Conception et analyse de schémas non-linéaires pour la résolution de problèmes paraboliques : application aux écoulements en milieux poreux

Ahmed Ait Hammou Oulhaj

Université de Lille 1 SCIENCES ET TECHNOLOGIES

Laboratoire Paul Painlevé

Inria INVENTEURS DU MONDE NUMÉRIQUE

École Doctorale Sciences Pour l'Ingénieur

THÈSE

Pour obtenir le grade de docteur délivré par

L'Université de Lille 1

Spécialité doctorale "Mathématiques"

présentée et soutenue publiquement par

Ahmed AIT HAMMOU OULHAJ

le 11 Décembre 2017

Conception et analyse de schémas non-linéaires pour la résolution de problèmes paraboliques : application aux écoulements en milieux poreux

Directeurs de thèse : **Claire Chainais-Hillairet et Clément Cancès**

Jury

M Brahim Amaziane,	Examinateur	Université de Pau
M. Clément Cancès,	Directeur	Inria Lille
Mme Claire Chainais-Hillairet,	Directrice	Université de Lille 1
Mme Catherine Choquet,	Examinatrice	Université de la Rochelle
M. Robert Eymard,	Rapporteur	Université Paris-Est Marne-la-Vallée
M. Bogdan Matioc,	Rapporteur	Université de Hannover
Mme Carole Rosier,	Examinatrice	Université de Calais U.L.C.O

Thèse effectuée au sein du **Laboratoire Paul Painlevé (UMR CNRS 8524)**
de l'Université de Lille 1
Cité Scientifique
59655 Villeneuve d'Ascq Cedex
France

Résumé. L'objectif de cette thèse est de concevoir et d'analyser des schémas numériques performants pour la simulation d'écoulements complexes en milieux poreux. Ces écoulements représentent un domaine de recherche à la fois d'un grand intérêt scientifique et d'une grande importance pour la pratique de part leur vaste domaine d'applications.

Dans un premier temps nous proposons un schéma CVFE (Control Volume Finite Element) non-linéaire pour approcher la solution de l'équation de Richards anisotrope. Les degrés de liberté sont affectés aux sommets du maillage triangulaire primal tandis que les équations de bilan sont discrétisées sur un maillage dual barycentrique. La mobilité d'arête est gérée à l'aide d'une procédure de décentrement qui autorise les transmissivités négatives et donc les pertes de monotonie. Nous montrons d'abord que ce schéma est non-linéairement stable grâce à un contrôle de l'énergie physique, qu'il admet (au moins) une solution discrète et que la saturation est bornée entre 0 et 1. Ensuite nous montrons sous hypothèse faible de régularité sur le maillage que la solution discrète converge vers la solution faible du problème continu. Enfin, en vue de mettre en évidence l'efficacité, la stabilité et la robustesse de la méthode, nous réalisons des tests numériques dans des cas isotropes et anisotropes.

Dans un second temps nous nous intéressons à un modèle d'intrusion saline. Pour ce modèle nous discrétisons la partie temporelle à l'aide du schéma d'Euler implicite et la discrétisation en espace repose sur un schéma volumes finis à deux points avec un décentrement des mobilités d'arêtes. Pour ce schéma nous montrons qu'il préserve au niveau discret les principales propriétés du problème continu : l'existence de solutions discrètes positives, la décroissance de l'énergie et le contrôle de l'entropie et sa dissipation. En se basant sur ces résultats, nous montrons que la solution discrète converge vers la solution faible du problème continu. De plus, nous illustrons numériquement le comportement du modèle.

Le dernier chapitre est dédié à l'étude du comportement en temps long d'un modèle d'intrusion saline. Il s'agit d'identifier les états d'équilibres qui sont les minimiseurs d'une énergie convexe. Nous montrons pour le problème continu l'existence et l'unicité des minimiseurs de la fonctionnelle d'énergie, que les minimiseurs sont des états stationnaires et que ces états stationnaires sont radiaux et uniques. Pour différents jeux de paramètres, nous donnons une illustration numérique des états stationnaires et nous exhibons le taux de convergence.

Mots clefs: schémas Volumes Finis non-linéaires, schéma CVFE, écoulements en milieu poreux, intrusion saline, comportement en temps long.

Abstract. This thesis is focused on the design and the analysis of efficient numerical schemes for the simulation of complex flows in porous media.

First, we propose a nonlinear Control Volume Finite Element scheme (CVFE) in order to approximate the solution of Richards equation with anisotropy. In particular the diffusion terms are discretized by means of a conforming piecewise linear finite element method on a primal triangular mesh, and the others terms are discretized by means of an upwind finite volume method on a barycentric dual mesh. This scheme is based on a suitable upwinding of the mobility which allows the negative transmissibility coefficients. First of all, we prove the nonlinear stability of the scheme thanks to an energy estimate, that there exists (at least) one discrete solution and that the saturation belongs to the interval $[0, 1]$. Moreover, the convergence of the method is proved as the discretization steps tend to 0. We give some numerical experiments on isotropic and anisotropic cases illustrate the efficiency of the method.

Second, We consider a degenerate parabolic system modelling the flow of fresh and saltwater in a porous medium in the context of seawater intrusion. We propose and analyze a finite volume scheme based on two-point flux approximation with upwind mobilities. The scheme preserves at the discrete level the main features of the continuous problem, namely the nonnegativity of the solutions, the decay of the energy and the control of the entropy and its dissipation. Based on these non-linear stability results, we show that the scheme converges towards a weak solution to the problem. Numerical results are provided to illustrate the behavior of the model and of the scheme.

Finally, the large time behaviour of the seawater intrusion model is studied. The goal is to identify the equilibrium states which are the minimizers of a convex energy. We prove for the continuous problem the existence and uniqueness of the minimizers of the energy functional, that the minimizers are stationary states and that these stationary states are radial and unique. For different viscosity ratios, we give numerical illustrations of the stationary states and we exhibit the convergence rate.

Keywords: non-linear Finite Volume schemes, CVFE scheme, flows in porous media, seawater intrusion , large-time behaviour.

Remerciements. J'exprime tout d'abord mon entière reconnaissance et ma gratitude à mes directeurs de recherche Claire Chainais-Hillairet et Clément Cancès qui n'ont ménagé aucun effort dans l'encadrement de cette thèse. Ils m'ont, par leurs suggestions, leurs différentes orientations, apporté un précieux soutien dans la réalisation de ce travail. Ils m'ont communiqué leur passion pour la recherche en mathématiques appliquées.

Je tiens à remercier les rapporteurs Robert Eymard et Bogdan Matioc pour avoir accepté d'apporter un jugement objectif et scientifique au travail effectué. J'adresse mes remerciements à Brahim Amaziane, à Catherine Choquet et à Carole Rosier pour avoir accepté de faire partie de mon jury de thèse en tant qu'examinateurs.

Je tiens à remercier également Philippe Laurençot pour les échanges fructueux qui ont permis l'élaboration du dernier chapitre de ce travail. Je lui suis très reconnaissant, ainsi que très fier d'avoir pu collaborer avec lui.

J'adresse un grand merci aux membres des équipes ANEDP (Laboratoire Paul Painlevé), RAPSODI et MEPHYSTO (Inria) pour leur accueil et leur gentillesse. Plus généralement je remercie les membres du laboratoire Paul Painlevé au sein duquel je me suis épanoui durant trois ans.

Pendant cette thèse, j'ai eu le plaisir d'échanger dans la bonne humeur avec de nombreux enseignants chercheurs et avec de nombreux doctorants, en particulier Fadil, Oussama, Iheb, Loïc, Haithem, Yasser, Thomas, Moussa, Abdullatif, Moudhaffar, Diala, Meryem, Pallavi, Aya, Joanna, Mohammed, Marwa et Abdelrahman. Je ne peux m'empêcher d'avoir une pensée très particulière pour Florent Dewez, Van Ha Hoang et Geoffrey Boutard avec qui j'ai partagé mon bureau, sans oublier bien sûr Céline Esser, Sara, Pierre et Simon. Merci pour leur soutien, leur bonne humeur, leur sympathie et les bons moments passés ensemble. Je remercie également mes anciens et nouveaux co-bureaux (Inria): Pierre-Louis, Flore, Claire, Antoine, Pierre et David.

Au delà des horizons Lillois, je tiens à remercier Moussa B., Aissa A., Nassim A., Mustapha R., Youcef M., Mohammed S., Mofdi El, Moha R., Nouh I., Mohamed M., Nawfal S., El Houssaine Q., Somaya G., Mohamed F., Marc W., El Mostafa S, Larbi T., Saïd T., Naji A., Naji Y., Najib K., Saïd A, Youssef Z., Soumaya S., Rabab A., Zakariya B., Saleh, Ait Icho, Mohammed L., Abdellatif C et Thierry T.

Le bon fonctionnement du monde de la recherche repose largement sur les secrétariats, services informatiques ou autres services administratifs. Je profite de ces quelques lignes pour leur témoigner ma reconnaissance.

Je conclus mes remerciements par remercier ma famille, mes proches notamment mes parents (Touda et Mohamed), mes frères (Abdellah, Abderrahman, Abdelaali et Youssef), la famille Ben Haïda, Ait Ouchen (Abderrahman, Ali,...) et mes amis qui m'ont apporté leur soutien au cours de ce travail.

Contents

Chapitre 1

Introduction

L'objectif de cette thèse est de concevoir et d'analyser des schémas numériques performants pour la simulation d'écoulements complexes en milieux poreux. Les modèles mathématiques les plus fréquemment utilisés pour décrire des écoulements en milieux poreux insaturés consistent en des systèmes de lois de conservation paraboliques dégénérés. Ils nécessitent la mise au point de méthodes numériques dédiées où la conservation de certaines quantités et la préservation de certaines propriétés (positivité de la solution, contrôle de l'entropie et sa dissipation, décroissance de l'énergie...) s'avèrent fondamentales. Les problèmes d'écoulements en milieu poreux constituent un domaine de recherche à la fois d'un grand intérêt scientifique et pratique, de par leur vaste domaine d'applications. On peut citer par exemple l'écoulement de l'eau dans les nappes aquifères qui sont essentielles à l'alimentation des villes en eau potable. D'autres applications sont issues de l'industrie pétrolière, comme la simulation de réservoir, la modélisation de bassin, ou la simulation de l'injection et du stockage du $CO2$. Ces écoulements mettent en jeu des processus complexes et font notamment intervenir des phénomènes non linéaires. Cela rend nécessaire la compréhension profonde des écoulements, ainsi que le développement d'outils adaptés pour les simulations numériques. Cette thèse aborde d'une part des questions autour de l'équation de Richards qui décrit l'écoulement de l'eau dans un milieu poreux insaturé et d'autre part sur un modèle d'intrusion saline et en particulier de son comportement en temps long.

1.1 Modèles physiques

On commence par une présentation des modèles étudiés dans la thèse, mais tout d'abord rappelons quelques définitions.

1.1.1 Quelques définitions

Les modèles étudiés sont des modèles d'écoulements en milieu poreux. Un milieu poreux est constitué à l'échelle microscopique d'une phase solide et d'une phase de "vide" appelée "pore" dans laquelle circule un fluide (voir Figure 1.1). Le milieu poreux est dit saturé en eau si les pores sont remplis exclusivement en eau. Si une autre phase (par exemple gazeuse) y est également présente, le milieu est dit non saturé en eau. Nous définissons un certain nombre de grandeurs permettant de décrire les écoulements en milieu poreux.

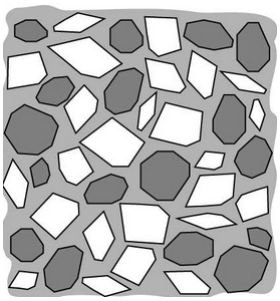


FIGURE 1.1 – Exemple d'un milieu poreux

Porosité. Un des paramètres principaux décrivant un milieu poreux est la porosité φ. Pour un volume élémentaire représentatif donné, centré en un point x du milieu, la porosité $\varphi(x)$ est le rapport (sans dimension) entre le volume occupé par les pores et le volume total :

$$\varphi = \frac{\text{Volume des pores}}{\text{Volume total}}.$$

C'est le ratio du volume disponible pour un écoulement fluide.

Perméabilité intrinsèque. La perméabilité intrinsèque Λ ne dépend que de la géométrie du milieu et indique l'aptitude de celui-ci à être traversé par un fluide. Elle ne dépend que de la structure et de la connectivité des pores et non du fluide qui se déplace dans les pores. Lorsque le milieu est *isotrope*, la perméabilité $\Lambda = \lambda\mathbf{Id}, \lambda \in \mathbb{R}$ est indépendante de la direction de l'écoulement. Dans le cas où le milieu est *anisotrope*, Λ prend la forme d'un tenseur symétrique défini positif, soit en deux dimensions :

$$\begin{pmatrix} \Lambda_{xx} & \Lambda_{xy} \\ \Lambda_{xy} & \Lambda_{yy} \end{pmatrix}.$$

On notera que selon les coefficients de la matrice, le fluide privilégiera des directions propres correspondant aux plus grandes valeurs propres.

Teneur en eau. La teneur en eau θ de la phase α est le ratio volumique de la phase α dans le volume donné du milieu poreux :

$$\theta = \frac{\text{Volume d'eau}}{\text{Volume total}}.$$

Saturation. La saturation s_α de la phase α est le ratio volumique de la phase α dans le volume des pores :

$$s_\alpha = \frac{\theta}{\varphi} = \frac{\text{Volume d'eau}}{\text{Volume des pores}}.$$

Elle est comprise entre 0 et 1.

Perméabilité relative. La perméabilité relative k_α de la phase α décrit, selon la saturation d'un fluide dans le milieu, sa capacité à s'écouler. Elle dépend de la saturation, elle est croissante et elle vérifie : $k_\alpha(0) = 0$, $k_\alpha(1) = 1$ et elle est sous-linéaire, i.e. $k_\alpha(s) \leq s$ (voir Figure 1.2).

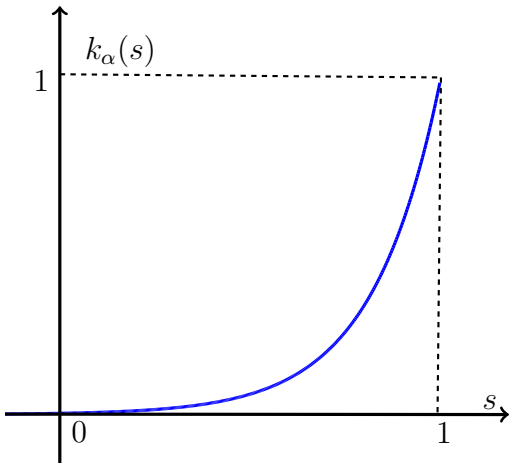


FIGURE 1.2 – Perméabilité relative

Conductivité hydraulique. La conductivité hydraulique $\mathbb{K}$ traduit la facilité avec laquelle le fluide traverse le milieu. Elle dépend de la perméabilité intrinsèque du milieu et de la viscosité cinématique du fluide :

$$\mathbb{K} = \frac{\rho g}{\mu}\Lambda = \frac{g}{\nu}\Lambda,$$

où ρ est la densité du fluide, μ la viscosité dynamique du fluide, ν la viscosité cinématique du fluide et g la constante de gravité.

Charge hydraulique. La charge hydraulique Φ d'un fluide incompressible est donnée par :

$$\Phi = \frac{v^2}{2g} + \frac{p}{\rho g} + z,$$

où v est la vitesse du fluide, g la constante de gravité, p la pression du fluide, ρ sa masse volumique et z l'altitude du point considéré. Nous considérons le cas où la vitesse du fluide est suffisamment faible pour qu'on puisse négliger l'énergie cinétique. Ainsi la charge hydraulique s'écrit :

$$\Phi = \frac{p}{\rho g} + z.$$

Aquifère. Un aquifère est une formation géologique contenant de l'eau et à travers laquelle, en conditions normales, une quantité significative de cette eau s'écoule. On distinguera deux types d'aquifère (voir Figure 4.1) : les aquifères confinés et les aquifères non confinés.

Aquifère confiné. Il contient une nappe confinée par le bas et par le haut, reposant sur un plancher (ou un *substratum*) peu perméable et limité par un toit (ou un *superstratum*) peu perméable.

Aquifère non confiné. Il contient une nappe à surface libre, reposant sur un plancher peu perméable, mais non confinée par le haut.

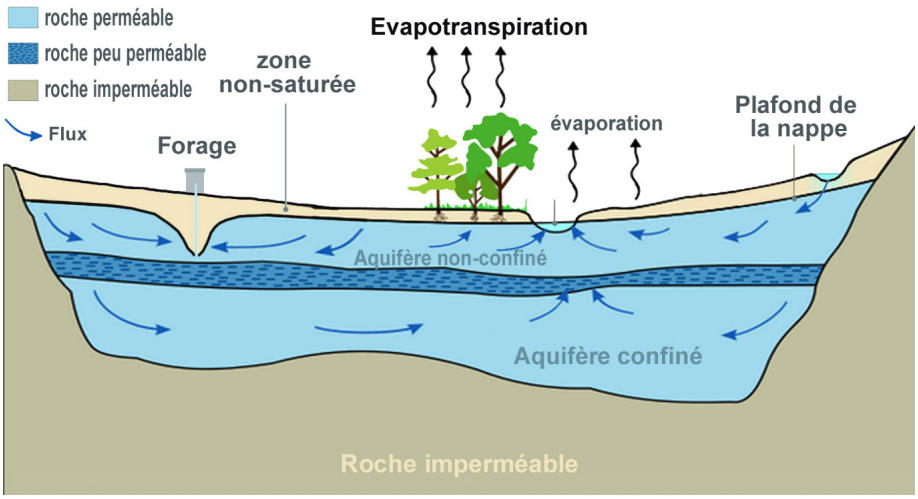


FIGURE 1.3 – Exemple d'un aquifère (`https://fr.wikipedia.org/wiki/Loi_de_Darcy`).

1.1.2 Écoulements en milieu poreux insaturé

On cherche à décrire l'écoulement diphasique eau-air dans un milieu poreux. On se place dans un domaine $\Omega \subset \mathbb{R}^2$, correspondant à un milieu poreux tel que présenté ci-dessus. L'équation de conservation de la masse dans Ω appliquée à chaque phase $\alpha \in \{\mathrm{w}, \mathrm{a}\}$ (ici en l'absence de termes sources) décrit la dynamique de l'écoulement :

$$\partial_t(\varphi s_\alpha \rho_\alpha) + \nabla\cdot(\rho_\alpha v_\alpha) = 0. \tag{1.1}$$

Les vitesses v_α de chaque phase s'expriment à l'aide de la loi de Darcy généralisée [60] :

$$v_\alpha = -\Lambda \frac{k_\alpha}{\mu_\alpha} \nabla(p_\alpha + \rho_\alpha g z).$$

Le système (1.1) admet six inconnues : $s_w, s_n, p_w, p_n, \rho_w, \rho_n$. Des relations supplémentaires sont nécessaires pour fermer le système. La première dit simplement que le volume poreux est totalement occupé par le mélange eau-air, ce qui se traduit par

$$s_w + s_a = 1.$$

La deuxième traduit la compressibilité des fluides et relie la densité de la phase α à sa pression :

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

La troisième est la relation pression capillaire. Elle relie la différence entre les pressions de phase à la composition du fluide.

$$p_w - p_a \in p_c(s_w).$$

Sur la figure 1.4, on représente la fonction $p \mapsto s(p)$ qui est l'inverse du graphe p_c.

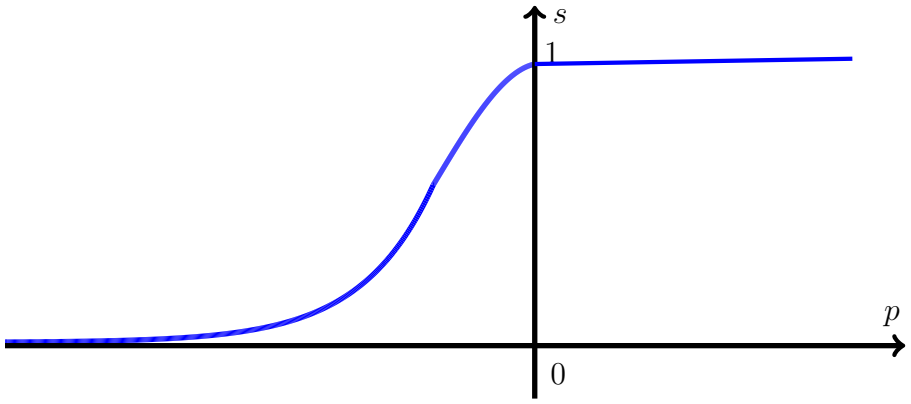


FIGURE 1.4 – Relation saturation-pression

Il est classique pour un écoulement en milieu insaturé de supposer la pression de l'air $p_a = 0$, ainsi l'équation (1.1) pour l'eau peut être résolue indépendamment de celle de l'air. Supposons de plus que :

- l'eau est homogène et incompressible, alors la masse volumique ρ_w est constante (ne dépend pas de la variable temporelle) et uniforme (ne dépend pas de la variable spatiale),
- le milieu poreux est indéformable, donc φ est constante.

On obtient à partir de (1.1) :

$$\varphi \partial_t s_w - \nabla \cdot \Big(\Lambda \frac{k_w(s_w)}{\mu_w} \nabla (p_w + \rho_w g z) \Big) = 0, \tag{1.2}$$

appelée équation de Richards [19, 50, 139]. Comme la saturation s et la perméabilité relative k ne sont pas constantes en milieu non saturé, alors on peut exprimer s et k en fonction de la pression. En adimensionnant on peut supposer $\rho g = 1$ et (1.2) peut se réécrire avec p l'inconnue comme (en omettant l'indice w) :

$$\varphi \partial_t s(p) - \nabla \cdot \Big(\Lambda \eta(s(p)) \nabla \big(p + z \big) \Big) = 0, \tag{1.3}$$

où $\eta(s) = \dfrac{k(s)}{\mu}$. Cette forme reste valable dans les milieux hétérogènes et éventuellement saturés, où elle dégénère en l'équation de Darcy. On peut également réécrire l'équation (1.3) en choisissant comme inconnue s tant que le milieu n'est pas saturé, le problème devenant elliptique. D'autres formulations existent, par exemple en introduisant la transformée de Kirchhoff :

$$F : \mathbb{R} \longrightarrow \mathbb{R}, \quad F(p) = \int_0^p \eta(s(a)) \, \mathrm{d}a. \tag{1.4}$$

F est inversible sur $\{p |\ s(p) > 0\}$, on pose $u = F(p)$ et $c(u) = c(F(p)) = s(p)$, avec c qualitativement similaire à s. De plus on a $\nabla F = \eta(s(p)) \nabla p = \nabla u$, alors (1.3) devient :

$$\varphi \partial_t c(u) - \nabla \cdot \Big(\Lambda \nabla u + \Lambda \eta(c(u)) \nabla z \Big) = 0. \tag{1.5}$$

L'équation (1.3) a fait l'objet de nombreux travaux de recherche, en particulier dans le domaine de l'analyse numérique. Un schéma Différences Finies conservatif, avec des restrictions sur la grille et sur le tenseur de diffusion Λ, a été étudié numériquement dans [151]. Cependant, à notre connaissance, il n'existe pas de preuve de convergence pour le schéma présenté dans [151]. Dans un milieu isotrope (i.e., $\Lambda = \lambda \mathbf{I}$) et sans dégénérescence des mobilités (i.e., $\eta(s) \geq \eta_\star > 0$), un schéma Volumes Finis à deux points sur maillage admissible a été étudié dans [84], nécessitant l'introduction de la transformée de Kirchhoff (1.4) et dans [82], exprimé en variables physiques (saturation et pression). D'autres schémas localement conservatifs ont été largement utilisés pour l'approximation de l'équation de Richards, comme par exemple les Éléments Finis Mixtes. Mentionnons [15, 137, 138] où les auteurs ont réussi à donner une estimation d'erreur. Cependant les schémas dans

[15, 137, 138] et le schéma dans [30] nécessitent l'introduction de la transformée de Kirchhoff et une régularisation du problème dans [138] pour surmonter les difficultés liées à la dégénérescence. On mentionne également les schémas Volumes Finis MPFA (Multi-Point Flux Approximations) étudiés dans le cas de l'équation de Richards [29, 117]. Notons que les Éléments Finis Mixtes et les Volumes Finis MPFA peuvent produire des oscillations non physiques sur la saturation, avec en particulier des valeurs négatives ou des valeurs supérieures à 1.

1.1.3 Intrusion saline

Sur terre l'eau salée représente un peu plus de 97% et l'eau douce un peu moins de 3%. Cette dernière est présente dans les rivières, les lacs, les glaciers, les tourbières, etc. L'eau douce sous forme de glace représente un peu moins de deux tiers de la quantité totale et les réserves d'eaux souterraines un tiers. La quantité faible restante représente les eaux de surface, qui malheureusement aujourd'hui ne sont plus suffisantes pour répondre à la consommation humaine et aux activités industrielles. De ce fait le recours aux puisements dans les réserves d'eaux souterraines est indispensable.

Dans les zones côtières, l'intrusion de l'eau de mer est due à deux causes : l'une est de l'ordre de l'humain, lorsque ce dernier surexploite les eaux souterraines pour des fins de consommation, pour l'agriculture et l'industrie. L'autre cause est simplement naturelle, du fait que l'eau salée (le fluide le plus lourd) est en dessous de l'eau douce (le fluide le plus léger). Nous avons besoin de modèles efficaces et précis pour simuler le transport du front de l'eau salée dans l'aquifère côtier. On distingue deux cas importants : le cas de l'aquifère libre (non confiné) et le cas de l'aquifère confiné. Dans ces deux cas, la nappe est délimitée par deux couches, la couche inférieure étant toujours supposée imperméable. Pour l'aquifère confiné, la surface supérieure de l'aquifère est imperméable et pour l'aquifère libre, la surface supérieure est une couche perméable constituée par exemple de graviers, ou de sable.

Beaucoup de recherches ont été faites sur le problème d'intrusion saline et on peut classer les modèles selon quatre approches : modèle sans interface, avec interface diffuse, avec interface nette et un modèle avec interface mixte.

Modèle sans interface : il s'agit d'un modèle d'écoulements de deux fluides miscibles. Cette approche ne permet pas de suivre l'interface de l'eau saline et de l'eau douce (voir [55]), d'autant plus qu'il est difficile de définir la zone de transition entre l'eau douce et l'eau salée. Bien qu'elle décrive réellement le problème, elle est néanmoins très lourde d'un point de vue théorique et numérique.

Modèle avec interface diffuse : cette approche quant à elle suppose l'existence d'interfaces diffuses entre l'eau douce et l'eau salée en interaction et entre les milieux saturés et insaturés. Cette zone de transition est caractérisée par les variations de la

concentration en sel. Ce modèle a fait par exemple l'objet d'étude dans [54, 143, 144].

Modèle avec interface nette ou abrupte : on considère deux fluides immiscibles. On suppose que chaque fluide est confiné dans une région et qu'il n'y a pas de transfert de masse entre les deux fluides, d'où l'existence de l'interface les séparant. C'est une approche qui a une réalité physique lorsque la taille de l'aquifère est très grande par rapport à sa largeur. Elle nous donne des informations sur le mouvement des interfaces (voir [10, 19, 20, 21, 108, 145, 146]).

Modèle avec interface mixte : l'objectif est de combiner l'efficacité du modèle avec interface nette au réalisme des modèles avec interface diffuse. Un tel modèle a été proposé dernièrement dans [57] et étudié dans [56].

Dans cette thèse nous sommes partis du modèle [108] qui modélise l'évolution des interfaces de l'eau salée et de l'eau douce dans un aquifère non confiné (voir Figure 1.5), en supposant que les deux fluides sont immiscibles et donc chaque que fluide est confiné dans une région séparée par une interface. Cette interface est en réalité une zone de transition (car ces deux fluides sont en fait miscibles), mais qu'on néglige du fait que l'épaisseur de celle-ci est très petite comparée à la taille de l'aquifère. Ceci permet d'établir une équation de la conservation de la masse dans chaque domaine en 3D. Ensuite on considère l'hypothèse de Dupuit : les écoulements sont quasi-horizontaux et de ce fait la charge hydraulique Φ est indépendante de z. Cette hypothèse permet d'intégrer les équations verticalement et se ramener à un problème 2D. Enfin on suppose la continuité de la pression à travers les interfaces.

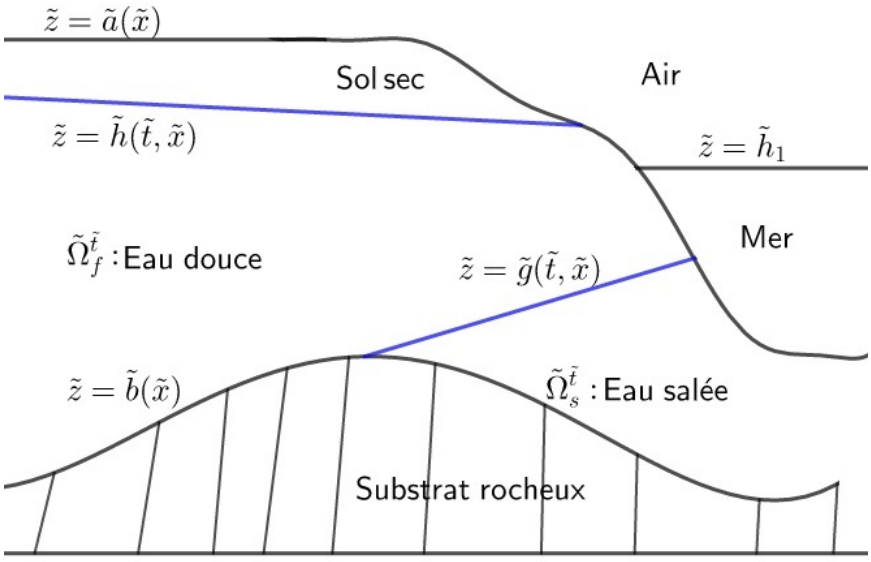


FIGURE 1.5 – Aquifère non confiné

On note $\alpha = f$ pour l'eau douce et $\alpha = s$ pour l'eau salée.

$$\tilde{\Omega}_\alpha = \cup_{\{\tilde{t}>0\}}\{\tilde{t}\} \times \tilde{\Omega}_\alpha^{\tilde{t}},$$

où $\tilde{\Omega}_\alpha^{\tilde{t}}$ est la région occupée par chaque fluide α (voir Figure 1.5). On suppose de plus que la porosité du milieu poreux est constante et quitte à changer l'échelle de temps, on peut supposer $\varphi = 1$. Le milieu est isotrope ici (i.e., $\Lambda = \lambda \mathbf{Id},\ \lambda \in \mathbb{R}$).

Le point de départ se déduit de (1.1) :

$$\frac{\partial \tilde{\rho}_\alpha}{\partial \tilde{t}} + \tilde{\mathrm{div}}(\tilde{\rho}_\alpha \tilde{v}_\alpha) = 0 \quad \text{dans } \tilde{\Omega}_\alpha, \tag{1.6}$$

où $\tilde{v}_\alpha = -\dfrac{\Lambda}{\mu}\tilde{\nabla}(\tilde{p} + \tilde{\rho}_\alpha \tilde{z}) = -\dfrac{\lambda}{\mu}\tilde{\nabla}(\tilde{p} + \tilde{\rho}_\alpha \tilde{z})$ sur $\tilde{\Omega}_\alpha$. On suppose l'existence d'un petit paramètre $\varepsilon > 0$ tel que :

$$\begin{cases} \tilde{x} = x,\ \tilde{z} = \varepsilon z,\ \tilde{t} = t/\varepsilon,\ \tilde{h_1} = \varepsilon h_1, \\ \tilde{a}(\tilde{x}) = \varepsilon a(x),\ \tilde{b}(\tilde{x}) = \varepsilon b(x). \end{cases}$$

Le paramètre ε est le ratio entre la longueur de l'aquifère et sa largeur. Ensuite on considère le scaling suivant :

$$\begin{cases} \tilde{h}(\tilde{t},\tilde{x}) = \varepsilon h^\varepsilon(x,t),\ \tilde{g}(\tilde{t},\tilde{x}) = \varepsilon g^\varepsilon(x,t),\ \tilde{p}(\tilde{t},\tilde{x}) = \varepsilon p^\varepsilon(x,t), \\ \tilde{v}_\alpha^{\tilde{x}}(\tilde{t},\tilde{x}) = \varepsilon v_\alpha^{x,\varepsilon}(x,t),\ \tilde{v}_\alpha^{\tilde{z}}(\tilde{t},\tilde{x}) = \varepsilon^2 v_\alpha^{z,\varepsilon}(x,t). \end{cases} \tag{1.7}$$

Notons que dans (1.7) $\tilde{v}_\alpha^{\tilde{x}} \gg \tilde{v}_\alpha^{\tilde{z}}$ avec $\tilde{v}_\alpha^{\tilde{x}}$ (resp. $\tilde{v}_\alpha^{\tilde{z}}$) la vitesse horizontale (resp. la vitesse verticale), en accord avec l'hypothèse de Dupuit. Supposons l'existence du développement asymptotique formel :

$$\begin{cases} h^\varepsilon & = h & + O(\varepsilon) \\ g^\varepsilon & = g & + O(\varepsilon) \\ p^\varepsilon & = p & + O(\varepsilon) \\ v_\alpha^{x,\varepsilon} & = v_\alpha^x & + O(\varepsilon) \\ v_\alpha^{z,\varepsilon} & = v_\alpha^z & + O(\varepsilon) \end{cases}$$

et définissons

$$\mathcal{K}_\alpha(x,z) = \frac{\lambda \rho_s}{\mu}\int_0^z \mathrm{d}\bar{z} = \frac{\lambda \rho_s}{\mu} z.$$

En intégrant (1.6) verticalement on obtient :

$$\begin{cases} \partial_t(\Phi_f(.,h) - \Phi_f(.,g)) - \mathrm{div}\Big(\Big[\mathcal{K}_f(x,z)\Big]_{z=g}^{z=h}\nabla_x(p + (1-\varepsilon_0)h)\Big) & = 0 \text{ dans} \quad \Omega_f, \\ \partial_t(\Phi_s(.,g)) - \mathrm{div}\Big(\Big[\mathcal{K}_f(x,z)\Big]_{z=b}^{z=g}\nabla_x(p + (1-\varepsilon_0)h + \varepsilon_0 g)\Big) & = 0 \text{ dans} \quad \Omega_s, \end{cases} \tag{1.8}$$

avec $\varepsilon_0 = \dfrac{\rho_s - \rho_f}{\rho_s} \in (0,1)$ et γ_α la densité du fluide α. Quitte à changer l'échelle en temps on peut supposer :

$$\frac{\rho_s \lambda}{\mu} = 1, \quad \text{et} \ \ \Phi_\alpha(x,z) = \int_0^z \varphi \, \mathrm{d}\bar{z} = z.$$

De plus dans le cas d'un aquifère non confiné on a $p = 0$. On obtient au final le système suivant :

$$\begin{cases} \partial_t(h-g) - \operatorname{div}\Big((h-g)\nabla_x((1-\varepsilon_0)h)\Big) = 0 & \text{dans} \quad \Omega_f, \\ \partial_t g - \operatorname{div}\Big((g-b)\nabla_x((1-\varepsilon_0)h + \varepsilon_0 g)\Big) = 0 & \text{dans} \quad \Omega_s. \end{cases} \tag{1.9}$$

Posons $u = h - g$, $v = g - b$ et $\nu = 1 - \varepsilon_0$ alors on a le système d'équations paraboliques dégénérées avec diffusion croisée suivant :

$$\begin{cases} \partial_t u - \operatorname{div}\Big(\nu u \nabla_x(u + v + b)\Big) = 0 & \text{dans} \quad \Omega, \\ \partial_t v - \operatorname{div}\Big(v \nabla_x(\nu u + v + b)\Big) = 0 & \text{dans} \quad \Omega, \end{cases} \tag{1.10}$$

où $\Omega \subset \mathbb{R}^2$ est le milieu poreux. Ce modèle est similaire à celui de Muskat [128, 129, 130] (en prenant $b = 0$), qui modélise le mouvement de deux fluides avec différentes densités et viscosités dans un milieu poreux.

Mentionnons que les systèmes de diffusion croisée sont largement présents dans différents domaines, comme en écologie, biologie, chimie etc. Pour un système de matériaux granulaires on peut se réferer à [97], à [51, 52, 150] pour un modèle de dynamique des populations, à [53] pour un modèle des semi-conducteurs, à [43, 105] pour un modèle en chimiotactisme (Patlak-Keller-Segel), à [115] pour un modèle de croissance tumorale, à [64, 112] pour les systèmes de réaction-diffusion. L'existence de solutions globales dans le cas de l'approche interface abrupte est étudiée dans [58]. D'autres auteurs se sont intéressés à des problèmes similaires. Dans [72, 73, 74, 119, 120] les auteurs ont étudié les solutions classiques et faibles du problème de Muskat.

1.2 Comportement en temps long

Dans cette thèse, il est également question d'étude du comportement en temps long. Elle est basée sur la méthode d'entropie-dissipation, qui a fait l'objet de nombreux travaux de recherche. Les premiers travaux concernaient le domaine des équations cinétiques de Boltzmann et Landau [147]. Elle a été ensuite étendue à d'autres problèmes, comme les équations de Fokker-Planck linéaires [44], l'équation des milieux poreux [45], les systèmes de réactions diffusion [62, 63, 99], les systèmes de dérive-diffusion pour les semi-conducteurs [94, 95, 96], les modèles de type films minces

[42], les modèles de type coagulation-fragmentation [90]. Pour un panorama plus complet de cette méthode et de ses domaines d'applications, on pourra se reporter à [16, 113] et aux références citées dans ces références. Des résultats similaires ont été obtenu dans [12], basés sur l'interprétation des modèles EDP comme flot gradient pour la distance de Wasserstein. On peut se référer à cette liste non exhaustive [25, 26, 111, 134].

L'objectif est d'identifier un état d'équilibre u_∞ d'une EDP et une fonctionnelle d'entropie (d'énergie) convexe $\mathfrak{E}$ qui atteigne son minimum à l'équilibre u_∞. L'écart entre une fonction u et l'équilibre u_∞ est mesuré par l'entropie relative :

$$\mathfrak{E}(u|u_\infty) := \mathfrak{E}(u) - \mathfrak{E}(u_\infty).$$

On aimerait montrer

$$\mathfrak{E}(u(t)|u_\infty) \longrightarrow 0,\ t \to +\infty, \tag{1.11}$$

et ensuite en déduire des résultats de convergence dans des espaces L^p. Pour cela on étudie $\dfrac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u(t))$ de telle façon à avoir ceci :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u(t)) + I(u(t)) = 0,$$

où $I \geq 0$ est la dissipation d'entropie. Si on arrive à avoir un contrôle de I par $\mathfrak{E}(u|u_\infty)$ du type :

$$I(u) \geq \lambda \mathfrak{E}(u|u_\infty) \quad \text{pour un } \lambda > 0,$$

alors on aura une inégalité du type :

$$-\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u|u_\infty) \geq \lambda \mathfrak{E}(u|u_\infty),$$

qui permet d'en déduire (1.11) avec une vitesse de convergence exponentielle. Avec des inégalités fonctionnelles de type Csiszar-Kullback [59, 118] on obtient la convergence dans L^1 avec une vitesse $\lambda/2$ pour l'entropie de Boltzmann.

Nous allons présenter à présent un exemple d'étude de comportement en temps long pour un exemple issu de la physique, qui s'inscrit dans le cadre général ci-dessus et que nous cherchons à étendre au cas du modèle (1.10) dans le chapitre 4. Cela concerne l'équation des milieux poreux. L'écoulement d'un gaz parfait dans un milieu poreux homogène peut être décrit classiquement par la solution du problème de Cauchy suivant

$$\begin{cases} \partial_t v = \Delta v^m & \text{sur } \mathbb{R}^N \times (0,T), \\ v(x,0) = v_0(x) & \text{sur } \mathbb{R}^N, \end{cases} \tag{1.12}$$

où v est la densité de gaz dans le milieu poreux et $m > 1$ une constante physique. Dans le cas où $m = 2$ et $N = 2$, l'équation (1.12) est une simplification du modèle d'intrusion saline. Maintenant en passant en variables auto-similaires (voir [45, 149])

$$u(t,x) = e^{Nt} v\Big(k(e^{t/k} - 1), x e^t\Big), \tag{1.13}$$

où $k = \dfrac{1}{N(m-1)+2}$, le problème (1.12) peut se réécrire sous la forme de l'équation de Fokker-Planck non-linéaire suivante :

$$\begin{cases} \partial_t u = \operatorname{div}(xu + \nabla u^m) & \text{sur } \mathbb{R}^N \times (0,T), \\ u(x,0) = u_0(x) & \text{sur } \mathbb{R}^N. \end{cases} \tag{1.14}$$

Si v est solution de (1.12) alors u donnée par (1.13) est solution de (1.14) et réciproquement si u est solution de (1.14) alors v donnée par :

$$v(t,x) = \left(1+\frac{t}{k}\right)^{-Nk} u\left(k\log\left(1+\frac{t}{k}\right), x\left(1+\frac{t}{k}\right)^{-k}\right), \tag{1.15}$$

est solution de (1.12).
Dans [45] les auteurs étudient le comportement en temps long de l'équation des milieux poreux en utilisant la reformulation (1.14). L'entropie pour ce problème est

$$\mathfrak{E}(u) = \int_{\mathbb{R}^N} \left(|x|^2 u + \frac{2}{m-1}u^m\right) \mathrm{d}x. \tag{1.16}$$

Ils prouvent que l'unique solution stationnaire de (1.14) est la distribution de Barenblatt-Pattle suivante :

$$u_\infty(x) = \left(\beta - \frac{m-1}{2m}|x|^2\right)_+^{1/m-1}, \tag{1.17}$$

où β est telle que $\int_{\mathbb{R}^N} u_\infty(x)\,\mathrm{d}x = \int_{\mathbb{R}^N} u_0\,\mathrm{d}x$ et $y_+ = \max(y,0)$. L'écart entre u et u_∞ est mesuré par l'entropie relative :

$$\mathfrak{E}(u|u_\infty) := \mathfrak{E}(u) - \mathfrak{E}(u_\infty) \geq 0.$$

La dissipation d'entropie pour $\mathfrak{E}(u|u_\infty)$ est donnée par $2I(u)$, où

$$I(u) = \int_{\mathbb{R}^N} u\left|x + \frac{m}{m-1}\nabla u^{m-1}\right|^2 \mathrm{d}x. \tag{1.18}$$

Sous certaines conditions sur u_0 on a

$$\mathfrak{E}(u(t)|u_\infty) \longrightarrow 0,\ t\to+\infty \quad \text{et} \quad I(u(t)) \longrightarrow 0,\ t\to+\infty.$$

De plus on montre

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u(t)|u_\infty) = -2I(u(t)), \tag{1.19}$$

et l'équation pour la dissipation d'entropie

$$\frac{\mathrm{d}}{\mathrm{d}t}I(u(t)) = -2I(u(t)) - R(t), \tag{1.20}$$

où $R(t) \geq 0$. En combinant (1.19) et (1.20) on obtient

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(u(t)|u_\infty) = \frac{\mathrm{d}}{\mathrm{d}t}I(u(t)) + R(t). \tag{1.21}$$

Intégrons (1.21) entre $t > 0$ et $+\infty$ on obtient

$$0 \leq \mathfrak{E}(u(t)|u_\infty) \leq I(u(t)), \quad t > 0. \tag{1.22}$$

En substituant (1.22) dans (1.19) on conclut que l'entropie relative converge exponentiellement avec une vitesse de convergence 2.

Dans [121], les auteurs ont étudié le comportement en temps long en 1D du problème de Muskat, qui est similaire à (1.10). Il a été question tout d'abord de classifier les solutions auto-similaires qui sont des états stationnaires, de montrer leur unicité et enfin de montrer la convergence de la solution vers une solution auto-similaire en étudiant une entropie relative. La vitesse de convergence n'a pas été explicitée dans cet article.

1.3 Schémas numériques pour les problèmes paraboliques

Afin de concevoir des schémas numériques performants pour la simulation d'écoulements complexes en milieu poreux, il est important de comprendre le processus d'écoulement. Rappelons que, dans les applications pratiques, le milieu poreux est souvent hétérogène et anisotrope et il peut avoir une géométrie complexe. Ces difficultés doivent être gérées par les méthodes numériques utilisées. Au cours des dernières années, le développement de méthodes numériques, de type Volumes Finis pour les équations de diffusion a fait l'objet d'efforts importants de la part de la communauté mathématique. La motivation majeure était de proposer des méthodes robustes par rapport à l'anisotropie et l'hétérogénéité de l'opérateur de diffusion et par rapport à la régularité des maillages. Comme cela apparaît clairement dans les résultats présentés dans différents benchmarks, aucune méthode linéaire ne préserve sur les maillages généraux le principe du maximum vérifié par le problème continu [36, 116, 132]. De plus, la décroissance de certaines entropies mathématiques n'est pas assurée au niveau discret. Dans ce contexte, la motivation pour la conception et l'analyse de schémas non-linéaires est grande.

Rappelons qu'en l'absence d'anisotropie, on peut envisager de discrétiser les équations par la méthode des Volumes Finis à deux points [76, 78] que nous allons décrire brièvement dans le cas de l'approximation du problème de Poisson suivant :

$$\begin{cases} -\Delta u = f & \text{sur } \Omega, \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{1.23}$$

où Ω est un ouvert borné de $\mathbb{R}^2$ et $f \in L^2(\Omega)$.
On se donne un maillage admissible (voir [78, Definition 9.1]) $\mathcal{T}$ de Ω constitué de volumes de contrôle (ou cellules) K (voir Figure 1.6). On intègre l'équation (1.23)

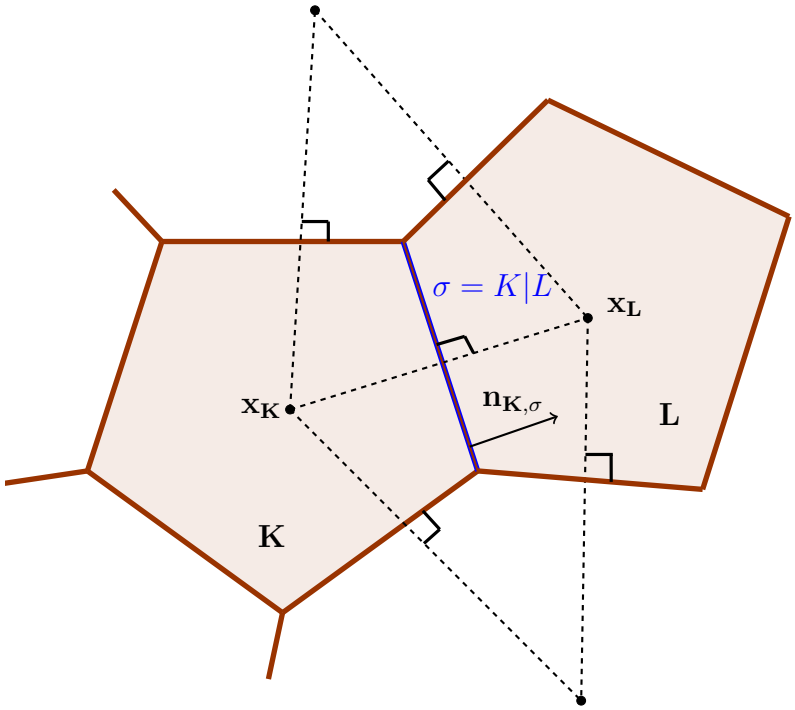


FIGURE 1.6 – Cellules admissibles

sur une cellule K

$$\int_K -\Delta u \, \mathrm{d}x = \int_K f \, \mathrm{d}x.$$

Par la formule de la divergence on obtient

$$\int_{\partial K} -\nabla u \cdot n_{K,\sigma} = \int_K f,$$

qui devient sur chacune des mailles

$$\sum_{\sigma \subset \partial K} \int_\sigma -\nabla u \cdot n_{K,\sigma} = \int_K f,$$

où la somme porte sur les arêtes du volume de contrôle K. Construire un schéma de discrétisation revient alors à approcher le flux $\int_\sigma \nabla u \cdot n_{K,\sigma}$. Un choix simple pour calculer $\nabla u \cdot n_{K,\sigma}$ sur une arête située entre deux mailles K et L est le suivant :

$$\nabla u \cdot n_{K,\sigma} \approx |\sigma| \frac{u(x_L) - u(x_K)}{|x_K - x_L|},$$

où x_K et x_L sont des points des mailles K et L. Cette approximation est consistante (voir [78]) à condition que la droite $(x_K x_L)$ soit perpendiculaire à l'arête commune

à K et L (voir Figure 1.6). Le schéma volumes finis obtenu en suivant ce principe simple d'approximation consiste ainsi à trouver des inconnues discrètes $(u_K)_{K\in\mathcal{T}}$ vérifiant le système de bilans discrets

$$\forall K \in \mathcal{T}, \quad -\sum_{\sigma} F_{K,\sigma} = \int_K f, \tag{1.24}$$

où le flux discret $F_{K,\sigma}$ est donné par

$$F_{K,\sigma} = |\sigma| \frac{u_L - u_K}{|x_K - x_L|}.$$

Cette méthode possède des propriétés remarquables, comme en particulier un principe du maximum discret. De plus elle est peu coûteuse, car elle n'utilise qu'une seule inconnue par maille. Néanmoins ce schéma ne s'applique que sur des maillages admissibles (i.e., vérifiant une condition d'orthogonalité). Ainsi sur des maillages plus généraux et en présence d'opérateurs diffusifs anisotropes, on perd la consistance.

Pour s'affranchir de ces contraintes, plusieurs méthodes numériques ont été développées au cours de ces dernières années dans le but de proposer des discrétisations qui gèrent les problèmes d'anisotropie et d'hétérogénéité sur des maillages généraux : méthodes de Galerkine discontinues [65], de différences finies mimétiques [33, 34], de Volumes Finis à flux multi-points [1, 5, 6, 70], de Volumes Finis en dualité discrète [27, 66, 104] , de Volumes Finis hybrides [79, 80, 81], de Volumes Finis mixtes [68, 69]. On peut se référer à [67] et aux benchmarks sur la diffusion anisotrope et hétérogène organisés à l'occasion des conférences FVCA 5 [102] et FVCA 6 [85]. D'autres méthodes combinent l'application d'une méthode de type Éléments Finis pour la discrétisation du terme de diffusion avec un schéma de Volumes Finis pour la discrétisation de tous les autres termes, en construisant également un maillage dual [87, 88, 91, 92].

Le schéma CVFE (Control Volume Finite Element) a été introduit dans le contexte d'écoulements en milieux poreux par Forsyth [91, 92]. Il est basé sur une discrétisation très simple des termes de diffusion par la méthode des Éléments Finis avec condensation de masse et une discrétisation Volumes Finis pour les autres termes. L'objectif est de reconstruire la solution approchée aux sommets du maillage primal. Les degrés de liberté sont affectés aux sommets du maillage triangulaire primal, tandis que les équations de bilan sont discrétisées sur un maillage dual barycentrique (Voir Figure 1.7). On note $\mathcal{V}$ l'ensemble des sommets du maillage triangulaire primal, $\mathcal{E}$ l'ensemble des arêtes σ de la triangulation $\mathcal{T}$, $\mathcal{E}_K$ le sous-ensemble de $\mathcal{E}$ constitué des arêtes admettant le sommet K comme une extrémité. Une arête joignant deux sommets K et L est notée par σ_{KL} et ω_K est le volume de contrôle associé au sommet K. $V_{\mathcal{T}}$ est l'espace usuel des éléments finis $\mathbb{P}_1$ sur $\mathcal{T}$ muni de la base canonique $(e_K)_{K\in\mathcal{T}}$, avec $e_K(x_L) = \delta_{KL}$. $X_{\mathcal{M}}$ est l'espace des fonctions constantes par morceaux sur le maillage dual.

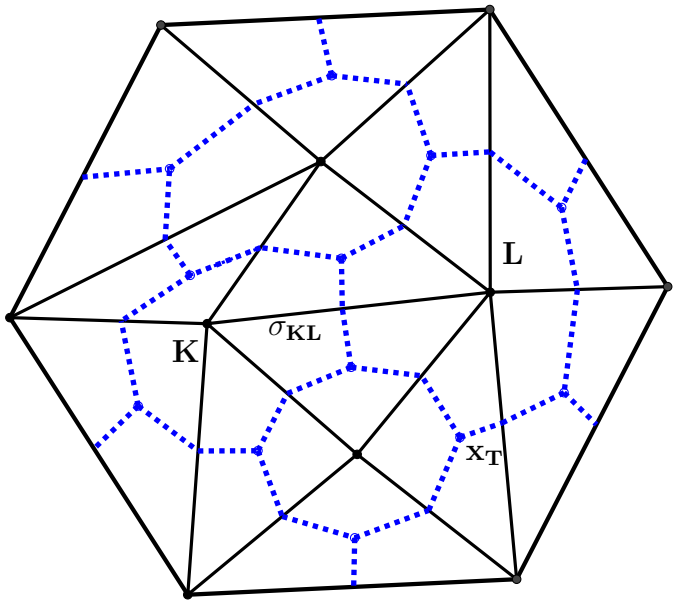


FIGURE 1.7 – Le maillage triangulaire primal $\mathcal{T}$ (*trait plein noir*) et le maillage dual barycentrique $\mathcal{M}$ (*trait pointillé bleu*)

Nous allons décrire brièvement ce schéma dans le cas du problème elliptique suivant

$$\begin{cases} -\mathrm{div}\big(\Lambda\nabla u\big) + \alpha u = f & \text{sur } \Omega, \\ \Lambda\nabla u \cdot \mathbf{n} = 0 & \text{sur } \partial\Omega, \end{cases} \tag{1.25}$$

où Ω est un ouvert borné de $\mathbb{R}^2$, $f \in L^2(\Omega)$, Λ la perméabilité intrinsèque anisotrope et $\alpha > 0$. Multiplions (1.25) par e_K et intégrons sur Ω

$$\int_\Omega \Big[-\mathrm{div}\big(\Lambda\nabla u\big) + \alpha u \Big] e_K \,\mathrm{d}\boldsymbol{x} = \int_\Omega f e_K \,\mathrm{d}\boldsymbol{x}.$$

Par la formule de la divergence on obtient

$$\int_\Omega -\mathrm{div}\big(\Lambda\nabla u\big) e_K \,\mathrm{d}\boldsymbol{x} = \int_\Omega \Lambda\nabla u \cdot \nabla e_K \,\mathrm{d}\boldsymbol{x}.$$

Puisque

$$\sum_{K\in\mathcal{V}} \nabla e_L = \nabla e_K + \sum_{L\neq K} \nabla e_L = 0,$$

alors

$$\nabla\Big(\sum_L u_L e_L\Big) = u_K \nabla e_K + \sum_{L\neq K} u_L \nabla e_L = \sum_{L\neq K} (u_L - u_K)\nabla e_L.$$

Par conséquent

$$\begin{aligned}\int_\Omega \Lambda\nabla u\cdot\nabla e_K\,\mathrm{d}\boldsymbol{x} &= \int_\Omega \Lambda\nabla\Big(\sum_L u_L e_L\Big)\cdot\nabla e_K\,\mathrm{d}\boldsymbol{x}\\ &= \int_\Omega \Lambda\sum_{L\neq K}(u_L-u_K)\nabla e_L\cdot\nabla e_K\,\mathrm{d}\boldsymbol{x}\\ &= \sum_{L\neq K}(u_K-u_L)\bigg(-\int_\Omega \Lambda\nabla e_K\cdot\nabla e_L\,\mathrm{d}\boldsymbol{x}\bigg).\end{aligned}$$

Le schéma CVFE pour (1.25) est le suivant

$$\sum_{\sigma_{KL}\in\mathcal{E}_K} a_{KL}(u_K-u_L)+\alpha m_K u_K = m_K f_K, \tag{1.26}$$

avec

$$a_{KL}=a_{LK}=-\int_\Omega \Lambda\nabla e_K(\boldsymbol{x})\cdot\nabla e_L(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\quad m_K=\int_\Omega e_K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}=\int_{\omega_K}\mathrm{d}\boldsymbol{x},$$

$$u_K=\frac{1}{m_K}\int_{\omega_K}u(x)\,\mathrm{d}\boldsymbol{x}\quad\text{et}\quad f_K=\frac{1}{m_K}\int_{\omega_K}f(x)\,\mathrm{d}\boldsymbol{x}.$$

Le schéma (1.26) est un système d'équations linéaires d'inconnues $(u_K)_{K\in\mathcal{V}}$. Notons $U=(u_K)_{K\in\mathcal{V}}$ le vecteur inconnu. Le schéma s'écrit $(A+\alpha B)U=b$ avec $b_K=m_K f_K\ \forall K\in\mathcal{V}$, $A_{KL}=A_{LK}=-a_{KL}$, $A_{KK}=\sum_{K\neq L}a_{KL}$, $B_{KK}=m_K$ et $B_{KL}=B_{LK}=0$. La matrice B est inversible. La matrice A est symétrique définie positive. Si

$$a_{KL}\geq 0\quad\forall\,\sigma_{KL}\in\mathcal{E}, \tag{1.27}$$

alors A est une $M-$matrice et par conséquent l'opérateur de diffusion anisotrope discret est monotone. Dans le cas où Λ est isotrope la propriété (1.27) est vérifiée si tous les angles des triangles de la triangulation $\mathcal{T}$ sont aigus [93] ou plus généralement dans le cas de la triangulation de Delaunay [109]. Néanmoins, pour des triangulations générales $\mathcal{T}$ de Ω et/ou dans le cas où Λ est anisotrope, il est bien connu que les coefficients a_{KL} peuvent être négatifs. Par conséquent A n'est plus une matrice définie positive (on perd la monotonie de l'opérateur de diffusion discret).

Dans cette thèse, nous nous intéressons à la convergence du schéma CVFE pour le problème de Richards (1.3) et du schéma à deux points pour le modèle d'intrusion saline (1.10). Nous sommes également intéressés par l'étude du comportement en temps long du modèle d'intrusion saline et du schéma associé. La question de la préservation de certaines propriétés vérifiées par le problème continu est essentielle pour un schéma numérique. Il est important de concevoir des schémas numériques pour lesquels on a par exemple la positivité de la solution, la conservation de la masse, la décroissance de l'entropie et le contrôle de sa dissipation. Beaucoup de travaux ont été consacrés à l'étude du comportement en temps long des schémas

numériques. Dans [90] l'auteur s'est intéressé à l'étude schéma Volumes Finis pour un modèle de coagulation-fragmentation et a mené une investigation numérique pour le comportement en temps long. On peut se référer à [18] pour l'équation de Fokker-Planck, à [101] pour l'équation des milieux poreux, à [47, 114] pour les équations de diffusion non linéaires, à [99, 100] pour les systèmes de réaction diffusion et à [22, 49] pour les systèmes de dérive diffusion.

1.4 Organisation du manuscrit

Les travaux effectués dans cette thèse correspondent dans un premier temps à l'étude d'un schéma numérique pour l'équation de Richards avec anisotropie. Ce schéma est convergent sans restriction sur le maillage du domaine spatial ni sur les coefficients de transmissibilité. Dans un second temps nous proposons et analysons un schéma Volumes Finis non-linéaire pour un modèle d'intrusion saline, pour finir par étudier le comportement en temps long d'un modèle similaire.

1.4.1 Un schéma CVFE non-linéaire pour l'équation de Richards anisotrope

Dans le chapitre 2, on s'intéresse à un schéma CVFE (Control Volume Finite Element) [107], pour l'approximation numérique de la solution exacte de l'équation de Richards (1.3), décrivant l'écoulement de l'eau dans les milieux insaturés (voir [84]) :

$$\begin{cases} \partial_t s(p) - \nabla.(\eta(s(p))\Lambda(\nabla p - \varrho \mathbf{g}) = 0 & \text{dans} \quad Q_{t_f}, \\ s(p)_{t=0} = s_0 & \text{dans} \quad \Omega, \\ \eta(s(p))\Lambda(\nabla p - \varrho \mathbf{g}).\mathbf{n} = 0 & \text{sur} \quad \partial\Omega \times (0, t_f), \end{cases} \tag{1.28}$$

où le milieu poreux Ω est un ouvert borné de $\mathbb{R}^2$. L'inconnue est la pression p, $s(p)$ est la saturation, $\eta(s)$ est la fonction de mobilité, Λ est la perméabilité intrinsèque et $\mathbf{g}$ la gravité.

La méthode que nous proposons a été conçue pour :

(a) supporter des tenseurs anisotropes et hétérogènes,

(b) ne pas introduire des quantités non physiques dans la conception du schéma numérique, comme par exemple la transformée de Kirchhoff,

(c) préserver les bornes physiques sur la saturation,

(d) conserver localement la masse du fluide,

(e) converger vers la solution du problème continu.

Nous proposons le schéma CVFE (Control Volume Finite Element) non-linéaire. C'est une extension au cas de l'équation de Richards du problème étudié dans [38, 39]. Il est basé sur une discrétisation très simple des termes de diffusion par la méthode des Éléments Finis avec condensation de masse et une discrétisation Volumes Finis pour les autres termes. Les degrés de liberté sont affectés aux sommets du maillage triangulaire primal, tandis que les équations de bilan sont discrétisées sur un maillage dual barycentrique. La mobilité d'arête η est gérée à l'aide d'une procédure de décentrement qui autorise les transmissivités négatives et donc les pertes de monotonie. Avec les mêmes notations que dans la Section 1.3 on a :

$$\frac{s(p_K^{n+1}) - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} \eta_{KL}^{n+1} a_{KL} (u_K^{n+1} - u_L^{n+1}) = 0,$$

où $s_K^0 = \dfrac{1}{m_K} \displaystyle\int_{w_K} s_0(x)\,\mathrm{d}x, \quad s_K^n = s(p_K^n)$ si $n \geq 1, \quad u_K^{n+1} = p_K^{n+1} - z_K$, $z_K = z(x_K)$.
On pose $a_{KL} = -\displaystyle\int_\Omega \Lambda(x) \nabla e_K(x) \cdot \nabla e_L(x)\,\mathrm{d}x$.

$$\eta_{KL}^{n+1} = \begin{cases} \eta(s_K^{n+1}) & \text{si} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) \geq 0, \\ \eta(s_L^{n+1}) & \text{si} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) < 0. \end{cases}$$

Ce schéma possède quelques propriétés remarquables. En particulier il permet de traiter le cas anisotrope contrairement au schéma proposé dans [84]. Nous avons démontré que ce schéma est non-linéairement stable, grâce à un contrôle de l'énergie physique, qu'il admet (au moins) une solution discrète et que la saturation est bornée entre 0 et 1 sans aucune restriction sur le maillage et sur le tenseur d'anisotropie. Ensuite nous avons montré, sous l'hypothèse faible de régularité sur le maillage, que la solution discrète converge vers la solution faible du problème continu (1.28). Enfin, en vue de mettre en évidence l'efficacité, la stabilité et la robustesse du schéma, nous avons présenté des tests numériques dans des cas isotropes et anisotropes.

La Figure 1.8 met en évidence la convergence du schéma proposé, avec un ordre 1 en espace. La Figure 1.9 montre la décroissance de l'énergie libre discrète au cours du temps. La Figure 1.10 montre que notre schéma non-linéaire ne produit pas des undershoots sur la saturation.

Notons une difficulté liée à la résolution du système issu du schéma numérique, du fait de sa non-linéarité, qu'on résout par la méthode de Newton. Cette méthode est extrêmement précise. En revanche, elle nécessite une initialisation relativement proche de la solution que l'on cherche. Utiliser celle-ci à partir d'un point quelconque peut conduire à des résultats numériquement instables. Pour contrevenir aux instabilités dues à l'initialisation de la méthode de Newton, on applique une procédure d'adaptation du pas de temps.

Ce travail est accepté et va paraître dans le journal *ESAIM :M2AN* (Mathematical Modelling and Numerical Analysis) ; il est disponible en ligne (voir [9]).

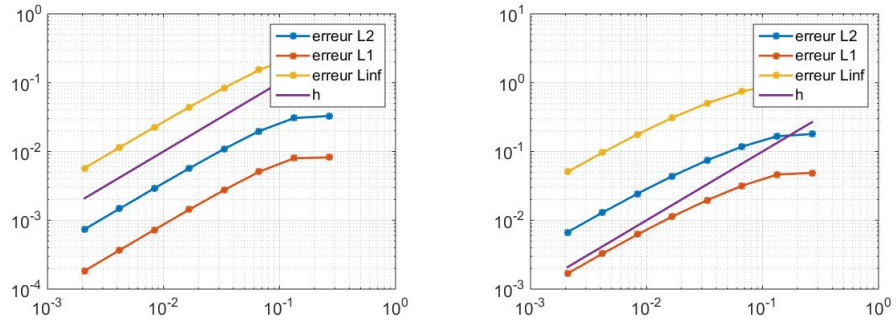


FIGURE 1.8 – Erreur de convergence dans le cas isotrope (*à gauche*) et anisotrope (*à droite*)

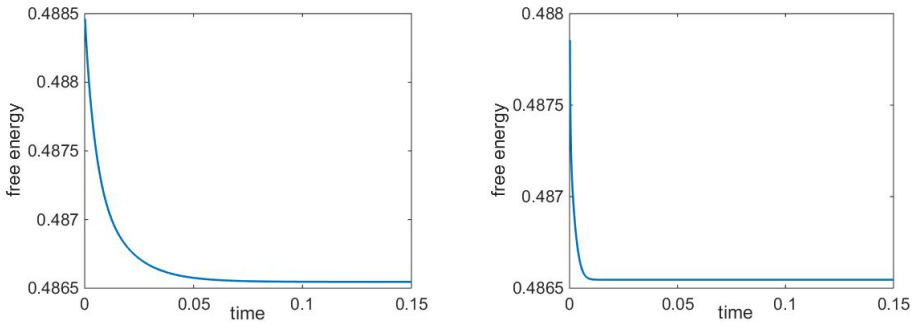


FIGURE 1.9 – Décroissance de l'énergie libre au cours du temps dans le cas isotrope (*à gauche*) et anisotrope (*à droite*)

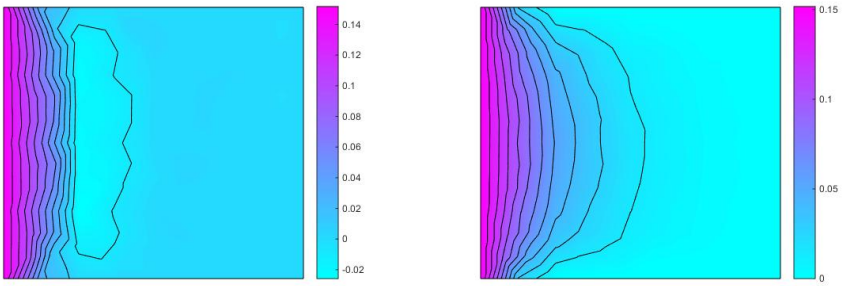


FIGURE 1.10 – Schéma quasi-linéaire (*à gauche*) et non-linéaire (*à droite*)

1.4.2 Analyse numérique d'un schéma Volumes Finis pour un modèle d'intrusion saline

Dans le chapitre 3 on s'intéresse à l'intrusion de l'eau de mer dans un aquifère non confiné (cf. Figure 1.11). On considère le modèle (1.10) obtenu par Jazar-Monneau [108], dans lequel on suppose les déplacements quasi-horizontaux et des interfaces nettes. De plus, par intégration verticale on a un problème 2D.

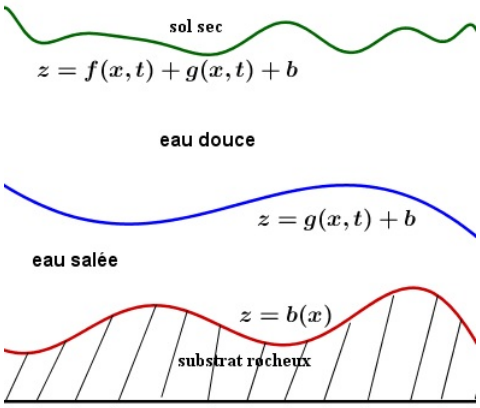


FIGURE 1.11 – Aquifère non confiné

On se propose donc d'étudier et d'analyser un schéma Volumes Finis pour le système d'équations paraboliques dégénérées avec diffusion croisée (1.10) :

$$\begin{cases} \partial_t f - \nabla.\Big(\nu f \nabla(f+g+b)\Big) = 0 \quad \text{dans} \quad \Omega \times (0,T), \\ \partial_t g - \nabla.\Big(g \nabla(\nu f+g+b)\Big) = 0 \quad \text{dans} \quad \Omega \times (0,T), \end{cases} \tag{1.29}$$

où le ratio des densités $\nu \in (0,1)$, $\Omega \subset \mathbb{R}^2$ un ouvert borné et $T > 0$. On ajoute les conditions initiales

$$f_{|t=0} = f_0, \quad g_{|t=0} = g_0,$$

et on impose un flux nul au bord

$$\nabla f \cdot \mathbf{n} = \nabla g \cdot \mathbf{n} = 0, \quad \text{sur } \partial\Omega \times (0,T), \tag{1.30}$$

où $\mathbf{n}$ est le vecteur normal au bord $\partial\Omega$. On suppose que f_0, g_0 sont positives et bornées.

On propose une discrétisation du système (4.9) : la discrétisation en temps repose sur le schéma d'Euler implicite et la discrétisation en espace repose sur un schéma Volumes Finis à deux points [78] avec un décentrement des mobilités d'arêtes. On note

$$f_{\mathcal{D}}^0 = \sum_{K \in \mathcal{T}} f_K^0 \mathbf{1}_K, \quad \text{où } f_K^0 = \frac{1}{m(K)} \int_K f_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T},$$

$$g_{\mathcal{D}}^0 = \sum_{K\in\mathcal{T}} g_K^0 \mathbf{1}_K, \quad \text{où } g_K^0 = \frac{1}{m(K)} \int_K g_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T},$$

où $\mathbf{1}_K$ est la fonction caractéristique de K. On note f_K^n et g_K^n les approximations des valeurs moyennes de $f(.,t^n)$ et $g(.,t^n)$ sur K, respectivement. La discrétisation du problème (4.9) est donnée par les equations non linéaires suivantes :

$$m(K)\frac{f_K^{n+1}-f_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma f_\sigma^{n+1} \nu \left((f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) \right) = 0,$$

et

$$m(K)\frac{g_K^{n+1}-g_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\rm int}} \tau_\sigma g_\sigma^{n+1} \left(\nu(f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) \right) = 0,$$

pour tout $K \in \mathcal{T}$ and $0 \le n \le M_T - 1$, où

$$f_\sigma^{n+1} = \begin{cases} (f_K^{n+1})^+ & \text{si } (f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) \ge 0, \\ (f_L^{n+1})^+ & \text{si } (f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) < 0. \end{cases}$$

et

$$g_\sigma^{n+1} = \begin{cases} (g_K^{n+1})^+ & \text{si } \nu(f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) \ge 0, \\ (g_L^{n+1})^+ & \text{si } \nu(f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L) < 0, \end{cases}$$

où $x^+ = \max(0,x)$.

Pour ce schéma nous démontrons dans le chapitre 3 qu'il préserve au niveau discret les principales propriétés du problème continu : l'existence de solutions discrètes positives, la décroissance de l'énergie et le contrôle de l'entropie et sa dissipation. En se basant sur ces résultats, nous avons montré que la solution discrète converge vers la solution faible du problème continu (4.9). Enfin, nous avons mis au point un code Matlab qui permet de résoudre numériquement le problème (4.9). Notons que nous procédons de la même manière qu'au chapitre 2 pour l'adaptation du pas de temps.

La Figure 1.12 met en évidence la décroissance de l'énergie au cours du temps et le contrôle de l'entropie. La Figure 3.3 montre l'évolution du modèle, pour lequel on a la convergence vers un état d'équilibre, avec des interfaces horizontales comme attendu.

Ce travail est accepté et va paraître dans le journal *NMPDE* (Numerical Methods for Partial Differential Equations) ; il est disponible en ligne (voir [8]). Il a donné lieu à un acte pour la conférence Finite Volumes for Complex Applications 8 (voir [7]).

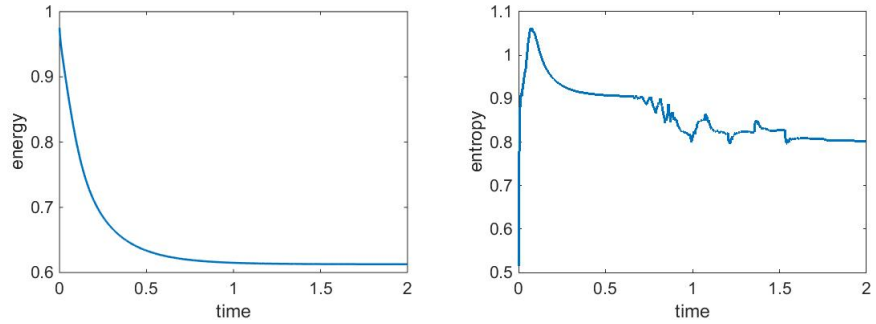


FIGURE 1.12 – Décroissance de l'énergie au cours du temps (*à gauche*) et le contrôle de l'entropie (*à droite*)

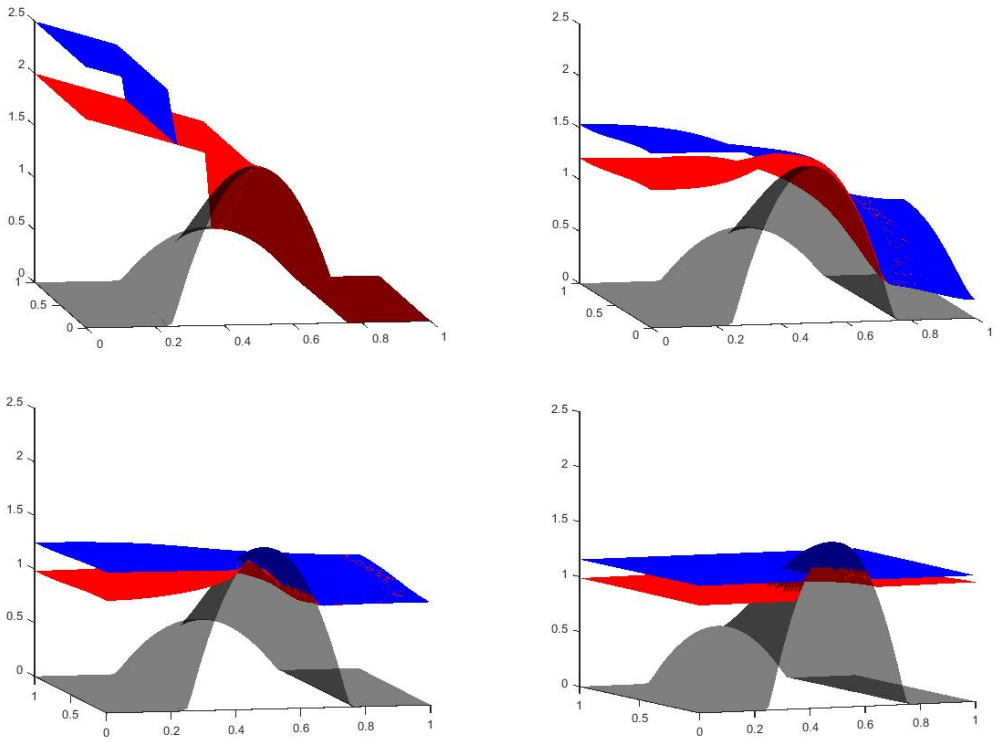


FIGURE 1.13 – Comportement du modèle aux temps $t = 0,\ t = 0.2,\ t = 0.72,\ t = 12$. Le substrat rocheux en noir, l'interface entre l'eau salée et l'eau douce en rouge et en bleu l'interface entre l'eau douce et le sol sec.

1.4.3 Comportement en temps long d'un modèle d'intrusion saline

L'objectif du chapitre 4 est l'étude du comportement en temps long d'un modèle d'intrusion saline. C'est un modèle qui décrit le mouvement de deux fluides avec des densités et viscosités différentes dans un aquifère non confiné, où le substrat rocheux (bedrock) est placé à $z = 0$. Il s'agit du modèle

$$\begin{cases} \partial_t \tilde{f} - \nabla.\Big(\nu\mu\tilde{f}\nabla(\tilde{f}+\tilde{g})\Big) = 0 & \text{dans} \quad (0,T)\times\mathbb{R}^2, \\ \partial_t \tilde{g} - \nabla.\Big(\tilde{g}\nabla(\nu\tilde{f}+\tilde{g})\Big) = 0 & \text{dans} \quad (0,T)\times\mathbb{R}^2, \\ \tilde{f}_{|t=0} = \tilde{f}_0, \quad \tilde{g}_{|t=0} = \tilde{g}_0 & \text{sur} \ \ \mathbb{R}^2, \end{cases} \tag{1.31}$$

où ν (resp. μ) est le ratio des densités (resp. viscosités). Ce système est une généralisation de l'équation des milieux poreux. Comme dans la section 1.2, en passant en variables auto-similaires

$$(f,g) = \frac{1}{(1+t)^{1/2}}(\tilde{f},\tilde{g})(\log(1+t), \frac{x}{(1+t)^{1/4}}),$$

le système (1.31) se réécrit comme

$$\begin{cases} \partial_t f - \nabla.\Big(\nu\mu f\nabla(f+g+b/\mu)\Big) = 0 & \text{dans} \quad (0,T)\times\mathbb{R}^2, \\ \partial_t g - \nabla.\Big(g\nabla(\nu f+g+b)\Big) = 0 & \text{dans} \quad (0,T)\times\mathbb{R}^2, \\ f_{|t=0} = f_0, \quad g_{|t=0} = g_0 & \text{sur} \ \ \mathbb{R}^2, \end{cases} \tag{1.32}$$

où $b(x) = \dfrac{1-\nu}{8}|x|^2$. L'étude du comportement en temps long consiste en l'identification d'un état d'équilibre (F,G) et une fonctionnelle d'énergie convexe $\mathfrak{E}$. On mesure l'écart entre (f,g) et (F,G) par l'énergie relative

$$\mathfrak{E}\Big(f,g|F,G\Big) = \mathfrak{E}(f,g) - \mathfrak{E}(F,G).$$

Nous allons étendre les résultats de [121] au cas de $\mathbb{R}^2$ et de plus exhiber numériquement la vitesse de convergence. Nous devons utiliser un schéma numérique pour lequel nous avons au niveau discret certaines propriétés vérifiées par le problème continu (en particulier la positivité de la solution, conservation de la masse, décroissance de l'énergie) et qui converge vers un état stationnaire. Ainsi nous pouvons utiliser le schéma Volumes Finis étudié dans le chapitre 3. Mais remarquons d'abord que le problème (1.32) est une généralisation à deux phases de l'équation des milieux poreux écrite en variables auto-similaires

$$\partial_t f = \operatorname{div}\left(f\nabla\Big(f+\frac{|x|^2}{8}\Big)\right), \quad (t,x)\in(0,\infty)\times\mathbb{R}^2. \tag{1.33}$$

On a vu dans la section 1.2 que les solutions faibles de (1.33) convergent exponentiellement vers les solutions de Barenblatt. Ce résultat sur la décroissance de l'entropie relative. Nous allons étendre l'approche rappelée dans la section 1.2 au cas de notre problème et montrer la convergence vers un état d'équilibre et exhiber numériquement la vitesse de convergence.

Soit $\mathfrak{E}$ la fonctionnelle d'énergie strictement convexe

$$\mathfrak{E}(f,g) = \int_\Omega E(f,g)\,\mathrm{d}x,$$

où

$$E(f,g) = \frac{\nu}{2}(f+g)^2 + \frac{1-\nu}{2}g^2 + b\Big(\frac{\nu}{\mu}f + g\Big). \tag{1.34}$$

Considérons le problème de minimisation suivant

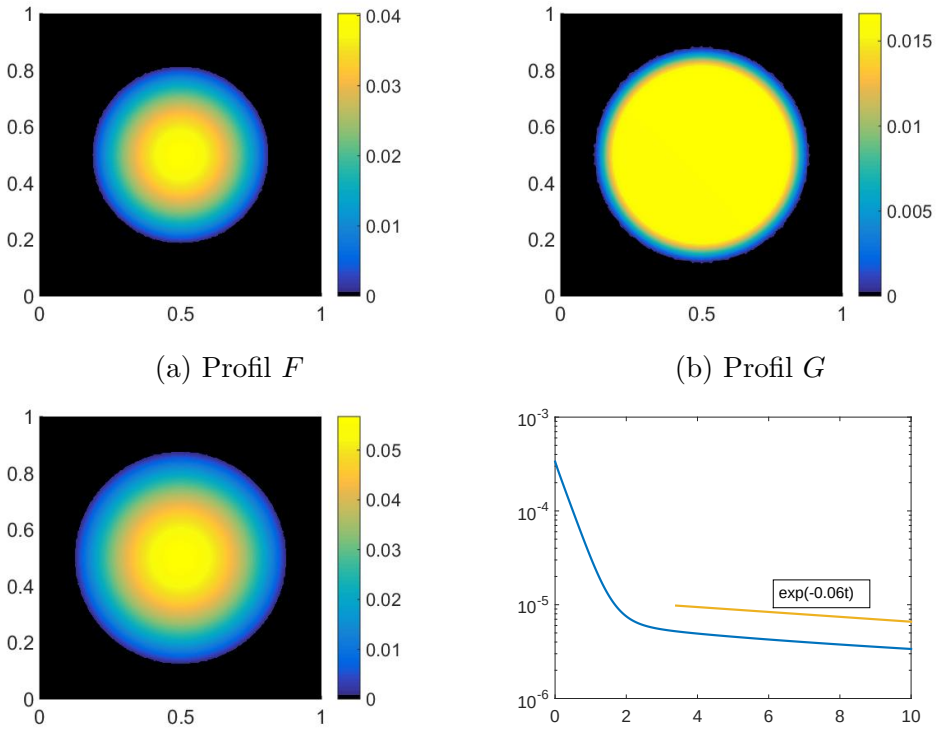
$$\inf_{(f,g)\in\mathcal{K}_f\times\mathcal{K}_g} \mathfrak{E}(f,g), \tag{1.35}$$

avec

$$\mathcal{K}_h := \Big\{ h \in L^2(\mathbb{R}^2)\cap L^1(\mathbb{R}^2,(1+|x|^2\,\mathrm{d}x)) : h \geq 0 \ \text{ p.p et} \\ \int_{\mathbb{R}^2} h(x)\,\mathrm{d}x = \int_{\mathbb{R}^2} h_0(x)\,\mathrm{d}x = M_h \Big\}.$$

Nous assurons l'existence et l'unicité des minimiseurs de $\mathfrak{E}$. De plus ils sont des états d'équilibre, autrement dit (F,G) est un état d'équilibre de (1.32) si et seulement si c'est un minimiseur de $\mathfrak{E}$ dans $\mathcal{K}_f \times \mathcal{K}_g$. Ensuite nous montrons que ces états d'équilibres sont uniques, radiaux, lipschitziens et à support compact. De plus, nous caractérisons les états stationnaires en faisant varier μ en fonction de ν, M_f et M_g. Les valeurs critiques de μ sont des valeurs seuils pour la topologie des supports $E_F = \{x \in \mathbb{R}^2 \mid F(x) > 0\}$ et $E_G = \{x \in \mathbb{R}^2 \mid G(x) > 0\}$. Les valeurs $\mu = \nu$ et $\mu = 1$ sont des valeurs critiques pour la courbure de l'interface entre les fluides. On présente dans la Figure 1.14 les états stationnaires et la décroissance de l'énergie pour $\mu = \nu = 0.9$. Les différentes configurations seront présentées au chapitre 4.

Ce travail est réalisé en collaboration avec C. Cancès, C. Chainais-Hillairet et P. Laurençot.



(a) Profil F

(b) Profil G

(c) Profil $F + G$

(d) L'énergie relative discrète

FIGURE 1.14 – Les états stationnaires et l'énergie relative pour $\mu = 0.90$

Chapitre 2

Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy

Abstract : We extend the nonlinear Control Volume Finite Element scheme of [C. Cancès and C. Guichard, *Math. Comp.*, 85 (298) : 549-580, **2016**] to the discretization of Richards equation. This scheme ensures the preservation of the physical bounds without any restriction on the mesh and on the anisotropy tensor. Moreover, it does not require the introduction of the so-called Kirchhoff transform in its definition. It also provides a control on the capillary energy. Based on this nonlinear stability property, we show that the scheme converges towards the unique solution to Richards equation when the discretization parameters tend to 0. Finally we present some numerical experiments to illustrate the behavior of the method.

2.1 Introduction

2.1.1 Presentation of the continuous problem

We are interested in the numerical approximation of Richards equation. It is a degenerate nonlinear parabolic equation modeling unsaturated flow in porous media. The diffusion terms can be anisotropic and heterogeneous. In order to ease the reading, we restrict our study to the case of a two-dimensional porous medium. However, the extension of our purpose to the three-dimensional framework does not lead to any theoretical difficulty.

Let Ω be a polygonal connected open bounded subset of $\mathbb{R}^2$, and $t_f > 0$ a finite time horizon. We define $Q_{t_f} = \Omega \times (0, t_f)$. The Richards equation writes :

$$\begin{cases} \partial_t s(p) - \nabla \cdot (\eta(s(p))\Lambda(\nabla p - \rho \mathbf{g})) = 0 & \text{in} \quad Q_{t_f}, \\ s(p)_{t=0} = s_0 & \text{in} \quad \Omega, \\ \eta(s(p))\Lambda(\nabla p - \rho \mathbf{g}) \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0, t_f). \end{cases} \tag{2.1}$$

In (2.1), p denotes the water pressure, s the water content, η the water mobility function, Λ the intrinsic permeability tensor, and $\mathbf{g}$ is the gravity. We do the following assumptions on the data of the continuous problem (2.1) :

(**A1**) The function $s : \mathbb{R} \to [0, 1]$ is increasing on $\mathbb{R}_-$ and takes the value 1 on $\mathbb{R}_+$. We assume that there exists $p_\star \in [-\infty, 0)$ such that $s(p_\star) = 0$, and that $s \in L^1(p_\star, 0)$. Figure 2.1 shows two typical profiles of the function s.

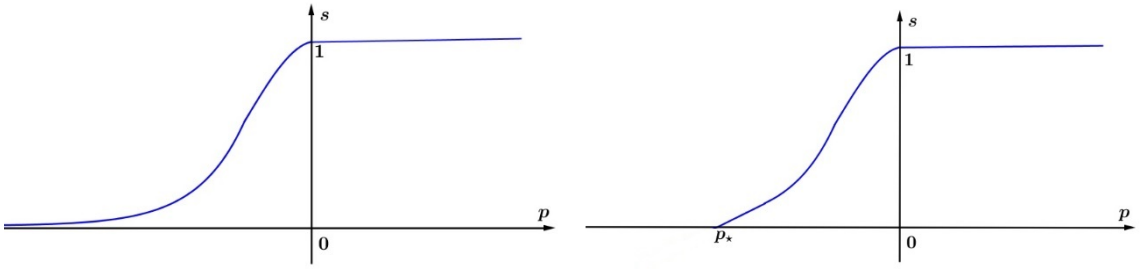


FIGURE 2.1 – Typical water content functions. Two distinct behaviors are allowed in our study : *(left)* either the function s remains strictly positive on $\mathbb{R}$ but tends to 0 as p tends to $-\infty$ and $p_\star = -\infty$, or *(right)* there exists a finite value of $p_\star$ such that $s(p_\star) = 0$.

(**A2**) The water mobility function $\eta : [0, 1] \longrightarrow \mathbb{R}^+$ is assumed to be bounded, continuous, nondecreasing, and to fulfill (cf. Figure 2.2)

$$\eta(0) = 0 \quad \text{and} \quad \eta(s) > 0 \ \text{ if } \ s \neq 0. \tag{2.2}$$

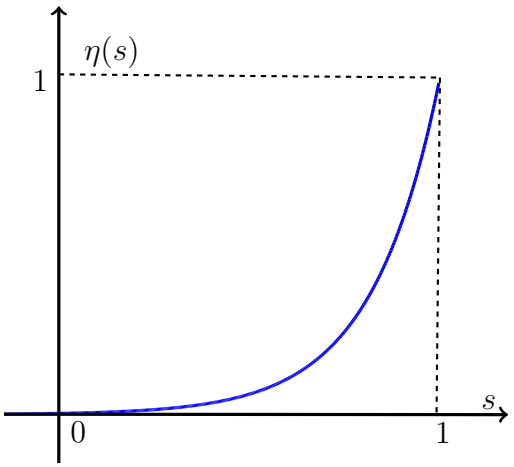


FIGURE 2.2 – Typical water mobility function η.

Moreover, we assume all along in this chapter that

$$\xi_\star := \int_{p_\star}^{0} \sqrt{\eta(s(a))}\,\mathrm{d}a < +\infty. \tag{2.3}$$

Remark that (2.3) is trivially satisfied if $p_\star > -\infty$.

(**A3**) The permeability tensor Λ belongs $(L^\infty(\Omega))^{2\times 2}$, and it is supposed to be symmetric and uniformly elliptic on Ω, i.e, there exists $(\overline{\Lambda}, \underline{\Lambda}) \in \mathbb{R}_+^* \times \mathbb{R}_+^*$ such that

$$\underline{\Lambda}|\mathbf{v}|^2 \leq \Lambda(\mathbf{x})\mathbf{v}\cdot\mathbf{v} \leq \overline{\Lambda}|\mathbf{v}|^2, \quad \forall \mathbf{v} \in \mathbb{R}^2, \quad \text{for a.e.} \quad \mathbf{x} \in \Omega.$$

(**A4**) The initial data s_0 is supposed to belong to $L^\infty(\Omega;[0,1])$, and we assume

$$0 < \overline{s}_0 := \frac{1}{\mathrm{meas}(\Omega)} \int_\Omega s_0(x)\,\mathrm{d}x < 1. \tag{2.4}$$

Since s is continuous and increasing on $[p_\star, 0]$, there exists a continuous and increasing function $s^{-1} : [0,1] \to [p_\star, 0]$ such that, $s \circ s^{-1}(\zeta) = \zeta$ for all $\zeta \in [0,1]$. Simple calculations (see for instance [41]) show that

$$\|s^{-1}\|_{L^1(0,1)} = \|s\|_{L^1(p_\star,0)} \leq C. \tag{2.5}$$

thanks to (**A1**).

Remark 2.1.1. *The assumptions (***A1***) and (***A2***) impose some constraints on the nonlinearities $p \mapsto s(p)$ and $s \mapsto \eta(s)$. Let us connect these constraints to two very classical models.*

- van Genuchten-Mualem *model [127, 148] (see also [135]) :* $p_\star = -\infty$ *and the function* s *is chosen as*

$$s(p) = \begin{cases} (1+\alpha|p|^n)^{\frac{1-n}{n}} & \text{if } p < 0, \\ 1 & \text{if } p \geq 0, \end{cases}$$

where $\alpha > 0$ *is a fixed parameter. The condition (***A1***) is fulfilled if* $n > 2$*. The function* $s^{-1} : (0,1] \to (-\infty, 0]$ *is then given by*

$$s^{-1}(s) = -\left(\frac{s^{-\frac{n}{n-1}} - 1}{\alpha}\right)^{1/n}, \qquad \forall s \in (0,1].$$

whereas η *is given by*

$$\eta(s) = \kappa\sqrt{s}\left(\int_0^s \frac{1}{s^{-1}(a)}\mathrm{d}a\right)^2, \qquad \forall s \in [0,1]$$

for some $\kappa > 0$*. Condition (***A2***) —in particular* (2.3)*— is fulfilled.*

- Brooks-Corey *model [35] : here again,* $p_\star = -\infty$*. Let* $p_b < 0$ *and* $\lambda > 0$ *be given, then the function* s *is chosen as*

$$s(p) = \begin{cases} \left(\frac{p+p_b}{p_b}\right)^{-\lambda} & \text{if } p < 0, \\ 1 & \text{if } p \geq 0. \end{cases}$$

The integrability condition on s *in (***A1***) is fulfilled as soon as* $\lambda > 1$*. The mobility function* η *is then chosen as*

$$\eta(s) = \kappa s^{3+\frac{2}{\lambda}}, \qquad \forall s \in [0,1]$$

for some $\kappa > 0$*. Here again, Condition* (2.3) *of (***A2***) is fulfilled.*

We define the function $\Gamma : \mathbb{R} \to \mathbb{R}_+$ (called *capillary energy function*) by

$$\Gamma(p) = \int_0^p a s'(a)\,\mathrm{d}a. \tag{2.6}$$

The function $\Gamma \circ s^{-1}$ is convex on $[0,1]$, and it follows from the definition (2.6) that

$$\partial_t \Gamma(p) = p \partial_t s(p). \tag{2.7}$$

In order to give a proper mathematical sense to the solution of (2.1), we need to introduce the Lipschitz continuous increasing function $\xi : \mathbb{R} \longrightarrow \mathbb{R}$ defined by

$$\xi(p) = \int_0^p \sqrt{\eta(s(a))}\,\mathrm{d}a, \quad \forall\, p \in \mathbb{R}. \tag{2.8}$$

Since $\sqrt{\eta \circ s \circ \xi^{-1}}$ is uniformly continuous, it admits a modulus of continuity i.e., there exits a continuous function $\mu : \mathbb{R}_+ \longrightarrow \mathbb{R}_+$ such that

$$\mu(0) = 0, \quad |\sqrt{\eta \circ s \circ \xi^{-1}}(x) - \sqrt{\eta \circ s \circ \xi^{-1}}(y)| \leq \mu(|x-y|), \ \forall x, y \in [\xi_\star, +\infty). \quad (2.9)$$

We introduce the so-called hydraulic head u defined by

$$u(\boldsymbol{x}, t) = \frac{p(\boldsymbol{x}, t)}{\rho g} + z(\boldsymbol{x}) \quad \text{for all} \quad (\boldsymbol{x}, t) \in Q_{t_f}, \quad \text{for all} \quad t_f > 0,$$

where g denotes the modulus of $\mathbf{g}$ and the function $z(\boldsymbol{x})$ is the projection of the point $\boldsymbol{x}$ on the vertical axis, oriented upward by $-\mathbf{g}/g$. With a simple adimensionalization, we can assume that $\rho g = 1$. The system (2.1) then rewrites :

$$\begin{cases} \partial_t s(p) - \nabla \cdot (\eta(s(p)) \Lambda \nabla u) = 0 & \text{in} \quad Q_{t_f}, \\ s(p)_{t=0} = s_0 & \text{in} \quad \Omega, \\ \eta(s(p)) \Lambda \nabla u \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times (0, t_f), \\ u = p + z & \text{in} \quad Q_{t_f}. \end{cases} \quad (2.10)$$

Remark 2.1.2. *In §2.5.1, we will present a test case without gravity.* Stricto sensu, *this case is not included in our study. But it corresponds to the simpler case $p = u$ (mainly carried out in [39]) for which our analysis can be straightforwardly adapted.*

Multiplying (formally) the equation (2.10) by u and integrating on Ω yields the classical energy/dissipation property :

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_\Omega (\Gamma(p) + s(p) z(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x} + \int_\Omega \eta(s(p)) \Lambda \nabla u \cdot \nabla u \, \mathrm{d}\boldsymbol{x} = 0, \qquad \forall \, t \in (0, t_f). \quad (2.11)$$

This allows in particular to show that the capillary energy remains bounded and that the function $\xi(p)$ belongs to $L^2((0,T); H^1(\Omega))$, i.e.,

$$\int_\Omega \Gamma(p(\boldsymbol{x}, t)) \mathrm{d}\boldsymbol{x} + \int_0^t \int_\Omega |\nabla \xi(p(\boldsymbol{x}, \tau))|^2 \mathrm{d}\boldsymbol{x} \mathrm{d}\tau \leq C, \qquad \forall t \in (0, t_f). \quad (2.12)$$

Remark 2.1.3. *It is clear that the quantity $\int_\Omega s(p) z(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$ represents the* gravitational (potential) energy *of the fluid. Therefore, it follows from* (2.11) *that the quantity $\int_\Omega \Gamma(p) \mathrm{d}\boldsymbol{x}$ also corresponds to an energy. Since it originates from capillary (or suction) effects, we call it* capillary energy. *The* free energy $\int_\Omega (\Gamma(p) + s(p) z(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x}$ *is then obtained as the sum of the capillary and gravitational energies. It is supposed to decay with time thanks to* (2.11) *(see also §2.5.4 for a numerical evidence).*

Definition 2.1.4 (weak solution)**.** *A measurable function $p : Q_{t_f} \to \mathbb{R}$ is said to be a weak solution of (2.1) if $p \geq p_\star$ a.e. in Q_{t_f}, if $\xi(p)$ belongs to $L^2((0, t_f); H^1(\Omega))$, and if, for all $\psi \in \mathcal{C}_c^\infty(\overline{\Omega} \times [0, t_f))$, one has*

$$\begin{aligned} &\iint_{Q_{t_f}} s(p) \partial_t \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t + \int_\Omega s_0 \psi(., 0) \, \mathrm{d}\boldsymbol{x} \\ &- \iint_{Q_{t_f}} \sqrt{\eta(s(p))} \Lambda \nabla \xi(p) \cdot \nabla \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \iint_{Q_{t_f}} \eta(s(p)) \rho g \Lambda \nabla z \cdot \nabla \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t = 0. \end{aligned} \quad (2.13)$$

The notion of weak solution is motivated by the following theorem.

Theorem 2.1.5. *Under assumptions (**A1**)–(**A4**), there exists a unique weak solution to the problem (2.1) in the sense of Definition 2.1.4.*

The existence of a solution is a by-product of the convergence of the scheme proved in §3.4. It can also be obtained by compactness arguments following the program of Alt and Luckhaus [11]. Concerning the uniqueness, since we consider no-flux boundary conditions, we can not directly apply Otto's result [133], where Dirichlet boundary conditions are imposed. However, a slight adaptation of Otto's proof detailed in appendix (cf. Proposition 2.7.4) allows us to extend the uniqueness result to our framework.

2.1.2 Goal and positioning of the work

Because of its broad interest in the environmental studies, the Richards equation [140] has been the purpose of many research papers, especially in the field of numerical analysis. Richards equation is locally conservative and a particular effort was made to preserve this property in most of the contributions.

A conservative Finite Difference scheme has been studied numerically in [151]. However, there is up to our knowledge no convergence proof for the scheme presented in [151]. Moreover, restrictive conditions have to be prescribed on the grid and on the permeability tensor Λ. The convergence of Two-Point Flux Approximation Finite Volume schemes have been studied in [84] for a scheme that requires the introduction of the Kirchhoff transform, and in [82] for a scheme expressed in physical variables (saturation and pressure), but under the non-physical assumption that the mobility function was not degenerated (i.e., $\eta(s) \geq \eta_\star > 0$ for all s). In both [84] and [82], it was moreover required that the porous medium was isotropic (i.e., $\Lambda = \lambda \mathrm{Id}$) and that the mesh satisfies the so-called orthogonality condition (see, e.g., [78, Definition 9.1] and [77]) so that the two-point flux approximation is consistent. Since they are naturally locally conservative, Mixed Finite Elements have been widely used for the approximation of Richards equation. Let us for instance mention [15, 137, 138] where the authors managed to provide an error estimate. Nevertheless, the schemes studied in [15, 137, 138] rely on the introduction of the Kirchhoff transform, and on a regularization of the problem in [138] to overcome the difficulties due the degeneracy. Let us also mention the extension of Multi-Point Flux Approximation Finite Volume schemes to the context of Richards equation in [29, 117]. Note that Mixed Finite Elements and Multi-Point Flux Approximation Finite Volumes may produce over- and undershoots on the saturation. We refer to [67] for a review of the numerous Finite Volume methods developed in the last decades that can be applied to the discretization of Richards equation.

The method we study here was designed on the following specifications :

(a) to handle anisotropic and heterogeneous anisotropy tensors ;

(b) to avoid the introduction of non-physical quantities like, e.g., the Kirchhoff transform ;

(c) to preserve the physical bounds on the saturation ;

(d) to conserve locally the mass of fluid ;

(e) to converge towards the solution to the continuous problem (mathematical proof and numerical evidence).

The scheme we propose belongs to the family of the so-called Control Volume Finite Element schemes introduced in the context of porous media flows by Forsyth [91, 92]. Roughly speaking, it consists in an interpretation of Finite Elements with mass lumping as a locally conservative method on dual cells. It was already noticed in [91] that the grid had to fulfill some restrictive condition unless the transmissivities may become negative. It results that the reconstructed numerical flux goes the opposite sense to the physical one. More precisely, the triangular grid has to fulfill a so-called Delaunay condition in the two-dimensional isotropic case $\Lambda = \lambda \mathrm{Id}$. But in the case where Λ is a spatially varying full tensor, there is no algorithm up to our knowledge to build a triangulation such that the transmissivity remain nonnegative. As it will be proved in the sequel, the method we propose still converges even in the case where negative transmissivities appear. Our scheme is an extension of the one studied in [38, 39]. It is based on a suitable upwinding of the mobility (i.e., w.r.t. the numerical flux and not w.r.t. the physical one) that allows to preserve the physical bounds (but not the monotonicity as in [92]). Moreover, we show that our method provides a control on the capillary energy and that this control is sufficient to perform a convergence proof based on compactness arguments. Alternatively, the convergence of the Finite Volume approximation can be obtained by means of error estimates (see [136] in the case where $\mathbf{g} = 0$).

The chapter is organized as follows. In Section 4.5.1, we introduce the scheme and we state the main results of the chapter. Theorem 2.2.4 states the existence of a solution to scheme which preserves the physical bounds and for which the capillary energy and the energy dissipation are bounded uniformly w.r.t. the grid. Theorem 2.2.5 states the convergence of a sequence of approximate solutions given by the scheme to the unique weak solution to (2.10) (its uniqueness is proved in Appendix). In Section 2.3, we derive a priori estimates on the discrete solution. They allow us to prove in §2.3.3 the existence of a discrete solution to the nonlinear system corresponding to the scheme. Section 3.4 is devoted to the convergence proof of the scheme. This proof is based first on the compactness of the sequence of approximate solutions and then on the identification of the limit. We finally present numerical

experiments in Section 3.5, which confirm the theoretical results we proved. We take care to fairly present the advantages and the drawbacks of the method from a computational point of view.

2.2 The numerical scheme

2.2.1 Discretization of Q_{t_f}

Discretizations of Ω

The CVFE method requires the introduction of two different space discretizations of Ω : a *primal triangular mesh* and a *dual barycentric mesh*.

The *primal triangular mesh* is denoted by $\mathcal{T}$. It is a conformal triangular discretization of the polygonal domain Ω, consisting in open bounded separated triangles satisfying $\bigcup_{T\in\mathcal{T}} \overline{T} = \overline{\Omega}$. For $T \in \mathcal{T}$, we denote by $\boldsymbol{x}_T$ the center of gravity of T, by h_T the diameter of the triangle T, and by ρ_T the diameter of the largest ball inscribed in the triangle T. Then, we define the mesh diameter h and the mesh regularity $\theta_{\mathcal{T}}$ by

$$h = \max_{T\in\mathcal{T}} h_T, \quad \theta_{\mathcal{T}} = \max_{T\in\mathcal{T}} \frac{h_T}{\rho_T}.$$

We denote by $\mathcal{V}$ the set of the vertices of the discretization $\mathcal{T}$, located at positions $(\boldsymbol{x}_K)_{K\in\mathcal{V}}$. The set $\mathcal{E}$ of the edges of $\mathcal{T}$ is made of straight segments σ joining two vertices of $\mathcal{V}$. Given $T, T' \in \mathcal{T}$, we assume that $\overline{T}\cap\overline{T'}$ is either empty, or it is reduced to $\boldsymbol{x}_K$ for some $K \in \mathcal{V}$, or it consists in an edge σ belonging $\mathcal{E}$. For $T \in \mathcal{T}$, we denote by $\mathcal{E}_T$ the set of the edges of T : $\bigcup_{\sigma\in\mathcal{E}_T} \overline{\sigma} = \partial T$. We assume that $\mathcal{E} = \bigcup_{T\in\mathcal{T}} \mathcal{E}_{\mathcal{T}}$. Given two vertices $K, L \in \mathcal{V}$ of a triangle T, then the edge joining $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ is denoted by σ_{KL}. For $K \in \mathcal{V}$, one denotes by $\mathcal{T}_K$ the subset of $\mathcal{T}$ made the triangles admitting K as a vertex, by $\mathcal{E}_K$ the set of edges having the vertex K as an extremity, and by $\mathcal{V}_K$ the subset of $\mathcal{V}$ such that, if $L \in \mathcal{V}_K$, then $[\boldsymbol{x}_K, \boldsymbol{x}_L]$ is an edge of $\mathcal{E}_K$.

Once the *primal triangular mesh* has been built, we can define its *dual barycentric mesh* $\mathcal{M}$ as follows. To each $K \in \mathcal{V}$, we associate a cell ω_K whose vertices are the isobarycenters $\boldsymbol{x}_T$ of the triangles $T \in \mathcal{T}_K$ and the isobarycenters $\boldsymbol{x}_\sigma$ of the edges $\sigma \in \mathcal{E}_K$. Note that $\overline{\Omega} = \bigcup_{K\in\mathcal{V}} \overline{\omega}_K$. We refer to Figure 2.3 for an illustration of the primary and dual barycentric meshes. The 2-dimensional Lebesgue measure of ω_K is denoted by m_K.

Let us now introduce some useful functional spaces. The space $V_{\mathcal{T}} \subset \mathcal{C}(\overline{\Omega})$ is made of piecewise affine functions on the primal mesh, i.e.,

$$V_{\mathcal{T}} = \{f \in H^1(\Omega) \mid f_{|T} \text{ is affine}, \ \forall T \in \mathcal{T}\}.$$

For all $K \in \mathcal{V}$, we denote by e_K the unique element of $V_{\mathcal{T}}$ such that $e_K(\boldsymbol{x}_K) = 1$

and $e_K(\boldsymbol{x}_L) = 0$ if $L \in \mathcal{V} \setminus \{K\}$. The geometrical construction of ω_K ensures that

$$\int_\Omega e_K(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\omega_K} \mathrm{d}\boldsymbol{x} =: m_K, \qquad \forall K \in \mathcal{V}.$$

We can also define the set of the piecewise constant functions on $\mathcal{M}$, $X_\mathcal{M}$, by

$$X_\mathcal{M} = \{f : \Omega \longrightarrow \mathbb{R} \text{ measurable} \mid f_{|\omega_K} \in \mathbb{R} \text{ is constant}, \quad \forall K \in \mathcal{V}\}.$$

Given a vector $(u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}$, there exists a unique $u_\mathcal{T} \in V_\mathcal{T}$ and a unique $u_\mathcal{M} \in X_\mathcal{M}$ such that $u_\mathcal{T}(\boldsymbol{x}_K) = u_\mathcal{M}(\boldsymbol{x}_K) = u_K$ for all $K, L \in \mathcal{V}$. Let us note that $u_\mathcal{T} = \sum_{K\in\mathcal{V}} u_K e_K$. Moreover, for all $q \in [1, \infty)$, there exist C_1 and C_2 depending only on q and on $\theta_\mathcal{T}$ such that

$$C_1\|u_\mathcal{T}\|_{L^q(\Omega)} \leq \|u_\mathcal{M}\|_{L^q(\Omega)} \leq C_2\|u_\mathcal{T}\|_{L^q(\Omega)}, \qquad \forall\ (u_K)_{K\in\mathcal{V}} \in \mathbb{R}^{\#\mathcal{V}}. \tag{2.14}$$

A proof of the above inequalities can be found for instance in [40, Lemma A.6].

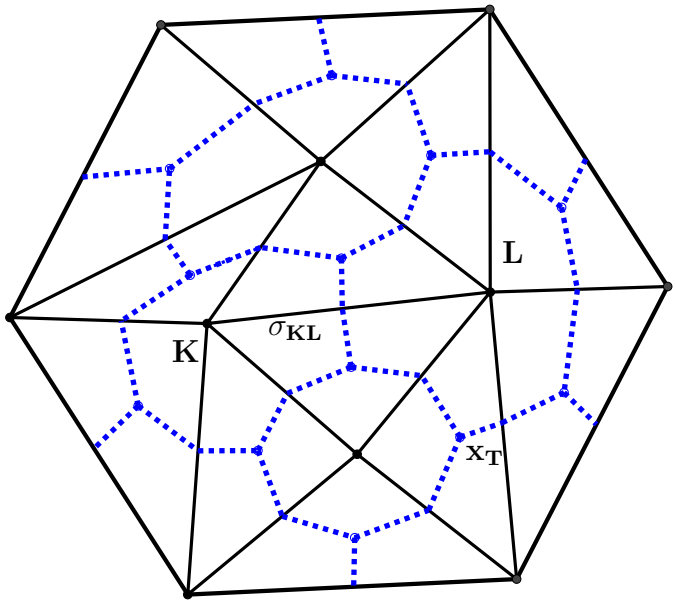


FIGURE 2.3 – The triangular mesh $\mathcal{T}$ (solid line) and its corresponding dual barycentric dual mesh $\mathcal{M}$ (dashed line).

Space-time discretizations

In order to avoid heavier notations, we restrict our study to the case of a uniform time discretization of $(0, t_f)$. However, all the results presented in this chapter can be extended to general time discretizations without any technical difficulty. In what

follows, we assume that the spatial mesh is fixed and does not change with the time step.

Let N be a nonnegative integer, then we define $\Delta t = \dfrac{t_f}{N+1}$, and $t_n = n\Delta t$ for all $n \in \{0, ..., N+1\}$, so that $t_0 = 0$, and $t_{N+1} = t_f$.

We define the space and time discrete spaces $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ as the set of piecewise constant functions in time with values in $V_{\mathcal{T}}$ and $X_{\mathcal{M}}$ respectively :

$$\begin{aligned} V_{\mathcal{T},\Delta t} =& \{f : Q_{t_f} \to \overline{\mathbb{R}} \mid f(x,t) = f(x,t^{n+1}) \in V_{\mathcal{T}}, \quad \forall t \in (t_n, t_{n+1}]\}, \\ X_{\mathcal{M},\Delta t} =& \{f : Q_{t_f} \to \overline{\mathbb{R}} \mid f(x,t) = f(x,t^{n+1}) \in X_{\mathcal{M}}, \quad \forall t \in (t_n, t_{n+1}]\}. \end{aligned}$$

For a given $(u_K^{n+1})_{n\in\{0,...,N\},K\in\mathcal{V}} \in \mathbb{R}^{(N+1)\#\mathcal{V}}$, we denote by $u_{\mathcal{T},\Delta t}$ and $u_{\mathcal{M},\Delta t}$ the unique elements of $V_{\mathcal{T},\Delta t}$ and $X_{\mathcal{M},\Delta t}$ respectively such that

$$u_{\mathcal{T},\Delta t}(x_K, t) = u_{\mathcal{M},\Delta t}(x_K, t) = u_K^{n+1}, \quad \forall K \in \mathcal{V},\ \forall t \in (t_n, t_{n+1}]. \tag{2.15}$$

2.2.2 Finite elements

The method we propose, and more generally the CVFE method, is based on P_1-finite elements. We introduce in this section the technical material that is needed in order to define the scheme and to perform its analysis. We define the transmissibility coefficients

$$a_{KL}^T = -\int_T \Lambda \nabla e_K \cdot \nabla e_L \, \mathrm{d}\boldsymbol{x} = a_{LK}^T, \quad \forall T \in \mathcal{T},\ \forall (K,L) \in \mathcal{V}^2, \tag{2.16}$$

and

$$a_{KL} = a_{LK} = -\int_\Omega \Lambda \nabla e_K \cdot \nabla e_L \, \mathrm{d}\boldsymbol{x} = \sum_{T\in\mathcal{T}} a_{KL}^T, \quad \forall (K,L) \in \mathcal{V}^2. \tag{2.17}$$

Note that $a_{KL} = 0$ unless $\sigma_{KL} \in \mathcal{E}$. Moreover, since $\sum\limits_{K\in\mathcal{V}} \nabla e_K = 0$, we have that :

$$-a_{KK} = \sum_{L\neq K} a_{KL} > 0. \tag{2.18}$$

As a consequence of (2.17)-(2.18), given $u_{\mathcal{T}}$ and $v_{\mathcal{T}}$ two elements of $V_{\mathcal{T}}$, one has

$$\begin{aligned} \int_\Omega \Lambda \nabla u_{\mathcal{T}} \cdot \nabla v_{\mathcal{T}} \, \mathrm{d}\mathbf{x} &= \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(u_K - u_L)(v_K - v_L) \\ &= \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T (u_K - u_L)(v_K - v_L). \end{aligned} \tag{2.19}$$

The following lemma plays a crucial role in the numerical analysis carried out in this chapter.

Lemma 2.2.1. *There exists C_3 depending only on $\theta_{\mathcal{T}}$, $\Lambda_\star$ and $\Lambda^\star$ such that, for all $u_{\mathcal{T}} \in V_{\mathcal{T}}$, one has*

$$\sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|(u_K - u_L)^2 \leq \sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(u_K - u_L)^2 \leq C_3 \int_\Omega \Lambda \nabla u_{\mathcal{T}} \cdot \nabla u_{\mathcal{T}} \,\mathrm{d}\boldsymbol{x}.$$

Proof. We reproduce the proof of [39, Lemma 3.2] .

In order to prove Lemma 2.2.1, it only remains to prove that

$$\sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(u_K - u_L)^2 \leq C\|\nabla u_{\mathcal{T}}\|^2_{L^2(\Omega)^2},$$

since one has using assumption (**A3**) on Λ, that

$$\|\nabla u_{\mathcal{T}}\|^2_{L^2(\Omega)^2} = \int_\Omega \Lambda \nabla u_{\mathcal{T}} \cdot \nabla u_{\mathcal{T}} \,\mathrm{d}\boldsymbol{x} \leq \frac{1}{\Lambda_\star} \int_\Omega \Lambda \nabla u_{\mathcal{T}} \cdot \nabla u_{\mathcal{T}} \,\mathrm{d}\boldsymbol{x}.$$

Thanks to Cauchy-Schwarz inequality, we have

$$|a_{KL}^T| \leq \Lambda^\star \|\nabla e_K\|_{L^2(T)^2} \|\nabla e_L\|_{L^2(T)^2}.$$

Using a classical inequality stemming from finite element properties (see [31, 71]), one has

$$\|\nabla e_K\|^2_{L^2(T)^2} \leq c\theta_{\mathcal{T}} \frac{|T|}{(h_{\mathcal{T}})^2}, \quad \forall K \in \mathcal{V},\ \forall T \in \mathcal{T},$$

where c is an absolute constant, so that

$$|a_{KL}^T| \leq c\theta_{\mathcal{T}}\Lambda^\star \frac{|T|}{(h_{\mathcal{T}})^2}, \quad \forall K \in \mathcal{V},\ \forall T \in \mathcal{T}.$$

This implies that

$$\sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(u_K - u_L)^2 \leq C \frac{|T|}{(h_T)^2} \sum_{\sigma_{KL}\in\mathcal{E}_T} (u_K - u_L)^2, \quad \text{with } C = c\theta_{\mathcal{T}}\Lambda^\star.$$

Finally, it follows from the analysis carried out for example in [71] that for all $T \in \mathcal{T}$, with K, L, M being its vertices

$$\frac{|T|}{(h_T)^2}\Big((u_K - u_L)^2 + (u_K - u_M)^2 + (u_L - u_M)^2\Big) \leq C\|\nabla u_{\mathcal{T}}\|^2_{L^2(\Omega)^2}.$$

Since $\sigma_{KL} \in \mathcal{E}_T$ is shared by at most two triangles, we have

$$\sum_{T\in\mathcal{T}} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(u_K - u_L)^2 \leq C \sum_{T\in\mathcal{T}} \|\nabla u_{\mathcal{T}}\|^2_{L^2(T)^2} = C\|\nabla u_{\mathcal{T}}\|^2_{L^2(\Omega)^2},$$

this concludes the proof of the Lemma. □

2.2.3 The nonlinear CVFE scheme

In this section, we explicit the discretization of the problem (2.1) we will study in this chapter. The time discretization relies on backward Euler scheme, while the space discretization relies on finite elements with mass lumping and a suitable upwinding of the mobility.

The discretization $s^0_{\mathcal{M}} \in X_{\mathcal{M}}$ of the initial data is defined by

$$s^0_K = \frac{1}{m_K} \int_{\omega_K} s_0(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \qquad \forall K \in \mathcal{V}. \tag{2.20}$$

In the sequel, we will make use of the shortened notation

$$z_K = z(\boldsymbol{x}_K), \qquad \forall K \in \mathcal{V}.$$

Let us now introduce the scheme. For all $n \in \{0, ..., N\}$, a solution $\left(p_K^{n+1}\right)_{K \in \mathcal{V}}$ to the scheme at the time step $n+1$ has to satisfy the following equations : for all $K \in \mathcal{V}$,

$$\frac{s(p_K^{n+1}) - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} \eta_{KL}^{n+1} a_{KL} (u_K^{n+1} - u_L^{n+1}) = 0, \tag{2.21a}$$

$$u_K^{n+1} = p_K^{n+1} + \rho g z_K, \tag{2.21b}$$

$$s_K^{n+1} = s(p_K^{n+1}), \tag{2.21c}$$

$$\eta_{KL}^{n+1} = \begin{cases} \eta(s_K^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) \geq 0, \\ \eta(s_L^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1} - u_L^{n+1}) < 0. \end{cases} \tag{2.21d}$$

Remark 2.2.2. *It follows from the monotonicity of the mobility and water content functions η and s that* (2.21d) *is equivalent to*

$$\eta_{KL}^{n+1} = \begin{cases} \max\limits_{p \in I_{KL}^{n+1}} \eta(s(p)) & \text{if} \quad a_{KL}(p_K^{n+1} - p_L^{n+1})(u_K^{n+1} - u_L^{n+1}) \geq 0, \\ \min\limits_{p \in I_{KL}^{n+1}} \eta(s(p)) & \text{if} \quad a_{KL}(p_K^{n+1} - p_L^{n+1})(u_K^{n+1} - u_L^{n+1}) \leq 0, \end{cases} \tag{2.22}$$

where

$$I_{KL}^{n+1} = [\min(p_K^{n+1}, p_L^{n+1}), \max(p_K^{n+1}, p_L^{n+1})].$$

It is then worth noticing that the monotonicity assumption on η can be bypassed if one enforces (2.22) *directly instead of* (2.21d) *for the definition of the upwind mobility.*

This scheme, whose construction is based on finite elements *via* (2.17), can be interpreted as a finite volume scheme. Indeed denoting by

$$F_{KL}^{n+1} = a_{KL} \eta_{KL}^{n+1} (u_K^{n+1} - u_L^{n+1}),$$

the scheme (2.21) can be rewritten under the locally conservative form on the dual cells ω_K :

$$\begin{cases} F_{KL}^{n+1} + F_{LK}^{n+1} = 0, & \text{for all } \sigma_{KL} \in \mathcal{E}_K \\ \dfrac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \displaystyle\sum_{\sigma_{KL} \in \mathcal{E}_K} F_{KL}^{n+1} = 0, & \text{for all } K \in \mathcal{V}. \end{cases} \tag{2.23}$$

As a straightforward consequence, we can claim that the scheme (2.21) is globally conservative, i.e.,

$$\sum_{K \in \mathcal{V}} m_K s_K^{n+1} = \sum_{K \in \mathcal{V}} m_K s_K^n = \int_\Omega s_0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad \forall n \geq 0. \tag{2.24}$$

Remark 2.2.3. *It will appear in the analysis that the discrete pressures p_K^{n+1} are always bounded (see Lemmas 2.3.10 and 2.3.11). Therefore, all the terms appearing in the scheme are finite, hence the products and sums are well defined.*

2.2.4 Main results

The scheme (2.21) amounts to a nonlinear system to be solved at each time step. The existence of a solution to this system is therefore non trivial. The first result we highlight is thus the existence of a solution to the scheme (2.21) and the stability in terms of the discrete capillary energy.

Theorem 2.2.4. *There exists (at least) one solution $\left(p_K^{n+1}\right)_{K \in \mathcal{V}, n \in \{0,\dots,N\}}$ to the scheme* (2.21a)*. Moreover, $0 \leq s_K^n \leq 1$ for all $K \in \mathcal{V}$ and for all $n \in \{0, \dots, M\}$, and there exists C depending only on $\theta_{\mathcal{T}}$, Λ, Ω, t_f, $\|s\|_{L^1(p_\star,0)}$, and $\|\eta\|_\infty$ such that*

$$\sup_{n \in \{0,\dots,N\}} \sum_{K \in \mathcal{V}} m_K \Gamma(p_K^{n+1}) + \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 \leq C.$$

Once we have the discrete solution $\left(p_K^{n+1}\right)_{K \in \mathcal{V}, n \in \{0,\dots,N\}}$ at hand for all meshes and all time discretizations, then we can study the convergence of the scheme when the discretization parameters tend to 0. More precisely, consider a sequence $(\mathcal{T}_m)_{m \geq 1}$ of triangulations of Ω such that

$$h_m = \max_{T \in \mathcal{T}_m} \operatorname{diam}(T) \underset{m \to \infty}{\longrightarrow} 0, \tag{2.25}$$

and such that there exists $\theta^\star > 0$ such that

$$\theta_{\mathcal{T}_m} \leq \theta^\star, \quad \forall m \geq 1. \tag{2.26}$$

A sequence of dual meshes $(\mathcal{M}_m)_{m \geq 1}$ corresponding to the triangular meshes $(\mathcal{T}_m)_{m \geq 1}$ is built as in §2.2.1. Let $(N_m)_{m \geq 1}$ be an increasing sequence of integers, then we define the corresponding sequence of time steps $\Delta t_m = \frac{t_f}{N_m + 1}$ tending to 0 as m tends

to ∞. To this sequence of discretizations of Q_{t_f} corresponds a sequence of solutions $\left(p_K^{n+1}\right)_{K\in\mathcal{V}_m,n\in\{0,\dots,N_m\}}$ to the scheme. Thanks to these solutions, we can construct the functions $s_{\mathcal{M}_m,\Delta t_m} \in X_{\mathcal{M}_m,\Delta t_m}$ and $\xi_{\mathcal{T}_m,\Delta t_m} \in V_{\mathcal{T}_m,\Delta t_m}$ defined by

$$s_{\mathcal{M}_m,\Delta t_m}(\boldsymbol{x}_K, t_{n+1}) = s(p_K^{n+1}) = s_K^{n+1}, \qquad \forall K \in \mathcal{V}_m,\ \forall n \in \{0,\dots,N_m\}, \tag{2.27}$$

and

$$\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_K, t_{n+1}) = \xi(p_K^{n+1}) = \xi_K^{n+1}, \qquad \forall K \in \mathcal{V}_m,\ \forall n \in \{0,\dots,N_m\}. \tag{2.28}$$

Once these sequences of discrete functions at hand, we can state the second main result of this chapter, namely the convergence of the scheme (2.21).

Theorem 2.2.5. *Let $(\mathcal{T}_m)_{m\geq 1}$ be a sequence conformal triangular discretization of Ω such that* (2.25) *and* (2.26) *hold. Let $(s_{\mathcal{M}_m,\Delta t_m})_m$ and $(\xi_{\mathcal{T}_m,\Delta t_m})_m$ be the functions reconstructed from the solutions $\left(\left(p_K^{n+1}\right)_{K,n}\right)_m$ to the scheme* (2.21) *thanks to formulas* (2.27)–(2.28)*. Then*

$$s_{\mathcal{M}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} s(p) \quad \text{a.e in } Q_{t_f},$$
$$\xi_{\mathcal{T}_m,\Delta t_m} \underset{m\to+\infty}{\longrightarrow} \xi(p) \quad \text{weakly in } L^2((0,t_f);H^1(\Omega)) \text{ and strongly in } L^2(Q_{t_f}),$$

where p is the unique solution to the continuous problem (2.1)*.*

The proof of Theorem 2.2.4 is addressed in §2.3. The convergence of the scheme towards a weak solution is the purpose of §3.4, while the uniqueness of the weak solution is proved in appendix, cf. Proposition 2.7.4. Numerical illustrations are provided in §3.5.

2.3 Discrete properties, *a priori* estimates and existence

In this section, we establish *a priori* estimates, among which the positivity of the saturation and the stability of the capillary energy. These estimates allow to prove the existence of a solution to the nonlinear system (2.21). They are also keystones in order to perform the convergence analysis later on.

2.3.1 A uniform L^∞-estimate on $s_{\mathcal{M},\Delta t}$

In what follows, $(p_K^{n+1})_{K\in\mathcal{V},n\geq 0}$ denotes a solution to the scheme (2.21) (whose existence will be established later). This allows to define the quantities $s_K^{n+1} = s(p_K^{n+1})$ and $\xi_K^{n+1} = \xi(p_K^{n+1})$ for all $K \in \mathcal{V}$ and all $n \in \{0,\dots,N\}$.

Proposition 2.3.1. *For all $K \in \mathcal{V}$, and all $n \in \{0, ..., N\}$, one has*

$$0 \leq s_K^n \leq 1. \tag{2.29}$$

Equivalently, one has

$$p_\star \leq p_K^{n+1}, \qquad \forall K \in \mathcal{V},\ \forall n \in \{0, \dots, N\}. \tag{2.30}$$

Proof. First of all, note that there is nothing to prove if $p_\star = -\infty$. Therefore, we restrict our attention to the case of a finite $p_\star$. The property (2.29) holds for $n = 0$ thanks to the discretization (2.20) of the initial data. Assume now (2.29) holds at time step n. It is equivalent to prove $p_K^{n+1} \geq p_\star$. Assume that

$$p_{K_m}^{n+1} = \min_{L \in \mathcal{V}} p_L^{n+1} < p_\star \quad \Leftrightarrow \quad s_{K_m}^{n+1} < 0. \tag{2.31}$$

In view of the definition (2.22) of $\eta_{K_mL}^{n+1}$, and of the fact that $\eta(s) = 0$ if $s < 0$, it follows from (2.21d) that

$$\eta_{K_mL}^{n+1} = 0 \quad \text{if} \quad a_{K_mL}(u_{K_m}^{n+1} - u_L^{n+1}) \geq 0.$$

Therefore, the scheme (2.21) at vertex K_m rewrites

$$s_{K_m}^{n+1} = s_{K_m}^n - \frac{\Delta t}{m_{K_m}} \sum_{\sigma_{K_mL} \in \mathcal{E}} \eta_{K_mL}^{n+1} a_{K_mL}(u_{K_m}^{n+1} - u_L^{n+1}) \geq 0.$$

This yields a contradiction with (2.31). Hence, the L^∞ estimate (2.29) holds at the time step $n + 1$, thus for all n. □

2.3.2 Capillary energy estimate and the control of the dissipation

The goal of this section is to get an *a priori* control for the capillary energy of the discrete solution and to derive some estimates coming from the dissipation of the energy. We were not able to derive the discrete counterpart of the energy/dissipation estimate (2.11). However, we can prove a discrete counterpart of (2.12) (cf. Proposition 2.3.2) that appears to be sufficient to establish Theorems 2.2.4 and 2.2.5. In what follows, we assume that $(s_K^n)_{K \in \mathcal{V}}$ is known and $\left(p_K^{n+1}\right)_{K \in \mathcal{V}}$ denotes an arbitrary solution to the scheme (2.21).

Proposition 2.3.2. *There exists C_4 depending only on $\theta_\mathcal{T}$, Λ, Ω, t_f, $\|s\|_{L^1(p_\star,0)}$, and $\|\eta\|_\infty$ such that*

$$\sup_{n \in \{0,\dots,N\}} \sum_{K \in \mathcal{V}} m_K \Gamma(p_K^{n+1}) + \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} a_{KL} \left(\xi(p_K^{n+1}) - \xi(p_L^{n+1})\right)^2 \leq C_4.$$

The proof of Proposition 2.3.2 is based on several Lemmas stated below. This section also contains technical lemmas that will be useful in the convergence proof of §3.4.

Lemma 2.3.3. *There exists C_5 depending only on Ω, s such that, for all $\nu \in \{0,\dots,N\}$, one has*

$$\sum_{K\in\mathcal{V}} m_K\Gamma(p_K^{\nu+1}) + \sum_{n=0}^{\nu}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1})(p_K^{n+1}-p_L^{n+1}) \le C_5. \quad (2.32)$$

Proof. We multiply the scheme (2.21a) by $\Delta t p_K^{n+1}$ and sum on $K \in \mathcal{V}$. This yields :

$$A + B = 0,$$

where

$$A = \sum_{K\in\mathcal{V}} m_K(s_K^{n+1}-s_K^n)p_K^{n+1}, \quad B = \Delta t \sum_{K\in\mathcal{V}}\sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1})p_K^{n+1}.$$

Since $a_{KL} = a_{LK}$ and $\eta_{KL}^{n+1} = \eta_{LK}^{n+1}$, we can rewrite

$$B = \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1})(p_K^{n+1}-p_L^{n+1}).$$

By convexity of $\Gamma\circ s^{-1}$ one deduces this estimation

$$A \ge \sum_{K\in V} m_K(\Gamma(p_K^{n+1}) - \Gamma\circ s^{-1}(s_K^n)).$$

Summing over $n \in \{0,...,\nu\}$ provides

$$\sum_{K\in\mathcal{V}} m_K\Gamma(p_K^{\nu+1}) + \sum_{n=0}^{\nu}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1}-u_L^{n+1})(p_K^{n+1}-p_L^{n+1}) \le \sum_{K\in\mathcal{V}} m_K\Gamma\circ s^{-1}(s_K^0). \quad (2.33)$$

It remains to check that for $b \in [0,1]$,

$$0 \le \Gamma\circ s^{-1}(b) = \int_0^{s^{-1}(b)} as'(a)\,\mathrm{d}a = \int_1^b s^{-1}(a)\,\mathrm{d}a \le \|s^{-1}\|_{L^1(0,1)} < +\infty,$$

ensuring that

$$\sum_{K\in\mathcal{V}} m_K\Gamma\circ s^{-1}(s_K^0) \le \int_\Omega \Gamma\circ s^{-1}(s_0)\,\mathrm{d}\boldsymbol{x} \le |\Omega|\|s^{-1}\|_{L^1(0,1)}$$

thanks to Jensen's inequality and to (2.5). □

From the previous lemma, we can get an estimate on the spatial variations of the function $\xi_{\mathcal{T},\Delta t}$. In order to ease the reading, we use the shortened notation

$$\xi_K^{n+1} = \xi(p_K^{n+1}), \qquad \forall K \in \mathcal{V},\ \forall n \in \{0,\dots,N\},$$

and we define

$$\tilde{\eta}_{KL}^{n+1} = \begin{cases} \left(\dfrac{\xi_K^{n+1} - \xi_L^{n+1}}{p_K^{n+1} - p_L^{n+1}}\right)^2 & \text{if} \quad p_K^{n+1} \neq p_L^{n+1}, \\ \eta(s_K^{n+1}) & \text{if} \quad p_K^{n+1} = p_L^{n+1}. \end{cases} \tag{2.34}$$

Lemma 2.3.4. *There exists C_6 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\iint_{Q_{t_f}} \Lambda \nabla \xi_{\mathcal{T},\Delta t} \cdot \nabla \xi_{\mathcal{T},\Delta t} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(\xi_K^{n+1} - \xi_L^{n+1})^2 \leq C_6. \tag{2.35}$$

Proof. The definition (2.22) of the mobilities η_{KL}^{n+1} has been chosen so that

$$\begin{aligned} C_5 &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) \\ &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\check{\eta}_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}), \end{aligned}$$

where $\check{\eta}_{KL}^{n+1} = \eta(s(p_{KL}))$ whatever $p_{KL} \in I_{KL}^{n+1}$. Therefore, using the definition (2.21b) of u_K^{n+1} and Young's inequality leads to

$$\begin{aligned} C_5 &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\check{\eta}_{KL}^{n+1}\left((p_K^{n+1} - p_L^{n+1})^2 + (p_K^{n+1} - p_L^{n+1})(z_K - z_L)\right) \\ &\geq \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\check{\eta}_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 \\ &\quad - \frac{\alpha}{2}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\check{\eta}_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 - \frac{\|\eta\|_\infty}{2\alpha}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|(z_K - z_L)^2 \end{aligned}$$

where α is a positive parameter to be fixed. We choose $\check{\eta}_{KL}^{n+1} = \tilde{\eta}_{KL}^{n+1}$ defined in (2.34), leading to

$$\begin{aligned} &\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 - \frac{\alpha}{2}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \\ &\qquad \leq C_5 + \frac{\|\eta\|_\infty}{2\alpha}\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\left(z_K - z_L\right)^2. \end{aligned}$$

Using Lemma 2.2.1, we get that

$$\left(1 - \frac{\alpha C_3}{2}\right)\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \leq C_5 + \frac{\|\eta\|_\infty}{2\alpha} t_f C_3 |\Omega|.$$

We conclude the proof by setting $\alpha = \frac{1}{C_3}$. □

The function Γ takes non-negative values, hence so does the first term in (2.32). But since a_{KL} may become negative, we are not able to claim that the second term is non-negative (this would end the proof of Proposition 2.3.2). Nevertheless, we can prove that this term is uniformly bounded. This information, combined with Lemma 2.3.4, is sufficient to conclude the proof of Proposition 2.3.2.

Lemma 2.3.5. *There exists C_7 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}|u_K^{n+1} - u_L^{n+1}||p_K^{n+1} - p_L^{n+1}| \le C_7.$$

Proof. Since $|x| = x + 2x^-,\ x^- = \max(-x, 0)$, one has

$$\begin{aligned}
&\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}|u_K^{n+1} - u_L^{n+1}||p_K^{n+1} - p_L^{n+1}| \\
&\qquad = \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1}) \\
&\qquad\quad + 2\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \eta_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^-. \quad (2.36)
\end{aligned}$$

It follows from the definition (2.22) of η_{KL}^{n+1} that

$$\eta_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^- \le \tilde{\eta}_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^-,$$

with $\tilde{\eta}_{KL}^{n+1}$ defined by (2.34). Moreover, using the definition (2.21b) of u_K^{n+1} together with Young's inequality, we obtain that

$$\begin{aligned}
&\tilde{\eta}_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^- \\
&\le \tilde{\eta}_{KL}^{n+1}|a_{KL}|(p_K^{n+1} - p_L^{n+1})^2 + \tilde{\eta}_{KL}^{n+1}|a_{KL}||z_K - z_L||p_K^{n+1} - p_L^{n+1}| \\
&\le |a_{KL}|(\xi_K^{n+1} - \xi_L^{n+1})^2 + \frac{1}{2}|a_{KL}|\tilde{\eta}_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 + \frac{1}{2}|a_{KL}|\tilde{\eta}_{KL}^{n+1}(z_K - z_L)^2 \\
&\le \frac{3}{2}|a_{KL}|(\xi_K^{n+1} - \xi_L^{n+1})^2 + \frac{\|\eta\|_\infty}{2}|a_{KL}|(z_K - z_L)^2.
\end{aligned}$$

We deduce from Lemma 2.2.1 that

$$\begin{aligned}
&2\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} \eta_{KL}^{n+1}[a_{KL}(u_K^{n+1} - u_L^{n+1})(p_K^{n+1} - p_L^{n+1})]^- \\
&\qquad\qquad \le 3C_3 \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}(\xi_K^{n+1} - \xi_L^{n+1})^2 + C_3\|\eta\|_\infty t_f|\Omega|. \quad (2.37)
\end{aligned}$$

Then we combine (2.36), (2.37), Lemma 2.3.4, and Lemma 2.3.3 to conclude. □

The *a priori* estimate of Proposition 2.3.2 follows easily from Lemmas 2.3.3, 2.3.4, and 2.3.5. It is sufficient to prove the existence of a solution to the scheme (2.21) (see §2.3.3). Nevertheless, before going to this existence proof, we still provide some additional *a priori* estimates to be used later on in §3.4.

Lemma 2.3.6. *There exists C_8 depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}$, and η such that*

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 \le C_8, \tag{2.38}$$

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})^2 \le C_8. \tag{2.39}$$

Proof. The definition (2.21b) of u_K^{n+1} yields

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma\in\mathcal{E}_{KL}} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 = A + B,$$

where

$$A = \sum_{n=0}^{N} \Delta t \sum_{\sigma\in\mathcal{E}_{KL}} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})(u_K^{n+1} - u_L^{n+1}),$$

$$B = -\sum_{n=0}^{N} \Delta t \sum_{\sigma\in\mathcal{E}_{KL}} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})(z_K^{n+1} - z_L^{n+1}).$$

Thanks to Lemma 2.3.5, one has $A \le C_7$. Moreover, combining once again Young inequality with Lemma 2.2.1, we get that

$$B \le \frac{1}{2}\sum_{n=0}^{N} \Delta t \sum_{\sigma\in\mathcal{E}_{KL}} |a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1} - p_L^{n+1})^2 + C_3\|\eta\|_\infty t_f|\Omega|,$$

hence (2.38) holds with $C_8 = 2C_7 + 2C_3\|\eta\|_\infty t_f|\Omega|$. The proof of (2.39) is similar. □

The last lemma of this section is devoted to the control of the L^2 norm of $\xi_{\mathcal{T},\Delta t}$. Lemma 2.3.4 only provides a control on the gradient of $\xi_{\mathcal{T},\Delta t}$, but not on $\xi_{\mathcal{T},\Delta t}$ directly. The control on $\xi_{\mathcal{T},\Delta t}$ is provided by an argument *à la* Poincaré, cf. Appendix 2.7.1.

Lemma 2.3.7. *There exists C_9 depending only on $\Omega, t_f, s, \Lambda, \theta_{\mathcal{T}}, \eta, \overline{s}_0$, and $\xi_\star$ such that*

$$\|\xi_{\mathcal{T},\Delta t}\|_{L^2(Q_{t_f})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \le C_9, \tag{2.40}$$

$$\|\xi_{\mathcal{M},\Delta t}\|_{L^2(Q_{t_f})}\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \le C_9. \tag{2.41}$$

Proof. Let us first establish (2.41). Thanks to Assumption (2.4), we know that $\int_\Omega s_0 \,\mathrm{d}\boldsymbol{x} < \mathrm{meas}(\Omega)$. The global conservativity property (2.24) allows to claim that

$$\sum_{K\in\mathcal{V}} s_K^{n+1} m_K = \overline{s}_0 = \int_\Omega s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} < \mathrm{meas}(\Omega)$$

for any $n \in \{0,\dots,N\}$. Using that $\xi_K^{n+1} < 0$ if and only if $s_K^{n+1} < 1$ (recall that $\xi(p) < 0$ iff $p < 0$ iff $s < 1$), one gets

$$\mathrm{meas}\left\{\xi_{\mathcal{M},\Delta t}(\cdot, t_{n+1}) < 0\right\} \geq \mathrm{meas}(\Omega) - \overline{s}_0 > 0. \tag{2.42}$$

Denote by $\xi_K^{+,n+1} = \max(0, \xi_K^{n+1})$, and by $\xi^+_{\mathcal{M},\Delta t}$ and $\xi^+_{\mathcal{T},\Delta t}$ the unique elements of $X_{\mathcal{M},\Delta t}$ and $V_{\mathcal{T},\Delta t}$ respectively such that

$$\xi^+_{\mathcal{M},\Delta t}(\boldsymbol{x}_K, t_{n+1}) = \xi^+_{\mathcal{T},\Delta t}(\boldsymbol{x}_K, t_{n+1}) = \xi_K^{+,n+1}, \qquad \forall K \in \mathcal{V},\ \forall n \in \{0,\dots,N\}.$$

Note that $\xi^+_{\mathcal{T},\Delta t} \neq (\xi_{\mathcal{T},\Delta t})^+ = \max\left(0, \xi_{\mathcal{T},\Delta t}\right)$ in general, but that $\xi^+_{\mathcal{M},\Delta t} = (\xi_{\mathcal{M},\Delta t})^+$ and that $\xi^-_{\mathcal{M},\Delta t} = (\xi_{\mathcal{M},\Delta t})^- = \max\left(0, -\xi_{\mathcal{M},\Delta t}\right)$. Using Assumption (**A3**), the 1-Lipschitz continuity of $x \mapsto x^+$, and Lemmas 2.2.1 and 2.3.4, we obtain

$$\begin{aligned}
\iint_{Q_{t_f}} |\nabla \xi^+_{\mathcal{T},\Delta t}|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t &\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL} \left(\xi_K^{+,n+1} - \xi_L^{+,n+1}\right)^2 \\
&\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}| \left(\xi_K^{+,n+1} - \xi_L^{+,n+1}\right)^2 \\
&\leq \frac{1}{\underline{\Lambda}} \sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}| \left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2 \leq \frac{C_3 C_6}{\underline{\Lambda}}.
\end{aligned}$$

Therefore, we can apply Lemma 2.7.3 stated in appendix. This provides

$$\iint_{Q_{t_f}} \left(\xi^+_{\mathcal{M},\Delta t}\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C. \tag{2.43}$$

On the other hand, because of (2.3), we know that $\xi^-_{\mathcal{M},\Delta t} \leq \xi_\star$ a.e. in Q_{t_f}, hence

$$\iint_{Q_{t_f}} \left(\xi^-_{\mathcal{M},\Delta t}\right)^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq (\xi_\star)^2 \,\mathrm{meas}(\Omega) t_f. \tag{2.44}$$

Combining (2.43) with (2.44) provides (2.41). In order to recover (2.40), in only remains to use (2.14) and (2.41). □

2.3.3 Existence of a discrete solution

In order to prove the existence of a solution $(p_K^{n+1})_K$ to the scheme (2.21), we need an additional mesh-depending estimate on the solution. Following [39], we introduce now the notion of *transmissive path*.

Definition 2.3.8. *A transmissive path w joining $K_i \in \mathcal{V}$ to $K_f \in \mathcal{V}$ consists in a list of vertices $(K_q)_{0\le q\le M}$ such that $K_i = K_0, K_f = K_M$, with $K_q \neq K_\ell$ if $q \neq \ell$, and such that $\sigma_{K_qK_{q+1}} \in \mathcal{E}$ with $a_{K_qK_{q+1}} > 0$ for all $q \in \{0, ..., M-1\}$. We denote by $\mathcal{W}(K_i, K_f)$ the set of the transmissive path joining $K_i \in \mathcal{V}$ to $K_f \in \mathcal{V}$.*

We now state a result which is proved in [39, Lemma 3.5].

Lemma 2.3.9. *For all $(K_i, K_f) \in \mathcal{V}^2$ there exists a transmissive path $w \in \mathcal{W}(K_i, K_f)$.*

Lemma 2.3.10. *There exists $C_\star > -\infty$ depending only on $\mathcal{T}, \Delta t, \Omega, s, \overline{s}_0, t_f, \Lambda$, $\theta_\mathcal{T}$, η and z such that*

$$p_K^{n+1} \ge C_\star, \qquad \forall K \in \mathcal{V}, \quad \forall n \in \{0, ..., N\}.$$

Proof. Let us prove that $p_K^{n+1} \ge C_\star$. Assume first that $p_\star > -\infty$, then we can choose $C_\star = p_\star$ thanks to (2.30), so that we can focus on the case $p_\star = -\infty$.

In view of the global conservation property (2.24), one has that

$$\sum_{K\in\mathcal{V}} (s_K^{n+1} - \overline{s}_0)m_K = 0.$$

This ensures the existence of at least one vertex K_i such that $s_{K_i}^{n+1} \ge \overline{s}_0 > 0$. In particular,

$$-\infty < s^{-1}(\overline{s}_0) \le p_{K_i}^{n+1}. \tag{2.45}$$

Let $K_f \in \mathcal{V}\setminus\{K_i\}$, then thanks to Lemma 2.3.9, there exists a transmissive path $w \in \mathcal{W}(K_i, K_f) = (K_q)_{0\le q\le M}$ of finite length in the sense of Definition 2.3.8. Let us show that for all $p_{K_q}^{n+1} > -\infty$ for all $q \in \{0, ..., M\}$.

First, we have checked in (2.45) that $p_{K_0}^{n+1} > -\infty$. Assume now that $p_{K_q}^{n+1} > -\infty$ for some $q \in \{0, \dots, M-1\}$, then it follows from Lemma 2.3.5 that

$$\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL}\in\mathcal{E}} |a_{KL}|\eta_{KL}^{n+1}|u_K^{n+1} - u_L^{n+1}||p_K^{n+1} - p_L^{n+1}| \le C_6.$$

This ensures in particular that

$$\Delta t a_{K_qK_{q+1}}\eta_{K_qK_{q+1}}^{n+1}(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \le C_6.$$

Thanks to the definition (2.22) of $\eta_{K_qK_{q+1}}^{n+1}$, one has

$$a_{K_qK_{q+1}}\eta_{K_qK_{q+1}}^{n+1}(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \\ \ge a_{K_qK_{q+1}}\eta(s_{K_q}^{n+1})(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}).$$

Since $a_{K_qK_{q+1}} > 0$, we obtain that

$$(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \\ = (p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})^2 + (p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})(z_{K_q} - z_{K_{q+1}}) \ \leq \frac{C_6}{\Delta t a_{K_qK_{q+1}} \eta(s(p_{K_q}^{n+1}))}.$$

Using Young inequality one has

$$(u_{K_q}^{n+1} - u_{K_{q+1}}^{n+1})(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1}) \geq \frac{1}{2}(p_{K_q}^{n+1} - p_{K_{q+1}}^{n+1})^2 - \frac{1}{2}(z_{K_q} - z_{K_{q+1}})^2,$$

thus

$$p_{K_{q+1}}^{n+1} \geq p_{K_q}^{n+1} - \sqrt{(z_{K_q} - z_{K_{q+1}})^2 + \frac{2C_6}{\Delta t a_{K_qK_{q+1}} \eta(s(p_{K_q}^{n+1}))}}.$$

This ensures that $p_{K_{q+1}}^{n+1} > -\infty$.

We have proved the existence of a finite quantity $(C_{K_i,K_f,w})_{K_f \in \mathcal{V}}$ (depending on the data of the continuous problem $\Omega, s, \overline{s}_0, t_f, \Lambda, \theta_{\mathcal{T}}, \eta$ but also on the discretization $\mathcal{T}$ and on Δt) such that

$$s(p_{K_i}^{n+1}) \geq \overline{s}_0 \quad \Longrightarrow \quad p_{K_f}^{n+1} \geq -C_{K_i,K_f,w}.$$

As a consequence, since there exists a finite number of transmissive paths between two vertices, we get the estimate

$$p_K^{n+1} \geq -\max_{K_i \in \mathcal{V}} \max_{K_f \in \mathcal{V}} \min_{w \in \mathcal{W}(K_i,K_f)} C_{K_i,K_f,w} > -\infty, \quad \forall K \in \mathcal{V}, \ \forall n \in \{0, ..., N\}.$$

□

In the previous lemma, we managed to bound the $\{p_K^{n+1}\}$ from below. The next lemma provides a bound from above.

Lemma 2.3.11. *There exists $p^\star < \infty$ depending only on $\mathcal{T}, \Delta t, \Omega, t_f, s, \Lambda, \eta, \overline{s}_0$ and $\xi_\star$ such that*

$$p_K^{n+1} \leq p^\star \quad \forall K \in \mathcal{V}, \ \ \forall n \in \{0, ..., N\}.$$

Proof. Since $s(p) = 1$ if $p \geq 0$, one has $\xi(p) = p\sqrt{\eta(1)}$ if $p \geq 0$. By (2.41), one has

$$\Delta t m_K \xi(p_K^{n+1})^2 \leq \|\xi_{\mathcal{M},\Delta t}\|^2_{L^2(Q_{t_f})} \leq (C_9)^2.$$

Therefore, we get $p_K^{n+1} \leq \dfrac{C_9}{\sqrt{\Delta t m_K}} \dfrac{1}{\eta(1)}$. □

Now, one can apply the same strategy as in [39, Lemma 3.11] for proving the existence of a solution to the scheme (2.21).

Proposition 2.3.12. *Let $(s_K^n)_{K\in\mathcal{V}} \in [0,1]^{\#\mathcal{V}}$ be such that $\sum_{K\in\mathcal{V}} m_K s_K^n = \mathrm{meas}(\Omega)\overline{s}_0$, there exists (at least) one solution $(p_K^{n+1})_{K\in\mathcal{V}} \in [p_\star, p^\star]^{\#\mathcal{V}}$ of the scheme (2.21). Moreover, it satisfies $\sum_{K\in\mathcal{V}} m_K s_K^{n+1} = \mathrm{meas}(\Omega)\overline{s}_0$.*

The proof of Proposition 2.3.12 is not detailed here since it mimics the one of [39, Lemma 3.11]. Let us just mention that it is based on a topological degree argument [61, 122].

2.4 Convergence towards a weak solution

The proof of the convergence properties stated in Theorem 2.2.5 is based on compactness arguments. As a first step, we show in §2.4.1 the appropriate compactness properties on the reconstructed discrete solutions. Then we identify in §3.4.2 the limit value (whose existence is ensured thanks to the compactness properties) as the unique weak solution to the problem (2.1).

2.4.1 Compactness properties of discrete solutions

As it is classical for unsteady problems, we need to prove some time-compactness for the approximate solutions. Because of the degeneracy of the problem we consider, we cannot use a strategy *à la* Aubin-Simon [142] (see [98] for an extension of this strategy to the discrete setting). A classical way to circumvent this problem is to estimate the time-translates (see [11] in the continuous setting and [78] in the discrete setting). This strategy could have been used here, but we rather make use of the time-compactness result for degenerate parabolic equations proposed in [14]. To this end, we need the following lemma.

Lemma 2.4.1. *There exists C_{10} depending only on $\Omega, s, t_f, \Lambda, \theta_{\mathcal{T}}, z$ and η such that*

$$\sum_{n=0}^{N}\sum_{K\in\mathcal{V}} (s_K^{n+1} - s_K^n)\psi(\boldsymbol{x}_K, t_{n+1}) m_K \le C_{10}\|\nabla\psi\|_\infty, \quad \forall\, \psi \in \mathcal{C}_c^\infty(Q_{t_f}). \tag{2.46}$$

Proof. For the sake of readability, we denote by $\psi_K^{n+1} = \psi(\boldsymbol{x}_K, t_{n+1})$ for all $K \in \mathcal{V}$ and all $n \in \{0,\dots,M\}$. We multiply (2.21a) by $\Delta t \psi_K^{n+1}$ and sum for $K \in \mathcal{V}$, for $n \in \{0, ..., N\}$. This yields

$$A = B,$$

where

$$A = \sum_{n=0}^{N}\sum_{K\in\mathcal{V}} m_K(s_K^{n+1} - s_K^n)\psi_K^{n+1},$$

$$B = -\sum_{n=0}^{N}\Delta t \sum_{\sigma_{KL}\in\mathcal{E}} a_{KL}\eta_{KL}^{n+1}(u_K^{n+1} - u_L^{n+1})(\psi_K^{n+1} - \psi_L^{n+1}).$$

Using the Cauchy-Schwarz inequality, we get

$$|B|^2 \leq \left(\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}| \eta_{KL}^{n+1} (u_K^{n+1} - u_L^{n+1})^2 \right) \times \left(\sum_{n=0}^{N} \Delta t \sum_{\sigma_{KL} \in \mathcal{E}} |a_{KL}| \eta_{KL}^{n+1} (\psi_K^{n+1} - \psi_L^{n+1})^2 \right).$$

Using Lemma 2.3.6, the boundedness of η and Lemma 2.2.1, we obtain that

$$|B|^2 \leq \|\eta\|_\infty C_8 C_3 \iint_{Q_{t_f}} \Lambda \nabla \psi_{\mathcal{T},\Delta t} \cdot \nabla \psi_{\mathcal{T},\Delta t} \leq \|\eta\|_\infty C_8 C_3 \text{meas}(\Omega) t_f \overline{\Lambda} \|\nabla \psi\|_\infty^2.$$

Therefore (2.46) holds with $C_{10} = \sqrt{\|\eta\|_\infty C_8 C_3 \text{meas}(\Omega) t_f \overline{\Lambda}}$. □

We can now state the expected compactness properties.

Proposition 2.4.2. *There exists a measurable function $p : Q_{t_f} \longrightarrow [p_\star, p^\star]$ such that, up to an unlabeled subsequence, one has*

$$s_{\mathcal{M}_m,\Delta t_m} \underset{m \to +\infty}{\longrightarrow} s(p) \quad \textit{a.e in } Q_{t_f},$$
$$\xi_{\mathcal{T}_m,\Delta t_m} \underset{m \to +\infty}{\longrightarrow} \xi(p) \quad \textit{weakly in } L^2((0,t_f); H^1(\Omega)).$$

Proof. Thanks to (2.35), the sequence $(\nabla \xi_{\mathcal{T}_m,\Delta t_m})_{m \geq 1}$ is bounded in $(L^2(Q_{t_f}))^2$. Moreover, it follows from (2.40) that $(\xi_{\mathcal{T}_m,\Delta t_m})_{m \geq 1}$ is uniformly bounded in $L^2(Q_{t_f})$, providing the boundedness of $(\xi_{\mathcal{T}_m,\Delta t_m})_{m \geq 1}$ in $L^2((0,t_f); H^1(\Omega))$. Therefore, there exists $\Xi \in L^2((0,t_f); H^1(\Omega))$ such that

$$\xi_{\mathcal{T}_m,\Delta t_m} \underset{m \to +\infty}{\longrightarrow} \Xi \quad \text{weakly in } L^2((0,t_f); H^1(\Omega)).$$

By (2.29) we obtain directly that $0 \leq s_{\mathcal{M}_m,\Delta t_m} \leq 1$, ensuring the L^∞-weak-$\star$ convergence of an unlabeled subsequence towards $s \in L^\infty(Q_{t_f}; [0,1])$. Thanks to Lemma 2.4.1, we can apply [14, Theorem 3.9]. It gives the existence of $p : Q_{t_f} \longrightarrow [p_\star, p^\star]$ such that, up to an unlabeled subsequence,

$$s_{\mathcal{M}_m,\Delta t_m} \underset{m \to +\infty}{\longrightarrow} s(p) \quad \text{a.e in } Q_{t_f},$$

and $\Xi = \xi(p)$. □

2.4.2 Identification as a weak solution

Proposition 2.4.3. *Let p be as in Proposition 2.4.2, then p is the unique weak solution to* (2.1) *in the sense of Definition 2.1.4.*

Proof. Let $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_f))$, and denote by $\psi_K^n = \psi(\boldsymbol{x}_K, t_n)$, for all $K \in \mathcal{V}_m$ and all $n \in \{0, ..., N_m\}$. We multiply (2.21a) by $\Delta t_m \psi_K^n$ and sum over $n \in \{0, ..., N_m\}$ and $K \in \mathcal{V}_m$ to obtain

$$A_m + B_m + C_m + D_m = 0, \tag{2.47}$$

where, denoting by $\xi_K^{n+1} = \xi(p_K^{n+1})$, we have set

$$\begin{aligned}
A_m &= \sum_{n=0}^{N_m} \sum_{K \in \mathcal{V}_m} (s_K^{n+1} - s_K^n) \psi_K^n m_K, \\
B_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \Big(\eta_{KL}^{n+1} (p_K^{n+1} - p_L^{n+1}) - \sqrt{\eta_{KL}^{n+1}} (\xi_K^{n+1} - \xi_L^{n+1}) \Big) (\psi_K^n - \psi_L^n), \\
C_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \sqrt{\eta_{KL}^{n+1}} \left(\xi_K^{n+1} - \xi_L^{n+1} \right) (\psi_K^n - \psi_L^n), \\
D_m &= \sum_{n=0}^{N_m} \Delta t_m \sum_{\sigma_{KL} \in \mathcal{E}_m} a_{KL} \eta_{KL}^{n+1} \left(z_K - z_L \right) (\psi_K^n - \psi_L^n).
\end{aligned}$$

Note that $\psi_K^{N_m+1} = 0$ for all $K \in \mathcal{V}_m$, then a discrete integration parts yields

$$\begin{aligned}
A_m &= - \sum_{n=0}^{N_m} \Delta t_m \sum_{K \in \mathcal{V}_m} s_K^{n+1} \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m} m_K - \sum_{K \in \mathcal{V}_m} s_K^0 \psi_K^0 m_K \\
&= - \iint_{Q_{t_f}} s_{\mathcal{M}_m, \Delta t_m} \delta\psi_{\mathcal{M}_m, \Delta t_m} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t - \int_\Omega s_{\mathcal{M}_m}^0 \psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x},
\end{aligned}$$

where the function $\delta\psi_{\mathcal{M}_m, \Delta t_m}$ of $X_{\mathcal{M}_m, \Delta t_m}$ is defined by

$$\delta\psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, t) = \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m} \quad \text{if } (\boldsymbol{x}, t) \in \omega_K \times (t_n, t_{n+1}).$$

Thanks to the regularity of ψ, the function $\delta\psi_{\mathcal{M}_m, \Delta t_m}$ converges uniformly towards $\partial_t \psi$ on Q_{t_f}. Moreover, we have

$$s_{\mathcal{M}_m, \Delta t_m} \longrightarrow s(p) \quad \text{in } L^r(Q_{t_f}) \text{ as } m \to \infty,$$

for all $r \in [1, \infty)$ thanks to Proposition 2.4.2. Therefore,

$$\iint_{Q_{t_f}} s_{\mathcal{M}_m, \Delta t_m} \delta\psi_{\mathcal{M}_m, \Delta t_m} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \longrightarrow \iint_{Q_{t_f}} s(p) \partial_t \psi \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}t \quad \text{as } m \to \infty. \tag{2.48}$$

Moreover, $s_{\mathcal{M}_m}^0$ converges strongly in $L^1(\Omega)$ towards the initial data s_0 and $\psi_{\mathcal{M}_m, \Delta t_m}(\cdot, 0)$ converges uniformly towards $\psi(\cdot, 0)$. Therefore, we get that

$$\int_\Omega s_{\mathcal{M}_m}^0(\boldsymbol{x}) \psi_{\mathcal{M}_m, \Delta t_m}(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x} \longrightarrow \int_\Omega s_0(\boldsymbol{x}) \psi(\boldsymbol{x}, 0) \, \mathrm{d}\boldsymbol{x} \quad \text{as } m \to \infty. \tag{2.49}$$

We deduce from (2.48) and (2.49) that

$$A_m \longrightarrow -\iint_{Q_{t_f}} s(p)\partial_t\psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t - \int_\Omega s_0\psi(.,0)\,\mathrm{d}\boldsymbol{x} \quad \text{as } m\to\infty. \tag{2.50}$$

The term B_m rewrites

$$B_m = \sum_{n=0}^{N_m}\Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m} a_{KL}\sqrt{\eta_{KL}^{n+1}}\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\tilde\eta_{KL}^{n+1}}\right)(p_K^{n+1}-p_L^{n+1})(\psi_K^n-\psi_L^n),$$

where $\tilde\eta_{KL}^{n+1}$ is defined by (2.34). Using the Cauchy-Schwarz inequality, we get

$$|B_m|^2 \le \left(\sum_{n=0}^{N_m}\Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m}|a_{KL}|\eta_{KL}^{n+1}(p_K^{n+1}-p_L^{n+1})^2\right) \times \underbrace{\left(\sum_{n=0}^{N_m}\Delta t_m \sum_{\sigma_{KL}\in\mathcal{E}_m}|a_{KL}|\left(\sqrt{\eta_{KL}^{n+1}}-\sqrt{\tilde\eta_{KL}^{n+1}}\right)^2(\psi_K^n-\psi_L^n)^2\right)}_{:=R_m}. \tag{2.51}$$

The first term in the right-hand side of (2.51) is bounded by C_8 thanks to Lemma 2.3.6. Therefore, in order to prove that $\lim_{m\to\infty} B_m = 0$, it suffices to prove that $\lim_{m\to\infty} R_m = 0$. For $T\in\mathcal{T}_m$, we denote by

$$\overline{\xi}_T^{n+1} = \max_{\boldsymbol{x}\in T}\left(\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t_{n+1})\right), \qquad \underline{\xi}_T^{n+1} = \min_{\boldsymbol{x}\in T}\left(\xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t_{n+1})\right),$$

and we define the piecewise constant functions $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}$ and $\underline{\xi}_{\mathcal{T}_m,\Delta t_m}$ by

$$\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) = \overline{\xi}_T^{n+1} \quad\text{and}\quad \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) = \underline{\xi}_T^{n+1} \quad \text{if } (\boldsymbol{x},t)\in T\times(t_n,t_{n+1}),$$

We can estimate

$$\left|\sqrt{\eta_{KL}^{n+1}}-\sqrt{\tilde\eta_{KL}^{n+1}}\right| \le \mu\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right), \quad \forall\sigma_{KL}\in\mathcal{E}_T, \tag{2.52}$$

where μ is the continuity modulus defined in (2.9). Using (2.52) in the definition (2.51) of R_m, we get

$$0\le R_m \le \sum_{n=0}^{N_m}\Delta t_m\sum_{T\in\mathcal{T}_m}\mu\left(\overline{\xi}_T^{n+1}-\underline{\xi}_T^{n+1}\right)^2\sum_{\sigma_{KL}\in\mathcal{E}_T}|a_{KL}^T|(\psi_K^n-\psi_L^n)^2. \tag{2.53}$$

Following the proof of Lemma 2.2.1 (cf. [39, Lemma 3.2]), we can prove that

$$\sum_{\sigma_{KL}\in\mathcal{E}_T}|a_{KL}^T|(\psi_K^n-\psi_L^n)^2 \le C_3\overline{\Lambda}\|\nabla\psi\|_\infty^2\mathrm{meas}(T), \qquad \forall T\in\mathcal{T}. \tag{2.54}$$

Therefore, we deduce from (2.53) that

$$0\le R_m \le C\iint_{Q_{t_f}}\mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m}-\underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2\mathrm{d}\boldsymbol{x}\,\mathrm{d}t, \tag{2.55}$$

where C depends only on Λ, θ and ψ. Since μ is bounded (as η is bounded), continuous, with $\mu(0) = 0$, it suffices to show that $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t)$ tends to 0 almost everywhere in Q_{t_f} as $m \to \infty$ (up to an unlabeled subsequence). Thanks to [39, Lemma A.1] and to Lebesgue's dominated convergence theorem, one has

$$\iint_{Q_{t_f}} \left|\overline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t)\right| \mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\ \leq Ch_m \iint_{Q_{t_f}} |\nabla \xi_{\mathcal{T}_m,\Delta t_m}|\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \underset{m\to+\infty}{\longrightarrow} 0. \quad (2.56)$$

As a consequence of (2.51), (2.55) and (2.56), and still up to the extraction of an unlabeled subsequence, one has

$$\lim_{m\to\infty} B_m = \lim_{m\to\infty} R_m = 0. \quad (2.57)$$

Let us now focus on the term C_m. We define the piecewise constant functions $\Xi_{\mathcal{T}_m,\Delta t_m}$ and $H_{\mathcal{T}_m,\Delta t_m}$ by

$$\Xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x},t) = \xi_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_T,t), \quad \forall \boldsymbol{x} \in T, \forall t \in (t_n, t_{n+1}),$$

$\boldsymbol{x}_T$ being the center of mass of the triangle T, and by $H_{\mathcal{T}_m,\Delta t_m} = \eta \circ s \circ \xi^{-1}(\Xi_{\mathcal{T}_m,\Delta t_m})$. Clearly, one has

$$\underline{\xi}_{\mathcal{T}_m,\Delta t_m} \leq \Xi_{\mathcal{T}_m,\Delta t_m} \leq \overline{\xi}_{\mathcal{T}_m,\Delta t_m}.$$

It follows from (2.56) that both $\underline{\xi}_{\mathcal{T}_m,\Delta t_m}$ and $\overline{\xi}_{\mathcal{T}_m,\Delta t_m}$ converge almost everywhere to $\xi(p)$, hence so does $\Xi_{\mathcal{T}_m,\Delta t_m}$. This provides that

$$H_{\mathcal{T}_m,\Delta t_m} \longrightarrow \eta(s(p)) \quad \text{ in } L^1(Q_{t_f}) \text{ as } m \to \infty. \quad (2.58)$$

We introduce the term

$$C'_m = \iint_{Q_{t_f}} \sqrt{H_{\mathcal{T}_m,\Delta t_m}} \Lambda \nabla \xi_{\mathcal{T}_m,\Delta t_m} \cdot \nabla \psi_{\mathcal{T}_m,\Delta t_m}(.,t-\Delta t_m)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

Since $\nabla \xi_{\mathcal{T}_m,\Delta t_m}$ converges weakly in $L^2(Q_{t_f})$ towards $\nabla \xi(p)$, since $\nabla \psi_{\mathcal{T}_m,\Delta t_m}$ converges uniformly towards $\nabla \psi$, and thanks to (2.58), we obtain that

$$\lim_{m\to\infty} C'_m = \iint_{Q_{t_f}} \sqrt{\eta(s(p))} \Lambda \nabla \xi(p) \cdot \nabla \psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \quad (2.59)$$

Let us now check that $|C_m - C'_m|$ tends to 0 as m tends to ∞. We denote by

$$\eta_T^{n+1} = H_{\mathcal{T}_m,\Delta t_m}(\boldsymbol{x}_T, t_{n+1}), \quad \forall T \in \mathcal{T}_m, \forall n \in \{0, ..., N_m\}.$$

The term C'_m can be rewritten

$$C'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sqrt{\eta_T^{n+1}} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T \left(\xi_K^{n+1} - \xi_L^{n+1}\right)(\psi_K^n - \psi_L^n),$$

so that

$$C_m - C'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T \left(\sqrt{\eta_{KL}^{n+1}} - \sqrt{\eta_T^{n+1}}\right) \left(\xi_K^{n+1} - \xi_L^{n+1}\right) (\psi_K^n - \psi_L^n).$$

For all $n \in \{0, ..., N_m\}$, for all $T \in \mathcal{T}_m$, and for all $\sigma_{KL} \in \mathcal{E}_T$, one has

$$\left|\sqrt{\eta_{KL}^{n+1}} - \sqrt{\eta_T^{n+1}}\right| \leq \mu(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}) \tag{2.60}$$

where μ is the continuity modulus defined in (2.9). Then one obtains that

$$|C_m - C'_m| \leq \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \Bigg[\mu(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}) \times \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T| \left|\xi_K^{n+1} - \xi_L^{N+1}\right| |\psi_K^n - \psi_L^n|\Bigg].$$

The Cauchy-Schwarz inequality provides

$$|C_m - C'_m|^2 \leq \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T| (\psi_K^n - \psi_L^n)^2 \times \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T| \left(\xi_K^{n+1} - \xi_L^{n+1}\right)^2.$$

Using Lemma 2.2.1 and Lemma 2.3.4, together with (2.54), one deduces that

$$|C_m - C'_m|^2 \leq C \iint_{Q_{t_f}} \mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m} - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}t,$$

thus $|C_m - C'_m|$ tends to 0 thanks to the arguments already developed to prove that R_m tends to 0. Finally, we obtain that

$$\lim_{m\to\infty} C_m = \iint_{Q_{t_f}} \sqrt{\eta(s(p))}\Lambda\nabla\xi(p)\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{2.61}$$

Let us focus on the last term D_m. We introduce the term

$$D'_m = \iint_{Q_{t_f}} H_{\mathcal{T}_m,\Delta t_m}\Lambda\nabla z\cdot\nabla\psi_{\mathcal{T}_m,\Delta t_m}(.,t-\Delta t_m)\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

It follows from (2.58) and from the uniform convergence of $\nabla\psi_{\mathcal{T}_m,\Delta t_m}$ towards $\nabla\psi$ as m tends to $+\infty$ that

$$\lim_{m\to\infty} D'_m = \iint_{Q_{t_f}} \eta(s(p))\Lambda\nabla z\cdot\nabla\psi\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t.$$

We will now check that $|D_m - D'_m| \longrightarrow 0$ as $m \to \infty$. The term D'_m rewrites

$$D'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \eta_T^{n+1} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T (z_K - z_L)(\psi_K^n - \psi_L^n),$$

so that

$$D_m - D'_m = \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} a_{KL}^T \left(\eta_{KL}^{n+1} - \eta_T^{n+1}\right)(z_K - z_L)(\psi_K^n - \psi_L^n).$$

For all $\sigma_{KL} \in \mathcal{E}_T$, one has

$$\left|\eta_{KL}^{n+1} - \eta_T^{n+1}\right| \leq \left|\sqrt{\eta_{KL}^{n+1}} - \sqrt{\eta_T^{n+1}}\right| \left(\sqrt{\eta_{KL}^{n+1}} + \sqrt{\eta_T^{n+1}}\right) \leq 2\|\eta\|_\infty \mu(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}).$$

Therefore, using the Cauchy-Schwarz inequality, one has

$$|D_m - D'_m|^2 \leq 4\|\eta\|_\infty^2 \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \mu\left(\overline{\xi}_T^{n+1} - \underline{\xi}_T^{n+1}\right)^2 \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(\psi_K^n - \psi_L^n)^2$$
$$\times \sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(z_K - z_L)^2.$$

We use Lemma 2.2.1 to get

$$\sum_{n=0}^{N_m} \Delta t_m \sum_{T\in\mathcal{T}_m} \sum_{\sigma_{KL}\in\mathcal{E}_T} |a_{KL}^T|(z_K - z_L)^2 \leq C_3 \iint_{Q_{t_f}} \Lambda \nabla z \cdot \nabla z \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \leq C.$$

We deduce from (2.54) that

$$|D_m - D'_m|^2 \leq C \iint_{Q_{t_f}} \mu\left(\overline{\xi}_{\mathcal{T}_m,\Delta t_m} - \underline{\xi}_{\mathcal{T}_m,\Delta t_m}\right)^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}t \underset{m\to\infty}{\longrightarrow} 0,$$

and then that

$$\lim_{m\to\infty} D_m = \iint_{Q_{t_f}} \eta(s(p))\Lambda\nabla z \cdot \nabla\psi \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t. \tag{2.62}$$

Putting (2.50), (2.57), (2.61) and (2.62) together in (2.47) provides that p satisfies the weak formulation (2.13), then it is the unique weak solution to the problem (cf. Theorem 2.1.5). □

Finally, let us remark that since the weak solution p is unique, all the convergence in functional space that were proved to occur up to the extraction of a subsequence are valid for the whole sequences. Concerning the almost everywhere convergence, we cannot do better than saying that it holds up to a subsequence.

2.5 Numerical results

Let us provide some illustrations of the behavior of the numerical scheme (2.21). The scheme leads to a nonlinear system that we solve thanks to the Newton-Raphson method with Matlab. As proved in Proposition 2.3.1, the approximate pressure remains greater than $p_\star$. Therefore, we project the discrete solution at each Newton

iteration on the set $\{p \geq p_\star\}$. We refer to [28, 123] for a study on iterative methods for solving Richard's equation.

In all our test cases, the domain is the unit square, i.e., $\Omega = (0,1)^2$. We use meshes coming from the 2D benchmark on anisotropic diffusion problems [102]. An illustration of the meshes is given in Figure 2.4. These triangle meshes show no symmetry which could artificially increase the convergence rate. All angles are acute, so that, in the case of an isotropic tensor Λ, the coefficients a_{KL} are all non-negative. This is no longer the case when Λ is chosen to be anisotropic. To be more precise concerning the diffusion tensor, we have considered constant diagonal tensors

$$\Lambda = \begin{pmatrix} \Lambda_{xx} & 0 \\ 0 & \Lambda_{yy} \end{pmatrix}$$

where Λ_{xx} and Λ_{yy} are chosen constant in Ω, and the gravity acceleration $\mathbf{g}$ is defined by $\mathbf{g} = (g, 0)^T$ for all $\boldsymbol{x} \in \Omega$ with $g \in \mathbb{R}_+$.

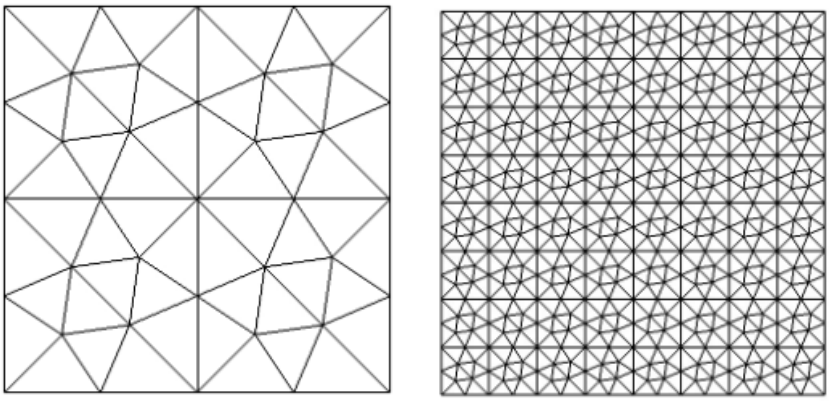


FIGURE 2.4 – Second and fourth meshes used in the numerical tests

The numerical analysis of the scheme was carried out for a uniform time discretization of $(0, t_f)$ only in order to avoid heavy notations. In order to increase the robustness of the algorithm and to ensure the convergence of the Newton-Raphson iterative procedure, we used an adaptive time step procedure in the practical implementation. More precisely, to each mesh, we associate a maximal time step Δt_k, k being the index of the mesh (1 for the coarsest, 8 for the finest). If the Newton-Raphson method fails to converge after 30 iterations —we choose that the ℓ^∞ norm of the residual has to be smaller than 10^{-7} as stopping criterion—, the time step is divided by two. If the Newton-Raphson method converges, the time step is multiplied by two and projected on $[0, \Delta t_k]$.

In sections §2.5.1, §2.5.2 and §2.5.3, we give evidence of the convergence of scheme (2.21) on test cases where exact analytical solutions are known. We are

interested in the convergence speed of our method when the discretization parameters h and Δt tend to 0. We focus on the error caused by the spatial discretization (the time discretization is a classical first order accurate backward Euler method). As we will see, our scheme is at most first order accurate. In order to be sure that the error caused by the time discretization will not be of leading order, we choose $\Delta t_{k+1} = \Delta t_k/4$ while $h_{k+1} = h_k/2$, h_k being the size the mesh $k \in \{1, \dots, 8\}$. The first time step Δt_1 to 0.01024 in all the test cases presented below.

The test cases we chose to present here do not perfectly match with the assumptions presented at the beginning of the chapter. They rather isolate the difficulties of the problem and give a better view of the behavior of the scheme. More precisely, the so called Hornung-Messing problem presented in §2.5.1 aims to illustrate the behavior of the scheme when an elliptic degeneracy occurs. The linear Fokker-Planck problem of §2.5.2 illustrates the behavior of the scheme for a stiff problem when $p_\star = -\infty$. The porous medium equation with drift presented in §2.5.3 allows to illustrate the behavior of the scheme near a hyperbolic degeneracy at $s(p) = 0$. The test case presented in §2.5.4 is there to illustrate numerically the decay of the free energy. Finally, we illustrate the behavior of Newton's method in §2.5.5. Let us stress that the numerical analysis we developed in the chapter can be adapted without any major modification to all the cases we present here.

In the case where $p_\star = -\infty$, it was proved in Lemma 2.3.10 that there exists $C_\star > -\infty$ depending only on $\mathcal{T}, \Delta t, \Omega, s, \overline{s}_0, t_f, \Lambda, \theta_\mathcal{T}, \eta$ and z such that

$$p_K^{n+1} \geq C_\star, \qquad \forall K \in \mathcal{V}, \quad \forall n \in \{0, ..., N\}.$$

Therefore we initialize the Newton-Raphson algorithm by

$$p_K^{n+1,0} = \max(s^{-1}(\epsilon), p_K^n), \quad \text{where } \epsilon = 10^{-14}.$$

Let us mention that in the tests 2, 3, and 4, we considered problems without elliptic degeneracy. The corresponding functions s are increasing on $(p_\star, +\infty)$. Therefore, we can choose $S = s(p)$ rather that p as a primary unknown in these cases. Denoting by $p = s^{-1}$, the problem solved numerically in §2.5.2, §2.5.3 and §2.5.4 can then be written

$$\partial_t S - \nabla \cdot (\Lambda \eta(S)(\nabla p(S) - \mathbf{g})) = 0 \quad \text{in } Q_{t_f}. \tag{2.63}$$

Finally, we have set the gravity $\mathbf{g} = \mathbf{e}_x$ horizontal from the left to the right in the tests 2, 3, and 4. As a consequence, the scheme we considered in §2.5.2, §2.5.3, and

§2.5.4 is

$$\frac{s_K^{n+1}-s_K^n}{\Delta t}m_K + \sum_{\sigma_{KL}\in\mathcal{E}_K}\eta_{KL}^{n+1}a_{KL}(u_K^{n+1}-u_L^{n+1}) = 0, \tag{2.64a}$$

$$u_K^{n+1} = p_K^{n+1} - x_K, \tag{2.64b}$$

$$p_K^{n+1} = p(s_K^{n+1}), \tag{2.64c}$$

$$\eta_{KL}^{n+1} = \begin{cases}\eta(s_K^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1}-u_L^{n+1}) \geq 0,\\ \eta(s_L^{n+1}) & \text{if} \quad a_{KL}(u_K^{n+1}-u_L^{n+1}) < 0.\end{cases} \tag{2.64d}$$

2.5.1 Test 1 : A test case with saturated zones

The first test-case we propose here is the so-called Hornung-Messing problem [106]. In this problem, gravity is neglected (i.e. $g = 0$ and $u_K^{n+1} = p_K^{n+1}$ for all $K \in \mathcal{V}$ and $n \geq 0$). We consider the following nonlinearities

$$\eta(p) = \begin{cases}\dfrac{2}{1+p^2} & \text{if } p<0,\\ 2 & \text{if } p\geq 0,\end{cases} \qquad s(p) = \begin{cases}\left(\dfrac{\pi^2}{4} - \arctan^2(p)\right)(\Lambda_{xx}+\Lambda_{yy}) & \text{if } p<0,\\ \dfrac{\pi^2}{4}(\Lambda_{xx}+\Lambda_{yy}) & \text{if } p\geq 0.\end{cases}$$

and the exact solution to the Richards equation

$$p_{\text{ex}} = \begin{cases}-\dfrac{x-y-t}{2} & \text{if } x-y-t<0,\\ -\tan\left(\dfrac{e^{x-y-t}-1}{e^{x-y-t}+1}\right) & \text{if } x-y-t\geq 0,\end{cases} \qquad \forall (x,y)\in\Omega,\ \forall t\in(0,t_f), \tag{2.65}$$

where t_f was set to 0.05. This exact solution does not satisfies the no-flux boundary conditions. Therefore, we prescribe the exact solution p_{ex} as Dirichlet boundary conditions on $\partial\Omega\times(0,t_f)$. In Tables 2.1 and 2.2, we report the errors

$$err_{L^p} = \|p_{\mathcal{M},\Delta t} - p_{\text{ex}}\|_{L^p(Q_{t_f})} \quad \text{for } p = 1, 2, \infty$$

for 7 successively refined meshes in the Isotropic case $\Lambda = \text{Id}$ and in the anisotropic case $\Lambda = \text{diag}(1, 10^{-3})$.

We observe that numerical order of convergence is close to 1 for the three norms whatever the anisotropy tensor on this test case.

2.5.2 Test 2 : Linear Fokker-Planck equation

In this test case, we study the behavior of our scheme on the problem (2.1) with the choice of nonlinearities $s(p) = \exp(p)$ and $\eta(s) = s$. The function s does not fulfill

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate
0.500	12	0.343E-3	-	0.548E-4	-	0.352E-2	-
0.250	37	0.218E-3	0.651	0.472E-4	0.215	0.197E-2	0.838
0.125	129	0.141E-3	0.629	0.329E-4	0.522	0.113E-2	0.801
0.063	481	0.769E-4	0.886	0.185E-4	0.844	0.607E-3	0.907
0.031	1857	0.399E-4	0.927	0.967E-5	0.912	0.306E-3	0.966
0.016	7297	0.202E-4	1.025	0.493E-5	1.019	0.154E-3	1.041
0.008	28929	0.102E-4	0.989	0.249E-5	0.986	0.771E-4	0.996
0.004	115201	0.512E-5	0.994	0.125E-5	0.993	0.386E-4	0.997

TABLEAU 2.1 – Test 1, isotropic case $\Lambda = \mathrm{Id}$.

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate
0.500	12	0.382E-3	-	0.581E-4	-	0.384E-2	-
0.250	37	0.368E-3	0.057	0.682E-4	-0.231	0.396E-2	-0.044
0.125	129	0.225E-3	0.710	0.475E-4	0.522	0.218E-2	0.861
0.063	481	0.120E-3	0.911	0.268E-4	0.838	0.112E-2	0.974
0.031	1857	0.621E-4	0.933	0.141E-4	0.904	0.522E-3	1.075
0.016	7297	0.315E-4	1.026	0.721E-5	1.012	0.260E-3	1.052
0.008	28929	0.159E-4	0.990	0.365E-5	0.983	0.130E-3	1.003
0.004	115201	0.796E-5	0.993	0.183E-5	0.992	0.647E-4	1.002

TABLEAU 2.2 – Test 1, anisotropic case with $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 10^{-3}$

Assumption (**A1**) since s is not constant on $\mathbb{R}_+$. Since s in injective, we can use $S = s(p)$ as a primary unknown, leading to the problem

$$\begin{cases} \partial_t S - \nabla\cdot(S\Lambda(\nabla \log(S) - \mathbf{e}_x)) = 0 & \text{in } Q_{t_f}, \\ S\Lambda(\nabla \log(S) - \mathbf{e}_x)\cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0,T), \\ S_{|t=0} = s_0 & \text{in } \Omega, \end{cases} \tag{2.66}$$

that turns out to be the linear convection diffusion equation

$$\begin{cases} \partial_t S - \nabla\cdot(\Lambda(\nabla S - S\mathbf{e}_x) = 0 & \text{in } Q_{t_f}, \\ \Lambda(\nabla S - S\mathbf{e}_x)\cdot \mathbf{n} = 0 & \text{on } \partial\Omega \times (0,T), \\ S_{|t=0} = s_0 & \text{in } \Omega. \end{cases} \tag{2.67}$$

We compare the results obtained with the nonlinear CVFE scheme (2.64) with the following *linear scheme* where the convection is discretized thanks to centered fluxes :

$$\frac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL}\in\mathcal{E}_K} a_{KL}\left((s_K^{n+1} - s_L^{n+1}) + (x_K - x_L)\frac{s_K^{n+1} + s_L^{n+1}}{2}\right) = 0 \tag{2.68}$$

for all $K \in \mathcal{V}$ and for all $n \in \{0, ..., N\}$.

The schemes (2.64) and (2.68) are compared on the following analytical solution built from a 1D case :

$$s_{\text{ex}}(x, y, t) = \exp(-\alpha t + \frac{x}{2}) \left(\pi \cos(\pi x) + \frac{1}{2} \sin(\pi x) \right) + \pi \exp(x - \frac{1}{2}) \text{ in } Q_{t_f},$$

where $\alpha = \Lambda_{xx}(\pi^2 + \frac{1}{4})$, and where the final time has been fixed to 0.05. This analytical solution is nonnegative and satisfies homogeneous Neumann boundary conditions.

In Tables 2.3 to 2.6, we report the $L^1(Q_{t_f})$, $L^2(Q_{t_f})$, and $L^\infty(Q_{t_f})$ on the variable S, i.e.,

$$err_{L^p} = \|s_{\mathcal{M},\Delta t} - s_{\text{ex}}\|_{L^p(Q_{t_f})} \quad \text{for } p = 1, 2, \infty$$

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.328E-01	-	0.820E-02	-	0.232E+00	-	0
0.250	37	0.306E-01	0.0979	0.798E-02	0.0389	0.239E+00	-0.0466	0
0.125	129	0.198E-01	0.6320	0.508E-02	0.6519	0.153E+00	0.6477	0
0.063	481	0.109E-01	0.8674	0.276E-02	0.8911	0.841E-01	0.8722	0
0.031	1857	0.570E-02	0.9130	0.143E-02	0.9237	0.441E-01	0.9101	0
0.016	7297	0.292E-02	1.0152	0.729E-03	1.0214	0.226E-01	1.0123	0
0.008	28929	0.147E-02	0.9845	0.368E-03	0.9893	0.114E-01	0.9831	0
0.004	115201	0.741E-03	0.9923	0.185E-03	0.9937	0.575E-02	0.9913	0

Tableau 2.3 – Test 2, nonlinear scheme (2.64), with an isotropic tensor $\Lambda = \text{Id}$.

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.294E-01	-	0.372E-02	-	0.484E+00	-	0
0.250	37	0.829E-02	1.8267	0.198E-02	1.6352	0.166E+00	1.5428	0
0.125	129	0.218E-02	1.9286	0.349E-03	1.8389	0.426E-01	1.9639	0
0.063	481	0.548E-03	2.0138	0.859E-04	1.9863	0.108E-01	2.0069	0
0.031	1857	0.137E-03	1.9521	0.216E-04	1.9310	0.274E-02	1.9310	0
0.016	7297	0.343E-04	2.0956	0.542E-05	2.0675	0.697E-03	2.0675	0
0.008	28929	0.858E-05	1.9998	0.135E-05	1.9994	0.178E-03	1.9720	0
0.004	115201	0.214E-05	2.0000	0.339E-06	1.9998	0.453E-04	1.9719	0

Tableau 2.4 – Test 2, linear scheme (2.68) with an isotropic tensor $\Lambda = \text{Id}$.

The numerical order of convergence of the linear scheme (2.68) is close to 2. However, the more the anisotropy ratio is important, the more we observe oscillations and undershoots (see in particular Table 2.6). The nonlinear scheme (2.64) preserves

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.179E+00	-	0.488E-01	-	1.022E+00	-	0
0.250	37	0.166E+00	0.1080	0.462E-01	0.0792	0.959E+00	0.0930	0
0.125	129	0.118E+00	0.4947	0.318E-01	0.5396	0.744E+00	0.3659	0
0.063	481	0.746E-01	0.6685	0.197E-01	0.7008	0.504E+00	0.5689	0
0.031	1857	0.439E-01	0.7498	0.113E-01	0.7755	0.309E+00	0.6880	0
0.016	7297	0.243E-01	0.8904	0.621E-02	0.9118	0.177E+00	0.8416	0
0.008	28929	0.130E-01	0.9087	0.327E-02	0.9229	0.964E-01	0.8793	0
0.004	115201	0.672E-02	0.9481	0.169E-02	0.9571	0.506E-01	0.9304	0

TABLEAU 2.5 – Test 2 : nonlinear scheme (2.64) with an anisotropic tensor $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 20$.

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.566E-01	-	0.115E-01	-	0.376E+00	-	0
0.250	37	0.222E-01	1.3523	0.427E-02	1.4290	0.250E+00	0.5878	0
0.125	129	0.613E-02	1.8553	0.119E-02	1.8469	0.883E-01	1.5036	-2.1867E-03
0.063	481	0.155E-02	2.0021	0.300E-03	2.0053	0.247E-01	1.8621	-9.3704e-04
0.031	1857	0.390E-03	1.9506	0.755E-04	1.9468	0.647E-02	1.8859	-2.6687e-04
0.016	7297	0.976E-04	2.0948	0.189E-04	2.0952	0.168E-02	2.0358	-6.9729e-05
0.008	28929	0.244E-04	1.9997	0.472E-05	1.9997	0.437E-03	1.9470	-1.7741e-05
0.004	115201	0.610E-05	1.9999	0.118E-05	1.9999	0.113E-03	1.9495	-4.4696e-06

TABLEAU 2.6 – Test 2, linear scheme (2.68) with an anisotropic tensor : $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 20$.

the positivity of the solution whatever the anisotropy, but this property has a cost. Indeed, the numerical diffusion introduced by the nonlinear scheme (2.64) becomes very important when the anisotropy ratio is large. This yields a loss of accuracy. The method (2.64) seems to be first order accurate, i.e.,

$$err_{L^p} \leq C_p(\Lambda, \theta)h, \quad p \in \{1, 2, \infty\}, \tag{2.69}$$

but with constants $C_p(\Lambda, \theta)$ that strongly depend on the anisotropy ratio and of the regularity of the mesh.

2.5.3 Test 3 : Porous medium equation with drift

In this third test case, we set $s(p) = p/2$ if $p \geq 0$ and $\eta(s) = s$. Choosing $S = s(p)$ as a primary variable, we obtain the degenerate parabolic equation

$$\partial_t S - \nabla \cdot (\Lambda(2|S|\nabla S - S\mathbf{e}_x)) = 0 \quad \text{in } Q_{t_f},$$

or equivalently

$$\partial_t S - \nabla \cdot (\Lambda(\nabla\varphi(S) - S\mathbf{e}_x)) = 0 \quad \text{in } Q_{t_f}, \quad \text{where } \varphi(S) = |S|S. \tag{2.70}$$

The function s_{ex} defined by

$$s_{\text{ex}}(x, y, t) = \max(\beta t - x, 0),\ \forall((x, y), t) \in Q_{t_f}, \tag{2.71}$$

with $\beta = 2\Lambda_{xx}$ satisfies the equation (2.70). As in Test 1, we complement (2.70) by Dirichlet boundary conditions and an initial condition prescribed by (2.71). The final time t_f has been set to 0.05.

The nonlinear scheme (2.64) is adapted to the case of Dirichlet boudary conditions : (2.64a) is assumed to hold only for $K \in \mathcal{V}_{\text{int}} = \{K \in \mathcal{V} \mid \boldsymbol{x}_K \notin \partial\Omega\}$. The equations (2.64b) and (2.64c) are enforced for all $K \in \mathcal{V}$, and (2.64d) is enforced for all $\sigma_{KL} \in \mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E} \mid \sigma \not\subset \partial\Omega\}$. In order to close the system, we impose $s_K^{n+1} = s_{\text{ex}}(\boldsymbol{x}_K, t_{n+1})$ for all K such that $\boldsymbol{x}_K \in \partial\Omega$.

The numerical results obtained thanks to our scheme are compared with those obtained thanks to a so-called *quasilinear scheme* where (2.64a) has been replaced by

$$\frac{s_K^{n+1} - s_K^n}{\Delta t} m_K + \sum_{\sigma_{KL} \in \mathcal{E}_K} a_{KL} \left(\varphi(s_K^{n+1}) - \varphi(s_L^{n+1}) + (x_K - x_L) \frac{s_K^{n+1} + s_L^{n+1}}{2} \right) = 0. \tag{2.72}$$

The analytical solution s_{ex} defined by (2.71) belongs to $C([0, t_f), H^{3/2-\epsilon}(\Omega))$ for all $\epsilon > 0$. Therefore, we expect for the quasilinear scheme (2.72) a convergence order close to 1.5 in the $L^2(Q_{t_f})$ norm, as observed in Tables 2.8 and 2.10.

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.174E-02	-	0.215E-03	-	0.284E-01	-	0
0.250	37	0.238E-02	-0.4509	0.224E-03	-0.0573	0.559E-01	-0.9751	0
0.125	129	0.168E-02	0.5062	0.160E-03	0.4883	0.305E-01	0.8754	0
0.063	481	0.100E-02	0.7489	0.889E-04	0.8544	0.237E-01	0.3645	0
0.031	1857	0.609E-03	0.7049	0.486E-04	0.8522	0.174E-01	0.4369	0
0.016	7297	0.359E-03	0.7994	0.259E-04	0.9509	0.114E-01	0.6459	0
0.008	28929	0.206E-03	0.8043	0.136E-04	0.9315	0.734E-02	0.6301	0
0.004	115201	0.115E-03	0.8445	0.703E-05	0.9511	0.460E-02	0.6751	0

TABLEAU 2.7 – Nonlinear scheme, with an isotropic tensor : $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 1$

We observe that, as expected, that the nonlinear scheme (2.64) has a smaller order of convergence (less than 1) when Λ is isotropic, cf. Table 2.5. Here again, as in Test 2, the accuracy is strongly affected by the anisotropy. The numerical diffusion

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.990E-03	-	0.120E-03	-	0.166E-01	-	0
0.250	37	0.148E-02	-0.5805	0.129E-03	-0.1076	0.383E-01	-1.2026	0
0.125	129	0.825E-03	0.8427	0.720E-04	0.8424	0.176E-01	1.1253	0
0.063	481	0.356E-03	1.2265	0.268E-04	1.4434	0.106E-01	0.7307	0
0.031	1857	0.151E-03	1.2052	0.998E-05	1.3912	0.582E-02	0.8507	0
0.016	7297	0.581E-04	1.4499	0.320E-05	1.7193	0.296E-02	1.0232	-1.3853e-18
0.008	28929	0.214E-04	1.4403	0.950E-06	1.7531	0.149E-02	0.9921	-6.9053e-17
0.004	115201	0.711E-05	1.4722	0.270E-06	1.8125	0.743E-03	1.0008	-2.1592e-18

TABLEAU 2.8 – Quasilinear scheme, with an isotropic tensor : $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 1$

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.672E-02	-	0.983E-03	-	0.829E-01	-	0
0.250	37	0.664E-02	0.0178	0.102E-02	-0.0551	0.101E00	-0.2802	0
0.125	129	0.552E-02	0.2663	0.862E-03	0.2439	0.831E-01	0.2774	0
0.063	481	0.441E-02	0.3286	0.647E-03	0.4191	0.699E-01	0.2526	0
0.031	1857	0.345E-02	0.3471	0.458E-03	0.4876	0.625E-01	0.1586	0
0.016	7297	0.260E-02	0.4284	0.310E-03	0.5954	0.514E-01	0.2946	0
0.008	28929	0.189E-02	0.4608	0.200E-03	0.6241	0.410E-01	0.3266	0
0.004	115201	0.132E-02	0.5141	0.125E-03	0.6794	0.318E-01	0.3666	0

TABLEAU 2.9 – Nonlinear scheme, with an anisotropic tensor : $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$

introduced by the scheme increases with the anisotropy ratio. But the solutions to the scheme (2.21) do not present undershoots (up to the precision of the nonlinear solver), on the contrary to the solutions to the quasilinear scheme (2.72), cf. Table 2.10. In order to illustrate the overdiffusive behavior of the nonlinear scheme (2.64) as well as the undershoots produced by the quasilinear scheme (2.72), we present in Figure 2.5 the snapshots of both numerical solutions at time $t = t_f$.

2.5.4 Decay of discrete free energy

Let us denote by $\mathfrak{M}(Q_{t_f})$ the set of the measurable functions mapping Q_{t_f} to $\mathbb{R}$. The *free energy* functional [110] $\mathfrak{E} : \mathfrak{M}(Q_{t_f}) \to \mathbb{R} \cup \{+\infty\}$, defined by

$$\mathfrak{E}(p) = \int_\Omega \Big(\Gamma(p) + s(p)\mathbf{g} \cdot \boldsymbol{x} \Big) \, \mathrm{d}\boldsymbol{x}, \qquad \forall p \in \mathfrak{M}(Q_{t_f}), \tag{2.73}$$

consists in the sum of the capillary energy (2.6), and the gravitational energy. We have formally the classical energy/dissipation property (2.11), and in particular $t \longmapsto$

h	$\#\mathcal{V}$	err_{L^2}	rate	err_{L^1}	rate	err_{L^∞}	rate	S_{min}
0.500	12	0.976E-02	-	0.159E-01	-	0.111E+00	-	-5.4034E-02
0.250	37	0.722E-02	0.4337	0.110E-02	0.5325	0.108E+00	0.0424	-3.5579E-02
0.125	129	0.414E-02	0.8015	0.583E-03	0.9103	0.589E-01	0.8736	-2.5825E-02
0.063	481	0.179E-02	1.2215	0.198E-03	1.5786	0.419E-01	0.4968	-1.1696E-02
0.031	1857	0.779E-03	1.1765	0.746E-03	1.3747	0.220E-01	0.9062	-5.8549E-03
0.016	7297	0.336E-02	1.2698	0.262E-04	1.5806	0.118E-01	0.9376	-2.9309E-03
0.008	28929	0.140E-03	1.2662	0.876E-05	1.5822	0.636E-02	0.8980	-1.4663E-03
0.004	115201	0.565E-04	1.3073	0.282E-05	1.6351	0.333E-02	0.9341	-7.3339E-04

TABLEAU 2.10 – Quasilinear scheme, with an anisotropic tensor : $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$

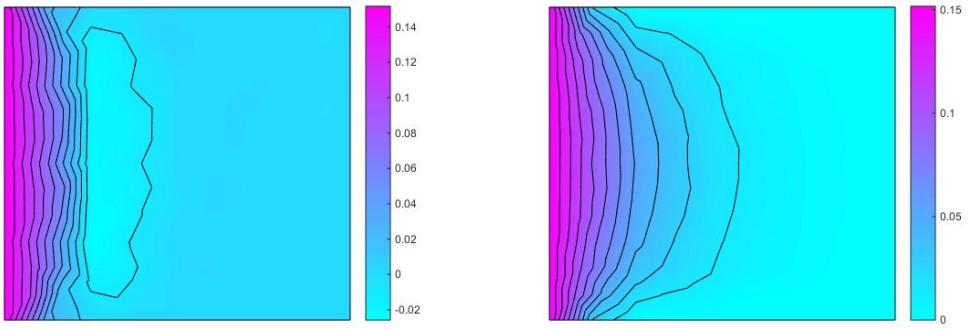


FIGURE 2.5 – Test 3 : 2nd mesh and anisotropic tensor $\Lambda_{xx} = 1$ and $\Lambda_{yy} = 100$. Discrete solutions $s_{\mathcal{M},\Delta t}(\cdot, t_f)$ and their iso-values. *Left :* Quasilinear scheme (2.72). *Right :* Nonlinear scheme (2.64).

$\mathfrak{E}(p)(t)$ is decreasing. The discrete counterpart of the *free energy* is

$$\mathfrak{E}(p^n_{\mathcal{M}}) = \sum_{K\in\mathcal{V}} m_K \Big[\Gamma(p^n_K) + gs(p^n_K)x_K\Big].$$

We have not succeeded to prove the decay of the discrete free energy contrarily to [39]. Let us provide a numerical evidence of this energy/dissipation property. Define the nonlinearities

$$s(p) = \begin{cases} \dfrac{1}{1+p^2} & \text{if } p < 0, \\ 1 & \text{if } p \geq 0, \end{cases} \qquad \eta(s) = s^2,$$

and set $\mathbf{g} = \mathbf{e}_x$, and

$$p_0 = \begin{cases} -\dfrac{x-y}{2} & \text{if } x - y < 0, \\ -\tan\left(\dfrac{e^{x-y}-1}{e^{x-y}+1}\right) & \text{if } x - y \geq 0. \end{cases}$$

We solve the scheme (2.21) and we remark (cf. Figure 2.6) that $(\mathfrak{E}(p^n_{\mathcal{M}}))_{n\geq 0}$ is decreasing. As already noticed on the previous test cases, the scheme (2.21) suffers from an excessive numerical diffusion, in particular when the anisotropy ratio is high. The origins of faster convergence towards the equilibrium in the anisotropic case illustrated by Figure 2.6 are twofold. The anisotropy favors the convergence towards the equilibrium at the continuous level. But the additional numerical diffusion introduced by the scheme also accelerates this convergence.

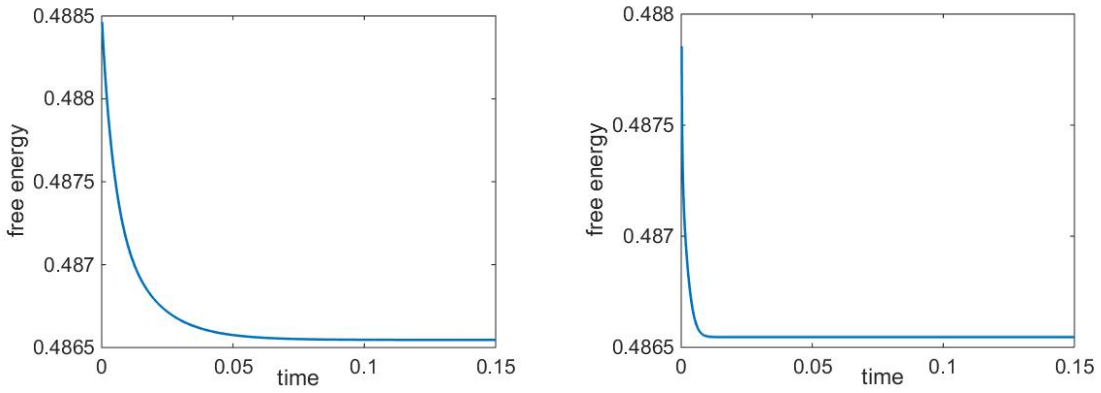


FIGURE 2.6 – Evolution of the free energy along time, for $\Lambda_{xx} = 1, \Lambda_{yy} = 1$ *(on the left)* and $\Lambda_{xx} = 1, \Lambda_{yy} = 100$ *(on the right)*.

2.5.5 Newton-Raphson method and adaptive time stepping

In order to illustrate the behavior of Newton's method, we consider again Test 2 of §2.5.2 solved with the nonlinear scheme (2.64) on the 5^{th} mesh in the anisotropic case $\Lambda_{yy} = 20$. The final time t_f for the simulation is set to 0.1. The maximal time step $\Delta t_{\max}$ mesh is fixed to 0.01, and the initial time step is chosen equal to $\Delta t_{\max}$. We observe on Figure 2.7 that the Newton's method fails to converge a first step. Four successive time step refinements are required. But there is no need to further refine the time step in the next steps. The time step increases until it reaches the maximal value $\Delta t_{\max}$.

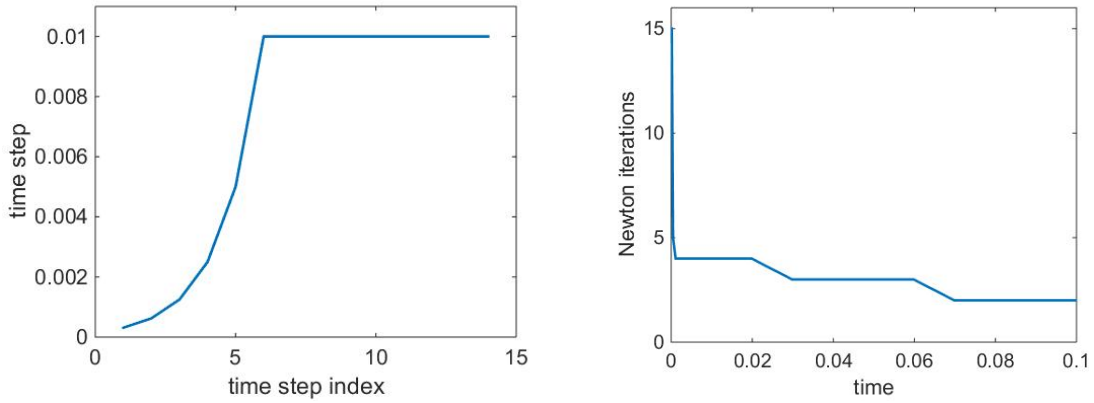


FIGURE 2.7 – Adaptive time step *(left)* and Newton iterations *(right)*.

2.6 Conclusion

We proposed and analyzed a nonlinear energy stable scheme for solving the Richards equation. Moreover, the definition of the scheme only rely on physical quantities and not on artificial ones like for instance the Kirchhoff transform. We were able to carry out a full convergence analysis based on compactness arguments. Contrarily to classical schemes, this new nonlinear scheme produces no undershoot. As far as we know, our scheme is the first one to ensure that the discrete solution remains in the physical range even in the case of strong anisotropy.

However, it appears in the numerical simulations that in the case of strong anisotropy ratio, the scheme introduces an excessive numerical diffusion that makes its convergence very slow. This shall motivate the design of some new more robust schemes (for instance based on [40]) that preserve the main advantages of the scheme studied in this chapter, namely the formulation in physical variables, the preservation of the physical range, and then control of the physical energy.

2.7 Appendix

2.7.1 Some inequalities of Sobolev's type

Lemma 2.7.1. *Let $q \geq 1$, and let $u \in W^{1,q}(\Omega)$ be such that*

$$u \geq 0 \quad \text{and} \quad \lambda\left(\{u = 0\}\right) \geq \alpha > 0, \tag{2.74}$$

where λ denotes the 2-dimensional Lebesgue measure. Define $q^\star = 2q/(2-q)$ if $q < 2$ and $q^\star = +\infty$ if $q \geq 2$, then, for all $r \leq q^\star$ if $q \neq 2$ and $r < \infty$ if $q = 2$, there exists C depending only on Ω, r, and α such that

$$\|u\|_{L^r(\Omega)} \leq C \|\nabla u\|_{L^q(\Omega)^2}.$$

Proof. Define the mean $\langle u \rangle$ value of u by

$$\langle u \rangle = \frac{1}{\lambda(\Omega)} \int_\Omega u(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \geq 0.$$

Due to the properties (2.74) of u, one has

$$\int_\Omega |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} = \int_{\{u=0\}} \langle u \rangle \, \mathrm{d}\boldsymbol{x} + \int_{\{u>0\}} |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} \geq \alpha \langle u \rangle.$$

On the other hand, thanks to Poincaré's inequality (see, e.g., [3]), one has

$$\int_\Omega |u - \langle u \rangle| \, \mathrm{d}\boldsymbol{x} \leq \frac{\mathrm{diam}(\Omega)}{2} \int_\Omega |\nabla u| \, \mathrm{d}\boldsymbol{x} \leq \frac{\mathrm{diam}(\Omega)}{2} \lambda(\Omega)^{\frac{q-1}{q}} \|\nabla u\|_{L^q(\Omega)}.$$

Therefore, we get that

$$\langle u \rangle \leq \frac{\mathrm{diam}(\Omega)}{2\alpha} \lambda(\Omega)^{\frac{q-1}{q}} \|\nabla u\|_{L^q(\Omega)}.$$

Combining this estimate with Sobolev's inequality (see, e.g., [4]) yields

$$\|u\|_{L^r(\Omega)} \leq \|u - \langle u \rangle\|_{L^r(\Omega)} + \lambda(\Omega) \langle u \rangle \leq C \|\nabla u\|_{L^q(\Omega)^d}$$

where C depends only the prescribed quantities. □

In the next Lemma, we prove a discrete Sobolev inequality. Note that the proof takes advantage of the existence of a conformal $V_\mathcal{T}$, leading to a much simpler proof than in [81] or [23].

Lemma 2.7.2. *Let $\mathcal{T}$ and $\mathcal{M}$ be a primal and a dual discretizations of Ω as prescribed in §2.2.1. Let $(u_K)_{K \in \mathcal{V}}$ be an arbitrary element of $\mathbb{R}^{\#\mathcal{V}}$, and denote by*

$$\langle u_\mathcal{M} \rangle = \frac{1}{\lambda(\Omega)} \int_\Omega u_\mathcal{M} \, \mathrm{d}\boldsymbol{x} = \frac{1}{\lambda(\Omega)} \int_\Omega u_\mathcal{T} \, \mathrm{d}\boldsymbol{x}.$$

Then there exists C depending only on r, q Ω, and $\theta_\mathcal{T}$ such that

$$\|u_\mathcal{M} - \langle u_\mathcal{M} \rangle\|_{L^r(\Omega)} \leq C \int_\Omega |\nabla u_\mathcal{T}|^q \, \mathrm{d}\boldsymbol{x}, \qquad \forall r \in [1, \infty), \, \forall q \geq \min\left(1, \frac{2r}{2+r}\right).$$

Proof. Since $u_\mathcal{T}$ is Lipschitz continuous, the classical Sobolev inequality (cf. [4]) gives that

$$\|u_\mathcal{T} - \langle u_\mathcal{M} \rangle\|_{L^r(\Omega)} \leq C \int_\Omega |\nabla u_\mathcal{T}|^q \, \mathrm{d}\boldsymbol{x}, \qquad \forall r \in [1, \infty), \, \forall q \geq \min\left(1, \frac{2r}{2+r}\right).$$

It only remains to use (2.14) to conclude the proof. □

With that discrete Sobolev inequality at hand (cf. Lemma 2.7.2), we can now easily adapt the proof of Lemma 2.7.1 to the discrete setting, leading to the following statement, whose proof is left to the reader.

Lemma 2.7.3. *Let $(v_K)_{K\in\mathcal{V}}$, and let $v_\mathcal{M}$ and $v_\mathcal{T}$ the corresponding discrete functions belonging to $X_\mathcal{M}$ and $V_\mathcal{T}$ respectively. Assume that*

$$v_\mathcal{M} \geq 0 \quad \textit{and} \quad \lambda_d(\{v_\mathcal{M} = 0\}) \geq \alpha > 0,$$

Define $q^\star = qd/(d-q)$ if $q < d$ and $q^\star = +\infty$ if $q \geq d$, then, for all finite *$r \leq q^\star$, there exists C depending only on Ω, $\theta_\mathcal{T}$, r, and α such that*

$$\|v_\mathcal{M}\|_{L^r(\Omega)} \leq C\|\nabla v_\mathcal{T}\|_{L^q(\Omega)^d}.$$

2.7.2 Uniqueness of the weak solution

Proposition 2.7.4. *Under Assumptions (**A1**)–(**A4**), there exists a unique weak solution to the problem* (2.1) *in the sense of Definition 2.1.4.*

Proof. First, define the full Kirchhoff transform $\varphi : \mathbb{R} \to \mathbb{R}$ by

$$\varphi(p) = \int_0^p \eta(s(a))\mathrm{d}a, \qquad \forall p \in \mathbb{R}.$$

It follows from Assumptions (**A1**) and (**A2**) that φ is Lipschitz continuous, increasing, and fulfills $p\varphi(p) > 0$ for all $p \neq 0$. Since η is assumed to be bounded, one has

$$\nabla\varphi(p) = \sqrt{\eta(s(p))}\nabla\xi(p) \in L^2(Q_{t_f})^d$$

for any $p : Q_{t_f} \to \mathbb{R}$ such that $\xi(p) \in L^2((0,T); H^1(\Omega))$ (thus in particular for any weak solution). Therefore, any weak solution p satisfies

$$\iint_{Q_{t_f}} s(p)\partial_t\psi\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega s_0\psi(\cdot,0)\mathrm{d}\boldsymbol{x} + \iint_{Q_{t_f}} (\eta(s(p))\mathbf{g} - \nabla\varphi(p)) \cdot \Lambda\nabla\psi\mathrm{d}\boldsymbol{x}\mathrm{d}t = 0 \tag{2.75}$$

for all $\psi \in C_c^\infty(\overline{\Omega} \times [0, t_f))$. Mimicking Otto's uniqueness proof for degenerate parabolic-elliptic problems [133], we obtain that, given two weak solutions p and $\widehat{p}$ corresponding to the same initial data s_0, one has

$$\int_\Omega |s(p(\boldsymbol{x},t)) - s(\widehat{p}(\boldsymbol{x},t))|\mathrm{d}\boldsymbol{x} \leq 0 \quad \text{for a.e. } t \geq 0, \tag{2.76}$$

hence $s(p) = s(\widehat{p})$. Moreover, the mass being conserved, it follows from Assumption (**A4**) that

$$0 < \int_\Omega s(p(\boldsymbol{x},t))\mathrm{d}\boldsymbol{x} = \int_\Omega s_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \overline{s}_0\,\mathrm{meas}(\Omega) < \mathrm{meas}(\Omega) \quad \text{for a.e. } t \geq 0.$$

Therefore, defining

$$\mathcal{U}(t) := \{\boldsymbol{x} \in \Omega \mid s(p(\cdot,t)) < 1\} \quad \text{for a.e. } t \geq 0,$$

one has

$$\text{meas}\,(\mathcal{U}(t)) \geq (1-\overline{s}_0)\text{meas}(\Omega) > 0 \quad \text{for a.e. } t \geq 0. \tag{2.77}$$

Since s is increasing on $[p_\star, 0]$, one gets that $p(\cdot,t) = \widehat{p}(\cdot,t)$ on $\mathcal{U}(t)$ for a.e. $t \geq 0$.

Subtracting the weak formulation (2.75) corresponding to p to the one for $\widehat{p}$ then yields

$$\iint_{Q_{t_f}} \nabla\Big(\varphi(p) - \varphi(\widehat{p})\Big) \cdot \Lambda \nabla \psi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0, \quad \forall \psi \in C_c^\infty(\overline{\Omega} \times [0, t_f)),$$

and thus for all ψ in $L^2((0,T); H^1(\Omega))$ thanks to a density argument. Choosing $\psi = \varphi(p) - \varphi(\widehat{p})$ and using Assumption (**A3**) yields

$$\|\nabla\Big(\varphi(p(\cdot,t)) - \varphi(\widehat{p}(\cdot,t))\Big)\|_{L^2(\Omega)^d} = 0 \quad \text{for a.e. } t \geq 0.$$

The function $\varphi(p) - \varphi(\widehat{p})$ is identically equal to 0 on $\mathcal{U}(t)$, we can apply Lemma 2.7.1 to infer that

$$\|\varphi(p(\cdot,t)) - \varphi(\widehat{p}(\cdot,t))\|_{L^2(\Omega)} = 0 \quad \text{for a.e. } t \geq 0.$$

Since φ is increasing, one obtains that $p = \widehat{p}$ a.e. in Q_{t_f}. □

Chapitre 3

Numerical analysis of a finite volume scheme for a seawater intrusion model with cross-diffusion in an unconfined aquifer

Abstract : We consider a degenerate parabolic system modeling the flow of fresh and saltwater in a porous medium in the context of seawater intrusion. We propose and analyze a finite volume scheme based on two-point flux approximation with upwind mobilities. The scheme preserves at the discrete level the main features of the continuous problem, namely the nonnegativity of the solutions, the decay of the energy and the control of the entropy and its dissipation. Based on these nonlinear stability results, we show that the scheme converges towards a weak solution to the problem. Numerical results are provided to illustrate the behavior of the model and of the scheme.

3.1 Introduction

3.1.1 Presentation of the continuous problem

We are interested in the study of seawater intrusion problem in coastal regions. If they are densely populated areas, the intensive extraction of freshwater yields to local water table depression and saltwater from the sea can enter the ground and replace the freshwater. This causes sea intrusion problem. In these zones, the optimal exploitation of freshwater and the limitation of seawater intrusion are a challenge for the future. Since freshwater and saltwater are miscible fluids, we have a transition zone separating them caused by hydrodynamic dispersion. In the literature, there exists several modelling approaches. The first approach is to assume that the fluids are immiscible and the domains occupied by each fluid are separated by an interface called sharp interface. Moreover, we consider that no mass transfer occurs between the fresh and the saltwater and the so-called Dupuit approximation. Physically, this approach is not totally correct but enables to follow the saltwater front. We refer to [10, 19, 20, 21, 145, 146] for more details about this first approach. The second approach consists in considering the existence of a transition zone with variable concentrations of salt. It is difficult to tackle this approach from theorical and numerical points of view (see [54, 143, 144]). The third approach is to assume that the fluids are miscible and no interface between these fluids. This modelling approach is physically correct, but it has the disadvantage that it is impossible to follow explicitly the interface (see [55]). The fourth approach is a mixed approach (sharp-diffuse). Recently in [57] the authors derive this model for seawater intrusion phenomena in unconfined aquifer. It combines the efficiency of the sharp interface approach with the physical realism of the diffuse interface one. For mathematical analysis of sharp-diffuse interfaces see [56].

In this chapter, we consider the first approach, by focusing on the seawater intrusion model in an unconfined aquifer, obtained in [108] considering the formal asymptotic limit as the aspect ratio between the thickness and the horizontal length of the porous medium tends to zero. In our setting ξ is a nonnegative function expressing the height of the interface between the saltwater and the freshwater while $h \geq \xi$ is the height of the interface separating the freshwater and the dry soil. We assume that the bottom of the porous medium, which is located at $b \neq 0$, is impermeable, cf. Figure 3.1. Moreover we assume quasi-horizontal displacements (Dupuit approximation), hence we get a 2D-vertically averaged model. This assumption is reasonable when the thickness of the aquifer is small compared to the horizontal length of the aquifer. The evolution of ξ and h is given by the following cross-diffusion system of

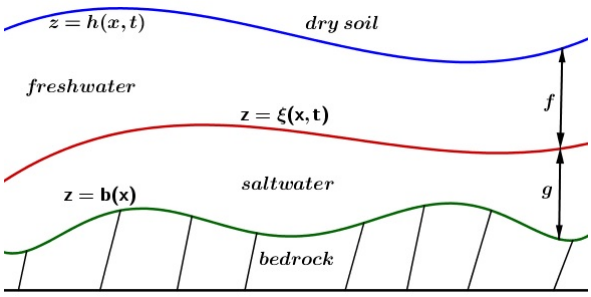


FIGURE 3.1 – Description of an unconfined aquifer

degenerate parabolic equations

$$\begin{cases} \partial_t(h-\xi) - \nabla.\Big((h-\xi)\nabla((1-\varepsilon_0)h)\Big) &= 0 \quad \text{in} \quad \Omega\times(0,T) =: \Omega_T \\ \partial_t\xi - \nabla.\Big((\xi-b)\nabla((1-\varepsilon_0)h+\varepsilon_0\xi)\Big) &= 0 \quad \text{in} \quad \Omega\times(0,T), \end{cases} \tag{3.1}$$

where $\Omega \subset \mathbb{R}^2$ is a polygonal open bounded subset and $T > 0$ is a finite time horizon. The parameter ε_0 is given by

$$\varepsilon_0 = \frac{\rho_s - \rho_f}{\rho_s},$$

where ρ_s (resp. ρ_f) is the mass density of the fluid saltwater (resp. freshwater) (assumed to be constant with $0 < \rho_f < \rho_s$). We set $f = h - \xi,\ g = \xi - b$ and

$$\nu = 1 - \varepsilon_0 = \frac{\rho_f}{\rho_s} \in (0,1). \tag{3.2}$$

The system (4.1) then rewrites :

$$\begin{cases} \partial_t f - \nabla.\Big(\nu f\nabla(f+g+b)\Big) &= 0 \quad \text{in} \quad \Omega\times(0,T), \\ \partial_t g - \nabla.\Big(g\nabla(\nu f+g+b)\Big) &= 0 \quad \text{in} \quad \Omega\times(0,T). \end{cases} \tag{3.3}$$

It is supplemented with no-flux boundary conditions

$$\nabla f\cdot\mathbf{n} = \nabla g\cdot\mathbf{n} = 0, \quad \text{on} \ \ \partial\Omega\times(0,T), \tag{3.4}$$

where $\mathbf{n}$ is the unit normal to the boundary $\partial\Omega$. Initial data are

$$f_{|t=0} = f_0, \quad g_{|t=0} = g_0, \tag{3.5}$$

with $f_0, g_0 \in L^\infty(\Omega)$ and

$$f_0, g_0 \geq 0, \quad \text{a.e } x \in \Omega. \tag{3.6}$$

Before describing more precisely our results, let us mention that the problems of kind (4.9), have been the object of several studies. The authors in [73, 74] studied

the classical solutions of the system (4.9) (with $b = 0$). Moreover, weak solutions are established under different assumptions in [72, 119, 120].

We recall (see [72, 119, 120]) the definition of entropy functional :

$$\mathfrak{H}(f,g) = \int_\Omega \left[\Gamma(g) + \frac{1}{\nu}\Gamma(f)\right] \mathrm{d}\mathbf{x}, \quad \text{where } \Gamma(s) = s\log s - s + 1,$$

and of the energy functional :

$$\mathfrak{E}(f,g) = \int_\Omega \left[\frac{\nu}{2}(f+g+b)^2 + \frac{1-\nu}{2}(g+b)^2\right] \mathrm{d}\mathbf{x}.$$

Multiplying (formally) the first equation of (4.9) by $\frac{1}{\nu}\log f$ and the second equation by $\log g$, integrating over Ω and summing both equations, yields the classical entropy/dissipation property :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{H}(f,g) + \frac{1-\nu}{2}\int_\Omega \left[(\nabla f)^2 + (\nabla g)^2\right] \mathrm{d}\mathbf{x} \leq \frac{1}{2(\nu+1)}\int_\Omega (\nabla b)^2 \,\mathrm{d}\mathbf{x}. \tag{3.7}$$

Moreover multiplying (formally) the first equation of (4.9) by $\nu(f+g+b)$ and the second equation by $\nu f + g + b$, integrating over Ω and summing both equations, yields that the energy functional decreases along time :

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathfrak{E}(f,g) + \int_\Omega \left[\nu^2 f(\nabla(f+g+b))^2 + g(\nabla(\nu f+g+b))^2\right] \mathrm{d}\mathbf{x} = 0. \tag{3.8}$$

Let us mention that the cross-diffusion systems are extensively presented in different domain as ecology, biology, chemistry and others. In [89] the author propose and analyze a finite volume scheme for the Patlak-Keller-Segel (PKS) chemotaxis model. In [24] the authors studie the PKS model with additional cross diffusion. We refer to [13] for the analysis of a finite volume method for a cross diffusion model in population dynamics. See [124, 131] for the numerical analysis for a seawater intrusion problem in a unconfined aquifer with finite element method approximation. In [2] the authors propose a finite element method and a finite volume method and compare the results given by these two methods. In [58] the authors address the question of global existence for the sharp interface approach. For an analysis of a finite volume scheme for two-phase immiscible flow in porous media, used in petroleum engineering, we can refer to these papers [86, 126].

In this work, we propose a finite volume scheme for the problem (4.9). This scheme is based on a two-point flux approximation with upwind mobilities. It is designed in order to preserve at the discrete level the main features of the continuous problem : the nonnegativity of the solutions, the decay of the energy (3.8), the control of the entropy and its dissipation (3.7).

3.1.2 The numerical scheme

In this section, we explicit the discretization of the problem (4.9)-(3.4) we will study in this chapter. The time discretization relies on backward Euler scheme, while the space discretization relies on a finite volume approach (see e.g [78]), with two-point flux approximation and upstream mobility.

Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, polygonal subset. An admissible mesh of Ω is given by a family $\mathcal{T}$ of a control volumes (open and convex polygons), a family $\mathcal{E}$ of edges and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 9.1 in [78]. This definition implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$.

We distinguish the interior edges $\sigma \in \mathcal{E}_{\text{int}}$ and the boundary edges $\sigma \in \mathcal{E}_{\text{ext}}$. The set of edges $\mathcal{E}$ equals the union $\mathcal{E}_{\text{int}} \cup \mathcal{E}_{\text{ext}}$. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, by $\mathcal{E}_{K,\text{int}}$ the set of its interior edges and by $\mathcal{E}_{K,\text{ext}}$ the set of edges of K included in $\partial\Omega$.

Furthermore, we denote by d the distance in $\mathbb{R}^2$ and by m the Lebesgue measure in $\mathbb{R}^2$ or $\mathbb{R}$. We assume that the family of meshes satisfies the following regularity requirement : there exists $\zeta > 0$ such that for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{\text{int},K}$ with $\sigma = K|L$, it holds

$$d(x_K, \sigma) \geq \zeta d(x_K, x_L) \tag{3.9}$$

For all $\sigma \in \mathcal{E}_{\text{int},K}$ with $\sigma = K|L$, we define $d_\sigma = d(x_K, x_L)$ and the transmissibility coefficient

$$\tau_\sigma = \frac{m(\sigma)}{d_\sigma}, \quad \sigma \in \mathcal{E}. \tag{3.10}$$

The size of the mesh is defined by

$$\delta = \max_{K\in\mathcal{T}}(\text{diam}(K)).$$

In order to avoid heavier notations, we restrict our study to the case of a uniform time discretization of $(0, T)$. However, all the results presented in this chapter can be extended to general time discretizations without any technical difficulty. In what follows, we assume that the spatial mesh is fixed and does not change with the time step. Let $T > 0$ be some final time and M_T the number of time steps. Then the time step size and the time points are given by, respectively,

$$\Delta t = \frac{T}{M_T}, \quad t^n = n\Delta t, \quad 0 \leq n \leq M_T.$$

We denote by $\mathcal{D}$ an admissible space-time discretization of $\Omega_T = \Omega \times (0, T)$ composed of an admissible mesh $\mathcal{T}$ of Ω and the values Δt and M_T. The size of this space-time discretization $\mathcal{D}$ is defined by $\eta = \max(\delta, \Delta t)$.

The initial conditions are discretized by

$$f_{\mathcal{T}}^0 = \sum_{K\in\mathcal{T}} f_K^0 \mathbf{1}_K, \quad \text{where } f_K^0 = \frac{1}{m(K)}\int_K f_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T}, \tag{3.11}$$

$$g_{\mathcal{T}}^0 = \sum_{K\in\mathcal{T}} g_K^0 \mathbf{1}_K, \quad \text{where } g_K^0 = \frac{1}{m(K)}\int_K g_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T}, \tag{3.12}$$

and $\mathbf{1}_K$ is the characteristic function on K. Denoting by f_K^n and g_K^n approximations of the mean value of $f(.,t^n)$ and $g(.,t^n)$ on K, respectively. Taking for b_K the value of b in a fixed point of K (for instance, the center of gravity of K), where b is a regular function and assume that $b_K \geq 0 \quad \forall K \in \mathcal{T}$. The discretization of problem (4.9) is given by the following set of nonlinear equations :

$$m(K)\frac{f_K^{n+1}-f_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\text{int}},K} \tau_\sigma f_\sigma^{n+1}\nu\left((f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\right) = 0, \tag{3.13}$$

and

$$m(K)\frac{g_K^{n+1}-g_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\text{int}},K} \tau_\sigma g_\sigma^{n+1}\left(\nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\right) = 0, \tag{3.14}$$

for $K \in \mathcal{T}$ and $0 \leq n \leq M_T - 1$, where

$$f_\sigma^{n+1} = \begin{cases} (f_K^{n+1})^+ & \text{if } (f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L) \geq 0, \\ (f_L^{n+1})^+ & \text{if } (f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L) < 0. \end{cases} \tag{3.15}$$

and

$$g_\sigma^{n+1} = \begin{cases} (g_K^{n+1})^+ & \text{if } \nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L) \geq 0, \\ (g_L^{n+1})^+ & \text{if } \nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L) < 0, \end{cases} \tag{3.16}$$

where $x^+ = \max(0,x)$. We next define the numerical approximation $(f_{\mathcal{D}}, g_{\mathcal{D}})$ of (f,g) by

$$f_{\mathcal{D}}(x,t) = \sum_{\substack{K\in\mathcal{T} \\ 0\leq n\leq M_T-1}} f_K^{n+1}\mathbf{1}_{K\times(t^n,t^{n+1}]}(x,t), \quad \text{and}$$

$$g_{\mathcal{D}}(x,t) = \sum_{\substack{K\in\mathcal{T} \\ 0\leq n\leq M_T-1}} g_K^{n+1}\mathbf{1}_{K\times(t^n,t^{n+1}]}(x,t).$$

We also define approximations of the gradients $\nabla^{\mathcal{D}} f_{\mathcal{D}}$ and $\nabla^{\mathcal{D}} g_{\mathcal{D}}$ of f and g, respectively. To this end, we introduce that : for $K \in \mathcal{T}$

- If $\sigma = K|L \in \mathcal{E}_{\text{int},K}$, $\mathfrak{D}_{K,L}$ is the cell ("diamond") whose vertices are given by x_K, x_L and the end points of the edge $\sigma = K|L$.
- $\mathfrak{D}_{K,\sigma} = \mathfrak{D}_{K,L} \cap K$ is the cell ("triangle") whose vertices are given by x_K and the end points of the edge $\sigma = K|L$.

The approximate gradient $\nabla^{\mathcal{D}} S_{\mathcal{D}}$ (with $S = f$, or $S = g$) is a piecewise constant function, defined in Ω_T by

$$\nabla^{\mathcal{D}} S_{\mathcal{D}}(x,t) = -\frac{m(\sigma)}{m(\mathfrak{D}_{K,L})}(S_K^{n+1} - S_L^{n+1})\nu_{K,L}, \quad x \in \mathfrak{D}_{K,L},\ t \in (t^n, t^{n+1}], \tag{3.17}$$

where $\nu_{K,L}$ is the unit vector normal to σ and outward to K.

3.1.3 Main results and outline of the chapter

The scheme (4.40)-(4.43) amounts to a nonlinear system to be solved at each time step. The existence of a solution to this system is therefore non trivial. The first result we highlight is thus the existence of a nonnegative solution to the scheme (4.40)-(4.43), the stability in terms of the discrete entropy and the decay of the discrete energy.

Theorem 3.1.1. *There exists (at least) one solution* $\left(f_K^{n+1}, g_K^{n+1}\right)_{K\in\mathcal{T}, n\in\{0,\dots,M_T-1\}}$ *to the scheme (4.40)-(4.43). Moreover,* $f_K^n \geq 0$, $g_K^n \geq 0$ *for all* $K \in \mathcal{T}$ *and for all* $n \in \{0,\dots,M_T\}$ *and there exists* C *depending only on* Ω, f_0, g_0, ν *and* b *such that*

$$\sup_{n\in\{0,\dots,M_T-1\}} \mathfrak{H}^{n+1} + \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma \Big[(f_K^{n+1} - f_L^{n+1})^2 + (g_K^{n+1} - g_L^{n+1})^2\Big] \leq C(1+T),$$

and

$$\begin{aligned}
&\sup_{n\in\{0,\dots,M_T-1\}} \mathfrak{E}^{n+1} \\
&\quad + \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1} \nu^2 \Big((f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2 \\
&\quad + \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma g_\sigma^{n+1} \Big(\nu(f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2 \leq C.
\end{aligned}$$

Our second result concerns the convergence of the scheme to a weak solution of (4.9)-(3.4). Let $(\mathcal{D}_m)_{m>0}$ be a family of admissible space-time discretization of Ω_T. We denote by $(\mathcal{T}_m)_{m>0}$ the corresponding meshes of Ω, with $\text{size}(\mathcal{T}_m) = \delta_m \longrightarrow 0$, as $m \to 0$. We define $(f_m, g_m) := (f_{\mathcal{D}_m}, g_{\mathcal{D}_m})$ the sequence of approximate solutions constructed on the discretization $\mathcal{D}_m$. We set $\nabla^m := \nabla^{\mathcal{D}_m}$.

Theorem 3.1.2. *Let $(\mathcal{D}_m)_{m>0}$ be a sequence of admissible discretizations satisfying (4.36) uniformly in m and $\lim_{m\to+\infty}\eta_m = 0$. Let (f_m, g_m) be a sequence of finite volume solutions to (4.40)-(4.43). Then there exists (f, g) such that, up a subsequence,*

$$\begin{aligned} f_m &\longrightarrow f \quad \text{in } L^r(\Omega_T), \forall r < 4, \text{ and } \nabla^m f_m \longrightarrow \nabla f \quad \text{weakly in } L^2(\Omega_T)^2, \\ g_m &\longrightarrow g \quad \text{in } L^r(\Omega_T), \forall r < 4, \text{ and } \nabla^m g_m \longrightarrow \nabla g \quad \text{weakly in } L^2(\Omega_T)^2, \end{aligned}$$

and $(f, g) \in L^2(0, T; H^1(\Omega))^2$ is a weak solution to (4.9)-(3.4) in the following sense

$$\int_0^T \int_\Omega (f\partial_t\psi - \nu f\nabla(f + g + b)\cdot\nabla\psi)\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega f_0\psi(.,0)\,\mathrm{d}x = 0, \tag{3.18}$$

$$\int_0^T \int_\Omega (g\partial_t\psi - g\nabla(\nu f + g + b)\cdot\nabla\psi)\,\mathrm{d}x\,\mathrm{d}t + \int_\Omega g_0\psi(.,0)\,\mathrm{d}x = 0, \tag{3.19}$$

for all test functions $\psi \in C_0^\infty(\Omega \times [0, T))$.

The chapter is organized as follows. The existence of nonnegative solution is shown in Section 3.2. Discrete counterparts of the entropy/entropy-dissipation (3.7) and energy/energy-dissipation (3.8) relations are established in Section 3.3. Section 3.4 is devoted to the convergence proof of the scheme. This proof is based first on the compactness of the sequence of approximate solutions and then on the identification of the limit. We finally present numerical experiments in Section 3.5, to illustrate the behaviour of the model and of the scheme.

3.2 Existence of a nonnegative discrete solutions

First all, we prove the positivity of the discrete solutions. This estimate allows to prove the existence of a solution to the nonlinear system (4.40)-(4.43).

Proposition 3.2.1. *For all $K \in \mathcal{T}, n \geq 0$,*

$$f_K^n \geq 0, \quad g_K^n \geq 0, \tag{3.20}$$

hence

$$\sum_{K\in\mathcal{T}} m(K) f_K^n = \sum_{K\in\mathcal{T}} m(K) f_K^0 = \|f_0\|_{L^1(\Omega)}, \tag{3.21}$$

and

$$\sum_{K\in\mathcal{T}} m(K) g_K^n = \sum_{K\in\mathcal{T}} m(K) g_K^0 = \|g_0\|_{L^1(\Omega)}. \tag{3.22}$$

Proof. The property (3.20) clearly holds for $n = 0$ thanks to (3.6). Assume now that (3.20) holds at time step n and assume that

$$f_K^{n+1} < 0, \quad \text{for some } K \in \mathcal{T}.$$

In view of the definition (4.42) of f_σ^{n+1} one has that

$$f_K^{n+1} = -\frac{\nu\Delta t}{m(K)}\sum_{\sigma\in\mathcal{E}_{\rm int},K}\tau_\sigma\Bigg(\underbrace{(f_K^{n+1})^+}_{=0}\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\Big]^+ \\ -(f_L^{n+1})^+\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\Big]^-\Bigg)+f_K^n\geq 0,$$

yielding a contradiction, ensuring that

$$f_K^{n+1}\geq 0,\quad \forall K\in\mathcal{T},\forall n\geq 0.$$

Proving that $g_K^n\geq 0$ for all $K\in\mathcal{T},\forall n\geq 0$, is similar. □

We will prove the existence of a solution, we follow the methodology proposed in [75], using a topological degree argument [61, 122].

Proposition 3.2.2. *Let $\mathcal{D}$ be an admissible discretization of $\Omega\times(0,T)$. There exists (at least) one solution to the scheme (4.40)-(4.43).*

Proof. Let $\mu\in[0,1]$ and define $(f_{K,\mu}^{n+1},g_{K,\mu}^{n+1})_K$ as the solution of the scheme : $\forall K\in\mathcal{T}$

$$m(K)\frac{f_{K,\mu}^{n+1}-f_K^n}{\Delta t} \\ +\mu\sum_{\sigma\in\mathcal{E}_{\rm int},K}\tau_\sigma f_{\sigma,\mu}^{n+1}\nu\Big((f_{K,\mu}^{n+1}-f_{L,\mu}^{n+1})+(g_{K,\mu}^{n+1}-g_{L,\mu}^{n+1})+(b_K-b_L)\Big)=0,$$

$$m(K)\frac{g_{K,\mu}^{n+1}-g_K^n}{\Delta t} \\ +\mu\sum_{\sigma\in\mathcal{E}_{\rm int},K}\tau_\sigma g_{\sigma,\mu}^{n+1}\Big(\nu(f_{K,\mu}^{n+1}-f_{L,\mu}^{n+1})+(g_{K,\mu}^{n+1}-g_{L,\mu}^{n+1})+(b_K-b_L)\Big)=0.$$

Reproducing the proof of Proposition 3.2.1, one can show that

$$f_{K,\mu}^{n+1}\geq 0,\ \text{ and }\ g_{K,\mu}^{n+1}\geq 0\quad\forall\mu\in[0,1],$$

hence

$$\sum_{K\in\mathcal{T}}m(K)f_{K,\mu}^{n+1}=\sum_{K\in\mathcal{T}}m(K)f_K^0=\|f_0\|_{L^1(\Omega)},$$

and

$$\sum_{K\in\mathcal{T}}m(K)g_{K,\mu}^{n+1}=\sum_{K\in\mathcal{T}}m(K)g_K^0=\|g_0\|_{L^1(\Omega)}.$$

Therefore, for all $K\in\mathcal{T}$, one has

$$0\leq f_{K,\mu}^{n+1}\leq\frac{\|f_0\|_{L^1(\Omega)}}{m(K)}\leq\frac{\|f_0\|_{L^1(\Omega)}}{\min\limits_{K\in\mathcal{T}}m(K)}:=m_f,\tag{3.23}$$

and

$$0 \leq g_{K,\mu}^{n+1} \leq \frac{\|g_0\|_{L^1(\Omega)}}{m(K)} \leq \frac{\|g_0\|_{L^1(\Omega)}}{\min_{K\in\mathcal{T}} m(K)} := m_g. \tag{3.24}$$

Define the compact subset $\mathcal{K} = [-1, m_f + 1]^{\#\mathcal{T}} \times [-1, m_g + 1]^{\#\mathcal{T}}$ of $\mathbb{R}^{\#\mathcal{T}} \times \mathbb{R}^{\#\mathcal{T}}$ and define the function $\mathcal{H}((f_K, g_K)_K, \mu) : \mathbb{R}^{\#\mathcal{T}} \times \mathbb{R}^{\#\mathcal{T}} \times [0,1] \to \mathbb{R}^{\#\mathcal{T}} \times \mathbb{R}^{\#\mathcal{T}}$ by : $\forall K \in \mathcal{T}$,

$$\begin{aligned}
\mathcal{H}((f_K, g_K)_K, \mu) = \Bigg(& m(K)\frac{f_{K,\mu}^{n+1} - f_K^n}{\Delta t} \\
& + \mu \sum_{\sigma\in\mathcal{E}_{\rm int},K} \tau_\sigma f_{\sigma,\mu}^{n+1} \nu \Big((f_{K,\mu}^{n+1} - f_{L,\mu}^{n+1}) + (g_{K,\mu}^{n+1} - g_{L,\mu}^{n+1}) + (b_K - b_L) \Big), \\
m(K)\frac{g_{K,\mu}^{n+1} - g_K^n}{\Delta t} + \mu \sum_{\sigma\in\mathcal{E}_{\rm int},K} & \tau_\sigma g_{\sigma,\mu}^{n+1} \Big(\nu (f_{K,\mu}^{n+1} - f_{L,\mu}^{n+1}) + (g_{K,\mu}^{n+1} - g_{L,\mu}^{n+1}) + (b_K - b_L) \Big) \Bigg).
\end{aligned}$$

The function $\mathcal{H}$ is uniformly continuous on $\mathcal{K} \times [0,1]$ and it follows from (3.23) that for all $\mu \in [0,1]$, the nonlinear system

$$\mathcal{H}((f_K, g_K)_K, \mu) = (0,0), \tag{3.25}$$

cannot admit any solution on $\partial\mathcal{K}$. Therefore, the corresponding topological degree $\delta(\mathcal{H}, \mathcal{K})(\mu)$ is constant w.r.t μ. For $\mu = 0$ the linear system $\mathcal{H}((f_K, g_K)_K, 0)$ admits a unique solution and the topological degree is equal to 1. Hence, the nonlinear system (3.25) admits at least one solution for $\mu = 1$, ensuring the existence of a solution to the scheme (4.40)-(4.43). □

3.3 Entropy and energy estimates

The goal of this section is to establish discrete counterparts to the entropy/entropy-dissipation estimate (3.7) and energy/energy-dissipation estimate (3.8). In what follows, (f_K^n, g_K^n) denotes a solution to the scheme (4.40)-(4.43). The proof of Theorem 3.1.1 is based on suitable estimates, which are shown below. This section also contains some results that will be useful in the convergence proof of Section 3.4.

3.3.1 Discrete $L^2(0,T; H^1(\Omega))$ semi-norm

We first have to define the space $\mathcal{X}(\mathcal{D})$ the solution belongs to and the discrete $L^2(0,T; H^1(\Omega))$ semi-norm.

Definition 3.3.1. *We denote by $\mathcal{X}(\mathcal{D})$ the functional space :*

$$\mathcal{X}(\mathcal{D}) = \left\{ \begin{array}{c} u \in L^\infty(\Omega_T)/u \text{ is constant on } K \times (t^n, t^{n+1}] \\ \forall K \in \mathcal{T}, \forall n \in \{0, \cdots, M_T - 1\} \end{array} \right\}.$$

Definition 3.3.2. *(**Discrete** $L^2(0,T;H^1(\Omega))$ **semi-norm**) We define the discrete $L^2(0,T;H^1(\Omega))$ semi-norm on $\mathcal{X}(\mathcal{D})$ by :*

$$|u|_{1,\mathcal{D}} = \left(\sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma (u_K^{n+1} - u_L^{n+1})^2 \right)^{1/2}.$$

Remark 3.3.3. *Note that*

$$\|\nabla^{\mathcal{D}} u\|_{L^2(\Omega_T)} = \sqrt{2}|u|_{1,\mathcal{D}}. \tag{3.26}$$

3.3.2 Entropy estimate

We introduce a discrete version of entropy functional :

$$\mathfrak{H}^n := \mathfrak{H}(f_K^n, g_K^n) = \sum_{K\in\mathcal{T}} m(K)\Big(\frac{1}{\nu}\Gamma(f_K^n) + \Gamma(g_K^n)\Big).$$

Proposition 3.3.4. *(Entropy stability) For all $n \in \{0,...,M_T-1\}$, one has*

$$\begin{aligned}\mathfrak{H}^{n+1} - \mathfrak{H}^n + \frac{1-\nu}{2}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma\Big[(f_K^{n+1} - f_L^{n+1})^2 + (g_K^{n+1} - g_L^{n+1})^2\Big]\\ \leq \frac{\Delta t}{2(\nu+1)} \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma (b_K - b_L)^2.\end{aligned} \tag{3.27}$$

Proof. We multiply (4.40) by $\Delta t \dfrac{\log f_K^{n+1}}{\nu}$ and summing over $K \in \mathcal{T}$ and (4.41) by $\Delta t \log g_K^{n+1}$ and summing over $K \in \mathcal{T}$, provides that :

$$A + B + C = 0,$$

where

$$\begin{aligned}
A &= \sum_{K\in\mathcal{T}} m(K)\Big[\frac{1}{\nu}(f_K^{n+1} - f_K^n)\log f_K^{n+1} + (g_K^{n+1} - g_K^n)\log g_K^{n+1}\Big],\\
B &= \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} \tau_\sigma f_\sigma^{n+1}\Big((f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)\log f_K^{n+1},\\
C &= \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} \tau_\sigma g_\sigma^{n+1}\Big(\nu(f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)\log g_K^{n+1}.
\end{aligned}$$

By the convexity of Γ, we find that

$$\mathfrak{H}^{n+1} - \mathfrak{H}^n = \sum_{K\in\mathcal{T}} m(K)\Big[\frac{1}{\nu}(\Gamma(f_K^{n+1}) - \Gamma(f_K^n)) + \Gamma(g_K^{n+1}) - \Gamma(g_K^n)\Big] \leq A.$$

We can rewrite B and C as :

$$B = \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}\Big((f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\Big)\times \\ (\log f_K^{n+1}-\log f_L^{n+1}),$$

$$C = \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma g_\sigma^{n+1}\Big(\nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\Big)\times \\ (\log g_K^{n+1}-\log g_L^{n+1}).$$

It follows from the convexity of exp that

$$a(\log a-\log b)\geq a-b\geq b(\log a-\log b)\quad \forall a,b\in[0,+\infty[,$$

where we have used the convention $\log(0)=-\infty$ and $0\log(0)=0$. Hence, in view of the definition (4.42) and (4.43) of the upwind mobilities, one has

$$B \geq \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma\Big((f_K^{n+1}-f_L^{n+1})^2+(g_K^{n+1}-g_L^{n+1})(f_K^{n+1}-f_L^{n+1}) \\ +(b_K-b_L)(f_K^{n+1}-f_K^{n+1})\Big),$$

$$C \geq \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma\Big((g_K^{n+1}-g_L^{n+1})^2+\nu(g_K^{n+1}-g_L^{n+1})(f_K^{n+1}-f_L^{n+1}) \\ +(b_K-b_L)(g_K^{n+1}-g_L^{n+1})\Big).$$

Combining these inequalities, one deduces that

$$\mathfrak{H}^{n+1}-\mathfrak{H}^n+\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma(f_K^{n+1}-f_L^{n+1})^2+\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma(g_K^{n+1}-g_L^{n+1})^2 \\ +(\nu+1)\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma(f_K^{n+1}-f_L^{n+1})(g_K^{n+1}-g_L^{n+1})\leq D,$$

where

$$D=-\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma(b_K-b_L)\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})\Big].$$

Using the Young inequality, one has

$$D\leq\frac{1}{2\epsilon}\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma(b_K-b_L)^2+\frac{\epsilon}{2}\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})\Big]^2,$$

for all $\epsilon > 0$. We choose $\epsilon = 1 + \nu$, we have

$$D \leq \frac{1}{2(\nu+1)}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (b_K - b_L)^2 + \frac{\nu+1}{2}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma \Big[(f_K^{n+1} - f_L^{n+1})^2 + (g_K^{n+1} - g_L^{n+1})^2\Big] + (\nu+1)\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma \Big[(f_K^{n+1} - f_L^{n+1})(g_K^{n+1} - g_L^{n+1})\Big].$$

Finally, one has

$$\mathfrak{H}^{n+1} - \mathfrak{H}^n + \frac{1-\nu}{2}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (f_K^{n+1} - f_L^{n+1})^2 + \frac{1-\nu}{2}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (g_K^{n+1} - g_L^{n+1})^2 \leq \frac{1}{2(\nu+1)}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (b_K - b_L)^2,$$

where $1 - \nu > 0$ thanks to (3.2). □

Corollary 3.3.5. *There exists C_{11} depending only on T, Ω, f_0, g_0, ν and b such that*

$$\mathfrak{H}^{M_T} + \frac{1-\nu}{2}\Big(|f_\mathcal{D}|_{1,\mathcal{D}}^2 + |g_\mathcal{D}|_{1,\mathcal{D}}^2\Big) \leq C_{11}. \tag{3.28}$$

Proof. Summing (3.27) over $n = 0, ..., M_T - 1$ provides

$$\mathfrak{H}^{M_T} + \frac{1-\nu}{2}\sum_{n=0}^{M_T-1}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma \Big[(f_K^{n+1} - f_L^{n+1})^2 + (g_K^{n+1} - g_L^{n+1})^2\Big] \leq \mathfrak{H}^0 + \frac{1}{2(\nu+1)}\sum_{n=0}^{M_T-1}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (b_K - b_L)^2.$$

As a consequence of Jensen's inequality, one has

$$\mathfrak{H}^0 = \int_\Omega \left[\frac{1}{\nu}\Gamma(f_K^0) + \Gamma(g_K^0)\right] \mathrm{d}x \leq \int_\Omega \left[\frac{1}{\nu}\Gamma(f_0) + \Gamma(g_0)\right] \mathrm{d}x < +\infty,$$

and

$$\sum_{n=0}^{M_T-1}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (b_K - b_L)^2 \leq T|\Omega|\|\nabla b\|_\infty^2,$$

concluding the proof of Corollary 3.3.5. □

We obtain immediately thanks to (3.26), the following discrete $L^2(\Omega_T)$ estimate on the discrete gradients :

$$\|\nabla^\mathcal{D} f_\mathcal{D}\|_{L^2(\Omega_T)}^2 + \|\nabla^\mathcal{D} g_\mathcal{D}\|_{L^2(\Omega_T)}^2 \leq 2C_{11}. \tag{3.29}$$

3.3.3 Energy estimate

The current subsection is devoted to the proof of the discrete energy estimate. We introduce a discrete version of energy functional :

$$\mathfrak{E}^n := \mathfrak{E}(f_K^n, g_K^n) = \sum_{K\in\mathcal{T}} m(K)\Big(\frac{\nu}{2}(f_K^n + g_K^n + b_K)^2 + \frac{1-\nu}{2}(g_K^n + b_K)^2\Big).$$

Proposition 3.3.6. *For all $n \in \{0, ..., M_T - 1\}$, one has*

$$\mathfrak{E}^{n+1} + \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}\nu^2\Big((f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2$$
$$+ \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma g_\sigma^{n+1}\Big(\nu(f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2 \le \mathfrak{E}^n. \quad (3.30)$$

Proof. We multiply (4.40) (resp. (4.41)) by $\Delta t\nu(f_K^{n+1}+g_K^{n+1}+b_K)$ (resp. $\Delta t(\nu f_K^{n+1} + g_K^{n+1} + b_K)$) and sum over $K \in \mathcal{T}$. Summing both equalities and reorganizing the sums, we get $A + B = 0$, where

$$A = \sum_{K\in\mathcal{T}} m(K)\Big[\nu(f_K^{n+1} - f_K^n)(f_K^{n+1} + g_K^{n+1} + b_K)\Big]$$
$$+ \sum_{K\in\mathcal{T}} m(K)\Big[(g_K^{n+1} - g_K^n)(\nu f_K^{n+1} + g_K^{n+1} + b_K)\Big],$$

$$B = \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}\nu^2\Big((f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2$$
$$+ \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma g_\sigma^{n+1}\Big(\nu(f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big)^2.$$

One has

$$A = \sum_{K\in\mathcal{T}} m(K)\Big[\nu\Big((f_K^{n+1} + g_K^{n+1} + b_K) - (f_K^n + g_K^n + b_K)\Big)\Big(f_K^{n+1} + g_K^{n+1} + b_K\Big)\Big]$$
$$+ \sum_{K\in\mathcal{T}} m(K)\Big[(1-\nu)\Big((g_K^{n+1} + b_K) - (g_K^n + b_K)\Big)\Big(g_K^{n+1} + b_K\Big)\Big].$$

We use the following inequality : $(a-b)a \ge \frac{1}{2}(a^2 - b^2)$, $\forall a, b \in \mathbb{R}$, to get

$$A \ge \sum_{K\in\mathcal{T}} m(K)\Big[\frac{\nu}{2}\Big((f_K^{n+1} + g_K^{n+1} + b_K)^2 - (f_K^n + g_K^n + b_K)^2\Big)\Big]$$
$$+ \sum_{K\in\mathcal{T}} m(K)\Big[\frac{1-\nu}{2}\Big((g_K^{n+1} + b_K)^2 - (g_K^n + b_K)^2\Big)\Big] = \mathfrak{E}^{n+1} - \mathfrak{E}^n.$$

□

Remark 3.3.7. *In the discrete counterpart (3.30) of (3.8), the equality is remplaced by an inequality. But as well as in the continuous setting, the function* $\mathfrak{E}$ *decreases along time.*

Corollary 3.3.8. *There exists* C_{12} *depending only on* f_0, g_0, b, ν *and* Ω *such that*

$$\|f_{\mathcal{D}}\|_{L^\infty(0,T;L^2(\Omega))} + \|g_{\mathcal{D}}\|_{L^\infty(0,T;L^2(\Omega))} \leq C_{12}.$$

Proof. Summing (3.30) over $n = 0, ..., M_T - 1$, we obtain immediately thanks to the positivity of f_σ^{n+1} and g_σ^{n+1} that

$$\mathfrak{E}^n \leq \mathfrak{E}^0.$$

Using the Cauchy-Schwarz inequality, we obtain

$$\mathfrak{E}^0 \leq \sum_{K\in\mathcal{T}} m(K)\left(\frac{\nu+\nu^2}{2}(f_K^0)^2 + 2(g_K^0)^2 + 2b_K^2\right).$$

Moreover, for $s = f$, or g one has

$$\sum_{K\in\mathcal{T}} m(K)(s_K^0)^2 \leq \|s_0\|^2_{L^2(\Omega)}, \quad \text{and} \quad \sum_{K\in\mathcal{T}} m(K)b_K^2 \leq |\Omega|\|b\|^2_\infty.$$

Hence

$$\mathfrak{E}^0 \leq \|f_0\|^2_{L^2(\Omega)} + 2\|g_0\|^2_{L^2(\Omega)} + 2|\Omega|\|b\|^2_\infty < +\infty.$$

On the other hand since b_K, f_K and g_K are nonnegative for all $K \in \mathcal{T}$, then we have

$$\mathfrak{E}^n \geq \sum_{K\in\mathcal{T}} m(K)\frac{\nu}{2}\Big[(f_K^n)^2 + (g_K^n)^2\Big].$$

We deduce that

$$\sum_{K\in\mathcal{T}} m(K)\Big[(f_K^n)^2 + (g_K^n)^2\Big] \leq \frac{2}{\nu}\mathfrak{E}^n \leq \frac{2}{\nu}\mathfrak{E}^0,$$

concluding the proof of Corollary 3.3.8. □

Then we deduce from Proposition 3.3.6 that

Corollary 3.3.9. *There exists* C_{13} *depending only on* f_0, g_0, b, ν *and* Ω *such that*

$$\sum_{n=0}^{M_T-1}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}\nu^2\Big((f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L)\Big)^2 +$$
$$\sum_{n=0}^{M_T-1}\Delta t \sum_{\substack{\sigma\in\mathcal{E}_{int}\\ \sigma=K|L}} \tau_\sigma g_\sigma^{n+1}\Big(\nu(f_K^{n+1}-f_L^{n+1}) + (g_K^{n+1}-g_L^{n+1}) + (b_K-b_L)\Big)^2 \leq C_{13}.$$

3.4 Convergence analysis

This section is devoted to the compactness of the approximate solution. Our goal is to show the strong compactness of the sequences $(f_m)_{m>0}$ and $(g_m)_{m>0}$ in $L^2(\Omega_T)$ and the weak compactness in $L^2(\Omega_T)^2$ of the approximate gradient of f_m and g_m defined by (3.17). As a first step, we show in §3.4.1 the appropriate compactness properties on the reconstructed discrete solutions. Then we identify in §3.4.2 the limit value (whose existence is ensured thanks to the compactness properties) as the weak solution to the problem (4.9).

3.4.1 Compactness properties of discrete solutions

As it is classical for unsteady problems, we need to prove some time-compactness for the approximate solutions. We make use of the time-compactness result for degenerate parabolic equations proposed in [14], as an alternative to the classical technique that consists in estimating the time-translates (see [11] in the continuous setting and [78] in the discrete setting).

Lemma 3.4.1. *There exists C_{14} depending only on ζ, T, f_0, g_0, ν and b such that*

$$\sum_{n=0}^{M_T-1}\sum_{K\in\mathcal{T}} m(K)(f_K^{n+1}-f_K^n)\varphi(x_K,t_{n+1}) \le C_{14}\|\nabla\varphi\|_{L^\infty(\Omega_T)}, \quad \varphi\in\mathcal{C}_c^\infty(\Omega_T). \quad (3.31)$$

$$\sum_{n=0}^{M_T-1}\sum_{K\in\mathcal{T}} m(K)(g_K^{n+1}-g_K^n)\varphi(x_K,t_{n+1}) \le C_{14}\|\nabla\varphi\|_{L^\infty(\Omega_T)}, \quad \varphi\in\mathcal{C}_c^\infty(\Omega_T). \quad (3.32)$$

Proof. For the sake of readability, we denote by $\varphi_K^{n+1}=\varphi(x_K,t_{n+1})$ for all $K\in\mathcal{T}$ and all $n\in\{0,...,M_T-1\}$. We multiply the scheme (4.40) by $\Delta t\varphi_K^{n+1}$ and sum for $K\in\mathcal{T}$, for $n\in\{0,...,M_T-1\}$. This yields :

$$A=B,$$

where

$$A=\sum_{n=0}^{N}\sum_{K\in\mathcal{T}} m(K)(f_K^{n+1}-f_K^n)\varphi_K^{n+1},$$

and

$$B=-\nu\sum_{n=0}^{M_T-1}\Delta t\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}}\tau_\sigma f_\sigma^{n+1}\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1}) + (b_K-b_L)\Big](\varphi_K^{n+1}-\varphi_L^{n+1}).$$

Using the Cauchy-Schwarz inequality, we get

$$|B|^2 \leq \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}(\varphi_K^{n+1}-\varphi_L^{n+1})^2$$
$$\times \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}\nu^2\Big[(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\Big]^2.$$

Moreover $f_\sigma^{n+1} \in \Big[\min(f_K^{n+1}, f_L^{n+1}), \max(f_K^{n+1}, f_L^{n+1})\Big]$, hence

$$0 \leq f_\sigma^{n+1} \leq f_K^{n+1} + f_L^{n+1}, \quad \forall \sigma \in \mathcal{E},\ \forall n \in \{0, ..., M_T-1\}. \tag{3.33}$$

Using Corollary 3.3.9 , we get

$$|B|^2 \leq C_{13} \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (f_K^{n+1}+f_L^{n+1}) d(x_K,x_L)^2 \|\nabla\varphi\|^2_{L^\infty(\Omega_T)}.$$

Observe that in two space dimensions,

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{{\rm int},K}} d_\sigma m(\sigma) \leq 2 \sum_{K\in\mathcal{T}} m(K). \tag{3.34}$$

By (4.37), (4.36) one has

$$\sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma (f_K^{n+1}+f_L^{n+1}) d(x_K,x_L)^2$$
$$\leq \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} m(\sigma) d(x_K,x_L)(f_K^{n+1}+f_L^{n+1})$$
$$\leq \sum_{n=0}^{M_T-1} \Delta t \sum_{K\in\mathcal{T}} f_K^{n+1} \sum_{\sigma\in\mathcal{E}_{{\rm int},K}} m(\sigma) d(x_K,x_L)$$
$$\leq \frac{2T}{\zeta} \sum_{K\in\mathcal{T}} m(K) f_K^{n+1} \leq \frac{2T}{\zeta} \|f_0\|_{L^1(\Omega)},$$

thanks to the mass conservation (3.21). Hence

$$\sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} m(\sigma) d(x_K,x_L) f_\sigma^{n+1} \leq \frac{2T}{\zeta} \|f_0\|_{L^1(\Omega)}. \tag{3.35}$$

This concludes the proof of (3.31). The proof of (3.32) is similar. □

We can apply the [14, Theorem 3.9], we conclude that

Proposition 3.4.2. *Let $(\mathcal{D}_m)_{m\geq 1}$ be a family of admissible space-time discretization of Ω_T such that (4.36) holds. Let $(f_m)_m$ and $(g_m)_m$ be the corresponding sequence of discrete solution to the scheme (4.40)-(4.43), then up to an unlabeled subsequence, there exist $f \in L^2(0,T;H^1(\Omega))$ and $g \in L^2(0,T;H^1(\Omega))$ such that*

$$f_m \underset{m\to+\infty}{\longrightarrow} f, \quad a.e \text{ in } \Omega_T, \quad \text{and} \quad g_m \underset{m\to+\infty}{\longrightarrow} g, \quad a.e \text{ in } \Omega_T.$$

$(f_m)_{m>0}$ and $(g_m)_{m>0}$ are uniformly bounded in $L^\infty(0;T,L^2(\Omega))$ thanks to Corollary 3.3.8. By Corollary 3.3.5 and Sobolev embedding, $(f_m)_{m>0}$ and $(g_m)_{m>0}$ are bounded in $L^2(0,T;L^p(\Omega))$, with $p<+\infty$. Then, thanks to Riez-Thorin theorem, $(f_m)_{m>0}$ and $(g_m)_{m>0}$ are bounded in $L^r(\Omega_T)$ with $2<r<4$. Hence $(f_m)_{m>0}$ and $(g_m)_{m>0}$ are equi-integrable in $L^r(\Omega_T)$. Applying the Vitali's convergence theorem we deduce that

Lemma 3.4.3. *Keeping the assumption and notations of Proposition 3.4.2 , one has*

$$f_m \underset{m\to+\infty}{\longrightarrow} f, \quad \text{strongly in } L^r(\Omega_T), \quad \text{for all } r<4,$$

and

$$g_m \underset{m\to+\infty}{\longrightarrow} g, \quad \text{strongly in } L^r(\Omega_T), \quad \text{for all } r<4.$$

We show now the weak compactness in $L^2(\Omega_T)^2$ of the approximate gradient of f_m and g_m defined in (3.17) :

Proposition 3.4.4. *Keeping the assumption and notations of Proposition 3.4.2 , one has*

$$\nabla^m f_m \underset{m\to+\infty}{\longrightarrow} \nabla f \text{ weakly in } L^2(\Omega_T)^2, \text{ and } \nabla^m g_m \underset{m\to+\infty}{\longrightarrow} \nabla g \text{ weakly in } L^2(\Omega_T)^2.$$

Proof. Thanks to (3.29) $\nabla^{\mathcal{D}} f_{\mathcal{D}}$ and $\nabla^{\mathcal{D}} g_{\mathcal{D}}$ are bounded in $L^2(\Omega_T)^2$, then there exists a subsequence of $\nabla^{\mathcal{D}} f_{\mathcal{D}}$ and of $\nabla^{\mathcal{D}} g_{\mathcal{D}}$ (still labeled $\nabla^{\mathcal{D}} f_D$ and $\nabla^{\mathcal{D}} g_{\mathcal{D}}$) and two function $\Theta, \Xi \in L^2(\Omega_T)^2$ such that

$$\nabla^m f_m \underset{m\to+\infty}{\longrightarrow} \Theta, \quad \text{weakly in } L^2(\Omega_T)^2,$$

and

$$\nabla^m g_m \underset{m\to+\infty}{\longrightarrow} \Xi, \quad \text{weakly in } L^2(\Omega_T)^2.$$

We refer to [48, 83] to prove that $\Theta = \nabla f$ and $\Xi = \nabla g$. □

3.4.2 Identification as a weak solution

Proposition 3.4.5. *Let (f, g) be as in Proposition 3.4.2, then f and g are the weak solution to (4.9)-(3.4) in the sense of (3.18) and (3.19).*

Proof. Let $\psi \in C_0^\infty(\overline{\Omega} \times [0, T))$ be a test function and $\psi_K^{n+1} = \psi(x_K, t_{n+1})$ for all $K \in \mathcal{T}$ and $n \in \{0, ..., M_T - 1\}$. We first establish (3.18) from (4.40) and to obtain (3.19) from (4.41) is similar. In order to prove that f is a weak solution, we multiply (4.40) by $\Delta t_m \psi_K^n$ and sum over $n \in \{0, ..., M_T - 1\}$ and $K \in \mathcal{T}$, we obtain

$$A_m + B_m = 0,$$

where

$$A_m = \sum_{n=0}^{M_T-1} \sum_{K\in\mathcal{T}} m(K)(f_K^{n+1} - f_L^{n+1})\psi_K^n,$$
$$B_m = \nu \sum_{n=0}^{M_T-1} \Delta t_m \sum_{K\in\mathcal{T}} \psi_K^n \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} \tau_\sigma f_\sigma^{n+1} \Big[(f_K^{n+1} - f_L^{n+1}) + (g_K^{n+1} - g_L^{n+1}) + (b_K - b_L)\Big].$$

Note that $\psi_K^{M_T} = 0$ for all $K \in \mathcal{T}$, then a discrete integration part yields

$$\begin{aligned} A_m &= -\sum_{n=0}^{M_T-1} \Delta t_m \sum_{K\in\mathcal{T}} m(K) \frac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m} f_K^{n+1} - \sum_{K\in\mathcal{T}} m(K) f_K^0 \psi_K^0 \\ &= -\int_0^T \int_\Omega f_m (\delta\psi)_m \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega f_m^0 \psi_m(.,0)\,\mathrm{d}x, \end{aligned}$$

where the function $\delta\psi_m(x_K, t) = \dfrac{\psi_K^{n+1} - \psi_K^n}{\Delta t_m}$, if $(x_K, t) \in K \times (t^n, t^{n+1})$. Thanks to the regularity of ψ, the function $\delta\psi_m$ converges uniformly towards $\partial_t\psi$ on Ω_T. Moreover, we have

$$f_m \longrightarrow f \quad \text{in } L^2(\Omega_T) \text{ as } m \to \infty.$$

Therefore

$$\int_0^T \int_\Omega f_m (\delta\psi)_m \,\mathrm{d}x\,\mathrm{d}t \longrightarrow \int_0^T \int_\Omega f(x)\partial_t\psi \,\mathrm{d}x\,\mathrm{d}t \quad \text{as } m \to \infty. \tag{3.36}$$

Moreover, f_m^0 converges strongly in $L^1(\Omega)$ towards the initial data f_0 and $\psi_m(.,0)$ converges uniformly towards $\psi(.,0)$. Therefore, we get that

$$\int_\Omega f_m^0 \psi_m(.,0)\,\mathrm{d}x \longrightarrow \int_\Omega f_0(x)\psi(.,0)\,\mathrm{d}x \quad m \to \infty. \tag{3.37}$$

We deduce from (3.36) and (3.37) that

$$A_m \longrightarrow -\int_0^T \int_\Omega f(x)\partial_t\psi \,\mathrm{d}x\,\mathrm{d}t - \int_\Omega f_0(x)\psi(.,0)\,\mathrm{d}x \quad m \to \infty.$$

We introduce the term

$$E_m = \int_{\Omega_T} \overline{f}_{\mathcal{D}_m} \nabla^m u_m \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t,$$

where

$$\overline{f}_{\mathcal{D}}(x,t) = f_\sigma^{n+1} \quad \forall (t,x) \in (t^n, t^{n+1}] \times \mathfrak{D}_{K,L}, \text{ and } u = f + g + b.$$

We have $f_m \underset{m\to+\infty}{\longrightarrow} f$, strongly in $L^2(\Omega_T)$. Let us to prove that

$$\overline{f}_{\mathcal{D}_m} := \overline{f}_m \underset{m\to+\infty}{\longrightarrow} f, \quad \text{strongly in } L^2(\Omega_T).$$

Since $\|\overline{f}_m - f\|_{L^2(\Omega_T)} \leq \|\overline{f}_m - f_m\|_{L^2(\Omega_T)} + \|f_m - f\|_{L^2(\Omega_T)}$, it is sufficient to prove that $\|\overline{f}_m - f_m\|_{L^2(\Omega_T)} \longrightarrow 0$, as $\delta_m \underset{m\to+\infty}{\longrightarrow} 0$. One has

$$\begin{aligned}
\|\overline{f}_m - f_m\|^2_{L^2(\Omega_T)} &= \sum_{n=0}^{M_T-1} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} m(\mathfrak{D}_{K,\sigma})(f_K^{n+1} - f_\sigma^{n+1})^2 \\
&\leq \sum_{n=0}^{M_T-1} \Delta t \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} m(\mathfrak{D}_{K,\sigma})(f_K^{n+1} - f_L^{n+1})^2 \\
&\leq \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}} \\ \sigma=K|L}} m(\mathfrak{D}_{K,L})(f_K^{n+1} - f_L^{n+1})^2 \\
&\leq \frac{1}{2} \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}} \\ \sigma=K|L}} \tau_\sigma (f_K^{n+1} - f_L^{n+1})^2 d_\sigma^2 \leq \frac{C_{11}}{2}\delta_m^2.
\end{aligned}$$

Since $\nabla^m u_m$ converges weakly in $L^2(\Omega_T)$ to ∇u, since $\overline{f}_m$ converges strongly in $L^2(\Omega_T)$ to f, we have

$$E_m \longrightarrow \int_{\Omega_T} f\nabla u \cdot \nabla\psi \,\mathrm{d}x\,\mathrm{d}t \quad \text{as } m \to \infty.$$

Let us to prove $E_m - \overline{E}_m$ tends to 0 as $m \to \infty$, where

$$\overline{E}_m = \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}} \\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}(u_K^{n+1} - u_L^{n+1})(\psi_K^n - \psi_L^n).$$

Using the definition of discrete gradient (3.17), we have

$$E_m = \sum_{n=0}^{M_T-1} \sum_{\substack{\sigma\in\mathcal{E}_{\mathrm{int}} \\ \sigma=K|L}} f_\sigma^{n+1} \int_{t^n}^{t^{n+1}} \int_{\mathfrak{D}_{K,L}} \frac{m(\sigma)}{m(\mathfrak{D}_{K,L})}(u_K^{n+1} - u_L^{n+1})\nabla\psi \cdot \nu_{L,K} \,\mathrm{d}x\,\mathrm{d}t.$$

Therefore by the definition of τ_σ,

$$E_m - \overline{E}_m = \sum_{n=0}^{M_T-1} \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} m(\sigma) f_\sigma^{n+1}(u_K^{n+1}-u_L^{n+1}) \int_{t^n}^{t^{n+1}} \left(\frac{\psi_K-\psi_L}{d(x_K,x_L)} - \frac{1}{m(\mathfrak{D}_{K,L})}\int_{\mathfrak{D}_{K,L}} \nabla\psi\cdot\nu_{L,K}\,\mathrm{d}x \right)\mathrm{d}t.$$

On the one hand, since the straight line (x_K, x_L) is orthogonal to σ, we have $x_K - x_L = d(x_K,x_L)\nu_{L,K}$. It follows from the regularity of ψ that

$$\begin{aligned}\frac{\psi_K^n-\psi_L^n}{d(x_K,x_L)} =& \nabla\psi(t^n,x_L)\cdot\nu_{L,K} + O(\delta)\\ =& \nabla\psi(t,x)\cdot\nu_{L,K} + O(\eta), \quad \forall (t,x)\in(t^n,t^{n+1})\times\mathfrak{D}_{K,L}.\end{aligned}$$

By taking the mean value over $\mathfrak{D}_{K,L}$, there exists a constant $C_{15}>0$, depending only on ψ, such that

$$\left| \int_{t^n}^{t^{n+1}} \left(\frac{\psi_K^n-\psi_L^n}{d(x_K,x_L)} - \frac{1}{m(\mathfrak{D}_{K,L})}\int_{\mathfrak{D}_{K,L}} \nabla\psi\cdot\nu_{L,K}\,\mathrm{d}x \right)\mathrm{d}t \right| \le C_{15}\Delta t\eta.$$

On the other hand, one has $m(\sigma) = \sqrt{\tau_\sigma}\sqrt{m(\sigma)d(x_K,x_L)}$. Hence by Cauchy-schwarz inequality, we have

$$\begin{aligned}&\left| \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} m(\sigma) f_\sigma^{n+1}(u_K^{n+1}-u_L^{n+1}) \right|^2\\ &\quad\le \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} \tau_\sigma f_\sigma^{n+1}(u_K^{n+1}-u_L^{n+1})^2 \times \sum_{n=0}^{M_T-1} \Delta t \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} m(\sigma)d(x_K,x_L) f_\sigma^{n+1}.\end{aligned}$$

Using Corollary 3.3.9 and (3.35) we conclude that

$$|E_m - \overline{E}_m|^2 \le \frac{2T}{\zeta}\|f_0\|_{L^1(\Omega)} C_{12} C_{15}^2 \eta^2 \longrightarrow 0, \quad \text{as } \eta\to 0.$$

This ensures that B_m converges towards $\nu\int_{\Omega_T} f\nabla u\cdot\nabla\psi\,\mathrm{d}x\,\mathrm{d}t$, as $m\to+\infty$. □

3.5 Numerical results

Let us provide some illustrations of the behavior of the numerical scheme (4.40)-(4.43). The scheme leads to a nonlinear system that we solve thanks to the Newton-Raphson method. In our test case, the domain is the unit square, i.e., $\Omega = (0,1)^2$. We

consider an admissible triangular mesh made of 14336 triangles. An illustration of a mesh type used here is given in Figure 3.2. The numerical analysis of the scheme was carried out for a uniform time discretization of $(0,T)$ only in order to avoid heavy notations. In order to increase the robustness of the algorithm and to ensure the convergence of the Newton-Raphson iterative procedure, we used an adaptive time step procedure in the practical implementation. More precisely, we associate a maximal time step $\Delta t_{\max} = 0.00004$ for the mesh . If the Newton-Raphson method fails to converge after 30 iterations —we choose that the ℓ^∞ norm of the residual has to be smaller than 10^{-10} as stopping criterion—, the time step is divided by two. If the Newton-Raphson method converges, the time step is multiplied by two and projected on $[0, \Delta t_{\max}]$. The first time step Δt is equal to $\Delta t_{\max}$ in the test case presented below. We perform the numerical experiments with the following data

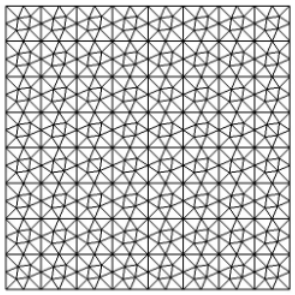


FIGURE 3.2 – Mesh type used in the numerical test

$$b(x,y) = \max\left(0, \frac{1}{2}\Big(1 - 16(x-1/2)^2\Big)\Big((\cos(\pi y) + 2\Big)\right), \quad \nu = 0.9.$$

As an initial condition we take

$$f_0(x,y) = \begin{cases} \dfrac{1}{2} & \text{if } x \le \dfrac{1}{4}, \\ 0 & \text{elsewhere,} \end{cases} \qquad g_0(x,y) = \begin{cases} b\Big(\dfrac{1}{2}, 0\Big) - b(x,y) - \Big(x - \dfrac{1}{2}\Big) & \text{if } x \le \dfrac{1}{2}, \\ 0 & \text{elsewhere.} \end{cases}$$

Figure 3.3 shows the evolution of $\xi(x,t) = b(x) + g(x,t)$ (*red*) and $h(x,t) = b(x) + g(x,t) + f(x,t)$ (*blue*) at time $t = 0, t = 0.2, t = 0.72$ and $t = 12$. There is convergence towards an equilibrium state, with horizontal interfaces as expected (see [72]).

Figure 3.4 shows the evolution of the energy along time

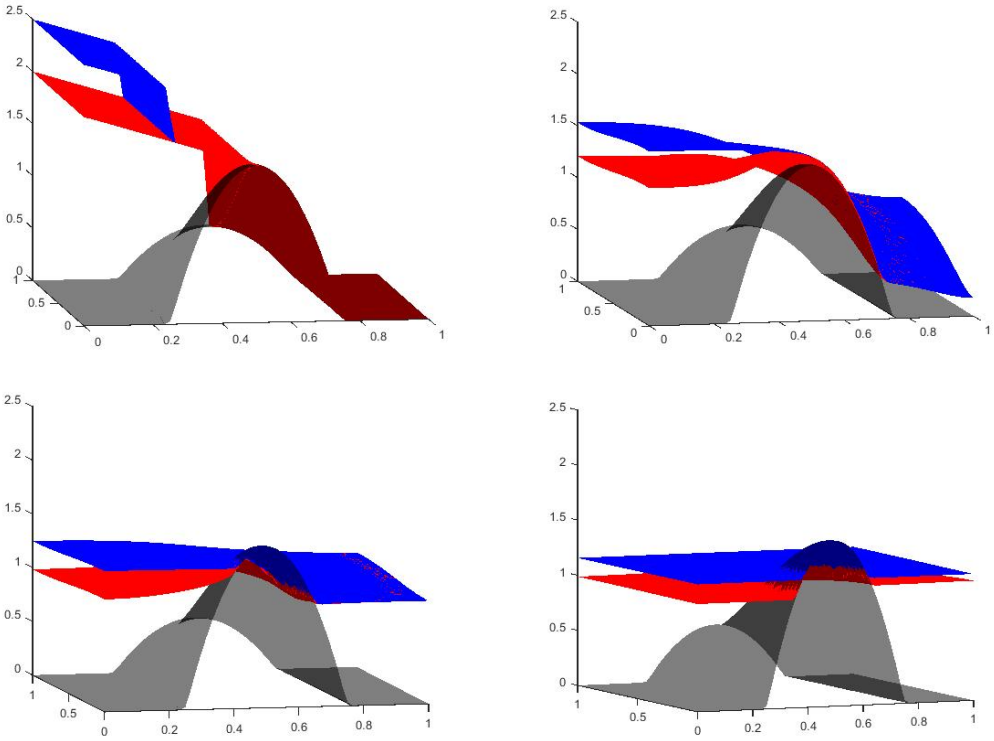


FIGURE 3.3 – Behaviour of the model at $t = 0,\ t = 0.2,\ t = 0.72,\ t = 12$

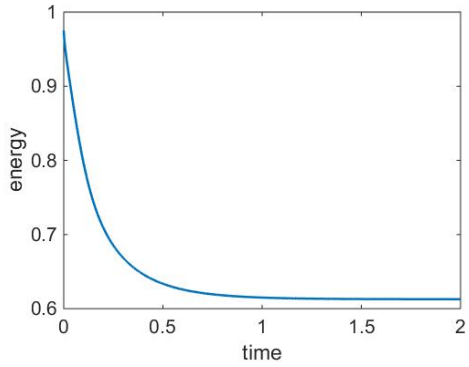


FIGURE 3.4 – Evolution of the energy along time

3.6 Conclusion

We proposed and analyzed a finite volume scheme for solving the seawater intrusion model. It preserves at the discrete level the main features of the continuous problem : the nonnegativity of the solutions, the decay of the energy, the control of the entropy and its dissipation. Moreover, we were able to carry out a full convergence analysis based on compactness arguments.

Let us mention that to derive the problem (4.9), the authors in [108] assume that, for simplification, the porous medium is isotropic. Moreover in our work the mesh satisfies the so-called orthogonality condition (see, e.g., [78, Definition 9.1] so that the two-point flux approximation is consistent. We know that the finite volume method with two-point flux approximation, used here, does not allow to handle the anisotropic case. Nevertheless in order to treat this case, we can use for instance a Control Volume Finite Element scheme (CVFE) [9, 39, 91].

Chapitre 4

The large time behaviour of a seawater intrusion model

Abstract : The large time behaviour of the solutions to a seawater intrusion model is studied. The goal is to identify the equilibrium states which are the minimizers of a convex energy. We prove for the continuous problem the existence and uniqueness of the minimizers of the energy functional, that the minimizers are stationary states, that these stationary states are radial and unique. For different viscosities, we give numerical illustrations of the stationary states and we exhibit the convergence rate.

4.1 Introduction

4.1.1 Presentation of the continuous problem

The purpose of this work is to investigate the large time behaviour of a seawater intrusion model. In densely populated coastal regions, the intensive extraction of freshwater yields local water table depression and saltwater from the sea can enter the ground and replace the freshwater, leading to the so-called seawater intrusion problem. There are several approaches to model this phenomenon. In this chapter we are interested in a mathematical model describing the time evolution of two immiscible fluids (freshwater and saltwater) in an unconfined aquifer, assuming the sharp interface (the fluids occupy disjointed regions) and the Dupuit approximation (quasi-horizontal displacements). We consider the formal asymptotic limit as the aspect ratio between the thickness and the horizontal length of the porous medium tends to zero. Notice that we have a 2D model obtained from 3D model after applying a vertical integration thanks to Dupuit approximation (see Section 1.1.3). More precisely the interface between the saltwater and the bedrock is set at $\{z = 0\}$. We denote the height of the freshwater (resp. saltwater) layer by $\{z = \tilde{f}(t,x)\}$ (resp. $\{z = \tilde{g}(t,x)\}$), see Figure 4.1. The seawater intrusion problem in an unconfined aquifer can be described by the following model

$$\begin{cases} \partial_t \tilde{f} - \nabla.\Big(\nu\mu\tilde{f}\nabla(\tilde{f}+\tilde{g})\Big) = 0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2, \\ \partial_t \tilde{g} - \nabla.\Big(\tilde{g}\nabla(\nu\tilde{f}+\tilde{g})\Big) = 0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2, \\ \tilde{f}_{|t=0} = \tilde{f}_0, \quad \tilde{g}_{|t=0} = \tilde{g}_0 & \text{in} \quad \mathbb{R}^2, \end{cases} \tag{4.1}$$

with the density ratio ν is given by $\nu = \dfrac{\rho_{\text{fresh}}}{\rho_{\text{salt}}} \in (0,1)$, and the viscosity ratio by $\mu = \dfrac{\mu_{\text{salt}}}{\mu_{\text{fresh}}}$. The model (4.1) has been derived by Monneau and Jazar in [108]. The

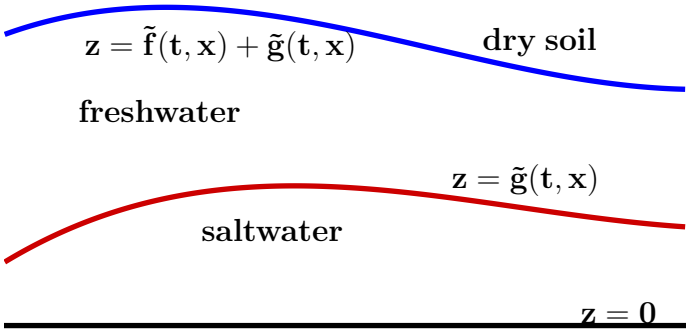


FIGURE 4.1 – The physical setting

authors in [73, 74] studied the classical solutions of system (4.1). Moreover, weak solutions are established under different assumptions in [58, 72, 119, 120].

The characteristic time corresponding to the aquifer dynamics is large. Therefore, understanding the large-time behaviour of system (4.1) is of great interest. The so-called entropy method [17, 113] provides a powerful approach to study the long time behavior of different systems of PDEs. It has been developed firstly for the kinetic equations (Boltzmann and Landau [147]). Then it was extended to other problems, as the linear Fokker-planck equation [44], the porous medium equation (PME) [45, 149], the reaction-diffusion systems [62, 63, 99], the drift-diffusion systems for semiconductor devices [94, 95, 96], the thin film models [42], coagulation-fragmentation models [90]. For more details about this method and its application domains, one can refer to [17, 113] and the references therein. Similar results were obtained in [25, 26, 111, 134, 152] based on the interpretation of the PDE models as the gradient flow of a certain energy functional with respect the Wasserstein metric. We refer for instance to the monographs for an extensive discussion on this topic [12, 141].

In order to capture numerically the long-time behavior, it is important to design numerical schemes for these systems which preserve at the discrete level the main features of the continuous problem as the positivity of densities, the conservation of mass and decay of the energy. The question of the large time behavior of numerical schemes has been investigated in several works. We refer for instance to [18, 37] for Fokker-Planck, to [101] for the porous medium equation, to [90] for coagulation-fragmentation models, to [47, 114] for nonlinear diffusion equations, to [99, 100] for reaction-diffusion systems, to [22, 49] for drift-diffusion systems. See [46] for an overview on the discrete entropy method.

In [121], Laurençot and Matioc studied the large-time behavior of the system (4.1) in the one-dimensional case. In their paper a classification of self-similar solutions is first provided : there is always a unique even self-similar solution while a continuum of non-symmetric self similar solutions exists for certain fluid configurations. The authors proved the convergence of all nonnegative weak solutions towards a self-similar solution. Nevertheless nothing is known about the rate of convergence. Surprisingly, the situation is simpler in the 2D case, as it will appear in the sequel.

The system (4.1) can be interpreted as a two-phase generalization of the porous medium equation (PME) . In order to explain the principles of the entropy method, let us consider the following PME

$$\begin{cases} \partial_t v = \Delta v^2 & \text{on } \mathbb{R}^2 \times (0,\infty), \\ v(x,0) = v_0(x) \geq 0 & \text{on } \mathbb{R}^2, \end{cases} \tag{4.2}$$

A further transformation of (4.2) involves the so-called self-similar variables (see [45, 149]) and reads

$$u(t,x) = e^{2t} v\Big(\frac{1}{4}(e^{4t}-1), x e^t\Big), \tag{4.3}$$

then we transform (4.2) into the nonlinear Fokker-Planck equation

$$\begin{cases} \partial_t u = \operatorname{div}(xu + \nabla u^2) & \text{on } \mathbb{R}^2 \times (0,\infty), \\ u(x,0) = u_0(x) \geq 0 & \text{on } \mathbb{R}^2. \end{cases} \tag{4.4}$$

In [45] the authors study the large time behavior of the PME (4.2) using the rescaled equation (4.4). The energy corresponding to (4.4) writes

$$H(u) = \int_{\mathbb{R}^2} \left(|x|^2 u + 2u^2 \right) \mathrm{d}x.$$

They prove that the unique stationary solution of (4.4) is given by the Barenblatt-Pattle type formula

$$u_\infty(x) = \left(\beta - \frac{1}{4}|x|^2 \right)_+,$$

where β is given by $\int_{\mathbb{R}^2} u_\infty(x)\,\mathrm{d}x = \int_{\mathbb{R}^2} u_0(x)\,\mathrm{d}x$ and $y_+ = \max(y, 0)$. The relative entropy of u w.r.t u_∞ is then defined by

$$H(u|u_\infty) := H(u) - H(u_\infty) \geq 0,$$

whereas the entropy production for $H(u|u_\infty)$ is given by

$$I(u) = \int_{\mathbb{R}^2} u \left| x + 2\nabla u \right|^2 \mathrm{d}x.$$

Under some mild assumptions on u_0 one has

$$H(u(t)|u_\infty) \longrightarrow 0,\ t \to +\infty \quad \text{and} \quad I(u(t)) \longrightarrow 0,\ t \to +\infty.$$

Moreover, $H(u(t)|u_\infty)$ and $I(u(t))$ are linked by the relation

$$\frac{\mathrm{d}}{\mathrm{d}t} H(u(t)|u_\infty) = -2I(u(t)), \tag{4.5}$$

and

$$\frac{\mathrm{d}}{\mathrm{d}t} I(u(t)) = -2I(u(t)) - R(t), \tag{4.6}$$

where $R(t) \geq 0$. Combining (4.5) and (4.6) one obtains

$$\frac{\mathrm{d}}{\mathrm{d}t} H(u(t)|u_\infty) = \frac{\mathrm{d}}{\mathrm{d}t} I(u(t)) + R(t). \tag{4.7}$$

Integrating (4.7) between $t > 0$ and $+\infty$, we get

$$0 \leq H(u(t)|u_\infty) \leq I(u(t)), \quad t > 0, \tag{4.8}$$

and substituting (4.8) into (4.5), one concludes with the exponential decay of the relative entropy to zero at a rate 2.

In this chapter we study the convergence in relative energy of the evolutive solutions towards an equilibrium state. We also want to recover numerically the rate of convergence which appears in different tests cases. To this end, we need a numerical scheme which preserves at the discrete level the main features of the continuous problem (in particular the nonnegative of the solutions, conservation of mass and decay of energy) and which converges towards a stationary state. The Finite Volume scheme studied in Chapter 3 is a natural candidate.

The outline of the chapter is as follows. In the next section we state the main results of our chapter. As a preliminary step, we introduce a rescaled version (4.9) of the system (4.1) which relies in particular on the introduction of self-similar variables. In Theorem 4.2.1 we state the existence and uniqueness of nonnegative stationary solutions to (4.9), which are moreover radial and Lipschitz continuous. Some key properties of the continuous problem are studied in Section 4.3 and we prove Theorem 4.2.1. In Section 4.4 we give a classification of the self-similar profiles and we exhibit critical values of the parameter μ for which the shape of the stationary profile changes. We finally present in Section 4.5 numerical simulations for different values of μ in order to observe the stationary solutions and the decay of the relative energy.

4.2 Main results

We introduce the following closed convex set

$$\mathcal{K}_{M_h} := \Big\{ h \in L^2(\mathbb{R}^2) \cap L^1\Big(\mathbb{R}^2, (1+|x|^2)\,\mathrm{d}x\Big) : h \geq 0 \ \text{ a.e. and } \int_{\mathbb{R}^2} h(x)\,\mathrm{d}x = \int_{\mathbb{R}^2} h_0(x) := M_h > 0 \Big\}.$$

4.2.1 Self-similar solutions

The main contribution of this chapter is the classification of the nonnegative self-similar solutions to (4.1). Let us transform (4.1) using the so-called self-similar variables [45, 149], i.e.

$$(f,g) = \frac{1}{(1+t)^{1/2}} (\tilde{f},\tilde{g})\Big(\log(1+t), \frac{x}{(1+t)^{1/4}} \Big).$$

Then setting $(\tilde{f}_0, \tilde{g_0}) := (f_0, g_0)$, we end up with the following rescaled system

$$\begin{cases} \partial_t f - \nabla.\Big(\nu\mu f \nabla(f+g+b/\mu)\Big) = 0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2, \\ \partial_t g - \nabla.\Big(g\nabla(\nu f+g+b)\Big) = 0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2, \\ f_{|t=0} = f_0, \quad g_{|t=0} = g_0 & \text{in} \quad \mathbb{R}^2, \end{cases} \tag{4.9}$$

where $b(x) = \dfrac{1-\nu}{8}|x|^2$. The change of variables preserves the gradient flow structure, but for a modified energy. Indeed the system (4.9) can be interpreted as the gradient flow with respect to the 2−Wasserstein metric of the following energy

$$\mathfrak{E}(f,g) = \int_{\mathbb{R}^2} E(f,g)\,\mathrm{d}x, \tag{4.10}$$

where

$$E(f,g) = \frac{\nu}{2}(f+g)^2 + \frac{1-\nu}{2}g^2 + b\Big(\frac{\nu}{\mu}f + g\Big). \tag{4.11}$$

A corner stone of our study is that, if $(\tilde{f},\tilde{g})$ is a self-similar solution to (4.1) whose the form is

$$(\tilde{f},\tilde{g}) = t^{-1/2}(F,G)(xt^{-1/4}), \quad (t,x) \in (0,\infty)\times\mathbb{R}^2, \tag{4.12}$$

then the corresponding self-similar profile (F,G) is a stationary solution to (4.9). We will see in what follows that (F,G) is the unique minimizer of $\mathfrak{E}$ in $\mathcal{K}_{M_f}\times\mathcal{K}_{M_g}$ and that it satisfies the following system

$$\begin{cases} F\nabla(F+G+b/\mu) &= 0 \quad \text{in} \quad \mathbb{R}^2, \\ G\nabla(\nu F+G+b) &= 0 \quad \text{in} \quad \mathbb{R}^2. \end{cases}$$

Let $(F,G) \in \mathcal{K}_{M_f}\times\mathcal{K}_{M_g}$ be a stationary solution to (4.9), we define the positivity sets E_F and E_G of F and G by

$$E_F = \{x\in\mathbb{R}^2 \ : \ F(x)>0\}, \quad E_G = \{x\in\mathbb{R}^2 \ : \ G(x)>0\}.$$

We notice that E_F and E_G are both nonempty as

$$\int_{\mathbb{R}^2} F\,\mathrm{d}x = M_f > 0 \quad \text{and} \quad \int_{\mathbb{R}^2} G\,\mathrm{d}x = M_g > 0. \tag{4.13}$$

4.2.2 Main results

The viscosity ratio μ appears to play a central role in the characterization of the stationary profiles, as well as in the convergence speed towards the stationary state. Therefore, we will suppose that M_f, M_g and $\nu \in (0,1)$ are fixed and we will study the dependency on the large-time behaviour w.r.t μ.

Let $(F,G) \in \mathcal{K}_{M_f}\times\mathcal{K}_{M_g}$ be a stationary solution of (4.9).

Theorem 4.2.1 (Self-similar profiles)**.** *There exists a unique stationary solution (F,G) of (4.9). It is radial, Lipschitz continuous and satisfies*

$$\begin{cases} F\nabla(F+G+b/\mu) &= 0 \quad \textit{in} \quad \mathbb{R}^2, \\ G\nabla(\nu F+G+b) &= 0 \quad \textit{in} \quad \mathbb{R}^2. \end{cases}$$

Moreover, E_F and E_G are connected sets and

$$(F,G) \in \operatorname*{arg\,min}_{(f,g)\in\mathcal{K}_{M_f}\times\mathcal{K}_{M_g}} \mathfrak{E}(f,g).$$

Thanks to the uniqueness of stationary solutions (F, G), one can find explicitly and easily (F, G) contrary to the 1D case [121]. As it will be explicated in Section 4.4, the shape of F and G strongly depends on μ and the topology of E_F and E_G changes following the values of μ.

4.3 Mathematical study of the continuous problem

Let $\rho \in L^1_{\text{loc}}(\mathbb{R}^2; \mathbb{R}_+)$, then we denote by $L^r_\rho(\mathbb{R}^2)$ for $r \geq 1$, the set of the measurable functions ψ such that

$$\int_{\mathbb{R}^2} |\psi(y)|^r \rho(y)\, \mathrm{d}y < \infty.$$

The set $L^r_\rho(\mathbb{R}^2)$ is equipped with the norm

$$\|\psi\|_{L^2_\rho(\mathbb{R}^2)} = \left(\int_{\mathbb{R}^2} |\psi(y)|^r \rho(y)\, \mathrm{d}y \right)^{1/r}.$$

We set $\phi_f = f + g + b/\mu$ and $\phi_g = \nu f + g + b$.

Definition 4.3.1 (weak solution)**.** *A pair $(f, g) : \mathbb{R}^2 \times \mathbb{R}_+ \to \mathbb{R}^2_+$ is said to be a weak solution to the problem (4.9) if*

(i) f, g belong to $L^\infty(\mathbb{R}_+; L^2(\mathbb{R}^2))$ and to $L^\infty(\mathbb{R}_+; L^1_{1+|x|^2}(\mathbb{R}^2))$,

(ii) $\nabla f, \nabla g$ belong to $L^2_{\text{loc}}(\mathbb{R}_+; L^2(\mathbb{R}^2))^2$,

(iii) $\nabla\phi_f, \nabla\phi_g \in L^2_{\text{loc}}(\mathbb{R}_+; L^2(\mathbb{R}^2))^2$ and to $L^2_f(\mathbb{R}^2 \times \mathbb{R}_+)$ and $L^2_g(\mathbb{R}^2 \times \mathbb{R}_+)$ respectively,

(iv) for all $\xi \in C^\infty_c(\mathbb{R}^2 \times \mathbb{R}_+)$, there holds

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}^2} f\partial_t\xi \,\mathrm{d}x\,\mathrm{d}t + \int_{\mathbb{R}^2} f_0\xi(\cdot,0)\,\mathrm{d}x - \int_{\mathbb{R}_+}\int_{\mathbb{R}^2} \mu\nu f\nabla\phi_f\cdot\nabla\xi\,\mathrm{d}x\,\mathrm{d}t = 0,$$

$$\int_{\mathbb{R}_+}\int_{\mathbb{R}^2} g\partial_t\xi \,\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}^2} g_0\xi(\cdot,0)\,\mathrm{d}x - \int_{\mathbb{R}_+}\int_{\mathbb{R}^2} g\nabla\phi_g\cdot\nabla\xi\,\mathrm{d}x\,\mathrm{d}t = 0.$$

The existence of a weak solutions of the problem (4.9) and the decay of the energy $\mathfrak{E}$ is given by the following theorem.

Theorem 4.3.2. *There exists a weak solution in the sense of the above definition. Moreover, it satisfies the following energy inequality :*

$$\mathfrak{E}(f,g)(t) + \int_s^t I(f,g)(\tau)\,\mathrm{d}\tau \leq \mathfrak{E}(f,g)(s), \quad \forall t \geq s \geq 0, \tag{4.14}$$

with I the energy dissipation

$$I(f,g) = \int_{\mathbb{R}^2} \left(\mu\nu^2 f|\nabla\phi_f|^2 + g|\nabla\phi_g|^2 \right) \mathrm{d}x. \tag{4.15}$$

The existence of a weak solution was proven by Laurençot and Matioc in [119, 120] by proving the convergence of a JKO scheme, but the proof can be extended in the presence of a quadratic confining potential b without particular difficulties. The L^2_{loc} estimates on ∇f, ∇g, $\nabla \phi_f$ and $\nabla \phi_g$ are obtained thanks to the flow interchange technique of Matthes *et al.* [125].

We consider the minimization problem

$$\inf_{(f,g)\in \mathcal{K}_{M_f}\times \mathcal{K}_{M_g}} \mathfrak{E}(f,g). \tag{4.16}$$

In order to prove the uniqueness of the minimizer, we need the strict convexity of the energy functional $\mathfrak{E}$.

Proposition 4.3.3. *$\mathfrak{E}$ is a strictly convex function on $\mathcal{K}_{M_f} \times \mathcal{K}_{M_g}$.*

Proof. If E strictly convex then $\mathfrak{E}$ is strictly convex. We denote by D^2E the hessian matrix of E. Then

$$D^2E = \begin{pmatrix} \nu & \nu \\ \nu & 1 \end{pmatrix}.$$

The matrix D^2E is symmetric. We have

$$\det(D^2E) = \nu(1-\nu) > 0 \quad \text{and} \quad \operatorname{tr}(D^2E) = 1 + \nu > 0.$$

Since $\det(D^2E) = \lambda_1\lambda_2$ and $\operatorname{tr}(D^2E) = \lambda_1 + \lambda_2$, where λ_1, λ_2 are the eigenvalues of D^2E, then $\lambda_1, \lambda_2 > 0$. We deduce that D^2E is definite positive and hence E is strictly convex. □

In order to prove the existence of the minimizer, we need the lower semi-continuity of the energy functional $\mathfrak{E}$ for the weak topology. We recall the definition of lower semi-continuity.

Definition 4.3.4. *Let X be a topological space. $f : X \longrightarrow]-\infty, +\infty]$ lower semi-continuous (l.s.c.) function if and only if we have these equivalent properties*

- $\forall a \in \mathbb{R} \quad \{x \in X |\ f(x) \leq a\}$ *is a closed set,*
- $\forall x_n \in X$ *such that* $x_n \longrightarrow x$ *one has* $\liminf_{n\to\infty} f(x_n) \geq f(x)$.

Since $\mathfrak{E}$ is convex, we will need the following proposition proved in [32].

Proposition 4.3.5. *Let $f : X \longrightarrow]-\infty, +\infty]$ be a l.s.c. convex function for the strong topology. Then f is l.s.c. for the weak topology. In particular if $x_n \rightharpoonup x$ weakly in X, then*

$$f(x) \leq \liminf_{n\to\infty} f(x_n).$$

The l.s.c. of $\mathfrak{E}$ for the weak topology is given by the following proposition.

Proposition 4.3.6. *The functional $\mathfrak{E}$ is l.s.c. for the weak topology of $L^2(\mathbb{R}^2)$. In particular if $f_n \to f$ weakly in L^2 and $g_n \to g$ weakly in L^2 then*

$$\mathfrak{E}(f,g) \leq \liminf_{n\to\infty} \mathfrak{E}(f_n, g_n).$$

Proof. Let $f_n \to f$ strongly in L^2 and $g_n \to g$ strongly in L^2, there exist two subsequences $(f_n)_n$ and $(g_n)_n$ such that $(f_n, g_n) \to (f, g)$ a.e. But E is convex, then l.s.c. and in particular

$$E(f,g) \leq \liminf_{n\to\infty} E(f_n, g_n).$$

Using Fatou's Lemma, one has

$$\mathfrak{E}(f,g) = \int_{\mathbb{R}^2} E(f,g) \leq \int_{\mathbb{R}^2} \liminf_{n\to\infty} E(f_n,g_n) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^2} E(f_n,g_n) = \liminf_{n\to\infty} \mathfrak{E}(f_n,g_n).$$

Therefore $\mathfrak{E}$ is l.s.c. for the L^2 strong topology. Thanks to Proposition 4.3.5 $\mathfrak{E}$ is l.s.c. for the L^2 weak topology since $\mathfrak{E}$ is convex. □

We can prove that there exists C (resp. C_1) depending only on ν (resp. on ν, μ and b) such that

$$C(\|f\|_{L^2(\mathbb{R}^2)} + \|g\|_{L^2(\mathbb{R}^2)}) \leq \mathfrak{E}(f,g) \leq C_1(\|f\|_{L^2(\mathbb{R}^2)} + \|g\|_{L^2(\mathbb{R}^2)} + 1). \tag{4.17}$$

In fact, since f, g and b are nonnegative, thus

$$E(f,g) = \frac{\nu}{2}(f+g)^2 + \frac{1-\nu}{2}g^2 + b(\frac{\nu}{\mu}f + g) \geq \frac{\nu}{2}(f+g)^2 \geq \frac{\nu}{2}(f^2+g^2),$$

which implies $\frac{\nu}{2}(\|f\|_{L^2(\mathbb{R}^2)} + \|g\|_{L^2(\mathbb{R}^2)}) \leq \mathfrak{E}(f,g)$. On the other hand, using Young inequality, we have $E(f,g) \leq \frac{3\nu}{2}f^2 + \frac{2+\nu}{2}g^2 + b^2\Big(\frac{1}{2} + \frac{\nu}{2\mu^2}\Big)$. We deduce that

$$\mathfrak{E}(f,g) \leq C_1(\|f\|_{L^2(\mathbb{R}^2)} + \|g\|_{L^2(\mathbb{R}^2)} + 1).$$

We are giving a classification of a nonnegative self-similar solutions to (4.1). We recall that the main feature of (4.9) is that, if $(\tilde{f}, \tilde{g})$ is a self-similar solution of (4.1), whose the form is given in (4.12) then the corresponding self-similar profile (F, G) is a stationary solution to (4.9) and by (4.14) one has

$$I(F,G) = \int_{\mathbb{R}^2} \left(\mu\nu^2 F|\nabla\phi_F|^2 + G|\nabla\phi_G|^2\right) \mathrm{d}x = 0,$$

where $\phi_F = F + G + b/\mu$ and $\phi_G = \nu F + G + b$. Hence (F, G) satisfy the equations

$$\begin{cases} F\nabla(F+G+b/\mu) &= 0 \quad \text{in} \quad \mathbb{R}^2, \\ G\nabla(\nu F+G+b) &= 0 \quad \text{in} \quad \mathbb{R}^2. \end{cases} \tag{4.18}$$

We first prove the existence and uniqueness of the minimizers of $\mathfrak{E}$.

Lemma 4.3.7. *There exists a unique minimizer (F,G) of $\mathfrak{E}$ in $\mathcal{K}_{M_f}\times\mathcal{K}_{M_g}$. Moreover it satisfy (4.18).*

Proof. The uniqueness of the minimizer follows from the strict convexity of the energy functional $\mathfrak{E}$ proved in Proposition 4.3.3.
Let us now proving the existence of a minimizer. To this end, pick a minimizing sequence $(f_k,g_k)_{k\geq 1}$. Thanks to (4.17) there exists a constant $C>0$ such that

$$\|f_k\|_{L^2}+\|g_k\|_{L^2}\leq C,\ \forall k\geq 1. \tag{4.19}$$

We obtain that there exist $(F,G)\in L^2(\mathbb{R}^2;\mathbb{R})^2$ and a subsequence of $(f_k,g_k)_{k\geq 1}$ (denoted again by $(f_k,g_k)_{k\geq 1}$) such that

$$f_k\to F \ \text{ weakly in } \ L^2 \quad \text{and} \quad g_k\to G \ \text{ weakly in } \ L^2.$$

Thanks to Proposition 4.3.6, $\mathfrak{E}$ is l.s.c., which implies that

$$\mathfrak{E}(F,G)\leq \liminf_{k\to\infty}\mathfrak{E}(f_k,g_k),$$

so that (F,G) is a minimizer of $\mathfrak{E}$. Let $(\check{f},\check{g})$ a solution of the following system

$$\begin{cases}\partial_t\check{f}-\nabla.\Big(\nu\mu\check{f}\nabla(\check{f}+\check{g}+b/\mu)\Big)=0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2,\\ \partial_t\check{g}-\nabla.\Big(\check{g}\nabla(\nu\check{f}+\check{g}+b)\Big)=0 & \text{in} \quad (0,\infty)\times\mathbb{R}^2,\\ \check{f}_{|t=0}=F,\quad \check{g}_{|t=0}=G & \text{in} \quad \mathbb{R}^2.\end{cases} \tag{4.20}$$

Using (4.14) one has

$$\mathfrak{E}(\check{f},\check{g})(t)+\int_0^t I(\check{f},\check{g})(\tau)\,\mathrm{d}\tau\leq\mathfrak{E}(F,G),$$

with

$$I(\check{f},\check{g})=\int_{\mathbb{R}^2}\left(\mu\nu^2\check{f}|\nabla\phi_{\check{f}}|^2+\check{g}|\nabla\phi_{\check{g}}|^2\right)\mathrm{d}x,$$

where $\phi_{\check{f}}=\check{f}+\check{g}+b/\mu$ and $\phi_{\check{g}}=\nu\check{f}+\check{g}+b$. Since (F,G) is a minimizer of $\mathfrak{E}$ and $I\geq 0$, then $I(\check{f},\check{g})=0$ a.e and hence $\partial_t\check{f}=\partial_t\check{g}=0$. We deduce that $\check{f}=F,\check{g}=G$ and $I(F,G)=0$. Hence (F,G) satisfy the system (4.18). We conclude that the minimizers of $\mathfrak{E}$ are stationary solutions of (4.9) and satisfy (4.18). □

Let $(F,G)\in\mathcal{K}_{M_f}\times\mathcal{K}_{M_g}$ be a solution of (4.18) and we recall the positivity sets E_F and E_G of F and G defined above

$$E_F=\{x\in\mathbb{R}^2\ :\ F(x)>0\},\quad E_G=\{x\in\mathbb{R}^2\ :\ G(x)>0\}.$$

Lemma 4.3.8. *Let (F,G) be a solution of (4.18). Then*

(i) $E_F \cap E_G \neq \emptyset$,

(ii) the stationary states (F, G) *of (4.9) are locally Lipschitz continuous and radial.*

Proof. Assume for contradiction that $E_F \cap E_G = \emptyset$. Then $F\nabla G = G\nabla F = 0$. Therefore (F, G) satisfy the equations

$$\begin{cases} F\nabla(F + b/\mu) &= 0 \quad \text{in} \quad \mathbb{R}^2, \\ G\nabla(G + b) &= 0 \quad \text{in} \quad \mathbb{R}^2, \end{cases}$$

hence F and G are Barenblatt solution centred at 0. Thus $0 \in E_F \cap E_G = \emptyset$, yielding a contradiction.

Let us prove that (F, G) are locally Lipschitz continuous and radial. We have

$$\mathbb{R}^2 = (E_F \cap E_G) \cup (E_F \cap E_G^c) \cup (E_F^c \cap E_G) \cup (E_F^c \cap E_G^c).$$

Let (F, G) be a solution to (4.18). Using Stampacchia's theorem that says that if $u \in H^1$, then $\nabla u = 0$ a.e on $\{u = k \in \mathbb{R}\}$, thereby one has

$$\nabla F = 0 \text{ on } E_F^c, \quad \text{and } \nabla G = 0 \text{ on } E_G^c.$$

Therefore

$$\nabla G = -\nabla b \text{ a.e on } E_F^c \cap E_G, \ \nabla F = -\nabla b/\mu \text{ a.e on } E_G^c \cap E_F, \tag{4.21}$$

$$\nabla F = \frac{\mu - 1}{\mu(1 - \nu)}\nabla b \text{ a.e on } E_F \cap E_G, \ \nabla G = \frac{\nu - \mu}{\mu(1 - \nu)}\nabla b \text{ a.e on } E_F \cap E_G. \tag{4.22}$$

Since $\nabla b(x) = \dfrac{1 - \nu}{4} x \in L^\infty_{\text{loc}}(\mathbb{R}^2)$, ∇F and ∇G are collinear to x, thus F and G are radial and locally Lipschitz continuous. □

According to the discussion above the profiles (F, G) of self-similar solutions of (4.1) defined in (4.12) are stationary solutions of (4.9) and satisfy (4.18). Moreover (F, G) are radial. Thanks to (4.21)-(4.22) one has :

- On $E_F \cap E_G$, there are $(C_1, C_2) \in \mathbb{R}^2$ such that

$$F(r) = C_1 + \frac{\mu - 1}{8\mu} r^2, \qquad G(r) = C_2 + \frac{\nu - \mu}{8\mu} r^2. \tag{4.23}$$

- On $E_F \cap E_G^c$, there is $C_3 \in \mathbb{R}$ such that

$$F(r) = C_3 + \frac{\nu - 1}{8\mu} r^2, \qquad G(r) = 0. \tag{4.24}$$

- On $E_G \cap E_F^c$, there is $C_4 \in \mathbb{R}$ such that

$$G(r) = C_4 + \frac{\nu - 1}{8} r^2, \qquad F(r) = 0. \tag{4.25}$$

Lemma 4.3.9. *We have $0 \in E_F \cup E_G$, E_F and E_G are connected sets and bounded.*

Proof. Thanks to formulas (4.23)-(4.25), either F or G is nonincreasing. Assume that F is nonincreasing. The case where G is nonincreasing is similar. If $0 \notin E_F$, then $F = 0$ and $M_f = 0$. This is contradicts the assumption (4.13). Since F is nonincreasing, E_F is connected (it is an interval containing 0).

Assume for contradiction that E_G is not connected. Thanks to formulas (4.23)-(4.25), G is decreasing on $[r_1, r_2]$ and increasing on $[r_3, r_4]$ with $r_1 < r_2 \leq r_3 < r_4$. We have $(r_3, r_4) \subset E_F$, but $(r_1, r_2) \subset E_F^c$. This contradicts the fact that E_F is an interval containing 0.

Now we will prove that every E_F and E_G are bounded. One has

$$\begin{aligned}\int_{\mathbb{R}^2} F(x)\,\mathrm{d}x = \int_{E_F} F(x)\,\mathrm{d}x = M_f < +\infty \\ = \int_{E_F \cap E_G} F(x)\,\mathrm{d}x + \int_{E_F \cap E_G^c} F(x)\,\mathrm{d}x.\end{aligned}$$

Using (4.23)-(4.24), one has

$$\int_{\mathbb{R}^2} F(x)\,\mathrm{d}x = \int_{E_F \cap E_G} C_1 + \frac{\mu - 1}{8\mu} r^2\,\mathrm{d}x + \int_{E_F \cap E_G^c} C_3 + \frac{\nu - 1}{8\mu} r^2\,\mathrm{d}x.$$

Then E_F cannot have an unbounded connected component. For E_G the argument is similar. □

As a consequence of Lemmas 4.3.8 and 4.3.9 we get that F and G are compactly supported and globally Lipchitz continuous.

4.4 Self-similar profiles

In this section we will classify the stationary solutions. We define the critical values of μ by

$$\mu_1^\star = \frac{\nu^2 M_f}{M_g + \nu(M_f - M_g)}, \quad \mu_2^\star = \frac{\nu M_f + M_g}{M_f + M_g}, \quad \mu_3^\star = 1 + (1 - \nu)\frac{M_f}{M_g}. \tag{4.26}$$

It is easy to check that

$$0 < \mu_1^\star < \nu < \mu_2^\star < 1 < \mu_3^\star.$$

It follows from the above discussion that only four configurations are possible for the stationary solutions. Figure 4.2 illustrate these four configurations. We will now justify the values of the viscosity μ corresponding to each configuration.

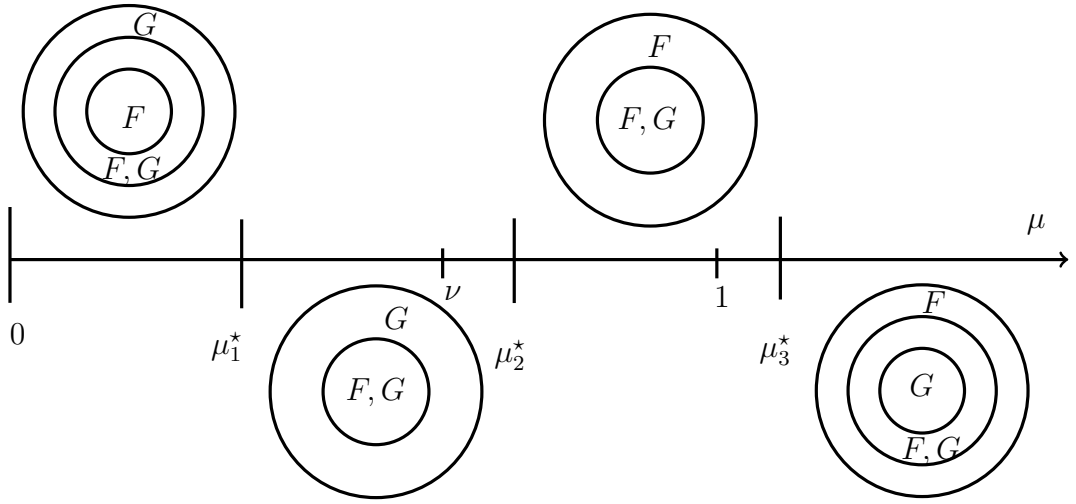


FIGURE 4.2 – Different values of μ and stationary profiles corresponding

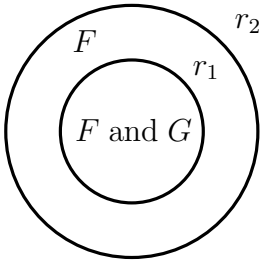


FIGURE 4.3 – The first configuration

First case : the first configuration we consider is shown on Figure 4.3. We note that $0 \in E_F \cap E_G$. In this situation we have : $F(0) = C_1 > 0$, $F(r_2) = 0$, F is continuous in $r = r_1$ and

$$\int_{\mathbb{R}^2} F \,\mathrm{d}x = \int_{E_F} F \,\mathrm{d}x = \int_{E_F \cap E_G} F \,\mathrm{d}x + \int_{E_F \cap E_G^c} F \,\mathrm{d}x = M_f.$$

This implies that

$$C_3 = \frac{1-\nu}{8\mu} r_2^2, \tag{4.27}$$

$$C_1 + \frac{\mu - 1}{8\mu} r_1^2 = C_3 - \frac{1-\nu}{8\mu} r_1^2, \tag{4.28}$$

and by passing in polar coordinates we have

$$\int_0^{r_1} \left[C_1 + \frac{\mu-1}{8\mu} r^2 \right] r \,\mathrm{d}r + \int_{r_1}^{r_2} \left[C_3 - \frac{1-\nu}{8\mu} r^2 \right] r \,\mathrm{d}r = \frac{M_f}{2\pi}.$$

Then one has

$$\frac{C_1}{2} r_1^2 + \frac{\mu-1}{32\mu} r_1^4 + \frac{C_3}{2}[r_2^2 - r_1^2] - \frac{1-\nu}{32\mu}[r_2^4 - r_1^4] = \frac{M_f}{2\pi}. \tag{4.29}$$

We have also : G decrease then $\nu < \mu$ and

$$\int_{\mathbb{R}^2} G \,\mathrm{d}x = \int_{E_G} G \,\mathrm{d}x = \int_{E_F \cap E_G} G \,\mathrm{d}x + \int_{E_G \cap E_F^c} G \,\mathrm{d}x = M_g.$$

By passing in polar coordinates one has that

$$\int_0^{r_1} \left[C_2 + \frac{\nu - \mu}{8\mu} r^2 \right] r \,\mathrm{d}r = \frac{C_2}{2} r_1^2 + \frac{\nu - \mu}{32\mu} r_1^4 = \frac{M_g}{2\pi}. \tag{4.30}$$

Moreover $G(r_1) = 0$, then

$$C_2 = \frac{\mu - \nu}{8\mu} r_1^2. \tag{4.31}$$

Using (4.31) and (4.30) one has

$$r_1^4 = \frac{16\mu M_g}{\pi(\mu - \nu)} > 0. \tag{4.32}$$

Multiplying (4.28) by $\frac{r_1^2}{2}$, we get

$$\frac{C_1}{2} r_1^2 - \frac{C_3}{2} r_1^2 = \frac{\nu - \mu}{16\mu} r_1^4,$$

and using (4.27) and (4.29) we obtain

$$r_2^4 = r_1^4 + \frac{16\mu}{\pi(1-\nu)} \left[M_f + M_g \frac{\mu - 1}{\mu - \nu} \right] = \frac{16\mu}{\pi(1-\nu)} (M_f + M_g). \tag{4.33}$$

Then

$$r_2^4 > r_1^4 \Longleftrightarrow \mu > \frac{\nu M_f + M_g}{M_f + M_g}. \tag{4.34}$$

Using (4.27), (4.28) and (4.32), we get

$$C_1 = \frac{1-\mu}{8\mu} r_1^2 + \frac{1-\nu}{8\mu}(r_2^2 - r_1^2),$$

and by (4.33) one has

$$C_1 > 0 \Longleftrightarrow \mu < 1 + (1-\nu)\frac{M_f}{M_g}. \tag{4.35}$$

We conclude that we have the first case if and only if

$$\mu_2^\star < \mu < \mu_3^\star,$$

with

$$\mu_2^\star = \frac{\nu M_f + M_g}{M_f + M_g} \quad \text{and} \quad \mu_3^\star = 1 + (1-\nu)\frac{M_f}{M_g}.$$

Remark that if $\mu = \mu_2^\star$ then $r_1 = r_2$.

Second case : the second configuration we consider is shown on Figure 4.4. We note that $0 \in E_F \cap E_G$. This case is similar to the first case, where the roles of F

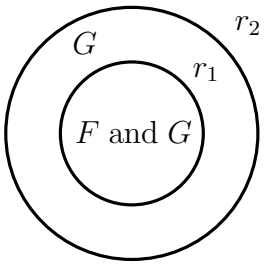


FIGURE 4.4 – The second configuration

and G have been exchanged. Following the same computations, we get

$$r_1^4 = \frac{16\mu M_f}{\pi(1-\mu)} > 0,$$

$$r_2^4 = r_1^4 + \frac{16 M_g}{\pi(1-\nu)} + \frac{\nu-\mu}{\mu(1-\nu)} r_1^4 = r_1^4 + \frac{16}{\pi(1-\nu)}\Big[M_g + M_f \frac{\nu-\mu}{1-\mu}\Big],$$

and

$$C_2 = \frac{\mu-\nu}{8\mu} r_1^2 + \frac{1-\nu}{8}(r_2^2 - r_1^2).$$

We conclude that we have the second case if and only if

$$\mu_1^\star < \mu < \mu_2^\star,$$

with

$$\mu_1^\star = \frac{\nu^2 M_f}{M_g + \nu(M_f - M_g)} \quad \text{and} \quad \mu_2^\star = \frac{\nu M_f + M_g}{M_f + M_g}.$$

Third case : we consider the configuration shown on Figure 4.5. We note that $0 \in E_F$. In this situation we have : F is nonincreasing and G is nondecreasing on

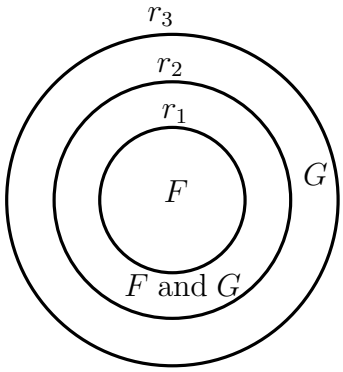


FIGURE 4.5 – The third configuration

$E_F \cap E_G$, then $\mu < \nu < 1$. Since

$$\frac{\nu^2 M_f}{M_g + \nu(M_f - M_g)} < \nu,$$

and that the case

$$\frac{\nu^2 M_f}{M_g + \nu(M_f - M_g)} < \mu < \nu,$$

corresponds to the second case, then we must necessarily have the third case, which corresponds to

$$\mu \leq \frac{\nu^2 M_f}{M_g + \nu(M_f - M_g)} = \mu_1^\star.$$

Fourth case : we consider the last configuration shown on Figure 4.6. We note that $0 \in E_G$.

In this situation we have : G is nonincreasing and F is nondecreasing on $E_F \cap E_G$, then $\nu < \mu$ and $\mu > 1$. Since

$$1 + (1 - \nu)\frac{M_f}{M_g} > 1,$$

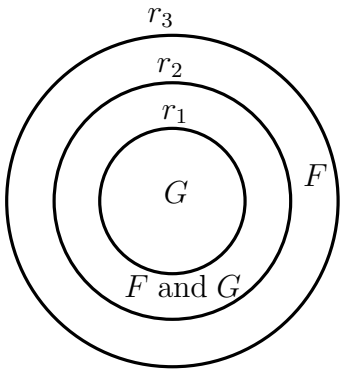


FIGURE 4.6 – The fourth configuration

and that the case

$$1 \leq \mu < 1 + (1-\nu)\frac{M_f}{M_g},$$

corresponds to the first case, then we must necessarily have the fourth case, which corresponds to

$$\mu \geq 1 + (1-\nu)\frac{M_f}{M_g} = \mu_3^\star.$$

4.5 Numerical investigation

In this section we present the results of several numerical simulations realized in the context of the rescaled system (4.9) in order to study the large time behaviour. It consists into the study of the convergence in relative energy of the evolutive solutions towards the equilibrium state. Moreover, we want to recover the rate of convergence towards equilibrium which appears in different tests cases. The existence of a relative energy turns out to be an essential ingredient in the convergence towards a stationary state. Indeed, we can define the relative energy $\mathfrak{E}(f,g|F,G)$ of a solution (f,g) to (4.9) w.r.t the stationary state (F,G) by

$$\mathfrak{E}(f,g|F,G) = \int_{\mathbb{R}^2} \Big(E(f,g) - E(F,G) \Big)\, \mathrm{d}x = \mathfrak{E}(f,g) - \mathfrak{E}(F,G).$$

Since (F,G) is a minimizer of $\mathfrak{E}$, then

$$\mathfrak{E}(f,g|F,G) = \mathfrak{E}(f,g) - \mathfrak{E}(F,G) \geq 0.$$

4.5.1 The numerical scheme

We explicit the discretization of the problem (4.9) we will use for the numerical simulations. The time discretization relies on backward Euler scheme, while the

space discretization relies on a finite volume approach (see e.g [78]), with a two-points flux approximation and an upstream mobility.

Let $\Omega \subset \mathbb{R}^2$ be an open, bounded, polygonal subset. An admissible mesh of Ω is given by a family $\mathcal{T}$ of a control volumes (open and convex polygons), a family $\mathcal{E}$ of edges and a family of points $(x_K)_{K\in\mathcal{T}}$ which satisfy Definition 9.1 in [78]. This definition implies that the straight line between two neighboring centers of cells (x_K, x_L) is orthogonal to the edge $\sigma = K|L$.

We denote by $\sigma \in \mathcal{E}_{\text{int}}$ the interior edges. For a control volume $K \in \mathcal{T}$, we denote by $\mathcal{E}_K$ the set of its edges, by $\mathcal{E}_{K,\text{int}}$ the set of its interior edges.

Furthermore, we denote by d the distance in $\mathbb{R}^2$ and by m the Lebesgue measure in $\mathbb{R}^2$ or $\mathbb{R}$. We assume that the family of meshes satisfies the following regularity requirement : there exists $\zeta > 0$ such that for all $K \in \mathcal{T}$ and all $\sigma \in \mathcal{E}_{\text{int},K}$ with $\sigma = K|L$, it holds

$$d(x_K, \sigma) \geq \zeta d(x_K, x_L) \tag{4.36}$$

For all $\sigma \in \mathcal{E}_{\text{int},K}$ with $\sigma = K|L$, we define $d_\sigma = d(x_K, x_L)$, and the transmissibility coefficient

$$\tau_\sigma = \frac{m(\sigma)}{d_\sigma}, \quad \sigma \in \mathcal{E}. \tag{4.37}$$

The size of the mesh is defined by

$$\delta = \max_{K\in\mathcal{T}}(\text{diam}(K)).$$

In order to avoid heavier notations, we restrict our study to the case of a uniform time discretization of $(0, T)$. However, all the results presented in this paper can be extended to general time discretizations without any technical difficulty. In what follows, we assume that the spatial mesh is fixed and does not change with the time step. Let $T > 0$ be some final time and M_T the number of time steps. Then the time step size and the time points are given by, respectively,

$$\Delta t = \frac{T}{M_T}, \quad t^n = n\Delta t, \quad 0 \leq n \leq M_T.$$

We denote by $\mathcal{D}$ an admissible space-time discretization of $\Omega_T = \Omega \times (0, T)$ composed of an admissible mesh $\mathcal{T}$ of Ω and the values Δt and M_T. The size of this space-time discretization $\mathcal{D}$ is defined by $\eta = \max(\delta, \Delta t)$.

The initial conditions are discretized by

$$f_\mathcal{T}^0 = \sum_{K\in\mathcal{T}} f_K^0 \mathbf{1}_K, \quad \text{where } f_K^0 = \frac{1}{m(K)} \int_K f_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T}, \tag{4.38}$$

$$g_\mathcal{T}^0 = \sum_{K\in\mathcal{T}} g_K^0 \mathbf{1}_K, \quad \text{where } g_K^0 = \frac{1}{m(K)} \int_K g_0(x)\,\mathrm{d}x, \quad \forall K \in \mathcal{T}, \tag{4.39}$$

and $\mathbf{1}_K$ is the characteristic function on K. Denoting by f_K^n and g_K^n approximations of the mean value of $f(.,t^n)$ and $g(.,t^n)$ on K, respectively. Taking for b_K the value of b in a fixed point of K (for instance, the center of gravity of K), where b is a regular function and assume that $b_K \geq 0 \quad \forall K \in \mathcal{T}$. The discretization of problem (4.9) is given by the following set of nonlinear equations :

$$m(K)\frac{f_K^{n+1}-f_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\text{int},K}} \tau_\sigma f_\sigma^{n+1}\nu\mu\left((f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+\frac{1}{\mu}(b_K-b_L)\right)=0, \tag{4.40}$$

and

$$m(K)\frac{g_K^{n+1}-g_K^n}{\Delta t} + \sum_{\sigma\in\mathcal{E}_{\text{int},K}} \tau_\sigma g_\sigma^{n+1}\left(\nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\right)=0, \tag{4.41}$$

for $K \in \mathcal{T}$ and $0 \leq n \leq M_T - 1$, where

$$f_\sigma^{n+1} = \begin{cases} (f_K^{n+1})^+ & \text{if } (f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+\dfrac{1}{\mu}(b_K-b_L)\geq 0, \\ (f_L^{n+1})^+ & \text{if } (f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+\dfrac{1}{\mu}(b_K-b_L)< 0. \end{cases} \tag{4.42}$$

and

$$g_\sigma^{n+1} = \begin{cases} (g_K^{n+1})^+ & \text{if } \nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)\geq 0, \\ (g_L^{n+1})^+ & \text{if } \nu(f_K^{n+1}-f_L^{n+1})+(g_K^{n+1}-g_L^{n+1})+(b_K-b_L)< 0, \end{cases} \tag{4.43}$$

where $x^+ = \max(0,x)$. The discretization of problem (4.18) is given by the following set of nonlinear equations :

$$\sum_{\sigma\in\mathcal{E}_{\text{int},K}} \tau_\sigma F_\sigma^{n+1}\nu\mu\left((F_K^{n+1}-F_L^{n+1})+(G_K^{n+1}-G_L^{n+1})+\frac{1}{\mu}(b_K-b_L)\right)=0, \tag{4.44}$$

and

$$\sum_{\sigma\in\mathcal{E}_{\text{int},K}} \tau_\sigma G_\sigma^{n+1}\left(\nu(F_K^{n+1}-F_L^{n+1})+(G_K^{n+1}-G_L^{n+1})+(b_K-b_L)\right)=0, \tag{4.45}$$

where F_σ^{n+1} and G_σ^{n+1} are defined as below. We next define the numerical approximation $(f_\mathcal{D}, g_\mathcal{D})$ of (f,g) by

$$f_\mathcal{D}(x,t) = \sum_{\substack{K\in\mathcal{T}\\ 0\leq n\leq M_T-1}} f_K^{n+1}\mathbf{1}_{K\times(t^n,t^{n+1}]}(x,t), \quad \text{and}$$

$$g_\mathcal{D}(x,t) = \sum_{\substack{K\in\mathcal{T}\\ 0\leq n\leq M_T-1}} g_K^{n+1}\mathbf{1}_{K\times(t^n,t^{n+1}]}(x,t).$$

We introduce a discrete version of energy functional :

$$\mathfrak{E}^n := \mathfrak{E}(f_K^n, g_K^n) = \sum_{K\in\mathcal{T}} m(K)\left[\frac{\nu}{2}\Big(f_K^n + g_K^n\Big)^2 + \frac{1-\nu}{2}(g_K^n)^2 + b_K\Big(\frac{\nu}{\mu}f_K^n + g_K^n\Big)\right].$$

In Chapter 3 we have proved that this scheme preserves at the discrete level the main features of the continuous problem, namely the nonnegativity of the solutions, the decay of the energy and the control of the entropy and its dissipation. And based on these nonlinear stability results, it has been shown that this scheme converges towards a weak solution to the problem.

4.5.2 Numerical simulations

Our scheme leads to a nonlinear system that we solve thanks to the Newton-Raphson method. In our test case, the domain is the unit square, i.e., $\Omega = (0,1)^2$. We consider an admissible triangular mesh made of 14336 triangles. We use a meshe coming from the 2D benchmark on anisotropic diffusion problems [103]. For the evolutive solutions and in order to increase the robustness of the algorithm and to ensure the convergence of the Newton-Raphson iterative procedure, we used an adaptive time step procedure in the practical implementation. More precisely, we associate a maximal time step $\Delta t_{\max} = 0.0002$ for the mesh . If the Newton-Raphson method fails to converge after 30 iterations —we choose that the ℓ^∞ norm of the residual has to be smaller than 10^{-9} as stopping criterion—, the time step is divided by two. If the Newton-Raphson method converges, the first time step is multiplied by two and projected on $[0, \Delta t_{\max}]$. The first time step Δt is equal to $\Delta t_{\max}$ in the test case presented below.

We perform the numerical experiments with the following data

$$b(x,y) = 30\frac{1-\nu}{8}\Big((x-1/2)^2 + (y-1/2)^2\Big), \quad \nu = 0.9,$$

and as an initial condition we take

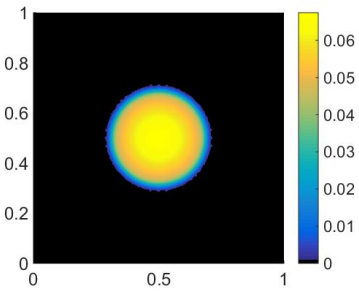
$$f_0(x,y) = \Big(\frac{1}{16} - (x-2/7)^2 - (y-2/7)^2\Big)_+,\ g_0(x,y) = \Big(\frac{1}{16} - (x-5/7)^2 - (y-5/7)^2\Big)_+.$$

In this case we have $M_f = M_g$ then

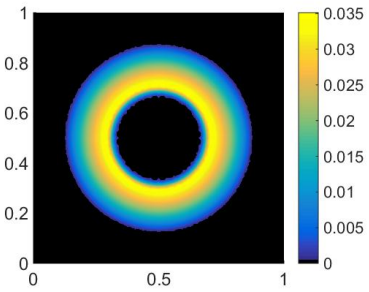
$$\mu_1^\star = 0.81, \quad \mu_2^\star = 0.95 \quad \text{and} \quad \mu_3^\star = 1.1.$$

Note that the f_0 and g_0 are not radial.

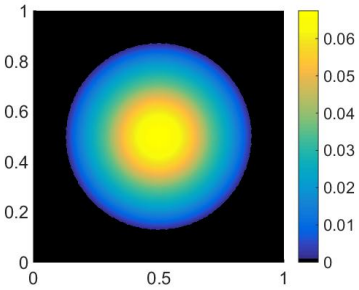
We represent in Figure 4.7 to 4.17 the self-similar profiles and the decay of the discrete relative energy. Following the values of μ these figures confirm the discussion above on the shape of the steady states.



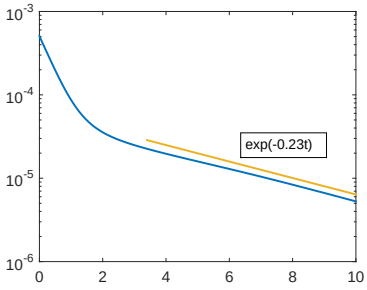
(a) Profile F



(b) Profile G

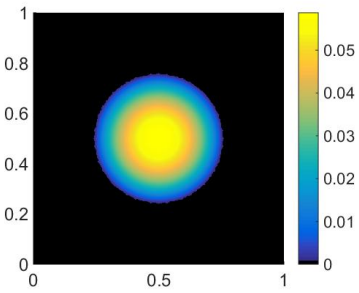


(c) Profile $F + G$

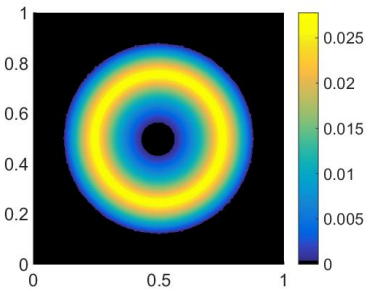


(d) Discrete relative energy

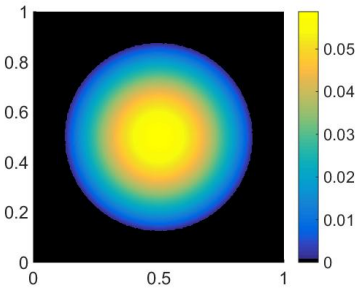
FIGURE 4.7 – Self-similar profiles and the relative energy for $\mu = 0.50$



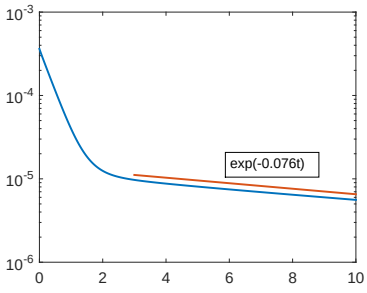
(a) Profile F



(b) Profile G

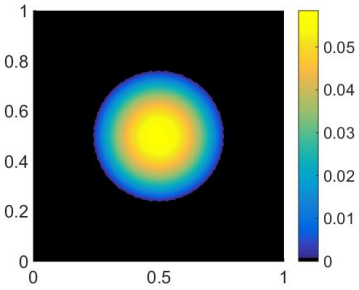


(c) Profile $F + G$

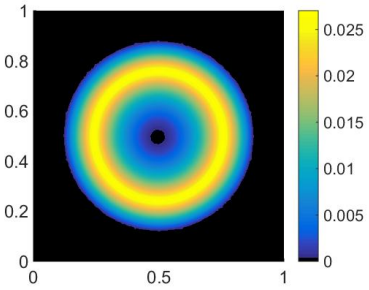


(d) Discrete relative energy

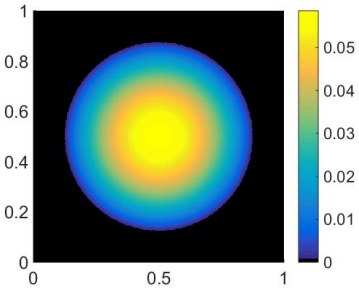
FIGURE 4.8 – Self-similar profiles and the relative energy for $\mu = 0.80$



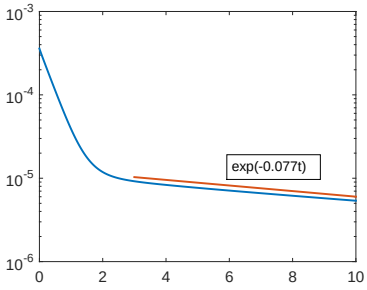
(a) Profile F



(b) Profile G

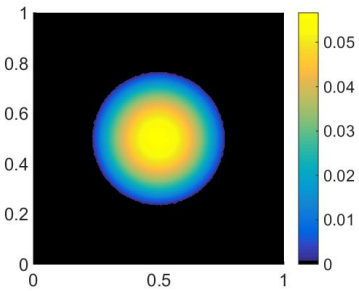


(c) Profile $F + G$

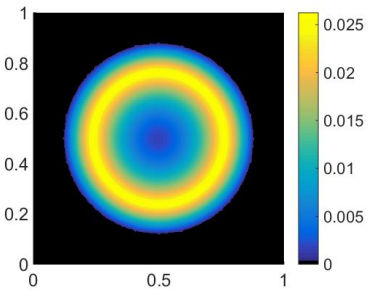


(d) Discrete relative energy

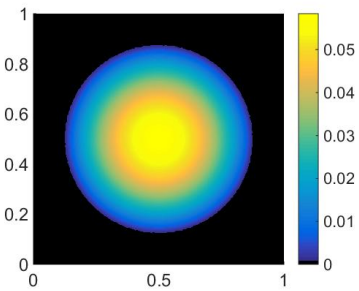
FIGURE 4.9 – Self-similar profiles and the relative energy for $\mu = \mu_1^\star = 0.81$



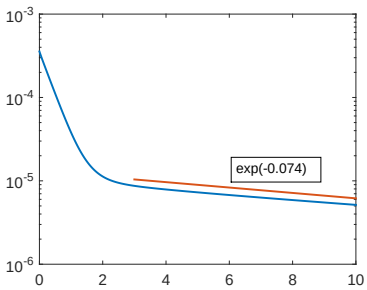
(a) Profile F



(b) Profile G

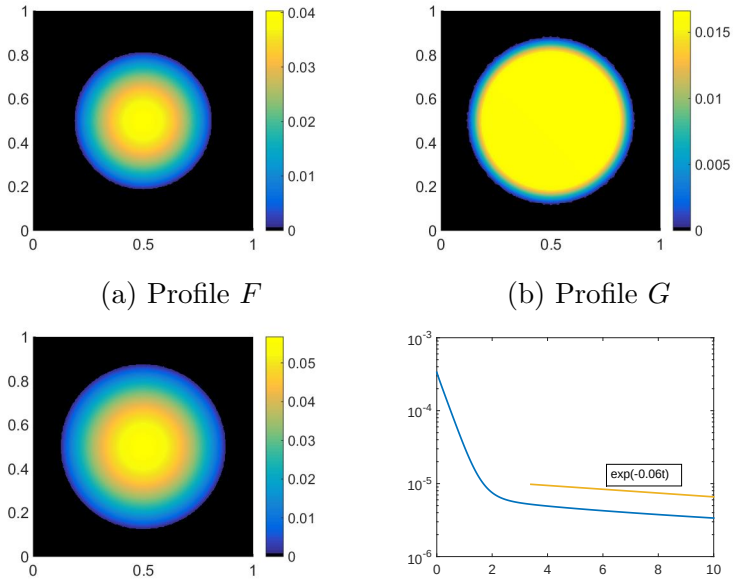


(c) Profile $F + G$



(d) Discrete relative energy

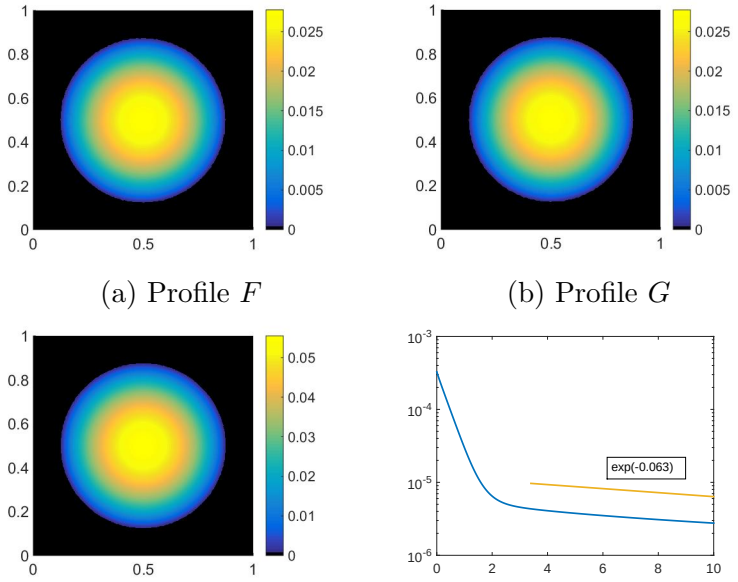
FIGURE 4.10 – Self-similar profiles and the relative energy for $\mu = 0.82$



(a) Profile F (b) Profile G

(c) Profile $F + G$ (d) Discrete relative energy

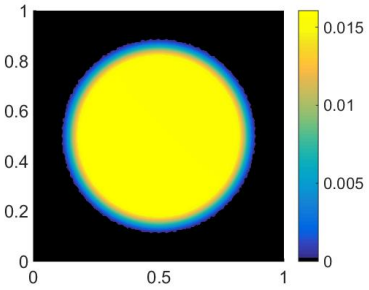
FIGURE 4.11 – Self-similar profiles and the relative energy for $\mu = \nu = 0.90$



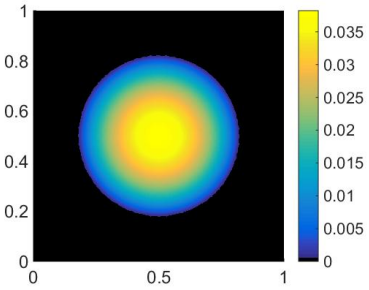
(a) Profile F (b) Profile G

(c) Profile $F + G$ (d) Discrete relative energy

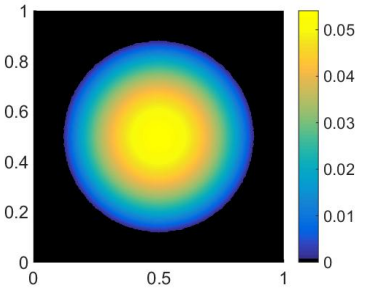
FIGURE 4.12 – Self-similar profiles and the relative energy for $\mu = \mu_2^\star = 0.95$. It is the case when $r_1 = r_2$ in the first and second case



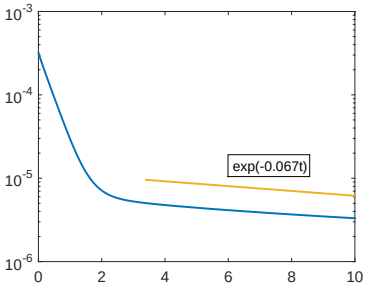
(a) Profile F



(b) Profile G

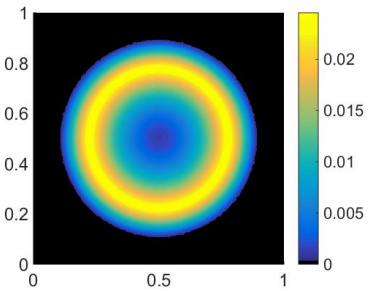


(c) Profile $F + G$

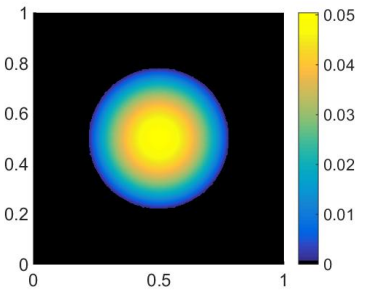


(d) Discrete relative energy

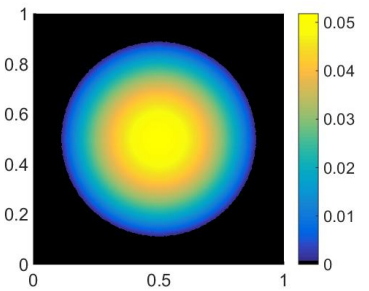
FIGURE 4.13 – Self-similar profiles and the relative energy for $\mu = 1$



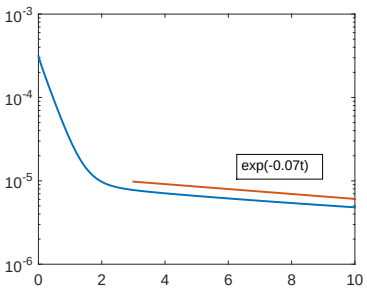
(a) Profile F



(b) Profile G

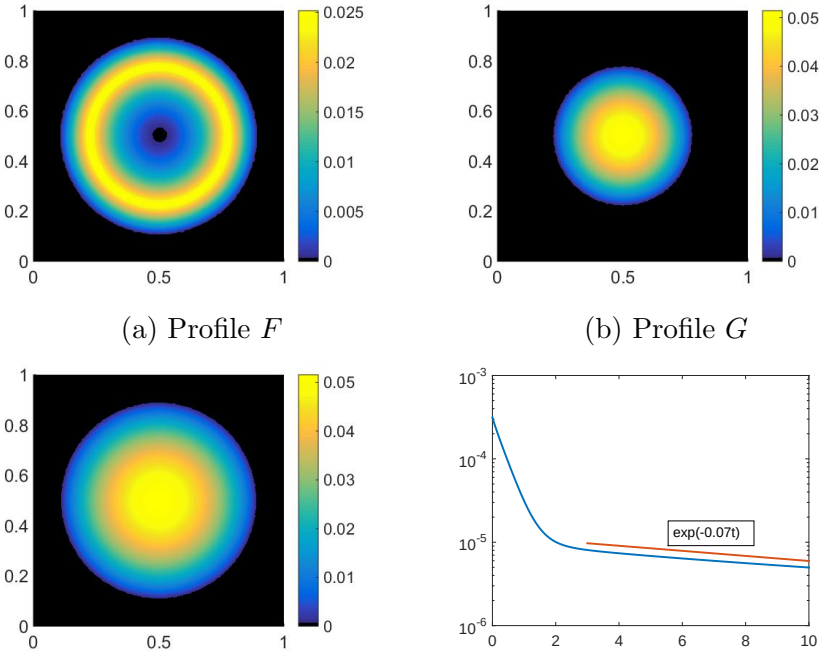


(c) Profile $F + G$



(d) Discrete relative energy

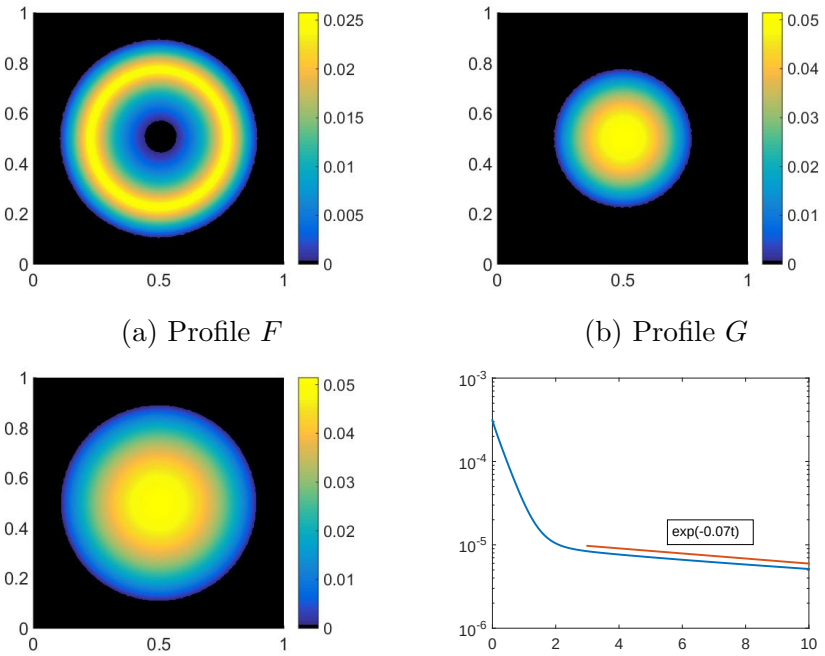
FIGURE 4.14 – Self-similar profiles and the relative energy for $\mu = 1.09$



(a) Profile F (b) Profile G

(c) Profile $F + G$ (d) Discrete relative energy

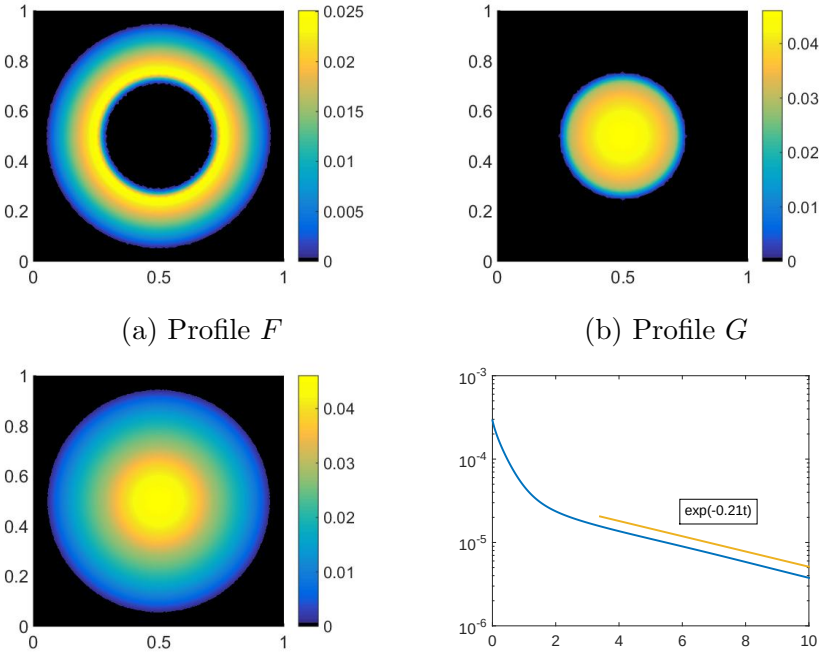
FIGURE 4.15 – Self-similar profiles and the relative energy for $\mu = \mu_3^\star = 1.1$



(a) Profile F (b) Profile G

(c) Profile $F + G$ (d) Discrete relative energy

FIGURE 4.16 – Self-similar profiles and the relative energy for $\mu = 1.11$



(a) Profile F (b) Profile G

(c) Profile $F + G$ (d) Discret relative energy

FIGURE 4.17 – Self-similar profiles and the relative energy for $\mu = 2$

Figure 4.7 to 4.17 suggest that the convergence of the discrete solution to the scheme towards the discrete equilibrium occurs at exponential rate. More precisely we have

$$\mathfrak{E}(f, g|F, G) \leq C \exp(-p(\mu)t), \tag{4.46}$$

where the rate $p(\mu)$ strongly depends on μ. We plot on Figure 4.18 the function $\mu \mapsto p(\mu)$ obtained experimentally. At its minimum, the function p is close to 0. This prohibits to conclude to the exponential convergence whatever $\mu \in (0, +\infty)$ for the continuous model.

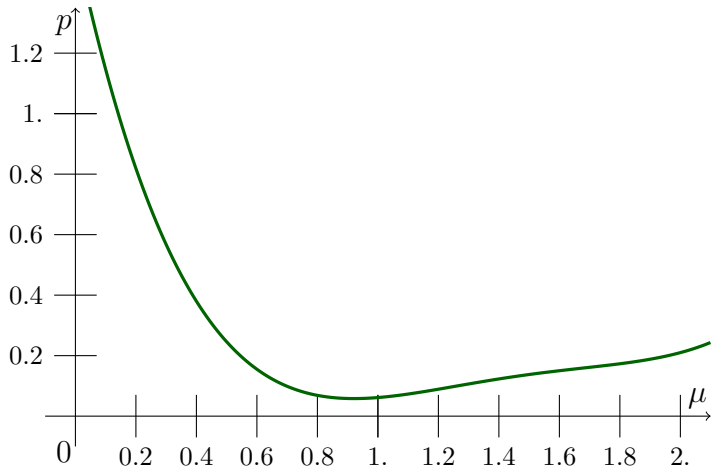


FIGURE 4.18 – The rate of convergence p in (4.46) following the values of μ

Bibliographie

[1] I. Aavatsmark, T. Barkve, O. . Boe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. I. Derivation of the methods. *SIAM J. Sci. Comput.*, 19(5) :1700–1716, 1998.

[2] A. Abudawia and C. Rosier. Numerical analysis for a seawater intrusion problem in a confined aquifer. *Math. Comput. Simulation*, 118 :2–16, 2015.

[3] G. Acosta and R. G. Durán. An optimal Poincaré inequality in L^1 for convex domains. *Proc. Amer. Math. Soc.*, 132(1) :195–202 (electronic), 2004.

[4] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Academic press (Elsevier), 2nd edition, 2003.

[5] L. Agélas, D. A. Di Pietro, and J. Droniou. The G method for heterogeneous anisotropic diffusion on general meshes. *M2AN Math. Model. Numer. Anal.*, 44(4) :597–625, 2010.

[6] L. Agélas, C. Guichard, and R. Masson. Convergence of finite volume MPFA O type schemes for heterogeneous anisotropic diffusion problems on general meshes. *Int. J. Finite Vol.*, 7(2) :33, 2010.

[7] A. Ait Hammou Oulhaj. A finite volume scheme for a seawater intrusion model with cross-diffusion. In *Finite volumes for complex applications VIII—methods and theoretical aspects*, volume 199 of *Springer Proc. Math. Stat.*, pages 421–429. Springer, Cham, 2017.

[8] A. Ait Hammou Oulhaj. Numerical analysis of a finite volume scheme for a seawater intrusion model with cross-diffusion in an unconfined aquifer. HAL : hal-01432197, submitted, 2017.

[9] A. Ait Hammou Oulhaj, C. Cancès, and C. Chainais-Hillairet. Numerical analysis of a nonlinearly stable and positive Control Volume Finite Element scheme for Richards equation with anisotropy. HAL : hal-01372954, 2016.

[10] A. Al Bitar. Modélisation des écoulements en milieu poreux hétérogenes 2d/3d, avec couplages surface/souterrain et densitaires. 2007. Thèse Institut National Polytechnique de Toulouse.

[11] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3) :311–341, 1983.

[12] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[13] B. Andreianov, M. Bendahmane, and R. Ruiz-Baier. Analysis of a finite volume method for a cross-diffusion model in population dynamics. *Math. Models Methods Appl. Sci.*, 21(02) :307–344, 2011.

[14] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic–elliptic PDEs. *J. Funct. Anal.*, 273(12) :3633–3670, 2017.

[15] T. Arbogast, M. F. Wheeler, and N.-Y. Zhang. A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media. *SIAM J. Numer. Anal.*, 33(4) :1669–1687, 1996.

[16] A. Arnold, J. A. Carrillo, L. Desvillettes, J. Dolbeault, A. Jüngel, C. Lederman, P. A. Markowich, G. Toscani, and C. Villani. Entropies and equilibria of many-particle systems : an essay on recent research. *Monatsh. Math.*, 142(1-2) :35–43, 2004.

[17] A. Arnold, P. Markowich, G. Toscani, and A. Unterreiter. On convex Sobolev inequalities and the rate of convergence to equilibrium for Fokker-Planck type equations. *Comm. Partial Differential Equations*, 26(1-2) :43–100, 2001.

[18] A. Arnold and A. Unterreiter. Entropy decay of discretized Fokker-Planck equations. I. Temporal semidiscretization. *Comput. Math. Appl.*, 46(10-11) :1683–1690, 2003.

[19] J. Bear. *Dynamic of fuilds in Porous media*. American Elsevier, New York, 1972.

[20] J. Bear. Hydraulics of groundwater, McGraw-Hill series in water resources and environmental engineering. 1979.

[21] J. Bear, A. H.-D. Cheng, S. Sorek, D. Ouazar, and I. Herrera. *Seawater intrusion in coastal aquifers : concepts, methods and practices*, volume 14. Springer Science & Business Media, 1999.

[22] M. Bessemoulin-Chatard and C. Chainais-Hillairet. Exponential decay of a finite volume scheme to the thermal equilibrium for drift–diffusion systems. *Journal of Numerical Mathematics*, 2016.

[23] M. Bessemoulin-Chatard, C. Chainais-Hillairet, and F. Filbet. On discrete functional inequalities for some finite volume schemes. *IMA J. Numer. Anal.*, 35 :1125–1149, 2015.

[24] M. Bessemoulin-Chatard and A. Jüngel. A finite volume scheme for a Keller-Segel model with additional cross-diffusion. *IMA J. Numer. Anal.*, 34(1) :96–122, 2014.

[25] F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8) :2430–2457, 2012.

[26] F. Bolley, I. Gentil, and A. Guillin. Uniform convergence to equilibrium for granular media. *Arch. Ration. Mech. Anal.*, 208(2) :429–445, 2013.

[27] F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Numer. Anal.*, 46(6) :3032–3070, 2008.

[28] K. Brenner and C. Cancès. Improving Newton's method performance by parametrization : the case of Richards equation. HAL : hal-01342386, submitted for publication.

[29] K. Brenner, D. Hilhorst, and H. C. Vu Do. A gradient scheme for the discretization of Richards equation. In *Finite volumes for complex applications. VII. Elliptic, parabolic and hyperbolic problems*, volume 78 of *Springer Proc. Math. Stat.*, pages 537–545. Springer, Cham, 2014.

[30] K. Brenner, D. Hilhorst, and H.-C. Vu-Do. The generalized finite volume SUSHI scheme for the discretization of Richards equation. *Vietnam J. Math.*, 44(3) :557–586, 2016.

[31] K. Brenner and R. Masson. Convergence of a Vertex centred Discretization of Two-Phase Darcy flows on General Meshes. working paper or preprint, Nov. 2012.

[32] H. Brezis. *Analyse fonctionnelle*. Masson, Paris, 1983. Théorie et applications.

[33] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 43(5) :1872–1896, 2005.

[34] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 15(10) :1533–1551, 2005.

[35] R. H. Brooks and A. T. Corey. Hydraulic properties of porous media and their relation to drainage design. *Transactions of the ASAE*, 7(1) :0026–0028, 1964.

[36] C. Buet and S. Cordier. On the non existence of monotone linear schema for some linear parabolic equations. *C. R. Math. Acad. Sci. Paris*, 340(5) :399–404, 2005.

[37] C. Cancès, C. Chainais-Hillairet, and S. Krell. A nonlinear discrete duality finite volume scheme for convection-diffusion equations. In *Finite volumes for complex applications VIII—methods and theoretical aspects*, volume 199 of *Springer Proc. Math. Stat.*, pages 439–447. Springer, Cham, 2017.

[38] C. Cancès and C. Guichard. Entropy-diminishing CVFE scheme for solving anisotropic degenerate diffusion equations. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 187–196. Springer, Cham, 2014.

[39] C. Cancès and C. Guichard. Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.*, 85(298) :549–580, 2016.

[40] C. Cancès and C. Guichard. Numerical analysis of a robust free energy-diminishing Finite Volume scheme for degenerate parabolic equations with gradient structure. *Found. Comput. Math.*, 2016. online first, DOI : 10.1007/s10208-016-9328-6.

[41] C. Cancès and M. Pierre. An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.*, 44(2) :966–992, 2012.

[42] E. A. Carlen and S. Ulusoy. An entropy dissipation-entropy estimate for a thin film type equation. *Commun. Math. Sci.*, 3(2) :171–178, 2005.

[43] J. A. Carrillo, S. Hittmeir, and A. Jüngel. Cross diffusion and nonlinear diffusion preventing blow up in the Keller-Segel model. *Math. Models Methods Appl. Sci.*, 22(12) :1250041, 35, 2012.

[44] J. A. Carrillo and G. Toscani. Exponential convergence toward equilibrium for homogeneous Fokker-Planck-type equations. *Math. Methods Appl. Sci.*, 21(13) :1269–1286, 1998.

[45] J. A. Carrillo and G. Toscani. Asymptotic L^1-decay of solutions of the porous medium equation to self-similarity. *Indiana Univ. Math. J.*, 49(1) :113–142, 2000.

[46] C. Chainais-Hillairet. Entropy method and asymptotic behaviours of finite volume schemes. In *Finite volumes for complex applications. VII. Methods and theoretical aspects*, volume 77 of *Springer Proc. Math. Stat.*, pages 17–35. Springer, Cham, 2014.

[47] C. Chainais-Hillairet, A. Jüngel, and S. Schuchnigg. Entropy-dissipative discretization of nonlinear diffusion equations and discrete Beckner inequalities. *ESAIM Math. Model. Numer. Anal.*, 50(1) :135–162, 2016.

[48] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *M2AN Math. Model. Numer. Anal.*, 37(2) :319–338, 2003.

[49] M. Chatard. Asymptotic behavior of the Scharfetter-Gummel scheme for the drift-diffusion model. In *Finite volumes for complex applications VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proc. Math.*, pages 235–243. Springer, Heidelberg, 2011.

[50] G. Chavent and J. Jaffré. Dynamics of fluids in porous media, 1986.

[51] L. Chen and A. Jüngel. Analysis of a multidimensional parabolic population model with strong cross-diffusion. *SIAM J. Math. Anal.*, 36(1) :301–322, 2004.

[52] L. Chen and A. Jüngel. Analysis of a parabolic cross-diffusion population model without self-diffusion. *J. Differential Equations*, 224(1) :39–59, 2006.

[53] L. Chen and A. Jüngel. Analysis of a parabolic cross-diffusion semiconductor model with electron-hole scattering. *Comm. Partial Differential Equations*, 32(1-3) :127–148, 2007.

[54] Z. Chen and R. Ewing. Mathematical analysis for reservoir models. *SIAM J. Math. Anal.*, 30(2) :431–453, 1999.

[55] C. Choquet. Parabolic and degenerate parabolic models for pressure-driven transport problems. *Math. Models Methods Appl. Sci.*, 20(4) :543–566, 2010.

[56] C. Choquet, M. M. Diédhiou, and C. Rosier. Mathematical analysis of a sharp–diffuse interfaces model for seawater intrusion. *J. Differential Equations*, 259(8) :3803–3824, 2015.

[57] C. Choquet, M. M. Diédhiou, and C. Rosier. Derivation of a sharp-diffuse interfaces model for seawater intrusion in a free aquifer. numerical simulations. *SIAM J. Appl. Math.*, 76(1) :138–158, 2016.

[58] C. Choquet, J. Li, and C. Rosier. Global existence for seawater intrusion models : comparison between sharp interface and sharp-diffuse interface approaches. *Electron. J. Differential Equations*, pages No. 126, 27, 2015.

[59] I. Csiszár. Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2 :299–318, 1967.

[60] H. Darcy. *Les fontaines publiques de la ville de Dijon : exposition et application...* Victor Dalmont, 1856.

[61] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.

[62] L. Desvillettes and K. Fellner. Exponential decay toward equilibrium via entropy methods for reaction-diffusion equations. *J. Math. Anal. Appl.*, 319(1) :157–176, 2006.

[63] L. Desvillettes and K. Fellner. Duality and entropy methods for reversible reaction-diffusion equations with degenerate diffusion. *Math. Methods Appl. Sci.*, 38(16) :3432–3443, 2015.

[64] L. Desvillettes, T. Lepoutre, A. Moussa, and A. Trescases. On the entropic structure of reaction-cross diffusion systems. *Comm. Partial Differential Equations*, 40(9) :1705–1747, 2015.

[65] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69. Springer Science & Business Media, 2011.

[66] K. Domelevo and P. Omnes. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *M2AN Math. Model. Numer. Anal.*, 39(6) :1203–1249, 2005.

[67] J. Droniou. Finite volume schemes for diffusion equations : introduction to and review of modern methods. *Math. Models Methods Appl. Sci.*, 24(8) :1575–1619, 2014.

[68] J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105(1) :35–71, 2006.

[69] J. Droniou and R. Eymard. Study of the mixed finite volume method for Stokes and Navier-Stokes equations. *Numer. Methods Partial Differential Equations*, 25(1) :137–171, 2009.

[70] M. G. Edwards and C. F. Rogers. Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Comput. Geosci.*, 2(4) :259–290 (1999), 1998.

[71] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[72] J. Escher, P. Laurençot, and B.-V. Matioc. Existence and stability of weak solutions for a degenerate parabolic system modelling two-phase flows in porous media. *Ann. Inst. H. Poincaré Anal. Non Linéaire.*, 28(4) :583–598, 2011.

[73] J. Escher, A.-V. Matioc, and B.-V. Matioc. Modelling and analysis of the Muskat problem for thin fluid layers. *J. Math. Fluid Mech.*, 14(2) :267–277, 2012.

[74] J. Escher and B.-V. Matioc. Existence and stability of solutions for a strongly coupled system modelling thin fluid films. *NoDEA Nonlinear Differential Equations Appl.*, 20(3) :539–555, 2013.

[75] R. Eymard, T. Gallouët, M. Ghilani, and R. Herbin. Error estimates for the approximate solutions of a nonlinear hyperbolic equation given by finite volume schemes. *IMA J. Numer. Anal.*, 18(4) :563–594, 1998.

[76] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. Tp or not tp, that is the question. *Computational Geosciences*, 18(3) :285–296, Aug 2014.

[77] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. *Comput. Geosci.*, 18 :285–296, 2014.

[78] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In *Handbook of numerical analysis, Vol. VII*, Handb. Numer. Anal., VII, pages 713–1020. North-Holland, Amsterdam, 2000.

[79] R. Eymard, T. Gallouët, and R. Herbin. A new finite volume scheme for anisotropic diffusion problems on general grids : convergence analysis. *C. R. Math. Acad. Sci. Paris*, 344(6) :403–406, 2007.

[80] R. Eymard, T. Gallouët, and R. Herbin. Discretization schemes for linear diffusion operators on general non-conforming meshes. In *Finite volumes for complex applications V*, pages 375–382. ISTE, London, 2008.

[81] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI : a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4) :1009–1043, 2010.

[82] R. Eymard, T. Gallouët, R. Herbin, M. Gutnic, and D. Hilhorst. Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies. *Math. Models Methods Appl. Sci.*, 11(09) :1505–1528, 2001.

[83] R. Eymard and T. Gallouët. H-convergence and numerical schemes for elliptic problems. *SIAM J. Numer. Anal.*, 41(2) :539–562, 2003.

[84] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. *Comput. Geosci.*, 3(3-4) :259–294 (2000), 1999.

[85] R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klofkorn, and G. Manzini. 3D Benchmark on Discretization Schemes for Anisotropic Diffusion Problems on General Grids. In *Finite Volume for Complex Applications VI*, pages 895–930, Praha, Czech Republic, June 2011. Springer.

[86] R. Eymard, R. Herbin, and A. Michel. Mathematical study of a petroleum-engineering scheme. *M2AN Math. Model. Numer. Anal.*, 37(6) :937–972, 2003.

[87] R. Eymard, D. Hilhorst, and M. Vohralí k. A combined finite volume–nonconforming/mixed-hybrid finite element scheme for degenerate parabolic problems. *Numer. Math.*, 105(1) :73–131, 2006.

[88] R. Eymard, D. Hilhorst, and M. Vohralí k. A combined finite volume-finite element scheme for the discretization of strongly nonlinear convection-diffusion-reaction problems on nonmatching grids. *Numer. Methods Partial Differential Equations*, 26(3) :612–646, 2010.

[89] F. Filbet. A finite volume scheme for the Patlak-Keller-Segel chemotaxis model. *Numer. Math.*, 104(4) :457–488, 2006.

[90] F. Filbet. An asymptotically stable scheme for diffusive coagulation-fragmentation models. *Commun. Math. Sci.*, 6(2) :257–280, 2008.

[91] P. A. Forsyth. A control volume finite element approach to NAPL groundwater contamination. *SIAM J. Sci. Statist. Comput.*, 12(5) :1029–1057, 1991.

[92] P. A. Forsyth and M. C. Kropinski. Monotonicity considerations for saturated–unsaturated subsurface flow. *SIAM J. Sci. Comput.*, 18(5) :1328–1354, 1997.

[93] P. A. Forsyth, Jr. and P. H. Sammon. Practical considerations for adaptive implicit methods in reservoir simulation. *J. Comput. Phys.*, 62(2) :265–281, 1986.

[94] H. Gajewski and K. Gärtner. On the discretization of van Roosbroeck's equations with magnetic field. *Z. Angew. Math. Mech.*, 76(5) :247–264, 1996.

[95] H. Gajewski and K. Gröger. On the basic equations for carrier transport in semiconductors. *J. Math. Anal. Appl.*, 113(1) :12–35, 1986.

[96] H. Gajewski and K. Gröger. Semiconductor equations for variable mobilities based on Boltzmann statistics or Fermi-Dirac statistics. *Math. Nachr.*, 140 :7–36, 1989.

[97] G. Galiano, A. Jüngel, and J. Velasco. A parabolic cross-diffusion system for granular materials. *SIAM J. Math. Anal.*, 35(3) :561–578, 2003.

[98] T. Gallouët and J.-C. Latché. Compactness of discrete approximate solutions to parabolic PDEs—application to a turbulence model. *Commun. Pure Appl. Anal.*, 11(6) :2371–2391, 2012.

[99] A. Glitzky. Exponential decay of the free energy for discretized electro-reaction-diffusion systems. *Nonlinearity*, 21(9) :1989–2009, 2008.

[100] A. Glitzky. Uniform exponential decay of the free energy for Voronoi finite volume discretized reaction-diffusion systems. *Math. Nachr.*, 284(17-18) :2159–2174, 2011.

[101] L. Gosse and G. Toscani. Identification of asymptotic decay to self-similarity for one-dimensional filtration equations. *SIAM J. Numer. Anal.*, 43(6) :2590–2606, 2006.

[102] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.

[103] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In *Finite volumes for complex applications V*, pages 659–692. ISTE, London, 2008.

[104] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, 160(2) :481–499, 2000.

[105] S. Hittmeir and A. Jüngel. Cross diffusion preventing blow-up in the two-dimensional Keller-Segel model. *SIAM J. Math. Anal.*, 43(2) :997–1022, 2011.

[106] U. Hornung and W. Messing. *Poröse medien : methoden und simulation.* Verlag Beiträge zur Hydrologie, 1984.

[107] R. Huber and R. Helmig. Node-centered finite volume discretizations for the numerical simulation of multiphase flow in heterogeneous porous media. *Comput. Geosci.*, 4(2) :141–164, 2000.

[108] M. Jazar and R. Monneau. Derivation of seawater intrusion models by formal asymptotics. *SIAM J. Appl. Math.*, 74(4) :1152–1173, 2014.

[109] B. Joe. Delaunay triangular meshes in convex polygons. *SIAM J. Sci. Statist. Comput.*, 7(2) :514–539, 1986.

[110] R. Jordan, D. Kinderlehrer, and F. Otto. Free energy and the Fokker-Planck equation. *Phys. D*, 107(2-4) :265–271, 1997. Landscape paradigms in physics and biology (Los Alamos, NM, 1996).

[111] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1) :1–17, 1998.

[112] A. Jüngel. The boundedness-by-entropy method for cross-diffusion systems. *Nonlinearity*, 28(6) :1963–2001, 2015.

[113] A. Jüngel. *Entropy methods for diffusive partial differential equations*. SpringerBriefs in Mathematics. Springer, [Cham], 2016.

[114] A. Jüngel and S. Schuchnigg. Entropy-dissipating semi-discrete Runge-Kutta schemes for nonlinear diffusion equations. *Commun. Math. Sci.*, 15(1) :27–53, 2017.

[115] A. Jüngel and I. V. Stelzer. Entropy structure of a cross-diffusion tumor-growth model. *Math. Models Methods Appl. Sci.*, 22(7) :1250009, 26, 2012.

[116] E. Keilegavlen, J. M. Nordbotten, and I. Aavatsmark. Sufficient criteria are necessary for monotone control volume methods. *Appl. Math. Lett.*, 22(8) :1178–1180, 2009.

[117] R. A. Klausen, F. A. Radu, and G. T. Eigestad. Convergence of MPFA on triangulations and for Richards' equation. *Int. J. Numer. Meth. Fl.*, 58(12) :1327–1351, 2008.

[118] S. Kullback. A lower bound for discrimination information in terms of variation (corresp.). *IEEE Trans. Inf. Theor.*, 13(1) :126–127, Sept. 2006.

[119] P. Laurençot and B.-V. Matioc. A gradient flow approach to a thin film approximation of the Muskat problem. *Calc. Var. Partial Differential Equations*, 47(1-2) :319–341, 2013.

[120] P. Laurençot and B.-V. Matioc. A thin film approximation of the Muskat problem with gravity and capillary forces. *J. Math. Soc. Japan*, 66(4) :1043–1071, 2014.

[121] P. Laurençot and B.-V. Matioc. Self-Similarity in a Thin Film Muskat Problem. *SIAM J. Math. Anal.*, 49(4) :2790–2842, 2017.

[122] J. Leray and J. Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup. (3)*, 51 :45–78, 1934.

[123] F. List and F. A. Radu. A study on iterative methods for solving Richards' equation. *Comput. Geosci.*, 20(2) :341–353, 2016.

[124] P. Marion, K. Najib, and C. Rosier. Numerical simulations for a seawater intrusion problem in a free aquifer. *Appl. Numer. Math.*, 75 :48–60, 2014.

[125] D. Matthes, R. J. McCann, and G. Savaré. A family of nonlinear fourth order equations of gradient flow type. *Comm. Partial Differential Equations*, 34(10-12) :1352–1397, 2009.

[126] A. Michel. A mathematical comparison of two finite volume methods for two phase flow in porous media. In *Finite volumes for complex applications, III (Porquerolles, 2002)*, pages 213–220. Hermes Sci. Publ., Paris, 2002.

[127] Y. Mualem. A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water Resour. Res.*, 12(3) :513–522, 1976.

[128] M. Muskat. Two fluid systems in porous media. the encroachment of water into an oil sand. *Journal of Applied Physics*, 5 :250–264, Sept. 1934.

[129] M. Muskat. The flow of homogeneous fluids through porous media. *Soil Science*, 46(2) :169, 1938.

[130] M. Muskat. The flow of homogeneous fluids through porous media. Technical report, 1946.

[131] K. Najib and C. Rosier. On the global existence for a degenerate elliptic–parabolic seawater intrusion problem. *Math. Comput. Simulation*, 81(10) :2282–2295, 2011.

[132] J. M. Nordbotten, I. Aavatsmark, and G. T. Eigestad. Monotonicity of control volume methods. *Numer. Math.*, 106(2) :255–288, 2007.

[133] F. Otto. L^1-contraction and uniqueness for quasilinear elliptic-parabolic equations. *J. Differential Equations*, 131 :20–38, 1996.

[134] F. Otto. The geometry of dissipative evolution equations : the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2) :101–174, 2001.

[135] I. S. Pop. Error estimates for a time discretization method for the Richards' equation. *Comput. Geosci.*, 6 :141–160, 2002.

[136] I. S. Pop, M. Sepúlveda, F. A. Radu, and O. P. Vera Villagrán. Error estimates for the finite volume discretization for the porous medium equation. *J. Comput. Appl. Math.*, 234(7) :2135–2142, 2010.

[137] F. Radu, I. S. Pop, and P. Knabner. Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation. *SIAM J. Numer. Anal.*, 42(4) :1452–1478, 2004.

[138] F. A. Radu, I. S. Pop, and P. Knabner. Newton-type methods for the mixed finite element discretization of some degenerate parabolic equations. In *Numerical mathematics and advanced applications*, pages 1192–1200. Springer, Berlin, 2006.

[139] L. A. Richards. Capillary conduction of liquids through porous mediums. *Physics*, 1(5) :318–333, 1931.

[140] L. A. Richards. Capillary conduction of liquids through porous mediums. *Journal of Applied Physics*, 1(5) :318–333, 1931.

[141] F. Santambrogio. *Optimal Transport for Applied Mathematicians : Calculus of Variations, PDEs, and Modeling.* Progress in Nonlinear Differential Equations and Their Applications 87. Birkhäuser Basel, 1 edition, 2015.

[142] J. Simon. Compact sets in the space $L^p(0,T;B)$. *Ann. Mat. Pura Appl. (4)*, 146 :65–96, 1987.

[143] O. Strack, R. Barnes, and A. Verruijt. Vertically integrated flows, discharge potential, and the Dupuit-Forchheimer approximation. *Ground water*, 44(1) :72–75, 2006.

[144] O. D. Strack. A Dupuit-Forchheimer model for three-dimensional flow with variable density. *Water Resources Research*, 31(12) :3007–3017, 1995.

[145] M. E. A. Talibi and M. H. Tber. Existence of solutions for a degenerate seawater intrusion problem. *Electron. J. Differential Equations*, 2005(72) :1–14, 2005.

[146] M. H. Tber, M. E. A. Talibi, and D. Ouazar. Parameters identification in a seawater intrusion model using adjoint sensitive method. *Math. Comput. Simulation*, 77(2) :301–312, 2008.

[147] G. Toscani and C. Villani. On the trend to equilibrium for some dissipative systems with slowly increasing a priori bounds. *J. Statist. Phys.*, 98(5-6) :1279–1309, 2000.

[148] M. T. Van Genuchten. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil science society of America journal*, 44(5) :892–898, 1980.

[149] J. L. Vázquez. *The porous medium equation.* Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford, 2007. Mathematical theory.

[150] N. Zamponi and A. Jüngel. Analysis of degenerate cross-diffusion population models with volume filling. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 34(1) :1–29, 2017.

[151] R. L. Zarba, E. T. Bouloutas, and M. Celia. General mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 26(7) :1483–1496, 1990.

[152] J. Zinsl and D. Matthes. Exponential convergence to equilibrium in a coupled gradient flow system modeling chemotaxis. *Anal. PDE*, 8(2) :425–466, 2015.