Méthodes numériques pour les écoulements en milieu poreux : estimations a posteriori et stratégie d'adaptation

Vincent Baron

Université de Nantes
Faculté des sciences et des techniques

École doctorale :
Sciences et technologies de l'information et mathématiques (STIM)

Année 2015

Méthodes numériques pour les écoulements en milieu poreux : estimations *a posteriori* et stratégie d'adaptation

---

THÈSE DE DOCTORAT

Discipline : Mathématiques et leurs interactions
Spécialité : Analyse numérique

Présentée et soutenue publiquement par

Vincent BARON

Le 20 mai 2015, devant le jury ci-dessous

Rapporteur	Daniele DI PIETRO, Professeur, Université de Montpellier
Examinateurs	Florence HUBERT, Maître de conférence, Aix-Marseille Université
	Pascal OMNÈS, Professeur Associé, Université Paris XIII
	Mazen SAAD, Professeur, École centrale Nantes
Directeur de thèse	Yves COUDIÈRE, Professeur, Université Bordeaux I
Encadrant	Pierre SOCHALA, Ingénieur de recherche, BRGM

Rapporteurs du manuscrit :

Daniele DI PIETRO, Professeur, Université de Montpellier
Mario OHLBERGER, Professeur, Université de Münster

Remerciements

Mes premières pensées vont vers ma famille, leur soutien inconditionnel tout au long de ce marathon, leurs relectures attentives (merci Éric!) et leur amour. Je salue l'effort de la délégation des Hautes-Alpes dont la présence à ma soutenance m'a beaucoup touché et a contribué à rendre ce moment spécial.

Je remercie Yves Coudière et Pierre Sochala de m'avoir accordé leur confiance pour cette thèse. Ils ont su se montrer patients, disponibles et accordés dans leur guidage. Les connaissances théoriques et l'expérience d'Yves ont été indispensables; de son côté, Pierre m'a aidé à comprendre le volet physique et m'a permis de partager son bureau ainsi que son tableau pendant un an et demi. Les occasions de se creuser la tête ensemble furent nombreuses, stimulantes et agréables.

Je suis reconnaissant à Daniele Di Pietro et Mario Ohlberger d'avoir rapporté consciencieusement ma thèse, ainsi qu'à chaque membre du jury pour leur présence, questions et remarques. Merci en particulier à Pascal Omnès que j'ai croisé avec grand plaisir au gré des conférences. Toute mon affection va à Florence Hubert pour tellement de choses depuis la préparation à l'agrégation, et sans laquelle je n'aurais d'ailleurs pas fait cette thèse.

Merci également à Martin Vohralík d'avoir patiemment répondu à mes questions les plus naïves sur l'estimation par flux équilibrés, et à Alexandre Ern d'avoir pris deux heures de son temps lors d'un CEMRACS pour comprendre mon problème et suggérer des pistes de travail pertinentes.

Le laboratoire Jean Leray est un rare microcosme de convivialité mathématique qu'on a peine à quitter. Merci à Annick, Brigitte et Stéphanie pour leur disponibilité malgré la charge de travail. Merci à l'ensemble des chercheurs, et en particulier à l'équipe d'analyse numérique, accueillante, dynamique et motivante. Au rez-de-chaussée, on trouve de nombreux spécimens de thésards. Une espèce farfelue mais très sociable, qui aime manger en réfléchissant à des problèmes de prisonniers, des questions pointues de grammaire ou de théorie du jeu de taroinche. Merci aux anciens, Alexandre, Céline, Tristan, Carlos, Anne, Salim, Gilberto, Carl pour ces bons moments. Une pensée particulière à Alexandre et Vivien du bureau cube (qui est évidemment le meilleur bureau), au regretté Blob, au sans doute regretté Blub. À Thomas, seul frère de la fournée 2011, qui m'a accompagné à plein de formations, parfois douteuses mais souvent intéressantes, en tout cas plus marrantes avec lui (courage, la fin approche!). Merci aux moins anciens Christophe et Moudhaffar pour m'avoir hébergé quelques jours, à Virgile, Antoine, Ilaria, mais aussi Florian, Thomas, Marjorie, Pierre, le clan des V. qui savent perpétuer les bonnes traditions. Et à tous ceux qui n'ont pas été cités!

Je n'oublie pas l'unité DRP/RSE du BRGM à laquelle j'ai été rapidement intégré, Farid à qui je pouvais parler librement de tout ce que je faisais, qui m'a consacré du temps et m'a aidé plus d'une fois. Je le remercie chaleureusement.

Enfin, merci au LTN de Polytech de m'avoir permis de terminer ma thèse dans de bonnes conditions, à Yann et David pour leurs bons conseils avant la soutenance, et à Marie pour avoir relu mon manuscrit.

Table des matières

Liste des Figures

Liste des Tableaux

Chapitre 1

Introduction

On s'intéresse dans ce manuscrit à la simulation numérique des écoulements souterrains en milieu poreux, ainsi qu'à l'optimisation du temps de calcul nécessaire à une telle simulation. Ces écoulements mettent en jeu des processus complexes, et font notamment intervenir des phénomènes non linéaires. Dans ce chapitre d'introduction, on commence par motiver l'intérêt de notre étude, avant de mettre en place l'équation de Richards, qui est utilisée dans ce travail pour décrire un écoulement en milieu poreux. On poursuit avec une synthèse des méthodes couramment rencontrées pour la discrétisation d'une telle équation, ainsi que des estimations *a posteriori* que l'on peut mettre en place pour avoir un contrôle de l'erreur entre la solution exacte et la solution approchée. Enfin, on détaille l'apport de ce manuscrit et l'articulation de ses différentes parties.

1.1 Motivations

La simulation numérique en mécanique des fluides s'est considérablement développée ces quarante dernières années. Elle est actuellement utilisée comme un outil d'étude des écoulements dans de nombreux domaines industriels : aéronautique, nucléaire, automobile, industrie pétrolière, etc. En particulier, le groupement de recherche MoMaS (Modélisations Mathématiques et Simulations Numériques liées aux problèmes de gestion des déchets nucléaires) coordonne un certain nombre de projets centrés sur la simulation d'écoulements dans les sous-sols géologiques. Ces dernières années s'est également développé un intérêt pour le stockage géologique de CO_2. Le principe est de transporter une partie du gaz issu d'installations fortement émettrices (centrales à combustibles fossiles, aciéries, cimenteries, etc.), puis de l'injecter dans des formations géologiques adaptées, idéalement de façon définitive et sûre; ce qui contribue à lutter contre l'effet de serre. Une simulation numérique dans ce cadre permet d'évaluer l'avenir d'un tel

stockage et les risques associés.
Un sous-sol géologique est schématisé par un milieu poreux, aux propriétés souvent complexes. Simuler un écoulement dans un tel milieu demande donc des techniques élaborées. L'objectif de ce travail est de développer des outils permettant une simulation :

- **précise** : la solution obtenue doit être assez proche de la solution exacte du problème continu, sur des maillages comportant un nombre de cellules raisonnable. Dans le cadre des écoulements en milieu poreux, modélisés par des équations de diffusion non linéaires, il est également important de garantir une évaluation précise des flux;
- **robuste** : la simulation doit produire une solution de même qualité dans des situations variées, que ce soit au niveau du maillage, des conditions de bord ou encore des propriétés physiques;
- **efficace** : l'utilisation des ressources informatiques disponibles doit être optimisée, de façon à garantir l'obtention d'une solution précise en un temps CPU raisonnable.

Dans cet esprit, notre travail s'articule en deux parties. On met tout d'abord en place une discrétisation complète de l'équation de Richards, qui permet de décrire l'écoulement d'un mélange eau-air dans lequel la pression de l'air est constante. Nous considérons ici un domaine à deux dimensions d'espace. Le schéma utilisé est d'ordre 2, et reste précis sur des maillages généraux. On élabore ensuite, à maillage fixé, un algorithme d'adaptation du pas de temps, proche de celui introduit dans [31]. Une simulation produit une erreur provenant à la fois de la discrétisation du domaine de calcul, et de la discrétisation de l'intervalle de temps d'observation. Partant de cette remarque, le principe est alors de choisir un pas de temps, éventuellement variable au cours de la simulation, qui équilibre ces deux sources d'erreur. En effet, l'erreur en espace est fixée par le choix du maillage, donc si le pas de temps est mal choisi par l'utilisateur, on arrive soit à une perte de précision de la solution (si le pas de temps est trop grand, c'est alors l'erreur en temps qui domine), soit à un surcoût de temps CPU inutile (si le pas de temps est trop petit, c'est alors l'erreur en espace qui domine). L'équilibrage des deux sources d'erreur vise donc à satisfaire la contrainte d'efficacité définie ci-dessus. La stratégie mise en oeuvre s'appuie sur une analyse *a posteriori* de l'erreur produite par le schéma. Notons enfin que dans notre cas, la présence de non-linéarités suppose la mise en place d'une procédure de linéarisation à chaque temps de calcul. Ainsi, la simulation produit également une erreur due à ces linéarisations. L'algorithme d'adaptation propose donc un critère d'arrêt pour cette procédure de linéarisation, de manière que l'erreur de linéarisation soit négligeable devant les erreurs de discrétisation, sans pour autant nuire au temps CPU.

1.2 Modèle physique

Dans cette section, on présente l'équation de Richards étudiée dans toute la suite du manuscrit. On commence par préciser la schématisation d'un milieu poreux, puis on met en place l'équation de Richards en partant des équations de conservation de la masse décrivant un écoulement diphasique non miscible eau-air. Le dernier paragraphe liste les lois de fermeture les plus utilisées en hydrologie.

1.2.1 Milieu poreux

Un milieu poreux est un matériau dont la phase solide, fortement imbriquée avec la phase fluide, est fixe. Ainsi des sols, des couches sédimentaires, de la plupart des roches, et de certains matériaux vivants. On modélise donc classiquement un sol par une matrice solide (le *squelette*) et un espace interstitiel (constitué de *pores*) perméable à travers lequel s'effectuent des échanges de masse fluide. Cet espace est connexe par arcs, deux points de la partie fluide sont liés par un trajet entièrement intérieur à l'espace interstitiel (voir la figure 1.1). Deux grandeurs macroscopiques décrivent un milieu poreux :

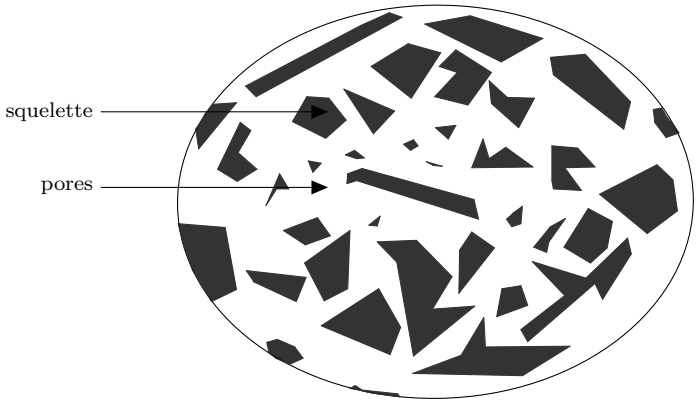


FIGURE 1.1: Schéma d'un milieu poreux.

- Pour un volume élémentaire donné, centré en un point $\mathbf{x}$ du milieu, la *porosité* («fraction de vide») $\varphi(\mathbf{x})$ est le rapport (sans dimension) entre le volume occupé par les pores et le volume total élémentaire.
- La *perméabilité intrinsèque* $\overline{\mathbb{K}}$ ne dépend que de la géométrie du milieu, et indique l'aptitude de celui-ci à être traversé par un écoulement. Lorsque le milieu est *isotrope*, la perméabilité $\overline{\mathbb{K}}$ est indépendante de la direction, et une perméabilité scalaire suffit à le décrire. Sinon, le milieu est dit *anisotrope*, et la perméabilité $\overline{\mathbb{K}}$

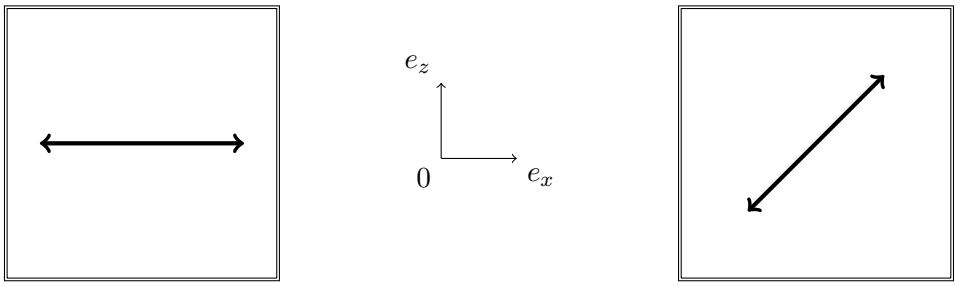
prend la forme d'un tenseur symétrique [45], soit en deux dimensions :

$$\overline{\mathbb{K}} = \begin{pmatrix} k_{xx} & k_{xz} \\ k_{xz} & k_{zz} \end{pmatrix}.$$

Par exemple, un tenseur de la forme $\overline{\mathbb{K}} = \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix}$, avec $\alpha < 1$, traduit une direction préférentielle horizontale pour l'écoulement. Un tenseur de la forme

$$\overline{\mathbb{K}} = R_{\pi/4} \begin{pmatrix} 1 & 0 \\ 0 & \alpha \end{pmatrix} R_{\pi/4}^{\mathsf{t}}, \quad \text{avec} \quad R_{\pi/4} = \begin{pmatrix} \cos(\pi/4) & -\sin(\pi/4) \\ \sin(\pi/4) & \cos(\pi/4) \end{pmatrix},$$

traduit une direction préférentielle oblique pour l'écoulement (voir la figure 1.2). On doit à Matheron [60] l'essentiel des démonstrations des propriétés de $\overline{\mathbb{K}}$. On notera en particulier que c'est un tenseur symétrique, défini et positif.



(A) Horizontale. (C) Oblique.

FIGURE 1.2: Direction préférentielle pour l'écoulement.

Décrire l'écoulement d'un fluide de masse volumique ρ (en $kg.m^{-3}$) et de viscosité dynamique μ (en $Pa.s$) données, c'est connaître en chaque point du milieu et à chaque instant :

- sa pression p (en Pa),
- le volume occupé par ce fluide relativement au volume des pores : c'est la saturation s. Par définition, on a : $0 < s < 1$.

On définit également la *perméabilité relative* $k(s)$ d'un fluide, qui décrit selon sa saturation dans le milieu sa capacité à s'écouler.

1.2.2 Équation de Richards

On cherche à décrire un écoulement diphasique non miscible eau-air (phases notées respectivement w et a par la suite) dans un milieu poreux. On se place dans un domaine $\Omega \subset \mathbb{R}^2$, entièrement constitué d'un milieu poreux tel que présenté ci-dessus. L'équation de conservation de la masse dans Ω (que l'on suppose fermé, sans échange avec l'extérieur), appliquée à chaque phase $\alpha \in \{w, a\}$, décrit la dynamique de l'écoulement :

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

Les dépendances spatiales et temporelles sont ici omises par souci de clarté. Les vitesses v_α de chaque phase s'expriment à l'aide de la loi de Darcy généralisée [28] :

$$v_\alpha = -\overline{\mathbb{K}} \frac{k_\alpha(s_\alpha)}{\mu_\alpha} \left[\nabla \left(p_\alpha + \rho_\alpha g z\right)\right]. \tag{1.2}$$

Notons en particulier que, si la direction du gradient de pression n'est pas une des directions principales du tenseur $\overline{\mathbb{K}}$, alors la direction d'écoulement du fluide et la direction du gradient ne sont pas colinéaires.
En l'état, il s'agit d'un système de deux équations à quatre inconnues : p_w, p_a, s_w et s_a. Deux contraintes sont nécessaires pour fermer ce système. La première découle directement de la définition des saturations :

$$s_a + s_w = 1.$$

Nous avons donc une seule inconnue de saturation $s := s_w$, l'autre se déduisant de la formule précédente. La deuxième contrainte traduit le fait que l'on sait exprimer la différence des deux pressions inconnues par :

$$p_a - p_w = \delta p.$$

La fonction δp ainsi que ses dépendances seront précisées plus loin. On fait alors l'hypothèse, classique en hydrogéologie, que l'air circule librement dans le milieu poreux : sa viscosité μ_a est considérée nulle (de l'ordre de $10^{-6} Pa.s$ en réalité), si bien que, d'après la loi de Darcy (1.2) appliquée à l'air, on a :

- soit $k_a(s) = 0$, ce qui implique que le milieu est entièrement saturé en eau. Il y a alors dégénérescence du système en deux équations de Darcy;

- soit $\nabla\left(p_a + \rho_a g z\right) = 0$. La répartition de la pression d'air est alors hydrostatique : $p_a = p_0 - \rho_a g z$, où p_0 est la pression de l'air à la surface ($z = 0$), c'est-à-dire la pression atmosphérique, que l'on considère constante durant les temps d'observation.

Dans ce deuxième cas, on connaît alors entièrement la pression de l'air p_a, ainsi que la saturation en air s_a, sous réserve de connaître la saturation s. Ainsi, on ne s'intéresse plus désormais qu'à l'équation de continuité en eau (1.1), qui s'écrit :

$$\varphi\rho_w\partial_t s + \varphi s\partial_t\rho_w + \rho_w s\partial_t\varphi - \nabla\cdot\left[\rho_w\overline{\mathbb{K}}\frac{k_w(s)}{\mu_w}\nabla\left(p_w - p_a + p_a + \rho_w g z\right)\right] = 0.$$

À ce stade, on effectue deux hypothèses supplémentaires (nous rappelons que «constant» signifie indépendant de la variable temporelle, et «uniforme» indépendant de la variable d'espace) :

- l'eau est considérée homogène et incompressible. Ainsi, la masse volumique ρ_w est constante et uniforme.
- le squelette est indéformable : la porosité φ est constante.

En remplaçant la pression p_a par son expression $p_a = p_0 - \rho_a g z$ et en introduisant le différentiel de masse volumique $\delta\rho = \rho_a - \rho_w$, on obtient alors :

$$\varphi\rho_w\partial_t(s) - \rho_w\nabla\cdot\left[\overline{\mathbb{K}}\frac{k_w(s)}{\mu_w}\nabla\left(p_0 - \delta p - \delta\rho g z\right)\right] = 0.$$

On utilise le fait que la masse volumique de l'air est négligeable devant celle de l'eau ($\rho_w \simeq 1000\rho_a$), si bien que $\delta\rho \simeq -\rho_w$. On décompose alors les deux termes intervenant dans le flux, pour aboutir à une première forme de l'équation de Richards, dont l'inconnue est la saturation s :

$$\varphi\partial_t s - \frac{\rho_w g}{\mu_w}\nabla\cdot\left(\overline{\mathbb{K}}k_w(s)\nabla z\right) = -\frac{1}{\mu_w}\nabla\cdot\left[\overline{\mathbb{K}}k_w(s)\nabla\left(\delta p\right)\right]. \tag{1.3}$$

Sous cette forme, on voit clairement apparaître les effets gravitaires (terme convectif) et les effets capillaires (terme diffusif). Le différentiel δp est déterminé expérimentalement. Il s'agit d'une fonction de la saturation s nommée *pression capillaire* et notée p_c. Elle est monotone en zone insaturée, donc inversible. On remarque que l'on a $\nabla(\delta p) = \nabla(p_c) \simeq -\nabla(p_w)$. Ainsi, l'équation de Richards (1.3) peut se réécrire :

$$\varphi\partial_t(s) - \frac{\rho_w g}{\mu_w}\nabla\cdot\left(\overline{\mathbb{K}}k_w(s)\nabla\left(\frac{p_w}{\rho_w g} + z\right)\right) = 0.$$

On introduit maintenant quelques grandeurs spécifiques à l'hydrogéologie :

- la *teneur en eau* θ définie par $\theta(s) = \varphi s$,
- la conductivité hydraulique $\mathbb{K}$ $[m.s^{-1}]$ définie par $\mathbb{K}(s) = \overline{\mathbb{K}} k_w(s) \rho_w g / \mu_w$,
- la charge hydraulique ψ $[m]$ définie par $\psi = p_w/(\rho_w g)$.

La pression capillaire p_c étant une fonction inversible, on peut exprimer la teneur en eau et la perméabilité relative comme des fonctions de la pression, puis de la charge hydraulique. On notera toujours par abus de notation θ et k de telles fonctions (on omet l'indice w par la suite), et aussi $\mathbb{K}(\psi) = \overline{\mathbb{K}} k(\psi)$. On arrive ainsi à la forme suivante de l'équation de Richards, où l'inconnue est la charge hydraulique ψ :

$$\boxed{\partial_t(\theta(\psi)) - \nabla \cdot (\mathbb{K}(\psi) \nabla (\psi + z)) = 0.} \tag{1.4}$$

Nous travaillerons dans toute la suite du manuscrit sur cette équation. D'autres formes peuvent être proposées mais présentent certains inconvénients. On peut notamment faire l'hypothèse que la teneur en eau θ est une fonction dérivable de la charge hydraulique ψ et développer par composition la dérivée en temps, mais on perd alors la conservativité. On peut également réécrire l'équation (1.4) en choisissant comme inconnue θ, mais cette formulation n'est pas applicable lorsque le milieu devient saturé. On trouvera une discussion à propos de ces différentes formulations dans [47]. La présente forme est conservative et reste valable dans des milieux hétérogènes éventuellement saturés, où elle dégénère en l'équation de Darcy. Enfin, il est possible d'utiliser une transformée de Kirchhoff pour obtenir une équation dont le terme elliptique est linéaire. Cette stratégie permet notamment de simplifier l'analyse et d'obtenir des résultats de convergence (voir par exemple [44]). Nous avons cependant décidé de rejeter cette approche, d'une part parce que l'équation n'est alors plus conservative, et d'autre part parce que l'interprétation physique de la nouvelle inconnue $\int_0^\psi K(u)\, \mathrm{d}u$ est délicate.

1.2.3 Lois de fermeture

L'inversibilité de la pression capillaire p_c n'est valable qu'en milieu insaturé, donc nous ne disposons pas *a priori* de l'expression de la teneur en eau θ et de la perméabilité relative k à saturation. La façon la plus courante de remédier à ce problème est de prolonger par continuité ces lois à saturation par leurs valeurs maximales respectives :

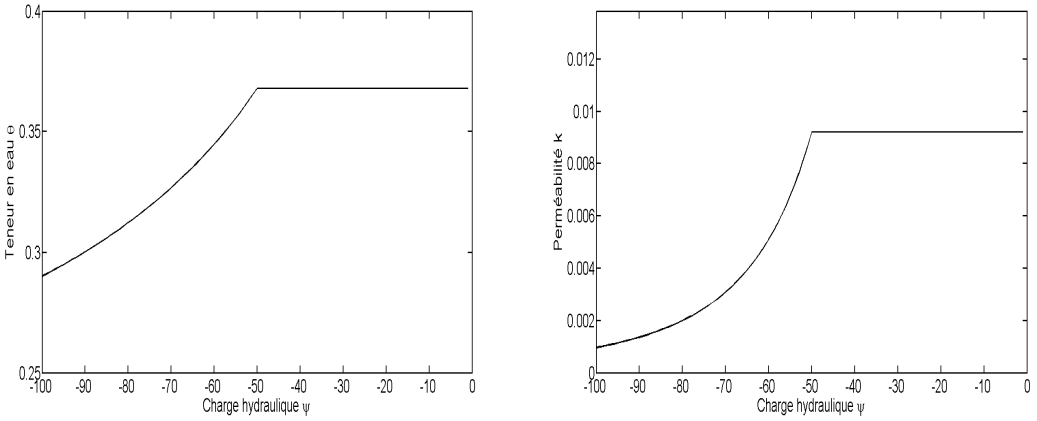
$$\theta(\psi) = \begin{cases} \theta(\psi) & \text{si } \psi < -\psi_e \\ \theta_\mathrm{s} & \text{si } \psi \geq -\psi_e \end{cases}, \quad k(\psi) = \begin{cases} k(\psi) & \text{si } \psi < -\psi_e \\ k_\mathrm{s} & \text{si } \psi \geq -\psi_e \end{cases}.$$

La valeur $\psi_e \geq 0$ désigne la pression d'entrée d'air. Au-delà de cette valeur, le milieu est saturé en eau. La valeur à saturation θ_s ne vaut pas exactement la porosité φ, car il peut rester des bulles d'air emprisonnées dans les pores les plus petits. Ainsi, en milieu insaturé, l'équation de Richards (1.4) est parabolique non linéaire en temps et en espace. À saturation, elle dégénère en une équation elliptique linéaire. Nous présentons ci-dessous quelques modèles parmi les plus utilisés en hydrologie. La valeur résiduelle θ_r correspond à la valeur de la teneur en eau lorsque le milieu est *désaturé* ($\psi \to -\infty$). Cette valeur est non nulle en raison de phénomènes d'adsorption, dépendant de la nature du sol. Pour une synthèse récente des mécanismes de rétention d'eau, on pourra se référer à [71]. Les coefficients dont la valeur n'est pas précisée dépendent du sol considéré.

Modèle de Brooks et Corey [18]

$$\theta(\psi) = \theta_r + (\theta_s - \theta_r)\left(\frac{|\psi|}{\psi_e}\right)^{-\lambda} \quad \text{et} \quad k(\psi) = k_s\left(\frac{|\psi|}{\psi_e}\right)^{-n}.$$

Ici, $n = (2+l)\lambda + 2$, l est un paramètre traduisant la connectivité des pores, et λ est inversement proportionnel à leur taille moyenne. L'inconvénient de ce modèle est qu'il présente une rupture de pente au niveau de la pression d'entrée d'air ψ_e, ce qui peut nuire à la convergence des schémas numériques (voir la figure 1.3).



(A) Teneur en eau.

(B) Perméabilité relative.

FIGURE 1.3: Lois de fermeture pour le modèle de Brooks et Corey.

Modèle de Haverkamp [50]

$$\theta(\psi) = \theta_r + \frac{\theta_s - \theta_r}{1 + |\tilde{\alpha}\psi|^{\beta}} \quad \text{et} \quad k(\psi) = \frac{k_s}{1 + |\tilde{A}\psi|^{\gamma}}.$$

Ce modèle est plus robuste que le précédent, il sera utilisé dans la partie numérique du chapitre 2. Les paramètres $\tilde{\alpha}$, $\tilde{A}$, β et γ sont empiriques, on renvoie à la sous-section 2.4.1 pour un exemple de jeu de valeurs.

Modèle de Van Genuchten [76]

$$\theta(\psi) = \theta_{\mathrm{r}} + \frac{(\theta_{\mathrm{s}} - \theta_{\mathrm{r}})}{\left(1 + (\xi|\psi|)^{\beta}\right)^{\gamma}} \quad \text{et} \quad k(\psi) = k_{\mathrm{s}} \frac{\left(1 - (\xi|\psi|)^{\beta-1}(1 + (\xi|\psi|)^{\beta})^{-\gamma}\right)^2}{\left(1 + (\xi|\psi|)^{\beta}\right)^{\frac{\gamma}{2}}}.$$

Là encore, les paramètres β, γ et ξ sont empiriques; on utilisera ce modèle dans la sous-section 2.4.2. La loi de perméabilité induit de plus fortes variations que dans le modèle d'Haverkamp. Une modification pour les sols de texture fine a été proposée dans [78]. On peut également citer les modèles de Broadbridge [17], Gardner [46] ou Mualem [61]. Comme nous l'avons vu ci-dessus, l'équation de Richards provient d'une simplification du modèle diphasique eau-air en une seule équation, ce qui présente un intérêt évident. Cependant, sa résolution nécessite la connaissance des lois de fermeture $\theta(\psi)$ et $k(\psi)$, souvent empiriques, qui font elles-mêmes intervenir un certain nombre de paramètres empiriques. Leur détermination expérimentale est entachée d'incertitudes, ce qui constitue une limitation du modèle. La prise en compte de ces incertitudes est un domaine de recherche récent et en pleine expansion [75]. Notons également qu'on ne tient pas compte des variations de comportement du sol suivant que l'eau s'infiltre dans le sol (imbibition) ou s'en exfiltre (drainage). En particulier, on observe expérimentalement que $\theta(\psi)$ varie suivant le cas : c'est le phénomène d'hystérésis [49] (voir la figure 1.4).

1.2.4 Modélisation du sous-sol

Simuler un écoulement en milieu poreux en résolvant numériquement l'équation de Richards (1.4) comporte certaines contraintes :

- Le milieu est en général hétérogène (un sous-sol consiste souvent en l'empilement d'une multitude de couches géologiques, comme on peut le voir sur la figure 1.5), et anisotrope (on observe souvent une direction préférentielle de l'écoulement). Dans ce cas, la perméabilité intrinsèque $\overline{\mathbb{K}}$ du milieu est un tenseur différent de l'identité, et dépend de la variable d'espace.
- À l'interface entre plusieurs couches géologiques, les propriétés physiques peuvent présenter des discontinuités, en particulier la conductivité $\mathbb{K}$.

Prendre en compte ces contraintes dans notre modélisation impose de travailler avec des maillages non structurés, qui reflètent la forme du milieu et ses éventuelles discontinuités.

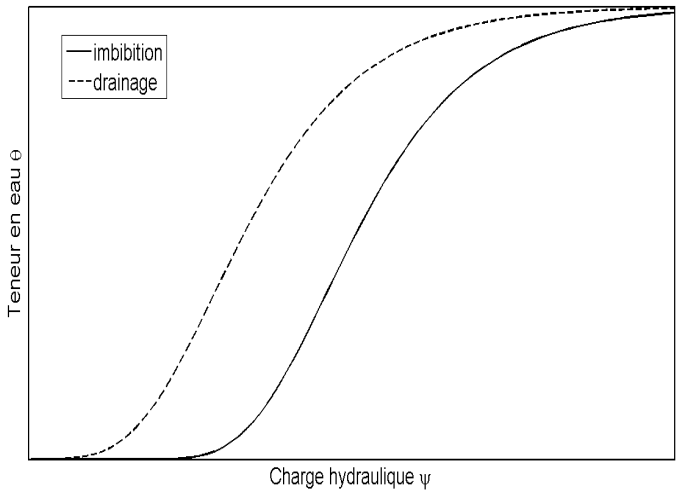
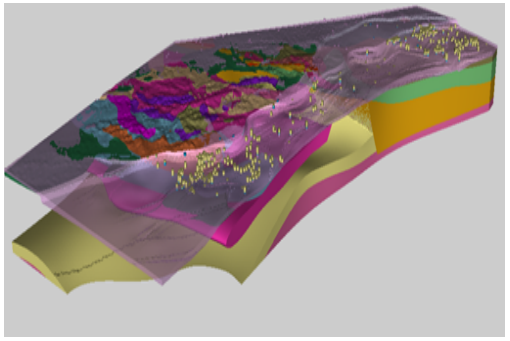
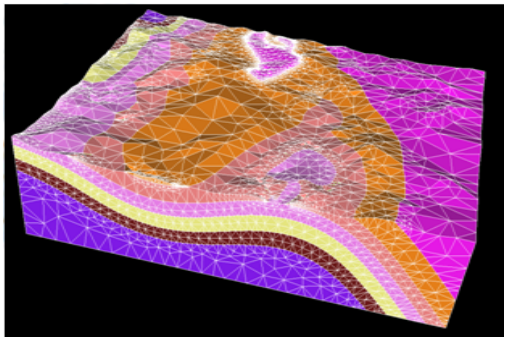


FIGURE 1.4: Phénomène d'hystérésis.



(A) Sous-sol géologique.



(B) Maillage non structuré.

FIGURE 1.5: Schématisation d'un sous-sol géologique et exemple de maillage.

1.3 Méthodes numériques pour les écoulements en milieu poreux

On présente ici rapidement le cœur de ce travail de thèse. La modélisation des écoulements en milieu poreux nécessite l'utilisation d'un schéma précis et robuste vis-à-vis de maillages généraux et des non-linéarités présentes dans l'équation (1.4). Notre choix se porte sur un schéma DDFV (acronyme de *Discrete Duality Finite Volume*) pour la discrétisation en espace, et le schéma BDF2 (pour *Backward Differentiation Formula of order 2*) pour la discrétisation en temps. On cherche ensuite à réduire le temps CPU

en adaptant le pas de temps et le nombre d'itérations de linéarisation au cours de la simulation à l'aide de la théorie des estimations *a posteriori*.

1.3.1 Motivations

Les contraintes précédentes nous poussent à utiliser un schéma de discrétisation choisi :

- conservatif : on travaille avec la forme conservative de l'équation de Richards;
- précis et robuste dans le sens donné dans la section 1.1. En particulier, il doit être valable sur des maillages généraux sans dégradation de l'ordre de convergence, adapté à la discrétisation de flux non linéaires et anisotropes, et capable de traiter des hétérogénéités, voire des discontinuités spatiales des propriétés physiques.

Enfin, comme on l'a vu dans la sous-section 1.2.3, les lois constitutives θ et k sont non linéaires. On doit donc traiter un problème d'évolution en temps, qui nécessite des linéarisations pour se ramener à la résolution de systèmes linéaires à chaque itération en temps. Le modèle requiert donc des moyens de calcul efficaces, une contrainte d'autant plus pertinente si l'on souhaite, par exemple, quantifier les incertitudes de modélisation. Il est particulièrement intéressant dans ce cas de questionner l'optimalité du pas de temps choisi ainsi que du critère d'arrêt de la procédure de linéarisation.

1.3.2 Discrétisation du flux diffusif

On synthétise ici quelques schémas numériques utilisés pour la discrétisation de flux diffusifs non linéaires, notamment celui présent dans l'équation de Richards.
Parmi les schémas de type éléments finis, Knabner [55] a développé une **méthode d'éléments finis mixtes (MFE)** pour la discrétisation de l'équation de Richards, et parvient à un système dont la matrice est définie positive. Cependant, la mise en place est peu aisée et le surcoût de calcul par rapport aux méthodes d'éléments finis classiques, non négligeable.
Les **méthodes de Galerkin discontinues (DG)** forment une classe de schémas populaires pour bon nombre de problèmes d'applications industrielles. Ce sont techniquement des méthodes d'éléments finis, mais qui bénéficient également de propriétés normalement propres aux volumes finis (conservativité locale due à un calcul des flux aux interfaces du maillage, flexibilité dans l'utilisation de maillages non conformes). En particulier, une discrétisation de l'équation de Richards utilisant la version à pénalisation intérieure symétrique de ces méthodes (SIPG) est considérée dans [73].
Dans la famille des méthodes volumes finis, un premier choix «simple» est **le schéma**

à deux points [39] (également appelé TPFA ou VF4), pour lequel on se donne une inconnue locale par volume de contrôle. Nous verrons cependant dans le chapitre 2 qu'il n'est pas adapté à la modélisation en milieu hétérogène et anisotrope.

Le schéma diamant [24] s'appuie sur le maillage éponyme, construit en reliant les sommets d'une arête donnée du maillage de départ aux centres des mailles qui se partagent cette arête. Les valeurs de la fonction inconnue aux sommets de l'arête sont interpolées à partir de ses valeurs aux centres des mailles associées par la méthode des moindres carrés. Ces quatre valeurs servent ensuite à définir les flux numériques. Le schéma est valable sur maillages généraux, mais sa construction rend l'analyse délicate. En particulier, les estimations d'erreur de [24] dépendent d'une coercivité obtenue sous des conditions géométriques exprimées sur le maillage diamant. De plus, le système linéaire associé n'est pas symétrique. Il reste que le schéma est régulièrement utilisé car robuste, et permet de monter en ordre assez facilement.

Les schémas MPFA [1] utilisent des inconnues intermédiaires localisées aux centres des arêtes pour reconstruire un gradient discret sur des sous-volumes du maillage. La solution est supposée affine sur chacun de ces sous-volumes, et les inconnues intermédiaires sont éliminées en exprimant la continuité de la solution ainsi que la conservativité des flux. Le choix d'expression de ces équations conduit à différentes versions : ce sont entre autres les schémas O [1],L [3],G [4] et U [2]. On obtient au final des flux numériques consistants, conservatifs et leur expression fait seulement intervenir les inconnues associées à chaque volume de contrôle. Le principal inconvénient à retenir est la perte de coercivité et/ou de monotonie dans certains cas (fortes hétérogénéités, anisotropie).

Les schémas mimétiques MFD [16] se résument à la recherche d'un produit scalaire discret vérifiant certaines propriétés de coercivité et de consistance. En utilisant la formule de Green, et en construisant un opérateur de flux approché en dualité avec la divergence discrète naturelle (ces opérateurs discrets *miment* les opérateurs continus), on arrive à un système linéaire de type point-selle; la monotonie n'est pas assurée. La méthode est cependant inconditionnellement coercive sur maillages généraux. Leur généralisation à des problèmes non linéaires n'est pas évidente.

Les schémas SUCCES/SUSHI placent des inconnues scalaires au centre de chaque maille et de chaque arête, permettant d'écrire un gradient discret. Le schéma SUCCES [40] élimine ensuite les inconnues d'arêtes en les exprimant comme des combinaisons convexes des inconnues de maille. Le nombre d'inconnues est alors restreint, mais des instabilités numériques peuvent se produire lorsque le tenseur est discontinu. C'est ainsi qu'est apparu le schéma SUSHI [41], qui fait le choix de garder les inconnues d'arêtes à travers lesquelles le tenseur n'est pas régulier, et de fermer le système en écrivant la continuité des flux discrets à travers ces arêtes. La précision du schéma s'en trouve améliorée, au prix d'un nombre d'inconnues plus élevé. Notons qu'une approche unifiée des méthodes volumes finis mixtes, MFD et SUCCES a été développée dans [35].

Les schémas VAG [42] utilisent des inconnues supplémentaires situées aux nœuds du maillage pour définir un gradient discret sur des sous-volumes de chaque maille, qui est ensuite stabilisé afin d'éviter des dégénérescences indésirables. On associe à chaque nœud un volume dépendant des propriétés physiques autour de ce nœud, puis on écrit une équation exprimant un bilan de flux nul sur ce volume. À ce prix, on obtient un schéma inconditionnellement coercif sur maillages généraux. Le schéma a été implémenté pour des écoulements diphasiques incluant l'équation de Richards dans [43]. Notons cependant que le choix de la redistribution des volumes de nœud reste relativement empirique.

Les schémas DDFV ont été introduits dans [52]. Leur principe est, comme pour les schémas VAG, d'ajouter des inconnues aux sommets du maillage. Un maillage *secondaire* est construit à partir du maillage de départ (dit *primaire*), chacun de ses volumes étant associé à un nœud unique. L'équation continue est alors intégrée sur chaque volume de ces deux maillages, ce qui conduit après application de la formule de Green à approcher les composantes normale et tangentielle du flux continu. Ces schémas ont été utilisés sur des problèmes variés ces dernières années, allant de la diffusion scalaire [34] à l'électrocardiologie [23], entre autres. La symétrie du problème continu est préservée (sauf dans le cas de conditions de bord mixtes Dirichlet/Neumann, que l'on étudiera au chapitre 2), et les résultats numériques exhibent une approximation particulièrement précise du gradient, comme cela a été noté dans [51]. Pour ces raisons, auxquelles s'ajoute sa simplicité d'appréhension et de mise en oeuvre sur des domaines à deux dimensions, en tant qu'extension «naturelle» du schéma à deux points, c'est ce choix que l'on retiendra dans ce manuscrit.

1.3.3 Discrétisation du terme instationnaire

L'équation de Richards (1.4) est instationnaire, il est nécessaire de discrétiser la dérivée en temps $\partial_t \theta(\psi)$. Ayant choisi un schéma d'ordre 2 pour la discrétisation du flux diffusif, il semble naturel d'avoir la même exigence pour le terme en temps. Remarquons pour commencer qu'il ne nous est pas permis d'utiliser un schéma explicite ici, comme cela a été mentionné dans [44]. En effet, la teneur en eau θ est constante à saturation, donc non inversible, et le schéma associé est mal posé. Le schéma de Crank-Nicolson est un choix populaire, particulièrement adapté à l'analyse *a posteriori* présentée au chapitre 3 [33]. Il peut néanmoins se révéler instable et introduire des oscillations parasites sur la solution lorsque la condition initiale n'est pas régulière [65, 70], comme dans certains cas tests présentés dans les chapitres 2 et 4. Il est également possible d'utiliser un schéma de Runge-Kutta, mais la présence de plusieurs étapes de calcul ralentit l'exécution du code, tout particulièrement lorsqu'il est couplé avec un solveur non linéaire. Il a également

été relevé dans [73] que le schéma de Runge-Kutta diagonalement implicite d'ordre 3 (DIRK3) est inutilisable lorsqu'une partie du domaine est saturée. En effet, la formule définissant la seconde étape du schéma est de nouveau mal posée à cause de l'absence de terme implicite. La simulation de cas tests raides (voir notamment le cas test de Polmann, sous-section 2.4.2) nous demande enfin d'être exigeants quant à la stabilité du schéma retenu. Le comportement des schémas en temps sur des problèmes raides est bien décrit par l'étude de la A-stabilité [58]. Parmi les méthodes linéaires multipas, on rejette donc les méthodes d'Adams-Moulton et d'Adams-Bashforth au profit de la méthode BDF2 [25], la seule à être A-stable. Il est par ailleurs notable que la montée en ordre dégrade cette A-stabilité : on peut montrer qu'un schéma implicite linéaire, multipas et A-stable est au plus d'ordre 2 (il s'agit de la fameuse seconde barrière de Dahlquist, voir [27]).

1.3.4 Estimations *a posteriori*

La seconde partie de ce travail concerne les estimations *a posteriori*. Celles-ci visent de manière générale à obtenir une borne de l'erreur entre la solution exacte du problème continu et la solution approchée issue de la simulation numérique, et ce à partir de la solution approchée seulement. Le résultat est fort : même si on ne connaît pas la solution exacte (ce qui arrive dans l'immense majorité des cas), on est capable de donner un majorant, voire dans certains cas un minorant de l'erreur commise par la résolution numérique. Lorsque cette borne est entièrement calculable, on parle d'estimations *garanties*. Notre objectif est d'avoir plus précisément un contrôle sur la distribution de l'erreur, et d'utiliser cette information pour élaborer une stratégie d'adaptation du pas de temps et du nombre d'itérations de linéarisation, le maillage étant fixé par ailleurs; nous avons donc besoin de bornes locales en temps. Le principe général est de choisir à chaque itération en temps, un pas de temps «optimal», au sens où il équilibre les sources d'erreur provenant de la discrétisation en temps et de la discrétisation en espace. On détermine aussi un critère d'arrêt pour la procédure de linéarisation qui contraint les erreurs de linéarisation à être négligeables devant les erreurs en temps et en espace.
Plusieurs stratégies sont à notre disposition pour établir des estimations *a posteriori*. La théorie a débuté avec les travaux de Babuška et Rheinboldt [9] dans les années 1970, qui ont proposé des *estimations par résidu* étendues ensuite par Verfürth [77]. Cette méthode borne l'erreur entre les solutions exacte et approchée dans une norme d'énergie. Elle utilise une évaluation de résidus locaux relatifs à la forme forte de l'équation à résoudre, mais font généralement intervenir une constante multiplicative difficile à implémenter [79]. *L'estimateur de Zienkiewicz-Zhu* (voir [80]), donne une estimation locale du gradient de la solution exacte à l'aide de la solution approchée. Cette méthode est

populaire grâce à sa simplicité de mise en oeuvre et son faible coût de calcul, mais ne permet pas de bénéficier d'une estimation garantie. Les *méthodes hiérarchiques* enrichissent l'espace d'approximation (par raffinement en temps et en espace) pour calculer les estimateurs. Elles sont souvent utilisées pour adapter le maillage. La *méthode par résidus équilibrés* résout des problèmes aux limites locaux, où les données de bord sont choisies de manière à équilibrer le résidu intérieur. L'article de Bank et Weiser [10] contient des idées fondamentales (notamment l'hypothèse de saturation ainsi que l'équilibrage des données de bord) reprises par beaucoup de travaux sur les méthodes hiérarchiques et par résidus équilibrés. L'ensemble de ces techniques est synthétisé dans [6] pour les éléments finis.

Concernant les méthodes de volumes finis, Ohlberger obtient une estimation de l'erreur en norme L^1 pour des équations de convection-diffusion non linéaires, respectivement pour des volumes finis centrés par maille dans [62] et centrés par sommet dans [63]. Une stratégie d'adaptation de maillage y est également proposée. Plus spécifiquement pour un schéma DDFV, Omnes [64] propose une estimation de l'erreur en norme d'énergie, en passant par une formulation variationnelle du schéma qui permet d'utiliser des techniques bien connues pour les éléments finis. Des estimations sont obtenues à la fois sur le maillage primaire et sur le maillage secondaire, et font appel à une décomposition de Helmholtz-Hodge.

Notre travail rentre dans la catégorie des *méthodes de flux équilibrés* [68], fondées sur la reconstruction de flux continus à partir des flux discrets du schéma, équilibrés en un certain sens. Ce type de méthode a reçu une attention particulière dans de nombreuses études ces dernières années : des problèmes d'élasticité traités par éléments finis ainsi que l'équation de Poisson dans [30, 57], des schémas DG pour une équation de convection-réaction-diffusion dans [37], des volumes finis pour les écoulements multiphasiques dans [32]. Plus récemment, des développements théoriques ont unifié une classe générale de discrétisations en espace, d'abord pour l'équation de la chaleur [38], puis pour un problème parabolique non linéaire [33]. Dans ce dernier article, des estimations (bornes supérieure et inférieure) robustes et complètement calculables sont écrites à l'aide d'une norme duale espace-temps. D'autres estimations pour l'équation de Richards peuvent être trouvées dans [13], mais utilisent la transformée de Kirchhoff. On s'intéresse pour notre part à la formulation de l'équation de Richards relative à la charge hydraulique, sans application de la transformée de Kirchhoff; un schéma conservatif tel que DDFV est justement conçu pour travailler avec les variables physiques.

Notre travail s'organise en trois parties. On commence par écrire une borne supérieure dans l'esprit de [33, 68]; puis, on propose des reconstructions espace-temps adaptées à la discrétisation DDFV-BDF2 de l'équation de Richards, qui nécessitent une reformulation du schéma BDF2. La non-linéarité du terme en temps nécessite un traitement particulier : on équilibre alors le flux temporel aussi bien que le flux spatial, en cohérence

avec la norme espace-temps utilisée dans les estimateurs. Enfin, on distingue plusieurs composantes d'erreur dans notre estimation : une erreur provenant de la discrétisation en temps, une erreur provenant de la discrétisation en espace et une erreur de linéarisation. On propose un algorithme adaptatif qui permet de choisir un critère d'arrêt pour la procédure de linéarisation, ainsi qu'un pas de temps qui équilibre les erreurs en temps et en espace. Soulignons que les résultats présentés restent généraux et peuvent s'adapter à d'autres discrétisations.

1.4 Contributions de la thèse

L'ensemble des travaux présentés dans ce manuscrit comporte trois parties originales, qui correspondent aux chapitres 2, 3 et 4 du manuscrit.

- On étend une discrétisation DDFV présentée notamment dans [53, 56], au cas d'un flux diffusif de la forme $K(\psi)\nabla\psi$, où le tenseur dépend de l'inconnue comme c'est le cas dans l'équation de Richards (1.4). Une telle extension a été proposée dans [22] pour un flux diffusif de la forme $K(\nabla\psi)\nabla\psi$, comprenant notamment le cas de l'équation du p-laplacien. Lorsque les conditions aux limites sont mixtes de type Dirichlet/Neumann (non homogène), on propose un traitement des arêtes situées à l'interface entre ces deux types de conditions qui permet d'éviter des oscillations numériques indésirables. La discrétisation DDFV-BDF2 de l'équation de Richards est également originale, ce qui justifie des tests numériques qui permettent d'évaluer l'ordre de convergence de la méthode, ainsi que sa stabilité dans différentes configurations.

- Le deuxième volet s'articule autour des estimations *a posteriori* par la méthode dite des flux équilibrés. On étend les travaux récents de [33], qui traitent le cas d'une équation parabolique générique, éventuellement non linéaire en espace, à l'équation de Richards qui présente également une non-linéarité dans le terme instationnaire. On obtient alors une borne supérieure de l'erreur en norme duale. Cela implique notamment l'ajout d'un estimateur dans la borne supérieure déjà connue. Les reconstructions en espace nécessaires à l'écriture de ces estimations sont précisées pour le schéma DDFV utilisé, et on introduit également un terme de correction qui provient notamment de la non-linéarité de la loi de saturation. Une autre nouveauté est la définition des reconstructions en temps adaptées au schéma BDF2, moins naturelles *a priori* que dans le cas où l'on utilise un schéma à un pas.

- Enfin, on applique numériquement les estimations précédentes en proposant un algorithme d'adaptation du pas de temps et du critère d'arrêt des linéarisations, à maillage fixé. À l'ajout près des nouveaux estimateurs provenant de la loi de saturation, cet algorithme est similaire à celui de [31], où il est appliqué à un modèle d'écoulement compositionnel multiphasique. On analyse l'influence des paramètres de l'algorithme, et on regarde le gain en termes de temps CPU et de nombre d'itérations de linéarisation effectuées par rapport à une simulation sans adaptation.

On souligne également que les résultats numériques présentés dans ce manuscrit ont nécessité l'élaboration d'un code de calcul complet, réalisé à l'aide du logiciel Matlab. Une attention particulière a été portée sur les performances; en particulier, l'ensemble du code a été vectorisé (voir l'annexe), accélérant considérablement son exécution.

1.5 Plan du manuscrit

Le manuscrit est organisé de la manière suivante.
Le chapitre 2 présente la résolution numérique de l'équation de Richards (1.4), en espace par une méthode DDFV et en temps par le schéma BDF2. On commence par justifier qualitativement l'utilisation d'un schéma qui reconstruit les deux composantes du flux, plutôt que de se contenter de sa composante normale avec le flux à deux points (schéma volumes finis standard, VF4). La méthodologie DDFV est ensuite introduite sur l'équation de Poisson avec conditions de Dirichlet homogènes, puis étendue successivement à de la diffusion non linéaire avec conditions aux limites mixtes Dirichlet/Neumann et à l'équation de Richards. Puis on construit la dérivée intervenant dans la formule en temps BDF2 à l'aide d'un polynôme d'interpolation. Le système obtenu demeure non linéaire à cause de la forme des lois de saturation et de conductivité hydraulique, on met donc en place une procédure de linéarisation à l'ordre zéro en espace (de type point fixe) et à l'ordre un en temps (de type Newton). On donne des précisions concernant l'assemblage pratique en Matlab ainsi que la résolution efficace du système linéaire et l'initialisation de la procédure de linéarisation. La partie numérique comporte trois cas tests; outre la validation du code à l'aide d'une solution analytique, on teste la robustesse et la précision de la discrétisation sur un cas d'infiltration raide, puis sur une infiltration en milieu hétérogène anisotrope.
Le chapitre 3 est consacré à la mise en place théorique des estimations *a posteriori*. On commence par décrire le cadre fonctionnel adapté à la définition d'une solution faible de l'équation (1.4), et on définit l'erreur que l'on va chercher à contrôler. Puis on précise le cadre d'utilisation de nos estimations, notamment concernant la discrétisation en

temps. On montre comment le schéma BDF2 s'inscrit dans ce cadre, moyennant une reformulation sous la forme d'un schéma à un pas. On écrit ensuite une première borne supérieure globale en espace et locale en temps, qui ne tient pas compte des différentes linéarisations. Les estimateurs obtenus sont intégrés en espace et en temps; ils font intervenir des fonctions abstraites, liées par une relation de *flux équilibrés*, qui demande que ces fonctions vérifient localement le système discret fourni par le schéma. En pratique, de telles fonctions sont reconstruites à partir de la solution approchée. Puis, on précise une nouvelle borne calquée sur la précédente, tenant cette fois compte des linéarisations nécessaires à la résolution de l'équation de Richards (1.4). Cette estimation est calculée à chaque temps de simulation et à chaque itération de linéarisation. La fin du chapitre est consacrée à l'écriture de reconstructions espace-temps adaptées à notre discrétisation DDFV-BDF2.

Dans le chapitre 4, on propose à partir des estimations locales précédentes notre algorithme d'adaptation qui cherche à équilibrer les différentes sources d'erreur. On propose dans cette optique de rassembler les différents estimateurs en trois groupes. Les estimateurs relatifs au résidu et au terme source donnent une estimation de l'erreur due à la discrétisation temporelle, tandis que les estimateurs de flux approchent l'erreur spatiale. Enfin, les estimateurs de linéarisation sont naturellement associés à l'erreur due à la linéarisation du système discret. Notre stratégie d'adaptation consiste alors à choisir un pas de temps pour lequel l'erreur en temps est proche de l'erreur en espace, tout en rendant les erreurs de linéarisation négligeables. Différents cas tests sont ensuite étudiés afin d'observer les performances de cette stratégie d'adaptation. Le cas test analytique proposé au chapitre 2 nous permet dans un premier temps de valider l'algorithme en analysant le comportement du pas de temps adapté au cours de la simulation. On étudie également sa robustesse vis-à-vis des paramètres d'entrée choisis par l'utilisateur en début de simulation, à savoir le pas de temps initial, et les critères d'équilibrage entre les différentes sources d'erreur. Des cas tests plus complexes confirment cette robustesse et sont l'occasion de quantifier le surcoût dû au calcul des estimateurs, ainsi que le gain de temps CPU sur l'ensemble de la simulation par rapport à une simulation «classique», c'est-à-dire sans adaptation.

Chapitre 2

Schéma DDFV pour l'équation de Richards

On s'intéresse à la discrétisation de l'équation de Richards posée sur un domaine spatial à deux dimensions, écrite avec des conditions de bord mixtes Dirichlet/Neumann. Le milieu est éventuellement hétérogène et anisotrope, et on suppose que les discontinuités du tenseur se situent le long des arêtes du maillage. De telles hypothèses restreignent le choix du schéma utilisé pour la discrétisation en espace. Ainsi, le schéma volumes finis standard (VF4), implémenté dans le logiciel de calcul TOUGH2 dédié à la simulation d'écoulements multiphasiques [69], n'est pas adapté dans ce cas. Il faut se tourner vers des reconstructions de gradient plus précises, qui cherchent à reconstituer un gradient complet et non seulement sa composante normale. Notre choix se porte ici sur une discrétisation DDFV, qui peut être vue sous sa forme classique comme une généralisation du schéma VF4 aux maillages généraux. Des inconnues supplémentaires sont ajoutées aux sommets du maillage et permettent d'approcher la composante tangentielle du flux continu. Plusieurs versions de DDFV ont été développées depuis le début des années 2000. D'abord introduites pour des problèmes de diffusion scalaire [34, 52], elles ont été étendues à des équations elliptiques non linéaires de type Leray-Lions [8, 14], à des équations non linéaires en trois dimensions [21], mais aussi au problème de Stokes [29, 56], à l'électrocardiologie [23] ou encore aux équations de Maxwell [54]. On utilise ici une version du schéma DDFV telle que celle utilisée dans [53, 56]. Notons qu'on utilise ici un maillage dual barycentrique pour améliorer la convergence : la définition du gradient par demi-diamant est légèrement différente, mais les estimations *a priori* restent valables. La principale nouveauté consiste en l'extension de cette discrétisation à un tenseur dépendant de l'inconnue. De plus, il faut être particulièrement vigilant sur la discrétisation au bord du domaine lorsque des conditions mixtes Dirichlet/Neumann sont imposées, sous peine de dégrader l'approximation. On propose ainsi un traitement particulier des

arêtes du maillage situées à la frontière entre ces deux types de conditions aux limites.
En ce qui concerne le schéma en temps, on se tourne vers la formule de différentiation rétrograde d'ordre 2 (BDF2), à la fois moins coûteuse qu'un schéma de Runge-Kutta et plus stable que celui de Crank-Nicolson. L'équation de Richards étant non linéaire, à la fois en temps et en espace, on propose un algorithme itératif décrit dans [59] qui couple une linéarisation de type Newton pour le terme en temps, et de type Picard pour le terme en espace.
Enfin, on détaille notre traitement des arêtes mixtes Dirichlet/Neumann qui permet d'éviter l'apparition d'oscillations parasites. Quelques précisions relatives à la résolution numérique du système discret sont également données, avant de réaliser différentes simulations visant à valider la discrétisation DDFV-BDF2 et à étudier son comportement.

2.1 Diffusion en milieu hétérogène anisotrope

On commence par illustrer les limitations du schéma VF4 sur l'équation de Poisson. La convergence peut être perdue sur un maillage non admissible (déformé par exemple). Si la non conformité est restreinte à une interface, on peut montrer que le schéma converge, simplement avec un ordre de 1/2 au lieu de 1 en norme H^1 [19].
Puis, on détaille la mise en place du schéma DDFV. Afin de pouvoir prendre en compte les éventuelles discontinuités du tenseur de conductivité à travers les arêtes du maillage, des inconnues artificielles ψ_σ sont ajoutées au milieu de chaque arête σ, et permettent de définir un gradient complet de chaque côté de l'arête. Chaque inconnue ψ_σ est éliminée algébriquement en imposant la continuité normale du flux numérique à travers σ.
Dans la suite, on se donne un domaine Ω à deux dimensions, ainsi qu'une partition $\mathcal{T}$ de ce domaine en ouverts polygonaux disjoints K appelés *volumes de contrôle*. La partition $\mathcal{T}$ sera appelée *maillage primaire*. On a ainsi $\bar{\Omega} = \underset{K\in\mathcal{T}}{\cup} \bar{K}$. Un point générique de Ω sera noté $\mathbf{x} = (x, z)$.

2.1.1 Limitations du schéma VF4 pour l'équation de Poisson

On considère l'équation de Poisson avec conditions de Dirichlet homogènes :

$$\begin{cases} -\Delta\psi(\mathbf{x}) = f(\mathbf{x}) & \text{dans } \Omega, \\ \psi(s) = 0 & \text{sur } \partial\Omega, \end{cases} \tag{2.1}$$

où $\partial\Omega$ désigne le bord du domaine Ω. La stratégie volumes finis consiste à se donner pour chaque volume de contrôle K un point $\mathbf{x}_K$ (typiquement, le centre de masse de K) et à chercher une approximation précise ψ_K de $\psi(\mathbf{x}_K)$, où ψ est la solution de (2.1).

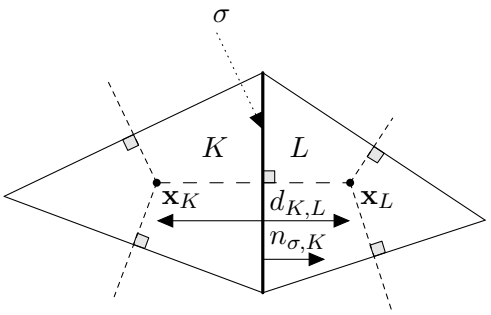


FIGURE 2.1: Mailles respectant la condition d'orthogonalité $\sigma \perp [\mathbf{x}_K\mathbf{x}_L]$.

Pour cela, on approche la relation $\displaystyle\int_K f(\mathbf{x})\ \mathrm{d}\mathbf{x} = \int_K -\Delta\psi(\mathbf{x})\ \mathrm{d}\mathbf{x}$ de la manière suivante (avec les notations de la figure 2.1) :

$$\int_K f(\mathbf{x})\ \mathrm{d}\mathbf{x} = \int_K -\Delta\psi(\mathbf{x})\ \mathrm{d}\mathbf{x} = -\sum_{\sigma\subset\partial K}\int_\sigma \nabla\psi(s)\cdot n_{\sigma,K}\ \mathrm{d}s \simeq -\sum_{\sigma\subset\partial K}|\sigma|\frac{\psi(\mathbf{x}_L)-\psi(\mathbf{x}_K)}{d_{K,L}}.$$

Le quotient différentiel $[\psi(\mathbf{x}_L)-\psi(\mathbf{x}_K)]/d_{K,L}$ est une approximation consistante du flux normal $\nabla\psi\cdot n_{\sigma,K}$ sur l'arête σ si le vecteur normal unitaire $n_{\sigma,K}$ est colinéaire au vecteur $\overrightarrow{\mathbf{x}_K\mathbf{x}_L}$ (condition d'orthogonalité à σ). Lorsque c'est le cas pour chaque arête de $\mathcal{T}$, on parle de maillage *admissible*. Le schéma VF4 associé au problème (2.1) s'écrit alors pour tout K dans $\mathcal{T}$:

$$-\sum_{\sigma\subset\partial K} F_{\sigma,K} = \int_K f(\mathbf{x})\ \mathrm{d}\mathbf{x}, \quad \text{où} \quad F_{\sigma,K} = \begin{cases} |\sigma|\dfrac{\psi_L-\psi_K}{d_{K,L}} \text{ si } \sigma \text{ est à l'intérieur de } \Omega, \\ |\sigma|\dfrac{-\psi_K}{d_{\sigma,K}} \text{ si } \sigma \text{ est au bord de } \Omega. \end{cases}$$

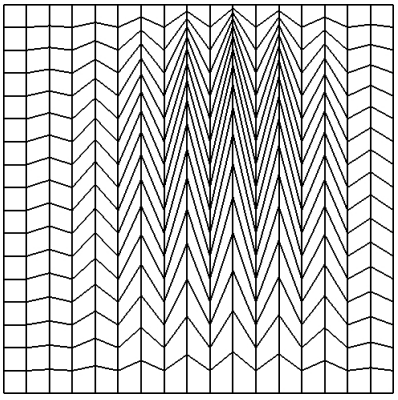
Lorsque l'arête σ se trouve au bord de Ω, on a noté $d_{\sigma,K}$ la distance entre $\mathbf{x}_K$ et le milieu de σ. Dans le cas d'un maillage non admissible, la convergence peut être ralentie (imposant l'utilisation de maillages très fins pour obtenir des erreurs convenables), voire perdue sur un maillage déformé de type Kershaw. Par exemple, la solution exacte $\psi_{\text{ex}}(x,z) = sin(\pi x)sin(\pi z)$ vérifie l'équation (2.1) avec le terme source $f(x,z) = 2\pi^2 sin(\pi x)sin(\pi z)$. Après utilisation du schéma VF4[1] sur le maillage déformé $\mathcal{T}_{\text{déf}}$ représenté sur la figure 2.2, on a obtenu un vecteur discret de valeurs $(\psi_K)_{K\in\mathcal{T}_{\text{déf}}}$. On a alors tracé la projection constante par maille $\overline{\psi_{\text{ex}}}$ de ψ_{ex} sur $\mathcal{T}_{\text{déf}}$, ainsi que la

[1]Code réalisé par F. Boyer et S. Krell à l'aide du logiciel libre SCILAB, disponible à l'adresse : `http://math.unice.fr/~krell/index.php?page=CodeScilab`

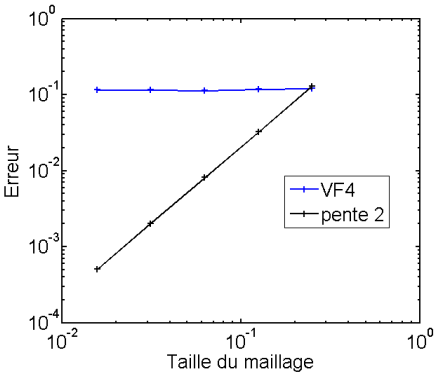
reconstruction constante par maille $\overline{\psi_{\text{app}}}$ définies pour tout $K \in \mathcal{T}_{\text{déf}}$ par :

$$\overline{\psi_{\text{ex}}}|_K = \psi_{\text{ex}}(x_K)\mathbb{1}_K \quad \text{et} \quad \overline{\psi_{\text{app}}}|_K = \psi_K \mathbb{1}_K,$$

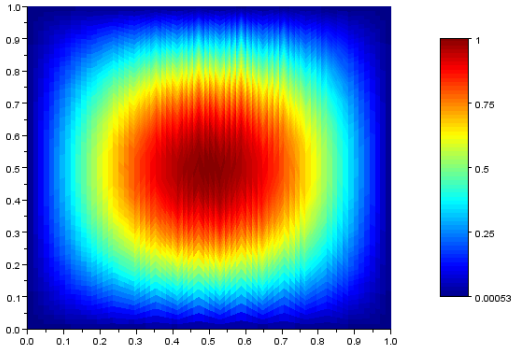
où $\mathbb{1}_K$ désigne la fonction indicatrice de K. La solution approchée est de très mauvaise qualité, et on s'aperçoit après raffinement du maillage $\mathcal{T}_{\text{déf}}$ que la convergence est perdue dans ce cas. L'usage du schéma VF4 est donc compromis dans le cadre d'une modélisation en milieu hétérogène et/ou anisotrope, où la condition d'orthogonalité ne peut pas être vérifiée.



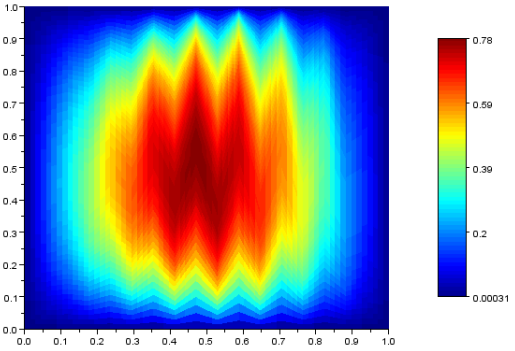
(A) Maillage $\mathcal{T}_{\text{déf}}$.



(B) Erreur L2.



(C) Solution exacte projetée $\overline{\psi_{\text{ex}}}$.



(D) Solution approchée $\overline{\psi_{\text{app}}}$.

FIGURE 2.2: Résolution par le schéma VF4 du problème (2.1).

2.1.2 Schéma DDFV pour la diffusion linéaire

On se place maintenant dans le cadre d'une équation de diffusion linéaire, assortie de conditions de Dirichlet homogènes, pour présenter le schéma DDFV utilisé dans la suite. Il s'agit de résoudre le problème suivant :

$$\begin{cases} -\nabla \cdot (\mathbb{K}(\mathbf{x})\nabla\psi(\mathbf{x})) = f(\mathbf{x}) & \text{dans } \Omega, \\ \psi(s) = 0 & \text{sur } \partial\Omega. \end{cases} \tag{2.2}$$

Principe général Ici, le tenseur de conductivité $\mathbb{K}$ est matriciel, et éventuellement discontinu à travers les arêtes du maillage $\mathcal{T}$. Comme on l'a vu ci-dessus, tenter de reconstruire seulement la composante normale du gradient demande une condition d'orthogonalité portant sur $\mathcal{T}$ assez restrictive en pratique. L'idée des schémas DDFV est de considérer les sommets du maillage comme des inconnues, afin de pouvoir également reproduire la composante tangentielle du gradient continu, et ainsi de construire un gradient discret consistant sur des maillages généraux. En utilisant les notations de la figure 2.3, on cherche ainsi de manière très naturelle un gradient discret ∇_σ vérifiant :

$$\begin{cases} \nabla_\sigma\psi \cdot \dfrac{\overrightarrow{\mathbf{x}_K\mathbf{x}_L}}{\left\|\overrightarrow{\mathbf{x}_K\mathbf{x}_L}\right\|} = \dfrac{\psi_L - \psi_K}{|\sigma^\star|} \simeq \nabla\psi \cdot \dfrac{\overrightarrow{\mathbf{x}_K\mathbf{x}_L}}{\left\|\overrightarrow{\mathbf{x}_K\mathbf{x}_L}\right\|}, \\ \nabla_\sigma\psi \cdot \dfrac{\overrightarrow{\mathbf{x}_A\mathbf{x}_B}}{\left\|\overrightarrow{\mathbf{x}_A\mathbf{x}_B}\right\|} = \dfrac{\psi_B - \psi_A}{|\sigma|} \simeq \nabla\psi \cdot \dfrac{\overrightarrow{\mathbf{x}_A\mathbf{x}_B}}{\left\|\overrightarrow{\mathbf{x}_A\mathbf{x}_B}\right\|}, \end{cases} \tag{2.3}$$

où $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^2$. Pour des mailles non dégénérées, on a $\det(\mathbf{x}_L - \mathbf{x}_K, \mathbf{x}_B - \mathbf{x}_A) \neq 0$, donc ces deux conditions définissent un unique gradient discret. Ces nouvelles inconnues doivent bien entendu être attachées à des équations supplémentaires, pour ne pas aboutir à un système sous-déterminé. Ainsi, de la même manière que la partition *primaire* $\mathcal{T}$ de Ω est en bijection avec l'ensemble des points $\mathbf{x}_K$, on définit un maillage *secondaire* $\mathcal{S}$ dont chaque élément est centré sur un sommet du maillage $\mathcal{T}$. Pour un sommet $\mathbf{x}_A$ donné, l'élément A de $\mathcal{S}$ correspondant est obtenu en reliant les centres des mailles voisines ainsi que les milieux des arêtes se partageant le sommet $\mathbf{x}_A$ (voir la figure 2.4). Le maillage $\mathcal{S}$ décrit ici est *barycentrique* car il permet une meilleure convergence que la version *directe* qui relie directement les centres de masse entre eux. Cela a notamment été souligné dans [56]. Notons enfin que l'acronyme DDFV provient du fait que la divergence discrète «naturelle» associée à la discrétisation volumes finis de l'équation (2.2) est en dualité avec le gradient discret via un analogue discret de la formule de Green. On se réfèrera par exemple à [29, 34] pour plus de détails, ou à [23] pour la version du gradient par demi-diamant utilisée ici.

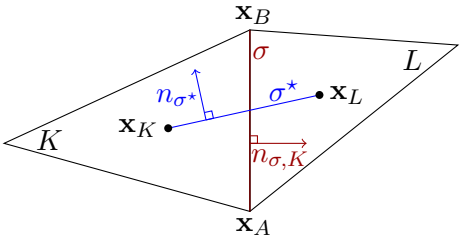
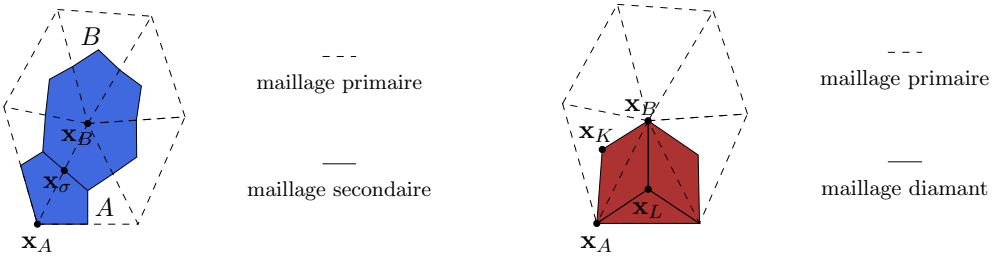


FIGURE 2.3: Principe général du gradient DDFV.



(A) Maillage secondaire.

(B) Maillage diamant.

FIGURE 2.4: Différents maillages pour DDFV.

Notations On rassemble ici les notations précédentes ainsi que celles utilisées dans la suite du chapitre.

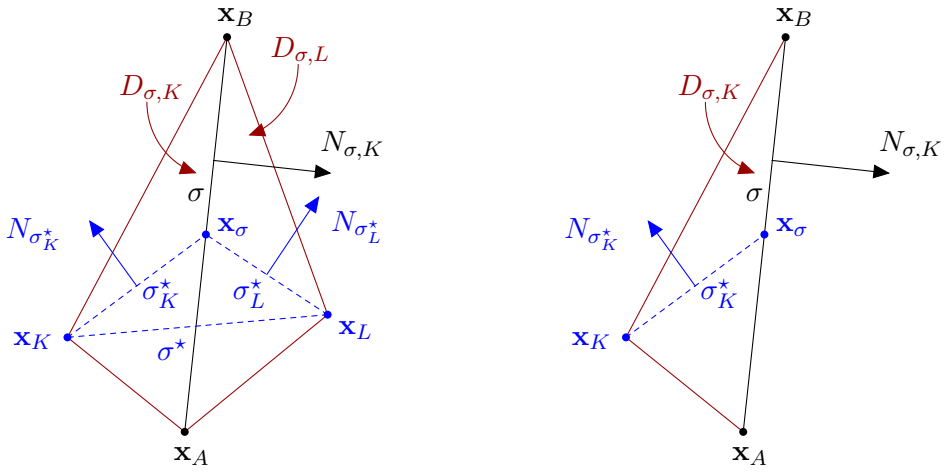
- Pour un élément K de $\mathcal{T}$, ∂K est sa frontière;
- On note $\mathcal{F}$ l'ensemble des arêtes du maillage primaire, que l'on partitionne en arêtes internes et de Dirichlet : $\mathcal{F} = \mathcal{F}^{\mathrm{I}} \cup \mathcal{F}^{\mathrm{D}}$. Pour une arête interne donnée σ de $\mathcal{F}^{\mathrm{I}}$, on peut trouver des mailles K et L de $\mathcal{T}$ telles que $\sigma = \partial K \cap \partial L$. Une arête du bord délimite une unique maille primaire que l'on notera K.
- Pour σ dans $\mathcal{F}$, on note $\mathbf{x}_\sigma$ son centre, $\sigma^\star$ le segment $[\mathbf{x}_K \mathbf{x}_L]$, et $\sigma_K^\star$ le segment $[\mathbf{x}_K \mathbf{x}_\sigma]$.
- Pour K dans $\mathcal{T}$, on note $\mathbf{x}_K$ un point de K (par exemple son centre de masse), et $\mathcal{F}_K$ le sous-ensemble de $\mathcal{F}$ formé des arêtes σ vérifiant $\partial K = \underset{\sigma \in \mathcal{F}_K}{\cup} \sigma$.
- Le maillage secondaire $\mathcal{S}$ se compose de mailles centrées sur un sommet intérieur, et de celles, dégénérées, centrées sur un sommet du bord : $\mathcal{S} = \mathcal{S}^{\mathrm{I}} \cup \mathcal{S}^{\mathrm{D}}$.
- Pour A dans $\mathcal{S}$, on note $\mathbf{x}_A$ le sommet de $\mathcal{T}$ intérieur à A et $\mathcal{F}_A$ le sous-ensemble de $\mathcal{F}$ formé des arêtes σ qui ont $\mathbf{x}_A$ pour extrémité.

- Pour deux mailles voisines K et L telles que $\sigma = \partial K \cap \partial L$, $n_{\sigma,K}$ est la normale unitaire à σ orientée de K vers L, et on définit $N_{\sigma,K} := |\sigma| n_{\sigma,K}$.
- Pour σ dans $\mathcal{F}$ d'extrémités $\mathbf{x}_A$ et $\mathbf{x}_B$, $n_{\sigma_K^\star}$ est la normale unitaire à $\sigma_K^\star$ orientée de A vers B, et on définit $N_{\sigma_K^\star} := |\sigma_K^\star| n_{\sigma_K^\star}$.
- Pour σ dans $\mathcal{F}^{\mathrm{I}}$, le quadrilatère $[\mathbf{x}_K \mathbf{x}_A \mathbf{x}_L \mathbf{x}_B]$ est appelé *diamant* et noté D_σ. Il est lui-même composé des triangles $[\mathbf{x}_K \mathbf{x}_A \mathbf{x}_B]$ et $[\mathbf{x}_A \mathbf{x}_L \mathbf{x}_B]$, appelés *demi-diamants* et notés respectivement $D_{\sigma,K}$ et $D_{\sigma,L}$. Pour une arête du bord, D_σ est seulement constitué du demi-diamant $D_{\sigma,K}$ (voir la figure 2.5). On peut remarquer qu'un diamant donné D_σ est associé à une unique arête primaire σ et à une unique arête secondaire $\sigma^\star$. On note $\mathcal{D}$ l'ensemble des demi-diamants, qui forme une partition du domaine Ω.
- On note ψ_C la valeur approchée de la solution continue sur la maille C, où C appartient à l'ensemble $\{K, L, A, B\}$.
- Enfin, l'inconnue discrète fournie par le schéma DDFV est $\Psi := (\psi_K, \psi_A)_{K \in \mathcal{T}, A \in \mathcal{S}^{\mathrm{I}}}$. Seules les mailles secondaires intérieures sont prises en compte, puisque les mailles secondaires du bord sont associées à des sommets du bord, où la valeur de ψ est connue (imposée par la condition de Dirichlet). L'inconnue Ψ peut être vue comme un vecteur de $\mathbb{R}^{|\mathcal{T}|+|\mathcal{S}^{\mathrm{I}}|}$, où $|\mathcal{T}|$ (respectivement $|\mathcal{S}^{\mathrm{I}}|$) désigne le cardinal de l'ensemble $\mathcal{T}$ (respectivement $\mathcal{S}^{\mathrm{I}}$).

Description du schéma Suivant la stratégie développée dans [23, 56], pour σ dans $\mathcal{F}^{\mathrm{I}}$, on introduit en $\mathbf{x}_\sigma$ une inconnue auxiliaire ψ_σ qui sera éliminée algébriquement par la suite. On est ainsi amené à définir un gradient discret pour chaque demi-diamant $D_{\sigma,K}$ et $D_{\sigma,L}$, ce qui nous permet d'éviter des pertes de convergence dans le cas où le tenseur $\mathbb{K}$ présente des discontinuités à travers σ. Pour σ dans $\mathcal{F}^{\mathrm{D}}$, ψ_σ est donné par la condition de Dirichlet, et il n'y a qu'un gradient à définir. Les gradients $\nabla_{\sigma,K}$ et $\nabla_{\sigma,L}$ sont constants sur $D_{\sigma,K}$ et $D_{\sigma,L}$ et définis par :

$$\begin{cases} \nabla_{\sigma,K}\Psi = \dfrac{1}{2|D_{\sigma,K}|}\left[(\psi_\sigma - \psi_K)N_{\sigma,K} + (\psi_B - \psi_A)N_{\sigma_K^\star}\right], \\ \nabla_{\sigma,L}\Psi = \dfrac{1}{2|D_{\sigma,L}|}\left[(\psi_\sigma - \psi_L)N_{\sigma,L} + (\psi_B - \psi_A)N_{\sigma_L^\star}\right]. \end{cases} \tag{2.4}$$

Il suffit pour arriver à ces formules d'adapter les conditions de consistance (2.3) à chaque demi-diamant. Cela consiste simplement, par exemple pour définir $\nabla_{\sigma,K}\Psi$, à remplacer le point $\mathbf{x}_L$ (respectivement la valeur ψ_L) par le point $\mathbf{x}_\sigma$ (respectivement la valeur ψ_σ). Si l'on intègre l'équation (2.2) sur une maille primaire K, on obtient à l'aide de la formule de Green :



(A) Arête intérieure. (B) Arête du bord.

FIGURE 2.5: Description d'un diamant.

$$-\sum_{\sigma\in\mathcal{F}_K}\int_\sigma \mathbb{K}(s)\nabla\psi(s)\cdot n_{\sigma,K}\,\mathrm{d}s=\int_K f(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

L'équation DDFV correspondante s'écrit :

$$-\sum_{\sigma\in\mathcal{F}_K}\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi\cdot N_{\sigma,K}=|K|f_K, \tag{2.5}$$

où $\mathbb{K}_{\sigma,K}$ (respectivement f_K) approche $\mathbb{K}$ sur le demi-diamant $D_{\sigma,K}$ (respectivement f sur la maille K) :

$$\mathbb{K}_{\sigma,K}\simeq\frac{1}{|D_{\sigma,K}|}\int_{D_{\sigma,K}}\mathbb{K}(\mathbf{x})\,\mathrm{d}\mathbf{x}\qquad\text{et}\qquad f_K\simeq\frac{1}{|K|}\int_K f(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

D'autres choix sont possibles pour $\mathbb{K}_{\sigma,K}$. Par exemple, si le tenseur $\mathbb{K}$ est continu par morceaux sur Ω, on peut prendre $\mathbb{K}_{\sigma,K}\simeq 1/|\sigma|\int_\sigma\overline{\mathbb{K}_K}(s)\,\mathrm{d}s$, où $\overline{\mathbb{K}_K}$ désigne le prolongement continu de $\mathbb{K}|_{\overline{K}}$ à Ω.

De même, l'équation DDFV correspondant à une maille A de $\mathcal{S}^{\mathrm{I}}$ s'écrit :

$$-\sum_{\sigma\in\mathcal{F}_A}\left(\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi\cdot N_{\sigma_K^\star}+\mathbb{K}_{\sigma,L}\nabla_{\sigma,L}\Psi\cdot N_{\sigma_L^\star}\right)=|A|f_A. \tag{2.6}$$

On définit alors les flux à travers les arêtes primaires et secondaires par :

$$\begin{cases} F_{\sigma,K} := -\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi \cdot N_{\sigma,K}, \\ F_{\sigma,L} := -\mathbb{K}_{\sigma,L}\nabla_{\sigma,L}\Psi \cdot N_{\sigma,L}, \\ F_{\sigma^\star} := -\left(\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi \cdot N_{\sigma^\star_K} + \mathbb{K}_{\sigma,L}\nabla_{\sigma,L}\Psi \cdot N_{\sigma^\star_L}\right). \end{cases} \tag{2.7}$$

Le flux $F_{\sigma^\star}$ est continu à travers l'interface $\sigma^\star$ entre deux mailles secondaires A et B par construction. En revanche, les flux $F_{\sigma,K}$ et $F_{\sigma,L}$ ne se raccordent *a priori* pas de manière continue à travers l'arête primaire σ, alors que le flux $-\mathbb{K}\nabla\psi \cdot n_{\sigma,K}$ y est continu. On impose cette continuité au niveau discret à l'aide de l'inconnue auxiliaire : ψ_σ est la solution de l'équation de continuité $F_{\sigma,K} + F_{\sigma,L} = 0$. On notera désormais $F_\sigma := F_{\sigma,K} = -F_{\sigma,L}$ et $N_\sigma := N_{\sigma,K} = -N_{\sigma,L}$. Le système DDFV complet (2.5)-(2.6) s'écrit donc :

$$\begin{cases} \forall K \in \mathcal{T},\ \sum_{\sigma\in\mathcal{F}_K} F_\sigma = |K| f_K, \\ \forall A \in \mathcal{S}^{\mathrm{I}},\ \sum_{\sigma\in\mathcal{F}_A} F_{\sigma^\star} = |A| f_A. \end{cases} \tag{2.8}$$

Pour terminer, on détaille le calcul des flux F_σ et $F_{\sigma^\star}$. On se place dans un diamant intérieur D_σ avec les notations de la figure 2.5. On a :

$$\begin{aligned} F_\sigma = F_{\sigma,K} &= -\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi \cdot N_\sigma, \\ &= \frac{1}{2|D_{\sigma,K}|}\mathbb{K}_{\sigma,K}\left[(\psi_K - \psi_\sigma)N_\sigma \cdot N_\sigma + (\psi_A - \psi_B)N_{\sigma^\star_K} \cdot N_\sigma\right], \\ &= a_K(\psi_K - \psi_\sigma) + b_K(\psi_A - \psi_B); \end{aligned}$$

de même,

$$\begin{aligned} F_{\sigma,L} &= -\mathbb{K}_{\sigma,L}\nabla_{\sigma,L}\Psi \cdot N_{\sigma,L}, \\ &= \frac{1}{2|D_{\sigma,L}|}\mathbb{K}_{\sigma,L}\left[(\psi_L - \psi_\sigma)N_{\sigma,L} \cdot N_{\sigma,L} + (\psi_A - \psi_B)N_{\sigma^\star_L} \cdot N_{\sigma,L}\right], \\ &= \frac{1}{2|D_{\sigma,L}|}\mathbb{K}_{\sigma,L}\left[(\psi_L - \psi_\sigma)N_\sigma \cdot N_\sigma + (\psi_B - \psi_A)N_{\sigma^\star_L} \cdot N_\sigma\right], \\ &= a_L(\psi_L - \psi_\sigma) + b_L(\psi_B - \psi_A). \end{aligned}$$

Enfin,

$$\begin{aligned}
F_{\sigma^\star} &= -\left(\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi\cdot N_{\sigma_K^\star} + \mathbb{K}_{\sigma,L}\nabla_{\sigma,L}\Psi\cdot N_{\sigma_L^\star}\right),\\
&= \frac{1}{2|D_{\sigma,K}|}\mathbb{K}_{\sigma,K}\left[(\psi_K-\psi_\sigma)N_\sigma\cdot N_{\sigma_K^\star} + (\psi_A-\psi_B)N_{\sigma_K^\star}\cdot N_{\sigma_K^\star}\right]\\
&+ \frac{1}{2|D_{\sigma,L}|}\mathbb{K}_{\sigma,L}\left[(\psi_\sigma-\psi_L)N_\sigma\cdot N_{\sigma_L^\star} + (\psi_A-\psi_B)N_{\sigma_L^\star}\cdot N_{\sigma_L^\star}\right],\\
&= b_K(\psi_K-\psi_\sigma) + b_L(\psi_\sigma-\psi_L) + (c_K+c_L)(\psi_A-\psi_B).
\end{aligned}$$

Les coefficients a_K, b_K, c_K, a_L, b_L et c_L ci-dessus sont définis à l'aide des matrices de Gram suivantes :

$$\begin{pmatrix} a_K & b_K\\ b_K & c_K\end{pmatrix} := \frac{1}{2|D_{\sigma,K}|}\mathbb{K}_{\sigma,K}\begin{pmatrix} N_\sigma\cdot N_\sigma & N_\sigma\cdot N_{\sigma_K^\star}\\ N_{\sigma_K^\star}\cdot N_\sigma & N_{\sigma_K^\star}\cdot N_{\sigma_K^\star}\end{pmatrix},$$

et

$$\begin{pmatrix} a_L & b_L\\ b_L & c_L\end{pmatrix} := \frac{1}{2|D_{\sigma,L}|}\mathbb{K}_{\sigma,L}\begin{pmatrix} N_\sigma\cdot N_\sigma & N_\sigma\cdot N_{\sigma_L^\star}\\ N_{\sigma_L^\star}\cdot N_\sigma & N_{\sigma_L^\star}\cdot N_{\sigma_L^\star}\end{pmatrix}.$$

L'équation de continuité des flux normaux s'écrit alors :

$$a_K(\psi_K-\psi_\sigma) + b_K(\psi_A-\psi_B) = a_L(\psi_\sigma-\psi_L) + b_L(\psi_A-\psi_B),$$

ce qui donne :

$$\psi_\sigma = \frac{a_K\psi_K + a_L\psi_L}{a_K+a_L} + \frac{b_K-b_L}{a_K+a_L}(\psi_A-\psi_B).$$

En substituant cette valeur dans l'expression des flux, on trouve finalement :

$$\begin{pmatrix} F_\sigma\\ F_{\sigma^\star}\end{pmatrix} = \mathcal{A}_\sigma\begin{pmatrix}\psi_K-\psi_L\\ \psi_A-\psi_B\end{pmatrix},$$

avec

$$\mathcal{A}_\sigma = \begin{pmatrix} \dfrac{a_Ka_L}{a_K+a_L} & \dfrac{a_Kb_L+a_Lb_K}{a_K+a_L}\\[2ex] \dfrac{a_Kb_L+a_Lb_K}{a_K+a_L} & c_K+c_L-\dfrac{(b_K-b_L)^2}{a_K+a_L}\end{pmatrix}.$$

Dans le cas d'un diamant dégénéré du bord $D_{\sigma,K}$, on a : $F_{\sigma^\star} = -\mathbb{K}_{\sigma,K}\nabla_{\sigma,K}\Psi\cdot N_{\sigma_K^\star}$,

d'où :

$$\begin{pmatrix} F_\sigma \\ F_{\sigma^\star} \end{pmatrix} = \mathcal{A}_\sigma^\partial \begin{pmatrix} \psi_K - \psi_\sigma \\ \psi_A - \psi_B \end{pmatrix}, \text{ avec } \mathcal{A}_\sigma^\partial = \begin{pmatrix} a_K & b_K \\ b_K & c_K \end{pmatrix}. \tag{2.9}$$

Seule l'équation portant sur la maille primaire K fait partie du système, donc la contribution du flux $F_{\sigma^\star}$ n'est pas prise en compte. Après assemblage de ces contributions locales, on aboutit au système linéaire suivant :

$$\mathcal{A}\Psi = \mathcal{B}, \tag{2.10}$$

où $\mathcal{A}$ est une matrice symétrique positive [23], et $\mathcal{B}$ rassemble les contributions du terme source f et de la condition de Dirichlet. La matrice $\mathcal{A}$ possède une structure bloc particulière due au couplage des équations primaires et secondaires :

$$\mathcal{A} = \left(\begin{array}{c|c} \mathcal{A}_{K,L} & \mathcal{A}_{K,A} \\ \hline \mathcal{A}_{A,K} & \mathcal{A}_{A,B} \end{array}\right).$$

2.1.3 Extension à la diffusion non linéaire et aux conditions aux limites mixtes Dirichlet/Neumann

On ajoute un degré de difficulté supplémentaire en autorisant le tenseur $\mathbb{K}$ à dépendre également de l'inconnue ψ, ainsi qu'en considérant des conditions de bord mixtes Dirichlet/Neumann non homogènes. Le problème à résoudre est le suivant :

$$\begin{cases} -\nabla \cdot (\mathbb{K}(\psi, \mathbf{x})\nabla\psi(\mathbf{x})) = f(\mathbf{x}) & \text{dans } \Omega, \\ \psi(s) = \psi_{\mathrm{D}}(s) & \text{sur } \partial\Omega^{\mathrm{D}}, \\ -\mathbb{K}(\psi, s)\nabla\psi(s)\cdot\nu(s) = g(s) & \text{sur } \partial\Omega^{\mathrm{N}}, \end{cases} \tag{2.11}$$

où $\partial\Omega^{\mathrm{D}}$ (respectivement $\partial\Omega^{\mathrm{N}}$) est la partie de la frontière de Ω sur laquelle une condition de Dirichlet est appliquée (respectivement condition de Neumann) et ν est la normale unitaire extérieure à $\partial\Omega = \partial\Omega^{\mathrm{D}} \cup \partial\Omega^{\mathrm{N}}$. Quelques modifications doivent être apportées afin de pouvoir appliquer la stratégie précédente.

Tout d'abord, le tenseur dépend maintenant de l'inconnue, il s'agit donc de trouver une nouvelle approximation de $\mathbb{K}$ sur le demi-diamant $D_{\sigma,K}$. Différents choix sont discutés dans la section 2.3. On suppose pour l'instant qu'on dispose d'une telle approximation $\mathbb{K}_{\sigma,K}(\Psi)$. Les coefficients a_K, b_K, c_K, a_L, b_L et c_L sont donc désormais des fonctions de l'inconnue discrète Ψ. En particulier, l'équation de continuité $F_{\sigma,K} + F_{\sigma,L} = 0$ n'est plus linéaire en Ψ. On a alors recours à une linéarisation de $\mathbb{K}_{\sigma,K}(\Psi)$, on renvoie à la sous-section 2.2.3 pour les détails. Ici, on omet cette dépendance en Ψ par souci de clarté. Elle sera néanmoins conservée dans la matrice locale $\mathcal{A}_\sigma(\Psi)$.

Ensuite, le traitement des arêtes sur lesquelles est imposée une condition de Neumann est particulier; principalement, parce qu'on ne connaît pas les valeurs ponctuelles de la charge ψ sur ces arêtes-là. Ainsi, on va devoir intégrer l'équation (2.11) également sur les mailles secondaires centrées sur un sommet appartenant au bord Neumann. Introduisons quelques notations supplémentaires :

- $\mathcal{F}^{\mathrm{N}}$ est l'ensemble des arêtes sur lesquelles une condition de Neumann est imposée. Ainsi, on a $\mathcal{F} = \mathcal{F}^{\mathrm{I}} \cup \mathcal{F}^{\mathrm{D}} \cup \mathcal{F}^{\mathrm{N}}$. De même, $\mathcal{S}$ est maintenant partitionné en $\mathcal{S} = \mathcal{S}^{\mathrm{I}} \cup \mathcal{S}^{\mathrm{D}} \cup \mathcal{S}^{\mathrm{N}}$.
- Pour une arête σ de $\partial\Omega^{\mathrm{N}}$ d'extrémités $\mathbf{x}_A$ et $\mathbf{x}_B$, on note G_σ une approximation de $\displaystyle\int_\sigma g(s)\,\mathrm{d}s$, et $G_{\sigma,A}$ une approximation de $\displaystyle\int_{[\mathbf{x}_\sigma \mathbf{x}_A]} g(s)\,\mathrm{d}s$. On a donc $G_\sigma = G_{\sigma,A} + G_{\sigma,B}$.
- Pour une maille primaire K de $\mathcal{T}$, on note $\mathcal{F}_K^N = \mathcal{F}_K \cap \mathcal{F}^{\mathrm{N}}$ l'ensemble des arêtes de ∂K appartenant au bord Neumann, et $\mathcal{F}_K^{\mathrm{I}} = \mathcal{F}_K \setminus \mathcal{F}^{\mathrm{N}}$ son complémentaire. On note de même $\mathcal{F}_A^N = \mathcal{F}_A \cap \mathcal{F}^{\mathrm{N}}$ l'ensemble des arêtes ayant $\mathbf{x}_A$ pour extrémité et appartenant au bord Neumann, et $\mathcal{F}_A^{\mathrm{I}} = \mathcal{F}_A \setminus \mathcal{F}^{\mathrm{N}}$ son complémentaire.

Le système d'équations (2.5)-(2.6) prend la forme :

$$\begin{cases} \forall K \in \mathcal{T}, \quad -\displaystyle\sum_{\sigma \in \mathcal{F}_K^{\mathrm{I}}} \mathbb{K}_{\sigma,K} \nabla_{\sigma,K}\Psi \cdot N_\sigma + \sum_{\sigma \in \mathcal{F}_K^N} G_\sigma = |K| f_K, \\ \forall A \in \mathcal{S}^{\mathrm{I}} \cup \mathcal{S}^{\mathrm{N}}, \quad -\displaystyle\sum_{\sigma \in \mathcal{F}_A^{\mathrm{I}}} \left(\mathbb{K}_{\sigma,K} \nabla_{\sigma,K}\Psi \cdot N_{\sigma_K^\star} + \mathbb{K}_{\sigma,L} \nabla_{\sigma,L}\Psi \cdot N_{\sigma_L^\star} \right) \\ \qquad\qquad\qquad -\displaystyle\sum_{\sigma \in \mathcal{F}_A^N} \left(\mathbb{K}_{\sigma,K} \nabla_{\sigma,K}\Psi \cdot N_{\sigma_K^\star} - G_{\sigma,A} \right) = |A| f_A. \end{cases}$$

Détaillons la troisième ligne. Pour une arête σ de $\mathcal{F}^{\mathrm{N}}$ donnée, il y a une contribution qui provient de l'arête $\sigma_K^\star$ (il n'y a pas d'arête $\sigma_L^\star$ puisqu'on se trouve sur une arête du bord), et une contribution qui provient de l'arête $[\mathbf{x}_\sigma \mathbf{x}_A]$ qui ferme la maille secondaire correspondante (voir la figure 2.6). Cette «demi-arête» est incluse dans σ, il est donc logique de lui appliquer la condition de Neumann idoine. Pour une arête σ de $\mathcal{F}^{\mathrm{N}}$, on notera encore $F_{\sigma^\star}$ le flux $-\mathbb{K}_{\sigma,K}\nabla\Psi \cdot N_{\sigma_K^\star}$. Son expression fait toujours intervenir l'inconnue auxiliaire ψ_σ, qui n'est cette fois pas éliminée par une condition de continuité mais par la condition de Neumann :

$$F_\sigma = a_K(\psi_K - \psi_\sigma) + b_K(\psi_A - \psi_B) = G_\sigma.$$

De même que l'équation de continuité des flux normaux, cette équation est non linéaire en Ψ et sa résolution nécessite des linéarisations de a_K et b_K, qui seront explicitées plus

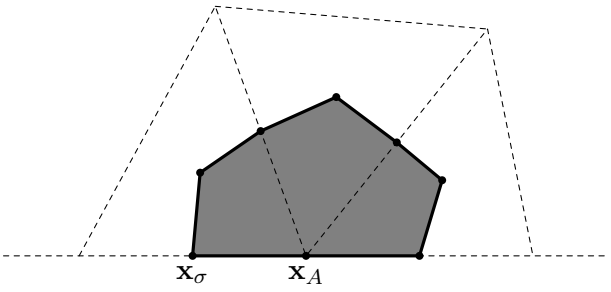


FIGURE 2.6: Une maille secondaire centrée sur un bord Neumann.

loin. Si l'on dispose de telles linéarisations, que l'on appelle encore a_K et b_K, on trouve alors en réinjectant :

$$F_{\sigma^\star} = \left(c_K - b_K^2/a_K\right)(\psi_A - \psi_B) + b_K/a_K G_\sigma.$$

Le système précédent se synthétise de la manière suivante :

$$\begin{cases} \forall K \in \mathcal{T},\ \displaystyle\sum_{\sigma \in \mathcal{F}_K^{\mathrm{I}}} F_\sigma + \sum_{\sigma \in \mathcal{F}_K^N} G_\sigma = |K| f_K, \\ \forall A \in \mathcal{S}^{\mathrm{I}} \cup \mathcal{S}^{\mathrm{N}},\ \displaystyle\sum_{\sigma \in \mathcal{F}_A} F_{\sigma^\star} + \sum_{\sigma \in \mathcal{F}_A^N} G_{\sigma,A} = |A| f_A. \end{cases}$$

Traitement des conditions de bord mixtes Dirichlet/Neumann Il reste à préciser le traitement de l'intersection Dirichlet/Neumann, notamment où fixer la frontière entre les deux types de conditions.

- **Frontière primaire** Soit on décide de fixer la frontière entre les deux types de conditions aux limites en un sommet du maillage primaire. Ainsi, une arête primaire appartient clairement soit à $\mathcal{F}^{\mathrm{D}}$, soit à $\mathcal{F}^{\mathrm{N}}$. C'est un choix naturel : si l'on travaille sur un domaine de calcul rectangulaire et que l'on souhaite, par exemple, imposer une condition de Neumann (respectivement Dirichlet) sur les bords verticaux (respectivement horizontaux), ce sont alors les sommets formant les coins du maillage qui délimitent la frontière entre les deux types de conditions aux limites. Il s'agit alors de décider quelle condition appliquer en un tel sommet.

- Choix «Dirichlet invasif» : on considère qu'il s'agit d'un nœud de type Dirichlet; dans ce cas, la maille secondaire correspondante ne participe pas au système. Cela pose problème, car cette maille est bordée par une demi-arête sur laquelle une condition de Neumann est imposée, qui n'est alors pas prise en compte (voir la figure 2.7).
- Choix «Neumann invasif» : on considère qu'il s'agit d'un nœud de type Neumann, et la maille secondaire correspondante participe alors au système. Ce n'est pas satisfaisant non plus, cette maille étant à son tour bordée par une demi-arête sur laquelle une condition de Dirichlet est imposée.

En pratique, on observe dans les deux cas des oscillations parasites pour le cas test du five spot, qui sera décrit dans la sous-section 2.4.3.

- **Frontière secondaire** Une meilleure tentative consiste à délimiter les deux types de conditions aux limites au centre d'une arête primaire. Cette fois, le traitement de chaque maille secondaire est bien défini; c'est au centre $\mathbf{x}_M$ de l'arête mixte σ_M dont une extrémité est un sommet Dirichlet (sommet $\mathbf{x}_A$), et l'autre Neumann (sommet $\mathbf{x}_B$), qu'il faut faire un choix (on utilise les notations de la figure 2.7). On propose de prescrire la valeur de la charge ψ en ce nœud par la moyenne arithmétique des valeurs prises aux deux extrémités de l'arête mixte concernée :

$$\psi(x_M) = \frac{\psi(x_A) + \Psi_B}{2}.$$

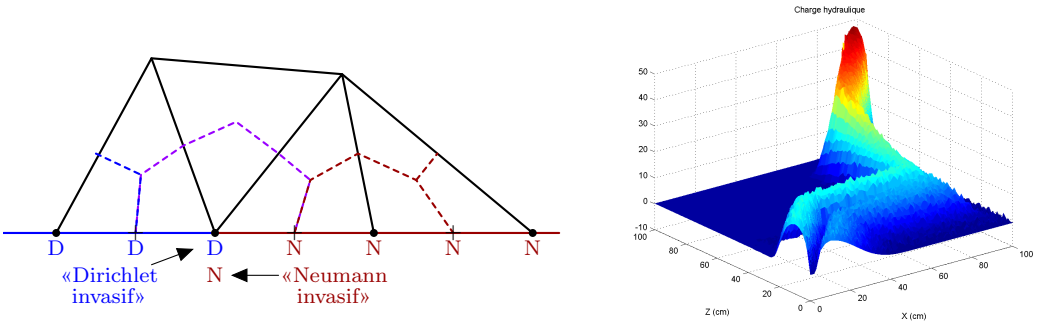
Ce nœud est en un sens «mixte», puisqu'il contient à la fois l'information donnée par la condition de Dirichlet (via la valeur $\psi(x_A)$) , et celle donnée par la condition de Neumann (via l'inconnue Ψ_B). La valeur $\psi(x_M)$ est entièrement déterminée par l'inconnue Ψ_B, elle n'ajoute pas d'inconnue supplémentaire au système discret.

- L'arête mixte σ_M est de type Dirichlet : les flux F_{σ_M} et $F_{\sigma_M^\star}$ sont calculés à l'aide des formules (2.9).
- La maille secondaire A, centrée sur un sommet de type Dirichlet, ne participe pas au système.
- Pour la maille secondaire B, centrée en un sommet de type Neumann, on impose que la contribution de la demi-arête $[\mathbf{x}_M\mathbf{x}_B]$ vale $F_{\sigma_M}/2$. L'équation correspondant à la maille secondaire B est alors :

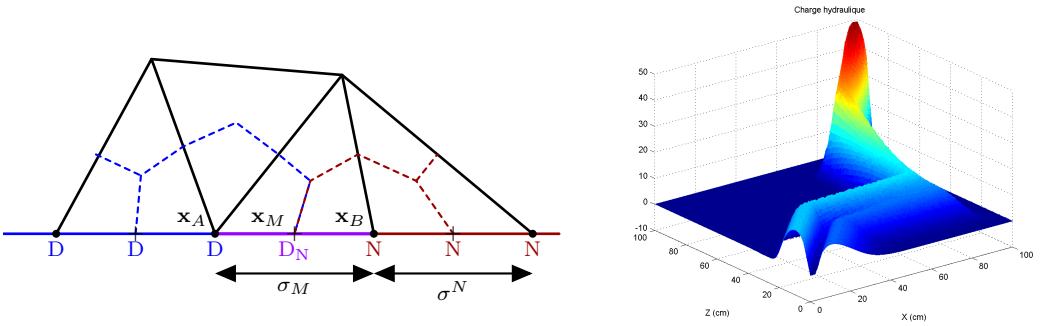
$$\sum_{\sigma\in\mathcal{F}_B} F_{\sigma^\star} + G_{\sigma^N,B} + \frac{F_{\sigma_M}}{2} = |B| f_B,$$

où σ^N est l'unique arête Neumann contribuant à la maille secondaire B.

Cette manière de procéder permet d'éviter les instabilités numériques. Pour terminer, le système complet se met sous la forme $\mathcal{A}(\Psi)\Psi = \mathcal{B}$, où la dépendance en Ψ de la matrice $\mathcal{A}$ provient de celle du tenseur $\mathbb{K}_{\sigma,K}(\Psi)$. À cause des arêtes mixtes, la matrice $\mathcal{A}$ n'est plus symétrique.



(A) Frontière primaire.



(B) Frontière secondaire.

FIGURE 2.7: Deux traitements pour les arêtes mixtes.

2.2 Équation de Richards

On arrive enfin à la résolution numérique complète de l'équation de Richards, assortie de conditions aux limites mixtes Dirichlet/Neumann :

$$\begin{cases} \partial_t \theta(\psi(\mathbf{x},t)) - \nabla \cdot (\mathbb{K}(\psi,\mathbf{x})(\nabla\psi(\mathbf{x},t) + e_z)) = f(\mathbf{x},t) & \text{dans } \Omega\times]0,T], \\ \psi(\mathbf{x},0) = \psi^0(\mathbf{x}) & \text{dans } \Omega \times \{0\}, \\ \psi(s,t) = \psi_{\mathrm{D}}(s,t) & \text{sur } \partial\Omega^{\mathrm{D}}\times]0,T], \\ -\mathbb{K}(\psi,s)(\nabla\psi(s,t) + e_z)\cdot\nu(s) = g(s,t) & \text{sur } \partial\Omega^{\mathrm{N}}\times]0,T]. \end{cases} \tag{2.12}$$

Le paramètre T désigne le temps final de simulation, ψ^0 la donnée initiale et $e_z = (0\ 1)^{\mathrm{t}}$. Les hypothèses de régularité nécessaires à l'existence et à l'unicité d'une solution seront précisées dans le chapitre 3.

2.2.1 Discrétisation en espace

Dans cette sous-section, on cherche une semi-discrétisation en espace de l'équation (2.12). Pour cette raison, on omet la dépendance de la charge hydraulique ψ en temps. Intégrer l'équation (2.12) sur une maille primaire K conduit à :

$$\frac{d}{dt}\int_K \theta(\psi(\mathbf{x}))\ \mathrm{d}\mathbf{x} - \sum_{\sigma\in\mathcal{F}_K}\int_\sigma \mathbb{K}(s)\nabla(\psi(s)+e_z)\cdot n_{\sigma,K}\ \mathrm{d}s = \int_K f(\mathbf{x})\ \mathrm{d}\mathbf{x}.$$

La discrétisation DDFV décrite dans la sous-section 2.1.3 reste valable ici, la seule différence étant l'ajout du terme de gravité $-\nabla\cdot(\mathbb{K}(\psi,\mathbf{x})e_z)$. Le système discret s'écrit :

$$\begin{cases} \forall K\in\mathcal{T},\ |K|\theta(\psi_K) + \sum_{\sigma\in\mathcal{F}_K^{\mathrm{I}}} F_\sigma + \sum_{\sigma\in\mathcal{F}_K^{N}} G_\sigma = |K| f_K, \\ \forall A\in\mathcal{S}^{\mathrm{I}}\cup\mathcal{S}^{\mathrm{N}},\ |A|\theta(\psi_A) + \sum_{\sigma\in\mathcal{F}_A} F_{\sigma^\star} + \sum_{\sigma\in\mathcal{F}_A^{N}} G_{\sigma,A} = |A| f_A. \end{cases}$$

Les flux à travers les arêtes primaires et secondaires, F_σ et $F_{\sigma^\star}$, sont définis par :

$$\begin{aligned} F_\sigma &= -\mathbb{K}_{\sigma,K}(\nabla_{\sigma,K}\Psi + e_z)\cdot N_\sigma \\ &= a_K(\psi_K-\psi_\sigma) + b_K(\psi_A-\psi_B) + \alpha_K \end{aligned}$$

$$\begin{aligned} F_{\sigma^\star} &= -\left(\mathbb{K}_{\sigma,K}(\nabla_{\sigma,K}\Psi + e_z)\cdot N_{\sigma_K^\star} + \mathbb{K}_{\sigma,L}(\nabla_{\sigma,L}\Psi + e_z)\cdot N_{\sigma_L^\star}\right) \\ &= b_K(\psi_K-\psi_\sigma) + b_L(\psi_\sigma-\psi_L) + (c_K+c_L)(\psi_A-\psi_B) + \beta_K + \beta_L. \end{aligned}$$

Les coefficients a_K, b_K et c_K s'écrivent toujours :

$$\begin{pmatrix} a_K & b_K \\ b_K & c_K \end{pmatrix} := \frac{1}{2|D_{\sigma,K}|}\mathbb{K}_{\sigma,K}\begin{pmatrix} N_\sigma\cdot N_\sigma & N_\sigma\cdot N_{\sigma_K^\star} \\ N_{\sigma_K^\star}\cdot N_\sigma & N_{\sigma_K^\star}\cdot N_{\sigma_K^\star} \end{pmatrix},$$

et les coefficients relatifs aux termes de gravité sont définis par :

$$\begin{pmatrix} \alpha_K \\ \beta_K \end{pmatrix} := -\mathbb{K}_{\sigma,K}e_z\cdot\begin{pmatrix} N_\sigma \\ N_{\sigma_K^\star} \end{pmatrix}.$$

De même pour a_L, b_L, c_L, α_L et β_L. On omet également dans la suite leur dépendance en Ψ. Pour un diamant D_σ intérieur, l'équation de continuité des flux normaux s'écrit :

$$a_K(\psi_\sigma - \psi_K) + b_K(\psi_B - \psi_A) + \alpha_K = a_L(\psi_L - \psi_\sigma) + b_L(\psi_B - \psi_A) + \alpha_L,$$

ce qui donne (après linéarisations) :

$$\psi_\sigma = \frac{a_K\psi_K + a_L\psi_L}{a_K + a_L} + \frac{b_K - b_L}{a_K + a_L}(\psi_A - \psi_B) + \frac{\alpha_K - \alpha_L}{a_K + a_L},$$

puis pour les flux :

$$\begin{pmatrix} F_\sigma \\ F_{\sigma^\star} \end{pmatrix} = \mathcal{A}_\sigma(\Psi) \begin{pmatrix} \psi_K - \psi_L \\ \psi_A - \psi_B \end{pmatrix} + \mathcal{G}_\sigma(\Psi),$$

avec

$$\mathcal{A}_\sigma(\Psi) = \begin{pmatrix} \dfrac{a_K a_L}{a_K + a_L} & \dfrac{a_K b_L + a_L b_K}{a_K + a_L} \\ \dfrac{a_K b_L + a_L b_K}{a_K + a_L} & c_K + c_L - \dfrac{(b_K - b_L)^2}{a_K + a_L} \end{pmatrix}$$

et

$$\mathcal{G}_\sigma(\Psi) = \begin{pmatrix} \dfrac{a_K\alpha_L + a_L\alpha_K}{a_K + a_L} \\ \beta_K + \beta_L + \dfrac{(b_L - b_K)(\alpha_K - \alpha_L)}{a_K + a_L} \end{pmatrix}.$$

Dans le cas d'un diamant dégénéré du bord $D_{\sigma,K}$, on a :

$$\begin{pmatrix} F_\sigma \\ F_{\sigma^\star} \end{pmatrix} = \mathcal{A}_\sigma^\partial(\Psi) \begin{pmatrix} \psi_K - \psi_\sigma \\ \psi_A - \psi_B \end{pmatrix} + \mathcal{G}_\sigma^\partial(\Psi),$$

avec

$$\mathcal{A}_\sigma^\partial(\Psi) = \begin{pmatrix} a_K & b_K \\ b_K & c_K \end{pmatrix} \quad \text{et} \quad \mathcal{G}_\sigma^\partial(\Psi) = \begin{pmatrix} \alpha_K \\ \beta_K \end{pmatrix}.$$

Dans le cas d'une arête σ de $\mathcal{F}^{\mathrm{N}}$, la condition de Neumann s'écrit :

$$F_\sigma = a_K(\psi_K - \psi_\sigma) + b_K(\psi_A - \psi_B) + \alpha_K = G_\sigma,$$

ce qui donne en réinjectant :

$$F_{\sigma^\star} = \left(c_K - \frac{b_K^2}{a_K}\right)(\psi_A - \psi_B) + \frac{b_K}{a_K}(G_\sigma - \alpha_K).$$

Le système semi-discret se met finalement sous la forme :

$$M\frac{d}{dt}\Theta(\Psi) + \mathcal{A}(\Psi)\Psi + \mathcal{G}(\Psi) = \mathcal{B}, \tag{2.13}$$

où la matrice diagonale de masse M contient les aires des mailles primaires et secondaires, $\mathcal{G}$ contient les termes gravitaires et $\mathcal{B}$ rassemble les contributions du terme source f et des conditions de bord.

2.2.2 Discrétisation en temps

On se donne une suite d'instants discrets $t^0 = 0 < t^1 < \cdots < t^n < \cdots < T$, d'espacement constant δt tel que $N_T := T/\delta t$ soit entier, et on cherche à approcher la dérivée d'une fonction du temps y à une itération $n \geq 2$ donnée, $y'(t^n)$. Pour tout $0 \leq k \leq n$, on note $y^k := y(t^k)$. Une manière de procéder est d'utiliser des polynômes d'interpolation, suivant la technique classique pour établir les formules BDF [25] :

- Une première approximation est $y'(t^n) \simeq P_1'(t^n)$, où P_1 est la fonction affine qui interpole y en t^{n-1} et en t^n. Or on a pour tout t :

  $$P_1(t) = y^n + \frac{y^n - y^{n-1}}{\delta t}(t - t^n), \quad \text{soit} \quad P_1'(t) = \frac{y^n - y^{n-1}}{\delta t}.$$

  D'où $P_1'(t^n) = (y^n - y^{n-1})/\delta t$, et on retrouve la dérivée discrète apparaissant dans la méthode d'Euler; la méthode d'Euler implicite est également nommée BDF1.

- L'approximation $y'(t^n) \simeq P_2'(t^n)$, où P_2 est le trinôme qui interpole y en t^{n-2}, t^{n-1} et en t^n, conduit au schéma BDF2. Le polynôme P_2 s'écrit :

  $$P_2(t) = y^n + \frac{y^n - y^{n-1}}{\delta t}(t - t^n) + \frac{y^n - 2y^{n-1} + y^{n-2}}{2\delta t^2}(t - t^n)(t - t^{n-1}),$$

  ce qui donne

  $$P_2'(t) = \frac{y^n - y^{n-1}}{\delta t} + \frac{y^n - 2y^{n-1} + y^{n-2}}{2\delta t^2}(2t - t^n - t^{n-1}),$$

  soit

  $$P_2'(t^n) = \frac{1}{\delta t}\left(\frac{3}{2}y^n - 2y^{n-1} + \frac{1}{2}y^{n-2}\right).$$

Après substitution dans le système (2.13), on obtient le système discret suivant :

$$\forall n \geq 2, \quad \frac{3}{2\delta t}M\Theta(\Psi^n) + \mathcal{A}(\Psi^n)\Psi^n + \mathcal{G}(\Psi^n) = \mathcal{B} + \frac{2}{\delta t}M\Theta(\Psi^{n-1}) - \frac{1}{2\delta t}M\Theta(\Psi^{n-2}). \tag{2.14}$$

Pour l'initialisation $n = 1$, on peut utiliser un schéma de Crank-Nicolson, par exemple. Il est important de choisir un schéma d'ordre 2 pour l'initialisation. Sinon, on perd le bénéfice de la précision du schéma BDF2 en commettant à la première itération une erreur en temps du même ordre que l'erreur cumulée aux itérations suivantes.

2.2.3 Linéarisation

Le système (2.14) est maintenant complètement discret, mais toujours non linéaire, à la fois en temps (loi de saturation θ) et en espace (tenseur de conductivité $\mathbb{K}$). Nous mettons donc en place une procédure itérative pour le résoudre numériquement. Nous suivons ici [59]. On fixe $n \geq 2$ et on cherche une suite $(\Psi^{n,m})_{m\geq 0}$ vérifiant :

$$\begin{cases} \Psi^{n,0} = \Psi^{n-1}, \\ \forall m \geq 1, \quad \dfrac{3}{2\delta t} M\overline{\Theta(\Psi^{n,m})} + \overline{\mathcal{A}(\Psi^{n,m})\Psi^{n,m}} + \overline{\mathcal{G}(\Psi^{n,m})} \\ \qquad\qquad = \mathcal{B} + \dfrac{2}{\delta t} M\Theta(\Psi^{n-1}) - \dfrac{1}{2\delta t} M\Theta(\Psi^{n-2}), \end{cases} \tag{2.15}$$

où les quantités $\overline{\Theta(\Psi^{n,m})}$, $\overline{\mathcal{A}(\Psi^{n,m})\Psi^{n,m}}$ et $\overline{\mathcal{G}(\Psi^{n,m})}$, définies ci-après, approchent respectivement $\Theta(\Psi^{n,m})$, $\mathcal{A}(\Psi^{n,m})\Psi^{n,m}$ et $\mathcal{G}(\Psi^{n,m})$. Une initialisation plus élaborée que $\Psi^{n,0} = \Psi^{n-1}$ sera discutée dans la sous-section 2.3.3. Pour $m \geq 1$, on utilise une linéarisation d'ordre 1 en temps (de type Newton), et une linéarisation d'ordre 0 en espace (de type Picard) :

$$\begin{cases} \overline{\Theta(\Psi^{n,m})} := \Theta\left(\Psi^{n,m-1}\right) + \Theta'(\Psi^{n,m-1}) \cdot \underbrace{(\Psi^{n,m} - \Psi^{n,m-1})}_{:=\delta\Psi^{n,m}}, \\ \overline{\mathcal{A}(\Psi^{n,m})\Psi^{n,m}} := \mathcal{A}(\Psi^{n,m-1})\Psi^{n,m}, \quad \overline{\mathcal{G}(\Psi^{n,m})} := \mathcal{G}(\Psi^{n,m-1}). \end{cases} \tag{2.16}$$

Localement, cela correspond, pour une maille primaire K et une maille secondaire A données, à effectuer ces linéarisations sur la loi de saturation et le tenseur :

$$\begin{cases} \overline{\theta(\psi_C^{n,m})} = \theta(\psi_C^{n,m-1}) + \theta'(\psi_C^{n,m-1})(\psi_C^{n,m} - \psi_C^{n,m-1}), \quad C \in \{K, A\}, \\ \overline{\mathbb{K}_{\sigma,K}(\Psi^{n,m})} = \mathbb{K}_{\sigma,K}(\Psi^{n,m-1}). \end{cases}$$

Ainsi, pour $n \geq 2$, $m \geq 1$, le système linéaire d'inconnue $\delta\Psi^{n,m}$ est le suivant :

$$\left(\frac{3}{2}M\Theta'(\Psi^{n,m-1}) + \delta t\mathcal{A}(\Psi^{n,m-1})\right)\delta\Psi^{n,m}$$
$$= -M\left(\frac{3}{2}\Theta(\Psi^{n,m-1}) - 2\Theta(\Psi^{n-1}) + \frac{1}{2}\Theta(\Psi^{n-2})\right)$$
$$- \delta t\left(\mathcal{A}(\Psi^{n,m-1})\Psi^{n,m-1} + \mathcal{G}(\Psi^{n,m-1}) - \mathcal{B}\right). \quad (2.17)$$

Pour $n = 1, m \geq 1$, il s'agit de résoudre (en utilisant le schéma de Crank-Nicolson) :

$$\left(M\partial_\psi\Theta(\Psi^{1,m-1}) + \frac{\delta t}{2}\mathcal{A}(\Psi^{1,m-1})\right)\delta\Psi^{1,m} = -M\left(\Theta(\Psi^{1,m-1}) - \Theta(\Psi^0)\right)$$
$$- \frac{\delta t}{2}\left(\mathcal{A}(\Psi^{1,m-1})\Psi^{1,m-1} + \mathcal{G}(\Psi^{1,m-1}) + \mathcal{A}(\Psi^0)\Psi^0 + \mathcal{G}(\Psi^0) - 2\mathcal{B}\right). \quad (2.18)$$

Sous cette forme, la suite des itérés $(\Psi^{n,m})_{m\geq 0}$ est définie pour $m \geq 1$ par la relation de récurrence $\Psi^{n,m} = \Psi^{n,m-1} + \delta\Psi^{n,m}$.

Critère d'arrêt des itérations non linéaires Si la procédure converge, on peut trouver un itéré m_∞ tel que Ψ^{n,m_∞} vérifie le critère $\|\delta\Psi^{n,m_\infty}\|\,\|\Psi^{n-1}\|^{-1} \leq \varepsilon$, où ε est une tolérance choisie par l'utilisateur, et $\|\cdot\|$ désigne la norme euclidienne sur $\mathbb{R}^{|\mathcal{T}|+|\mathcal{S}^{\mathrm{I}}\cup\mathcal{S}^{\mathrm{N}}|}$. On arrête alors la procédure et on fixe $\Psi^n := \Psi^{n,m_\infty}$.

2.2.4 Évaluation du tenseur

On se place en un itéré en temps n, et en un itéré de linéarisation $m \geq 1$. On connaît donc les vecteurs $\Psi^{n,i}$, pour $1 \leq i \leq m-1$, et on cherche $\Psi^{n,m}$ solution du système (2.17)-(2.18). Pour σ dans $\mathcal{F}$, il reste encore à trouver une approximation $\mathbb{K}_{\sigma,K}(\Psi^{n,m-1})$. On se place sur la maille primaire K associée à l'arête σ, le côté L se traite de manière identique. On cherche cette approximation sous la forme $\mathbb{K}_{\sigma,K}(\Psi^{n,m-1}) = \mathbb{K}(\psi^{n,m-1}_{\star,K}, \mathbf{x}_{\star,K})$, où $\psi^{n,m-1}_{\star,K}$ et $\mathbf{x}_{\star,K}$ sont à définir. On propose plusieurs solutions :

- Soit on définit $\psi^{n,m-1}_{\star,K}$ et $\mathbf{x}_{\star,K}$ directement. Parmi les choix possibles :

  Choix 1
  $$\begin{cases} \psi^{n,m-1}_{\star,K} = \psi^{n,m-1}_K, \\ \mathbf{x}_{\star,K} = \mathbf{x}_K. \end{cases}$$

Choix 2
$$\begin{cases} \psi_{\star,K}^{n,m-1} = \dfrac{1}{2}\left(\psi_A^{n,m-1} + \psi_B^{n,m-1}\right), \\ \mathbf{x}_{\star,K} = \dfrac{1}{2}\left(\mathbf{x}_A + \mathbf{x}_B\right). \end{cases}$$

Choix 3
$$\begin{cases} \psi_{\star,K}^{n,m-1} = \dfrac{1}{3}\left(\psi_K^{n,m-1} + \psi_A^{n,m-1} + \psi_B^{n,m-1}\right), \\ \mathbf{x}_{\star,K} = \dfrac{1}{3}\left(\mathbf{x}_K + \mathbf{x}_A + \mathbf{x}_B\right). \end{cases}$$

Ces choix ont le mérite d'être des prescriptions directes, donc peu coûteuses.

- Soit on prescrit seulement $\mathbf{x}_{\star,K}$, par exemple le barycentre du demi-diamant $D_{\sigma,K}$: $\mathbf{x}_{\star,K} = 1/3\,(\mathbf{x}_K + \mathbf{x}_A + \mathbf{x}_B)$. Une façon plus élaborée de procéder pour le choix de la variable $\psi_{\star,K}^{n,m-1}$ est alors de demander à ce qu'elle soit solution de l'équation de continuité des flux en $(n, m-1)$ suivante, d'inconnue ξ :

$$\begin{aligned} &-\mathbb{K}(\xi, \mathbf{x}_{\star,K})\left[\frac{1}{2|D_{\sigma,K}|}\left[(\xi - \psi_K^{n,m-1})N_\sigma + (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_K^\star}\right] + e_z\right]\cdot N_\sigma \\ &-\mathbb{K}(\xi, \mathbf{x}_{\star,L})\left[\frac{1}{2|D_{\sigma,L}|}\left[(\psi_L^{n,m-1} - \xi)N_\sigma + (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_L^\star}\right] + e_z\right]\cdot N_\sigma = 0. \end{aligned} \tag{2.19}$$

La variable $\psi_{\star,K}^{n,m-1}$, en intervenant dans la définition des gradients discrets, joue alors le rôle de $\psi_\sigma^{n,m-1}$ dans l'équation (2.19), ce qui est logique puisque $\psi_\sigma^{n,m-1}$ est justement le degré de liberté servant à exprimer la continuité des flux normaux en $(n, m-1)$. L'équation (2.19) est non linéaire en ξ, et on a de nouveau recours à une procédure itérative pour la résoudre. Spécifiquement, on construit une suite d'itérés $(\xi^j)_{j\geq 1}$ définie par :

$$\begin{cases} \xi^0 = \psi_\sigma^{n,m-1} \quad \text{(par exemple)}, \\ \forall j \geq 1, \quad -\mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,K})\Lambda_{\sigma,K}^{n,m-1,j}\cdot N_\sigma - \mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,L})\Lambda_{\sigma,L}^{n,m-1,j}\cdot N_\sigma = 0, \end{cases}$$

avec

$$\begin{cases} \Lambda_{\sigma,K}^{n,m-1,j} = \dfrac{1}{2|D_{\sigma,K}|}\left[(\xi^j - \psi_K^{n,m-1})N_\sigma + (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_K^\star}\right] + e_z, \\ \Lambda_{\sigma,L}^{n,m-1,j} = \dfrac{1}{2|D_{\sigma,L}|}\left[(\psi_L^{n,m-1} - \xi^j)N_\sigma + (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_L^\star}\right] + e_z. \end{cases}$$

On a effectué une linéarisation de type Picard sur le terme $\mathbb{K}(\xi^j, \mathbf{x}_{\star,K})$, identique à celle de la sous-section 2.2.3. En notant $\delta\xi^j = \xi^j - \xi^{j-1}$, on trouve après calculs :

$$\delta\xi^j = \frac{1}{\left(\left[\frac{1}{2|D_{\sigma,K}|}\mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,K}) + \frac{1}{2|D_{\sigma,L}|}\mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,L})\right] N_\sigma\right) \cdot N_\sigma} \times$$

$$\left(\mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,K})\overline{\Lambda}_{\sigma,K}^{n,m-1,j} + \mathbb{K}(\xi^{j-1}, \mathbf{x}_{\star,L})\overline{\Lambda}_{\sigma,L}^{n,m-1,j}\right) \cdot N_\sigma.$$

avec

$$\begin{cases} \overline{\Lambda}_{\sigma,K}^{n,m-1,j} = \dfrac{1}{2|D_{\sigma,K}|}\left[(\psi_K^{n,m-1} - \xi^{j-1})N_\sigma - (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_K^\star}\right] - e_z, \\ \overline{\Lambda}_{\sigma,L}^{n,m-1,j} = \dfrac{1}{2|D_{\sigma,L}|}\left[(\psi_L^{n,m-1} - \xi^{j-1})N_\sigma + (\psi_B^{n,m-1} - \psi_A^{n,m-1})N_{\sigma_L^\star}\right] + e_z. \end{cases}$$

Le choix retenu pour les tests réalisés dans la section 2.4 est le choix 3 par prescription directe. C'est en effet la version qui allie faible coût de calcul, efficacité en terme de nombre d'itérations de linéarisation du système discret et enfin robustesse vis-à-vis des problèmes raides.

2.3 Résolution numérique

Dans cette section, on rentre un peu plus dans le détail de l'implémentation pratique. Après avoir fait quelques remarques sur l'assemblage en lui-même, on précisera comment est résolu le système linéaire (2.17)-(2.18). On parlera ensuite de l'initialisation de la procédure de linéarisation (2.15).

2.3.1 Assemblage du système linéaire

- Le choix de l'orientation des points $\mathbf{x}_K$, $\mathbf{x}_L$, $\mathbf{x}_A$ et $\mathbf{x}_B$ pour une arête σ d'extrémités $\mathbf{x}_A$, $\mathbf{x}_B$ et située entre deux mailles primaires K et L est arbitraire et doit être fait avant l'assemblage. Dans la section 2.1, l'orientation a été choisie de manière à vérifier $\det(\mathbf{x}_L - \mathbf{x}_K, \mathbf{x}_B - \mathbf{x}_A) > 0$.

- Chaque diamant D_σ est associé à une et une seule arête σ du maillage primaire. De plus, l'ensemble des diamants forme une partition du domaine de calcul Ω. Ceci, indépendamment de la nature des polygones utilisés pour construire le maillage primaire. En conséquence, on réalise l'assemblage par arête primaire; c'est une

implémentation très flexible puisqu'elle ne dépend pas de la structure du maillage primaire.

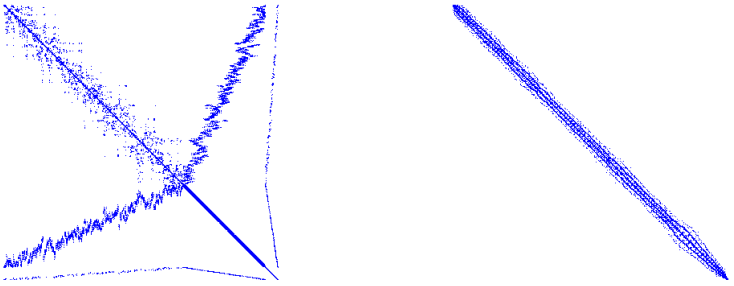
- Suivant la remarque précédente, la seule connectivité dont on a besoin est relative aux demi-diamants. En pratique, on définit trois tableaux : le premier contient les coordonnées de chaque sommet du maillage primaire, le deuxième celles des centres de masse, et le troisième rassemble les numéros des extrémités de chaque arête σ, ainsi que les numéros des mailles qui se partagent cette arête (une seule maille s'il s'agit d'une arête du bord) :

$$\begin{bmatrix} A_1 & \cdots & A_1^\partial & \cdots \\ B_1 & \cdots & B_1^\partial & \cdots \\ K_1 & \cdots & K_1^\partial & \cdots \\ L_1 & \cdots & 0 & \cdots \end{bmatrix}$$

  Pour une colonne de ce tableau, relative à une arête σ, on calcule les flux F_σ et $F_{\sigma^\star}$ associés et leurs contributions aux mailles primaires et secondaires concernées.

2.3.2 Résolution du système linéaire

On choisit de résoudre le système linéaire (2.17)-(2.18) par une méthode directe, celui-ci ne comportant que quelques milliers d'inconnues dans notre étude. Pour des sytèmes posés sur des maillages plus fins, on aurait recours à une méthode itérative, comme le gradient conjugué ou l'algorithme GMRES selon la symétrie de la matrice. Lorsque la matrice $\mathcal{A}$ est symétrique, ce qui est le cas seulement quand les conditions de bord ne sont pas mixtes comme cela a été noté dans la sous-section 2.1.3, on utilise la factorisation de Cholesky. Sinon, on utilise une factorisation LU (librairie UMFPACK).
La matrice du système (2.17)-(2.18) garde la même structure tout au long de la simulation, et elle doit être assemblée un grand nombre de fois, notamment à cause de la procédure de linéarisation. C'est pourquoi on utilise une méthode de renumérotation, qui est appliquée une seule fois sur les centres et les sommets des mailles primaires (sauf les sommets du bord de type Dirichlet, dont les mailles secondaires associées ne participent pas au système), en prétraitement. On choisit ici la permutation inversée de Cuthill et McKee [26]. On perd la structure par bloc de la matrice initiale, mais les coefficients non nuls de la nouvelle matrice sont resserrés autour de la diagonale (voir la figure 2.8), conduisant à une largeur de bande minimale et à un gain en temps CPU significatif.



(A) Avant renumérotation. (B) Après renumérotation.

FIGURE 2.8: Structure de la matrice du système linéaire.

2.3.3 Initialisation de la procédure de linéarisation

Dans la sous-section 2.2.3, on a initialisé à chaque itéré en temps $n \geq 1$ la procédure de linéarisation (équation (2.15)) par la solution obtenue au temps discret précédent : $\Psi^{n,0} = \Psi^{n-1}$. Ce n'est en aucun cas une nécessité, et des initialisations de plus grand stencil peuvent être considérées [74]. Une façon naturelle de prédire la solution à un itéré donné n est d'utiliser les polynômes d'interpolation de Lagrange :

- **Ordre 0** Il s'agit de l'initialisation la plus simple : $\Psi^{n,0} = \Psi^{n-1}$.
- **Ordre 1** On choisit $\Psi^{n,0} = P_1(t^n)$, où P_1 est la fonction affine qui interpole Ψ en t^{n-2} et en t^{n-1}, rencontrée dans la sous-section 2.2.2. On trouve $\Psi^{n,0} = 2\Psi^{n-1} - \Psi^{n-2}$.
- **Ordre 2** On choisit $\Psi^{n,0} = P_2(t^n)$, ce qui donne : $\Psi^{n,0} = 3\Psi^{n-1} - 3\Psi^{n-2} + \Psi^{n-3}$.
- Des initialisations d'ordre quelconque $d \geq 0$ peuvent ainsi être implémentées, en utilisant le polynôme d'interpolation de Lagrange qui interpole les vecteurs $\Psi^{n-1}, \cdots, \Psi^{n-d-1}$ en $t^{n-1}, \cdots, t^{n-d-1}$. Les coefficients du polynôme servant à l'initialisation sont ceux du triangle de Pascal avec signes alternés : pour une initialisation d'ordre d, on a : $\Psi^{n,0} = \sum_{i=1}^{d+1} (-1)^{i+1} \binom{d+1}{i} \Psi^{n-i}$.

Pour les tous premiers temps de simulation, on peut bien sûr être limité par le nombre d'itérés précédents disponibles, et il faut alors restreindre l'ordre d'initialisation en

conséquence. Ainsi, une initialisation d'ordre 2 s'écrit :

$$\begin{cases} \Psi^{1,0} = \Psi^0, \\ \Psi^{2,0} = 2\Psi^1 - \Psi^0, \\ \forall n \geq 3,\ \Psi^{n,0} = 3\Psi^{n-1} - 3\Psi^{n-2} + \Psi^{n-3}. \end{cases}$$

On expose dans le Tableau 2.1 le temps CPU obtenu pour différents ordres d'initialisation et différents maillages (définis dans la sous-section 2.4.1), pour le cas test analytique présenté dans la sous-section 2.4.1. L'initialisation d'ordre 0 est prise comme référence, avec un temps CPU normalisé à 100 secondes. Les initialisations d'ordre 1 et 2 sem-

Maillage	0	1	2	3	4
M_3	100	67	77	67	80
M_4	100	69	73	74	89
M_5	100	66	67	71	88
M_6	100	62	62	68	85

TABLEAU 2.1: Temps CPU pour différents ordres d'initialisation.

blent donner des performances équivalentes sur maillage fin, tandis que des initialisations d'ordre plus élevé sont moins performantes. L'initialisation d'ordre 2 semble naturelle, puisqu'elle provient du trinôme d'interpolation à l'origine de la formule BDF2. Cependant, il est important de noter qu'en cas de variations trop rapides de la solution, une initialisation d'ordre $d > 0$ peut être très éloignée de la solution cherchée et faire diverger l'algorithme de linéarisation. Par exemple, pour le cas test de Polmann étudié dans la sous-section 2.4.2, une initialisation d'ordre 1 à l'itéré $n = 2$ et en un sommet $\mathbf{x}_A$ situé sur le bord haut $\mathbb{H}$ de la colonne donne : $\psi_A^{2,0} = 2\psi_A^1 - \psi_A^0 = 2 \times (-75) - (-1000) = 850 \gg 0$. Dans la section 2.4, on se restreindra à un ordre d'initialisation de 0 pour éviter ce type de problèmes.

2.4 Cas tests

Dans cette section, on présente les résultats obtenus après discrétisation en espace par la méthode DDFV décrite dans la section 2.1, et par la formule BDF2 de la sous-section 2.2.2 pour la discrétisation en temps. On propose trois cas tests de difficulté croissante; cette difficulté est dictée par la conductivité $\mathbb{K}$ qui prend la forme :

$$\mathbb{K}(\psi, \mathbf{x}) = \frac{\rho g}{\nu} k(\psi) \overline{\mathbb{K}}(\mathbf{x}),$$

où on rappelle que k désigne la perméabilité relative et $\overline{\mathbb{K}}$ la perméabilité intrinsèque du sol; ρ, g et ν représentent respectivement la densité du fluide, la constante de gravité et la viscosité dynamique. On commence par étudier deux cas tests d'infiltration verticale dans un milieu homogène et isotrope (*i.e.* $\overline{\mathbb{K}} = \mathbb{I}$). Le premier possède une solution analytique connue, ce qui nous permet de vérifier l'ordre de convergence théorique de la méthode, mais aussi d'étudier plus en détail la matrice du système. Le deuxième décrit la propagation d'un front nettement plus raide. Enfin, nous testons le problème classique quarter five spot en milieu hétérogène et anisotrope (*i.e.* $\overline{\mathbb{K}} \neq \mathbb{I}$ et dépend de la variable d'espace).

2.4.1 Infiltration en milieu homogène isotrope : validation (TC1)

Le domaine étudié est $\Omega = [0, 4] \times [0, 20]$ (en centimètres) et le temps final de simulation est $T = 2\ min$. La solution analytique (voir la figure 2.9) est donnée par :

$$\psi(z,t) = 20.4 \tanh\left[0.5\left(z + \frac{t}{12} - 15\right)\right] - 41.1.$$

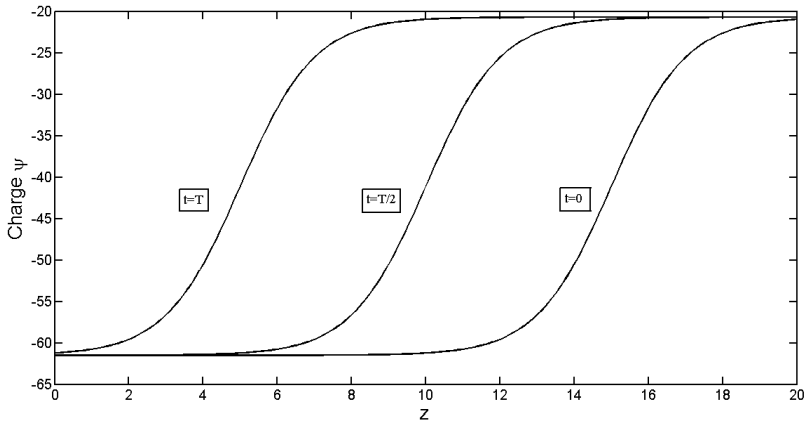


FIGURE 2.9: Profil de la solution exacte à différents instants.

On utilise cette formule pour calculer la condition initiale, ainsi que les termes de bord (condition de Dirichlet) sur $\partial\Omega$ et le terme source adéquats pour que l'équation (2.12) soit vérifiée. La loi de saturation et la perméabilité relative sont données par les relations d'Haverkamp [50] :

$$\theta(\psi) = \frac{\theta_s - \theta_r}{1 + |\tilde{\alpha}\psi|^\beta} + \theta_r \quad \text{et} \quad k(\psi) = \frac{k_s}{1 + |\tilde{A}\psi|^\gamma},$$

avec les paramètres :

$$\theta_s = 0.287, \quad \theta_r = 0.075, \quad \tilde{\alpha} = 0.0271\ cm^{-1}, \quad \beta = 3.96,$$
$$k_{\mathrm{s}} = 9.44 \cdot 10^{-3}\ cm.s^{-1}, \quad \tilde{A} = 0.0524\ cm^{-1}, \quad \gamma = 4.74.$$

Pour un maillage donné, le nombre d'inconnues est $N_{\mathrm{u}} = N_{\mathrm{t}} + N_{\mathrm{n}} - N_{\mathrm{n}}^D$, où N_{t} est le nombre de triangles, N_{n} le nombre total de sommets et N_{n}^D est le nombre de sommets de type Dirichlet (c'est-à-dire tous les sommets de $\partial\Omega$ ici). Pour comparaison, à maillage triangulaire fin fixé (*i.e.* $N_{\mathrm{n}} \gg N_{\mathrm{n}}^D$), le nombre d'inconnues pour DDFV est plus élevé de moitié que le nombre d'inconnues utilisé pour le schéma VF4. En effet, on a $N_{\mathrm{t}} + N_{\mathrm{n}} - N_{\mathrm{n}}^D \simeq N_{\mathrm{t}} + N_{\mathrm{n}} \simeq 3/2 N_{\mathrm{t}}$, car $N_{\mathrm{t}} \simeq 2N_{\mathrm{n}}$ d'après les relations d'Euler. On a construit une famille de maillages non structurés et conformes, composés de triangles, $(\mathrm{M}_i)_{1\leq i\leq 6}$. On rassemble dans le Tableau 2.2 les valeurs prises par $N_{\mathrm{n}}, N_{\mathrm{t}}$ et N_{u} pour chaque maillage. Sont également indiqués le nombre total de coefficients non nuls nnz, le nombre moyen de coefficients non nuls par ligne $\overline{nnz}$ ($= nnz/N_{\mathrm{u}}$) et la largeur de bande bw de la matrice du système (après renumérotation). Pour des maillages fins (typiquement M_5 ou M_6), on a :

$$nnz \simeq 7N_{\mathrm{t}} + 13N_{\mathrm{n}}, \quad \text{d'où} \quad \overline{nnz} \simeq \frac{7N_{\mathrm{t}} + 13N_{\mathrm{n}}}{N_{\mathrm{t}} + N_{\mathrm{n}} - N_{\mathrm{n}}^D} \simeq 9.$$

Maillage	N_{n}	N_{t}	N_{u}	nnz	$\overline{nnz}$	bw
M_1	28	34	42	234	5.6	6
M_2	79	118	159	1127	7.1	16
M_3	253	430	609	4897	8.0	34
M_4	917	1688	2461	21009	8.5	58
M_5	3380	6474	9570	83742	8.8	145
M_6	13233	25896	39031	348595	8.9	267

TABLEAU 2.2: TC1 - Propriétés de la matrice du système.

En effet, si l'on considère un maillage primaire conforme constitué exclusivement de triangles équilatéraux, le stencil du schéma DDFV pour l'équation associée à une maille primaire (respectivement secondaire) intérieure est de 7 (respectivement 13). Ce stencil est représenté sur la figure 2.10. En écho à la figure 2.8, la colonne indiquant la largeur de bande bw témoigne de la creusité de la matrice du système et de l'efficacité de la permutation de Cuthill et McKee. Afin d'observer l'ordre de convergence du schéma, on définit les fonctions suivantes :

(A) Maille primaire. (B) Maille secondaire.

FIGURE 2.10: Stencil du schéma DDFV.

- La fonction ψ^n est la solution exacte évaluée à l'instant t^n :

$$\forall\ x \in \Omega,\ \psi^n(x) := \psi(x, t^n).$$

- La fonction ψ_h^n est reconstruite à partir du vecteur discret produit par le schéma à l'instant t^n, Ψ^n. Elle est choisie affine sur chaque demi-diamant et interpolant les valeurs prises par Ψ^n aux sommets de ce demi-diamant :

$$\forall\ D_{\sigma,K} \in \mathcal{D}, \quad \psi_h^n|_{D_{\sigma,K}} = \sum_{P \in \{K,A,B\}} \psi_P^n \lambda_{\sigma,K,\mathbf{x}_P},$$

  où $\{\lambda_{\sigma,K,\mathbf{x}_P}\}$ est la fonction nodale sur $D_{\sigma,K} = \mathbf{x}_A\mathbf{x}_B\mathbf{x}_K$ relative au sommet $\mathbf{x}_P$.

- La vitesse exacte v est définie par :

$$\forall\ (x,t) \in \Omega \times [0,T],\ v(x,t) = -\mathbb{K}(\psi(x,t))(\nabla\psi(x,t) + e_z).$$

- Enfin, la vitesse v_h est construite constante par morceaux en temps sur chaque intervalle $[t^{n-1}, t^n]$, et constante par morceaux en espace sur chaque demi-diamant $D_{\sigma,K}$ de $\mathcal{D}$, où elle est définie par le flux numérique :

$$\forall\ D_{\sigma,K} \in \mathcal{D},\ \forall\ x \in D_{\sigma,K}, v_h(\cdot, t^n) = -\mathbb{K}_{\sigma,K}(\Psi^n)(\nabla_{\sigma,K}\Psi^n + e_z).$$

Maillage	δt	$e_{\psi,\Omega}$		$e_{v,\Omega}$	
		erreur	ordre	erreur	ordre
M_1	8	1.00e-2		1.15e-1	
M_2	4	3.34e-3	2.46	4.18e-2	2.26
M_3	2	1.00e-3	1.90	1.91e-2	1.23
M_4	1	2.51e-4	2.23	9.41e-3	1.14
M_5	1/2	6.29e-5	2.03	4.59e-3	1.05
M_6	1/4	1.58e-5	1.99	2.28e-3	1.01

TABLEAU 2.3: TC1 - Résultats de convergence.

Le Tableau 2.3 présente alors les erreurs sur la charge hydraulique $e_{\psi,\Omega}$ et sur la vitesse $e_{v,\Omega}$ pour chaque maillage, qui sont définies par :

$$e_{\psi,\Omega} = \frac{\max\limits_{1\leq n\leq N_T} ||\psi^n - \psi_h^n||_{L^2(\Omega)}}{\max\limits_{1\leq n\leq N_T} ||\psi^n||_{L^2(\Omega)}} \quad \text{et} \quad e_{v,\Omega} = \frac{||v - v_h||_{L^2(\Omega\times[0,T])}}{||v||_{L^2(\Omega\times[0,T])}}.$$

Les résultats sont conformes aux attentes, puisqu'on observe une convergence du second ordre sur la charge hydraulique, et du premier ordre sur la vitesse. Notons qu'une comparaison avec la méthode DG à pénalisation intérieure symétrique pondérée (SWIP) a été réalisée dans [11].

2.4.2 Infiltration en milieu homogène isotrope : cas raide (TC2)

Ce cas test d'infiltration fait suite aux travaux de Polmann, et a été proposé dans [20]. Le domaine est $\Omega = [0, 20] \times [0, 100]$ (en centimètres) et le temps final $T = 48\ h$. La condition initiale est choisie constante, et on impose une condition de Dirichlet sur les bords haut $\mathbb{H}$ et bas $\mathbb{B}$ de la colonne. On complète par une condition de Neumann homogène sur les bords latéraux $\mathbb{L}$ (voir la figure 2.11) :

$$\begin{cases} \psi^0(\mathbf{x}) = -10\ m & \text{dans } \Omega, \\ \psi(s,t) = -10\ m & \text{sur } \mathbb{B}\times]0,T], \\ \psi(s,t) = -75\ cm & \text{sur } \mathbb{H}\times]0,T], \\ \mathbb{K}(\psi,s)(\nabla\psi(s,t) + e_z)\cdot n(s) = 0 & \text{sur } \mathbb{L}\times]0,T]. \end{cases}$$

La teneur en eau θ et la perméabilité relative k sont données par les relations de Van Genuchten [76], et tracées sur la figure 2.12 :

$$\theta(\psi) = \frac{(\theta_s - \theta_r)}{(1 + (\xi|\psi|)^\beta)^\gamma} + \theta_r \quad \text{et} \quad k(\psi) = k_s \frac{\left(1 - (\xi|\psi|)^{\beta-1}(1 + (\xi|\psi|)^\beta)^{-\gamma}\right)^2}{(1 + (\xi|\psi|)^\beta)^{\frac{\gamma}{2}}},$$

avec pour paramètres

$$\theta_s = 0.368, \quad \theta_r = 0.102, \quad k_s = 9.22 \cdot 10^{-3} cm.s^{-1},$$
$$\xi = 0.0335\ cm^{-1}, \quad \beta = 2, \quad \gamma = 1 - \frac{1}{\beta}.$$

La raideur de ce test tient dans la forte surpression (égale à $9.25m$) imposée en haut de la colonne, ce qui induit de fortes variations de la perméabilité relative : on a en effet $k(-75cm)/k(-10m) = 8.92 \cdot 10^4$. Comme les conditions initiale et de bord ne dépendent

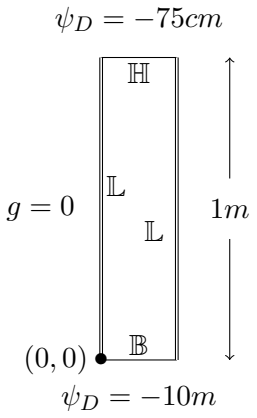


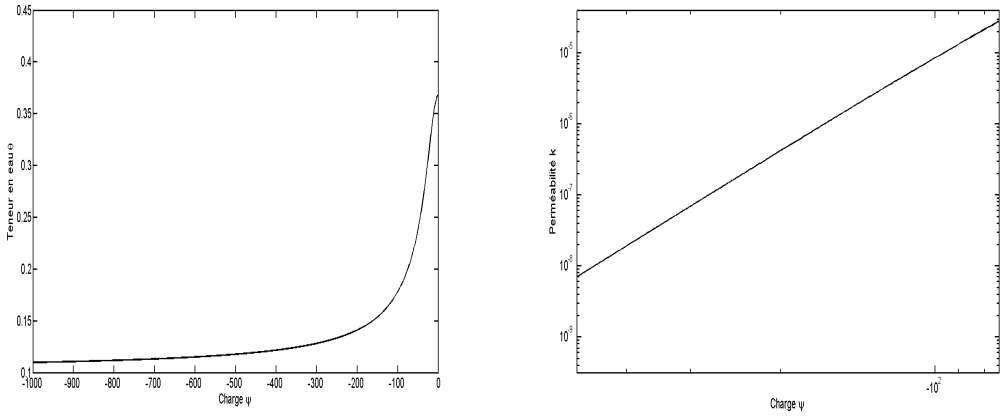
FIGURE 2.11: TC2 - Cas test de Polmann.

pas de la coordonnée x, laquelle n'apparaît par ailleurs pas explicitement dans l'équation de Richards (2.12), la solution exacte est une fonction du temps et de la seule coordonnée verticale, z. A l'issue de la simulation, on s'intéresse donc à la valeur moyenne selon l'axe horizontal de la charge reconstruite, $\overline{\psi_h}(z)$:

$$\forall z \in [0, 100], \quad \overline{\psi_h}(z) = \frac{1}{20} \int_0^{20} \psi_h(x, z)\ \mathrm{d}x.$$

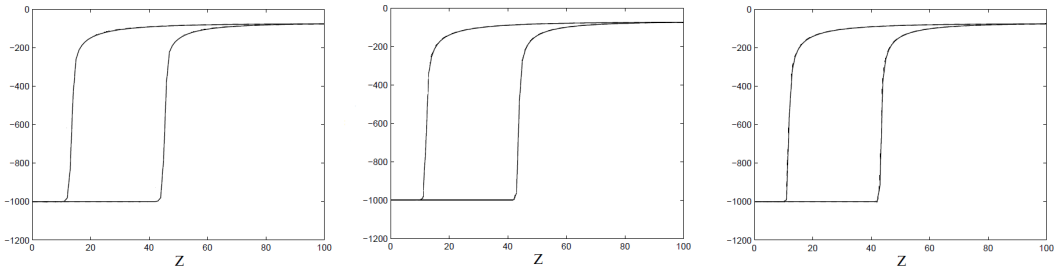
On trace sur la figure 2.13 les valeurs de $\overline{\psi_h}$ au bout de $24h$ et $48h$ de simulation, obtenues avec les maillages M_4, M_5 et M_6. Pour un maillage suffisamment fin, le schéma fournit un profil de charge en accord avec les résultats trouvés dans la littérature [20, 59]. Il est

intéressant de constater que, même pour ce cas test raide, la solution reste monotone et bornée dans l'intervalle $[-1000, -75]$.



(A) Teneur en eau θ. (B) Perméabilité k (échelle logarithmique).

FIGURE 2.12: TC2 - Relations constitutives $\theta(\psi)$ et $k(\psi)$.



(A) M_4. (B) M_5. (C) M_6.

FIGURE 2.13: TC2 - Charge moyenne $\overline{\psi_h}$ à $24h$ et $48h$ sur différents maillages.

2.4.3 Quarter five spot en milieu hétérogène anisotrope (TC3)

Ce dernier cas test s'inspire de la configuration five spot étudiée dans [72], qui reproduit une cellule élémentaire constituée d'un réseau périodique de sources et de puits. Le domaine $\Omega = [0, 1]^2$ (en mètre) est découpé en quatre parties (voir la figure 2.14),

$$\Omega_1 = \Omega \cap \{x + z \leq 0.5\}, \quad \Omega_2 = \Omega \cap \{0.5 < x + z \leq 1\},$$
$$\Omega_3 = \Omega \cap \{1 < x + z \leq 1.5\}, \quad \Omega_4 = \Omega \cap \{1.5 < x + z\}.$$

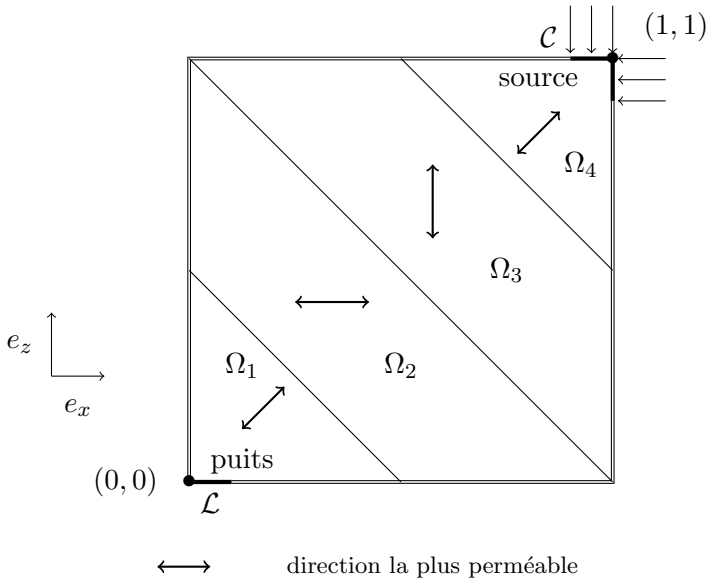


FIGURE 2.14: TC3 - Problème quarter five spot en milieu hétérogène anisotrope.

Les propriétés du sol sont celles définies par Van Genuchten, à l'exception de la perméabilité intrinsèque $\overline{\mathbb{K}}$, qui est constante par morceaux et définie par :

$$\mathbb{K}(\mathbf{x}) = \sum_{i=1}^{4} \mathbb{1}_{\Omega_i}(\mathbf{x}) R_{\omega_i} D R_{\omega_i}^{\mathsf{t}},$$

où la matrice D est diagonale et R_{ω_i} désigne la matrice de rotation associée à Ω_i :

$$D = \begin{pmatrix} 1 & 0 \\ 0 & 10^{-3} \end{pmatrix} \quad \text{and} \quad R_\omega = \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}.$$

Les angles de rotation valent :

$$\omega_1 = \pi/4, \quad \omega_2 = 0, \quad \omega_3 = \pi/2, \quad \omega_4 = \pi/4.$$

Cette perméabilité induit une direction préférentielle de l'écoulement dans chaque partie du domaine, comme illustré sur la figure 2.14. Le temps final est $T = 6h$. On applique une condition initiale hydrostatique, ainsi qu'une condition de Neumann homogène sur l'ensemble du domaine; à l'exception du coin supérieur droit $\mathcal{C}$ le long duquel on impose un flux entrant g non nul, et sur le segment $\mathcal{L} = [0, 0.025] \times \{0\}$ sur lequel on applique

une condition de Dirichlet homogène :

$$\begin{cases} \psi(\mathbf{x},0) = -z & \text{dans } \Omega, \\ \psi(s,t) = 0 & \text{sur } \mathcal{L}\times]0,T], \\ -\mathbb{K}(\psi,s)(\nabla\psi(s,t)+e_z)\cdot n(s) = g(s,t) & \text{sur } \mathcal{C}\times]0,T], \\ -\mathbb{K}(\psi,s)(\nabla\psi(s,t)+e_z)\cdot n(s) = 0 & \text{sur } \Gamma\times]0,T], \end{cases}$$

avec $\mathcal{C} = \{1\}\times[0.975,1]\cup[0.975,1]\times\{1\}$ et $\Gamma = \partial\Omega\setminus(\mathcal{L}\cup\mathcal{C})$. La fonction g (en $cm.s^{-1}$) est donnée par (voir la figure 2.15) :

$$g(s,t) = \begin{cases} -5\cdot 10^{-3}\dfrac{t}{1800} & \text{si } t \le 0.5h, \\ -5\cdot 10^{-3} & \text{si } 0.5h < t \le 4h, \\ 0 & \text{si } t > 4h. \end{cases}$$

La valeur $-5\cdot 10^{-3} cm.s^{-1}$ correspond à un débit d'injection de $9kg$ par heure. En plus

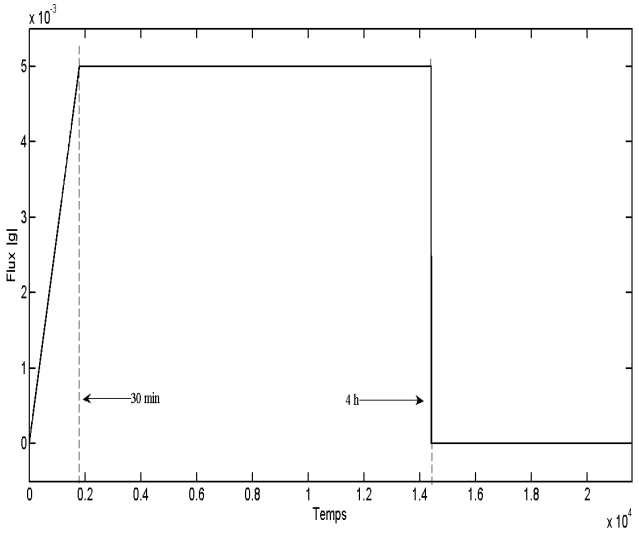


FIGURE 2.15: TC3 - Flux de Neumann à l'entrée du domaine.

du maillage isotrope M_5 utilisé dans les cas tests précédents, on génère à l'aide du logiciel FreeFem [67], un maillage hétérogène anisotrope, adapté à la direction préférentielle de l'écoulement, et de taille proche de celle de M_5. La figure 2.16 montre que les triangles formant le maillage primaire appartiennent à un unique sous-domaine à la fois. Les isolignes de la surpression $\psi_h^T - \psi_h^0$ au temps final de la simulation sont tracées sur la figure 2.17. On s'aperçoit que le schéma produit un profil proche dans les deux cas.

Ainsi, DDFV semble peu sensible au choix du maillage dans le cas d'une perméabilité hétérogène anisotrope.

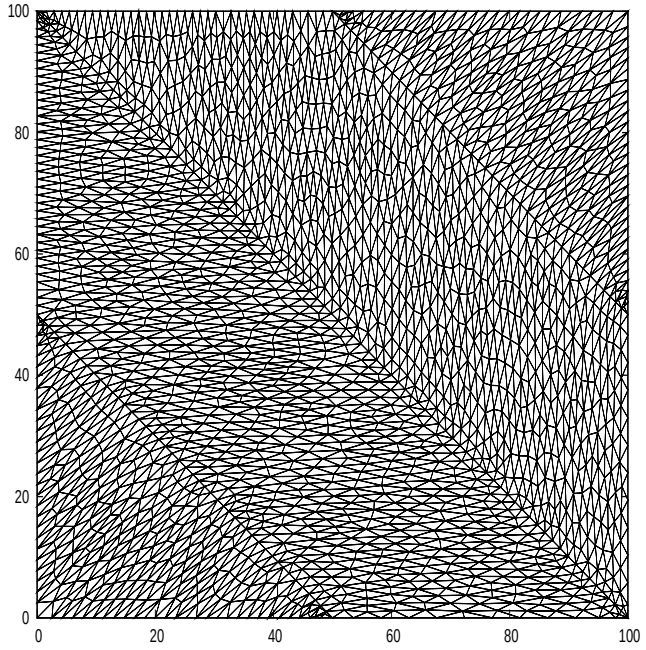
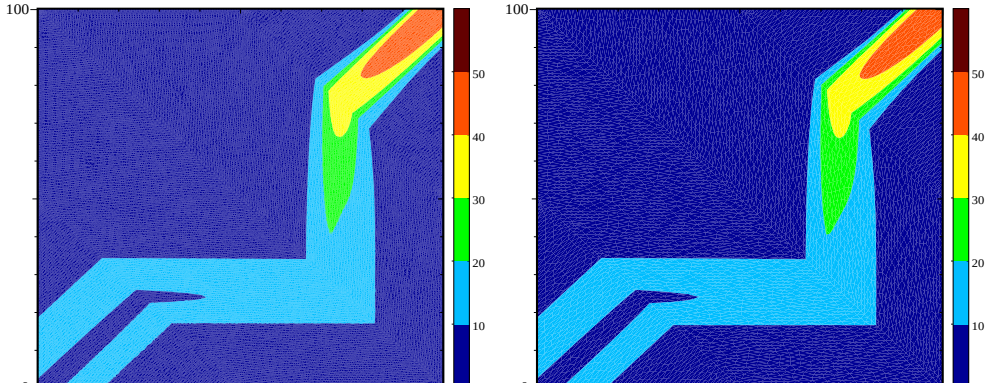


FIGURE 2.16: TC3 - Exemple de maillage hétérogène anisotrope.



(A) Maillage isotrope. (B) Maillage anisotrope.

FIGURE 2.17: TC3 - Isolignes de la surpression à $T = 6h$.

Chapitre 3

Estimations *a posteriori* pour l'équation de Richards

Dans le souci d'efficacité décrit dans le chapitre 1, on souhaite avoir, à maillage fixé, un contrôle sur l'erreur commise par notre résolution numérique, et pouvoir distinguer une erreur provenant de la discrétisation en espace, une erreur due à la discrétisation en temps, et enfin une erreur associée aux linéarisations du système discret. En effet, puisque le maillage est fixé pour l'ensemble de la simulation, l'erreur due à la discrétisation en espace l'est aussi, et on souhaite donc obtenir une solution approchée de qualité tout en évitant des calculs inutiles, qui ne permettraient pas d'améliorer significativement l'erreur totale. L'objectif est donc de mettre en place une stratégie d'adaptation du pas de temps et du critère d'arrêt des linéarisations qui équilibre les différentes sources d'erreur, en un sens qui sera précisé dans le chapitre 4. Dans ce chapitre, on établit les estimations *a posteriori* nécessaires au bon fonctionnement d'une telle stratégie.
Le plan est le suivant : on commence par décrire le cadre fonctionnel de notre étude, et on précise en quel sens on va mesurer l'erreur entre la solution exacte et la solution approchée. Puis, au niveau discret, on indique la classe des schémas en temps que l'on considère ici. Le point important ici est la reformulation de la formule BDF2, laquelle, mise sous la forme d'un schéma à un pas à l'aide d'une formule de récurrence, pourra être associée à notre stratégie d'adaptation. On donne ensuite deux bornes supérieures pour notre erreur. La première suppose que l'on sait résoudre de manière exacte le système discret non linéaire, et nous permet de clarifier les mécanismes mis en jeu dans cette estimation. L'ingrédient clé est la reconstruction des différents flux vérifiant une hypothèse d'équilibrage locale, liée au schéma. La seconde étend ensuite le résultat précédent au cas où des linéarisations sont présentes, et c'est cette estimation qui sera utilisée dans le chapitre 4. Enfin, on propose des reconstructions adaptées à notre discrétisation DDFV-BDF2 présentée dans le chapitre 2.

3.1 Problème continu

On s'intéresse toujours à l'équation de Richards (2.12). Pour rappel :

$$\begin{cases} \partial_t \theta(\psi(\mathbf{x},t)) - \nabla \cdot (\mathbb{K}(\psi,\mathbf{x})(\nabla\psi(\mathbf{x},t) + e_z)) = f(\mathbf{x},t) & \text{dans } \ Q_T := \Omega\times]0,T], \\ \psi(\mathbf{x},0) = \psi^0(\mathbf{x}) & \text{dans } \ \Omega \times \{0\}, \\ \psi(s,t) = \psi_{\mathrm{D}}(s,t) & \text{sur } \ \partial\Omega^{\mathrm{D}}\times]0,T], \\ -\mathbb{K}(\psi,s)(\nabla\psi(s,t) + e_z)\cdot\nu(s) = g(s,t) & \text{sur } \ \partial\Omega^{\mathrm{N}}\times]0,T]. \end{cases}$$

Dans cette section, on indique les conditions de régularité nécessaires à l'existence et à l'unicité d'une solution faible. On précise les espaces fonctionnels associés, et on définit le résidu de l'équation, qui sera notre mesure de l'erreur à contrôler.

3.1.1 Hypothèses sur les données

On fait les hypothèses de régularité suivantes :

- L'ouvert Ω est connexe, borné, et sa frontière $\partial\Omega$ est lipschitzienne.
- Le temps final T vérifie $T > 0$.
- La teneur en eau $\theta : \mathbb{R} \to \mathbb{R}$ est une fonction croissante et lipschitzienne sur $\mathbb{R}$.
- Le tenseur $\mathbb{K}$ prend la forme $\mathbb{K}(\psi,\mathbf{x}) = \overline{\mathbb{K}}(\mathbf{x})k(\psi)$, où :
  - $k : \mathbb{R} \to [k_{\min}, k_{\max}]$, avec $0 < k_{\min} \leq k_{\max}$, est une fonction continue et bornée sur $\mathbb{R}$.
  - $\overline{\mathbb{K}} : \Omega \to \mathcal{M}_2$, $\mathcal{M}_2$ représentant l'ensemble des matrices carrées de taille 2 à coefficients réels, est une fonction mesurable, bornée et uniformément elliptique sur Ω :

    $$\forall\, \mathbf{x} \in \Omega,\ \forall\, \xi \in \mathbb{R}^2, \quad \overline{\mathbb{K}}_{\min}|\xi|^2 \leq \xi^{\mathrm{t}}\overline{\mathbb{K}}(\mathbf{x})\xi \leq \overline{\mathbb{K}}_{\max}|\xi|^2, \text{ avec } 0 < \overline{\mathbb{K}}_{\min} \leq \overline{\mathbb{K}}_{\max}.$$

- Le terme source f est dans $L^2(Q_T)$.
- La donnée initiale ψ^0 est dans $L^2(\Omega)$.
- La donnée au bord de Dirichlet ψ_{D} est dans $L^2(0,T;H^1(\Omega)) \cap L^\infty(Q_T)$.
- La donnée au bord de Neumann g est dans $H^{\frac{1}{2}}(\partial\Omega)$.

Notons que l'hypothèse $k_{\min} > 0$ est importante pour les résultats d'existence et d'unicité évoqués ci-après. Elle évite d'éventuels problèmes de dégénérescence lorsque le milieu est complètement désaturé.

3.1.2 Solution faible, définition du résidu

Pour une fonction ψ suffisamment régulière et une fonction test ϕ choisie dans l'espace :

$$Y := \{\phi \in H^1(Q_T) \mid \phi(.,T) = 0,\ \phi(x,t) = 0\ \forall\ (x,t) \in \partial\Omega^{\mathrm{D}} \times]0,T]\},$$

on a, en intégrant l'équation (2.12) sur Q_T :

$$\int_0^T \{(f,\phi) + (\theta(\psi), \partial_t\phi) - (\mathbb{K}(\psi)\nabla(\psi+z), \nabla\phi) + (\mathbb{K}(\psi)\nabla(\psi+z)\cdot\nu, \phi)_{\partial\Omega}\}(t)\ \mathrm{d}t \\ + (\theta(\psi^0), \phi(.,0)) - (\theta(\psi(.,T)), \phi(.,T)) = 0,$$

où $(.,.)$ désigne le produit scalaire usuel dans $L^2(\Omega)$. On a intégré par parties le terme de dérivée en temps et appliqué le théorème de Green au terme de divergence. En utilisant la condition de Neumann imposée sur $\partial\Omega^{\mathrm{N}}$ et la relation $\phi(.,T) = 0$, on arrive à la formulation faible suivante de l'équation (2.12) :

$$\int_0^T \{(f,\phi) + (\theta(\psi), \partial_t\phi) - (\mathbb{K}(\psi)\nabla(\psi+z), \nabla\phi) - (g,\phi)_{\partial\Omega^{\mathrm{N}}}\}(t)\ \mathrm{d}t + (\theta(\psi^0), \phi(.,0)) = 0. \tag{3.1}$$

L'équation (3.1) a un sens pour ψ dans l'espace $X := L^2(0,T;W) \cap L^\infty(Q_T)$, où l'espace W est défini par : $W := \{f \in H^1(\Omega) \mid f = 0 \text{ sur } \partial\Omega^{\mathrm{D}}\}$. Une solution faible de (2.12) est donc une application $\psi : \mathbb{R}^2 \times]0,+\infty[\to \mathbb{R}$ vérifiant :

- $\psi - \psi_{\mathrm{D}}$ est dans X,
- Pour toute fonction ϕ dans Y,

$$\int_0^T \{(f,\phi) + (\theta(\psi), \partial_t\phi) - (\mathbb{K}(\psi)\nabla(\psi+z), \nabla\phi) - (g,\phi)_{\partial\Omega^{\mathrm{N}}}\}(t)\ \mathrm{d}t + (\theta(\psi^0), \phi(.,0)) = 0.$$

Notons que les hypothèses sur le tenseur $\mathbb{K}$ garantissent que $\mathbb{K}(\psi)\nabla(\psi+z)$ est dans $[L^2(Q_T)]^2$. Sous ces hypothèses, on peut prouver l'existence [7] et l'unicité [66] d'une telle solution faible lorsque la condition de Neumann est homogène : $g = 0$. On définit

alors, pour $\psi \in X$, la quantité $R(\psi)$ par la formule suivante :

$$\langle R(\psi), \phi \rangle := \int_0^T \{(f,\phi) + (\theta(\psi), \partial_t \phi) - (\mathbb{K}(\psi)\nabla(\psi + z), \nabla\phi) - (g,\phi)_{\partial\Omega^N}\}(t)\ \mathrm{d}t \\ + (\theta(\psi^0), \phi(.,0)), \quad \forall \phi \in Y. \quad (3.2)$$

Cette expression définit une unique forme linéaire continue sur Y, appelée *résidu de* ψ. Notons qu'une solution faible est alors définie par :

- $\psi - \psi_D \in X$,
- $R(\psi) = 0$.

Pour une fonction conforme ψ_{ht} de X provenant d'une solution discrète Ψ, il est donc naturel de définir le résidu $R(\psi_{ht})$ comme mesure de l'erreur d'approximation de la solution ψ de (2.12).

3.2 Schémas en temps et linéarisations

La discrétisation en espace donne lieu à un système d'équations différentielles, que l'on discrétise à l'aide d'un schéma en temps à un ou plusieurs pas. On considèrera à titre d'exemple les schémas d'Euler implicite, de Crank-Nicolson et BDF2. Par souci de cohérence avec la section 3.3 qui porte sur l'estimation *a posteriori* par une méthode de flux équilibrés, il est utile d'avoir une formulation générique, indépendante du schéma, et prenant la forme d'un schéma à un pas de type Euler implicite. Cette section est dédiée à l'expression de cette formulation, et notamment à la réécriture du schéma BDF2, qui fait intervenir un second membre défini par une formule de récurrence.

3.2.1 Cadre discret général

On se donne un cadre abstrait pour fixer les idées. On définit une suite d'instants discrets $t^0 = 0 < t^1 < \ldots < t^{n-1} < t^n < \ldots < t^N = T$ et on note $I^n := [t^{n-1}, t^n]$. Pour la discrétisation du domaine Ω, on considère un maillage $\mathcal{T}$ constitué de triangles. On désigne par h_K le diamètre d'un triangle K de $\mathcal{T}$ et par $h = \max_{K\in\mathcal{T}} h_K$ celui du maillage. On suppose disposer d'un schéma volumes finis qui aboutit à la forme semi-discrétisée suivante de l'équation (2.12), pour tout triangle K :

$$|K|\frac{d}{dt}(\theta(\psi_K)) + \sum_{\sigma\in\mathcal{F}_K} |\sigma| F_{\sigma,K} = |K| f_K, \quad (3.3)$$

où ψ_K approche la solution faible ψ de l'équation de Richards (2.12) sur le triangle K, le flux numérique $F_{\sigma,K}$ est une approximation du flux continu $-\mathbb{K}(\psi)\nabla(\psi+z)$ sur une arête σ bordant K, et f_K approche le terme source f sur K. Ces équations locales peuvent être assemblées en un système global et associées à une condition initiale, ce qui donne le problème de Cauchy suivant :

$$\begin{cases} \dfrac{d}{dt}\Theta(\Psi) = V(\Psi), \\ \Psi(0) = \Psi^0, \end{cases} \tag{3.4}$$

où $\Psi \in \mathbb{R}^P$ rassemble les P degrés de liberté en espace de la solution discrète, $\Psi^0 \in \mathbb{R}^P$ est la condition initiale et :

$$\begin{cases} \Theta : \Psi \in \mathbb{R}^P \mapsto \Theta(\Psi) = (|K|\theta(\psi_K))_{K\in\mathcal{T}} \in \mathbb{R}^P, \\ V : \Psi \in \mathbb{R}^P \mapsto V(\Psi) \in \mathbb{R}^P \end{cases}$$

sont des applications non linéaires, suffisamment régulières pour que le problème (3.4) soit bien posé. Avec les notations du chapitre 2, on rappelle que V peut se mettre sous la forme $V(\Psi) = \mathcal{A}(\Psi)\Psi + \mathcal{G}(\Psi) + \mathcal{B}$.

Remarque 1. *La forme $\Theta(\Psi) = (|K|\theta(\psi_K))_{K\in\mathcal{T}}$ est spécifique aux discrétisations en espace mises en œuvre ici, pour lesquelles ψ_K peut être vu comme une approximation de la solution continue ψ au point $\mathbf{x}_K$.*

On suppose également que l'on dispose d'une linéarisation de chacune de ces deux fonctions en tout $\Psi \in \mathbb{R}^P$, c'est-à-dire que l'on dispose de deux applications $d\Theta, dV : \mathbb{R}^P \to \mathcal{M}_P$ ($\mathcal{M}_P$ désignant l'ensemble des matrices carrées de taille P à coefficients dans $\mathbb{R}$) vérifiant :

$$\forall\ \Phi, \Psi \in \mathbb{R}^P,\ \begin{cases} \Theta(\Psi) - \Theta(\Phi) \simeq d\Theta(\Phi)(\Psi - \Phi), \\ V(\Psi) - V(\Phi) \simeq dV(\Phi)(\Psi - \Phi). \end{cases}$$

On dira que $d\Theta(\Phi)$ et $dV(\Phi)$ sont les matrices de linéarisation de Θ et V en $\Phi \in \mathbb{R}^P$. On note alors

$$\forall\ \Phi, \Psi \in \mathbb{R}^P,\ \begin{cases} \delta\Theta_{\text{lin}}(\Psi, \Phi) := \Theta(\Psi) - \Theta(\Phi) - d\Theta(\Phi)(\Psi - \Phi), \\ \delta V_{\text{lin}}(\Psi, \Phi) := V(\Psi) - V(\Phi) - dV(\Phi)(\Psi - \Phi), \end{cases}$$

les erreurs de linéarisation correspondantes. En pratique, on choisit

$$\begin{cases} d\Theta(\Phi) := \text{diag}((\Theta'(\phi_K))_{K\in\mathcal{T}}), \\ dV(\Phi) := \mathcal{A}(\Phi), \end{cases}$$

où pour $\Phi \in \mathbb{R}^P$, diag(Φ) est la matrice diagonale dont les coefficients diagonaux sont donnés par les coordonnées du vecteur Φ. Ces linéarisations correspondent à une itération de Newton pour définir $d\Theta(\Phi)$ et à une itération de Picard pour définir $dV(\Phi)$.
Un schéma en temps définit alors une suite d'approximations $\Psi^1, \cdots, \Psi^N$ de la solution Ψ du problème (3.4) aux instants discrets $t^1, \cdots, t^N$. Ces approximations sont définies par une récurrence non linéaire. Ainsi, nous allons montrer de quelle manière les trois schémas cités dans l'introduction de cette section peuvent se mettre sous la forme :

$$\begin{cases} \Psi^0 \in \mathbb{R}^P, \\ \forall\ 1 \leq n \leq N,\ \Theta(\Psi^n) - \Theta(\Psi^{n-1}) = \delta t^n V^n(\Psi^n). \end{cases} \tag{3.5}$$

On voit bien qu'en dehors du schéma d'Euler implicite pour lequel on a clairement $V^n(\Psi^n) = V(\Psi^n)$, l'application $V^n : \Psi \in \mathbb{R}^P \mapsto V^n(\Psi) \in \mathbb{R}^P$ doit être précisée. C'est l'objet de la section suivante.
On suppose que l'écriture (3.5) permet également de définir deux matrices de linéarisation $dV^n(\Phi)$ et $d\Theta^n(\Phi)$ associées respectivement à V^n et Θ, à partir des matrices de linéarisation $d\Theta(\Phi)$ et $dV(\Phi)$. On note alors

$$\forall\ \Phi, \Psi \in \mathbb{R}^P,\ \begin{cases} \delta V^n_{\text{lin}}(\Psi, \Phi) & := V^n(\Psi) - V^n(\Phi) - dV^n(\Phi)(\Psi - \Phi), \\ \delta \Theta^n_{\text{lin}}(\Psi, \Phi) & := \Theta(\Psi) - \Theta(\Phi) - d\Theta^n(\Phi)(\Psi - \Phi), \end{cases}$$

les erreurs de linéarisation correspondantes. La dépendance en n de la matrice de linéarisation $d\Theta^n$ sera visible dans le cas du schéma BDF2, qui sera reformulé dans la sous-section 3.2.3. Puisqu'on ne peut pas résoudre de manière exacte l'équation définissant le schéma (3.5) d'inconnue Ψ^n, on calcule plutôt, pour tout $1 \leq n \leq N$, la suite $(\Psi^{n,m})_{m \geq 0}$ obtenue en remplaçant dans l'équation non linéaire $\Theta(\Psi^{n,m})$ et $V^n(\Psi^{n,m})$ par leur linéarisation en $\Psi^{n,m-1}$:

$$\begin{cases} \Theta(\Psi^{n,m}) \simeq \Theta(\Psi^{n,m-1}) + d\Theta^n(\Psi^{n,m-1})(\Psi^{n,m} - \Psi^{n,m-1}), \\ V^n(\Psi^{n,m}) \simeq V^n(\Psi^{n,m-1}) + dV^n(\Psi^{n,m-1})(\Psi^{n,m} - \Psi^{n,m-1}). \end{cases}$$

L'étape m de l'itération n en temps consiste en le système linéaire suivant, d'inconnue $\Psi^{n,m}$:

$$\begin{cases} \Psi^{n,0} = \Psi^{n-1} \text{ (par exemple)}, \\ \Theta(\Psi^{n,m}) - \Theta(\Psi^{n-1}) = \delta t^n (V^n(\Psi^{n,m}) - \delta V^n_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1})) \\ \qquad\qquad\qquad\qquad\qquad\qquad + \delta \Theta^n_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1}). \end{cases}$$

Sous cette forme, on voit apparaître les modifications induites par les linéarisations sur la forme non linéaire (3.5). Elle sera également plus facile à manipuler pour la reformulation du schéma BDF2. On convient ensuite d'arrêter les itérations de linéarisation pour

$m = m_\infty(n)$ et de noter $\Psi^n := \Psi^{n,m_\infty(n)}$. De cette manière, la suite $(\Psi^1, \Psi^2, \cdots, \Psi^N)$ est entièrement définie par :

$$\begin{cases} \Psi^0 \in \mathbb{R}^P, \\ \forall\ 1 \leq n \leq N,\ \Psi^n = \Psi^{n,m_\infty(n)},\ \text{avec} \\ \quad \begin{cases} \Psi^{n,0} = \Psi^{n-1}, \\ \forall\ 1 \leq m \leq m_\infty(n), \\ \Theta(\Psi^{n,m}) - \Theta(\Psi^{n-1}) = \delta t^n (V^n(\Psi^{n,m}) - \delta V^n_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1})) \\ \qquad\qquad\qquad\qquad + \delta\Theta^n_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1}). \end{cases} \end{cases} \tag{3.6}$$

On suppose ici que l'on utilise un solveur linéaire direct; ainsi, on ne tient pas compte des éventuelles erreurs dues à la résolution du système linéaire [32].

3.2.2 Schémas d'Euler implicite et de Crank-Nicolson

On montre dans cette section comment les schémas d'Euler implicite et de Crank-Nicolson se mettent sous les formes (3.5) (forme non linéaire) et (3.6) (forme linéarisée).

Euler implicite Dans ce cas, le schéma a directement la forme (3.5) avec $V^n(\Psi) = V(\Psi)$. De plus, les linéarisations associées à V^n et Θ^n sont simplement $dV^n(\Phi) = dV(\Phi)$ et $d\Theta^n(\Phi) = d\Theta(\Phi)$; la forme (3.6) est valable avec ces notations pour V^n, dV^n et $d\Theta^n$.

Crank-Nicolson Le schéma non linéaire s'écrit dans ce cas :

$$\Theta(\Psi^n) - \Theta(\Psi^{n-1}) = \frac{\delta t^n}{2}\left(V(\Psi^n) + V(\Psi^{n-1})\right),$$

ce qui correspond à la forme (3.5) avec $V^n(\Psi) := \frac{1}{2}(V(\Psi) + V(\Psi^{n-1}))$. On a de plus :

$$\begin{aligned} V^n(\Psi) &= V^n(\Phi) + \frac{1}{2}(V(\Psi) - V(\Phi)) \\ &\simeq V^n(\Phi) + \frac{1}{2} dV(\Phi)(\Psi - \Phi). \end{aligned}$$

On définit donc la matrice de linéarisation de V^n par $dV^n(\Phi) = \frac{1}{2} dV(\Phi)$, et avec de plus $d\Theta^n(\Phi) = d\Theta(\Phi)$, on peut bien écrire le schéma linéarisé sous la forme (3.6).

3.2.3 Schéma BDF2

La discrétisation qui nous intéresse fait intervenir la formule BDF2, qui est un schéma à deux pas. Puisqu'on souhaite adapter le pas de temps au cours de la simulation, on considère la formule BDF2 à pas variable, que l'on propose d'initialiser avec la condition initiale $\Psi^0 \in \mathbb{R}^P$ et le vecteur Ψ^1 obtenu à l'aide du schéma de Crank-Nicolson :

$$\begin{cases} \Psi^0 \in \mathbb{R}^P, \\ \Theta(\Psi^1) - \Theta(\Psi^0) = \dfrac{\delta t^1}{2}(V(\Psi^1) + V(\Psi^0)), \\ \forall\ 2 \leq n \leq N, \quad \displaystyle\sum_{i=0}^{2} \alpha_i^n \Theta(\Psi^{n-i}) = \delta t^n V(\Psi^n), \end{cases} \tag{3.7}$$

avec

$$\alpha_0^n := \frac{1+2r^n}{1+r^n}, \quad \alpha_1^n := -(1+r^n), \quad \text{et} \quad \alpha_2^n := \frac{(r^n)^2}{1+r^n}. \tag{3.8}$$

Le ratio $r^n := \delta t^n / \delta t^{n-1}$ est le rapport entre deux pas de temps consécutifs. On admet ici l'existence et l'unicité de la suite $(\Psi^n)_{0\leq n\leq N}$, qui est un autre problème, renvoyant à la cohérence de la discrétisation en espace. La condition de stabilité $r^n \leq 1+\sqrt{2}$ de ce schéma (voir [48]) permet d'augmenter de manière significative le pas de temps si nécessaire.

Notre but est d'écrire des estimations *a posteriori* locales en temps, c'est-à-dire mesurant une erreur sur chaque intervalle $I^n = [t^{n-1}, t^n]$. Un schéma à deux pas n'est donc pas, sous cette forme, adapté; on en propose ici une réécriture sous les formes (3.5) et (3.6).

Cas linéaire Pour simplifier, on commence par s'intéresser à la situation où pour tout Ψ, $\theta(\Psi) = \Psi$ et $V(\Psi) = \mathbb{V}\Psi$, où $\mathbb{V}$ est une matrice fixée dans $\mathcal{M}_P$. La discrétisation (3.7) se simplifie alors en le système suivant :

$$\begin{cases} \Psi^0 \in \mathbb{R}^P, \\ \Psi^1 - \Psi^0 = \dfrac{\delta t^1}{2}\mathbb{V}(\Psi^1 + \Psi^0), \\ \forall\ 2 \leq n \leq N, \quad \displaystyle\sum_{i=0}^{2} \alpha_i^n \Psi^{n-i} = \delta t^n \mathbb{V}\Psi^n. \end{cases} \tag{3.9}$$

Dans ce cas, on se ramène à la forme (3.5) à l'aide du résultat suivant :

Lemme 3.1. *La suite $(\Psi^n)_{0\leq n\leq N}$ définie par (3.9) est égale à la suite $(\overline{\Psi}^n)_{0\leq n\leq N}$ définie par :*

$$\begin{cases} \overline{\Psi}^0 = \Psi^0, \\ \forall\ 1 \leq n \leq N, \quad \overline{\Psi}^n - \overline{\Psi}^{n-1} = \delta t^n V^n(\overline{\Psi}^n), \end{cases}$$

où la fonction V^n est définie pour tout $\overline{\Psi}$ dans $\mathbb{R}^P$ par :

$$\begin{cases} V^1(\overline{\Psi}) = \dfrac{\mathbb{V}}{2}(\overline{\Psi} + \overline{\Psi}^0), \\ \forall\ 2 \leq n \leq N, \quad V^n(\overline{\Psi}) = \omega^n \mathbb{V}\overline{\Psi} + (1-\omega^n) V^{n-1}(\overline{\Psi}^{n-1}), \quad où \quad \omega^n := \dfrac{1+r^n}{1+2r^n}. \end{cases}$$

On dit que les deux schémas ainsi définis sont équivalents.

Preuve. On procède par récurrence forte sur n.

- Pour $n = 0$, on a $\Psi^0 = \overline{\Psi}^0$ par définition.
- Pour $n = 1$, on a :

$$\begin{aligned} \Psi^1 &= \Psi^0 + \frac{\delta t^1}{2}\mathbb{V}(\Psi^1 + \Psi^0), \\ &= \overline{\Psi}^0 + \delta t^1 \frac{\mathbb{V}}{2}(\Psi^1 + \overline{\Psi}^0), \\ &= \overline{\Psi}^0 + \delta t^1 V^1(\Psi^1), \end{aligned}$$

  qui est exactement l'équation définissant $\overline{\Psi}^1$, d'où $\Psi^1 = \overline{\Psi}^1$.

- Supposons que pour tout $0 \leq k \leq n-1$, on ait $\Psi^k = \overline{\Psi}^k$. Alors Ψ^n est défini par :

$$\alpha_0^n \Psi^n + \alpha_1^n \Psi^{n-1} + \alpha_2^n \Psi^{n-2} = \delta t^n \mathbb{V}\Psi^n,$$

  ce qui est équivalent à :

$$\alpha_0^n (\Psi^n - \Psi^{n-1}) - \alpha_2^n (\Psi^{n-1} - \Psi^{n-2}) = \delta t^n \mathbb{V}\Psi^n,$$

  car $\alpha_0^n + \alpha_1^n + \alpha_2^n = 0$. Par hypothèse de récurrence, on a $\Psi^{n-2} = \overline{\Psi}^{n-2}$ et $\Psi^{n-1} = \overline{\Psi}^{n-1}$, d'où :

$$\frac{1+2r^n}{1+r^n}\left(\Psi^n - \Psi^{n-1}\right) - \frac{(r^n)^2}{1+r^n}\left(\Psi^{n-1} - \Psi^{n-2}\right) = \delta t^n \mathbb{V}\Psi^n$$

  équivaut à :

$$\frac{1+2r^n}{1+r^n}\left(\Psi^n - \overline{\Psi}^{n-1}\right) - \frac{(r^n)^2}{1+r^n}\left(\overline{\Psi}^{n-1} - \overline{\Psi}^{n-2}\right) = \delta t^n \mathbb{V}\Psi^n,$$

  puis à

$$\frac{1+2r^n}{1+r^n}\left(\Psi^n - \overline{\Psi}^{n-1}\right) - \frac{(r^n)^2}{1+r^n}\left(\delta t^{n-1} V^{n-1}(\overline{\Psi}^{n-1})\right) = \delta t^n \mathbb{V}\Psi^n,$$

par définition de $\overline{\Psi}^{n-1}$. Le reste suit naturellement.

$$\begin{aligned}\Psi^n - \overline{\Psi}^{n-1} &= \frac{1+r^n}{1+2r^n}\delta t^n \mathbb{V}\Psi^n + \frac{(r^n)^2}{1+2r^n}\delta t^{n-1} V^{n-1}(\overline{\Psi}^{n-1}),\\ &= \delta t^n \left(\frac{1+r^n}{1+2r^n}\mathbb{V}\Psi^n + \frac{r^n}{1+2r^n}V^{n-1}(\overline{\Psi}^{n-1})\right),\\ &= \delta t^n V^n(\Psi^n).\end{aligned}$$

Ainsi, $\Psi^n = \overline{\Psi}^n$, ce qui achève la récurrence.

□

Remarque 2. *Dans ce cas linéaire, les matrices de linéarisation de Θ et V sont simplement définies pour tout Φ dans $\mathbb{R}^P$ par $d\Theta(\Phi) = I_P$, où I_P désigne la matrice identité de $\mathcal{M}_P$, et par $dV(\Phi) = \mathbb{V}$. D'après l'expression de V^n, on voit également que l'on a $dV^n(\Phi) = \omega^n \mathbb{V}$.*

Cas non linéaire Dans le cas qui nous intéresse, les fonctions Θ et V sont non linéaires en Ψ. Le schéma (3.7) prend alors la forme linéarisée suivante :

$$\begin{cases}\Psi^0 \in \mathbb{R}^P,\\ \forall\ 1 \le n \le N,\ \Psi^n = \Psi^{n,m_\infty(n)},\ \text{avec}\\ \quad\begin{cases}\Psi^{1,0} = \Psi^0,\\ \forall\ 1 \le m \le m_\infty(1),\\ \Theta(\Psi^{1,m}) - \Theta(\Psi^0) = \dfrac{\delta t^1}{2}(V(\Psi^{1,m}) + V(\Psi^0) - \delta V_{\text{lin}}(\Psi^{1,m}, \Psi^{1,m-1}))\\ \qquad\qquad +\delta\Theta_{\text{lin}}(\Psi^{1,m}, \Psi^{1,m-1}),\end{cases}\\ \text{et},\ \forall\ 2 \le n \le N,\\ \quad\begin{cases}\Psi^{n,0} = \Psi^{n-1},\\ \forall\ 1 \le m \le m_\infty(n),\\ \alpha_0^n\Theta(\Psi^{n,m}) + \displaystyle\sum_{i=1}^{2}\alpha_i^n\Theta(\Psi^{n-i}) = \delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1}))\\ \qquad\qquad +\alpha_0^n\delta\Theta_{\text{lin}}(\Psi^{n,m}, \Psi^{n,m-1}).\end{cases}\end{cases} \tag{3.10}$$

On se ramène alors à la forme (3.6) à l'aide du résultat suivant :

Lemme 3.2. *La suite $(\Psi^{n,m})_{\substack{0\le n\le N\\ 0\le m\le m_\infty(n)}}$ définie par (3.10) est égale à la suite*

$(\overline{\Psi}^{n,m})_{\substack{0\leq n\leq N\\ 0\leq m\leq m_\infty(n)}}$ *définie par :*

$$\begin{cases} \overline{\Psi}^0 = \Psi^0, \\ \forall\ 1\leq n\leq N,\ \overline{\Psi}^n = \overline{\Psi}^{n,m_\infty(n)},\ \textit{avec} \\ \quad \begin{cases} \overline{\Psi}^{n,0} = \overline{\Psi}^{n-1}, \\ \forall\ 1\leq m\leq m_\infty(n), \\ \Theta(\overline{\Psi}^{n,m}) - \Theta(\overline{\Psi}^{n-1}) = \delta t^n (V^n(\overline{\Psi}^{n,m}) - \delta V^n_{lin}(\overline{\Psi}^{n,m}, \overline{\Psi}^{n,m-1})) \\ \qquad\qquad\qquad\qquad +\delta\Theta^n_{lin}(\overline{\Psi}^{n,m}, \overline{\Psi}^{n,m-1}), \end{cases} \end{cases} \tag{3.11}$$

où les fonctions V^n, δV^n_{lin} et $\delta\Theta^n_{lin}$ sont définies pour tout $\overline{\Psi}, \overline{\Phi}$ dans $\mathbb{R}^P$ par :

$$\begin{cases} V^1(\overline{\Psi}) = \dfrac{1}{2}(V(\overline{\Psi}) + V(\overline{\Psi}^0)), \\ \delta V^1_{lin}(\overline{\Psi}, \overline{\Phi}) = \dfrac{1}{2}\delta V_{lin}(\overline{\Psi}, \overline{\Phi}), \\ \delta\Theta^1_{lin}(\overline{\Psi}, \overline{\Phi}) = \delta\Theta_{lin}(\overline{\Psi}, \overline{\Phi}), \end{cases}$$

et pour tout $2\leq n\leq N$:

$$\begin{cases} V^n(\overline{\Psi}) = \omega^n V(\overline{\Psi}) + (1-\omega^n)V^{n-1}(\overline{\Psi}^{n-1}), \\ \delta V^n_{lin}(\overline{\Psi}, \overline{\Phi}) = \omega^n \delta V_{lin}(\overline{\Psi}, \overline{\Phi}) + (1-\omega^n)\delta V^{n-1}_{lin}(\overline{\Psi}^{n-1}, \overline{\Psi}^{n-1,m_\infty(n-1)-1}), \\ \delta\Theta^n_{lin}(\overline{\Psi}, \overline{\Phi}) = \delta\Theta_{lin}(\overline{\Psi}, \overline{\Phi}) + (1-\omega^n) r^n \delta\Theta^{n-1}_{lin}(\overline{\Psi}^{n-1}, \overline{\Psi}^{n-1,m_\infty(n-1)-1}). \end{cases} \tag{3.12}$$

Preuve. On procède de nouveau par récurrence forte sur $0\leq n\leq N$.

- Pour $n=0$, on a $\Psi^0 = \overline{\Psi}^0$ par définition.
- Pour $n=1$, on procède par récurrence simple sur $0\leq m\leq m_\infty(1)$:
  m=0 On a par définition : $\Psi^{1,0} = \Psi^0 = \overline{\Psi}^0 = \overline{\Psi}^{1,0}$.
  Hérédité Si $\Psi^{1,m-1} = \overline{\Psi}^{1,m-1}$, alors $\Psi^{1,m}$ est la solution du système linéaire :

$$\begin{aligned} \Theta(\Psi^{1,m}) &= \Theta(\Psi^0) + \frac{\delta t^1}{2}(V(\Psi^{1,m}) + V(\Psi^0) - \delta V_{\text{lin}}(\Psi^{1,m}, \Psi^{1,m-1})) \\ &\qquad +\delta\Theta_{\text{lin}}(\Psi^{1,m}, \Psi^{1,m-1}), \\ &= \Theta(\overline{\Psi}^0) + \frac{\delta t^1}{2}(V(\Psi^{1,m}) + V(\overline{\Psi}^0) - \delta V_{\text{lin}}(\Psi^{1,m}, \overline{\Psi}^{1,m-1})) \\ &\qquad +\delta\Theta_{\text{lin}}(\Psi^{1,m}, \overline{\Psi}^{1,m-1}), \end{aligned}$$

qui, avec les définitions de V^1_{lin}, δV^1_{lin} et $\delta\Theta^1_{\text{lin}}$, est exactement l'équation définissant $\overline{\Psi}^{1,m}$, d'où $\Psi^{1,m} = \overline{\Psi}^{1,m}$.

- Supposons que pour tout $0 \leq k \leq n-1$, on ait $\Psi^k = \overline{\Psi}^k$, et procédons de nouveau par récurrence sur $0 \leq m \leq m_\infty(n)$.
  m=0 On a par définition : $\Psi^{n,0} = \Psi^{n-1} = \overline{\Psi}^{n-1} = \overline{\Psi}^{n,0}$.
  Hérédité Si $\Psi^{n,m-1} = \overline{\Psi}^{n,m-1}$, alors $\Psi^{n,m}$ est solution du système linéaire :

$$\alpha_0^n\Theta(\Psi^{n,m}) + \sum_{i=1}^{2}\alpha_i^n\Theta(\Psi^{n-i}) = \delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m},\Psi^{n,m-1})) + \alpha_0^n\delta\Theta_{\text{lin}}(\Psi^{n,m},\Psi^{n,m-1}),$$

qui se réécrit, en suivant le même calcul que dans le cas linéaire :

$$\frac{1+2r^n}{1+r^n}\left(\Theta(\Psi^{n,m}) - \Theta(\Psi^{n-1})\right) - \frac{(r^n)^2}{1+r^n}\left(\Theta(\Psi^{n-1}) - \Theta(\Psi^{n-2})\right) = \delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m},\Psi^{n,m-1})) + \frac{1+2r^n}{1+r^n}\delta\Theta_{\text{lin}}(\Psi^{n,m},\Psi^{n,m-1}),$$

soit, par hypothèse de récurrence :

$$\frac{1+2r^n}{1+r^n}\left(\Theta(\Psi^{n,m}) - \Theta(\overline{\Psi}^{n-1})\right) - \frac{(r^n)^2}{1+r^n}\left(\Theta(\overline{\Psi}^{n-1}) - \Theta(\overline{\Psi}^{n-2})\right) = \delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1})) + \frac{1+2r^n}{1+r^n}\delta\Theta_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1}),$$

ce qui donne, en remplaçant $\Theta(\overline{\Psi}^{n-1}) - \Theta(\overline{\Psi}^{n-2})$ par son expression (on rappelle que $\overline{\Psi}^{n-1} = \overline{\Psi}^{n-1,m_\infty(n-1)}$) :

$$\frac{1+2r^n}{1+r^n}\left(\Theta(\Psi^{n,m}) - \Theta(\overline{\Psi}^{n-1})\right) - \frac{(r^n)^2}{1+r^n}\Big(\delta t^{n-1}(V^{n-1}(\overline{\Psi}^{n-1}) - \delta V^{n-1}_{\text{lin}}(\overline{\Psi}^{n-1},\overline{\Psi}^{n-1,m_\infty(n-1)-1})) + \delta\Theta^{n-1}_{\text{lin}}(\overline{\Psi}^{n-1},\overline{\Psi}^{n-1,m_\infty(n-1)-1})\Big) = \delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1})) + \frac{1+2r^n}{1+r^n}\delta\Theta_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1}).$$

On en tire :

$$\Theta(\Psi^{n,m}) - \Theta(\overline{\Psi}^{n-1}) = \frac{(r^n)^2}{1+2r^n}\Big(\delta t^{n-1}(V^{n-1}(\overline{\Psi}^{n-1}) - \delta V^{n-1}_{\text{lin}}(\overline{\Psi}^{n-1},\overline{\Psi}^{n-1,m_\infty(n-1)-1})) + \delta\Theta^{n-1}_{\text{lin}}(\overline{\Psi}^{n-1},\overline{\Psi}^{n-1,m_\infty(n-1)-1})\Big) + \frac{1+r^n}{1+2r^n}\delta t^n(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1})) + \delta\Theta_{\text{lin}}(\Psi^{n,m},\overline{\Psi}^{n,m-1}),$$

puis,

$$\Theta(\Psi^{n,m}) - \Theta(\overline{\Psi}^{n-1}) = \delta t^n \left(\frac{1+r^n}{1+2r^n}(V(\Psi^{n,m}) - \delta V_{\text{lin}}(\Psi^{n,m}, \overline{\Psi}^{n,m-1})) \right.$$
$$\left. + \frac{r^n}{1+2r^n}(V^{n-1}(\overline{\Psi}^{n-1}) - \delta V_{\text{lin}}^{n-1}(\overline{\Psi}^{n-1}, \overline{\Psi}^{n-1,m_\infty(n-1)-1})) \right)$$
$$+ \delta\Theta_{\text{lin}}(\Psi^{n,m}, \overline{\Psi}^{n,m-1}) + \frac{(r^n)^2}{1+2r^n}\delta\Theta_{\text{lin}}^{n-1}(\overline{\Psi}^{n-1}, \overline{\Psi}^{n-1,m_\infty(n-1)-1}),$$

qui, avec les définitions de V_{lin}^n, δV_{lin}^n et $\delta\Theta_{\text{lin}}^n$, est exactement l'équation définissant $\overline{\Psi}^{n,m}$, d'où $\Psi^{n,m} = \overline{\Psi}^{n,m}$.

□

3.2.4 Synthèse

On a montré dans cette section comment la discrétisation en temps de l'équation locale semi-discrétisée en espace (3.3) pouvait, pour tout triangle K de $\mathcal{T}$, se mettre sous la forme non linéaire :

$$\begin{cases} \psi_K^0 \in \mathbb{R}, \\ \forall\ 1 \le n \le N,\ |K|(\theta(\psi_K^n) - \theta(\psi_K^{n-1})) + \delta t^n \displaystyle\sum_{\sigma\in\mathcal{F}_K} |\sigma| F_{\sigma,K}^n = \delta t^n |K| f_K^n, \end{cases} \tag{3.13}$$

ou sous la forme linéarisée :

$$\begin{cases} \psi_K^0 \in \mathbb{R}, \\ \forall\ 1 \le n \le N,\ \psi_K^n = \psi_K^{n,m_\infty(n)},\ \text{avec} \\ \quad \begin{cases} \psi_K^{n,0} = \psi_K^{n-1}, \\ \forall\ 1 \le m \le m_\infty(n), \\ |K|(\theta(\psi_K^{n,m}) - \theta(\psi_K^{n-1})) + \delta t^n \left(\displaystyle\sum_{\sigma\in\mathcal{F}_K} |\sigma| F_{\sigma,K}^{n,m} - \delta F_{\text{lin}}^n(\psi_K^{n,m}, \psi_K^{n,m-1}) \right) \\ \qquad\qquad = \delta t^n |K| f_K^n + |K| \delta\theta_{\text{lin}}^n(\psi_K^{n,m}, \psi_K^{n,m-1}). \end{cases} \end{cases} \tag{3.14}$$

dans le cas des schémas d'Euler implicite, de Crank-Nicolson et BDF2. Il est aisé de généraliser ce résultat à d'autres schémas multipas, à l'aide de formules de récurrence adaptées.

3.3 Estimations *a posteriori* par flux équilibrés

Cette section est dédiée à l'écriture d'estimations *a posteriori* pour l'équation de Richards (2.12), suivant la technique de *flux équilibrés* présentée dans [33]. On souligne qu'à ce stade, on a seulement besoin d'une discrétisation espace-temps du domaine $\Omega \times [0,T]$. Notre objectif est de contrôler, à un instant t^n donné, les erreurs provenant d'une part de la discrétisation en temps, et d'autre part de la linéarisation du système discret, par rapport à l'erreur due à la discrétisation en espace. Pour atteindre ce but, on peut étudier la norme du résidu, en choisissant des fonctions tests à support compact soit dans $[0,t^n[$ (on cherche alors à contrôler des erreurs cumulées sur tous les temps discrets), soit dans chaque intervalle de temps $]t^{n-1},t^n[$ (pour estimer des erreurs localisées à l'instant discret t^n). Pour une fonction ψ_{ht} de X donnée, on définit ainsi deux erreurs :

$$\mathcal{E}^n_{\text{glob}}(\psi_{ht}) := \sup\{\langle R(\psi_{ht}),\phi\rangle \;|\; \phi\in Y,\ \|\phi\|_Y=1,\ \operatorname{supp}\phi \subset [0,t^n[\}, \tag{3.15}$$

$$\mathcal{E}^n_{\text{loc}}(\psi_{ht}) := \sup\{\langle R(\psi_{ht}),\phi\rangle \;|\; \phi\in Y,\ \|\phi\|_Y=1,\ \operatorname{supp}\phi \subset]t^{n-1},t^n[\}, \tag{3.16}$$

où la norme sur l'espace Y est définie par

$$\|\phi\|_Y = \left(\sum_{n=1}^{N}\sum_{K\in\mathcal{T}} \|\phi\|^2_{Y,K\times I^n}\right)^{1/2},$$

avec

$$\|\phi\|_{Y,K\times I^n} := \left(\|\phi\|^2_{K\times I^n} + h_K^2\|\nabla\phi\|^2_{K\times I^n} + (\delta t^n)^2\|\partial_t\phi\|^2_{K\times I^n}\right)^{1/2}. \tag{3.17}$$

On a utilisé ici la notation $\|\cdot\|_{K\times I^n} := \|\cdot\|_{L^2(K\times I^n)}$.

3.3.1 Notion d'équilibrage de flux

Dans l'équation (2.12), les fonctions $\theta(\psi(\mathbf{x},t))$ et $-\mathbb{K}(\psi,\mathbf{x})(\nabla\psi(\mathbf{x},t)+e_z)$ sont des flux exacts en temps et en espace. Des flux dits *équilibrés* seront deux fonctions θ_{ht} et $\mathbf{t}_{ht}$ qui vérifient localement l'équation intégrale écrite sur le maillage espace-temps. Cet équilibre local fait intervenir une approximation f_{ht} du terme source f.

Définition 3.3 (Flux équilibrés)**.** Étant donné une fonction $f_{ht}\in L^2(Q_T)$, on dit que les fonctions $\theta_{ht}: Q_T\to\mathbb{R}$ et $\mathbf{t}_{ht}: Q_T\to\mathbb{R}^2$ sont des flux équilibrés pour f_{ht} si :

$$\theta_{ht}\in L^2(Q_T),\ \partial_t\theta_{ht}\in L^2(Q_T),\ \mathbf{t}_{ht}\in L^2([0,T],H(\operatorname{div},\Omega)),$$

$$\text{et}\quad \forall\, K\in\mathcal{T},\ \forall\, 1\le n\le N,\quad \int_{K\times I^n}\{f_{ht}-\partial_t\theta_{ht}-\operatorname{div}(\mathbf{t}_{ht})\}\ \mathrm{d}\mathbf{x}\,\mathrm{d}t = 0. \tag{3.18}$$

Cette relation d'équilibrage est centrale pour l'obtention des estimations présentées dans la sous-section 3.3.3.

3.3.2 Lien avec les schémas

On remarque qu'avec les hypothèses de régularité décrites dans la définition précédente, l'égalité d'équilibrage (3.18) est équivalente à :

$$\int_K \theta_{ht}(\mathbf{x}, t^n)\ \mathrm{d}\mathbf{x} - \int_K \theta_{ht}(\mathbf{x}, t^{n-1})\ \mathrm{d}\mathbf{x} + \sum_{\sigma\in\mathcal{F}_K} \int_{\sigma\times I^n} \mathbf{t}_{ht}(s,t)\cdot n_{\sigma,K}\ \mathrm{d}s\,\mathrm{d}t = \int_{K\times I^n} f_{ht}(\mathbf{x},t)\ \mathrm{d}\mathbf{x}\,\mathrm{d}t, \quad (3.19)$$

où, pour une maille K donnée, $\mathcal{F}_K$ désigne l'ensemble des arêtes de K, et pour une telle arête σ, $n_{\sigma,K}$ est la normale unitaire à σ sortant de K. Cette relation est de la forme (3.13); on précise ce lien dans le théorème suivant :

Théorème 3.4 (Lien avec les schémas)**.** *Avec les notations précédentes, supposons que, pour tout élément K de $\mathcal{T}$ et toute itération $1 \le n \le N$, il existe une relation du type :*

$$|K|(\theta_K^n - \theta_K^{n-1}) + \delta t^n \sum_{\sigma\in\mathcal{F}_K} |\sigma| F_{\sigma,K}^n = \delta t^n |K| f_K^n,$$

où $(\theta_K^n)_{K\in\mathcal{T},0\le n\le N}$, $(F_{\sigma,K}^n)_{K\in\mathcal{T},\sigma\in\mathcal{F}_K,1\le n\le N}$ et $(f_K^n)_{K\in\mathcal{T},1\le n\le N}$ sont des suites données de réels, et que, pour une arête σ bordant deux mailles voisines K et L et une itération $1 \le n \le N$ donnée, on ait $F_{\sigma,K}^n + F_{\sigma,L}^n = 0$.
Alors tout couple de fonctions $(\theta_{ht}, \mathbf{t}_{ht})$ vérifiant :

$$\theta_{ht} \in L^2(Q_T),\ \partial_t\theta_{ht} \in L^2(Q_T),\ \mathbf{t}_{ht} \in L^2([0,T], H(\mathrm{div},\Omega))$$

$$\text{et} \quad \begin{cases} \forall\ K\in\mathcal{T}, \forall\ 0\le n\le N,\ \displaystyle\int_K \theta_{ht}(\mathbf{x},t^n)\ \mathrm{d}\mathbf{x} = |K|\theta_K^n \\ \forall\ K\in\mathcal{T}, \forall\ \sigma\in\mathcal{F}_K, \forall\ 1\le n\le N,\ \displaystyle\int_{\sigma\times I^n} \mathbf{t}_{ht}(s,t)\cdot n_{\sigma,K}\ \mathrm{d}s\,\mathrm{d}t = \delta t^n|\sigma|F_{\sigma,K}^n \end{cases}$$

est un couple de flux équilibrés pour tout terme source f_{ht} vérifiant :

$$\int_{K\times I^n} f_{ht}(\mathbf{x},t)\ \mathrm{d}\mathbf{x}\,\mathrm{d}t = \delta t^n |K| f_K^n.$$

Preuve. Cela découle directement de l'équation (3.19). □

Remarque 3. *C'est ce théorème qui va guider notre choix de reconstructions pour les flux spatial et temporel, que l'on verra dans la section 3.4.*

3.3.3 Estimations pour un solveur non linéaire exact

On commence par énoncer une inégalité de Poincaré espace-temps, introduite dans [33], qui est un ingrédient important de l'estimation qui va suivre. Elle est obtenue à partir d'une inégalité de Poincaré optimale pour les domaines convexes, voir par exemple [12].

Lemme 3.5 (Inégalité de Poincaré espace-temps)**.** *Pour ϕ dans $H^1(Q_T)$, on note ϕ_{ht} sa projection L^2-orthogonale sur les fonctions constantes par morceaux sur chaque intervalle espace-temps $K \times I^n$ ($K \in \mathcal{T}$ et $1 \leq n \leq N$) :*

$$\phi_{ht} := \sum_{n=1}^{N} \sum_{K \in \mathcal{T}} \phi_K^n \mathbb{1}_K^n \in L^2(Q_T),$$

où ϕ_K^n est la moyenne de ϕ sur $K \times I^n$ et $\mathbb{1}_K^n$ désigne la fonction indicatrice de $K \times I^n$. Pour tout $K \in \mathcal{T}$ et $1 \leq n \leq N$, on a :

$$\|\phi - \phi_{ht}\|_{K \times I^n} \leq C^P \left(h_K^2 \|\nabla \phi\|_{K \times I^n}^2 + (\delta t^n)^2 \|\partial_t \phi\|_{K \times I^n}^2 \right)^{1/2}. \tag{3.20}$$

La constante de Poincaré vaut ici $C^P = \dfrac{1}{\pi}$.

On présente maintenant deux estimations qui s'appuient sur l'hypothèse centrale (3.18), la première écrite sur tout l'intervalle de temps $[0, t^n]$, pour $1 \leq n \leq N$, la seconde localisée sur chaque intervalle de temps $[t^{k-1}, t^k]$, pour $1 \leq k \leq n$.

Théorème 3.6. *Soit $f_{ht} \in L^2(Q_T)$ et $\psi_{ht} \in X$. Soit θ_{ht} et $\mathbf{t}_{ht}$ des flux équilibrés pour le terme source f_{ht}. Alors, pour tout $1 \leq n \leq N$, on a les estimations suivantes :*

$$\mathcal{E}_{glob}^n(\psi_{ht}) \leq \eta_{res,glob}^n + \eta_{f,glob}^n + \eta_{\theta,glob}^n + \eta_{\mathbf{t},glob}^n + \eta_{BD,glob}^n + \eta_{IC}$$

et

$$\mathcal{E}_{loc}^n(\psi_{ht}) \leq \eta_{res,loc}^n + \eta_{f,loc}^n + \eta_{\theta,loc}^n + \eta_{\mathbf{t},loc}^n + \eta_{BD,loc}^n$$

avec

$$\eta_{\bullet,glob}^n := \left(\sum_{k=1}^{n} \sum_{K \in \mathcal{T}} (\eta_{\bullet,K}^k)^2 \right)^{\frac{1}{2}} \quad \eta_{BD,glob}^n := \left(\sum_{k=1}^{n} \sum_{\sigma \subset \partial\Omega^N} (\eta_{BD,\sigma}^k)^2 \right)^{\frac{1}{2}}$$

$$\eta_{\bullet,loc}^n := \left(\sum_{K \in \mathcal{T}} (\eta_{\bullet,K}^n)^2 \right)^{\frac{1}{2}} \quad \eta_{BD,loc}^n := \left(\sum_{\sigma \subset \partial\Omega^N} (\eta_{BD,\sigma}^n)^2 \right)^{\frac{1}{2}} \quad \eta_{IC} := \left(\sum_{K \in \mathcal{T}} (\eta_{IC,K})^2 \right)^{\frac{1}{2}}$$

où $\bullet \in \{res, f, \theta, \mathbf{t}\}$*, et pour* $\sigma \subset \partial\Omega^{\mathrm{N}}$*,* $K \in \mathcal{T}$ *et* $1 \leq k \leq n$*, les estimateurs locaux sont définis par :*

$$
\begin{aligned}
\eta^k_{res,K} &:= C^P \|f_{ht} - \partial_t \theta_{ht} - \operatorname{div}(\mathbf{t}_{ht})\|_{K\times I^k},\\
\eta^k_{f,K} &:= \|f - f_{ht}\|_{K\times I^k},\\
\eta^k_{\theta,K} &:= (\delta t^k)^{-1} \|\theta(\psi_{ht}) - \theta_{ht}\|_{K\times I^k},\\
\eta^k_{\mathbf{t},K} &:= h_K^{-1} \| - \mathbb{K}(\psi_{ht})\nabla(\psi_{ht} + z) - \mathbf{t}_{ht}\|_{K\times I^k},\\
\eta^k_{BD,\sigma} &:= \left(C^{Agm}_{\sigma,K}\right)^{1/2} h_K^{-1/2} \|g - \mathbf{t}_{ht}\cdot n_{\sigma,K}\|_{\sigma\times I^k},\\
\eta_{IC,K} &:= (\delta t^1)^{-1/2}\sqrt{2}\|\theta(\psi^0) - \theta_{ht}(\cdot,0)\|_K.
\end{aligned}
$$

On utilise ici les notations $\|\cdot\|_{K\times I^k} = \|\cdot\|_{L^2(K\times I^k)}$*,* $\|\cdot\|_K = \|\cdot\|_{L^2(K)}$*, et* $\|\cdot\|_{\sigma\times I^k} = \|\cdot\|_{L^2(\sigma\times I^k)}$*.*

Preuve. En partant de la définition (3.15) de l'erreur globale à l'instant t^n, on a, pour toute fonction test ϕ dans Y, à support compact dans $[0, t^n[:= I_0^n$ et vérifiant $\|\phi\|_Y = 1$:

$$
\begin{aligned}
\langle R(\psi_{ht}), \phi\rangle &= \langle R(\psi_{ht}), \phi\rangle + \int_{\Omega\times I_0^n} f_{ht}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} f_{ht}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&+ \int_{\Omega\times I_0^n} \theta_{ht}\partial_t\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} \theta_{ht}\partial_t\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\Omega\times I_0^n} \mathbf{t}_{ht}\nabla\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} \mathbf{t}_{ht}\nabla\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&= \int_{\Omega\times I_0^n} f\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\Omega\times I_0^n} \theta(\psi_{ht})\partial_t\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} \mathbb{K}(\psi_{ht})\nabla(\psi_{ht} + z)\nabla\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&- \int_{\partial\Omega\times I_0^n} g\phi \,\mathrm{d}s\,\mathrm{d}t + \int_\Omega \theta(\psi^0)\phi(\mathbf{x},0) \,\mathrm{d}\mathbf{x} + \int_{\Omega\times I_0^n} f_{ht}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} f_{ht}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&- \int_{\Omega\times I_0^n} \partial_t\theta_{ht}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_\Omega \theta_{ht}(\mathbf{x},0)\phi(\mathbf{x},0) \,\mathrm{d}\mathbf{x} - \int_{\Omega\times I_0^n} \theta_{ht}\partial_t\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&- \int_{\Omega\times I_0^n} \operatorname{div}(\mathbf{t}_{ht})\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\partial\Omega\times I_0^n} \mathbf{t}_{ht}\cdot\nu\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{\Omega\times I_0^n} \mathbf{t}_{ht}\nabla\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t,
\end{aligned}
$$

où on a pu utiliser le théorème de Green puisque $\mathbf{t}_{ht} \in L^2(0,t^n;H(\mathrm{div},\Omega))$, et intégrer par parties en temps puisque $\partial_t\theta_{ht} \in L^2(\Omega\times I_0^n)$. On réorganise les termes pour obtenir :

$$\begin{aligned}\langle R(\psi_{ht}),\phi\rangle &= \int_{\Omega\times I_0^n}(f_{ht}-\partial_t\theta_{ht}-\mathrm{div}(\mathbf{t}_{ht}))\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\Omega\times I_0^n}(f-f_{ht})\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &+\int_{\Omega\times I_0^n}(\theta(\psi_{ht})-\theta_{ht})\partial_t\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{\Omega\times I_0^n}(-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht}+z)-\mathbf{t}_{ht})\nabla\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t\\ &-\int_{\partial\Omega\times I_0^n}(g-\mathbf{t}_{ht}\cdot\nu)\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_\Omega(\theta(\psi^0)-\theta_{ht}(\mathbf{x},0))\phi(\mathbf{x},0)\ \mathrm{d}\mathbf{x}.\end{aligned}$$

Finalement, on peut écrire

$$\langle R(\psi_{ht}),\phi\rangle = \sum_{k=1}^n\sum_{K\in\mathcal{T}}\left(R_{\mathrm{res}}^k+R_f^k+R_\theta^k+R_{\mathbf{t}}^k\right)+\sum_{k=1}^n\sum_{\sigma\subset\partial\Omega}R_{\mathrm{bd}}^k+\sum_{K\in\mathcal{T}}R_{\mathrm{ic}},$$

où l'on distingue les différentes contributions :

- du résidu, $R_{\mathrm{res}}^k = \displaystyle\int_{K\times I^k}(f_{ht}-\partial_t\theta_{ht}-\mathrm{div}(\mathbf{t}_{ht}))\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t$;
- du terme source, $R_f^k = \displaystyle\int_{K\times I^k}(f-f_{ht})\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t$;
- du flux en temps, $R_\theta^k = \displaystyle\int_{K\times I^k}(\theta(\psi_{ht})-\theta_{ht})\partial_t\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t$;
- du flux en espace, $R_{\mathbf{t}}^k = \displaystyle\int_{K\times I^k}(-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht}+z)-\mathbf{t}_{ht})\nabla\phi\ \mathrm{d}\mathbf{x}\,\mathrm{d}t$;
- de la condition de Neumann, $R_{\mathrm{bd}}^k = -\displaystyle\int_{\sigma\times I^k}(g-\mathbf{t}_{ht}\cdot n_{\sigma,K})\phi\ \mathrm{d}s\,\mathrm{d}t$;
- de la condition initiale, $R_{\mathrm{ic}} = \displaystyle\int_K(\theta(\psi^0)-\theta_{ht}(\mathbf{x},0))\phi(\mathbf{x},0)\ \mathrm{d}\mathbf{x}$.

Dans le cas où l'on s'intéresse à l'erreur locale $\mathcal{E}_{\mathrm{loc}}^n(\psi_{ht})$, on restreint le support de la fonction test ϕ à un compact inclus dans l'intervalle $]t^{n-1},t^n[$. Dans ce cas, le terme provenant de la condition initiale disparaît, de même que la sommation sur les itérations en temps; on trouve :

$$\langle R(\psi_{ht}),\phi\rangle = R_{\mathrm{res}}^n+R_f^n+R_\theta^n+R_{\mathbf{t}}^n+R_{\mathrm{bd}}^n.$$

On va maintenant estimer chacune de ces erreurs locales. Pour contrôler R_{res}^k, on utilise le fait que θ_{ht} et $\mathbf{t}_{ht}$ sont équilibrés pour la fonction f_{ht} (voir la définition 3.3). Ainsi, en notant ϕ_{ht} la projection L^2-orthogonale définie dans le lemme 3.5, on a :

$$\int_{K\times I^k}(f_{ht}-\partial_t\theta_{ht}-\mathrm{div}(\mathbf{t}_{ht}))\phi_{ht}\ \mathrm{d}\mathbf{x}\,\mathrm{d}t = 0,$$

puisque la fonction ϕ_{ht} est constante sur $K \times I^k$. On peut alors appliquer l'inégalité de Poincaré (3.20) pour obtenir :

$$\begin{aligned} R^k_{\mathrm{res}} &= \int_{K\times I^k} (f_{ht} - \partial_t\theta_{ht} - \mathrm{div}(\mathbf{t}_{ht}))(\phi - \phi_{ht}) \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \\ &\leq \|f_{ht} - \partial_t\theta_{ht} - \mathrm{div}(\mathbf{t}_{ht})\|_{K\times I^k} C^{\mathrm{P}} \left(h_K^2\|\nabla\phi\|^2_{K\times I^k} + (\delta t^k)^2\|\partial_t\phi\|^2_{K\times I^k}\right)^{1/2} \\ &\leq \eta^k_{\mathrm{res},K}\|\phi\|_{Y,K\times I^k}. \end{aligned}$$

Pour les trois termes suivants, l'inégalité de Cauchy-Schwarz donne directement, grâce à la définition de la norme locale sur l'espace Y, $\|\cdot\|_{Y,K\times I^k}$:

$$\begin{aligned} R^k_f &\leq \|f - f_{ht}\|_{K\times I^k}\|\phi\|_{K\times I^k} \leq \eta^k_{f,K}\|\phi\|_{Y,K\times I^k}, \\ R^k_\theta &\leq \|\theta(\psi_{ht}) - \theta_{ht}\|_{K\times I^k}\|\partial_t\phi\|_{K\times I^k} \leq \eta^k_{\theta,K}\|\phi\|_{Y,K\times I^k}, \\ R^k_{\mathbf{t}} &\leq \| - \mathbb{K}(\psi_{ht})\nabla(\psi_{ht} + z) - \mathbf{t}_{ht}\|_{K\times I^k}\|\nabla\phi\|_{K\times I^k} \leq \eta^k_{\mathbf{t},K}\|\phi\|_{Y,K\times I^k}. \end{aligned}$$

Pour le terme de bord, on utilise l'inégalité dite de Agmon (voir [5] pour une preuve) :

$$\begin{aligned} \|\phi\|^2_\sigma &\leq C^{\mathrm{Agm}}_{\sigma,K}\left(h_K^{-1}\|\phi\|^2_K + h_K\|\nabla\phi\|^2_K\right), \\ &= C^{\mathrm{Agm}}_{\sigma,K}h_K^{-1}\left(\|\phi\|^2_K + h_K^2\|\nabla\phi\|^2_K\right), \end{aligned}$$

pour toute arête σ d'un triangle K de $\mathcal{T}$ donné. La constante $C^{\mathrm{Agm}}_{\sigma,K}$ dépend en général de la forme de l'élément K, et peut être bornée par $|\sigma|h_K/|K|$ (voir [36]). Après intégration en temps, on en déduit une version espace-temps de l'inégalité de Agmon :

$$\begin{aligned} \|\phi\|^2_{\sigma\times I^k} &\leq C^{\mathrm{Agm}}_{\sigma,K}h_K^{-1}\left(\|\phi\|^2_{K\times I^k} + h_K^2\|\nabla\phi\|^2_{K\times I^k}\right), \\ &\leq C^{\mathrm{Agm}}_{\sigma,K}h_K^{-1}\|\phi\|^2_{Y,K\times I^k}. \end{aligned}$$

La contribution du terme de bord s'écrit donc :

$$R^k_{\mathrm{bd}} \leq \|g - \mathbf{t}_{ht}\cdot n_{\sigma,K}\|_{\sigma\times I^k}\|\phi\|_{\sigma\times I^k} \leq \eta^k_{\mathrm{BD},\sigma}\|\phi\|_{Y,K\times I^k}.$$

Il reste à estimer la contribution due à la condition initiale, pour laquelle on utilise une version unidimensionnelle de l'inégalité de Agmon. En partant de la relation

$$\phi(\cdot,0) = \phi(\cdot,t) - \int_0^t \partial_t\phi(\cdot,s)\,\mathrm{d}s, \quad \text{on écrit :}$$

$$\frac{1}{2}|\phi(\cdot,0)|^2 \leq |\phi(\cdot,t)|^2 + \left|\int_0^t \partial_t\phi(\cdot,s)\,\mathrm{d}s\right|^2 \leq |\phi(\cdot,t)|^2 + t\int_0^t |\partial_t\phi(\cdot,s)|^2\,\mathrm{d}s,$$

où on a utilisé l'inégalité de Cauchy-Schwarz sur la deuxième intégrale. Puis, on intègre cette dernière équation pour t dans l'intervalle $[0, t^1]$, ce qui donne :

$$\begin{aligned}\frac{1}{2}|\phi(\cdot,0)|^2\delta t^1 &\leq \int_0^{t^1} |\phi(\cdot,t)|^2 \,\mathrm{d}t + \int_0^{t^1}\int_0^t t|\partial_t\phi(\cdot,s)|^2 \,\mathrm{d}s\,\mathrm{d}t \\ &\leq \int_0^{t^1} |\phi(\cdot,t)|^2 \,\mathrm{d}t + \int_0^{t^1} |\partial_t\phi(\cdot,s)|^2 \int_s^{t^1} t\,\mathrm{d}t\,\mathrm{d}s \\ &\leq \int_0^{t^1} |\phi(\cdot,t)|^2 \,\mathrm{d}t + \frac{(\delta t^1)^2}{2}\int_0^{t^1} |\partial_t\phi(\cdot,s)|^2 \,\mathrm{d}s,\end{aligned}$$

et, finalement,

$$\|\phi(\cdot,0)\|_K^2 \leq 2\left((\delta t^1)^{-1}\|\phi\|_{K\times I^1}^2 + \delta t^1\|\partial_t\phi\|_{K\times I^1}^2\right) \leq 2(\delta t^1)^{-1}\|\phi\|_{Y,K\times I^1}^2.$$

Ainsi, la contribution provenant de la condition intiale est :

$$R_{\mathrm{ic}} \leq \|\theta(\psi^0) - \theta_{ht}(\cdot,0)\|_K\|\phi(\cdot,0)\|_K \leq \eta_{\mathrm{IC},K}\|\phi\|_{Y,K\times I^1}.$$

En rassemblant les résultats précédents, on obtient enfin :

$$\begin{aligned}\langle R(\psi_{ht}),\phi\rangle \leq \sum_{k=1}^n \sum_{K\in\mathcal{T}} \left(\eta_{\mathrm{res},K}^k + \eta_{f,K}^k + \eta_{\theta,K}^k + \eta_{\mathbf{t},K}^k\right)\|\phi\|_{Y,K\times I^k} \\ + \sum_{k=1}^n \sum_{\sigma\subset\partial\Omega} \eta_{\mathrm{BD},\sigma}\|\phi\|_{Y,K\times I^k} + \sum_{K\in\mathcal{T}} \eta_{\mathrm{IC},K}\|\phi\|_{Y,K\times I^1}\end{aligned}$$

si la fonction test ϕ est à support compact dans $[0, t^n[$, et

$$\langle R(\psi_{ht}),\phi\rangle \leq \sum_{K\in\mathcal{T}} \left(\eta_{\mathrm{res},K}^n + \eta_{f,K}^n + \eta_{\theta,K}^n + \eta_{\mathbf{t},K}^n\right)\|\phi\|_{Y,K\times I^n} + \sum_{\sigma\subset\partial\Omega} \eta_{\mathrm{BD},\sigma}\|\phi\|_{Y,K\times I^n}$$

si la fonction test ϕ est à support compact dans $]t^{n-1}, t^n[$.
On applique enfin une dernière fois l'inégalité de Cauchy-Schwarz; par exemple, si ϕ est

à support compact dans $[0, t^n[$, cela donne (on rappelle que $\|\phi\|_Y = 1$) :

$$\begin{aligned}
\langle R(\psi_{ht}), \phi\rangle \leq & \sum_{\bullet} \left(\sum_{k=1}^{n} \sum_{K\in\mathcal{T}} \left(\eta^k_{\bullet,K}\right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{n} \sum_{K\in\mathcal{T}} \|\phi\|^2_{Y,K\times I^k} \right)^{\frac{1}{2}} \\
& + \left(\sum_{k=1}^{n} \sum_{\sigma\subset\partial\Omega^{\mathrm{N}}} \left(\eta^k_{\mathrm{BD},\sigma}\right)^2 \right)^{\frac{1}{2}} \left(\sum_{k=1}^{n} \sum_{K\in\mathcal{T}} \|\phi\|^2_{Y,K\times I^k} \right)^{\frac{1}{2}} \\
& + \left(\sum_{K\in\mathcal{T}} \left(\eta_{\mathrm{IC},K}\right)^2 \right)^{\frac{1}{2}} \left(\sum_{K\in\mathcal{T}} \|\phi\|^2_{Y,K\times I^1} \right)^{\frac{1}{2}} \\
\leq & (\eta^n_{\mathrm{res,glob}} + \eta^n_{\theta,\mathrm{glob}} + \eta^n_{\mathbf{t},\mathrm{glob}} + \eta^n_{f,\mathrm{glob}} + \eta^n_{\mathrm{BD,glob}} + \eta_{\mathrm{IC}}) \|\phi\|_Y \\
\leq & \eta^n_{\mathrm{res,glob}} + \eta^n_{\theta,\mathrm{glob}} + \eta^n_{\mathbf{t},\mathrm{glob}} + \eta^n_{f,\mathrm{glob}} + \eta^n_{\mathrm{BD,glob}} + \eta_{\mathrm{IC}},
\end{aligned}$$

où $\bullet \in \{\mathrm{res}, f, \theta, \mathbf{t}\}$. Le résultat du théorème suit. □

Remarque 4. *L'estimateur de flux temporel $\eta^n_{\theta,loc}$ quantifie la violation de la relation définissant la teneur en eau sur l'intervalle I^n; l'estimateur de flux spatial $\eta^n_{\mathbf{t},loc}$ concerne quant à lui le non-respect de la loi de Darcy. L'estimateur de résidu $\eta^n_{res,loc}$ mesure l'écart vis-à-vis de l'équation de conservation de la masse. L'estimateur de terme source $\eta^n_{f,loc}$ apparaît lorsque f n'est pas polynomial par morceaux en espace et en temps, et enfin l'estimateur de bord $\eta^n_{BD,loc}$ provient de la condition de Neumann.*

3.3.4 Estimation pour le problème linéarisé

En réalité, le schéma espace-temps posé sur le domaine discret $\Omega \times [0, T]$ produit un système non linéaire, c'est pourquoi on fait appel à des linéarisations. Souhaitant tenir compte dans nos estimations des erreurs dues à ces linéarisations, on modifie légèrement notre hypothèse d'équilibrage.

Définition 3.7. Étant donné des fonctions $f_{ht},\ \delta\theta$ et $\delta\mathbf{t}$ dans $L^2(Q_T)$, on dit que les fonctions $\theta_{ht} : Q_T \to \mathbb{R}$ et $\mathbf{t}_{ht} : Q_T \to \mathbb{R}^2$ sont des flux équilibrés pour le terme source f_{ht} et les fonctions $\delta\theta$ et $\delta\mathbf{t}$ si :

$$\theta_{ht} \in L^2(Q_T),\ \partial_t\theta_{ht} \in L^2(Q_T),\ \mathbf{t}_{ht} \in L^2([0,T], H(\mathrm{div}, \Omega)),$$

et

$$\forall\ K \in \mathcal{T},\ \forall\ 1 \leq n \leq N, \quad \int_{K\times I^n} \{f_{ht} - \partial_t\theta_{ht} - \mathrm{div}(\mathbf{t}_{ht}) + \delta\theta + \delta\mathbf{t}\}\ \mathrm{d}\mathbf{x}\,\mathrm{d}t = 0. \quad (3.21)$$

On étend ainsi aisément le lien établi dans le théorème 3.4 avec les schémas linéarisés de la forme (3.14). On écrit alors une nouvelle estimation qui intègre ces changements. Bien que la borne globale reste valable, on se contente d'énoncer la borne localisée à l'instant t^n, puisque c'est elle que l'on utilisera en pratique.

Théorème 3.8. *Soit $f_{ht} \in L^2(Q_T)$, $\psi_{ht} \in X$, $\delta\theta$ et $\delta\mathbf{t}$ des fonctions telles que définies dans 3.7. Soit θ_{ht} et $\mathbf{t}_{ht}$ des flux équilibrés pour f_{ht}, $\delta\theta$ et $\delta\mathbf{t}$. Alors, pour tout $1 \leq n \leq N$, on a l'estimation suivante :*

$$\mathcal{E}^n(\psi_{ht}) \leq \eta^n_{res} + \eta^n_f + \eta^n_\theta + \eta^n_{\mathbf{t}} + \eta^n_{BD} + \eta^n_{\theta_{\text{lin}}} + \eta^n_{\mathbf{t}_{lin}},$$

avec

$$\eta^n_\bullet := \left(\sum_{K\in\mathcal{T}} (\eta^n_{\bullet,K})^2\right)^{\frac{1}{2}} \qquad \eta^n_{BD} := \left(\sum_{\sigma\subset\partial\Omega^{\text{N}}} (\eta^n_{BD,\sigma})^2\right)^{\frac{1}{2}},$$

où $\bullet \in \{res, f, \theta, \mathbf{t}, \theta_{lin}, \mathbf{t}_{lin}\}$, et pour $\sigma \subset \partial\Omega$, $K \in \mathcal{T}$ et $1 \leq k \leq n$, les estimateurs locaux sont définis par :

$$\begin{aligned}
\eta^n_{res,K} &:= C^P \|f_{ht} - \partial_t\theta_{ht} - \operatorname{div}(\mathbf{t}_{ht}) + \delta\theta + \delta\mathbf{t}\|_{K\times I^n},\\
\eta^n_{f,K} &:= \|f - f_{ht}\|_{K\times I^n},\\
\eta^n_{\theta,K} &:= (\delta t^n)^{-1}\|\theta(\psi_{ht}) - \theta_{ht}\|_{K\times I^n},\\
\eta^n_{\mathbf{t},K} &:= h_K^{-1}\| - \mathbb{K}(\psi_{ht})\nabla(\psi_{ht} + z) - \mathbf{t}_{ht}\|_{K\times I^n},\\
\eta^n_{BD,\sigma} &:= \left(C^{Agm}_{\sigma,K}\right)^{1/2} h_K^{-1/2}\|g - \mathbf{t}_{ht}\cdot n_{\sigma,K}\|_{\sigma\times I^n},\\
\eta^n_{\theta_{\text{lin,K}}} &:= \|\delta\theta\|_{K\times I^n}, \quad \eta^n_{\mathbf{t}_{lin,K}} := \|\delta\mathbf{t}\|_{K\times I^n}.
\end{aligned}$$

Preuve. La démonstration suit exactement les mêmes étapes que celle décrite dans la sous-section 3.3.3. On ajoute également les termes

$$\int_{K\times I^n} \delta\theta\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{K\times I^n} \delta\theta\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t + \int_{K\times I^n} \delta\mathbf{t}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t - \int_{K\times I^n} \delta\mathbf{t}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t,$$

et l'estimation du terme R^n_{res} s'écrit donc cette fois :

$$\begin{aligned}
R^n_{\text{res}} &= \int_{K\times I^n} (f_{ht} - \partial_t\theta_{ht} - \operatorname{div}(\mathbf{t}_{ht}) + \delta\theta + \delta\mathbf{t})(\phi - \phi_{ht}) \,\mathrm{d}\mathbf{x}\,\mathrm{d}t\\
&\leq \eta^n_{\text{res},K}\|\phi\|_{Y,K\times I^n}.
\end{aligned}$$

Apparaissent également les nouvelles contributions dues aux linéarisations :

$$R^n_{\theta,\mathrm{lin}} = -\int_{K\times I^n} \delta\theta\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq \eta^n_{\theta_{\mathrm{lin},K}} \|\phi\|_{Y,K\times I^n}, \quad \text{et} \quad R^n_{\mathbf{t},\mathrm{lin}} = -\int_{K\times I^n} \delta\mathbf{t}\phi \,\mathrm{d}\mathbf{x}\,\mathrm{d}t \leq \eta^n_{\mathbf{t}_{\mathrm{lin},K}} \|\phi\|_{Y,K\times I^n}.$$

Le reste de la démonstration est identique. □

Remarque 5. *En pratique, c'est cette estimation qui sera utilisée pour élaborer notre stratégie d'adaptation, à chaque itération de linéarisation m d'un instant t^n.*

3.4 Choix de reconstructions pour le schéma DDFV-BDF2

L'estimation que l'on vient d'écrire pour une fonction ψ_{ht} quelconque dans l'espace X s'appuie sur la relation clé d'équilibrage des flux (3.21). Cette borne est valable dès qu'on a pu trouver des fonctions θ_{ht} et $\mathbf{t}_{ht}$ vérifiant les hypothèses de la définition 3.7. En pratique cependant, on cherche à estimer l'erreur commise par la solution approchée, et il est naturel de souhaiter que la borne obtenue soit proche de l'erreur $\mathcal{E}^n(\psi_{ht})$, où la fonction ψ_{ht} de X est reconstruite à partir de ladite solution approchée. Sa pertinence dépend de la précision des estimateurs, et donc du choix des reconstructions ψ_{ht}, θ_{ht} et $\mathbf{t}_{ht}$. En particulier, les fonctions θ_{ht} et $\mathbf{t}_{ht}$ doivent approcher précisément $\theta(\psi_{ht})$ et $-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht}+z)$ respectivement, de manière que les différents estimateurs quantifient bien les erreurs attendues (voir la remarque 4). En ce sens, il est pertinent de faire le lien entre l'équation d'équilibrage (3.21) et l'expression du schéma, comme on l'a vu dans le théorème 3.4.

On indique dans cette section des exemples de reconstructions, élaborées en accord avec le théorème 3.4, pour la discrétisation DDFV en espace vue au chapitre 2, associée à un schéma en temps de la forme (3.13). Il est important de noter que dans le cas du schéma DDFV, des équations sont écrites sur le maillage primaire $\mathcal{T}$ ainsi que sur un maillage secondaire $\mathcal{S}$ associé. Ici, on ne tient pas compte de ces équations supplémentaires, et on se contente donc d'écrire des estimations sur le maillage $\mathcal{T}$.

Dans chaque cas, on définit d'abord des reconstructions en espace, indexées par le pas du maillage h; puis on explicite des fonctions en temps qui s'appuient sur ces reconstructions en espace. On obtient des fonctions reconstruites en espace et en temps, indexées par ht. En pratique, à chaque itération de linéarisation m d'un instant discret t^n, on intègre ces fonctions sur l'intervalle de temps I^n, donc on se contente de définir leur restriction sur un tel intervalle. Dans la suite, on note $\mu_K(\zeta)$ la valeur moyenne sur l'élément K d'une fonction $\zeta \in L^1(K)$, $\mu_K(\zeta) := \dfrac{1}{|K|}\displaystyle\int_K \zeta(\mathbf{x})\,\mathrm{d}\mathbf{x}$.

3.4.1 Reconstructions de la charge

Pour le schéma DDFV, les degrés de liberté de l'inconnue principale Ψ sont situés aux sommets ainsi qu'aux centres de masse du maillage primaire :

$$\Psi = ((\psi_K)_{K\in\mathcal{T}}, (\psi_A)_{A\in\mathcal{S}^{\mathrm{I}}}) \in \mathbb{R}^{|\mathcal{T}|+|\mathcal{S}^{\mathrm{I}}|},$$

avec les notations du chapitre 2. On a toute latitude *a priori* pour reconstruire une fonction solution ψ_h à partir de ce vecteur de réels, représentant dans le cadre de notre problème la charge hydraulique. Cependant, la qualité de notre estimation dépend fortement de ce choix. Dans le cas du schéma DDFV, on propose ici deux manières de procéder, qui aboutissent à des fonctions conformes, $\psi_h \in H^1(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$. Elles seront comparées numériquement dans le chapitre 4.

Reconstruction $\mathbb{P}1$-Lagrange par demi-diamant Une première reconstruction naturelle et peu coûteuse consiste à définir ψ_h comme la fonction affine par demi-diamant $D_{\sigma,K}$ qui interpole les valeurs de la charge aux sommets de $D_{\sigma,K}$, c'est-à-dire :

$$\forall\ D_{\sigma,K} \in \mathcal{D}, \quad \psi_h|_{D_{\sigma,K}} = \sum_{P\in\{A,B,K\}} \psi_P \lambda_{\sigma,K,\mathbf{x}_P},$$

où $\{\lambda_{\sigma,K,\mathbf{x}_P}\}$ est la fonction nodale sur $D_{\sigma,K} = \mathbf{x}_A\mathbf{x}_B\mathbf{x}_K$ qui vaut 1 au sommet $\mathbf{x}_P$ et 0 aux deux autres sommets de $D_{\sigma,K}$. L'inconvénient de cette reconstruction est que son gradient sur chaque demi-diamant $D_{\sigma,K}$ ne vaut pas le gradient numérique, qui fait notamment intervenir la valeur ψ_σ qui assure la continuité des flux normaux.

Reconstruction $\mathbb{P}1$-Lagrange par quart de diamant Une autre idée est de découper chaque demi-diamant en deux quarts de diamant, séparés par la médiane issue du centre du volume de contrôle correspondant (voir la figure 3.1). La fonction ψ_h est alors choisie comme la fonction affine sur chaque quart de diamant qui interpole les valeurs de la charge aux sommets de ce quart de diamant. L'approximation du gradient numérique est meilleure, puisque la moyenne des gradients de ψ_h sur chaque demi-diamant coïncide avec le gradient DDFV.
À un instant t^n et une itération m de la procédure de linéarisation, on peut donc reconstruire en espace une fonction $\psi_h^{n,m}$. On définit alors la reconstruction espace-temps ψ_{ht} affine en temps, interpolant les reconstructions en espace obtenues en t^{n-1}et en t^n :

$$\forall\ 1 \le n \le N, \quad \psi_{ht}^m|_{I^n}(t) = \rho^n(t)\psi_h^{n,m} + (1-\rho^n(t))\psi_h^{n-1}, \quad \text{où} \quad \rho^n(t) := \frac{t-t^{n-1}}{\delta t^n}.$$

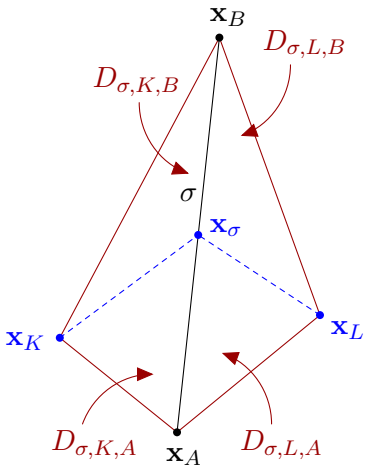


FIGURE 3.1: Découpage en quarts de diamant.

3.4.2 Reconstructions de la teneur en eau

Soit $\hat{\theta}_h$ une fonction polynomiale par morceaux en espace satisfaisant les conditions :

$$\forall\, D_{\sigma,K} \in \mathcal{D}, \quad \forall\, 1 \leq i \leq N_{\mathrm{p}}, \quad \hat{\theta}_h|_{D_{\sigma,K}}(\mathbf{x}_i^*) = \theta(\psi_h)|_{D_{\sigma,K}}(\mathbf{x}_i^*), \tag{3.22}$$

où les $\{\mathbf{x}_i^*\}_{1\leq i\leq N_{\mathrm{p}}}$ sont les N_{p} points d'intégration sur $D_{\sigma,K}$ servant au calcul des estimateurs.

Définition 3.9. La reconstruction en espace $\theta_h \in L^2(\Omega)$ de la teneur en eau est la somme de la fonction polynomiale $\hat{\theta}_h$ définie par (3.22) et d'une fonction bulle assurant la condition de moyenne $\mu_K(\theta_h) = \theta(\psi_K)$:

$$\forall\, K \in \mathcal{T}, \quad \theta_h|_K := \hat{\theta}_h|_K + \beta_K \lambda_K, \quad \text{avec} \quad \beta_K := \frac{\theta(\psi_K) - \mu_K(\hat{\theta}_h)}{\mu_K(\lambda_K)},$$

où $\lambda_K := \lambda_{K,\mathbf{x}_A}\lambda_{K,\mathbf{x}_B}\lambda_{K,\mathbf{x}_C}$ et $\{\lambda_{K,\mathbf{x}_P}\}$ désigne chaque fonction nodale sur K.

On a ainsi clairement : $\mu_K(\theta_h) = \theta(\psi_K)$. Dans la définition 3.9, le choix de la fonction $\hat{\theta}_h$ ne nous contraint pas à calculer le polynôme entier, puisque seules les valeurs situées aux points d'intégration sont nécessaires pour évaluer les estimateurs η_θ et $\eta_{\theta_{\mathrm{lin}}}$.
Pour la reconstruction en temps, on fait de même que pour la charge reconstruite ψ_{ht} :

$$\forall\, 1 \leq n \leq N, \quad \theta_{ht}^m|_{I^n}(t) = \rho^n(t)\theta_h^{n,m} + (1-\rho^n(t))\theta_h^{n-1}.$$

3.4.3 Reconstructions du flux spatial

Le flux spatial $\mathbf{t}_h$ est construit dans un espace de Raviart-Thomas-Nédélec [15] : on définit $\mathbb{RTN}_l(\mathcal{T}) := \{\zeta \in H(\mathrm{div},\Omega) \mid \forall K \in \mathcal{T},\ \zeta|_K \in \mathbb{RTN}_l(K)\}$, avec $\mathbb{RTN}_l(K) := [\mathbb{P}_l(K)]^2 + \mathbf{x}\tilde{\mathbb{P}}_l(K)$, où $\mathbb{P}_l$ sont les polynômes de degré au plus égal à l et $\tilde{\mathbb{P}}_l$ est l'espace des polynômes homogènes de degré l.
On introduit maintenant le flux numérique approchant la vitesse continue $\mathbf{v}(\psi) := -\mathbb{K}(\psi)\,(\nabla\psi + e_z)$. Le flux discret $\mathbf{v}_h \in L^2(\Omega)$ associé au schéma DDFV est la fonction constante par morceaux, sur le maillage $\mathcal{D}$ constitué des demi-diamants, et qui est définie par :

$$\forall D_{\sigma,K} \in \mathcal{D}, \quad \mathbf{v}_h|_{D_{\sigma,K}} = -\mathbb{K}(\psi_{\sigma,K})\,(\nabla_{\sigma,K}\Psi + e_z)\,\mathbb{1}_{D_{\sigma,K}}. \tag{3.23}$$

En commettant un léger abus de notation, on identifie dans la suite la fonction $\mathbf{v}_h|_{D_{\sigma,K}}$ et sa valeur constante sur $D_{\sigma,K}$.

Définition 3.10. Le flux reconstruit $\mathbf{t}_h \in \mathbb{RTN}_0(\mathcal{T})$ associé au schéma DDFV est défini de manière unique sur le maillage $\mathcal{T}$ par les conditions surfaciques :

$$\forall K \in \mathcal{T},\ \forall \sigma \in \partial K, \quad \int_\sigma \mathbf{t}_h \cdot n_\sigma \,\mathrm{d}s = \mathbf{v}_h|_{D_{\sigma,K}} \cdot N_\sigma. \tag{3.24}$$

Définition 3.11. Le flux reconstruit $\mathbf{t}_h \in \mathbb{RTN}_1(\mathcal{T})$ associé au schéma DDFV est défini de manière unique sur le maillage $\mathcal{T}$ par les conditions surfaciques (3.24) et les conditions volumiques suivantes :

$$\forall K \in \mathcal{T},\ \forall \sigma \in \partial K, \quad \int_\sigma s\ \mathbf{t}_h \cdot n_\sigma \,\mathrm{d}s = \frac{1}{2}\mathbf{v}_h|_{D_{\sigma,K}} \cdot N_\sigma, \tag{3.25}$$

$$\int_K \mathbf{t}_h \,\mathrm{d}\mathbf{x} = \sum_{\sigma\subset\partial K} |D_{\sigma,K}|\ \mathbf{v}_h|_{D_{\sigma,K}}, \tag{3.26}$$

où s représente l'abscisse curviligne.

Le flux $\mathbf{t}_h$ prend la forme $\mathbf{t}_h(x,z) = (a\ b)^\mathsf{t} + c(x\ z)^\mathsf{t}$ dans $\mathbb{RTN}_0(K)$; ses trois degrés de liberté sont les flux normaux (3.24) sur chaque face, comme illustré sur la figure 3.2b. Il prend la forme $\mathbf{t}_h(x,z) = (ax + bz + c \quad dx + ez + f)^\mathsf{t} + g(x^2\ xz)^\mathsf{t} + h(xz\ z^2)^\mathsf{t}$ dans $\mathbb{RTN}_1(K)$; ses huit degrés de liberté correspondent aux flux normaux (3.24) ainsi qu'à leur premier moment (3.25) sur chaque face, et à deux conditions de moyenne sur chaque élément K. On peut remarquer que la condition (3.26) s'écrit aisément avec le schéma DDFV, grâce au fait que le gradient discret est localisé sur chaque triangle $D_{\sigma,K}$.
Enfin, la reconstruction en temps $\mathbf{t}_{ht}$ s'écrit cette fois :

$$\forall 2 \le n \le N, \quad \mathbf{t}_{ht}^m|_{I^n}(t) = 2\rho^n(t)\mathbf{t}_h^{n,m} + (1 - 2\rho^n(t))\mathbf{t}_{ht}^m(t^{n-1}).$$

3.4.4 Reconstructions du terme source

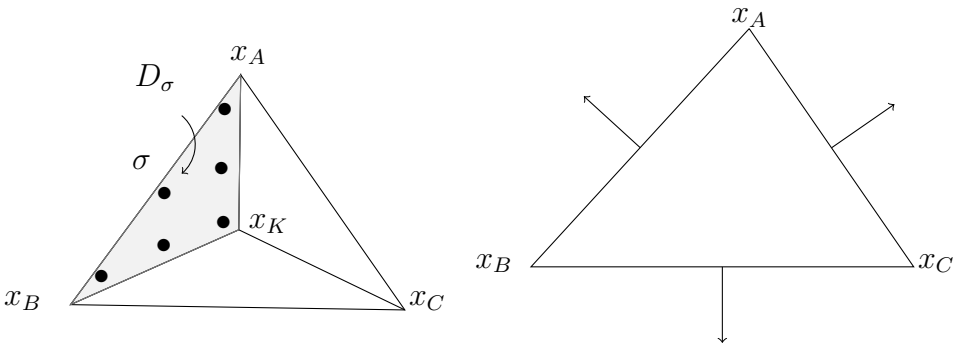
Définition 3.12. La reconstruction en espace $f_h^n \in L^2(\Omega)$ du terme source à l'itération n est la fonction constante par morceaux sur le maillage primaire définie par :

$$\forall K \in \mathcal{T}, \quad f_h^n|_K = f_K^n,$$

où f_K^n est l'approximation de $\mu_K(f(\cdot, t^n))$ utilisée dans le schéma. La reconstruction en temps f_{ht} est alors la fonction constante par morceaux suivante :

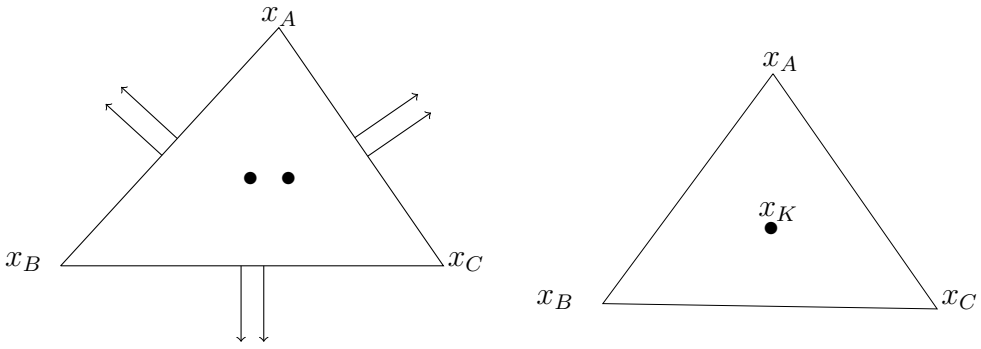
$$\forall\ 1 \leq n \leq N, \quad f_{ht}|_{I^n}(t) = f_h^n.$$

Remarque 6. *Comme on peut le vérifier facilement à l'aide du théorème 3.4, les reconstructions θ_{ht} et $\mathbf{t}_{ht}$ ainsi définies vérifient l'hypothèse d'équilibrage 3.7.*



(A) Teneur en eau (pour $N_{\mathrm{p}} = 6$).

(B) Flux reconstruit dans $\mathbb{RTN}_0(K)$.



(C) Flux reconstruit dans $\mathbb{RTN}_1(K)$.

(D) Terme source.

FIGURE 3.2: Degrés de liberté pour les reconstructions en espace sur un triangle ABC.

Chapitre 4

Algorithme et résultats numériques

On applique les résultats théoriques du chapitre 3 à la mise en place d'une stratégie adaptative. On travaille à maillage fixé, et on construit un algorithme d'adaptation du pas de temps et du nombre d'itérations de linéarisation au cours de la simulation. Pour cela, on forme trois groupes d'estimateurs, qui représentent l'erreur due à la discrétisation en temps, l'erreur due à la discrétisation en espace et l'erreur due aux linéarisations. À chaque temps discret, les itérations de linéarisation sont stoppées lorsque l'erreur de linéarisation devient négligeable devant les deux autres sources d'erreur. Puis, le pas de temps est ajusté de manière à équilibrer l'erreur due à la discrétisation en temps et l'erreur due à la discrétisation en espace. On commence par décrire cette stratégie d'adaptation, avant de tester notre algorithme sur différents cas tests. On reprend tout d'abord un cas test analytique qui nous permet de valider le comportement du pas de temps adapté, dans le cas d'un écoulement à vitesse constante, puis dans le cas d'un écoulement accéléré. On règle alors les paramètres d'entrée de l'algorithme de manière à produire une solution proche de celle obtenue avec une simulation classique, c'est-à-dire sans adaptation. Puis, on teste la robustesse de notre approche sur le cas test de Polmann étudié au chapitre 2, ainsi que sur un cas test modélisant un écoulement en sol hétérogène. On quantifie également le gain en termes de nombre d'itérations de Picard effectuées et de temps CPU obtenu avec notre algorithme d'adaptation sur ces deux simulations.

4.1 Algorithme proposé

On se donne un maillage fixe tout au long de la simulation. On se place en un itéré en temps n, et en une itération de linéarisation m. L'équation (3.8) a fourni une borne supérieure de l'erreur en norme duale, locale en temps (sur l'intervalle $I^n = [t^{n-1}, t^n]$) et en espace (sur chaque maille primaire K). La localité en espace pourrait être mise à profit dans le cadre d'une adaptation de maillage, ce qu'on ne considère pas ici. Cette borne supérieure est la somme de plusieurs estimateurs. La stratégie d'adaptation que l'on propose s'articule en trois étapes :

1. On classe les estimateurs en trois groupes : un groupe estime l'erreur due aux linéarisations; un groupe caractérise l'erreur due à la discrétisation en espace; le dernier quantifie l'erreur due à la discrétisation en temps.
2. On stoppe les itérations de linéarisation lorsque la composante d'erreur correspondante devient négligeable devant les deux autres sources d'erreur.
3. On ajuste le pas de temps utilisé pour l'itéré en temps suivant de manière à équilibrer les composantes d'erreur en temps et en espace.

Classification des estimateurs On souhaite réécrire la borne (3.8) sous la forme :

$$\mathcal{E}^n_{\mathrm{loc}}(\psi^m_{ht}) \leq \eta^{n,m}_{\mathrm{esp}} + \eta^{n,m}_{\mathrm{tps}} + \eta^{n,m}_{\mathrm{lin}}.$$

On ne tient pas compte dans ce chapitre de l'estimateur de bord $\eta^{n,m}_{\mathrm{BD}}$, les cas tests présentés ne comportant pas de condition de Neumann non homogène. La définition de la composante comprenant les estimateurs de linéarisation est immédiate : $\eta^{n,m}_{\mathrm{lin}} := \eta^{n,m}_{\theta_{\mathrm{lin}}} + \eta^{n,m}_{\mathbf{t}_{\mathrm{lin}}}$. En revanche, tous les estimateurs étant intégrés en temps et en espace, la définition de $\eta^{n,m}_{\mathrm{esp}}$ et $\eta^{n,m}_{\mathrm{tps}}$ n'a rien d'évident. Ainsi, les estimateurs comprennent à la fois une erreur due à la discrétisation en temps et une erreur due à la discrétisation en espace. Il paraît cependant naturel d'associer l'estimateur de flux $\eta^{n,m}_{\mathbf{t}}$ à une erreur en espace, et l'estimateur de résidu $\eta^{n,m}_{\mathrm{res}}$ à une erreur en temps (avec notre choix de reconstruction affine en temps de la charge), comme cela a été fait dans [33]. Pour les deux derniers, à savoir $\eta^{n,m}_{\theta}$ et η^n_f, notre choix est ici guidé par le comportement des simulations adaptatives. On qualifiera ainsi de *robuste à la condition initiale* un algorithme qui produit un pas de temps adapté indépendant du pas de temps initial choisi (après un temps de simulation suffisamment long). Or, seul un choix pour les composantes d'erreur spatiale et temporelle permet d'observer cette robustesse à la condition initiale :

$$\eta^{n,m}_{\mathrm{esp}} := \eta^{n,m}_{\theta} + \eta^{n,m}_{\mathbf{t}}, \quad \eta^{n,m}_{\mathrm{tps}} := \eta^{n,m}_{\mathrm{res}} + \eta^n_f.$$

Critère d'arrêt des linéarisations Il est naturel de contraindre l'erreur de linéarisation à être petite devant les deux autres sources d'erreur. C'est particulièrement important dans des cas raides, où un critère d'arrêt des itérations de linéarisation trop lâche peut détériorer de manière significative la solution approchée (voir la sous-section 4.2.2). Ce critère d'arrêt s'écrit ici :

$$\eta_{\text{lin}}^{n,m} \leq \gamma_{\text{lin}}(\eta_{\text{tps}}^{n,m} + \eta_{\text{esp}}^{n,m}), \quad \text{où} \quad \gamma_{\text{lin}} \ll 1. \tag{4.1}$$

Critère d'équilibrage L'autre critère concerne l'équilibrage des erreurs spatiale et temporelle : lorsque le critère d'arrêt de linéarisation est vérifié, à une itération notée m_∞, on considère que le pas de temps δt^n utilisé pour calculer la solution approchée Ψ^n est optimal si $\eta_{\text{tps}}^{n,m_\infty} = \lambda_{\text{eq}} \eta_{\text{esp}}^{n,m_\infty}$, où le *ratio d'équilibre* λ_{eq}, que l'on souhaite proche de 1, est un paramètre défini par l'utilisateur en début de simulation. Bien entendu, cette égalité n'est presque jamais vérifiée en pratique. Ainsi, on accepte la solution approchée Ψ^n, et on passe à l'itéré en temps $n+1$, si le critère d'équilibrage suivant est vérifié :

$$\frac{\eta_{\text{tps}}^{n,m_\infty}}{\eta_{\text{esp}}^{n,m_\infty}} =: \lambda^{n,m_\infty} \leq \lambda^{\max}, \tag{4.2}$$

où $\lambda^{\max} > \lambda_{\text{eq}}$ est le ratio maximal autorisé.

Stratégie d'adaptation À l'issue d'une boucle de linéarisation (stoppée à l'aide du critère d'arrêt (4.1)), la stratégie est la suivante (voir la figure 4.1) :

- si $\lambda^{n,m_\infty} \leq \lambda_{\text{eq}}$, on accepte la solution approchée Ψ^n puisque l'erreur en temps est plus petite que souhaitée. On peut alors déraffiner le pas de temps sur l'intervalle I^{n+1} en choisissant $\delta t^{n+1} \geq \delta t^n$;
- si $\lambda_{\text{eq}} < \lambda^{n,m_\infty} \leq \lambda^{\max}$, on accepte la solution Ψ^n même si l'erreur en temps est plus grande que souhaitée. Cependant, on raffine le pas de temps sur l'intervalle I^{n+1} en choisissant $\delta t^{n+1} < \delta t^n$;
- si $\lambda^{n,m_\infty} > \lambda^{\max}$, on rejette la solution Ψ^n, le critère d'équilibrage (4.2) n'étant pas vérifié. On diminue alors le pas de temps δt^n et on recommence la simulation sur l'intervalle I^n.

À cause de la dernière condition (lorsque la condition d'équilibrage (4.2) n'est pas vérifiée), on a ainsi affaire à une boucle d'équilibrage des erreurs spatiale et temporelle. À une itération $l \geq 1$ de cette boucle d'équilibrage, il s'agit de modifier le pas de temps

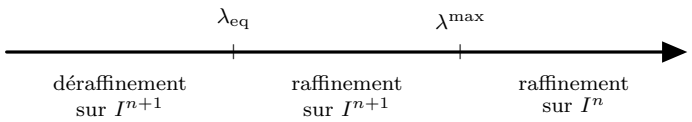


FIGURE 4.1: Stratégie d'adaptation du pas de temps selon la valeur de λ^{n,m_∞}.

courant $\delta t^{n,l}$ (initialisé par $\delta t^{n,0} = \delta t^n$), pour définir le pas de temps $\delta t^{n,l+1}$ (qui deviendra le pas de temps δt^{n+1} si le critère (4.2) est vérifié). On définit ainsi le coefficient de pondération $r^{n,l}$ par la formule suivante :

$$r^{n,l} := \begin{cases} 2 & \text{si } \lambda^{n,m_\infty} < 0.5\lambda_{\text{eq}} \\ -2\lambda^{n,m_\infty}/\lambda_{\text{eq}} + 3 & \text{si } 0.5\lambda_{\text{eq}} \leq \lambda^{n,m_\infty} \leq \lambda_{\text{eq}} \\ -0.5\lambda^{n,m_\infty}/\lambda_{\text{eq}} + 1.5 & \text{si } \lambda_{\text{eq}} \leq \lambda^{n,m_\infty} \leq \lambda^{\max} \\ 0.5 & \text{si } \lambda^{n,m_\infty} > \lambda^{\max} \end{cases}, \tag{4.3}$$

et on pose $\delta t^{n,l+1} := r^{n,l}\delta t^{n,l}$. Dans (4.3), le ratio $r^{n,l}$ vaut 1 lorsque la condition d'équilibrage $\lambda^{n,m_\infty} = \lambda_{\text{eq}}$ est satisfaite, et varie de manière affine par morceaux entre ses valeurs minimale et maximale, respectivement 1/2 et 2 (voir la figure 4.2). On décide ainsi de doubler le pas de temps courant $\delta t^{n,l}$ au maximum (rappelons que la condition de A-stabilité du schéma BDF2 impose $r^{n+1} = \delta t^{n+1}/\delta t^n \leq 1 + \sqrt{2}$), lorsque $\lambda^{n,m_\infty} \leq 0.5\lambda_{\text{eq}}$, et de le diminuer de moitié au minimum, lorsque $\lambda^{n,m_\infty} \geq \lambda^{\max}$. D'autres choix sont bien entendu possibles.

Remarque 7. *Il paraît raisonnable d'imposer une tolérance maximale $\lambda^{max} = 2\lambda_{eq}$. Ainsi, la solution obtenue à l'instant courant est refusée dès que le rapport entre l'erreur temporelle et l'erreur spatiale atteint le double du ratio d'équilibre λ_{eq}, paramètre défini par l'utilisateur. On divise alors le pas de temps courant par 2, et on recommence l'itération en temps concernée. Dans les cas tests de la section 4.2, on se concentrera sur les valeurs prises par la tolérance λ^{max}, étant entendu que le ratio d'équilibre vaut $\lambda_{eq} = 0.5\lambda^{max}$.*

Algorithme d'adaptation On présente maintenant l'algorithme utilisé pour adapter le nombre d'itérations de linéarisation, ainsi que le pas de temps, au cours de la simulation. On omet la dépendance en l des différentes variables et fonctions, qui sont recalculées à chaque nouvelle itération de la boucle d'équilibrage sans garder trace de l'itéré $l-1$; ceci, à l'exception notable du pas de temps $\delta t^{n,l}$ et du ratio $r^{n,l}$. Pour alléger la présentation, on définit les procédures suivantes :

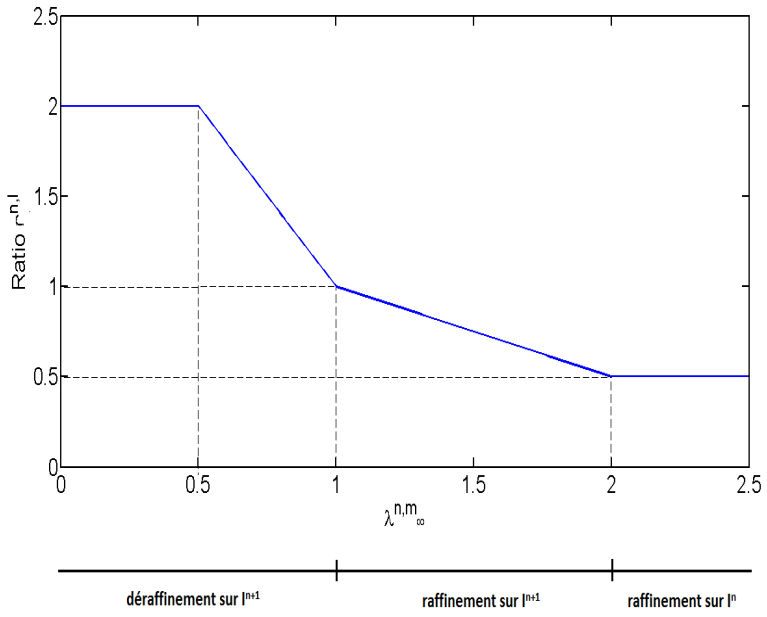


FIGURE 4.2: Valeur du ratio $r^{n,l}$, pour $\lambda_{\text{eq}} = 1$ et $\lambda^{\max} = 2$.

- `Coefficients_BDF2`$(\delta t^{n,l}, \delta t^{n-1})$ fait référence à l'implémentation des coefficients du schéma BDF2 suivant (3.7) avec le ratio $r^n := \delta t^{n,l}/\delta t^{n-1}$. Elle renvoie les coefficients α_0^n, α_1^n et α_2^n suivant les formules (3.8).

- `Richards_itération_NL`$(t^{n-1}, \delta t^{n,l}, \Psi^{n,m-1})$ désigne une résolution du problème discret linéarisé au temps discret t^{n-1} avec le pas de temps $\delta t^{n,l}$. L'argument de sortie est le vecteur discret $\Psi^{n,m}$.

- `Erreurs_lin`$(\Psi^{n,m}, \Psi^{n,m-1})$ calcule les erreurs de linéarisation $\delta\theta^{n,m}$ et $\delta\mathbf{v}^{n,m}$.

- `Modifications_unpas`$(\mathbf{v}_h^{n,m}, f_h^n, \delta\theta^{n,m}, \delta\mathbf{v}^{n,m})$ modifie les flux, terme source et erreurs de linéarisation suivant les formules de récurrence (3.12). On gardera les mêmes notations après modification.

- `Reconstructions`$(\Psi^{n,m}, \mathbf{v}_h^{n,m}, f_h^n)$ rassemble les reconstructions espace-temps intervenant dans les estimateurs. On obtient les fonctions $\psi_{ht}^{n,m}$, $\theta_{ht}^{n,m}$, $\mathbf{t}_{ht}^{n,m}$ et f_{ht}^n.

- `Estimateurs`$(\psi_{ht}^{n,m}, \theta_{ht}^{n,m}, \mathbf{t}_{ht}^{n,m}, f_{ht}^n, \delta\theta^{n,m}, \delta\mathbf{v}^{n,m})$ calcule les groupes d'estimateurs définis ci-dessus, à savoir $\eta_{\text{esp}}^{n,m}$, $\eta_{\text{tps}}^{n,m}$ et $\eta_{\text{lin}}^{n,m}$.

- `Ratio`$(\eta_{\text{esp}}^{n,m}, \eta_{\text{tps}}^{n,m}, \lambda_{\text{eq}})$ fait référence à l'évaluation du ratio $r^{n,l}$ défini par (4.3).

Pour $2 \leq n \leq N$, le principe de l'algorithme ci-dessous sur l'intervalle I^n est le suivant. Au temps t^{n-1}, on initialise le pas de temps courant avec la valeur δt^n issue de l'algorithme appliqué à l'intervalle I^{n-1}. L'initialisation de l'inconnue correspond à la valeur obtenue à l'instant t^{n-1} (*i.e.* $\Psi^{n,0} = \Psi^{n-1}$). Les coefficients du schéma BDF2 sont également calculés avant de rentrer dans la boucle de linéarisation. À l'intérieur de cette boucle, il s'agit de résoudre l'équation de Richards à l'itération de linéarisation donnée, de calculer les erreurs de linéarisation, de modifier les flux et le terme source suite à la reformulation du schéma en temps, de définir des reconstructions espace-temps et enfin de calculer les estimateurs. Les itérations de linéarisation sont stoppées lorsque l'erreur de linéarisation devient petite devant les autres sources d'erreur (critère (**??**)). Ensuite, on évalue la nouvelle valeur du ratio $r^{n,l}$. Pour terminer, le pas de temps actualisé $\delta t^{n,l+1}$ remplace le pas de temps courant si le critère (4.2) n'est pas vérifié; sinon, il initialise le pas de temps utilisé à l'instant discret suivant. Notons que les initialisations $\eta_{\text{esp}}^{n,0} = 1$ et $\eta_{\text{tps}}^{n,0} = 1 + \lambda^{\max}$ servent à entrer dans la boucle d'équilibrage, tandis que l'initialisation $\eta_{\text{lin}}^{n,0} = \gamma_{\text{lin}}(3 + \lambda^{\max})$ sert à entrer dans la boucle de linéarisation.

Algorithme 1 Algorithme d'adaptation sur l'intervalle de temps I^n.

Entrée : t^{n-1}, δt^n, Ψ^{n-1}, $\lambda^{\max}$, γ_{lin}, λ_{eq}

$l = 0$, $\delta t^{n,1} = \delta t^n$, $\eta_{\text{esp}}^{n,0} = 1$, $\eta_{\text{tps}}^{n,0} = 1 + \lambda^{\max}$

tant que $\eta_{\text{tps}}^{n,m} > \lambda^{\max}\eta_{\text{esp}}^{n,m}$ **faire**

$l \leftarrow l + 1$

$m = 0$, $\Psi^{n,0} = \Psi^{n-1}$, $\eta_{\text{lin}}^{n,0} = \gamma_{\text{lin}}(3 + \lambda^{\max})$

$(\alpha_0^n, \alpha_1^n, \alpha_2^n) \leftarrow$ `Coefficients_BDF2`$(\delta t^{n,l}, \delta t^{n-1})$

tant que $\eta_{\text{lin}}^{n,m} > \gamma_{\text{lin}}(\eta_{\text{tps}}^{n,m} + \eta_{\text{esp}}^{n,m})$ **faire**

$m \leftarrow m + 1$

$\Psi^{n,m} \leftarrow$ `Richards_itération_NL`$(t^{n-1}, \delta t^{n,l}, \Psi^{n,m-1})$

$(\delta\theta^{n,m}, \delta\mathbf{v}^{n,m}) \leftarrow$ `Erreurs_lin`$(\Psi^{n,m}, \Psi^{n,m-1})$

$(\mathbf{v}_h^{n,m}, f_h^n, \delta\theta^{n,m}, \delta\mathbf{v}^{n,m}) \leftarrow$ `Modifications_unpas`$(\mathbf{v}_h^{n,m}, \delta\theta^{n,m}, \delta\mathbf{v}^{n,m}, f_h^n)$

$(\psi_{ht}^{n,m}, \theta_{ht}^{n,m}, \mathbf{t}_{ht}^{n,m}, f_{ht}^n) \leftarrow$ `Reconstructions`$(\Psi^{n,m}, \mathbf{v}_h^{n,m}, f_h^n)$

$(\eta_{\text{esp}}^{n,m}, \eta_{\text{tps}}^{n,m}, \eta_{\text{lin}}^{n,m}) \leftarrow$ `Estimateurs`$(\psi_{ht}^{n,m}, \theta_{ht}^{n,m}, \mathbf{t}_{ht}^{n,m}, f_{ht}^n, \delta\theta^{n,m}, \delta\mathbf{v}^{n,m})$

fin tant que

$r^{n,l} \leftarrow$ `Ratio`$(\eta_{\text{esp}}^{n,m}, \eta_{\text{tps}}^{n,m}, \lambda_{\text{eq}})$

$\delta t^{n,l+1} \leftarrow r^{n,l}\delta t^{n,l}$

fin tant que

$t^n \leftarrow t^{n-1} + \delta t^{n,l}$, $\delta t^{n+1} \leftarrow \delta t^{n,l+1}$, $\Psi^n \leftarrow \Psi^{n,m}$

Sortie : t^n, δt^{n+1}, Ψ^n

4.2 Cas tests

On teste plusieurs simulations d'infiltration, qui permettent d'étudier en détail le comportement de l'algorithme d'adaptation décrit ci-dessus.
Sur un premier cas test analytique, similaire à celui présenté au chapitre 2, on commence par comparer qualitativement les flux reconstruits dans les espaces $\mathbb{RTN}_0$ et $\mathbb{RTN}_1$, dont la différence de précision nous conduit à préférer la reconstruction dans $\mathbb{RTN}_1$. Ensuite, dans le cas d'un écoulement à vitesse constante, on regarde l'influence des paramètres initiaux, à savoir le pas de temps initial δt^1 et les tolérances $\lambda^{\max}$ et γ_{lin}, sur le pas de temps adapté au cours de la simulation ainsi que sur l'erreur entre la solution exacte et la solution approchée. On choisit des valeurs pour ces paramètres qui permettent à l'algorithme de produire une solution approchée proche de celle obtenue avec une simulation sans adaptation, dont le pas de temps (constant) a été choisi de manière à équilibrer l'erreur L^2 provenant de la discrétisation en espace, et l'erreur provenant de la discrétisation en temps. Enfin, on applique la même méthodologie à un écoulement accéléré, qui confirme la validité de l'algorithme.
On poursuit avec deux cas tests physiquement plus réalistes : l'infiltration (raide) de Polmann présentée au chapitre 2, et un cas test d'infiltration (2D) dans un milieu hétérogène à frontière curviligne. Ces simulations, de difficulté supérieure, nous permettent à la fois de poursuivre la validation de notre stratégie d'adaptation, mais aussi d'estimer le gain de temps CPU par rapport à une simulation effectuée à pas de temps constant.
Pour tous ces cas tests, on utilise la reconstruction de la charge par quart de diamant vue dans la sous-section 3.4.1; on verra en effet que la reconstruction par demi-diamant est insuffisante dans le cas raide de Polmann.

4.2.1 Cas test analytique

On reprend le cas test analytique étudié au chapitre 2, dont on modifie légèrement les caractéristiques; la solution exacte est donnée par :

$$\psi(z,t) = 20\tanh\left(0.17z + a(t)\frac{t}{300} - 13\right) - 40, \quad (z,t) \in [0,100]\times[0,T].$$

La fonction $a(t)$ permet d'augmenter la vitesse de propagation du front, et donc de tester l'algorithme d'adaptation. On considèrera par la suite deux simulations :

- la première, dite sans raideur : $\forall\, t \in [0,T],\ a(t) = 1$;
- la seconde, avec raideur en temps : $\forall\, t \in [0,T],\ a(t) = 1 + \dfrac{t}{4T}$.

Sauf mention contraire, les simulations sont effectuées avec le maillage M_4 défini dans le chapitre 2.

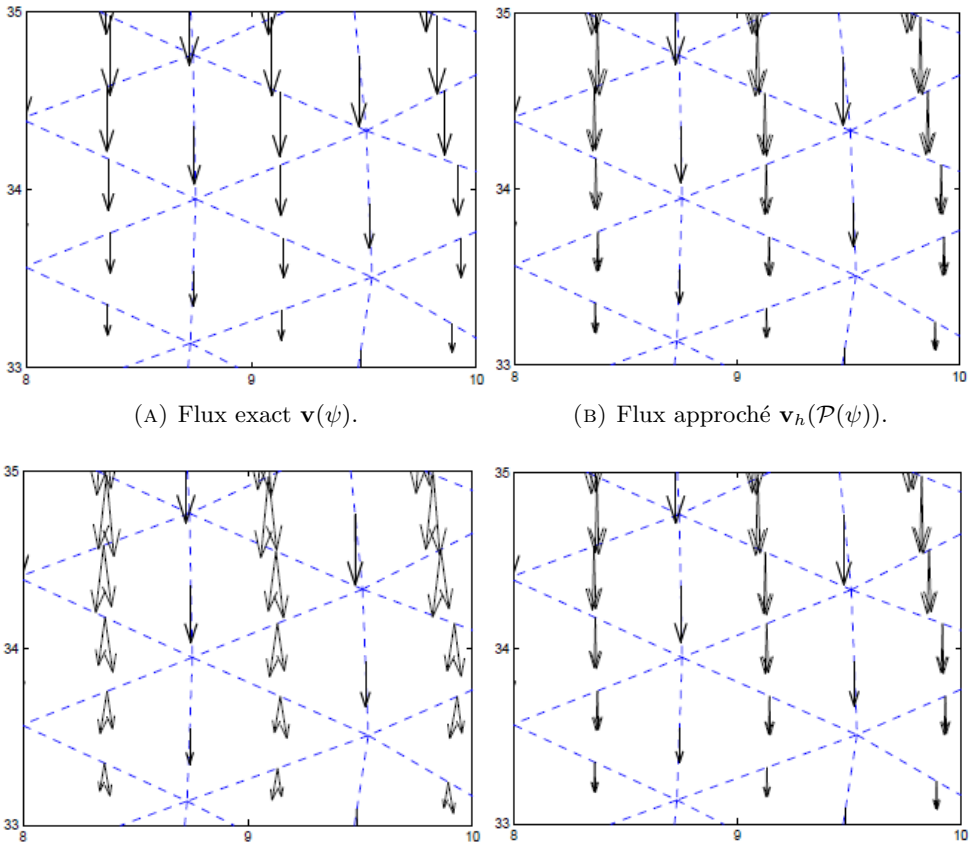
4.2.1.1 Espace de reconstruction du flux Hdiv

Notre premier travail est de faire le choix de l'espace de Raviart-Thomas-Nédélec dans lequel on reconstruit le flux en espace $\mathbf{t}_h$: $\mathbb{RTN}_0$ ou $\mathbb{RTN}_1$. On se place dans le cas sans raideur. Sur la figure 4.3, on a tracé le flux exact $\mathbf{v}(\psi)$, le flux numérique appliqué à la solution exacte (plus précisément au projeté $\mathcal{P}(\psi)$ de la solution exacte sur l'espace des degrés de liberté définis par le schéma DDFV), $\mathbf{v}_h(\mathcal{P}(\psi))$, ainsi que le flux $\mathbf{t}_h$ reconstruit en espace, dans $\mathbb{RTN}_0$ et dans $\mathbb{RTN}_1$. Chaque flux est évalué au centre de chaque arête du maillage primaire. Le flux numérique et les flux reconstruits sont différents de chaque côté de l'arête (maille K et maille L), d'où la présence de deux flèches par arête. Néanmoins, la solution exacte ne dépendant que de la coordonnée verticale z, on s'attend à ce que les flux soient verticaux. Ce n'est clairement pas le cas du flux reconstruit dans $\mathbb{RTN}_0$, ce qui témoigne de son manque de précision dans notre étude. En effet, un flux reconstruit de la forme $(a\ b)^{\mathsf{t}} + c(x\ z)^{\mathsf{t}}$ ne permet pas d'approcher de manière satisfaisante un flux variable suivant la coordonnée verticale, c'est-à-dire du type $(0\ \alpha z)^{\mathsf{t}}$, avec α réel. C'est pourquoi dans les simulations numériques présentées ci-après, on s'intéresse uniquement à des reconstructions dans $\mathbb{RTN}_1$. Celles-ci sont plus coûteuses, mais le temps CPU utilisé par l'ensemble des reconstructions demeure inférieur au coût d'assemblage et de résolution d'un système linéaire (voir les sous-sections 4.2.2 et 4.2.3).

4.2.1.2 Simulation sans raideur

On considère pour commencer une simulation sans raideur, c'est-à-dire $a(t) = 1$ pour tout t. Le temps final de simulation est fixé à $T = 40min$. La charge se propage dans le domaine à vitesse constante, comme indiqué sur la figure 4.4.

Détermination du pas de temps initial δt^1 On commence par lancer des simulations sans adapter ni le pas de temps, ni le critère d'arrêt des linéarisations (on choisit un critère ε fixe et très faible, de l'ordre de 10^{-6}). Le but est de déterminer le pas de temps, que l'on qualifiera d'«optimal», δt_{opt}, qui équilibre les erreurs L^2 provenant de la discrétisation du domaine spatial et de la discrétisation du domaine temporel. Pour cela, on procède en deux étapes :



(A) Flux exact $\mathbf{v}(\psi)$.

(B) Flux approché $\mathbf{v}_h(\mathcal{P}(\psi))$.

(C) Flux $\mathbf{t}_h$ reconstruit dans $\mathbb{RTN}_0$.

(D) Flux $\mathbf{t}_h$ reconstruit dans $\mathbb{RTN}_1$.

FIGURE 4.3: Flux exact, DDFV et reconstruit dans $\mathbb{RTN}_0$ et $\mathbb{RTN}_1$.

1. en choisissant un pas de temps constant très petit, on estime l'erreur e_{esp} due uniquement à la discrétisation du domaine spatial, l'erreur due à la discrétisation en temps étant alors négligeable devant celle-ci;

2. une fois l'erreur e_{esp} identifiée, on définit δt_{opt} comme le pas de temps qui produit une erreur entre la solution exacte et la solution approchée égale à $2e_{\text{esp}}$: l'erreur en espace et l'erreur en temps sont alors égales.

L'erreur est ici définie par :

$$e_{\Omega,T} = \frac{||\psi_{\text{ex}} - \psi_{\text{app}}||_{L^2(\Omega\times[0,T])}}{||\psi_{\text{app}}||_{L^2(\Omega\times[0,T])}},$$

où ψ_{ex} est la fonction constante par morceaux sur chaque maille primaire K et chaque intervalle de temps I^n, définie à l'aide de la solution exacte par :

$$\psi_{\text{ex}}|_{K\times I^n}(x,t) := \psi(x_K, t^n)\mathbb{1}_{K\times I^n},$$

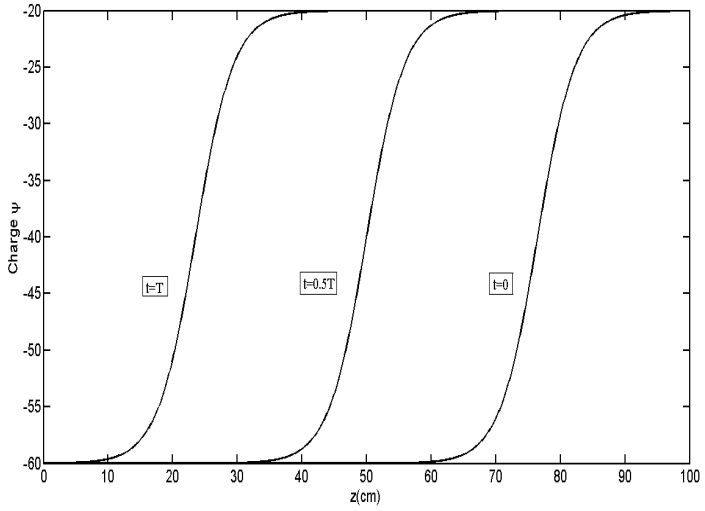


FIGURE 4.4: Profil de la charge à différents instants.

et ψ_{app} est la fonction constante par morceaux sur chaque maille primaire K et chaque intervalle de temps I^n, définie à l'aide de la solution approchée Ψ^n par :

$$\psi_{\text{app}}|_{K\times I^n}(x,t) := \Psi_K^n \mathbb{1}_{K\times I^n}.$$

On a ainsi tracé sur la figure 4.5 les valeurs prises par $e_{\Omega,T}$ en fonction du pas de temps choisi pour la simulation. À l'aide de ces tracés, on identifie l'erreur e_{esp} due exclusivement à la discrétisation en espace : $e_{\text{esp}} \simeq 3.9e^{-4}$. Le pas de temps δt_{opt} pour lequel on a égalité des erreurs dues à la discrétisation en temps et à la discrétisation en espace est alors $\delta t_{\text{opt}} = 40s$. Ainsi, puisque la propagation du front s'effectue à vitesse constante, on s'attend à ce qu'une simulation adaptative conserve un pas de temps à peu près constant, égal à δt_{opt}, tout au long de la simulation. On réalise maintenant des simulations adaptatives, et on cherche à estimer les paramètres d'entrée de l'algorithme, à savoir $\lambda^{\max}$, γ_{lin} et δt^1. Le paramètre δt^1 est pour l'instant choisi égal à δt_{opt}, on verra par la suite que ce choix n'est de toute façon pas important, puisque l'algorithme est robuste à la condition initiale.

Influence de $\lambda^{\max}$ Notre objectif étant de rendre les erreurs de linéarisation assez petites pour avoir une influence négligeable sur les résultats de la simulation (on ne tient pas encore compte du temps de calcul), on donne pour l'instant une valeur volontairement faible à la tolérance de linéarisation. On choisit donc $\gamma_{\text{lin}} = 1/1000$, et on observe l'influence de la valeur de $\lambda^{\max}$ sur le pas de temps adapté au cours de la simulation,

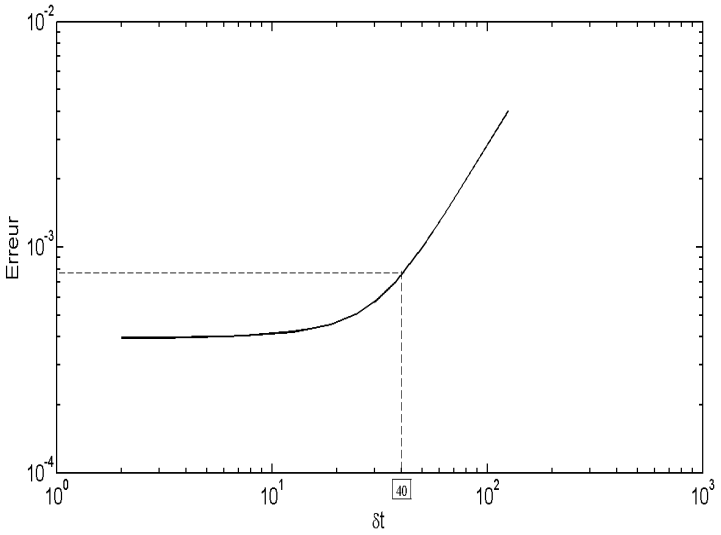
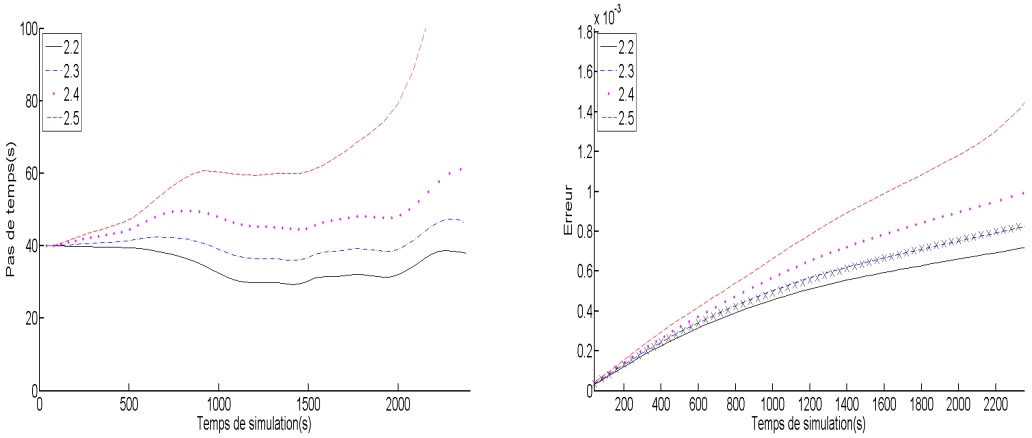


FIGURE 4.5: Erreur $e_{\Omega,T}$ en fonction du pas de temps δt choisi.

ainsi que sur l'erreur $e_{\Omega,T}$ (voir la figure 4.6). Comme attendu, on voit que plus la



(A) Pas de temps adapté.

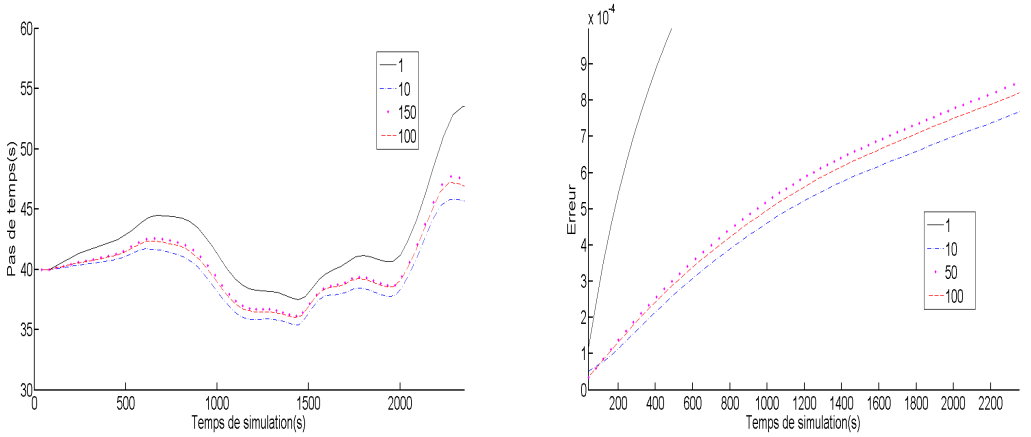
(B) Erreur L^2.

FIGURE 4.6: Pas de temps adapté et erreur pour différentes valeurs de $\lambda^{\max}$. L'erreur obtenue sans adaptation est indiquée par les croix noires.

tolérance $\lambda^{\max}$ est restrictive, plus le pas de temps adapté au cours de la simulation est petit. Notre choix se porte naturellement sur la valeur de $\lambda^{\max}$ qui permet de produire une erreur $e_{\Omega,T}$ proche de celle obtenue sans adaptation avec le pas de temps δt_{opt} : $\lambda^{\max} = 2.3$. Cette valeur correspond à un ratio d'équilibre $\lambda_{\text{eq}} = 1.15 \simeq 1$; ainsi, une

simulation adaptative effectuée avec un ratio d'équilibre proche de 1 équilibre les erreurs dues à la discrétisation en temps et en espace.

Influence de γ_{lin} Le paramètre $\lambda^{\max}$ étant maintenant estimé, on trace (figure 4.7) l'influence de la tolérance de linéarisation γ_{lin} sur le pas de temps adapté au cours de la simulation. On observe qu'une tolérance $\gamma_{\text{lin}} = 1/50$ semble suffisante pour rendre les



(A) Pas de temps adapté.

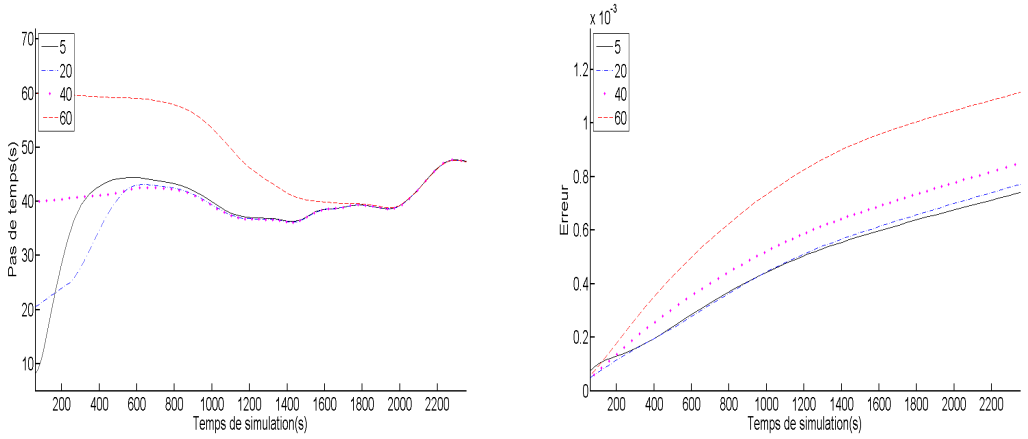
(B) Erreur L^2.

FIGURE 4.7: Pas de temps adapté et erreur pour différentes valeurs de $1/\gamma_{\text{lin}}$.

effets des erreurs de linéarisation négligeables.

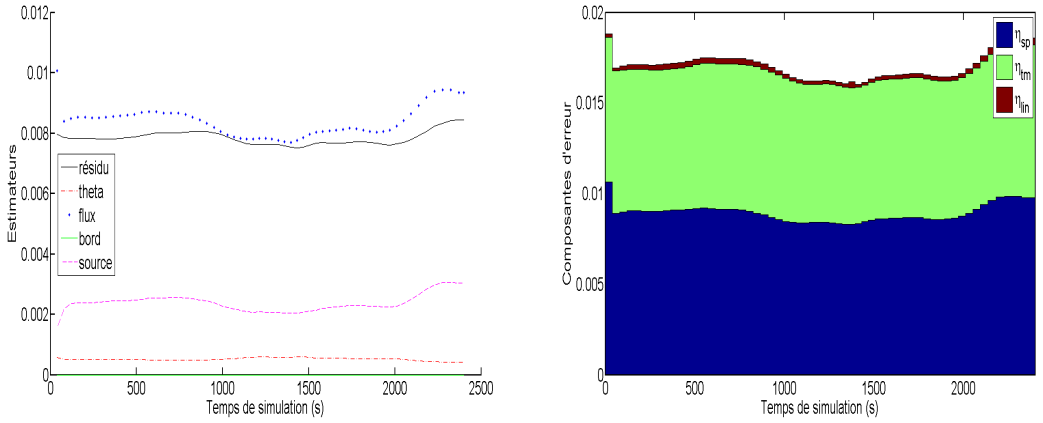
Influence de δt^1 Enfin, on fixe $\gamma_{\text{lin}} = 1/50$, $\lambda^{\max} = 2.3$, et on observe l'influence du pas de temps initial sur le pas de temps adapté au cours de la simulation (voir la figure 4.8). On exhibe ainsi la robustesse de l'algorithme d'adaptation, puisque le choix du pas de temps initial n'a aucune influence sur le pas de temps adapté en fin de simulation, dans une large mesure.
Les paramètres d'entrée de l'algorithme d'adaptation désormais fixés ($\delta t^1 = 40s$, $\gamma_{\text{lin}} = 1/50$, $\lambda^{\max} = 2.3$), on regarde l'évolution de chaque estimateur au cours de la simulation, séparément puis rassemblés en composantes d'erreur en temps, espace et de linéarisation (figure 4.9). On s'aperçoit que les estimateurs de résidu et de flux en espace, η_{res} et $\eta_{\mathbf{t}}$, dominent les estimateurs sur le flux en temps et le terme source, η_θ et η_f.

Influence du maillage Pour terminer, on souhaite observer l'influence de la taille du maillage sur l'estimation du paramètre $\lambda^{\max}$. Ayant travaillé jusqu'à présent avec



(A) Pas de temps adapté. (B) Erreur L^2.

FIGURE 4.8: Pas de temps adapté et erreur pour différentes valeurs de δt^1.



(A) Estimateurs. (B) Composantes d'erreur.

FIGURE 4.9: Évolution des estimateurs.

le maillage M_4, on considère successivement les maillages M_2, M_3 et M_5. Pour chacun, on applique la même méthodologie que précédemment pour l'estimation de δt_{opt}, puis de $\lambda^{\max}$. Les deux autres paramètres de l'algorithme adaptatif sont fixés : $\gamma_{\text{lin}} = 1/50$ et $\delta t^1 = \delta t_{\text{opt}}$. On a synthétisé dans le Tableau 4.1 les valeurs obtenues pour δt_{opt} et $\lambda^{\max}$. Il ressort que plus le pas du maillage est petit, plus la tolérance $\lambda^{\max}$ doit être choisie petite pour obtenir une erreur proche de celle obtenue sans adaptation, avec un pas de temps constant égal à δt_{opt}. Il semble de plus que la valeur $\lambda^{\max}$ converge asymptotiquement vers une valeur proche de 2, ce qui correspond à un ratio d'équilibre $\lambda_{\text{eq}} = 1$. On peut alors se demander si pour un maillage suffisamment fin, l'ordre de

convergence du schéma est respecté avec une tolérance $\lambda^{\max} \simeq 2$ fixe pour tous les maillages. On a ainsi calculé l'erreur $e_{\Omega,T}$ obtenue après une simulation adaptative, avec les paramètres $\gamma_{\text{lin}} = 1/50$, $\lambda^{\max} = 2.1$ et $\delta t^1 = \delta t_{\text{opt}}$ pour chaque maillage. Les résultats indiqués dans le Tableau 4.2 laissent supposer que l'ordre 2 est effectivement préservé.

Maillage	M_2	M_3	M_4	M_5
$e_{\Omega,T}$	5.2e-3	1.6e-3	3.9e-4	1.1e-4
$\delta t_{\text{opt}}(s)$	230	94	40	19
$\lambda^{\max}$	4.2	2.9	2.3	2.2

TABLEAU 4.1: Valeur de $\lambda^{\max}$ selon le maillage.

Maillage	M_2	M_3	M_4	M_5
$\delta t^1(s)$	230	94	40	19
$e_{\Omega,T}$	1e-2	3.2e-3	8.4e-4	2.1e-4
ordre	-	1.8	2.2	2

TABLEAU 4.2: Erreur $e_{\Omega,T}$ obtenue pour une simulation adaptative avec $\lambda^{\max} = 2.1$.

4.2.1.3 Simulation avec raideur en temps

On définit maintenant la fonction de raideur a par $a(t) = 1 + \dfrac{t}{4T}$ pour tout t. On augmente ainsi la vitesse de propagation du front avec le temps, comme on peut le voir sur la figure 4.10. On s'attend à ce que l'erreur en temps augmente au cours d'une simulation sans adaptation, et donc à ce que le pas de temps optimal δt_{opt} soit une fonction décroissante du temps, prenant des valeurs inférieures à 40 secondes (cas sans raideur). Plus précisément, on applique la méthodologie précédente à des subdivisions de l'intervalle de temps $[0, T]$; sur chaque sous-intervalle de la forme $[t^i, t^{i+1}]$, on calcule l'erreur suivante :

$$e_{\Omega,t^i} = \frac{||\psi_{\text{ex}} - \psi_{\text{app}}||_{L^2(\Omega\times[t^i,t^{i+1}])}}{||\psi_{\text{app}}||_{L^2(\Omega\times[t^i,t^{i+1}])}}.$$

On obtient un pas de temps $\delta t_{\text{opt},t^i}$ qui équilibre l'erreur due à la discrétisation en espace et l'erreur due à la discrétisation en temps, sur chaque intervalle $[t^i, t^{i+1}]$. Les valeurs sont rassemblées dans le Tableau 4.3, et montrent que le pas de temps optimal diminue au cours de la simulation, ce qu'on attendait. On utilise maintenant l'algorithme d'adaptation, et on veut de nouveau estimer les paramètres δt^1, $\lambda^{\max}$ et γ_{lin}. Une

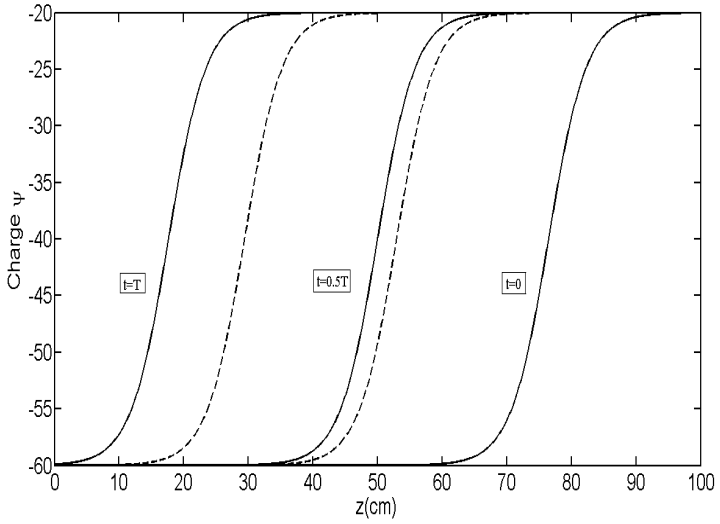


FIGURE 4.10: Profil de la charge à différents instants (sans raideur en pointillés).

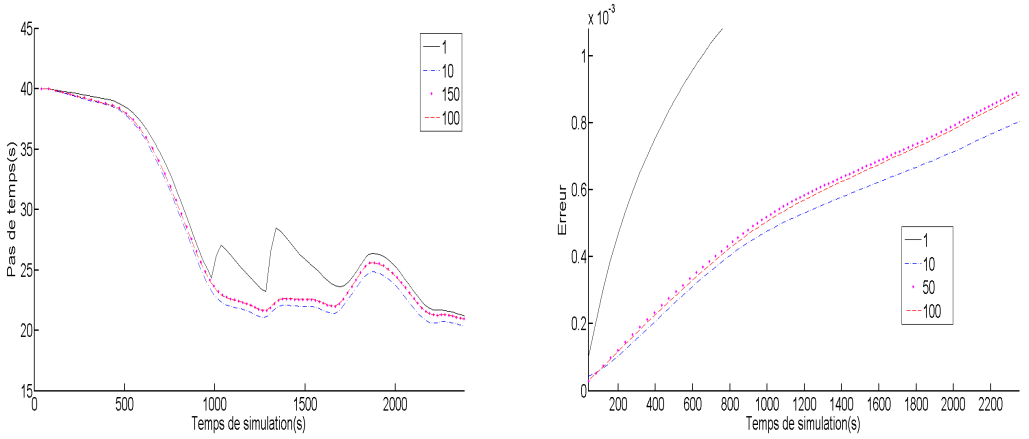
Temps t^i	0	$T/6$	$T/3$	$T/2$	$2T/3$	$5T/6$
e_{esp,t^i}	3e-4	1.9e-4	1.7e-4	1.5e-4	1.4e-4	1.5e-4
$\delta t_{\text{opt},t^i}(s)$	40	34	31	29	27	25

TABLEAU 4.3: Choix du pas de temps optimal $\delta t_{\text{opt},t^i}$.

tolérance de linéarisation $\gamma_{\text{lin}} = 1/50$ est encore suffisante pour ne pas influencer les résultats numériques, en particulier le pas de temps adapté au cours de la simulation (voir la figure 4.11). En remarquant que $a(0) = 1$, la solution exacte avec et sans raideur coïncident à l'instant initial. On choisit donc un pas de temps initial $\delta t^1 = 40s$ (comme dans le cas sans raideur), et on estime la valeur de $\lambda^{\max}$ telle que l'erreur $e_{\Omega,T}$ soit proche de celle obtenue sans adaptation (voir la figure 4.13). La valeur obtenue est $\lambda^{\max} = 2.1$, ce qui est proche du cas sans raideur. Enfin, sur la figure 4.13, on vérifie également la robustesse de l'algorithme à la condition initiale dans ce cas de difficulté supérieure.

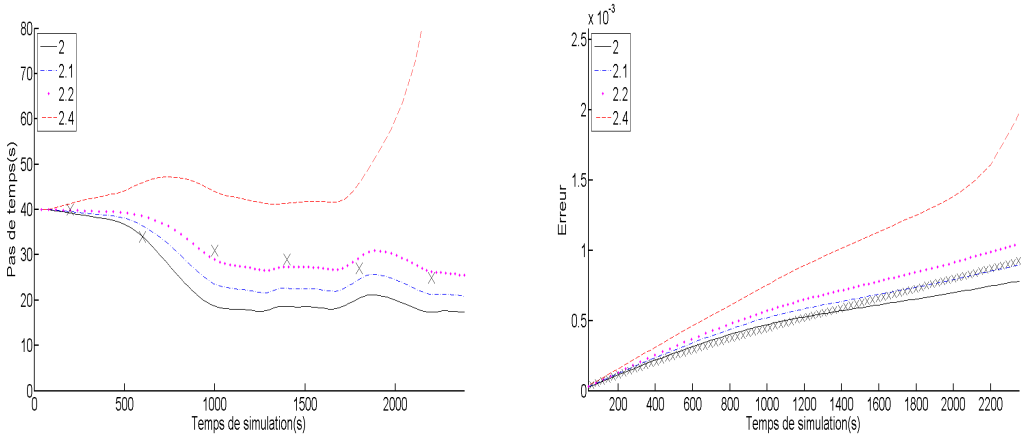
4.2.2 Cas test de Polmann

On reprend le cas test d'écoulement raide présenté au chapitre 2. On rappelle que la difficulté de la simulation tient dans les premiers pas de temps, du fait de la forte discontinuité entre la condition initiale et la condition de Dirichlet. L'objectif est double :



(A) Pas de temps adapté. (B) Erreur L^2.

FIGURE 4.11: Pas de temps adapté et erreur $e_{\Omega,T}$ pour différentes valeurs de $1/\gamma_{\text{lin}}$.

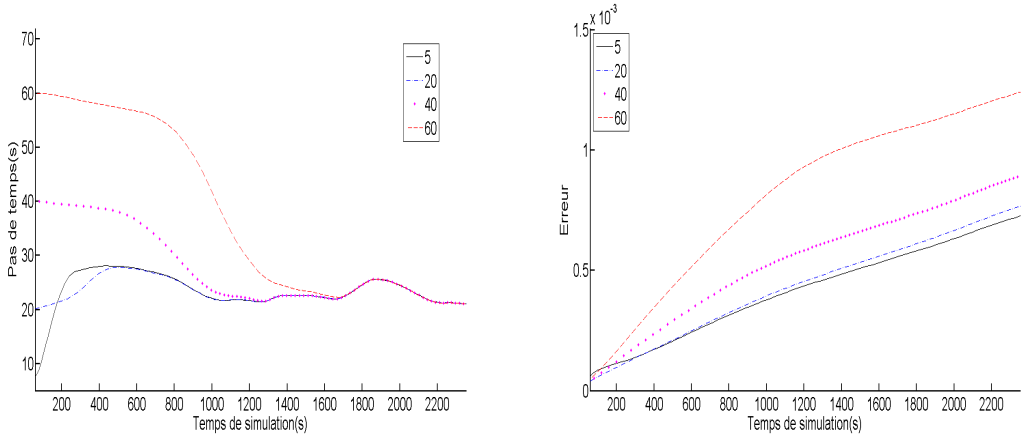


(A) Pas de temps adapté. (B) Erreur L^2.

FIGURE 4.12: Pas de temps adapté et erreur pour différentes valeurs de $\lambda^{\max}$. Les pas de temps $\delta t_{\text{opt},t^i}$ ainsi que l'erreur sans adaptation sont indiqués par les croix noires.

on souhaite poursuivre la validation de notre stratégie d'adaptation, mais aussi tester ses performances. La propagation du front étant plus aisée loin des premiers instants, on s'attend en effet à une augmentation du pas de temps et donc à un gain significatif de temps de calcul.

On commence par remarquer que la reconstruction de la charge affine par demi-diamant n'est pas suffisante pour ce cas test, ce qui justifie l'emploi de la reconstruction par quart de diamant (voir la sous-section 3.4.1), associée à une nouvelle évaluation du tenseur. On regarde alors de nouveau l'influence de chacun des paramètres $\lambda^{\max}$, γ_{lin} et δt^1 sur la



(A) Pas de temps adapté. (B) Erreur L^2.

FIGURE 4.13: Pas de temps adapté et erreur $e_{\Omega,T}$ pour différentes valeurs de δt^1.

qualité de la simulation ainsi que sur les temps de calcul, ce qui nous permet de choisir un triplet $(\lambda^{\max}, \gamma_{\text{lin}}, \delta t^1)$ et de comparer la simulation obtenue avec le cas classique (sans adaptation). Le maillage primaire utilisé pour la simulation est le maillage M_5 défini au chapitre 2 (6474 triangles). Notons également que dans ce cas test ainsi que le suivant (sous-section 4.2.3), le coefficient multiplicatif d'adaptation du pas de temps, $r^{n,l}$, est toujours affine par morceaux, mais ses valeurs minimale et maximale sont fixées respectivement à 1.05 et 0.95 (au lieu de 2 et 0.5). En atténuant les variations du pas de temps au cours de la simulation, on évite ainsi des instabilités de l'algorithme lors de passages raides (en début de simulation dans le cas présent).

Nécessité de la reconstruction par quart de diamant L'estimateur le plus sensible aux différentes reconstructions présentées dans le chapitre 3 est l'estimateur portant sur le flux, $\eta_{\mathbf{t}}$. Cela est particulièrement visible sur des cas tests raides comme celui de Polmann. On rappelle qu'à un instant discret t^n et un itéré de linéarisation m, cet estimateur s'écrit sous la forme suivante :

$$\eta_{\mathbf{t},K}^{n,m} := h_K^{-1} \left\| \mathbb{K}(\psi_{ht}) \nabla(\psi_{ht} + z) + \mathbf{t}_{ht}^m \right\|_{K \times I^n} .$$

À cause des non-linéarités, deux choix doivent être faits avec précision pour garantir un lien étroit entre le flux numérique $\mathbf{t}_{ht}^m$ (reconstruit dans un espace de Raviart-Thomas-Nédélec idoine) et le flux continu $-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht} + z)$ (appliqué à la solution approchée reconstruite en temps et en espace, ψ_{ht}) :

- la reconstruction en espace ψ_h, à laquelle il est naturel de demander que son gradient soit proche du gradient DDFV. Ce choix doit être fait **après** la résolution du système linéaire;

- l'évaluation du tenseur, $\mathbb{K}_{\sigma,K}(\Psi^{n,m-1})$, intervenant **pendant** l'assemblage du système linéaire. Plusieurs possibilités ont été listées dans la sous-section 2.2.4. Ce choix est plus délicat, puisqu'il doit être fait en adéquation avec la reconstruction ψ_h, afin de rendre le tenseur exact $\mathbb{K}(\psi_{ht})$ proche de l'approximation $\mathbb{K}_{\sigma,K}(\Psi^{n,m-1})$ intervenant dans les conditions définissant le flux $\mathbf{t}_{ht}^m$. La non-linéarité de $\mathbb{K}$ constitue bien sûr une difficulté supplémentaire.

Des choix peu pertinents peuvent aboutir dans des cas raides à une surévaluation de l'estimateur $\eta_{\mathbf{t}}$, comme on le verra sur la figure 4.16. Ainsi, la reconstruction interpolante affine par demi-diamant échoue sur ce cas test (quelle que soit la méthode choisie pour l'évaluation du tenseur par ailleurs), victime notamment des fortes variations de la perméabilité relative. Le flux $-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht}+z)$ est en effet très imprécis en haut du domaine de calcul dès les premiers instants de simulation. Les flux tracés sur la figure 4.14 sont évalués au barycentre de chaque demi-diamant pour la première reconstruction, et au barycentre de chaque quart de diamant pour la seconde. La reconstruction par quart de diamant, elle, se montre plus robuste lorsqu'on l'associe à la nouvelle évaluation du tenseur suivante :

$$\mathbb{K}_{\sigma,K}(\Psi^{n,m-1}) = \frac{1}{2}\left(\mathbb{K}(\psi_{\star,K,A}^{n,m-1}, \mathbf{x}_{\star,K}) + \mathbb{K}(\psi_{\star,K,B}^{n,m-1}, \mathbf{x}_{\star,K})\right),$$

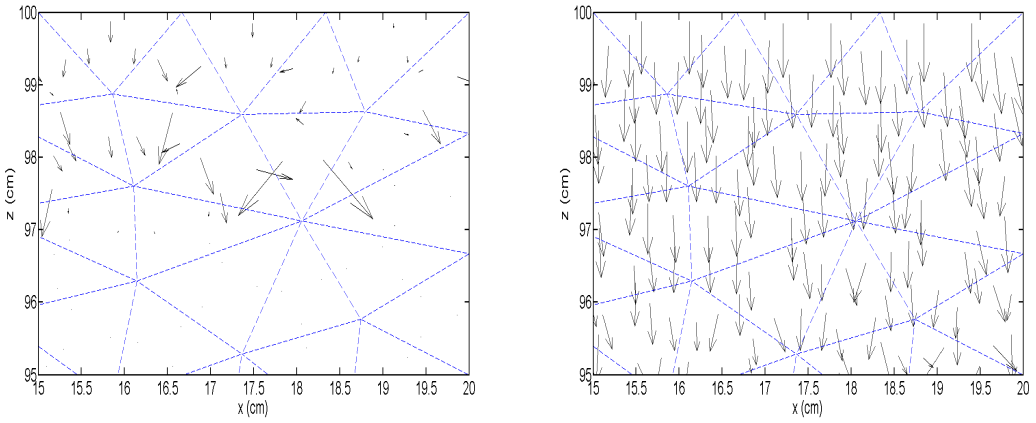
avec :

$$\begin{cases} \psi_{\star,K,A}^{n,m-1} = \dfrac{1}{3}\left(\psi_A^{n,m-1} + \psi_K^{n,m-1} + \psi_\sigma^{n,m-1}\right), \\ \psi_{\star,K,B}^{n,m-1} = \dfrac{1}{3}\left(\psi_B^{n,m-1} + \psi_K^{n,m-1} + \psi_\sigma^{n,m-1}\right), \\ \mathbf{x}_{\star,K} = \dfrac{1}{3}\left(\mathbf{x}_K + \mathbf{x}_A + \mathbf{x}_B\right). \end{cases}$$

C'est pourquoi dans tous les cas tests de ce chapitre, on a retenu la reconstruction par quart de diamant, associée à l'évaluation ci-dessus pour le tenseur.

Choix des paramètres

- Influence de $\lambda^{\max}$ On étudie l'influence de $\lambda^{\max}$, ayant fixé $\gamma_{\text{lin}} = 1/50$ et $\delta t^1 = 20s$, qui est le pas de temps choisi sans adaptation (figure 4.15). Logiquement, plus on

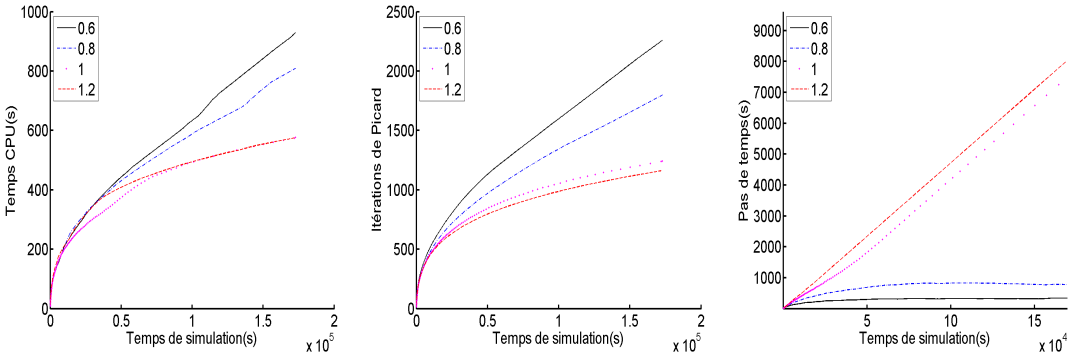


(A) Par demi-diamant. (B) Par quart de diamant.

FIGURE 4.14: Flux $-\mathbb{K}(\psi_{ht})\nabla(\psi_{ht}+z)$ selon la reconstruction choisie.

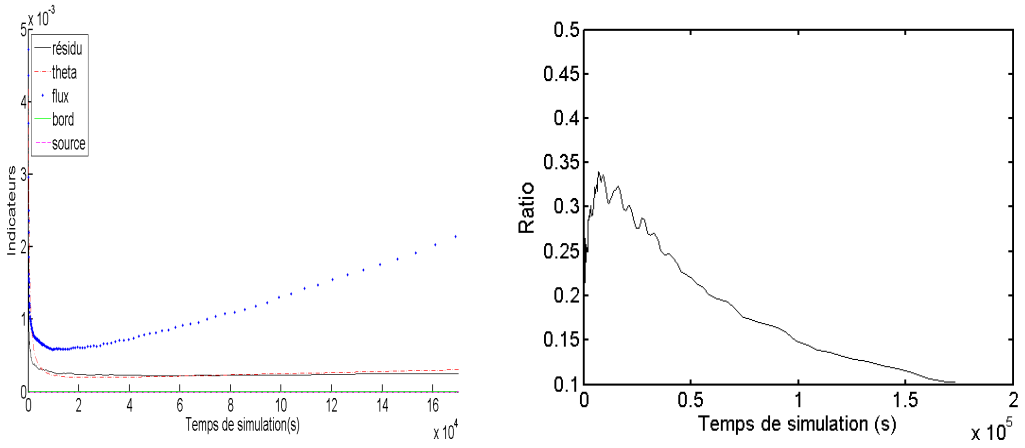
est lâche sur la tolérance $\lambda^{\max}$, plus le pas de temps adapté est élevé et plus la simulation est économique en temps de calcul et en nombre cumulé d'itérations de Picard. Cependant, dans les cas $\lambda^{\max}=1$ et $\lambda^{\max}=1.2$, le pas de temps est toujours augmenté au maximum autorisé d'une itération sur l'autre, pour atteindre un pas de temps final d'environ $8000s$. Plus précisément, l'indicateur sur le flux $\eta_{\mathbf{t}}$ augmente plus rapidement que les autres estimateurs, ce qui conduit l'algorithme à considérer que le quotient de l'erreur en temps et de l'erreur en espace λ^{n,m_∞} (défini en (4.3)), devient de plus en plus petit, alors que ça devrait être le contraire puisque l'on augmente le pas de temps tout au long de la simulation (voir la figure 4.16). Dans ce cas, on parlera de *divergence de l'algorithme d'adptation.* Cela nous conduit à choisir $\lambda^{\max}=0.8$. Ainsi, les observations sont similaires à celles du cas test analytique; la différence principale est que la valeur «optimale» de $\lambda^{\max}$ est inférieure dans ce cas, ce qu'on attribue principalement à l'absence de terme source, dont l'estimateur associé est donc nul.

- Influence de γ_{lin} On fixe le pas de temps initial $\delta t^1=20s$ et la tolérance espace-temps $\lambda^{\max}=0.8$, et on regarde l'influence de la valeur de γ_{lin} sur le temps de calcul, le nombre cumulé d'itérations de Picard et le pas de temps adapté au cours de la simulation (figure 4.17). Comme attendu, le nombre cumulé d'itérations de Picard ainsi que le temps CPU augmentent avec γ_{lin}. De son côté, le pas de temps adapté au cours de la simulation a une allure concave. En effet, la discontinuité initiale induit une erreur en temps significative, qui s'estompe par la suite; c'est pourquoi le pas de temps a tendance à augmenter, puis à se stabiliser lorsque



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

FIGURE 4.15: Comparaison selon le choix de la tolérance $\lambda^{\max}$.

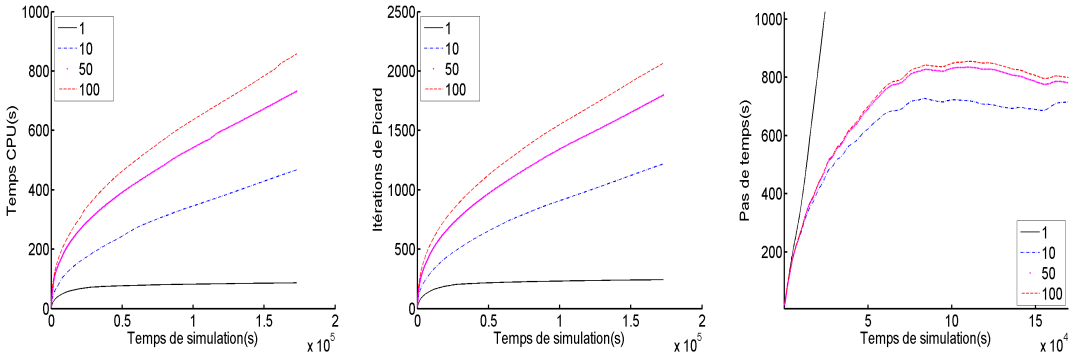


(A) Estimateurs. (B) Ratio λ^{n,m_∞}.

FIGURE 4.16: Évolution des estimateurs et du ratio λ^{n,m_∞} lorsque $\lambda^{\max} \geq 1$.

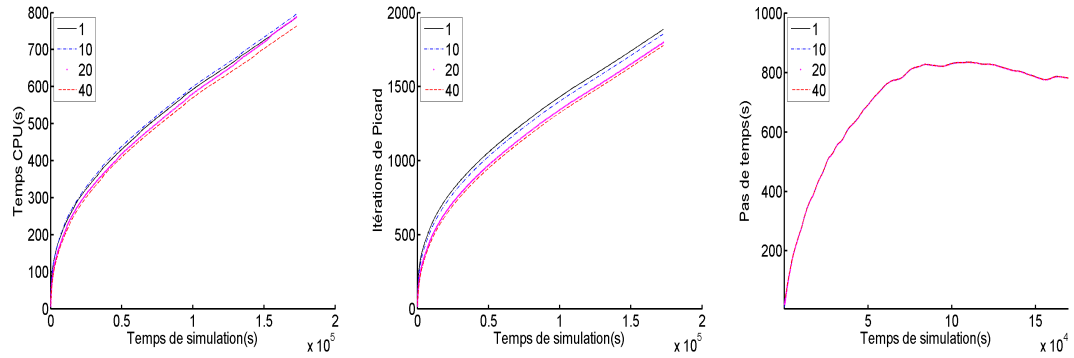
l'erreur en temps et l'erreur en espace s'équilibrent. Le choix $\gamma_{\text{lin}} = 1/50$ semble de nouveau suffisant pour ce cas test.

- Influence de δt^1 On fixe cette fois les deux tolérances en prenant $\gamma_{\text{lin}} = 1/50$ et $\lambda^{\max} = 0.8$, et on fait varier le pas de temps initial (figure 4.18). Les graphiques obtenus illustrent une nouvelle fois la stabilité de l'algorithme d'adaptation à la condition initiale, puisque le choix du pas de temps initial semble avoir une influence négligeable sur les calculs : l'algorithme ajuste rapidement le pas de temps indépendamment du choix initial δt^1. Notons enfin que l'algorithme diverge au-delà de $\delta t^1 = 40s$, ce qui est également le cas sans stratégie d'adaptation. Par la suite, on prendra (par exemple) $\delta t^1 = 20s$.



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

FIGURE 4.17: Comparaison selon le choix de la tolérance $1/\gamma_{\text{lin}}$.



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

FIGURE 4.18: Comparaison selon le choix du pas de temps initial δt^1.

Comparaison avec la solution obtenue sans adaptation On fixe désormais $\gamma_{\text{lin}} = 1/50$, $\delta t^1 = 20s$ et $\lambda^{\max} = 0.8$, et on quantifie le gain par rapport à une simulation sans adaptation effectuée à pas de temps constant $\delta t = 20s$. Le coût des reconstructions nécessaires à l'adaptation n'est pas négligeable, principalement pour deux raisons : les diverses reconstructions liées à la charge ψ et à la saturation θ sont effectuées par quart de diamant, qui est un maillage six fois plus fin que le maillage primaire associé; de plus, la construction d'un flux dans l'espace de Raviart-Thomas-Nédélec $\mathbb{RTN}_1$ est significativement plus chère que dans l'espace $\mathbb{RTN}_0$ (8 degrés de liberté par maille primale contre 3). La figure 4.19 montre cependant que le coût total de ces reconstructions et du calcul des estimateurs reste inférieur au coût d'assemblage et de résolution du système linéaire associé à la discrétisation de l'équation de Richards. On quantifie ensuite le gain en termes de temps de calcul et de nombre d'itérations de Picard effectuées au cours de

la simulation (figure 4.20). On observe que la simulation adaptative permet de réduire le nombre d'itérations de Picard d'un facteur dix, et le temps CPU d'un facteur cinq environ.

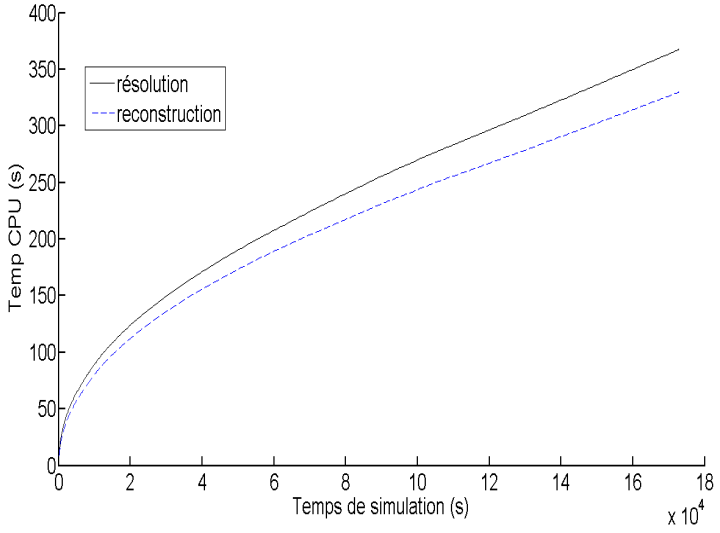
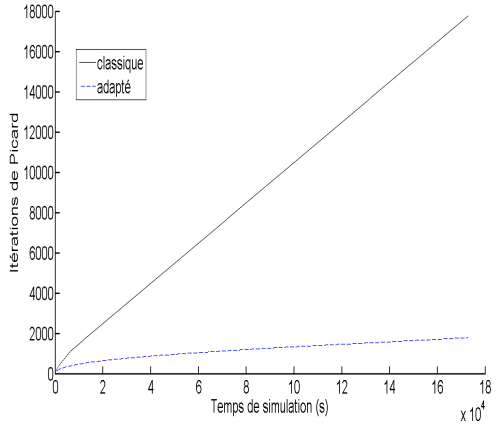
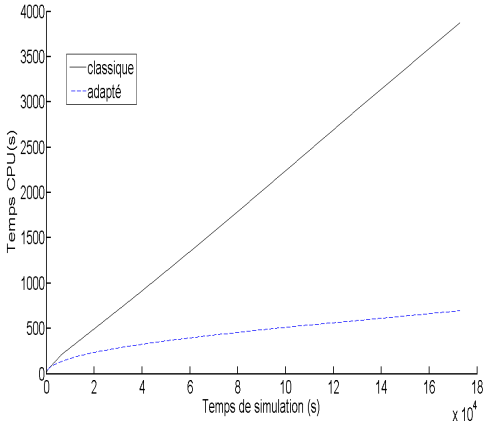


FIGURE 4.19: Coût de la résolution du système linéaire et du calcul des estimateurs.



(A) Itérations de Picard cumulées.



(B) Temps CPU cumulé(s).

FIGURE 4.20: Gain par rapport à une simulation sans adaptation.

4.2.3 Sol hétérogène à frontière curviligne

On définit un domaine consistant en l'empilement de deux sols de mêmes caractéristiques, données par les relations de Van Genuchten. Seule la perméabilité relative à saturation change; elle vaut $\overline{k_s} = 9.44 \cdot 10^{-5} cm.s^{-1}$ dans la partie supérieure, et $\underline{k_s} = 2\overline{k_s}$ dans la partie inférieure. Le domaine est $\Omega = [0, 100] \times [0, 100]$ (en centimètres) et le temps final $T = 72\ h$. La condition initiale est hydrostatique, et on impose une condition de Dirichlet homogène sur les bords haut et bas du domaine. On complète par une condition de Neumann homogène sur les bords latéraux. La discontinuité entre la pression initiale hydrostatique et la condition de Dirichlet provoque un écoulement du haut vers le bas. La frontière entre les deux sols est curviligne, décrite par la courbe d'équation : $\zeta(x) = 100\,[0.1\,(1 - \cos(\pi x/100)) + 0.45]$; elle permet de tester les performances de l'algorithme sur un cas test à deux dimensions (voir la figure 4.21). Le profil du front pour une simulation classique est indiqué sur la figure 4.22). On suit la même méthodologie que pour le cas test de Polmann : on étudie séparément l'influence de γ_{lin}, δt^1 et $\lambda^{\max}$ sur les performances de l'algorithme, puis on quantifie le gain procuré par une simulation adaptative avec les paramètres retenus.

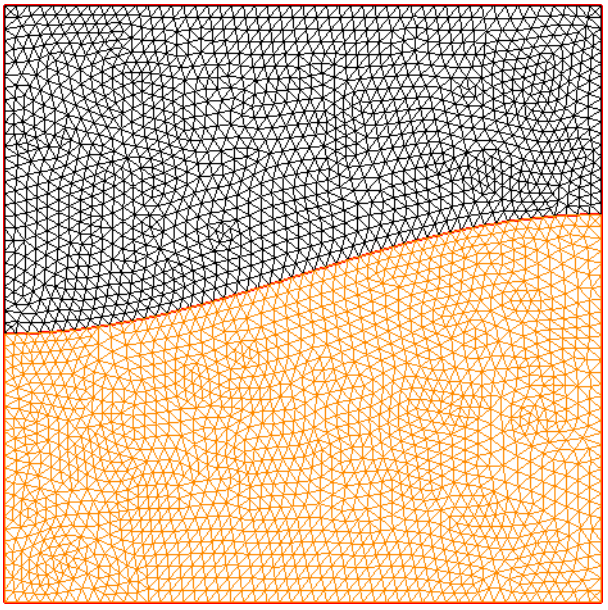
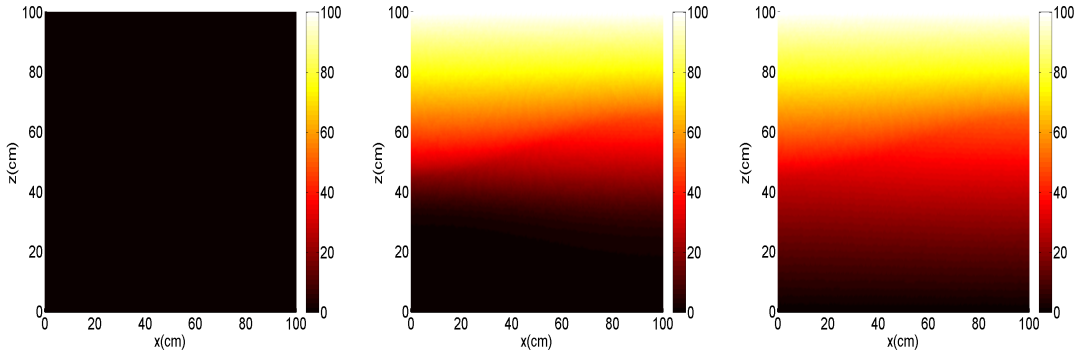


FIGURE 4.21: Exemple de maillage pour le domaine à frontière curviligne.



(A) $t = 0$. (B) $t = 36h$. (C) $t = 72h$.

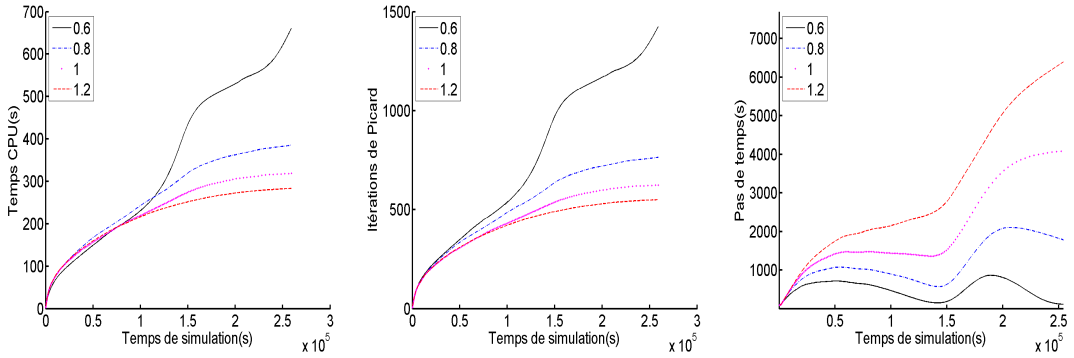
FIGURE 4.22: Profil de la surpression à différents instants.

Choix des paramètres

- Influence de $\lambda^{\max}$ On fait varier la tolérance $\lambda^{\max}$, fixant $\gamma_{\text{lin}} = 1/50$ et $\delta t^1 = 100s$ (pas de temps pour une simulation sans adaptation). L'allure du pas de temps adapté au cours de la simulation (voir la figure 4.23) nous permet d'estimer la tolérance d'équilibrage. Le pas de temps adapté associé à la valeur $\lambda^{\max} = 1$ suit quatre phases : une première phase de croissance qui répond, comme dans le cas test de Polmann, à la détente qui suit le choc provoqué par la discontinuité initiale; puis, une stabilisation du pas de temps qui correspond à un équilibre entre l'erreur en temps et l'erreur en espace. Lorsque le front arrive à l'interface curviligne entre les deux sols, l'erreur due à la discrétisation en espace augmente; pour retrouver un équilibre entre les erreurs en temps et en espace, il faut donc augmenter le pas de temps. On observe pour terminer une nouvelle phase de stabilisation. Ce choix de $\lambda^{\max}$ paraît le plus pertinent, il est proche de la valeur obtenue avec le cas test de Polmann.

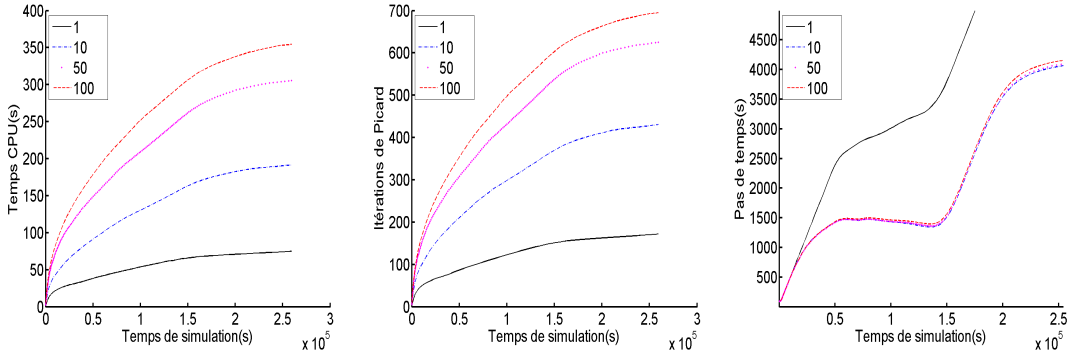
- Influence de γ_{lin} On fixe $\delta t^1 = 100s$, $\lambda^{\max} = 1$ et on fait varier γ_{lin}. Là encore, une tolérance $\gamma_{\text{lin}} = 1/50$ suffit pour cette simulation : le pas de temps reste inchangé, de même que le profil de la charge (figure 4.24).

- Influence de δt^1 Choisissant $\gamma_{\text{lin}} = 1/50$ et $\lambda^{\max} = 1$, on fait varier le pas de temps initial (figure 4.25). La stabilité de l'algorithme d'adaptation est confirmée, là encore le choix du pas de temps initial δt^1 a peu d'importance, pourvu qu'il soit choisi dans un intervalle raisonnable qui permette à la fois la convergence de l'algorithme sur les premiers pas de temps, et de ne pas gaspiller inutilement du temps de calcul. On choisit ici $\delta t^1 = 100s$, comme pour une simulation sans adaptation.



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

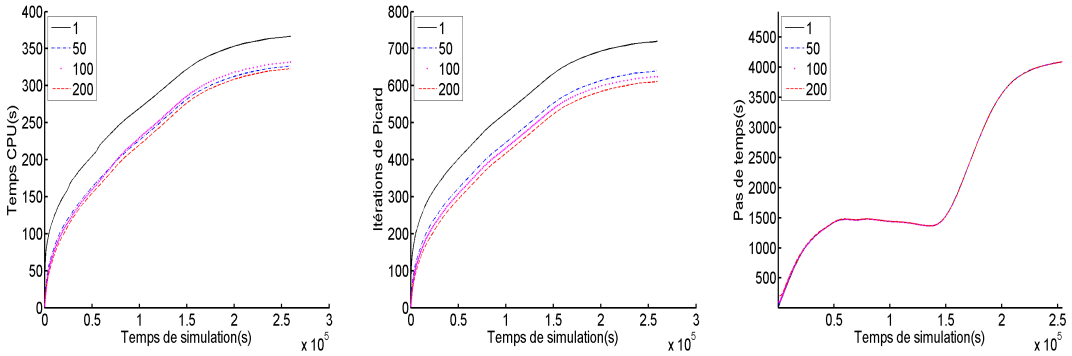
FIGURE 4.23: Comparaison selon le choix de la tolérance $\lambda^{\max}$.



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

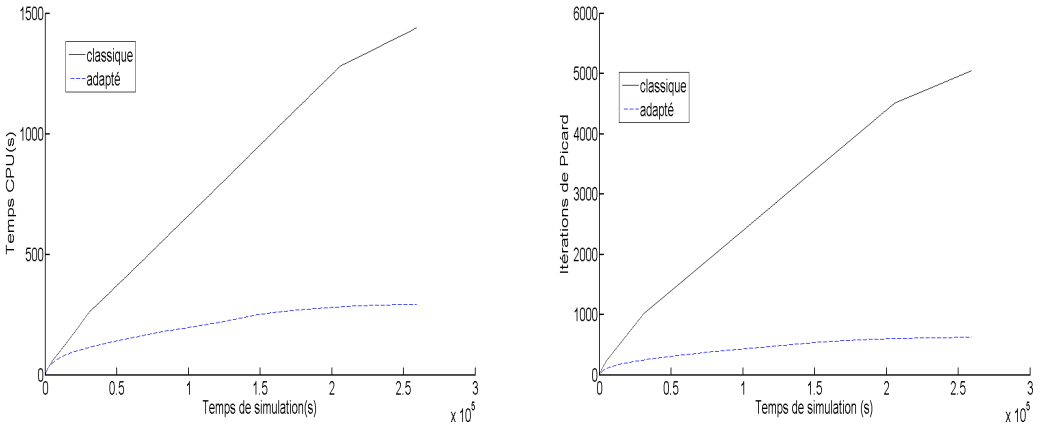
FIGURE 4.24: Comparaison selon le choix de la tolérance $1/\gamma_{\text{lin}}$.

Comparaison avec la solution obtenue sans adaptation On fixe $\gamma_{\text{lin}} = 1/50$, $\delta t^1 = 100s$ et $\lambda^{\max} = 1$, et on quantifie le gain par rapport à une simulation sans adaptation effectuée à pas de temps constant $\delta t = 100s$. Les résultats présentés sur la figure 4.26 sont similaires à ceux obtenus avec le cas test de Polmann. Le gain est moindre, puisque la discontinuité initiale est moins forte.



(A) Temps CPU(s). (B) Itérations de Picard. (C) Pas de temps(s).

FIGURE 4.25: Comparaison selon le choix du pas de temps initial δt^1.



(A) Temps CPU cumulé(s). (B) Itérations de Picard cumulées.

FIGURE 4.26: Gain par rapport à une simulation sans adaptation.

Conclusion et perspectives

Dans ce mémoire, on a discrétisé une équation de type parabolique non linéaire utilisée pour la simulation d'écoulements en milieu poreux. Des estimations *a posteriori* adaptées à cette discrétisation nous ont également permis d'établir une stratégie d'adaptation du pas de temps et du critère d'arrêt des linéarisations.

Dans le chapitre 2, on a ainsi étendu une discrétisation DDFV à un flux non linéaire de la forme $K(\psi)\nabla\psi$. Le schéma obtenu autorise la présence de discontinuités des propriétés physiques à l'interface entre deux volumes de contrôle. Cette discrétisation a été couplée avec la formule BDF2 pour résoudre numériquement l'équation de Richards. On propose également différents moyens pour évaluer le tenseur. Numériquement, la présence de conditions aux limites mixtes de type Dirichlet/Neumann non homogène demande un traitement particulier afin d'éviter l'apparition d'oscillations numériques indésirables. Les tests numériques montrent une superconvergence du schéma DDFV-BDF2 sur un cas test analytique, ainsi qu'une bonne robustesse dans le cas d'écoulements raides ou en milieu hétérogène et anisotrope.

Dans le chapitre 3, on s'est placé dans la cadre des estimations *a posteriori* utilisant la technique des flux équilibrés. On a étendu les résultats originaux de [33], qui donnent un cadre théorique général pour obtenir des estimations garanties pour une équation parabolique non linéaire en espace, à l'équation de Richards qui présente également une non-linéarité dans le terme en temps. La borne supérieure obtenue de l'erreur en norme duale fait intervenir des estimateurs mesurés dans une norme espace-temps. Sa démonstration s'appuie sur une hypothèse d'équilibrage des flux, qui demande en pratique de s'assurer que les reconstructions faites à partir de la solution approchée soient en adéquation avec le système discret. On a expliqué comment définir de telles reconstructions pour notre discrétisation DDFV-BDF2, bien que notre approche reste valable pour d'autres types de discrétisations en espace et une large classe de schémas en temps qui peuvent s'écrire sous la forme de schémas à un pas.

Enfin, le chapitre 4 a été l'occasion d'appliquer les estimations locales précédentes à une stratégie d'adaptation du pas de temps et du critère d'arrêt des linéarisations du système

discret. On a proposé un algorithme qui s'appuie sur un équilibrage des différents estimateurs, regroupés selon leur sensibilité à la discrétisation du maillage, de l'intervalle de temps de simulation, ou des linéarisations. On a observé sur l'ensemble des simulations que l'algorithme est robuste à la condition initiale, c'est-à-dire qu'il produit un pas de temps adapté (loin des premiers instants de simulation) indépendant du pas de temps initial choisi. Le choix de la tolérance d'arrêt des linéarisations du système discret permettant des erreurs de linéarisation négligeables, semble relativement indépendant du cas test considéré. Enfin, l'équilibrage des erreurs en temps et en espace à réaliser pour produire une solution approchée proche de celle obtenue sans adaptation varie selon le cas test; elle semble en particulier dépendre de la présence d'un terme source. Une fois ces paramètres fixés, l'algorithme d'adaptation permet de gagner un temps de calcul considérable, notamment sur des cas raides où le pas de temps peut être augmenté significativement après les premiers instants de simulation.

La bonne mise en œuvre de cette adaptation est conditionnée à la proximité entre les estimateurs et l'erreur commise lors de la simulation. Un point crucial concerne la pertinence des différentes reconstructions, en particulier de la charge hydraulique. Ainsi, même la reconstruction par quart de diamant qui a été utilisée pour les cas tests du chapitre 4 se révèle insuffisante dans certains cas. Par exemple, le problème five spot étudié au chapitre 2 pose des problèmes d'adaptation, l'estimateur de flux explosant lorsque le front de pression atteint le bas du domaine. Une autre reconstruction, non présentée dans ce manuscrit, permet de stabiliser l'algorithme dans ce cas. Son avantage principal est que le gradient de la fonction reconstruite coïncide avec le gradient numérique sur chaque demi-diamant (la reconstruction par quart de diamant ne vérifie cette propriété qu'en moyenne sur chaque demi-diamant). Cependant, cette reconstruction n'est pas conforme, ce qui nécessiterait l'introduction d'un estimateur de non-conformité, et se montre inadaptée pour le cas test de Polmann. La question d'une reconstruction pertinente dans tous les cas reste donc ouverte, et constitue à ce jour un frein pour l'obtention d'une borne inférieure.

Il est également important de comprendre plus en profondeur le rôle joué par chaque estimateur dans l'adaptation proposée. Les composantes liées à une erreur en espace (respectivement en temps) sont difficiles à identifier, notamment parce que chaque estimateur est intégré localement en espace et en temps, et varie donc avec le pas du maillage et le pas de temps. Notre choix s'est porté sur le regroupement qui donnait à l'algorithme sa propriété de robustesse à la condition initiale, mais une analyse théorique permettrait de s'affranchir de cette heuristique. Une première extension naturelle de ces

travaux consisterait à profiter de la localité en espace des estimations pour adapter le maillage au cours de la simulation. Il serait également intéressant d'établir une borne inférieure de l'erreur en norme duale, qui garantirait ainsi la pertinence des estimateurs et de l'adaptation associée. Le cadre d'estimation par flux équilibrés intégrés à la fois en espace et en temps nous paraît le cadre élégant et naturel pour obtenir une telle borne. Néanmoins, sa démonstration n'est pas évidente pour des schémas volumes finis, une difficulté de nouveau liée au choix de la reconstruction de la charge, conjointement avec l'évaluation du tenseur lors de la discrétisation. D'un point de vue numérique, le passage en trois dimensions d'espace est une étape importante pour les simulations actuelles. Des cas tests plus réalistes, en deux ou trois dimensions, permettraient aussi de poursuivre la validation de la discrétisation et de l'algorithme d'adaptation associé, et de questionner la possibilité d'une implémentation en milieu industriel. Cela impliquerait des maillages comportant un grand nombre de volumes de contrôle (de l'ordre du million), et donc probablement le recours à un solveur itératif pour résoudre chaque système linéaire. L'erreur algébrique qui en résulterait peut être prise en compte dans notre stratégie d'adaptation, comme cela a par exemple été réalisé récemment dans [31]. Enfin, plus spécifiquement pour DDFV, ces estimations peuvent être étendues en prenant également en compte les équations écrites sur le maillage secondaire, écartées ici.

.

Programmation efficace en MATLAB

On dresse ici une liste de remarques utiles concernant la mise en oeuvre du code Matlab, tant pour l'assemblage et la résolution du système discret que pour le calcul des estimateurs et l'algorithme d'adaptation. Le tout dans un souci d'optimisation du temps de calcul, l'espace mémoire ne faisant pas défaut pour les maillages considérés dans ce mémoire (quelques milliers de cellules).

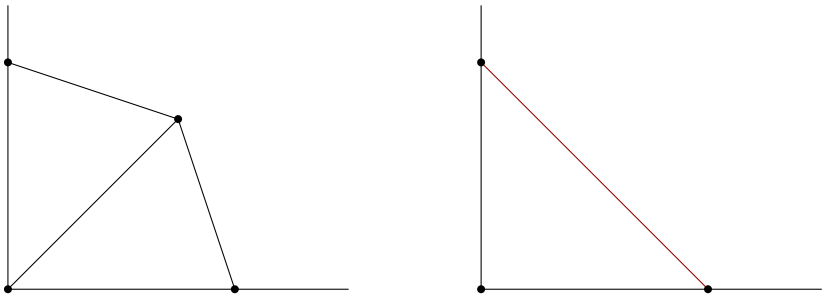
- Les performances du logiciel Matlab sont optimales lorsque les calculs sont vectorisés. Pour mener à bien cette vectorisation, les arêtes sont groupées selon leur type, c'est-à-dire selon la nature de chacune de leurs extrémités (sommet intérieur, Dirichlet ou Neumann). On suppose pour simplifier que deux arêtes formant un coin du maillage bordent (au moins) deux mailles primaires distinctes. Ainsi, une arête dont les deux extrémités se trouvent au bord du maillage est forcément une arête du bord (voir la figure 27). Une arête $\sigma = [\mathbf{x}_A \mathbf{x}_B]$ donnée appartient donc nécessairement à l'un des types suivants :

  1. $\mathbf{x}_A$ et $\mathbf{x}_B$ sont des sommets intérieurs,
  2. $\mathbf{x}_A$ est un sommet intérieur, $\mathbf{x}_B$ un sommet Dirichlet,
  3. $\mathbf{x}_A$ est un sommet intérieur, $\mathbf{x}_B$ un sommet Neumann,
  4. $\mathbf{x}_A$ est un sommet Neumann, $\mathbf{x}_B$ un sommet Dirichlet,
  5. $\mathbf{x}_A$ et $\mathbf{x}_B$ sont des sommets Neumann,
  6. $\mathbf{x}_A$ et $\mathbf{x}_B$ sont des sommets Dirichlet.

  Dans le même souci de clarté, nous n'avons pas tenu compte dans la liste précédente de l'ordre des extrémités. À chaque type d'arête correspond une contribution spécifique des flux à la matrice du système et à son second membre.

- La matrice du système linéaire est creuse (voir le Tableau 2.2), l'assembler comme telle en utilisant la fonction **sparse** de Matlab et utiliser les algorithmes de résolution spécifiques font gagner un temps de calcul considérable et permettent

d'économiser de l'espace mémoire. Une matrice creuse est entièrement décrite par la donnée de trois vecteurs, contenant respectivement les abscisses des coefficients non nuls, leurs ordonnées et leur valeur. Les deux premiers vecteurs peuvent être assemblés en prétraitement, puisque la matrice du système garde la même structure itération après itération.



(A) Autorisé (B) Non autorisé

FIGURE 27: Maillage primaire en un coin du domaine Ω.

- Dans le chapitre 4, la charge hydraulique reconstruite en espace ψ_h est affine sur chaque quart de diamant. Pour cette raison, le calcul des estimateurs locaux sur chaque maille primaire est décomposé sur chaque quart de diamant formant cette maille (il y en a six). L'intégration est réalisée à l'aide d'une quadrature de Gauss. Comme il s'agit d'une intégration en temps et en espace (sur chaque quart de diamant et chaque intervalle de temps), le coût de calcul peut rapidement devenir prohibitif. En pratique, un point de Gauss (le centre de masse du quart de diamant en espace, le point milieu en temps) suffit pour ne pas influencer qualitativement les différents estimateurs.

- Le flux reconstruit $\mathbf{t}_h$ dans l'espace $\mathbb{RTN}_1$ prend la forme générique suivante sur chaque maille primaire K :

$$\forall\ (x,z) \in K,\ \mathbf{t}_h(x,z) = \begin{pmatrix} ax + bz + c \\ dx + ez + f \end{pmatrix} + g \begin{pmatrix} x^2 \\ xz \end{pmatrix} + h \begin{pmatrix} xz \\ z^2 \end{pmatrix}.$$

  Les dépendances en la maille K sont ici omises. Le calcul des différents estimateurs fait intervenir l'évaluation de $\mathbf{t}_h$ en chacun des points d'intégration choisis. Cette évaluation se compose *a priori* de deux étapes :

  1. On calcule la solution $X = (a\ b\ c\ d\ e\ f\ g\ h)$ du système linéaire défini par les conditions (3.24), (3.25) et (3.26). En écrivant ce système linéaire

sous la forme $M_1 X = B$, il est aisé de voir que la matrice M_1 ne dépend que de données géométriques. Ainsi, pour une maille K donnée, plutôt que d'inverser le système linéaire à chaque itéré en temps n et à chaque itération de linéarisation m, on garde en mémoire la matrice inverse M_1^{-1} calculée en début de simulation. L'inversion de systèmes linéaires locaux (de taille 8 ici) s'est simplifiée en un simple produit matrice-vecteur : $X = M_1^{-1} B$.

2. On évalue le flux $\mathbf{t}_h$ en chaque point d'intégration; dans notre cas, le centre de masse de chaque quart de diamant. On appelle M_2 la matrice canoniquement associée à l'application linéaire, qui à partir de l'octuplet (a, b, c, d, e, f, g, h) renvoie les valeurs du flux au centre de masse des six quarts de diamant formant la maille primaire K. Les différentes valeurs cherchées s'obtiennent alors grâce au produit $M_2(a\ b\ c\ d\ e\ f\ g\ h)^{\mathsf{t}}$. La matrice M_2 dépend elle aussi seulement de données géométriques et peut de ce fait être calculée en prétraitement.

Finalement, il suffit pour chaque maille primaire de calculer le produit $M_2 M_1^{-1} B$. On assemble en prétraitement la matrice $M := M_2 M_1^{-1}$, puis, à chaque couple d'itérés (n, m), on calcule le second membre B (qui dépend du flux numérique) et on effectue le produit MB.

.

Bibliographie

[1] I. Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Computational Geosciences*, 6(3-4):405–432, 2002.

[2] I. Aavatsmark, T. Barkve, O. Bøe, and T. Mannseth. Discretization on unstructured grids for inhomogeneous, anisotropic media. part i: Derivation of the methods. *SIAM Journal on Scientific Computing*, 19(5):1700–1716, 1998.

[3] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, and J.M. Nordbotten. A compact multipoint flux approximation method with improved robustness. *Numerical Methods for Partial Differential Equations*, 24(5):1329–1360, 2008.

[4] L. Agélas, D.A. Di Pietro, and J. Droniou. The G method for heterogeneous anisotropic diffusion on general meshes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(04):597–625, 2010.

[5] S. Agmon. *Lectures on Elliptic Boundary Value Problems*. Van Nostrand, 1965.

[6] M. Ainsworth and J.T. Oden. *A posteriori error estimation in finite element analysis*, volume 37. John Wiley & Sons, 2011.

[7] H. Alt Wilhelm and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Mathematische Zeitschrift*, 183(3):311–341, 1983.

[8] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general 2D meshes. *Numerical Methods for Partial Differential Equations*, 23(1):145–195, 2007.

[9] I. Babuška and W.C. Rheinboldt. A-posteriori error estimates for the finite element method. *Int. J. Numer. Meth. Eng.*, 12(10):1597–1615, 1978.

[10] R.E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Mathematics of Computation*, 44(170):283–301, 1985.

[11] V. Baron, Y. Coudière, and P. Sochala. Comparison of DDFV and DG methods for flow in anisotropic heterogeneous porous media. *Oil Gas Sci. Technol. – Rev. IFP En. nouvelles*, 2013.

[12] M. Bebendorf. A note on the poincaré inequality for convex domains. *Z. Anal. Anwendungen22, 4*, 22(4):751–756, 2003.

[13] C. Bernardi, L. El Alaoui, and Z. Mghazli. A posteriori analysis of a space and time discretization of a nonlinear model for the flow in variably saturated porous media. *IMA Journal of Numerical analysis*, 2013.

[14] F. Boyer and F. Hubert. Finite volume method for 2D linear and nonlinear elliptic problems with discontinuities. *SIAM J. Numer. Anal.*, 46(6):3032–3070, 2008.

[15] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. New York: Springer-Verlag, 1991.

[16] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM Journal on Numerical Analysis*, 43(5):1872–1896, 2005.

[17] P. Broadbridge and I. White. Constant rate rainfall infiltration: A versatile nonlinear model: 1. analytic solution. *Water Resources Research*, 24(1):145–154, 1988.

[18] R.H. Brooks and A.T. Corey. Hydraulic properties of porous media. *Hydrology Papers. Colorado State University*, (3), 1964.

[19] R. Cautrès, R. Herbin, and F. Hubert. The Lions domain decomposition algorithm on non matching cell-centered finite volume meshes. *IMA Journal Numerical Analysis*, 24:465–490, 2004.

[20] M.A. Celia, E.T. Bouloutas, et al. A general mass-conservative numerical solution for the unsaturated flow equation. *Water Resour. Res.*, 26(7):1483–1496, 1990.

[21] Y. Coudière and F. Hubert. A 3D discrete duality finite volume method for nonlinear elliptic equations. *SIAM Journal on Scientific Computing*, 33(4):1739–1764, 2011.

[22] Y. Coudière, F. Hubert, and G. Manzini. A CeVeFE DDFV scheme for discontinuous anisotropic permeability tensors. In *Finite Volumes for Complex Applications VI Problems & Perspectives*, pages 283–291. Springer, 2011.

[23] Y. Coudière, C. Pierre, et al. A 2D/3D Discrete Duality Finite Volume scheme. Application to ECG simulation. *IJFV*, 6(1):1–24, 2009.

[24] Y. Coudière, J.P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 33(03):493–516, 1999.

[25] C.F. Curtiss and J.O. Hirschfelder. Integration of stiff equations. *Proceedings of the National Academy of Sciences of the United States of America*, 38(3):235, 1952.

[26] E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th national conference*, pages 157–172. ACM, 1969.

[27] G.G. Dahlquist. A special stability problem for linear multistep methods. *BIT Numerical Mathematics*, 3(1):27–43, 1963.

[28] H. Darcy. *Les fontaines publiques de la ville de Dijon: exposition et application...* Victor Dalmont, 1856.

[29] S. Delcourte. *Développement de méthodes de volumes finis pour la mécanique des fluides.* Thèse de Doctorat, Université Paul Sabatier-Toulouse III, 2007.

[30] P. Destuynder and B. Métivet. Explicit error bounds in a conforming finite element method. *Math. Comp. of the Amer. Math. Soc.*, 68(228):1379–1396, 1999.

[31] D.A. Di Pietro, E. Flauraud, M. Vohralik, S. Yousef, et al. A posteriori error estimates, stopping criteria, and adaptivity for multiphase compositional Darcy flows in porous media. 2013.

[32] D.A. Di Pietro, M. Vohralík, S. Yousef, et al. A posteriori error estimates with application of adaptive mesh refinement for thermal multiphase compositional flows in porous media. *Computers and Mathematics with Applications*, 2013.

[33] V. Dolejší, A. Ern, and M. Vohralík. A framework for robust a posteriori error control in unsteady nonlinear advection-diffusion problems. *SIAM J. Numer. Anal.*, 51(2):773–793, 2013.

[34] K. Domelevo and P. Omnes. A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids. *ESAIM: Mathematical Modelling and Numerical Analysis*, 39(06):1203–1249, 2005.

[35] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Mathematical Models and Methods in Applied Sciences*, 20(02):265–295, 2010.

[36] A. Ern, A. F. Stephansen, and M. Vohralík. Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection–diffusion–reaction problems. *Journal of computational and applied mathematics*, 234(1):114–130, 2010.

[37] A. Ern, A.F. Stephansen, and M. Vohralík. Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection–diffusion–reaction problems. *Journal of computational and applied mathematics*, 234(1):114–130, 2010.

[38] A. Ern and M. Vohralík. A posteriori error estimation based on potential and flux reconstruction for the heat equation. *SIAM J. Numer. Anal.*, 48(1):198–223, 2010.

[39] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. *Handbook of numerical analysis*, 7:713–1018, 2000.

[40] R. Eymard, T. Gallouët, and R. Herbin. Cell centred discretisation of non linear elliptic problems on general multidimensional polyhedral grids. *Journal of Numerical Mathematics*, 17(3):173–193, 2009.

[41] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes sushi: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, page drn084, 2009.

[42] R. Eymard, C. Guichard, R. Herbin, and R. Masson. Vertex-centred discretization of multiphase compositional darcy flows on general meshes. *Computational Geosciences*, 16(4):987–1005, 2012.

[43] R. Eymard, C. Guichard, R. Herbin, and R. Masson. Gradient schemes for two-phase flow in heterogeneous porous media and richards equation. *ZAMM-Journal of Applied Mathematics and Mechanics*, 2013.

[44] R. Eymard, M. Gutnic, and D. Hilhorst. The finite volume method for Richards equation. *Computational Geosciences*, 3(3-4):259–294, 1999.

[45] J. Ferrandon. Les lois de l'écoulement de filtration. *Génie civil*, 125(24), 1948.

[46] W.R. Gardner. Calculation of capillary conductivity from pressure plate outflow data. *Soil Science Society of America Journal*, 20(3):317–320, 1956.

[47] V.E. Ginting. *Computational upscaled modeling of heterogeneous porous media flow utilizing finite volume method.* Thèse de Doctorat, Texas A&M University, 2004.

[48] R.D. Grigorieff. Stability of multistep-methods on variable grids. *Numerische Mathematik*, 42(3):359–377, 1983.

[49] S.M. Hassanizadeh and W.G. Gray. Thermodynamic basis of capillary pressure in porous media. *Water Resources Research*, 29(10):3389–3405, 1993.

[50] R. Haverkamp, M. Vauclin, et al. A comparison of numerical simulation models for one-dimensional infiltration. *Soil Sci. Soc. Am. J.*, 41(2):285–294, 1977.

[51] R. Herbin, F. Hubert, et al. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. *Finite volumes for complex applications V*, pages 659–692, 2008.

[52] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *Journal of computational Physics*, 160(2):481–499, 2000.

[53] F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, 192(16-18):1939–1959, 2003.

[54] F. Hermeline, S. Layouni, and P. Omnes. A finite volume method for the approximation of Maxwell's equations in two space dimensions on arbitrary meshes. *Journal of Computational Physics*, 227(22):9365–9388, 2008.

[55] P. Knabner and E. Schneid. Adaptive hybrid mixed finite element discretization of instationary variably saturated flow in porous media. *High Performance Scientific and Engineering Computing*, 29:37–44, 2002.

[56] S. Krell. *Schémas Volumes Finis en mécanique des fluides complexes.* Thèse de Doctorat, Université de Provence-Aix-Marseille I, 2010.

[57] P. Ladevèze. *Comparaison de modèles de milieux continus.* Thèse de Doctorat, Université Pierre et Marie Curie (Paris 6), 1975.

[58] R.J. LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*, volume 98. Siam, 2007.

[59] G. Manzini and S. Ferraris. Mass-conservative finite volume methods on 2-D unstructured grids for the Richards' equation. *Adv. Water Res.*, 27(12):1199–1215, 2004.

[60] G. Matheron. *Eléments pour une théorie des milieux poreux*. Masson, 1967.

[61] Y. Mualem. A new model for predicting the hydraulic conductivity of unsaturated porous media. *Water resources research*, 12(3):513–522, 1976.

[62] M. Ohlberger. A posteriori error estimate for finite volume approximations to singularly perturbed nonlinear convection-diffusion equations. *Numerische Mathematik*, 87(4):737–761, 2001.

[63] M. Ohlberger. A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 35(02):355–387, 2001.

[64] P. Omnes, Y. Penel, and Y. Rosenbaum. A posteriori error estimation for the discrete duality finite volume discretization of the Laplace equation. *SIAM Journal on Numerical Analysis*, 47(4):2782–2807, 2009.

[65] O. Østerby. Five ways of reducing the Crank – Nicolson oscillations. *BIT Numerical Mathematics*, 43(4):811–822, 2003.

[66] F. Otto. L1-contraction and uniqueness for quasilinear elliptic–parabolic equations. *Journal of differential equations*, 131(1):20–38, 1996.

[67] O. Pironneau, Hecht, et al. FreeFEM. *URL: http://www. freefem. org*, 2006.

[68] W. Prager and J.L. Synge. Approximations in elasticity based on the concept of function space. *Quart. Appl. Math*, 5(3):1–21, 1947.

[69] K. Pruess. TOUGH2: A general-purpose numerical simulator for multiphase fluid and heat flow. *Lawrence Berkeley Laboratory Report LBL-29400*, 1991.

[70] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media, 2008.

[71] S. Salager. *Etude de la rétention d'eau et de la consolidation des sols dans un cadre thermo-hydro-mécanique*. Thèse de Doctorat, Université Montpellier 2, sciences et techniques du Languedoc, 2007.

[72] J. Simmons, B.L. Landrum, J.M. Pinson, and P.B. Crawford. Swept areas after breakthrough in vertically fractured five-spot patterns. *Trans. AIME*, 216:73, 1959.

[73] P. Sochala. *Méthodes numériques pour les écoulements souterrains et couplage avec le ruissellement*. Thèse de Doctorat, Ecole Nationale des Ponts et Chaussées, 2008.

[74] P. Sochala, A. Ern, and S. Piperno. Mass conservative BDF-discontinuous Galerkin/explicit finite volume schemes for coupling subsurface and overland flows. *Computer Methods in Applied Mechanics and Engineering*, 198(27):2122–2136, 2009.

[75] P. Sochala and O.P. Le Maître. Polynomial Chaos expansion for subsurface flows with uncertain soil parameters. *Advances in Water Resources*, 62:139–154, 2013.

[76] M.T. Van Genuchten. A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.*, 44(5):892–898, 1980.

[77] R. Verfürth. A review of a posteriori error estimation and adaptive mesh-refinement techniques. Wiley & Teubner, 1996.

[78] T. Vogel, M. Th. Van Genuchten, and M. Cislerova. Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions. *Advances in Water Resources*, 24(2):133–144, 2000.

[79] M Vohralík. *A posteriori error estimates, Stopping Criteria, and Inexpensive Implementations for Error Control and Efficiency in Numerical Simulations.* Habilitation à diriger les recherches, Université Pierre et Marie Curie, 2010.

[80] O.C. Zienkiewicz and J.Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *International Journal for Numerical Methods in Engineering*, 24(2):337–357, 1987.

Méthodes numériques pour les écoulements en milieu poreux : estimations *a posteriori* et stratégie d'adaptation

Résumé On cherche à améliorer l'efficacité de la résolution numérique de l'équation de Richards, qui est une équation parabolique non linéaire utilisée dans la modélisation d'écoulements souterrains. Dans une première partie, on propose une discrétisation DDFV, valable sur maillages généraux, couplée au schéma BDF2 pour la discrétisation du terme instationnaire. Des tests numériques confirment l'ordre élevé de la méthode, ainsi que sa stabilité dans différentes configurations. La deuxième partie s'articule autour des estimations *a posteriori* par la méthode des flux équilibrés. On obtient une borne supérieure garantie de l'erreur en norme duale, qui fait intervenir des estimateurs intégrés en espace et en temps. Cette borne repose sur une relation d'équilibrage de flux, qui nécessite en pratique de reconstruire des approximations pertinentes des flux continus à partir de la solution approchée. De telles reconstructions adaptées à la discrétisation DDFV-BDF2 sont présentées. Elles incluent un terme de correction spécifique au schéma DDFV, et une réécriture de la formule BDF2 à pas variable sous la forme d'un schéma à un pas. Enfin, on applique numériquement l'estimation précédente en proposant, à maillage fixé, un algorithme d'adaptation du critère d'arrêt des linéarisations et du pas de temps qui vise à équilibrer les différentes sources d'erreur. On analyse l'influence des paramètres de l'algorithme, et on observe le gain en termes de nombre d'itérations de linéarisation et de temps CPU par rapport à une simulation classique.

Mots-clés Équation de Richards, schéma DDFV, formule BDF2, estimations d'erreur *a posteriori*, flux equilibrés, algorithme adaptatif.

Numerical methods for subsurface flows : *a posteriori* error estimates and adaptive strategy

Abstract This work is devoted to improving the efficiency of the numerical resolution of the Richards equation, which is a nonlinear parabolic equation used for the simulation of subsurface flows. The first part is dedicated to the derivation of a DDFV scheme, which is valid on general meshes, coupled with the BDF2 formula used for the discretization of the instationary term. Numerical tests confirm the high-order accuracy of the method, as well as its stability in a variety of configurations. In the second part, we derive *a posteriori* estimates using the equilibrated fluxes method. A guaranteed upper bound is achieved, involving space-time estimators. This bound relies on an equilibrated fluxes relation built from relevant approximations of the continuous fluxes, reconstructed from the approximate solution. We present how to perform such reconstructions adapted to our DDFV-BDF2 discretization. They involve a correction term specific to the DDFV scheme used, as well as a rephrasing of the variable BDF2 formula under the form of a one-step scheme. Lastly, we apply the aforementioned estimate in some numerical tests by developing an adaptive algorithm, which, on a fixed mesh, aims to equilibrate the various error sources by adapting the linearization stopping criterium and the time step. We analyze the influence of the parameters of this algorithm, and observe the number of linearization iterations and CPU time gained as compared to a classical simulation.

Keywords Richards equation, DDFV scheme, BDF2 formula, *a posteriori* error estimates, equilibrated fluxes, adaptive algorithm.