (1979). *Petroleum reservoir simulation*. Applied Science Publisher LTD, London.

[18] Bachmat, Y. and Bear, J. (1986). Macroscopic modelling of transport phenomena in porous media. 1: The continuum approach. *Transport in porous media*, 1(3):213–240.

[19] Baliga, B. and Patankar, S. (1980). A new finite-element formulation for convection-diffusion problems. *Numerical Heat Transfer*, 3(4):393–409.

[20] Barrett, J. W. and Knabner, P. (1997). Finite element approximation of the transport of reactive solutes in porous media. Part II: Error estimates for equilibrium adsorption processes. *SIAM Journal on Numerical Analysis*, 34(2):455–479.

[21] Bastian, P. (1999). *Numerical computation of multiphase flow in porous media*. PhD thesis, habilitationsschrift Univeristät Kiel.

[22] Bear, J. (1972). *Dynamics of fluids in porous media*. American Elsevier.

[23] Bendahmane, M., Khalil, Z., and Saad, M. (2014). Convergence of a finite volume scheme for gas–water flow in a multi-dimensional porous medium. *Mathematical Models and Methods in Applied Sciences*, 24(01):145–185.

[24] Bessemoulin-Chatard, M., Chainais-Hillairet, C., and Filbet, F. (2014). On discrete functional inequalities for some finite volume schemes. *IMA Journal of Numerical Analysis*, 35(3):1125–1149.

[25] Bessemoulin-Chatard, M. and Filbet, F. (2012). A finite volume scheme for nonlinear degenerate parabolic equations. *SIAM Journal on Scientific Computing*, 34(5):B559–B583.

[26] Binning, P. and Celia, M. A. (1999). Practical implementation of the fractional flow approach to multi-phase flow simulation. *Advances in Water Resources*, 22(5):461–478.

[27] Brenner, K. and Cancès, C. (2017). Improving Newton's Method Performance by Parametrization: The Case of the Richards Equation. *SIAM Journal on Numerical Analysis*, 55(4):1760–1785.

[28] Brenner, K. and Masson, R. (2013). Convergence of a vertex centred discretization of two-phase darcy flows on general meshes. *International Journal on Finite Volumes*, 10:1–37.

[29] Brenner, S. and Scott, R. (2007). *The mathematical theory of finite element methods*, volume 15. Springer Science & Business Media.

[30] Brezis, H. (2010). *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media.

[31] Brooks, R. and Corey, T. (1964). Hydraulic properties of porous media. Tech. Report, Hydrology Paper 3, Colorado State University, Fort Collins, Colorado, USA.

[32] Brull, S. (2008). Two compressible immiscible fluids in porous media: The case where the porosity depends on the pressure. *Advances in Differential Equations*, 13(7-8):781–800.

[33] Buckley, S. E. and Leverett, M. (1942). Mechanism of fluid displacement in sands. *Transactions of the AIME*, 146(01):107–116.

[34] Cai, Z. (1990). On the finite volume element method. *Numerische Mathematik*, 58(1):713–735.

[35] Cai, Z., Mandel, J., and McCormick, S. (1991). The finite volume element method for diffusion equations on general triangulations. *SIAM Journal on Numerical Analysis*, 28(2):392–402.

[36] Cai, Z. and McCormick, S. (1990). On the accuracy of the finite volume element method for diffusion equations on composite grids. *SIAM Journal on Numerical Analysis*, 27(3):636–655.

[37] Camier, J.-S. and Hermeline, F. (2016). A monotone nonlinear finite volume method for approximating diffusion operators on general meshes. *International Journal for Numerical Methods in Engineering*, 107(6):496–519.

[38] Cancès, C., Cathala, M., and Le Potier, C. (2013). Monotone corrections for generic cell-centered finite volume approximations of anisotropic diffusion equations. *Numerische Mathematik*, 125(3):387–417.

[39] Cancès, C., Chainais-Hillairet, C., and Krell, S. (2018). Numerical analysis of a nonlinear free-energy diminishing discrete duality finite volume scheme for convection diffusion equations. *Computational Methods in Applied Mathematics*, 18(3):407–432.

[40] Cancès, C. and Guichard, C. (2016). Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Mathematics of Computation*, 85(298):549–580.

[41] Cancès, C. and Guichard, C. (2017). Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Foundations of Computational Mathematics*, 17(6):1525–1584.

[42] Cancès, C., Ibrahim, M., and Saad, M. (2017). Positive nonlinear cvfe scheme for degenerate anisotropic keller-segel system. *SMAI Journal of Computational Mathematics*, 3:1–28.

[43] Cancès, C., Pop, I., and Vohralík, M. (2014). An a posteriori error estimate for vertex-centered finite volume discretizations of immiscible incompressible two-phase flow. *Mathematics of Computation*, 83(285):153–188.

[44] Cancès, C. and Pierre, M. (2012). An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM Journal on Mathematical Analysis*, 44(2):966–992.

[45] Cariaga, E., Concha, F., Pop, I. S., and Sepúlveda, M. (2010). Convergence analysis of a vertex-centered finite volume scheme for a copper heap leaching model. *Mathematical Methods in the Applied Sciences*, 33(9):1059–1077.

[46] Chainais-Hillairet, C., Krell, S., and Mouton, A. (2015). Convergence analysis of a DDFV scheme for a system describing miscible fluid flows in porous media. *Numerical Methods for Partial Differential Equations*, 31(3):723–760.

[47] Chavent, G. and Jaffré, J. (1986). *Mathematical models and finite elements for reservoir simulation: single phase, multiphase and multicomponent flows through porous media.* 17, North-Holland Publishing Comp.

[48] Chen, Z. (2001). Degenerate two-phase incompressible flow: I. existence, uniqueness and regularity of a weak solution. *Journal of Differential Equations*, 171(2):203–232.

[49] Chen, Z. (2002). Degenerate two-phase incompressible flow II: Regularity, stability and stabilization. *Journal of Differential Equations*, 186(2):345–376.

[50] Chen, Z. (2006). On the control volume finite element methods and their applications to multiphase flow. *Networks and Heterogeneous Media*, 1(4):689.

[51] Chen, Z., Ewing, R., and Espedal, M. (1994). Multiphase flow simulation with various boundary conditions. *Computational Methods in Water Resources*, pages 925–932.

[52] Chen, Z. and Ewing, R. E. (1997). Fully discrete finite element analysis of multiphase flow in groundwater hydrology. *SIAM Journal on Numerical Analysis*, 34(6):2228–2253.

[53] Choquet, C. (2008). On a fully coupled nonlinear parabolic problem modelling miscible compressible displacement in porous media. *Journal of Mathematical Analysis and Applications*, 339(2):1112–1133.

[54] Ciarlet, P. G. (2002). *The finite element method for elliptic problems*. SIAM.

[55] Corey, A. T. (1994). *Mechanics of immiscible fluids in porous media*. Water Resources Publication.

[56] Coudière, Y. and Hubert, F. (2011). A 3D discrete duality finite volume method for nonlinear elliptic equations. *SIAM Journal on Scientific Computing*, 33(4):1739–1764.

[57] Coudière, Y. and Manzini, G. (2010). The discrete duality finite volume method for convection-diffusion problems. *SIAM Journal on Numerical Analysis*, 47(6):4163–4192.

[58] Darcy, H. (1856). *Les fontaines publiques de la ville de Dijon*. Victor Dalmont.

[59] Delcourte, S., Domelevo, K., and Omnès, P. (2005). Discrete duality finite volume method for second order elliptic problems. In *Finite Volumes for Complex Applications IV*, page 447–458. Hermes Science publishing.

[60] Domelevo, K. and Omnès, P. (2005). A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *ESAIM: Mathematical Modelling and Numerical Analysis*, 39(6):1203–1249.

[61] Droniou, J. (2014). Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Mathematical Models and Methods in Applied Sciences*, 24(08):1575–1619.

[62] Droniou, J., Eymard, R., Gallouët, T., Guichard, C., and Herbin, R. (2018). *The gradient discretisation method*, volume 82. Springer.

[63] Ebmeyer, C. (1998). Error estimates for a class of degenerate parabolic equations. *SIAM Journal on Numerical Analysis*, 35(3):1095–1112.

[64] Ern, A. and Guermond, J.-L. (2013). *Theory and practice of finite elements*, volume 159. Springer Science & Business Media.

[65] Evans, L. C. (2010). *Partial differential equations*, volume 19. American Mathematical Society, 2 edition.

[66] Eymard, R., Féron, P., Gallouët, T., Herbin, R., and Guichard, C. (2013). Gradient schemes for the Stefan problem. *International Journal On Finite Volumes*, 10.

[67] Eymard, R. and Gallouët, T. (1993). Convergence d'un schéma de type éléments finis-volumes finis pour un systeme couplé elliptique-hyperbolique, RAIRO Modél. *Math. Anal. Numér*, 27:843–861.

[68] Eymard, R., Gallouët, T., Guichard, C., Herbin, R., and Masson, R. (2014). TP or not TP, that is the question. *Computational Geosciences*, 18(3-4):285–296.

[69] Eymard, R., Gallouët, T., and Herbin, R. (2000). Finite volume methods. In *Handbook of Numerical Analysis*, volume 7, pages 713–1018. Elsevier.

[70] Eymard, R., Gallouët, T., Herbin, R., and Michel, A. (2002). Convergence of a finite volume scheme for nonlinear degenerate parabolic equations. *Numerische Mathematik*, 92(1):41–82.

[71] Eymard, R., Gallouët, T., Hilhorst, D., and Slimane, Y. N. (1998). Finite volumes and nonlinear diffusion equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 32(6):747–761.

[72] Eymard, R., Herbin, R., and Michel, A. (2003). Mathematical study of a petroleum-engineering scheme. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(6):937–972.

[73] Eymard, R., Hilhorst, D., and Vohralík, M. (2006). A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numerische Mathematik*, 105(1):73–131.

[74] Fabrie, P. and Gallouët, T. (2000). Modeling wells in porous media flow. *Mathematical Models and Methods in Applied Sciences*, 10(05):673–709.

[75] Fadimba, K. B. (2007). On existence and uniqueness for a coupled system modeling immiscible flow through a porous medium. *Journal of Mathematical Analysis and Applications*, 328(2):1034–1056.

[76] Feistauer, M., Felcman, J., and Lukacova-Medvid'ova, M. (1997). On the convergence of a combined finite volume-finite element method for nonlinear convection-diffusion problems. *Numerical Methods for Partial Differential Equations*, 13(2):163–190.

[77] Feng, X. (1994). Strong solutions to a nonlinear parabolic system modeling compressible miscible displacement in porous media. *Nonlinear Analysis: Theory, Methods & Applications*, 23(12):1515–1531.

[78] Feng, X. (1995). On existence and uniqueness results for a coupled system modeling miscible displacement in porous media. *Journal of Mathematical Analysis and Applications*, 194(3):883–910.

[79] Filbet, F. (2006). A finite volume scheme for the Patlak–Keller–Segel chemotaxis model. *Numerische Mathematik*, 104(4):457–488.

[80] Forchheimer, P. (1901). Wasserbewegung durch boden. *Z. Ver. Deutsch, Ing.*, 45:1782–1788.

[81] Gagneux, G. and Madaune-Tort, M. (1994). Unicité des solutions faibles d' équations de diffusion-convection. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 318(10):919–924.

[82] Gagneux, G. and Madaune-Tort, M. (1995). *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, volume 22. Springer Science & Business Media.

[83] Galusinski, C. and Saad, M. (2008a). A nonlinear degenerate system modelling water-gas flows in porous media. *Discrete and Continuous Dynamical Systems Series B*, 9(2):281.

[84] Galusinski, C. and Saad, M. (2008b). Two compressible immiscible fluids in porous media. *Journal of Differential Equations*, 244(7):1741–1783.

[85] Galusinski, C., Saad, M., et al. (2004). On a degenerate parabolic system for compressible, immiscible, two-phase flows in porous media. *Advances in Differential Equations*, 9(11-12):1235–1278.

[86] Gao, Z. and Wu, J. (2015). A second-order positivity-preserving finite volume scheme for diffusion equations on general meshes. *SIAM Journal on Scientific Computing*, 37(1):A420–A438.

[87] Helmig, R. (1997). *Multiphase flow and transport processes in the subsurface: a contribution to the modeling of hydrosystems.* Springer, Berlin.

[88] Herbin, R. and Hubert, F. (2008). Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite Volumes for Complex Applications V*, pages 659–692. Wiley.

[89] Hermeline, F. (1998). Une méthode de volumes finis pour les équations elliptiques du second ordre. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 326(12):1433–1436.

[90] Hermeline, F. (2000). A finite volume method for the approximation of diffusion operators on distorted meshes. *Journal of Computational Physics*, 160(2):481–499.

[91] Huber, R. and Helmig, R. (2000). Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media. *Computational Geosciences*, 4(2):141–164.

[92] Ibrahim, M. and Saad, M. (2014). On the efficacy of a control volume finite element method for the capture of patterns for a volume-filling chemotaxis model. *Computers & Mathematics with Applications*, 68(9):1032–1051.

[93] Jäger, W. and Kačur, J. (1991). Solution of porous medium type systems by linear approximation schemes. *Numerische Mathematik*, 60(1):407–427.

[94] Karlsen, K., Risebro, N., and Towers, J. (2002). Upwind difference approximations for degenerate parabolic convection–diffusion equations with a discontinuous coefficient. *IMA Journal of Numerical Analysis*, 22(4):623–664.

[95] Khalil, Z. and Saad, M. (2011a). Degenerate two-phase compressible immiscible flow in porous media: The case where the density of each phase depends on its own pressure. *Mathematics and Computers in Simulation*, 81(10):2225–2233.

[96] Khalil, Z. and Saad, M. (2011b). On a fully nonlinear degenerate parabolic system modeling immiscible gas–water displacement in porous media. *Nonlinear Analysis: Real World Applications*, 12(3):1591–1615.

[97] Krell, S. and Manzini, G. (2012). The Discrete Duality Finite Volume method for the Stokes equations on 3-D polyhedral meshes. *SIAM Journal on Numerical Analysis*, 50(2):808–837.

[98] Kroener, D. and Luckhaus, S. (1984). Flow of oil and water in a porous medium. *Journal of Differential Equations*, 55(2):276–288.

[99] Lazarov, R., Mishev, I. D., and Vassilevski, P. S. (1996). Finite volume methods for convection-diffusion problems. *SIAM Journal on Numerical Analysis*, 33(1):31–55.

[100] Lehmann, F. and Ackerer, P. (1998). Comparison of iterative methods for improved solutions of the fluid flow equation in partially saturated porous media. *Transport in Porous Media*, 31(3):275–292.

[101] Liakopoulos, A. (1965). Darcy's coefficient of permeability as symmetric tensor of second rank. *Hydrological Sciences Journal*, 10(3):41–48.

[102] Lipnikov, K., Shashkov, M., Svyatskiy, D., and Vassilevski, Y. (2007). Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227(1):492–512.

[103] Michel, A. (2003). A finite volume scheme for two-phase immiscible flow in porous media. *SIAM Journal on Numerical Analysis*, 41(4):1301–1317.

[104] Muskat, M., Wyckoff, R. D., et al. (1937). *Flow of homogeneous fluids through porous media*. McGraw-Hill Book Company, Inc.

[105] Nochetto, R., Schmidt, A., and Verdi, C. (2000). A posteriori error estimation and adaptivity for degenerate parabolic problems. *Mathematics of Computation of the American Mathematical Society*, 69(229):1–24.

[106] Ohlberger, M. (1997). Convergence of a mixed finite element: finite volume method for the two phase flow in porous media. *East West Journal of Numerical Mathematics*, 5:183–210.

[107] Osher, S. and Solomon, F. (1982). Upwind difference schemes for hyperbolic systems of conservation laws. *Mathematics of Computation*, 38(158):339–374.

[108] Peaceman, D. (1977). *Fundamentals of numerical reservoir engineering*. Elsevier Applied Science.

[109] Pinder, G. F. and Gray, W. G. (2008). *Essentials of multiphase flow and transport in porous media*. John Wiley & Sons, Inc., Hoboken, New Jersey.

[110] Radu, F. A., Kumar, K., Nordbotten, J. M., and Pop, I. S. (2017). A robust mass conservative scheme for two-phase flow in porous media including Hölder continuous nonlinearities. *IMA Journal of Numerical Analysis*, 38(2):884–920.

[111] Radu, F. A., Nordbotten, J. M., Pop, I. S., and Kumar, K. (2015). A robust linearization scheme for finite volume based discretizations for simulation of two-phase flow in porous media. *Journal of Computational and Applied Mathematics*, 289:134–141.

[112] Radu, F. A., Pop, I., and Knabner, P. (2006). Newton-Type Methods for the Mixed Finite Element Discretization of Some Degenerate Parabolic Equations. In *Numerical mathematics and advanced applications*, pages 1192–1200. Springer.

[113] Saad, B. and Saad, M. (2013). Study of full implicit petroleum engineering finite-volume scheme for compressible two-phase flow in porous media. *SIAM Journal on Numerical Analysis*, 51(1):716–741.

[114] Saad, B. and Saad, M. (2015). A combined finite volume–nonconforming finite element scheme for compressible two phase flow in porous media. *Numerische Mathematik*, 129(4):691–722.

[115] Saad, M. (2014). Slightly compressible and immiscible two-phase flow in porous media. *Nonlinear Analysis: Real World Applications*, 15:12–26.

[116] Van Genuchten, M. (1980). A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Science Society of America Journal*, 44(5):892–898.

[117] Vázquez, J. L. (2007). *The porous medium equation: mathematical theory*. Oxford University Press.

[118] Vohralík, M. and Wheeler, M. F. (2013). A posteriori error estimates, stopping criteria, and adaptivity for two-phase flows. *Computational Geosciences*, 17(5):789–812.

[119] Whitaker, S. (1986). Flow in porous media I: A theoretical derivation of Darcy's law. *Transport in Porous Media*, 1(1):3–25.

[120] Yeh, L.-M. (2006). Hölder continuity for two-phase flows in porous media. *Mathematical Methods in the Applied Sciences*, 29(11):1261–1289.

[121] Yotov, I. (1997). A mixed finite element discretization on non-matching multiblock grids for a degenerate parabolic equation arising in porous media flow. *East West J. Numer. Math.*, 5:211–230.

[122] Yuan, G. and Sheng, Z. (2008). Monotone finite volume schemes for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 227(12):6288–6312.

[123] Zheng, C. and Bennett, G. D. (2002). *Applied contaminant transport modeling*, volume 2. Wiley-Interscience New York.

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

## Titre : Volumes finis/Eléments finis pour des écoulements diphasiques compressibles en milieux poreux hétérogènes et anisotropes

**Mots clés :** milieux poreux, diphasique compressible, immiscible, volumes finis, éléments finis, positif, DDFV.

**Résumé :** Cette thèse est centrée autour du développement et de l'analyse des schémas volumes finis robustes afin d'approcher les solutions du modèle diphasique compressible en milieux poreux hétérogènes et anisotropes. Le modèle à deux phases compressibles comprend deux équations paraboliques dégénérées et couplées dont les variables principales sont la saturation du gaz et la pression globale. Ce système est discrétisé à l'aide de deux méthodes différentes (CVFE et DDFV) qui font partie de la famille des volumes finis.

La première classe à laquelle on s'intéresse consiste à combiner la méthode des volumes finis et celle des éléments finis. Dans un premier temps, on considère un schéma volume finis upwind pour la partie convective et un schéma de type éléments finis conformes pour la diffusion capillaire. Sous l'hypothèse que les coefficients de transmissibilités sont positifs, on montre que la saturation vérifie le principe du maximum et on établit des estimations d'énergies permettant de démontrer la convergence du schéma. Dans un second temps, on a mis en place un schéma positif qui corrige le précédent. Ce schéma est basé sur une approximation des flux diffusifs par le schéma de Godunov. L'avantage est d'établir la bornitude des solutions approchées ainsi que les estimations uniformes sur les gradients discrets sans aucune contrainte ni sur le maillage ni sur la perméabilité. En utilisant des arguments classiques de compacité, on prouve rigoureusement la converge du schéma. Chaque schéma est validé par des simulations numériques qui montrent bien le comportement attendu d'une telle solution.

Concernant la deuxième classe, on s'intéressera tout d'abord à la construction et à l'étude d'un nouveau schéma de type DDFV (Discrete Duality Finite Volume) pour une équation de diffusion non linéaire dégénérée. Cette méthode permet d' avantage de prendre en compte des maillages très généraux et des perméabilités quelconques. L'idée clé de cette discrétisation est d'approcher les flux dans la direction normale par un schéma centré et d'utiliser un schéma décentré dans la direction tangentielle. Par conséquent, on démontre que la solution approchée respecte les bornes physiques et on établit aussi des estimations d'énergie. La convergence du schéma est également établie. Des résultats numériques confirment bien ceux de la théorie. Ils exhibent en outre que la méthode est presque d'ordre deux.

## Title : Finite volume/finite element schemes for compressible two-phase flows inheterogeneous and anisotropic porous media

**Keywords :** porous media, two-phase flow, compressible, immiscible, finite volumes, finite elements, positive, DDFV monotone.

**Abstract :** The objective of this thesis is the development and the analysis of robust and consistent numerical schemes for the approximation of compressible two-phase flow models in anisotropic and heterogeneous porous media. A particular emphasis is set on the anisotropy together with the geometric complexity of the medium. The mathematical problem is given in a system of two degenerate and coupled parabolic equations whose main variables are the nonwetting saturation and the global pressure. In view of the difficulties manifested in the considered system, its cornerstone equations are approximated with two different classes of the finite volume family.

The first class consists of combining finite elements and finite volumes. Based on standard assumptions on the space discretization and on the permeability tensor, a rigorous convergence analysis of the scheme is carried out thanks to classical arguments. To dispense with the underlined assumptions on the anisotropy ratio and on the mesh, the model has to be first formulated in the factional flux formulation. Moreover, the diffusive term is discretized by a Godunov-like scheme while the convective fluxes are approximated using an upwind technique. The resulting scheme preserves the physical ranges of the computed solution and satisfies the coercivity property. Hence, the convergence investigation holds. Numerical results show a satisfactory qualitative behavior of the scheme even if the medium of interest is anisotropic.

The second class allows to consider more general meshes and tensors. It is about a new positive nonlinear discrete duality finite volume method. The main point is to approximate a part of the fluxes using a nonstandard technique. The application of this idea to a nonlinear diffusion equation yields surprising results. Indeed, not only is the discrete maximum property fulfilled but also the convergence of the scheme is established. Practically, the proposed method shows great promises since it provides a positivity-preserving and convergent scheme with optimal convergence rates.