Couplage réactions-transport pour la modélisation et la simulation du stockage géologique de CO2

Elodie Tillier

UNIVERSITÉ DE MARNE-LA-VALLÉE
École Doctorale Information, Communication, Modélisation et Simulation
en collaboration avec

L'INSTITUT FRANÇAIS DU PÉTROLE
Direction Technologie, Informatique et Mathématiques Appliquées

THÈSE

pour obtenir le grade de
Docteur de l'université de Marne-La-Vallée

Discipline : Mathématiques Appliquées

présentée et soutenue publiquement par
Elodie TILLIER
le 25 septembre 2007

COUPLAGE REACTIONS-TRANSPORT POUR LA MODELISATION ET LA SIMULATION DU STOCKAGE GÉOLOGIQUE DE CO2

devant le jury composé de :

M. Robert EYMARD	Directeur de thèse
M. Fayssal BENKHALDOUN	Rapporteur
Mme. Danielle HILHORST	Rapporteur
M. Frédéric COQUEL	Examinateur
M. Damien LAMBERTON	Examinateur
M. Laurent TRENTY	Responsable IFP

UNIVERSITÉ DE MARNE-LA-VALLÉE
École Doctorale Information, Communication, Modélisation et Simulation
en collaboration avec

L'INSTITUT FRANÇAIS DU PÉTROLE
Direction Technologie, Informatique et Mathématiques Appliquées

THÈSE

pour obtenir le grade de
Docteur de l'université de Marne-La-Vallée

Discipline : Mathématiques Appliquées

présentée et soutenue publiquement par
Elodie TILLIER
le 25 septembre 2007

COUPLAGE REACTIONS-TRANSPORT POUR LA MODELISATION ET LA SIMULATION DU STOCKAGE GÉOLOGIQUE DE CO2

devant le jury composé de :

M. Robert EYMARD	Directeur de thèse
M. Fayssal BENKHALDOUN	Rapporteur
Mme. Danielle HILHORST	Rapporteur
M. Frédéric COQUEL	Examinateur
M. Damien LAMBERTON	Examinateur
M. Laurent TRENTY	Responsable IFP

Remerciements

Je tiens à remercier tout d'abord Robert Eymard, mon directeur de thèse, qui m'a accompagnée pendant ces trois ans. Sa disponibilité, sa patience et ses encouragements m'ont beaucoup aidée tout au long de cette thèse. Il a toujours su me remotiver quand c'était nécessaire, et me convaincre de la qualité de mes travaux. Nos séances de travail à la fac, au téléphone, ou sous le soleil de Marrakech resteront pour moi de très bons souvenirs.

Je remercie également Laurent Trenty et Anthony Michel qui m'ont encadrée au quotidien à l'IFP. Ils ont toujours pris le temps de répondre à mes questions et m'ont beaucoup appris. Leur bonne humeur a rendu cette thèse très agréable et j'espère avoir de nouveau l'occasion de travailler avec eux.

Je remercie Danielle Hilhorst et Fayssal Benkhaldoun d'avoir accepté d'être les rapporteurs de ma thèse ainsi que Frédéric Coquel et Damien Lamberton d'avoir accepté de participer à mon jury. Leur remarques sympathiques et leur bienveillance m'ont permis d'apprécier le grand jour de la soutenance sans trop de stress.

Je souhaite également remercier Roland Masson de m'avoir intégrée dans la division Mathématiques Appliquées de l'IFP et je remercie l'ensemble des membres du département pour leur accueil, et plus particulièrement tous les thésards qui ont largement contribué à rendre ces trois années sympatiques. Je n'oublierai pas nos longues conversations de la pause déjeuner et de tout ce qu'elles m'ont apprises (et oui, les pingouins volent ...). J'ai une pensée toute particulière pour Séverine avec qui j'ai débuté et terminé ma thèse. Je suis convaincue que la fin de nos thèses ne sera pas la fin de notre amitié. Un grand merci pour Greg, patient correcteur de mes nombreuses fautes d'ortographe. Il en retrouvera probablement quelques unes dans ces remerciements.

Finalement, je remercie mes proches, famille et amis, tout particulièrement mes parents pour leur soutien constant au cours de mes études. Ils ont toujours cru en ma réussite (je suis toujours passée pour l'intello de la famille, et le choix de faire une thèse n'a rien arrangé), sans eux tout cela aurait été beaucoup plus difficile et mes relations avec mon banquier seraient sans doute compliquées.

Un dernier mot pour Alexis, la fin de cette thèse va enfin me permettre d'aller poser mes valises chez toi et c'est peut-être pour ça que je suis si heureuse de la terminer.

Table des matières

Introduction

Le stockage du co_2

Les risques de changement climatique ont fait l'objet de nombreux débats au cours de ces dernières années. Actuellement, la plupart des experts estiment que ces risques sont réels et directement reliés aux émissions de gaz à effet de serre, et tout particulièrement de co_2. Les émissions de co_2 ont fortement augmenté au cours des récentes décennies, entraînant une croissance de la teneur en co_2 dans l'atmosphère. Cette augmentation de la concentration serait responsable de la tendance au réchauffement climatique déjà observée, et pourrait avoir dans l'avenir des conséquences dramatiques si aucune mesure n'est prise.

Le co_2 n'est pas le seul gaz à effet de serre. Le protocole de Kyoto en prend en compte 5 autres, dont notamment le méthane et le protoxyde d'azote. Cependant, le co_2 constitue le principal contributeur à l'effet de serre lié aux activités humaines, en raison des importantes quantités rejetées dans l'atmosphère.

La réduction des émissions de co_2, en particulier grâce à la réduction des consommations et aux technologies de capture, de transport et de stockage du co_2, constitue un défi technologique important. Les solutions pour lutter contre les émissions de co_2 sont de trois types :

- Réduction des consommations d'énergie.
- Mise en oeuvre de combustibles ou de carburants émettant moins de co_2 par unité d'énergie produite.
- Capture et stockage du co_2.

On s'intéresse dans cette thèse à cette troisième solution. Le principe est de capturer et stocker le co_2 dans des formations géologiques souterraines. Cette option est applicable à des installations fixes de production concentrée d'énergie.

Après sa capture, il faut pouvoir stocker le co_2 pour des durées importantes, pouvant au minimum couvrir la période pendant laquelle le problème des émissions de co_2 risque de demeurer critique, période qui ne devrait pas dépasser un à deux siècles. Par mesure de précaution, on envisage des solutions qui permettent d'effectuer le stockage sur des périodes pouvant atteindre des milliers d'années. C'est principalement en cela que la problématique du stockage géologique du co_2 diffère considérablement de celle du stockage de déchets tels que les éléments radioactifs. Le stockage au fond des océans fait partie des options qui ont été envisagées, mais une telle solution présente deux inconvénients majeurs : d'une part, le devenir à long terme d'un tel stockage demeure difficile à modéliser et incertain. D'autre part, on connaît mal l'impact d'une augmentation de la concentration en co_2 sur les écosystèmes marins. Le stockage géologique dans le sous-sol constitue donc la solution généralement préférée. Les principales options possibles sont alors les suivantes :

- Stockage dans des gisements de pétrole et de gaz épuisés.
- Stockage dans les veines de houille non exploitées.
- Stockage dans les aquifères salins profonds.

Dans ce contexte, la modélisation de l'injection et du stockage du co_2 a un rôle important à jouer, notamment pour l'étude des mécanismes d'écoulement, la caractérisation des sites et l'étude de la sécurité à long terme du stockage.

La phase de stockage se décompose en deux étapes. La première étape est une étape d'injection de l'ordre de quelques dizaines d'années, suivie d'une étape de stockage de l'ordre du millier d'années. Lors de l'étape d'injection, le domaine impacté par l'injection de co_2 est de faible étendue, car les vitesses de transport sont faibles (échelle réservoir). Lors de l'étape de stockage, la zone impactée peut être de grande étendue (échelle bassin).

Différents phénomènes sont mis en jeu lors de l'injection et du stockage du co_2 :

- Les phénomènes de transport : convection, diffusion et dispersion
- Les phénomènes chimiques : dissolution du co_2 dans l'eau, acidification de l'eau environnante, réactions entre l'eau et la roche.

Ces différents phénomènes se produisent à des échelles de temps très différentes (réactions chimiques instantanées, réactions cinétiques lentes, transport à faible vitesse).

Il existe un couplage très fort entre ces différents phénomènes. L'injection de co_2 et sa dissolution dans l'eau influencent fortement l'écoulement. L'équilibre chimique du milieu est modifié par la dissolution du co_2 dans l'eau, ce qui peut provoquer de nombreuses réactions chimiques, notamment des réactions de dissolution et de précipitation de la roche. Ces réactions peuvent modifier la porosité, la perméabilité, et ainsi changer les caractéristiques du milieu poreux, et donc les propriétés de l'écoulement.

Les modèles utilisés doivent être capables de simuler l'évolution du système sur différentes échelles de temps et d'espace (cf. figure 1). De plus, les simulations sont en 3D, avec un grand nombre d'inconnues (nécessaires pour la représentation d'un modèle géochimique) et un grand nombre de mailles.

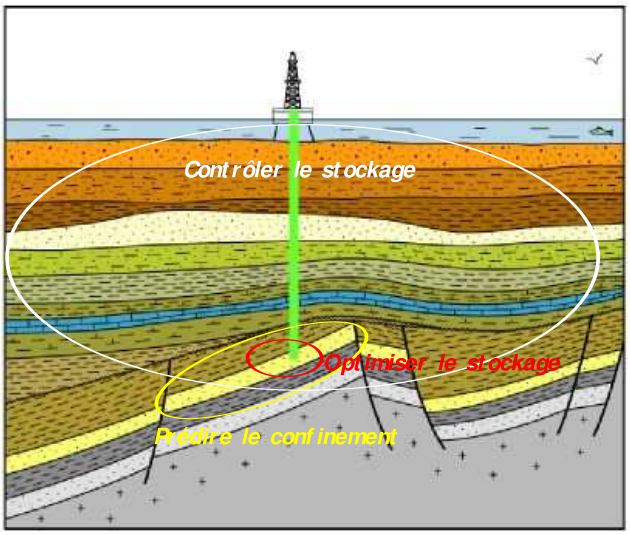


FIG. 1: Représentation schématique de la zone de stockage et des environs

Pour modéliser l'ensemble de ces phénomènes, on utilise généralement, de façon couplée, des modèles d'écoulements multiphasiques et des modèles géochimiques. On obtient ainsi un modèle dit modèle d'écoulements multiphasiques réactifs.

L'IFP développe actuellement un logiciel industriel, nommé COORES, permettant de simuler le stockage du co_2. C'est dans ce contexte que s'inscrit le travail de thèse. L'objectif est d'étudier différents schémas de couplage possibles, et tout particulièrement le schéma de couplage développé dans le logiciel COORES.

Solutions numériques

Deux grands types de méthodes existent pour résoudre les problèmes d'écoulement multiphasique réactif. Il s'agit des méthodes globales et des méthodes de splitting.

Le principe d'une méthode globale est de résoudre l'ensemble du problème de façon simultanée. L'avantage d'une résolution globale est de ne pas introduire d'erreur de splitting. En revanche ces méthodes ont l'inconvénient d'être coûteuses. En effet, pour résoudre le problème de façon simultanée, il faut linéariser l'ensemble du système. La matrice jacobienne ainsi obtenue est donc de très grande taille. Elle est difficile à stocker et coûteuse à inverser. Néanmoins, certaines études actuelles proposent des méthodes de résolution de type Newton-Krylov, qui ne nécessitent plus ni d'inverser ni de stocker la matrice jacobienne. Le second inconvénient des méthodes globales est qu'elles ne sont pas modulaires. La résolution étant simultanée, il est nécessaire de développer un code spécifique au problème traité.

Le principe d'une méthode de splitting est de séparer les différents opérateurs afin de pouvoir les résoudre séparément. Les méthodes de splitting permettent de réduire le coût de simulation et sont modulaires mais elles nécessitent de gérer l'erreur introduite par le splitting. En développant des méthodes générales de couplage de modèles, on peut considérer qu'il est plus facile d'ajouter un nouveau module à un code existant, et ceci permet également de réutiliser les logiciels préexistants. La qualité et le coût de la simulation dépend de la méthode utilisée pour coupler les différents modèles. Les différents phénomènes étant fortement couplés, la réduction de l'erreur de splitting peut s'avérer difficile et coûteuse.

La méthode principalement étudiée au cours de cette thèse est une méthode de splitting séquentielle non itérative appelée (Rs-Tr). Le principe de cette méthode est d'effectuer un splitting qui repose sur des hypothèses physiques empiriques. On suppose que les échanges eau-gaz influencent fortement l'écoulement et que par conséquent le calcul des flux ne peut être découplé du calcul correspondant à ces échanges. A l'inverse, on suppose que les cinétiques de précipitation-dissolution et les réactions en phase aqueuse agissent faiblement et lentement sur l'écoulement et par conséquent il est possible de les calculer séparément. Le splitting est donc un schéma à deux étapes, premièrement l'étape réservoir, (Rs), qui consiste à calculer les flux et les échanges eau-gaz, deuxièmement une étape de transport réactif, (Tr), qui consiste à calculer les autres réactions chimiques dans un champ de vitesse donné.

La deuxième caractéristique de cette méthode est la façon dont sont couplés les deux modèles (Rs) et (Tr). Il s'agit d'une méthode de couplage par pénalisation. Le principe est de pénaliser l'un des modèles par l'erreur de splitting. On réduit ainsi progressivement l'erreur de splitting sans avoir besoin d'itérer entre les deux étapes (Rs) et (Tr).

Organisation et contenu de la thèse

Cette thèse se divise en cinq parties. La première partie est consacrée à la modélisation. Dans la deuxième partie, les schémas numériques étudiés au cours de cette thèse sont présentés. La troisième partie contient la mise en oeuvre de ces différents schémas numériques soit dans des situations d'injection, soit dans des cas simplifiés permettant d'étudier des propriétés des schémas. L'ensemble des résultats numériques obtenus est décrit dans cette partie. Finalement la quatrième partie est consacrée à l'étude mathématique d'un système d'équations hyperboliques pour la modélisation des écoulements multiphasiques miscibles. Plus précisément, le contenu des différentes partie est :

- **Partie I**

  La première partie de cette thèse est consacrée à la modélisation des écoulements multiphasiques en milieu poreux et à la modélisation mathématique d'un système géochimique.

  Le premier chapitre présente les caractéristiques usuelles d'un milieu poreux, les lois de caractérisation des fluides, puis les équations de bilan habituellement utilisées pour la modélisation des

écoulements des fluides en milieu poreux.

Le second chapitre de cette partie est une présentation de la modélisation d'un système géochimique sous une forme mathématique, et la façon dont il est couplé avec le modèle d'écoulement. La modélisation d'un système géochimique consiste à décrire la composition des différents fluides, la composition de la roche et les réactions qui peuvent ce produire entre ces différents constituants chimiques. Il faut également définir comment modéliser ces réactions afin de les inclure dans le modèle d'écoulement. Les réactions sont de deux types, réaction cinétique ou réaction équilibrée. Le modèle géochimique ainsi obtenu peut être réduit en déterminant des espèces primaires, secondaires et cinétiques. Cette réduction se fait à l'aide d'opérations matricielles sur la matrice de réaction. Le principe est d'abord d'éliminer les réactions redondantes, puis d'utiliser ces réactions pour exprimer certaines espèces en fonction des autres. Le but est d'obtenir un système de taille réduite, mais suffisant pour représenter l'ensemble du système.

La réunion des deux modèles présentés, c'est à dire le modèle d'écoulement et le modèle géochimique, permet d'obtenir un modèle couplé pour la modélisation des écoulements multiphasiques réactifs. Ce modèle est dit "couplé" car il intègre les interactions de la géochimie sur les écoulements.

- **Partie II**

La deuxième partie de cette thèse est consacrée à l'étude de plusieurs algorithmes pour résoudre un problème couplé d'écoulements multiphasiques réactifs.

Le premier chapitre est une revue bibliographique d'un panel de solutions habituellement utilisées pour résoudre ce type de problèmes. Les deux grandes catégories de solutions sont soit des méthodes globales (résolution simultanée), soit des méthodes de splitting.

Le deuxième chapitre présente l'écriture d'un schéma numérique permettant une résolution globale du problème. L'élimination des inconnues secondaires est détaillée.

Le troisième chapitre porte sur la deuxième méthode étudiée au cours de cette thèse, à savoir la méthode de splitting (Rs-Tr). Il s'agit de la méthode actuellement développée à l'IFP dans un contexte industriel. Les hypothèses physiques qui ont déterminé les choix faits pour découpler les différents phénomènes seront présentées, ainsi que le schéma numérique obtenu. La résolution numérique est alors effectuée en couplant deux modèles différents, et dépend donc de la façon de résoudre chaque modèle et de la façon de coupler ces modèles (transmission des informations, réduction de l'erreur de splitting). Nous présenterons également dans ce chapitre la méthode de couplage par pénalisation. Cette méthode consiste à ajouter un terme de pénalisation sur le modèle (Rs). Ce terme est égal à l'erreur de splitting multiplié par un coefficient λ qui permet de paramétrer la rapidité à laquelle on souhaite corriger l'erreur.

Finalement, le quatrième chapitre de cette partie est une analyse de convergence du schéma pénalisé pour un problème simplifié. L'étude se limite à un problème diffusif et réactif, et permet de montrer que le schéma pénalisé converge bien vers la même solution faible que le schéma global. La preuve de convergence est obtenue à l'aide d'estimations qui permettent de passer à la limite sur le paramètre λ.

- **Partie III**

La troisième partie regroupe différentes mises en oeuvre des schémas présentés dans la partie précédente pour des situations d'injection de co_2 gazeux, ou pour des cas simplifiés. On présente l'ensemble des résultats numériques obtenus dans ces différentes cas.

On s'intéresse dans le premier chapitre à la validation des différentes hypothèses. On souhaite tout

d'abord valider les simplifications faites lors de la modélisation d'un système géochimique. En étudiant un système géochimique réaliste décrit dans Gunter et al. (1997), on montre que malgré les simplifications réalisées, le modèle utilisé permet d'obtenir des résultats satisfaisants. Dans un deuxième temps, on étudie des situations permettant de valider les hypothèses empiriques qui ont guidé les choix faits pour le découplage du problème d'écoulement multiphasique réactif.

Le deuxième chapitre est une étude de quelques solutions de couplage des modèles sur un problème 0D, pour lequel on dispose d'une solution analytique. La solution actuellement choisie à l'IFP est la correction d'erreur par pénalisation. Nous étudions le fonctionnement de cette méthode, son effet sur le système, et nous la comparons avec une méthode itérative classique. Les résultats obtenus suggèrent d'apporter une amélioration à cette méthode. Nous proposons alors de combiner la méthode par pénalisation avec une méthode itérative, le but étant d'associer la rapidité de la méthode de correction par pénalisation avec la précision de la méthode itérative.

Dans un troisième chapitre, on étudie les performances et les qualités respectives de la méthode globale et de la méthode de splitting (*Rs-Tr*), sur un cas 1D d'injection de co_2. On étudie la convergence numérique en temps et en espace de chacune des deux méthodes.

Le dernier chapitre concerne l'étude des phénomènes de diffusion-dispersion. On distingue la diffusion moléculaire, et la dispersion. Le terme de dispersion est un terme particulièrement difficile à appréhender, tant d'un point de vue physique que numérique. La dispersion est liée à l'homogénéisation des lois de convection et aux hétérogénéités du milieu poreux lui-même (aussi bien d'un point de vue microscopique que macroscopique). D'un point de vue numérique, il s'agit d'un terme difficile à discrétiser car c'est un phénomène qui ne se produit pas dans la direction des axes, mais dans la direction de la convection. La quatrième partie de cette thèse met en avant l'importance de ce terme et la difficulté de le prendre en compte dans un schéma numérique découplé.

- **Partie IV**

Les différentes situations physiques rencontrées au cours de cette thèse nous ont amené à réfléchir sur un problème d'écoulement multiphasique miscible. L'écriture des lois de conservation correspondantes permet d'obtenir un système d'équations hyperboliques très particulier. L'étude de ce système a fait l'objet d'un article, écrit en collaboration avec Robert Eymard, qui constitue la dernière partie de cette thèse.

Le problème étudié modélise la dissolution dans l'eau et la migration d'une bulle de co_2 gazeux, dans un système 1D vertical. L'écriture des lois de conservation, de l'eau et du co_2, permet d'obtenir un système hyperbolique particulier dont les inconnues sont la quantité totale de co_2 (sous forme de gaz et dissous dans l'eau), et le gradient de pression. La quantité d'eau s'écrit alors comme une fonction non linéaire de la quantité de co_2. Le système hyperbolique étudié est particulier car il ne contient pas de dérivée en temps par rapport à la seconde inconnue.

Si l'on choisit une solubilité nulle du co_2, on retrouve le cas habituel des écoulements immiscibles, traité de façon abondante dans la littérature. Les difficultés du problème étudié ici, proviennent du fait que cette solubilité n'est pas nulle. En effet, lorsque la solubilité du co_2 est nulle, la quantité d'eau est une fonction linéaire de la quantité de co_2, et il est possible de résoudre le système d'équation de façon classique par élimination de l'inconnue correspondant au gradient de pression.

Nous donnons une définition d'une solution faible, inspirée de la condition de Liu pour les chocs admissibles, et des paires entropiques de Krushkov. Nous prouvons ensuite, dans le cas d'une généralisation naturelle du problème de Riemann, l'existence d'une solution faible. Cette propriété découle de l'existence d'une certaine fonction permettant d'obtenir simultanément une solution faible entropique classique pour chacune des deux équations de façon couplée. L'existence d'une telle fonction est démontrée à l'aide de la convergence d'un algorithme de recherche d'enveloppes

convexes et concaves simultanées pour chacune des deux équations.

Nous prouvons ensuite l'existence d'une solution faible pour un ensemble général de données initiales, grâce à la preuve de convergence d'un schéma volume fini. Le principe de ce schéma est de calculer le flux de Godunov numérique à chaque interface, tout en respectant la conservation simultanée des deux équations.

L'ensemble de ce travail n'a pas permis de prouver l'unicité de la solution, ni de traiter le problème des conditions aux limites. Le schéma proposé est en effet un schéma défini sur un domaine infini. Un travail commun avec Robert Eymard et Julien Vovelle est en cours sur la prise en compte des conditions limites dans le schéma numérique. Quelques pistes de réflexion sont données dans le second chapitre.

Première partie

Modélisation

Pour modéliser le stockage géologique de co_2, on s'intéresse à l'influence d'une injection de co_2 sur le milieu poreux. Lorsqu'on injecte du co_2 dans un milieu poreux, on crée un déplacement des fluides en place. De plus, on modifie l'équilibre chimique du milieu, notamment sous l'effet de la dissolution du co_2 gazeux dans l'eau. Ceci signifie que la composition de l'eau est modifiée et qu'elle peut alors réagir avec la roche. En effet, la roche est composée de plusieurs minéraux qui peuvent précipiter ou se dissoudre selon la composition de l'eau environnante. La précipitation et la dissolution des minéraux modifie alors la porosité du milieu poreux. On a donc une interaction forte entre les écoulements et les réactions chimiques.

Pour modéliser l'ensemble de ces phénomènes, il faut tout d'abord modéliser les écoulements de fluides dans le milieu poreux et l'influence des réactions chimiques sur ces écoulements. Les écoulements sont modélisés par des équations de bilan, auxquelles s'ajoute l'effet des réactions chimiques sous la forme d'un terme source.

La deuxième étape est de décrire précisément la composition des fluides, la composition de la roche et les réactions qui peuvent se produire entre ces divers constituants chimiques. On définit également comment modéliser ces réactions de façon à pouvoir les intégrer dans les équations d'écoulement. L'ensemble de ces données constitue un système géochimique.

La modélisation des écoulements et la modélisation du système géochimique débouche alors sur un problème continu couplé, appelé problème d'écoulement multiphasique réactif. Ce problème est dit "couplé" car il intègre les interactions de la géochimie sur les écoulements.

On présente dans le chapitre 1 la modélisation des écoulements, puis la modélisation du système géochimique dans le chapitre 2, pour obtenir finalement le problème continu couplé (section 2.5).

Chapitre 1

Modélisation des écoulements multiphasiques en milieu poreux

Un milieu poreux est un matériau formé d'une partie solide (la matrice) et de vides (les pores). Les milieux auxquels on s'intéresse en géologie, sont les roches qui composent les couches supérieures du sous-sol. La partie poreuse est remplie par différents fluides, dont on souhaite modéliser l'écoulement. On utilise pour cela des lois homogènes issues de la mécanique des fluides, ce qui permet d'écrire des équations de conservation de masse. Ces équations dépendent des caractéristiques du milieu poreux (perméabilité, viscosité), des caractéristiques des fluides (composition, viscosité, densité) et des interactions entre les différents fluides et la roche.

La roche est composée de différentes phases minérales, caractérisées par leur composition ou leur cristallographie. On suppose que la roche est incompressible et immobile (pas de phénomène de compaction). Seules les réactions chimiques de type précipitation-dissolution peuvent modifier la porosité du milieu.

On suppose également que la température est constante.

1.1 Caractéristiques du milieu poreux

Une caractéristique essentielle d'un milieu poreux est la porosité, notée Φ. La porosité est le rapport entre le volume de pore et le volume total.

$$\Phi = \frac{\text{Volume de pore}}{\text{Volume total}} \tag{1.1}$$

L'autre caractéristique importante d'un milieu poreux est la perméabilité, notée K. Elle caractérise l'aptitude de la roche à laisser circuler à travers ses pores un fluide dont elle est saturée. Pour un gradient de pression donné, plus la roche est perméable, plus le flux sera important. La perméabilité s'exprime en m^2 (unité S.I). Il existe de nombreux modèles reliant la perméabilité à la porosité, mais on suppose ici que la perméabilité est constante.

La roche est composée de minéraux, chaque minéral m occupant une certaine fraction du volume de roche, notée F_m et définie par

$$F_m = \frac{V_{\text{minéral}}}{V_{\text{roche}}}. \tag{1.2}$$

On a donc $\sum_m F_m = 1$. Habituellement, on préfère utiliser la fraction volumique ϕ_m d'un minéral m pour caractériser le volume d'un minéral plutôt que F_m. La fraction volumique est définie par

$$\phi_m = \frac{V_{\text{minéral}}}{V_{\text{total}}}. \tag{1.3}$$

L'équation de fermeture du volume total s'écrit alors

$$\Phi + \sum_m \phi_m = 1 \tag{1.4}$$

La densité molaire d'un minéral m, notée ξ_m, est le nombre de moles par unité de volume. Elle permet de traduire en terme de quantité de matière, le volume occupé par un minéral.

La partie poreuse est occupée par différents fluides. Chacun occupe une certaine fraction du volume poreux. On appelle saturation de la phase α, et on note S_α la fraction de volume poreux occupée par le fluide α. Comme le volume poreux est saturé par les fluides, on a

$$\sum_{N_\alpha} S_\alpha = 1, \tag{1.5}$$

où N_α est le nombre de phases fluides.

Chaque phase, qu'il s'agisse d'une phase fluide ou d'une phase solide, est caractérisée par sa composition.

Néanmoins, on ne considère dans ce travail que des minéraux purs (un seul constituant). La composition d'une phase minérale est donc triviale.

La composition d'une phase fluide est caractérisée à l'aide des fractions molaires de chaque espèce constituant le fluide. La fraction molaire d'une espèce i, notée x_i, est le rapport du nombre de moles de l'espèce i et du nombre total de moles dans la phase à laquelle elle appartient. On note $\alpha(i)$, la phase à laquelle appartient l'espèce i. Pour chaque fluide α, l'équation de fermeture suivante est vérifiée

$$\sum_{i,\alpha(i)=\alpha} x_i = 1 \tag{1.6}$$

La répartition du volume entre partie fluide et partie solide et la décomposition de chacune de ces deux parties peut être représentée par la figure suivante :

Volume Total				
$1-\Phi$			Φ	
F_1	$\cdots$	F_m	S_w	S_g
$(x_i)=1$	$(x_i)=1$	$(x_i)=1$	$(x_i)=w_i$	$(x_i)=y_i$

1.2 Les fluides

1.2.1 Caractéristiques

Un fluide α est caractérisé essentiellement par sa pression P_α, sa viscosité dynamique et sa densité volumique.

La viscosité dynamique d'un fluide α, notée μ_α, désigne la capacité du fluide à s'écouler. Plus elle augmente, plus la capacité du fluide à s'écouler diminue. Elle s'exprime en Pa.s (unité S.I).

On note ρ_α la densité volumique d'un fluide α, qui est une masse par unité de volume. On utilise également la notion de densité molaire, notée ξ_α, qui correspond au nombre de moles par unité de volume. La densité molaire est reliée à la masse volumique par l'intermédiaire de la composition du fluide et des masses molaires de ses constituants, selon l'expression suivante :

$$\xi_\alpha = \frac{\rho_\alpha}{\sum\limits_{i,\alpha(i)=\alpha} x_i M_i} \tag{1.7}$$

Ces propriétés se calculent différemment selon le type de fluide. On indique ici une façon de calculer ces propriétés pour l'eau et le gaz.

L'eau

La viscosité de l'eau dépend essentiellement de la quantité de co_2 dissous, de la salinité et de la température. Cette loi est le résultat d'une interpolation de données expérimentales (ATHOS).

Dans le modèle développé au cours de cette thèse, la masse volumique de l'eau est calculée en fonction de la composition, de la pression, de la température et de la salinité, à l'aide d'une loi d'interpolation (ATHOS).

Le gaz

La viscosité du gaz dépend de la température, de la pression et aussi de la quantité de co_2 dans le gaz. On utilise ici pour la calculer la loi de Lohrenz-Bray-Clark (Lohrenz et al. (1964)). La loi des gaz parfaits est utilisée pour calculer ρ_g.

1.2.2 La loi de Darcy

La vitesse d'écoulement d'un fluide α s'exprime à l'aide de la loi de Darcy. C'est une loi fondamentale pour décrire les écoulements dans un milieu poreux. Il s'agit d'une loi expérimentale. Elle établit que le débit volumique Q, à travers une surface A est proportionnel au gradient de pression et au rapport de la perméabilité sur la viscosité

$$Q = -A\frac{K}{\mu_\alpha}\nabla P_\alpha \tag{1.8}$$

Le signe indique qu'en l'absence de gravité, le fluide se déplace toujours de la plus grande pression vers la plus petite. On appelle vitesse de Darcy la quantité Q/A, qui est un flux volumique rapporté à une section du milieu, et non une vitesse réelle du fluide.

Cette loi se généralise au cas multiphasique, et en présence de gravité. La vitesse de Darcy de la phase α, notée v_α, s'écrit dans ce cas

$$v_\alpha = -\frac{K k_{r,\alpha}(S_\alpha)}{\mu_\alpha}\left(\nabla P - \rho_\alpha g\right) \tag{1.9}$$

où $k_{r,\alpha}$ est la perméabilité relative de la phase α et g le vecteur gravité.

La notion de perméabilité relative exprime le fait qu'un fluide en contact avec un autre ne se déplace pas de la même façon que s'il était seul. La perméabilité relative d'un fluide dépend de sa saturation et des propriétés de mouillabilité et de densité des autres fluides en contact.

Pour les simulations numériques, dans un cas diphasique eau-gaz, on utilisera souvent dans cette thèse des perméabilités relatives en croix, c'est à dire

$$k_{r,w} = S_w \quad \text{et} \quad k_{r,g} = S_g. \tag{1.10}$$

Parfois, on utilisera plutôt des perméabilités relatives données en fonction des saturations.

L'interaction entre deux fluides se traduit également par la pression capillaire. La pression capillaire, notée p_c, est la différence de pression entre deux points infiniment voisins de part et d'autre d'une interface séparant deux fluides. Elle peut être déterminée par la loi de Laplace. En première approximation, la pression capillaire d'une phase α par rapport à l'eau est une fonction de la saturation du fluide α. L'équation de pression capillaire s'écrit alors

$$P_\alpha = P_w + p_c(S_\alpha). \tag{1.11}$$

1.3 Les équations de bilan

On introduit les équations de bilan. Pour chaque espèce chimique i, on peut faire un bilan de la matière qui s'écrit,

$$\frac{\partial}{\partial t}\left(\pi_i(\Phi, \xi_{\alpha(i)}, S_{\alpha(i)})x_i\right) + \nabla \cdot J_i = Q_i + R_i \tag{1.12}$$

où Q_i est un terme source traduisant l'apport par injection de l'espèce i, R_i, le taux de production de l'espèce par réaction chimique, et J_i le flux de l'espèce i.

La fonction $\pi_i(\Phi, \xi_{\alpha(i)}, S_{\alpha(i)})$, est donnée par

$$\pi_i = \begin{cases} \Phi S_w \xi_w & \text{si } \alpha(i) = w \\ \Phi S_g \xi_g & \text{si } \alpha(i) = g \\ \xi_{m(i)}\phi_{m(i)} & \text{si } \alpha(i) = m \end{cases} \tag{1.13}$$

Le flux J_i est donné par

$$J_i = \begin{cases} 0 & \text{si } \alpha(i) = m \\ \xi_{\alpha(i)} v_{\alpha(i)} x_i - \xi_{\alpha(i)} S_{\alpha(i)} D_i \cdot \nabla x_i & \text{sinon} \end{cases} \tag{1.14}$$

Le flux est nul pour les espèces minérales (qui sont des espèces immobiles). Pour les espèces mobiles, il est donné par la somme d'un terme de convection et d'un terme de diffusion-dispersion, D_i étant le tenseur de diffusion-dispersion. On détaille la forme du terme D_i dans le chapitre 10, car l'étude du terme de diffusion-dispersion fait l'objet d'un chapitre complet de cette thèse.

Le taux de production de l'espèce i correspond au nombre de moles par unité de volume et de temps apparues ou disparues de cette espèce par l'intermédiaire des différentes réactions chimiques. Le chapitre 2 est consacrée à la modélisation d'un système géochimique, et on reviendra par conséquent sur la modélisation du terme R_i.

Pour simplifier l'écriture des équations de bilan, on pose

$$Ł_i = \frac{\partial}{\partial t}\left(\pi_i(\Phi, \xi_{\alpha(i)}, S_{\alpha(i)}) x_i\right) + \nabla \cdot J_i - Q_i. \tag{1.15}$$

Les équations de bilan des espèces s'écrivent alors sous la forme simplifiée suivante :

$$Ł_i = R_i \tag{1.16}$$

Chapitre 2

Modélisation d'un système géochimique

La réactivité d'un système compositionnel peut être mesurée à l'aide des taux de production des réactions chimiques. Pour modéliser les réactions chimiques, il faut connaître la composition des fluides et de la roche ainsi que les réactions chimiques possibles. C'est l'ensemble de ces données qui forme un système géochimique.

Pour décrire un système géochimique, il faut donc définir les constituants chimiques qui le composent, les réactions chimiques entre ces constituants et la manière de modéliser ces réactions. Dans un système chimique réel, on distingue deux types de réactions, les réactions rapides (modélisées comme des équilibres) et les réactions cinétiques.

Les différentes réactions chimiques possibles relient les constituants chimiques entre eux. Ceci permet de réduire la taille du système à résoudre on exprimant certaines espèces en fonction des autres. On parle alors d'espèces primaires et secondaires. C'est ainsi que l'on pourra éliminer les taux de production qu'on ne peut pas exprimer de façon explicite.

La présentation qui suit est une compilation des travaux de Lichtner (1996), de Missen and Smith (1998), de Holstad (2000) et de Nourtier (2003). On présente dans un premier temps les notions nécessaires à la définition d'un système géochimique (définitions et lois régissant les réactions), puis dans un second temps, la méthode utilisée pour réduire la dimension de ce système.

Pour clarifier cette présentation, on utilisera un exemple de système géochimique simple, qu'on détaillera au fur et à mesure de la présentation pour expliciter les notions présentées.

2.1 Espèces, éléments et composition

Les objets qui servent à décrire un système géochimique sont les constituants chimiques, également appelés espèces. On note $\mathcal{S}$ l'ensemble des espèces, de cardinal $n(\mathcal{S})$.

En général, une espèce n'est pas pure, mais est composée de constituants de base, également appelés éléments. On note $\mathcal{E}$ l'ensemble des éléments, de cardinal $n(\mathcal{E})$. Chaque espèce possède donc une formule de composition en éléments. Cette propriété peut s'écrire de façon matricielle. On obtient ainsi une matrice de composition, notée Γ, de taille $n(\mathcal{E}) \times n(\mathcal{S})$. Chaque coefficient $\gamma_{i,j}$ de la matrice correspond au nombre d'éléments i dans l'espèce j. La formule de composition en éléments n'est qu'une propriété et non pas une caractérisation d'une espèce. Deux espèces distinctes peuvent avoir la même composition.

Thermodynamiquement, une espèce chimique est caractérisée par son potentiel chimique, qu'on manipule en pratique à l'aide de la notion d'activité. Chaque espèce appartient à une phase unique. Ceci signifie que par exemple le co_2 dans l'eau et le co_2 dans le gaz sont deux espèces distinctes. On note $\alpha(i)$, la phase de l'espèce E_i, et $\mathcal{S}_p$ la restriction de l'ensemble $\mathcal{S}$ aux espèces de la phase p. On a donc $\mathcal{S}_p = \{E_i \in \mathcal{S}, \alpha(i) = p\}$. Ainsi, les phases forment une partition de l'ensemble des espèces.

On distingue les espèces mobiles des espèces non mobiles ou fixées. On considère ici que la mobilité des espèces coïncide avec la mobilité des phases. Ainsi, les espèces appartenant à une phase fluide (eau, gaz) sont mobiles alors que les espèces appartenant à une phase solide sont immobiles. Cette distinction est nécessaire, car les espèces mobiles peuvent être déplacées contrairement aux espèces fixées. On note $\mathcal{S}_{mob}$ l'ensemble des espèces mobiles et $\mathcal{S}_{fix}$ l'ensemble des espèces fixées.

***E*xemple 1 :**

L'exemple présenté ici est repris tout au long de ce chapitre pour expliciter la présentation.

On souhaite dans cette exemple définir un système géochimique simple qui permette de traiter le problème suivant. On injecte du dioxyde de carbone dans un aquifère carbonaté dont la roche est majoritairement composée de calcite. Dans les conditions usuelles de température et de pression, le co_2 se trouve sous la forme d'une phase supercritique, qu'on appelle pour simplifier "gaz".

On veut définir le système géochimique minimum permettant de prendre en compte ce cas. La première étape est de définir les espèces et éléments nécessaires. On définit les ensembles d'éléments et d'espèces suivants :

$$\begin{array}{lll} \mathcal{E} & = \{H, C, Ca, O, \oplus\} & n(\mathcal{E}) = 5 \\ \mathcal{S} & = \left\{h^+, hco_3^-, ca^{2+}, co_2^w, h_2o, co_2^g, caco_3\right\} & n(\mathcal{S}) = 7 \end{array} \tag{2.1}$$

L'élément $\oplus$ représente l'électron. Il est utile si l'on souhaite modéliser des réactions de types oxydo-réduction. Dans le cas contraire, il permet juste de vérifier que les réactions chimiques sont équilibrées en terme de charge.

Le système géochimique comporte ici trois phases, deux phases fluides (eau (w) et gaz(g)), et une unique phase solide. On peut partitionner l'ensemble des espèces selon leur phase, ce qui donne :

$$\begin{array}{lll} \mathcal{S}_w & = \left\{h^+, hco_3^-, ca^{2+}, co_2^w, h_2o\right\} & n(\mathcal{S}_w) = 5 \\ \mathcal{S}_g & = \{co_2^g\} & n(\mathcal{S}_g) = 1 \\ \mathcal{S}_s & = \{caco_3\} & n(\mathcal{S}_s) = 1 \end{array} \tag{2.2}$$

La matrice de composition est la suivante :

$\Gamma =$

	h^+	hco_3^-	ca^{2+}	co_2^w	h_2o	co_2^g	$caco_3$
H	1	1	0	0	2	0	0
C	0	1	0	1	0	1	1
Ca	0	0	1	0	0	0	1
O	0	3	0	2	1	2	3
$\oplus$	1	-1	2	0	0	0	0

(2.3)

2.2 Réactions chimiques

Dans un système chimique, les espèces ne sont pas conservées, même dans un système fermé. Les espèces échangent entre elles des éléments au cours des réactions. Par principe, les éléments sont des quantités indivisibles. Par conséquent, seules les réactions conservant les éléments sont admissibles. Cette contrainte est une contrainte d'orthogonalité entre la matrice de composition et la matrice de stoechiométrie des réactions, notée $\mathcal{R}$. Cette matrice est formée de tous les coefficients stoechiométriques de chaque réaction. On considère ici un ensemble de R réactions et on note $\nu_{i,r}$ le coefficient stoechiométrique de l'espèce E_i dans la réaction r. L'ensemble des réactions s'écrit donc de la façon suivante :

$$\left[E_0 \cdots E_{n(\mathcal{S})}\right] \begin{bmatrix} \nu_{11} & \cdots & \nu_{1R} \\ & \ddots & \\ \vdots & \nu_{ir} & \vdots \\ & & \ddots \\ \nu_{n(\mathcal{S})1} & \cdots & \nu_{n(\mathcal{S})R} \end{bmatrix} = 0 \tag{2.4}$$

On distingue parmi ces réactions, celles qui décrivent un équilibre local (réactions équilibrées) de celles qui sont contrôlées par une cinétique (réactions cinétiques). On appelle taux de production d'une réaction r, et on note I_r, le taux de production (ou de disparition) associée à la réaction r. Il s'exprime en $\text{mol.m}^{-3}.\text{s}^{-1}$ (unité S.I). Pour connaître le taux de production d'une espèce i par la réaction, il suffit de pondérer le terme I_r par le coefficient stoechiométrique $\nu_{i,r}$.

Le taux de production est ici exprimé par unité de volume pour être homogène aux équations de bilan. En réalité, le taux d'avancement d'une réaction chimique s'exprime uniquement comme un nombre de moles par unité de temps.

Toutes les réactions ne sont pas linéairement indépendantes. On note R_{eq} le nombre de relations équilibrées linéairement indépendantes, et R_{kin} le nombre de relations cinétiques indépendantes (entre elles et également indépendantes des réactions équilibrées). Ils vérifient

$$\begin{aligned} R_{\text{eq}} &= \text{rg}(\mathcal{R}_{eq}) \\ R_{\text{kin}} &= \text{rg}(\mathcal{R}) - \text{rg}(\mathcal{R}_{eq}) \end{aligned} \tag{2.5}$$

où $\mathcal{R}_{eq}$ désigne la restriction de la matrice $\mathcal{R}$ aux réactions équilibrées, et $\text{rg}(A)$ désigne le rang d'une matrice A.

Exemple 1 - suite :

Dans notre exemple, on peut supposer que les réactions suivantes sont admissibles :

$$\begin{array}{lllllllll} & & co_2^g & \leftrightharpoons & co_2^w, & & & & \\ hco_3^- & + & h^+ & \leftrightharpoons & co_2^w & + & h_2o & & \\ 2hco_3^- & + & 2h^+ & \leftrightharpoons & 2co_2^w & + & 2h_2o & & \\ caco_3 & + & h^+ & \Longleftrightarrow & ca^{2+} & + & hco_3^- & & \\ caco_3 & + & 2h^+ & \Longleftrightarrow & ca^{2+} & + & co_2^w & + & h_2o \end{array} \tag{2.6}$$

Le symbole $\Longleftrightarrow$ désigne une réaction cinétique, et le symbole $\leftrightharpoons$ une réaction équilibrée. Dans ce cas, on a $R = 5$.

Par convention, on note avec un signe $(-)$, les coefficients stoechiométriques des espèces intervenant à gauche de la réaction, et avec un signe $(+)$, les coefficients des espèces intervenant à droite. La matrice de réaction s'écrit

$$\mathcal{R} = \begin{bmatrix} 0 & -1 & -2 & -1 & -2 \\ 0 & -1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 & 1 \\ 0 & 1 & 2 & 0 & 1 \\ -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & -1 \end{bmatrix} \tag{2.7}$$

Le rang de cette matrice est $\text{rg}(\mathcal{R}) = 3$. Le rang de la matrice restreinte aux réactions équilibrées (*i.e.* privée de ses deux dernières colonnes) est $\text{rg}(\mathcal{R}_{eq}) = 2$. On a donc $R_{\text{eq}} = 2$ et $R_{\text{kin}} = 1$. Ceci signifie que parmi toutes les réactions proposées, il suffit d'en prendre trois pour représenter correctement le système réactif. Le choix des réactions n'est évidemment pas arbitraire. La méthode présentée dans la section 2.4 permet de faire ce choix.

2.2.1 Les réactions à l'équilibre

Certaines réactions sont suffisamment rapides pour être considérées comme des équilibres. C'est le cas des réactions entre espèces aqueuses et des réactions d'échange entre l'eau et le gaz. Ces équilibres obéissent à la loi d'action de masse, ainsi pour chaque réaction r, on a

$$K_r = \prod_{i=1}^{n(\mathcal{S})} (a_i)^{\nu_{i,r}} \tag{2.8}$$

où K_r est la constante d'équilibre de la réaction r et a_i l'activité de l'espèce i. On détaillera dans la section 2.3, ce qu'est l'activité et comment la calculer.

Le taux de production I_r, associé à une réaction équilibrée r, n'est pas explicite. En effet, la réaction étant instantanée, il n'existe pas de formule permettant de connaître à l'avance combien de moles d'une espèce seront nécessaires pour maintenir cette réaction à l'équilibre.

2.2.2 Les cinétiques de précipitation-dissolution

Certaines réactions ne sont pas assez rapides pour être considérées comme étant à l'équilibre. C'est le cas des réactions hétérogènes entre l'eau et les minéraux. Il faut alors définir une vitesse de réaction. En géochimie, les modèles utilisés pour le calcul de la vitesse de réaction sont issus de la "théorie de l'état transitoire" (Lasaga (1981), Aagaard and Helgeson (1982)). En pratique, ces modèles ont toujours un caractère empirique afin de mieux reproduire les données expérimentales. On peut trouver une description complète dans Nourtier (2003). Dans cette thèse nous utilisons un modèle simplifié. Seule cette forme simplifiée sera présentée ici.

La vitesse de réaction dépend de l'indice de saturation du minéral défini par

$$\Omega_r = \frac{Q_r}{K_r} \tag{2.9}$$

où Q_r est le produit de solubilité et K_r la constante d'équilibre de la réaction. Le produit de solubilité vaut

$$Q_r = \prod_{i=1}^{n(\mathcal{S})} (a_i)^{\nu_{i,r}}$$

La vitesse de réaction est définie de façon différente selon que la réaction soit orientée dans le sens de la précipitation ou de la dissolution.

Si $Q_r > K_r$, la solution est sur-saturée par rapport au minéral, et la réaction qui peut se produire est une réaction de précipitation. La vitesse de réaction est

$$v_r = k_r^p s_r \left(\frac{Q_r}{K_r} - 1 \right) \tag{2.10}$$

où k_r^p est la constante de précipitation du minéral (en mol.m^{-2}, unité S.I) liée à la réaction r, et s_r sa surface réactive (en m^2.m^{-3}, unité S.I).

Si $Q_r < K_r$, la solution est sous-saturée, et la réaction qui peut se produire est une réaction de dissolution. La vitesse de réaction est

$$v_r = -k_r^d s_r \left(1 - \frac{Q_r}{K_r} \right) \tag{2.11}$$

où k_r^d est la constante de dissolution du minéral (en mol.m^{-2}, unité S.I) liée à la réaction r, et s_r sa surface réactive.

Dans le cas d'une réaction cinétique r, le taux de production I_r, est égale à la vitesse de réaction v_r.

2.3 Modèle d'activité

L'activité d'une espèce E_i est l'influence de la quantité de cette espèce sur l'énergie libre du système. Elle est définie comme le rapport de la fugacité de l'espèce dans un état donné et de la fugacité de la même espèce dans un état de référence. C'est une propriété thermodynamique de l'espèce, et elle s'écrit donc différemment selon la phase à laquelle appartient l'espèce considérée.

2.3.1 Espèce gazeuse

Pour les gaz, l'état de référence est l'état gaz parfait. La fugacité de référence f_0 est égale à la pression de référence P_0. Pour un gaz parfait, la fugacité est égale à la pression partielle, soit $f = Px_i$, où x_i est la fraction molaire de l'espèce E_i. Pour généraliser cette notion d'activité à tous les gaz, on utilise un coefficient d'activité, noté λ_i, qui corrige la fraction molaire de l'espèce. On obtient donc

$$a_i = \frac{f_i}{f_0} = \frac{P}{P_0} \lambda_i x_i \tag{2.12}$$

2.3.2 Espèce aqueuse

L'activité d'une espèce aqueuse en solution s'exprime à l'aide de la concentration, c_i, et d'un coefficient d'activité, λ_i, de la façon suivante :

$$a_i = \lambda_i c_i \tag{2.13}$$

Dans le cas d'une solution idéale, c'est à dire infiniment diluée, le coefficient d'activité vaut 1. Dans les saumures faiblement concentrées, c'est la concentration en ions, autrement dit la force ionique qui influence principalement le coefficient d'activité (lois de Davis, Debye-Hückel, BDot). Pour des saumures plus concentrées, ces modèles sont jugés insuffisants et on utilise en général des modèles d'interaction de type Pitzer.

L'activité du solvant (h_2o pour les solutions aqueuses) se calcule de façon différente. L'état de référence étant celui du solvant pur, l'activité s'écrit simplement

$$a_i = \lambda_i \tag{2.14}$$

Pour des solutions plus concentrées, l'activité du solvant se calcule à l'aide du coefficient osmotique, calculé en fonction de la force ionique de la solution. On peut trouver le détail de ces lois dans Nourtier (2003).

2.3.3 Les minéraux

On ne considère comme "solide" que des minéraux. Or un minéral pur est une phase déterminée et par conséquent son activité est égale à 1. On ne s'intéresse pas ici au cas des solutions solides.

2.3.4 Choix en pratique

Dans le prototype développé pour cette thèse, on prend le parti d'utiliser des modèles d'activité idéaux. Pour les espèces gazeuses, on retient l'hypothèse de gaz parfait. L'activité d'une espèce gazeuse est donc égale à sa pression partielle. Pour les espèces aqueuses, on retient l'hypothèse d'une solution infiniment diluée. Par conséquent, l'activité d'une espèce, hors solvant, est égale à la concentration, et l'activité du solvant est égale à 1. On suppose également que le solvant est l'eau (h_2o). Le modèle d'activité est donc le suivant :

$$\begin{cases} a_i = Px_i, & \text{si } \alpha(i) = g, \\ a_i = c_i, & \text{si } \alpha(i) = w,\ E_i \neq h_2o, \\ a_{(h_2o)} = 1 & \\ a_i = 1, & \text{si } \alpha(i) = s. \end{cases} \tag{2.15}$$

2.4 Réduction du système

Le taux de production de l'espèce i, R_i, est égal à la somme des taux de production de chaque réaction à laquelle participe l'espèce i pondéré par le coefficient stoechiométrique de l'espèce i dans la réaction considérée. On a donc

$$R_i = \sum_{r=1}^{R} \nu_{i,r} I_r \tag{2.16}$$

Or parmi les R réactions, on a vu que certaines étaient redondantes, car linéairement dépendantes (section 2.2) et il est donc intéressant de les éliminer. De plus, pour les réactions équilibrées, le taux de production I_r n'est pas directement calculable, par conséquent, pour pouvoir résoudre le système (1.16), il faut préalablement éliminer les termes I_r.

Pour faire cela, on utilise les réactions équilibrées pour exprimer certaines espèces (les espèces secondaires) en fonction des autres (les espèces primaires). C'est cette étape que l'on appelle réduction du système réactif. Elle permet de réécrire la matrice de réaction sous une forme particulière, qui permet d'éliminer les taux de production, et qui permet de ne retenir que les réactions indispensables à la description du système.

2.4.1 Décomposition en espèces primaires, secondaires et cinétiques

La première étape pour réduire le système est de déterminer les espèces primaires, secondaires et cinétiques. Les espèces secondaires peuvent alors s'exprimer en fonction de ces espèces primaires. Il existe différentes approches pour réaliser cette décomposition. On s'intéresse ici à l'approche présentée dans Lichtner (1996).

On note N_{le} le nombre d'espèces qui participent aux réactions équilibrées. On a $N_{le} \leq n(\mathcal{S})$. Les espèces restantes sont appelées espèces cinétiques. On note N_k le nombre d'espèces cinétiques. On a donc $N_k = n(\mathcal{S}) - N_{le}$. On partitionne les N_{le} espèces qui participent aux réactions équilibrées en espèces primaires et secondaires. Le nombre d'espèces primaires est $N_p = N_{le} - R_{\text{eq}}$ et le nombre d'espèces secondaires est $N_s = R_{\text{eq}}$.

Pour améliorer la clarté de la présentation, on réordonne la matrice de réaction. Les R_{eq} premières colonnes correspondent à R_{eq} réactions équilibrées indépendantes, les R_{kin} colonnes suivantes correspondent aux réactions cinétiques indépendantes, et enfin les dernières colonnes sont les réactions dépendantes des $(R_{\text{eq}} + R_{\text{kin}})$ premières. On réordonne ensuite les espèces dans l'ordre suivante : Espèces primaires, puis espèces secondaires puis espèces cinétiques. Le système (2.4) s'écrit alors

$$[E_p, E_s, E_k] \begin{bmatrix} \Upsilon_{1,1} & \Upsilon_{1,2} & \Upsilon_{1,3} \\ \Upsilon_{2,1} & \Upsilon_{2,2} & \Upsilon_{2,3} \\ 0 & \Upsilon_{3,2} & \Upsilon_{3,3} \end{bmatrix} = 0 \tag{2.17}$$

où $\Upsilon_{1,1} \in \mathbb{R}^{N_p \times R_{\text{eq}}}$, $\Upsilon_{1,2} \in \mathbb{R}^{N_p \times R_{\text{kin}}}$, $\Upsilon_{1,3} \in \mathbb{R}^{N_p \times (R - R_{\text{eq}} - R_{\text{kin}})}$, $\Upsilon_{2,1} \in \mathbb{R}^{N_s \times R_{\text{eq}}}$, $\Upsilon_{2,2} \in \mathbb{R}^{N_s \times R_{\text{kin}}}$, $\Upsilon_{2,3} \in \mathbb{R}^{N_s \times (R - R_{\text{eq}} - R_{\text{kin}})}$, $\Upsilon_{3,2} \in \mathbb{R}^{N_k \times R_{\text{kin}}}$, et $\Upsilon_{3,3} \in \mathbb{R}^{N_k \times (R - R_{\text{eq}} - R_{\text{kin}})}$. Les blocs $\Upsilon_{i,j}$ sont formés des mêmes coefficients que la matrice $\mathcal{R}$, réordonnés selon le type de l'espèce (primaire, secondaire, ou cinétique) et selon le type de réaction. Le bloc de zéros correspond à la définition choisie des espèces cinétiques, à savoir les espèces non impliquées dans une ou plusieurs réactions équilibrées. Le fait que $N_k = 0$, ne signifie pas qu'il n'y a pas de réaction cinétique, mais signifie que toutes les espèces impliquées dans des réactions cinétiques sont également impliquées dans des réactions équilibrées.

Le choix des espèces primaires est arbitraire. La seule restriction est que ces espèces soient indépendantes entre elles. Cette condition se traduit par le fait que la matrice $\Upsilon_{2,2}$ doit être inversible.

Il y a donc en général plusieurs choix possibles. En pratique, on choisit plutôt des espèces aqueuses, et parmi les espèces aqueuses, celles qui sont majoritaires (pour des raisons de précision numérique). Choisir un minéral peut poser problème car il n'est pas forcément présent dans l'ensemble du système.

2.4.2 Réécriture de la matrice de réaction

Il est possible de réduire le système réactif, en éliminant les équations linéairement indépendantes, et en exprimant les espèces secondaires en fonction des espèces primaires. Pour cela, on réécrit le système (2.17) sous la forme de 3 équations :

$$\begin{cases} E_p\Upsilon_{1,1} + E_s\Upsilon_{2,1} = 0 \\ E_p\Upsilon_{1,2} + E_s\Upsilon_{2,2} + E_k\Upsilon_{3,2} = 0 \\ E_p\Upsilon_{1,3} + E_s\Upsilon_{2,3} + E_k\Upsilon_{3,3} = 0 \end{cases} \tag{2.18}$$

Par construction, la matrice $\Upsilon_{2,1}$ est inversible, par conséquent le système peut se réécrire

$$\begin{cases} E_s = -E_p\Upsilon_{1,1}\Upsilon_{2,1}^{-1} \\ E_p\left(\Upsilon_{1,2} - \Upsilon_{1,1}\Upsilon_{2,1}^{-1}\Upsilon_{2,2}\right) + E_k\Upsilon_{3,2} = 0 \\ E_p\left(\Upsilon_{1,3} - \Upsilon_{1,1}\Upsilon_{2,1}^{-1}\Upsilon_{2,3}\right) + E_k\Upsilon_{3,3} = 0 \end{cases} \tag{2.19}$$

De plus, les blocs de la 3$^{\text{e}}$ colonne de $\mathcal{R}$ correspondent aux équations linéairement dépendantes, ils peuvent donc s'exprimer comme une combinaison linéaires des autres, ce qui s'écrit :

$$\begin{bmatrix} \Upsilon_{1,3} \\ \Upsilon_{2,3} \\ \Upsilon_{3,3} \end{bmatrix} = \begin{bmatrix} \Upsilon_{1,1} & \Upsilon_{1,2} \\ \Upsilon_{2,1} & \Upsilon_{2,2} \\ 0 & \Upsilon_{3,2} \end{bmatrix} \times \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} \tag{2.20}$$

où $\alpha_1 \in \mathbb{R}^{R_{\text{eq}} \times (R - R_{\text{eq}} - R_{\text{kin}})}$ et $\alpha_2 \in \mathbb{R}^{R_{\text{kin}} \times (R - R_{\text{eq}} - R_{\text{kin}})}$, sont les matrices des coefficients des combinaisons linéaires.

En remplaçant les blocs de la 3$^{\text{e}}$ colonne par leurs combinaisons linéaires, la troisième équation du système (2.19) devient

$$\left(E_p \left(\Upsilon_{1,2} - \Upsilon_{1,1}\Upsilon_{2,1}^{-1}\Upsilon_{2,2}\right) + E_k\Upsilon_{3,2}\right)\alpha_2 = 0, \tag{2.21}$$

ce qui est équivalent, d'après la deuxième équation de (2.19) à $(0 = 0)$. Le système se réduit alors à

$$\begin{cases} E_s = -E_p\Upsilon_{1,1}\Upsilon_{2,1}^{-1} \\ E_p \left(\Upsilon_{1,2} - \Upsilon_{1,1}\Upsilon_{2,1}^{-1}\Upsilon_{2,2}\right) + E_k\Upsilon_{3,2} = 0 \end{cases} \tag{2.22}$$

On peut le réécrire sous forme matricielle

$$[E_p \mid E_s \mid E_k] \begin{bmatrix} \Lambda_{1,1} & \Lambda_{1,2} \\ Id & 0 \\ 0 & \Lambda_{3,2} \end{bmatrix} = 0 \tag{2.23}$$

avec $\Lambda_{1,1} = \Upsilon_{1,1}\Upsilon_{2,1}^{-1}$, $\Lambda_{1,2} = \Upsilon_{1,2} - \Upsilon_{1,1}\Upsilon_{2,1}^{-1}\Upsilon_{2,2}$ et $\Lambda_{3,2} = \Upsilon_{3,2}$.

Dans la suite du document, pour simplifier l'écriture, on notera toujours $\nu_{i,r}$ les coefficients de la matrice de réaction modifiée, bien qu'ils soient différents des coefficients de la matrice de réaction d'origine.

2.4.3 Utilisation de la matrice de réaction modifié

On utilise maintenant la matrice de réaction modifiée pour exprimer le taux de production de l'espèce i,

$$R_i = \sum_{i=1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_r \tag{2.24}$$

Cette écriture indique simplement que la dimension du système réactif est passée de $(n(\mathcal{S}) \times R)$ à $(n(\mathcal{S}) \times (R_{\text{eq}} + R_{\text{kin}}))$. En utilisant maintenant la décomposition des espèces en espèces primaires secondaires et cinétique, on a

$$\begin{aligned} R_i &= \sum_{r=1}^{R_{\text{eq}}} \nu_{i,r} I_{r,e} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}, && \forall i = 1, \cdots, N_p \\ R_j &= I_{(j-N_p),eq}, && \forall j = N_p + 1, \cdots, N_p + N_s \\ R_k &= \sum_{r=R_{\text{eq}}+1}^{R_{\text{kin}}} \nu_{k,r} I_{r,k}, && \forall k = N_p + N_s + 1, \cdots, n(\mathcal{S}) \end{aligned} \tag{2.25}$$

où l'indice k ajouté à $I_{r,k}$ signifie qu'il s'agit d'une réaction cinétique, et l'indice eq qu'il s'agit d'une réaction équilibrée. L'écriture du système (2.25) utilise le fait que les espèces secondaires interviennent uniquement dans une seule réaction équilibrée avec un coefficient stoechiométrique égal à 1 (bloc identité dans la matrice de réaction), et n'interviennent pas dans les réactions cinétiques (bloc nul dans la matrice de réaction). Pour les espèces cinétiques, on utilise le fait qu'elles n'interviennent pas dans les réactions équilibrées (bloc nul dans la matrice de réaction).

En utilisant ces expressions pour remplacer le terme R_i dans les équations bilan (1.16), on obtient le système d'équations suivant

$$\begin{cases} Ł_i = \sum_{r=1}^{R_{\text{eq}}} \nu_{i,r} I_{r,eq} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}, & \forall i = 1, \cdots, N_p \\ Ł_j = I_{\{j-N_p\},eq}, & \forall j = N_p + 1, \cdots, N_p + N_s \\ Ł_k = \sum_{r=R_{\text{eq}}+1}^{R_{\text{kin}}} \nu_{i,r} I_{r,k}, & \forall k = N_s + N_p + 1, \cdots, n(\mathcal{S}) \end{cases} \tag{2.26}$$

Les termes ne pouvant être calculés directement sont les taux de production des réactions équilibrées, soit les R_{eq} termes $I_{r,eq}$. Or, comme $N_s = R_{\text{eq}}$, la deuxième équation du système (2.26) permet d'exprimer

tous les termes $I_{r,eq}$. On peut donc remplacer $I_{r,eq}$ par cette expression dans les équations de bilan des espèces primaires. Les équations de bilan des espèces secondaires ainsi supprimées sont remplacées par les lois d'action de masse correspondantes. On remplace ainsi des équations aux dérivées partielles non locales par des équations algébriques locales.

2.4.4 Ajout du point de vue thermodynamique

La présentation faite dans ce chapitre d'un système géochimique est une vision purement stoechiométrique du problème. En effet, la réduction du système est faite uniquement en fonction des réactions chimiques, et de la composition des espèces. Néanmoins, il faut rappeler que l'on s'intéresse à un problème thermodynamique, et que la notion d'appartenance à une phase a une grande importance. En effet, même si une espèce appartenant à une phase α n'est pas liée aux autres espèces de la même phase par une réaction équilibrée, elle l'est par l'intermédiaire de son potentiel chimique, et donc par son activité.

De plus, toutes les réactions chimiques n'ont pas la même influence sur l'écoulement. En effet, les réactions équilibrées internes à une phase n'ont pas le même effet sur l'écoulement que les réactions d'échanges entre deux phases.

C'est la raison pour laquelle on souhaite traiter séparément les réactions équilibrées homogènes des réactions équilibrées hétérogènes. Notamment, les réactions équilibrées entre une espèce aqueuse et une espèce gazeuse ne sont pas modélisées par la loi d'action de masse. On utilise une équation d'état où la constante d'équilibre est calculée par un rapport des fugacités des espèces. On choisit ici d'utiliser la loi de Henry.

La loi de Henry, s'exprime de la façon suivante

$$w = \frac{y}{H} \tag{2.27}$$

où y est la fraction molaire de l'espèce en phase gazeuse, w la fraction molaire de l'espèce en phase aqueuse et H la constante de Henry.

Cette loi remplace la loi d'action de masse écrite pour l'équilibre eau-gaz. La constante de Henry est une constante apparente, plus facilement accessible que la constante d'équilibre de la réaction. Ceci signifie qu'il est inutile de connaître une description fine du système géochimique pour mesurer cette constante. Or dans le cadre de simulation réservoir classique, on ne dispose pas d'une description fine de la chimie, puisqu'on ne s'intéresse pas aux réactions chimiques au sein de l'eau et avec la roche.

Pour pouvoir utiliser la loi de Henry au lieu de la loi d'action de masse pour les réactions équilibrées faisant intervenir des espèces de phases différentes, il faut différentier ces réactions des autres. Parmi les R_{eq} équilibres on notera $I_{r,\alpha}$, les taux de production des R^{α}_{eq} équilibres faisant intervenir des espèces de phases différentes et $I_{r,\beta}$ les taux de production des R^{β}_{eq} équilibres ne faisant intervenir que des espèces d'une même phase.

Pour pouvoir déterminer R^{α}_{eq} et R^{β}_{eq}, il faut que les réactions chimiques soient écrites sous la "bonne" forme, c'est à dire que l'on cherche l'écriture qui minimise R^{α}_{eq}. Ceci peut se faire par une méthode de réduction similaire à celle présentée ci-dessus. Une façon simple d'obtenir directement cette écriture est de choisir uniquement des espèces primaires appartenant à la phase aqueuse. Il suffit alors de décompter le nombre de réactions faisant intervenir des espèces gazeuses pour obtenir R^{α}_{eq} et R^{β}_{eq}.

On ordonne maintenant les réactions équilibrées dans l'ordre suivant : équilibres hétérogènes, puis homogènes. Le système (2.26) s'écrit alors

$$Ł_i = \sum_{r=1}^{R^{\alpha}_{\text{eq}}} \nu_{i,r} I_{r,\alpha} + \sum_{r=R^{\alpha}_{\text{eq}}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{kin}}} \nu_{i,r} I_{r,k}, \quad \forall i = 1, \cdots, N_p \tag{2.28}$$

$$Ł_j = I_{\{j-N_p\},\alpha}, \quad \forall j = N_p + 1, \cdots, N_p + R^{\alpha}_{\text{eq}} \tag{2.29}$$

$$Ł_j = I_{\{j-N_p\},\beta}, \quad \forall j = N_p + R^{\alpha}_{\text{eq}} + 1, \cdots, N_p + N_s \tag{2.30}$$

$$Ł_k = \sum_{r=R_{\text{eq}}+1}^{R_{\text{kin}}} \nu_{i,r} I_{r,k}, \quad \forall k = N_s + N_p + 1, \cdots, n(\mathcal{S}) \tag{2.31}$$

Exemple 1 - suite :

Dans cet exemple, les espèces participant aux relations équilibrées sont hco_3^-, co_2^w, co_2^g, h_2o et h^+. On a donc $N_{le} = 5$ et $N_k = 7-5 = 2$. Les 2 espèces cinétiques sont ca^{2+} et $caco_3$. Le nombre d'espèces primaires est $N_p = 5 - 2 = 3$. Et finalement, $N_s = 2$. Les espèces primaires et secondaires sont à choisir parmi les 5 espèces participant aux relations équilibrées. Il faut choisir des espèces primaires indépendantes. On commence par choisir la première espèce, qui est h^+. On lui ajoute une deuxième espèce hco_3^-. Le rang de la matrice de composition restreinte à ces deux espèces est 2. Ces deux espèces sont indépendantes, il reste à trouver la troisième espèce. On ajoute co_2^w. Le rang de la matrice de composition restreinte à ces trois espèces est 3. On a donc trouvé trois espèces primaires. Finalement :

$$\begin{array}{ll} \mathcal{S}^p = \left\{h^+, hco_3^-, co_2^w\right\}, & N_p = 3 \\ \mathcal{S}^s = \left\{h_2o, co_2^g\right\}, & N_s = 2 \\ \mathcal{S}^k = \left\{caco_3, ca^{2+}\right\}, & N_k = 2 \end{array} \tag{2.32}$$

La figure 2.1 montre une représentation de ce système.

En réordonnant les espèces et les réactions, la matrice de réaction s'écrit alors

$$\mathcal{R} = \left[\begin{array}{ccc|cc} 0 & -1 & -1 & -2 & -2 \\ 0 & -1 & 1 & 0 & -2 \\ 1 & 1 & 0 & 1 & 2 \\ \hline 0 & 1 & 0 & 1 & 2 \\ -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & -1 & 0 \\ 0 & 0 & 1 & 1 & 0 \end{array}\right] \tag{2.33}$$

La matrice sous sa forme réduite s'écrit

$$\mathcal{R} = \left[\begin{array}{cc|c} -1 & 0 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \\ \hline \multicolumn{2}{c|}{Id} & 0 \\ \hline \multicolumn{2}{c|}{0} & \begin{array}{c} -1 \\ 1 \end{array} \end{array}\right] \tag{2.34}$$

Ajouter un point de vue thermodynamique à cette décomposition des espèces revient à séparer l'espèce co_2^g des autres, car elle appartient à la phase gazeuse et à regrouper l'espèce ca^{2+} aux espèces aqueuses. Ceci ne change rien à leur statut d'espèce primaire, secondaire ou cinétique. Le point de vue uniquement compositionnel et le point de vue thermodynamique sont représentés, respectivement, par les figures 2.1 et 2.2.

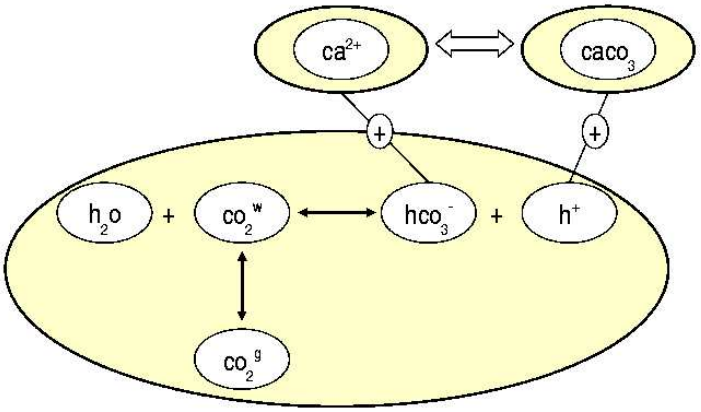


FIG. 2.1: Système réactif

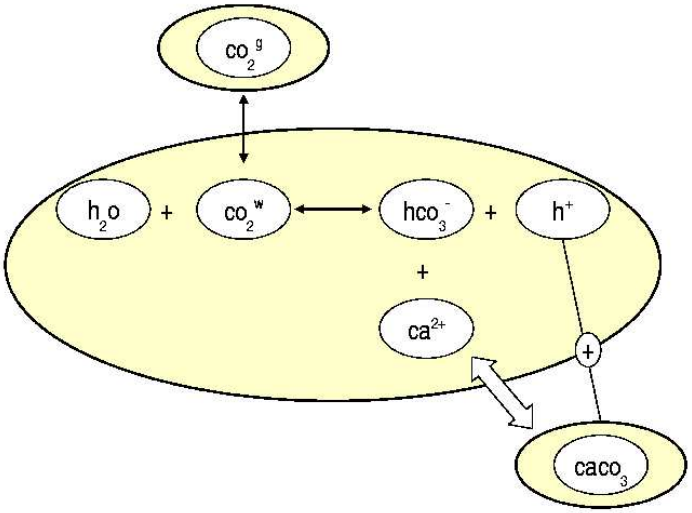


FIG. 2.2: Point de vue thermodynamique

2.5 Le problème continu

En intégrant le système géochimique dans le modèle d'écoulement multiphasique, on obtient un problème d'écoulement multiphasique réactif. Ce problème est dit "couplé" car il intègre à la fois le déplacement des fluides, le transport des constituants et les réactions chimiques, ce qui permet de mesurer l'influence de la chimie sur les écoulements et réciproquement.

Les équations à résoudre sont

- les équations de conservation (2.26) $n(\mathcal{S})$ équations,
- les équations de pression capillaire (1.11) $(N_\alpha - 1)$ équations,
- les lois de Henry (2.27) R_{eq}^{α} équations,
- les lois d'action de masse (2.8) R_{eq}^{β} équations,
- la fermeture du volume total (1.4) 1 équation,
- la fermeture du volume fluide (1.5) 1 équation,
- la fermeture de la composition des phases fluides (1.6) N_α équations.

Au total, le système comporte $n(\mathcal{S})$ équations aux dérivées partielles et $(2 \times N_\alpha + 1 + R_{\text{eq}})$ équations algébriques locales.

Les inconnues principales du problème continu couplé sont

- la porosité — 1 inconnue,
- les pressions de chaque phase fluide — N_α inconnues,
- les saturations de chaque phase fluide — N_α inconnues,
- les fractions molaires de toutes les espèces mobiles — $n(\mathcal{S}_{mob})$ inconnues,
- les fractions volumiques des minéraux — $n(\mathcal{S}_{fix})$ inconnues.

Les inconnues secondaires sont les taux de production des réactions équilibrées, soit R_{eq} inconnues.

Au total, le système comporte donc $(2 \times N_\alpha + n(\mathcal{S}) + 1)$ inconnues principales et R_{eq} inconnues secondaires.

Le système est donc bien défini. Il s'écrit

$$\left\{\begin{array}{ll}
Ł_i = \displaystyle\sum_{r=1}^{R^\alpha_{\text{eq}}} \nu_{i,r} I_{r,\alpha} + \sum_{r=R^\alpha_{\text{eq}}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{kin}}} \nu_{i,r} I_{r,k}, & \forall i \in \mathcal{S} \\
P_\alpha = P_w + P_{c,\alpha} & \forall \alpha = 1, \cdots, N_\alpha - 1 \\
y = x\, H_r & \forall r = 1, \cdots, R^\alpha_{\text{eq}} \\
\displaystyle\prod_{i=1}^{n(\mathcal{S})} (a_i)^{\nu_{i,r}} = K_r & \forall r = R^\alpha_{\text{eq}} + 1, \cdots, R_{\text{eq}} \\
\Phi + \displaystyle\sum_{i \in \mathcal{S}_{fix}} \phi_i = 1 & \\
\displaystyle\sum_{i=1}^{N_\alpha} S_\alpha = 1 & \\
\displaystyle\sum_{i,\alpha(i)=\alpha} x_i = 1 & \forall \alpha = 1, \cdots, N_\alpha
\end{array}\right. \tag{2.35}$$

Deuxième partie

Schémas de couplage et pénalisation

Le but de cette thèse est d'étudier des méthodes de résolution numérique pour les problèmes d'écoulement multiphasique réactif bien adaptées à la problématique du stockage de co_2.

Il existe plusieurs approches pour résoudre le système d'équations (2.35) présenté dans la section 2.5. On peut les classer en deux types, les approches globales et les approches séquentielles, ces dernières étant soit itératives, soit non itératives.

On présente ces différentes approches dans le chapitre 3. Nous présenterons ensuite dans les chapitres 4 et 5, deux approches que nous avons étudié pour la résolution du problème (2.35).

Parmi les méthodes étudiées, cette thèse met l'accent sur la méthode (Rs-Tr), qui est une méthode séquentielle non itérative à deux pas, proposée par Michel et Trenty (Michel and Trenty (2005)). Il s'agit d'un splitting particulier, qui repose sur des hypothèses liées à la physique du stockage de co_2. On présente cette méthode dans le chapitre 5.

Pour étudier les résultats obtenus, cette méthode est comparée avec une méthode de résolution globale implicite, présentée au chapitre 4.

Pour chacune de ces méthodes, on expliquera comment éliminer les variables secondaires. Les différences entre les différents schémas numériques présentés se situent au niveau de la discrétisation en temps. La discrétisation en espace est une discrétisation de type volumes finis classique. On utilise le schéma dit "des pétroliers" qui consiste à faire un décentrage amont phase par phase. Par conséquent, on négligera l'aspect discrétisation en espace dans l'écriture des schémas afin de mettre en évidence les particularités de la discrétisation en temps.

Finalement, le chapitre 6 de cette partie est consacré à l'analyse de convergence du schéma pénalisé dans un cas simplifié. On limite l'étude à un problème diffusif et réactif. L'objectif de ce chapitre est de montrer que le schéma pénalisé est convergent et qu'il converge vers la même solution faible que le schéma global. Cette étude est réalisée à l'aide d'estimations obtenues sur le schéma pénalisé qui permettent de passer à la limite sur les différents paramètres, à savoir le pas de temps et le pas d'espace, ainsi que le paramètre λ, coefficient du terme de pénalisation.

Chapitre 3

Les différents types d'approche

Il existe différents types d'approche pour résoudre un problème d'écoulement multiphasique réactif. Si on se réfère à Yeh and Tripathi (1989) et à Steefel and MacQuarrie (1996), on peut distinguer deux grandes catégories, les approches globales implicites (GIA = global implicit approach) et les approches séquentielles, SIA (sequential iterative approach), ou SNIA (sequential non iterative approach).

Pour bien situer chacune des méthodes, on écrit chacun des schémas présentés sur une équation simple, représentant un système avec une seule espèce :

$$\frac{\partial c}{\partial t} + L(c) = R(c) \tag{3.1}$$

où L est l'opérateur de transport et R l'opérateur de chimie.

3.1 Approche globale implicite

L'approche globale implicite (GIA), également appelée "fully-coupled" ou "one-step", consiste à résoudre la totalité du système en une seule étape. Pour l'équation (3.1), le schéma s'écrit

$$\frac{c^{n+1} - c^n}{\delta t} + L(c^{n+1}) - R(c^{n+1}) = 0 \tag{3.2}$$

L'équation (3.2) étant non linéaire, elle est résolue à l'aide d'une méthode de Newton. Par conséquent, si la dimension du système est importante, la matrice jacobienne obtenue lors de la linéarisation sera de très grande taille. Le stockage de la matrice nécessite donc une place mémoire importante, et de plus le calcul et l'inversion de cette matrice sont deux étapes très coûteuses.

Ce schéma est donc bien adapté pour un petit système, c'est à dire un système avec peu de constituants ou un système dans lequel les réactions n'introduisent pas de couplage entre les espèces. Dans ce cas, les schémas associés à chaque espèce peuvent être résolus séparément. Mais, en général, pour un système multi-constituants, les différentes espèces sont couplées entre elles via les réactions, ce qui accroît de façon considérable la taille de la matrice jacobienne et le nombre de termes non nuls dans cette matrice. Ce type d'approche a été fortement critiqué par Yeh and Tripathi (1989), en raison d'un coût de simulation et de mémoire important. Néanmoins, grâce au développement des capacités de calcul, l'approche GIA est de nouveau envisagée pour la résolution de problèmes industriels (par exemple Nghiem et al. (2004b), PARTRAN Hammond (2003)).

Le système global peut être résolu de deux façons. La première consiste à résoudre simultanément les EDP et les équations algébriques locales. Cette méthode est appelée DAE (différential and algebric equation). La deuxième solution consiste à substituer directement les équations algébriques locales dans les EDP. Cette méthode s'appelle DSA (direct substitution approach). Elle permet d'obtenir un système de taille inférieure mais limite la prise en compte de certains phénomènes chimiques.

Pour contrer les difficultés liées à la taille du problème à résoudre, plusieurs solutions sont envisageables. Par exemple, il est possible d'utiliser des méthodes de type Newton-Krylov sans jacobienne, appelée Jacobian Free Newton Krylov (Mousseau et al. (2000), Amir and Kern (2006)). L'avantage de ces méthodes est qu'il n'est pas nécessaire de calculer et stocker la jacobienne à chaque itération de Newton.

En effet, seul le produit matrice vecteur est utilisé et il est calculé par différences finies. Cette méthode est par exemple utilisée dans le logiciel PARTRAN, Hammond (2003), (avec un préconditionnement par splitting d'opérateur), qui est un logiciel prévu pour résoudre des problèmes industriels.

3.2 Approche séquentielle

L'approche séquentielle (ou méthode de splitting d'opérateurs) consiste à résoudre le problème en deux étapes. Le principe est de résoudre séparément l'opérateur de transport non local et l'opérateur de chimie local. Les méthodes de splitting peuvent être classées en deux grandes catégories, les méthodes non itératives (SNIA) et les méthodes itératives (SIA).

L'un des attraits importants de ces méthodes est la modularité. En effet, comme l'opérateur de transport est résolu séparément de l'opérateur de chimie, ces méthodes permettent de coupler des codes de résolution existants.

3.2.1 Approche séquentielle non itérative (SNIA)

L'approche non itérative (SNIA) consiste à résoudre le transport et la chimie sur un pas de temps donné. Le principe est d'effectuer une unique étape de transport suivie d'une unique étape de chimie. La deuxième étape utilise les concentrations obtenues lors de la première étape. Pour l'équation (3.1), le schéma s'écrit

$$\begin{cases} \dfrac{c^\star - c^n}{\delta t} & = L(c^\star) \\ \dfrac{c^{n+1} - c^\star}{\delta t} & = R(c^\star) \end{cases} \tag{3.3}$$

Cette approche est attractive en raison de sa simplicité. L'équation de transport peut être résolue espèce par espèce car elles ne sont pas couplées par les réactions. L'équation de chimie peut être résolue de façon locale, car il n'y figure plus aucun terme de transport.

L'erreur introduite par le découplage, souvent appelée "erreur de splitting", est le principal inconvénient de cette méthode. Elle est proportionnelle au pas de temps et au taux de réaction (Hammond (2003)). Pour que la méthode fonctionne bien, il faut donc que le transport soit suffisamment rapide par rapport aux réactions (Steefel and MacQuarrie (1996)). Il est possible de réduire l'erreur de splitting, en utilisant une méthode centrée en temps ("strang splitting", Strang (1968)), qui consiste à effectuer une demi-étape de transport, suivie d'une étape de chimie, suivie d'une demi-étape de transport.

3.2.2 Approche séquentielle itérative (SIA)

L'approche itérative permet de corriger l'erreur de splitting en itérant entre les étapes de transport et de chimie. Le principe est d'obtenir à $(n+1)$, une solution couplée implicite, à la fois pour le terme de transport et le terme de chimie. Pour cela, on ne résout pas les deux opérateurs simultanément comme dans l'approche globale, mais on itère entre transport et chimie. La première étape est une étape de transport implicite avec un terme de réaction explicite, fourni par l'itération précédente. La deuxième étape est une étape de chimie implicite, avec un terme de transport explicite fourni par l'étape de transport précédente. Les itérations continuent tant que les concentrations des espèces obtenues par chacune des deux étapes ne coïncident pas, avec une certaine tolérance. Sur le problème (3.1), ce schéma s'écrit

$$\begin{cases} \dfrac{c^{n+1,k} - c^n}{\delta t} & -L(c^{n+1,k}) & = R(c^{n+1,k-1}) \\ \dfrac{c^{n+1,k+1} - c^n}{\delta t} & -R(c^{n+1,k+1}) & = L(c^{n+1,k}) \end{cases} \tag{3.4}$$

Il existe plusieurs variantes de ce schéma, mais le principe est toujours le même.

Cette méthode est celle préconisée par Yeh and Tripathi (1989), car c'est selon eux la seule qui puisse permettre de simuler des cas réalistes (en 2D et 3D). Elle est attractive en raison de sa modularité et parce qu'elle permet de réduire l'erreur de splitting grâce au processus itératif. Mais dans certains cas, la procédure itérative de la méthode peut avoir des difficultés à converger (Engesgaard and Kipp

(1992), Steefel and MacQuarrie (1996)), notamment pour des problèmes quasi-stationnaires (par exemple déplacement lent d'un front de précipitation-dissolution).

Une comparaison des méthodes SNIA et SIA effectuée par Steefel and MacQuarrie (1996), indique que pour obtenir une précision équivalente, il est plus coûteux de réduire le pas de temps avec l'approche SNIA, que d'effectuer des itérations avec un pas de temps plus grand pour la méthode SIA.

3.3 Un splitting non itératif avec pénalisation

Le premier objectif visé à l'IFP est de construire une méthode permettant de prendre en compte le couplage d'un écoulement multiphasique compositionnel avec la géochimie, tout en conservant des temps de simulation du même ordre de grandeur que ceux obtenus sans la géochimie. C'est la raison pour laquelle l'utilisation d'une méthode itérative comme dans l'approche SIA ne semble pas convenir.

Le deuxième objectif visé est d'obtenir une méthode modulaire qui permette d'utiliser les logiciels existants. On souhaite donc à priori favoriser une méthode de splitting d'opérateurs.

On propose donc une méthode qui s'inspire des méthodes de splitting présentées ci-dessus dans la section 3.2, et qui permette de satisfaire les deux objectifs. Cette méthode est une méthode de splitting non itérative symétrique. C'est à dire que l'étape de transport comporte un terme de réaction explicite, et l'étape de chimie un terme de transport explicite (comme dans l'approche SIA).

Cependant il est important de réduire l'erreur de splitting pour garder la cohérence du modèle physique couplé. La solution étudiée dans cette thèse consiste à ajouter un terme de pénalisation afin de réduire progressivement l'erreur au cours du temps. Le principe est de pénaliser l'équation de transport en ajoutant l'erreur faite sur les concentrations.

Pour le problème (3.1), le schéma s'écrit

$$\left\{ \begin{array}{ll} \dfrac{c^{n+1,\star} - c^{n,\star}}{\delta t} - L(c^{n+1,\star}) & = R(c_i^n) + \lambda \left(c^{n,\star} - c^n\right) \\ \dfrac{c^{n+1} - c^n}{\delta t} - R(c^{n+1}) & = L(c^{n+1,\star}) \end{array} \right. \tag{3.5}$$

Le schéma écrit pour le problème (3.1) donne une idée du type de méthode que l'on souhaite utiliser. Mais le problème à traiter étant multiphasique, les réactions chimiques peuvent se produire entre espèces appartenant à deux phases différentes (par exemple les réactions d'équilibre eau-gaz), et la méthode étudiée au cours de cette thèse, appelée (*Rs-Tr*) et inspirée du schéma (3.5), tient compte de cette problématique. On la présente en détail au chapitre 5, et elle sera étudiée par comparaison avec une méthode globale présentée au chapitre 4.

Chapitre 4

Approche globale implicite

L'approche globale implicite consiste à résoudre simultanément toutes les équations, c'est à dire les équations algébriques locales et les équations aux dérivées partielles. On utilise une méthode de Newton classique pour linéariser le problème.

Le schéma numérique est implicite sur tous les termes.

4.1 Discrétisation des équations de bilan

On pose

$$Ł_i^{n+1} = \begin{cases} \dfrac{N_i^{n+1} - N_i^n}{\delta t} + L\left(v_{\alpha(i)}^{n+1}\right) c_i^{n+1} - (Q_i)^{n+1} & \text{si } E_i \in \mathcal{S}_{mob} \\ Ł_i^{n+1} = \dfrac{N_i^{n+1} - N_i^n}{\delta t} & \text{si } E_i \in \mathcal{S}_{fix} \end{cases} \tag{4.1}$$

avec

$$N_i^{n+1} = \pi_i\left(\Phi^{n+1}, \xi_{\alpha(i)}^{n+1}, \mathcal{S}_{\alpha(i)}^{n+1}\right) x_i^{n+1} \tag{4.2}$$

Le schéma numérique utilisé pour la résolution du système d'équation (2.35) est alors

$$\begin{aligned} &\bullet\ Ł_i^{n+1} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\alpha}^{n+1} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n+1} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1} = 0 \quad \forall i = 1, \cdots, N_p + N_s, \\ &\bullet\ Ł_k^{n+1} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1} = 0 \quad \forall k = N_p + N_s + 1, \cdots, N_k. \end{aligned} \tag{4.3}$$

Comme on est dans une approche globale, tous les taux de production sont implicites. Il est nécessaire d'éliminer toutes les inconnues secondaires, c'est à dire tous les taux de production correspondant à des réactions équilibrées. On utilise pour cela la méthode de réduction du système présentée dans la section 2.4. Les équations (2.29) et (2.30) permettent d'exprimer les taux de production des réactions équilibrées en fonction des espèces secondaires, et on obtient alors les équations de bilan suivantes :

Pour $i = 1, \cdots, N_p$

$$Ł_i^{n+1} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1} = 0 \tag{4.4}$$

Pour $k = N_p + N_s + 1, \cdots, N_k$,

$$Ł_k - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{k,r} I_{r,k}^{n+1} = 0 \tag{4.5}$$

4.2 Schéma numérique complet

Le schéma numérique pour résoudre le problème (2.35) est composé des équations (4.6) à (4.12)

- les équations de bilan

$$\begin{cases} Ł_i^{n+1} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1} = 0, \quad \forall i = 1, \cdots, N_p \\ Ł_k - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{k,r} I_{r,k}^{n+1} = 0, \quad \forall k = N_p + N_s + 1, \cdots, N_k \end{cases} \tag{4.6}$$

- les équations de pression capillaire

$$P_{\alpha}^{n+1} = P_w^{n+1} + P_{c,\alpha}(S_{\alpha}^{n+1}), \quad \forall \alpha = 1, \cdots, N_{\alpha} - 1 \tag{4.7}$$

- les lois d'action de masse

$$\prod_{i=1}^{N_p+N_s} \left(a_i(x_i^{n+1})\right)^{\nu_{i,r}} = K_r, \quad \forall r = R_{\text{eq}}^{\alpha} + 1, \cdots, R_{\text{eq}} \tag{4.8}$$

- les lois de Henry

$$y^{n+1} = x^{n+1} H_r^{n+1}, \quad \forall r = 1, \cdots, R_{\text{eq}}^{\alpha} \tag{4.9}$$

- l'équation de fermeture du volume total

$$\Phi^{n+1} + \sum_{i \in \mathcal{S}_{fix}} \phi_i^{n+1} = 1 \tag{4.10}$$

- l'équation de fermeture du volume fluide

$$\sum_{i=1}^{N_{\alpha}} S_{\alpha}^{n+1} = 1 \tag{4.11}$$

- les équations de fermeture des compositions des phases fluides

$$\sum_{i, \alpha(i)=\alpha} x_i^{n+1} = 1, \quad \forall \alpha = 1, \cdots, N_{\alpha} \tag{4.12}$$

Seules les équations de bilan sur les espèces mobiles sont des équations non locales. Il y a donc $(n(\mathcal{S}_{fix}) + R_{\text{eq}} + 2N_{\alpha} + 1)$ équations non locales à résoudre, et $(N_p + N_k - n(\mathcal{S}_{fix}))$ équations locales.

On utilise une méthode de Newton classique (avec stockage de la jacobienne) pour résoudre ce système. Les équations locales sont éliminées au niveau du solveur afin de réduire la dimension de la matrice à inverser à chaque itération.

Chapitre 5

Résolution par splitting, la méthode (Rs-Tr)

Dans le cadre de la simulation du stockage de co_2, l'IFP souhaite développer des méthodes de simulation permettant de simuler le couplage d'un écoulement multiphasique compositionnel avec la géochimie.

La méthode doit permettre de conserver des temps de simulation du même ordre de grandeur que ceux obtenus lors d'un calcul d'écoulement sans prise en compte de la géochimie. C'est la raison pour laquelle, l'utilisation d'une méthode itérative ne semble pas convenir.

Le deuxième objectif visé est d'obtenir une méthode modulaire qui permette d'utiliser les logiciels existants. On souhaite donc à priori favoriser une méthode de splitting d'opérateur.

La méthode étudiée au cours de cette thèse, proposée par Michel and Trenty (2005), tente de répondre à ces deux objectifs. Elle s'appelle (*Rs-Tr*). Il s'agit d'un splitting séquentiel non itératif à deux pas. La première étape, (*Rs*), est une étape de résolution du problème d'écoulement multiphasique et la seconde, (*Tr*), est une étape de transport réactif.

La méthode doit également tenir compte de la problématique particulière du stockage du co_2, c'est à dire prendre en compte le fait que l'écoulement est multiphasique. En effet, les réactions d'équilibre entre phases fluides ont un effet important sur l'écoulement, et le découplage des opérateurs doit en tenir compte. Le découplage repose donc sur des hypothèses physiques liées au problème. Elles sont présentées dans la section 5.1. Les hypothèses sur le découplage guident l'écriture des schémas utilisés pour la résolution des deux modèles (*Rs*) et (*Tr*). Ces schémas diffèrent essentiellement par le choix d'impliciter ou d'expliciter chacun des termes. Ils sont présentés dans la section 5.2.

La réduction de l'erreur de splitting, qui peut être vue comme une stratégie de couplage des modèles (*Rs*) et (*Tr*), repose sur une méthode de pénalisation non itérative présentée dans la section 5.3.

Un atout de la méthode (*Rs-Tr*) est de pouvoir utiliser des pas de temps différents pour chacune des deux étapes (section 5.4). On considère en effet que c'est le calcul de transport réactif qui contraint le pas de temps. La méthode permet d'utiliser des sous pas de temps pour la résolution de (*Tr*), sans contraindre le pas de temps pour la résolution de (*Rs*).

Le transport réactif étant la partie la plus contraignante et la plus coûteuse à résoudre, on s'intéresse finalement aux différentes méthodes de résolution possibles dans la section (5.5).

5.1 Principe et hypothèses

Le schéma de splitting étudié, appelé (*Rs-Tr*), est un schéma de splitting à deux pas. La première étape, notée (*Rs*), consiste à résoudre un modèle réservoir permettant de calculer les vitesses et volumes des différentes phases, ainsi que les quantités de matière échangées entre l'eau et le gaz. La deuxième étape, notée (*Tr*), est un modèle de transport réactif, qui tient compte des réactions de spéciation et des cinétiques de réaction eau-roche dans un champ de vitesse donné par le modèle réservoir.

Le découplage du problème en deux modèles résulte des hypothèses suivantes :

Hypothèse H1

Les échanges eau-gaz influencent fortement l'écoulement.

Hypothèse H2
Les spéciations et les cinétiques de dissolution-précipitation influencent faiblement et lentement l'écoulement.

L'hypothèse H1 indique qu'il n'est pas judicieux de découpler le calcul des flux du calcul des réactions équilibrées entre l'eau et le gaz. En effet, si ces réactions influencent fortement l'écoulement, l'erreur de splitting sera probablement importante. On choisit donc de résoudre simultanément le calcul des flux et les réactions équilibrées entre l'eau et le gaz. A l'inverse, l'hypothèse H2 indique qu'il est possible de découpler le calcul des réactions équilibrées au sein de l'eau et le calcul des cinétiques du calcul des flux. En effet, comme on suppose que ces réactions influencent peu l'écoulement et de façon lente, l'erreur de splitting sera probablement faible. Les hypothèses H1 et H2 sont des hypothèses heuristiques qui permettent de comprendre les choix initiaux qui ont guidé l'écriture du schéma (*Rs*-*Tr*). Dans le chapitre 7, on essaiera de valider ces hypothèses à l'aide de résultats numériques.

Chacun des deux modèles possèdent son propre jeux d'inconnues. Une inconnue du modèle (*Rs*) peut être transmise au modèle (*Tr*) et vice-versa. Pour distinguer les inconnues des deux modèles, on indice les variables par (*Rs*) ou (*Tr*). Lorsqu'il n'y a pas d'ambiguïté, les variables ne portent pas d'indice.

5.2 Le schéma numérique

Les schémas (5.3) et (5.10) sont tous les deux similaires au schéma (4.3) dans leur formulation. Pour chacun de ces schémas, on fait le choix d'impliciter ou d'expliciter certains termes des équations.

- Le schéma (4.3) est complètement implicite.
- Le schéma (5.3) est explicite pour les cinétiques et les équilibres au sein d'une même phase.
- Le schéma (5.10) est explicite sur la vitesse, les saturations, la porosité et les équilibres eau-gaz.

Le choix d'impliciter ou d'expliciter les différents termes modifie la manière d'éliminer les inconnues secondaires, puisque seules les inconnues secondaires implicites ont besoin d'être éliminées. En effet, les inconnues secondaires explicites sont déduites des étapes précédentes.

5.2.1 Le modèle (*Rs*)

On pose

$$Ł_i^{n+1,Rs} = \begin{cases} \dfrac{N_i^{n+1,Rs} - N_i^{n,Rs}}{\delta t} + L\left(v_{\alpha(i)}^{n+1,Rs}\right) c_i^{n+1,Rs} - (Q_i)^{n+1,Rs} & \text{si } E_i \in \mathcal{S}_{mob} \\ \dfrac{N_i^{n+1,Rs} - N_i^{n,Rs}}{\delta t} & \text{si } E_i \in \mathcal{S}_{fix} \end{cases} \tag{5.1}$$

avec

$$N_i^{n+1,Rs} = \pi_i\left(\Phi^{n+1,Rs}, \xi_{\alpha(i)}^{n+1,Rs}, S_{\alpha(i)}^{n+1,Rs}\right) x_i^{n+1,Rs} \tag{5.2}$$

Le schéma numérique pour le modèle (*Rs*) s'écrit :

$$\begin{aligned} &\bullet\ Ł_i^{n+1,Rs} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\alpha}^{n+1,Rs} = \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n,Tr} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr} \qquad \forall i = 1, \cdots, N_p + N_s, \\ &\bullet\ Ł_i^{n+1,Rs} = \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr} \qquad \forall i = N_p + N_s + 1, \cdots, N_k. \end{aligned} \tag{5.3}$$

Les taux de production au sein d'une même phase sont explicites et fournis par le modèle (*Tr*), par conséquent il n'est pas nécessaire de les éliminer. Les seules inconnues secondaires à éliminer sont donc les taux de production des équilibres entre phase. On obtient ainsi

Pour $i = 1, \cdots, N_p$

$$Ł_i^{n+1,Rs} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1,Rs} = \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n,Tr} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr} \tag{5.4}$$

Pour $j = N_p + 1, \cdots, R_{\text{eq}}^{\alpha}$

$$Ł_i^{n+1,Rs} = I_{i-N_p,\alpha}^{n,Tr} \tag{5.5}$$

Pour $k = N_p + N_s + 1, \cdots, n(\mathcal{S})$

$$Ł_k^{n+1,Rs} = \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{k,r} I_{r,k}^{n,Tr} \tag{5.6}$$

Les équilibres au sein d'une phase étant explicites, il est inutile de résoudre les lois d'actions de masse. De plus, le but du modèle (Rs) étant de calculer les vitesses, les saturations et les équilibres entre phases fluides, on constate que la résolution des équations de bilan liées aux espèces fixes n'a aucune utilité. On considère donc que les espèces fixes ne sont pas des inconnues du modèle (Rs).

Finalement, les inconnues du modèle (Rs) sont les pressions (N_α inconnues), les saturations (N_α inconnues), la porosité et les fractions molaires des espèces mobiles (N_{mob} inconnues). Soit, au total, $(N_{fix} + 2N_\alpha + 1)$ équations. Les équations associées sont les $\left(N_{mob} - R_{eq}^{\alpha}\right)$ équations de bilan, les R_{eq}^{α} lois de Henry, l'équation de fermeture du volume fluide, l'équation de fermeture du volume total, les N_α équations de fermeture de la composition des phases fluides et les $(N_\alpha - 1)$ équations de pression capillaire.

Le système à résoudre est

$$\left\{\begin{array}{ll}
Ł_i^{n+1,Rs} - \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1,Rs} = \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\alpha}^{n,Tr} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr}, & \forall i = 1, \cdots, N_p, \\
Ł_i^{n+1,Rs} = I_{i-N_p,\beta}^{n,Tr}, & \forall j = N_p + R_{\text{eq}}^{\alpha} + 1, \cdots, R_{\text{eq}}, E_j \in \mathcal{S}_{mob} \\
Ł_k^{n+1,Rs} = \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{k,r} I_{r,k}^{n,Tr}, & \forall k = N_p + N_s + 1, \cdots, n(\mathcal{S}), E_k \in \mathcal{S}_{mob} \\
y^{n+1,Rs} = x^{n+1,Rs} H_r, & \forall r = 1, \cdots, R_{\text{eq}}^{\alpha} \\
\sum_{i=1}^{N_\alpha} S_\alpha^{n+1} = 1 & \\
\Phi^{n+1} + \sum_{i \in \mathcal{S}_{fix}} \phi_i^n = 1 & \\
\sum_{i,\alpha(i)=\alpha} x_i^{n+1,Rs} = 1, & \forall \alpha = 1, \cdots, N_\alpha \\
P_\alpha^{n+1} = P_w^{n+1} + P_{c,\alpha}, & \forall \alpha = 1, \cdots, N_\alpha - 1.
\end{array}\right. \tag{5.7}$$

5.2.2 Le modèle (Tr)

On pose

$$Ł_i^{n+1,Tr} = \left\{\begin{array}{ll}
\dfrac{N_i^{n+1,Tr} - N_i^{n,Tr}}{\delta t} + L\left(v_{\alpha(i)}^{n+1,Rs}\right) c_i^{n+1,Tr} - (Q_i)^{n+1,Tr} & \text{si } E_i \in \mathcal{S}_{mob} \\
Ł_i^{n+1,Tr} = \dfrac{N_i^{n+1,Tr} - N_i^{n,Tr}}{\delta t} & \text{si } E_i \in \mathcal{S}_{fix}
\end{array}\right. \tag{5.8}$$

avec

$$N_i^{n+1,Tr} = \pi_i\left(\Phi^{n+1,Rs}, \xi_{\alpha(i)}^{n+1,Tr}, S_{\alpha(i)}^{n+1,Rs}\right) x_i^{n+1,Tr} \tag{5.9}$$

Le schéma numérique pour le modèle ($\mathcal{T}r$) s'écrit :

$$\begin{aligned}
&\bullet\ \forall i = 1, \cdots, N_p + N_s, \\
&Ł_i^{n+1,\mathcal{T}r} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n+1,\mathcal{T}r} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1,\mathcal{T}r} = \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\alpha}^{n+1,\mathcal{R}s}, \\
&\bullet\ \forall i = N_p + N_s + 1, \cdots, N_k, \\
&Ł_i^{n+1,\mathcal{T}r} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1,\mathcal{T}r} = 0.
\end{aligned} \tag{5.10}$$

Les taux de production des réactions entre phases sont explicites et fournis par le modèle ($\mathcal{R}s$), par conséquent il n'est pas nécessaire de les éliminer. Les seules inconnues secondaires à éliminer sont donc les taux de production des équilibres au sein d'une même phase. On obtient ainsi

Pour $i = 1, \cdots, N_p$

$$Ł_i^{n+1,\mathcal{T}r} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} Ł_{\{r+N_p\}}^{n+1,\mathcal{T}r} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n+1,\mathcal{T}r} = \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\beta}^{n+1,\mathcal{R}s} \tag{5.11}$$

Pour $i = R_{\text{eq}}^{\alpha} + 1, \cdots, R_{\text{eq}}$

$$Ł_i^{n+1,\mathcal{T}r} = I_{i-N_p,\beta}^{n+1,\mathcal{R}s} \tag{5.12}$$

Pour $k = N_p + N_s + 1, \cdots, n(\mathcal{S})$

$$Ł_k^{n+1,\mathcal{R}s} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{k,r} I_{r,k}^{n+1,\mathcal{T}r} = 0 \tag{5.13}$$

Les vitesses et les volumes étant donnés, les pressions, les saturations et la porosité ne sont pas des inconnues du modèle de transport réactif. Par conséquent, les équations de pression capillaire, l'équation de fermeture du volume fluide, et l'équation de fermeture du volume total ne sont pas utiles. Seules les fractions molaires des espèces mobiles (N_{mob} inconnues) et les fractions volumiques des espèces fixes (N_{fix}) sont des inconnues du modèle ($\mathcal{T}r$).

D'autre part, les échanges entre phases fluides étant explicites, les équations de bilan des espèces de chaque phase fluide peuvent être résolues séparément. Si une phase fluide est pure, ou non réactive (pas de réaction homogène dans cette phase), et si ses constituants ne participent pas aux cinétiques, il n'est pas utile de résoudre les équations de bilan liées aux espèces de cette phase. Dans les situations étudiées au cours de cette thèse, la phase gazeuse se trouve dans ce cas.

Les équations à résoudre sont donc les $\left(N - R_{eq}^{\beta}\right)$ équations de bilan, les R_{eq}^{β} lois d'action de masse et les N_α équations de fermeture de la composition des phases fluides.

On constate alors que le système est sur-déterminé, car on dispose de $(N + N_\beta)$ équations pour N inconnues. Ceci découle directement de la façon de concevoir le problème. On impose un changement de volume qui ne tient pas compte du changement de porosité lié aux réactions de précipitation-dissolution, mais les hypothèses physiques suggèrent que ces changements de porosité sont négligeables et très lents. Pour ne pas devoir choisir arbitrairement de supprimer l'une des équations, on ajoute N_α inconnues supplémentaires (une par phase fluide), notées c_α, qui permettent de compenser la variation de volume due aux réactions de précipitation dissolution. Ceci revient à dire que le volume d'une phase fluide α est $V_\alpha = V\Phi S_\alpha c_\alpha$ au lieu de $V_\alpha = V\Phi S_\alpha$, où V est le volume géométrique. Si l'hypothèse faite, à savoir, les variations dues à la précipitation-dissolution sont négligeables, les inconnues c_α restent très proches de 1.

5.3 Stratégie de couplage

Nous avons découplé le problème en deux modèles ($\mathcal{R}s$) et ($\mathcal{T}r$). Chaque modèle utilise des termes transmis par l'autre, mais il possède son propre jeux d'inconnues. Il n'y a donc aucune raison que les compositions des phases obtenues avec chacun des deux modèles coïncident. Il est donc nécessaire de

définir une stratégie pour réduire l'erreur entre les inconnues de (Rs) et (Tr), c'est à dire réduire la quantité

$$E = N_i^{Rs} - N_i^{Tr} \tag{5.14}$$

Pour réduire cette erreur, la méthode que nous choisissons d'étudier est une méthode de correction d'erreur non itérative. Le principe est d'ajouter un terme de pénalisation pour faire décroître progressivement l'erreur au cours du temps. Le terme de pénalisation est paramétré par un coefficient $\lambda > 0$, homogène à l'inverse d'un temps. Ce coefficient détermine la rapidité avec laquelle l'erreur sera corrigée.

Le terme de pénalisation peut être ajouté aussi bien sur le modèle (Rs) que sur le modèle (Tr), ou de façon symétrique sur les deux modèles. La solution la plus robuste est de l'ajouter sur le modèle (Tr) plutôt que sur le modèle (Rs) car on peut alors le faire de façon implicite, la quantité $N_i^{n+1,Rs}$ ayant déjà été calculée.

Néanmoins, deux arguments penchent en faveur d'un ajout du terme de pénalisation sur le modèle (Rs). Premièrement, on souhaite introduire par la suite une stratégie de simplification du système géochimique utilisé par le modèle (Rs). Cela signifie que le modèle (Rs) ne connaît qu'une description grossière de la géochimie, alors que le modèle (Tr) utilise une description fine. Or, il est difficile de reconstruire une description fine à partir d'une description grossière alors que l'inverse est plus facile. Deuxièmement, le modèle réservoir est réputé plus stable, et on pense à priori qu'il sera moins perturbé par l'ajout du terme de pénalisation.

Le schéma (Rs) s'écrit alors :

$$\begin{aligned}
&\bullet\ \forall i = 1, \cdots, N_p + N_s, \\
&Ł_i^{n+1,Rs} + \lambda\left(N_i^{n,Rs} - N_i^{n,Tr}\right) = \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\alpha}^{n+1,Rs} + \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n,Tr} + \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr}, \\
&\bullet\ \forall i = N_p + N_s + 1, \cdots, N_k, \\
&Ł_i^{n+1,Rs} + \lambda\left(N_i^{n,Rs} - N_i^{n,Tr}\right) = \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,Tr}.
\end{aligned} \tag{5.15}$$

Le terme de pénalisation étant ajouté de façon explicite, le paramètre λ ne doit pas être trop grand afin de ne pas déstabiliser le schéma.

L'effet du terme de pénalisation sur le schéma sera étudié sur un problème simplifié 0D, ainsi que d'autres stratégies de couplage. De plus, le choix d'ajouter le terme de pénalisation sur le modèle (Rs) parait raisonnable, mais les autres choix méritent d'être considérés, notamment sur des cas théoriques plus simples (cf. chapitre 8).

5.4 Utilisation de sous-pas de temps

Le découplage du problème en deux modèles permet d'utiliser des pas de temps différents pour la résolution de (Rs) et pour la résolution de (Tr). Il suffit de vérifier

$$(\delta t)_{Rs}^{n} = \sum_{k} (\delta t)_{Tr}^{n,k} \tag{5.16}$$

où $(\delta t)_{Rs}^{n}$ est le pas de temps du modèle (Rs) à l'étape n et $(\delta t)_{Tr}^{n,k}$ le k^{e} pas de temps de la même étape n pour le modèle (Tr).

Le schéma (Tr), avec sous-pas de temps, s'écrit de la manière suivante. Pour $k = 0$, on pose $t_{n,k} = t_n$ et $u^{n,k,Tr} = u^{n,Tr}$ pour toutes variables u de (Tr). Tant que $t_{n,k+1} \neq t_{n+1}$, on résout

$$\begin{aligned}
&\bullet\ \forall i = 1, \cdots, N_p + N_s, \\
&Ł_i^{n,k+1,Tr} - \sum_{r=R_{\text{eq}}^{\alpha}+1}^{R_{\text{eq}}} \nu_{i,r} I_{r,\beta}^{n,k+1,Tr} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,k+1,Tr} = \frac{(\delta t)_{Tr}^{n,k}}{(\delta t)_{rs}^{n}} \sum_{r=1}^{R_{\text{eq}}^{\alpha}} \nu_{i,r} I_{r,\alpha}^{n+1,Rs}, \\
&\bullet\ \forall i = N_p + N_s + 1, \cdots, N_k, \\
&Ł_i^{n,k+1,Tr} - \sum_{r=R_{\text{eq}}+1}^{R_{\text{eq}}+R_{\text{kin}}} \nu_{i,r} I_{r,k}^{n,k+1,Tr} = 0.
\end{aligned} \tag{5.17}$$

avec

$$Ł_i^{n,k+1,\mathcal{T}r} = \begin{cases} \dfrac{N_i^{n,k+1,\mathcal{T}r} - N_i^{n,k,\mathcal{T}r}}{\delta t} + L\left(v_{\alpha(i)}^{n+1,\mathcal{R}s}\right) c_i^{n,k+1,\mathcal{T}r} - (Q_i)^{n,k+1,\mathcal{T}r} & \text{si } E_i \in \mathcal{S}_{mob} \\ Ł_i^{n,k+1,\mathcal{T}r} = \dfrac{N_i^{n,k+1,\mathcal{T}r} - N_i^{n,k,\mathcal{T}r}}{\delta t} & \text{si } E_i \in \mathcal{S}_{fix} \end{cases} \tag{5.18}$$

et

$$N_i^{n,k+1,\mathcal{T}r} = \pi_i\left(\Phi^{n+1,\mathcal{R}s}, \xi_{\alpha(i)}^{n,k+1,\mathcal{T}r}, S_{\alpha(i)}^{n+1,\mathcal{R}s}\right) x_i^{n,k+1,\mathcal{T}r} \tag{5.19}$$

La variation de volume entre t_n et t_{n+1} est fixée par le modèle ($\mathcal{R}s$) par l'intermédiaire des saturations et de la porosité. Cette variation de volume est alors imposée sur le premier sous pas de temps du modèle ($\mathcal{T}r$). En fait, le modèle ($\mathcal{R}s$) calcule une variation sur un intervalle de temps $(t_{n+1} - t_n)$, mais le modèle ($\mathcal{T}r$) doit absorber cette variation sur le premier sous pas de temps correspondant à un intervalle de temps $(t_{n,1} - t_{n,0})$.

Les échanges équilibrés entre phases sont calculés par le modèle ($\mathcal{R}s$) par unité de temps, afin d'être imposés progressivement sur le modèle ($\mathcal{T}r$).

5.5 Stratégie de résolution du modèle de transport réactif

On peut utiliser différentes méthodes pour la résolution du problème de transport réactif. Ce choix est important car c'est à priori la résolution du transport réactif qui contraint le plus le pas de temps et qui est la plus coûteuse en terme de temps calcul et de place mémoire.

On se retrouve face aux mêmes choix que pour la résolution du problème complet, soit résolution globale implicite ou par splitting.

Une résolution par splitting présente l'avantage important de pouvoir résoudre les termes liés à la chimie de façon locale. Ceci est important, car cela permet d'utiliser des sous-pas de temps locaux. En effet, en présence d'un front, seules quelques mailles localisées nécessitent de petits pas de temps. De plus, il est beaucoup moins coûteux de résoudre N systèmes de taille n que de résoudre un système de taille $N \times n$ lorsque N est grand.

Pour une résolution globale implicite, il se pose à nouveau le problème de la taille de la matrice jacobienne. De plus, cette méthode ne permet pas l'utilisation de sous pas de temps locaux.

Une autre solution, essayée avec succès à l'IFP, est de tirer partie de la linéarité de l'opérateur de convection dans le modèle ($\mathcal{T}r$). Ceci permet de résoudre le modèle ($\mathcal{T}r$) de façon locale en suivant l'écoulement maille par maille. Cette méthode, que l'on appelle méthode des caractéristiques, est très attractive car elle est permet une résolution locale et par conséquent l'utilisation de sous-pas de temps locaux, et ceci sans faire d'erreur de splitting. L'inconvénient majeur de cette méthode est qu'elle n'est applicable que si l'écoulement est purement convectif. En particulier, s'il existe des flux de diffusion-dispersion du même ordre de grandeur que les flux de convection, la méthode n'est plus applicable. C'est pourquoi, nous consacrerons un chapitre de cette thèse (chapitre 10) à l'étude de la diffusion-dispersion, afin d'évaluer son importance.

Chapitre 6

Étude mathématique de la convergence du schéma pénalisé

Nous avons présenté dans la section 5.3, une méthode de correction d'erreur par pénalisation. Cette méthode semble attractive car elle évite un processus itératif coûteux. On souhaite étudier la convergence de ce schéma, et vérifier qu'il converge vers la même solution qu'un schéma global.

On se limite pour cette étude à un problème simplifié, de type diffusion-réaction. On considère du co_2 sous forme gazeuse capable de se dissoudre dans l'eau. Le co_2 dissous dans l'eau peut alors réagir avec un minéral (précipitation et/ou dissolution). On note $X \in [0, \overline{X}]$, la fraction molaire du co_2 dans l'eau et S la saturation de la phase aqueuse. L'équilibre eau-gaz se modélise par la loi de Henry, c'est à dire que pour (x,t) donné,

- soit $S(x,t) = 1$ et $X(x,t) \in [0, \overline{X}]$
- soit $S(x,t) < 1$ et $X(x,t) = \overline{X}$.

Cette alternative se traduit par la relation suivante

$$\left(X(x,t) \leq \overline{X} \text{ et } S(x,t) = 1\right) \text{ ou } \left(X = \overline{X} \text{ et } S(x,t) < 1\right). \tag{6.1}$$

Comme la phase gazeuse ne peut être que pure, la diffusion n'est possible qu'en phase aqueuse. Dans ce modèle simplifié, on néglige la compressibilité du gaz, et on suppose que la porosité est constante et uniforme. La réaction avec le minéral est modélisée par une loi cinétique. On note w, le rapport entre la fraction volumique de minéral et la porosité. Si $X > \widetilde{X}$, alors le minéral se dissout à la vitesse $\mu\left(f(w)\left(\widetilde{X} - X\right)\right)$, avec $f(w) = 0$ si $w < 0$, $f(w) = w/\varepsilon$ si $w \in [0, \varepsilon]$, et $f(w) = 1$ sinon, avec $\varepsilon > 0$. Si $X < \widetilde{X}$, alors le minéral précipite à la vitesse $\mu\left(X - \widetilde{X}\right)$. Le paramètre μ est la constante de précipitation et de dissolution du minéral. Le terme de réaction s'écrit donc de la façon suivante

$$F(w,X) = \mu\left(f(w)\left(\widetilde{X} - X\right)^{+} - \left(\widetilde{X} - X\right)^{-}\right) \tag{6.2}$$

avec $x^{+} = \max(x,0)$ et $x^{-} = \max(-x,0)$. Cette écriture du terme de réaction est reprise des travaux de Bouillard (2006), mais a été légèrement modifiée de façon à être continue. On souhaite ainsi éviter les problèmes dus à la discontinuité, qui ne sont pas l'objet de notre étude. Si l'on fait tendre ε vers 0, on retrouve alors une écriture discontinue. En général, on a $\widetilde{X} << \overline{X}$. Sous ces hypothèses, les équations de conservation du co_2 sous forme aqueuse, du co_2 sous forme gazeuse et du minéral s'écrivent (sous leur forme adimensionnée)

$$\begin{cases} (SX)_t - \operatorname{div}(S\nabla X) + \mathcal{T}^{wg} = F(w,X) \\ \xi\,(1-S)_t - \mathcal{T}^{wg} = 0 \\ w_t = -F(w,X) \end{cases}$$

ou ξ est le rapport entre la densité molaire du gaz et celle de l'eau. Le terme d'échange eau-gaz (adimensionné), noté $\mathcal{T}^{wg}$, est une inconnue secondaire du problème. Il correspond à la quantité de co_2 échangée

entre l'eau et le gaz (en sommant les deux premières équations de conservation, on retrouve une équation de conservation classique sur la totalité du co_2, quelle que soit la phase dans laquelle il est présent). Cette écriture est nécessaire pour pouvoir écrire la méthode de splitting en temps que l'on souhaite étudier.

L'équation (6.1) permet d'écrire les relations suivantes

$$\begin{array}{l} SX = X + \overline{X}\,(S-1) \\ S\nabla X = \nabla X \text{ p.p} \end{array}$$

En effet, soit $S = 1$ et c'est évident, soit $S < 1$ et alors $\nabla X = 0$. On peut alors réécrire les équations de conservation du co_2 sous la forme

$$\begin{array}{l} \left(X_t + \overline{X} S_t\right) - \triangle X + \mathcal{T}^{wg} = F(w, X) \\ \xi\,(1-S)_t - \mathcal{T}^{wg} = 0 \end{array}$$

On utilise cette dernière écriture des équations pour définir le problème sous sa forme couplée et sous sa forme découplée. Le problème couplé s'écrit

$$\left\{ \begin{array}{l} \left(X_t + \overline{X} S_t\right) - \triangle X + \mathcal{T}^{wg} = F(w, X) \\ \xi\,(1-S)_t - \mathcal{T}^{wg} = 0 \\ w_t = -F(w, X) \\ \left(X(x,t) \leq \overline{X} \text{ et } S(x,t) = 1\right) \text{ ou } \left(X = \overline{X} \text{ et } S(x,t) < 1\right). \end{array} \right.$$

Pour écrire le problème découplé, on définit une fraction molaire "réservoir", notée X^{Rs} et une fraction molaire "transport réactif", notée X^{Tr}. On sépare alors l'équilibre eau-gaz, calculé sur X^{Rs}, de la cinétique, calculée sur X^{Tr}. La méthode de splitting étant non itérative l'erreur est corrigée par un terme de pénalisation. Le problème s'écrit donc

$$\left\{ \begin{array}{l} \left(X_t^{Rs} + \overline{X} S_t\right) - \triangle X^{Rs} + \mathcal{T}^{wg} = F(w, X^{Rs}) \\ \left(X_t^{Tr} + \overline{X} S_t\right) - \triangle X^{Tr} + \mathcal{T}^{wg} + \lambda\left(X^{Tr} - X^{Rs}\right) = F(w, X^{Tr}) \\ \xi\,(1-S)_t - \mathcal{T}^{wg} = 0 \\ w_t = -F(w, X^{Tr}) \\ \left(X^{Rs}(x,t) \leq \overline{X} \text{ et } S(x,t) = 1\right) \text{ ou } \left(X^{Rs} = \overline{X} \text{ et } S(x,t) < 1\right). \end{array} \right.$$

On pose $u = SX^{Rs} + \xi(1-S)$ et $v = X^{Tr}$. On a $u = X^{Rs} + (S-1)(\overline{X} - \xi)$ et S et X^{Rs} sont maintenant des fonctions de u.

6.1 Formulation mathématique du problème

Soit Ω un ouvert borné de $\mathbb{R}^n$. On considère le problème couplé, noté (P), suivant

$$(P) \left\{ \begin{array}{l} u_t - \triangle\psi(u) = F\left(w, \psi(u)\right) \text{ dans } \Omega \times (0,T) \\ w_t = -F\left(w, \psi(u)\right) \text{ dans } \Omega \times (0,T), \end{array} \right.$$

avec les conditions initiales et limites suivantes

$$\left\{ \begin{array}{l} u(x,t) = 0, \text{ dans } \partial\Omega \times (0,T) \\ u(.,0) = u_0(x) \text{ dans } \Omega \\ w(.,0) = w_0(x) \text{ dans } \Omega \text{ et } w_0(x) \geq 0 \forall x \in \Omega. \end{array} \right.$$

Le problème découplé, noté $(\overline{P}_\lambda)$, s'écrit

$$(\overline{P}_\lambda) \left\{ \begin{array}{l} u_t - \triangle\psi(u) = F\left(w, v\right) \text{ dans } \Omega \times (0,T) \\ v_t - \triangle v + (u_t - (\psi(u))_t) + \lambda\left(v - \psi(u)\right) = F\left(w, v\right) \text{ dans } \Omega \times (0,T) \\ w_t = -F\left(w, v\right) \text{ dans } \Omega \times (0,T), \end{array} \right.$$

avec les conditions initiales et limites suivantes

$$\left\{ \begin{array}{l} u(x,t) = 0, \text{ dans } \partial\Omega \times (0,T) \\ v(x,t) = 0, \text{ dans } \partial\Omega \times (0,T) \\ u(.,0) = u_0(x) \text{ dans } \Omega \\ v(.,0) = v_0(x) \text{ dans } \Omega \\ w(.,0) = w_0(x) \text{ dans } \Omega \text{ et } w_0(x) \geq 0 \forall x \in \Omega \end{array} \right.$$

En soustrayant les deux premières équations et en posant $z = \psi(u) - v$, on obtient le problème (P_λ), équivalent au problème $(\overline{P}_\lambda)$,

$$(P_\lambda)\begin{cases} u_t - \triangle\psi(u) = F(w, \psi(u) - z) \text{ dans } \Omega \times (0,T) \\ z_t - \triangle z + \lambda z = 0 \text{ dans } \Omega \times (0,T) \\ w_t = -F(w, \psi(u) - z) \text{ dans } \Omega \times (0,T) \end{cases}$$

avec les conditions initiales et limites suivantes

$$\begin{cases} u(x,t) = 0, \text{ dans } \partial\Omega \times (0,T) \\ z(x,t) = 0, \text{ dans } \partial\Omega \times (0,T) \\ u(.,0) = u_0(x) \text{ dans } \Omega \\ z(.,0) = z_0(x) \text{ dans } \Omega \\ w(.,0) = w_0(x) \text{ dans } \Omega \text{ et } w_0(x) \geq 0 \forall x \in \Omega \end{cases}$$

Si on pose $\widetilde{z} = e^{\lambda t} z$, on constate que $\widetilde{z}$ vérifie une équation de la chaleur. Par conséquent, si $z_0 = 0$, la solution pour z est identiquement nulle, et le problème (P_λ) est équivalent au problème (P). Mais, en pratique, chacune des équations du problème est résolue par un code différent et par conséquent l'erreur initiale n'est pas nulle. Les deux problèmes ne sont alors plus équivalents.

Hypothèse 6.1.1
La fonction $F : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ vérifie les propriétés suivantes

- *F est lipschitzienne par rapport à ses deux arguments. On note L_{Fa} sa constante de Lipschitz par rapport à son premier argument et L_{Fb} sa constante de Lipschitz par rapport à son deuxième argument.*
- *$F(a,b)$ est croissante en a et décroissante en b, pour tout $a, b \in \mathbb{R}$.*
- *$F(0,b) = 0$, pour tout $b \in \mathbb{R}$, tel que $b \leq \widetilde{X}$ et $F(a, \widetilde{X}) = 0$, pour tout $a \in \mathbb{R}$.*
- *La fonction F vérifie :*

$$\begin{array}{lll} F(a,b) = F(0,b) = 0 & \forall a \in]-\infty, 0[& et \quad \forall b \in [0, +\infty[\\ F(a,b) = F(0,0) = 0 & \forall a \in]-\infty, 0[& et \quad \forall b \in]-\infty, 0[\\ F(a,b) = F(a,0) & \forall a \in [0, +\infty[& et \quad \forall b \in]-\infty, 0[\end{array}$$

Ces hypothèses impliquent que pour $b \leq \widetilde{X}$, $F(a,b) \leq 0$ et pour $b \geq \widetilde{X}$, $F(a,b) \geq 0$.

Hypothèse 6.1.2
La fonction $\psi : \mathbb{R} \to \mathbb{R}$ est une fonction croissante lipschitzienne de constante L_ψ, telle que $\psi(0) = 0$ et telle que $|\psi(a)|_\infty \leq \overline{X}$, pour tout $a \in \mathbb{R}$.

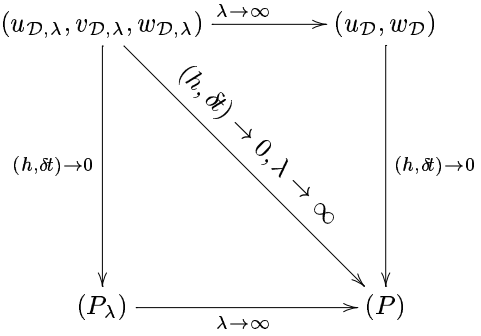
Définition 6.1.3
Sous les hypothèses 6.1.2 et 6.1.1, soit $L \in \mathbb{R}^+$, définie par $L = \max(L_{Fa}, L_\psi)$, où L_{Fb} est la constante de Lipschitz de la fonction F par rapport à son deuxième argument et L_ψ la constante de Lipschitz de la fonction ψ.

Hypothèse 6.1.4
On définit

1. *Soit Ω une ouvert borné de $\mathbb{R}^n$, et $T \in (0,T)$.*
2. *Soient $u_0(x) \in L^\infty(\Omega)$, $v_0(x) \in L^\infty(\Omega)$ et $w_0(x) \in L^\infty(\Omega)$*

Dans cette étude, on cherche à étudier les différents schémas possibles pour résoudre le problème couplé (P). On peut le résoudre directement à l'aide d'un schéma numérique sur (u,w) ou le résoudre de façon découplée, en introduisant le problème pénalisé (P_λ), pour lequel on définit un schéma numérique sur (u,v,w). On souhaite montrer toutes les convergences représentées sur le diagramme suivant par des

flèches. Pour faire cela, on cherche sur les deux schémas, soit des estimations indépendantes de λ, soit des estimations en $\frac{1}{\lambda}$.



(6.3)

Définition 6.1.5 (solution faible du problème (P))
Sous les hypothèses 6.1.4 et 6.1.1 et 6.1.2, un couple (u,w), $u \in L^\infty(\Omega\times(0,T))$ tel que $\psi(u) \in L^2(0,T;H_0^1(\Omega))$ et $w \in L^2(\Omega\times(0,T))$ est une solution faible du problème (P) si, $\forall\varphi \in C_c^\infty(\Omega\times(0,T)$

$$\begin{cases} \displaystyle\int_0^T\int_\Omega \begin{pmatrix} u(x,t)\varphi_t(x,t) - \nabla\psi(u(x,t))\nabla\varphi(x,t) \\ +F\left(w(x,t),\psi(u(x,t))\right)\varphi(x,t) \end{pmatrix} dxdt + \displaystyle\int_\Omega u_0(x)\varphi(x,0)dx = 0 \\ \displaystyle\int_0^T\int_\Omega w(x,t)\varphi_t(x,t)dxdt + \int_\Omega w_0(x)\varphi(x,0)dx = \int_0^T\int_\Omega \varphi(x,t)F\left(w(x,t),\psi(u(x,t))\right)dxdt \end{cases} \tag{6.4}$$

Définition 6.1.6 (solution faible du problème (P_λ))
Sous les hypothèses 6.1.4, 6.1.1 et 6.1.2, un triplet $(u_\lambda, v_\lambda, w_\lambda)$, $u \in L^\infty(\Omega\times(0,T))$ tel que $\psi(u) \in L^2(0,T;H_0^1(\Omega))$, $v \in L^2(0,T;H_0^1(\Omega))$ et $w \in L^2(\Omega\times(0,T))$ est une solution faible du problème (P_λ) si, $\forall\varphi \in C_c^\infty(\Omega\times(0,T))$

$$\begin{cases} \displaystyle\int_0^T\int_\Omega \begin{pmatrix} u_\lambda(x,t)\varphi_t(x,t) - \nabla\psi(u_\lambda(x,t))\nabla\varphi(x,t) \\ +F\left(w_\lambda(x,t),v_\lambda(x,t)\right)\varphi(x,t) \end{pmatrix} dxdt + \displaystyle\int_\Omega u_0(x)\varphi(x,0)dx = 0 \\ \displaystyle\int_0^T\int_\Omega \begin{pmatrix} \varphi_t(x,t)\left(v_\lambda(x,t) + u_\lambda(x,t) - \psi(u_\lambda(x,t))\right) \\ -\nabla v_\lambda(x,t)\nabla\varphi(x,t) - \lambda\varphi(x,t)(v_\lambda(x,t) - \psi(u_\lambda(x,t))) \\ +F\left(w_\lambda(x,t),v_\lambda(x,t)\right)\varphi(x,t) \end{pmatrix} dxdt \\ + \displaystyle\int_\Omega \varphi(x,0)\left(v_0(x) + u_0(x) - \psi(u_0(x))\right)dx = 0 \\ \displaystyle\int_0^T\int_\Omega w_\lambda(x,t)\varphi_t(x,t)dxdt + \int_\Omega w_0(x)\varphi(x,0)dx = \int_0^T\int_\Omega F\left(w_\lambda(x,t),v_\lambda(x,t)\right)\varphi(x,t)dxdt \end{cases} \tag{6.5}$$

Définition 6.1.7
On définit

1. *Soit Ω une ouvert borné de $\mathbb{R}^n$, et $T > 0$.*
2. *Soit $\delta t \in \mathbb{R}^+$ tel que $\delta t < T$.*
3. *Soit $\mathcal{N}_{\delta t} \in \mathbb{N}^*$ tel que $\mathcal{N}_{\delta t} = \max\{n \in \mathbb{N}, n\delta t < T\}$.*
4. *Soit $\mathcal{T}$ un maillage admissible au sens défini dans Eymard et al. (2000).*

6.1.1 Schéma discret pour la résolution de (P)

Le schéma discret, noté $P_{\mathcal{D}}$, écrit pour la résolution couplée du problème (P), est un schéma volume fini semi-implicite défini par les équations (6.6) à (6.9)

$$m_K \frac{u_K^{n+1} - u_K^n}{\delta t} - \sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1}) \right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} \psi(u_K^{n+1}) = m_K F\left(w_K^n, \psi(u_K^n)\right), \quad (6.6)$$

$$w_K^{n+1} = \max\left(w_K^{n+1} - \delta t F\left(w_K^n, \psi(u_K^{n+1})\right), 0\right). \quad (6.7)$$

$$u_K^0 = \frac{1}{m_K} \int_K u_0(x) \mathrm{d}x \quad (6.8)$$

$$w_K^0 = \frac{1}{m_K} \int_K w_0(x) \mathrm{d}x \quad (6.9)$$

Le schéma $P_{\mathcal{D}}$ est utilisé pour construire une approximation de la solution, $(u_{\mathcal{D}}, w_{\mathcal{D}})$, définie par

$$\left\{ \begin{array}{ll} u_{\mathcal{D}} & = u_K^n, \quad \forall x \in K, \forall t \in [n\delta t, (n+1)\delta t[, \forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t} + 1]\!], \\ w_{\mathcal{D}} & = w_K^n, \quad \forall x \in K, \forall t \in [n\delta t, (n+1)\delta t[, \forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t} + 1]\!]. \end{array} \right. \quad (6.10)$$

6.1.2 Schéma discret pour P_λ

On écrit les équations suivantes

$$m_K \frac{u_K^{n+1} - u_K^n}{\delta t} - \sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1}) \right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} \psi(u_K^{n+1}) = m_K F\left(w_K^n, v_K^n\right) \quad (6.11)$$

$$\begin{array}{l} m_K \dfrac{v_K^{n+1} - v_K^n}{\delta t} - \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(v_L^{n+1} - v_K^{n+1} \right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} v_K^{n+1} + m_K \frac{u_K^{n+1} - u_K^n}{\delta t} \\ \quad - m_K \dfrac{\psi(u_K^{n+1}) - \psi(u_K^n)}{\delta t} + m_K \lambda \left(\psi(u_K^{n+1}) - v_K^{n+1} \right) = m_K F(w_K^n, v_K^{n+1}) \end{array} \quad (6.12)$$

$$w_K^{n+1} = \max\left(w_K^n - \delta t F\left(w_K^n, v_K^{n+1}\right), 0\right) \quad (6.13)$$

On pose $z_K^n = \psi(u_K^n) - v_K^n$, puis en soustrayant les deux premières équations, on obtient alors le schéma $P_{\mathcal{D},\lambda}$, décrit par les équations (6.14) à (6.19),

$$z_K^0 = \frac{1}{m_K} \int_K z_0(x) \mathrm{d}x \quad (6.14)$$

$$u_K^0 = \frac{1}{m_K} \int_K u_0(x) \mathrm{d}x \quad (6.15)$$

$$w_K^0 = \frac{1}{m_K} \int_K w_0(x) \mathrm{d}x \quad (6.16)$$

$$\begin{array}{l} m_K \dfrac{z_K^{n+1} - z_K^n}{\delta t} - \displaystyle\sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(z_L^{n+1} - z_K^{n+1} \right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} z_K^{n+1} + m_K \lambda z_K^{n+1} \\ = m_K \left(F\left(w_K^n, \psi(u_K^n) - z_K^n\right) - F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right) \right) \end{array} \quad (6.17)$$

$$m_K \frac{u_K^{n+1} - u_K^n}{\delta t} - \sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1}) \right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} \psi(u_K^{n+1}) = m_K F\left(w_K^n, \psi(u_K^n) - z_K^n\right) \quad (6.18)$$

$$w_K^{n+1} = \max\left(w_K^n - \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right), 0\right) \tag{6.19}$$

Le schéma $P_{(\mathcal{D},\lambda)}$ est utilisé pour construire une approximation de la solution, $(u_{\mathcal{D},\lambda}, v_{\mathcal{D},\lambda}, w_{\mathcal{D},\lambda})$, définie par

$$\left\{\begin{array}{ll} u_{\mathcal{D},\lambda} = u_K^n, & \forall x \in K, \forall t \in [n\delta t, (n+1)\delta t[\,, \forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t}+1]\!]\,, \\ v_{\mathcal{D},\lambda} = \psi(u_K^n) - z_K^n, & \forall x \in K, \forall t \in [n\delta t, (n+1)\delta t[\,, \forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t}+1]\!]\,, \\ w_{\mathcal{D},\lambda} = w_K^n, & \forall x \in K, \forall t \in [n\delta t, (n+1)\delta t[\,, \forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t}+1]\!]\,. \end{array}\right. \tag{6.20}$$

6.2 Les estimations sur le schéma pénalisé

Proposition 6.2.1 (Estimation L^∞)
Sous les hypothèses 6.1.7, 6.1.4, 6.1.1, et 6.1.2, soient $u_{\mathcal{D},\lambda}$, $v_{\mathcal{D},\lambda}$, $z_{\mathcal{D},\lambda}$, $w_{\mathcal{D},\lambda}$, définis par le schéma (6.17) à (6.20), alors,

1. *il existe $Z \in \mathbb{R}$, dépendant uniquement de $||z^0||_\infty$, T, L et $\overline{X}$ tel que*

$$||z_{\mathcal{D},\lambda}||_{L^\infty(\Omega\times[0,T])} \leq Z, \tag{6.21}$$

2. *il existe $W \in \mathbb{R}$, dépendant uniquement de $||z^0||_\infty$, $||w^0||_\infty$, T, L, et $\overline{X}$ tel que*

$$||w_{\mathcal{D},\lambda}||_{L^\infty(\Omega\times[0,T])} \leq W. \tag{6.22}$$

3. *il existe $U \in \mathbb{R}$ dépendant uniquement de $||z^0||_\infty$, $||w^0||_\infty$, $||u^0||_\infty$, T, L et $\overline{X}$, tel que*

$$||u_{\mathcal{D},\lambda}||_{L^\infty(\Omega\times[0,T])} \leq U, \tag{6.23}$$

4. *il existe $V \in \mathbb{R}$ dépendant uniquement de $||z^0||_\infty$, T, L et $\overline{X}$, tel que*

$$||v_{\mathcal{D},\lambda}||_{L^\infty(\Omega\times[0,T])} \leq V, \tag{6.24}$$

Preuve.

Preuve de l'item 1
On réécrit le schéma (6.17) sous la forme

$$\begin{aligned} z_K^{n+1}(1+\delta t\lambda) = z_K^n \quad & + \frac{\delta t}{m_K}\left(\sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(z_L^{n+1} - z_K^{n+1}\right) - \sum_{\sigma\in\mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma} z_K^{n+1}\right) \\ & + \delta t\left(F(w_K^n, \psi(u_K^n) - z_K^n) - F(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1})\right) \end{aligned}$$

On choisit K tel que $z_K^{n+1} = \max_{L\in\mathcal{T}} z_L^{n+1}$. On a alors $\left(z_L^{n+1} - z_K^{n+1}\right) \leq 0$. On a donc

$$z_K^{n+1}(1+\delta t\lambda) \leq z_K^n + \delta t\tau\left(w_K^n, \psi(u_K^n) - z_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)\left(\psi(u_K^{n+1}) - \psi(u_K^n) + z_K^n - z_K^{n+1}\right)$$

avec $\tau(a,b_1,b_2)$ défini par $\tau(a,b_1,b_2) = \frac{F(a,b_1)-F(a,b_2)}{b_2-b_1}$ si $b_1 \neq b_2$, $\tau(a,b_1,b_2) = L$, sinon. Les propriétés de F permettent d'écrire que

$$0 \leq \tau(a,b_1,b_2) \leq L \tag{6.25}$$

et donc

$$z_K^{n+1}(1+\delta t\lambda+\delta t\tau) \leq z_K^n(1+\delta t\tau) + \delta t L\left|\psi(u_K^{n+1}) - \psi(u_K^n)\right|$$

On a alors

$$z_K^{n+1} \leq |z_K^n|\frac{(1+\delta t\tau)}{(1+\delta t\lambda+\delta t\tau)} + \delta t\frac{L}{(1+\delta t\lambda+\delta t\tau)}\left|\psi(u_K^{n+1}) - \psi(u_K^n)\right|.$$

Or $\frac{(1+\delta t\tau)}{(1+\delta t\lambda+\delta t\tau)} \leq 1$ et comme $0 \leq \tau \leq L$, $\frac{1}{(1+\delta t\lambda+\delta t\tau)} \leq 1$, donc

$$z_K^{n+1} \leq |z_K^n| + \delta t\tau\left|\psi(u_K^{n+1}) - \psi(u_K^n)\right|.$$

Et finalement, ψ est borné donc

$$z_K^{n+1} \leq |z_K^n| + 2\delta t L\overline{X}.$$

Le même travail sur le minimum permet d'écrire l'inégalité suivante

$$\max_{K\in\mathcal{T}} \left|z_K^{n+1}\right| \leq \max_{K\in\mathcal{T}} |z_K^n| + 2\delta t L\overline{X},$$

et par récurrence sur n, on obtient finalement

$$\max_{K\in\mathcal{T}} \left|z_K^{n+1}\right| \leq \max_{K\in\mathcal{T}} z_K^0 + 2(n+1)\delta t L\,\overline{X}\ \forall n \in [\![0, \mathcal{N}_{\delta t}]\!].$$

Sous les hypothèses 6.1.7, comme $\mathcal{N}_{\delta t}\delta t < T$ et $\delta t < T$, on obtient l'estimation (6.21), avec $Z = 4TL\overline{X} + ||z_0||_\infty$.

Preuve de l'item 2

La positivité de $w_K^{n+1}, \forall n \in [\![0, \mathcal{N}_{\delta t}]\!]$, est immédiate par (6.19). Supposons maintenant que $F(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}) \geq 0$. Dans ce cas le schéma (6.19) donne

$$w_K^{n+1} \leq w_K^n \tag{6.26}$$

Supposons maintenant que $F(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}) \leq 0$. Comme ψ est borné et grâce à l'estimation précédente, on a $\left(\psi(u_K^{n+1}) - z_K^{n+1}\right) \leq \overline{X} + Z$, et de plus, $w_K^n \geq 0$. Comme F est croissante par rapport à son premier argument et décroissante par rapport à son deuxième argument, on obtient que $F(w_K^n, \psi(u_K^{n+1}), z_K^{n+1}) \geq F(0, \overline{X} + Z)$. Le schéma (6.19) donne alors

$$w_K^{n+1} \leq w_K^n - \delta t F(0, \overline{X} + Z) \tag{6.27}$$

On pose $\overline{F}_1 = \max(0, -F(0, \overline{X} + Z))$, et alors

$$w_K^{n+1} \leq w_K^n + \delta t \overline{F}_1$$

En réunissant les deux inégalités (6.26) et (6.27), on obtient alors que quelque soit l'hypothèse sur le signe de F,

$$w_K^{n+1} \leq w_K^n + \delta t \overline{F}_1$$

Par récurrence sur n, on obtient alors

$$w_K^{n+1} \leq w_K^0 + (n+1)\delta t \overline{F}_1$$

Comme $(n+1)\delta t < 2T$, on obtient finalement l'estimation (6.22) avec $W = w_K^0 + 2T\overline{F}_1$.

Preuve de l'item 3 On réécrit l'équation (6.18) sous la forme

$$u_K^{n+1} = u_K^n + \frac{\delta t}{m_K}\left(\sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right) - \sum_{\sigma\in\mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma}\psi(u_K^{n+1})\right) + \delta t F\left(w_K^n, \psi(u_K^n) - z_K^n\right) \tag{6.28}$$

Soit $K \in \mathcal{T}$ tel que $u_K^{n+1} = \max_{L\in\mathcal{T}} u_L^{n+1}$. Comme ψ est une fonction croissante, le second terme à droite de l'équation est négatif, on a alors

$$u_K^{n+1} \leq u_K^n + \delta t F\left(w_K^n, \psi(u_K^n) - z_K^n\right) \tag{6.29}$$

Comme ψ est une fonction positive et bornée, en utilisant les estimations (6.21) et (6.22) et la monotonie de F, on a que

$$F(0, Z + \overline{X}) \leq F(w_K^n, \psi(u_K^n) - z_K^n) \leq F(W, -Z) \tag{6.30}$$

Le schéma (6.18) donne alors

$$u_K^{n+1} \leq u_K^n + \delta t F\left(W, -Z\right) \tag{6.31}$$

On pose $\overline{F}_2 = F(W, -Z)$. Une récurrence sur n permet d'obtenir

$$u_K^{n+1} \leq u_K^0 + \delta t(n+1)\overline{F}_2 \tag{6.32}$$

Le même travail sur le minimum permet d'obtenir

$$\max_{K\in\mathcal{T}} |u_K^{n+1}| \leq \max_{K\in\mathcal{T}} |u_K^0| + (n+1)\delta t\overline{F}_2.$$

Sous les hypothèses (6.1.7), comme $\mathcal{N}_{\delta t}\delta t < T$ et $\delta t < T$, on obtient l'estimation (6.23) avec $U = ||u^0||_\infty + 2T\overline{F}_2$.

Preuve de l'item 4
Comme $v_K^{n+1} = \psi(u_K^{n+1}) - z_K^{n+1}$, on a, $\forall K \in \mathcal{T}, \forall n \in [\![0, \mathcal{N}_{\delta t}+1]\!]$,

$$|v_K^{n+1}| \leq V \tag{6.33}$$

avec $V = Z + \overline{X}$.
□

Définition 6.2.2
Soit $\overline{F} \in \mathbb{R}^+$ défini par

$$\overline{F} = \max(0, -F(0,-Z), F(W,-Z)) \tag{6.34}$$

où Z et W sont définis par

$$Z = 4TL\overline{X} + ||z_0||_\infty$$
$$W = w_K^0 + 2T\max(0, -F(0, \overline{X}+Z))$$

Proposition 6.2.3 (Estimations L^2)
Sous les hypothèse de la proposition 6.2.1, si $\delta t \leq 1$ et $\lambda \geq 2L$,

1. *il existe $C_1 \in \mathbb{R}$ et $C_2 \in \mathbb{R}$ ne dépendant que de $||z^0||_\infty$, $||u^0||_\infty$, Ω, T, $\widetilde{X}$ et $\overline{X}$, tels que*

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \left(\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L} \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right)^2 + \sum_{\sigma\in\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{n+1})^2 \right) \leq C_1, \tag{6.35}$$

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K(u_K^{n+1} - u_K^n)^2 \leq C_1, \tag{6.36}$$

et

$$\sum_{K\in\mathcal{T}} m_K(u_K^{n+1})^2 \leq C_2, \tag{6.37}$$

2. *il existe $C_3 \in \mathbb{R}$ et $C_4 \in \mathbb{R}$, ne dépendant que de $||z^0||_\infty$, Ω, T, et $\overline{X}$, tels que*

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K \delta t \left(z_K^{(n+1)}\right)^2 \leq \frac{C_3}{\lambda^2} + m(\Omega)\frac{||z_0||_\infty}{\lambda} \tag{6.38}$$

et

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \left(\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L} \left(z_K^{n+1} - z_L^{n+1}\right)^2 + \delta t \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}(z_K^{n+1})^2 \right) \leq C_4\frac{\delta t}{\lambda} + m(\Omega)||z_0||_\infty, \tag{6.39}$$

Preuve.

Preuve de l'item 1

On multiplie (6.18) par u_K^{n+1} et on somme sur n et sur K

$$\begin{aligned}&\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K \left((u_K^{n+1})^2 - (u_K^n)^2 + (u_K^{n+1}-u_K^n)^2\right)\\ &+\delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_L^{n+1})-\psi(u_K^{n+1})\right)\left(u_L^{n+1}-u_K^{n+1}\right) \qquad (6.40)\\ &+\delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}\psi(u_K^{n+1})u_K^{n+1} = 2\delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K u_K^{n+1} F\left(w_K^n, \psi(u_K^n)-z_K^n\right)\end{aligned}$$

On pose $T_1 = \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K \left((u_K^{n+1})^2 - (u_K^n)^2\right)$. On a

$$T_1 = \sum_{K\in\mathcal{T}} m_K (u_K^{\mathcal{N}_{\delta t}+1})^2 - \sum_{K\in\mathcal{T}} m_K (u_K^0)^2 \qquad (6.41)$$

On pose $T_2 = \delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_L^{n+1})-\psi(u_K^{n+1})\right)\left(u_L^{n+1}-u_K^{n+1}\right) + \delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}\psi(u_K^{n+1})u_K^{n+1}$.
Comme la fonction $\psi(u)$ est lipschitzienne de constante L et monotone, on a

$$T_2 \geq L\delta t \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \left(\sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_L^{n+1})-\psi(u_K^{n+1})\right)^2 + \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}\psi(u_K^{n+1})^2 \right)$$

On pose $T_3 = 2\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \delta t m_K u_K^{n+1} F\left(w_K^n, \psi(u_K^n)-z_K^n\right)$. Or, sous les hypothèses 6.1.2, et grâce aux estimations (6.21) et (6.22), on a $F\left(w_K^{n+1}, \psi(u_K^n)-z_K^n\right) \leq \overline{F}$. L'estimation (6.23) donne alors

$$T_3 \leq 4TU\, m(\Omega)\overline{F}$$

En regroupant les estimations précédentes, on obtient

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \left(\sum_{L\in\mathcal{N}(K)} \delta t \tau_{K,L}\left(\psi(u_L^{n+1})-\psi(u_K^{n+1})\right)^2 + \sum_{\sigma\in\mathcal{E}_K} \delta t \tau_{K,\sigma}\psi(u_K^{n+1})^2 \right) \leq \begin{pmatrix} \frac{4}{L}TU m(\Omega)\overline{F} \\ +\frac{1}{L}\sum_{K\in\mathcal{T}} m_K (u_K^0)^2 \end{pmatrix} \qquad (6.42)$$

En posant $C_1 = \frac{1}{L}\left(4T\overline{F}m(\Omega)U + m(\Omega)U^2\right)$, on obtient l'estimation (6.35), puis l'estimation (6.36) à partir de l'inégalité (6.40).

Une autre conséquence de l'estimation (6.40) est

$$\sum_{K\in\mathcal{T}} m_K (u_K^{n+1})^2 \leq 2\delta t U m(\Omega)\overline{F} + \sum_{K\in\mathcal{T}} m_K (u_K^n)^2$$

En utilisant l'estimation (6.23), on obtient l'estimation (6.37) avec $C_2 = 2TUm(\Omega)\overline{F} + m(\Omega)U^2$.

Preuve de l'item 2

En multipliant (6.17) par z_K^{n+1} et en sommant sur $K \in \mathcal{T}$, on obtient $T_4^{n+1} + T_5^{n+1} + T_6^{n+1} = T_7^{n+1}$, avec

$$T_4^{n+1} = \sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right) z_K^{n+1}, \qquad (6.43)$$

$$T_5^{n+1} = \frac{\delta t}{2} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(z_K^{n+1} - z_L^{n+1}\right)^2 + \frac{\delta t}{2} \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \tau_{K,\sigma}(z_K^{n+1})^2, \qquad (6.44)$$

$$T_6^{n+1} = \sum_{K\in\mathcal{T}} \delta t\, m_K \lambda \left(z_K^{n+1}\right)^2, \tag{6.45}$$

$$T_7^{n+1} = \sum_{K\in\mathcal{T}} \delta t\, m_K z_K^{n+1} \left(F\left(w_K^n, \psi(u_K^n) - z_K^n\right) - F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)\right). \tag{6.46}$$

Clairement, $T_5^{n+1} > 0$. D'autre part,

$$T_4^{n+1} = \frac{1}{2}\sum_{K\in\mathcal{T}} m_K \left((z_K^{n+1})^2 - (z_K^n)^2 + (z_K^{n+1} - z_K^n)^2\right)$$

et donc

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} T_4^{n+1} = \frac{1}{2}\sum_{K\in\mathcal{T}} m_K (z_K^{\mathcal{N}_{\delta t}+1})^2 - \frac{1}{2}\sum_{K\in\mathcal{T}} m_K (z_K^0)^2 + \frac{1}{2}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right)^2$$

L'inégalité de Young $\left(ab \le \frac{\alpha}{2}a^2 + \frac{1}{2\alpha}b^2\right)$, avec $\alpha = \lambda$ donne

$$T_7^{n+1} \le \delta t \frac{\lambda}{2}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1}\right)^2 + \frac{\delta t}{2\lambda}\sum_{K\in\mathcal{T}} m_K \left(F\left(w_K^n, \psi(u_K^n) - z_K^n\right) - F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)\right)^2$$

Comme la fonction F est lipschitzienne (par rapport à tous ses arguments) de constante L, on a

$$T_7^{n+1} \le \delta t \frac{\lambda}{2}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1}\right)^2 + L\frac{\delta t}{2\lambda}\sum_{K\in\mathcal{T}} m_K \left(\left(\psi(u_K^n) - \psi(u_K^{n+1})\right) + \left(z_K^{n+1} - z_K^n\right)\right)^2$$

Comme $(a+b)^2 \le 2a^2 + 2b^2$, on a

$$T_7^{n+1} \le \delta t \frac{\lambda}{2}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1}\right)^2 + L\frac{\delta t}{\lambda}\sum_{K\in\mathcal{T}} m_K \left(\psi(u_K^n) - \psi(u_K^{n+1})\right)^2 + L\frac{\delta t}{\lambda}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right)^2$$

Comme la fonction ψ est lipschitzienne de constante L, on a

$$T_7^{n+1} \le \delta t \frac{\lambda}{2}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1}\right)^2 + L\frac{\delta t}{\lambda}\sum_{K\in\mathcal{T}} m_K \left(u_K^n - u_K^{n+1}\right)^2 + L\frac{\delta t}{\lambda}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right)^2$$

On regroupe toutes les expressions précédentes et on somme sur n.

$$\left(\begin{array}{l} \frac{1}{2}\sum_{K\in\mathcal{T}} m_K (z_K^{\mathcal{N}_{\delta t}+1})^2 - \frac{1}{2}\sum_{K\in\mathcal{T}} m_K (z_K^0)^2 + \sum_{0}^{\mathcal{N}_{\delta t}} T_5^{n+1} \\ +\frac{1}{2}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right)^2 + \sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} \delta t m_K \lambda \left(z_K^{n+1}\right)^2 \end{array}\right) \le \left(\begin{array}{l} \delta t \frac{\lambda}{2}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1}\right)^2 \\ +L\frac{\delta t}{\lambda}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(z_K^n - z_K^{n+1}\right)^2 \\ +L\frac{\delta t}{\lambda}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(u_K^n - u_K^{n+1}\right)^2 \end{array}\right)$$

Une réécriture et l'estimation (6.36) donne

$$\left(\frac{1}{2} - L\frac{\delta t}{\lambda}\right)\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \left(z_K^{n+1} - z_K^n\right)^2 + \sum_{0}^{\mathcal{N}_{\delta t}} T_5^{n+1} + \frac{1}{2}\sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} \delta t m_K \lambda \left(z_K^{n+1}\right)^2 \le L\frac{\delta t}{\lambda}C_1 + \frac{1}{2}\sum_{K\in\mathcal{T}} m_K (z_K^0)^2 \tag{6.47}$$

Comme $\delta t \le 1$ et $\lambda \ge 2L$, on a $\left(\frac{1}{2} - L\frac{\delta t}{\lambda}\right) \ge 0$, et donc

$$\frac{2}{\lambda}\sum_{0}^{\mathcal{N}_{\delta t}} T_5^{n+1} + \sum_{0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} \delta t m_K \left(z_K^{n+1}\right)^2 \le 2L\frac{\delta t}{\lambda^2}C_1 + \frac{1}{\lambda}\sum_{K\in\mathcal{T}} m_K (z_K^0)^2 \tag{6.48}$$

Comme $T_5^{n+1} \geq 0$, on a

$$\sum_{0}^{\mathcal{N}_{\delta t}} \sum_{K \in \mathcal{T}} \delta t m_K \left(z_K^{n+1}\right)^2 \leq 2LC_1 \frac{\delta t}{\lambda^2} + m(\Omega) \frac{||z_0||_\infty}{\lambda} \tag{6.49}$$

En prenant $C_3 = 2LC_1$, on obtient l'estimation (6.38). Le deuxième terme de gauche de l'inégalité (6.48) est positif, donc en posant $C_4 = 2LC_1$, on obtient l'estimation (6.39). □

Les lemmes suivants sont des conséquences classiques des estimations précédentes, à savoir l'écriture des translations en espace et en temps, nécessaire pour la preuve de convergence.

Lemme 6.2.4 (translation en espace pour $u_{\mathcal{D},\lambda}$, $z_{\mathcal{D},\lambda}$ et $v_{\mathcal{D},\lambda}$)
Soit $\xi \in \mathbb{R}^N$. Sous les hypothèses de la proposition 6.2.1, si $\delta t \leq 1$ et $\lambda \geq 2L$,

1.
$$\int_0^T \int_{\mathbb{R}^d} \left(\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 dxdt \leq \frac{1}{2} C_1 \left(|\xi| + C_5\ size(\mathcal{T})\right) |\xi| , \tag{6.50}$$

2.
$$\int_0^T \int_{\mathbb{R}^d} \left(z_{\mathcal{D},\lambda}(x+\xi,t) - z_{\mathcal{D},\lambda}(x,t)\right)^2 dxdt \leq \frac{1}{2} C_4 \left(|\xi| + C_5\ size(\mathcal{T})\right) |\xi| , \tag{6.51}$$

3.
$$\int_0^T \int_{\mathbb{R}^d} \left(v_{\mathcal{D},\lambda}(x+\xi,t) - v_{\mathcal{D},\lambda}(x,t)\right)^2 dxdt \leq (C_1 + C_4) \left(|\xi| + C_5\ size(\mathcal{T})\right) |\xi| , \tag{6.52}$$

4. *Les translatés en espace de $w_{\mathcal{D},\lambda}$ convergent uniformément vers 0, lorsque $\xi \to 0$*

Preuve.
Preuve de l'item 1
On définit $\chi_{K|L}$ de $\mathbb{R}^d \times \mathbb{R}^d$ dans $[\![0,1]\!]$ par

$$\chi_{K|L}(x,y) = \begin{cases} 1 & \text{si } [x,y] \cap K|L \neq \varnothing \\ 0 & \text{si } [x,y] \cap K|L = \varnothing \end{cases}$$

Soit $\xi \in \mathbb{R}^d$, $\xi \neq 0$, on a

$$|\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))| \leq \sum_{K|L \in \mathcal{E}} \chi_{K|L}(x+\xi,x) \left|\psi(u_K^n) - \psi(u_L^n)\right|$$

$$|\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))| \leq \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+\xi,x) \left|D_\sigma \psi(u)\right|$$

où

$$D_\sigma \psi(u) = \begin{cases} \psi(u_L) - \psi(u_K) & \text{si } \sigma = K|L \in \mathcal{E}_{\text{int}} \\ \psi(u_K) & \text{si } \sigma \in \mathcal{E}_{\text{ext}} \cap \mathcal{E}_K \end{cases}$$

On pose $c_\sigma = \left|n_\sigma \cdot \frac{\xi}{|\xi|}\right|$, où n_σ est la normale unité extérieure à σ. On réécrit l'équation précédente de la façon suivante

$$|\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))| \leq \sum_{\sigma \in \mathcal{E}} \left(\sqrt{\chi_\sigma(x+\xi,x)} \frac{|D_\sigma \psi(u)|}{\sqrt{d_\sigma c_\sigma}}\right) \left(\sqrt{\chi_\sigma(x+\xi,x)} \sqrt{d_\sigma c_\sigma}\right)$$

En utilisant Cauchy Schwarz, on obtient

$$|\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))|^2 \leq \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+\xi,x) \frac{|D_\sigma \psi(u)|^2}{d_\sigma c_\sigma} \sum_{\sigma \in \mathcal{E}} \chi_\sigma(x+\xi,x) d_\sigma c_\sigma$$

Selon Eymard et al. (2000) (preuve du lemme 3.3), il existe $C_5 > 0$ tel que, pour presque tout $x \in \mathbb{R}^d$,

$$\sum_{\sigma\in\mathcal{E}} \chi_\sigma(x+\xi,x) d_\sigma c_\sigma \leq |\xi| + C_5 \text{ size}(\mathcal{T}). \tag{6.53}$$

D'autre part,

$$\int_{\mathbb{R}^d} \chi_\sigma(x,x+\xi)\mathrm{d}x \leq m_\sigma c_\sigma |\xi|,$$

aussi, en intégrant l'inégalité (6.2) sur $\mathbb{R}^d$ et en utilisant (6.53), on a

$$\int_{\mathbb{R}^d} \left(\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 \mathrm{d}x \leq (|\xi| + C_5 \text{ size}(\mathcal{T}))\, |\xi| \sum_{\sigma\in\mathcal{E}} \tau_{K|L} \left(D_\sigma \psi(u)\right)^2.$$

Si on intègre maintenant l'inégalité (6.2) sur $(0,T)$, il vient,

$$\int_0^T \int_{\mathbb{R}^d} \left(\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 \mathrm{d}x\mathrm{d}t \leq (|\xi| + C_5 \text{ size}(\mathcal{T}))\, |\xi| \sum_{n=0}^{\mathcal{N}_\delta t} \delta t \sum_{\sigma\in\mathcal{E}} \tau_{K|L} \left(D_\sigma \psi(u)\right)^2,$$

et en utilisant l'estimation (6.35), on obtient (6.50).

Preuve de l'item 2

De façon similaire, en utilisant l'estimation (6.39), on obtient (6.51).

Preuve de l'item 3

Comme $v_{\mathcal{D},\lambda} = z_{\mathcal{D},\lambda} + \psi(v_{\mathcal{D},\lambda})$, on a

$$\int_0^T \int_{\mathbb{R}^d} \left(v_{\mathcal{D},\lambda}(x+\xi,t) - v_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t = \int_0^T \int_{\mathbb{R}^d} \left(\begin{array}{l} z_{\mathcal{D},\lambda}(x+\xi,t) - z_{\mathcal{D},\lambda}(x,t) \\ +\psi(u_{\mathcal{D},\lambda}(x+\xi,t)) - \psi(u_{\mathcal{D},\lambda}(x,t)) \end{array} \right)^2 \mathrm{d}x\mathrm{d}t,$$

et comme $(a+b)^2 \leq 2a^2 + 2b^2$, en utilisant les estimations (6.50) et (6.51), on obtient l'estimation (6.52).

Preuve de l'item 4

Pour montrer ce dernier item, on utilise une preuve reprise des travaux de Bouillard (2006). Soit $\xi \in \mathbb{R}^n$, and $\Omega_\xi = \{x \in \Omega, [x, x+\xi] \subset \Omega\}$. Soit $(x,t) \in \Omega_\xi \times (0,T)$ donnés. On note $K \in \mathcal{T}$ et $L \in \mathcal{T}$ les volumes de contrôle tels que $x \in K$ et $x+\xi \in L$ (ces volumes de contrôle existent pour presque tout $x \in \Omega_\xi$). On a alors $w_{\mathcal{D},\lambda}(x,t) - w_{\mathcal{D},\lambda}(x+\xi,t) = w_K^n - w_L^n$ avec n choisi tel que $t \in [n\delta t, (n+1)\delta t[$, et (6.19) donne

$$w_L^{n+1} - w_K^{n+1} = \max\left(w_L^n - \delta t F\left(w_L^n, v_L^{n+1}\right), 0\right) - \max\left(w_K^n - \delta t F\left(w_K^n, v_K^{n+1}\right), 0\right)$$

et donc

$$\left|w_L^{n+1} - w_K^{n+1}\right| \leq |w_L^n - w_K^n| + \delta t \left|F\left(w_L^n, v_L^{n+1}\right) - F\left(w_K^n, v_K^{n+1}\right)\right|$$

qu'on réécrit

$$\left|w_L^{n+1} - w_K^{n+1}\right| \leq |w_L^n - w_K^n| + \delta t \left|F\left(w_L^n, v_L^{n+1}\right) - F\left(w_L^n, v_K^{n+1}\right)\right| + \delta t \left|F\left(w_L^n, v_K^{n+1}\right) - F\left(w_K^n, v_K^{n+1}\right)\right|.$$

Comme F lipschitzienne par rapport à ses deux arguments,

$$\left|w_L^{n+1} - w_K^{n+1}\right| \leq (1 + \delta t L) |w_L^n - w_K^n| + \delta t L \left|v_L^{n+1} - v_L^{n+1}\right|$$

Une récurrence sur n et l'inégalité $(1+x) \leq e^x$, vérifiée $\forall x \in \mathbb{R}$, donne

$$\left|w_L^{n+1} - w_K^{n+1}\right| \leq e^{(n+1)\delta t L} \left|w_L^0 - w_K^0\right| + \delta t \sum_{p=0}^{n} e^{(n-p)\delta t L} \left|v_L^{p+1} - v_K^{p+1}\right|$$

Comme $n \in [\![0, \mathcal{N}_{\delta t}]\!]$, on a $(n+1)\delta t \leq 2T$, et donc

$$\left|w_L^{n+1} - w_K^{n+1}\right| \leq e^{2TL} \left|w_L^0 - w_K^0\right| + \delta t \sum_{p=0}^{n} e^{2TL} \left|v_L^{p+1} - v_K^{p+1}\right|$$

en élevant au carré et grâce à la relation $(a+b)^2 \leq 2a^2 + 2b^2$

$$\left|w_L^{n+1} - w_K^{n+1}\right|^2 \leq 2e^{4TL} \left|w_L^0 - w_K^0\right|^2 + 2e^{4TL}\delta t^2 \left(\sum_{p=0}^{n} \left|v_L^{p+1} - v_K^{p+1}\right|\right)^2$$

puis, grâce à l'inégalité de Cauchy-Schwarz

$$\left|w_L^{n+1} - w_K^{n+1}\right|^2 \leq 2e^{4TL} \left|w_L^0 - w_K^0\right|^2 + 2e^{4TL}\delta t(n+1) \sum_{p=0}^{n} \delta t \left|v_L^{p+1} - v_K^{p+1}\right|^2$$

Si on intègre l'inégalité précédente sur Ω_ξ, il vient

$$\begin{aligned} \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}^{n+1}(x+\xi) - w_{\mathcal{D},\lambda}^{n+1}(x)\right)^2 \mathrm{d}x \leq \quad & 2e^{4TL} \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}^{0}(x+\xi) - w_{\mathcal{D},\lambda}^{0}(x)\right)^2 \mathrm{d}x \\ & + 4Te^{4TL} \sum_{p=0}^{n} \delta t \int_{\Omega_\xi} \left(v_{\mathcal{D},\lambda}^{p+1}(x+\xi) - v_{\mathcal{D},\lambda}^{p+1}(x)\right)^2 \mathrm{d}x \end{aligned}$$

On somme sur n

$$\begin{aligned} \sum_{n=0}^{\mathcal{N}_{\delta t}} \delta t \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}^{n+1}(x+\xi) - w_{\mathcal{D},\lambda}^{n+1}(x)\right)^2 \mathrm{d}x \leq \quad & 2e^{4TL} \sum_{n=0}^{\mathcal{N}_{\delta t}} \delta t \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}^{0}(x+\xi) - w_{\mathcal{D},\lambda}^{0}(x)\right)^2 \mathrm{d}x \\ & + 4Te^{4TL} \sum_{n=0}^{\mathcal{N}_{\delta t}} \delta t \sum_{p=0}^{n} \delta t \int_{\Omega_\xi} \left(v_{\mathcal{D},\lambda}^{p+1}(x+\xi) - v_{\mathcal{D},\lambda}^{p+1}(x)\right)^2 \mathrm{d}x \end{aligned}$$

Et alors

$$\begin{aligned} \int_0^T \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}(x+\xi,t) - w_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t \leq \quad & 4Te^{4TL} \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}^{0}(x+\xi) - w_{\mathcal{D},\lambda}^{0}(x)\right)^2 \mathrm{d}x \\ & + 4Te^{4TL} \int_0^T \int_{\Omega_\xi} \left(v_{\mathcal{D},\lambda}(x+\xi,t) - v_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t \end{aligned}$$

D'où l'estimation (6.52)

$$\begin{aligned} \int_0^T \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}(x+\xi,t) - w_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t \leq \quad & 4Te^{4TL} \int_{\Omega_\xi} \left(w_{\mathcal{D},\lambda}(x+\xi,0) - w_{\mathcal{D},\lambda}(x,0)\right)^2 \mathrm{d}x \\ & + 4Te^{4TL} \left(C_1 + C_4\right) \left(|\xi| + C_5 \ \mathrm{size}(\mathcal{T})\right) |\xi| \end{aligned}$$

Comme $w_{\mathcal{D},\lambda}^0$ converge fortement vers w^0, on peut appliquer la réciproque du théorème de Kolmogorov, pour obtenir que les translations en espace de $w_{\mathcal{D},\lambda}^0$ convergent uniformément vers 0. Et par conséquent l'inégalité ci-dessus indique que les translatés en espace de $w_{\mathcal{D}}$ convergent uniformément vers 0. On obtient ainsi la convergence forte de w_D dans $L^2(\Omega \times [0,T])$ vers w.

□

Lemme 6.2.5 (Lemme technique)
Soit $T > 0$, $\tau \in (0,T)$, $k \in (0,T)$ et $(a^n)_{n\in\mathbb{N}}$ une famille de valeurs réelles non négatives. Alors

$$\int_0^{T-\tau} \sum_{n=[t/k]+1}^{[(t+\tau)/k]} a^{n+1} dt \leq \tau \sum_{n=0}^{[T/k]} a^{n+1}$$

et $\forall \zeta \in [0,\tau]$

$$\int_0^{T-\tau} \sum_{n=[t/k]+1}^{[(t+\tau)/k]} a^{[(t+\zeta)/k]+1} dt \leq \tau \sum_{n=0}^{[T/k]} a^{n+1} \tag{6.54}$$

On peut trouver une démonstration de ce lemme dans Eymard et al. (2001).

Lemme 6.2.6 (translation en temps pour $u_{\mathcal{D},\lambda}$, $z_{\mathcal{D},\lambda}$ et $v_{\mathcal{D},\lambda}$)
Sous les hypothèses de la proposition 6.2.1, si $\delta t \leq 1$ et $\lambda \geq 2L$,

1. *Il existe C_6 ne dépendant que de $||z^0||_\infty$, $||u^0||_\infty$, T, $\widetilde{X}$ et $\overline{X}$, tel que,*

$$\int_0^{T-\tau} \int_\Omega \left(\psi(u_{\mathcal{D},\lambda}(x,t+\tau)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 dxdt \leq \tau C_6, \tag{6.55}$$

2. *Il existe C_7 ne dépendant que de $||z^0||_\infty$, $||u^0||_\infty$, T, $\widetilde{X}$ et $\overline{X}$, tel que,*

$$\int_0^{T-\tau} \int_\Omega \left(z_{\mathcal{D},\lambda}(x,t+\tau) - z_{\mathcal{D},\lambda}(x,t)\right)^2 dxdt \leq \tau C_7, \tag{6.56}$$

3.

$$\int_0^{T-\tau} \int_\Omega \left(v_{\mathcal{D},\lambda}(x,t+\tau) - v_{\mathcal{D},\lambda}(x,t)\right)^2 dxdt \leq 2\tau\left(C_6 + C_7\right), \tag{6.57}$$

4. *Il existe C_8 ne dépendant que de $||z^0||_\infty$, $||u^0||_\infty$, T, $\widetilde{X}$ et $\overline{X}$, tel que,*

$$\int_0^{T-\tau} \int_\Omega \left(w_{\mathcal{D},\lambda}(x,t+\tau) - w_{\mathcal{D},\lambda}(x,t)\right)^2 dxdt \leq \tau C_8, \tag{6.58}$$

Preuve.
preuve de l'item 1 Comme la fonction ψ est croissante et lipschitzienne de constante L, on a

$$\int_0^{T-\tau} \int_\Omega \left(\psi(u_{\mathcal{D},\lambda}(x,t+\tau)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 \mathrm{d}x\mathrm{d}t \leq \int_0^{T-\tau} A(t)\mathrm{d}t$$

avec

$$A(t) = L\int_\Omega \left(\psi(u_{\mathcal{D},\lambda}(x,t+\tau)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)\left(u_{\mathcal{D},\lambda}(x,t+\tau) - u_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x$$

On a

$$A(t) = L\sum_K m_K \left(\psi(u_K^{a_\tau+1}) - \psi(u_K^{a_0+1})\right)\left(u_K^{a_\tau+1} - u_K^{a_0+1}\right)$$

qu'on peut réécrire

$$A(t) = L\sum_K m_K \left(\psi(u_K^{a_\tau+1}) - \psi(u_K^{a_0+1})\right) \sum_{n=a_0+1}^{a_\tau} \left(u_K^{n+1} - u_K^n\right)$$

En utilisant le schéma (6.18), on obtient

$$A(t) = L\sum_K m_K \left(\psi(u_K^{a_\tau+1}) - \psi(u_K^{a_0+1})\right) \sum_{n=a_0+1}^{a_\tau} \delta t \left(\begin{array}{l} \displaystyle\sum_{L\in\mathcal{N}(K)} \frac{\tau_{K,L}}{m_K}\left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right) \\ \displaystyle -\sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\tau_{K,\sigma}}{m_K}\psi(u_K^{n+1}) + \delta t f(w_K^n)\left(\widetilde{X} - v_K^n\right) \end{array} \right)$$

qu'on réécrit

$$A(t) = L\left(A_0(t,\tau) - A_0(t,0) + A_1(t,\tau) - A_1(t,0)\right)$$

avec, pour $\xi = 0$ ou $\xi = \tau$,

$$A_0(t,\xi) = \sum_{n=a_0+1}^{a_\tau}\sum_K \left(\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L}\psi(u_K^{a_\xi+1})\left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right) - \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{a_\xi+1})\psi(u_K^{n+1}) \right)$$

et

$$A_1(t,\xi) = \sum_{n=a_0+1}^{a_\tau}\sum_K m_K\delta t\psi(u_K^{a_\xi+1})f(w_K^n)\left(\widetilde{X} - v_K^n\right)$$

- Traitement du terme A_0
  En faisant la somme sur K, on obtient

$$A_0(t,\xi) = \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K \left(\begin{array}{l} \displaystyle\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L}\left(\psi(u_K^{a_\xi+1}) - \psi(u_L^{a_\xi+1})\right)\left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right) \\ \displaystyle -\sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{a_\xi+1})\psi(u_K^{n+1}) \end{array} \right)$$

  L'inégalité de Young donne

$$\begin{aligned} |A_0(t,\xi)| \leq\ & \frac{1}{4}\sum_{n=a_0+1}^{a_\tau}\sum_K\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L}\left(\left(\psi(u_K^{a_\xi+1}) - \psi(u_L^{a_\xi+1})\right)^2 + \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right)^2\right) \\ & + \frac{1}{4}\sum_{n=a_0+1}^{a_\tau}\sum_K\sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{a_\xi+1})^2 + \frac{1}{4}\sum_{n=a_0+1}^{a_\tau}\sum_K\sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{n+1})^2 \end{aligned}$$

  On intègre sur $(0, T-\tau)$ et on applique le lemme 6.2.5,

$$|A_0(t,\xi)| \leq \frac{\tau}{2}\sum_{n=0}^{[T/k]}\sum_K \left(\sum_{L\in\mathcal{N}(K)} \delta t\tau_{K,L}\left(\psi(u_K^{n+1}) - \psi(u_L^{n+1})\right)^2 + \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \delta t\tau_{K,\sigma}\psi(u_K^{n+1})^2 \right)$$

  En posant $k = \delta t$, l'estimation (6.35) permet d'obtenir

$$|A_0(t,\xi)| \leq \tau\frac{C_1}{2}.$$

- Traitement du terme A_1
  On applique l'inégalité de Young

$$|A_1(t,\xi)| \leq \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K\delta t\psi(u_K^{a_\xi+1})^2 + \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K\delta t f(w_K^n)^2\left(\widetilde{X} - v_K^n\right)^2$$

  En utilisant les bornes L^∞,

$$|A_1(t,\xi)| \leq \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K\delta t\psi(u_K^{a_\xi+1})^2 + \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K\delta t\left(\widetilde{X} + V\right)^2$$

On intègre sur $(0, T-\tau)$ et on applique le lemme 6.2.5 pour obtenir

$$|A_1(t,\xi)| \leq \frac{\tau}{2} \sum_{n=0}^{[T/k]} \sum_K m_K \delta t \psi(u_K^{n+1})^2 + \frac{\tau}{2} m(\Omega) \left(\widetilde{X} + V\right)^2 \sum_{n=0}^{[T/k]} \delta t$$

On choisit $k = \delta t$ et on utilise le fait que ψ est borné pour obtenir

$$|A_1(t,\xi)| \leq \tau T m(\Omega) \left(\overline{X}^2 + \left(\widetilde{X} + V\right)^2\right)$$

On regroupe les estimations précédentes et on pose $C_6 = LC_1 + 2LTm(\Omega)\left(\overline{X}^2 + \left(\widetilde{X} + V\right)^2\right)$ pour obtenir l'estimation (6.55).

preuve de l'item 2

On a

$$\int_0^{T-\tau} \int_\Omega \left(z_{\mathcal{D},\lambda}(x, t+\tau) - z_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t = \int_0^{T-\tau} A(t)\mathrm{d}t$$

avec $A(t) = \sum_K m_K \left(z_K^{a_\tau+1} - z_K^{a_0+1}\right)^2$, avec $a_\xi = [(t+\xi)/k]$. On a

$$A(t) = \sum_K m_K \left(z_K^{a_\tau+1} - z_K^{a_0+1}\right) \sum_{n=a_0+1}^{a_\tau} \left(z_K^{n+1} - z_K^n\right)$$

En utilisant le schéma (6.17), on obtient

$$A(t) = \sum_K m_K \left(z_K^{a_\tau+1} - z_K^{a_0+1}\right) \sum_{n=a_0+1}^{a_\tau} \left(\begin{array}{l} \frac{\delta t}{m_K} \left(\sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(z_L^{n+1} - z_K^{n+1}\right) - \sum_{\sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K} \tau_{K,\sigma} z_K^{n+1} \right) \\ -\delta t \lambda z_K^{n+1} + \delta t f(w_K^n) \left(z_K^n - z_K^{n+1} + \psi(u_K^n) - \psi(u_K^{n+1})\right) \end{array} \right)$$

qu'on réécrit de la façon suivante

$$A(t) = (A_0(t,\tau) - A_0(t,0)) + (A_1(t,\tau) - A_1(t,0)) + (A_2(t,\tau) - A_2(t,0))$$

avec

$$A_0(t,\xi) = \sum_K m_K z_K^{a_\xi+1} \sum_{n=a_0+1}^{a_\tau} \frac{\delta t}{m_K} \left(\sum_{L \in \mathcal{N}(K)} \tau_{K,L} \left(z_L^{n+1} - z_K^{n+1}\right) - \sum_{\sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K} \tau_{K,\sigma} z_K^{n+1} \right),$$

$$A_1(t,\xi) = -\sum_K m_K z_K^{a_\xi+1} \sum_{n=a_0+1}^{a_\tau} \delta t \lambda z_K^{n+1},$$

et

$$A_2(t,\xi) = -\sum_K m_K z_K^{a_\xi+1} \sum_{n=a_0+1}^{a_\tau} \delta t f(w_K^n) \left(z_K^n - z_K^{n+1} + \psi(u_K^n) - \psi(u_K^{n+1})\right)$$

avec $\xi = 0$ ou $\xi = \tau$.

- Traitement du terme A_0

  En sommant sur K, on obtient

$$A_0(t,\xi) = \sum_{n=a_0+1}^{a_\tau} \sum_K \left(\sum_{L \in \mathcal{N}(K)} \delta t \tau_{K,L} \left(z_K^{a_\xi+1} - z_L^{a_\xi+1}\right) \left(z_L^{n+1} - z_K^{n+1}\right) - \sum_{\sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K} \tau_{K,\sigma} z_K^{a_\xi+1} z_K^{n+1} \right),$$

En appliquant l'inégalité de Young

$$|A_0(t,\xi)| \leq \sum_{n=a_0+1}^{a_\tau} \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{a_\xi+1}-z_K^{a_\xi+1}\right)^2 + \sum_{n=a_0+1}^{a_\tau} \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{n+1}-z_K^{n+1}\right)^2$$
$$+ \sum_{n=a_0+1}^{a_\tau} \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{a_\xi+1})^2 + \sum_{n=a_0+1}^{a_\tau} \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{n+1})^2$$

On intègre $A_0(t,\xi)$ sur $(0,T-\tau)$,

$$\int_0^{T-\tau} |A_0(t,\xi)|\,\mathrm{d}t \leq \int_0^{T-\tau} \sum_{n=a_0+1}^{a_\tau} \left(\begin{array}{l} \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{a_\xi+1}-z_K^{a_\xi+1}\right)^2 + \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{a_\xi+1})^2 \\ + \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{n+1}-z_K^{n+1}\right)^2 + \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{n+1})^2, \end{array} \right) \mathrm{d}t$$

L'utilisation du lemme 6.2.5 donne,

$$\int_0^{T-\tau} |A_0(t,\xi)|\,\mathrm{d}t \leq \tau \sum_{n=0}^{[T/k]+1} \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{n+1}-z_K^{n+1}\right)^2 + \tau \sum_{n=0}^{[T/k]+1} \sum_K \sum_{L\in\mathcal{N}(K)} \frac{\delta t}{2}\tau_{K,L}\left(z_L^{n+1}-z_K^{n+1}\right)^2$$
$$+\tau \sum_{n=0}^{[T/k]+1} \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{n+1})^2 + \tau \sum_{n=0}^{[T/k]+1} \sum_K \sum_{\sigma\in\mathcal{E}_{ext}\cap\mathcal{E}_K} \frac{\delta t}{2}\tau_{K,\sigma}(z_K^{n+1})^2,$$

et finalement, en prenant $k=\delta t$, grâce à l'estimation (6.39), on a

$$\int_0^{T-\tau} |A_0(t,\xi)|\,\mathrm{d}t \leq \tau\left(C_4 + ||z_0||_\infty\right).$$

- Traitement du terme A_1
  L'inégalité de Young donne

$$|A_1(t,\xi)| \leq \sum_K \sum_{n=a_0+1}^{a_\tau} \frac{\delta t}{2}\lambda m_K\left(z_K^{a_\xi+1}\right)^2 + \sum_K \sum_{n=a_0+1}^{a_\tau} m_K\frac{\delta t}{2}\lambda\left(z_K^{n+1}\right)^2$$

On intègre $A_0(t,\xi)$ sur $(0,T-\tau)$,

$$\int_0^{T-\tau} |A_1(t,\xi)|\,\mathrm{d}t \leq \int_0^{T-\tau} \sum_K \sum_{n=a_0+1}^{a_\tau} \frac{\delta t}{2}\lambda m_K\left(z_K^{a_\xi+1}\right)^2 \mathrm{d}t + \int_0^{T-\tau} \sum_K \sum_{n=a_0+1}^{a_\tau} m_K\frac{\delta t}{2}\lambda\left(z_K^{n+1}\right)^2 \mathrm{d}t$$

L'utilisation du lemme 6.2.5 donne,

$$\int_0^{T-\tau} |A_1(t,\xi)|\,\mathrm{d}t \leq \tau \sum_K \sum_{n=0}^{[T/k]+1} \frac{\delta t}{2}\lambda m_K\left(z_K^{n+1}\right)^2 + \tau \sum_K \sum_{n=0}^{[T/k]+1} m_K\frac{\delta t}{2}\lambda\left(z_K^{n+1}\right)^2$$

et finalement, en prenant $k=\delta t$, grâce à l'estimation (6.38), on a

$$\int_0^{T-\tau} |A_1(t,\xi)|\,\mathrm{d}t \leq \tau\left(C_3 + ||z_0||_\infty\right)$$

- Traitement du terme A_2
  On utilise l'inégalité de Young avec $\alpha=\lambda$,

$$|A_2(t,\xi)| \leq \lambda \sum_{n=a_0+1}^{a_\tau} \sum_K \frac{\delta t}{2} m_K (z_K^{a_\xi+1})^2 + \sum_{n=a_0+1}^{a_\tau} \sum_K \frac{\delta t}{2\lambda} m_K\left(z_K^n - z_K^{n+1} + \psi(u_K^n) - \psi(u_K^{n+1})\right)^2$$

on intègre sur $(0, T-\tau)$, puis on applique le lemme 6.2.5,

$$|A_2(t,\xi)| \leq \tau\lambda \sum_{n=0}^{[T/k]} \sum_K \frac{\delta t}{2} m_K (z_K^{n+1})^2 + \frac{\tau}{\lambda} 2Tm(\Omega) \left(Z + \overline{X}\right)^2$$

On utilise l'estimation (6.38), et comme $\lambda > 1$,

$$|A_2(t,\xi)| \leq \frac{\tau}{2} \left(C_3 + ||z_0||_\infty\right) + 2\tau Tm(\Omega) \left(Z + \overline{X}\right)^2$$

Et finalement, en regroupant les estimations précédentes, on obtient l'estimation (6.56) avec $C_7 = 3C_3 + 2C_4 + 5\,||z_0||_\infty + 4Tm(\Omega)\left(Z + \overline{X}\right)^2$.

preuve de l'item 3

Comme $v_{\mathcal{D},\lambda} = z_{\mathcal{D},\lambda} + \psi(v_{\mathcal{D},\lambda})$, on a

$$\int_0^T \int_{\mathbb{R}^d} \left(v_{\mathcal{D},\lambda}(x,t+\tau) - v_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t = \int_0^T \int_{\mathbb{R}^d} \left(z_{\mathcal{D},\lambda}(x,t+\tau) - z_{\mathcal{D},\lambda}(x,t) + \psi(u_{\mathcal{D},\lambda}(x,t+\tau)) - \psi(u_{\mathcal{D},\lambda}(x,t))\right)^2 \mathrm{d}x\mathrm{d}t,$$

et comme $(a+b)^2 \leq 2a^2 + 2b^2$, en utilisant les estimations (6.55) et (6.56), on obtient l'estimation (6.57).

preuve de l'item 4

On a

$$\int_0^{T-\tau} \int_\Omega \left(w_{\mathcal{D},\lambda}(x,t+\tau) - w_{\mathcal{D},\lambda}(x,t)\right)^2 \mathrm{d}x\mathrm{d}t = \int_0^{T-\tau} A(t)\mathrm{d}t$$

avec $A(t) = \sum_K m_K \left(w_K^{a_\tau+1} - w_K^{a_0+1}\right)^2$. On a

$$A(t) = \sum_K m_K \left(w_K^{a_\tau+1} - w_K^{a_0+1}\right) \sum_{n=a_0+1}^{a_\tau} \left(w_K^{n+1} - w_K^n\right)$$

On utilise le schéma (6.19).

- Cas $\left(w_K^n - \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)\right) > 0$

  On a alors

  $$w_K^{n+1} - w_K^n = -\delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right),$$

  et donc

  $$A(t) = A_0(t,\tau) - A_0(t,0)$$

  avec

  $$A_0(t,\xi) = -\sum_{n=a_0+1}^{a_\tau} \sum_K m_K w_K^{a_\xi+1} \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right),$$

- Cas $\left(w_K^n - \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)\right) \leq 0$

  On a alors $w_K^{n+1} = 0$ et donc

  $$w_K^{n+1} - w_K^n = -w_K^n$$

  ce qui donne

  $$A(t) = A_0(t,\tau) - A_0(t,0)$$

  avec

  $$|A_0(t,\xi)| = \sum_{n=a_0+1}^{a_\tau} \sum_K m_K w_K^{a_\xi+1}(w_K^n)$$

  Dans ce cas, on a $0 \leq w_K^n \leq \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)$ et donc

  $$|A_0(t,\xi)| \leq \sum_{n=a_0+1}^{a_\tau} \sum_K m_K w_K^{a_\xi+1} \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)$$

Les deux cas aboutissent sur une inégalité identique. En appliquant alors l'inégalité de Young, on a

$$|A_0(t,\xi)| \leq \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K \delta t \left(w_K^{a_\xi+1}\right)^2 + \frac{1}{2}\sum_{n=a_0+1}^{a_\tau}\sum_K m_K \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)^2$$

On applique le lemme 6.2.5

$$|A_0(t,\xi)| \leq \frac{\tau}{2}\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_K m_K \delta t \left(w_K^{n+1}\right)^2 + \frac{\tau}{2}\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_K m_K \delta t F\left(w_K^n, \psi(u_K^{n+1}) - z_K^{n+1}\right)^2$$

et donc les estimations (6.22) et (6.21), donnent

$$|A_0(t,\xi)| \leq \tau T m(\Omega)\left(W^2 + \overline{F}^2\right)$$

En posant $C_8 = Tm(\Omega)\left(W^2 + \overline{F}^2\right)$, on obtient l'estimation (6.58). □

6.3 Étude des convergences

Proposition 6.3.1
Sous les hypothèses 6.1.7, 6.1.4, 6.1.1, et 6.1.2, soient $u_{\mathcal{D},\lambda}$, $v_{\mathcal{D},\lambda}$, $z_{\mathcal{D},\lambda}$, $w_{\mathcal{D},\lambda}$, soient $(u_{\mathcal{D}_m,\lambda}, v_{\mathcal{D}_m,\lambda}, w_{\mathcal{D}_m,\lambda})$ données par le schéma $P_{(\mathcal{D},\lambda)}$, ((6.17) à (6.20)). Alors, si $\delta t \leq 1$, lorsque $\lambda \to \infty$, $(u_{\mathcal{D}_m,\lambda}, w_{\mathcal{D}_m,\lambda})$ converge vers $(u_{\mathcal{D}_m}, w_{\mathcal{D}_m})$ solution du schéma discret $P_{\mathcal{D}}$ ((6.6) à (6.10)) et $z_{\mathcal{D}_m,\lambda}$ converge vers 0.

Preuve. Les fonctions $(u_{\mathcal{D}_m,\lambda}, z_{\mathcal{D}_m,\lambda}, w_{\mathcal{D}_m,\lambda})$ vérifient les estimations de la proposition 6.2.3. En particulier $z_{\mathcal{D}_m,\lambda}$ vérifie l'estimation (6.38). Comme $\lambda \to \infty$, pour δt et h fixé, l'estimation peut se réécrire de la façon suivante

$$\lim_{\lambda\to\infty}\left(\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K \delta t \left(z_K^{(n+1)}\right)^2\right) = 0.$$

Ceci signifie que, $\forall K \in \mathcal{T}$ et $\forall n \in [\![0, \mathcal{N}_{\delta t}]\!]$

$$\lim_{\lambda\to\infty} z_K^{n+1} = 0. \tag{6.59}$$

De plus, l'égalité (6.59) implique que $\lim_{\lambda\to\infty} v_K^n = \psi(u_K^n)$, car par définition $z_K^n = \psi(u_K^n) - v_K^n$. L'équation (6.18) peut alors s'écrire

$$m_K \frac{u_K^{n+1} - u_K^n}{\delta t} - \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right) + \sum_{\sigma\in\mathcal{E}_{K,\text{ext}}} \tau_{K,\sigma}\psi(u_K^{n+1}) = m_K F\left(w_K^n, \psi(u_K^n)\right) \tag{6.60}$$

et l'équation (6.19)

$$w_K^{n+1} = \max\left(w_K^n - \delta t F\left(w_K^n, \psi(u_K^{n+1})\right), 0\right) \tag{6.61}$$

Le schéma $P_{(\mathcal{D},\lambda)}$ devient donc lorsque $\lambda \to \infty$ le schéma $P_{\mathcal{D}}$. □

Les estimations sur le schéma $P_{\mathcal{D}}$ sont déduites directement de celles sur le schéma $P_{(\mathcal{D},\lambda)}$, par passage à la limite en λ.

Proposition 6.3.2 (Estimation L^∞)
Sous les hypothèses 6.1.7, 6.1.4, 6.1.1, et 6.1.2, soient $u_{\mathcal{D}}$ et $w_{\mathcal{D}}$, définies par le schéma (6.6) à (6.10), alors, si $\delta t \leq 1$,

1. *il existe $W \in \mathbb{R}$, défini par la proposition 6.2.1, ne dépendant que de $||w^0||_\infty$, T, L, $\overline{F}$ et $\overline{F}$ tel que*

$$||w_{\mathcal{T},\delta t}||_{L^\infty(\Omega\times[0,T])} \leq W. \tag{6.62}$$

2. *il existe* $U \in \mathbb{R}$, *défini par la proposition 6.2.1, ne dépendant que de* $||w^0||_\infty$, $||u^0||_\infty$, T *et* $\overline{X}$, *tel que*

$$||u_{\mathcal{T},\delta t}||_{L^\infty(\Omega\times[0,T])} \leq U, \tag{6.63}$$

Proposition 6.3.3 (Estimations L^2)
Sous les hypothèses de la proposition 6.3.2, si $\delta t \leq 1$,

1. *il existe* $C_1 \in \mathbb{R}$, *défini par la proposition 6.2.3, ne dépendant que de* $||z^0||_\infty$, $||u^0||_\infty$, Ω, T, $\widetilde{X}$ *et* $\overline{X}$, *tel que*

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \left(\sum_{L\in\mathcal{N}(K)} \delta t \tau_{K,L} \left(\psi(u_L^{n+1}) - \psi(u_K^{n+1})\right)^2 + \sum_{\sigma\in\mathcal{E}_K} \delta t \tau_{K,\sigma} \psi(u_K^{n+1})^2 \right) \leq C_1. \tag{6.64}$$

2. *et tel que*

$$\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K (u_K^{n+1} - u_K^n)^2 \leq C_1, \tag{6.65}$$

3. *il existe* $C_2 \in \mathbb{R}$, *défini par la proposition 6.2.3, ne dépendant que de* $||z^0||_\infty$, $||u^0||_\infty$, Ω, T, $\widetilde{X}$ *et* $\overline{X}$, *tel que*

$$\sum_{K\in\mathcal{T}} m_K (u_K^{n+1})^2 \leq C_2, \tag{6.66}$$

Lemme 6.3.4 (translation en espace pour $u_\mathcal{D}$ et $v_\mathcal{D}$)
Soit $\xi \in \mathbb{R}^N$. *Sous les hypothèses de la proposition 6.3.2, si* $\delta t \leq 1$,

1.

$$\int_0^T \int_{\mathbb{R}^d} \left(\psi(u_\mathcal{D}(x+\xi,t)) - \psi(u_\mathcal{D}(x,t))\right)^2 dxdt \leq \frac{1}{2} C_1 \left(|\xi| + C_5\ size(\mathcal{T})\right) |\xi|, \tag{6.67}$$

2. *Les translatés en espace de* $w_\mathcal{D}$ *convergent uniformément vers* 0, *lorsque* $\xi \to 0$

Lemme 6.3.5 (translation en temps pour $u_\mathcal{D}$ et $v_\mathcal{D}$)
Sous les hypothèses de la proposition 6.3.2,

1. *Il existe* C_6, *défini par la proposition 6.2.6, ne dépendant que de* $||z^0||_\infty$, $||u^0||_\infty$, T, $\widetilde{X}$ *et* $\overline{X}$, *tel que,*

$$\int_0^{T-\tau} \int_\Omega \left(\psi(u_\mathcal{D}(x,t+\tau)) - \psi(u_\mathcal{D}(x,t))\right)^2 dxdt \leq \tau C_6, \tag{6.68}$$

2. *Il existe* C_8, *défini par la proposition 6.2.6, ne dépendant que de* $||z^0||_\infty$, $||u^0||_\infty$, T, $\widetilde{X}$ *et* $\overline{X}$, *tel que,*

$$\int_0^{T-\tau} \int_\Omega \left(w_\mathcal{D}(x,t+\tau) - w_\mathcal{D}(x,t)\right)^2 dxdt \leq \tau C_8, \tag{6.69}$$

Proposition 6.3.6 (convergence)
Soit $m \in \mathbb{N}$, *soit* $\mathcal{T}_m$ *un maillage admissible et* δt_m *une discrétisation en temps. On suppose que* $size(\mathcal{T}_m) \to 0$ *quand* $m \to \infty$. *Soient* $u_{\mathcal{D}_m,\lambda} \in L^\infty(\Omega\times(0,T))$, $z_{\mathcal{D}_m,\lambda} \in L^2(\Omega\times(0,T))$, $w_{\mathcal{D}_m,\lambda} \in L^2(\Omega\times(0,T))$ *définies par le schéma ((6.18), (6.17), (6.19), (6.15), (6.16)) et soit* $v_{\mathcal{D}_m,\lambda} \in L^2(\Omega\times(0,T))$ *définie par* $v_{\mathcal{D}_m,\lambda} = \psi(u_{\mathcal{D}_m,\lambda}) - z_{\mathcal{D}_m,\lambda}$. *Alors, sous les hypothèses (6.1.7), à* λ *fixé, tel que* $\lambda \geq 2L$, *quand* $m \to \infty$,

1. $u_{\mathcal{D}_m,\lambda}$ *converge vers* u_λ *pour la topologie faible-*$\star$ *de* $L^\infty(\Omega\times(0,T))$, *et* $\psi(u_\lambda) \in L^2(0,T;H^1(\Omega))$,
2. $v_{\mathcal{D}_m,\lambda}$ *converge vers* v_λ *dans* $L^2(\Omega\times(0,T))$ *et* $v_\lambda = \psi(u_\lambda) - z_\lambda$,
3. $z_{\mathcal{D}_m,\lambda}$ *converge vers* z_λ *dans* $L^2(\Omega\times(0,T))$,

4. $w_{\mathcal{D}_m,\lambda}$ *converge vers* w_λ *dans* $L^2(\Omega\times(0,T))$.

De plus, le triplet $(u_\lambda, v_\lambda, w_\lambda)$ *est une solution faible du problème* (P_λ) *au sens de la définition 6.1.6. Les fonctions* u_λ, v_λ, w_λ *et* z_λ *vérifient les estimations (6.21) à (6.24), et (6.35) à (6.39).*

Preuve.

L'estimation (6.23) indique que $u_{\mathcal{D}_m,\lambda}$ est bornée dans $L^\infty(\Omega\times(0,T))$ et par conséquent il existe une sous suite, encore notée $u_{\mathcal{D}_m,\lambda}$, qui converge vers u_λ, pour la topologie faible-$\star$ de $L^\infty(\Omega\times(0,T))$. Les estimations sur les translations en espace du lemme 6.2.4 et en temps (lemme 6.2.6), sur $\psi(u_{\mathcal{D}_m,\lambda})$, $v_{\mathcal{D}_m,\lambda}$, $z_{\mathcal{D}_m,\lambda}$ et $w_{\mathcal{D}_m,\lambda}$ permettent d'obtenir la compacité nécessaire à l'application du théorème de Kolmogorov (cf. Eymard et al. (2000)). On obtient ainsi la convergence forte pour $\psi(u_{\mathcal{D}_m,\lambda})$, $v_{\mathcal{D}_m,\lambda}$, $z_{\mathcal{D}_m,\lambda}$ et $w_{\mathcal{D}_m,\lambda}$, vers, respectivement, $\overline{\psi}$, v_λ, z_λ et w_λ dans $L^2(\Omega\times(0,T))$ quand $m\to\infty$. La convergence faible de $u_{\mathcal{D}_m,\lambda}$ vers u_λ, et la convergence forte de $\psi(u_{\mathcal{D}_m,\lambda})$ vers $\overline{\psi}$, indiquent que $\overline{\psi}=\psi(u_\lambda)$. On obtient ainsi la convergence forte de $\psi(u_{\mathcal{D}_m,\lambda})$ vers $\psi(u_\lambda)$.

Il reste à prouver que le triplet $(u_\lambda, v_\lambda, w_\lambda)$ est bien la solution faible du problème (P_λ) au sens de la définition 6.1.6. Pour cela, pour toute fonction test $\varphi(x,t)\in C_c^\infty(\Omega\times(0,T))$, on multiplie le schéma par $\delta t\varphi(x_K,(n+1)\delta t)$, puis on le somme sur n puis sur K. Avec l'équation (6.18), on obtient $T_8+T_9=T_{10}$, avec

$$T_8=\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K(u_K^{n+1}-u_K^n)\varphi(x_K,(n+1)\delta t)$$

$$T_9=\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}}\delta t\varphi(x_K,(n+1)\delta t)\sum_{L\in\mathcal{N}(K)}\tau_{K,L}\left(\psi(u_K^{n+1})-\psi(u_L^{n+1})\right)$$

$$T_{10}=\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K\delta t\varphi(x_K,(n+1)\delta t)F\left(w_K^n,\psi(u_K^n)-z_K^n\right)$$

De façon classique (cf. Eymard et al. (2000)), on obtient que

$$\lim_{\delta t\to 0,n\to\infty} T_8=-\int_0^T\int_\Omega u_\lambda\varphi_t\mathrm{d}x\mathrm{d}t-\int_\Omega u_0(x)\varphi(x,0)\mathrm{d}x\mathrm{d}t, \tag{6.70}$$

et que

$$\lim_{\delta t\to 0,n\to\infty} T_9=\int_0^T\int_\Omega \psi(u_\lambda(x,t))\triangle\varphi(x,t)\mathrm{d}x\mathrm{d}t, \tag{6.71}$$

car $\psi(u_{\mathcal{D},\lambda})$ converge fortement vers $\psi(u_\lambda)$. Et finalement, comme $z_{\mathcal{D},\lambda}$ converge fortement vers z_λ et $w_{\mathcal{D},\lambda}$ vers w_λ, on a

$$\lim_{\delta t\to 0,n\to\infty} T_{10}=\int_0^T\int_\Omega \varphi(x,t)F\left(w_\lambda,\psi(u_\lambda)-z_\lambda\right). \tag{6.72}$$

De plus, comme $v_\lambda=\psi(u_\lambda)-z_\lambda$, on obtient

$$\lim_{\delta t\to 0,n\to\infty} T_{10}=\int_0^T\int_\Omega \varphi(x,t)F\left(w_\lambda,v_\lambda\right). \tag{6.73}$$

De façon similaire, avec l'équation (6.17), on a $T_{11}+T_{12}+T_{13}=T_{14}$, avec

$$\lim_{\delta t\to 0,n\to\infty} T_{11}=\lim_{\delta t\to 0,n\to\infty}\sum_{n=0}^{\mathcal{N}_{\delta t}}\sum_{K\in\mathcal{T}} m_K(z_K^{n+1}-z_K^n)\varphi(x_K,(n+1)\delta t)=-\int_0^T\int_\Omega z\varphi_t\mathrm{d}x\mathrm{d}t-\int_\Omega z_0(x)\varphi(x,0)\mathrm{d}x\mathrm{d}t,$$

$$\lim_{\delta t\to 0,n\to\infty} T_{12} = \lim_{\delta t\to 0,n\to\infty} \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \delta t\varphi(x_K,(n+1)\delta t) \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(z_K^{n+1}) - \psi(z_L^{n+1})\right) = \int_0^T \int_\Omega z(x,t)\triangle\varphi(x,t)\mathrm{d}x\mathrm{d}t$$

$$\lim_{\delta t\to 0,n\to\infty} T_{13} = \lim_{\delta t\to 0,n\to\infty} \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \delta t m_K \lambda z_K^{n+1}\varphi(x_K,(n+1)\delta t) = \lambda \int_0^T \int_\Omega z(x,t)\varphi(x,t)\mathrm{d}x\mathrm{d}t$$

Il est clair que $\lim_{\delta t\to 0,n\to\infty} T_{14} = 0$.

On obtient, de façon similaire, en utilisant l'équation (6.19), la troisième équation de la formulation faible du problème (P_λ). □

Corollaire 6.3.7 (Existence d'une solution faible particulière du problème (P_λ))
Sous les hypothèses 6.1.1, 6.1.2 et 6.1.4, pour un réel $\lambda \in [2L,+\infty[$, il existe une solution faible $(u_\lambda, v_\lambda, w_\lambda)$ du problème (P_λ) au sens de la définition (6.1.6). De plus, si on définit z_λ par $z_\lambda = \psi(u_\lambda) - v_\lambda$, les estimations suivantes sont vérifiées

1.
$$||w_\lambda||_{L^\infty(\Omega\times(0,T))} \leq W. \tag{6.74}$$

2.
$$||u_\lambda||_{L^\infty(\Omega\times(0,T))} \leq U, \tag{6.75}$$

3.
$$||z_\lambda||_{L^\infty(\Omega\times(0,T))} \leq Z, \tag{6.76}$$

4.
$$||v_\lambda||_{L^\infty(\Omega\times(0,T))} \leq V, \tag{6.77}$$

5.
$$||\nabla\psi(u_\lambda)||_{L^2(\Omega\times(0,T))} \leq C_1. \tag{6.78}$$

6.
$$||u_\lambda||_{L^2(\Omega\times(0,T))} \leq C_2, \tag{6.79}$$

7.
$$||z_\lambda||_{L^2(\Omega\times(0,T))} \leq \frac{C_3}{\lambda^2}, \tag{6.80}$$

8.
$$||\nabla z_\lambda||_{L^2(\Omega\times(0,T))} \leq \frac{C_4}{\lambda}, \tag{6.81}$$

Preuve. La proposition 6.3.6, qui donne la convergence d'un schéma vers une solution faible suffit à montrer l'existence d'au moins une solution vérifiant les estimations (6.21) à (6.24), et (6.35) à (6.39). En passant à la limite, $\delta t_m \to 0$ et size$(\mathcal{T}_m) \to 0$, dans ces estimations, on obtient les estimations (6.74) à (6.81). □

Proposition 6.3.8
Soit $(u_\lambda, v_\lambda, w_\lambda) \in L^2(\Omega\times(0,T))$ une solution faible de (P_λ) vérifiant les estimations (6.74) à (6.81). Alors, lorsque $\lambda\to\infty$, le triplet $(u_\lambda, v_\lambda, w_\lambda)$ converge vers le triplet $(u,v,w) \in L^2(\Omega\times(0,T))$ avec $v = \psi(u)$ et avec (u,w) solution faible de (P) au sens de la définition 6.1.5.

Preuve. Les estimations sur les translations en espace et en temps décrites par les lemmes 6.2.4 et 6.2.6, sont soit indépendantes de λ, soit fonction de $\frac{1}{\lambda}$. Dans le premier cas, elles sont toujours valables lorsque $\lambda\to\infty$, dans le second cas, elles sont améliorées. On a donc la compacité nécessaire pour appliquer le théorème de Kolmogorov. On obtient ainsi la convergence forte de u_λ, v_λ, et w_λ vers, respectivement u, v et w dans $L^2(\Omega\times(0,T))$. Les estimations obtenues sur z_λ ont permis de montrer que $||z_\lambda||_{L^2(\Omega\times(0,T))}$, $\lambda||z_\lambda||_{L^2(\Omega\times(0,T))}$ et $||\nabla z_\lambda||_{L^2(\Omega\times(0,T))}$ tendent uniformément vers 0. Il est donc possible de passer à la limites dans (6.5). On obtient alors (6.4). □

Proposition 6.3.9 (convergence)
Soit $m \in \mathbb{N}$, soit $\mathcal{T}_m$ un maillage admissible et δt_m une discrétisation en temps. On suppose que $size(\mathcal{T}_m) \to 0$ quand $m \to \infty$. Soient $u_{\mathcal{D}_m} \in L^\infty(\Omega \times (0,T))$ et $w_{\mathcal{D}_m} \in L^2(\Omega \times (0,T))$ définies par le schéma ((6.6) à (6.10)). Alors, sous les hypothèses 6.1.7, 6.1.1, et 6.1.2, quand $m \to \infty$,

1. *$u_{\mathcal{D}_m}$ converge faiblement vers u pour la topologie faible-$\star$ de $L^\infty(\Omega\times(0,T))$, et $\psi(u) \in L^2(0,T;H^1(\Omega))$,*
2. *$w_{\mathcal{D}_m}$ converge fortement vers w dans $L^2(\Omega \times (0,T))$.*

De plus, le couple (u,w) est une solution faible du problème (P) au sens de la définition 6.1.5. Les fonctions u et w vérifient les estimations (6.63) à (6.66).

Preuve. La preuve de convergence est classique. On utilise le lemme de convergence en temps et en espace 6.3.5 pour appliquer le théorème de Kolmogorov et obtenir ainsi la compacité nécessaire. On montre également de façon classique que les fonctions convergent vers la solution faible du problème (P). On peut trouver cette preuve dans Eymard et al. (2000). □

Proposition 6.3.10
Soit $m \in \mathbb{N}$, soit $\mathcal{T}_m$ un maillage admissible et δt_m une discrétisation en temps. On suppose que $size(\mathcal{T}_m) \to 0$ quand $m \to \infty$. Soient $u_{\mathcal{D}_m,\lambda} \in L^\infty(\Omega\times(0,T))$, $z_{\mathcal{D}_m,\lambda} \in L^2(\Omega\times(0,T))$, $w_{\mathcal{D}_m,\lambda} \in L^2(\Omega\times(0,T))$ définies par le schéma ((6.17) à (6.20)) et soit $v_{\mathcal{D}_m,\lambda} \in L^2(\Omega \times (0,T))$ définie par $v_{\mathcal{D}_m,\lambda} = \psi(u_{\mathcal{D}_m,\lambda}) - z_{\mathcal{D}_m,\lambda}$. Alors, Sous les hypothèses 6.1.7, 6.1.4, 6.1.1, et 6.1.2, lorsque $\lambda \to \infty$ et lorsque $m \to \infty$,

1. *$u_{\mathcal{D}_m,\lambda}$ converge faiblement vers u pour la topologie faible-$\star$ de $L^\infty(\Omega\times(0,T))$, et $\psi(u_\lambda) \in L^2(0,T;H^1(\Omega))$,*
2. *$v_{\mathcal{D}_m,\lambda}$ converge fortement vers $v = \psi(u)$ dans $L^2(\Omega \times (0,T))$,*
3. *$z_{\mathcal{D}_m,\lambda}$ converge uniformément vers 0,*
4. *$w_{\mathcal{D}_m,\lambda}$ converge fortement vers w dans $L^2(\Omega \times (0,T))$.*

De plus, le couple (u,w) est une solution faible du problème (P) au sens de la définition 6.1.5.

Preuve. L'estimation (6.23) permet d'obtenir la convergence de $u_{\mathcal{D}_m,\lambda}$ vers u pour la topologie faible-$\star$ de $L^\infty(\Omega\times(0,T))$. Les estimations sur les translations en espace (lemme 6.2.4) et en temps (lemme 6.2.6), sur u, v, z et w permettent d'obtenir la compacité nécessaire à l'application du théorème de Kolmogorov (cf. Eymard et al. (2000)). On obtient ainsi la convergence forte pour $\psi(u_{\mathcal{D}_m,\lambda})$, $v_{\mathcal{D}_m,\lambda}$, $z_{\mathcal{D}_m,\lambda}$ et $w_{\mathcal{D}_m,\lambda}$, vers, respectivement, $\psi(u)$, v, z et w dans $L^2(\Omega \times (0,T))$ quand $m \to \infty$ et quand $\lambda \to \infty$.

On a vu que $z_{\mathcal{D}_m,\lambda}$ converge vers z dans $L^2(\Omega \times (0,T))$ quand $m \to \infty$. De plus, z vérifie l'estimation (6.38), qui devient lorsque $\lambda \to \infty$,

$$||z||_{L^2(\Omega\times(0,T))} = 0. \tag{6.82}$$

On a donc bien vérifié que $z_{\mathcal{D}_m,\lambda}$ converge vers 0.

Il reste à prouver que le triplet (u,w) est bien une solution faible du problème (P) au sens de la définition 6.1.5. Pour cela, pour toute fonction test $\varphi(x,t) \in C_c^\infty(\Omega \times (0,T))$, on multiplie le schéma par $\delta t \varphi(x_K,(n+1)\delta t)$, puis on le somme sur n puis sur K. L'équation (6.18) donne $T_{15} + T_{16} = T_{17}$, avec

$$T_{15} = \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K (u_K^{n+1} - u_K^n)\varphi(x_K,(n+1)\delta t)$$

$$T_{16} = \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} \delta t \varphi(x_K,(n+1)\delta t) \sum_{L\in\mathcal{N}(K)} \tau_{K,L}\left(\psi(u_K^{n+1}) - \psi(u_L^{n+1})\right)$$

$$T_{17} = \sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K\in\mathcal{T}} m_K \delta t \varphi(x_K,(n+1)\delta t) f(w_K^n)\left(\widetilde{X} - v_K^n\right)$$

De façon classique (cf. Eymard et al. (2000)), on obtient que $\lim_{\delta t\to 0, n\to\infty} T_{15} = -\int_0^T \int_\Omega u\varphi_t \mathrm{d}x\mathrm{d}t - \int_\Omega u_0(x)\varphi(x,0)\mathrm{d}x\mathrm{d}t$, et que $\lim_{\delta t\to 0, n\to\infty} T_{16} = \int_0^T \int_\Omega \psi(u(x,t))\Delta\varphi(x,t)\mathrm{d}x\mathrm{d}t$. En effet, $\psi(u_{\mathcal{D},\lambda})$ converge fortement vers $\psi(u)$. Et finalement, comme $v_{\mathcal{D},\lambda}$ converge fortement vers v et $w_{\mathcal{D},\lambda}$ vers w, on a $\lim_{\delta t\to 0, n\to\infty} T_{17} = \int_0^T \int_\Omega \varphi(x,t) F(w,v)$.

De plus, comme $v_{\mathcal{D},\lambda} = \psi(u_{\mathcal{D},\lambda}) - z_{\mathcal{D},\lambda}$, on a $v = \psi(u) - z$, et comme $||z|| = 0$, on obtient finalement que $v = \psi(u)$. Par conséquent, $\lim_{\delta t \to 0, n \to \infty} T_{17} = \int_0^T \int_\Omega \varphi(x,t) F(w, \psi(u))$. La fonction u vérifie donc bien la définition 6.1.5 de la solution faible de (P).

□

6.4 Résultats numériques

6.4.1 Un cas 1D

On choisit les valeurs numériques suivantes

$$\overline{X} = 0.7 \qquad \widetilde{X} = 0.3 \qquad \varepsilon = 0.1 \qquad \mu = 5 \qquad T = 0.5 \tag{6.83}$$

On définit le domaine $\Omega = [0,1]$, discrétisé avec un pas d'espace constant $h = 0.005$ et on utilise un pas de temps constant $\delta t = 10\,h$. Les conditions initiales et limites sont

$$\begin{cases} u(x,t) &= 0.5, & \forall x \in \partial\Omega, t \in [0,T] \\ u_0(x) &= 1, & \forall x \in \Omega \\ w_0(x) &= 0.1, & \forall x \in \Omega \end{cases}$$

Le but de ces tests est de montrer que pour h et δt fixés, la solution du schéma $P_{(\mathcal{D},\lambda)}$ converge vers la solution du schéma $P_{\mathcal{D}}$. Pour $\lambda = 10$, on observe un décalage entre les solutions, notamment sur w (cf. figure 6.1). Le décalage sur u est plus faible (cf. figure 6.2). En faisant tendre λ vers l'infini, on vérifie que la norme de $z_{\mathcal{D},\lambda}$ tend vers 0 (cf. figure 6.3), et que l'erreur entre $w_{\mathcal{D}}$ et $w_{\mathcal{D},\lambda}$ tend également vers 0 (cf. figure 6.4).

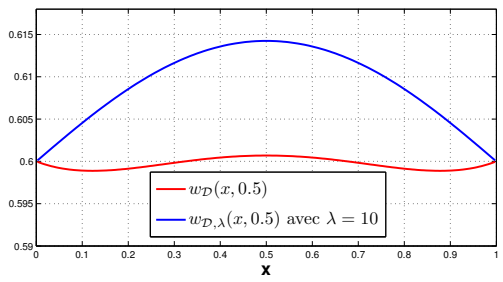


FIG. 6.1: Profil de $w_{\mathcal{D},\lambda}$ et $w_{\mathcal{D}}$ pour $t = 0.5$

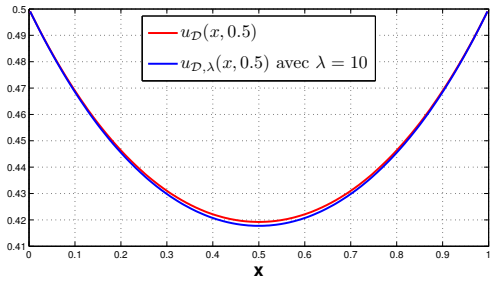


FIG. 6.2: Profil de $u_{\mathcal{D},\lambda}$ et $u_{\mathcal{D}}$ pour $t = 0.5$

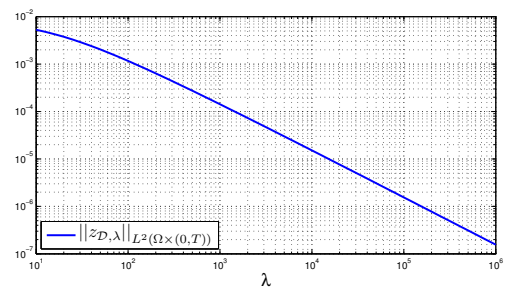


FIG. 6.3: Norme L^2 de $z_{\mathcal{D},\lambda}$ en fonction de λ

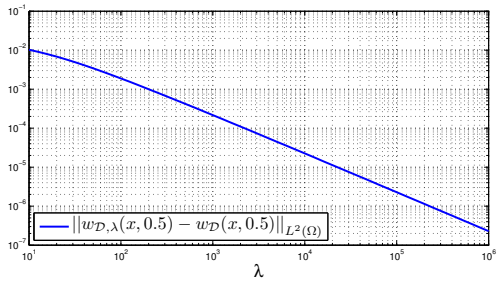


FIG. 6.4: Norme L^2 de l'erreur entre $w_{\mathcal{D},\lambda}$ et $w_{\mathcal{D}}$ en fonction de λ

6.4.2 Un cas 2D

On définit $\Omega = [0,1]\times[0,1]$, discrétisé par un maillage cartésien régulier $(n\times n)$, ce qui donne $m_K = \frac{1}{n^2}$. Les valeurs numériques des différents paramètres sont données par (6.83).

Les conditions limites et initiales sont

$$\begin{array}{ll} u(x,t) & = 0.5 \ \ \forall x \in \partial\Omega,\ t \in [0,T] \\ u_0(x) & = \left\{ \begin{array}{ll} 1 & \text{si } (x,y) \in ([0.3,\ 0.7] \times [0.3,\ 0.7]) \\ 0 & \text{sinon} \end{array} \right. \\ w_0(x) & = 0.1, \quad \forall (x,y) \in \Omega \end{array} \tag{6.84}$$

L'erreur initiale $z_0(x,0)$ est initialisée par une loi normale tronquée (cf. figure 6.5). La solution (u,w) et la fonction z, pour $t = 0.5$ et $t = 1$, sont représentées sur les figures 6.6 à 6.11.

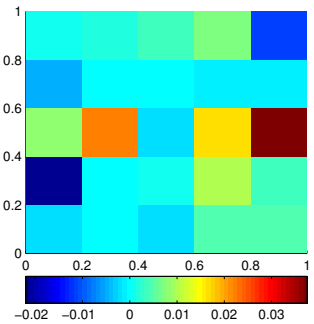


FIG. 6.5: Erreur initiale : $z_0(x)$

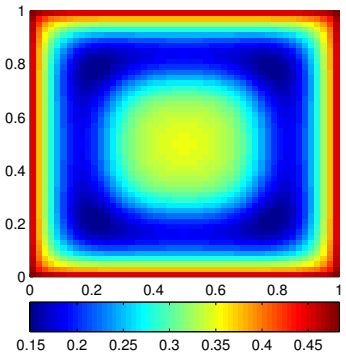


FIG. 6.6: Profil de u à $t = 0.5$

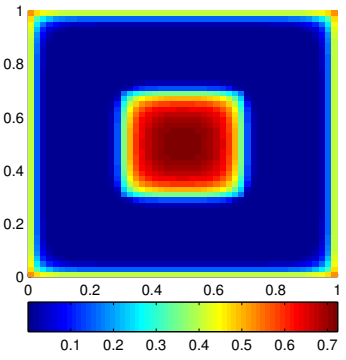


FIG. 6.7: Profil de w à $t = 0.5$

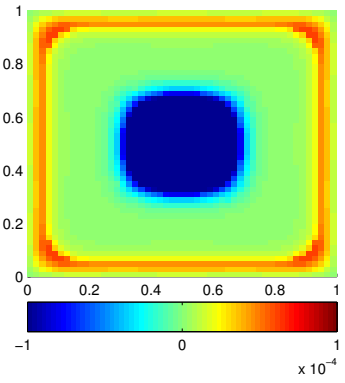


FIG. 6.8: Profil de z à $t = 0.5$

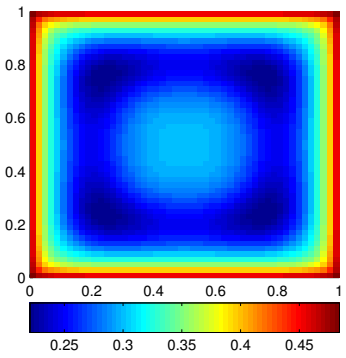


FIG. 6.9: Profil de u à $t = 1$

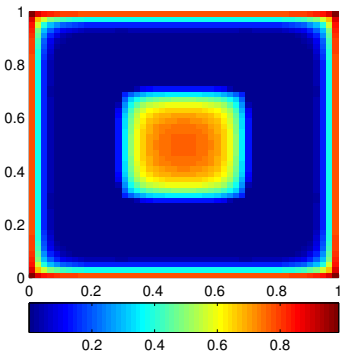


FIG. 6.10: Profil de w à $t = 1$

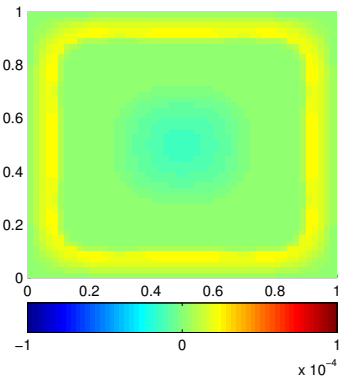


FIG. 6.11: Profil de z à $t = 1$

On étudie l'erreur entre les solutions obtenues avec les deux schémas numériques. A convergence en espace et en temps ($m_K = 10^{-4}$, $\delta t = 10 m_K$), on calcule pour différente valeur de λ (cf. figure 6.12) l'erreur entre les solutions définie par

$$E(u_\lambda, u) = \left(\sum_{n=0}^{\mathcal{N}_{\delta t}} \sum_{K \in \mathcal{T}} m_K \delta t \, (u_\lambda - u)^2 \right)^{\frac{1}{2}}. \tag{6.85}$$

A partir de $\lambda = 10$, qui est la valeur minimale de λ permettant de respecter les hypothèses des différents théorèmes de convergence, l'erreur entre les solutions obtenues avec chacun des deux schémas diminue. On calcule également la norme L^2 de $z_{\mathcal{D},\lambda}$ lorsque $(m_K, 10m_K) \to 0$, pour différentes valeurs de λ (cf. figure 6.13). On constate que pour des valeurs de λ trop faibles, cette norme ne tend pas vers 0.

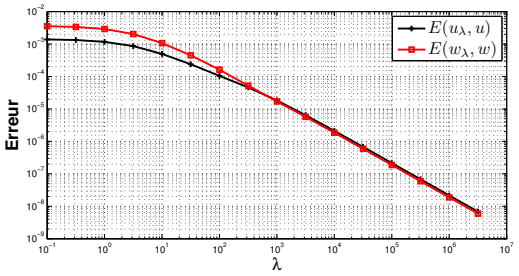


FIG. 6.12: Erreur entre les solutions

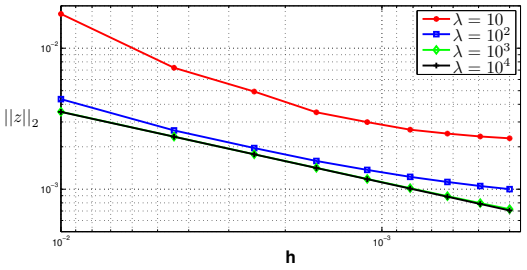


FIG. 6.13: Norme L^2 de z

On définit,

$$E_z(h) = \left(\sum_{K \in \mathcal{T}} m_K \delta t z_0(x)^2 \right)^{\frac{1}{2}} \tag{6.86}$$

avec $h = \frac{1}{n^2}$ et $\delta t = 10\, m_K$. Lorsque $\lambda \to 0$, on vérifie que $|| \, z_{\mathcal{D},\lambda} \, ||_{L^2} \to Ez$ (cf. figure 6.14).

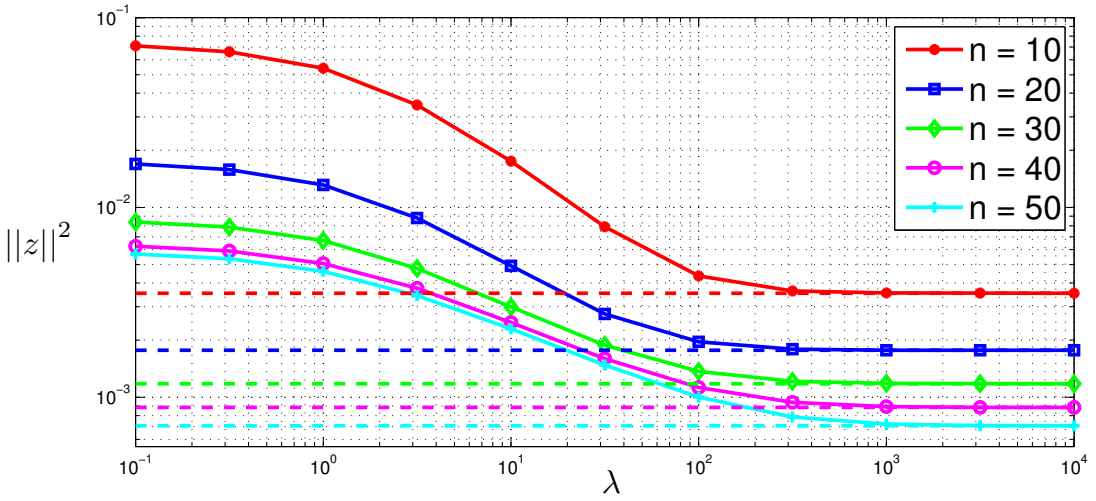


FIG. 6.14: Norme L^2 de z en fonction de $m_K, \delta t$ et λ

Troisième partie

Mise en oeuvre des différents schémas numériques

Cette partie est consacrée à l'étude des résultats numériques.

Nous cherchons tout d'abord à montrer que les différentes hypothèses faites lors de la modélisation et lors de l'écriture du schéma sont raisonnables. La première étape est de montrer que malgré les hypothèses de simplification faites pour la modélisation d'un système géochimique, le modèle final retenu est capable de reproduire des résultats corrects. Deuxièmement, on s'intéressera aux hypothèses empiriques qui ont guidé les choix faits pour le découplage du problème d'écoulement multiphasique réactif.

L'une des originalités de la méthode (*Rs-Tr*) est la correction d'erreur par pénalisation. Une étude portera sur l'efficacité de cette méthode pour corriger l'erreur de splitting, et une seconde étude permettra de montrer la convergence d'un tel schéma dans un cas simplifié.

Enfin, une étude permettra de comparer les résultats obtenus avec le schéma global implicite par rapport à ceux obtenus avec le schéma (*Rs-Tr*). Cette étude permet de valider de façon qualitative le schéma (*Rs-Tr*).

Chapitre 7

Validation des hypothèses

7.1 Validation de la modélisation d'un système géochimique

De nombreuses hypothèses de simplification sont utilisées pour la modélisation d'un système géochimique (cf. chapitre 2), notamment pour le modèle d'activité (modèle idéal) et également pour la modélisation des cinétiques de précipitation et dissolution. Afin de valider la modélisation utilisée et de montrer que les hypothèses de simplification sont raisonnables, on étudie la mise à l'équilibre d'un système géochimique donné.

Le système géochimique étudié ici provient d'une étude réalisée par Gunter et al. (1997). Le choix de ce système a été guidé par plusieurs raisons. Premièrement, le système minéralogique étudié est réaliste puisqu'il s'agit d'une minéralogie représentative de l'aquifère d'Alberta, dans lequel il est envisagé d'injecter et de stocker du co_2. De plus, les différentes données nécessaires (constantes d'équilibre, constantes de précipitation et de dissolution, etc.) sont très bien décrites dans l'article de Gunter et al. (1997), et par conséquent il est facile d'essayer de reproduire les résultats obtenus par les auteurs et de les comparer. Cette étude a également été reprise dans Azaroual et al. (2003) qui ont obtenu des résultats similaires. Nous comparerons nos propres résultats avec les graphiques issus de Azaroual et al. (2003).

La situation étudiée est une situation sans transport. Le but est de comparer l'équilibre chimique obtenu lorsque le système géochimique est déstabilisé par la présence de co_2. Pour simuler ce cas, on effectue donc une simulation avec une seule maille dans laquelle le co_2 est présent sous forme gazeuse tout au long de la simulation avec une pression partielle de 90bars. Initialement, la composition de l'eau est celle de l'aquifère d'Alberta, acidifiée par la présence du co_2. La simulation est faite sur une durée de 100 ans à une température de 105°C.

Les minéraux présents sont les suivants : quartz (87%), k-feldspath (2%), albite (1%), kaolinite (2%), sidérite (1%), annite (5%), dolomite (1%), calcite (1%) et muscovite (0%). La muscovite n'est pas initialement présente mais peut apparaître par la suite. Les espèces et éléments choisis sont :

$$\begin{aligned} \mathcal{E} &= \{H, C, Ca, O, Si, Na, Al, Mg, K, Fe\} \\ &\quad n(\mathcal{E}) = 10, \\ \mathcal{S} &= \left\{ \begin{array}{l} h_2o, h^+, hco_3^-, co_2^w, ca^{2+}, sio_2^w, na^+, al^{3+}, mg^{2+}, k^+, fe^{2+}, co_2^g, \\ calcite, quartz, kaolinite, k-feldspath, albite, annite, dolomite, sidrite, muscovite \end{array} \right\} \\ &\quad n(\mathcal{S}) = 21 \end{aligned}$$

Le système comporte donc 11 espèces en phase aqueuse, 1 espèce en phase gazeuse et 9 espèces minérales, correspondant à 9 phases solides pures. Une unique réaction équilibrée en phase aqueuse est définie : $co_2^w + h_2o \leftrightharpoons h^+ + hco_3^-$. Les réactions de dissolution et précipitation des minéraux s'écrivent :

calcite	+		h^+	$\Longleftrightarrow$		ca^{2+}	+		hco_3^-								
quartz				$\Longleftrightarrow$		sio_2^w											
kaolinite	+	6	h^+	$\Longleftrightarrow$	2	al^{3+}	+	2	sio_2^w	+	5	h_2o					
k-feldspath	+	4	h^+	$\Longleftrightarrow$		al^{3+}	+		k^+	+	2	h_2o	+	3	sio_2^w		
albite	+	4	h^+	$\Longleftrightarrow$		al^{3+}	+		na^+	+	2	h_2o	+	3	sio_2^w		
annite	+	10	h^+	$\Longleftrightarrow$		al^{3+}	+		k^+	+	3	fe^{2+}	+	3	sio_2^w	+ 6	h_2o
dolomite	+	2	h^+	$\Longleftrightarrow$		ca^{2+}	+		mg^{2+}	+	2	hco_3^-					
sidérite	+		h^+	$\Longleftrightarrow$		fe^{2+}	+		hco_3^-								
muscovite	+	10	h^+	$\Longleftrightarrow$		k^+	+	3	al^{3+}	+	3	sio_2^w	+	6	h_2o		

Le coefficient de dissolution, k_r^d, d'un minéral tient compte d'un coefficient de dissolution en milieu acide, k_r^a, d'un coefficient de dissolution en milieu neutre, k_r^n, et d'un coefficient prenant en compte la quantité de co_2 dissous. Cela correspond à choisir k_r^d de la façon suivante

$$k_r^d = k_r^a a_{(h^+)} + k_r^n + k_r^{co_2} P_{co_2} \tag{7.1}$$

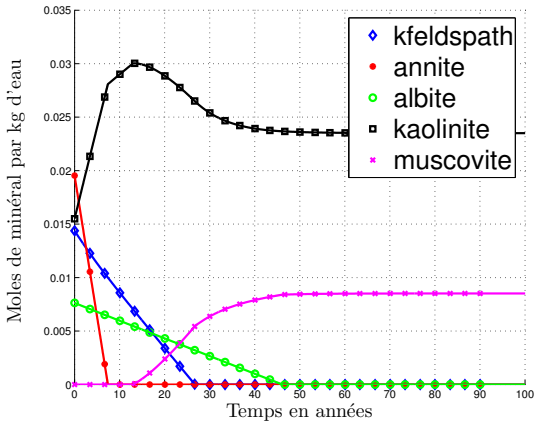
P_{co_2} est la pression partielle du CO_2 (ici 90 bars). Pour tous les minéraux, le coefficient de précipitation est

$$k_r^p = 1 \tag{7.2}$$

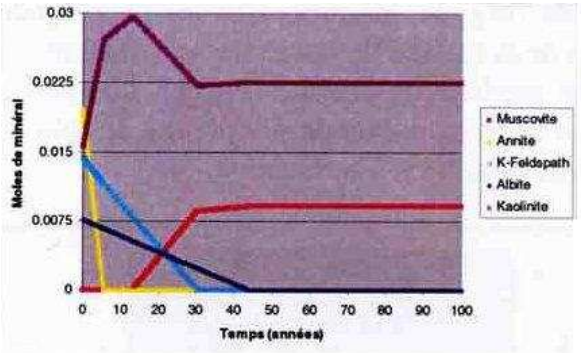
Choisir $k_r^p = 1$ revient à précipiter très rapidement.

Les constantes de cinétique des minéraux, les constantes d'équilibre des réactions, ainsi que la composition de l'eau initiale sont décrites précisément dans l'annexe A.

Les résultats obtenus sont similaires à ceux obtenus par Gunter et al. (1997) et par Azaroual et al. (2003). On représente sur la figure 7.4(a) l'évolution du quartz au cours du temps. L'évolution du quartz ne figure pas dans les résultats de Gunter et al. (1997). La figure 7.1(a) représente l'évolution des minéraux à cinétique lente (à comparer avec la figure 7.1(b)), puis les minéraux à cinétique rapide (les carbonates : calcite, dolomite, sidérite) sont représentés sur la figure 7.2(a) en échelle logarithmique pour le temps (à comparer avec la figure 7.2(b)). La dolomite et la sidérite sont à nouveaux représentés sur la figure 7.3(a), mais sans échelle logarithmique (à comparer avec la figure 7.3(b)).

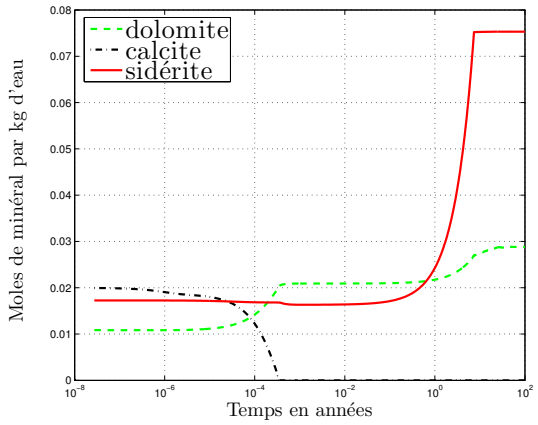


(a) Minéraux à cinétique lente

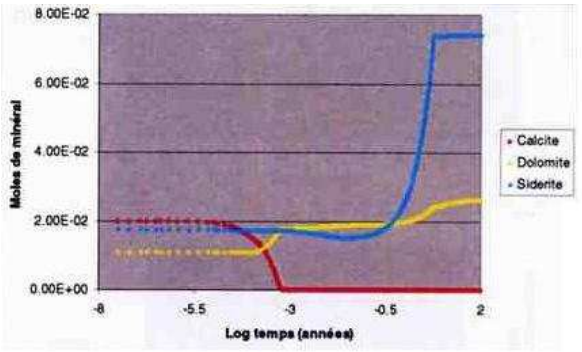


(b) Minéraux à cinétique lente, résultats de Azaroual et al. (2003)

FIG. 7.1: Mise à l'équilibre du système minéralogique de l'Alberta, comparaison avec les résultats de Azaroual et al. (2003)

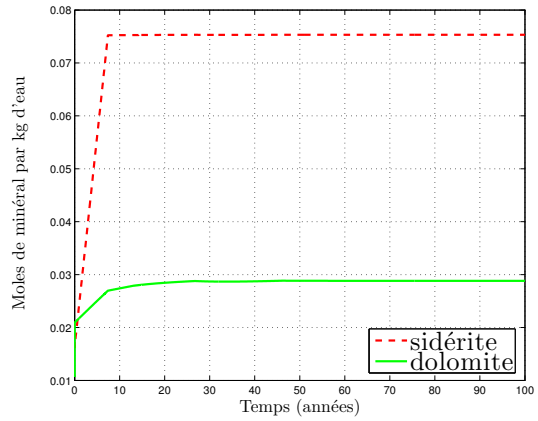


(a) Minéraux à cinétique rapide

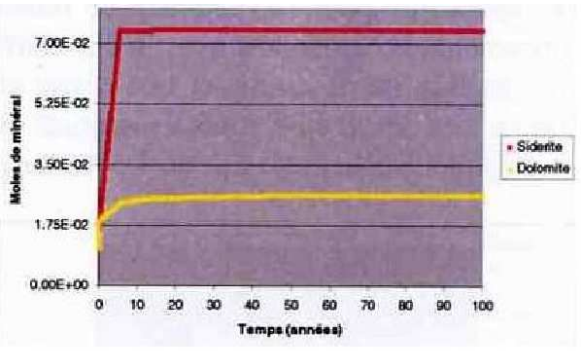


(b) Minéraux à cinétique rapide, résultats de Azaroual et al. (2003)

FIG. 7.2: Mise à l'équilibre du système minéralogique de l'Alberta, comparaison avec les résultats de Azaroual et al. (2003)

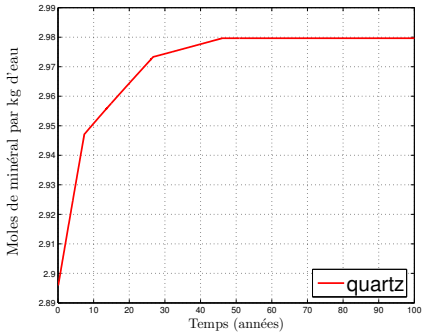


(a) Dolomite et sidérite



(b) Dolomite et sidérite, résultats de Azaroual et al. (2003)

FIG. 7.3: Mise à l'équilibre du système minéralogique de l'Alberta, comparaison avec les résultats de Azaroual et al. (2003)



(a) Quartz

FIG. 7.4: Mise à l'équilibre du système minéralogique de l'Alberta, comparaison avec les résultats de Azaroual et al. (2003)

Le système met 46 ans pour atteindre un état d'équilibre. Les phénomènes les plus rapides sont la dissolution de la calcite (dissolution complète en 3.5 ans), accompagnée d'une précipitation de la dolomite. Le quartz précipite jusqu'à ce que l'état d'équilibre soit atteint. Il précipite avec différentes vitesses en fonction de la cinétique des autres minéraux. L'annite se dissout complètement en 7 ans. Parallèlement la kaolinite précipite. Le comportement de la kaolinite s'inverse lorsque la muscovite commence à apparaître (début de précipitation à 13 ans). Le k-feldspath est complètement dissous en 26.5 ans, l'albite en 46 ans. Une fois que l'albite a totalement disparu, le système est à l'équilibre et n'évolue plus.

Ces résultats, en accord avec ceux de Gunter et al. (1997) et de Azaroual et al. (2003), permettent de montrer que malgré les simplifications réalisées, le modèle utilisé au cours de cette thèse est suffisant pour reproduire une mise à l'équilibre d'un système géochimique. Notamment, le choix d'un modèle idéal pour le calcul des activités et une loi simplifiée de cinétique, n'influencent pas l'équilibre chimique obtenu.

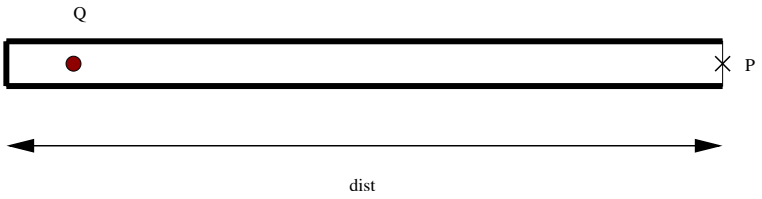
7.2 Validité de l'hypothèse de découplage

On souhaite étudier la validité de l'hypothèse de découplage H1, c'est à dire justifier le fait d'associer le calcul des échanges eau-gaz et le calcul d'écoulement. Pour faire cela, on modifie l'équilibre eau-gaz en faisant varier la solubilité du co_2 dans l'eau, et on observe les changements induits sur la vitesse de Darcy, et donc sur l'écoulement. Cette étude est réalisée sur un modèle 1D où le système géochimique se limite à une unique réaction eau-gaz.

7.2.1 Description générale

On modélise un écoulement diphasique avec dissolution du gaz dans l'eau dans un aquifère horizontal de longueur L. On injecte du dioxyde de carbone avec un débit constant égal à $0,01\text{m}^3$ par jour, en condition de fond, pendant vingt ans entre les dates $t = 2$ ans et $t = 22$ ans. On observe l'évolution du système jusqu'à la date $t = 30$ ans.

FIG. 7.5: Schéma de la colonne 1D horizontale



Les bords du domaine sont imperméables, à l'exception du bord vertical à droite. La pression est imposée, $P = 100$ bars. Initialement, il n'y a pas de gaz dans le domaine. La porosité est uniforme ($\Phi = 0.2$), ainsi que la perméabilité ($K = 100$mD). Il n'y a pas de pression capillaire. La densité de l'eau est constante ainsi que les viscosités de l'eau et du gaz. Le détail des paramètres utilisés pour la simulation se trouve dans l'annexe B.

Le système compositionnel utilisé pour ce test est très simplifié. Le gaz est constitué d'une seule espèce qui peut se dissoudre dans l'eau.

Adimensionnement

Dans ce cas, le problème peut se réécrire de la façon suivante :

$$\begin{aligned}
&\Phi\left(\xi_w S_w(1-X)\right)_t - \left(\xi_w(1-X)\tfrac{KS_w}{\mu_w}P_x\right)_x = 0\\
&\Phi\left(\xi_w S_w X + \xi_g(1-S_w)\right)_t - \left(\xi_w X\tfrac{KS_w}{\mu_w}P_x + \xi_g\tfrac{K(1-S_g)}{\mu_g}P_x\right)_x = Q
\end{aligned}$$

Il est possible d'adimensionner cette équation en divisant l'équation par ξ_w et en posant $t = \frac{t}{\Phi T}$, $x = \frac{x}{L}$ et $P = \frac{P}{\overline{P}}$. On obtient alors les équations adimensionnées suivantes

$$\begin{aligned}
&\left(S_w(1-X)\right)_t - \tfrac{KT\overline{P}}{L^2\mu_w}\left((1-X)S_w P_x\right)_x = 0\\
&\left(S_w X + \xi(1-S_w)\right)_t - \tfrac{KT\overline{P}}{L^2\mu_w}\left(X S_w P_x + \xi\tfrac{(1-S_g)}{\mu}P_x\right)_x = QT
\end{aligned}$$

La vitesse de Darcy de la phase aqueuse qui vaut, avant adimensionnement :

$$v_w = -\frac{K}{\mu_w}S_w P_x,$$

devient :

$$\overline{v_w} = -\left(\frac{K}{\mu_w}S_w P_x\right)\frac{T}{L}.$$

Le débit d'injection adimensionné est alors

$$\overline{Q} = Q\frac{T}{L} \simeq 0.11.$$

7.2.2 Les résultats

Pour faire varier le solubilité du co_2 dans l'eau, on fait varier la constante de Henry (de 50 à 10^3). Plus la constante de henry est forte, moins le gaz est soluble dans le l'eau (figure 7.2.2). On s'intéresse alors au profil de la vitesse de Darcy (adimensionnée) obtenue pour la phase aqueuse sur une interface au centre du système (x = 0.5) au cours du temps (cf 7.2.2). On constate que la vitesse de Darcy de la phase aqueuse est fortement corrélée avec la solubilité du co_2.

Il est possible de distinguer plusieurs phases dans l'évolution de la vitesse.

1. Tout d'abord, avant le début de l'injection le système est à l'équilibre, la vitesse est nulle.
2. Lorsque l'injection commence, la vitesse augmente rapidement pour atteindre un palier. Lors de cette période, le gaz injecté se dissout dans la partie du domaine précédant le point d'observation. La valeur de la vitesse atteinte et la durée de cette période dépendent de la constante de Henry.

   Plus la constante de Henry est forte, plus la valeur de la vitesse est grande. En effet, la solubilité étant plus faible, on dissout peu de gaz, et la quantité de gaz est donc plus grande. Or, étant donné les différences entre la densité de l'eau et celle du gaz (une tonne de $co2$ sous forme gazeuse occupe près de 1,5 m^3, alors qu'une tonne de $co2$ dissout dans l'eau occupe 1 m^3 seulement), le volume à déplacer est alors plus important et par conséquent la vitesse est plus grande.

   Plus la constante de Henry est forte, plus la durée de cette période est courte. En effet, la fin de cette période correspond au moment où toute l'eau dans la partie du domaine précédant le point d'observation est saturée en gaz. Or, plus la solubilité est faible, plus cela se produit rapidement.

H	$v_{\max}$	date fin période 2
50	0.08	0.1
100	0.07	0.122
500	0.048	0.244
1000	0.04	0.374

3. Lorsque toute l'eau dans la partie du domaine située en amont du point d'observation est saturée en gaz, il n'y a plus de dissolution possible et par conséquent les fluides se comportent comme des fluides immiscibles. Ce moment correspond à l'apparition de la phase gazeuse dans la maille précédent l'interface observée (cf figure 7.2.2). La solubilité n'a alors plus d'influence sur l'écoulement. Les vitesses obtenues pour les différentes valeurs de Henry ont alors toutes le même profil. Elles correspondent à un débit à travers l'interface d'observation identique au débit d'injection (cf figure 7.2.2). En effet, comme il n'y a plus de dissolution, il n'y a plus de changement de volume. Néanmoins la vitesse n'est pas constante car la quantité de gaz est de plus en plus grande et par conséquent, bien que le débit total soit alors constant, le débit de gaz augmente et le débit d'eau diminue (et donc la vitesse).
4. La dernière période commence lorsque l'injection est stoppée. Comme il n'y a pas de gravité (domaine horizontal) et pas de diffusion, le système est à l'équilibre et la vitesse est alors nulle.

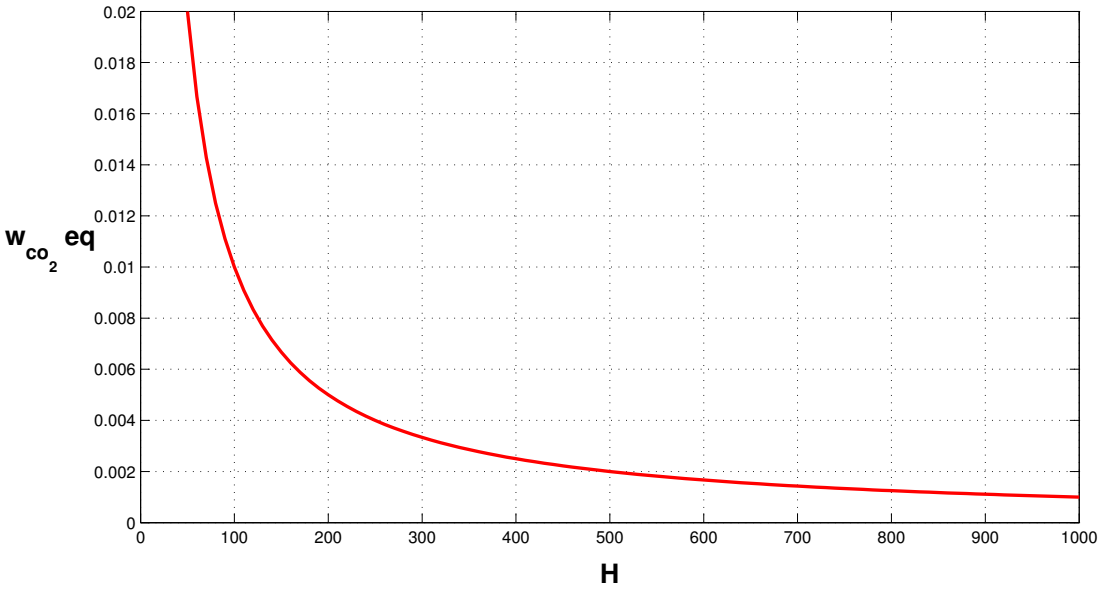


FIG. 7.6: Fraction molaire d'équilibre en fonction de la constante de Henry

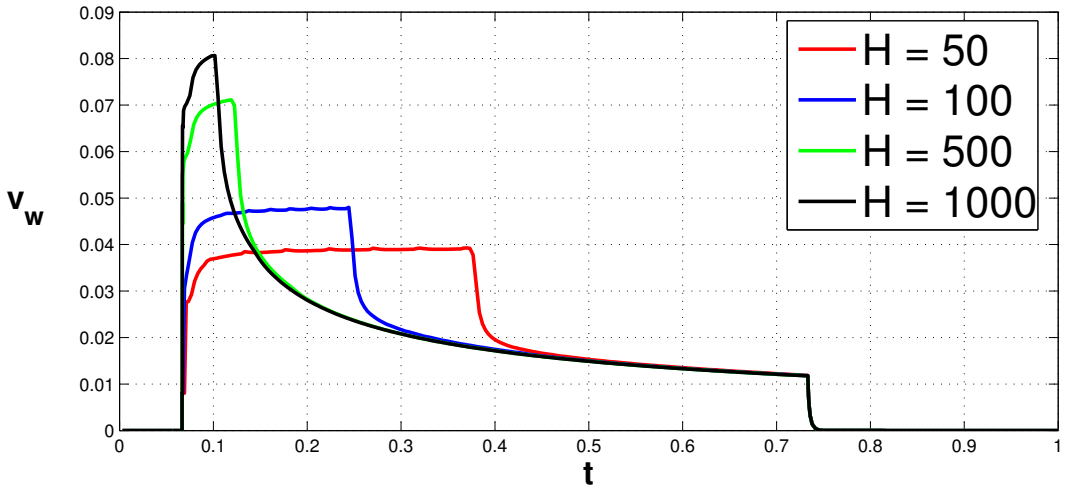


FIG. 7.7: Vitesse de Darcy (adimensionnée) de la phase aqueuse.

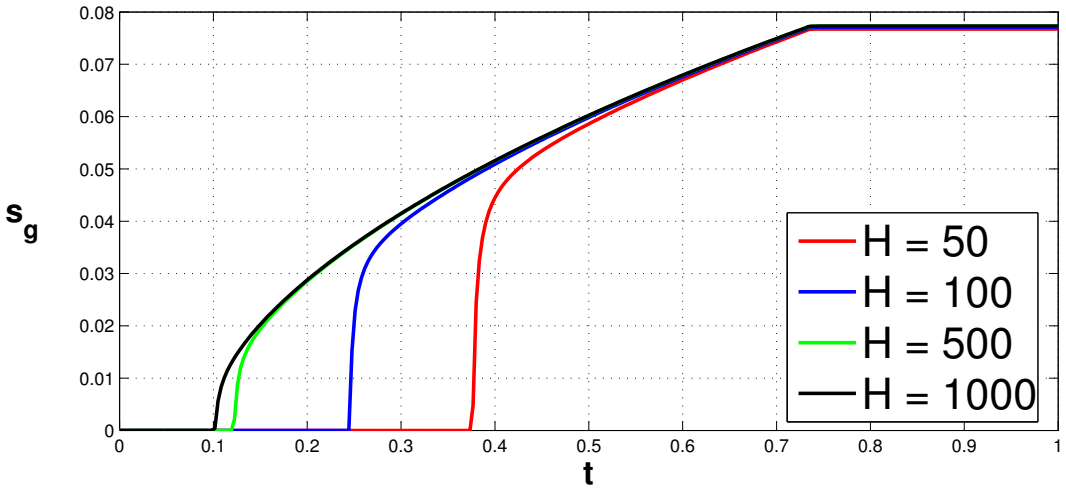


FIG. 7.8: Saturation de gaz au cours du temps dans la maille précédant le point d'observation.

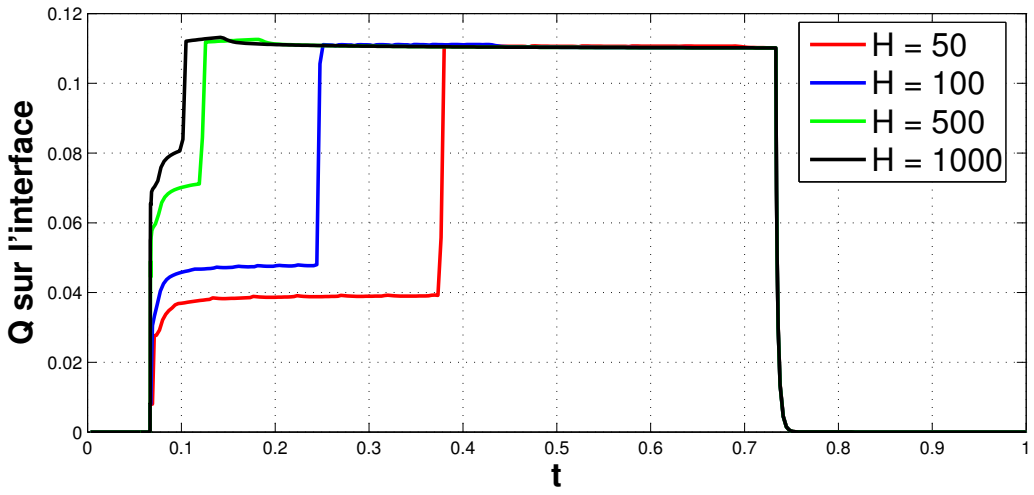


FIG. 7.9: Débit (adimensionnée) traversant la face considérée au cours du temps.

Chapitre 8

Correction d'erreur par pénalisation

Ce chapitre est consacré à la correction d'erreur par pénalisation. On souhaite étudier l'effet de l'ajout d'un terme de pénalisation sur le schéma.

Premièrement, on souhaite savoir si la pénalisation permet bien de réduire l'erreur due au splitting. On proposera également d'autres méthodes de correction d'erreur afin de mieux évaluer l'efficacité de la méthode de correction d'erreur par pénalisation. De même, il est également intéressant d'étudier la possibilité d'ajouter le terme de pénalisation sur l'un ou l'autre des modèles, même si on n'envisage pas de le faire sur des problèmes complexes. Deuxièmement, on se demande si le schéma pénalisé proposé est un schéma convergent, et s'il converge vers la même solution qu'un schéma global implicite.

Deux problèmes simplifiés serviront pour cette étude. Le premier est un problème 0D simplifié contenant les principales caractéristiques du problème d'écoulement multiphasique réactif (opérateur de transport non linéaire, deux réactions en compétition dont l'une est fortement couplée au terme de transport). Le second est un problème de type diffusion-réaction, pour lequel nous montrerons la convergence du schéma pénalisé vers une solution faible.

8.1 Écriture d'un problème 0D simplifié

Pour étudier la correction d'erreur par pénalisation, on étudie un problème 0D simplifié qui reprend les caractéristiques principales du problème complet, à savoir un terme de transport non linéaire, et une compétition entre deux termes de réaction, T_α et T_β. L'intérêt de ce problème est qu'il permet de comparer les solutions obtenues numériquement avec une solution analytique, ce qui permet de bien visualiser le fonctionnement de la correction d'erreur par pénalisation.

Le système s'écrit

$$\begin{cases} u_t + L\left(u, v(t,u)\right) + T_\alpha + T_\beta = 0 \\ u(0) = u_0 \end{cases} \tag{8.1}$$

où $L(u_1, v(t,u_2)) = (u_1 v(t,u_2) - \overline{u} v(t,\overline{u}))$, $v(t,u)$ représente la vitesse d'écoulement et $\overline{u}$ est la condition limite.

Le schéma global implicite, noté (Sc), pour la résolution de (8.1) s'écrit

$$(Sc) \quad \frac{u^{n+1} - u^n}{\delta t} + L\left(u^{n+1}, v(t^{n+1}, u^{n+1})\right) + T_\alpha^{n+1} + T_\beta^{n+1} = 0 \tag{8.2}$$

Par analogie avec le problème complet d'écoulement multiphasique réactif (2.35), on peut écrire le schéma (Rs-Tr) pour le problème (8.1). On considère que T_α est un terme de réaction qui influence l'écoulement et qui est donc pris en implicite dans le modèle (Rs) et en explicite dans le modèle (Tr). Le terme de réaction T_β est un terme de réaction qui influence peu l'écoulement, il est donc pris en explicite dans le modèle (Rs) et en implicite dans le modèle (Tr). La vitesse est calculée par le modèle (Rs). On

obtient alors le schéma suivant :

$$\begin{array}{ll} (Rs) & \dfrac{u_{Rs}^{n+1}-u_{Rs}^{n}}{\delta t}+L\left(u_{Rs}^{n+1},v(t^{n+1},u_{Rs}^{n+1})\right)+T_{\alpha}^{n+1,Rs}+T_{\beta}^{n,Fr}+\lambda\left(u_{Rs}^{n}-u_{Fr}^{n}\right)=0 \\ (Fr) & \dfrac{u_{Fr}^{n+1}-u_{Fr}^{n}}{\delta t}+L\left(u_{Fr}^{n+1},v(t^{n+1},u_{Rs}^{n+1})\right)+T_{\alpha}^{n+1,Rs}+T_{\beta}^{n+1,Fr}=0 \end{array} \tag{8.3}$$

Le terme $T_{\alpha}^{n+1,Rs}$ utilisé par le modèle (Fr) est déduit de la résolution du modèle (Rs). On a donc, pour tout $n \geq 0$,

$$T_{\alpha}^{n+1,Rs}=-\frac{u_{Rs}^{n+1}-u_{Rs}^{n}}{\delta t}-L\left(u_{Rs}^{n+1},v(t^{n+1},u_{Rs}^{n+1})\right)+T_{\beta}^{n,Fr}-\lambda\left(u_{Rs}^{n}-u_{Fr}^{n}\right) \tag{8.4}$$

Le terme $T_{\beta}^{n,Fr}$ utilisé par le modèle (Rs) est déduit de la résolution du modèle (Fr), si c'est possible, c'est à dire si $n > 0$. En effet, pour $n = 0$, aucune résolution de (Fr) n'a été effectuée. Le terme $T_{\beta}^{n,Fr}$ s'écrit donc

$$\begin{cases} \text{pour } n=0, & T_{\beta}^{0,Fr}=0 \\ \text{pour } n>0, & T_{\beta}^{n,Fr}=-\dfrac{u_{Fr}^{n+1}-u_{Fr}^{n}}{\delta t}-L\left(u_{Fr}^{n+1},v(t^{n+1},u_{Rs}^{n+1})\right)-T_{\alpha}^{n+1,Rs} \end{cases} \tag{8.5}$$

8.2 Solution analytique et résultats numériques

On pose $v(t,u)=u$, si $(t_1 \leq t \leq t_2)$ et $v(t,u)=0$, sinon. On pose $T_{\alpha}=c(u-\alpha)$ et $T_{\beta}=d(u-\beta)$. Ces expressions de v, T_{α} et T_{β} permettent de trouver une solution analytique.

Les paramètres sont $(c=0.7, d=0.4, \alpha=0.6, \beta=0.1, t_1=100, t_2=150)$, la condition initiale est $u^0=0.5$ et la condition limite est $\overline{u}=0.8$.

8.2.1 Écriture de la solution analytique

Le problème s'écrit maintenant

$$\begin{cases} u_t+u^2+u(c+d)+\left(\overline{u}^2-c\alpha-d\beta\right)=0 & \text{si } t_1 \leq t \leq t_2 \\ u_t+u(c+d)-(c\alpha+d\beta)=0 & \text{sinon} \end{cases} \tag{8.6}$$

On pose $u_{eq}=\frac{c\alpha+d\beta}{c+d}$. La solution analytique (cf. figure 8.1) s'écrit

$$u(t)=\begin{cases} u_{eq}+(u_0-u_{eq})e^{-(c+d)t} & \text{si } 0 \leq t < t_1 \\ \dfrac{-ks_1e^{(s_1-s_2)t}+s_2}{1-ke^{(s_1-s_2)t}} & \text{si } t_1 \leq t \leq t_2 \\ u_{eq}+(u_2-u_{eq})e^{-(c+d)t} & \text{si } t_2 \leq t \leq T \end{cases} \tag{8.7}$$

avec $u_1=u_{eq}+(u_0-u_{eq})e^{-(c+d)t_1}$, $u_2 == \frac{-ks_1e^{(s_1-s_2)t_2}+s_2}{1-ke^{(s_1-s_2)t_2}}$, $k=\frac{s_2-u_1}{s_1-u_1}e^{(s_2-s_1)t_1}$, $s_1=\frac{-(c+d)-\sqrt{(c+d)^2-4(\overline{u}^2-c\alpha-d\beta)}}{2}$ et $s_2=\frac{-(c+d)+\sqrt{(c+d)^2-4(\overline{u}^2-c\alpha-d\beta)}}{2}$.

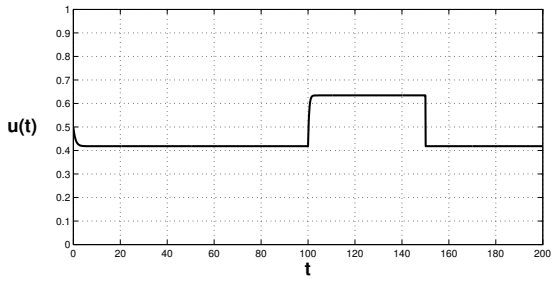


FIG. 8.1: Solution analytique

8.2.2 Résultats numériques

Pour étudier l'influence de la pénalisation, on fait varier le paramètre λ de 0 à 3.5. Le paramètre λ ne peut être choisi trop grand afin de conserver un schéma stable. L'erreur entre deux solutions u_1 et u_2 est calculée de la façon suivante :

$$E(u_1, u_2) = \frac{1}{T} \sum_{n=0}^{N} \delta t \, |u_1^n - u_2^n| \tag{8.8}$$

On s'intéresse à l'erreur commise par rapport à la solution obtenue avec le schéma couplé implicite, et par rapport à la solution analytique $u(t)$. On constate que toutes les erreurs diminuent jusqu'à approximativement $\lambda = 2$ (cf. figure 8.6). Ensuite, l'erreur augmente. Ceci signifie que le schéma est devenu instable et par conséquent la solution oscille. De plus, on constate que quelque soit la valeur choisie pour λ, le schéma global implicite est toujours celui qui donne l'erreur la plus faible.

Lorsque $\lambda = 0$, les schémas (Rs) et (Tr) évoluent séparément (cf. figure 8.2). Ce cas correspond à un splitting sans correction d'erreur. Plus λ augmente, plus les deux solutions se rapprochent rapidement. Mais on constate également de plus fortes oscillations au début (cf. figures 8.3 et 8.4). Et finalement, lorsque λ est trop grand, les solutions oscillent fortement (cf. figure 8.5).

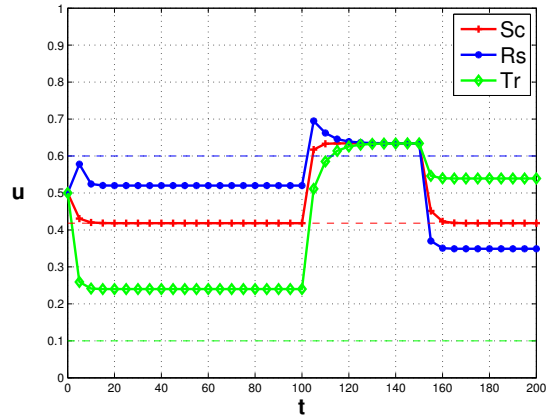


FIG. 8.2: Solution pour $\lambda = 0$

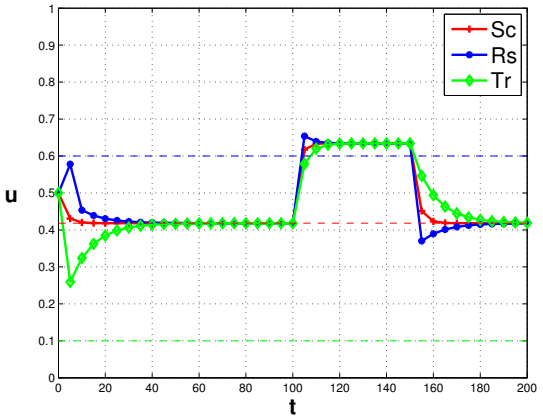


FIG. 8.3: Solution pour $\lambda = 1$

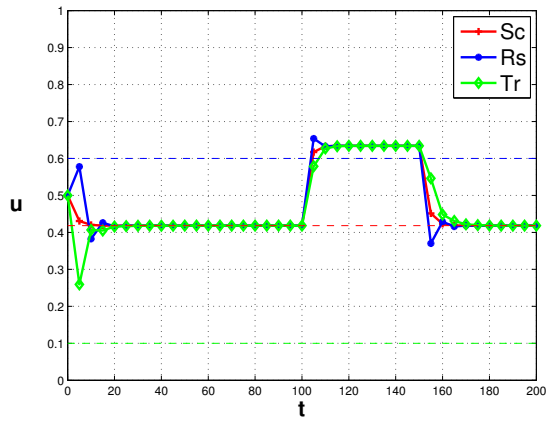


FIG. 8.4: Solution pour $\lambda = 2$

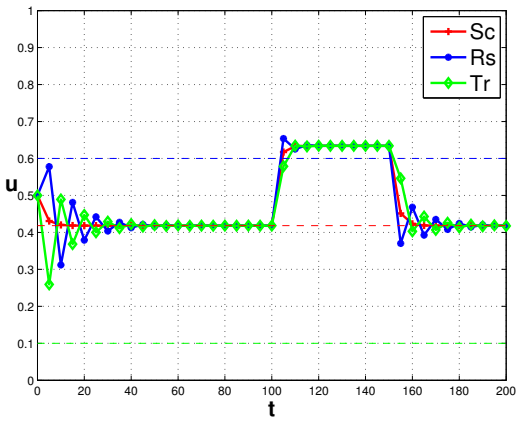


FIG. 8.5: Solution pour $\lambda = 3$

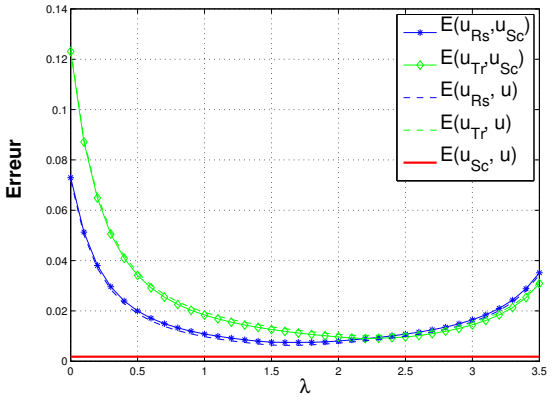


FIG. 8.6: Erreur en fonction de λ

8.2.3 Importance de l'initialisation

Initialement, le premier terme de réaction utilisé par le modèle (Rs), $T_\beta^{0,Tr}$ est nul, car dans un cas général, aucune information ne permet d'ajuster ce terme. Dans ce cas simplifié, il est possible d'initialiser correctement T_β^0 car on dispose d'une formule analytique, ce qui donne

$$T_\beta^0 = d(u_0 - \beta). \tag{8.9}$$

Il est donc possible d'utiliser (8.9) pour initialiser le modèle (Rs). Cette initialisation est évidemment de meilleure qualité. On constate une amélioration importante sur la qualité de la solution (cf. figure 8.7 et figure 8.8) d'un point de vue de la régularité initiale. En terme de précision, pour $\lambda = 0$, l'erreur par rapport à la solution exacte est diminuée d'un facteur 2 environ, mais pour $\lambda = 2$, l'erreur est du même ordre de grandeur. En effet, sans correction d'erreur (*i.e.* $\lambda = 0$), le fait de partir dès le début dans la bonne direction grâce à une bonne initialisation permet de réduire l'erreur de $t = 0$ jusqu'au début de la période d'injection. Mais, lorsque $\lambda = 2$, seules les premières itérations sont influencées par l'initialisation. C'est pourquoi, l'effet sur l'erreur tout au long de la simulation est réduit.

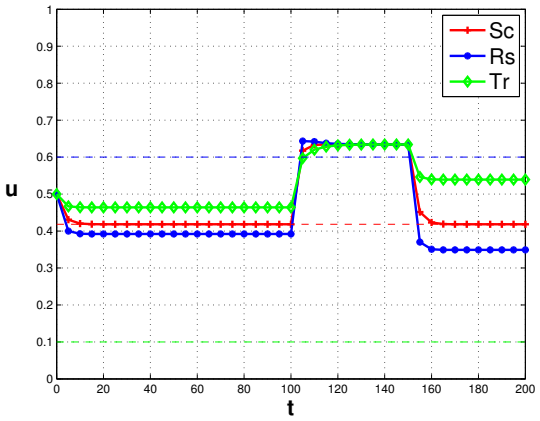


FIG. 8.7: Solution pour $\lambda = 0$, et $T_\beta^0 = d(u^0 - \beta)$

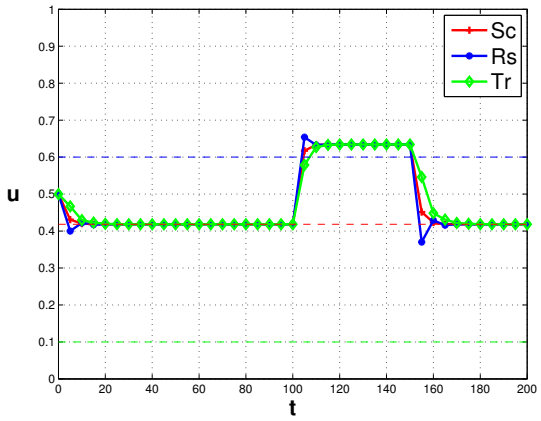


FIG. 8.8: Solution pour $\lambda = 2$, et $T_\beta^0 = d(u^0 - \beta)$

Dans le problème complet, il n'est pas possible d'initialiser le modèle (Rs) par un calcul simple. Une étape préalable de chimie locale permettrait de faire une initialisation correcte.

8.2.4 Pénalisation du modèle (Rs) ou du modèle (Tr) ?

Comme on l'a dit dans la section 5.3, le choix de pénaliser le modèle (Rs) plutôt que le modèle (Tr) est motivé par l'impossibilité de pénaliser un modèle disposant d'une description fine de la géochimie, à partir d'une solution obtenue avec une description grossière de la géochimie. Néanmoins sur ce cas simplifié, il est possible d'étudier ces différentes possibilités.

Pénaliser le modèle (Tr) semble attractif car cela permet d'impliciter le terme de pénalisation. On évite ainsi les problèmes de stabilité du schéma si λ est trop grand. Si l'initialisation est correcte, c'est à dire si (Rs) est initialisé avec (8.9), lorsque $\lambda \to \infty$, le schéma ($Rs - Tr$) converge vers le schéma couplé semi-explicite suivant

$$(S_c^{\text{ex}}) \quad \frac{u^{n+1} - u^n}{\delta t} + L\left(u^{n+1}, v(t^{n+1}, u^{n+1})\right) + T_\alpha^{n+1,Rs} + T_\beta^{n,Tr} = 0 \tag{8.10}$$

En effet, si la pénalisation est mise sur (Tr) on force la solution de (Tr) à rejoindre celle de (Rs), ce qui explique qu'on converge vers le schéma (S_c^{ex}) et non vers le schéma (Sc). Lorsque $\lambda \to \infty$, l'erreur entre la solution de (Rs) et la solution de (S_c^{ex}), ainsi que l'erreur entre la solution de (Tr) et la solution de (S_c^{ex}), tend vers 0 (cf. figure 8.9). Or, le but étant de réduire l'erreur lorsqu'on utilise de grands pas de temps, cette stratégie ne semble pas être aussi intéressante qu'on pourrait le penser à priori. On constate d'ailleurs que l'erreur des solutions de (Rs) et de (Tr) par rapport à la solution exacte ne diminue pas lorsque $\lambda \to \infty$, à δt fixé (cf. figure 8.10).

A l'inverse, lorsque la pénalisation est sur le modèle (Rs), cela force la solution de (Rs) à rejoindre celle de (Tr) qui elle est implicite. Pour de grands pas de temps, bien que la valeur de λ soit limitée pour des raisons de stabilité, il est plus intéressant de pénaliser (Rs) que de pénaliser (Tr).

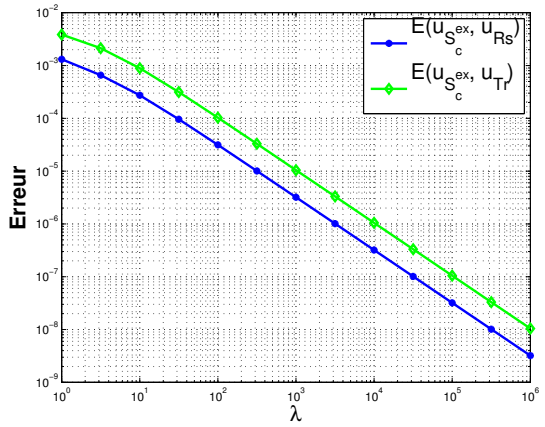


FIG. 8.9: Erreur par rapport à la solution de (S_c^{ex}) selon λ

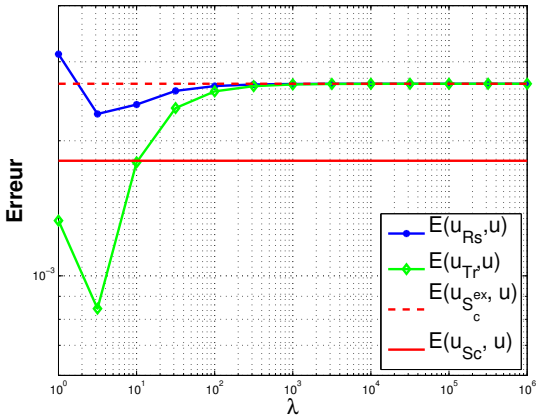


FIG. 8.10: Erreur par rapport à la solution exacte selon λ

8.2.5 Ajout d'une boucle itérative

L'un des objectifs de la correction d'erreur par pénalisation est de pouvoir réduire l'erreur de splitting sans utiliser un processus itératif trop coûteux. Les résultats précédents sur le cas simplifié montrent que la correction d'erreur permet bien de réduire l'erreur au cours du temps, mais que localement, la solution comporte des oscillations non physiques. C'est le cas, par exemple, au début de la simulation si le terme T_β^0 n'est pas initialisé correctement, et c'est aussi le cas au début et à la fin de l'injection, c'est à dire lorsqu'il y a un changement brusque dans la physique du problème.

Pour améliorer ce point, on souhaite étudier la possibilité de réduire l'erreur de splitting par un processus itératif. L'idée est de n'utiliser le processus itératif que lorsque la situation physique change rapidement. On suppose que lorsque c'est le cas, l'erreur entre la solution de (Rs) et la solution de (Tr) est importante. Lorsque l'erreur entre la solution de (Rs) et la solution de (Tr) est plus faible, on considère que

la correction d'erreur par pénalisation est suffisante. Il faut donc choisir un critère d'arrêt des itérations qui ne soit pas trop strict afin que les itérations ne s'effectuent que lorsque l'erreur entre les solutions est importante. Le schéma correspondant s'écrit

$$\begin{aligned}
(Rs\text{-}it)\quad & \frac{u_{Rs}^{n+1,k}-u_{Rs}^{n}}{\delta t}+L\left(u_{Rs}^{n+1,k},v(t^{n+1},u_{Rs}^{n+1,k})\right)+T_{\alpha}^{n+1,k,Rs}+T_{\beta}^{n+1,k-1,Tr}+\lambda\left(u_{Rs}^{n}-u_{Tr}^{n}\right)=0\\
(Tr\text{-}it)\quad & \frac{u_{Tr}^{n+1,k}-u_{Tr}^{n}}{\delta t}+L\left(u_{Tr}^{n+1,k},v(t^{n+1},u_{Rs}^{n+1,k})\right)+T_{\alpha}^{n+1,k,Rs}+T_{\beta}^{n+1,k,Tr}=0\\
& \text{tant que } \left|u_{Rs}^{n+1,k}-u_{Tr}^{n+1,k}\right|>\varepsilon
\end{aligned} \tag{8.11}$$

Afin de comparer les résultats obtenus avec cette méthode, on essayera également une méthode de réduction d'erreur purement itérative.

Les résultats numériques suivants correspondent au cas où $\lambda = 2$ et où le schéma est initialisé par $T_{\beta}^{0} = 0$. C'est en effet principalement les oscillations provoquées par cette mauvaise initialisation que l'on souhaite supprimer.

On fait varier ε de 1 à 10^{-4} pour étudier l'influence du processus itératif. Plus ε est petit plus l'erreur par rapport à la solution de (Sc) diminue. Pour une même valeur de ε, l'erreur est toujours plus faible lorsque les itérations sont associées à la pénalisation (cf. figure 8.11). De plus, le nombre d'itérations nécessaires est également plus faible lorsque l'erreur est corrigée par pénalisation en plus des itérations (cf. figure 8.12). On constate également que sans pénalisation, la méthode itérative a parfois des difficultés à converger.

Associer une méthode itérative à la pénalisation peut donc se révéler utile pour améliorer la qualité de la solution. Néanmoins, cela nécessite de trouver un compromis entre la précision voulue et le coût de calcul supplémentaire. Dans le cas de ce problème simplifié, les résultats montrent que la valeur $\varepsilon = 6.3 \times 10^{-2}$ permet de supprimer le problème des oscillations (cf. figure 8.13) avec un nombre d'itérations supplémentaires raisonnable (6 itérations supplémentaires pour un total de 40 sans méthode itérative).

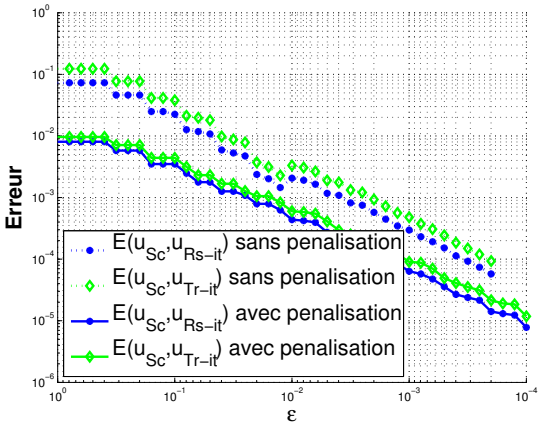


FIG. 8.11: Erreur par rapport à la solution de (Sc) selon ε

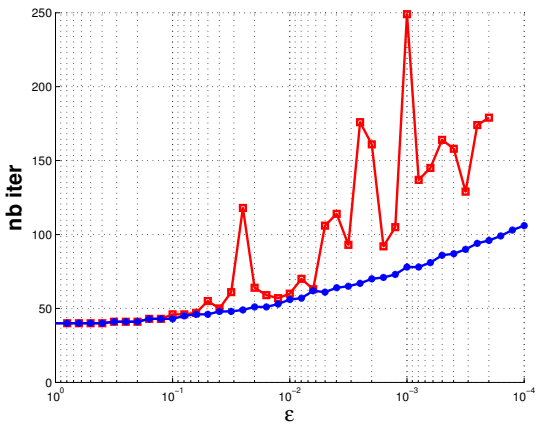


FIG. 8.12: Nombre d'itérations en fonction de ε

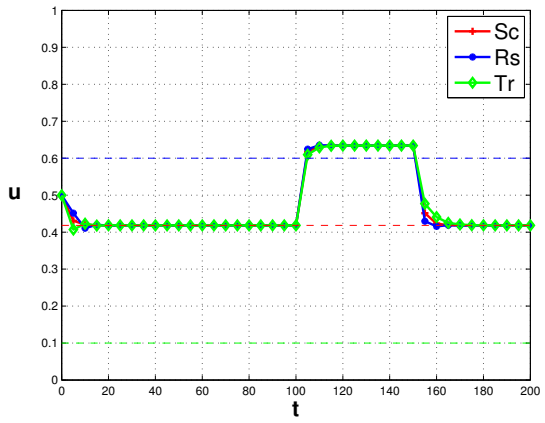


FIG. 8.13: Solution pour $\lambda = 2$ et $\varepsilon = 6.3 \times 10^{-2}$

Chapitre 9

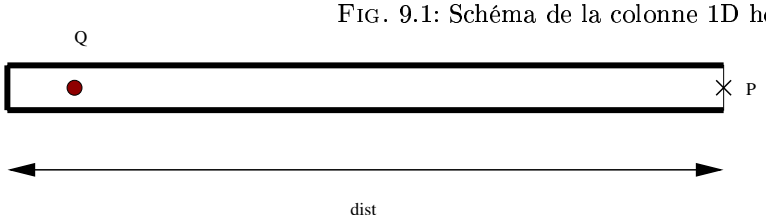
Comparaison de deux algorithmes

Le but de chapitre est de comparer les résultats obtenus avec les deux types de schéma sur un cas simple. Dans un premier temps, on souhaite étudier de façon qualitative l'aspect des solutions. Puis on étudie l'erreur entre les solutions lorsque le pas de temps diminue ainsi que le pas d'espace.

9.1 Description générale

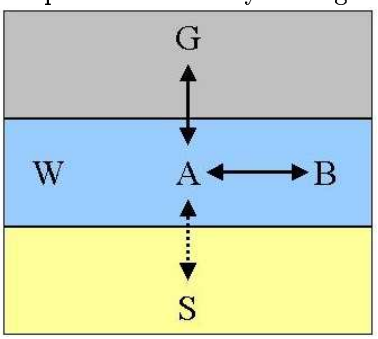
Pour faire cette étude, on définit un cas 1D avec une chimie simplifiée. On modélise un écoulement diphasique réactif dans un aquifère horizontal de longueur $L = 1000$ mètres ((9.1)). On injecte du dioxyde de carbone avec un débit constant égal à $0,01\text{m}^3$ par jour, en condition de fond, pendant cinq ans entre les dates $t = 20$ ans et $t = 25$ ans. On observe l'évolution du système jusqu'à la date $t = 300$ ans.

FIG. 9.1: Schéma de la colonne 1D horizontale



Le système géochimique défini pour cette étude n'est pas réaliste mais suffisant pour étudier les effets du découplage des différents phénomènes. Il est représenté sur la figure (9.2).

FIG. 9.2: Représentation du système géochimique



La description complète du cas étudié est en annexe C.

9.2 Les résultats

Le cas de la colonne 1D-CO_2 est utilisé pour faire une comparaison entre le schéma Sc, le schéma (Rs-Tr) et le schéma (Rs-Tr) avec itérations.

9.2.1 Correspondance des résultats

Dans un premier temps, on compare qualitativement (figure 9.3) le profil de la fraction molaire de co_2^a, à $t = 22.5$ ans, c'est à dire au milieu de l'injection et à $t = 50$ ans, c'est à dire 25 ans après la fin de l'injection. On observe que les résultats obtenus avec une résolution couplée et avec une résolution

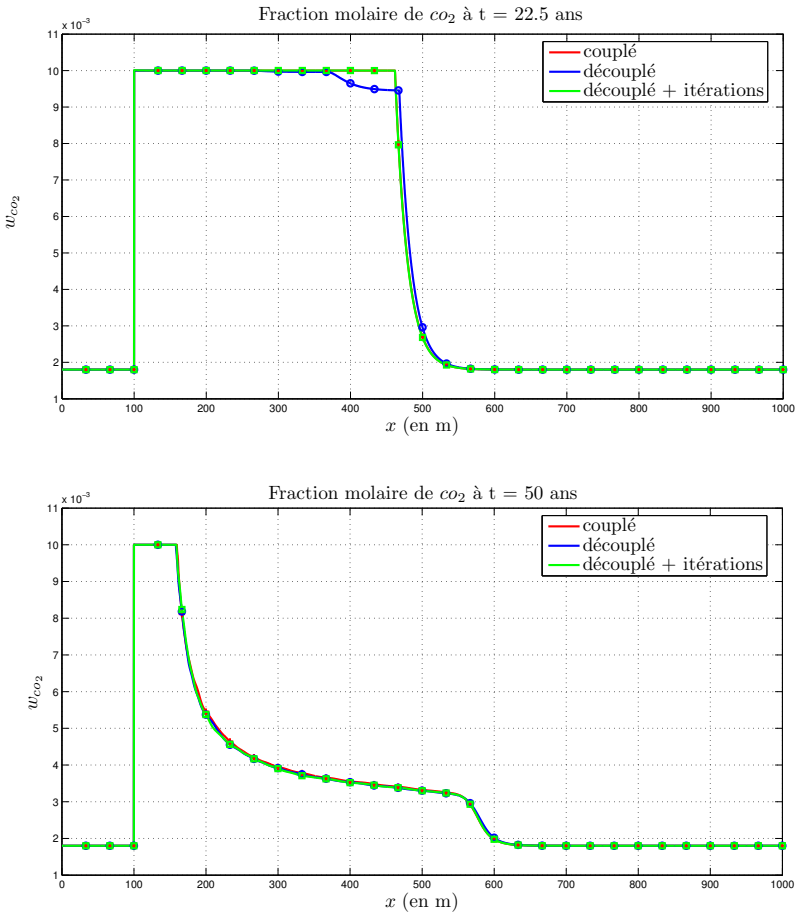


FIG. 9.3: Profil de la fraction molaire de $co2^a$

découplée avec itérations sont proches. En revanche, les résultats obtenus avec une résolution découplée sans itération sont légèrement différents. On observe comme prévu un "temps de retard" sur le profil (cf. figure 9.3(a)) visible au niveau de l'avancement du front de dissolution mais ce retard n'a pas d'influence notable sur le résultat final.

9.2.2 Calcul d'erreur

Pour analyser de façon plus précise la qualité des résultats, on mesure l'erreur entre une solution de référence et la solution obtenue pour différents pas de maillage.

En absence de solution analytique à ce problème, la solution de référence est une solution numérique obtenue par la méthode couplée avec un maillage fin de 6000 mailles. Cette solution est notée $\widetilde{u}$.

On définit l'erreur d'approximation à la date t, en norme L1 discrète sur la variable u, de la manière suivante :

$$\begin{aligned} E(t) &= \int_0^L |u(x,t) - \widetilde{u}(x,t)| \mathrm{d}x \\ &= \sum_{i=1}^{N} \left| u_i^n - \widetilde{u_i^n} \right| \end{aligned} \tag{9.1}$$

où n est choisi de manière à ce que $t \in \left[n\delta t^n, (n+1)\delta^{n+1} \right[$.

L'étude ne sera présentée que sur une variable, à savoir la fraction molaire du co_2^a dans l'eau à deux dates fixées : $t = 22.5$ ans et $t = 50$ ans. Néanmoins, le comportement observé est similaire pour les autres variables, et les deux dates choisies sont représentatives des différents comportements au cours de la simulation. En effet, à $t = 22.5$ ans l'injection est active (avancée rapide des fronts), à $t = 50$ ans l'injection est stoppée (phénomènes plus lents).

Les résultats obtenus sont reportés sur les figures ci dessous.

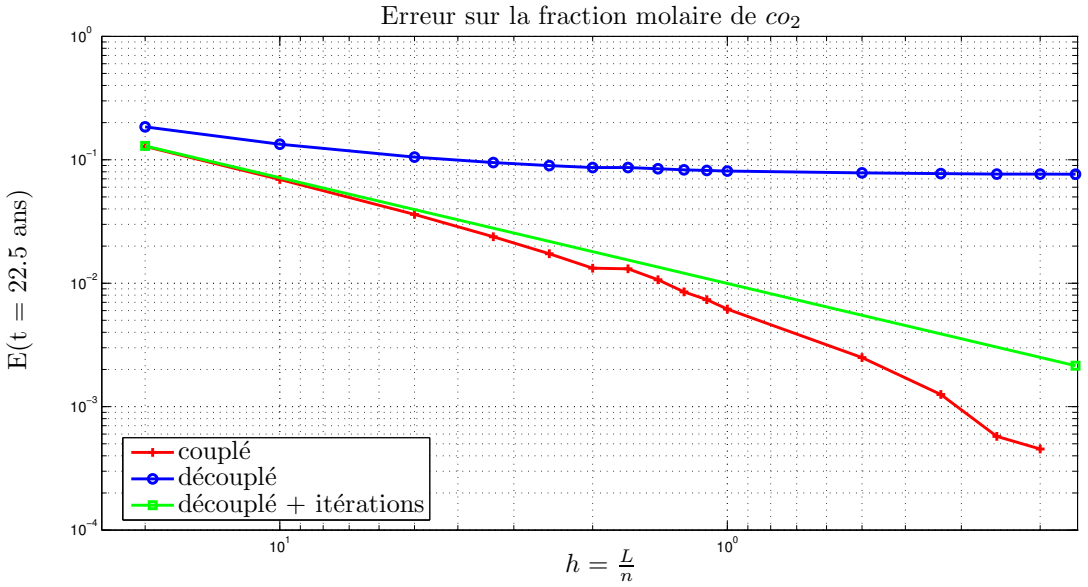


FIG. 9.4: Erreur en fonction du pas d'espace à t = 22.5 ans

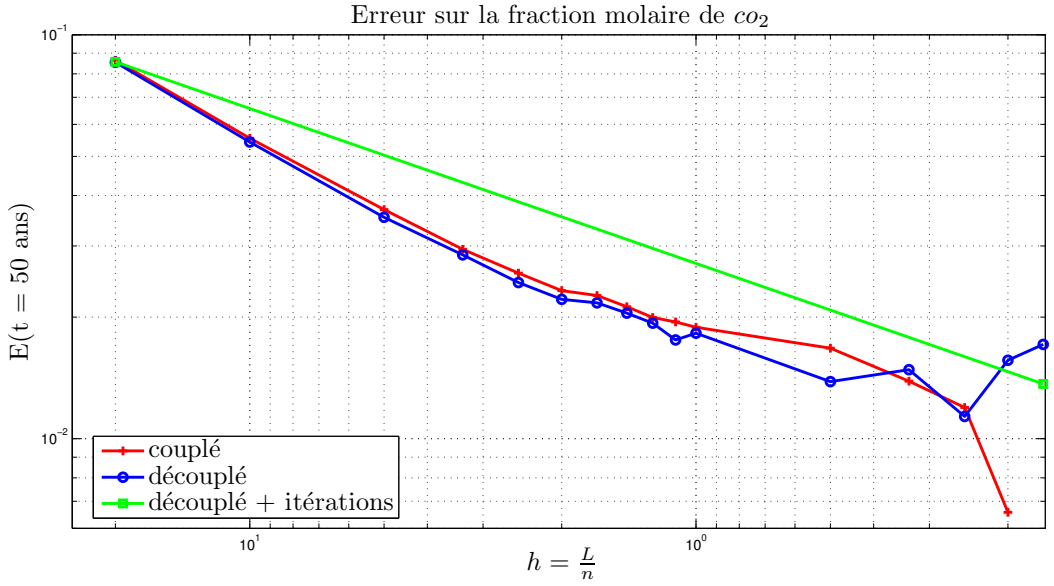


FIG. 9.5: Erreur en fonction du pas d'espace à t = 50 ans

On remarque que les courbes d'erreur correspondant au cas découplé sans itération ne tendent pas vers zéro comme on pourrait s'y attendre. En effet à un temps t fixé (lorsque le front de $co2^a$ avance), il existe toujours un décalage entre les solutions (cf. figure 9.3(a)). Raffiner le maillage n'a aucun effet sur cette erreur qui est due à un décalage en temps.

Comme le schéma est un splitting en temps, il est plus intéressant de s'intéresser à la convergence lorsqu'on diminue le pas de temps. Cependant, le pas de temps n'est pas constant au cours de la simulation. Il est géré de façon adaptative selon un paramètre p_{obj} et les variations constatées des variables au cours du pas de temps. Le principe est le suivant :

1. Pour chaque variable i, on calcule la variation relative, notée δ_i, par rapport à une valeur de référence $\overline{x_i}$, au cours du pas de temps.

$$\delta_i = \max_K \left(100 \frac{(x_K^{n+1})_i - (x_K^n)_i}{\overline{x_i}} \right)$$

2. On prend le pourcentage de variation maximum $\Delta = \max_i \delta_i$.
3. On estime un nouveau pas de temps selon la valeur de Δ
   - Premier cas : $\Delta < p_{\text{obj}} \quad \Longrightarrow$ on augmente le pas de temps

$$\delta t^{n+2} = \delta t^{n+1} \times \max\left(\frac{p_{\text{obj}}}{\Delta}, 1.8\right) \tag{9.2}$$

   - Deuxième cas : $p_{\text{obj}} < \Delta < 2p_{\text{obj}} \quad \Longrightarrow$ on ne fait rien

$$\delta t^{n+2} = \delta t^{n+1} \tag{9.3}$$

   - Troisième cas : $\Delta > 2p_{\text{obj}} \quad \Longrightarrow$ on diminue le pas de temps et on le refait

$$\delta t^{n+1} = 2\delta t^{n+1} \frac{p_{\text{obj}}}{\Delta} \tag{9.4}$$

Pour étudier la convergence en temps, le plus simple serait de prendre un pas de temps constant. Néanmoins, ceci n'est pas vraiment envisageable, car la simulation nécessite parfois de petits pas de temps (au début de l'injection par exemple) et parfois aussi de très grands pas de temps. Par conséquent, une étude avec un pas de temps constant serait très restrictive et peu représentative. On a donc choisit d'étudier la convergence en diminuant le paramètre p_{obj} (avec un maillage en espace constant de 1000 mailles).

Pour calculer l'erreur, on utilise une norme L^2 en espace par rapport à une solution de référence, et on s'intéresse à deux temps fixés : $t = 22.5$ ans et $t = 50$ ans. La solution de référence est réalisée avec le schéma couplé pour un maillage de 1000 mailles et $p_{\text{obj}} = 0.4\%$.

On constate tout d'abord que l'erreur diminue bien, avec les deux schémas, lorsque la contrainte sur le pas de temps est plus stricte (cf figures 9.2.2 et 9.2.2). L'erreur obtenue avec le schéma (*Rs-Tr*) décroît moins vite que celle obtenue avec le schéma couplé. Néanmoins ce comportement parait logique au vu de la méthode utilisée pour le calcul de l'erreur. La différence entre l'erreur obtenue avec les deux schémas est plus faible pour $t = 50$ ans que pour $t = 22.5$ ans. Ceci s'explique par la physique du problème qui est différente à ces deux dates. En effet, à $t = 22.5$ ans l'injection est active et par conséquent, il existe des fronts de concentration qui se déplacent vite. A $t = 50$ ans l'injection est stoppée depuis 25 ans, les phénomènes sont donc plus lents, et donc les résultats sont moins sensibles au splitting.

Pour avoir une idée de la précision obtenue par rapport à une solution fine, on réalise une interpolation linéaire sur l'erreur dans chacun des cas. Pour la méthode couplée, on obtient pour $t = 22.5$ ans, une pente de 2.4 et pour $t = 50$ ans, une pente de 2.1. Pour la méthode (*Rs-Tr*), on obtient pour $t = 22.5$ ans, une pente de 0.91 et pour $t = 50$ ans, une pente de 1.4. Ceci permet d'avoir une indication pour le choix du paramètre p_{obj} selon la précision voulue.

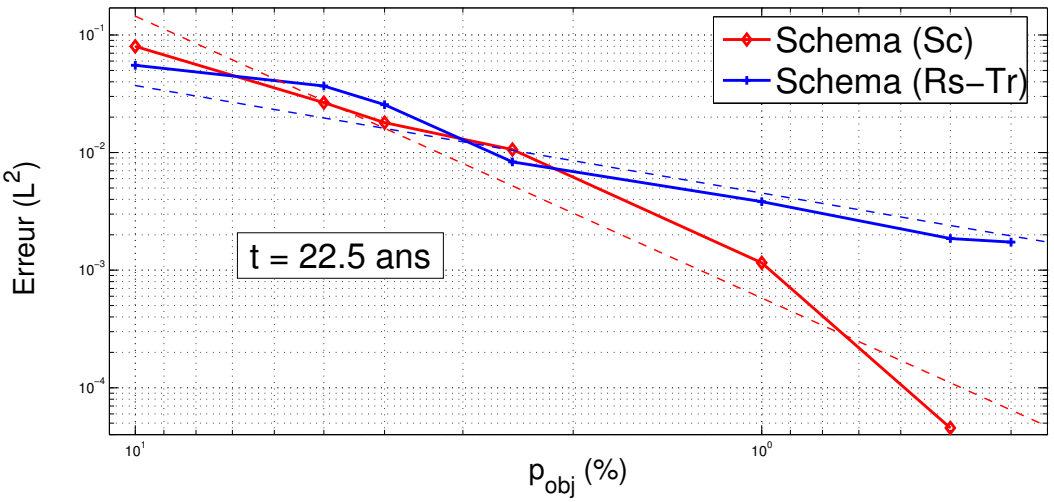


FIG. 9.6: Erreur pour $t = 22.5$ ans sur la fraction molaire de $co2^a$

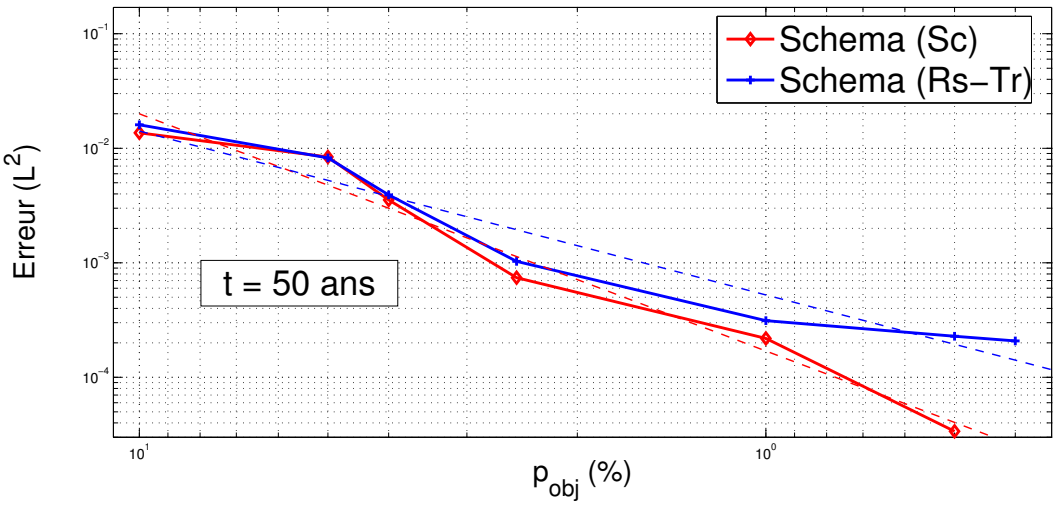


FIG. 9.7: Erreur pour $t = 50$ ans sur la fraction molaire de $co2^a$

Pour se rendre compte de l'effet de la diminution du pas de temps sur les résultats obtenus avec le schéma (*Rs-Tr*), il est intéressant de regarder comment évolue le décalage en temps observé sur le profil de la fraction molaire de $co2^a$ (cf. figure 9.2.2). On constate que ce décalage en temps se réduit progressivement lorsqu'on diminue le paramètre p_{obj}, pour finalement obtenir un profil qui approxime correctement le front.

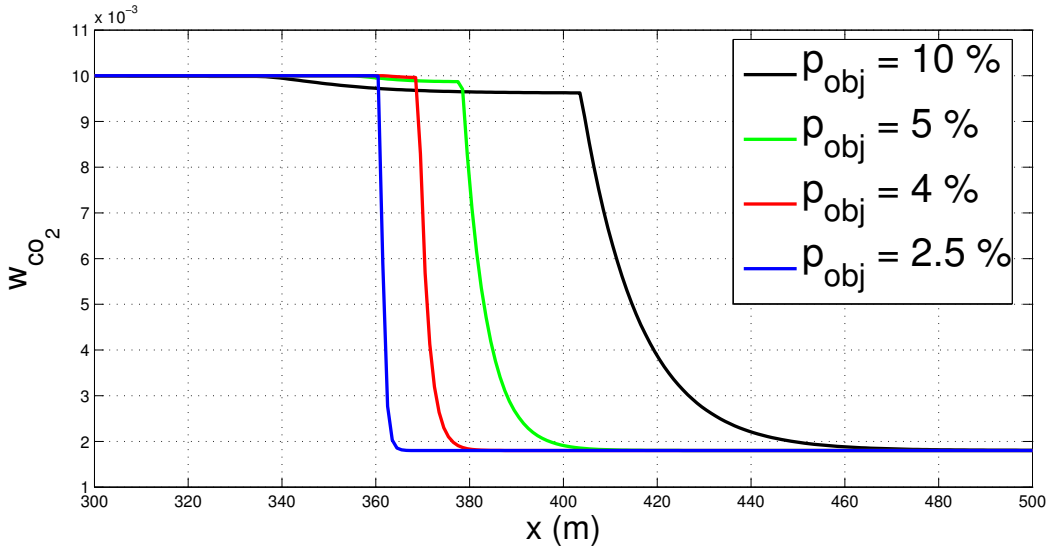


FIG. 9.8: Profil du $co2^a$ autour du front selon p_{obj}

Finalement, il faut maintenant s'intéresser à l'augmentation du coût numérique lorsque p_{obj} diminue. On constate que le nombre de pas de temps nécessaire pour réaliser la simulation complète augmente de façon exponentielle lorsque p_{obj} diminue (*cf.* figure 9.2.2). Il ne semble pas raisonnable de prendre une valeur inférieure à 2.

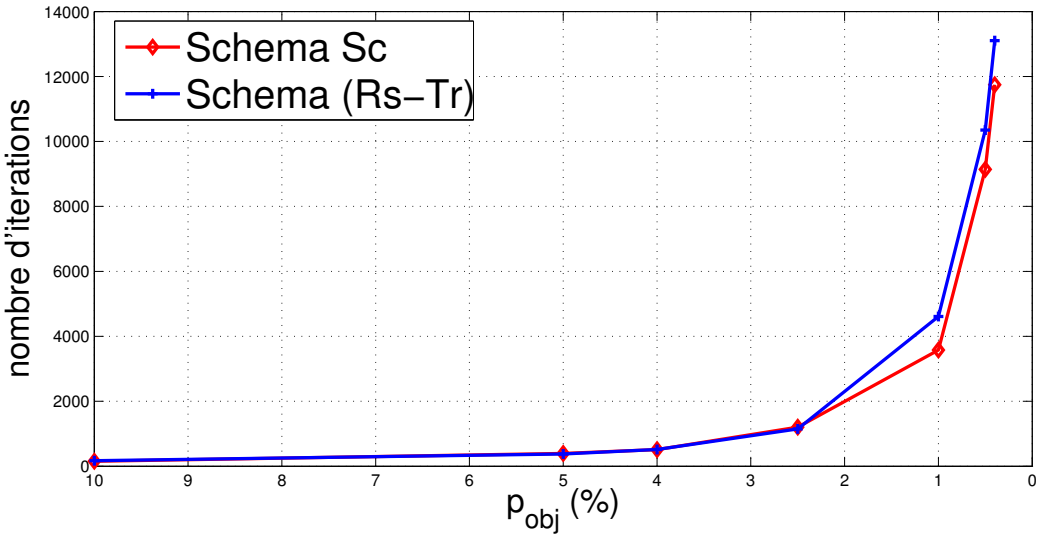


FIG. 9.9: Évolution du nombre de pas de temps selon p_{obj}

Chapitre 10

La dispersion

La diffusion moléculaire et la dispersion sont deux phénomènes qui tendent à réduire les gradients de concentration des espèces en solution. L'un des objectifs de la thèse est de pouvoir inclure la diffusion-dispersion dans le schéma (*Rs-Tr*), tout en conservant une résolution locale de (*Tr*). En effet, dans certains cas, si l'écoulement n'est pas purement convectif, il est nécessaire de traiter de façon particulière le terme de diffusion-dispersion si l'on souhaite résoudre le schéma (*Tr*) de façon locale. La première étape consiste donc à étudier, sur un cas représentatif d'une étude de stockage, l'ordre de grandeur des flux de diffusion-dispersion par rapport aux flux convectifs.

La diffusion moléculaire dépend du constituant et de la phase fluide qui le contient. La dispersion dépend essentiellement du milieu poreux lui-même.

On distingue la micro-dispersion et la macro-dispersion. La micro-dispersion est lié au changement d'échelle effectué lors de la modélisation. En effet, la loi de Darcy est une loi "moyenne", sur un VER (volume élémentaire représentatif), considéré homogène et isotrope. En réalité, les particules ne se déplacent pas toutes dans la même direction à la même vitesse. Certaines peuvent emprunter des chemins différents, et les frictions avec les grains de minéral peuvent ralentir le flux. La micro-dispersion traduit en fait l'incapacité de suivre en détail le mouvement du fluide. La macro-dispersion est liée à l'hétérogénéité du milieu poreux à l'échelle macroscopique.

Dans la suite, on limite notre étude à la macro-dispersion.

Le flux diffusif et dispersif s'exprime par une loi de Fick

$$J_i = \xi_{\alpha(i)} D_i \nabla x_i \tag{10.1}$$

où D_i est le tenseur de diffusion-dispersion composé d'un terme de diffusion moléculaire D_i^m et d'un terme de dispersion D_i^v.

$$D_i = D_i^m + D_i^v \tag{10.2}$$

Le tenseur de diffusion moléculaire s'écrit

$$D_i^m = \Phi S_{\alpha(i)} D_i^0 \, Id \tag{10.3}$$

et le tenseur de dispersion (non diagonal)

$$D_i^v = d_T \, ||v|| \, Id + (d_L - d_T) \, \frac{v \cdot v^T}{||v||} \tag{10.4}$$

Les coefficients d_L et d_T sont les coefficients de dispersion longitudinale et transversale, exprimés en m. Il est difficile d'estimer ces coefficients. Des relations empiriques ont été déduites à partir de mesures expérimentales (Neumann (1990)). Ces relations établissent que les coefficients de dispersion dépendent de l'échelle caractéristique du problème considéré. Le caractère universel de ces lois peut évidemment être remis en cause (Oelkers (1996)), notamment en raison du manque de données expérimentales sur certains types de roche.

La loi utilisée ici (Neumann (1990)) est

$$\begin{array}{l} d_L = \left\{ \begin{array}{ll} 0.0169\ L^{1.69} & \text{si } L \leq 100m \\ 0.32\ L^{0.83} & \text{sinon} \end{array} \right. \\ d_T = 0.33\ d_L \end{array} \tag{10.5}$$

L'échelle caractéristique du problème est une grandeur difficile à choisir. Dans le cas d'une injection de traceur, il peut s'agir de la distance entre le point d'injection et le point d'observation du traceur. Face à cette difficulté, on choisira pour notre étude une valeur moyenne, une valeur maximum et une valeur minium afin de recouvrir la gamme des valeurs possibles. Cette étude permettra de mesurer l'influence de ce paramètre sur les résultats.

L'objectif est de mesurer à l'aide d'un cas test, l'importance du flux diffusif, du flux dispersif et du flux convectif. Si l'écoulement est majoritairement convectif, le traitement de la diffusion-dispersion ne pose pas de problème. Mais si les phénomènes de diffusion-dispersion sont du même ordre de grandeur que la convection, il devient nécessaire de s'intéresser à la prise en compte de ce terme.

Premièrement, la dispersion s'effectue dans la direction de la vitesse ou dans sa direction normale, qui ne suit pas forcément les axes du maillage. On cherche donc à savoir s'il est nécessaire d'utiliser des schémas plus précis que les schémas classiques.

Deuxièmement, pour utiliser la méthode des caractéristiques (cf. 5.5), il faut que la diffusion-dispersion puisse être traitée de manière explicite sans contraindre le pas de temps de façon excessive.

Les objectifs de ce chapitre sont :

- Étudier la nécessité d'utiliser un schéma particulier pour la discrétisation de terme de diffusion-dispersion par rapport à un schéma deux points classiques,
- Évaluer l'importance de la diffusion et de la dispersion par rapport à la convection,
- Étudier la possibilité de prendre en compte ce terme dans le schéma découplé avec résolution de ($\mathcal{T}r$) par la méthode des caractéristiques.

10.1 Un schéma pour le terme de diffusion-dispersion

On choisit d'utiliser le schéma VF gradient (Eymard et al. (2006)) pour la discrétisation du terme de diffusion-dispersion. Ce schéma a été conçu pour discrétiser les termes de la forme $\text{div}\,(\Lambda \nabla u)$, où Λ est une matrice non diagonale.

Il nécessite de décomposer la matrice Λ sous la forme

$$\Lambda = (\Lambda - \alpha\, Id) + \alpha\, Id \tag{10.6}$$

pour assurer les conditions nécessaires à la convergence.

Le tenseur de diffusion-dispersion est naturellement sous cette forme, on a

$$\Lambda = \alpha\, Id + \overline{\Lambda} \tag{10.7}$$

avec $\alpha = \Phi S_w D_i^0 + d_T\, ||v||$ et $\overline{\Lambda} = (d_L - d_T)\, \frac{v \cdot v^T}{||v||}$.

10.1.1 Notations

Soit $\mathcal{T}$ un maillage admissible au sens donné dans Eymard et al. (2000). On note $\mathcal{E}$ l'ensemble des faces, $\mathcal{E}_{ext}$ l'ensemble des faces externes et $\mathcal{E}_K$ l'ensemble des faces d'une maille K. Le centre d'une maille K est noté x_K, et la mesure de cette maille s'écrit m_K. On note $K|L$ l'interface entre les mailles K et L. La mesure de l'interface $K|L$ est notée $m_{K|L}$. La distance entre les centres d'une maille K et d'une maille L est notée $d_{K|L}$. Le centre de gravité d'une interface $K|L$ est noté $x_{K|L}$. La transmissivité d'une face interne (resp. externe) s'écrit $\tau_{K|L} = \frac{m_{K|L}}{d_{K|L}}$ (resp $\tau_\sigma = \frac{m_\sigma}{d_\sigma}$.

10.1.2 Gradient discret

Le schéma VF gradient utilise un gradient discret par maille, défini par

$$(\nabla_{\mathcal{D}} u)_K = \frac{1}{m(K)} \left(\sum_{L \in \mathcal{N}_K} A_{K,L} \left(u_L - u_K\right) + \sum_{\sigma \in \mathcal{E}_{K,\text{ext}}} A_{K,\sigma} \left(u_\sigma - u_K\right) \right) \tag{10.8}$$

où $A_{KL} = \tau_{K|L} \left(x_{K|L} - x_K\right)$. En posant

$$\gamma_\sigma u = \begin{cases} \dfrac{d_{L,\sigma} u_K + d_{K,\sigma} u_L}{d_{L,\sigma} + d_{K,\sigma}} & \text{si } \sigma = K|L \\ u_\sigma & \text{si } \sigma \in \mathcal{E}_K \cap \mathcal{E}_{ext} \end{cases} \tag{10.9}$$

le gradient discret peut également s'écrire

$$(\nabla_{\mathcal{D}} u)_K = \frac{1}{m(K)} \left(\sum_{\sigma \in \mathcal{E}_K} m_\sigma (\gamma_\sigma u - u_K) \frac{x_\sigma - x_K}{d_{K,\sigma}} \right) \tag{10.10}$$

Le gradient discret ainsi défini est fortement et faiblement convergent ainsi que consistant.

10.1.3 Le schéma VF gradient - forme 1

Le produit scalaire s'écrit

$$< u, v >_{\mathcal{D},\alpha} = \sum_{K \in \mathcal{T}} \left(\begin{array}{l} m_K \left(\nabla u_K\right) \left(\Lambda_K - \alpha_K Id\right) \left(\nabla v_K\right) + \frac{1}{2} \sum\limits_{L \in \mathcal{N}(K)} \alpha_{K|L} \tau_{K|L} (u_L - u_K)(v_L - v_K) \\ + \sum\limits_{\sigma \in \mathcal{E}_K} \alpha_\sigma \tau_{K|\sigma} (u_\sigma - u_K)(v_\sigma - v_K) \end{array} \right) \tag{10.11}$$

On peut le réécrire de la façon suivante

$$< u, v >_{\mathcal{D},\alpha} = \sum_{K \in \mathcal{T}} \left(\begin{array}{l} m_K \left(\nabla u_K\right) \Lambda_K \left(\nabla v_K\right) - \alpha_K m_K \left(\nabla u_K\right) \left(\nabla v_K\right) \\ + \left(\sum\limits_{\sigma \in \mathcal{E}(K)} \alpha_{K|\sigma} \frac{m_\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)(\gamma_\sigma v - v_K) \right) \end{array} \right) \tag{10.12}$$

Pour obtenir le flux sur une interface, il suffit de factoriser le produit scalaire par $(v_K - v_L)$. On obtient

$$< u, v >_{\mathcal{D},\alpha} = \sum_{\sigma = K|L \in \mathcal{E}} F_{K|L} \left(v_K - v_L\right) + \sum_{\sigma \in \mathcal{E}_{ext}} F_{K|\sigma} \left(v_K - v_\sigma\right) \tag{10.13}$$

En effectuant la factorisation, on obtient que

$$F_{K|L} = \alpha_{K|L} \tau_{K|L} (u_K - u_L) + \left(\nabla u_L\right) \left(\Lambda_L - \alpha_L Id\right) A_{L,K} - \left(\nabla u_K\right) \left(\Lambda_K - \alpha_K Id\right) A_{K,L} \tag{10.14}$$

et

$$F_{K,\sigma} = \alpha_\sigma \tau_{K|\sigma} (u_K - u_\sigma) - \left(\nabla u_K\right) \left(\Lambda_K - \alpha_K Id\right) A_{K,\sigma} \tag{10.15}$$

10.1.4 Le schéma VF gradient - forme 2

Le schéma donné précédemment correspond au schéma donné dans Eymard et al. (2006). Il existe une autre façon d'écrire le schéma, équivalente pour les maillages vérifiant une certaine condition, mais qui a l'avantage d'être coercif quelque soit la valeur prise pour $\alpha \geq 0$. Dans ce cas, le schéma est défini à partir du produit scalaire suivant

$$< u, v >_{\mathcal{D},\alpha} = \sum_{K \in \mathcal{T}} \left(m_K \left(\nabla u_K\right) \Lambda_K \left(\nabla v_K\right) + \sum_{\sigma \in \mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma} u R_{K,\sigma} v \right) \tag{10.16}$$

avec

$$R_{K,\sigma}u = \alpha_\sigma \left(\frac{\gamma_\sigma u - u_K}{d_{K,\sigma}} - (\nabla u_K) \frac{x_\sigma - x_K}{d_{K,\sigma}} \right) \tag{10.17}$$

Les deux écritures sont équivalentes si le maillage vérifie la condition suivante

$$n_{K,\sigma} = \frac{x_\sigma - x_K}{d_{K,\sigma}} \tag{10.18}$$

et si

$$\alpha_{K,\sigma} = \sqrt{\alpha d} \tag{10.19}$$

Un maillage cartésien vérifie la condition (10.18). Pour obtenir l'équivalence, il faut développer le produit scalaire (10.16). On obtient

$$\begin{array}{l} \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma}u R_{K,\sigma}v = \\ \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} (\alpha_{K,\sigma})^2 \left(\frac{\gamma_\sigma u - u_K}{d_{K,\sigma}} - (\nabla u_K) \frac{x_\sigma - x_K}{d_{K,\sigma}} \right) \left(\frac{\gamma_\sigma v - v_K}{d_{K,\sigma}} - (\nabla v_K) \frac{x_\sigma - x_K}{d_{K,\sigma}} \right) \end{array} \tag{10.20}$$

et donc

$$\sum_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma}u R_{K,\sigma}v = \left(\begin{array}{l} \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d d_{K,\sigma}} (\alpha_{K,\sigma})^2 (\gamma_\sigma u - u_K)(\gamma_\sigma v - v_K) \\ - (\nabla u_K) \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d} (\alpha_{K,\sigma})^2 (\gamma_\sigma v - v_K) n_{K,\sigma} \\ - (\nabla v_K) \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d} (\alpha_{K,\sigma})^2 (\gamma_\sigma u - u_K) n_{K,\sigma} \\ + (\nabla u_K)^t \left(\sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} (\alpha_{K,\sigma})^2 n_{K,\sigma} (n_{K,\sigma})^t \right) (\nabla v_K) \end{array} \right) \tag{10.21}$$

De plus, d'après (10.19), on a

$$\sum_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma}u R_{K,\sigma}v = \alpha_K \left(\begin{array}{l} \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)(\gamma_\sigma v - v_K) \\ - (\nabla u_K) \sum\limits_{\sigma\in\mathcal{E}_K} m_\sigma (\gamma_\sigma v - v_K) n_{K,\sigma} \\ - (\nabla v_K) \sum\limits_{\sigma\in\mathcal{E}_K} m_\sigma (\gamma_\sigma u - u_K) n_{K,\sigma} \\ + (\nabla u_K)^t \left(\sum\limits_{\sigma\in\mathcal{E}_K} m_\sigma (x_\sigma - x_K)(n_{K,\sigma})^t \right) (\nabla v_K) \end{array} \right) \tag{10.22}$$

Or comme

$$\sum_\sigma m_\sigma (x_\sigma - x_K)(n_{K,\sigma})^t = m_K Id, \tag{10.23}$$

on obtient

$$\sum_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma}u R_{K,\sigma}v = \alpha_K \left(\begin{array}{l} \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)(\gamma_\sigma v - v_K) \\ - (\nabla u_K) \sum\limits_{\sigma\in\mathcal{E}_K} m_\sigma (\gamma_\sigma v - v_K) n_{K,\sigma} \\ - (\nabla v_K) \sum\limits_{\sigma\in\mathcal{E}_K} m_\sigma (\gamma_\sigma u - u_K) n_{K,\sigma} \\ + m_K (\nabla u_K)^t (\nabla v_K) \end{array} \right) \tag{10.24}$$

D'après (10.18) et (10.10), on a

$$\sum_{\sigma\in\mathcal{E}_K} \frac{m_\sigma d_{K,\sigma}}{d} R_{K,\sigma}u R_{K,\sigma}v = \alpha_K \left(\begin{array}{l} \sum\limits_{\sigma\in\mathcal{E}_K} \frac{m_\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)(\gamma_\sigma v - v_K) \\ -2 m_K (\nabla u_K) (\nabla v_K) \\ + m_K (\nabla u_K)^t (\nabla v_K) \end{array} \right) \tag{10.25}$$

et donc

$$\sum_{\sigma\in\mathcal{E}_K}\frac{m_\sigma d_{K,\sigma}}{d}R_{K,\sigma}uR_{K,\sigma}v=\alpha_K\left(\sum_{\sigma\in\mathcal{E}_K}\left(\frac{m_\sigma}{d_{K,\sigma}}(\gamma_\sigma u-u_K)(\gamma_\sigma v-v_K)\right)-m_K\left(\nabla u_K\right)\left(\nabla v_K\right)\right) \quad (10.26)$$

Le produit scalaire (10.16) s'écrit finalement

$$<u,v>_{\mathcal{D},\alpha}=\sum_{K\in\mathcal{T}}\left(m_K\left(\nabla u_K\right)\Lambda_K\left(\nabla v_K\right)-\alpha_K m_K\left(\nabla u_K\right)\left(\nabla v_K\right)+\alpha_K\sum_{\sigma\in\mathcal{E}_K}\frac{m_\sigma}{d_{K,\sigma}}(\gamma_\sigma u-u_K)(\gamma_\sigma v-v_K)\right) \quad (10.27)$$

On retrouve bien la forme précédente (10.12).

10.2 Le cas test étudié

Le cas étudié est un cas 2D. Afin de se rapprocher des contraintes des cas réalistes, on souhaite étudier un cas dans lequel la perméabilité n'est pas uniforme (mais constante) et on utilise également des perméabilités relatives particulières et des pressions capillaires.

La perméabilité est $K = 3000$ mD dans l'ensemble du domaine excepté sur une bande allant de $z = -870$ à $z = -880$ m, où $K = 10$ mD.

Le CO_2 est injecté pendant 10 ans, à raison de 0.02 millions de tonnes par an.

Le domaine (figure 10.1) a une longueur de 1 km, une profondeur de 100 m et une hauteur de 200 m. On discrétise le domaine avec un maillage $(83 \times 1 \times 26)$ non régulier, raffiné horizontalement à la verticale au dessus du point d'injection et raffiné verticalement en haut du domaine, et à proximité de la couche où la perméabilité est plus faible.

Les bords supérieur et inférieur sont imperméables. Ils sont situés, respectivement, à $z = -800$ m et $z = -1000$ m. Les bords verticaux ont une pression hydrostatique imposée (référence $P_w = 1$ bar pour $z = 0$), avec un ΔP de 0.02 bar sur le bord gauche afin de créer une vitesse de circulation dans l'aquifère. Ceci correspond à une vitesse de Darcy (horizontale) égale à 0.8 mm.j^{-1} dans la zone où la perméabilité est de 3000 mD et égale à 3×10^{-3} mm.j^{-1} dans la zone où la perméabilité est de 10 mD.

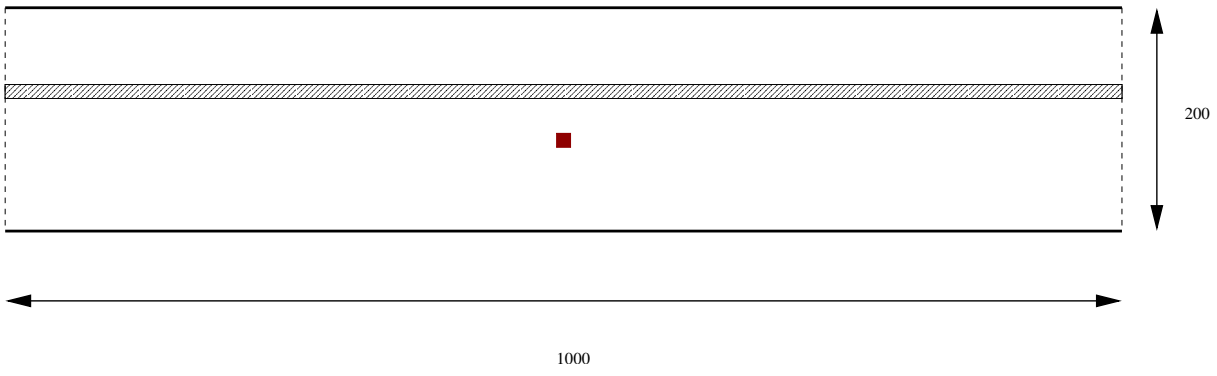


FIG. 10.1: Représentation du domaine géométrique

Les perméabilités relatives sont données par la figure 10.2. Les pressions capillaires sont différentes selon la valeur de la perméabilité (figure 10.3).

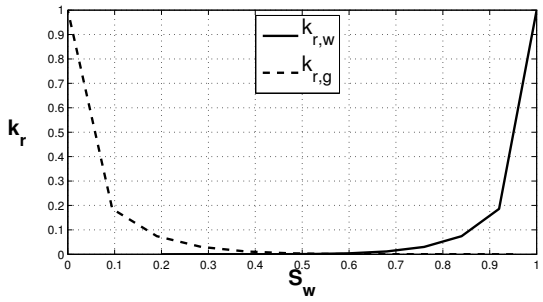


FIG. 10.2: Perméabilité relatives

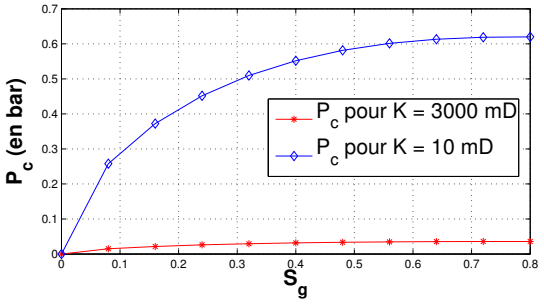


FIG. 10.3: Pression capillaire

Le but des simulations effectuées est d'étudier l'influence de la dispersion sur les résultats et d'évaluer son importance par rapport à la convection. Pour cela, on réalise des simulations avec et sans diffusion-dispersion afin de comparer les résultats. Les simulations avec dispersion seront réalisés avec 3 longueurs caractéristiques qui permettent de couvrir la gamme des valeurs possibles pour les coefficients de dispersion. Pour la valeur médiane, on réalisera les simulations avec un schéma classique et avec le schéma VF gradient (en essayant plusieurs valeurs du paramètre α).

La première remarque est que la diffusion moléculaire est négligeable par rapport à la dispersion. Par conséquent, la suite de l'étude est plutôt consacrée à la dispersion.

Pour comparer les résultats d'une simulation à l'autre, on compare l'évolution du bilan de carbone dans le système, et on effectue une comparaison qualitative des résultats à la fin de l'injection.

Le profil de la fraction molaire de co_2 et de la saturation de gaz donne une idée de l'influence de la dispersion sur les résultats. On constate que lorsqu'il y a de la dispersion, le profil de co_2 est beaucoup plus étalé (cf. figures 10.4 à 10.12). Le profil de la saturation de gaz est similaire avec ou sans dispersion pendant la période d'injection (cf. figures 10.13 à 10.15). En revanche, on constate qu'une plus grande quantité de gaz reste piégé dans la couche à perméabilité faible en l'absence de dispersion (cf. figures 10.16 à 10.18).

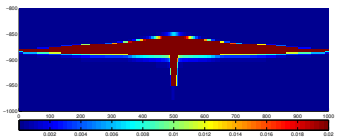


FIG. 10.4: Fraction molaire de co_2 sans dispersion à $t = 5$ ans

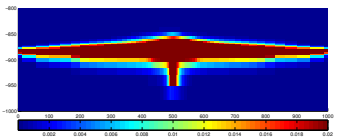


FIG. 10.5: Fraction molaire de co_2 avec dispersion à $t = 5$ ans

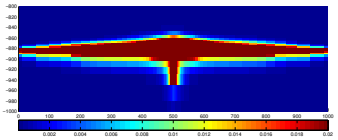


FIG. 10.6: Fraction molaire de co_2 avec dispersion à $t = 5$ ans, schéma VF gradient

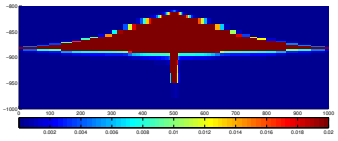


FIG. 10.7: Fraction molaire de co_2 sans dispersion à $t = 10$ ans

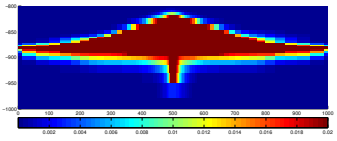


FIG. 10.8: Fraction molaire de co_2 avec dispersion à $t = 10$ ans

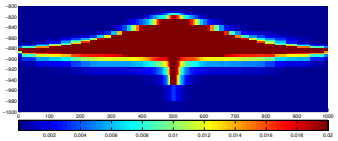


FIG. 10.9: Fraction molaire de co_2 avec dispersion à $t = 10$ ans, schéma VF gradient

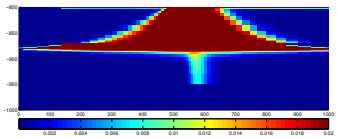


FIG. 10.10: Fraction molaire de co_2 sans dispersion à $t = 100$ ans

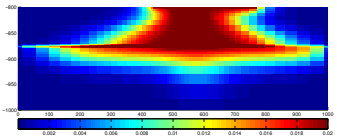


FIG. 10.11: Fraction molaire de co_2 avec dispersion à $t = 100$ ans

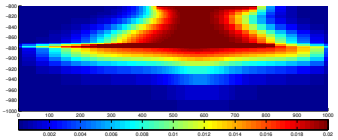


FIG. 10.12: Fraction molaire de co_2 avec dispersion à $t = 100$ ans, schéma VF gradient

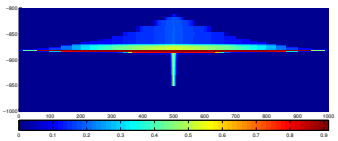


FIG. 10.13: Saturation de gaz sans dispersion à $t = 10$ ans

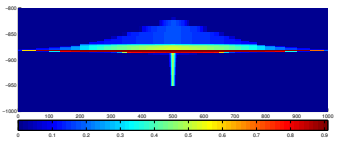


FIG. 10.14: Saturation de gaz avec dispersion à $t = 10$ ans

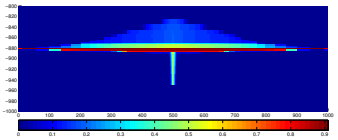


FIG. 10.15: Saturation de gaz avec dispersion à $t = 10$ ans, schéma VF gradient

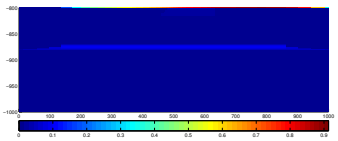


FIG. 10.16: Saturation de gaz sans dispersion à $t = 100$ ans

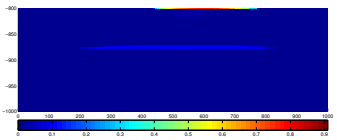


FIG. 10.17: Saturation de gaz avec dispersion à $t = 100$ ans

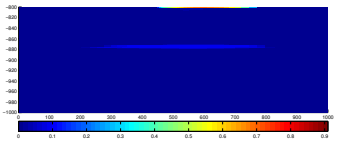


FIG. 10.18: Saturation de gaz avec dispersion à $t = 100$ ans, schéma VF gradient

On constate que le profil de co_2 obtenu dans le cas avec dispersion est similaire avec le schéma VF gradient et avec le schéma à deux points classique. Ceci s'explique par le fait que la vitesse de Darcy est orientée dans le sens des axes. Ce résultat peut être différent si l'on choisit mal la valeur du paramètre α. En effet, si on choisit une valeur trop grande, la dispersion est sur-estimée.

Afin d'évaluer l'influence de la dispersion sur les capacités de stockage, on mesure la quantité de carbone contenue dans le système sous ses différentes formes (dans l'eau, dans le gaz ou sous forme solide), la quantité injectée ainsi que la quantité sortie du domaine. (cf. figure 10.19 et figure 10.20). On constate que la dispersion modifie les bilans. Sans dispersion, à l'issue de la simulation 1.5 tonne de carbone est stockée dans l'eau, pour 2.2 tonnes avec dispersion (schéma classique).

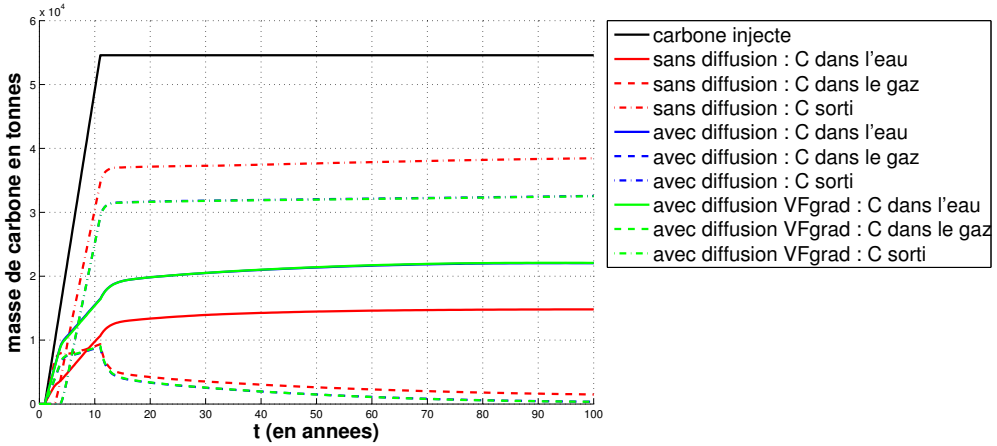


FIG. 10.19: Bilan de carbone (phases mobiles)

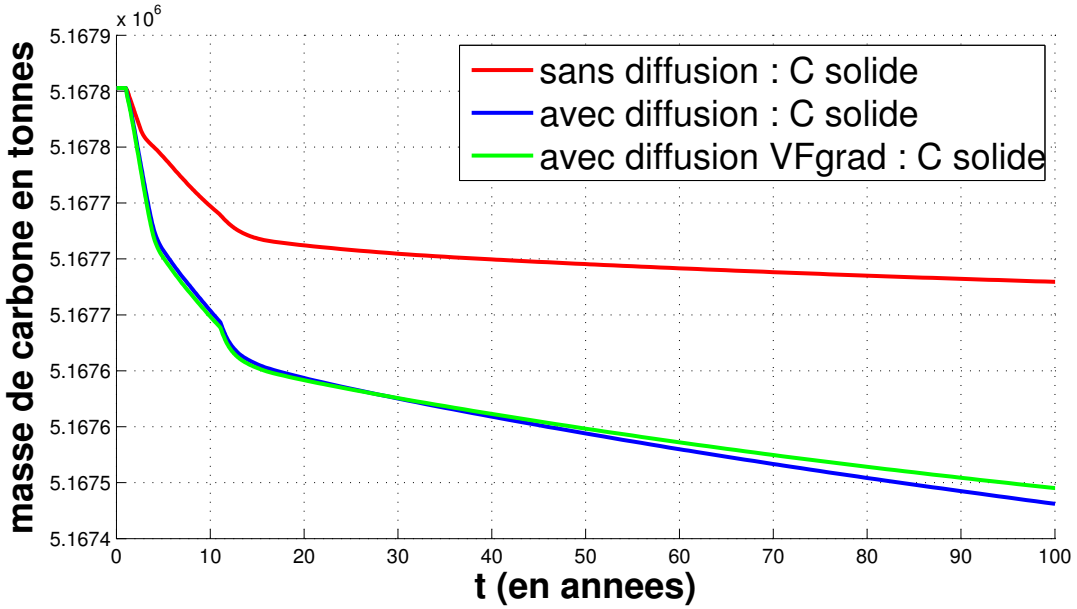


FIG. 10.20: Bilan de carbone (solide)

Pour étudier l'influence du choix de la longueur caractéristique, on réalise plusieurs simulations en faisant varier cette longueur. On étudie l'impact sur la quantité de carbone stockée dans l'eau. La longueur caractéristique varie de 1 m à 1000 m (cf. figure 10.21). Il n'est pas raisonnable d'aller au delà de 1000 m car c'est la longueur du domaine simulé. On constate que pour un facteur 1000 sur la longueur caractéristique, le bilan de carbone dans l'eau n'est modifié que d'un facteur 2 environ.

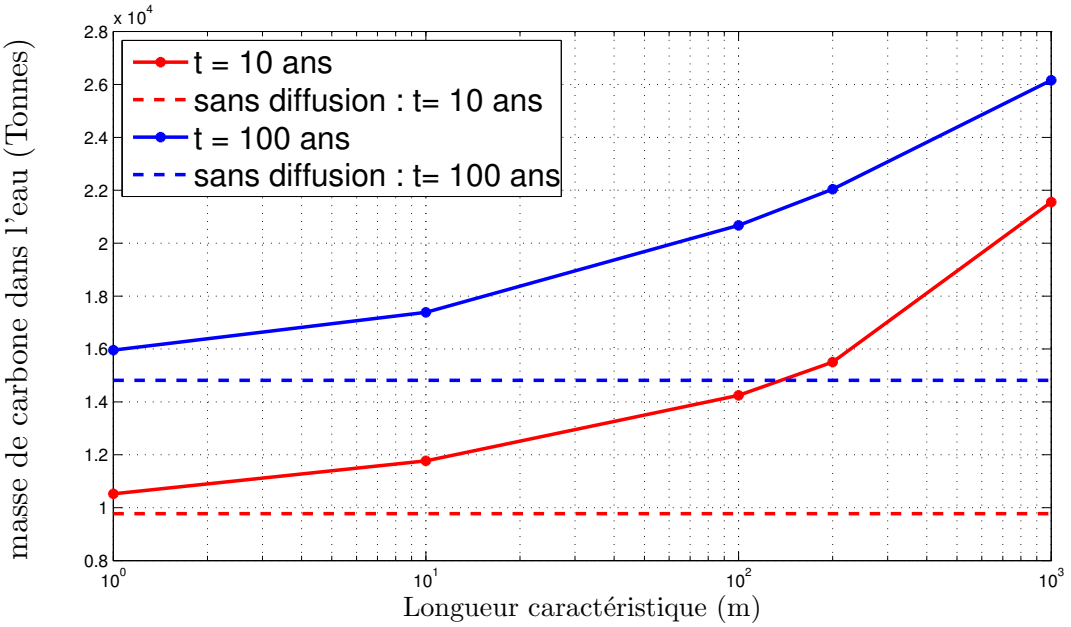


FIG. 10.21: Bilan de carbone en fonction de la longueur caractéristique

L'étude réalisée a permis de montrer l'importance du terme de dispersion. Étant parfois du même ordre de grandeur que le terme de convection, il parait difficile de pouvoir expliciter ce terme sans contraindre fortement le pas de temps. Si la contrainte sur le pas de temps est trop forte pour expliciter la dispersion, alors il n'est plus possible de résoudre le transport réactif de façon locale sans faire au préalable un splitting. Notamment, il n'est alors plus possible d'utiliser la méthode des caractéristiques (cf. 5.5). Pour vérifier ce point, on réalise la simulation avec la méthode (*Rs-Tr*), dans un premier temps, en implicitant la dispersion dans le modèle de transport réactif puis en l'explicitant. On constate alors que la contrainte sur le pas de temps du modèle de transport réactif est importante puisque le pas de temps moyen obtenu dans le cas implicite est de 3.75 jours contre 7.72×10^{-2} dans le cas explicite entre $t = 1$ ans et $t = 1.2$ ans (cf. figure 10.22). Ceci signifie que pour réaliser la simulation jusque $t = 1.2$ ans, il a fallu 963 pas de temps dans le cas explicite contre seulement 36 dans le cas implicite. Les pas de temps pour le modèle (*Rs*) sont du même ordre de grandeur. Ces résultats montrent qu'il est intéressant de regarder d'autres méthodes (par exemple une méthode avec splitting de la dispersion) afin de pouvoir à la fois tenir compte de la dispersion tout en résolvant le transport réactif de façon locale.

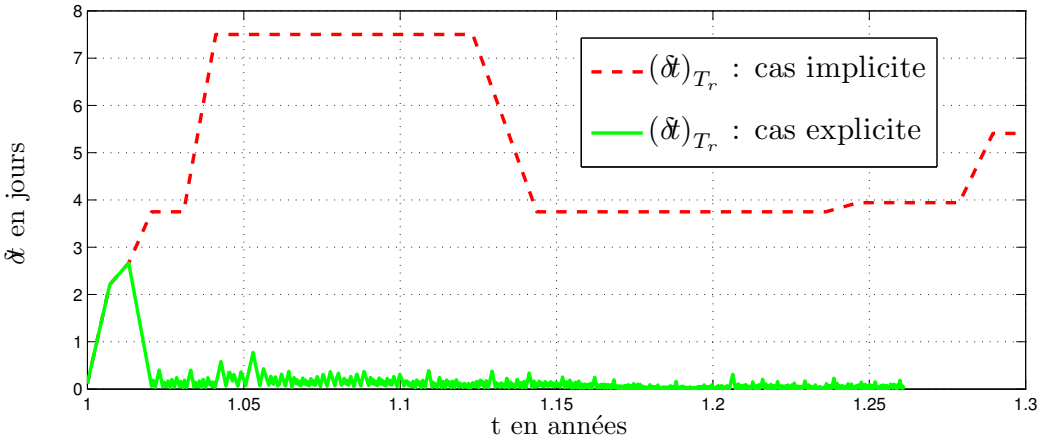


FIG. 10.22: Évolution du pas de temps de (Tr) (dispersion explicite ou implicite)

Quatrième partie

Analyse d'un problème simplifié

Le chapitre suivant est constitué d'un article écrit avec Robert Eymard, publié dans la revue "Journal of Equation Evolution" (Eymard and Tillier (2007)). Il s'agit de l'étude d'un système hyperbolique modélisant l'écoulement de fluides miscibles incompressibles dans un milieu poreux. Les différentes situations physiques rencontrées au cours de cette thèse nous ont amené à réfléchir à un tel problème.

Plus précisément, on étudie la dissolution dans l'eau et la migration d'une bulle de CO_2 gazeux, dans un système 1D vertical. L'écriture des lois de conservation, de l'eau et du CO_2, permet d'obtenir un système hyperbolique particulier dont les inconnues sont la quantité totale de CO_2 (sous forme de gaz et dissous dans l'eau), notée u, et le gradient de pression, noté v. La quantité d'eau s'écrit alors comme une fonction non linéaire de la quantité de CO_2, notée $f(u)$. La difficulté du problème traité est que l'on doit respecter simultanément la conservation de l'eau et du co_2. Ceci signifie qu'il faut résoudre le problème de façon symétrique par rapport à u et $f(u)$.

Si l'on choisit une solubilité nulle du CO_2, on retrouve le cas habituel des écoulements immiscibles, traité de façon abondante dans la littérature. Les difficultés du problème étudié ici proviennent du fait que cette solubilité n'est pas nulle. En effet, lorsque la solubilité du CO_2 est nulle, la quantité d'eau s'écrit comme une fonction linéaire de la quantité de CO_2, et il est possible de résoudre le système d'équations de façon classique par élimination de l'inconnue correspondant au gradient de pression.

Le système hyperbolique étudié est particulier car il ne contient pas de dérivée en temps par rapport à la seconde inconnue. En fait, le problème n'est pas symétrique par rapport à ses deux inconnues, qui jouent un rôle très différent.

Nous donnons une définition d'une solution faible, inspirée de la condition de Liu pour les chocs admissibles, et des paires entropiques de Krushkov. Nous prouvons ensuite, dans le cas d'une généralisation naturelle du problème de Riemann, l'existence d'une solution faible. Cette propriété découle de l'existence d'une certaine fonction permettant d'obtenir simultanément une solution faible entropique classique pour chacune des deux équations de façon couplée. L'existence d'une telle fonction est démontrée à l'aide de la convergence d'un algorithme qui permet d'obtenir des fonctions convexes et concaves simultanées pour chacune des deux équations.

Nous prouvons ensuite l'existence d'une solution faible dans un cas général, grâce à la preuve de convergence d'un schéma volume fini. Le principe de ce schéma est de calculer le flux de Godunov numérique à chaque interface, tout en respectant la conservation simultanée des deux équations.

L'ensemble de ce travail n'a pas permis de prouver l'unicité de la solution, ni de traiter le problème des conditions aux limites. Le schéma proposé est en effet un schéma défini sur un domaine infini. Un travail commun avec Robert Eymard et Julien Vovelle est en cours sur la prise en compte des conditions limites dans le schéma numérique. Quelques pistes de réflexion sont données dans le second chapitre.

Chapitre 11

Mathematical and numerical study of a system of conservation laws

Abstract

The system of equations $(f(u))_t - (a(u)v + b(u))_x = 0$ and $u_t - (c(u)v + d(u))_x = 0$, where the unknowns u and v are functions depending on $(x,t) \in \mathbb{R} \times \mathbb{R}_+$, arises within the study of some physical model of the flow of miscible fluids in a porous medium. We give a definition for a weak entropy solution (u,v), inspired by the Liu condition for admissible shocks and by Krushkov entropy pairs. We then prove, in the case of a natural generalization of the Riemann problem, the existence of a weak entropy solution only depending on x/t. This property results from the proof of the existence, by passing to the limit on some approximations, of a function g such that u is the classical entropy solution of $u_t - ((cg+d)(u))_x = 0$ and simultaneously $w = f(u)$ is the entropy solution of $w_t - ((ag+b)(f^{(-1)}(w)))_x = 0$. We then take $v = g(u)$, and the proof that (u,v) is a weak entropy solution of the coupled problem follows from a linear combination of the weak entropy inequalities satisfied by u and $f(u)$. We then show the existence of an entropy weak solution for a general class of data, thanks to the convergence proof of a coupled finite volume scheme. The principle of this scheme is to compute the Godunov numerical flux with some interface functions ensuring the symmetry of the finite volume scheme with respect to both conservation equations.

11.1 Introduction

The modelization of the injection of carbon dioxide (CO_2) into natural underground reservoirs, which seems to be a solution to some environmental problems, involves an increasing number of works (see Bachu et al. (2005); Brosse et al. (2002); Le Gallo et al. (2002); Holstad (2000); Lagneau et al. (2005); Nghiem et al. (2004a) and references therein for examples of modelization and simulation works). The present paper presents the mathematical analysis of such a model within a simplified framework. We consider that some CO_2 is injected at a given depth in a porous medium saturated with water. For simplicity, we focus on the onedimensional problem resulting from the competition between the migration of this gaseous species by gravity, and its dissolution into water. We denote by $x \in \mathbb{R}$ the vertical space variable, increasing with depth. We denote $X \in [0, \overline{X}]$ the molar fraction of CO_2 in water (always comprised between 0 and the maximum dissolution concentration $\overline{X} \in [0,1]$, assumed to be constant) and $S \in [0,1]$ the saturation of water (i.e. the volumic fraction of the porous medium filled by water, the volumic fraction $1-S$ being filled by gaseous CO_2). These quantities X and S, functions of the space variable x and the time variable t, are assumed to be such that, for a given (x,t) :

- either the gaseous phase is present, which means that $S(x,t) < 1$, and then $X(x,t) = \overline{X}$,
- or the gaseous phase is not present, and then $S(x,t) = 1$, and all values $X(x,t) \in [0, \overline{X}]$ are possible.

This alternative is summarized into the following relation.

$$(X(x,t) \leq \overline{X} \text{ and } S(x,t) = 1) \text{ or } (X(x,t) = \overline{X} \text{ and } S(x,t) \leq 1). \tag{11.1}$$

We then assume that the transport of CO_2 results only from two mechanisms : the flow of the gaseous phase, and the flow of water phase containing some dissolved gas. In this simplified model, we do not take into account the diffusion-dispersion phenomena of CO_2 within the water phase, nor the effects of compressibility of the gaseous phase and we assume that the capillary effects are negligible. Under such hypotheses, the conservation equations of the water component and of CO_2, given in a dimensionless form, read

$$\left\{ \begin{array}{lll} [S(1-X)]_t & -[(1-X)k_w(S)\,(P_x-1)]_x & = 0, \\ [SX+\xi(1-S)]_t & -\left[Xk_w(S)\,(P_x-1)+\frac{\xi}{\mu}k_g(S)\,(P_x-\rho)\right]_x & = 0, \end{array} \right. \tag{11.2}$$

where

- the lower index t (resp. x) denotes the partial derivative with respect to t (resp. x),
- ξ is the ratio between the molar density of the gas phase and that of water, ρ is the ratio between the bulk density of both phases, μ is the ratio of viscosity of the gaseous phase and that of the water phase (all these quantities are assumed to be given strictly positive constants),
- $k_w(S)$ and $k_g(S)$ are respectively the relative permeabilities of the water and the gaseous phases, assumed to be Lipschitz continuous functions of the saturation such that k_w is non decreasing with $k_w(0)=0$ and $k_w(1)=1$, and k_g is non increasing with $k_g(0)=1$ and $k_g(1)=0$, and such that $k_w(s)+k_g(s)$ is always strictly positive for all $s\in[0,1]$,
- $P(x,t)$ is the common pressure of both phases, function of the space and time variables.

In this model, the following hypothesis is assumed :

$$\overline{X} < \xi. \tag{11.3}$$

Its physical meaning is that a unit volume of gaseous phase contains more moles of CO_2 than a unit volume of water containing dissolved CO_2 at the maximum concentration. We then introduce the new unknown $u = SX + \xi(1-S)$. Thanks to Hypothesis (11.3) and using (11.1), we then express S and X as Lipschitz continuous functions $S(u)$ and $X(u)$ of $u\in[0,\xi]$, given by

$$\left\{ \begin{array}{l} X(u)=u \text{ and } S(u)=1,\ \forall u\in[0,\overline{X}], \\ X(u)=\overline{X} \text{ and } S(u)=\frac{\xi-u}{\xi-\overline{X}},\ \forall u\in[\overline{X},\xi]. \end{array} \right. \tag{11.4}$$

We substitute S and X by $S(u)$ and $X(u)$ in system (11.2), we introduce the unknown $v(x,t)=P_x(x,t)$ and we define the functions

$$\left\{ \begin{array}{l} f(u)=S(u)(1-X(u)), \\ a(u)=(1-X(u))k_w(S(u)), \\ b(u)=-(1-X(u))k_w(S(u)), \\ c(u)=X(u)k_w(S(u))+\frac{\xi}{\mu}k_g(S(u)), \\ d(u)=-X(u)k_w(S(u))-\frac{\xi}{\mu}k_g(S(u))\rho. \end{array} \right. \tag{11.5}$$

This provides the following system of equations :

$$\left\{ \begin{array}{lll} (f(u))_t & -(a(u)v+b(u))_x & = 0 \\ u_t & -(c(u)v+d(u))_x & = 0. \end{array} \right. \tag{11.6}$$

Let us first remark that system (11.6) is not an usual hyperbolic system of equations, since it contains no term v_t. The first idea to solve this system is to eliminate u_t between the two equations. Indeed, this could be achieved on a strong formulation of system (11.6), assuming that u is sufficiently regular and writing $(f(u))_t = f'(u)u_t$. It would then be easy to get v by solving a first order ordinary differential equation, and then one could find a function ψ such that $v(x,t)=\psi(u(x,t),t)$. But reporting such an expression in any of the two equations yields a nonlinear scalar hyperbolic equation with the unknown function u, the weak entropy solution of which is in general not continuous (recall that if the weak entropy solution is discontinuous, there does not exist any continuous weak solution). Hence the above method cannot be used for solving this problem in the general case.

Nevertheless, the elimination of u_t between the two equations can be done in the particular case where $f(u)$ is an affine function. This is the case if we assume that $\overline{X}=0$ (then $f(u)=1-u/\xi$), which then

leads to the classical incompressible immiscible two phase flow problem Aziz and Settari (1979). In this case, it is possible to eliminate the time derivative of the test functions between the weak formulations of both equations. We can then express v as a function of u after a simple integration with respect to x and then get the classical Buckley-Leverett equation. This nonlinear hyperbolic equation has been studied by many authors (see e.g. Aziz and Settari (1979); Brenier and Jaffré (1991); Eymard et al. (2000) and references therein) from the theoretical and numerical points of view. Let us notice that, in this case, discontinuous solutions u and v can be obtained even in the case of initial regular data, which shows the necessity to formulate the problem under a weak formulation.

But f is no longer affine taking $\overline{X} > 0$, since f is then a continuous piecewise affine function with a significant slope variation. It then becomes impossible to proceed to an elimination of the time derivative of the test functions on weak formulations of system (11.6), which prevents from expressing, at each time t, v as a function of u. In fact, one cannot expect that such an expression exists in the general case. Indeed, we show, on an analytical example given in section 11.2 (inspired by the problem presented in the beginning of this introduction), that we can observe the existence, for some times $t > 0$, of points x_1 and x_2 such that $v(x_1, t) \neq v(x_2, t)$ although $u(x_1, t) = u(x_2, t)$ holds.

System (11.6) must therefore be solved in a coupled way, including weak formulation senses in order to take into account discontinuous solutions. We obtain a first simple weak sense by multiplying the two equations of (11.6) by a regular test function, and integrating by parts. The particular case, obtained when f is an affine function, shows that this weak sense cannot be expected to characterize the solution. We have therefore used in section 11.2 some works of Liu (see Liu (1976, 1975, 1974)) for deriving a notion of entropy weak solution for this system. Then, considering a generalized Riemann problem (the situation is not symmetric with respect to u and v), we prove that system (11.6) can be solved thanks to the solution of two nonlinear scalar hyperbolic equations in u, the nonlinear functions in each of these equations being linked in order to provide the same shocks and characteristic velocities (theorem 11.2.5). Note that such a generalized Riemann problem has been studied in a case of multiphase flow in a porous medium, leading to the determination of the shocks and the rarefaction waves in some physical situations Bruining et al. (2003). Then, the Liu condition of admissibility of the shocks happens to result from a simple linear combination of the entropy inequalities for both nonlinear hyperbolic equations, using Krushkov entropy pairs Krushkov (1970). We then build analytically such two functions on some examples in section 11.2. We can then provide a proof of the existence, in the general case, of these two nonlinear functions (the uniqueness of which remains at this time an open problem), which relies on the construction of simultaneous convex and concave hulls for two functions (section 11.3) by passing to the limit on approximate piecewise affine hulls (theorem 11.2.6).

In order to extend this existence result to more general data than generalized Riemann problems, we then give a finite volume numerical scheme, the convergence of which to an entropy weak solution is proven in section 11.4. Since we have been able in section 11.2 to state the existence of the solution of a generalized Riemann problem, it would have been indeed natural to look for a numerical scheme obtained by averaging (after a discrete time step) the solution obtained from the analytical solution, deduced from the resolution of a sequence of generalized Riemann problems. Unfortunately, such an approach does not simultaneously respect both conservation equations, and its convergence properties do not seem to be clear. Thus we have developed an original numerical scheme, defined in such a way that values are defined for u at all the interfaces of the mesh using the Godunov numerical flux, simultaneously respecting the discrete balances resulting from a finite volume scheme applied to both equations. This choice enables the proof of a bounded variation estimate, and the proof of the convergence of the scheme to a weak solution of both equation can then be completed. The convergence property is then obtained for a strong topology for u, but only for a weak one for v. It is worth noticing that the proof of the L^∞ estimates on v, and that on u, cannot hold without the proof of the bounded variation estimate. Thanks to this convergence result, we are therefore able to prove the existence of a solution to system (11.6) for a large class of initial data (theorem 11.2.7).

So it is possible to compare the numerical results given by the scheme of section 11.4 and the analytical solution given in section 11.2. This is the aim of section 11.5, where an excellent agreement between these results seems to be a good indication for generalizing the numerical scheme studied here to less simplified models.

11.2 Entropy solutions and generalized Riemann problem

In order to give an entropy weak sense for a solution to system (11.6), we first state the following hypotheses on the functions a, b, c, d and f, always satisfied if these functions are given by (11.5) defining suitable prolongments.

$$\left\{\begin{array}{l} f,\ a,\ b,\ c,\ d \text{ are Lipschitz continuous functions defined on } \mathbb{R},\\ C_{\mathrm{Lip}} = \max\left(\|a'\|_\infty, \|b'\|_\infty, \|c'\|_\infty, \|d'\|_\infty, \|f'\|_\infty\right),\\ a,\ b,\ c,\ d \text{ are bounded on } \mathbb{R} \text{ and } C_{\max} = \max\left(\|a\|_\infty, \|b\|_\infty, \|c\|_\infty, \|d\|_\infty\right),\\ \text{there exists } f_m > 0 \text{ such that, for a.e. } s \in \mathbb{R},\ -f'(s) \geq f_m,\\ a(s) \geq 0 \text{ and } c(s) \geq 0 \text{ for all } s \in \mathbb{R},\\ \text{there exists } m_0 > 0 \text{ such that, for all } s \in \mathbb{R},\ a(s) + c(s) f_m \geq m_0. \end{array}\right. \tag{11.7}$$

Remarque 11.2.1
In the framework of the physical problem given in introduction to this paper, the monotony property of f is related to the fact that an increase of CO_2 must imply a decrease of water content. The hypothesis $a(s) + c(s) f_m \geq m_0 > 0$ expresses the fact that, whatever the water and CO_2 contents, the mixture of fluids must remain mobile.

Let us now define some hypotheses on the initial and boundary conditions.

$$\left\{\begin{array}{l} u_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R}) \text{ and } W_0 = \|u_0\|_{BV(\mathbb{R})},\\ \text{there exist } \overline{u}_0, \overline{M}_0 \in \mathbb{R} \text{ s.t. } u_0(x) = \overline{u}_0 \text{ for a.e. } x \in (-\infty, \overline{M}_0),\\ \text{(we then denote } U_m, U_M \in \mathbb{R} \text{ s.t. } U_m \leq u_0(x) \leq U_M \text{ for a.e. } x \in \mathbb{R}),\\ \overline{v}_0 \in L^\infty(\mathbb{R}_+) \text{ is given, and we denote } \overline{V}_0 = \|\overline{v}_0\|_{L^\infty(\mathbb{R}_+)}. \end{array}\right. \tag{11.8}$$

In (11.8), we classically define the set $BV(\mathbb{R})$ by $BV(\mathbb{R}) = \{u \in L^1_{\mathrm{loc}}(\mathbb{R}), \|u\|_{BV(\mathbb{R})} < \infty\}$ with $\|u\|_{BV(\mathbb{R})} = \sup\{\int_{\mathbb{R}} u(x)\varphi'(x)\,\mathrm{d}x,\ \varphi \in C^1_c(\mathbb{R}, [-1,1])\}$, where for all $E \subset \mathbb{R}^d$ with $d \in \mathbb{N}^\star$, for all $p \in \mathbb{N} \cup \{\infty\}$ and all $F \subset \mathbb{R}$, we denote by $C^p_c(E,F)$ the set of the restrictions to E of all C^p functions from $\mathbb{R}^d$ to F with a compact support.

Remarque 11.2.2
The hypothesis $u_0 \in BV(\mathbb{R})$ is strongly used in this paper. The other hypotheses are done in order to handle simple boundary conditions. The generalization of the results of this paper to more general boundary conditions, using in particular the results of Otto (1996) and Eymard et al. (2003), will be the object of further works. We assume below that $v(x,t)$ is equal to the boundary condition $\overline{v}_0(t)$ for small values x. We could as well give this boundary condition for large x, with minor changes in this paper.

Note that $U_m = \overline{u}_0 - W_0$ and $U_M = \overline{u}_0 + W_0$ always satisfy the third item of (11.8) from the two previous assumptions.

We now give the following definition.

Définition 11.2.3 (Weak entropy solution)
Under Hypotheses (11.7) and (11.8), the pair (u,v) is said to be an entropy weak solution of the problem :

$$\left\{\begin{array}{l} (f(u))_t - (a(u)v + b(u))_x = 0\\ u_t - (c(u)v + d(u))_x = 0\\ u(.,0) = u_0\\ v(x,.) = \overline{v}_0 \text{ for all } x \text{ small enough,} \end{array}\right. \tag{11.9}$$

if it is such that

- *the functions u and v satisfy $u, v \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$,*
- *the first three equations of (11.9) are satisfied in the following weak sense :*

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}} \big(f(u(x,t))\varphi_t(x,t) - (a(u(x,t))v(x,t) + b(u(x,t)))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} f(u_0(x))\varphi(x,0)\mathrm{d}x = 0,\\ \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}} \big(u(x,t)\varphi_t(x,t) - (c(u(x,t))v(x,t) + d(u(x,t)))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\varphi(x,0)\mathrm{d}x = 0,\\ \forall \varphi \in C^\infty_c(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}), \end{array} \tag{11.10}$$

- *the following entropy inequalities hold :*

$$
\begin{array}{l}
a(\kappa)\left(\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\left(\begin{array}{l}(u(x,t)\top\kappa-\kappa)\varphi_t(x,t)-\\ \left(\begin{array}{l}(c(u(x,t)\top\kappa)-c(\kappa))v(x,t)+\\ (d(u(x,t)\top\kappa)-d(\kappa))\end{array}\right)\varphi_x(x,t)\end{array}\right)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(u_0(x)\top\kappa-\kappa)\varphi(x,0)\mathrm{d}x\end{array}\right)\\
-c(\kappa)\left(\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\left(\begin{array}{l}(f(u(x,t)\top\kappa)-f(\kappa))\varphi_t(x,t)-\\ \left(\begin{array}{l}(a(u(x,t)\top\kappa)-a(\kappa))v(x,t)+\\ (b(u(x,t)\top\kappa)-b(\kappa))\end{array}\right)\varphi_x(x,t)\end{array}\right)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(f(u_0(x)\top\kappa)-f(\kappa))\varphi(x,0)\mathrm{d}x\end{array}\right)\geq 0,\\
-a(\kappa)\left(\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\left(\begin{array}{l}(u(x,t)\perp\kappa-\kappa)\varphi_t(x,t)-\\ \left(\begin{array}{l}(c(u(x,t)\perp\kappa)-c(\kappa))v(x,t)+\\ (d(u(x,t)\perp\kappa)-d(\kappa))\end{array}\right)\varphi_x(x,t)\end{array}\right)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(u_0(x)\perp\kappa-\kappa)\varphi(x,0)\mathrm{d}x\end{array}\right)\\
+c(\kappa)\left(\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\left(\begin{array}{l}(f(u(x,t)\perp\kappa)-f(\kappa))\varphi_t(x,t)-\\ \left(\begin{array}{l}(a(u(x,t)\perp\kappa)-a(\kappa))v(x,t)+\\ (b(u(x,t)\perp\kappa)-b(\kappa))\end{array}\right)\varphi_x(x,t)\end{array}\right)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(f(u_0(x)\perp\kappa)-f(\kappa))\varphi(x,0)\mathrm{d}x\end{array}\right)\geq 0,\\
\forall\kappa\in\mathbb{R},\ \forall\varphi\in C_c^\infty(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+),
\end{array}
\tag{11.11}
$$

  where we denote by $x\top y=\max(x,y)$ *and* $x\perp y=\min(x,y)$, *for all* $x,y\in\mathbb{R}$,
- *the fourth equation of (11.9) is satisfied in the following sense :*

$$
\forall T>0,\ \exists M\in\mathbb{R},\ v(x,t)=\overline{v}_0(t)\ \text{for a.e. } (x,t)\in(-\infty,M)\times(0,T). \tag{11.12}
$$

Let us comment the entropy weak sense (11.11) taken in the above definition. Let us assume that, at some given time, the solution is such that $u\to u_l$ and $v\to v_l$ for $x\to x_0$ with $x<x_0$ and that $u\to u_r$ and $v\to v_r$ for $x\to x_0$ with $x>x_0$. Then, the Rankine-Hugoniot relations deduced from system (11.9) gives the existence of some velocity V such that

$$
\begin{array}{ll}
V(f(u_l)-f(u_r)) & =-\big(a(u_l)v_l+b(u_l)-a(u_r)v_r-b(u_r)\big)\\
V(u_l-u_r) & =-\big(c(u_l)v_l+d(u_l)-c(u_r)v_r-d(u_r)\big).
\end{array}
$$

The Liu criterion, defining an admissible shock Liu (1976), expresses that the shock $u_l\to u_r$ cannot split in two shocks $u_l\to\kappa$, $\kappa\to u_r$, for any $\kappa\in\overline{\mathrm{I}}(u_l,u_r)$, where we define

$$
\forall s_1,s_2\in\mathbb{R},\ \overline{\mathrm{I}}(s_1,s_2)=[s_1,s_2]\ \text{ if } s_1\leq s_2,\ \text{ else } \overline{\mathrm{I}}(s_1,s_2)=[s_2,s_1]\,. \tag{11.13}
$$

This means that, if V_l, V_r and v_κ are reals such that

$$
\begin{array}{ll}
V_l(f(u_l)-f(\kappa)) & =-\big(a(u_l)v_l+b(u_l)-a(\kappa)v_\kappa-b(\kappa)\big)\\
V_l(u_l-\kappa) & =-\big(c(u_l)v_l+d(u_l)-c(\kappa)v_\kappa-d(\kappa)\big),\\
V_r(f(\kappa)-f(u_r)) & =-\big(a(\kappa)v_\kappa+b(\kappa)-a(u_r)v_r-b(u_r)\big)\\
V_r(\kappa-u_r) & =-\big(c(\kappa)v_\kappa+d(\kappa)-c(u_r)v_r-d(u_r)\big),
\end{array}
$$

then the properties $V_l\geq V$ and $V_r\leq V$ must hold. It is then easy to eliminate v_κ by multiplying the first and the third above equations by $c(\kappa)$, the second and the fourth by $a(\kappa)$, and then subtract the second to the first and the fourth to the third. The inequalities $V_l\geq V$ and $V_r\leq V$ can then be seen as Rankine-Hugoniot inequalities provided by weak formulation inequalities, in the same way as similar inequalities hold from the entropy weak formulation of a nonlinear scalar hyperbolic inequality using the entropy pairs of Krushkov (see Serre (1996a); Olejnik (1957); Otto (1996); Eymard et al. (2003)). Hence (11.11) can be deduced by analogy.

Remarque 11.2.4
In the case where there exist $\alpha, \beta \in \mathbb{R}$ with $f(u) = \alpha u + \beta$, we can easily deduce from definition (11.2.3) that $v(x,t)$ is obtained from $u(x,t)$ by

$$v(x,t) = \frac{(a(\overline{u}_0) - \alpha c(\overline{u}_0))\overline{v}_0(t) + b(\overline{u}_0) - b(u(x,t)) - \alpha(d(\overline{u}_0) - d(u(x,t)))}{a(u(x,t)) - \alpha c(u(x,t))}, \text{ for a.e. } (x,t) \in \mathbb{R} \times \mathbb{R}_+,$$

and u is the unique entropy solution of the equation

$$u_t - \left(c(u) \frac{(a(\overline{u}_0) - \alpha c(\overline{u}_0))\overline{v}_0(t) + b(\overline{u}_0) - b(u) - \alpha(d(\overline{u}_0) - d(u)))}{a(u) - \alpha c(u)} + d(u) \right)_x = 0,$$

with the initial condition $u(\cdot, 0) = u_0$ (it suffices to divide (11.11) by $a(\kappa) - \alpha c(\kappa)$).

Our aim is now to show that, under particular initial data called "generalized Riemann problem", we can exhibit a weak solution (u, v) to system (11.9) in the sense of Definition 11.2.3, only depending on x/t, permitting, in some case, to give the analytical expression of this solution. This generalized Riemann problem is defined by three reals u_l, u_r and g_l, and by setting $u_0(x) = u_l$ for a.e. $x < 0$, $u_0(x) = u_r$ for a.e. $x > 0$, and assuming that $v(x,t) = g_l$ for small values of x (we again consider a nonsymmetric condition for v). Let us recall that the concave (resp. convex) hull of a continuous functions $f : \mathbb{R} \to \mathbb{R}$ on the interval $[s_1, s_2]$, for given reals $s_1 \leq s_2$, is the function defined for all $s \in [s_1, s_2]$ by the infimum (resp. supremum) value in s of all functions $w \in C^2(\mathbb{R})$ such that $w'' \leq 0$ and $w \geq f$ (resp. $w'' \geq 0$ and $w \leq f$) on $[s_1, s_2]$. We also recall that, if f is Lipschitz continuous on $[s_1, s_2]$, these functions are Lipschitz continuous as well on $[s_1, s_2]$, with the same Lipschitz constant. We then state the following sufficient condition for an entropy weak solution to the generalized Riemann problem.

Théorème 11.2.5 (Generalized Riemann problem)
Under Hypotheses (11.7), using notation (11.13), let three reals g_l, u_l, u_r be given. Let $g : \overline{\mathrm{I}}(u_l, u_r) \to \mathbb{R}$ be a Lipschitz continuous function such that $g(u_l) = g_l$ and such that the functions μ, ν, defined by $\mu(u) = -(c(u)g(u) + d(u))$ and $\nu(f(u)) = -(a(u)g(u) + b(u))$ for all $u \in \overline{\mathrm{I}}(u_l, u_r)$, verify $\widehat{\nu}'(f(u)) = \widehat{\mu}'(u)$ for a.e. $u \in \overline{\mathrm{I}}(u_l, u_r)$, denoting by $\widehat{\mu}$ is the concave (resp. convex) hull of μ on $\overline{\mathrm{I}}(u_l, u_r)$ and by $\widehat{\nu}$ the convex (resp. concave) hull of ν on $\overline{\mathrm{I}}(f(u_l), f(u_r))$ if $u_l \geq u_r$ (resp. $u_l < u_r$). The existence of such a function is stated by theorem 11.2.6.

Let us define $V_M = \operatorname{ess\,sup}_{s \in \overline{\mathrm{I}}(u_l, u_r)} \widehat{\mu}'(s)$ and $V_m = \operatorname{ess\,inf}_{s \in \overline{\mathrm{I}}(u_l, u_r)} \widehat{\mu}'(s)$, let $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ be defined by

$$\begin{array}{l} u(x,t) = u_l, \ \forall t \in (0, +\infty), \quad \text{for a.e. } x \in (-\infty, tV_m) \\ x = t\widehat{\mu}'(u(x,t)), \ \forall t \in (0, +\infty), \quad \text{for a.e. } x \in (tV_m, tV_M) \\ u(x,t) = u_r, \ \forall t \in (0, +\infty), \quad \text{for a.e. } x \in (tV_M, +\infty), \end{array} \tag{11.14}$$

and let $v \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ be defined by $v(x,t) = g(u(x,t))$ for all $t \in (0, +\infty)$ and a.e. $x \in \mathbb{R}$.

Then (u, v) is an entropy weak solution of the system (11.9) in the sense of definition 11.2.3, where hypotheses (11.8) are satisfied setting $u_0(x) = u_l$ for a.e. $x < 0$ and $u_0(x) = u_r$ for a.e. $x > 0$, $W_0 = |u_l - u_r|$, $\overline{u}_0 = u_l$, $\overline{M}_0 = 0$ and $\overline{v}_0(t) = g_l$ for a.e. $t \in \mathbb{R}_+$.

Proof. Thanks to (11.14) and to the definition of $\widehat{\mu}$, we get from e.g. Serre (1996a) that u is the unique entropy weak solution of the problem

$$\begin{array}{l} u_t + (\mu(u))_x = u_t - (c(u)g(u) + d(u))_x = 0 \\ u(., 0) = u_0. \end{array} \tag{11.15}$$

Therefore it satisfies the classical weak sense, which is a consequence of (11.17),

$$\begin{array}{l} \displaystyle\int_{\mathbb{R}_+} \int_{\mathbb{R}} \big(u(x,t)\varphi_t(x,t) + \mu(u(x,t))\varphi_x(x,t)\big) \mathrm{d}x\mathrm{d}t + \int_{\mathbb{R}} u_0(x)\varphi(x,0)\mathrm{d}x = 0, \\ \forall \varphi \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}), \end{array} \tag{11.16}$$

and it also satisfies

$$\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\big((u(x,t)\top\kappa-\kappa)\varphi_t(x,t)+(\mu(u(x,t)\top\kappa)-\mu(\kappa))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(u_0(x)\top\kappa-\kappa)\varphi(x,0)\mathrm{d}x\geq 0,\\ \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\big((\kappa-u(x,t)\bot\kappa)\varphi_t(x,t)+(\mu(\kappa)-\mu(u(x,t)\bot\kappa))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(\kappa-u_0(x)\bot\kappa)\varphi(x,0)\mathrm{d}x\geq 0,\\ \forall\kappa\in\mathbb{R},\ \forall\varphi\in C_c^\infty(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+).\end{array}\tag{11.17}$$

Similarly, we get that $f(u)$ is the unique entropy weak solution of the problem

$$\begin{array}{l}(f(u))_t+(\nu(f(u)))_x=(f(u))_t-(a(u)g(u)+b(u))_x=0\\ u(.,0)=u_0,\end{array}\tag{11.18}$$

which implies

$$\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\big(f(u(x,t))\varphi_t(x,t)+\nu(f(u(x,t)))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t+\int_{\mathbb{R}}f(u_0(x))\varphi(x,0)\mathrm{d}x=0,\\ \forall\varphi\in C_c^\infty(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}),\end{array}\tag{11.19}$$

and

$$\begin{array}{l}\displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\big((f(u(x,t))\top f(\kappa)-f(\kappa))\varphi_t(x,t)+(\nu(f(u(x,t))\top f(\kappa))-\nu(f(\kappa)))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(f(u_0(x))\top f(\kappa)-f(\kappa))\varphi(x,0)\mathrm{d}x\geq 0,\\ \displaystyle\int_{\mathbb{R}_+}\int_{\mathbb{R}}\big((f(\kappa)-f(u(x,t))\bot f(\kappa))\varphi_t(x,t)+(\nu(f(\kappa))-\nu(f(u(x,t))\bot f(\kappa)))\varphi_x(x,t)\big)\mathrm{d}x\mathrm{d}t\\ \displaystyle+\int_{\mathbb{R}}(f(\kappa)-f(u_0(x))\bot f(\kappa))\varphi(x,0)\mathrm{d}x\geq 0,\\ \forall\kappa\in\mathbb{R},\ \forall\varphi\in C_c^\infty(\mathbb{R}\times\mathbb{R}_+,\mathbb{R}_+).\end{array}\tag{11.20}$$

We first note that replacing $\mu(u(x,t))$ by $-c(u(x,t))v(x,t)-d(u(x,t))$ in (11.16) and $\nu(f(u(x,t)))$ by $-a(u(x,t))v(x,t)-b(u(x,t))$ in (11.19) gives (11.10). We then remark that $f(s_1\top s_2)=f(s_1)\bot f(s_2)$, $f(s_1\bot s_2)=f(s_1)\top f(s_2)$ for all reals s_1,s_2. Defining $\mathrm{sign}^+(x)=1$ for all $x>0$ else $\mathrm{sign}^+(x)=0$ and $\mathrm{sign}^-(x)=-1$ for all $x<0$ else $\mathrm{sign}^-(x)=0$, we can write the relations

$$\begin{array}{l}\mu(u(x,t)\top\kappa)-\mu(\kappa)=-\mathrm{sign}^+(u(x,t)-\kappa)\,(c(u(x,t))v(x,t)+d(u(x,t))-c(\kappa)g(\kappa)-d(\kappa))\,,\\ \mu(\kappa)-\mu(u(x,t)\bot\kappa)=-\mathrm{sign}^-(u(x,t)-\kappa)\,(c(u(x,t))v(x,t)+d(u(x,t))-c(\kappa)g(\kappa)-d(\kappa))\,,\\ \nu(f(u(x,t))\top f(\kappa))-\nu(f(\kappa))=\mathrm{sign}^-(u(x,t)-\kappa)\,(a(u(x,t))v(x,t)+b(u(x,t))-a(\kappa)g(\kappa)-b(\kappa))\,,\\ \nu(f(\kappa))-\nu(f(u(x,t))\bot f(\kappa))=\mathrm{sign}^+(u(x,t)-\kappa)\,(a(u(x,t))v(x,t)+b(u(x,t))-a(\kappa)g(\kappa)-b(\kappa))\,.\end{array}$$

Using the above relations, we get that the sum of the first relation of (11.17) multiplied by $a(\kappa)$ and of the second relation of (11.20) multiplied by $c(\kappa)$ provides the first relation of (11.11), whereas the sum of the second relation of (11.17) multiplied by $a(\kappa)$ and of the first relation of (11.20) multiplied by $c(\kappa)$ provides the second relation of (11.11). Relation (11.12) is an immediate consequence of

$$\forall T>0,\ u(x,t)=u_l \text{ for a.e. } (x,t)\in(-\infty,\min(V_m\ T,0))\times(0,T),$$

and of $g(u_l)=g_l$. □

An analytical example

We now apply theorem 11.2.5 to some particular case, for which we can give the analytical expression of the solutions that we provide. The interest of this example is twofold. First, it follows as closely as possible realistic data in the case of injection of CO_2 in a porous medium saturated with water. Secondly, it shows that v cannot be expressed, in the general case, as a function of u. We consider the functions a, b, c, d, and f, defined by (11.4) and (11.5), and the following data :

$$\boxed{\xi = 0.07 \quad \overline{X} = 0.06 \quad \rho = 0.17 \quad \mu = 0.1 \quad k_w(S) = S \quad k_g(S) = (1-S)}$$

*F*irst example

We consider the case $u_l^{(1)} = 0$ and $u_r^{(1)} = S_0\overline{X} + \xi(1-S_0)$ with $S_0 = 0.1$. Hence the left part (the upper one) is fully saturated with pure water, whereas the right one (the lower one) is initially filled by the water phase, at the water saturation S_0, containing dissolved gas at the maximum concentration, and by the gaseous phase, at the gas saturation $1-S_0$. This case corresponds to a zoom on the top of the region in which some CO_2 has been previously injected. We assume that the water phase does not move at the top of the region, which corresponds to set the gradient of the pressure equal to the hydrostatic one $g_l^{(1)} = 1$. Let us denote by $\tau_{fs}^{(1)} = \frac{1-\overline{X}}{\xi-\overline{X}} = -f'(u)$, for all $u \in (\overline{X}, \xi)$. It is then possible to find the set of values $(g^{(1)}, u_s^{(1)}, g_s^{(1)}, V_s^{(1)}, Q_s^{(1)})$, with $u_s^{(1)} \in (\overline{X}, \xi)$, solution of the following system of equations :

$$\left\{\begin{array}{l} V_s^{(1)}\left(f(u_l^{(1)}) - f(u_s^{(1)})\right) = -\left(a(u_l^{(1)})g_l^{(1)} + b(u_l^{(1)}) - a(u_s^{(1)})g_s^{(1)} - b(u_s^{(1)})\right), \\ V_s^{(1)}\left(u_l^{(1)} - u_s^{(1)}\right) = -\left(c(u_l^{(1)})g_l^{(1)} + d(u_l^{(1)}) - c(u_s^{(1)})g_s^{(1)} - d(u_s^{(1)})\right), \\ Q_s^{(1)} = (a(u_s^{(1)}) + \tau_{fs}^{(1)}\ c(u_s^{(1)}))g_s^{(1)} + b(u_s^{(1)}) + \tau_{fs}^{(1)}\ d(u_s^{(1)}), \\ \forall u \in [\overline{X}, u_r^{(1)}],\ (a(u) + \tau_{fs}^{(1)}\ c(u))g^{(1)}(u) + b(u) + \tau_{fs}^{(1)}\ d(u) = Q_s^{(1)}, \\ g^{(1)}(u) = g_l^{(1)},\ \forall u \in [0, \overline{X}], \\ \mu'(u_s^{(1)}) = (cg^{(1)} + d)'(u_s^{(1)}) = V_s^{(1)}. \end{array}\right.$$

The function $(\widehat{\mu}^{(1)})'$ is then given by

$$\left\{\begin{array}{l} (\widehat{\mu}^{(1)})'(u) = V_s^{(1)},\ \forall u \in [u_l, u_s^{(1)}], \\ (\widehat{\mu}^{(1)})'(u) = (cg^{(1)} + d)'(u),\ \forall u \in [u_s^{(1)}, \xi] \end{array}\right.$$

Hence the solution is given by a shock between $u_l^{(1)}$ and $u_s^{(1)}$, which moves at the velocity $V_s^{(1)}$, and a rarefaction wave between $u_s^{(1)}$ and $u_r^{(1)}$. In this particular case, it is even possible to give an explicit value for $u_s^{(1)}$:

$$u_s^{(1)} = \frac{\xi\overline{X}(1-\mu) - (\overline{X}-\xi)(\mu\xi\overline{X})^{1/2}}{\xi - \mu\overline{X}}$$

and all the other values are then easily deduced. The functions $g^{(1)}$, $\mu^{(1)}$ and $\nu^{(1)}$ are represented on Figure 11.1. Since, by construction, in the case $u_l = u_l^{(1)}$, $u_r = u_r^{(1)}$, $g_l = g_l^{(1)}$, the function $g = g^{(1)}$ satisfies the hypotheses of Theorem 11.2.5, we thus obtain that the pair of functions $(u^{(1)}, v^{(1)})$ given by (11.14) with $g = g^{(1)}$, is an entropy weak solution of the system (11.9) in the sense of definition 11.2.3. We then get that the value of $g^{(1)}(u_r^{(1)})$, denoted by $g_r^{(1)}$, is given by

$$g_r^{(1)} = \frac{Q_s^{(1)} - b(u_r^{(1)}) - \tau_{fs}^{(1)} d(u_r^{(1)})}{a(u_r^{(1)}) + \tau_{fs}^{(1)} c(u_r^{(1)})}. \tag{11.21}$$

*S*econd example

Let now consider a second example, coupled with the preceding one (we give hereafter a situation where both examples are simultaneously encountered). In this second problem, we set $u_l^{(2)} = u_r^{(1)} = S_0\overline{X} + \xi(1-S_0)$, $u_r^{(2)} = u_l^{(1)} = 0$, and we set $g_l^{(2)} = g_r^{(1)}$. In this case, the left part (the upper one) is initially filled by the water phase with dissolved gas and by the gaseous phase, whereas the right one (the lower one) is initially filled by pure water. This case corresponds to a zoom on the bottom of the region in which some CO_2 has been previously injected. We then define $u_s^{(2)}$ by $u_s^{(2)} = \overline{X}$, this value being such that a shock occurs between $u_l^{(2)}$ and $u_s^{(2)}$. Then the velocity of the shock $V_s^{(2)}$ is given by the solution $(V_s^{(2)}, g_s^{(2)})$ of the system

$$\left\{\begin{array}{l} V_s^{(2)}\left(f(u_l^{(2)}) - f(u_s^{(2)})\right) = -\left(a(u_l^{(2)})g_l^{(2)} + b(u_l^{(2)}) - a(u_s^{(2)})g_s^{(2)} - b(u_s^{(2)})\right) \\ V_s^{(2)}\left(u_l^{(2)} - u_s^{(2)}\right) = -\left(c(u_l^{(2)})g_l^{(2)} + d(u_l^{(2)}) - c(u_s^{(2)})g_s^{(2)} - d(u_s^{(2)})\right). \end{array}\right. \tag{11.22}$$

Therefore $(V_s^{(2)}, g_s^{(2)})$ is given by

$$\begin{cases} V_s^{(2)} = \dfrac{1}{\mu}(\rho - g_l^{(2)}) \\ g_s^{(2)} = g_l^{(2)} \dfrac{S_0(\mu-1)+1}{\mu} + (1-S_0)(1-\dfrac{\rho}{\mu}). \end{cases} \tag{11.23}$$

We then define the function $g^{(2)}(u)$, for all $u \in [u_s^{(2)}, u_l^{(2)}]$ by

$$\begin{cases} V_s^{(2)} \left(f(u_l^{(2)}) - f(u) \right) = - \left(a(u_l^{(2)}) g_l^{(2)} + b(u_l^{(2)}) - a(u) g^{(2)}(u) - b(u) \right) \\ V_s^{(2)} \left(u_l^{(2)} - u \right) = - \left(c(u_l^{(2)}) g_l^{(2)} + d(u_l^{(2)}) - c(u) g^{(2)}(u) - d(u) \right). \end{cases}$$

Since, for all $u \in [u_r^{(2)}, u_s^{(2)}]$, the conservation equations are respectively linear with respect to u and $f(u)$, there is a contact discontinuity whose velocity $V_c^{(2)}$ is given by $V_c^{(2)} = (1 - g_s^{(2)})$. We then define $g^{(2)}(u) = g_s^{(2)}$ for all $u \in [u_r^{(2)}, u_s^{(2)}]$, and we set $g_r^{(2)} = g_s^{(2)}$. The function $(\widehat{\mu}^{(2)})'$ is then given by

$$\begin{cases} (\widehat{\mu}^{(2)})'(u) = V_s^{(2)}, \ \forall u \in (u_s^{(2)}, u_l^{(2)}), \\ (\widehat{\mu}^{(2)})'(u) = V_c^{(2)}, \ \forall u \in (u_r^{(2)}, u_s^{(2)}) \end{cases}$$

The functions $g^{(2)}$, $\mu^{(2)}$ and $\nu^{(2)}$ are represented on Figure 11.2. Since, in the case $u_l = u_l^{(2)}$, $u_r = u_r^{(2)}$, $g_l = g_l^{(2)}$, the function $g = g^{(2)}$ again satisfies the hypotheses of Theorem 11.2.5, we thus obtain that the pair of functions $(u^{(2)}, v^{(2)})$ given by (11.14) with $g = g^{(2)}$, is an entropy weak solution of the system (11.9) in the sense of definition 11.2.3.

A third example built with the two preceding ones

It is now possible to consider the case of Problem (11.9), where the function u_0 of (11.8) is given by $u_0(x) = u_l^{(1)}$ for all $x < 2/5$, $u_0(x) = u_r^{(1)} = u_l^{(2)}$ for $x \in (2/5, 4/5)$ and $u_0(x) = u_r^{(2)} = u_l^{(1)}$ for $x > 4/5$. We then assume that $\overline{v}_0(t) = g_l^{(1)}$, for a.e. $t \in \mathbb{R}_+$. These data correspond to the case where some CO_2 has been previously injected in the region given by $x \in (2/5, 4/5)$ (simulations of this case are also considered in section 11.5). We then examine the simultaneous displacement of the top of the bubble and its bottom, at least for a limited period of time. We consider the functions (u, v) given by

$$\begin{array}{l} u(x,t) = u^{(1)}(x - 2/5, t) \text{ and } v(x,t) = g^{(1)}(u^{(1)}(x-2/5,t)), \ \forall t \in (0,T), \text{ for a.e } x \in (-\infty, 3/5), \\ u(x,t) = u^{(2)}(x - 4/5, t) \text{ and } v(x,t) = g^{(2)}(u^{(2)}(x-4/5,t)), \ \forall t \in (0,T), \text{ for a.e } x \in (3/5, +\infty), \end{array}$$

with $T > 0$ small enough such that the two following conditions simultaneously hold

$$T \operatorname*{ess\,sup}_{u \in [u_l^{(1)}, u_r^{(1)}]} (\widehat{\mu}^{(1)})'(u) < 3/5 - 2/5 \text{ and } \quad -T \operatorname*{ess\,inf}_{u \in [u_r^{(2)}, u_l^{(2)}]} (\widehat{\mu}^{(2)})'(u) < 4/5 - 3/5.$$

These conditions on T ensure that the solution issued from the first generalized Riemann problem is equal to the initial data for $x > 3/5$, and that the solution issued from the second one is equal to the initial data for $x < 3/5$, for all time $t \leq T$. Then this pair (u, v) is an entropy weak solution of the system (11.9) in the sense of definition 11.2.3 until time T. We see that in this case, the values $v(x,t)$ for small values of x and large ones, correspond to the same value of $u(x,t)$ (which is equal to $u_l^{(1)} = u_r^{(2)} = 0$), are respectively equal to $g_l^{(1)} = 1$ and $g_r^{(2)} = g_s^{(2)}$, which are different values in the general case (see the corresponding values on the figures). Note that, although this analytical solution holds only for $t \leq T$, the numerical scheme used in section 11.5 allows to approximate the solution at any time $t > 0$.

We now state the existence result, which allows applying theorem 11.2.5 to any generalized Riemann problem in the sense given above.

Théorème 11.2.6 (A coupled convexity property)
Under Hypotheses (11.7), using notation (11.13), let three reals g_l, u_l, u_r be given.

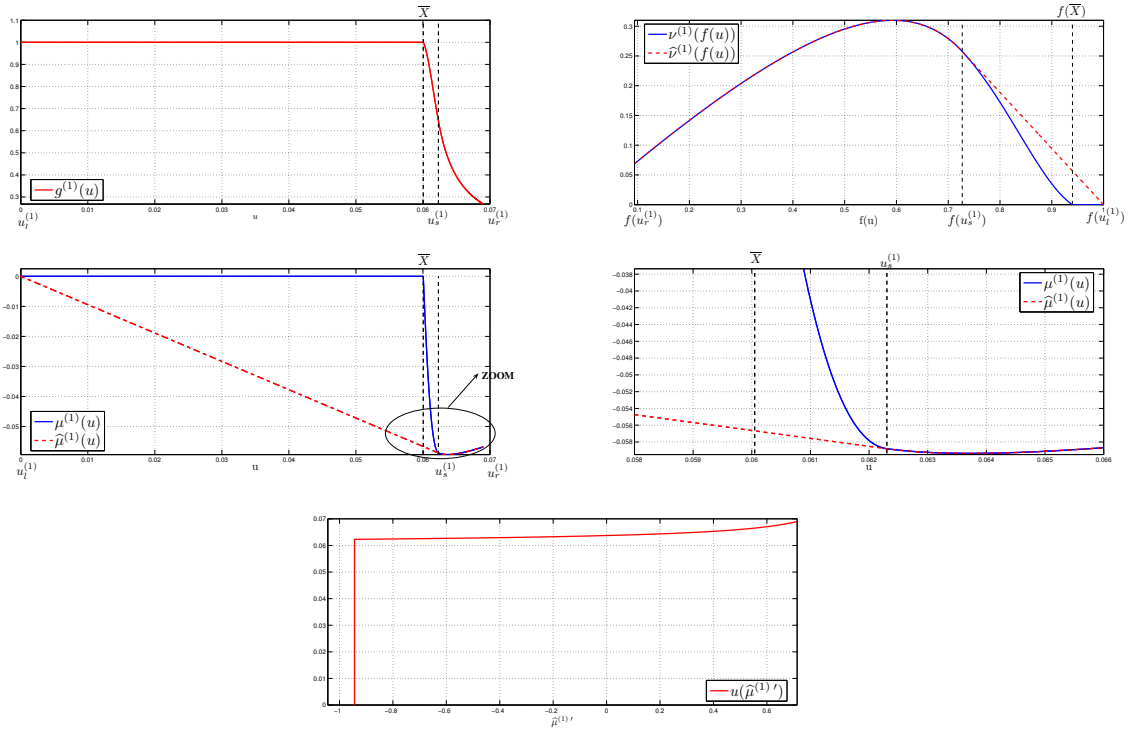


FIG. 11.1: Functions $g^{(1)}$ (top left), $\nu^{(1)}$ (top right) with $\widehat{\nu}^{(1)}$ (dashed line), $\mu^{(1)}$ (middle left) with $\widehat{\mu}^{(1)}$ (dashed line) and an enlargement of this function (middle right). and $u^{(1)}$ as a function of $(\widehat{\mu}^{(1)})'$, which gives the profile of the solution for all time t (bottom)

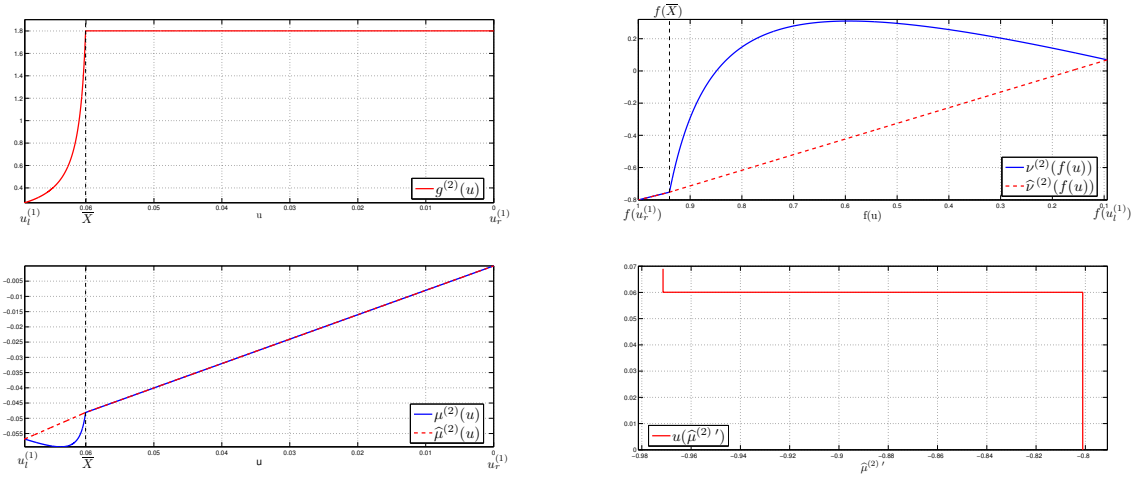


FIG. 11.2: Functions $g^{(2)}$ (top left), $\nu^{(2)}$ (top right) with $\widehat{\nu}^{(2)}$ (dashed line), $\mu^{(2)}$ (middle left) with $\widehat{\mu}^{(2)}$ (dashed line) and $u^{(2)}$ as a function of $(\widehat{\mu}^{(2)})'$, which gives the profile of the solution for all time t (bottom)

Then there exists at least one Lipschitz continuous function $g : \overline{\mathrm{I}}(u_l, u_r) \to \mathbb{R}$ *such that* $g(u_l) = g_l$ *and such that the functions* μ, ν, *defined by* $\mu(u) = -(c(u)g(u) + d(u))$ *and* $\nu(f(u)) = -(a(u)g(u) + b(u))$ *for all* $u \in \overline{\mathrm{I}}(u_l, u_r)$, *verify* $\widehat{\nu}'(f(u)) = \widehat{\mu}'(u)$ *for a.e.* $u \in \overline{\mathrm{I}}(u_l, u_r)$, *denoting by* $\widehat{\mu}$ *the concave (resp. convex) hull of* μ *on* $\overline{\mathrm{I}}(u_l, u_r)$ *and by* $\widehat{\nu}$ *the convex (resp. concave) hull of* ν *on* $\overline{\mathrm{I}}(f(u_l), f(u_r))$ *if* $u_r \leq u_l$ *(resp.* $u_r > u_l$*).*

The proof of theorem 11.2.6 is a straightforward consequence of Lemma 11.3.2 proven in section 11.3. Let us conclude this section with the following theorem.

Théorème 11.2.7 (Existence of an entropy weak solution of the system (11.9))
Under Hypotheses (11.7) and (11.8), there exists at least one entropy weak solution (u, v) *of the system (11.9) in the sense of definition 11.2.3, which moreover satisfies* $u - u_0 \in \mathrm{Lip}(\mathbb{R}_+; L^1(\mathbb{R})) \cap L^\infty(\mathbb{R}_+; BV(\mathbb{R}))$.

The proof of Theorem 11.2.7 is given in section 11.4, by passing to the limit in a finite volume scheme.

11.3 Proof of a coupled convexity property

*R*ecall of method for the decoupled case

In order to proceed to the proof of theorem 11.2.6 by passing to the limit in some approximation method, let us first recall a method to approximate, for two given reals u_l and u_r, the concave ($u_l \geq u_r$) or convex ($u_l \leq u_r$) hull $\widehat{\mu}$ of a function μ on the interval $\overline{\mathrm{I}}(u_l, u_r)$.

For all $i \leq j \in \mathbb{N}$, we denote $[\![i, j]\!] = \{k \in \mathbb{N}, i \leq k \leq j\}$. Let $N \in \mathbb{N}$ with $N \geq 2$ be given. We define the sequences $(\widehat{I}(j))_{j \in [\![0,N]\!]}$ and $(v_{\widehat{I}(j)})_{j \in [\![0,N]\!], \widehat{I}(j) < N}$ by

$$\left\{ \begin{array}{l} u_j = u_l + \frac{j}{N}(u_r - u_l),\ \forall j \in [\![0, N]\!], \\ \widehat{I}(0) = 0, \\ \forall j \in [\![0, N-1]\!], \\ \quad \text{if } \widehat{I}(j) < N \text{ then } \left\{ \begin{array}{l} v_{\widehat{I}(j)} = \min\limits_{k \in [\![\widehat{I}(j)+1, N]\!]} \dfrac{\mu(u_k) - \mu(u_{\widehat{I}(j)})}{u_k - u_{\widehat{I}(j)}}, \\ \widehat{I}(j+1) \text{ is any element of } \left\{ k \in [\![\widehat{I}(j)+1, N]\!],\ \dfrac{\mu(u_k) - \mu(u_{\widehat{I}(j)})}{u_k - u_{\widehat{I}(j)}} = v_{\widehat{I}(j)} \right\} \end{array} \right. \\ \quad \text{else } \widehat{I}(j+1) = N. \end{array} \right.$$

We next complete the definition of the sequence $(v_i)_{i \in [\![0, N-1]\!]}$ by

$$\left\{ \begin{array}{l} \widehat{N} = \min \left\{ j \in [\![0, N]\!], \widehat{I}(j) = N \right\} \\ \forall j \in [\![0, \widehat{N} - 1]\!],\ \forall k \in [\![\widehat{I}(j), \widehat{I}(j+1) - 1]\!],\ v_k = v_{\widehat{I}(j)}. \end{array} \right.$$

Then the piecewise constant function $\phi^{(N)}$, defined by

$$\phi^{(N)} : \overline{\mathrm{I}}(u_l, u_r) \to \mathbb{R},\ u \mapsto v_k, \quad \text{for a.e. } u \in \overline{\mathrm{I}}(u_k, u_{k+1}),\ \forall k \in [\![0, N-1]\!],$$

permits to define the piecewise affine continuous function $\widehat{\mu}^{(N)}$:

$$\widehat{\mu}^{(N)} : \overline{\mathrm{I}}(u_l, u_r) \to \mathbb{R},\ u \mapsto \mu(u_l) + \int_{u_l}^{u} \phi^{(N)}(s)\mathrm{d}s,$$

which is the concave hull ($u_l \geq u_r$) or the convex hull ($u_l \leq u_r$) of the function $\mu^{(N)}$ which is piecewise affine on all $\overline{\mathrm{I}}(u_k, u_{k+1})$, $k \in [\![0, N]\!]$, such that $\mu^{(N)}(u_k) = \mu(u_k)$ for all $k \in [\![0, N]\!]$. Then one can prove that $\mu^{(N)}$ uniformly converges to μ on $\overline{\mathrm{I}}(u_l, u_r)$ as $N \to \infty$, whereas $\widehat{\mu}^{(N)}$ also uniformly converges to $\widehat{\mu}$ on $\overline{\mathrm{I}}(u_l, u_r)$ as $N \to \infty$ (note that this result is indeed a consequence of Lemma 11.3.2 below, in the particular case where $c(u) = 0$ and $a(u) = 1$ for all $u \in \mathbb{R}$). The approximation method used below is then inspired by this one.

*S*ome functions related to the Rankine-Hugoniot relations

We now define two functions deduced from the Rankine-Hugoniot relations resulting from the conservation laws (11.6). Let us assume that there exists three reals u_1, u_2 and g_1 such that u tends to u_1 and v tends to g_1 for x tending to x_0 with $x < x_0$ and that u tends to u_2 for x tending to x_0 with $x > x_0$. Then, from the system of the two Rankine-Hugoniot relations, we can deduce the value g_2 to which tends v for x tending to x_0 with $x > x_0$, as well as the velocity of the shock. Indeed, this velocity and the value g_2 are therefore functions of g_1, u_1 and u_2, respectively denoted $V(g_1,u_1,u_2)$ and $G(g_1,u_1,u_2)$, solutions to the following linear system of equations in the case $u_1 \neq u_2$.

$$\left\{\begin{array}{ll} V(g_1,u_1,u_2)(f(u_1)-f(u_2)) & = -\big(a(u_1)g_1+b(u_1)-a(u_2)G(g_1,u_1,u_2)-b(u_2)\big) \\ V(g_1,u_1,u_2)(u_1-u_2) & = -\big(c(u_1)g_1+d(u_1)-c(u_2)G(g_1,u_1,u_2)-d(u_2)\big). \end{array}\right. \tag{11.24}$$

Indeed, thanks to hypotheses (11.7) and introducing the notation

$$\tau_h(u_1,u_2) = \frac{h(u_2)-h(u_1)}{u_2-u_1},\ \forall h \in C^0(\mathbb{R}),\ \forall u_1,u_2 \in \mathbb{R} \text{ with } u_1 \neq u_2, \tag{11.25}$$

(note that we have $-\tau_f(u_1,u_2) \geq f_m$) we get the following expressions for $V(g_1,u_1,u_2)$ and $G(g_1,u_1,u_2)$ (we prolong the latter by continuity for $u_1 = u_2$).

$$\left\{\begin{array}{l} V(g_1,u_1,u_2) = \dfrac{(c(u_2)\tau_a(u_1,u_2)-a(u_2)\tau_c(u_1,u_2))g_1+c(u_2)\tau_b(u_1,u_2)-a(u_2)\tau_d(u_1,u_2)}{a(u_2)-c(u_2)\tau_f(u_1,u_2)}, \\ G(g_1,u_1,u_2) = \dfrac{(a(u_1)-c(u_1)\tau_f(u_1,u_2))g_1+b(u_1)-b(u_2)-(d(u_1)-d(u_2))\tau_f(u_1,u_2)}{a(u_2)-c(u_2)\tau_f(u_1,u_2)}, \\ \forall g_1,u_1,u_2 \in \mathbb{R} \text{ with } u_1 \neq u_2, \\ \\ G(g_1,u_1,u_1) = g_1,\ \forall g_1,u_1 \in \mathbb{R}. \end{array}\right. \tag{11.26}$$

Some properties of these functions, used in the next proofs, are given in an appendix.

*A*pproximation in the case of the coupled problem

Let us now turn to the coupled problem considered in this paper. Let reals g_l,u_l,u_r be given. In order to prove theorem 11.2.6, we must show the existence of a Lipschitz continuous function $g : \overline{\mathrm{I}}(u_l,u_r) \to \mathbb{R}$ such that $g(u_l) = g_l$ and the functions μ, ν defined by : $\mu(u) = -(c(u)g(u)+d(u))$ and $\nu(f(u)) = -(a(u)g(u)+b(u))$ for all $u \in \overline{\mathrm{I}}(u_l,u_r)$ verify $\widehat{\nu}'(f(u)) = \widehat{\mu}'(u)$ for a.e. $u \in \overline{\mathrm{I}}(u_l,u_r)$, denoting by $\widehat{\mu}$ is the concave (resp. convex) hull of μ on $\overline{\mathrm{I}}(u_l,u_r)$ and by $\widehat{\nu}$ the convex (resp. concave) hull of ν on $\overline{\mathrm{I}}(f(u_l),f(u_r))$ if $u_l \geq u_r$ (resp. $u_l < u_r$). We then follow the lines of the approximation method given in introduction to this section. Let $N \in \mathbb{N}$ with $N \geq 2$ be given. Denoting for all $i \leq j \in \mathbb{N}$ by $[\![i,j]\!] = \{k \in \mathbb{N}, i \leq k \leq j\}$, we first define the sequences $(\widehat{I}(j))_{j\in[\![0,N]\!]}$, $(v_{\widehat{I}(j)})_{j\in[\![0,N]\!],\widehat{I}(j)<N}$ and $(g_{\widehat{I}(j)})_{j\in[\![0,N]\!]}$ (these sequences are used to define the approximations of the convex or concave hulls which are looking for) by

$$\left\{\begin{array}{l} u_j = u_l + \frac{j}{N}(u_r-u_l),\ \forall j \in [\![0,N]\!], \\ \widehat{I}(0) = 0,\ g_0 = g_l, \\ \forall j \in [\![0,N-1]\!], \\ \quad \text{if } \widehat{I}(j) < N \text{ then} \left\{\begin{array}{l} v_{\widehat{I}(j)} = \min\limits_{k\in[\![\widehat{I}(j)+1,N]\!]} V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k), \\ \widehat{I}(j+1) \text{ is any element of } \left\{k \in [\![\widehat{I}(j)+1,N]\!],\ V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k) = v_{\widehat{I}(j)}\right\}, \\ g_{\widehat{I}(j+1)} = G(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{\widehat{I}(j+1)}) \end{array}\right. \\ \quad \text{else } \widehat{I}(j+1) = N. \end{array}\right. \tag{11.27}$$

We next complete the definition of the sequences $(g_i)_{i\in[\![0,N]\!]}$ and $(v_i)_{i\in[\![0,N-1]\!]}$ by

$$\left\{\begin{array}{l} \widehat{N} = \min\left\{j \in [\![0,N]\!], \widehat{I}(j) = N\right\} \\ \forall j \in [\![0,\widehat{N}-1]\!],\ \forall k \in [\![\widehat{I}(j),\widehat{I}(j+1)-1]\!],\ g_{k+1} = G(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{k+1}) \text{ and } v_k = v_{\widehat{I}(j)}. \end{array}\right. \tag{11.28}$$

Thanks to the definition of sequences $(g_i)_{i\in[\![0,N]\!]}$ and $(v_i)_{i\in[\![0,N-1]\!]}$, we can now define the following piecewise constant functions :

$$\left\{\begin{array}{lllll} \phi^{(N)} & : & \overline{\mathrm{I}}(u_l,u_r)\to\mathbb{R}, & u\mapsto & v_k, \quad \text{for a.e. } u\in\overline{\mathrm{I}}(u_k,u_{k+1}),\ \forall k\in[\![0,N-1]\!], \\ \psi^{(N)} & : & \overline{\mathrm{I}}(f(u_l),f(u_r))\to\mathbb{R}, & w\mapsto & v_k, \quad \text{for a.e. } w\in\overline{\mathrm{I}}(f(u_k),f(u_{k+1})),\ \forall k\in[\![0,N-1]\!]. \end{array}\right. \tag{11.29}$$

The integration of these piecewise constant functions allows to define the following piecewise affine continuous functions :

$$\left\{\begin{array}{lll} \widehat{\mu}^{(N)} & : \overline{\mathrm{I}}(u_l,u_r)\to\mathbb{R}, & u\mapsto -(c(u_l)g_l+d(u_l))+\displaystyle\int_{u_l}^{u}\phi^{(N)}(s)\mathrm{d}s, \\ \widehat{\nu}^{(N)} & : \overline{\mathrm{I}}(f(u_l),f(u_r))\to\mathbb{R}, & w\mapsto -(a(u_l)g_l+b(u_l))+\displaystyle\int_{f(u_l)}^{w}\psi^{(N)}(s)\mathrm{d}s. \end{array}\right. \tag{11.30}$$

We then denote by $g^{(N)}$, $\mu^{(N)}$ the continuous functions which are piecewise affine on all $\overline{\mathrm{I}}(u_k,u_{k+1})$, $k\in[\![0,N]\!]$, and $\nu^{(N)}$ the continuous function which is piecewise affine on all $\overline{\mathrm{I}}(f(u_k),f(u_{k+1}))$, $k\in[\![0,N]\!]$, such that

$$\left\{\begin{array}{lll} g^{(N)} & : \overline{\mathrm{I}}(u_l,u_r)\to\mathbb{R}, & g^{(N)}(u_k)=g_k, \forall k\in[\![0,N]\!], \\ \mu^{(N)} & : \overline{\mathrm{I}}(u_l,u_r)\to\mathbb{R}, & \mu^{(N)}(u_k)=-(c(u_k)g_k+d(u_k)),\ \forall k\in[\![0,N]\!], \\ \nu^{(N)} & : \overline{\mathrm{I}}(f(u_l),f(u_r))\to\mathbb{R}, & \nu^{(N)}(f(u_k))=-(a(u_k)g_k+b(u_k)),\ \forall k\in[\![0,N]\!]. \end{array}\right. \tag{11.31}$$

We then have the following property.

Lemme 11.3.1
Under Hypotheses (11.7), let three reals g_l,u_l,u_r be given. Let $N\in\mathbb{N}$ with $N\geq 2$ be given and let $\widehat{N}\in\mathbb{N}$ and the sequences $(\widehat{I}(j))_{j\in[\![0,\widehat{N}]\!]}$, $(u_j)_{j\in[\![0,N]\!]}$, $(g_i)_{i\in[\![0,N]\!]}$ and $(v_i)_{i\in[\![0,N-1]\!]}$ be given by (11.27)-(11.28). Then the following properties hold :

1. *the sequence $(v_{\widehat{I}(j)})_{j\in[\![0,\widehat{N}-1]\!]}$, and therefore the sequence $(v_j)_{j\in[\![0,N-1]\!]}$, are non decreasing,*
2. *the following inequality holds*

$$\begin{array}{l} \forall j\in[\![0,\widehat{N}]\!],\ \forall k\in[\![\widehat{I}(j)+1,\widehat{I}(j+1)-1]\!], \\ V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k)\geq V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{\widehat{I}(j+1)})\geq V(g_k,u_k,u_{\widehat{I}(j+1)}), \end{array} \tag{11.32}$$

3. *if $u_r<u_l$ (resp. $u_r>u_l$) , then the functions $\widehat{\mu}^{(N)}$ is the concave (resp. convex) hull of $\mu^{(N)}$ on $\overline{\mathrm{I}}(u_l,u_r)$ and $\widehat{\nu}^{(N)}$ is the the convex (resp. concave) hull of $\nu^{(N)}$ on $\overline{\mathrm{I}}(f(u_l),f(u_r))$ (these functions are defined by (11.30) and (11.31)).*
4. *the function $g^{(N)}$ is bounded independently of N and is Lipschitz continuous on $\overline{\mathrm{I}}(u_l,u_r)$ with a constant independent of N.*
5. *the sequence $(v_i)_{i\in[\![0,N-1]\!]}$ is bounded independently of N.*

Proof.

Proof of item 1

Let us show that the sequence $(v_{\widehat{I}(j)})_{j\in[\![0,\widehat{N}]\!]}$ is non decreasing. Let $j\in[\![0,\widehat{N}-2]\!]$ and $k\in[\![\widehat{I}(j+1)+1,N]\!]$ be given. By definition of $v_{\widehat{I}(j)}$, we have $V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k)\geq v_{\widehat{I}(j)}=V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{\widehat{I}(j+1)})$. We apply Lemma 11.5.2 with $g_1=g_{\widehat{I}(j)}$, $u_1=u_{\widehat{I}(j)}$, $u_2=u_{\widehat{I}(j+1)}$ and $u_3=u_k$. We then get that the sign of $V(g_{\widehat{I}(j+1)},u_{\widehat{I}(j+1)},u_k)-V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k)$ is the same as that of $V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k)-V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{\widehat{I}(j+1)})$. Thus, the value $v_{\widehat{I}(j)}=V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_{\widehat{I}(j+1)})$, lower than or equal to $V(g_{\widehat{I}(j)},u_{\widehat{I}(j)},u_k)$ for $k\in[\![\widehat{I}(j+1)+1,N]\!]$, is lower than or equal to $V(g_{\widehat{I}(j+1)},u_{\widehat{I}(j+1)},u_k)$. This proves that $v_{\widehat{I}(j+1)}\geq v_{\widehat{I}(j)}$.

Proof of item 2

The left inequality in (11.32) results from (11.27). The right one is an immediate consequence of the left one and of Lemma 11.5.2 with $g_1=g_{\widehat{I}(j)}$, $u_1=u_{\widehat{I}(j)}$, $u_2=u_k$ and $u_3=u_{\widehat{I}(j+1)}$.

Proof of item 3

Let us now assume that $u_r < u_l$, and let us prove that the function $\widehat{\mu}^{(N)}$ is the concave hull of $\mu^{(N)}$ on $[u_r, u_l]$ (the case $u_r = u_l$ is straightforward, since it leads to constant functions, and the case $u_r > u_l$ can be handled in a similar way). For all $j \in [\![0, \widehat{N}]\!]$, we have $\mu^{(N)}(u_{\widehat{I}(j)}) = -\Big(c(u_{\widehat{I}(j)})\ g_{\widehat{I}(j)} + d(u_{\widehat{I}(j)})\Big)$. We have

$$\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_{\widehat{I}(j+1)}) = c(u_{\widehat{I}(j+1)})\ g_{\widehat{I}(j+1)} - c(u_{\widehat{I}(j)})\ g_{\widehat{I}(j)} + d(u_{\widehat{I}(j+1)}) - d(u_{\widehat{I}(j)}) \tag{11.33}$$

The algorithm gives $g_{\widehat{I}(j+1)} = G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{\widehat{I}(j+1)})$, so

$$\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_{\widehat{I}(j+1)}) = c(u_{\widehat{I}(j+1)})\ G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{\widehat{I}(j+1)}) - c(u_{\widehat{I}(j)})\ g_{\widehat{I}(j)} + d(u_{\widehat{I}(j+1)}) - d(u_{\widehat{I}(j)}), \tag{11.34}$$

using the definition of V, we obtain

$$\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_{\widehat{I}(j+1)}) = V\left(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{\widehat{I}(j+1)}\right)\left(u_{\widehat{I}(j)} - u_{\widehat{I}(j+1)}\right) \tag{11.35}$$

Hence, using the algorithm, we get that, for all $j \in [\![0, \widehat{N} - 1]\!]$,

$$\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_{\widehat{I}(j+1)}) = v_{\widehat{I}(j)}\left(u_{\widehat{I}(j)} - u_{\widehat{I}(j+1)}\right). \tag{11.36}$$

We have, by construction, $v_{\widehat{I}(j)} = v_k$, for all $k \in [\![\widehat{I}(j), \widehat{I}(j+1) - 1]\!]$ so

$$\begin{aligned} \mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_{\widehat{I}(j+1)}) &= v_{\widehat{I}(j)}\left(u_{\widehat{I}(j)} - u_{\widehat{I}(j+1)}\right) \\ &= \int_{u_{\widehat{I}(j+1)}}^{u_{\widehat{I}(j)}} v_{\widehat{I}(j)} = \int_{u_{\widehat{I}(j+1)}}^{u_{\widehat{I}(j)}} v_k = \int_{u_{\widehat{I}(j+1)}}^{u_{\widehat{I}(j)}} \phi^{(N)}(u) \end{aligned} \tag{11.37}$$

This leads to $\mu^{(N)}(u_{\widehat{I}(j)}) = \widehat{\mu}^{(N)}(u_{\widehat{I}(j)})$, thanks to definition of $\mu^{(N)}$. for all $j \in [\![0, \widehat{N}]\!]$. Since for all $k \in [\![\widehat{I}(j)+1, \widehat{I}(j+1) - 1]\!]$, we have $g_k = G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_k)$, we get

$$\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_k) = V(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_k)(u_{\widehat{I}(j)} - u_k).$$

Using (11.32), we obtain that $\mu^{(N)}(u_{\widehat{I}(j)}) - \mu^{(N)}(u_k) \geq \widehat{v}_j^{(N-1)}\left(u_{\widehat{I}(j)} - u_k\right)$. This proves that $\mu^{(N)}(u_k) \leq \widehat{\mu}^{(N)}(u_k)$, and therefore concludes the proof that $\widehat{\mu}^{(N)}$ is the concave hull of $\mu^{(N)}$ on $[u_r, u_l]$. Similarly, we get that $\nu^{(N)}(f(u_{\widehat{I}(j)})) = \widehat{\nu}^{(N)}(f(u_{\widehat{I}(j)}))$ for all $j \in [\![0, \widehat{N}]\!]$ and that

$$\nu^{(N)}(f(u_{\widehat{I}(j)})) - \nu^{(N)}(f(u_k)) = V(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_k)(f(u_{\widehat{I}(j)}) - f(u_k)).$$

Therefore, since f is strictly decreasing, we get that $\nu^{(N)}(f(u_k)) \geq \widehat{\nu}^{(N)}(f(u_k))$, which shows that $\widehat{\nu}^{(N)}$ is the convex hull of $\nu^{(N)}$ on $[f(u_l), f(u_r)]$. The case $u_r > u_l$ yields to similar conclusions.

Proof of item 4

We will first show this item about the sequence $g_{\widehat{I}(j)}$, $j \in [\![0, \widehat{N}]\!]$. Thanks to the definition of G, we have $g_{\widehat{I}(j)} = G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{\widehat{I}(j)})$, and by construction in the algorithm (11.27), $g_{\widehat{I}(j+1)} = G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{\widehat{I}(j+1)})$. By applying lemma (11.5.1), we get

$$\left|g_{\widehat{I}(j+1)} - g_{\widehat{I}(j)}\right| \leq C_{22}\left(\left|g_{\widehat{I}(j)}\right| + 1\right)\left|u_{\widehat{I}(j+1)} - u_{\widehat{I}(j)}\right|.$$

Using $g_{\widehat{I}(0)} = g_l$, and applying the discrete Gronwall's Lemma 11.5.3, we obtain,

$$\left|g_{\widehat{I}(j)}\right| \leq (|g_l| + 1)\exp\left[C_{22}\sum_{i=0}^{\widehat{N}-1}\left|u_{\widehat{I}(i+1)} - u_{\widehat{I}(i)}\right|\right],$$

Since $\sum_{i=0}^{\widehat{N}-1} \left| u_{\widehat{I}(i+1)} - u_{\widehat{I}(i)} \right| = |u_l - u_r|$, we get,

$$\left| g_{\widehat{I}(j)} \right| \leq C_9,$$

with $C_9 = (|g_l| + 1) \exp\left[C_{22} |u_l - u_r|\right] - 1$. We now turn to the study of the whole sequence $(g_k)_{k \in [\![0,N]\!]}$. We remark that, for all $j \in [\![0, \widehat{N}]\!]$ and for all k such that $\widehat{I}(j) \leq k \leq \widehat{I}(j+1) - 1$, we have

$$\begin{aligned} g_k &= G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_k) \\ g_{k+1} &= G(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_{k+1}). \end{aligned}$$

Hence, for all $k \in [\![\widehat{I}(j), \widehat{I}(j+1)]\!]$, we get

$$|g_{k+1} - g_k| \leq C_{22} \left(\left| g_{\widehat{I}(j)} \right| + 1 \right) |u_{k+1} - u_k| \tag{11.38}$$

Since $\left| g_{\widehat{I}(j)} \right| \leq C_9$, we conclude that g_k is Lipschitz continuous with the constant $C_{10} = C_{22}(C_9 + 1)$. From inequality (11.38) and thanks to an easy induction, we then get that

$$|g_k| \leq |g_l| + C_{22}(C_9 + 1) |u_r - u_l|, \ \forall k \in [\![0, N]\!], \tag{11.39}$$

which provides a bound for g, independent of N.

Proof of item 5

Thanks to (11.27)-(11.28), we get that for all $k \in [\![\widehat{I}(j), \widehat{I}(j+1) - 1]\!]$, $v_k = \min_{i \in [\![\widehat{I}(j)+1, N]\!]} V(g_{\widehat{I}(j)}, u_{\widehat{I}(j)}, u_i)$. Since $|g_k|$ is bounded independently of N for all $k \in [\![0, N]\!]$, we get from (11.26) and (11.39) that for all $s_1, s_2 \in \overline{\mathrm{I}}(u_l, u_r)$,

$$|V(g_k, s_1, s_2)| \leq 2 \frac{C_{\mathrm{Lip}} C_{\max}}{m_0} (|g_k| + 1) \leq 2 \frac{C_{\mathrm{Lip}} C_{\max}}{m_0} \left(|g_l| + C_{22}(C_9 + 1) |u_r - u_l| + 1 \right), \ \forall k \in [\![0, N]\!], \tag{11.40}$$

and the conclusion of the proof follows. □

Thanks to the estimates provided by Lemma 11.3.1, we can now state the following property, from which theorem 11.2.6 follows.

Lemme 11.3.2

Under Hypotheses (11.7), let three reals g_l, u_l, u_r be given. Let $\phi^{(N)}$, $\psi^{(N)}$, $\widehat{\mu}^{(N)}$, $\widehat{\nu}^{(N)}$, $\mu^{(N)}$, $\nu^{(N)}$ and $g^{(N)}$ for all $N \in \mathbb{N}$ with $N \geq 2$ be defined by (11.27)-(11.31). Then there exists a strictly increasing injection $\xi : \mathbb{N} \to \mathbb{N}$, such that the sequences $(\phi^{\xi(N)})_{N \in \mathbb{N}}$, $(\psi^{\xi(N)})_{N \in \mathbb{N}}$, $(\widehat{\mu}^{\xi(N)})_{N \in \mathbb{N}}$, $(\widehat{\nu}^{\xi(N)})_{N \in \mathbb{N}}$, $(\mu^{\xi(N)})_{N \in \mathbb{N}}$, $(\nu^{\xi(N)})_{N \in \mathbb{N}}$ and $(g^{\xi(N)})_{N \in \mathbb{N}}$ converge in the following sense :

1. *$(\phi^{\xi(N)})_{N \in \mathbb{N}}$ (resp. $(\psi^{\xi(N)})_{N \in \mathbb{N}}$) converge in $L^1(\overline{\mathrm{I}}(u_l, u_r))$ to some functions $\phi \in L^\infty(\overline{\mathrm{I}}(u_l, u_r)) \cap BV(\overline{\mathrm{I}}(u_l, u_r))$ (resp. $\psi \in L^\infty(\overline{\mathrm{I}}(u_l, u_r)) \cap BV(\overline{\mathrm{I}}(u_l, u_r))$),*
2. *$(\widehat{\mu}^{\xi(N)})_{N \in \mathbb{N}}$, $(\widehat{\nu}^{\xi(N)})_{N \in \mathbb{N}}$, $(\mu^{\xi(N)})_{N \in \mathbb{N}}$, $(\nu^{\xi(N)})_{N \in \mathbb{N}}$ and $(g^{\xi(N)})_{N \in \mathbb{N}}$ respectively uniformly converge to some Lipschitz continuous functions $\widehat{\mu}$, $\widehat{\nu}$, μ, ν and g, with $g(u_l) = g_l$, and $\mu(u) = -(c(u)g(u) + d(u))$ and $\nu(f(u)) = -(a(u)g(u) + b(u))$ for all $u \in \overline{\mathrm{I}}(u_l, u_r)$.*
3. *$\phi = \widehat{\mu}'$, $\psi = \widehat{\nu}'$ and $\psi(f(u)) = \phi(u)$ for a.e. $u \in \overline{\mathrm{I}}(u_l, u_r)$.*
4. *if $u_l \geq u_r$ (resp. $u_l \leq u_r$), the function $\widehat{\mu}$ is the concave (resp. convex) hull of μ on $\overline{\mathrm{I}}(u_l, u_r)$ and the function $\widehat{\nu}$ is the convex (resp. concave) hull of ν on $\overline{\mathrm{I}}(f(u_l), f(u_r))$.*

Proof.

Proof of item 1

In the case $u_r < u_l$, we see that for all $N \geq 2$, the functions $\phi^{(N)}$ and $\psi^{(N)}$ are respectively non increasing and non decreasing, and bounded independently on N. Hence, thanks to Helly's theorem, we can extract a subsequence such that item 1 holds (the proof is similar in the case $u_r > u_l$).

Proof of item 2

We get from Lemma 11.3.1 that, for all $N \geq 2$, the functions $g^{(N)}$ are Lipschitz continuous with constants independent of N and are bounded independently of N. Thanks to item 5 of lemma 11.3.1, we get that $\phi^{(N)}$ and $\psi^{(N)}$ are bounded independently of N, and the definition of $\widehat{\mu}$ and $\widehat{\nu}$ yields $\widehat{\mu}^{(N)'} = \phi^{(N)}$ and $\widehat{\nu}^{(N)'} = \psi^{(N)}$, which proves that these two functions are Lipschitz continuous on $\overline{\mathrm{I}}(u_l, u_r)$ with constants independent of N and are bounded independently of N. We have, for all $j, k \in [\![0, N]\!]$,

$$\begin{aligned} \left|\mu^{(N)}(u_k) - \mu^{(N)}(u_j)\right| &= \left|c(u_j)g_j - c(u_k)g_k + d(u_j) - d(u_k)\right| \\ &\leq \tfrac{1}{2}\left|c(u_j) - c(u_k)\right|\left|g_k + g_j\right| + \tfrac{1}{2}\left|c(u_j) + c(u_k)\right|\left|g_j - g_k\right| + \left|d(u_j) - d(u_k)\right|. \end{aligned}$$

From Lemma 11.3.1 and the above inequality, we easily deduce that $\mu^{(N)}$ is Lipschitz continuous on $\overline{\mathrm{I}}(u_l, u_r)$ with constant independent of N and is bounded independently of N. The same conclusion clearly holds for $\nu^{(N)}$.

We can therefore apply Ascoli's theorem, extracting a subsequence of that defined in the proof of item 1. Thanks to the definition of $\mu^{(N)}$ and $\nu^{(N)}$, we get that $\mu(u) = -(c(u)g(u) + d(u))$ and $\nu(f(u)) = -(a(u)g(u) + b(u))$ for all $u \in \overline{\mathrm{I}}(u_l, u_r)$.

Proof of item 3

This is an immediate consequence of (11.29).

Proof of item 4

Since the uniform limit of the convex (resp. concave) hull is the convex (resp. concave) hull of the uniform limit, this item is an immediate consequence of Lemma 11.3.1 and of item 2.

□

11.4 Study of a finite volume scheme

We now give a numerical scheme, which applies under Hypotheses (11.7) and (11.8) without restrictions. These hypotheses are therefore assumed in this section. Thanks to the proof of the convergence of the scheme given in this section (theorem 11.4.4), we then get the proof of Theorem 11.2.7.

Let $h > 0$ be given, which will be called the space step in the following. We define a finite volume discretization of $\mathbb{R}$ by $K_i = (ih, (i+1)h)$, for all $i \in \mathbb{Z}$. Let $\delta t > 0$ be given, which will be called the time step in the following. We set

$$u_i^{(0)} = \frac{1}{h}\int_{ih}^{(i+1)h} u_0(x)\mathrm{d}x, \ \forall i \in \mathbb{Z}. \tag{11.41}$$

Thanks to Hypothesis (11.8), the family $(u_i^{(0)})_{i\in\mathbb{Z}}$ is such that there exists $i_0^{(0)} \in \mathbb{Z}$ such that, for all $i \in \mathbb{Z}$ with $i \leq i_0^{(0)}$, $u_i^{(0)} = \overline{u}_0$. We then define the finite volume scheme by induction. Let $n \in \mathbb{N}$ be given and let us assume that $(u_i^{(n)})_{i\in\mathbb{Z}}$ is a given family of reals such that there exists $i_0^{(n)} \in \mathbb{Z}$ verifying

$$u_i^{(n)} = \overline{u}_0, \ \forall i \in \mathbb{Z} \text{ s.t. } i \leq i_0^{(n)}. \tag{11.42}$$

We then define

$$u_{i-\frac{1}{2}}^{(n)} = \overline{u}_0, \ v_{i-\frac{1}{2}}^{(n)} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t} \overline{v}_0(t)\mathrm{d}t, \quad \forall i \in \mathbb{Z} \text{ s.t. } i \leq i_0^{(n)}. \tag{11.43}$$

Let $i \geq i_0^{(n)}$. Let us assume that the values $v_{i-\frac{1}{2}}^{(n)}$ and $u_{i-\frac{1}{2}}^{(n)}$ are known. We define in the following the values $v_{i+\frac{1}{2}}^{(n)}$ and $u_{i+\frac{1}{2}}^{(n)}$, which permits to give the scheme by induction on $i \in \mathbb{Z}$. Let us define the function $\Phi_{i+\frac{1}{2}}^{(n)} : \mathbb{R}\times\mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} \Phi_{i+\frac{1}{2}}^{(n)}(u, v) = \ & f(u_i^{(n)}) + \frac{\delta t}{h}\left(a(u)v + b(u) - a(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - b(u_{i-\frac{1}{2}}^{(n)})\right) \\ & - f\left(u_i^{(n)} + \frac{\delta t}{h}\left(c(u)v + d(u) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right)\right). \end{aligned} \tag{11.44}$$

We remark that this function $\Phi^{(n)}_{i+\frac{1}{2}}$ is locally Lipschitz continuous with respect to its arguments and verifies

$$\partial_2\Phi^{(n)}_{i+\frac{1}{2}}(u,v) = \frac{\delta t}{h}(a(u) - c(u)f'(\widehat{u}^{(n)}_i(u,v))) \in [\frac{\delta t}{h}m_0, \frac{\delta t}{h}C_{\max}(1+C_{\rm Lip})], \quad \text{for a.e. } u,v \in \mathbb{R},$$

with $\widehat{u}^{(n)}_i(u,v) = u^{(n)}_i + \frac{\delta t}{h}\left(c(u)v + d(u) - c(u^{(n)}_{i-\frac{1}{2}})v^{(n)}_{i-\frac{1}{2}} - d(u^{(n)}_{i-\frac{1}{2}})\right)$. Thus, for all $u \in \mathbb{R}$, the function $v \to \Phi^{(n)}_{i+\frac{1}{2}}(u,v)$ is Lipschitz continuous one-to-one from $\mathbb{R}$ to $\mathbb{R}$, and its reciprocal function is Lipschitz continuous from $\mathbb{R}$ to $\mathbb{R}$. Hence we implicitly define the Lipschitz continuous function $g^{(n)}_{i+\frac{1}{2}}$ by

$$g^{(n)}_{i+\frac{1}{2}} \ : \ \mathbb{R} \to \mathbb{R}, \ u \mapsto v \text{ s.t. } \Phi^{(n)}_{i+\frac{1}{2}}(u,v) = \ 0. \tag{11.45}$$

Defining the Godunov flux Godunov (1959); Godlewski and Raviart (1996) $F_{\rm Go}(\mu, s_1, s_2)$, for all $\mu \in C^0(\mathbb{R},\mathbb{R})$ and for all $s_1, s_2 \in \mathbb{R}$, by

$$\begin{cases} \text{if } s_1 \geq s_2 & \text{then } F_{\rm Go}(\mu,s_1,s_2) = \max\limits_{s\in[s_2,s_1]}\mu(s) \\ \text{else if } s_1 < s_2 & \text{then } F_{\rm Go}(\mu,s_1,s_2) = \min\limits_{s\in[s_1,s_2]}\mu(s). \end{cases} \tag{11.46}$$

we define $u^{(n)}_{i+\frac{1}{2}} \in \mathbb{R}$ as a value associated with the Godunov scheme for the flux given by the function

$$\mu^{(n)}_{i+\frac{1}{2}} \ : \ \mathbb{R} \to \mathbb{R}, \ u \mapsto -(c(u)g^{(n)}_{i+\frac{1}{2}}(u) + d(u)), \tag{11.47}$$

the left value $u^{(n)}_i$ and the right value $u^{(n)}_{i+1}$, i.e.

$$u^{(n)}_{i+\frac{1}{2}} \text{ is any element of } \left\{ s \in \overline{\rm I}(u^{(n)}_i, u^{(n)}_{i+1}),\ \mu^{(n)}_{i+\frac{1}{2}}(s) = F_{\rm Go}(\mu^{(n)}_{i+\frac{1}{2}}, u^{(n)}_i, u^{(n)}_{i+1}) \right\}. \tag{11.48}$$

This allows to define the value $v^{(n)}_{i+\frac{1}{2}}$ by :

$$v^{(n)}_{i+\frac{1}{2}} = g^{(n)}_{i+\frac{1}{2}}(u^{(n)}_{i+\frac{1}{2}}). \tag{11.49}$$

Relations (11.44)-(11.45) and the fact that f is strictly decreasing imply that $f(u^{(n)}_{i+\frac{1}{2}})$ reaches the Godunov scheme for the flux given by the function

$$\nu^{(n)}_{i+\frac{1}{2}} \ : \ \mathbb{R} \to \mathbb{R}, \ w \mapsto -\left(a(f^{(-1)}(w))g^{(n)}_{i+\frac{1}{2}}(f^{(-1)}(w)) + b(f^{(-1)}(w))\right), \tag{11.50}$$

the left value $f(u^{(n)}_i)$ and the right value $f(u^{(n)}_{i+1})$, i.e.

$$\nu^{(n)}_{i+\frac{1}{2}}(f(u^{(n)}_{i+\frac{1}{2}})) = F_{\rm Go}(\nu^{(n)}_{i+\frac{1}{2}}, f(u^{(n)}_i), f(u^{(n)}_{i+1})). \tag{11.51}$$

Hence we have defined values $u^{(n)}_{i+\frac{1}{2}}$ and $v^{(n)}_{i+\frac{1}{2}}$, and the induction on $i \in \mathbb{Z}$ with $i \geq i^{(n)}_0$, used in the definition of the scheme, is now complete. These values are such that

$$\begin{aligned} &f(u^{(n)}_i) + \frac{\delta t}{h}\left(a(u^{(n)}_{i+\frac{1}{2}})v^{(n)}_{i+\frac{1}{2}} + b(u^{(n)}_{i+\frac{1}{2}}) - a(u^{(n)}_{i-\frac{1}{2}})v^{(n)}_{i-\frac{1}{2}} - b(u^{(n)}_{i-\frac{1}{2}})\right) \\ &= f\left(u^{(n)}_i + \frac{\delta t}{h}\left(c(u^{(n)}_{i+\frac{1}{2}})v^{(n)}_{i+\frac{1}{2}} + d(u^{(n)}_{i+\frac{1}{2}}) - c(u^{(n)}_{i-\frac{1}{2}})v^{(n)}_{i-\frac{1}{2}} - d(u^{(n)}_{i-\frac{1}{2}})\right)\right). \end{aligned} \tag{11.52}$$

We then set

$$u^{(n+1)}_i = u^{(n)}_i + \frac{\delta t}{h}\left(c(u^{(n)}_{i+\frac{1}{2}})v^{(n)}_{i+\frac{1}{2}} + d(u^{(n)}_{i+\frac{1}{2}}) - c(u^{(n)}_{i-\frac{1}{2}})v^{(n)}_{i-\frac{1}{2}} - d(u^{(n)}_{i-\frac{1}{2}})\right), \ \forall i \in \mathbb{Z}. \tag{11.53}$$

Note that, for all $i \in \mathbb{Z}$ with $i \leq i_0^{(n)} - 1$, we get $u_i^{(n+1)} = \overline{u}_0$. We also denote by

$$w_i^{(n)} = f(u_i^{(n)}),\ \forall i \in \mathbb{Z},\ \forall n \in \mathbb{N}. \tag{11.54}$$

Thanks to (11.52), we can write

$$w_i^{(n+1)} = w_i^{(n)} + \frac{\delta t}{h}\left(a(u_{i+\frac{1}{2}}^{(n)})v_{i+\frac{1}{2}}^{(n)} + b(u_{i+\frac{1}{2}}^{(n)}) - a(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - b(u_{i-\frac{1}{2}}^{(n)})\right),\ \forall i \in \mathbb{Z}. \tag{11.55}$$

We then see that the value

$$i_0^{(n+1)} = i_0^{(n)} - 1 \tag{11.56}$$

is such that (11.42) holds, replacing n by $n+1$, which allows the definition of the scheme by induction on n to hold.

Thanks to the definition of discrete values $u_i^{(n)}$ and $v_{i+\frac{1}{2}}^{(n)}$, for $n \in \mathbb{N}$ and $i \in \mathbb{Z}$, we can define the approximate functions $u_{h,\delta t} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ and $v_{h,\delta t} : \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}$ of u and v by

$$\begin{array}{l} u_{h,\delta t}(x,t) = u_i^{(n)},\ \forall x \in (ih,(i+1)h) \text{ and all } t \in [n\delta t,(n+1)\delta t),\ \forall i \in \mathbb{Z},\ \forall n \in \mathbb{N}, \\ v_{h,\delta t}(x,t) = v_{i-\frac{1}{2}}^{(n)},\ \forall x \in ((i-\frac{1}{2})h,(i+\frac{1}{2})h) \text{ and all } t \in [n\delta t,(n+1)\delta t),\ \forall i \in \mathbb{Z},\ \forall n \in \mathbb{N}. \end{array} \tag{11.57}$$

We then have the following lemmas, which provides the estimates on the approximate solutions, used below in the convergence theorem.

Lemme 11.4.1
Assume that Hypotheses (11.7) and (11.8) hold. Let $h > 0$ and $\delta t > 0$ be given, and, for a given $n \in \mathbb{N}$, let $(u_i^n)_{i\in\mathbb{Z}}$ be a sequence of reals such that (11.42) holds for some $i_0^{(n)} \in \mathbb{Z}$ and such that

$$\sum_{i\in\mathbb{Z}} |u_{i+1}^{(n)} - u_i^{(n)}| \leq W_0, \tag{11.58}$$

and

$$U_m \leq u_i^{(n)} \leq U_M,\ \forall i \in \mathbb{Z},\ \forall n \in \mathbb{N}. \tag{11.59}$$

Let $(w_i^n)_{i\in\mathbb{Z}}$, $(u_{i+\frac{1}{2}}^n)_{i\in\mathbb{Z}}$, $(v_{i+\frac{1}{2}}^n)_{i\in\mathbb{Z}}$, $(\mu_{i+\frac{1}{2}}^n)_{i\in\mathbb{Z}}$ and $(\nu_{i+\frac{1}{2}}^{(n)})_{i\in\mathbb{Z}}$ be defined from the values $(u_i^n)_{i\in\mathbb{Z}}$ by (11.43) - (11.55).

Then there exists C_{11}, C_{12}, C_{13}, C_{14} and C_{15}, only depending on $\overline{V}_0$, $\overline{u}_0$, U_m, U_M, W_0, $C_{\max}$, C_{Lip}, f_m, and m_0 such that

$$|v_{i+\frac{1}{2}}^{(n)}| \leq C_{11},\ \forall i \in \mathbb{Z}, \tag{11.60}$$

$$\|g_{i+\frac{1}{2}}^{(n)}\|_{L^\infty(U_m,U_M)} \leq C_{12},\ \forall i \in \mathbb{Z}, \tag{11.61}$$

$$\|(g_{i+\frac{1}{2}}^{(n)})'\|_{L^\infty(U_m,U_M)} \leq C_{13},\ \forall i \in \mathbb{Z}, \tag{11.62}$$

$$\|(\mu_{i+\frac{1}{2}}^{(n)})'\|_{L^\infty(U_m,U_M)} \leq C_{14},\ \forall i \in \mathbb{Z}, \tag{11.63}$$

$$\|(\nu_{i+\frac{1}{2}}^{(n)})'\|_{L^\infty(U_m,U_M)} \leq C_{15},\ \forall i \in \mathbb{Z}. \tag{11.64}$$

Proof. For a given $i \in \mathbb{Z}$, we set

$$\tau_{i,f}^{(n)} = \frac{f(u_i^{(n+1)}) - f(u_i^{(n)})}{u_i^{(n+1)} - u_i^{(n)}} \text{ if } u_i^{(n+1)} \neq u_i^{(n)} \text{ else } \tau_{i,f}^{(n)} = -f_m.$$

From (11.44)-(11.45) and definition (11.53) of u_i^{n+1}, we get

$$v_{i+\frac{1}{2}}^{(n)} = \frac{\left(a(u_{i-\frac{1}{2}}^{(n)}) - c(u_{i-\frac{1}{2}}^{(n)})\tau_{i,f}^{(n)}\right)v_{i-\frac{1}{2}}^{(n)} + b(u_{i-\frac{1}{2}}^{(n)}) - b(u_{i+\frac{1}{2}}^{(n)}) - (d(u_{i-\frac{1}{2}}^{(n)}) - d(u_{i+\frac{1}{2}}^{(n)}))\tau_{i,f}^{(n)}}{a(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i+\frac{1}{2}}^{(n)})\tau_{i,f}^{(n)}}.$$

Using again $C_{\rm Lip}$ defined by (11.7), we define

$$\tau_{i,a}^{(n)} = \frac{a(u_{i+\frac{1}{2}}^{(n)}) - a(u_{i-\frac{1}{2}}^{(n)})}{u_{i+\frac{1}{2}}^{(n)} - u_{i-\frac{1}{2}}^{(n)}} \text{ if } u_{i+\frac{1}{2}}^{(n)} \neq u_{i-\frac{1}{2}}^{(n)} \text{ else } \tau_{i,a}^{(n)} = C_{\rm Lip},$$

and we define similarly $\tau_{i,b}^{(n)}$, $\tau_{i,c}^{(n)}$ and $\tau_{i,d}^{(n)}$, replacing respectively a by b, c and d.

We can then write $v_{i+\frac{1}{2}}^{(n)} - v_{i-\frac{1}{2}}^{(n)} = (u_{i+\frac{1}{2}}^{(n)} - u_{i-\frac{1}{2}}^{(n)})(G_0 v_{i-\frac{1}{2}}^{(n)} + G_1)$ with

$$G_0 = \frac{\tau_{i,c}^{(n)}\tau_{i,f}^{(n)} - \tau_{i,a}^{(n)}}{a(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i+\frac{1}{2}}^{(n)})\tau_{i,f}^{(n)}} \text{ and } G_1 = \frac{\tau_{i,d}^{(n)}\tau_{i,f}^{(n)} - \tau_{i,b}^{(n)}}{a(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i+\frac{1}{2}}^{(n)})\tau_{i,f}^{(n)}},$$

and therefore, since $a(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i+\frac{1}{2}}^{(n)})\tau_{i,f}^{(n)} \geq m_0$, we get $|G_0| \leq C_{16}$ and $|G_1| \leq C_{16}$ with $C_{16} = \frac{C_{\rm Lip}^2 + C_{\rm Lip}}{m_0}$. We then have

$$\left|v_{i+\frac{1}{2}}^{(n)} - v_{i-\frac{1}{2}}^{(n)}\right| \leq C_{16}\left|u_{i+\frac{1}{2}}^{(n)} - u_{i-\frac{1}{2}}^{(n)}\right| \left(\left|v_{i-\frac{1}{2}}^{(n)}\right| + 1\right), \ \forall i \in \mathbb{Z}.$$

We can the apply the discrete Gronwall lemma 11.5.3 (see appendix 2), starting from $i \geq i_0^{(n)}$. We thus obtain

$$\forall i \in \mathbb{Z},\ i \geq i_0^{(n)} \quad \left|v_{i+\frac{1}{2}}^{(n)}\right| \leq \left(\left|\overline{V}_0\right| + 1\right) \exp\left(C_{16} \sum_{j=i_0^{(n)}}^{i} \left|u_{j+\frac{1}{2}}^{n} - u_{j-\frac{1}{2}}^{n}\right|\right) - 1.$$

Hence, thanks to (11.58) and (11.43), we have

$$\left|v_{i+\frac{1}{2}}^{(n)}\right| \leq \left(\left|\overline{V}_0\right| + 1\right) \exp\left(C_{16} W_0\right) - 1, \ \forall i \in \mathbb{Z},$$

which provides (11.60) with C_{11} defined by the right hand side of the above inequality. From definition (11.44)-(11.45), setting $\widehat{u}_i^{(n)} : \mathbb{R} \to \mathbb{R},\ s \mapsto u_i^{(n)} + \frac{\delta t}{h}\left(c(s)g_{i+\frac{1}{2}}^{(n)}(s) + d(s) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right)$, we get by derivation, for a.e. $u \in \mathbb{R}$,

$$\left(g_{i+\frac{1}{2}}^{(n)}\right)'(u) = \frac{f'(\widehat{u}_i^{(n)}(u))(c'(u)g_{i+\frac{1}{2}}^{(n)}(u) + d'(u)) - a'(u)g_{i+\frac{1}{2}}^{(n)}(u) - b'(u)}{a(u) - c(u)f'(\widehat{u}_i^{(n)}(u))},$$

which gives

$$\left|\left(g_{i+\frac{1}{2}}^{(n)}\right)'(u)\right| \leq C_{16}(|g_{i+\frac{1}{2}}^{(n)}(u)| + 1).$$

We then apply Gronwall's lemma, starting from $u = u_{i+\frac{1}{2}}^{(n)} \in [U_m, U_M]$ thanks to (11.59). Since we have $|g_{i+\frac{1}{2}}^{(n)}(u_{i+\frac{1}{2}}^{(n)})| = |v_{i+\frac{1}{2}}^{(n)}| \leq C_{11}$ by (11.60), we get (11.61) with $C_{12} = (C_{11}+1)\exp(C_{16}(U_M - U_m)) - 1$. This therefore gives (11.62) with $C_{13} = C_{16}(C_{12}+1)$. We thus get (11.63), with $C_{14} = C_{\rm Lip}C_{12} + C_{\max}C_{13} + C_{\rm Lip}$ and (11.64) with $C_{15} = C_{14}/m_0$. □

Lemme 11.4.2
Assume that Hypotheses (11.7) and (11.8) hold. Let C_{17} be defined by the relation

$$C_{17} = \frac{1}{2\max(C_{14}, C_{15})}, \tag{11.65}$$

where C_{14} and C_{15} are given by Lemma 11.4.1.

Then, for all $h > 0$ and $\delta t > 0$ be given such that

$$\delta t \leq C_{17} h, \tag{11.66}$$

the values $(u_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, $(w_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, $(u_{i+\frac{1}{2}}^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$ *and* $(v_{i+\frac{1}{2}}^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, *defined by (11.41) - (11.55), satisfy*

$$\sum_{i\in\mathbb{Z}} |u_{i+1}^{(n)} - u_i^{(n)}| \leq W_0, \ \forall n \in \mathbb{N}, \tag{11.67}$$

$$U_m \leq u_i^{(n)} \leq U_M, \ \forall i \in \mathbb{Z}, \ \forall n \in \mathbb{N}. \tag{11.68}$$

Proof. We first remark that (11.67) and (11.68) hold for $n = 0$ using definition (11.41). Let us assume that, for a given $n \in \mathbb{N}$, relations (11.67) and (11.68) hold. By the construction of the scheme, (11.42) holds, thus we can use the conclusions of Lemma 11.4.1. For a given $i \in \mathbb{Z}$, we define $\widetilde{u}_i^{(n+1)}$ by

$$\widetilde{u}_i^{(n+1)} = u_i^{(n)} - \frac{\delta t}{h}\left(\mu_{i+\frac{1}{2}}^{(n)}(u_{i+\frac{1}{2}}^{(n)}) - \mu_{i+\frac{1}{2}}^{(n)}(u_i^{(n)}) + \mu_{i-\frac{1}{2}}^{(n)}(u_i^{(n)}) - \mu_{i-\frac{1}{2}}^{(n)}(u_{i-\frac{1}{2}}^{(n)}) \right).$$

Using again the notations $w_i^{(n)} = f(u_i^{(n)})$ and $w_{i+\frac{1}{2}}^{(n)} = f(u_{i+\frac{1}{2}}^{(n)})$ for all $i \in \mathbb{Z}$, we define $\widetilde{w}_i^{(n+1)}$, for all $i \in \mathbb{Z}$, by

$$\widetilde{w}_i^{(n+1)} = w_i^{(n)} - \frac{\delta t}{h}\left(\nu_{i+\frac{1}{2}}^{(n)}(w_{i+\frac{1}{2}}^{(n)}) - \nu_{i+\frac{1}{2}}^{(n)}(w_i^{(n)}) + \nu_{i-\frac{1}{2}}^{(n)}(w_i^{(n)}) - \nu_{i-\frac{1}{2}}^{(n)}(w_{i-\frac{1}{2}}^{(n)}) \right).$$

Note that, in general, $\widetilde{w}_i^{(n+1)} \neq f(\widetilde{u}_i^{(n+1)})$. We now define $\tau_{i,+}^{(n,u)}$ by

$$\text{if } u_i^{(n)} \neq u_{i+\frac{1}{2}}^{(n)}, \text{ then } \tau_{i,+}^{(n,u)} = \frac{\mu_{i+\frac{1}{2}}^{(n)}(u_i^{(n)}) - \mu_{i+\frac{1}{2}}^{(n)}(u_{i+\frac{1}{2}}^{(n)})}{u_{i+\frac{1}{2}}^{(n)} - u_i^{(n)}}, \text{ else } \tau_{i,+}^{(n,u)} = C_{14}.$$

Note that, thanks to (11.48) and (11.63), we have $0 \leq \tau_{i,+}^{(n,u)} \leq C_{14}$. Similarly, we define $\tau_{i,-}^{(n,u)}$ by

$$\text{if } u_i^{(n)} \neq u_{i-\frac{1}{2}}^{(n)}, \text{ then } \tau_{i,-}^{(n,u)} = \frac{\mu_{i-\frac{1}{2}}^{(n)}(u_{i-\frac{1}{2}}^{(n)}) - \mu_{i-\frac{1}{2}}^{(n)}(u_i^{(n)})}{u_{i-\frac{1}{2}}^{(n)} - u_i^{(n)}}, \text{ else } \tau_{i,-}^{(n,u)} = C_{14}$$

(again, thanks to (11.48) and (11.63), we have $0 \leq \tau_{i,-}^{(n,u)} \leq C_{14}$). We now define $\tau_{i,+}^{(n,w)}$ by

$$\text{if } w_i^{(n)} \neq w_{i+\frac{1}{2}}^{(n)}, \text{ then } \tau_{i,+}^{(n,w)} = \frac{\nu_{i+\frac{1}{2}}^{(n)}(w_i^{(n)}) - \nu_{i+\frac{1}{2}}^{(n)}(w_{i+\frac{1}{2}}^{(n)})}{w_{i+\frac{1}{2}}^{(n)} - w_i^{(n)}}, \text{ else } \tau_{i,+}^{(n,u)} = C_{15}.$$

Note that, thanks to (11.51) and (11.64), we have $0 \leq \tau_{i,+}^{(n,w)} \leq C_{15}$. Similarly, we define $\tau_{i,-}^{(n,w)}$ by

$$\text{if } w_i^{(n)} \neq w_{i-\frac{1}{2}}^{(n)}, \text{ then } \tau_{i,-}^{(n,w)} = \frac{\nu_{i-\frac{1}{2}}^{(n)}(w_{i-\frac{1}{2}}^{(n)}) - \nu_{i-\frac{1}{2}}^{(n)}(w_i^{(n)})}{w_{i-\frac{1}{2}}^{(n)} - w_i^{(n)}}, \text{ else } \tau_{i,-}^{(n,w)} = C_{15}$$

(again, thanks to (11.51) and (11.64), we have $0 \leq \tau_{i,-}^{(n,w)} \leq C_{15}$). We then get

$$\widetilde{u}_i^{(n+1)} = u_i^{(n)}(1 - \frac{\delta t}{h}(\tau_{i,+}^{(n,u)} + \tau_{i,-}^{(n,u)})) + \frac{\delta t}{h}\tau_{i,+}^{(n,u)}u_{i+\frac{1}{2}}^{(n)} + \frac{\delta t}{h}\tau_{i,-}^{(n,u)}u_{i-\frac{1}{2}}^{(n)}, \tag{11.69}$$

and

$$\widetilde{w}_i^{(n+1)} = w_i^{(n)}(1 - \frac{\delta t}{h}(\tau_{i,+}^{(n,w)} + \tau_{i,-}^{(n,w)})) + \frac{\delta t}{h}\tau_{i,+}^{(n,w)}w_{i+\frac{1}{2}}^{(n)} + \frac{\delta t}{h}\tau_{i,-}^{(n,w)}w_{i-\frac{1}{2}}^{(n)}. \tag{11.70}$$

Defining C_{17} by (11.65), we obtain from (11.69) that

$$\widetilde{u}_i^{(n+1)} \in [\min(u_{i-\frac{1}{2}}^{(n)}, u_i^{(n)}, u_{i+\frac{1}{2}}^{(n)}), \max(u_{i-\frac{1}{2}}^{(n)}, u_i^{(n)}, u_{i+\frac{1}{2}}^{(n)})] \tag{11.71}$$

and from (11.70) that

$$\begin{aligned}\widetilde{w}_i^{(n+1)} \in\ & [\min(w_{i-\frac12}^{(n)}, w_i^{(n)}, w_{i+\frac12}^{(n)}), \max(w_{i-\frac12}^{(n)}, w_i^{(n)}, w_{i+\frac12}^{(n)})] \\ & = [f(\max(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)})), f(\min(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)}))].\end{aligned} \tag{11.72}$$

We now define the function $M_i^{(n)} : [U_m, U_M] \to \mathbb{R}$, $u \mapsto a(u_i^{(n)})u - c(u_i^{(n)})f(u)$. We then get from (11.71) and (11.72) that

$$M_i^{(n)}\left(\min(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)})\right) \le a(u_i^{(n)})\widetilde{u}_i^{(n+1)} - c(u_i^{(n)})\widetilde{w}_i^{(n+1)} \le M_i^{(n)}\left(\max(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)})\right). \tag{11.73}$$

From the definition of $\widetilde{u}_i^{(n+1)}$ and $\widetilde{w}_i^{(n+1)}$ and from (11.53)-(11.55), we get

$$M_i^{(n)}(u_i^{(n+1)}) = a(u_i^{(n)})\widetilde{u}_i^{(n+1)} - c(u_i^{(n)})\widetilde{w}_i^{(n+1)}, \tag{11.74}$$

since $a(u)\mu_{i+\frac12}^{(n)}(u) - c(u)\nu_{i+\frac12}^{(n)}(f(u)) = a(u)\mu_{i-\frac12}^{(n)}(u) - c(u)\nu_{i-\frac12}^{(n)}(f(u)) = -a(u)d(u) + c(u)b(u)$, for all $u \in [U_m, U_M]$. Thanks to (11.73) and (11.74), and to the strict monotony of the function $M_i^{(n)}$, we can write

$$u_i^{(n+1)} \in [\min(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)}), \max(u_{i-\frac12}^{(n)}, u_i^{(n)}, u_{i+\frac12}^{(n)})],$$

which shows (11.68) for $n+1$. Thanks to the above property, we can now apply Lemma 11.5.4, which permits to conclude

$$\sum_{i\in\mathbb{Z}} \left|u_i^{(n+1)} - u_{i+1}^{(n+1)}\right| \le \sum_{i\in\mathbb{Z}} \left|u_i^{(n)} - u_{i+1}^{(n)}\right|,$$

which completes the proof of (11.67) for $n+1$. □

Let us now classically deduce some bound for the variation in time, from that of the variation in space.

Lemme 11.4.3
Assume that Hypotheses (11.7) and (11.8) hold. Let C_{17} be defined by Lemma 11.4.2 (thus only depending on $\overline{V}_0$, $\overline{u}_0$, W_0, U_m, U_M, a, b, c, d and f). Let $h > 0$ and $\delta t > 0$ be given such that (11.66) holds. Then, there exists C_{18}, only depending on $\overline{V}_0$, $\overline{u}_0$, W_0, U_m, U_M, a, b, c, d and f such that the values $(u_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, $(w_i^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, $(u_{i+\frac12}^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$ and $(v_{i+\frac12}^n)_{i\in\mathbb{Z},n\in\mathbb{N}}$, defined by (11.43) - (11.55), satisfy

$$\sum_{n=\lfloor t_1/\delta t\rfloor}^{\lfloor t_2/\delta t\rfloor} \sum_{i\in\mathbb{Z}} h|u_i^{(n+1)} - u_i^{(n)}| \le (t_2 - t_1 + \delta t)C_{18},\ \forall t_2 \ge t_1 \ge 0, \tag{11.75}$$

where, for all $s \in \mathbb{R}$, we denote by $\lfloor s\rfloor$ the biggest integer lower or equal to s.

Proof. We get, setting $u = u_{i-\frac12}^{(n)}$ in (11.44)-(11.45), that $g_{i+\frac12}^{(n)}(u_{i-\frac12}^{(n)}) = g_{i-\frac12}^{(n)}(u_{i-\frac12}^{(n)})$. Using (11.53), we get

$$u_i^{(n+1)} = u_i^{(n)} + \frac{\delta t}{h}\left(c(u_{i+\frac12}^{(n)})g_{i+\frac12}^{(n)}(u_{i+\frac12}^{(n)}) + d(u_{i+\frac12}^{(n)}) - c(u_{i-\frac12}^{(n)})g_{i+\frac12}^{(n)}(u_{i-\frac12}^{(n)}) - d(u_{i-\frac12}^{(n)})\right),\ \forall i \in \mathbb{Z}.$$

Using the Lipschitz constant C_{14} defined in the proof of Lemma 11.4.2, this leads to

$$|u_i^{(n+1)} - u_i^{(n)}| \le \frac{\delta t}{h}C_{14}|u_{i+\frac12}^{(n)} - u_{i-\frac12}^{(n)}|,\ \forall i \in \mathbb{Z}.$$

Therefore, we get

$$\sum_{i\in\mathbb{Z}} h|u_i^{(n+1)} - u_i^{(n)}| \le \delta t C_{14}W_0,$$

which provides (11.75) with $C_{18} = C_{14}W_0$. □

It is now possible to state the convergence of the scheme. This is the aim of the following theorem.

Théorème 11.4.4
Assume that Hypotheses (11.7) and (11.8) hold. Let C_{17} be defined by Lemma 11.4.2 (thus only depending on $\overline{V}_0$, $\overline{u}_0$, W_0, U_m, U_M, a, b, c, d and f) and let $C_{19} \in (0, C_{17})$ be given. Let $(h_m, \delta t_m)_{m\in\mathbb{N}}$ be a sequence of pairs of positive reals such that $\lim_{m\to\infty} h_m = 0$ and such that $C_{19}h_m \leq \delta t_m \leq C_{17}h_m$ for all $m \in \mathbb{N}$. Then there exists a subsequence of $(h_m, \delta t_m)_{m\in\mathbb{N}}$, again denoted $(h_m, \delta t_m)_{m\in\mathbb{N}}$, such that the sequence of functions $(u_{h_m,\delta t_m}, v_{h_m,\delta t_m})_{m\in\mathbb{N}}$ defined by (11.43) - (11.57) satisfies

1. *there exists $u \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ with $u - u_0 \in \mathrm{Lip}(\mathbb{R}_+; L^1(\mathbb{R})) \cap L^\infty(\mathbb{R}_+; BV(\mathbb{R}))$ such that*
$$\forall T \in \mathbb{R}_+, \ \lim_{m\to\infty} \sup_{t\in[0,T]} \|u_{h_m,\delta t_m}(\cdot,t) - u(\cdot,t)\|_{L^1(\mathbb{R})} = 0, \tag{11.76}$$
2. *there exists $v \in L^\infty(\mathbb{R}\times\mathbb{R}_+)$ such that $(v_{h_m,\delta t_m})_{m\in\mathbb{N}}$ converges for the weak-$\star$ topology of $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ to v,*
3. *this pair of functions (u, v) is then an entropy weak solution of the system (11.9) in the sense of definition 11.2.3.*

Proof. Let us examine the first item of the above theorem. It is easy to see that, for all $m \in \mathbb{N}$, $\|u_{h_m,\delta t_m}(\cdot,0) - u_0\|_{L^1(\mathbb{R})} \leq h_m W_0$ thanks to (11.41). Thanks to Lemma 11.4.3, we get that, for all $t \in \mathbb{R}_+$, $\|u_{h_m,\delta t_m}(\cdot,t) - u_0\|_{L^1(\mathbb{R})} \leq h_m W_0 + (t+\delta t_m)C_{18}$, which is bounded by $tC_{18} + C_{20}$ for all $m \in \mathbb{N}$. Thanks to Lemma 11.4.2, we have $\|u_{h_m,\delta t_m}(\cdot,t) - u_0\|_{BV(\mathbb{R})} \leq 2W_0$. Helly's theorem proves that the set $A(t)$ of all the functions $v \in L^1(\mathbb{R})$ such that $\|v\|_{L^1(\mathbb{R})} \leq tC_{18} + C_{20}$ and $\|v\|_{BV(\mathbb{R})} \leq 2W_0$ is relatively compact in $L^1(\mathbb{R})$. We can then apply theorem 11.5.5 to the sequence $(u_{h_m,\delta t_m} - u_0)_{m\in\mathbb{N}}$, which allows to extract a sequence such that (11.76) holds.

Using (11.60), we again extract from this sequence a subsequence such that the second item holds.

Let us now prove the third item. Let $\varphi \in C_c^\infty(\mathbb{R}\times\mathbb{R}_+, \mathbb{R})$ be given. Let $R > 0$ and $T > 0$ such that $\mathrm{support}(\varphi) \subset [-R, R]\times[0, T]$. We get (11.12), using (11.43), (11.56) and the hypothesis $C_{19}h_m \leq \delta t_m$, for all $m \in \mathbb{N}$ (this hypothesis is only needed here).

Let us now show (11.11). Let $\varphi \in C_c^\infty(\mathbb{R}\times\mathbb{R}_+, \mathbb{R}_+)$ be given, and let $m \in \mathbb{N}$. For the simplicity of the notation, we omit some subscripts m in the following calculations. For a given $n \in \mathbb{N}$, relations (11.53)-(11.55) read

$$u_i^{(n+1)} = u_i^{(n)} - \frac{\delta t}{h}\left(F_{\mathrm{Go}}(\mu_{i+\frac12}^{(n)}, u_i^{(n)}, u_{i+1}^{(n)}) - F_{\mathrm{Go}}(\mu_{i-\frac12}^{(n)}, u_{i-1}^{(n)}, u_i^{(n)})\right), \ \forall i \in \mathbb{Z},$$

and

$$f(u_i^{(n+1)}) = f(u_i^{(n)}) - \frac{\delta t}{h}\left(F_{\mathrm{Go}}(\nu_{i+\frac12}^{(n)}, f(u_i^{(n)}), f(u_{i+1}^{(n)})) - F_{\mathrm{Go}}(\nu_{i-\frac12}^{(n)}, f(u_{i-1}^{(n)}), f(u_i^{(n)}))\right), \ \forall i \in \mathbb{Z}.$$

Let $\kappa \in \mathbb{R}$ be given. We get from the above relations, for all $i \in \mathbb{Z}$,

$$\begin{aligned} a(\kappa)u_i^{(n+1)} - c(\kappa)f(u_i^{(n+1)}) = \ & a(\kappa)u_i^{(n)} - c(\kappa)f(u_i^{(n)}) \\ & -a(\kappa)\frac{\delta t}{h}\left(F_{\mathrm{Go}}(\mu_{i+\frac12}^{(n)}, u_i^{(n)}, u_{i+1}^{(n)}) - F_{\mathrm{Go}}(\mu_{i-\frac12}^{(n)}, u_{i-1}^{(n)}, u_i^{(n)})\right) \\ & +c(\kappa)\frac{\delta t}{h}\left(F_{\mathrm{Go}}(\nu_{i+\frac12}^{(n)}, f(u_i^{(n)}), f(u_{i+1}^{(n)})) - F_{\mathrm{Go}}(\nu_{i-\frac12}^{(n)}, f(u_{i-1}^{(n)}), f(u_i^{(n)}))\right). \end{aligned}$$

We thus have the relation $\Psi(u_i^{(n+1)}) = \Phi(u_{i-1}^{(n+1)}, u_i^{(n)}, u_{i+1}^{(n)})$, where the functions Ψ and Φ are such that, for all $s \in \mathbb{R}$, $\Psi(s) = a(\kappa)s - c(\kappa)f(s)$ and

$$\begin{aligned} \forall s_1, s_2, s_3 \in \mathbb{R}, \ \Phi(s_1, s_2, s_3) = \ & a(\kappa)s_2 - c(\kappa)f(s_2) \\ & -a(\kappa)\frac{\delta t}{h}\left(F_{\mathrm{Go}}(\mu_{i+\frac12}^{(n)}, s_2, s_3) - F_{\mathrm{Go}}(\mu_{i-\frac12}^{(n)}, s_1, s_2)\right) \\ & +c(\kappa)\frac{\delta t}{h}\left(F_{\mathrm{Go}}(\nu_{i+\frac12}^{(n)}, f(s_2), f(s_3)) - F_{\mathrm{Go}}(\nu_{i-\frac12}^{(n)}, f(s_1), f(s_2))\right). \end{aligned}$$

Thanks to the relation

$$a(\kappa)\left(\mu_{i+\frac12}^{(n)}(\kappa) - \mu_{i-\frac12}^{(n)}(\kappa)\right) = c(\kappa)\left(\nu_{i+\frac12}^{(n)}(f(\kappa)) - \nu_{i-\frac12}^{(n)}(f(\kappa))\right),$$

we have $\Psi(\kappa) = \Phi(\kappa,\kappa,\kappa)$. We recall that, when $g \in C^0(\mathbb{R},\mathbb{R})$, the function $F_{\rm Go}(g,\cdot,\cdot)$ defined by (11.46) is Lipschitz continuous and non decreasing with respect to its second argument, Lipschitz continuous and non increasing with respect to its third argument, with the same Lipschitz constants as g. We then see that Φ is non decreasing with respect to its first and third arguments in $[U_m,U_M]^2$. Thanks to condition (11.66) in which C_{17} is given by (11.65), where C_{14} and C_{15} are respectively the Lipschitz constants of $\mu^{(n)}_{i\pm\frac12}$ and $\nu^{(n)}_{i\pm\frac12}$, we get that Φ is also non decreasing with respect to its second argument in $[U_m,U_M]$. Following a classical reasoning, we get that, if $\kappa \in [U_m,U_M]$, then

$$\Psi(u_i^{(n+1)}\top\kappa) \le \Phi(u_{i-1}^{(n)}\top\kappa, u_i^{(n)}\top\kappa, u_{i+1}^{(n)}\top\kappa).$$

This relation holds if $\kappa \ge U_M$ thanks to $\Psi(\kappa) = \Phi(\kappa,\kappa,\kappa)$, and it holds if $\kappa \le U_m$ thanks to $\Psi(u_i^{(n+1)}) = \Phi(u_{i-1}^{(n+1)}, u_i^{(n)}, u_{i+1}^{(n)})$. Hence it holds in the general case of $\kappa \in \mathbb{R}$, and we can write $\Psi(u_i^{(n+1)}\top\kappa) - \Psi(\kappa) \le \Phi(u_{i-1}^{(n)}\top\kappa, u_i^{(n)}\top\kappa, u_{i+1}^{(n)}\top\kappa) - \Phi(\kappa,\kappa,\kappa)$. This gives

$$\begin{aligned}
&a(\kappa)(u_i^{(n+1)}\top\kappa-\kappa) - c(\kappa)(f(u_i^{(n+1)}\top\kappa) - f(\kappa)) \le a(\kappa)(u_i^{(n)}\top\kappa-\kappa) - c(\kappa)f(u_i^{(n)}\top\kappa) - f(\kappa))\\
&\quad -a(\kappa)\frac{\delta t}{h}\left(F_{\rm Go}(\mu^{(n)}_{i+\frac12}, u_i^{(n)}\top\kappa, u_{i+1}^{(n)}\top\kappa) - F_{\rm Go}(\mu^{(n)}_{i-\frac12}, u_{i-1}^{(n)}\top\kappa, u_i^{(n)}\top\kappa)\right)\\
&+c(\kappa)\frac{\delta t}{h}\left(F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(u_i^{(n)}\top\kappa), f(u_{i+1}^{(n)}\top\kappa)) - F_{\rm Go}(\nu^{(n)}_{i-\frac12}, f(u_{i-1}^{(n)}\top\kappa), f(u_i^{(n)}\top\kappa))\right).
\end{aligned} \tag{11.77}$$

Thanks to the property $F_{\rm Go}(g,s,s) = g(s)$ for all $g \in C^0(\mathbb{R},\mathbb{R})$ and $s \in \mathbb{R}$, we have

$$\begin{aligned}
&a(\kappa)\left(F_{\rm Go}(\mu^{(n)}_{i+\frac12}, s\top\kappa, s\top\kappa) - F_{\rm Go}(\mu^{(n)}_{i+\frac12}, \kappa, \kappa)\right) -\\
&c(\kappa)\left(F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(s\top\kappa), f(s\top\kappa)) - F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(\kappa), f(\kappa))\right) = -G_0(s)g^{(n)}_{i+\frac12}(s) - G_1(s),\ \forall s\in\mathbb{R},
\end{aligned} \tag{11.78}$$

where we define

$$\begin{aligned}
G_0 &: \mathbb{R}\to\mathbb{R}, \quad s \mapsto a(\kappa)c(s\top\kappa) - c(\kappa)a(s\top\kappa),\\
G_1 &: \mathbb{R}\to\mathbb{R}, \quad s \mapsto a(\kappa)(d(s\top\kappa) - d(\kappa)) - c(\kappa)(b(s\top\kappa) - b(\kappa)),
\end{aligned}$$

where we remark that for all $s\in\mathbb{R}$, $-G_0(s)g^{(n)}_{i+\frac12}(s\top\kappa) - G_1(s) = -G_0(s)g^{(n)}_{i+\frac12}(s) - G_1(s)$, since $G_0(s) = G_1(s) = 0$ for all $s < \kappa$. Note that the following relation holds :

$$\begin{aligned}
&a(\kappa)F_{\rm Go}(\mu^{(n)}_{i+\frac12},\kappa,\kappa) - c(\kappa)F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(\kappa), f(\kappa)) = -a(\kappa)d(\kappa) + c(\kappa)b(\kappa)\\
&= a(\kappa)F_{\rm Go}(\mu^{(n)}_{i-\frac12},\kappa,\kappa) - c(\kappa)F_{\rm Go}(\nu^{(n)}_{i-\frac12}, f(\kappa), f(\kappa)).
\end{aligned} \tag{11.79}$$

We then get from (11.77), (11.78) and (11.79)

$$\begin{aligned}
&h\left(a(\kappa)(u_i^{(n+1)}\top\kappa-\kappa) - c(\kappa)(f(u_i^{(n+1)}\top\kappa) - f(\kappa)) - a(\kappa)(u_i^{(n)}\top\kappa-\kappa) + c(\kappa)(f(u_i^{(n)}\top\kappa) - f(\kappa))\right)\\
&-\delta t\left(G_0(u_{i+\frac12}^{(n)})v_{i+\frac12}^{(n)} + G_1(u_{i+\frac12}^{(n)}) - G_0(u_{i-\frac12}^{(n)})v_{i-\frac12}^{(n)} - G_1(u_{i-\frac12}^{(n)})\right) \le \delta t(W_i^n - W_{i-1}^{(n)}),
\end{aligned} \tag{11.80}$$

where we set

$$\begin{aligned}
W_i^n = \quad &-a(\kappa)\left(F_{\rm Go}(\mu^{(n)}_{i+\frac12}, u_i^{(n)}\top\kappa, u_{i+1}^{(n)}\top\kappa) - F_{\rm Go}(\mu^{(n)}_{i+\frac12}, u_{i+\frac12}^{(n)}\top\kappa, u_{i+\frac12}^{(n)}\top\kappa)\right)\\
&+c(\kappa)\left(F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(u_i^{(n)}\top\kappa), f(u_{i+1}^{(n)}\top\kappa)) - F_{\rm Go}(\nu^{(n)}_{i+\frac12}, f(u_{i+\frac12}^{(n)}\top\kappa), f(u_{i+\frac12}^{(n)}\top\kappa))\right).
\end{aligned} \tag{11.81}$$

We then multiply (11.80) by $\varphi((i+\frac12)h, n\delta t)$. Gathering the terms obtained by summing the result on $n\in\mathbb{N}$ and $i\in\mathbb{Z}$, we then obtain $T_{18}^{(m)} + T_{19}^{(m)} + T_{20}^{(m)} \le T_{21}^{(m)}$, with

$$T_{18}^{(m)} = -\sum_{n\in\mathbb{N}^\star}\sum_{i\in\mathbb{Z}} h\left(a(\kappa)(u_i^{(n)}\top\kappa-\kappa) - c(\kappa)(f(u_i^{(n)}\top\kappa) - f(\kappa))\right)(\varphi((i+\frac12)h, n\delta t) - \varphi((i+\frac12)h, (n-1)\delta t),$$

$$T_{19}^{(m)} = -\sum_{i\in\mathbb{Z}} h\left(a(\kappa)(u_i^{(0)}\top\kappa-\kappa)-c(\kappa)(f(u_i^{(0)}\top\kappa)-f(\kappa))\right)\varphi((i+\frac{1}{2})h,0),$$

$$T_{20}^{(m)} = \sum_{n\in\mathbb{N}^\star}\delta t\sum_{i\in\mathbb{Z}}\left(G_0(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)}+G_1(u_{i-\frac{1}{2}}^{(n)})\right)\left(\varphi((i+\frac{1}{2})h,n\delta t)-\varphi((i-\frac{1}{2})h,n\delta t)\right),$$

and

$$T_{21}^{(m)} = \sum_{n\in\mathbb{N}^\star}\delta t\sum_{i\in\mathbb{Z}} W_{i-1}^{(n)}\left(\varphi((i-\frac{1}{2})h,n\delta t)-\varphi((i+\frac{1}{2})h,n\delta t)\right)$$

We then classically obtain that

$$\lim_{m\to\infty} T_{18}^{(m)} = -\int_{\mathbb{R}_+}\int_{\mathbb{R}}(a(\kappa)(u(x,t)\top\kappa-\kappa)-c(\kappa)(f(u(x,t)\top\kappa)-f(\kappa)))\,\varphi_t(x,t)\mathrm{d}x\mathrm{d}t,$$

$$\lim_{m\to\infty} T_{19}^{(m)} = -\int_{\mathbb{R}}(a(\kappa)(u_0(x)\top\kappa-\kappa)-c(\kappa)(f(u_0(x)\top\kappa)-f(\kappa)))\,\varphi(x,0)\mathrm{d}x\mathrm{d}t,$$

and, thanks to (11.67), we get

$$\lim_{m\to\infty} T_{20}^{(m)} = \int_{\mathbb{R}_+}\int_{\mathbb{R}}(G_0(u(x,t))v(x,t)+G_1(u(x,t)))\,\varphi_x(x,t)\mathrm{d}x\mathrm{d}t.$$

Turning to the study of $T_{21}^{(m)}$, we get, from the Lipschitz continuity properties of F_{Go} and using $u_{i+\frac{1}{2}}^{(n)}\in \overline{\mathrm{I}}(u_i^{(n)},u_{i+1}^{(n)})$,

$$|W_i^n| \le C_{\max}(C_{14}+C_{15}C_{\mathrm{Lip}})\ \left|u_{i+1}^{(n)}-u_i^{(n)}\right|$$

Thanks to the above inequalities, we obtain that

$$\left|T_{21}^{(m)}\right| \le C_{\max}(C_{14}+C_{15}C_{\mathrm{Lip}})\sum_{n\in\mathbb{N}^\star}\delta t\sum_{i\in\mathbb{Z}}|u_{i-1}^{(n)}-u_i^{(n)}|\left|\varphi((i-\frac{1}{2})h,n\delta t)-\varphi((i+\frac{1}{2})h,n\delta t)\right|$$

which yields, setting $C_{21}=\|\partial_1\varphi\|_{L^\infty(\mathbb{R}\times\mathbb{R}_+)}C_{\max}(C_{14}+C_{15}C_{\mathrm{Lip}})$,

$$\left|T_{21}^{(m)}\right| \le hC_{21}\sum_{n=0}^{\lfloor T/\delta t\rfloor}\delta t\sum_{i\in\mathbb{Z}}\left|u_{i-1}^{(n)}-u_i^{(n)}\right|.$$

This gives, using (11.67) and reintroducing subscripts m, $\left|T_{21}^{(m)}\right| \le h_m(T+\delta t_m)C_{21}W_0$ and finally $\lim_{m\to\infty} T_{21}^{(m)} = 0$, which concludes the proof of the first inequality of (11.11). The second inequality is proven exactly in the same way.

The proof of (11.10) is similar and simpler. For a given $\varphi\in C_c^\infty(\mathbb{R}\times\mathbb{R}_+,\mathbb{R})$ and $m\in\mathbb{N}$, we multiply (11.53) and (11.55) by $h\varphi((i+\frac{1}{2})h_m,n\delta t_m)$, and we sum the result over $i\in\mathbb{Z}$ and $n\in\mathbb{N}$. We then pass to the limit $m\to\infty$, again using the convergence of $(u_{h_m,\delta t_m})_{m\in\mathbb{N}}$ to u in $L^1([-R,R]\times[0,T])$, the convergence of $(v_{h_m,\delta t_m})_{m\in\mathbb{N}}$ for the weak-$\star$ topology of $L^\infty(\mathbb{R}\times\mathbb{R}_+)$ to v, the estimate (11.67), and following the same reasoning as above. We then get that (11.10) is satisfied by (u,v). This concludes the proof that (u,v) is an entropy weak solution of the system (11.9) in the sense of definition 11.2.3.□

11.5 Numerical results

We provide in this section some numerical results, obtained with the generalized Godunov scheme presented in section 11.4. We consider the data of the third analytical example in section 11.2, corresponding to the simulation of a bubble initially present in the domain. The value for C_{17} provided by (11.65) is not sharp enough to be used for practically setting the value of the time step as a function of the space step. Hence, using the analytical values taken by the function v, it was possible to assess a much more

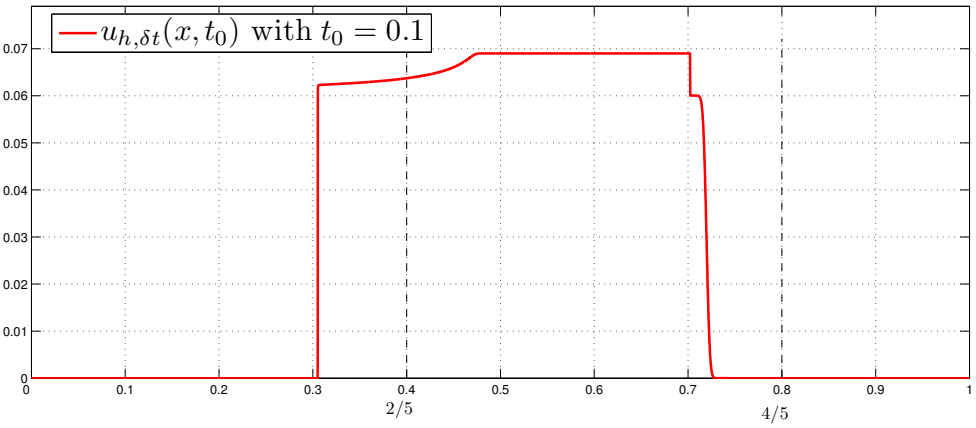


FIG. 11.3: $u_{h,\delta t}(x,t_0)$ with $t_0 = 0.1$

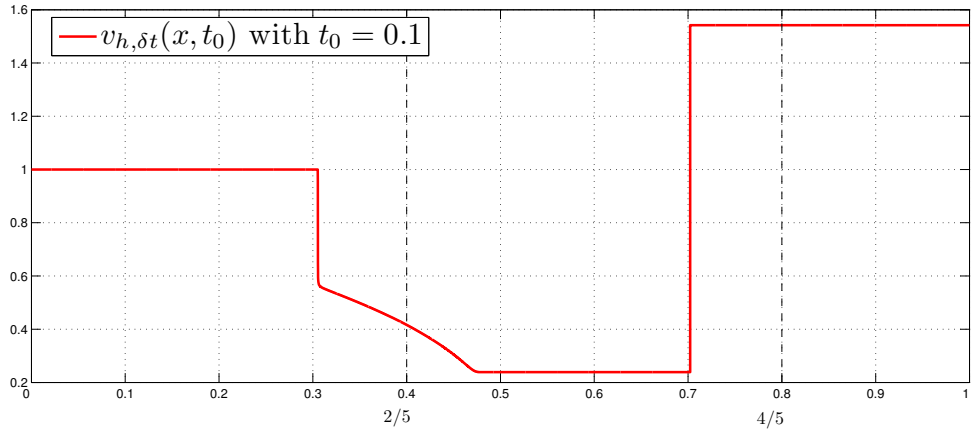


FIG. 11.4: $v_{h,\delta t}(x,t_0)$ with $t_0 = 0.1$

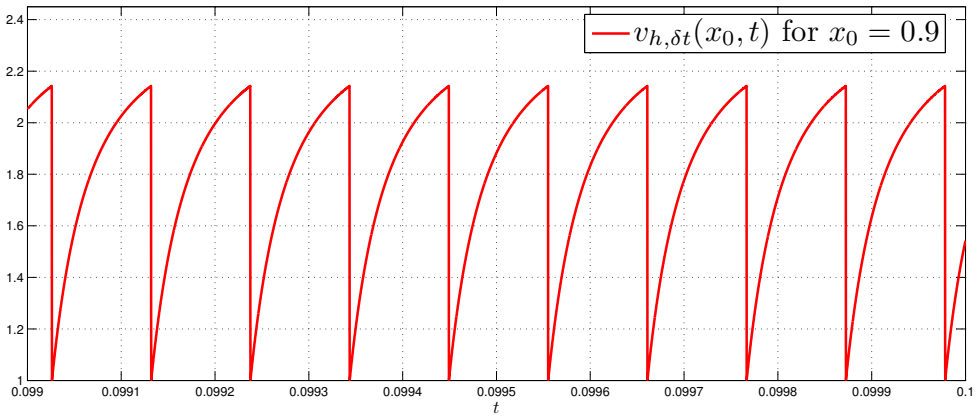


FIG. 11.5: $v_{h,\delta t}(x_0,t)$ for $x_0 = 0.9$

accurate value for $\delta t/h = \overline{C_{17}}$, ensuring the stability of the scheme. We have thus taken $\delta t = 10^{-7}$ for $h = 10^{-4}$. We show on figures 11.3 and 11.4 respectively the functions $u_{h,\delta t}(x,t_0)$ and $v_{h,\delta t}(x,t_0)$ at time $t_0 = 0.1$.

Let us first observe that the numerical solution is very close to the analytical one shown in figures 11.1-11.2. Classically, we note that the contact discontinuity is more subject to numerical diffusion than the shocks. We remark that the approximate solution $v_{h,\delta t}$ shows monotony properties with respect to x, but presents periodic oscillations with respect to t (see figure (11.5)). The period of these oscillations only depends on the space step, and numerically corresponds to the time needed for dissolving the gaseous component into the water phase at the maximum concentration in one control volume, starting with pure water (this time is also that of the apparition of the gaseous phase in this control volume). This is in agreement with the theoretical convergence properties in section 11.4, only based on the proof of the convergence of $v_{h,\delta t}$ to v for the weak-$\star$ convergence of $L^\infty(\mathbb{R}\times\mathbb{R}_+)$. Indeed, it is easy to see that some BV estimate could be considered for v with respect to the space variable, but the lack of term v_t prevents from obtaining a similar estimate with respect to the time variable. However, a time average of $v_{h,\delta t}(x,t)$ can be shown to accurately converge to the analytical value obtained in section 11.2. Let us compute the numerical error with respect to the analytical solution given in section 11.2. Since in this case, we have $u_{h,\delta t}(x,t) = u(x,t)$ for sufficiently large and small values of x, we can compute

$$E(h,\delta t,t) = \int_{\mathbb{R}} |u(x,t) - u_{h,\delta t}(x,t)|\,\mathrm{d}x. \tag{11.82}$$

We then provide $e(h) = E(h,\overline{C_{17}}h,0.1)$ as a function of h in figure 11.6 , showing a numerical convergence order about 0.5.

Appendix : some technical results

Lemme 11.5.1 (Lipschitz continuity of G)
Assuming Hypotheses (11.7), let G be the function defined by (11.25)-(11.26). Then there exists C_{22}, which only depends on C_{Lip}, $C_{\max}$, m_0 and f_m, such that

$$\forall s_1,s_2,s_3,g_1\in\mathbb{R},\begin{cases} |G(g_1,s_1,s_2)-G(g_1,s_1,s_3)| \le C_{22}\left(|g_1|+1\right)|s_3-s_2|\,, \\ |G(g_1,s_1,s_2)-G(g_1,s_3,s_2)| \le C_{22}\left(|g_1|+1\right)|s_3-s_1|\,. \end{cases} \tag{11.83}$$

Proof. Prolonging $\tau_f(s_1,s_2)$ by $\tau_f(s_1,s_2) = -C_{\mathrm{Lip}}$ for $s_1 = s_2$, and using (11.25)-(11.26), we can write

$$\forall s_1,s_2,g_1\in\mathbb{R},\ \ G(g_1,s_1,s_2) = g_1 + H_0(s_1,s_2)g_1 + H_1(s_1,s_2), \tag{11.84}$$

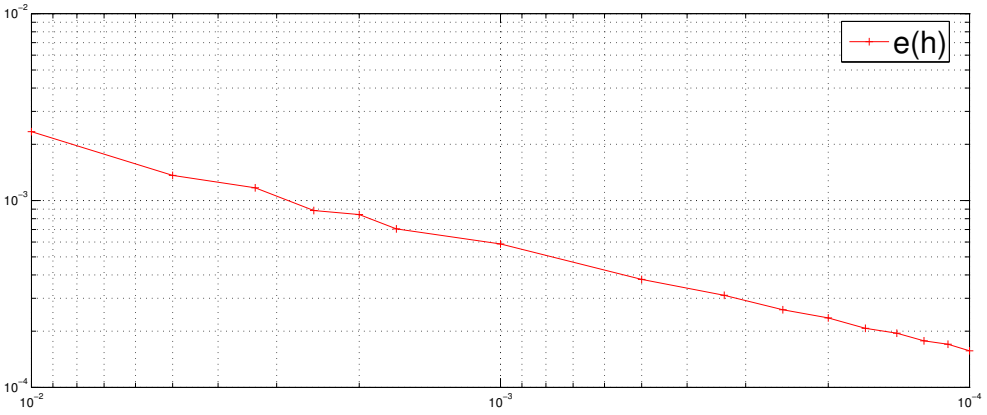


FIG. 11.6: $e(h)$

with

$$\forall (s_1,s_2)\in\mathbb{R}^2,\quad H_0(s_1,s_2)=\frac{(c(s_2)-c(s_1))\tau_f(s_1,s_2)-(a(s_2)-a(s_1))}{a(s_2)-c(s_2)\tau_f(s_1,s_2)},$$
$$H_1(s_1,s_2)=\frac{(d(s_2)-d(s_1))\tau_f(s_1,s_2)-(b(s_2)-b(s_1))}{a(s_2)-c(s_2)\tau_f(s_1,s_2)}.$$

Let us prove that the functions H_0 and H_1 are Lipschitz continuous on $\mathbb{R}^2$, which immediately gives (11.83) from (11.84). Let us first remark that these functions are clearly continuous in all $(s_1,s_2)\in\mathbb{R}^2$, with $s_1\neq s_2$. We now remark that, for all $(s_1,s_2)\in\mathbb{R}^2$, with $s_1\neq s_2$, we have

$$\frac{H_0(s_1,s_2)}{s_2-s_1}=\frac{\tau_c(s_1,s_2)\tau_f(s_1,s_2)-\tau_a(s_1,s_2)}{a(s_2)-c(s_2)\tau_f(s_1,s_2)},$$

which shows that

$$\forall (s_1,s_2)\in\mathbb{R}^2,\ |H_0(s_1,s_2)|\leq|s_2-s_1|\frac{C_{\mathrm{Lip}}^2+C_{\mathrm{Lip}}}{m_0}.$$

The above relation shows that H_0 is continuous on $\mathbb{R}^2$. The same conclusion holds for H_1. It now suffices to show that the partial derivatives of H_0 and H_1 are bounded almost everywhere. Let $(s_1,s_2)\in\mathbb{R}^2$ with $s_1\neq s_2$. We can write $H_0(s_1,s_2)=\widehat{H}_0(s_1,s_2,\tau_f(s_1,s_2))$, with

$$\forall (s_1,s_2)\in\mathbb{R}^2,\ \forall t\in[-C_{\mathrm{Lip}},-f_m],\ \widehat{H}_0(s_1,s_2,t)=\frac{(c(s_2)-c(s_1))t-(a(s_2)-a(s_1))}{a(s_2)-c(s_2)t}.$$

We then easily get that, for a.e. $(s_1,s_2)\in\mathbb{R}^2$ and $t\in[-C_{\mathrm{Lip}},-f_m]$,

$$|\partial_1\widehat{H}_0(s_1,s_2,t)|\leq\frac{C_{\mathrm{Lip}}+C_{\mathrm{Lip}}^2}{m_0},$$

$$|\partial_2\widehat{H}_0(s_1,s_2,t)|\leq\frac{(C_{\mathrm{Lip}}+C_{\mathrm{Lip}}^2)(C_{\max}+C_{\mathrm{Lip}}C_{\max})}{m_0^2},$$

$$|\partial_3\widehat{H}_0(s_1,s_2,t)|\leq|s_2-s_1|\left(\frac{C_{\mathrm{Lip}}+C_{\mathrm{Lip}}^2}{m_0}+\frac{C_{\max}(C_{\mathrm{Lip}}+C_{\mathrm{Lip}}^2)}{m_0^2}\right).$$

Since we get that

$$\text{for a.e. } (s_1,s_2)\in\mathbb{R}^2,\ \forall i=1,2,\ |\partial_i\tau_f(s_1,s_2)|\leq\frac{2C_{\mathrm{Lip}}}{|s_2-s_1|},$$

we conclude that H_0 is Lipschitz continuous on $\mathbb{R}^2$, since we have

$$\begin{array}{l}\text{for a.e. } (s_1,s_2)\in\mathbb{R}^2,\ \forall i=1,2,\\ \partial_i H_0(s_1,s_2)=\partial_i\widehat{H}_0(s_1,s_2,\tau_f(s_1,s_2))+\partial_i\tau_f(s_1,s_2)\partial_3\widehat{H}_0(s_1,s_2,\tau_f(s_1,s_2)).\end{array}$$

We proceed exactly in the same way for H_1, getting the same conclusion. □

Lemme 11.5.2 (Monotony property of V)
Under Hypotheses (11.7), let V and G be the functions defined by (11.25)-(11.26). Then

$$\begin{array}{l}\forall u_1,u_2,u_3,g_1\in\mathbb{R},\ (u_1-u_2)(u_2-u_3)>0\ \Rightarrow\\ \exists\alpha>0,\ (V(g_1,u_1,u_3)-V(g_1,u_1,u_2))=\alpha(V(G(g_1,u_1,u_2),u_2,u_3)-V(g_1,u_1,u_3)).\end{array}\tag{11.85}$$

Proof. Let $u_1,u_2,u_3,g_1\in\mathbb{R}$ be such that $(u_1-u_2)(u_2-u_3)>0$. Let us denote by $g_2=G(g_1,u_1,u_2)$, $v_2=V(g_1,u_1,u_2)$, $g_3=G(g_1,u_1,u_3)$, $v_3=V(g_1,u_1,u_3)$ and $g_3'=G(g_2,u_2,u_3)$, $v_3'=V(g_2,u_2,u_3)$. Using (11.24), we can write

$$\begin{array}{ll}v_2(f(u_1)-f(u_2)) &=-\big(a(u_1)g_1+b(u_1)-a(u_2)g_2-b(u_2)\big)\\ v_2(u_1-u_2) &=-\big(c(u_1)g_1+d(u_1)-c(u_2)g_2-d(u_2)\big),\\ &\\ v_3'(f(u_2)-f(u_3)) &=-\big(a(u_2)g_2+b(u_2)-a(u_3)g_3'-b(u_3)\big)\\ v_3'(u_2-u_3) &=-\big(c(u_2)g_2+d(u_2)-c(u_3)g_3'-d(u_3)\big).\\ &\\ v_3(f(u_1)-f(u_3)) &=-\big(a(u_1)g_1+b(u_1)-a(u_3)g_3-b(u_3)\big)\\ v_3(u_1-u_3) &=-\big(c(u_1)g_1+d(u_1)-c(u_3)g_3-d(u_3)\big),\end{array}$$

This provides, by addition of the first and second system, elimination of g_3', and by elimination of g_3 in the third system,

$$\begin{array}{l}v_3'\big(c(u_3)(f(u_2)-f(u_3))-a(u_3)(u_2-u_3)\big)+v_2\big(c(u_3)(f(u_1)-f(u_2))-a(u_3)(u_1-u_2)\big)\ =\\ v_3\big(c(u_3)(f(u_1)-f(u_3))-a(u_3)(u_1-u_3)\big).\end{array}$$

We thus get

$$\begin{array}{l}(v_3'-v_3)\big(c(u_3)(f(u_2)-f(u_3))-a(u_3)(u_2-u_3)\big)=\\ (v_3-v_2)\big(c(u_3)(f(u_1)-f(u_2))-a(u_3)(u_1-u_2)\big),\end{array}$$

which gives (11.85) thanks to Hypotheses (11.7). □

Lemme 11.5.3 (Discrete Gronwall's lemma)
Let $N\in\mathbb{N}^\star$ be given, and let $(g_k)_{k\in[\![0,N]\!]}$ and $(u_k)_{k\in[\![0,N]\!]}$ be discrete sequences of reals, such that there exists $C_G\geq 0$ with

$$|g_{k+1}-g_k|\leq C_G\left(|g_k|+1\right)|u_{k+1}-u_k|,\forall k\in[\![0,N-1]\!].\tag{11.86}$$

Then the following holds

$$|g_k|\leq(|g_0|+1)\exp\left[C_G\sum_{i=0}^{N-1}|u_{i+1}-u_i|\right]-1,\ \forall k\in[\![0,N]\!].\tag{11.87}$$

Proof. Let $k\in[\![0,N-1]\!]$. From (11.86), we get,

$$|g_{k+1}|-|g_k|\leq C_G\left(|g_k|+1\right)|u_{k+1}-u_k|,\tag{11.88}$$

which gives

$$|g_{k+1}|\leq|g_k|\left(1+C_G|u_{k+1}-u_k|\right)+C_G|u_{k+1}-u_k|,\tag{11.89}$$

and therefore

$$|g_{k+1}|+1\leq\left(|g_k|+1\right)\left(1+C_G|u_{k+1}-u_k|\right).\tag{11.90}$$

Since the above inequality holds for all $k \in [\![0, N-1]\!]$, we get

$$|g_{k+1}| + 1 \leq (|g_0| + 1) \prod_{i=0}^{k} (1 + C_G |u_{i+1} - u_i|) . \tag{11.91}$$

Hence we obtain,

$$\log [|g_{k+1}| + 1] \leq \log (|g_0| + 1) + \sum_{i=0}^{k} \log (1 + C_G |u_{i+1} - u_i|) . \tag{11.92}$$

Since $\forall s \geq 0, \log(1+s) \leq s$, we have

$$\log [|g_{k+1}| + 1] \leq \log (|g_0| + 1) + \sum_{i=0}^{k} C_G |u_{i+1} - u_i| . \tag{11.93}$$

Therefore, using $\sum_{i=0}^{k} |u_{i+1} - u_i| \leq \sum_{i=0}^{N-1} |u_{i+1} - u_i|$, we get

$$|g_{k+1}| + 1 \leq (|g_0| + 1) \exp \left[C_G \sum_{i=0}^{N-1} |u_{i+1} - u_i| \right] . \tag{11.94}$$

The conclusion of the proof follows. □

Lemme 11.5.4 (A Total Variation Diminution property)
Let $(u_i)_{i\in\mathbb{Z}}$, $(u_{i+\frac{1}{2}})_{i\in\mathbb{Z}}$ and $(\widehat{u}_i)_{i\in\mathbb{Z}}$ be sequences of reals such that

1. $\sum_{i\in\mathbb{Z}} |u_{i+1} - u_i| < \infty$,
2. *for all $i \in \mathbb{Z}$, $u_{i+\frac{1}{2}} \in \bar{\mathrm{I}}(u_i, u_{i+1})$,*
3. *for all $i \in \mathbb{Z}$, $\widehat{u}_i \in [\min(u_{i-\frac{1}{2}}, u_i, u_{i+\frac{1}{2}}), \max(u_{i-\frac{1}{2}}, u_i, u_{i+\frac{1}{2}})]$.*

Then

$$\sum_{i\in\mathbb{Z}} |\widehat{u}_{i+1} - \widehat{u}_i| \leq \sum_{i\in\mathbb{Z}} |u_{i+1} - u_i|.$$

Proof. Let us first remark that the property

$$\forall s_1, s_2, s_3, s_4 \in \mathbb{R},\ \ s_1 \in [\min(s_2, s_3, s_4), \max(s_2, s_3, s_4)] \Rightarrow (|s_3 - s_4| + |s_3 - s_2| \geq |s_1 - s_4| + |s_1 - s_2|)$$

can easily be shown by considering the different cases. We thus get that

$$\left|u_i - u_{i-\frac{1}{2}}\right| + \left|u_i - u_{i+\frac{1}{2}}\right| \geq \left|\widehat{u}_i - u_{i-\frac{1}{2}}\right| + \left|\widehat{u}_i - u_{i+\frac{1}{2}}\right|, \ \forall i \in \mathbb{Z}.$$

Using that, for all $i \in \mathbb{Z}$, we have $|u_i - u_{i+1}| = \left|u_i - u_{i+\frac{1}{2}}\right| + \left|u_{i+\frac{1}{2}} - u_{i+1}\right|$, since $u_{i+\frac{1}{2}} \in \bar{\mathrm{I}}(u_i, u_{i+1})$, and $|\widehat{u}_i - \widehat{u}_{i+1}| \leq \left|u_{i+\frac{1}{2}} - \widehat{u}_i\right| + \left|\widehat{u}_{i+1} - u_{i+\frac{1}{2}}\right|$, we conclude the proof of the lemma. □

Théorème 11.5.5 (A variant of Ascoli's theorem)
Let E be a Banach space, let $(u_n)_{n\in\mathbb{N}}$ be a sequence of functions from $\mathbb{R}_+ \to E$ such that, for all $t \in \mathbb{R}_+$, there exists a relatively compact subset $A(t)$ of E with $u_n(t) \in A(t)$ for all $n \in \mathbb{N}$. We assume that there exists a sequence $(\delta t_n)_{n\in\mathbb{N}}$ of non negative reals which converges to 0, and that there exists $C > 0$ with

$$\|u_n(t_2) - u_n(t_1)\|_E \leq C\ (|t_2 - t_1| + \delta t_n),\ \forall n \in \mathbb{N},\ \forall t_1, t_2 \in \mathbb{R}_+ . \tag{11.95}$$

Then there exists $u \in \mathrm{Lip}(\mathbb{R}_+; E)$ and a subsequence of $(u_n, \delta t_n)_{n\in\mathbb{N}}$, again denoted $(u_n, \delta t_n)_{n\in\mathbb{N}}$, such that

$$\forall T \in \mathbb{R}_+,\ \lim_{n\to\infty} \sup_{t\in[0,T]} \|u_n(t) - u(t)\|_E = 0. \tag{11.96}$$

Proof. The proof follows that of Ascoli's theorem. Let $(t_n)_{n\in\mathbb{N}}$ be a dense sequence in $\mathbb{R}_+$. One extracts from the sequence $(u_n, \delta_n)_{n\in\mathbb{N}}$, thanks to the diagonal process, a subsequence such that $(u_n(t_k))_{n\in\mathbb{N}}$ converges for all $k \in \mathbb{N}$.

Then the property (11.95) allows to show that, for all $t \in \mathbb{R}_+$, $(u_n(t))_{n\in\mathbb{N}}$ is a Cauchy sequence. Indeed, for $\varepsilon > 0$, one first chooses $k \in \mathbb{N}$ such that $|t - t_k| \leq \varepsilon$, then $n_0 \in \mathbb{N}$ such that $\delta_n \leq \varepsilon$ for all $n \geq n_0$, and $\|u_n(t_k) - u_p(t_k)\|_E \leq \varepsilon$ for all $n, p \geq n_0$. The inequality $\|u_n(t) - u_p(t)\|_E \leq \|u_n(t) - u_n(t_k)\|_E + \|u_n(t_k) - u_p(t_k)\|_E + \|u_p(t_k) - u_p(t)\|_E \leq (1 + 4C)\varepsilon$ for all $n, p \geq n_0$ follows.

One then defines, for all $t \in \mathbb{R}_+$, $u(t)$ as the limit of $(u_n(t))_{n\in\mathbb{N}}$. Passing to the limit $n \to \infty$ in (11.95) provides

$$\|u(t_2) - u(t_1)\|_E \leq C\ |t_2 - t_1|,\ \forall t_1, t_2 \in \mathbb{R}_+, \tag{11.97}$$

which shows that $u \in \mathrm{Lip}(\mathbb{R}_+; E)$. Then (11.96) is again an easy consequence of (11.95). Indeed, let $T \geq 0$ and $\varepsilon > 0$ be given. Since, for all $k = 0, \ldots, \lfloor T/\varepsilon \rfloor$ (where $\lfloor x \rfloor$ denotes the greater integer lower of equal to x), the sequence $(u_n(k\varepsilon))_{n\in\mathbb{N}}$ converges to $u(k\varepsilon)$, let $n_0 \in \mathbb{N}$ be such that $\|u_n(k\varepsilon) - u(k\varepsilon)\|_E \leq \varepsilon$ for all $k = 0, \ldots, \lfloor T/\varepsilon \rfloor$ and all $n \geq n_0$, and such that $\delta_n \leq \varepsilon$ for all $n \geq n_0$. Then, for all $t \in [0, T]$ and $n \geq n_0$, letting $k = \lfloor t/\varepsilon \rfloor$, we get using (11.97) and (11.95), $\|u(t) - u_n(t)\|_E \leq \|u(t) - u(k\varepsilon)\|_E + \|u(k\varepsilon) - u_n(k\varepsilon)\|_E + \|u_n(k\varepsilon) - u_n(t)\|_E \leq (1 + 3C)\varepsilon$, which concludes the proof of (11.96). □

Chapitre 12

Boundary conditions

In the previous chapter, the boundary conditions are not considered. We propose to give, in this chapter, some indications to treat the boundary conditions, althoug this problem is not currently solved.

We add to the problem (11.6) the initial condition

$$u(x,0) = u_0(x),\ x \in]A,B[\tag{12.1}$$

and the boundary conditions

$$u(A,t) = u_A(t),\ u(B,t) = u_B(t),\ v(x,t) = \overline{v}_A(t),\ t \in \mathbb{R}_+. \tag{12.2}$$

The hypotheses (11.7) are always valid. Let us now define some hypotheses on the initial and boundary conditions

$$\left\{\begin{array}{l} A,B \in \mathbb{R} \text{ with } A < B \\ u_0 \in L^\infty(\mathbb{R}) \cap BV(]A,B[) \text{ and } u_A, u_B \in L^\infty(\mathbb{R}_+) \cap BV(\mathbb{R}_+) \\ W_0 = \|u_0\|_{BV(]A,B[)}, W_A = \|u_A\|_{BV(\mathbb{R}_+)}, W_B = \|u_B\|_{BV(\mathbb{R}_+)}, \\ \text{(we then denote } U_m, U_M \in \mathbb{R} \text{ s.t. } u_0(x), u_A(t), u_B(t) \in [U_m, U_M] \text{ for a.e. } x \in]A,B[\text{ and } t \in \mathbb{R}_+), \\ v_A \in L^\infty(\mathbb{R}_+) \text{ is given, and we denote } \overline{v}_A = \|v_A\|_{L^\infty(\mathbb{R}_+)}. \end{array}\right. \tag{12.3}$$

12.1 The finite volume scheme

The idea of the following scheme is to keep the same scheme as the one presented in section 11.4 for all interfaces except for the first one. For the first interface, denoted by subscript $\frac{1}{2}$, we use the solution of the generalized Riemann problem for the given data (v_A, u_A^n, u_1^n).

Let $\delta t > 0$ be given, which will be called the time step in the following. Let $N \in \mathbb{N}^\star$ be given, and let $h = (B-A)/N$ which will be called the space step in the following. We denote $[\![1,N]\!] = \{n \in \mathbb{N}, 1 \leq n \leq N\}$ and we define $K_i = (ih, (i+1)h)$, for all $i \in [\![1,N]\!]$. We set

$$u_i^{(0)} = \frac{1}{h}\int_{ih}^{(i+1)h} u_0(x)\mathrm{d}x,\ \forall i \in [\![1,N]\!]. \tag{12.4}$$

We then define the finite volume scheme by induction on $n \in \mathbb{N}$. Let $n \in \mathbb{N}$ be given and let us assume that $(u_i^{(n)})_{i\in[\![1,N]\!]}$ is a given family of reals. We define

$$u_A^{(n)} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t} u_A(t)\mathrm{d}t,\ \ u_{N+1}^{(n)} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t} u_B(t)\mathrm{d}t,\ v_A^{(n)} = \frac{1}{\delta t}\int_{n\delta t}^{(n+1)\delta t} v_A(t)\mathrm{d}t,\ \forall n \in \mathbb{N}. \tag{12.5}$$

Let $g_{\frac{1}{2}}^{(n)} : \mathbb{R} \to \mathbb{R}$ be the fonction given by theorem (11.2.6) of Eymard and Tillier (2007) for the values $u_l = u_0^{(n)}$, $u_r = u_1^{(n)}$ and $g_l = v_A^{(n)}$:

$$g_{\frac{1}{2}}^{(n)} = g(u_0^{(n)}, u_1^{(n)}, v_A^{(n)}) \tag{12.6}$$

We then introduce

$$\mu_{\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \to \mathbb{R}, \ u \mapsto -(c(u)g_{\frac{1}{2}}^{(n)}(u) + d(u)), \tag{12.7}$$

and

$$\nu_{\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \to \mathbb{R}, \ w \mapsto -\left(a(f^{(-1)}(w))g_{\frac{1}{2}}^{(n)}(f^{(-1)}(w)) + b(f^{(-1)}(w))\right), \tag{12.8}$$

We can then define $u_{\frac{1}{2}}^{(n)}$ by

$$u_{\frac{1}{2}}^{(n)} \text{ is any element of } \left\{ s \in \overline{\mathrm{I}}(u_0^{(n)}, u_1^{(n)}), \ \mu_{\frac{1}{2}}^{(n)}(s) = F_{\mathrm{Go}}(\mu_{\frac{1}{2}}^{(n)}, u_0^{(n)}, u_1^{(n)}) \right\}. \tag{12.9}$$

This allows to define the value $v_{\frac{1}{2}}^{(n)}$ by :

$$v_{\frac{1}{2}}^{(n)} = g_{\frac{1}{2}}^{(n)}(u_{\frac{1}{2}}^{(n)}). \tag{12.10}$$

We now define the finite volume scheme by induction on $i \in [\![1, N]\!]$ (the following holds including the case $i = N$, which corresponds to the boundary condition in $x = B$), assuming that the values $v_{i-\frac{1}{2}}^{(n)}$ and $u_{i-\frac{1}{2}}^{(n)}$ are known. The principle of the scheme is to give the steps leading to $v_{i+\frac{1}{2}}^{(n)}$ and $u_{i+\frac{1}{2}}^{(n)}$, which permits to give the scheme by induction on $i \in [\![1, N]\!]$. Let us define the function $\Phi_{i+\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\begin{aligned} \Phi_{i+\frac{1}{2}}^{(n)}(u, v) = \quad & f(u_i^{(n)}) + \frac{\delta t}{h}\left(a(u)v + b(u) - a(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - b(u_{i-\frac{1}{2}}^{(n)})\right) \\ & - f\left(u_i^{(n)} + \frac{\delta t}{h}\left(c(u)v + d(u) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right)\right). \end{aligned} \tag{12.11}$$

We remark that this function $\Phi_{i+\frac{1}{2}}^{(n)}$ is locally Lipschitz continuous with respect to its arguments and verifies

$$\partial_2 \Phi_{i+\frac{1}{2}}^{(n)}(u, v) = \frac{\delta t}{h}(a(u) - c(u)f'(\widehat{u}_i^{(n)}(u, v))) \in [\frac{\delta t}{h} m_0, \frac{\delta t}{h} C_{\max}(1 + C_{\mathrm{Lip}})], \quad \text{for a.e. } u, v \in \mathbb{R},$$

with $\widehat{u}_i^{(n)}(u, v) = u_i^{(n)} + \frac{\delta t}{h}\left(c(u)v + d(u) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right)$. Thus, for all $u \in \mathbb{R}$, the function $v \to \Phi_{i+\frac{1}{2}}^{(n)}(u, v)$ is Lipschitz continuous one-to-one from $\mathbb{R}$ to $\mathbb{R}$, and its reciprocal function is Lipschitz continuous from $\mathbb{R}$ to $\mathbb{R}$. Hence we implicitly define the Lipschitz continuous function $g_{i+\frac{1}{2}}^{(n)}$ by

$$g_{i+\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \to \mathbb{R}, \ u \mapsto v \text{ s.t. } \Phi_{i+\frac{1}{2}}^{(n)}(u, v) = \ 0. \tag{12.12}$$

We define $u_{i+\frac{1}{2}}^{(n)} \in \mathbb{R}$ as a value associated with the Godunov scheme for the flux given by the function

$$\mu_{i+\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \to \mathbb{R}, \ u \mapsto -(c(u)g_{i+\frac{1}{2}}^{(n)}(u) + d(u)), \tag{12.13}$$

the left value $u_i^{(n)}$ and the right value $u_{i+1}^{(n)}$, i.e.

$$u_{i+\frac{1}{2}}^{(n)} \text{ is any element of } \left\{ s \in \overline{\mathrm{I}}(u_i^{(n)}, u_{i+1}^{(n)}), \ \mu_{i+\frac{1}{2}}^{(n)}(s) = F_{\mathrm{Go}}(\mu_{i+\frac{1}{2}}^{(n)}, u_i^{(n)}, u_{i+1}^{(n)}) \right\}. \tag{12.14}$$

This allows to define the value $v_{i+\frac{1}{2}}^{(n)}$ by :

$$v_{i+\frac{1}{2}}^{(n)} = g_{i+\frac{1}{2}}^{(n)}(u_{i+\frac{1}{2}}^{(n)}). \tag{12.15}$$

Relations (12.11)-(12.12) and the fact that f is strictly decreasing imply that $f(u_{i+\frac{1}{2}}^{(n)})$ reaches the Godunov scheme for the flux given by the function

$$\nu_{i+\frac{1}{2}}^{(n)} \ : \ \mathbb{R} \to \mathbb{R}, \ w \mapsto -\left(a(f^{(-1)}(w))g_{i+\frac{1}{2}}^{(n)}(f^{(-1)}(w)) + b(f^{(-1)}(w))\right), \tag{12.16}$$

the left value $f(u_i^{(n)})$ and the right value $f(u_{i+1}^{(n)})$, i.e.

$$\nu_{i+\frac{1}{2}}^{(n)}(f(u_{i+\frac{1}{2}}^{(n)})) = F_{\mathrm{Go}}(\nu_{i+\frac{1}{2}}^{(n)}, f(u_i^{(n)}), f(u_{i+1}^{(n)})). \tag{12.17}$$

Hence we have defined values $u_{i+\frac{1}{2}}^{(n)}$ and $v_{i+\frac{1}{2}}^{(n)}$, and the induction on $i \in [\![1, N]\!]$ is now complete. These values are such that

$$\begin{aligned} &f(u_i^{(n)}) + \frac{\delta t}{h}\left(a(u_{i+\frac{1}{2}}^{(n)})v_{i+\frac{1}{2}}^{(n)} + b(u_{i+\frac{1}{2}}^{(n)}) - a(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - b(u_{i-\frac{1}{2}}^{(n)})\right) \\ &= f\left(u_i^{(n)} + \frac{\delta t}{h}\left(c(u_{i+\frac{1}{2}}^{(n)})v_{i+\frac{1}{2}}^{(n)} + d(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right)\right). \end{aligned} \tag{12.18}$$

We then set

$$u_i^{(n+1)} = u_i^{(n)} + \frac{\delta t}{h}\left(c(u_{i+\frac{1}{2}}^{(n)})v_{i+\frac{1}{2}}^{(n)} + d(u_{i+\frac{1}{2}}^{(n)}) - c(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - d(u_{i-\frac{1}{2}}^{(n)})\right), \ \forall i \in [\![1, N]\!]. \tag{12.19}$$

Note that, for all $i \in [\![1, N]\!]$ with $i \leq i_0^{(n)} - 1$, we get $u_i^{(n+1)} = \overline{u}_0$. We also denote by

$$w_i^{(n)} = f(u_i^{(n)}), \ \forall i \in [\![0, N+1]\!], \ \forall n \in \mathbb{N}. \tag{12.20}$$

Thanks to (12.18), we can write

$$w_i^{(n+1)} = w_i^{(n)} + \frac{\delta t}{h}\left(a(u_{i+\frac{1}{2}}^{(n)})v_{i+\frac{1}{2}}^{(n)} + b(u_{i+\frac{1}{2}}^{(n)}) - a(u_{i-\frac{1}{2}}^{(n)})v_{i-\frac{1}{2}}^{(n)} - b(u_{i-\frac{1}{2}}^{(n)})\right), \ \forall i \in [\![1, N]\!]. \tag{12.21}$$

This allows the definition of the scheme by induction on n to hold.

Thanks to the definition of discrete values $u_i^{(n)}$ and $v_{i+\frac{1}{2}}^{(n)}$, for $n \in \mathbb{N}$ and $i \in [\![1, N]\!]$, we can define the approximate functions $u_{h,\delta t} :]A, B[\times\mathbb{R}_+ \to \mathbb{R}$ and $v_{h,\delta t} :]A, B[\times\mathbb{R}_+ \to \mathbb{R}$ of u and v by

$$\begin{aligned} &u_{h,\delta t}(x,t) = u_i^{(n)}, \ \forall x \in (A + (i-1)h, A + ih) \text{ and all } t \in [n\delta t, (n+1)\delta t), \ \forall i \in [\![1, N]\!], \ \forall n \in \mathbb{N}, \\ &v_{h,\delta t}(x,t) = v_{\frac{1}{2}}^{(n)}, \ \forall x \in (A, A + \tfrac{1}{2}\,h) \text{ and all } t \in [n\delta t, (n+1)\delta t), \ \forall n \in \mathbb{N}, \\ &v_{h,\delta t}(x,t) = v_{i+\frac{1}{2}}^{(n)}, \ \forall x \in (A + (i-\tfrac{1}{2})h, A + (i+\tfrac{1}{2})h) \text{ and all } t \in [n\delta t, (n+1)\delta t), \ \forall i \in [\![1, N-1]\!], \ \forall n \in \mathbb{N}, \\ &v_{h,\delta t}(x,t) = v_{N+\frac{1}{2}}^{(n)}, \ \forall x \in (A + (N-\tfrac{1}{2})h, B) \text{ and all } t \in [n\delta t, (n+1)\delta t). \end{aligned} \tag{12.22}$$

12.2 Numerical result

We choose, in this section, to modify the value of $\overline{X}$. We take $\overline{X} = 0.02$ instead of the previous value used in the section 11.5. This enables to reduce the solubility of the co_2 in water to keep more co_2 in gaz phase.

We choose the boundary conditions to close the top of the domain. It corresponds to take

$$\begin{cases} u_A = \xi_g \\ v_A = \rho_g \end{cases}$$

For the boundary condition at the bottom of the domain, we choose

$$u_B = 0$$

This choice corresponds to let pure water get in the system if necessary.

We show on figures 12.1 and 12.2 respectively the functions $u_{h,\delta t}(x, t_0)$ and $v_{h,\delta t}(x, t_0)$ at differents times t_0.

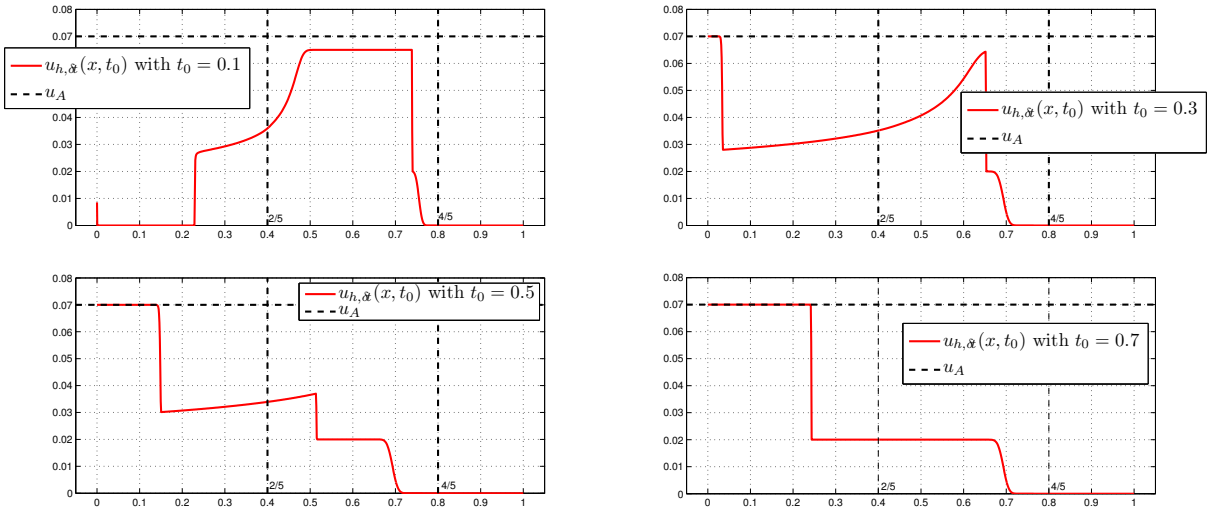


FIG. 12.1: $u_{h,\delta t}(w, t_0)$ with $t_0 = 0.1$, $t_0 = 0.3$, $t_0 = 0.5$ and $t_0 = 0.7$

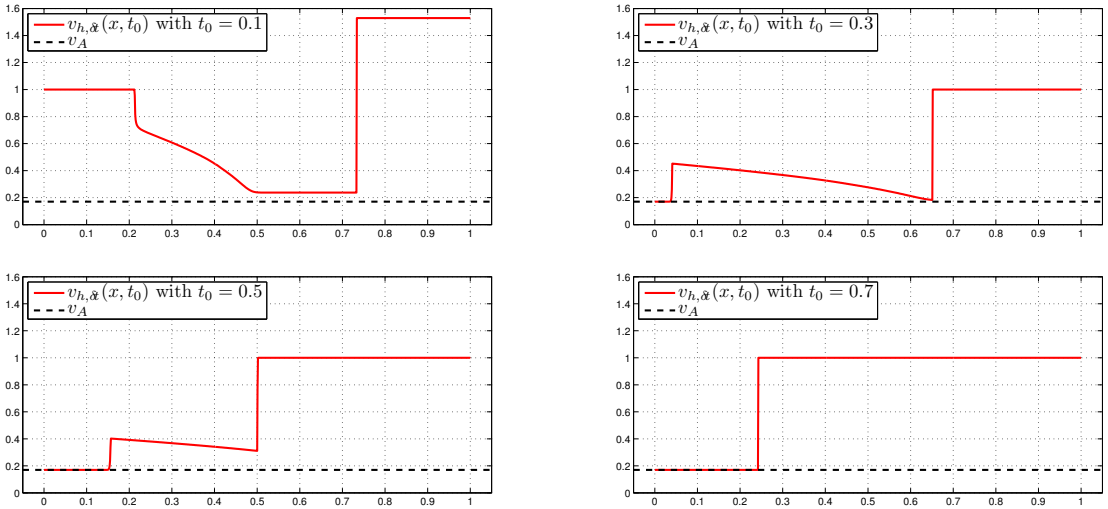


FIG. 12.2: $v_{h,\delta t}(w, t_0)$ with $t_0 = 0.1$, $t_0 = 0.3$, $t_0 = 0.5$ and $t_0 = 0.7$

The figure 12.3 shows the evolution of $u_{god}(t)$ and $v_{god}(t)$ according to the values of u_A, v_A and u_0^n.

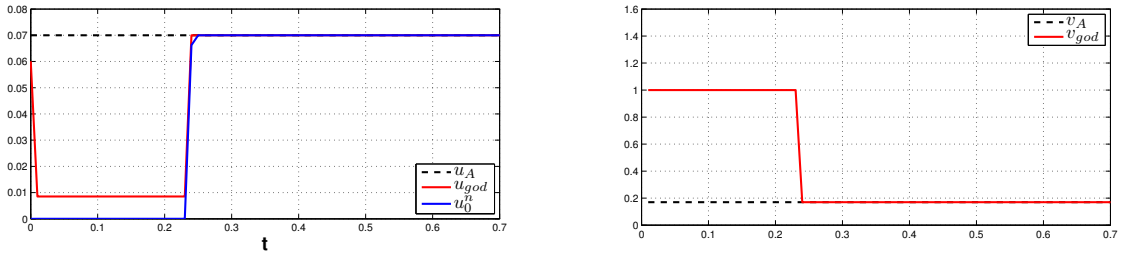


FIG. 12.3: $u_{god}(t)$ and $v_{god}(t)$

Conclusion

Conclusion

Au cours de cette thèse, nous avons étudié une méthode de couplage pour simuler les écoulements multiphasiques réactifs. Pour réaliser cette étude, un prototype a été développé. Ce prototype permet d'utiliser aussi bien une méthode globale que la méthode (*Rs-Tr*), ce qui permet de comparer les deux méthodes et d'étudier la validité de la méthode (*Rs-Tr*).

Après une validation du module géochimie du prototype et des hypothèses de découplage, nous avons effectué une comparaison des deux méthodes. Cette étude montre que les résultats obtenus avec (*Rs-Tr*) sont similaires à ceux obtenus avec la méthode globale. En considérant la solution obtenue avec la méthode globale comme une solution de référence, une étude numérique de l'erreur et de la convergence de la méthode (*Rs-Tr*) a été réalisée. On montre alors que l'erreur est faible et qu'en réduisant le pas de temps, la méthode (*Rs-Tr*) converge vers la méthode globale.

Une étude détaillée de la méthode de couplage a permis de montrer que la correction d'erreur par pénalisation est suffisante pour corriger l'erreur de splitting et qu'elle est moins coûteuse et plus stable qu'une méthode itérative, bien qu'elle engendre localement des oscillations non physiques sur la solution. Une combinaison de la méthode par pénalisation et d'une méthode itérative a été proposée pour corriger ces oscillations, sans ajouter un coût de simulation trop important. De plus, une analyse mathématique de convergence sur un cas simplifié a permis de montrer que le schéma pénalisé est convergent et qu'il converge vers la même solution que le schéma global.

L'un des points importants de la méthode (*Rs-Tr*) est d'offrir la possibilité de résoudre le modèle de transport réactif de façon locale avec des sous-pas de temps locaux. Or, dans le cas où les phénomènes de diffusion et de dispersion sont importants, on ne peut pas utiliser directement une méthode locale. C'est pourquoi nous avons réalisé une étude du poids de la diffusion et de la dispersion, qui a permis de mettre en évidence que la dispersion peut être du même ordre de grandeur que la convection. Une analyse numérique a ensuite permis d'étudier l'effet d'une dispersion importante sur la méthode (*Rs-Tr*). Cette analyse confirme qu'en présence d'une forte dispersion, il est nécessaire de discrétiser le terme de dispersion de façon implicite. Par conséquent, la méthode des caractéristiques utilisée pour la résolution de (*Tr*) ne fonctionne pas. Une solution possible est d'utiliser une méthode de splitting pour découpler la dispersion et les réactions. Une autre solution est d'utiliser des méthodes de décomposition de domaine car la partie du domaine sur laquelle la dispersion est importante est assez réduite.

Finalement, une partie de cette thèse a été consacrée à l'étude mathématique d'un système hyperbolique issu des lois de conservation de l'eau et du co_2 dans une situation simple. Il s'agit de la dissolution dans l'eau et de la migration par gravité d'une bulle de gaz. Ce système d'équation est très original et difficile à traiter en raison de la non linéarité du problème. L'une des idées principales de cette étude est que pour obtenir un schéma convergent, il est nécessaire de respecter chaque loi de conservation de façon simultanée, sans privilégier l'une des équations par rapport à l'autre. Cette étude a également mis en évidence la convergence faible du gradient de pression. Ceci signifie que le gradient de pression oscille au cours du temps avec une fréquence dépendante du maillage.

Perspectives

Pour poursuivre le travail de thèse, il serait intéressant d'étudier différentes méthodes pour traiter le terme de dispersion. Notamment, il faudrait essayer des méthodes de splitting ou des méthodes de décomposition de domaine.

De plus, les méthodes de décomposition de domaine sont également très intéressantes pour la partie réactive. En effet, les différentes simulations et études réalisées ont permis de constater que les fronts de réaction sont présents de façon locale. Pour une grande partie du domaine, les réactions sont quasiment inexistantes. L'utilisation d'une méthode de décomposition de domaine est donc une option intéressante dans l'évolution des méthodes de simulation du transport réactif appliquée au stockage de co_2.

Concernant l'étude mathématique du système hyperbolique, l'unicité et la prise en compte des conditions limites restent un problème ouvert. Nous proposons un schéma pour prendre en compte les conditions limites, mais actuellement la preuve de convergence n'est pas complète, certains points restent à démontrer. Concernant l'unicité de la solution, nous n'avons pas à l'heure actuelle de piste de réflexion.

Annexes

Annexe A

Données du test pour la validation du modèle géochimique

Caractéristiques de l'assemblage minéralogique

	proportion (%)	masse molaire ($g.mol^{-1}$)	volume molaire ($cm^3.mol^{-1}$)	masse initiale (g)	nb moles initial (mol/ kg(H_2O))
quartz	87	60.07	22.93	26.1	2.9
k-feldspath	2	278.28	108.70	0.6	1.44×10^{-2}
albite	1	262.18	100.07	0.3	7.63×10^{-3}
kaolinite	2	258.11	99.27	0.6	1.55×10^{-2}
sidérite	1	115.84	29.25	0.3	1.73×10^{-2}
annite	5	511.82	161.46	1.5	1.95×10^{-2}
dolomite	1	184.37	60.45	0.3	1.08×10^{-2}
calcite	1	100.07	36.93	0.3	2.00×10^{-2}
muscovite	0	398.31	14.23	0.0	0.00×10^{-2}

Système géochimique

Constantes cinétiques des minéraux

	log(K) 105°C	log(k_{neutre}) ($mol.m^{-2}.s^{-1}$)	log(k_{acide}) ($mol.m^{-2}.jour^{-1}$)	log($k_{(co_2)}$) ($mol.m^{-2}.s^{-1}$)
quartz	-3.02	-10.59	-	-
k-feldspath	-1.89	-10.11	-5.61	-
albite	0.15	-9.82	-7.11	-
kaolinite	0.70	-8.46	-5.69	-
sidérite	-1.56	-5.61	-1.85	-3.83
annite	19.17	-9.23	-	-
dolomite	-0.01	-5.61	-1.85	-3.83
calcite	0.72	-5.70	0.02	-1.92
muscovite	3.77	-	-	-

Modèle transport réactif

Espèce	Phase	Condition limite	Condition initiale
h_2o	w	à calculer	à calculer
h^+	w	3.76×10^{-6}	3.76×10^{-6}
co_2	w	0.0103	0.0103
ca^{2+}	w	0.0013	0.0013
hco_3^-	w	1.96×10^{-5}	1.96×10^{-5}
al^{3+}	w	0	0
k^+	w	2×10^{-4}	2×10^{-4}
sio_2	w	0	0
na^+	w	0.0226	0.0226
mg^{2+}	w	4.284×10^{-4}	4.284×10^{-4}
fe^{2+}	w	0	0
co_2^g	g	-	-

La fraction molaire de l'espèce h_2o est initialisée à l'aide des fractions molaires des autres espèces selon la loi de fermeture (1.6).

Réactions

équilibres homogènes	log K
$h_2o + co_2^w \longleftrightarrow h^+ + hco_3^-$	-6.4
équilibres hétérogènes	H
$co_2^w \longleftrightarrow co_2^g$	97.38

Annexe B

Données du test pour l'étude de l'hypothèse de découplage

Géométrie

Dimension ($L \times P \times H$)		$1000 \times 1 \times 1$
Maillage		$200 \times 1 \times 1$
hauteur max		$z = -999$ m
Condition limites	bord supérieur	flux nul
	bord inférieur	flux nul
	bord gauche	flux nul
	bord droit	$P_w = 100$ bars imposé
	injection	Q=0.01 $m^3.j^{-1}$ entre t = 2 ans et t = 22 ans

Propriétés pétrophysiques

Perméabilité	constante et uniforme	$K = 100$ mD
Perméabilité relative	en croix	$S_w^i = 0.01$
Pression capillaire	pas de pression capillaire	
Densité de l'eau	constante	$\xi_w = \frac{10^3}{18.015}$ mol.L^{-1}
Viscosité de l'eau	corrélation empirique	
Densité du gaz	gaz parfait	
Viscosité du gaz	Lohrenz-Bray-Clarck	

Les perméabilités relatives en croix sont calculés par $k_{rw}(S_w) = S_w^\star$ et $k_{rg}(S_w) = (1 - S_w^\star)$ avec $S_w^\star = \frac{S_w - S_w^i}{1 - S_w^i}$.

Système géochimique

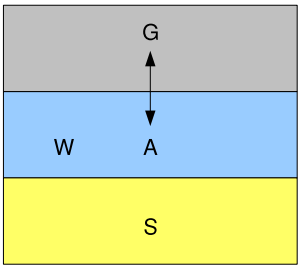


FIG. B.1: Représentation de la géocichimie

Modèle transport réactif

Espèce	Phase	Condition limite	Condition initiale
h_2o	w	à calculer	à calculer
co_2	w	10^{-12}	10^{-12}
co_2^g	g	-	-
$quartz$	m_2	-	0.8

La fraction molaire de l'espèce h_2o est initialisée à l'aide des fractions molaires des autres espèces selon la loi de fermeture (1.6).

Réactions

équilibres hétérogènes	H
$co_2^a \longleftrightarrow co_2^g$	*variable*

Annexe C

Données du test pour comparaison des schémas

Géométrie

Dimension ($L \times P \times H$)		$1000 \times 1 \times 1$
Maillage		$1000 \times 1 \times 1$
hauteur max		$z = -999$ m
Condition limites	bord supérieur	flux nul
	bord inférieur	flux nul
	bord gauche	flux nul
	bord droit	$P_w = 100$ bars imposé
	injection	Q=0.01 $m^3.j^{-1}$ entre t = 20 ans et t = 25 ans

Propriétés pétrophysiques

Perméabilité	constante et uniforme	$K = 100$ mD
Perméabilité relative	en croix	$S_w^i = 0.01$
Pression capillaire	pas de pression capillaire	
Densité de l'eau	constante	$\xi_w = \frac{10^3}{18.015}$ mol.L^{-1}
Viscosité de l'eau	corrélation empirique	
Densité du gaz	gaz parfait	
Viscosité du gaz	Lohrenz-Bray-Clarck	

Les perméabilités relatives en croix sont calculées par $k_{rw}(S_w) = S_w^\star$ et $k_{rg}(S_w) = (1 - S_w^\star)$ avec $S_w^\star = \frac{S_w - S_w^i}{1 - S_w^i}$.

Système géochimique

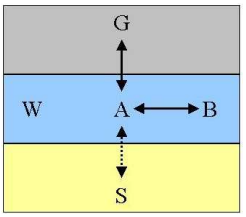


FIG. C.1: Représentation de la géochimie

Modèle transport réactif

Espèce	Phase	Condition limite	Condition initiale
h_2o	w	à calculer	à calculer
co_2^a	w	1.80×10^{-3}	1.80×10^{-3}
co_2^b	w	4.53×10^{-4}	4.53×10^{-4}
co_2^g	g	-	-
co_2^s	m_1	-	0.1
$quartz$	m_2	-	0.7

La fraction molaire de l'espèce h_2o est initialisée à l'aide des fractions molaires des autres espèces selon la loi de fermeture (1.6).

Modèle réservoir

(h_2o)	w
(co_2^w)	w
(co_2^g)	g

Réactions

cinétiques	log K	k_p	k_d
$co_2^a \Longleftrightarrow co_2^s$	-1	0.005	0.005
équilibres homogènes	log K		
$co_2^a \longleftrightarrow co_2^b$	-0.6		
équilibres hétérogènes	H		
$co_2^a \longleftrightarrow co_2^g$	50		

Annexe D

Données du test pour l'étude de la diffusion et de la dispersion

Géométrie

Dimension ($L \times P \times H$)		$1000 \times 1 \times 200$
Maillage		$83 \times 1 \times 24$
Hauteur max		$z = -800$ m
Conditions limites	bord supérieur	flux nul
	bord inférieur	flux nul
	bord gauche	pression hydrostatique + $\delta P = 0.02$ bars
	bord droit	pression hydrostatique
	injection	Q=0.02 $m^3.j^{-1}$ entre t = 1 an et t = 11 ans

Propriétés pétrophysiques

Perméabilité	constante mais non uniforme	$K = \begin{cases} 10 \text{ mD} & \text{si } -880 \leq z \leq -870 \\ 3000 \text{ mD} & \text{sinon} \end{cases}$
Perméabilité relative	cf. tableau D.1	$S_w^i = 0.01$
Pression capillaire	cf. tableau D.2	
Densité de l'eau	constante	$\xi_w = \frac{10^3}{18.015}$ mol.L^{-1}
Viscosité de l'eau	corrélation empirique	
Densité du gaz	gaz parfait	
Viscosité du gaz	Lohrenz-Bray-Clarck	

Les perméabilités relatives sont déterminées par les données du tableau D.1 et la pression capillaire par les données du tableau D.2.

S_w	k_{rw}	S_g	k_{rg}
0.20	0	0.05	0
0.28	5.07×10^{-7}	0.145	5.07×10^{-7}
0.36	2.31×10^{-5}	0.24	2.31×10^{-5}
0.44	0.00022	0.335	0.00022
0.52	0.00110	0.43	0.00110
0.60	0.00396	0.525	0.00396
0.68	0.01164	0.62	0.01169
0.76	0.03028	0.715	0.03028
0.84	0.07425	0.81	0.07425
0.92	0.18616	0.905	0.18616
1.	1.	1.	1.

TAB. D.1: Données perméabilités relatives

pour $K = 10$ mD		pour $K = 3000$ mD	
S_g	p_c	S_g	p_c
0	0	0	0
0.08	0.0149	0.08	0.258
0.16	0.0215	0.16	0.3724
0.24	0.0261	0.24	0.4518
0.32	0.0294	0.32	0.5096
0.4	0.0318	0.4	0.5517
0.48	0.0336	0.48	0.5816
0.56	0.0347	0.56	0.6015
0.64	0.0354	0.64	0.6133
0.72	0.0357	0.72	0.6188
0.8	0.0358	0.8	0.62

TAB. D.2: Données pressions capillaire

Système géochimique

Modèle transport réactif

Espèce	Phase	Condition limite	Condition initiale
h_2o	w	à calculer	à calculer
co_2^w	w	1.41×10^{-3}	1.41×10^{-3}
hco_3^-	w	7.07×10^{-7}	7.07×10^{-7}
mg^{2+}	w	10^{-9}	10^{-9}
$mgco_3^-$	w	10^{-9}	10^{-9}
h^+	w	1.8×10^{-12}	1.8×10^{-12}
ca^{2+}	w	1.6×10^{-6}	1.6×10^{-6}
co_2^g	g	-	-
calcite	m_1	-	0.45 si $K = 10$ mD 0.35 sinon
dolomite	m_2	-	0.45 si $K = 10$ mD 0.35 sinon

La fraction molaire de l'espèce h_2o est initialisée à l'aide des fractions molaires des autres espèces selon la loi de fermeture (1.6).

Modèle réservoir

Espèces	**Phase**
(h_2o)	w
(co_2^w)	w
(co_2^g)	g

Réactions

cinétiques	log K	$k_p = k_d$
$calcite + h^+ \Longleftrightarrow hco_3^- + ca^{2+}$	0.72	0.172
$dolomite + 2h^+ \Longleftrightarrow \begin{matrix} mg^{2+} + ca^{2+} \\ +2hco_3^- \end{matrix}$	-0.01	0.0212
équilibres homogènes	log K	
$h_2o + co_2^w \longleftrightarrow h^+ + hco_3^-$	-6.3011	
$mgco_3^- + h^+ \longleftrightarrow mg^{2+} + hco_3^-$	-7.1095	
équilibres hétérogènes	H	
$co_2^w \longleftrightarrow co_2^g$	50	

Bibliographie

Aagaard, P. and Helgeson, H. Thermodynamic and kinetic constraints on reaction rates among minerals and aqueous solutions ; i, theoretical considerations. *American Journal of Science*, 282 :237–285, March 1982.

Amir, L. and Kern, M. Newton-krylov methods for coupling transport with chemistry in porous media. In Binning, P., Engesgaard, P., Dahle, H., Pinder, G., and Gray, W. G., editors, *XVI International Conference on Computational Methods in Water ressources (CMWR XVI)*, Copenhagen, Denmark, 2006.

Azaroual, M., Baranger, P., Wustman, P., and Kervevan, C. Stockage de gaz acides dans les aquifères et les réservoirs pétroliers : les effets à long terme, volet 3 : modélisation des expériences de laboratoire. Technical report, BRGM, 2003.

Aziz, K. and Settari, A. Petroleum reservoir simulation. *Applied Science Publishers*, 1979.

Bachu, S., Celia, M., and Nordbotten, J. Injection and storage of co_2 in deep saline aquifers : Analytical solution $forco_2$ plume evolution during injection. *Transport in Porous Media*, 58(3) :339 – 360, March 2005.

Bear, J. *Dynamics of Fluids in Porous Materials.* American Elsevier, 1972.

Bielinski, A., Ennis-King, J., Fabriol, R., Le Gallo, Y., García, J., Jessen, K., Kovscek, T., Law, D., Lichtner, P., Oldenburg, C., Pawar, R., Rutqvist, J., Steefel, C., Travis, B., Tsang, C., White, S., and Xu, T. Code intercomparison builds confidence in numerical models for geologic disposal of co2 karsten pruess. Kyoto, October 2002. 6th International Conference on Greenhouse Gas Control Technologies.

Bildstein, O. *Modélisation géochimique des ineractions eau-gaz-roche, application à la diagenèse minérale dans les réservoirs géologiques.* PhD thesis, Thèse IFP - Université Louis Pasteur Strasbourg, 1998.

Bossie-Codreanu, D. and Le Gallo, Y. A simulation method for the rapid screening of potential depleted oil reservoirs for co2 sequestration. *Energy*, 29 :1237–1657, 6th International Conference on Greenhouse Gas Control Technologies,Kyoto, Japan, 1 - 4 October 2002.

Bouillard, N. *Développement de méthodes numériques pour le transport réactif.* PhD thesis, Université de Provence, 2006.

Brenier, Y. and Jaffré, J. Upstream differencing for multiphase flow in reservoir simulation. *SIAM J. Numer. Anal.*, 28(3) :685–696, 1991.

Brosse, E., Le Gallo, Y., and C.C., M. Long-term mineral trapping of co2 in aquifers and reservoirs : Integration of thermodynamics and kinetics in reservoir engineering and geological simulations. Houston, March 2002. proceedings of AAPG Conference.

Brosse, E., Potdevin, J., Bazin, B., and Le Gallo, Y. Simulation de la diagenese minérale, modélisation numérique couplée "réaction transport" dans les réservoirs gréseux : premiers choix pour la modélisation. *Rapport IFP*, 40804, 1993.

Bruining, J., Marchesin, D., and Van Duijn, C. J. Steam injection into water-saturated porous rock. *Comput. Appl. Math.*, 22(3) :359–395, 2003. ISSN 0101-8205.

Cassou, C. *Modélisation numérique des intéractions eau-roche.* PhD thesis, 2000.

Cossé, R. *Le gisement.* Institut Français du Pétrole, 1988.

Crouzeix, M. and Mignot, A. *Analyse numérique des équations différentielles. 2e éd. révisée et augmentée. (Numerical analysis of differential equations). 2e éd. révisée at augmentée.* Collection Mathématiques Appliquées pour la Maîtrise. Paris etc. : Masson. viii, 183 p. FF 121.00, 1989.

Duan, Z., Moller, N., and Weare, J. An equation of state for ch4,co2 and h2o i, pure systems from 0 to 8000c from 0 to 8000 bar. *Geochim. Cosmochim. Acta*, 56 :2605–2617, 1992.

Enchéry, G. *Modèles et schémas numériques pour la simulation de genèse de bassins sédimentaires.* PhD thesis, Institut Français du Pétrole et Université de Marne-La-Vallée, 2004.

Engesgaard, P. and Kipp, K. A geochemical transport model for redox-controlled movement of mineral fronts in groundwater flow systems : A case of nitrate removal by oxidation of pyrite. *Water Resources Research*, 28(10) :2829–2843, 1992.

Evans, L. *Partial differential equations.* Graduate Studies in Mathematics. 19. Providence, RI : American Mathematical Society (AMS). xvii, 662 p. $ 75.00 , 1998.

Eymard, R., Gallouët, T., and Herbin, R. Finite volume methods. In *Ciarlet, P. G. (ed.) et al., Handbook of numerical analysis. Vol. 7 : Solution of equations in* $\mathbb{R}^{\times}$ *(Part 3). Techniques of scientific computing (Part 3). Amsterdam : North-Holland/ Elsevier. 713-1020* . 2000.

Eymard, R., Gallouët, T., and Herbin, R. A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. *IMA J. Numer. Anal.*, 26(2) : 326–353, 2006.

Eymard, R., Gallouët, T., Herbin, R., Gutnic, M., and Hilhorst, D. Approximation by the finite volume method of an elliptic-parabolic equation arising in environmental studies. *Math. Models Methods Appl. Sci.*, 11(9) :1505–1528, 2001.

Eymard, R., Gallouët, T., Herbin, R., Hilhorst, D., and Mainguy, M. Instantaneous and noninstantaneous dissolution : Approximation by the finite volume method. *ESAIM, Proc.*, 6 :41–55, 1999.

Eymard, R., Gallouët, T., and Vovelle, J. Limit boundary conditions for finite volume approximations of some physical problems. *J. Comput. Appl. Math.*, 2003.

Eymard, R. and Tillier, E. Mathematical and numerical study of a system of conservation laws. *Journal of Evolution Equation*, 7(2), 2007 2007.

Ferrando, N., Lugo, R., and Mougin, P. Coupling activity coefficient models, henry constant equations, and equations of state to calculate vapor-liquid and solid-liquid equilibrium data. *Chemical Engineering and Processing*, 45 :773–782, september 2006.

Frolkovic, P. and Geiser, J. Discretization methods with discrete minimum and maximum property for convection dominated transport in porous media. Dimov, Ivan (ed.) et al., Numerical methods and applications. 5th international conference, NMA 2002, Borovets, Bulgaria, August 20–24, 2002. Revised papers. Berlin : Springer. Lect. Notes Comput. Sci. 2542, 445-453 (2003)., 2003.

Frolkovivic, P. Flux-based method of characteristics for contaminant transport in flowing groundwater. *Comput. Vis. Sci.*, 5(2) :73–83, 2002.

Godlewski, E. and Raviart, P.-A. *Hyperbolic systems of conservation laws.* New York, 1996.

Godunov, S. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb. (N.S.)*, 47 :271–306, 1959.

Gunter, W., Wiwehar, B., and Perkins, E. Aquifer disposal of co_2-rich greenhouse gases : Extension of the time scale of experiment for co_2-sequestering reactions by geochemical modelling. *Mineralogy and Petrology*, 59 :121 – 140, Mar 1997.

Hammond, G. *Innovative Methods for Solving Multicomponent Biogeochemical Groundwater Transport on Supercomputers.* PhD thesis, Univ. Illinois, 2003.

Holstad, A. A mathematical and numerical model for reactive fluid flow systems. *Computational Geosciences*, 4(2) :103–139, Jun 2000.

ATHOS. *reference manual.* IFP, March 1996.

Krushkov, S. First order quasilinear equations in several independent variables. *Math. USSR, Sb.*, 10 : 217–243, 1970.

Lacomme, P., Prins, C., and Sevaux, M. *Algorithmes de graphes.* Eyrolles, 2003.

Lagneau, V., Pipart, A., and Catalette, H. Reactive transport and long term behaviour of co_2 sequestration in saline aquifers. *Oil and Gas Science and Technology*, 60(2) :231–247, 2005.

Lasaga, A. Transition state theory. *Reviews in Mineralogy and Geochemistry*, pages 135–168, 1981.

Le Gallo, Y., Couillens, P., and Manaï, T. Co2 sequestration in depleted oil or gas reservoirs. Kuala Lumpur, Malaysia, March 2002. SPE International Conference on Health, Safety and Environment in Oil and Gas Exploration and Production.

Le Gallo, Y., Trenty, L., Michel, A., Vidal-Gilbert, S., Parra, T., and Jeannin, L. Ghgt-8, 8th international conference on greenhouse gas control technologies ,19 - 22 june 2006, trondheim, norway. Trondheim, Norway, June 2006. GHGT-8, 8th International Conference on Greenhouse Gas Control Technologies.

Lichtner, P. *Continuum formulation of multicomponent-multiphase reactive transport*, volume 34, chapter 1. Reviews in Mineralogy, 1996.

Liu, T.-P. The Riemann problem for general 2×2 conservation laws. *Trans. Am. Math. Soc.*, 199 :89–112, 1974.

Liu, T.-P. The Riemann problem for general systems of conservation laws. *J. Differ. Equations*, 18 : 218–234, 1975.

Liu, T.-P. The entropy condition and the admissibility of shocks. *Arch. Rat. Mech. Anal.*, pages 78–88, 1976.

Lohrenz, J., Bray, B., and Clark, C. Calculating viscosity of reservoir fluids from their composition. *JPT*, pages 1171–1176, October 1964.

Michel, A., Tillier, E., and Trenty, L. A finite volume scheme for the modeling of co_2 storage. In Benkhaldoun, F., Ouazar, D., and Raghay, S., editors, *Finite Volume for Complex Applications IV*, pages 701–710, 2005.

Michel, A. and Trenty, L. Brevet méthodologie de couplage. Technical report, IFP, n°0000006, 2005.

Missen, R. and Smith, W. *Chemical Reaction Stoichiometry (CRS), Tutorial.* University of Toronto, University of Guelph, Canada, 1998.

Mousseau, V., Knoll, D., and Rider, W. Physics-based preconditioning and the Newton-Krylov method for non-equilibrium radiation diffusion. *J. Comput. Phys.*, 160(2) :743–765, 2000.

Neumann, S. Universal scaling of hydraulic conductivities and dispersivities in geological media. *Water Resources Res.*, 26 :1749–1758, 1990.

Nghiem, L., Sammon, P., and Grabenstetter, J. Modeling co2 storage in aquifers with a fully-coupled geochemical eos compositional simulator. SPE/DOE Symposium on Improved Oil Recovery, 17-21 April, Tulsa, Oklahoma, 2004a.

Nghiem, L., Sammon, P., Grabenstetter, J., and Ohkuma, H. Modeling CO2 storage in aquifers with a fully-coupled geochemical EOS compositional simulator. In *SPE-DOE Symposium on Improved Oil Recovery*, volume 89474, 2004b.

Nourtier, E. *Modélisation géochimique et numérique des intéractions entre des solutions solides et une solution aqueuse.* PhD thesis, Ecole nationale supérieure des mines de Saint-Etienne et Université Jean Monnet, 2003.

Oelkers, E. *Physical and chemical properties of rocks and fluids for chemical mass transport calculations*, volume 34, chapter 3. Reviews in Mineralogy, 1996.

Olejnik, O. Discontinuous solutions of non-linear differential equations. Translated by George Biriuk. *Am. Math. Soc., Transl., II. Ser.*, 26 :95–172, 1957.

Otto, F. Initial-boundary value problem for a scalar conservation law. *C. R. Acad. Sci., Paris, Sér.*, I 322(8) :729–734, 1996.

Pruess, K. A general-purpose numerical simulator for multiphase fluid and heat flow. Technical report, LBNL-29400, UC-251, Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA, 1991.

Serre, D. *Systèmes de lois de conservation. I : Hyperbolicité, entropies, ondes de choc. (Systems of conservation laws. I : Hyperbolicity, entropy, shock waves).* Fondations. Paris : Diderot Editeur. xii, 298 p. FF 180.00 , 1996a.

Serre, D. *Systèmes de lois de conservation. I : Hyperbolicité, entropies, ondes de choc. (Systems of conservation laws. I : Hyperbolicity, entropy, shock waves).* Fondations. Paris : Diderot Editeur. xii, 298 p. FF 180.00 , 1996b.

Serre, D. *Systèmes de lois de conservation. II : Structures géométriques, oscillations et problèmes mixtes. (Systems of conservation laws. II : Geometric structures, oscillations and mixed problems).* Fondations. Paris : Diderot Editeur. iii, 306 p. FF 195.00 , 1996c.

Steefel, C. and MacQuarrie, K. *Approaches to modeling of reactive transport in porous media*, volume 34, chapter 2. Reviews in Mineralogy, 1996.

Strang, G. On the construstion and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5 :506–517, 1968.

Trenty, L., Michel, A., Le Gallo, Y., and Tillier, E. A sequential splitting strategy for co2 storage modeling. In *10th European Conference on the Mathematics of Oil Recovery*, Amsterdam, 2006.

Xu, T. and Pruess, K. Modeling multiphase nonisothermal fluid flow and reactive geochemical transport in variably saturated fractured rocks : 1. methodology. *American Journal of Science*, 301 :16–33, 2001.

Xu, T., Sonnenthal, E., Spycher, N., and Pruess, K. TOUGHREACT : A simulation program for non-isothermal multiphase reactive geochemical transport in variably saturated geologic media, computer and geoscience. *DOI information :10.1016/j.cageo.2005.06.014*, 32 :145–165, 2006.

Yeh, G. and Tripathi, V. A critical evaluation of recent developments in hydrogeochemical transport models of reactive multichemical components. *Water Resources Res. ?*, 25 :93–108, 1989.

Résumé

Cette thèse porte sur le couplage chimie-transport pour la modélisation et la simulation du stockage géologique de co2. Nous présentons un modèle d'écoulement multiphasique et un modèle géochimique permettant de décrire un modèle couplé d'écoulement multiphasique réactif. Nous proposons ensuite deux méthodes de résolution, l'une est une méthode globale, l'autre est une méthode de splitting utilisée à l'IFP dans le logiciel COORES. Le splitting effectué pour cette méthode repose sur des hypothèses physiques. La méthode de couplage utilisée est une méthode de couplage non itérative dans laquelle l'erreur de splitting est corrigée à l'aide d'un terme de pénalisation. Une étude de convergence sur un cas simplifié permet de montrer que le schéma pénalisé est convergent vers la même solution que le schéma global. Une partie de cette thèse est consacrée à l'étude des phénomènes de diffusion-dispersion. On s'intéresse particulièrement à ce terme car il ne peut être intégré facilement dans un schéma de splitting si l'on souhaite résoudre le modèle de transport réactif de façon locale (nécessaire pour l'utilisation de sous-pas de temps locaux). Après avoir mis en évidence l'importance de ce terme sur un cas test représentatif, nous montrons la difficulté de l'intégrer dans le schéma de splitting. Finalement, on étudie un problème d'écoulement miscible en 1D d'un point de vue mathématique. Les difficultés proviennent de la non linéarité due à la solubilité non nulle du gaz dans l'eau. Nous proposons une définition d'une solution faible pour ce problème dont l'existence est montrée à l'aide de la convergence d'un schéma volumes finis de type Godunov.

Mots-clés : méthode de couplage, écoulement multiphasique réactif, problème hyperbolique, schéma de volumes finis

---

Abstract

In this work, we present some results about the coupling between transport and geochemistry for the modelling and the simulation of co2 geological storage. We present a multiphase flow model and a geochemical model which enables to describe a coupled reactive multiphase flow problem. We then propose two methods of resolution, the first one is a global method, the other one is a splitting method which is used at the IFP in the software COORES. The splitting is based on physical assumptions. The coupling method used is a non iterative method, in which the splitting error is corrected by adding a penalisation term. A convergence study shows that this scheme converges to the same solution as the global scheme. A part of this PhD is dedicated to diffusion and dispersion phenomena. We are interested in this term because it can't be integrated easily in a splitting scheme, if the reactive transport is solved locally (which is necessary to use local time-step). After having highlighted the importance of this term on a representative test case, we show some difficulties encountered to integrate it in a splitting scheme. Finally, we study a miscible multiphase flow problem in 1D from a mathematical point of view. The difficulties arise with the non linearity due to the non zero gas solubility in water. We propose a definition for the weak solution of this problem and its existence is shown thanks to the convergence of a finite volume scheme.

Keywords : splitting method, multiphase reactive flow, hyperbolic problem, finite volume scheme