Schémas volumes finis multipoints pour grilles non orthogonales

Léo Agélas

UNIVERSITÉ PARIS-EST

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS-EST

Discipline : **MATHÉMATIQUES**

présentée et soutenue publiquement par

Léo AGÉLAS

le 22 Décembre 2009

Schémas volumes finis multipoints pour grilles non orthogonales

---

devant le jury composé de:

M. Robert EYMARD,	Directeur de thèse,
M. François DUBOIS,	Rapporteur,
Mme. Laurence HALPERN,	Rapporteur,
M. Alexandre ERN,	Examinateur,
M. Pascal OMNES,	Examinateur,
M. Roland MASSON,	Examinateur,

Remerciements

Je souhaite remercier en premier lieu mon directeur de thèse, Robert Eymard (Université Paris Est). Merci de m'avoir fait découvrir ce monde de recherche où s'entremêlent les défis, l'intuition, la créativité et l'esprit de concision. Merci d'avoir encadré mon travail et d'avoir toujours su susciter en moi ce goût pour la recherche avec ses idées originales et ses remarques pertinentes. Je lui suis également reconnaissant pour sa disponibilié, ses qualités pédagogiques et scientifiques. J'ai beaucoup appris à ses côtés et je l'en remercie.

Je remercie mon responsable scientifique de thèse IFP, Roland Masson pour m'avoir proposé ce sujet de thèse aussi passionnant, de m'avoir laissé une grande liberté dans mes recherches et d'avoir suivi mes travaux avec beaucoup d'intérêt. Je le remercie pour tout le temps qu'il m'a consacré et également d'avoir accepté d'être membre de mon jury.

Je voudrais remercier les rapporteurs de cette thèse, Laurence Halpern (Université Paris 13) et François Dubois (CNAM) pour l'intérêt qu'ils ont porté à mon travail et particulièrement Laurence Halpern pour ses suggestions et remarques au sujet de mon mémoire.

Je remercie Alexandre Ern (ENPC, CERMICS) pour avoir accepté la présidence de mon jury et Pascal Omnes (CEA) à la fois pour sa participation au jury et aussi pour ses remarques pertinentes sur mon mémoire.

Je remercie Jean Brac (IFP), Thibaut Mouton (à l'époque thésard IFP) pour leur aide dans l'élaboration de maillages 3D pour l'étude des schémas volumes finis. J'adresse également mes remerciements à Ivan Kapyrin (à l'époque post-doc IFP) pour son aide précieuse dans l'étude des schémas volumes finis sur des cas tests 3D. Je remercie Sylvie Pégaz (IFP) et Abdallah Bénali (IFP) pour leur aide et leurs conseils.

Je remercie Daniele Di Pietro (IFP) avec qui j'ai souvent eu l'occasion de discuter des schémas volumes finis, confronter mes idées.

Je remercie également Jérome Droniou pour ses conseils et notre fructueuse collaboration.

Mes remerciements les plus affectueux vont à mes frères et soeurs pour leur soutien, à ma mère qui m'a tant donné et m'a donné la possibilité de faire des études supérieures, à ma fille mon petit rayon de soleil et à ma femme qui au cours

de ces trois longues années d'études a été d'un grand soutien, d'une grande patience et d'une grande aide, sans lesquels je ne serai pas arrivé à bout de ce projet.

Table des matières

Table des figures

Chapitre I

Introduction

Le pétrole est une roche liquide carbonée, ou huile minérale. L'exploitation de cette énergie fossile a été, au 20ème siècle, l'un des piliers de l'économie industrielle, car le pétrole a fourni la quasi totalité des carburants liquides. En ce début de 21ème siècle, la question d'une énergie de substitution ne semble pas encore résolue. Le pétrole est le produit de l'histoire géologique d'une région, et particulièrement de la succession de trois conditions :

- L'accumulation de matière organique, végétale essentiellement :
  - sur notre planète, il y a, en permanence, des organismes qui meurent. Ces organismes (nous inclus) sont composés pour l'essentiel de carbone, d'hydrogène, d'azote et d'oxygène et à leur mort, ces délicats assemblages sont cassés - on parle de décomposition et l'essentiel est recyclé et réutilisé rapidement par la biosphère. Cependant, une petite minorité de la matière morte sédimente, c'est-à-dire qu'elle s'accumule par gravité et est enfouie au sein de la matière minérale. Le processus de sédimentation est un processus permanent au fond des océans et des lacs, qui produit certes peu d'effets à l'échelle d'une vie humaine, mais est d'une importance capitale à l'échelle des temps dits géologiques (quelques millions d'années à quelques milliards d'années).
- Sa maturation en hydrocarbures :
  - Au fur et à mesure que des couches de sédiments se déposent au-dessus de cette strate riche en matières organiques, la roche-mère ou roche-source, croît en température et en pression. Dans ces conditions, la matière organique se transforme en kérogène, un extrait sec disséminé dans la roche sous forme de petits grumeaux. Si la température devient suffisante (le seuil est à au moins 50°C, généralement plus selon la nature de la roche

et du kérogène), le kérogène subit une décomposition d'origine thermique, la pyrolyse.

- Dans un premier temps, cette décomposition expulse de l'eau et du CO_2 du kérogène. Ensuite, les températures devenant croissantes avec le temps, le kérogène expulse des hydrocarbures liquides (c'est le fameux *pétrole*, que l'on appelle encore huile) et du gaz naturel. On appelle cette expulsion la migration primaire dans le jargon pétrolier.

- Son piégeage :

  - Après avoir été expulsés de la roche mère, les hydrocarbures et le gaz (et l'eau) entament alors ce que l'on appelle une migration secondaire : ils suintent le long des couches perméables qui jouxtent les couches de roche mère, en se dirigeant vers la surface sous l'effet de la pression des couches de sédiment situées au-dessus, mais une minime quantité est piégée : elle se retrouve dans une zone perméable (généralement du sable, des carbonates ou des dolomites) qu'on appelle la roche-réservoir, et ne peut s'échapper à cause d'une couche imperméable au dessus d'elle (composée d'argile, de schiste et de gypse) formant une structure-piège .
  - Il existe plusieurs types de pièges. Les plus grands gisements sont en général logés dans des pièges anticlinaux (tournant leur courbure vers le bas).

Le pétrole et le gaz se trouvent donc dans les profondeurs du sous-sol, où ils se sont accumulés pendant des millions d'années. Pour y accéder, il n'existe qu'un seul moyen : forer des puits jusqu'aux gisements, souvent à plusieurs kilomètres sous terre. Depuis la surface, on peut essayer de prévoir où se trouvent les gisements et émettre des hypothèses. Mais on n'est jamais sûr qu'ils existent tant qu'on ne les a pas atteints avec un forage, tant que les hypothèses n'ont pas été vérifiées et le coût de leur vérification est élevé.

Aussi, la modélisation de bassins sédimentaires et de réservoir est de plus en plus utilisée par les compagnies pétrolières pour conduire leurs activités d'exploration afin d'en réduire les risques et les coûts. La modélisation de bassins sédimentaires vise à simuler, à des échelles de temps et d'espaces géologiques, la migration de l'huile et du gaz dans un bassin sédimentaire afin de prédire la localisation, la qualité et la quantité d'huile piégée dans le réservoir. La modélisation de réservoir quant à elle a pour but d'optimiser l'emplacement des puits et de prédire la production d'un réservoir.

L'impact environnemental le plus préoccupant du pétrole est l'émission de dioxyde de carbone résultant de sa combustion comme carburant. Pour limiter sa contribution à l'acidification des milieux et aux modifications climatiques, le stockage géologique du CO_2 est envisagé comme une des formes possibles de séquestration

du carbone alors que les forêts, tourbières et puits de carbone ne suffisent plus à absorber les émissions humaines de CO_2. Dans ce contexte, la modélisation de l'injection et du stockage du CO_2 a un rôle important à jouer, notamment pour l'étude des mécanismes d'écoulement, la caractérisation des sites et l'étude de la sécurité à long terme du stockage.

Ces trois exemples (simulation de la migration des hydrocarbures, simulation de leur production, simulation de l'enfouissement du CO_2) illustrent le besoin de pouvoir simuler des écoulements darcéens multiphasiques de fluides en milieux poreux et hétérogènes c'est à dire simuler l'écoulement de fluides (huile, gaz) dans des milieux poreux (bassins sédimentaires) où se sont accumulées pendant des dizaines de millions d'années des couches superposées de différentes roches (des sables, des carbonates ou des dolomites) tout en permettant de connaître la qualité et la quantité de fluides présents dans le bassin.

La précision de ces simulations dépend de la qualité des données, de la représentation des phénomènes physiques, de l'efficacité des algorithmes numériques. Une simulation numérique efficace exige dès lors de concevoir des schémas de discrétisation précis et robustes qui devront être adaptés aussi bien à la complexité du système des équations qu'à à la complexité géologique des milieux poreux hétérogènes. Pour cela, le maillage doit exactement décrire les caractéristiques stratigraphiques et structurales complexes du bassin telles que les couches stratigraphiques hétérogènes, les canaux, les érosions, et les failles. Il doit être localement raffiné autour des puits et être adapté aux modèles de puits complexes tels que les puits déviés ou multi-branches. Ceci exige l'utilisation de maillages complexes qui mélangent par exemple des grilles hexaédriques structurées présentant des contrastes de dimension (longueur, largeur, épaisseur) assez grands pour suivre la géométrie des couches avec des maillages localement non structurés.

Un ingrédient principal du schéma de discrétisation est la discrétisation du flux de diffusion intervenant dans la loi de Darcy qui régit les écoulements en milieux poreux, donné par $-K\nabla P_\alpha$ où K est le tenseur de perméabilité et ∇P_α, $\alpha = w, o, g$ est le gradient de pression de la phase eau, huile ou gaz. Les directions principales du champ de perméabilité suivent principalement celles des couches stratigraphiques, en présence d'hétérogénéités habituellement fortes entre les couches. En outre, les caractéristiques géologiques telles que les canaux, failles, mènent à des hétérogénéités fortes du champ de perméabilités. Les caractéristiques géologiques à l'échelle des hétérogénéités les plus fines ou des fracturations sont prises en compte, suivant le maillage utilisé pour la simulation des écoulements, par des tenseurs de perméabilités pleins non alignés avec la direction de la grille, avec dans certains cas de grands rapports d'anisotropie.

Pour toutes ces raisons, le schéma de discrétisation doit s'adapter à des maillages polyédriques généraux et à des tenseurs de perméabilités pleins présentant de fortes hétérogénéités, de grand rapports d'anisotropie et non alignés avec les direc-

tions du maillage. Dans ce contexte, l'approximation des flux de diffusion par le schéma à deux points couramment utilisé dans la plupart des simulateurs commerciaux conduit à de grandes erreurs numériques. De nombreuses méthodes existent pour obtenir des approximations plus précises des flux de diffusion sur des mailles polygonales et polyédriques et pour des milieux anisotropes hétérogènes, comme la méthode des éléments finis mixtes hybrides (cf [55]), la méthode des différences finies mimétiques(cf [26]), la méthode des volumes finis hybrides (cf [43]) et la méthode des MPFA (Multi Point Flux Approximation) introduite au milieu des années 90 (cf [4, 36]). Néanmoins, ces méthodes ne présentent pas toujours les propriétés souhaitées de précision, de stabilité et d'efficacité en terme de temps calcul.

Les objectifs de notre travail se sont inscrits dans ces directions en recherchant plus précisément des schémas numériques présentant les caractéristiques suivantes :

- linéarité des flux en fonction des inconnues,
- conservativité des flux, (c'est le principe de base des méthodes de volumes finis)
- consistance des flux, (cette notion devra être précisée)
- stabilité c'est à dire que la solution doit être bornée pour une certaine norme par la norme du terme source (dans le cadre défini dans [46], cela est obtenu à partir de la notion de coercivité, qui permet d'assurer l'existence, l'unicité, la stabilité de la solution et la convergence du schéma.

Pour un certain nombre de schémas présentés dans la littérature, il n'existe pas de preuve mathématique de leur convergence. Dans le présent travail, nous étudierons la convergence des schémas dans le cas des maillages généraux avec des hypothèses habituelles de régularité de forme, et de champs de perméabilités hétérogènes anisotropes discontinues ayant des valeurs propres uniformément bornées par au-dessus et en-dessous.

Nous rechercherons également la vérification de propriétés additionnelles. Le schéma de discrétisation devra fournir des solutions assez précises pour des coefficients de diffusion constants par maille même dans le cas de grands sauts et de grilles grossières fréquemment rencontrées dans les applications de l'industrie pétrolière.

Cette propriété est souvent satisfaite par l'utilisation additionnelle d'inconnues intermédiaires à chaque face interne du maillage afin de capturer la singularité de la solution à cet endroit. Ensuite, le stencil du flux traversant une face devra être compact : cela signifie que seules les proches cellules voisines et les faces de bords voisines de la face en question devront intervenir dans l'approximation du flux. Cette propriété permet des gains très significatifs de temps calcul lors de la résolution, sur des ordinateurs à architecture parallèle, des systèmes linéaires résultant de la discrétisation et de la linéarisation des écoulements darcéens multiphasiques.

Les méthodes numériques présentées dans ce mémoire réalisent ainsi chacune un certain nombre d'avancées dans ces directions. Nous avons développé chacune d'elle pour répondre plus particulièrement à l'une de ces propriétés recherchées, dont la prépondérance peut varier selon le contexte industriel. Pour les présenter nous avons divisé le mémoire en cinq parties :

- Présentation du problème modèle, espaces discrets et cadre mathématique nécessaires à l'analyse.
- Des exemples de schémas non symétriques.
- Des exemples de schémas symétriques.
- Synthèse et résultats numériques 2D, 3D.
- Annexe

Plus précisément le contenu de ces parties est :

- Dans la première partie, on expose le problème modèle que l'on va considérer, on introduit une discrétisation du domaine, des espaces d'approximation, puis un cadre mathématique pour l'analyse des schémas volumes finis inspiré de celui introduit dans [42,43]. On tire alors une preuve de convergence pour tous les schémas volumes finis, ce qui étend le résultat obtenu dans [45] aux schémas non symétriques, puis une estimation d'erreur est donnée. L'idée essentielle introduite dans [26, 46] est de mimer le problème continu II.1 exprimé en formulation faible II.5 afin d'obtenir la convergence des solutions discrètes des schémas vers la solution faible du problème II.5 sur maillages généraux, pour des tenseurs de diffusion anisotrope, hétérogène. Cette partie met en lumière les hypothèses nécessaires à la preuve de convergence, c'est à dire une hypothèse de stabilité et une hypothèse de consistance du schéma. L'hypothèse de consistance jusqu'ici reposait sur des fonctions tests appartenant à C^2. Cependant, il arrive que l'on ne puisse pas vérifier cette hypothèse pour des fonctions tests appartenant à C^2 même pour des tenseurs de diffusion régulier par morceaux, comme c'est le cas pour le schéma MPFA L introduit dans [6]. L'idée originale a été d'introduire comme espace des fonctions tests l'ensemble des fonctions régulières par morceaux suivant les zones de régularité du tenseur et de flux continus puis de montrer que l'hypothèse de consistance peut être reformulée en considérant cet espace de fonctions tests. Cette idée nous a permis alors d'obtenir la consistance du schéma L. Ces travaux font l'objet d'un article en préparation.
- Dans la deuxième partie, on fournit trois exemples de schémas non symétriques, dont un original. Le premier est le schéma MPFA O introduit dans [1], souvent

utilisé dans les codes industriels mais pour lequel il n'y avait pas de résultat de convergence sur des maillages généraux polygonaux et polyédriques, et aucun résultat de convergence prenant en compte des coefficients de diffusions discontinus qui sont essentiels dans les applications des compagnies pétrolières. L'avantage de ce schéma est d'être un schéma centré aux mailles donnant des flux au sens classique des schémas volumes finis avec en plus un stencil compact. A partir du cadre mathématique introduit dans la première partie, nous établissons la preuve de convergence du schéma MPFA O sur maillages généraux pour des tenseurs de diffusion discontinus pouvant même appartenir à L^∞ sous l'hypothèse de stabilité, ce qui constitue un résultat nouveau. Le deuxième schéma est le schéma G tiré d'une famille de schémas que nous avons introduits et qui généralisent la méthode MPFA L introduite dans [6]. L'idée est d'écrire le flux traversant une face comme une moyenne pondérée de flux de type L. Un choix approprié des poids permet d'améliorer la stabilité de la méthode. Le troisième schéma est Cell-Gradient method qui a montré sa robustesse sur les différents cas tests numériques, ce schéma a été conçu dans le but d'obtenir un stencil réduit par volume contrôle. La preuve de convergence pour tous ces schémas est obtenue via le cadre mathématique. Dans cette partie, on constate que l'inconvénient de ces schémas est que leur convergence dépend de la condition de stabilité qui ne peut être vérifiée que numériquement. Ces travaux ont fait l'objet de trois articles, l'un publié [15] et les deux autres soumis ([10], [12, §3.2]).

- Dans la troisième partie, on introduit une famille de schémas symétriques pour laquelle la condition de stabilité est vérifiée uniquement sous des hypothèses de régularité du maillage. A partir de cette famille, on a proposé deux schémas, le premier est le schéma VFSYM qui a l'avantage par rapport au schéma SUCCES de ne conduire qu'à des flux aux faces du maillage. Le second schéma est le schéma DIOPTRE qui réunit toutes les caractéristiques souhaitées pour un schéma volumes finis, mais dont la généralisation en 3D reste à étudier. Ces travaux ont fait l'objet de deux articles publiés ([13], [14]).

- Dans la quatrième partie, une synthèse est faite et une comparaison numérique en 2D et 3D est faite entre les différents schémas étudiés.

- La cinquième partie constitue l'annexe de la thèse. On donne deux autres exemples de reconstruction de gradients discrets pour la définition du schéma MPFA O, la preuve de densité de l'espace de fonctions tests dans $H_0^1(\Omega)$ suivant les idées de [34] et nous proposons un cadre unifié d'analyse pour de nombreuses méthodes non conformes, en particulier, les méthodes volume fini (FV) et discontinu Galerkin (dg). Bien que les analogies entre ces deux familles de méthodes de discrétisation ont souvent été mentionnées, l'analyse

unifiée actuelle est, au meilleur de notre connaissance, nouvelle. Le travail sur l'analyse unifiée a fait l'objet d'un article soumis [12].

Chapitre II

Résultats généraux pour l'étude des schémas volumes finis

Un ingrédient principal des schémas de discrétisation pour la simulation des écoulements des hydrocarbures en milieux poreux est la discrétisation du flux de diffusion intervenant dans la loi de Darcy qui régit les écoulements en milieux poreux, typiquement donné par $-K\nabla P_\alpha$ où K est le tenseur de perméabilité et ∇P_α, $\alpha = w, o, g$ est le gradient de pression de la phase eau, huile ou gaz. Comme évoqué dans l'introduction, le tenseur de perméabilité est un tenseur hétérogène, plein non aligné avec la direction de la grille, avec dans certains cas de grands rapports d'anisotropie. Pour obtenir des simulations précises, une bonne approximation des flux de diffusion sur des maillages polygonaux et polyédriques est alors primordiale. Dans ce chapitre, nous prouvons la convergence des schémas volumes finis pour des problèmes de diffusion hétérogènes et anisotropes. Les résultats nouveaux abordés dans ce chapitre, ont été de fournir un cadre mathématique inspiré de celui introduit dans [42,43], une preuve de convergence pour les schémas volumes finis sur maillages généraux avec tenseurs de diffusion L^∞, ce qui est essentiel dans de nombreuses applications et une étude par estimation d'erreur pour des tenseurs dans $W^{1,\infty}$ et cela sans régularité supplémentaire de la solution du problème continu. Ces travaux font l'objet d'un article en préparation.

- Dans la première partie de ce chapitre, nous introduirons le problème modèle considéré (II.1), caractérisant notre recherche d'une bonne approximation du flux de diffusion.

- Dans la deuxième partie de ce chapitre, nous introduirons la discrétisation et les espaces discrets nécessaires à la définition des schémas volumes finis.

- Dans la troisième partie, nous ferons l'étude de la convergence des schémas volumes finis à travers sa formulation variationnelle, puis une étude par estimation d'erreur.

II.1 The continuous problem

The problem is to find an approximation of $\bar{u}$, weak solution to the following equation:

$$\begin{cases} -\mathrm{div}\Lambda\nabla\bar{u} = f & \text{in } \Omega, \\ \overline{u} = 0 & \text{on } \partial\Omega, \end{cases} \tag{II.1}$$

under the following assumptions:

$$\Omega \text{ is an open bounded connected polygonal subset of } \mathbb{R}^d,\ d \in \mathbb{N}^\star, \tag{II.2}$$

$$\begin{array}{l} \Lambda \text{ is a measurable function from } \Omega \text{ to } \mathcal{M}_d(\mathbb{R}), \\ \text{where } \mathcal{M}_d(\mathbb{R}) \text{ denotes the set of } d \times d \text{ matrices,} \\ \\ \text{such that for a.e. } x \in \Omega, \Lambda(x) \text{ is symmetric,} \\ \text{and the set of its eigenvalues is included in } [\alpha(x), \beta(x)] \end{array} \tag{II.3}$$

where $\alpha, \beta \in L^\infty(\Omega)$ are such that $0 < \alpha_0 \leq \alpha(x) \leq \beta(x) \leq \beta_0$ for a.e. $x \in \Omega$, and

$$f \in L^2(\Omega). \tag{II.4}$$

Définition II.1.1 (Weak solution) *Under hypotheses (II.2)-(II.4), $\bar{u}$ is a weak solution of (II.1) if*

$$\begin{cases} \bar{u} \in H_0^1(\Omega), \\ \displaystyle\int_\Omega \Lambda(x)\nabla\bar{u}(x)\cdot\nabla v(x)\mathrm{d}x = \int_\Omega f(x)v(x)\mathrm{d}x, \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{II.5}$$

It is known that there exists a unique weak solution of (II.1).

In what follows, we shall provide the definition of a finite volume (FV) discretization of problem [II.1] as well as an analysis framework covering fairly general (possibly nonconforming) polygonal meshes.

II.2 Discretization for Finite volume scheme

Définition II.2.1 (Admissible space discretization) *Let Ω be an open bounded subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$, and $\partial\Omega = \overline{\Omega}\setminus\Omega$ its boundary. An admissible finite volume discretization of Ω, denoted by $\mathcal{D}$, is given by $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P}, \mathcal{V})$, where:*

- *$\mathcal{T}$ is the set of control volumes which corresponds to a finite family of non empty connected open disjoint subsets of Ω such that $\overline{\Omega} = \cup_{K\in\mathcal{T}}\overline{K}$ (see figure II.1). For any $K \in \mathcal{T}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K and $\mathrm{m}_K > 0$ denote the measure of K.*

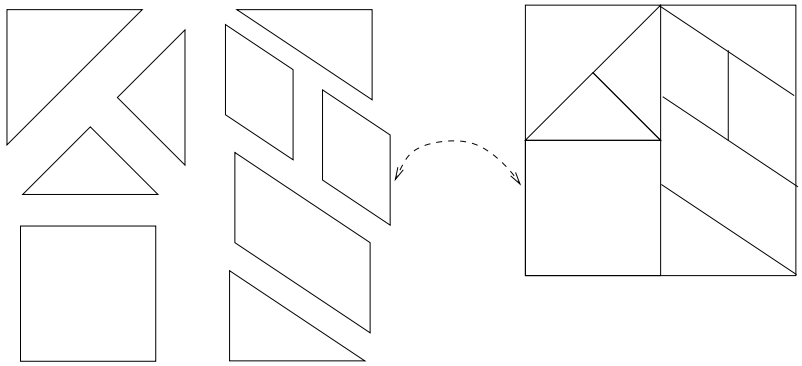


Figure II.1: $\mathcal{T}$, the set of control volumes.

- *$\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ such that, for all $\sigma \in \mathcal{E}$, σ is a non empty open subset of a hyperplane of $\mathbb{R}^d$, which has a measure $\mathrm{m}_\sigma > 0$ for the (d-1)-dimensional measure, for example in 2D, it is a set of segments (see figure II.2).*

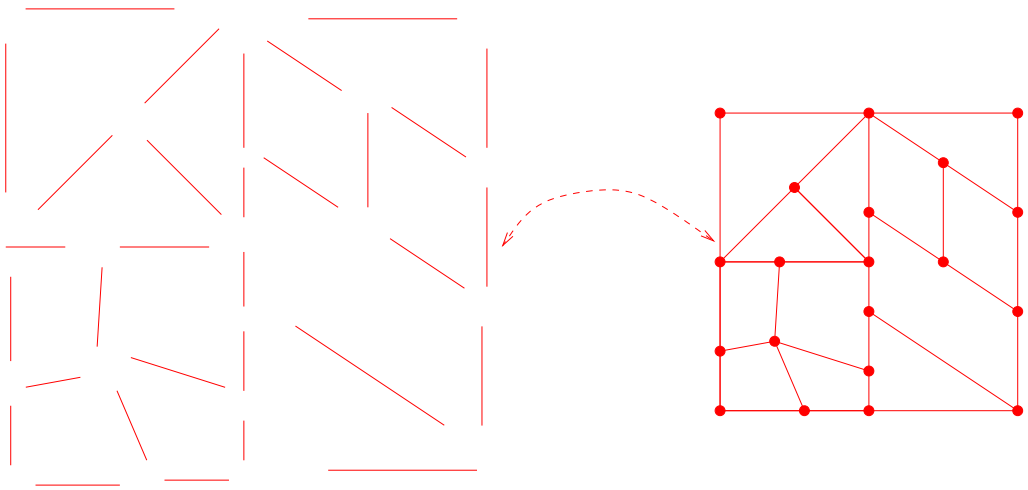


Figure II.2: $\mathcal{E}$, a set of segments in 2D, polygons in 3D

We assume that, for all $K \in \mathcal{T}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$ (see figure II.3).

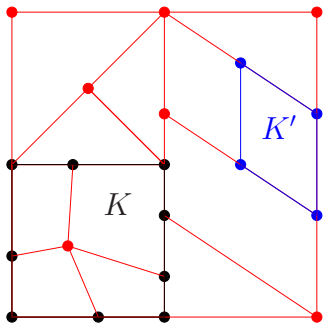
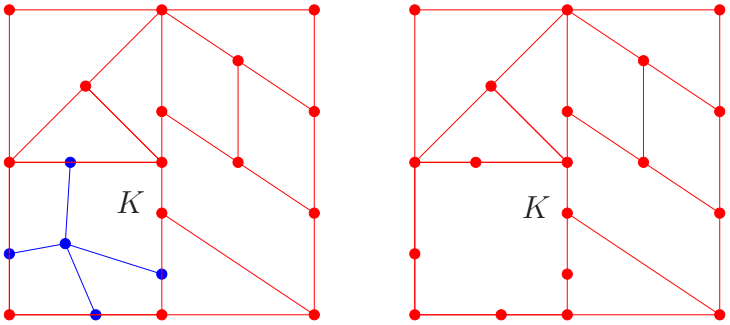


Figure II.3: $\mathcal{E}_K$(resp $\mathcal{E}_{K'}$) the set of edges of $K \in \mathcal{T}$ (resp $K' \in \mathcal{T}$).

Notice that the set $\cup_{K \in \mathcal{T}} \mathcal{E}_K$ is included in $\mathcal{E}$ (see figure II.4).



(a) $\mathcal{E}$ (b) $\cup_{K \in \mathcal{T}} \mathcal{E}_K$

Figure II.4: $\cup_{K \in \mathcal{T}} \mathcal{E}_K \subset \mathcal{E}$.

Usually, the set $\cup_{K \in \mathcal{T}} \mathcal{E}_K$ is known as the edges of the mesh, for the sake of simplicity, an element $\sigma \in \mathcal{E}$ is called edge.
The set of interior (resp. boundary) edges is denoted by $\mathcal{E}_{\text{int}}$ (resp. $\mathcal{E}_{\text{ext}}$), that is $\mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ (resp. $\mathcal{E}_{\text{ext}} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$).
For all $K \in \mathcal{T}$, we denote by $\mathcal{E}_{\text{int},K}$, the set of edges inside the control volume K which corresponds to the set $\{\sigma \in \mathcal{E}_{\text{int}} \backslash \mathcal{E}_K, \overline{\sigma} \subset \overline{K}\}$ (in figure II.4-(a), $\mathcal{E}_{\text{int},K}$ is the set of blue edges).
We then denote by $\mathcal{T}_\sigma$ the set of control volume sharing the edge σ which corresponds to the set $\{K \in \mathcal{T}, \sigma \subset \overline{K}\}$. For all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, we assume that either $\mathcal{T}_\sigma$ has exactly one element and then $\sigma \in \partial\Omega$ (boundary edge) or $\mathcal{T}_\sigma$ has exactly two elements and then $\sigma \in \mathcal{E}_{\text{int}}$ (interior edge). For all $\sigma \in \mathcal{E}$, we denote by x_σ the barycentre of σ.

- *$\mathcal{P}$ is a finite family of points of Ω indexed by $\mathcal{T}$, denoted by $\mathcal{P} = (x_K)_{K\in\mathcal{T}}$ such that $x_K \in K$ and K is strictly star-shaped with respect to x_K (see figure II.5).*

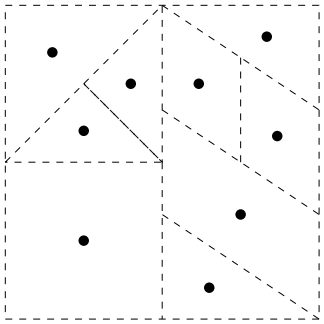


Figure II.5: $\mathcal{P}$, the centers of the mesh.

- *$\mathcal{V}$ is a finite family of points ("the vertices of the mesh") such that $\mathcal{V} \subset \cup_{K\in\mathcal{T}} \left(\cup_{H_K\subset\mathcal{E}_K,\,\mathrm{card}(H_K)\geq d} \left(\cap_{\sigma\in H_K}\overline{\sigma}\right)\right)$ (see figure II.6).*

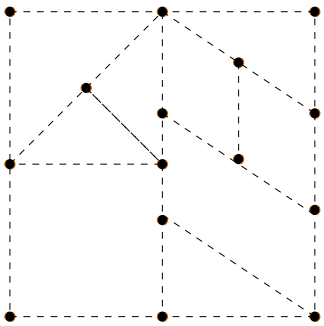
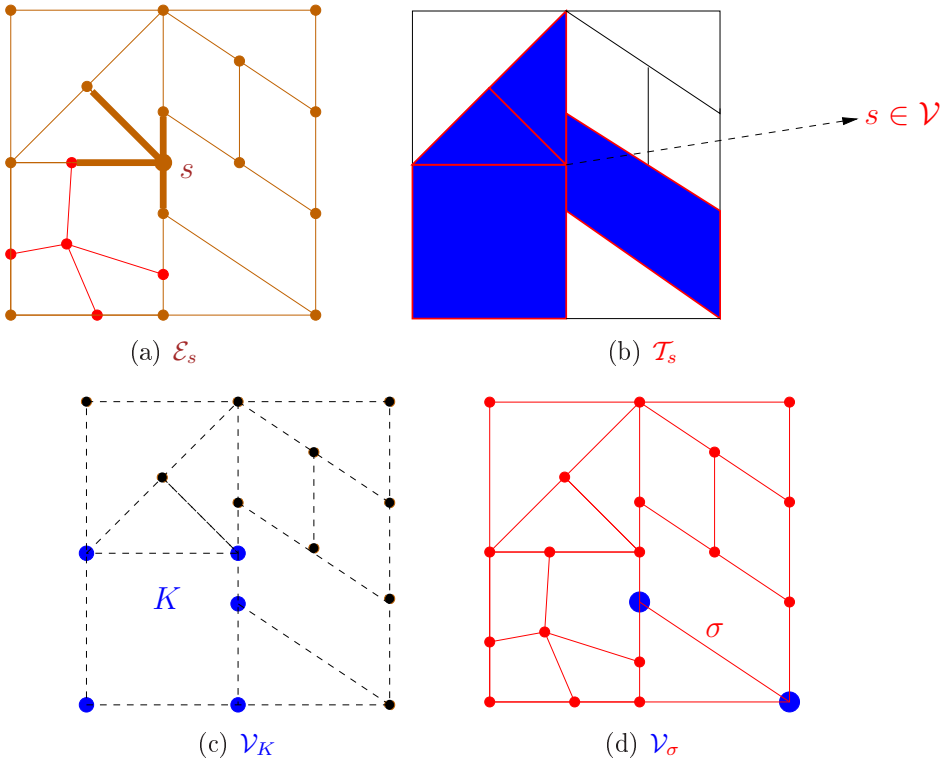


Figure II.6: $\mathcal{V}$, the vertices of the mesh.

For all $s \in \mathcal{V}$, we denote by $\mathcal{E}_s$ the set $\{\sigma \in \mathcal{E} \,|\, s \in \overline{\sigma}\}$ which is the set of edges sharing the vertex s (see figure II.7-(a)).
For all $s \in \mathcal{V}$, we denote by $\mathcal{T}_s$ the set $\{K \in \mathcal{T} \,|\, s \in \overline{K}\}$ which is the set of control volumes sharing the vertex s (see figure II.7-(b)).
For all $K \in \mathcal{T}$, the set $\mathcal{V}_K$ stands for $\{s \in \mathcal{V} \,|\, s \in \overline{K}\}$ which is the set of vertices belonging to $\overline{K}$ (see figure II.7-(c)), and for all $\sigma \in \mathcal{E}$ the set $\mathcal{V}_\sigma$ stands for $\{s \in \mathcal{V} \,|\, s \in \overline{\sigma}\}$, which is the set of vertices sharing the edge σ. (see figure II.7-(d)).

The following notations are used. For any point $x \in \mathbb{R}^d$, we write $x = (x^{(i)})_{i=1,\ldots,d}$. The size of the discretization is defined by:

$$h_{\mathcal{D}} = \sup\{\mathrm{diam}(K), K \in \mathcal{T}\}.$$



(a) $\mathcal{E}_s$ (b) $\mathcal{T}_s$

(c) $\mathcal{V}_K$ (d) $\mathcal{V}_\sigma$

Figure II.7:

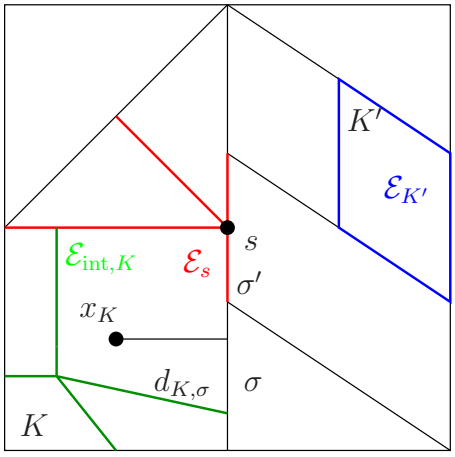


Figure II.8: Other Notations

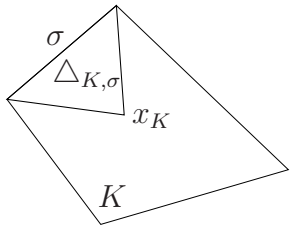


Figure II.9: Cone $\triangle_{K,\sigma}$ with vertex x_K and basis σ for $d = 2$.

For all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K.

Shape regularity of the mesh:

Hypothesis 1 $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ *is a family of meshes matching Definition II.2.1 such that*

(R1) For all $K \in \mathcal{T}_n$*, there exists a family of non-negative real* $(d_{K,\sigma})_{\sigma\in\mathcal{E}_K\cup\mathcal{E}_{\mathrm{int},K}}$ *such that,*

$$\sum_{\sigma\in\mathcal{E}_K\cup\mathcal{E}_{\mathrm{int},K}} \mathrm{m}_\sigma d_{K,\sigma} \le d\mathrm{m}_K. \qquad \text{[II.6]}$$

(R2) there exist $0 < \varrho_1 < +\infty$ *and* $0 < \varrho_2 < +\infty$ *independent of* n *s.t.*

$$\min_{\sigma\in\mathcal{E}_K,\, K\in\mathcal{T}_n} \frac{d_{K,\sigma}}{\mathrm{diam}(K)} \ge \varrho_1, \quad \min_{\sigma\in\mathcal{E}_{n,\mathrm{int}},\, \mathcal{T}_\sigma=\{K,L\}} \frac{\min(d_{K,\sigma}, d_{L,\sigma})}{\max(d_{K,\sigma}, d_{L,\sigma})} \ge \varrho_2. \qquad \text{[II.7]}$$

(R3) the number of faces sharing one node remains bounded as the mesh is refined, i.e. , there exists $\varrho_3 \in \mathbb{N}^*$ *such that*

$$\sup_{n\in\mathbb{N}} \max_{s\in\mathcal{V}_n} \mathrm{card}\mathcal{E}_s \le \varrho_3.$$

(R4) $\lim_{n\to+\infty} h_{\mathcal{D}_n} = 0.$

In what follows, when referring to a generic element $\mathcal{D}_n$ of an admissible family of discretizations $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$, the subscript n will be dropped for easiness of reading if no ambiguity arises. It is always assumed in the following that Hypotheses 1 hold.

Notations: For all vectors $x \in \mathbb{R}^d$, $d \in \mathbb{N}^*$, the Euclidean norm will be denoted by $|x| \stackrel{\text{def}}{=} \sqrt{x\cdot x}$; for all matrices $A \in \mathbb{R}^d \times \mathbb{R}^d$, $d \in \mathbb{N}^*$, we shall denote by $|A|$ the norm induced by the vector scalar product, *i.e.* , $|A| \stackrel{\text{def}}{=} \sup_{x\in\mathbb{R}^d} \frac{|Ax|}{|x|}$. The vector space of bounded linear operators from E to F will be denoted by $\mathcal{L}(E;F)$. For all $K \in \mathcal{T}$ and for all $\Phi \in L^1(K)$, we have set $\langle\Phi\rangle_K \stackrel{\text{def}}{=} \mathrm{m}_K^{-1} \int_K \Phi(x)\,\mathrm{d}x$. For all $\sigma \in \mathcal{E}$ and for all $\Phi \in L^1(\sigma)$, we have set $\langle\Phi\rangle_\sigma \stackrel{\text{def}}{=} \mathrm{m}_\sigma^{-1} \int_\sigma \Phi(x)\,\mathrm{d}x$. For vectorial functions, the notation has to be intended component-wise.

II.3 Discrete functional framework

The following definition introduces the space of piecewise constant functions on each cell K of the mesh that we need to describe our Finite Volume Schemes.

Définition II.3.1 *Let Ω be an open bounded subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^\star$. Let $\mathcal{D} = (\mathcal{T}, \mathcal{E}, \mathcal{P}, \mathcal{V})$ be an admissible finite volume discretization of Ω in the sense of definition II.2.1. We denote by $\mathbb{H}_\mathcal{T}(\Omega) \subset L^2(\Omega)$ the set of all $u \in L^2(\Omega)$ such that, for all $K \in \mathcal{T}$, there exists some real value denoted by $u_K \in \mathbb{R}$ such that $u(x) = u_K$ for a.e. $x \in K$. We then define $\gamma_\sigma u \in \mathbb{R}$ for all $\sigma \in \mathcal{E}$ by*

$$\begin{array}{l} \gamma_\sigma u = 0, \ \forall \sigma \in \mathcal{E}_{\rm ext}, \\ \gamma_\sigma u = u_K, \ \forall \sigma \in \mathcal{E}_{\rm int}, \ \text{with } \mathcal{T}_\sigma = \{K\}. \\ \dfrac{\gamma_\sigma u - u_K}{d_{K,\sigma}} + \dfrac{\gamma_\sigma u - u_L}{d_{L,\sigma}} = 0, \ \forall \sigma \in \mathcal{E}_{\rm int}, \ \text{with } \mathcal{T}_\sigma = \{K, L\}. \end{array} \tag{II.8}$$

The space $\mathbb{H}_\mathcal{T}(\Omega)$ is equipped with the following Euclidean structure. For $(v, w) \in (\mathbb{H}_\mathcal{T}(\Omega))^2$, we define the following inner product

$$[v, w]_\mathcal{T} = \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} (\gamma_\sigma v - v_K)(\gamma_\sigma w - w_K). \tag{II.9}$$

For $u \in \mathbb{H}_\mathcal{T}(\Omega)$, we define

$$\|u\|_\mathcal{T} = \left([u, u]_\mathcal{T}\right)^{1/2}.$$

We define by $\mathcal{H}_\mathcal{D} \subset \mathbb{R}^\mathcal{T} \times \mathbb{R}^\mathcal{E}$ the set of all $((u_K)_{K \in \mathcal{T}}, (u_\sigma)_{\sigma \in \mathcal{E}})$ such that $u_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\rm ext}$.
The space $\mathcal{H}_\mathcal{D}$ is equipped with the following Euclidean structure, for $(v, w) \in (\mathcal{H}_\mathcal{D})^2$, we define the following inner product,

$$[v, w]_\mathcal{D} = \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K \cup \mathcal{E}_{{\rm int},K}} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} (v_\sigma - v_K)(w_\sigma - w_K). \tag{II.10}$$

For $u \in \mathcal{H}_\mathcal{D}$, we define

$$\|u\|_\mathcal{D} = \left([u, u]_\mathcal{D}\right)^{1/2}.$$

For all $u \in \mathcal{H}_\mathcal{D}$, we denote by $P_\mathcal{T} u \in \mathbb{H}_\mathcal{T}(\Omega)$ the element defined by the values $(u_K)_{K \in \mathcal{T}}$.

For any partition $(\mathcal{F}, \overline{\mathcal{F}})$ of $\mathcal{E}$ and for a given family $(\Pi_\sigma)_{\sigma \in \overline{\mathcal{F}}}$ of linear forms on $\mathcal{H}_\mathcal{D}$ such that for all $u \in \mathcal{H}_\mathcal{D}$, $\Pi_\sigma u$ depends only on the values $(u_K)_{K \in \mathcal{T}}$, we denote by $\mathcal{H}_{\mathcal{D},\mathcal{F}} = \{v \in \mathcal{H}_\mathcal{D}$ such that $v_\sigma = \Pi_\sigma v$ for all $\sigma \in \overline{\mathcal{F}}\}$.

Denoting by $C_0(\overline{\Omega})$ the set of continuous functions which vanish on $\partial\Omega$, we define the interpolation operator $P_\mathcal{F} : \ C_0(\overline{\Omega}) \to \mathcal{H}_{\mathcal{D},\mathcal{F}}$, for all $\varphi \in C_0(\overline{\Omega})$ by $(P_\mathcal{F}\varphi)_K =$

$\varphi(x_K)$, for all $K \in \mathcal{T}$, $(P_{\mathcal{F}}\varphi)_\sigma = \varphi(x_\sigma)$ for all $\sigma \in \mathcal{F}$ and $(P_{\mathcal{F}}\varphi)_\sigma = \Pi_\sigma(P_{\mathcal{D}}\varphi)$ for all $\sigma \in \overline{\mathcal{F}}$, where $P_{\mathcal{D}}\varphi = \{\varphi(x_K)\}_{K\in\mathcal{T}}$.

Note that, thanks to (II.8), we have

$$\frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(\gamma_\sigma u - u_K)^2 + \frac{\mathrm{m}_\sigma}{d_{L,\sigma}}(\gamma_\sigma u - u_L)^2 = \min_{u_\sigma\in\mathbb{R}} \left(\frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(u_\sigma - u_K)^2 + \frac{\mathrm{m}_\sigma}{d_{L,\sigma}}(u_\sigma - u_L)^2 \right),$$

which implies

$$\|P_{\mathcal{T}} u\|_{\mathcal{T}} \leq \|u\|_{\mathcal{D}}, \ \forall u \in \mathcal{H}_{\mathcal{D}}. \qquad \text{[II.11]}$$

II.4 Finite volume scheme

A finite volume scheme is aimed to provide an approximation of the solution $\bar{u}$ of the system of equations (II.1) and an approximation of the fluxes $\int_\sigma \Lambda\nabla\bar{u}\cdot\mathbf{n}_{K,\sigma}$ where $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$. For this, thanks to Green's formula applied to Equation (II.1) on each cell $K \in \mathcal{T}$, we obtain:

$$-\sum_{\sigma\in\mathcal{E}_K} \int_\sigma \Lambda\nabla\bar{u}\cdot\mathbf{n}_{K,\sigma} = \int_K f. \qquad \text{[II.12]}$$

The flux $\int_\sigma \Lambda\nabla\bar{u}\cdot\mathbf{n}_{K,\sigma}$ is then approximated by a function $F_{K,\sigma}(u)$ depending on cell unknowns and face unknowns $u = ((u_K)_{K\in\mathcal{T}}, (u_\sigma)_{\sigma\in\mathcal{E}})$ such that the following discrete equation corresponding to (II.12) holds,

$$-\textstyle\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}(u) \ = \int_K f \text{ for all } K \in \mathcal{T}, \qquad \text{[II.13]}$$

In the spirit of the finite volume ideas, we may write the continuity for the discrete flux for all interior faces.

$$\begin{array}{lll} F_{K,\sigma}(u) + F_{L,\sigma}(u) & = 0 & \text{for } \sigma \in \mathcal{E}_{\mathrm{int}}, \text{ such that } \mathcal{T}_\sigma = \{K, L\}. \\ F_{K,\sigma}(u) & = 0 & \text{for } \sigma \in \mathcal{E}_{\mathrm{int}}, \text{ such that } \mathcal{T}_\sigma = \{K\}. \end{array} \qquad \text{[II.14]}$$

For the boundary faces, we have the condition,

$$u_\sigma = 0, \quad \text{for } \sigma \in \mathcal{E}_{\mathrm{ext}}. \qquad \text{[II.15]}$$

There are now $\mathrm{card}(\mathcal{T}) + \mathrm{card}(\mathcal{E})$ unknowns and equations. In some situations, it can be possible to express faces unknowns in terms of cell unknowns using the flux continuity equation (II.14) and then we can reduce the size of the linear system to solve it. We will notice also that the assumptions yielding convergence of a finite volume scheme under its hybrid form (where we have kept all the face unknowns) is more restrictive than the one where we have succeeded to express the face unknowns in terms of cell unknowns.

In order to study the convergence of finite volume scheme, we consider a more general discrete problem than (II.13)-(II.15).
Indeed, let $a_{\mathcal{D}}$ be a bilinear form defined from $\mathcal{H}_{\mathcal{D},\mathcal{F}} \times \mathcal{H}_{\mathcal{D},\mathcal{F}}$ to $\mathbb{R}$ and let us consider the following variational formulation given by

$$\text{Find } u \in \mathcal{H}_{\mathcal{D}} \text{ s.t. } a_{\mathcal{D}}(u,v) = \int_\Omega f\ P_{\mathcal{T}} v \,\mathrm{d}x, \text{ for all } v \in \mathcal{H}_{\mathcal{D}}. \qquad \text{[II.16]}$$

Then any finite volume scheme of type (II.13)-(II.15) can be written as a variational formulation of type (II.16), taking $a_{\mathcal{D}}$ defined for all $(u,v) \in \mathcal{H}_{\mathcal{D},\mathcal{F}} \times \mathcal{H}_{\mathcal{D},\mathcal{F}}$ by,

$$a_{\mathcal{D}}(u,v) = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K\cup\mathcal{E}_{\mathrm{int},K}} F_{K,\sigma}(u)(v_\sigma - v_K). \qquad \text{[II.17]}$$

Hence, we now focus on the study of the convergence of the solution of the variational formulation (II.16).

II.5 Convergence of Finite volume scheme

We introduce the two discrete gradients reconstructions $\nabla_{\mathcal{D}} \in \mathcal{L}(\mathcal{H}_{\mathcal{D}}; [\mathbb{H}_{\mathcal{T}}(\Omega)]^d)$ and $\widetilde{\nabla}_{\mathcal{D}} \in \mathcal{L}(\mathcal{H}_{\mathcal{D}}; [\mathbb{H}_{\mathcal{T}}(\Omega)]^d)$ s.t., for all $K \in \mathcal{T}$ and all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$,

$$(\nabla_{\mathcal{D}} v)_K \overset{\text{def}}{=} \nabla_{\mathcal{D}} v_{|K} = \frac{1}{\mathrm{m}_K} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma (v_\sigma - v_K)\mathbf{n}_{K,\sigma}. \qquad \text{[II.18]}$$

$$(\widetilde{\nabla}_{\mathcal{D}} v)_K \overset{\text{def}}{=} \widetilde{\nabla}_{\mathcal{D}} v_{|K} = \frac{1}{\mathrm{m}_K} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma (\gamma_\sigma v - v_K)\mathbf{n}_{K,\sigma}. \qquad \text{[II.19]}$$

Equation [II.6] together with Cauchy-Schwarz inequality yield

$$\|\nabla_{\mathcal{D}} v\|_{[L^2(\Omega)]^d} \le \sqrt{d}\|v\|_{\mathcal{D}} \quad \forall v \in \mathcal{H}_{\mathcal{D}}. \qquad \text{[II.20]}$$

$$\|\widetilde{\nabla}_{\mathcal{D}} v\|_{[L^2(\Omega)]^d} \le \sqrt{d}\|P_{\mathcal{T}} v\|_{\mathcal{T}} \quad \forall v \in \mathcal{H}_{\mathcal{D}}. \qquad \text{[II.21]}$$

Owing to (II.6), the following result can be deduced from [43, §5]:

Lemma 1 (Discrete Sobolev embeddings) *Let $\mathcal{D}$ be an element of a family of discretizations matching Definition II.2.1.*
Let $q \in [1,+\infty)$ if $d = 2$, and $q \in [1, 2d/(d-2)]$ if $d > 2$. Then, there exists a constant $\mathbb{C}_1 > 0$, depending only on Ω, q, ϱ_1 and ϱ_2 s.t.

$$\|u\|_{L^q(\Omega)} \le C_1 \|u\|_{\mathcal{T}} \quad \forall u \in \mathbb{H}_{\mathcal{T}}(\Omega).$$

Owing to Remark II.11, the following lemma can be deduced from [II.20] and the technique of proof of [43, Lemmata 5.6–5.7]:

Lemma 2 (Discrete Rellich theorem) *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a sequence of admissible discretizations matching Definition II.2.1 and s.t. $h_{\mathcal{D}_n} \to 0$ as $n \to \infty$, and let $\{v_n\}_{n\in\mathbb{N}}$ be a sequence of $\mathcal{H}_{\mathcal{D}_n,\mathcal{F}_n}$ s.t. there exists $C > 0$ with $\|v_n\|_{\mathcal{D}_n} \le C$ for all $n \in \mathbb{N}$. Then, there exist a subsequence of $\{v_n\}_{n\in\mathbb{N}}$ and a function $\bar{v} \in H_0^1(\Omega)$ s.t., as $n \to \infty$, (i) $P_{\mathcal{T}_n} v_n \to \bar{v}$ in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$); (ii) $\{\widetilde{\nabla}_{\mathcal{D}_n} v_n\}_{n\in\mathbb{N}}$ and $\{\nabla_{\mathcal{D}_n} v_n\}_{n\in\mathbb{N}}$ weakly converges to $\nabla\bar{v}$ in $[L^2(\Omega)]^d$.*

We also have the following result, proven in Appendix B.1

Lemma 3 (Density of a space of test-functions) *Under the condition that there exists a finite partition of Ω into open connected disjoint polygonal subsets, $P_\Omega \stackrel{\text{def}}{=} \{\Omega_i\}_{i=1\ldots N_\Omega}$ such that (s.t.) $\Lambda_{|\Omega_i} \in [C^2(\overline{\Omega_i})]^{d\times d}$ for all $i = 1 \ldots N_\Omega$, let $\mathcal{Q}$ be the space of functions $\varphi : \overline{\Omega} \to \mathbb{R}$ s.t.*

(i) (φ is continuous and piecewise regular*)* $\varphi \in C_0(\overline{\Omega})$ *and, for all* $i = 1, \ldots, N_\Omega$, $\varphi \in C^2(\overline{\Omega_i})$,

*(ii) (*the tangential derivatives of φ are continuous through the interfaces of P_Ω*) for all $i, j = 1, \ldots, N_\Omega$, for all vector $\mathbf{t}$ parallel to $\partial\Omega_i \cap \partial\Omega_j$, $(\nabla\varphi)_{|\overline{\Omega_i}} \cdot \mathbf{t} = (\nabla\varphi)_{|\overline{\Omega_j}} \cdot \mathbf{t}$, where $(\nabla\varphi)_{|\overline{\Omega_i}}$ refers to the value of $\nabla\varphi$ on $\partial\Omega_i$ computed from the values on $\overline{\Omega_i}$,*

*(iii) (*the flux of $\nabla\varphi$ directed by $\Lambda\mathbf{n}$ is continuous through the interfaces of P_Ω*) for all $i, j = 1, \ldots, N_\Omega$ s.t. $\partial\Omega_i \cap \partial\Omega_j$ has dimension $d-1$, $(\Lambda\nabla\varphi)_{|\overline{\Omega_i}} \cdot \mathbf{n}_i + (\Lambda\nabla\varphi)_{|\overline{\Omega_j}} \cdot \mathbf{n}_j = 0$ on $\partial\Omega_i \cap \partial\Omega_j$, where $\mathbf{n}_i$ is the outer normal to Ω_i.*
Then, $\mathcal{Q}$ is dense in $H_0^1(\Omega)$.

The assumptions yielding convergence of the finite volume schemes are gathered in the following

Hypothesis 2 *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a family of discretizations matching Definition II.2.1. We suppose that*

(P1) $a_{\mathcal{D}_n}$ is uniformly coercive, i.e. , there is $0 < \gamma_1 < +\infty$ independent of n s.t.

$$\forall v \in \mathcal{H}_{\mathcal{D}_n}, \quad a_{\mathcal{D}_n}(v, v) \ge \gamma_1 \|v\|_{\mathcal{D}_n}^2;$$

(P2) $a_{\mathcal{D}_n}$ is consistent in the sense that for all $\varphi \in C_0^2(\Omega)$ (space of test-functions), $\lim_{n\to\infty} \epsilon_{\mathcal{D}_n}(\varphi) = 0$, where we denote by

$$\epsilon_{\mathcal{D}}(\varphi) = \sup_{v\in\mathcal{H}_{\mathcal{D}}, \|v\|_{\mathcal{D}}=1} \left| a_{\mathcal{D}}(P_{\mathcal{F}}\varphi, v) - \int_\Omega \Lambda(x)\nabla\varphi(x) \cdot \nabla_{\mathcal{D}} v(x)\, dx \right|. \qquad [\text{II.22}]$$

Proposition 1 *For $a_{\mathcal{D}_n}$ of type* [II.17]*, owing to* [II.6]*, Property (P2) holds for strongly consistent numerical fluxes, i.e. fluxes s.t., for all $\varphi \in C_0^2(\Omega)$, there exists $0 < C_2 < +\infty$ independent of n s.t. for all $K \in \mathcal{T}_n$,*

$$\begin{array}{l} \forall \sigma \in \mathcal{E}_K, \quad |F_{K,\sigma}(P_{\mathcal{F}_n}\varphi) - \mathrm{m}_\sigma < \Lambda \nabla \varphi >_K \cdot \mathbf{n}_{K,\sigma}| \ \leq C_2 \mathrm{m}_\sigma h_{\mathcal{D}_n} \\ \forall \sigma \in \mathcal{E}_{\mathrm{int},K}, |F_{K,\sigma}(P_{\mathcal{F}_n}\varphi)| \leq C_2 \mathrm{m}_\sigma h_{\mathcal{D}_n} \end{array} \qquad \text{[II.23]}$$

Proof. Let $\varphi \in C_0^2(\Omega)$ and $u \in \mathcal{H}_{\mathcal{D}_n}$ such that $\|u\|_{\mathcal{D}_n} = 1$,

$$\begin{aligned} a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi, u) - \int_\Omega \Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}_n} u(x)\ dx \\ = \sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_K}\left[F_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \frac{\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x}{\mathrm{m}_K}\cdot\mathbf{n}_{K,\sigma}\right](u_\sigma - u_K) \\ + \sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}} F_{K,\sigma}(P_{\mathcal{F}}\varphi)(u_\sigma - u_K) \stackrel{\mathrm{def}}{=} T_1. \end{aligned}$$

Using Cauchy-Schwarz inequality and thanks to IV.31, II.6, we obtain that,

$$\begin{aligned} |T_1| \ &\leq \ \left(\sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_K}\frac{d_{K,\sigma}}{\mathrm{m}_\sigma}\left|F_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma\frac{\int_K\Lambda(x)\nabla\varphi(x)\,\mathrm{d}x}{\mathrm{m}_K}\cdot\mathbf{n}_{K,\sigma}\right|^2 + \right. \\ &\qquad \left.\sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}}\frac{d_{K,\sigma}}{\mathrm{m}_\sigma}\left|F_{K,\sigma}(P_{\mathcal{F}}\varphi)\right|^2\right)^{\frac{1}{2}} \|u\|_{\mathcal{D}_n} \\ &\leq \ C_2(d\mathrm{m}(\Omega))^{\frac{1}{2}}\ h_{\mathcal{D}_n}\ \|u\|_{\mathcal{D}_n} = C_2(d\mathrm{m}(\Omega))^{\frac{1}{2}}\ h_{\mathcal{D}_n}. \end{aligned}$$

Then, we deduce that,

$$\begin{aligned} \epsilon_{\mathcal{D}_n}(\varphi) \ &= \ \sup_{v\in\mathcal{H}_{\mathcal{D}_n}, \|v\|_{\mathcal{D}_n}=1}\left|a_{\mathcal{D}}(P_{\mathcal{F}_n}\varphi, v) - \int_\Omega \Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}_n}v(x)\ dx\right| \\ &\leq \ C_2(d\mathrm{m}(\Omega))^{\frac{1}{2}}\ h_{\mathcal{D}_n}, \end{aligned}$$

and therefore $\epsilon_{\mathcal{D}_n}(\varphi) \to 0$ as $n \to \infty$, which concludes the proof.
□

Remark 1 *Under the condition that there exists a finite partition of Ω into open connected disjoint polygonal subsets, $P_\Omega \stackrel{\mathrm{def}}{=} \{\Omega_i\}_{i=1\ldots N_\Omega}$ such that (s.t.) $\Lambda_{|\Omega_i} \in [C^2(\overline{\Omega_i})]^{d\times d}$ for all $i = 1 \ldots N_\Omega$, then thanks to the lemma 3, Hypothesis (P2) can be relaxed by considering test functions φ in the space $\mathcal{Q}$ instead of $C_0^2(\Omega)$. Indeed, the proof of Theorem II.5.1 uses only the regularity for each $K \in \mathcal{T}$ and the density in $H_0^1(\Omega)$ of the test functions space.*

Proposition 2 (Asymptotic stability of the interpolator) *Under Hypothesis 2, for all* $\varphi \in C_0^2(\Omega)$,

$$\|P_{\mathcal{F}}\varphi\|_{\mathcal{D}} \leq \frac{1}{\gamma_1}\left(\epsilon_{\mathcal{D}}(\varphi) + \beta_0\sqrt{d}\|\nabla\varphi\|_{L^2(\Omega)^d}\right),$$

where $\epsilon_{\mathcal{D}}(\varphi)$ *is defined by (II.22).*

Proof. Owing to [II.19], for all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, the following integration by parts formula holds:

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma \frac{\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x}{\mathrm{m}_K}\cdot\mathbf{n}_{K,\sigma}(v_\sigma - v_K)\\
&\quad= \sum_{K\in\mathcal{T}_n}\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x\cdot\frac{1}{\mathrm{m}_K}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\mathbf{n}_{K,\sigma}(v_\sigma - v_K) \qquad \text{[II.24]}\\
&\quad= \sum_{K\in\mathcal{T}_n}\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x\cdot(\nabla_{\mathcal{D}}v)_K = \int_\Omega \Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}}v(x)\,\mathrm{d}x.
\end{aligned}$$

The above result together with (P1) and (II.20) yield

$$\begin{aligned}
\gamma_1\|P_{\mathcal{F}}\varphi\|_{\mathcal{D}}^2 &\leq a_{\mathcal{D}}(P_{\mathcal{F}}\varphi, P_{\mathcal{F}}\varphi)\\
&= \left(a_{\mathcal{D}}(P_{\mathcal{F}}\varphi, P_{\mathcal{F}}\varphi) - \int_\Omega \Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}}P_{\mathcal{F}}\varphi(x)\,\mathrm{d}x\right)\\
&\quad+ \int_\Omega \Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}}P_{\mathcal{F}}\varphi(x)\,\mathrm{d}x\\
&\leq \epsilon_{\mathcal{D}}(\varphi)\|P_{\mathcal{F}}\varphi\|_{\mathcal{D}} + \beta_0\|\nabla\varphi\|_{L^2(\Omega)^d}\|\nabla_{\mathcal{D}}P_{\mathcal{F}}\varphi\|_{[L^2(\Omega)]^d}\\
&\leq \left(\epsilon_{\mathcal{D}}(\varphi) + \beta_0\sqrt{d}\|\nabla\varphi\|_{L^2(\Omega)^d}\right)\|P_{\mathcal{F}}\varphi\|_{\mathcal{D}}.
\end{aligned}$$

□

Lemma 4 (Uniform *a priori* estimate) *Assume that Hypothesis 2 holds. Then, problem* [II.16] *is well-posed for each* $n \in \mathbb{N}$ *in the sense that there exists one and only one solution to problem* [II.16]*, and the solutions* $u_n \in \mathcal{H}_{\mathcal{D}_n,\mathcal{F}_n}$ *satisfy the following uniform* a priori *estimate:*

$$\|u_n\|_{\mathcal{D}_n} \leq \frac{C_1}{\gamma_1}\|f\|_{L^r(\Omega)}. \qquad \text{[II.25]}$$

Proof. The well-posedness follows from (P1), which guarantees the non singularity of the linear system corresponding to [II.16]. Using (P1), Hölder's inequality, Lemma 1 and II.11, it is inferred that (with $r' \overset{\text{def}}{=} \frac{r}{r-1}$)

$$\begin{aligned}
\gamma_1\|u_n\|_{\mathcal{D}_n}^2 \leq a_{\mathcal{D}_n}(u_n,u_n) &= \textstyle\int_\Omega f P_{\mathcal{T}_n}u\,\mathrm{d}x \leq \|f\|_{L^r(\Omega)}\|P_{\mathcal{T}_n}u_n\|_{L^{r'}(\Omega)}\\
&\leq C_1\|f\|_{L^r(\Omega)}\|u_n\|_{\mathcal{D}_n}.
\end{aligned}$$

□

Théorème II.5.1 (Convergence) *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a family of discretizations satisfying Hypothesis 2 and s.t. $h_{\mathcal{D}_n}\to 0$ as $n\to\infty$. Then, as $n\to\infty$, the sequence of discrete solutions of problem* [II.16]*, say $\{P_{\mathcal{T}_n}u_n\}_{n\in\mathbb{N}}$, converges to the solution $\overline{u}$ of* [II.1.1] *in $L^q(\Omega)$ for all $q\in[1,2d/(d-2))$.*

Proof. Owing to the *a priori* estimate [II.25] together with Lemma 2, there is $\widetilde{u}\in H_0^1(\Omega)$ s.t., up to a subsequence, (i) $\{P_{\mathcal{T}_n}u_n\}_{n\in\mathbb{N}}$ converges to $\widetilde{u}$ in $L^q(\Omega)$ for all $q\in[1,2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d>2$) and (ii) $\{\widetilde{\nabla}_{\mathcal{D}_n}u_n\}_{n\in\mathbb{N}}$ weakly converges to $\nabla\widetilde{u}$ in $[L^2(\Omega)]^d$. It only remains to prove that $\widetilde{u}=\bar{u}$. Let $\varphi\in C_0^2(\Omega)$. Owing to [II.21] together with (II.11), (P1),

$$\begin{aligned}\|\widetilde{\nabla}_{\mathcal{D}_n}(u_n-P_{\mathcal{F}_n}\varphi)\|^2_{[L^2(\Omega)]^d}\le d\|u_n-P_{\mathcal{F}_n}\varphi\|^2_{\mathcal{D}_n} &\le \tfrac{d}{\gamma_1}a_{\mathcal{D}_n}(u_n-P_{\mathcal{F}_n}\varphi,u_n-P_{\mathcal{F}_n}\varphi)\\ &= \tfrac{d}{\gamma_1}\left(T_1+T_2\right),\end{aligned} \qquad \text{[II.26]}$$

where $T_1\overset{\text{def}}{=}\int_\Omega f(x)P_{\mathcal{T}_n}(u_n-P_{\mathcal{F}_n}\varphi)(x)\,\mathrm{d}x$ and $T_2\overset{\text{def}}{=}a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi,P_{\mathcal{F}_n}\varphi-u_n)$. By the integrability assumption on f and the weak convergence of $\{P_{\mathcal{T}_n}u_n\}_{n\in\mathbb{N}}$ to $\widetilde{u}$ in $L^q(\Omega)$ for all $q<+\infty$ if $d=2$ and for $q=\frac{2d}{d-2}$ if $d>2$,

$$T_1\to\int_\Omega f(x)(\widetilde{u}-\varphi)(x)\,\mathrm{d}x \text{ as } n\to\infty. \qquad \text{[II.27]}$$

We get that,

$$\begin{aligned}a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi,u_n)&=\left(a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi,u_n)-\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}_n}u_n(x)\;dx\right)\\&\quad+\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla_{\mathcal{D}_n}u_n(x)\;dx\\&\overset{\text{def}}{=}T_{2,1}+T_{2,2}.\end{aligned}$$

Using Cauchy-Schwarz inequality as in the proof of Proposition 2, we have $|T_{2,1}|\le\epsilon_{\mathcal{D}_n}(\varphi)\|u_n\|_{\mathcal{D}_n}$. Thanks to Lemma 4, $\|u_n\|_{\mathcal{D}_n,\mathcal{F}_n}$ is bounded uniformly with respect to n. Thus, by property (P2), $T_{2,1}\to 0$ as $n\to\infty$. Also, using the weak convergence of $\{\nabla_{\mathcal{D}_n}u_n\}_{n\in\mathbb{N}}$, we conclude that $T_{2,2}\to\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla\widetilde{u}(x)\,\mathrm{d}x$ as $n\to\infty$. By Proposition 2, $\|P_{\mathcal{F}_n}\varphi\|_{\mathcal{D}_n}$ is uniformly bounded with respect to n; since $P_{\mathcal{F}_n}\varphi$ obviously converges to φ, it is then easy, using Lemma 2, to see that $\nabla_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi)$ weakly converges to $\nabla\varphi$. Proceeding in a similar way as for $a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi,u_n)$, we can thus prove that $a_{\mathcal{D}_n}(P_{\mathcal{F}_n}\varphi,P_{\mathcal{F}_n}\varphi)\to\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla\varphi(x)\,\mathrm{d}x$ as $n\to\infty$. Therefore,

$$T_2\to\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla(\varphi-\widetilde{u})(x)\,\mathrm{d}x \text{ as } n\to\infty. \qquad \text{[II.28]}$$

Plugging [II.27] and [II.28] into the right hand side of [II.26] and using the weak convergence of $\widetilde{\nabla}_{\mathcal{D}_n}(u_n-P_{\mathcal{F}_n}\varphi)$, we conclude that, for all $\varphi\in C_0^2(\Omega)$,

$$\|\nabla(\widetilde{u}-\varphi)\|^2_{[L^2(\Omega)]^d}\le\frac{d}{\gamma_1}\left(\int_\Omega f(x)(\widetilde{u}-\varphi)(x)\,\mathrm{d}x+\int_\Omega\Lambda(x)\nabla\varphi(x)\cdot\nabla(\varphi-\widetilde{u})(x)\,\mathrm{d}x\right).$$

We can apply this inequality to a sequence $\{\varphi_m\}_{m\in\mathbb{N}} \in C_0^2(\Omega)$ which tends to $\overline{u}$ in $H_0^1(\Omega)$ and let $m \to \infty$; since $\overline{u}$ solves problem [II.1.1], we obtain

$$\begin{aligned}\|\nabla(\widetilde{u}-\overline{u})\|^2_{[L^2(\Omega)]^d} &\leq \frac{d}{\gamma_1}\left[\int_\Omega f(x)(\widetilde{u}(x)-\overline{u}(x))\,\mathrm{d}x - \int_\Omega \Lambda(x)\nabla\overline{u}(x)\cdot\nabla(\widetilde{u}-\overline{u})(x)\,\mathrm{d}x\right]\\ &= 0,\end{aligned}$$

i.e. , $\widetilde{u} = \overline{u}$. Due to the uniqueness of the solution of [II.1.1], we deduce that the entire sequence $\{u_n\}_{n\in\mathbb{N}}$ converges to $\overline{u}$ in $L^q(\Omega)$ for all $q \in [1, 2d/(d-2))$ (and weakly in $L^{2d/(d-2)}(\Omega)$ if $d > 2$). Observe that the order in which the limits for $n \to \infty$ and $m \to \infty$ are taken cannot be exchanged, the sequence $\{\|P_{\mathcal{F}_m}\varphi\|_{\mathcal{D}_n}\}_{m\in\mathbb{N}}$ being possibly unbounded. This concludes the proof.
□

In the chapters which follow, we will prove only the coercivity and the consistency of the studied scheme and the convergence will derive, thanks to Theorem II.5.1.
In Theorem II.5.2, we give an error estimate, this estimate is obtained thanks to the following Lemma.

Lemma 5 *Let us assume in addition that Ω is convex, $\Lambda = \lambda\ I_d$ such that $\lambda \in W^{1,\infty}(\Omega)$ and there exists $\overline{\lambda}_0 \geq 0$ such that for all $x \in \Omega$, $\|\nabla\lambda\|_{L^\infty(\Omega)} \leq \overline{\lambda}_0$, then there exists a unique solution $v \in H^2(\Omega)$ of the system of equations (II.29),*

$$\begin{cases} -\mathrm{div}(\Lambda\nabla v) = f & \text{in } \Omega,\\ v = 0 & \text{on } \partial\Omega,\end{cases} \qquad \text{[II.29]}$$

moreover there exists a real $C > 0$ depending only on d, Ω, α_0 and $\overline{\lambda}_0$ such that,

$$\|v\|_{H^2(\Omega)} \leq C\|f\|_{L^2(\Omega)}.$$

Proof. It is well known that there exists a unique solution $v \in H_0^1(\Omega)$ of the system of equations (II.29) and there exists a real $C > 0$ depending only on d, Ω and α_0 such that,

$$\|v\|_{H^1(\Omega)} \leq C\|f\|_{L^2(\Omega)}, \qquad \text{[II.30]}$$

moreover $\mathrm{div}(\lambda\nabla v) \in L^2(\Omega)$ with,

$$\|\mathrm{div}(\lambda\nabla v)\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}. \qquad \text{[II.31]}$$

Thanks to (II.30) and (II.31), we deduce that $\mathrm{div}(\lambda\nabla v) - \nabla v\cdot\nabla\lambda \in L^2(\Omega)$, with,

$$\|\mathrm{div}(\lambda\nabla v) - \nabla v\cdot\nabla\lambda\|_{L^2(\Omega)} \leq (1 + C\ \overline{\lambda}_0)\|f\|_{L^2(\Omega)}$$

this implies that $\lambda\Delta v \in L^2(\Omega)$, with,

$$\|\lambda\Delta v\|_{L^2(\Omega)} \leq (1 + C\ \overline{\lambda}_0)\|f\|_{L^2(\Omega)}. \qquad \text{[II.32]}$$

Then, we get that,

$$\|\Delta v\|_{L^2(\Omega)} \leq \frac{(1+C\ \overline{\lambda}_0)}{\alpha_0}\|f\|_{L^2(\Omega)}. \tag{II.33}$$

Thanks to (II.33) and the H^2-regularity of solution of the Homogeneous Dirichlet Problem (see [49], [48]) in convex domain, we obtain that $v \in H^2(\Omega)$ and there exists $C_1 > 0$ depending only on d, Ω, α_0 and $\overline{\lambda}_0$ such that,

$$\|v\|_{H^2(\Omega)} \leq C_1\|f\|_{L^2(\Omega)},$$

which concludes the proof.
□

Théorème II.5.2 *[error estimate] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition (II.2.1), we assume that $\mathcal{F} = \mathcal{E}$, for all $K \in \mathcal{T}$, $\mathcal{E}_{\mathrm{int},K} = \emptyset$, K is star-shaped with respect to all the points in a ball of radius $\alpha\mathrm{diam}(K)$ and $\Lambda = \lambda I_d$ such that $\lambda \in W^{1,\infty}(\Omega)$ and there exists $\overline{\lambda}_0 \geq 0$ such that for all $x \in \Omega$, $\|\nabla\lambda\|_{L^\infty(\Omega)} \leq \overline{\lambda}_0$. Let $\bar{u} \in H_0^1(\Omega)$ be the solution of (II.1) and $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$ be the solution of (II.16)-(II.17), then, there exists a real $C_0 > 0$ depending only on d, α, ϱ_3, β_0, $\overline{\lambda}_0$ and α_0 such that,*

$$\|\nabla_{\mathcal{D}}u - \nabla\bar{u}\|_{L^2(\Omega)^d} \leq \frac{d\sqrt{d}}{\gamma_1}\left(E_{\mathcal{D}}(\bar{u}) + C_0\|f\|_{L^2(\Omega)}h_{\mathcal{D}}\right), \tag{II.34}$$

where for all $v \in H_0^1(\Omega)$,

$$\begin{aligned} E_{\mathcal{D}}(v) &= \left(\sum_{K\in\mathcal{T}_n}\sum_{\sigma\in\mathcal{E}_K}\frac{d_{K,\sigma}}{\mathrm{m}_\sigma}\left|F_{K,\sigma}(v_{\mathcal{D}}) - \mathrm{m}_\sigma < \Lambda\nabla v > \cdot\mathbf{n}_{K,\sigma}\right|^2\right)^{\frac{1}{2}} \\ v_{\mathcal{D}} &= \{< v >_K, K \in \mathcal{T}, < v >_\sigma, \sigma \in \mathcal{E}\} \in \mathcal{H}_{\mathcal{D}}, \end{aligned} \tag{II.35}$$

moreover if there exists a real $C > 0$ such that for all $\varphi \in C_c^\infty(\Omega)$,

$$\forall\sigma \in \mathcal{E}_K,\ K \in \mathcal{T}, \quad \left|F_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma < \Lambda\nabla\varphi >_K \cdot\mathbf{n}_{K,\sigma}\right| \leq C\|\varphi\|_{H^2(\Omega)}\mathrm{m}_\sigma h_{\mathcal{D}}, \tag{II.36}$$

then, there exists a real $C_3 > 0$ depending only on d, α, ϱ_3, β_0, $\overline{\lambda}_0$ and α_0 such that

$$\|\nabla_{\mathcal{D}}u - \nabla\bar{u}\|_{L^2(\Omega)^d} \leq C_3\|f\|_{L^2(\Omega)}\ h_{\mathcal{D}}. \tag{II.37}$$

Proof. Thanks to Lemma 5, $\bar{u} \in H^2(\Omega)$ and there exists a real $C_1 > 0$ depending only on d, Ω, α_0 and $\overline{\lambda}_0$ such that,

$$\|\bar{u}\|_{H^2(\Omega)} \leq C_1\|f\|_{L^2(\Omega)}. \tag{II.38}$$

Let $\bar{u}_{\mathcal{D}} = \{< \bar{u} >_K, K \in \mathcal{T}, < \bar{u} >_\sigma, \sigma \in \mathcal{E}\} \in \mathcal{H}_{\mathcal{D}}$.

$$\begin{aligned} \|\nabla_{\mathcal{D}} u - \nabla \bar{u}\|_{L^2(\Omega)^d} &\leq \|\nabla_{\mathcal{D}}(u - \bar{u}_{\mathcal{D}})\|_{[L^2(\Omega)]^d} + \|\nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}} - \nabla \bar{u}\|_{[L^2(\Omega)]^d} \\ &\leq \sqrt{d}\|u - \bar{u}_{\mathcal{D}}\|_{\mathcal{D}} + \|\nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}} - \nabla \bar{u}\|_{[L^2(\Omega)]^d}. \end{aligned} \quad [\text{II.39}]$$

Thanks to the coercivity property, we get that,

$$\begin{aligned} \|u - \bar{u}_{\mathcal{D}}\|_{\mathcal{D}}^2 &\leq \tfrac{d}{\gamma_1} a_{\mathcal{D}}(u - \bar{u}_{\mathcal{D}}, u - \bar{u}_{\mathcal{D}}) \\ &= \tfrac{d}{\gamma_1} (T_1 + T_2), \end{aligned} \quad [\text{II.40}]$$

where $T_1 \overset{\text{def}}{=} \int_\Omega f(x) P_{\mathcal{T}}(u - \bar{u})(x)\, \mathrm{d}x$ and $T_2 \overset{\text{def}}{=} a_{\mathcal{D}}(\bar{u}_{\mathcal{D}}, \bar{u}_{\mathcal{D}} - u)$.
Let us introduce $v_{\mathcal{D}} = \bar{u}_{\mathcal{D}} - u \in \mathcal{H}_{\mathcal{D}}$. We decompose T_2 into two terms T_{21} and T_{22}, such that,

$$\begin{aligned} T_{21} &= a_{\mathcal{D}}(\bar{u}_{\mathcal{D}}, v_{\mathcal{D}}) - \sum_{K \in \mathcal{T}} (v_{\mathcal{D}})_K \sum_{\sigma \in \mathcal{E}_K} \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\ \mathrm{d}x \\ T_{22} &= \sum_{K \in \mathcal{T}} (v_{\mathcal{D}})_K \sum_{\sigma \in \mathcal{E}_K} \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\ \mathrm{d}x \end{aligned} \quad [\text{II.41}]$$

Since $\mathrm{div}(\lambda \nabla \bar{u}) \in L^2(\Omega)$, then the flux continuity holds, this implies that for all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{int}}$, $\mathcal{T}_\sigma = \{K, L\}$,

$$\int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\ \mathrm{d}x + \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/L} \cdot \mathbf{n}_{L,\sigma}\ \mathrm{d}x = 0$$

Then, we insert $(v_{\mathcal{D}})_\sigma$ and we can re-write T_{21} as follows,

$$\begin{aligned} T_{21} &= a_{\mathcal{D}}(\bar{u}_{\mathcal{D}}, v_{\mathcal{D}}) - \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} ((v_{\mathcal{D}})_K - (v_{\mathcal{D}})_\sigma) \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\ \mathrm{d}x \\ &= \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} \left(F_{K,\sigma}(\bar{u}_{\mathcal{D}}) - \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\, \mathrm{d}x \right) ((v_{\mathcal{D}})_K - (v_{\mathcal{D}})_\sigma) \\ &= T_{211} + T_{212}, \end{aligned}$$

where $T_{211} \overset{\text{def}}{=} \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} (F_{K,\sigma}(\bar{u}_{\mathcal{D}}) - \mathrm{m}_\sigma < \lambda(x) \nabla \bar{u}(x) >_K \cdot \mathbf{n}_{K,\sigma}) ((v_{\mathcal{D}})_K - (v_{\mathcal{D}})_\sigma)$ and
$T_{212} \overset{\text{def}}{=} \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} \left(\mathrm{m}_\sigma < \lambda(x) \nabla \bar{u}(x) >_K - \int_\sigma (\lambda(x) \nabla \bar{u}(x))_{/K}\, \mathrm{d}x \right) ((v_{\mathcal{D}})_K - (v_{\mathcal{D}})_\sigma) \cdot \mathbf{n}_{K,\sigma}$.
Using the same arguments as in the Proposition 1, we get

$$|T_{211}| \leq E_{\mathcal{D}}(\bar{u})\ \|v_{\mathcal{D}}\|_{\mathcal{D}}. \quad [\text{II.42}]$$

Using Cauchy-Schwarz inequality, we deduce,

$$T_{212} \leq \left(\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma d_{K,\sigma} \left| \frac{1}{\mathrm{m}_K} \int_K \lambda(x)\nabla\bar{u}(x)\,\mathrm{d}x - \frac{1}{\mathrm{m}_\sigma} \int_\sigma \lambda(x)\nabla\bar{u}(x)\,\mathrm{d}x \right|^2 \right)^{\frac{1}{2}} \|v_\mathcal{D}\|_\mathcal{D} \quad \text{[II.43]}$$

and thanks to the Lemma 6.6 in [35], for all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, we deduce that there exists $C > 0$ depending only on d and α such that,

$$|< \lambda\nabla\bar{u} >_K - < \lambda\nabla\bar{u} >_\sigma|^2 \leq \frac{C\mathrm{diam}(K)}{\mathrm{m}_\sigma} \int_K \sum_{j=1}^d \left| \nabla \left(\lambda(x) \frac{\partial\bar{u}(x)}{\partial x_j} \right) \right|^2 \mathrm{d}x. \quad \text{[II.44]}$$

For a.e $x \in \Omega$, we get,

$$\begin{aligned}
\sum_{j=1}^d \left| \nabla \left(\lambda(x) \frac{\partial\bar{u}(x)}{\partial x_j} \right) \right|^2 &= \sum_{j=1}^d \left| \nabla\lambda(x) \frac{\partial\bar{u}(x)}{\partial x_j} + \lambda(x)\nabla \frac{\partial\bar{u}(x)}{\partial x_j} \right|^2 \\
&\leq 2|\nabla\lambda(x)|^2 \sum_{j=1}^d \left| \frac{\partial\bar{u}(x)}{\partial x_j} \right|^2 + 2|\lambda(x)|^2 \sum_{j=1}^d \left| \nabla \frac{\partial\bar{u}(x)}{\partial x_j} \right|^2 \\
&\leq 2\overline{\lambda}_0^2 |\nabla\bar{u}(x)|^2 + 2\beta_0^2 |\nabla^2\bar{u}(x)|^2 \\
&\leq 2\max(\overline{\lambda}_0^2, \beta_0^2)(|\nabla\bar{u}(x)|^2 + |\nabla^2\bar{u}(x)|^2),
\end{aligned}$$

then, from (II.44), we deduce that,

$$|< \lambda\nabla\bar{u} >_K - < \lambda\nabla\bar{u} >_\sigma|^2 \leq 2C\max(\overline{\lambda}_0^2, \beta_0^2) \frac{\mathrm{diam}(K)}{\mathrm{m}_\sigma} \|\bar{u}\|^2_{H^2(K)}. \quad \text{[II.45]}$$

Using (II.43) and (II.45), we obtain,

$$T_{212} \leq \sqrt{2C\varrho_3}\max(\overline{\lambda}_0, \beta_0) h_\mathcal{D} \|\bar{u}\|_{H^2(\Omega)} \|v_\mathcal{D}\|_\mathcal{D} \quad \text{[II.46]}$$

Therefore, thanks to (II.42) and (II.46) , we deduce,

$$T_{21} \leq (E_\mathcal{D}(\bar{u}) + \sqrt{2C\varrho_3}\max(\overline{\lambda}_0, \beta_0)\|\bar{u}\|_{H^2(\Omega)}\ h_\mathcal{D}\ )\|v_\mathcal{D}\|_\mathcal{D}. \quad \text{[II.47]}$$

Since $\bar{u}$ is solution to (II.1), for all $K \in \mathcal{T}$, we get that,

$$\begin{aligned}
\sum_{\sigma\in\mathcal{E}_K} \int_\sigma (\lambda(x)\nabla\bar{u}(x))_{/K} \cdot \mathbf{n}_{K,\sigma}\ \mathrm{d}x &= \int_K \mathrm{div}(\lambda(x)\nabla\bar{u}(x))\,\mathrm{d}x \\
&= \int_K f(x)\,\mathrm{d}x.
\end{aligned}$$

then, we can re-write T_{22} as follows,

$$T_{22} = \int_\Omega f(x) P_\mathcal{T} v_\mathcal{D} \quad \text{[II.48]}$$

Gathering the results (II.40), (II.47) and (II.48), we obtain that,

$$\|\bar{u}_{\mathcal{D}} - u\|_{\mathcal{D}} \leq \frac{d}{\gamma_1}(E_{\mathcal{D}}(\bar{u}) + \sqrt{2C\varrho_3}\max(\overline{\lambda}_0, \beta_0)\|\bar{u}\|_{H^2(\Omega)}\ h_{\mathcal{D}}). \qquad \text{[II.49]}$$

It remains to estimate the term $\|\nabla_{\mathcal{D}}\bar{u}_{\mathcal{D}} - \nabla\bar{u}\|_{[L^2(\Omega)]^d}$, we get that,

$$\begin{aligned}
\|\nabla_{\mathcal{D}}\bar{u}_{\mathcal{D}} - \nabla\bar{u}\|^2_{[L^2(\Omega)]^d} &\leq 2\sum_{K\in\mathcal{T}} \mathrm{m}_K|(\nabla_{\mathcal{D}}\bar{u}_{\mathcal{D}})_K - < \nabla\bar{u} >_K|^2 \\
&\quad + 2\sum_{K\in\mathcal{T}}\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(x)|^2 dx \\
&= 2\sum_{K\in\mathcal{T}} \mathrm{m}_K \left| \frac{1}{\mathrm{m}_K}\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma(< \bar{u} >_\sigma - < \bar{u} >_K)\mathbf{n}_{K,\sigma} - < \nabla\bar{u} >_K \right|^2 \\
&\quad + 2\sum_{K\in\mathcal{T}}\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(x)|^2 dx \\
&= 2\sum_{K\in\mathcal{T}} \mathrm{m}_K \left| \frac{1}{\mathrm{m}_K}\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma < \bar{u} >_\sigma \mathbf{n}_{K,\sigma} - < \nabla\bar{u} >_K \right|^2 \\
&\quad + 2\sum_{K\in\mathcal{T}}\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(x)|^2 dx
\end{aligned}$$

since for all $K \in \mathcal{T}$, $\displaystyle\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma\mathbf{n}_{K,\sigma} = 0$, we get also that $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma < \bar{u} >_\sigma \mathbf{n}_{K,\sigma} = \sum_{\sigma\in\mathcal{E}_K}\int_\sigma \bar{u}(x)\mathbf{n}_{K,\sigma}\,\mathrm{d}\gamma(x)$ and thanks to Green's formula, we get

$$\sum_{\sigma\in\mathcal{E}_K}\int_\sigma \bar{u}(x)\mathbf{n}_{K,\sigma}d\gamma(x) = \int_K \nabla\bar{u}(x)\,\mathrm{d}\gamma(x),$$

therefore we deduce that

$$\|\nabla_{\mathcal{D}}\bar{u}_{\mathcal{D}} - \nabla\bar{u}\|^2_{[L^2(\Omega)]^d} \leq 2\sum_{K\in\mathcal{T}}\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(x)|^2 dx.$$

Since for any $K \in \mathcal{T}$, K is not convex, we can not conclude directly that there exists a real $Q > 0$ such that,

$$\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(x)|^2 dx \leq Q\ \mathrm{diam}(K)^2\|\bar{u}\|^2_{H^2(\overline{K})}. \qquad \text{[II.50]}$$

Then, let $y_K \in K$ a point of the ball for which K is star-shaped with respect to all the points of this ball, using twice the Taylor formula with the integral form of the remainder term, we deduce that there exists constants $Q_1 > 0$, $Q_2 > 0$ such that,

$$\int_K | < \nabla\bar{u} >_K - \nabla\bar{u}(y_K)|^2 dx \leq Q_1\ \mathrm{diam}(K)^2\|\bar{u}\|^2_{H^2(\overline{K})}, \qquad \text{[II.51]}$$

$$\int_K |\nabla \bar{u}(y_K) - \nabla \bar{u}(x)|^2 dx \leq Q_2 \ \mathrm{diam}(K)^2 \|\bar{u}\|^2_{H^2(\overline{K})}. \qquad \text{[II.52]}$$

Therefore, there exists a constant $Q_3 > 0$ such that,

$$\|\nabla_{\mathcal{D}} \bar{u}_{\mathcal{D}} - \nabla \bar{u}\|^2_{[L^2(\Omega)]^d} \leq Q_3 h_{\mathcal{D}} \|\bar{u}\|_{H^2(\Omega)}. \qquad \text{[II.53]}$$

Using (II.49) and (II.53), from (II.39), we deduce,

$$\|\nabla_{\mathcal{D}} u - \nabla \bar{u}\|_{L^2(\Omega)^d} \leq \frac{d\sqrt{d}}{\gamma_1}(E_{\mathcal{D}}(\bar{u}) + (\sqrt{2C\varrho_3}\max(\overline{\lambda}_0, \beta_0) + Q_3)\|\bar{u}\|_{H^2(\Omega)}\ h_{\mathcal{D}}). \qquad \text{[II.54]}$$

Thanks to (II.38) and (II.54), we infer that there exists a real $C_0 > 0$ depending only on d, α, ϱ_3, β_0, $\overline{\lambda}_0$ and α_0 such that,

$$\|\nabla_{\mathcal{D}} u - \nabla \bar{u}\|_{L^2(\Omega)^d} \leq \frac{d\sqrt{d}}{\gamma_1}(E_{\mathcal{D}}(\bar{u}) + C_0\|f\|_{L^2(\Omega)}\ h_{\mathcal{D}}), \qquad \text{[II.55]}$$

which concludes the first part of the proof. Since C_c^∞ is dense in $H_0^2(\Omega)$, there exists a sequence of functions of C_c^∞, $\{\varphi_p\}_{p\geq 0}$ such that

$$\|\varphi_p - \bar{u}\|_{H^2(\Omega)} \to 0 \text{ as } p \to \infty. \qquad \text{[II.56]}$$

Thanks to Cauchy-Schwarz inequality, for all $p \geq 0$, we can deduce that,

$$E_{\mathcal{D}}(\bar{u}) \leq E_{\mathcal{D}}(\bar{u} - \varphi_p) + E_{\mathcal{D}}(\varphi_p) \qquad \text{[II.57]}$$

Thanks to (II.36) and using the Proposition 1 with $C_2 = C\|\varphi_p\|_{H^2(\Omega)}$, we obtain that,

$$\begin{aligned} E_{\mathcal{D}}(\varphi_p) &\leq C\|\varphi_p\|_{H^2(\Omega)}(d\mathrm{m}(\Omega))^{\frac{1}{2}} h_{\mathcal{D}} \\ &\leq C(\|\bar{u}\|_{H^2(\Omega)} + \|\varphi_p - \bar{u}\|_{H^2(\Omega)})(d\mathrm{m}(\Omega))^{\frac{1}{2}} h_{\mathcal{D}} \end{aligned} \qquad \text{[II.58]}$$

and thanks to (II.38), we get,

$$E_{\mathcal{D}}(\varphi_p) \leq C(C_1\|f\|_{L^2(\Omega)} + \|\varphi_p - \bar{u}\|_{H^2(\Omega)})(d\mathrm{m}(\Omega))^{\frac{1}{2}} h_{\mathcal{D}} \qquad \text{[II.59]}$$

From (II.55), using (II.57) and (II.59), we deduce that there exists a real $C_3 > 0$ depending only on d, α, ϱ_3, β_0, $\overline{\lambda}_0$ and α_0 such that,

$$\|\nabla_{\mathcal{D}} u - \nabla \bar{u}\|_{L^2(\Omega)^d} \leq C_3(\|f\|_{L^2(\Omega)} + E_{\mathcal{D}}(\bar{u} - \varphi_p) + \|\varphi_p - \bar{u}\|_{H^2(\Omega)})\ h_{\mathcal{D}}. \qquad \text{[II.60]}$$

Owing to (II.56), we infer that $E_{\mathcal{D}}(\bar{u} - \varphi_p) \to 0$ as $p \to \infty$, therefore, taking the limit in (II.60) as $p \to \infty$, we obtain,

$$\|\nabla_{\mathcal{D}} u - \nabla \bar{u}\|_{L^2(\Omega)^d} \leq C_3\|f\|_{L^2(\Omega)}\ h_{\mathcal{D}}, \qquad \text{[II.61]}$$

which completes the proof.
□

Chapitre III

Les schémas volumes finis non symétriques

Ce chapitre réunit les résultats présentés dans trois articles, l'un publié [15] et les deux autres soumis ([10], [12, §3.2]). Il fournit d'abord une preuve originale de la convergence des schémas multipoints présentés dans [1], [2], [3] appelés (MPFA O) et donne une généralisation de ces schémas pour des maillages quelconques. Cette preuve repose sur les méthodes d'analyse communes à ce mémoire. Il présente ensuite un schéma original dont l'intérêt est de présenter un stencil compact au sens de l'introduction et de satisfaire des propriétés de stabilité dans plusieurs situations dans lesquelles le schéma MPFA O semble diverger. Ce schéma généralise le schéma MPFA L introduit dans [6] et donne un cadre mathématique d'étude de la convergence. Quant au troisième schéma que nous avons introduit, il a l'avantage d'être robuste vis à vis de l'hétérogénéité et l'anisotropie des tenseurs de diffusion, il a été conçu afin de garder un stencil limité par volume de contrôle, par exemple sur une grille cartésienne en 2D, le stencil est de 13 et en 3D, il est de 19.

III.1 MPFA O scheme

The usual MultiPoint Flux Approximation (MPFA) O method is a cell centered finite volume discretization of such second order elliptic equations described for example in [1] and [60]. It is a widely used scheme in the oil industry for the discretization of diffusion fluxes in multiphase Darcy porous media flow models (see for example [30], [53], and [58]).

Recent papers have studied the convergence of the MPFA O scheme but there is yet no convergence result on general polygonal and polyhedral meshes, and none taking into account discontinuous diffusion coefficients which are essential in oil industry applications. In [61], [5], [54], the convergence of the scheme is obtained on

quadrilateral meshes. The proofs are based on equivalences of the MPFA O scheme to mixed finite element methods using specific quadrature rules. The convergence of the scheme is obtained provided that a square d-dimensional matrix defined locally for each cell and each vertex of the cell, depending both on the distortion of cell and on the cell diffusion tensor, is uniformly positive definite. This analysis confirms the numerical experiments showing that the coercivity and convergence of the scheme is lost in the cases of strong distortion of the mesh and/or anisotropy of the diffusion tensor.

In [59] a mimetic finite difference scheme is introduced which is equivalent to the MPFA O scheme for simplicial and parallepipedic cells and a proper choice of the continuity points. This scheme has also been independently introduced in [28] in two dimensions. The numerical analysis in [59] provides a convergence result for such meshes with usual shape regularity assumptions and for smooth diffusion coefficients. In such specific cases, the MPFA O scheme is known to be symmetric and coercive whatever the diffusion tensor and the distortion of the mesh.

In this section, a discrete variational formulation is introduced using the framework described in [43], [46]. It involves the definition of two piecewise constant gradients and stability terms using residuals of the second gradient. The first gradient has a weak convergence property and is fixed in the construction. The second one is assumed to be consistent in the sense that it is exact on linear functions. For usual meshes such that each vertex of any cell K is shared by exactly d faces of the cell K, the stability terms are vanishing and our discrete variational formulation will be shown to be equivalent to the usual MPFA O scheme. It will in addition provide a generalization of the O scheme on more general polyhedral cells.

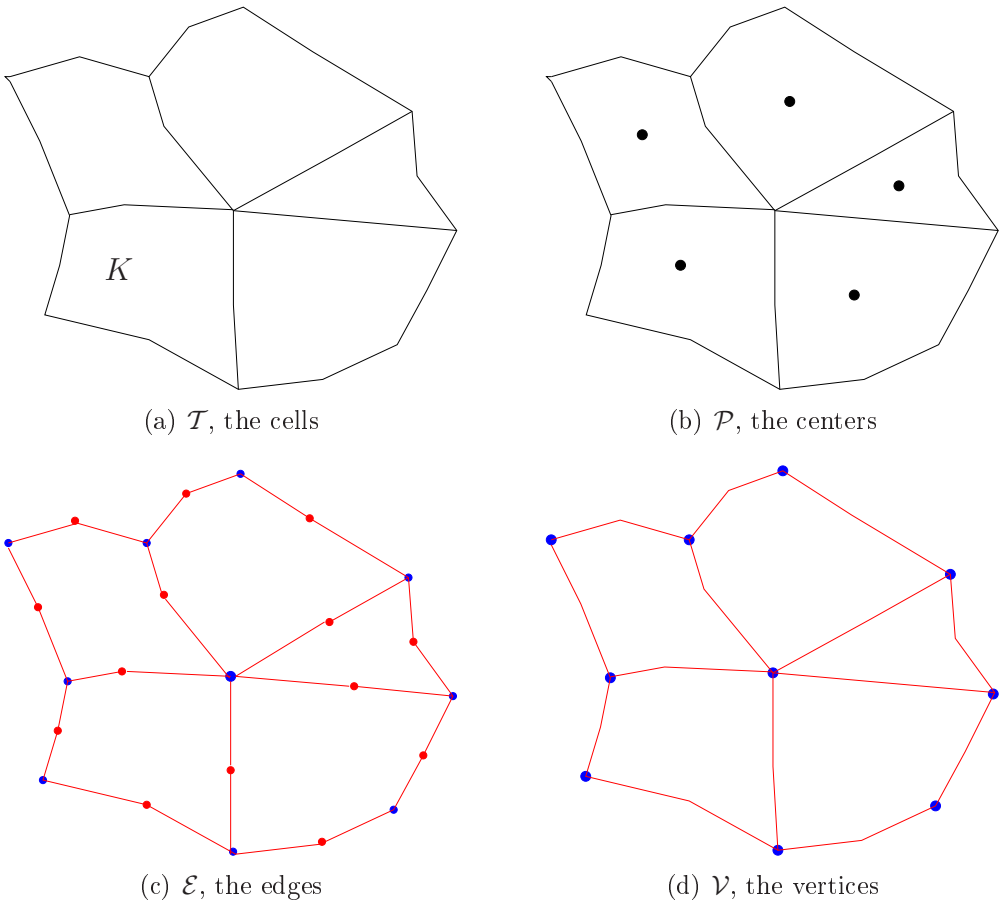
A sufficient local condition for the coercivity of the scheme is derived which will yield existence, and uniqueness of the solution. Under this coercivity condition, and a uniform stability assumption for the consistent gradient, the convergence of the scheme including the case of L^∞ diffusion coefficients can be proved.
In this section, we use the notations of chapter I. Hence we consider $\mathcal{D}$ an admissible discretization in the sense of Definition (II.2.1).
Notations: In the following, for any vectors $x, y \in \mathbb{R}^d$, we will denote by $x \cdot y$ their dot product $\sum_{i=1}^d x_i\ y_i$, and by $|x|$ the norm $\sqrt{x \cdot x}$. The notations $\lambda_{\max}(M)$ and $\lambda_{\min}(M)$ will stand for the maximum and minimum eigenvalues of any given square symmetric matrix M. For any matrix A, we denote by $|A|$ its norm defined by $\sup_{x \in \mathbb{R}^d} \frac{|Ax|}{|x|} = \sqrt{\lambda_{\max}(A^t A)}$.

III.1.1 The Finite Volume Scheme

To define the finite volume scheme MPFA O,

Hypothesis 3 *we further assume that (see figures III.1),*

- $\cup_{s\in\mathcal{V}}\mathcal{E}_s = \cup_{K\in\mathcal{T}}\mathcal{E}_K$,
- *for all* $K \in \mathcal{T}$, $s \in \mathcal{V}_K$, *the family of vectors* $(x_\sigma - x_K)_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}$ *spans* $\mathbb{R}^d$ *(see figure III.2),*
- *for all* $(s, s') \in \mathcal{V}\times\mathcal{V}$, $\mathcal{E}_s \cap \mathcal{E}_{s'} = \emptyset$, *which means that there exists no edge sharing at least two vertices of* $\mathcal{V}$ *(see figure III.3).*
- $\overline{\mathcal{F}} = \emptyset$, *which means that we use all the additional edge unknowns.*
- *for all* $K \in \mathcal{T}$, $\mathcal{E}_{\text{int},K} = \emptyset$, *which means that we have no "edge inside the cells".*
- *for all* $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, $d_{K,\sigma}$ *is the Euclidean distance from* x_K *to* σ.



(a) $\mathcal{T}$, the cells (b) $\mathcal{P}$, the centers

(c) $\mathcal{E}$, the edges (d) $\mathcal{V}$, the vertices

Figure III.1: Discretization of the domain Ω.

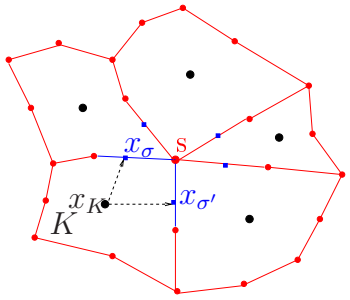


Figure III.2: $(x_\sigma - x_K)_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ spans $\mathbb{R}^2$.

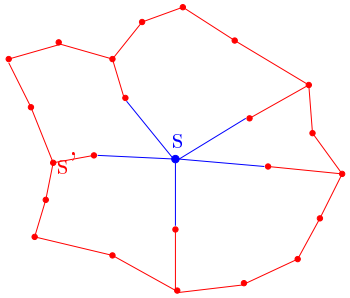


Figure III.3: $\mathcal{E}_s \cap \mathcal{E}_{s'} = \emptyset$.

Parameters of the MPFA O finite volume scheme: The volume of each cell $K \in \mathcal{T}$ is distributed to the vertices of the cell according to the sub-volumes m_K^s, $s \in \mathcal{V}_K$ defined by

$$\mathrm{m}_K^s = \frac{1}{d} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \mathrm{m}_\sigma d_{K,\sigma}, \qquad \text{[III.1]}$$

and which satisfy $\mathrm{m}_K = \sum_{s \in \mathcal{V}_K} \mathrm{m}_K^s$ for all $K \in \mathcal{T}$.

On each continuity point x_σ, the intermediate unknown u_σ is defined which will be used together with the cell unknowns u_K, $K \in \mathcal{T}$ for the construction of the finite volume scheme.

The usual MPFA O scheme introduced in [1],is a cell centered finite volume scheme with main degrees of freedom the cell unknowns u_K on each cell K of the mesh $\mathcal{T}$. The construction of the scheme uses additional degrees of freedom u_σ for each face σ. These unknowns will be locally eliminated as linear combinations of the neighbouring cell unknowns using the flux continuity equations.
The usual MPFA O scheme is built under the assumption that each vertex $s \in \mathcal{V}$ of any cell K is shared by exactly d faces of the cell K, this means $\mathrm{card}(\mathcal{E}_K \cap \mathcal{E}_s) = d$. More precisely, the construction of the usual MPFA O scheme is obtained as follows,

- For each vertex of the mesh $s \in \mathcal{V}$, for each control volume sharing this vertex,

$K \in \mathcal{T}_s$, we introduce the polytope K_s (polygon in 2D, polyhedron in 3D) built from s, the center of K, x_K and the centers of the edges of K sharing s, $(x_\sigma)_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ (see figure III.4). These polytopes are known as the subcells of the control volume (cell).

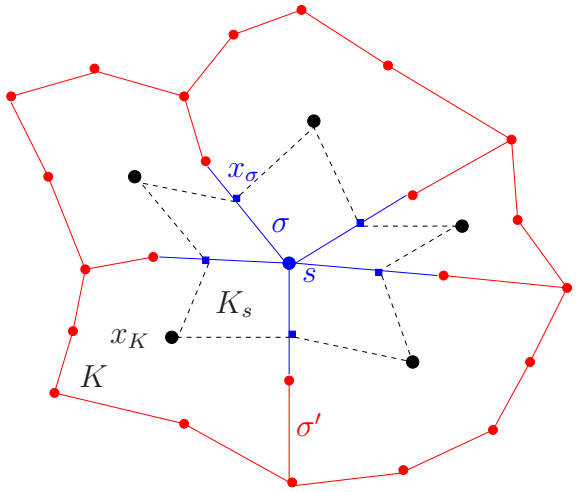


Figure III.4: The subcells around the vertex s and $\mathcal{E}_s$ the set of edges in blue.

- For any $K \in \mathcal{T}_s$, $s \in \mathcal{V}$, we assume that the approximation u of the solution of (II.1) is linear in the polytope K_s and continuous in x_σ, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$.

- Then, for any $K \in \mathcal{T}_s$, $s \in \mathcal{V}$, setting $u_K = u(x_K)$ and $u_\sigma = u(x_\sigma)$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ (see figure III.5), we can express $u(x)$ for all $x \in K_s$ as follows, $u(x) = u_K + g_K^s \cdot (x - x_K)$ with $g_K^s \in \mathbb{R}^d$.

- Since for any $K \in \mathcal{T}_s$, $s \in \mathcal{V}$, the family of vectors $\mathcal{B}_K^s = (x_\sigma - x_K)_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ is a basis of $\mathbb{R}^d$, there exists $(g_{K,\sigma})_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ a bi-orthogonal basis of $\mathcal{B}_K^s$ defined by Equations $g_{K,\sigma} \cdot (x_{\sigma'} - x_K) = \delta_{\sigma,\sigma'}$ for all $\sigma, \sigma' \in \mathcal{E}_K \cap \mathcal{E}_s$.

- Then, for any $K \in \mathcal{T}_s$, $s \in \mathcal{V}$, taking u in x_σ, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$, we obtain Equations $u_\sigma - u_K = g_K^s \cdot (x_\sigma - x_K)$ which imples that $\sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} (u_\sigma - u_K) g_{K,\sigma} = \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} g_K^s \cdot (x_\sigma - x_K) g_{K,\sigma} = g_K^s$.

- Moreover, we assume that the flux of u through the interior edges of the mesh is continuous. Then, we deduce that for each vertex of the mesh $s \in \mathcal{V}$, the

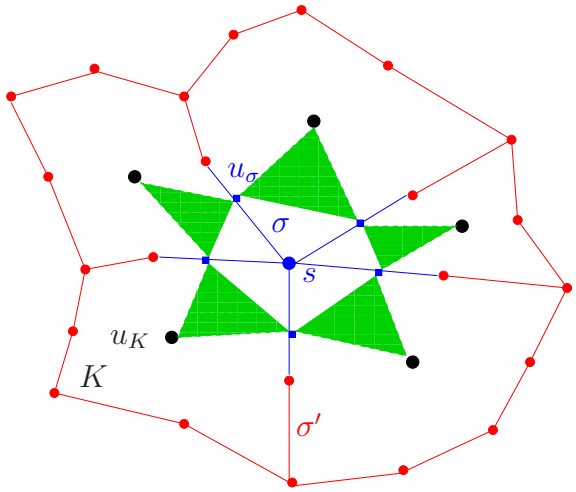


Figure III.5: Piecewise linear function u built from cell and edge unknowns

flux of u through the interior edges sharing the vertex s, $\sigma \in \mathcal{E}_s \cap \mathcal{E}_{\rm int}$ is continuous, which means that $\displaystyle\int_\sigma \Lambda_K \nabla u(x)_{/K} \cdot \mathbf{n}_{K,\sigma} + \int_\sigma \Lambda_L \nabla u(x)_{/L} \cdot \mathbf{n}_{L,\sigma} = 0$, with $\mathcal{T}_\sigma = \{K, L\}$.

- Therefore, we can eliminate the edge unknowns $(u_\sigma)_{\sigma\in\mathcal{E}_{\rm int}}$ in terms of $(u_K)_{K\in\mathcal{T}}$ using the flux continuity. Indeed, for any $\sigma \in \mathcal{E}_s \cap \mathcal{E}_{\rm int}$, $s \in \mathcal{V}$ with $\mathcal{T}_\sigma = \{K, L\}$, we get,

$$\begin{aligned} F_{K,\sigma} &= \int_\sigma \Lambda_K \nabla u(x)_{/K} \cdot \mathbf{n}_{K,\sigma} = \mathrm{m}_\sigma \Lambda_K g_K^s \cdot \mathbf{n}_{K,\sigma} \\ F_{L,\sigma} &= \int_\sigma \Lambda_L \nabla u(x)_{/L} \cdot \mathbf{n}_{L,\sigma} = \mathrm{m}_\sigma \Lambda_L g_L^s \cdot \mathbf{n}_{L,\sigma} \\ F_{K,\sigma} &+ F_{L,\sigma} = 0, \end{aligned} \qquad \text{[III.2]}$$

  and for any $\sigma \in \mathcal{E}_s \cap \mathcal{E}_{\rm ext}$, we impose

$$u_\sigma = 0. \qquad \text{[III.3]}$$

- It results that around each vertex $s \in \mathcal{V}$, the edge unknowns $(u_\sigma)_{\sigma\in\mathcal{E}_s\cap\mathcal{E}_{\rm int}}$ can be eliminated in terms of $(u_K)_{K\in\mathcal{T}_s}$ assuming the well-posedness of the linear system (III.2)-(III.3)

- Then, replacing the edge unknowns $(u_\sigma)_{\sigma\in\mathcal{E}}$ by their expression in terms of $(u_K)_{K\in\mathcal{T}}$ to the formula of the flux III.2, we deduce the cell unknowns $(u_K)_{K\in\mathcal{T}}$ by solving the following linear system,

$$-\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma} = \int_K f(x)\, dx, \;\; K \in \mathcal{T}. \qquad \text{[III.4]}$$

In our approach the finite volume scheme will be derived in section III.1.1 from a variational formulation defined on the space $\mathcal{H}_\mathcal{D}$ spanned by the cell and face unknowns and introduced below.

The definition of the finite volume scheme is based on a hybrid variational formulation on the space $\mathcal{H}_\mathcal{D}$ using the construction of two discrete gradients for each cell K of the mesh and each vertex s of the cell. The first gradient defined by

$$(\widetilde{\nabla}_\mathcal{D} u)_K^s = \frac{1}{\mathrm{m}_K^s} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \mathrm{m}_\sigma (u_\sigma - u_K) \mathbf{n}_{K,\sigma}, \qquad \text{[III.5]}$$

is built to have a weak convergence property using the Lemma 2, once averaged for each cell K over its vertices $s \in \mathcal{V}_K$ with the weights m_K^s. The second gradient is defined by

$$(\overline{\nabla}_\mathcal{D} u)_K^s = \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} (u_\sigma - u_K)\ g_{K,\sigma}, \qquad \text{[III.6]}$$

where the vectors $g_{K,\sigma} \in \mathbb{R}^d$ are given for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$. The gradient $(\overline{\nabla}_\mathcal{D} u)_K^s$ is built to be consistent in the sense that it is exact for linear functions. More precisely, thanks to the Lemma (6), the vectors $g_{K,\sigma}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ are assumed to satisfy the consistency property (III.7) defined within the following lemma.

Lemma 6 *[consistency of the gradient] Under Hypothesis (3), for all $K \in \mathcal{T}$, $s \in \mathcal{V}_K$, there exists vectors $g_{K,\sigma}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ such that for all vectors $v \in \mathbb{R}^d$ we have*

$$\sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} v \cdot (x_\sigma - x_K)\ g_{K,\sigma} = v. \qquad \text{[III.7]}$$

Proof. Thanks to Hypothesis (3), for all $K \in \mathcal{T}$, $s \in \mathcal{V}_K$, there exists $\mathcal{J}_{K,s} \subset \mathcal{E}_K \cap \mathcal{E}_s$ such that the family $(x_\sigma - x_K)_{\sigma \in \mathcal{J}_{K,s}}$ is a basis of $\mathbb{R}^d$.

Let us denote by $(\bar{g}_{K,\sigma})_{\sigma \in \mathcal{J}_{K,s}}$ the bi-orthogonal basis of the basis $(x_\sigma - x_K)_{\sigma \in \mathcal{J}_{K,s}}$ of $\mathbb{R}^d$. It is uniquely defined by Equations $\bar{g}_{K,\sigma} \cdot (x_{\sigma'} - x_K) = \delta_{\sigma,\sigma'}$ for all $\sigma, \sigma' \in \mathcal{J}_{K,s}$, then we deduce that for all $v \in \mathbb{R}^d$,

$$\sum_{\sigma \in \mathcal{J}_{K,s}} v \cdot (x_\sigma - x_K)\ \bar{g}_{K,\sigma} = v,$$

which conclude the proof.
□

Let us now define the bilinear form $a_\mathcal{D}$ on $\mathcal{H}_\mathcal{D} \times \mathcal{H}_\mathcal{D}$ by

$$\begin{aligned} a_\mathcal{D}(u,v) = \sum_{K \in \mathcal{T}} \sum_{s \in \mathcal{V}_K} \Bigg(& \mathrm{m}_K^s (\overline{\nabla}_\mathcal{D} u)_K^s \cdot \Lambda_K (\widetilde{\nabla}_\mathcal{D} v)_K^s \\ & + \alpha_K^s \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v) \Bigg) \end{aligned} \qquad \text{[III.8]}$$

for all $(u,v) \in \mathcal{H}_\mathcal{D} \times \mathcal{H}_\mathcal{D}$, with

$$\Lambda_K = \frac{1}{\mathrm{m}_K} \int_K \Lambda(x) dx,$$

for all $K \in \mathcal{T}$. In (III.8), the residual functions $R_{K,\sigma}$ are defined for all $u \in \mathcal{H}_\mathcal{D}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$, $s \in \mathcal{V}_K$, $K \in \mathcal{T}$, by

$$R_{K,\sigma}(u) = u_\sigma - u_K - (\overline{\nabla}_\mathcal{D} u)_K^s \cdot (x_\sigma - x_K), \qquad \text{[III.9]}$$

and the parameters α_K^s are real values such that

$$\mu_0 \leq \alpha_K^s \leq \gamma_0 \qquad \text{[III.10]}$$

for all $s \in \mathcal{V}_K$, $K \in \mathcal{T}$ with $\mu_0 > 0$ and $\gamma_0 > 0$. Note that instead of the scalar parameter α_K^s, we could have considered a more general positive definite matrix D_K^s of size $\mathrm{Card}(\mathcal{E}_K \cap \mathcal{E}_s)$ such that $\mu_0\, I \leq D_K^s \leq \gamma_0\, I$. The subsequent analysis will readily extends to this more general framework but we keep to the scalar term for the sake of simplicity in the notations.

The discretization of (II.5) on $\mathcal{D}$ is defined by the following discrete hybrid variational formulation: find $u_\mathcal{D} \in \mathcal{H}_\mathcal{D}$ such that

$$a_\mathcal{D}(u_\mathcal{D}, v) = \int_\Omega f(x) P_\mathcal{T} v(x) dx \quad \text{for all } v \in \mathcal{H}_\mathcal{D}. \qquad \text{[III.11]}$$

For all $u \in \mathcal{H}_\mathcal{D}$, let us introduce the following fluxes $F_{K,\sigma}(u)$ defined for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$, $s \in \mathcal{V}_K$, $K \in \mathcal{T}$ by

$$\begin{aligned} F_{K,\sigma}(u) \;&= -\mathrm{m}_\sigma \Lambda_K (\overline{\nabla}_\mathcal{D} u)_K^s \cdot \mathbf{n}_{K,\sigma} \\ &\quad - \alpha_K^s \left(\frac{\mathrm{m}_\sigma}{d_{K,\sigma}} R_{K,\sigma}(u) - g_{K,\sigma} \cdot \sum_{\sigma' \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}_{\sigma'}}{d_{K,\sigma'}} R_{K,\sigma'}(u)(x_{\sigma'} - x_K) \right), \end{aligned} \qquad \text{[III.12]}$$

in such a way that

$$a_\mathcal{D}(u,v) = \sum_{K \in \mathcal{T}} \sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u)(v_K - v_\sigma), \qquad \text{[III.13]}$$

for all $(u,v) \in \mathcal{H}_\mathcal{D} \times \mathcal{H}_\mathcal{D}$. Then, it is shown from (III.13) that the variational formulation (III.11) is equivalent to the following hybrid finite volume scheme: find $u_\mathcal{D} \in \mathcal{H}_\mathcal{D}$ such that

$$\left\{ \begin{array}{ll} \displaystyle\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma}(u_\mathcal{D}) = \int_K f(x) dx & \text{for all } K \in \mathcal{T}, \\ F_{K,\sigma}(u_\mathcal{D}) = -F_{L,\sigma}(u_\mathcal{D}) & \text{for all } \sigma \in \mathcal{E}_{\mathrm{int}},\ \mathcal{T}_\sigma = \{K, L\}. \end{array} \right. \qquad \text{[III.14]}$$

From definition (III.12) of the fluxes $F_{K,\sigma}(u)$, we can compute the coefficients $(T_K^s)_{\sigma,\sigma'}$, $\sigma' \in \mathcal{E}_s \cap \mathcal{E}_K$ such that

$$F_{K,\sigma}(u) = \sum_{\sigma' \in \mathcal{E}_s \cap \mathcal{E}_K} (T_K^s)_{\sigma,\sigma'}(u_K - u_{\sigma'}), \qquad \text{[III.15]}$$

for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ and $u \in \mathcal{H}_\mathcal{D}$. It results that around each vertex $s \in \mathcal{V}$, the face unknowns $(u_\sigma)_{\sigma \in \mathcal{E}_s}$ can be eliminated in terms of the $(u_K)_{K \in \mathcal{T}_s}$ assuming the well-posedness of the linear system

$$\begin{cases} F_{K,\sigma}(u_\mathcal{D}) + F_{L,\sigma}(u_\mathcal{D}) = 0 & \text{for all } \sigma \in \mathcal{E}_s \cap \mathcal{E}_{\rm int} \text{ with } \mathcal{T}_\sigma = \{K, L\}, \\ u_\sigma = 0 & \text{for all } \sigma \in \mathcal{E}_s \cap \mathcal{E}_{\rm ext}. \end{cases} \qquad \text{[III.16]}$$

Then, the hybrid finite volume scheme reduces to the cell centered finite volume scheme: find $u_\mathcal{T} \in H_\mathcal{T}(\Omega)$ such that for all $K \in \mathcal{T}$

$$\sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm int}, \mathcal{T}_\sigma = \{K,L\}} F_{K,L}(u_\mathcal{T}) \; + \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm ext}} F_\sigma(u_\mathcal{T}) = \int_K f(x)dx, \qquad \text{[III.17]}$$

where the inner fluxes $F_{K,L}(u_\mathcal{T})$, $\mathcal{T}_\sigma = \{K, L\}$, $\sigma \in \mathcal{E}_{\rm int}$, and the boundary fluxes $F_\sigma(u_\mathcal{T})$, $\sigma \in \mathcal{E}_{\rm ext}$, are linear combinations of the cell unknowns $(u_\mathcal{T})_M$ with $M \in \bigcup_{s \in \mathcal{V}_\sigma} \mathcal{T}_s$.

The well-posedness of the hybrid finite volume scheme (III.14), of the local linear systems (III.16), and of the cell centered scheme (III.17) is shown in the next section to result from the coercivity of the bilinear form $a_\mathcal{D}$ which will hold assuming a local coercivity assumption as stated in Proposition 3.

Equivalence with the usual MPFA O scheme

The MPFA O scheme described in the section III.1.1 is defined for polygonal and polyhedral meshes such that for all cells K and all vertices s of K, the cardinal of $\mathcal{E}_K \cap \mathcal{E}_s$ denoted by q_K^s is equal to the space dimension d, and such that the set of vectors $(x_\sigma - x_K)_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ spans $\mathbb{R}^d$. For such meshes, the MPFA O scheme from [1] or [60] is precisely defined by the hybrid finite volume formulation (III.14) using sub-fluxes $F_{K,\sigma}^s(u)$ given by $-{\rm m}_\sigma \Lambda_K (\nabla_\mathcal{D}^{MPFA} u)_K^s \cdot \mathbf{n}_{K,\sigma}$ where the gradient $(\nabla_\mathcal{D}^{MPFA} u)_K^s$ is the gradient of the unique linear function defined by its $d+1$ values u_K at point x_K and u_σ at points x_σ, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$.

In such cases, the equivalence between our hybrid finite volume scheme (III.14) and the MPFA O scheme defined in [1] and [60] readily results from the following lemma stating that $(\overline{\nabla}_\mathcal{D} u)_K^s = (\nabla_\mathcal{D}^{MPFA} u)_K^s$, and that $R_{K,\sigma}^s(u) = 0$ for all $u \in \mathcal{H}_\mathcal{D}$.

Lemma 7 *Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let $K \in \mathcal{T}$, $s \in \mathcal{V}_K$ be such that $q_K^s = d$ and such that the set of d vectors*

$(x_\sigma - x_K)_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}$ spans $\mathbb{R}^d$. Let us consider a discrete gradient $(\overline{\nabla}_\mathcal{D} u)_K^s$ given by (III.6) and satisfying the consistency hypothesis 6. Then, for all $u \in \mathcal{H}_\mathcal{D}$, the discrete gradient $(\overline{\nabla}_\mathcal{D} u)_K^s$ is the gradient of the unique linear function defined by its $d+1$ values u_K at point x_K and u_σ at points x_σ, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$, and the residuals $R_{K,\sigma}(u)$ vanish for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$.

PROOF. Let us denote by $(\bar{g}_{K,\sigma})_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}$ the bi-orthogonal basis of the basis $(x_\sigma - x_K)_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}$ of $\mathbb{R}^d$. It is uniquely defined by Equations $\bar{g}_{K,\sigma}\cdot(x_{\sigma'} - x_K) = \delta_{\sigma,\sigma'}$ for all $\sigma, \sigma' \in \mathcal{E}_K \cap \mathcal{E}_s$. Setting $v = \bar{g}_{K,\sigma}$ in (III.7) shows that $g_{K,\sigma} = \bar{g}_{K,\sigma}$ for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ and the gradient $\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}(u_\sigma - u_K)\bar{g}_{K,\sigma}$ is the unique gradient satisfying the consistency hypothesis 6. Let $u \in \mathcal{H}_\mathcal{D}$ be given and let φ be the unique linear function defined by its $d+1$ values u_K at point x_K and u_σ at points x_σ, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$. We have by definition $\nabla\varphi\cdot(x_\sigma - x_K) = u_\sigma - u_K$. Hence setting $v = \nabla\varphi$ in (III.7) it results that

$$\nabla\varphi = \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s} \nabla\varphi\cdot(x_\sigma - x_K)\bar{g}_{K,\sigma} = \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s} (u_\sigma - u_K)\bar{g}_{K,\sigma},$$

which proves the first part of the lemma. The second part results from the equation

$$\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s} R_{K,\sigma}(u)\ \bar{g}_{K,\sigma} = 0,$$

for all $u \in \mathcal{H}_\mathcal{D}$.
□

For cells such that $q_K^s > d$, they are several ways to define a gradient $(\overline{\nabla}_\mathcal{D} u)_K^s = (u_\sigma - u_K)\ g_{K,\sigma}$ satisfying the consistency hypothesis 6. In such cases, the residuals $R_{K,\sigma}(u)$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ satisfying the relation $\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s} R_{K,\sigma}(u)\ g_{K,\sigma} = 0$ for all $u \in \mathcal{H}_\mathcal{D}$ do not a priori vanish since the family $g_{K,\sigma}$, $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ is not free. They play the role of stabilization terms in the hybrid variational formulation (III.14) as shown in the following example. For $d = 3$, let us consider two pyramids K and L sharing a triangular face σ, and let $s \in \sigma$ denote the top of the two pyramids. We can build two consistent gradients $(\overline{\nabla}_\mathcal{D} u)_K^s$ and $(\overline{\nabla}_\mathcal{D} u)_L^s$ such that $g_{K,\sigma} = g_{L,\sigma} = 0$. Then, the residuals $R_{K,\sigma'}(u)$ and $R_{K,\sigma''}(u)$ vanish except for $\sigma' = \sigma'' = \sigma$. In this example, it is clear that only the residual terms in (III.14) can ensure the well-posedness of the system since the discrete gradients (III.6) do not depend on u_σ.

III.1.2 Coercivity of the scheme

The well-posedness of the hybrid finite volume scheme (III.14) and the cell centered finite volume scheme (III.17) will be derived from the coercivity of the bilinear form $a_\mathcal{D}$. This coercivity property depends on the finite volume discretization $\mathcal{D}$, on the diffusion tensor Λ, and on the parameters of the finite volume scheme. In the

following, we shall make the stronger assumption that the coercivity holds locally around each vertex s of the mesh. For a given discretization and diffusion tensor, this assumption can be numerically checked by computing the eigenvalues of a small linear system of size $2 \times \mathrm{Card}(\mathcal{T}_s)$ for each vertex $s \in \mathcal{V}$.

Let s be a given vertex in $\mathcal{V}$, and let $\mathcal{H}_{\mathcal{D}}^s$ be the subspace of

$$\{u_\sigma \in \mathbb{R}, u_K \in \mathbb{R},\ K \in \mathcal{T}_s, \sigma \in \mathcal{E}_K \cap \mathcal{E}_s\}$$

such that $u_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\mathrm{ext}}$. The space $\mathcal{H}_{\mathcal{D}^s}$ is endowed with the semi-norm

$$\|u\|_{\mathcal{D}^s} = \left(\sum_{K \in \mathcal{T}_s} \sum_{\sigma \in \mathcal{E}_s \cap \mathcal{E}_K} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} (u_\sigma - u_K)^2 \right)^2.$$

Let us also denote by $a_{\mathcal{D}^s}$ the bilinear form defined by

$$\begin{aligned} a_{\mathcal{D}^s}(u,v) &= \sum_{K \in \mathcal{T}_s} \mathrm{m}_K^s (\overline{\nabla}_{\mathcal{D}} u)_K^s \cdot \Lambda_K (\widetilde{\nabla}_{\mathcal{D}} v)_K^s \\ &\qquad + \alpha_K^s \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v) \\ &= \sum_{K \in \mathcal{T}_s} \sum_{\sigma,\sigma' \in \mathcal{E}_K \cap \mathcal{E}_s} (T_K^s)_{\sigma,\sigma'} (u_{\sigma'} - u_K)(u_\sigma - u_K), \end{aligned} \qquad \text{[III.18]}$$

for all $u, v \in \mathcal{H}_{\mathcal{D}^s}$, where we have used definition (III.15) of the coefficients $(T_K^s)_{\sigma,\sigma'}$, and the canonical injection from $\mathcal{H}_{\mathcal{D}^s}$ to $\mathcal{H}_{\mathcal{D}}$ to define the residual and the gradient functions on $\mathcal{H}_{\mathcal{D}^s}$.

Let us now define the following local coercivity measurement

$$\mathrm{coernode}(\mathcal{D}, \Lambda) = \min_{s \in \mathcal{V}} \inf_{\{u \in \mathcal{H}_{\mathcal{D}^s} \mid \|u\|_{\mathcal{D}^s} = 1\}} a_{\mathcal{D}^s}(u,u). \qquad \text{[III.19]}$$

It will be used to check the coercivity of the bilinear form $a_{\mathcal{D}}$ as stated in the following proposition:

Proposition 3 *Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that there exists $\theta_{\mathcal{D}} > 0$ such that* $\mathrm{coernode}(\mathcal{D}, \Lambda) \geq \theta_{\mathcal{D}}$. *Then, the bilinear form $a_{\mathcal{D}}$ is coercive in the sense that for all $u \in \mathcal{H}_{\mathcal{D}}$ we have*

$$a_{\mathcal{D}}(u,u) \geq \theta_{\mathcal{D}} \|u\|_{\mathcal{D}}^2. \qquad \text{[III.20]}$$

Proof. Using (III.15) we have for any $u \in \mathcal{H}_{\mathcal{D}}$

$$a_{\mathcal{D}}(u,u) = \sum_{s \in \mathcal{V}} \sum_{K \in \mathcal{T}_s} \sum_{\sigma,\sigma' \in \mathcal{E}_K \cap \mathcal{E}_s} (T_K^s)_{\sigma,\sigma'} (u_{\sigma'} - u_K)(u_\sigma - u_K).$$

Using definition (III.19), and the assumption coernode$(\mathcal{D},\Lambda) \geq \theta_{\mathcal{D}}$, the following estimate

$$a_{\mathcal{D}}(u,u) \geq \theta_{\mathcal{D}} \sum_{s\in\mathcal{V}} \sum_{K\in\mathcal{T}_s} \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(u_\sigma - u_K)^2,$$

is derived, which proves the lemma.
□
The following propositions state the well-posedness of the hybrid and cell centered finite volume schemes under the local coercivity assumption.

III.1.3 Consistency of the scheme

Hypothesis 4 *Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1. We always assume in the following that*

- *the local coercivity assumption* coernode$(\mathcal{D},\Lambda) \geq \theta_{\mathcal{D}} > 0$ *is satisfied, which ensures that there exists a unique solution $u_{\mathcal{D}} \in \mathcal{H}_{\mathcal{D}}$ to (III.14).*
- *there exist $0 < \varrho_5 < +\infty$ independent of the discretization $\mathcal{D}$ s.t.*

$$\min_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s, K\in\mathcal{T}_s, s\in\mathcal{V}} \frac{\mathrm{m}_\sigma}{\mathrm{m}_K^s |g_{K,\sigma}^s|} \geq \varrho_5. \qquad \text{[III.21]}$$

Lemma 8 *[Consistency of the discrete gradients] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that hypothesis 6 holds. Let φ be a given function in $C_c^\infty(\Omega)$. Then, there exists M_φ depending only on φ, such that for all $s \in \mathcal{V}_K$, $K \in \mathcal{T}$,*

$$|(\overline{\nabla}_{\mathcal{D}} P_{\mathcal{D}}\varphi)_K^s - \nabla\varphi(x_K)| \leq M_\varphi \, \frac{d}{\varrho_1\varrho_5} \, \mathrm{diam}(K),$$

Proof. Let $K \in \mathcal{T}$, $s \in \mathcal{V}_K$, $\varphi \in C_c^\infty(\Omega)$ be given. For all $\sigma \in \mathcal{E}_s \cap \mathcal{E}_K$, let us set $\epsilon_{K,\sigma}^s = \varphi(x_\sigma) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(x_\sigma - x_K)$. Since $\varphi \in C_c^\infty(\Omega)$, there exists a real $M_\varphi > 0$ depending only on φ such that $|\epsilon_{K,\sigma}^s| \leq M_\varphi |x_\sigma - x_K|^2$. From hypothesis 6, we have

$$(\overline{\nabla}_{\mathcal{D}} P_{\mathcal{D}}\varphi)_K^s - \nabla\varphi(x_K) = \sum_{\sigma\in\mathcal{E}_s\cap\mathcal{E}_K} \epsilon_{K,\sigma}^s \, g_{K,\sigma},$$

which ends the proof from the definitions of ϱ_5, m_K^s and ϱ_1.
□

Lemma 9 *[Consistency of the residual functions] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that hypothesis 6 holds. Let*

φ be a given function in $C_c^\infty(\Omega)$. Then, there exists a real $C>0$ depending only on φ, ϱ_1, ϱ_5 and Ω, such that for all $K\in\mathcal{T}$, $s\in\mathcal{V}_K$, $\sigma\in\mathcal{E}_s\cap\mathcal{E}_K$,

$$|R_{K,\sigma}(P_{\mathcal{D}}\varphi)|\leq C\left(1+\frac{d}{\varrho_1\varrho_5}\right)\ \mathrm{diam}(K)^2.$$

Proof. Let $K\in\mathcal{T}$, $s\in\mathcal{V}_K$ and $\varphi\in C_c^\infty(\Omega)$ be given. For all $\sigma\in\mathcal{E}_s\cap\mathcal{E}_K$, let us set $\epsilon^s_{K,\sigma}=\varphi(x_\sigma)-\varphi(x_K)-\nabla\varphi(x_K)\cdot(x_\sigma-x_K)$. Since $\varphi\in C_c^\infty(\Omega)$, there exists a real $M_\varphi>0$ already introduced in Lemma 8 and depending only on φ such that $|\epsilon^s_{K,\sigma}|\leq M_\varphi|x_\sigma-x_K|^2$. From the definition of the residual function we have

$$R_{K,\sigma}(P_{\mathcal{D}}\varphi)=\epsilon^s_{K,\sigma}-\left((\overline{\nabla}_{\mathcal{D}}P_{\mathcal{D}}\varphi)^s_K-\nabla\varphi(x_K)\right)\cdot(x_\sigma-x_K).$$

We deduce from Lemma 8, and the definition of ϱ_1 that

$$|R_{K,\sigma}(P_{\mathcal{D}}\varphi)|\leq M_\varphi\left(1+\frac{d}{\varrho_1\varrho_5}\right)\ \mathrm{diam}(K)^2,$$

which concludes the proof.
□

Lemma 10 *[Consistency of the flux functions] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that hypothesis 6 holds. Let φ be a given function in $C_c^\infty(\Omega)$. Then, there exists a real $C>0$ depending only on φ, ϱ_1, ϱ_5, β_0 and d, such that for all $K\in\mathcal{T}$, $s\in\mathcal{V}_K$, $\sigma\in\mathcal{E}_s\cap\mathcal{E}_K$,*

$$\left|F_{K,\sigma}(P_{\mathcal{D}}\varphi)-\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\int_K\Lambda(x)\nabla\varphi(x)dx\right|\leq C\ \mathrm{m}_\sigma\mathrm{diam}(K)$$

Proof. From the definition (III.12) of the flux functions $F_{K,\sigma}$, we deduce that,

$$\begin{aligned}\left|F_{K,\sigma}(P_{\mathcal{D}}\varphi)-\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\int_K\Lambda(x)\nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}\right|&\leq|F_{K,\sigma}(P_{\mathcal{D}}\varphi)-\mathrm{m}_\sigma\Lambda_K\nabla\varphi(x_K)\cdot\mathbf{n}_{K,\sigma}|\\&+\left|\mathrm{m}_\sigma\Lambda_K\nabla\varphi(x_K)\cdot\mathbf{n}_{K,\sigma}-\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\Lambda_K\int_K\nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}\right|\\&+\left|\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\Lambda_K\int_K\nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}-\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\int_K\Lambda(x)\nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}\right|\end{aligned}\qquad[\text{III.22}]$$

For the first term at the right hand side of the inequality (III.22), thanks to Lemmata 8 and 9, we infer that there exists a real $C_0>0$ depending only on φ, ϱ_1, ϱ_5, β_0 and d such that,

$$|F_{K,\sigma}(P_{\mathcal{D}}\varphi)-\mathrm{m}_\sigma\Lambda_K\nabla\varphi(x_K)\cdot\mathbf{n}_{K,\sigma}|\leq C_0\ \mathrm{m}_\sigma\mathrm{diam}(K).\qquad[\text{III.23}]$$

For the last two terms at the right hand side of the inequality (III.22), thanks to the regularity of $\varphi \in C_c^\infty(\Omega)$ and the assumption on Λ (see (II.3)), we deduce that there exists a real $C_\varphi > 0$ depending only on φ such that,

$$\left|\mathrm{m}_\sigma \Lambda_K \nabla\varphi(x_K)\cdot\mathbf{n}_{K,\sigma} - \frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\Lambda_K \int_K \nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}\right| \leq C_\varphi\beta_0\ \mathrm{m}_\sigma \mathrm{diam}(K) \quad \text{[III.24]}$$

and also

$$\left|\frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\Lambda_K \int_K \nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma} - \frac{\mathrm{m}_\sigma}{\mathrm{m}_K}\int_K \Lambda(x)\nabla\varphi(x)dx\cdot\mathbf{n}_{K,\sigma}\right| \leq C_\varphi\beta_0\ \mathrm{m}_\sigma\mathrm{diam}(K). \quad \text{[III.25]}$$

Gathering the results III.23-III.25, we conclude the proof from III.22.
□

III.2 The G method

In the present section we introduce a family of FV methods generalizing the MPFA L method of [6, 7]. The idea is to write the flux through a face as the weighted average of several L-type fluxes corresponding to different stencils. A proper choice of the weights allows to enhance the coercivity of the method, thereby improving robustness with respect to the skewdness of the mesh and to the anisotropy and heterogeneity of the permeability tensor. In this section, we assume that there exists a finite partition of Ω into open connected disjoint polygonal subsets,
$P_\Omega \stackrel{\text{def}}{=} \{\Omega_i\}_{i=1\ldots N_\Omega}$ such that (s.t.) $\Lambda_{|\Omega_i} \in [C^2(\overline{\Omega_i})]^{d\times d}$ for all $i = 1\ldots N_\Omega$. Then thanks to Lemma 3 and Remark 1, we can use the space $\mathcal{Q}$ as the space of test functions instead of $C_0^2(\Omega)$.

III.2.1 The Finite Volume scheme

Hypothesis 5 *Let $\{\mathcal{D}_n\}_{n\in\mathbb{N}}$ be a family of meshes matching Definition II.2.1 satisfying hypothese (1). We further suppose that*

(i) there exists a non-negative constant ϱ_4 independent of n such that

$$\max_{K\in\mathcal{T}_n}\max_{\sigma\in\mathcal{E}_K}\frac{\mathrm{diam}(K)^{d-1}}{\mathrm{m}_\sigma} \leq \varrho_4.$$

(ii) we assume that each σ contains at least one vertex (this could be false in dimension $d = 3$ if, for example, σ is only a piece of a "true face" of a cell);

(iii) $\widetilde{\mathcal{G}}$ is the finite family of face groups *defined as follows (see figure III.6-III.7):*

$$\widetilde{\mathcal{G}} \stackrel{\text{def}}{=} \{G \subset \mathcal{E}_K \cap \mathcal{E}_s,\ K\in\mathcal{T}_s,\ s\in\mathcal{V}_n,\ \mathrm{card}(G) = d\}.$$

For each $G \in \widetilde{\mathcal{G}}$, we let $\mathcal{T}_G = \{K\in\mathcal{T},\ G\cap\mathcal{E}_K \neq \emptyset\}$. We also arbitrarily select a cell, which we denote by K_G, such that $G \subset \mathcal{E}_{K_G}$.

(iv) $\mathcal{F} = \emptyset$,
(v) for all $K \in \mathcal{T}$, $\mathcal{E}_{\text{int},K} = \emptyset$,
(vi) for all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, $d_{K,\sigma}$ *is the Euclidian distance from* x_K *to* σ.

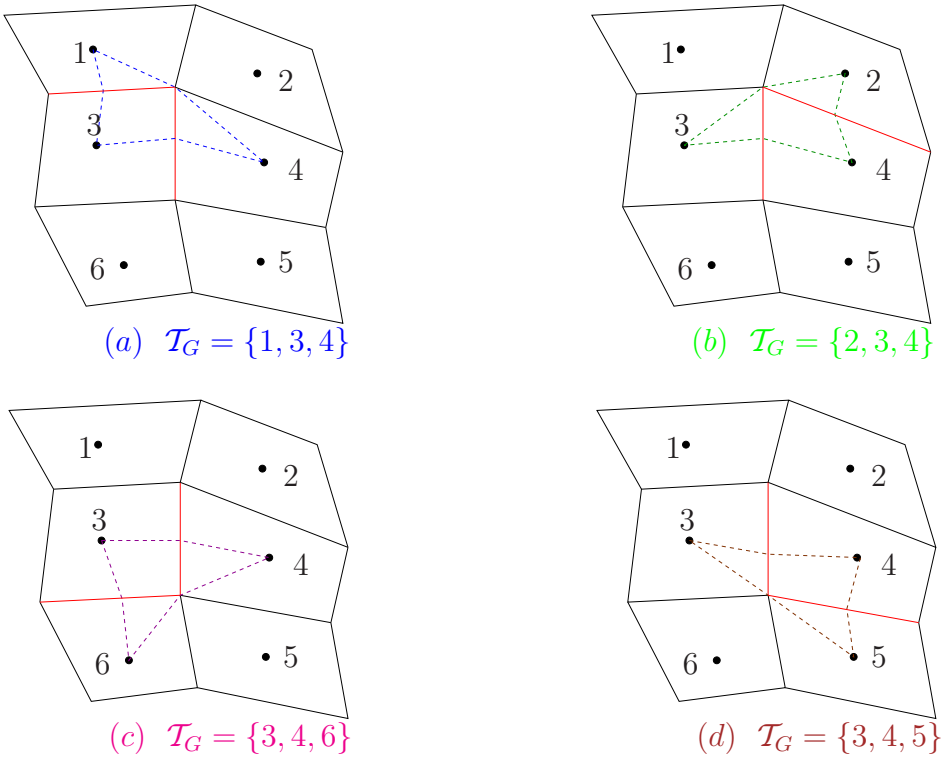
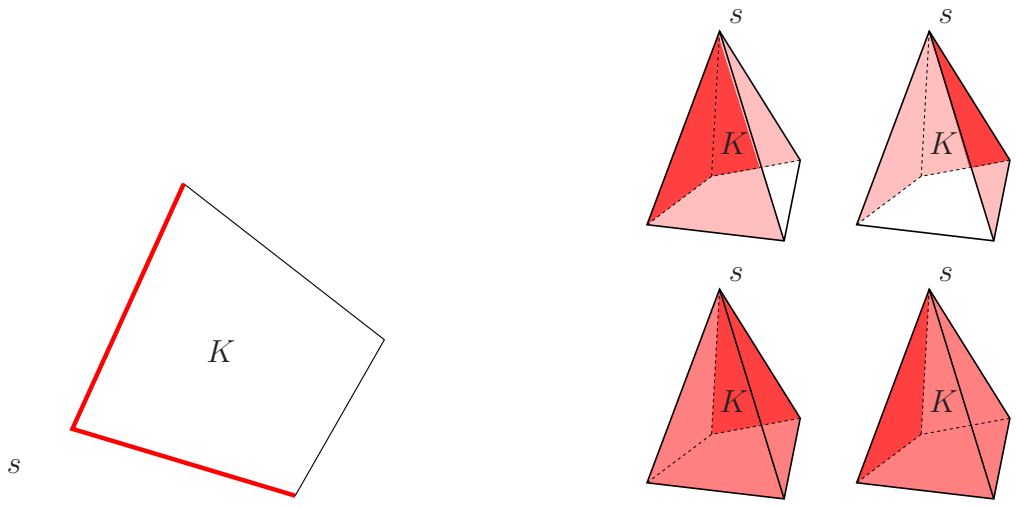


Figure III.6: Different groups of edges, $G \in \widetilde{\mathcal{G}}$ and group of cells $\mathcal{T}_G$ in the 2D case

The groups of edges $G \in \widetilde{\mathcal{G}}$ are the red edges in the figure III.6 and the cells of $\mathcal{T}_G$ are represented by numbers in the figure III.6.



(a) $d = 2$ (b) $d = 3$

Figure III.7: Groups of $\widetilde{\mathcal{G}}$ associated with a vertex s and a cell $K \in \mathcal{T}_s$

Remark 2 *In many cases, a given group G is contained in a unique $\mathcal{E}_K$ but, in some cases (especially if the discretization has non-convex cells), there can be multiple possible choices for K_G.*

Construction of group gradients

The idea of construction of the G scheme is based on the L-technique introduced in [6]. The L-technique allows to built a piecewise linear function on a connected domain, obtained from $d+1$ control volume and which takes into account the heterogeneous diffusion matrix. This $d+1$ control volume are such that there exists a control volume K, among the $d+1$ control volume, sharing a same vertex with the other and having a common edge with each other control volume (see figure III.8). Then, this $d+1$ control volume can be deduced from a group of edges $G \in \widetilde{\mathcal{G}}$ (see figure III.8), in other word they are the elements of the set $\mathcal{T}_G$. Therefore, for a given group of edges $G \in \widetilde{G}$, The L-technique allows to built a piecewise linear function, u from the values $(u_K)_{K\in\mathcal{T}_G}$ on the domain obtained from the set $\mathcal{T}_G$. The linear function u, must verify the three following conditions :

- for all $K \in \mathcal{T}_G$, $u(x_K) = u_K$ (= cell unknown)
- u is continuous on $\sigma \in G$.
- For any $\sigma \in G$, the flux through the edge σ is continuous, in other words $\int_\sigma \Lambda_K \nabla u_{/K}(x) \cdot \mathbf{n}_{K,\sigma} + \int_\sigma \Lambda_K \nabla u_{/L}(x) \cdot \mathbf{n}_{L,\sigma} = 0$, for which $\mathcal{T}_\sigma = \{K, L\}$.

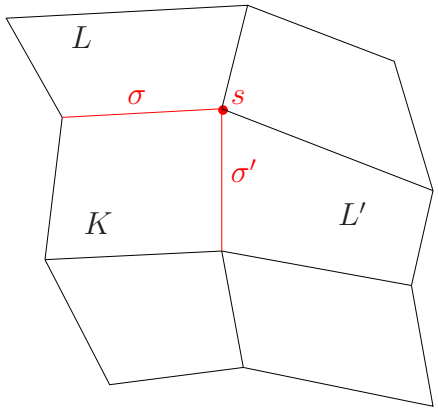


Figure III.8: s a vertex, $G = \{\sigma, \sigma'\}$, $\mathcal{T}_G = \{K, L, L'\}$ and $K_G = K$.

Since u is linear in each $K \in \mathcal{T}_G$, we can write u under the form : for all $x \in K$, $K \in \mathcal{T}_G$,

$$u(x) = u_K + (\nabla_{\mathcal{D}} u)_K^G \cdot (x - x_K)$$

where $(\nabla_{\mathcal{D}} u)_K^G$ is a vector of $\mathbb{R}^d$. The two last conditions yields to the following system of equations : for all $\sigma \in G$, with $\mathcal{T}_\sigma = \{K, L\}$,

$$\begin{cases} u_K + (\nabla_{\mathcal{D}} u)_K^G \cdot (x - x_{K_G}) = u_L + (\nabla_{\mathcal{D}} u)_L^G \cdot (x - x_L) \quad \forall x \in \sigma, \\ \Lambda_K (\nabla_{\mathcal{D}} u)_K^G \cdot \mathbf{n}_{K,\sigma} + \Lambda_L (\nabla_{\mathcal{D}} u)_L^G \cdot \mathbf{n}_{L,\sigma} = 0, \end{cases} \qquad \text{[III.26]}$$

The system of equations (III.26) corresponds to the system of equations III.27, where the Lemma 11 gives a solution to this system.

Thanks to the L-technique, from a set of cell unknowns $\{v_K \in \mathbb{R}, K \in \mathcal{T}\}$, for any group of edges $G \in \widetilde{\mathcal{G}}$, we can built the discrete gradient of $\mathbb{R}^d$, $(\nabla_{\mathcal{D}} v)_K^G$, for each cell, $K \in \mathcal{T}_G$ of the group G (see figure III.8 corresponding to III.6-(a)).
As, it is illustrated in the figure (III.6), a cell $K \in \mathcal{T}$ can belong to several groups of edges $G' \in \widetilde{\mathcal{G}}$. Therefore to write a flux $F_{K,\sigma}(v)$, $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, we take all the groups G sharing σ and such that $K \in \mathcal{T}_G$, we make a convex combination of the gradients, $(\nabla_{\mathcal{D}} v)_K^G$ and we use the result to built the flux, $F_{K,\sigma}(v)$.

Let us see now the details. Our idea is, for all $v \in H_{\mathcal{T}}(\Omega)$ which represents the set of cell unknowns $(v_K)_{K \in \mathcal{T}}$, all group $G \in \widetilde{\mathcal{G}}$, all $\sigma \in G$ and all $K \in \mathcal{T}_\sigma$, to build a "group gradient" $(\nabla_{\mathcal{D}} v)_K^{G,\sigma} \in \mathbb{R}^d$ and use it to define the flux $F_{K,\sigma}(u)$; this gradient could be understood as a *piece* of a full gradient of u on the pyramid $\triangle_{K,\sigma}$, the full gradient (and resulting flux) being obtained as a convex combination of these group gradients corresponding to all the groups G containing σ (see [III.32]).

First, for all $\sigma \in \mathcal{E}$, $\mathcal{T}_\sigma = \{K, L\}$, we require that, if $\mathcal{T}_\sigma = \{K, L\}$, the values v_K, v_L and the gradient reconstruction $(\nabla_\mathcal{D} v)_K^{G,\sigma}$, $(\nabla_\mathcal{D} v)_L^{G,\sigma}$ yield the same value of v on σ, that is to say

$$v_K + (\nabla_\mathcal{D} v)_K^{G,\sigma} \cdot (x - x_K) = v_L + (\nabla_\mathcal{D} v)_L^{G,\sigma} \cdot (x - x_L) \quad \forall x \in \sigma.$$

For boundary faces, we ask that the value obtained at $x = x_\sigma$, barycenter of σ be zero. Second, we would like the resulting fluxes to be conservative, *i.e.* ,

$$\Lambda_K (\nabla_\mathcal{D} v)_K^{G,\sigma} \cdot \mathbf{n}_{K,\sigma} + \Lambda_L (\nabla_\mathcal{D} v)_L^{G,\sigma} \cdot \mathbf{n}_{L,\sigma} = 0.$$

These two sets of equations are not sufficient to define uniquely the group gradients (and thus to estimate them, which is fundamental in the study of the numerical method). We therefore add another constraint, giving a particular role to the cell K_G selected for the group G: we ask that $(\nabla_\mathcal{D} v)_{K_G}^{G,\sigma}$ does not depend on $\sigma \in G$, and we denote by $(\nabla_\mathcal{D} v)_{K_G}^{G}$ the common value of this group gradient for all $\sigma \in G$. The discrete gradients are thus defined, as in [6, 7], by: For all $G \in \widetilde{\mathcal{G}}$ and all $\sigma \in G \cap \mathcal{E}_{\rm int}$, with $\mathcal{T}_\sigma = \{K_G, L\}$,

$$\begin{cases} v_{K_G} + (\nabla_\mathcal{D} v)_{K_G}^{G} \cdot (x - x_{K_G}) = v_L + (\nabla_\mathcal{D} v)_L^{G,\sigma} \cdot (x - x_L) \quad \forall x \in \sigma, \\ \Lambda_{K_G} (\nabla_\mathcal{D} v)_{K_G}^{G} \cdot \mathbf{n}_{K_G,\sigma} + \Lambda_L (\nabla_\mathcal{D} v)_L^{G,\sigma} \cdot \mathbf{n}_{L,\sigma} = 0, \end{cases} \qquad \text{[III.27]}$$

and for all $\sigma \in G \cap \mathcal{E}_{\rm ext}$,

$$v_{K_G} + (\nabla_\mathcal{D} v)_{K_G}^{G} \cdot (x_\sigma - x_{K_G}) = 0. \qquad \text{[III.28]}$$

Lemma 11 *The gradient reconstruction $(\nabla_\mathcal{D} v)_{K_G}^{G}$ defined by* [III.27] *and* [III.28] *can be obtained solving a linear system of the form*

$$\mathcal{A}_G X_G = \mathcal{B}_G(v), \qquad \text{[III.29]}$$

where the rows of $\mathcal{A}_G \in \mathbb{R}^{d \times d}$ are built from the following family of vectors of $\mathbb{R}^d$

$$\left(\left\{ \frac{\Lambda_L \mathbf{n}_{L,\sigma} \cdot \mathbf{n}_{L,\sigma}}{d_{L,\sigma}} (x_L - x_{K_G}) + \Lambda_{K_G} \mathbf{n}_{K_G,\sigma} + \Lambda_L \mathbf{n}_{L,\sigma} \right\}_{\sigma \in G \cap \mathcal{E}_{\rm int}}, \right.$$
$$\left. \left\{ \frac{\Lambda_{K_G} \mathbf{n}_{K_G,\sigma} \cdot \mathbf{n}_{K_G,\sigma}}{d_{K_G,\sigma}} (x_\sigma - x_{K_G}) \right\}_{\sigma \in G \cap \mathcal{E}_{\rm ext}} \right)$$

and $\mathcal{B}_G(v) \in \mathbb{R}^d$ is obtained from the family of vectors

$$\left(\left\{ \frac{\Lambda_L \mathbf{n}_{L,\sigma} \cdot \mathbf{n}_{L,\sigma}}{d_{L,\sigma}} (v_L - v_{K_G}) \right\}_{\sigma \in G \cap \mathcal{E}_{\rm int}}, \left\{ \frac{\Lambda_{K_G} \mathbf{n}_{K_G,\sigma} \cdot \mathbf{n}_{K_G,\sigma}}{d_{K_G,\sigma}} (-v_{K_G}) \right\}_{\sigma \in G \cap \mathcal{E}_{\rm ext}} \right).$$

PROOF. Let $v \in H_{\mathcal{T}}(\Omega)$, $G \in \mathcal{G}$, $\sigma \in G \cap \mathcal{E}_{\rm int}$ with $\mathcal{T}_\sigma = \{K_G, L\}$. Observe that, if $\mathbf{v} \stackrel{\rm def}{=} (\nabla_{\mathcal{D}} v)^G_{K_G} - (\nabla_{\mathcal{D}} v)^{G,\sigma}_L \neq 0$, the first equation of [III.27] is the equation of an hyperplane of $\mathbb{R}^d$ orthogonal to $\mathbf{v}$; satisfying this equation for all $x \in \sigma$ is equivalent to imposing that σ is contained in this hyperplane, and thus that $\mathbf{v}$ and $\mathbf{n}_{K_G,\sigma}$ are colinear (this is of course also true if $\mathbf{v} = 0$). As a consequence, taking $y_\sigma \in \sigma$, the first equation in [III.27] is equivalent to the following linear system (in which $\lambda^G_\sigma \in \mathbb{R}$ is an additional unknown):

$$\begin{cases} (\nabla_{\mathcal{D}} v)^G_{K_G} - (\nabla_{\mathcal{D}} v)^{G,\sigma}_L = \lambda^G_\sigma \mathbf{n}_{K_G,\sigma}, \\ v_{K_G} - v_L + (\nabla_{\mathcal{D}} v)^{G,\sigma}_L \cdot x_L - (\nabla_{\mathcal{D}} v)^G_{K_G} \cdot x_{K_G} = -\lambda^G_\sigma \mathbf{n}_{K_G,\sigma} \cdot y_\sigma. \end{cases}$$

Since $(y_\sigma - x_L)\cdot \mathbf{n}_{K_G,\sigma} = -d_{L,\sigma}$, solving for λ^G_σ we obtain

$$\begin{cases} \lambda^G_\sigma = -\frac{R_{L,\sigma}(v)}{d_{L,\sigma}}, \\ (\nabla_{\mathcal{D}} v)^{G,\sigma}_L = (\nabla_{\mathcal{D}} v)^G_{K_G} - \frac{R_{L,\sigma}(v)}{d_{L,\sigma}} \mathbf{n}_{L,\sigma}, \end{cases}$$

with $R_{L,\sigma}(v) \stackrel{\rm def}{=} v_L - v_{K_G} - (\nabla_{\mathcal{D}} v)^G_{K_G} \cdot (x_L - x_{K_G})$. Using these expressions, the second equation of [III.27] can be rewritten as

$$\left[\Lambda_{K_G} \mathbf{n}_{K_G,\sigma} + \Lambda_L \mathbf{n}_{L,\sigma} + \frac{\Lambda_L \mathbf{n}_{L,\sigma}\cdot\mathbf{n}_{L,\sigma}}{d_{L,\sigma}}(x_L - x_{K_G})\right] \cdot (\nabla_{\mathcal{D}} v)^G_{K_G} = \frac{\Lambda_L \mathbf{n}_{L,\sigma}\cdot\mathbf{n}_{L,\sigma}}{d_{L,\sigma}}(v_L - v_{K_G}).$$

Finally, the linear system [III.27]–[III.28] is equivalent to:

$$\begin{cases} (\nabla_{\mathcal{D}} v)^{G,\sigma}_L = (\nabla_{\mathcal{D}} v)^G_{K_G} - \dfrac{R_{L,\sigma}(v)}{d_{L,\sigma}} \mathbf{n}_{L,\sigma}\,, \quad \forall \sigma \in G \cap \mathcal{E}_{\rm int}\,,\ \mathcal{T}_\sigma = \{K_G, L\}, \\ \left[\dfrac{\Lambda_L \mathbf{n}_{L,\sigma}\cdot\mathbf{n}_{L,\sigma}}{d_{L,\sigma}}(x_L - x_{K_G}) + \Lambda_{K_G}\mathbf{n}_{K_G,\sigma} + \Lambda_L \mathbf{n}_{L,\sigma}\right] \cdot (\nabla_{\mathcal{D}} v)^G_{K_G} \\ \qquad\qquad = \dfrac{\Lambda_L \mathbf{n}_{L,\sigma}\cdot\mathbf{n}_{L,\sigma}}{d_{L,\sigma}}(v_L - v_{K_G})\,, \quad \forall \sigma \in G \cap \mathcal{E}_{\rm int}\,,\ \mathcal{T}_\sigma = \{K_G, L\}, \\ \dfrac{\Lambda_{K_G}\mathbf{n}_{K_G,\sigma}\cdot\mathbf{n}_{K_G,\sigma}}{d_{K_G,\sigma}} (\nabla_{\mathcal{D}} v)^G_{K_G} \cdot (x_\sigma - x_{K_G}) = \dfrac{\Lambda_{K_G}\mathbf{n}_{K_G,\sigma}\cdot\mathbf{n}_{K_G,\sigma}}{d_{K_G,\sigma}}(-v_{K_G}),\ \forall \sigma \in G \cap \mathcal{E}_{\rm ext}. \end{cases} \quad \text{[III.30]}$$

The assert follows.
□

In order to construct the gradient, we therefore need to consider the set of groups such that the matrix $\mathcal{A}_G$ is invertible, that is to say

$$\mathcal{G} \stackrel{\rm def}{=} \{G \in \widetilde{\mathcal{G}} \,|\, \mathcal{A}_G \text{ is invertible}\}.$$

We shall also need a symbol for the family of groups containing a give face $\sigma \in \mathcal{E}$. We thus let (see figure III.9)

$$\forall \sigma \in \mathcal{E}, \quad \mathcal{G}_\sigma \stackrel{\rm def}{=} \{G \in \mathcal{G} \,|\, \sigma \in G\},$$

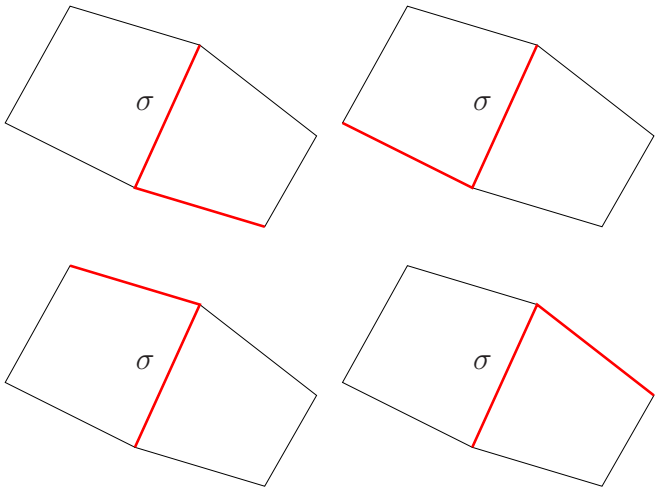


Figure III.9: Groups of $\mathcal{G}_\sigma$ for $d = 2$, $\sigma \in \mathcal{E}_{\rm int}$

and we assume throughout the rest of the present section that all the $\mathcal{G}_\sigma$ are non-empty.

Remark 3 *Figure III.10 shows two examples of groups respectively belonging and not belonging to $\mathcal{G}$ (in the case where Λ is constant). Indeed, since $\Lambda_{K_{G_2}} = \Lambda_L$, the terms $\Lambda_{K_{G_2}}\mathbf{n}_{K_{G_2},\sigma_i}$ and $\Lambda_L\mathbf{n}_{L,\sigma_i}$ cancel out each other in each line of $\mathcal{A}_{G_2}$ and, since the cell L on the other side of σ_1 and σ_2 is the same, both lines of $\mathcal{A}_{G_2}$ are colinear to $x_L - x_{K_{G_2}}$ (this matrix is therefore singular). The non-convexity of cells can be a cause to the singularity of some $\mathcal{A}_G$ (but this does not block the use of the G method since, even in this case, the non-emptiness of all $\mathcal{G}_\sigma$ often holds).*

Finally, we define in the following two lemmata the space playing the role of $\mathfrak{D}$ in Hypothesis 2 and state its density, and we establish the consistency on this space of the group gradients (this will give (P2)).

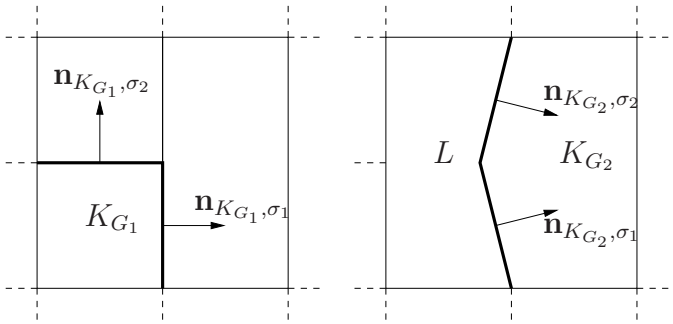
Numerical fluxes

We choose $\{\theta_\sigma^G\}_{\sigma\in\mathcal{E},G\in\mathcal{G}_\sigma}$ a set of weights such that

$$\text{For all } \sigma \in \mathcal{E}, \text{ for all } G \in \mathcal{G}_\sigma,\ 0 \le \theta_\sigma^G \le 1 \text{ and, for all } \sigma \in \mathcal{E},\ \sum_{G\in\mathcal{G}_\sigma} \theta_\sigma^G = 1. \quad \text{[III.31]}$$

The numerical fluxes are then defined as follows: For all $K \in \mathcal{T}$, for all $\sigma \in \mathcal{E}_K$,

$$F_{K,\sigma}(u) \overset{\rm def}{=} \sum_{G\in\mathcal{G}_\sigma} \theta_\sigma^G F_{K,\sigma}^G(u), \quad F_{K,\sigma}^G(u) \overset{\rm def}{=} {\rm m}\sigma\Lambda_K(\nabla_{\mathcal{D}}u)_K^{G,\sigma}\cdot\mathbf{n}_{K,\sigma}. \quad \text{[III.32]}$$



(a) $G_1 \in \mathcal{G}_\sigma$ (b) $G_2 \notin \mathcal{G}_\sigma$

Figure III.10: Face groups of $\widetilde{\mathcal{G}}$ respectively belonging and not belonging to $\mathcal{G}_\sigma$.

Remark 4 *Notice that the subfluxes $F^G_{K,\sigma}$ are conservative (second equation in* [III.27]*), and thus the whole fluxes $F_{K,\sigma}$ themselves are also conservative.*

Specific methods are obtained from [III.32] by defining a suitable criterion to compute the family of weights $\{\theta^G_\sigma\}_{\sigma\in\mathcal{E},G\in\mathcal{G}_\sigma}$.

Example 1 (MPFA L method) *The MPFA L method can be obtained as follows: For all $\sigma \in \mathcal{E}$, let $\widetilde{G} \in \mathcal{G}_\sigma$ be the group satisfying the criterion proposed in [7] and set $\theta^{\widetilde{G}}_\sigma = 1/\mathrm{card}(\{\mathrm{s} \in \mathcal{V}, \mathrm{s} \in \sigma\})$ and $\theta^G_\sigma = 0$ for $\widetilde{G} \neq G \in \mathcal{G}_\sigma$.*

Example 2 *The alternative choice used in the numerical examples of § V.3.2 is designed so as to enhance the coercivity of the method. For each group $G \in \mathcal{G}$, define the space $\mathcal{H}_{\mathcal{T}_G} \stackrel{\text{def}}{=} \{u_K \in \mathbb{R},\ K \in \mathcal{T}_G\}$ endowed with the semi-norm*

$$\|u\|^2_{\mathcal{T}_G} \stackrel{\text{def}}{=} \sum_{K\in\mathcal{T}_G}\sum_{\sigma\in\mathcal{E}_K\cap G} \frac{\mathrm{m}\sigma}{d_{K,\sigma}}(\gamma_\sigma u - u_K)^2.$$

For all $u,\ v \in \mathcal{H}_{\mathcal{T}_G}$ set $a_{\mathcal{T}_G}(u,v) = \sum_{K\in\mathcal{T}_G}\sum_{\sigma\in\mathcal{E}_K\cap G} F^G_{K,\sigma}(u)(\gamma_\sigma v - v_K)$. For each $G \in \mathcal{G}$ define

$$\gamma_2 \stackrel{\text{def}}{=} \inf_{\{u\in\mathcal{H}_{\mathcal{T}_G},\ \|u\|_{\mathcal{T}_G}=1\}} a_{\mathcal{T}_G}(u,u)$$

The computation of the parameter γ_2 requires to evaluate the eigenvalues of a local matrix of $\mathbb{R}^{d\times d}$ associated with the bilinear form $a_{\mathcal{T}_G}$, and its cost is negligible. The computation of the parameter γ_2 is given in Section Appendix.

III.2.2 Coercivity of the scheme

Property (P1) is only conditionally verified by non symmetric methods. We propose a computable criterion issued from the stronger assumption that coercivity holds

locally around each vertex $s \in \mathcal{V}$. For a given discretization and diffusion tensor, this assumption can be checked numerically by computing the eigenvalues of a small linear system of size card$\mathcal{E}_s \leq \varrho_3$ for each vertex $s \in \mathcal{V}$.

Let $s \in \mathcal{V}$, and set $\mathcal{H}_{\mathcal{T}_s} \stackrel{\text{def}}{=} \{u_K \in \mathbb{R},\ K \in \mathcal{T}_s\}$. The space $\mathcal{H}_{\mathcal{T}_s}$ is endowed with the semi-norm

$$\|u\|_{\mathcal{T}_s}^2 \stackrel{\text{def}}{=} \sum_{K \in \mathcal{T}_s} \sum_{\sigma \in \mathcal{E}_s \cap \mathcal{E}_K} \frac{\mathrm{m}\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)^2.$$

Denote by $a_{\mathcal{T}_s}$ the bilinear form defined as follows: For all $u,\ v \in \mathcal{H}_{\mathcal{T}_s}$,

$$a_{\mathcal{T}_s}(u,v) \stackrel{\text{def}}{=} \sum_{G \in \mathcal{G}, G \subset \mathcal{E}_s} \sum_{K \in \mathcal{T}_G} \sum_{\sigma \in \mathcal{E}_K \cap G} \theta_\sigma^G F_{K,\sigma}^G(u)(\gamma_\sigma v - v_K).$$

Lemma 12 *Let there be a positive constant γ_3 such that*

$$\min_{s \in \mathcal{V}} \inf_{\{v \in \mathcal{H}_{\mathcal{T}_s} \,|\, \|v\|_{\mathcal{T}_s} = 1\}} a_{\mathcal{T}_s}(v,v) \geq \gamma_3. \qquad \text{[III.33]}$$

Then, for all $u \in \mathcal{H}_{\mathcal{T}}$, $a_{\mathcal{T}}(u,u) \geq \gamma_3 \|u\|_{\mathcal{T}}^2$.

Proof. For all $u \in \mathcal{H}_{\mathcal{T}}$ and $s \in \mathcal{V}$, let $u_s \stackrel{\text{def}}{=} (u_K)_{K \in \mathcal{T}_s} \in \mathcal{H}_{\mathcal{T}_s}$. Since any given group G only belongs to one particular $\mathcal{E}_s$, it is easy to see that $a_{\mathcal{T}}(u,u) = \sum_{s \in \mathcal{V}} a_{\mathcal{T}_s}(u_s,u_s)$, and thus that $a_{\mathcal{T}}(u,u) \geq \gamma_3 \sum_{s \in \mathcal{V}} ||u_s||_{\mathcal{T}_s}^2$. The assert then follows from

$$\sum_{s \in \mathcal{V}} \|u_s\|_{\mathcal{T}_s}^2 \geq \|u\|_{\mathcal{T}}^2,$$

which is straightforward since, for all $K \in \mathcal{T}$ and for all $\sigma \in \mathcal{E}_K$, card$(\{s \in \mathcal{V} \,|\, \sigma \in \mathcal{E}_s\}) \geq 1$.
□

III.2.3 Consistency of the scheme

Hypothesis 6 *Let $\mathcal{D}$ be an admissible discretization in the sense of II.2.1, we furthermore suppose that there exists $\gamma_4 < +\infty$ such that*

$$\forall \sigma \in \mathcal{E}\,,\ \sum_{G \in \mathcal{G}_\sigma} \theta_\sigma^G |\mathcal{A}_G^{-1}| \leq \gamma_4. \qquad \text{[III.34]}$$

Lemma 13 *[Consistency of the group gradients] For all $\varphi \in \mathcal{Q}$, there exists a real $C_5 > 0$ which only depends on ϱ_1, ϱ_2, Λ and φ such that, for all $G \in \mathcal{G}$, all $\sigma \in G$ and all $K \in \mathcal{T}_\sigma$,*

$$|(\nabla_{\mathcal{D}} \varphi_{\mathcal{T}})_K^{G,\sigma} - \nabla\varphi(x_K)| \leq C_5\ (1 + |\mathcal{A}_G^{-1}|) \max_{K \in \mathcal{T}_G} \mathrm{diam} K.$$

PROOF. See Appendix B.2.
□

Lemma 14 *[consistency of the fluxes] Let φ be a given function in $\mathcal{Q}$. Then, there exists a real $C > 0$ depending only on φ, ϱ_1, γ_4, β_0 and d, such that for all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$,*

$$\left| F_{K,\sigma}(P_{\mathcal{D}}\varphi) - \frac{\mathrm{m}_\sigma}{\mathrm{m}_K} \int_K \Lambda(x)\nabla\varphi(x)dx \right| \leq C\ \mathrm{m}_\sigma \mathrm{diam}(K)$$

PROOF. For all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, [III.32], [III.31], Lemma 13 and [III.34] yield

$$\begin{aligned}
\left| F_{K,\sigma}(\varphi_{\mathcal{T}}) - \frac{1}{\mathrm{m}K} \int_K \Lambda(x)\nabla\varphi(x)\cdot\mathrm{m}\sigma\mathbf{n}_{K,\sigma} \right| &\leq |F_{K,\sigma}(\varphi_{\mathcal{T}}) - \Lambda_K\nabla\varphi(x_K)\mathrm{m}\sigma\mathbf{n}_{K,\sigma}| \\
&\quad + \left| \frac{1}{\mathrm{m}K} \int_K \Lambda(x)(\nabla\varphi(x) - \nabla\varphi(x_K))\cdot\mathrm{m}\sigma\mathbf{n}_{K,\sigma} \right| \\
&\leq \mathrm{m}_\sigma\beta_0 \sum_{G\in\mathcal{G}_\sigma} \theta_\sigma^G \left| (\nabla_{\mathcal{D}}\varphi_{\mathcal{T}_n})_K^{G,\sigma} - \nabla\varphi(x_K) \right| \\
&\quad + \mathrm{m}_\sigma\beta_0 \sup_{x\in K} |\nabla\varphi(x) - \nabla\varphi(x_K)| \\
&\leq (C_5\gamma_4 + C_3)\mathrm{m}_\sigma\beta_0 h_{\mathcal{D}},
\end{aligned}$$

where $C_3 = \sup_{x\in P_\Omega} |\varphi''(x)|$.
□

III.3 The Cell-Gradient method

Finite volume methods for diffusive problems on general polyhedral meshes have known an impetuous development in recent years. Multi-point schemes have been introduced in the middle of the 90s (see, e.g., [4,36]) to circumvent the mesh limitations of the classical two-points method. The key idea is to relax mesh requirements at the expense of a larger stencil while preserving second order convergence. A simpler, cell based gradient reconstruction yielding a convergent method has been proposed in [12, §3.2]. The resulting scheme, henceforth referred to as CG method, has a more straightforward implementation but a slightly larger stencil.

III.3.1 The Finite Volume scheme

Hypothesis 7 *We further assume that*

- *for all $K \in \mathcal{T}$, $\mathcal{E}_{\mathrm{int},K} = \emptyset$,*
- *$\mathcal{F} = \emptyset$,*

- *For all $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$, there exists a group of d faces $G_\sigma^K \subset \mathcal{E}_K$ sharing a same vertex with σ such that $\boldsymbol{A}_\sigma^K$ is invertible (cf definition below).*

For all $v \in \mathbb{H}_\mathcal{T}(\Omega)$, defined from the cell values $\{v_K\}_{K\in\mathcal{T}}$, we use the discrete gradient $(\nabla_\mathcal{D} v)_K^{G_\sigma^K}$ introduced in (III.2.1), which can be expressed thanks to Lemma 11, in terms of the cell values $\{v_K\}_{K\in\mathcal{T}}$ by solving the local linear system $\mathbf{A}_\sigma^K (\nabla_\mathcal{D} v)_K^{G_\sigma^K} = \mathbf{b}_\sigma^K(v)$ such that $\mathbf{A}_\sigma^K \in \mathbb{R}^{d\times d}$, $\mathbf{b}_\sigma^K(v) \in \mathbb{R}^d$,

$$\mathbf{A}_\sigma^K \overset{\text{def}}{=} \begin{bmatrix} \left\{ \dfrac{\Lambda_{L_{\sigma'}} \mathbf{n}_{L_{\sigma'},\sigma'} \cdot \mathbf{n}_{L_{\sigma'},\sigma'}}{d_{L,\sigma'}} (x_{L_{\sigma'}} - x_{K_\sigma}) + \Lambda_{K_\sigma} \mathbf{n}_{K_\sigma,\sigma'} + \Lambda_{L_{\sigma'}} \mathbf{n}_{L_{\sigma'},\sigma'} \right\}_{\sigma' \in G_\sigma^K \cap \mathcal{E}_{\text{int}}} \\ \left\{ \dfrac{\Lambda_{K_\sigma} \mathbf{n}_{K_\sigma,\sigma'} \cdot \mathbf{n}_{K_\sigma,\sigma'}}{d_{K_\sigma,\sigma'}} (x_\sigma - x_{K_\sigma}) \right\}_{\sigma' \in G_\sigma^K \cap \mathcal{E}_{\text{ext}}} \end{bmatrix}$$

and

$$\mathbf{b}_\sigma^K(v) \overset{\text{def}}{=} \begin{bmatrix} \left\{ \dfrac{\Lambda_{L_{\sigma'}} \mathbf{n}_{L_{\sigma'},\sigma'} \cdot \mathbf{n}_{L_{\sigma'},\sigma'}}{d_{L,\sigma'}} (v_{L_{\sigma'}} - v_{K_\sigma}) \right\}_{\sigma' \in G_\sigma^K \cap \mathcal{E}_{\text{int}}} \\ \left\{ \dfrac{\Lambda_{K_\sigma} \mathbf{n}_{K_\sigma,\sigma'} \cdot \mathbf{n}_{K_\sigma,\sigma'}}{d_{K_\sigma,\sigma'}} (-v_{K_\sigma}) \right\}_{\sigma' \in G_\sigma^K \cap \mathcal{E}_{\text{ext}}} \end{bmatrix}$$

where, for all $\sigma' \in G_\sigma^K \cap \mathcal{E}_{\text{int}}$, $L_{\sigma'}$ is such that $\mathcal{T}_{\sigma'} = \{K, L_{\sigma'}\}$. The local gradient $(\nabla_\mathcal{D} v)_K^{G_\sigma^K}$ allows to define the values $\Pi_\sigma^K v$, for all $v \in \mathbb{H}_\mathcal{T}(\Omega)$, $\sigma \in \mathcal{E}_K$, $K \in \mathcal{T}$:

$$\Pi_\sigma^K v \overset{\text{def}}{=} v_K + (\nabla_\mathcal{D} v)_K^{G_\sigma^K} \cdot (x_\sigma - x_K). \qquad \text{[III.35]}$$

We introduce the discrete gradient reconstruction $\overline{\nabla}_\mathcal{D} \in \mathcal{L}(\mathbb{H}_\mathcal{T}(\Omega); [\mathbb{H}_\mathcal{T}(\Omega)]^d)$ s.t., for all $K \in \mathcal{T}$ and all $v \in \mathbb{H}_\mathcal{T}(\Omega)$,

$$(\overline{\nabla}_\mathcal{D} v)_K \overset{\text{def}}{=} \overline{\nabla}_\mathcal{D} v_{|K} = \frac{1}{\mathrm{m}_K} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma (\Pi_\sigma^K v - v_K) \mathbf{n}_{K,\sigma}. \qquad \text{[III.36]}$$

Thanks to Lemma 13, it can be shown to be consistent for functions belonging to the space of test-functions, $\mathcal{Q}$.

The discrete bilinear form reads, for all u, $v \in \mathbb{H}_\mathcal{T}(\Omega)$,

$$a_\mathcal{D}(u,v) \overset{\text{def}}{=} \sum_{K\in\mathcal{T}} \mathrm{m}_K \Lambda_K \overline{\nabla}_\mathcal{D} u \cdot \widetilde{\nabla}_\mathcal{D} v + \sum_{K\in\mathcal{T}} \alpha_K \sum_{\sigma\in\mathcal{E}_K} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} R_{K,\sigma}(u) R_{K,\sigma}(v), \qquad \text{[III.37]}$$

where we have set, for all $v \in \mathbb{H}_\mathcal{T}(\Omega)$, $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$, $R_{K,\sigma}(v) \overset{\text{def}}{=} \Pi_\sigma^K v - v_K - (\overline{\nabla}_\mathcal{D} v)_K \cdot (x_\sigma - x_K)$. The discrete problem reads

$$\text{Find } u \in \mathbb{H}_\mathcal{T}(\Omega) \text{ s.t. } a_\mathcal{D}(u,v) = \int_K f v \,\mathrm{d}x \text{ for all } v \in \mathbb{H}_\mathcal{T}(\Omega). \qquad \text{[III.38]}$$

It can be shown from (III.37) that the discrete variational formulation (III.38) is equivalent to a finite volume scheme.
The consistency of the scheme in the sense that for all $\varphi \in \mathcal{Q}$,

$$\epsilon_{\mathcal{D}}(\varphi) \to 0 \text{ as } h_{\mathcal{D}} \to 0, \qquad \text{[III.39]}$$

derives from the consistency of the discrete gradient $\overline{\nabla}_{\mathcal{D}}$. The consistency of the gradient $\overline{\nabla}_{\mathcal{D}}$ is obtained thanks to Lemma 13 and the following property: For all $w \in \mathbb{R}^d$,

$$\sum_{\sigma \in \mathcal{E}_K} \mathrm{m}_\sigma w \cdot (x_\sigma - x_K)\mathbf{n}_{K,\sigma} = \mathrm{m}_K w \qquad \text{[III.40]}$$

More details about the coercivity and the consistency of the scheme can be found in [12, §3.2]

Chapitre IV

Une famille de schémas volumes finis symétriques

Notre recherche est motivée en particulier par la simulation numérique de fluides complexes en milieux poreux incluant des couplages avec des phénomènes thermodynamiques et ou chimiques. Comme nous avons pu le constater dans le chapitre II, la convergence d'un schéma dépend d'une condition de coercivité. Le chapitre III a montré que dans le cas de schémas non symétriques tels que le schéma MPFA O et G, la condition de coercivité ne peut être vérifiée que numériquement. Bien que de nombreux schémas aient été proposés (cf [52], chapter III), il n'existe pas encore de schéma ultime, c'est à dire centré aux mailles avec un stencil compact, qui respecterait les bornes physiques et donnerait de bonnes approximations même sur des maillages déformés non conformes avec de forts contrastes dans les perméabilités (ou coefficient de diffusions). La nouvelle famille de schémas que nous introduisons, a été conçue dans le cadre de cette recherche, et a les caractéristiques suivantes :

1. elle peut être utilisée sur n'importe quel maillage polygonal avec tenseur de diffusion hétérogène, anisotrope,

2. elle fournit la solution exacte dans le cas où $f = 0$, Λ est constante par morceaux dans des sous domaines polygonaux, polyédriques et u affine dans chacun de ces sous domaines

3. dans certains cas particuliers, elle donne le classique schéma à deux points,

4. elle est symétrique et coercive par rapport à une norme discrète adéquate, et la preuve de la convergence en découle.

A partir de cette famille, on a créé un schéma dit Dioptre qui permet de répondre aux caractéristiques recherchées avec en plus un stencil compact dans le cas de

maillages pas trop déformés (en un sens incluant le tenseur de diffusion), par exemple dans le cas de maillages quadrilatéraux pas trop déformés, il mène à un schéma neuf points. Ce dernier a ses inconnues aux centres des mailles, répond à notre recherche en combinant les avantages du schéma multi-points MPFA O et les schémas hybrides, respecte les hétérogénéités de la matrice de diffusion, et présente des propriétés de coercivité et de convergence. Un paramètre ajustable permet d'augmenter le poids des flux à deux points dans le schéma, ce qui peut permettre d'améliorer la monotonie du schéma. Le schéma est basé sur l'utilisation de points à l'interface entre milieux de matrices de diffusion différentes, où la formule de la moyenne harmonique est utilisable.

IV.1 The Finite Volume Scheme

Hypothesis 8 *Let $\mathcal{D}$ be a discretization matching Definition II.2.1. We further suppose that*

(H1) $\{\sigma \in \mathcal{E}_K,\ K \in \mathcal{T}\} \subset \mathcal{F}$.

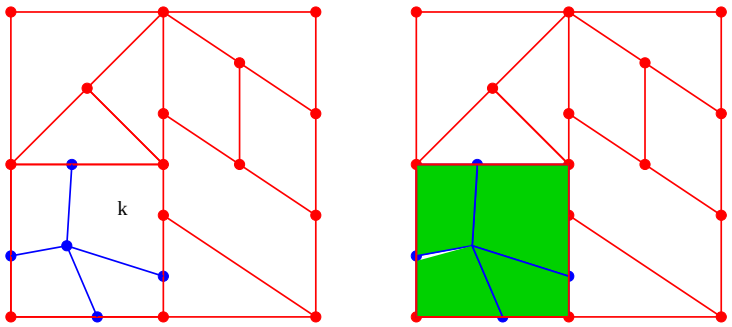
(H2) for all $K \in \mathcal{T}$, there exists $\mathcal{S}_K$, a finite family of non empty connex open disjoint subset of K (the sub-cells of K) such that $\cup_{\kappa\in\mathcal{S}_K}\overline{\kappa} = \overline{K}$ and for all $\kappa \in \mathcal{S}_K$, there exists $\mathcal{E}_\kappa$ a subset of $\mathcal{E}_{\mathrm{int},K} \cup \mathcal{E}_K$ such that $\partial\kappa = \cup_{\sigma\in\mathcal{E}_\kappa}\overline{\sigma}$ (see figure IV.1).

(H3) For all $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$, there exists $x_\kappa \in \kappa$ such that κ is strictly star-shaped with respect to x_κ.

(H4) We assume that there exists a real $C > 0$ independent of the discretization $\mathcal{D}$ such that for all $\sigma \in \overline{\mathcal{F}}$, for all $\varphi \in \mathcal{Q}$, $|\Pi_\sigma(P_\mathcal{F}\varphi) - \varphi(x_\sigma)| \le C\ h_\mathcal{D}^2$, we assume also that for all $K \in \mathcal{T}$, $\sigma \in \overline{\mathcal{F}}$, $\sigma \subset K$, there exists a family of real $(\pi_{\sigma,L})_{L\in\mathcal{N}_K\cup\{K\}} \cup (\pi_{\sigma,\sigma'})_{\sigma'\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}}$ such that $\displaystyle\sum_{L\in\mathcal{N}_K\cup\{K\}} \pi_{\sigma,L} + \sum_{\sigma'\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}} \pi_{\sigma,\sigma'} = 1$ and for all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, $\Pi_\sigma v = \displaystyle\sum_{L\in\mathcal{N}_K\cup\{K\}} \pi_{\sigma,L}\ v_L$.

Additional notations : Under Hypothesis 8, for all $\sigma \in \mathcal{E}_\kappa$, $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$, we denote $\mathbf{n}_{\kappa,\sigma}$ the unit vector normal to σ outward to κ and $d_{\kappa,\sigma}$ the Euclidean distance from x_κ to σ. For all $\sigma \in \mathcal{E}_{\mathrm{int},K}\cup\mathcal{E}_K$, we denote by $\mathcal{S}_{K,\sigma} = \{l \in \mathcal{S}_K, \sigma \in \mathcal{E}_l\}$, then either $\mathcal{S}_{K,\sigma}$ has exactly one element and then $\sigma \in \mathcal{E}_K$ or $\mathcal{S}_{K,\sigma}$ has exactly two elements and then $\sigma \in \mathcal{E}_{\mathrm{int},K}$. For all $K \in \mathcal{T}$, for all $\sigma \in \mathcal{E}_K$ with $S_{K,\sigma} = \{\kappa\}$, we set $d_{K,\sigma} = d_{\kappa,\sigma}$ and for all $\sigma \in \mathcal{E}_{\mathrm{int},K}$ with $S_{K,\sigma} = \{\kappa, l\}$, we set $d_{K,\sigma} = \min(d_{\kappa,\sigma}, d_{l,\sigma})$.

We define the finite volume scheme which is based on a variational formulation on the space $\mathcal{H}_{\mathcal{D},\mathcal{F}}$ using the discrete gradient for each sub-cell $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$ defined by



(a) $\mathcal{E}_{\text{int},K}$ (b) $\mathcal{S}_K$

Figure IV.1: $\mathcal{S}_K$ are the subcells of K built from $\mathcal{E}_{\text{int},K}$.

Let us now define the bilinear form $a_\mathcal{D}$ on $\mathcal{H}_{\mathcal{D},\mathcal{F}} \times \mathcal{H}_{\mathcal{D},\mathcal{F}}$ by

$$a_\mathcal{D}(u,v) = \sum_{K\in\mathcal{T}} \sum_{\kappa\in\mathcal{S}_K} \left(\mathrm{m}_\kappa (\nabla_\mathcal{D} u)_\kappa \cdot \Lambda_K (\nabla_\mathcal{D} v)_\kappa + \ \alpha_\kappa \sum_{\sigma\in\mathcal{E}_\kappa} \frac{\mathrm{m}_\sigma}{d_{\kappa,\sigma}} R_{\kappa,\sigma}(u) R_{\kappa,\sigma}(v) \right) \qquad \text{[IV.1]}$$

for all $(u,v) \in \mathcal{H}_{\mathcal{D},\mathcal{F}} \times \mathcal{H}_{\mathcal{D},\mathcal{F}}$, with

$$\Lambda_K = \frac{1}{\mathrm{m}_K} \int_K \Lambda(x)dx,$$

for all $K \in \mathcal{T}$. In (IV.1), the residual functions $R_{\kappa,\sigma}$ are defined for all $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, for all $\sigma \in \mathcal{E}_\kappa$, $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$, by

$$R_{\kappa,\sigma}(u) = u_\sigma - u_K - (\nabla_\mathcal{D} u)_\kappa \cdot (x_\sigma - x_K), \qquad \text{[IV.2]}$$

where the parameters α_κ are real such that

$$0 < \mu_0 \leq \alpha_\kappa \leq \gamma_0 \qquad \text{[IV.3]}$$

The subsequent analysis will readily extends to more general framework but we keep to the scalar term for the sake of simplicity in the notations.

The discretization of (II.5) on $\mathcal{D}$ is defined by the following discrete variational formulation: find $u_\mathcal{D} \in \mathcal{H}_{\mathcal{D},\mathcal{F},\mathcal{K}}$ such that

$$a_\mathcal{D}(u_\mathcal{D}, v) = \int_\Omega f(x) P_\mathcal{T} v(x) dx \quad \text{for all } v \in \mathcal{H}_{\mathcal{D},\mathcal{F},\mathcal{K}}. \qquad \text{[IV.4]}$$

Let us show that the variational formulation (IV.4) is equivalent to a finite volume scheme.

Proof.

Firstly, for all $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, we seek a family $(G_{\kappa,\sigma}(u))_{\sigma\in\mathcal{E}_\kappa,\kappa\in\mathcal{S}_K,K\in\mathcal{T}}$ such that,

$$\sum_{K\in\mathcal{T}}\sum_{\kappa\in\mathcal{S}_K}\sum_{\sigma\in\mathcal{E}_\kappa} G_{\kappa,\sigma}(u)(v_\sigma - v_K) = a_{\mathcal{D}}(u,v) \qquad \text{[IV.5]}$$

for all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$. From the definitions of the bilinear form $a_{\mathcal{D}}$ (IV.1), we can deduce,

$$\begin{aligned} G_{\kappa,\sigma}(u) &= \Lambda_K(\nabla_{\mathcal{D}}u)_\kappa \cdot \mathrm{m}_\sigma \mathbf{n}_{\kappa,\sigma} + \alpha_\kappa \frac{\mathrm{m}_\sigma}{d_{\kappa,\sigma}} R_{\kappa,\sigma}(u) \\ &\quad - \frac{1}{\mathrm{m}_\kappa}\mathrm{m}_\sigma \mathbf{n}_{\kappa,\sigma} \cdot \left(\sum_{\sigma'\in\mathcal{E}_\kappa} \alpha_\kappa \frac{\mathrm{m}_{\sigma'}}{d_{\kappa,\sigma'}} R_{\kappa,\sigma'}(u)(x_{\sigma'} - x_K) \right). \end{aligned} \qquad \text{[IV.6]}$$

We can re-write Equation (IV.5) as follows,

$$\begin{aligned} a_{\mathcal{D}}(u,v) &= \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K, \mathcal{S}_{K,\sigma}=\{\kappa\}} G_{\kappa,\sigma}(u)(v_\sigma - v_K) \\ &\quad + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}, \mathcal{S}_{K,\sigma}=\{\kappa,l\}} (G_{\kappa,\sigma}(u) + G_{l,\sigma}(u))(v_\sigma - v_K). \end{aligned} \qquad \text{[IV.7]}$$

For all $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, we introduce the family $(G_{K,\sigma}(u))_{\sigma\in\mathcal{E}_K\cup\mathcal{E}_{\mathrm{int},K},K\in\mathcal{T}}$ defined for all $K \in \mathcal{T}$ by,

$$\begin{aligned} &G_{K,\sigma}(u) = G_{\kappa,\sigma}(u) \text{ for all } \sigma \in \mathcal{E}_K \text{ with } \mathcal{S}_{K,\sigma} = \{\kappa\} \\ &G_{K,\sigma}(u) = G_{\kappa,\sigma}(u) + G_{l,\sigma}(u) \text{ for all } \sigma \in \mathcal{E}_{\mathrm{int},K} \text{ with } \mathcal{S}_{K,\sigma} = \{\kappa,l\}. \end{aligned} \qquad \text{[IV.8]}$$

Then, we get from (IV.7),

$$a_{\mathcal{D}}(u,v) = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K\cup\mathcal{E}_{\mathrm{int},K}} G_{K,\sigma}(u)(v_\sigma - v_K). \qquad \text{[IV.9]}$$

by splitting the set $\mathcal{E}_{\mathrm{int},K}$ into $\mathcal{E}_{\mathrm{int},K} \cap \mathcal{F}$ and $\mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}$, we deduce,

$$a_{\mathcal{D}}(u,v) = \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K\cup(\mathcal{E}_{\mathrm{int},K}\cap\mathcal{F})} G_{\kappa,\sigma}(u)(v_\sigma - v_K) + \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}\cap\overline{\mathcal{F}}} G_{K,\sigma}(u)(\Pi_\sigma v - v_K).$$

Using hypothesis 8, we get that for all $\sigma \in \mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}$, $K \in \mathcal{T}$, $\Pi_\sigma v - v_K = \sum_{L\in\mathcal{N}_K} \pi_{\sigma,L}(v_L - v_K) - \sum_{\sigma'\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}} \pi_{\sigma,\sigma'} v_K$, then we deduce,

$$\begin{aligned} a_{\mathcal{D}}(u,v) &= \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K\cup(\mathcal{E}_{\mathrm{int},K}\cap\mathcal{F})} G_{K,\sigma}(u)(v_\sigma - v_K) \\ &\quad + \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} \left(\sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}\cap\overline{\mathcal{F}}} \pi_{\sigma,L} G_{K,\sigma}(u) \right)(v_L - v_K) \\ &\quad - \sum_{K\in\mathcal{T}} v_K \sum_{\sigma'\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}} \left(\sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}\cap\overline{\mathcal{F}}} \pi_{\sigma,\sigma'} G_{K,\sigma}(u) \right). \end{aligned} \qquad \text{[IV.10]}$$

For all $L \in \mathcal{N}_K$, $K \in \mathcal{T}$, we set,

$$\widetilde{G}_{K,L}(u) = \sum_{\sigma \in \mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}} \pi_{\sigma,L} G_{K,\sigma}(u),$$

and for all $\sigma' \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}$, $K \in \mathcal{T}$, we set,

$$\widetilde{G}_{K,\sigma'}(u) = \sum_{\sigma \in \mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}} \pi_{\sigma,\sigma'} G_{K,\sigma}(u)$$

then we get,

$$\begin{aligned} \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} \left(\sum_{\sigma \in \mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}} \pi_{\sigma,L} G_{K,\sigma}(u) \right) (v_L - v_K) &= \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} \widetilde{G}_{K,L}(u)(v_L - v_K) \\ &= \sum_{K\in\mathcal{T}} \sum_{L\in\mathcal{N}_K} (\widetilde{G}_{L,K}(u) - \widetilde{G}_{K,L}(u)) v_K. \end{aligned} \quad \text{[IV.11]}$$

and

$$\sum_{K\in\mathcal{T}} v_K \sum_{\sigma' \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}} \left(\sum_{\sigma \in \mathcal{E}_{\mathrm{int},K} \cap \overline{\mathcal{F}}} \pi_{\sigma,\sigma'} G_{K,\sigma}(u) \right) = \sum_{K\in\mathcal{T}} v_K \sum_{\sigma' \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}} \widetilde{G}_{K,\sigma'}(u). \quad \text{[IV.12]}$$

Since for all $\sigma \in \mathcal{E}_{\mathrm{ext}}$, $v_\sigma = 0$, we get also,

$$\sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}(u) v_\sigma = \sum_{\sigma \in \mathcal{E}_{\mathrm{int}}, \mathcal{T}_\sigma = \{K,L\}} (G_{K,\sigma}(u) + G_{L,\sigma}(u)) v_\sigma \quad \text{[IV.13]}$$

Using (IV.10)-(IV.13), we obtain that for all $(u, v) \in \mathcal{H}_{\mathcal{D},\mathcal{F}} \times \mathcal{H}_{\mathcal{D},\mathcal{F}}$

$$\begin{aligned} a_{\mathcal{D}}(u,v) = &- \sum_{K\in\mathcal{T}} v_K \left(\sum_{L\in\mathcal{N}_K} F_{K,L}(u) + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}} F_{K,\sigma}(u) + \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}\cap\mathcal{F}} G_{K,\sigma}(u) \right) \\ &+ \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_{\mathrm{int},K}\cap\mathcal{F}} G_{K,\sigma}(u) v_\sigma \\ &+ \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}, \mathcal{T}_\sigma=\{K,L\}} (G_{K,\sigma}(u) + G_{L,\sigma}(u)) v_\sigma, \end{aligned} \quad \text{[IV.14]}$$

where for all $L \in \mathcal{N}_K$, $K \in \mathcal{T}$,

$$F_{K,L}(u) = \widetilde{G}_{K,L}(u) - \widetilde{G}_{L,K}(u) + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_L} G_{K,\sigma}(u) \quad \text{[IV.15]}$$

and for all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_{\rm ext}$,

$$F_{K,\sigma}(u) = G_{K,\sigma}(u) + \widetilde{G}_{K,\sigma}(u). \qquad \text{[IV.16]}$$

Then, thanks to (IV.14), we deduce that the variational formulation (IV.4) is equivalent to the following finite volume scheme : Find $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$ such that for all $K \in \mathcal{T}$,

$$\sum_{L\in\mathcal{N}_K} F_{K,L}(u) + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}} F_{K,\sigma}(u) = \int_K f(x)\,dx, \qquad \text{[IV.17]}$$

for all $\sigma \in \mathcal{E}_{{\rm int},K} \cap \mathcal{F}$,

$$G_{K,\sigma}(u) = 0, \qquad \text{[IV.18]}$$

and for all $\sigma \in \mathcal{E}_{\rm int}, \mathcal{T}_\sigma = \{K, L\}$,

$$G_{K,\sigma}(u) + G_{L,\sigma}(u) = 0, \qquad \text{[IV.19]}$$

Notice that thanks to Equation (IV.19), we get that for all $\sigma \in \mathcal{E}_{\rm int}$, with $\mathcal{T}_\sigma = \{K, L\}$, $F_{K,L}(u) + F_{L,K}(u) = 0$, which concludes the proof.
□

IV.2 Coercivity of the scheme

The coercivity of the scheme is obtained under the mesh regularity assumption.

Proposition 4 *[coercivity of the scheme] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1 we have the coercivity of the bilinear form $a_\mathcal{D}$ in the sense that there exists a real $C > 0$ depending only on $\varrho_1, \mu_0, \alpha_0$ and d such that*

$$a_\mathcal{D}(u, u) \geq C\|u\|^2_\mathcal{D}, \qquad \text{[IV.20]}$$

for all $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$.

Proof. From definition (IV.1) of the bilinear form $a_\mathcal{D}$, we have for all $u \in \mathcal{H}_\mathcal{D}$,

$$a_\mathcal{D}(u, u) = \sum_{K\in\mathcal{T}} \sum_{\kappa\in\mathcal{S}_K} \left({\rm m}_\kappa (\nabla_\mathcal{D} u)_\kappa \cdot \Lambda_K (\nabla_\mathcal{D} u)_\kappa + \sum_{\sigma\in\mathcal{E}_\kappa} \alpha_\kappa \frac{{\rm m}_\sigma}{d_{\kappa,\sigma}} R_{\kappa,\sigma}(u)^2 \right) \qquad \text{[IV.21]}$$

Using the following inequality

$$\mu(a-b)^2 \geq \frac{1}{2}\min(\mu,\lambda)a^2 - \lambda b^2, \ \text{ for all } (a, b, \mu, \lambda) \in (\mathbb{R}_+)^4,$$

with $\mu = \alpha_\kappa,\ a = u_\sigma - u_K,\ b = (\nabla_{\mathcal{D}} u)_\kappa \cdot (x_\sigma - x_K)$ and $\lambda = \rho_\kappa$ a positive real, for any $\sigma \in \mathcal{E}_\kappa$, $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$. Then, we obtain for all $\rho_\kappa \geq 0$, $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$, the lower bound,

$$\sum_{\sigma \in \mathcal{E}_\kappa} \alpha_\kappa \frac{\mathrm{m}_\sigma}{d_{\kappa,\sigma}} R_{\kappa,\sigma}(u)^2 \ \geq \frac{1}{2} \min(\rho_\kappa, \alpha_\kappa) \sum_{\sigma \in \mathcal{E}_\kappa} \frac{\mathrm{m}_\sigma}{d_{\kappa,\sigma}} (u_\sigma - u_K)^2 - \rho_\kappa (\nabla_{\mathcal{D}} u)_\kappa \cdot A_\kappa (\nabla_{\mathcal{D}} u)_\kappa, \tag{IV.22}$$

where the square matrix A_κ is defined by

$$A_\kappa = \sum_{\sigma \in \mathcal{E}_\kappa} \frac{\mathrm{m}_\sigma}{d_{\kappa,\sigma}} (x_\sigma - x_K)(x_\sigma - x_K)^t, \tag{IV.23}$$

and satisfies the bound, thanks to Hypothesis (II.7),

$$|A_\kappa| \leq \frac{d \mathrm{m}_\kappa}{\varrho_1^2}. \tag{IV.24}$$

Let us choose ρ_κ such that,

$$\rho_\kappa = \sup \left\{ \rho \in \mathbb{R},\ \mathrm{m}_\kappa\ \Lambda_K - \rho A_\kappa \geq 0 \right\}. \tag{IV.25}$$

Using the upper bound (IV.24), we can prove that ρ_κ defined by (IV.25) satisfies the lower bound

$$\rho_\kappa \geq \frac{\varrho_1^2\ \alpha_0}{d}, \tag{IV.26}$$

for all $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$. Using (IV.21), (IV.22), (IV.25), (IV.26) we obtain the lower bound

$$a_{\mathcal{D}}(u,u) \geq \frac{1}{2} \min \left(\mu_0, \frac{\varrho_1^2\ \alpha_0}{d} \right)\ \|u\|_{\mathcal{D}}^2, \tag{IV.27}$$

for all $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, which concludes the proof.
□

IV.3 Consistency of the scheme

Lemma 15 *[Consistency of the discrete gradients] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that Hypotheses 8 holds. Let φ be a given function in $\mathcal{Q}$ (definition of $\mathcal{Q}$ in Lemma 3). Then, there exists $C_0 > 0$ depending only on φ, ϱ_1 and d such that for all $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$,*

$$|(\nabla_{\mathcal{D}} P_{\mathcal{F}} \varphi)_\kappa - \nabla \varphi(x_K)| \leq C_0\ \mathrm{diam}(K).$$

PROOF. Let $\sigma \in \mathcal{E}_\kappa$, $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$ and $\varphi \in \mathcal{Q}$ be given. For all $\sigma \in \mathcal{E}_\kappa \cap \mathcal{F}$, let us set $\epsilon_{K,\sigma} = \varphi(x_\sigma) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(x_\sigma - x_K)$ and for all $\sigma \in \mathcal{E}_\kappa \backslash \mathcal{F}$, $\epsilon_{K,\sigma} = \Pi_\sigma(P_\mathcal{F}\varphi) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(x_\sigma - x_K)$. Thanks to the fourth property of Hypotheses 8 and since $\varphi \in \mathcal{Q}$, there exists a real $M_\varphi > 0$ depending only on φ such that,

$$\begin{array}{l} |\epsilon_{K,\sigma}| \leq M_\varphi |x_\sigma - x_K|^2 \\ |\nabla\varphi(x_K) - \nabla\varphi(x_\kappa)| \leq M_\varphi \mathrm{diam}(K) \end{array} \qquad \text{[IV.28]}$$

Since, we have the following property, for all $v \in \mathbb{R}^d$, for all $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$,

$$\sum_{\sigma\in\mathcal{E}_\kappa} v\cdot(x_\sigma - x_K)\ \mathrm{m}_\sigma \mathbf{n}_{\kappa,\sigma} = \mathrm{m}_\kappa v. \qquad \text{[IV.29]}$$

Then, thanks to (IV.29) with $v = \nabla\varphi(x_K)$, we get that,

$$(\nabla_\mathcal{D} P_\mathcal{F}\varphi)_\kappa - \nabla\varphi(x_K) = \frac{1}{\mathrm{m}_\kappa}\sum_{\sigma\in\mathcal{E}_\kappa} \epsilon_{K,\sigma}\ \mathrm{m}_\sigma \mathbf{n}_{\kappa,\sigma}.$$

Then, thanks to the first estimate (IV.28), (II.7) and since $\displaystyle\sum_{\sigma\in\mathcal{E}_\kappa} \mathrm{m}_\sigma d_{\kappa,\sigma} = d\mathrm{m}_\kappa$, we obtain the bound,

$$|(\nabla_\mathcal{D} P_\mathcal{F}\varphi)_\kappa - \nabla\varphi(x_K)| \leq M_\varphi\ \frac{d}{\varrho_1}\ \mathrm{diam}(K), \qquad \text{[IV.30]}$$

which concludes the proof.
□

Lemma 16 *[Consistency of the residual functions] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that hypotheses 8 holds. Let φ be a given function in $\mathcal{Q}$. Then, there exists $C_1 > 0$ depending only on φ, ϱ_1 and d such that for all $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$,*

$$R_{\kappa,\sigma}(P_\mathcal{F}\varphi) \leq C_1\ \mathrm{diam}(K)^2.$$

PROOF. For all $\sigma \in \mathcal{E}_\kappa \cap \mathcal{F}$, let us set $\epsilon_{K,\sigma} = \varphi(x_\sigma) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(x_\sigma - x_K)$ and for all $\sigma \in \mathcal{E}_\kappa\backslash\mathcal{F}$, $\epsilon_{K,\sigma} = \Pi_\sigma(P_\mathcal{F}\varphi) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(x_\sigma - x_K)$.

From the definition of the residual function we have

$$R_{\kappa,\sigma}(P_\mathcal{F}\varphi) = \epsilon_{K,\sigma} - ((\nabla_\mathcal{D} P_\mathcal{F}\varphi)_\kappa - \nabla\varphi(x_K))\cdot(x_\sigma - x_K).$$

Thanks to (IV.28) and (IV.30), we deduce that there exists M_φ depending only on φ such that,

$$\begin{array}{rcl} R_{\kappa,\sigma}(P_\mathcal{F}\varphi) & \leq & |\epsilon_{K,\sigma}| + |(\nabla_\mathcal{D} P_\mathcal{F}\varphi)_\kappa - \nabla\varphi(x_K)|\ |x_\sigma - x_K| \\ & \leq & M_\varphi\left(1 + \dfrac{d}{\varrho_1}\right)\mathrm{diam}(K)^2, \end{array}$$

which concludes the proof.
□

Proposition 5 *[Consistency of the flux functions] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1, and let us assume that hypotheses 8 holds. Let φ be a given function in $\mathcal{Q}$. Then, there exists $C_2 > 0$ depending only on φ, ϱ_1, γ_0 and d such that for all $K \in \mathcal{T}$,*

$$\begin{array}{l} \forall \sigma \in \mathcal{E}_K, \quad \left| G_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \dfrac{\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x}{\mathrm{m}_K}\cdot \mathbf{n}_{K,\sigma} \right| \leq C_2 \mathrm{m}_\sigma \mathrm{diam}(K) \\ \forall \sigma \in \mathcal{E}_{\mathrm{int},K}, |G_{K,\sigma}(P_{\mathcal{F}}\varphi)| \leq C_2 \mathrm{m}_\sigma \mathrm{diam}(K) \end{array} \quad \text{[IV.31]}$$

where $G_{K,\sigma}$ is defined in (IV.8)

Proof. For all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_K$ with $\mathcal{S}_{K,\sigma} = \{\kappa\}$, since $\displaystyle\sum_{\sigma' \in \sigma_\kappa} \mathrm{m}_{\sigma'} d_{\kappa,\sigma'} = d\mathrm{m}_\kappa$, then using (II.7), the lemmata (15) and (16), from (IV.6), we deduce that,

$$|G_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \Lambda_K \nabla\varphi(x_K)\cdot \mathbf{n}_{K,\sigma}| \leq \left(\beta_0 C_0 + (d+1)C_1 \frac{\gamma_0}{\varrho_1^2}\right) \mathrm{m}_\sigma \mathrm{diam}(K). \quad \text{[IV.32]}$$

Then, using (IV.32) and the same arguments as in the lemma (10), we get that there exists $C_\varphi > 0$ depending only on φ such that

$$\begin{aligned} &\left| G_{K,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \frac{\int_K \Lambda(x)\nabla\varphi(x)\,\mathrm{d}x}{\mathrm{m}_K}\cdot \mathbf{n}_{K,\sigma} \right| \\ &\qquad \leq \left(\beta_0 C_0 + (d+1)C_1 \frac{\gamma_0}{\varrho_1^2} + C_\varphi \beta_0\right) \mathrm{m}_\sigma \mathrm{diam}(K), \end{aligned}$$

which concludes the first part of the proof. For all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_{\mathrm{int},K}$, with $S_{K,\sigma} = \{\kappa, l\}$, using the same arguments as for (IV.32), from (IV.6), we get,

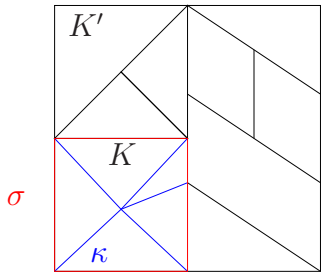
$$\begin{aligned} |G_{K,\sigma}(P_{\mathcal{F}}\varphi)| &= |G_{\kappa,\sigma}(P_{\mathcal{F}}\varphi) + G_{l,\sigma}(P_{\mathcal{F}}\varphi)| \\ &\leq |G_{\kappa,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \Lambda_K \nabla\varphi(x_K)\cdot \mathbf{n}_{\kappa,\sigma}| \\ &+ |G_{l,\sigma}(P_{\mathcal{F}}\varphi) - \mathrm{m}_\sigma \Lambda_K \nabla\varphi(x_K)\cdot \mathbf{n}_{l,\sigma}| \\ &\leq 2\left(\beta_0 C_0 + (d+1)C_1 \frac{\gamma_0}{\varrho_1^2}\right) \mathrm{m}_\sigma \mathrm{diam}(K), \end{aligned}$$

which completes the proof.
□

We will present now, two symmetric finite volume scheme built from the variational formulation IV.4.

IV.4 VFSYM

During our research, we have proposed a new family of finite volume discretization schemes. These are based on the discrete variational formulation framework introduced in IV.1. The use of a sub-grid, $\mathcal{S}_K$, for each cell $K \in \mathcal{T}$ of the mesh enables us to obtain fluxes only between cells sharing an edge as opposed to the cell centered finite volume scheme [43] for which fluxes are also defined between cells sharing only a vertex. The sub-grid, $\mathcal{S}_K$ of each cell K of the mesh is defined by the set of cones obtained joining the face $\sigma \in \mathcal{E}_K$ to the cell center x_K (see figure IV.2). For any $\sigma \in \mathcal{E}_{\text{int},K} \backslash \mathcal{F}$, $K \in \mathcal{T}$, the linear form Π_σ is defined as follows, first, we express x_K in terms of $(x_L)_{L\in\mathcal{N}_K} \cup (x_\sigma)_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\text{ext}}}$ and second, we use the same coefficients obtained, to write, for any $u \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, $\Pi_\sigma(P_{\mathcal{F}}u)$ in terms of $(u_L)_{L\in\mathcal{N}_K}$. The resulting finite volume schemes are cell centered, symmetric and coercive on general polygonal and polyhedral meshes and anisotropic heterogeneous media and can be proved to be convergent even for L^∞ diffusion coefficients under usual shape regularity assumptions. Using L type interpolation from [8], [7] for the definition of the linear forms $(\Pi_\sigma)_{\sigma\in\mathcal{E}_{\text{int},K}\backslash\mathcal{F}}$, $K \in \mathcal{T}$ as it is done in (III.35), enable us to take into account large jumps of the diffusion coefficients. Unfortunately, these properties on difficult anisotropic problems are obtained at the expense of larger scheme and flux stencils. For example, for a cartesian grid in 2D, the stencil by control volume is 21 and in 3D, 81.



(a) $\mathcal{T}_h$

Figure IV.2: $\kappa \in \mathcal{S}_K$ (the sub-cells of the cell K).

IV.5 A symmetric finite volume scheme with compact stencil in $\mathbb{R}^2$

In this section, we propose a cell-centred symmetric scheme which combines the advantages of MPFA (multi point flux approximation) schemes such as L or the O scheme and of hybrid schemes: it may be used on general non conforming meshes, it yields a 9-point stencil on two-dimensional quadrangular meshes, it takes into

account the heterogeneous diffusion matrix, and it is coercive and convergent. The scheme relies on the use of special points, called harmonic averaging points, located at the interfaces of heterogeneity.

Hypothesis 9 *We further assume that,*

- *For all* $(s, s') \in \mathcal{V} \times \mathcal{V}$, $\mathcal{E}_s \cap \mathcal{E}_{s'} = \emptyset$.
- *For all* $K \in \mathcal{T}$, $L \in \mathcal{N}_K$, K *and* L *are separated by a hyperplane.*
- $\cup_{s\in\mathcal{V}} \mathcal{E}_s = \cup_{K\in\mathcal{T}} \mathcal{E}_K$.
- *For all* $K \in \mathcal{T}$, $\{x_K\} = \cap_{\sigma\in\mathcal{E}_{\mathrm{int},K}} \overline{\sigma}$.

IV.5.1 Harmonic averaging points

Consider two domains K and L of $\mathbb{R}^2$ with different diffusion matrices (or permeabilities) Λ_K and Λ_L, separated by a hyperplane interface σ, and let $\boldsymbol{x}_K \in K$ and $\boldsymbol{x}_L \in L$. The harmonic averaging point between $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ is defined as the point of σ where one may compute the flux from $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ by the classical harmonic averaging formula, for a certain class of regular functions. If $\Lambda_K = \Lambda_L$, then this point is the intersection of the line $\boldsymbol{x}_K\boldsymbol{x}_L$ with the hyperplane interface between the two media.

Lemma 17 *Let* σ *be a hyperplane of* $\mathbb{R}^d$, *with* $d \in \mathbb{N}^\star$ *and let* K, L *be the two open half-spaces with the common boundary* σ. *Let* $\Lambda_K \in \mathcal{M}_d(\mathbb{R})$ *and* $\Lambda_L \in \mathcal{M}_d(\mathbb{R})$ *be two given symmetric definite positive matrices, let* $\boldsymbol{n}_{KL}$ *be the unit vector normal to* σ *oriented from* K *to* L, $\boldsymbol{x}_K \in K$ *and* $\boldsymbol{x}_L \in L$ *be given and* $d_{K,\sigma}$ (*resp* $d_{L,\sigma}$) *the distance from* x_K (*resp* x_L) *to* σ. *We write* $\boldsymbol{x}_K = -d_{K,\sigma}\boldsymbol{n}_{KL} + \boldsymbol{y}_K$ *with* $\boldsymbol{y}_K \in \sigma$, *and* $\boldsymbol{x}_L = d_{L,\sigma}\boldsymbol{n}_{KL} + \boldsymbol{y}_L$, *with* $\boldsymbol{y}_L \in \sigma$. *Let* $\boldsymbol{y}_\sigma \in \sigma$ *(called the harmonic averaging point) be defined by*

$$\boldsymbol{y}_\sigma = \frac{\lambda_L d_{K,\sigma}\boldsymbol{y}_L + \lambda_K d_{L,\sigma}\boldsymbol{y}_K}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} + \frac{d_{K,\sigma}d_{L,\sigma}}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}(\boldsymbol{\lambda}_K^\sigma - \boldsymbol{\lambda}_L^\sigma), \qquad \text{[IV.33]}$$

denoting by $\lambda_K = \boldsymbol{n}_{KL}\cdot\Lambda_K\boldsymbol{n}_{KL}$, $\boldsymbol{\lambda}_K^\sigma = (\Lambda_K - \lambda_K\mathrm{Id})\boldsymbol{n}_{KL}$, $\lambda_L = \boldsymbol{n}_{KL}\cdot\Lambda_L\boldsymbol{n}_{KL}$ *and* $\boldsymbol{\lambda}_L^\sigma = (\Lambda_L - \lambda_L\mathrm{Id})\boldsymbol{n}_{KL}$. *Then the following harmonic averaging formula holds, for all functions* u *defined on* $\mathbb{R}^d$, *affine in* K *and* L, *such that* u *is continuous on* σ, *and such that* $\Lambda_K\nabla u_{|K}\cdot\boldsymbol{n}_{KL} = \Lambda_L\nabla u_{|L}\cdot\boldsymbol{n}_{KL}$*:*

$$u(\boldsymbol{y}_\sigma) = \frac{\lambda_L d_{K,\sigma}u(\boldsymbol{x}_L) + \lambda_K d_{L,\sigma}u(\boldsymbol{x}_K)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}. \qquad \text{[IV.34]}$$

Let us sketch the proof of the lemma. We denote by $\boldsymbol{G}_K$ the gradient of u in K, with $\boldsymbol{G}_K = g_K \boldsymbol{n}_{KL} + \boldsymbol{G}_K^\sigma$, $\boldsymbol{G}_K^\sigma \cdot \boldsymbol{n}_{KL} = 0$ and by $\boldsymbol{G}_L = g_L \boldsymbol{n}_{KL} + \boldsymbol{G}_L^\sigma$, $\boldsymbol{G}_L^\sigma \cdot \boldsymbol{n}_{KL} = 0$ the gradient of u in L. The continuity property of u on σ first leads to $\boldsymbol{G}_K^\sigma = \boldsymbol{G}_L^\sigma = \boldsymbol{g}^\sigma$ and then to

$$d_{K,\sigma} g_K + d_{L,\sigma} g_L = u(\boldsymbol{x}_L) - u(\boldsymbol{x}_K) + (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^\sigma,$$

and the condition $\Lambda_K \boldsymbol{G}_K . \boldsymbol{n}_{KL} = \Lambda_L \boldsymbol{G}_L . \boldsymbol{n}_{KL}$ can be written $g_K \lambda_K - g_L \lambda_L = \boldsymbol{g}^\sigma . (\boldsymbol{\lambda}_L^\sigma - \boldsymbol{\lambda}_K^\sigma)$. Expressing g_K with respect to $\boldsymbol{g}^\sigma$ and using $u(\boldsymbol{y}) = u(\boldsymbol{x}_K) + \boldsymbol{G}_K \cdot (\boldsymbol{y} - \boldsymbol{x}_K)$, for all $\boldsymbol{y} \in \sigma$, allows us to write that

$$\begin{aligned} u(\boldsymbol{y}) &= u(\boldsymbol{x}_K) + d_{K,\sigma} \frac{\lambda_L (u(\boldsymbol{x}_L) - u(\boldsymbol{x}_K) + (\boldsymbol{y}_K - \boldsymbol{y}_L) \cdot \boldsymbol{g}^\sigma) + d_{L,\sigma} \boldsymbol{g}^\sigma \cdot (\boldsymbol{\lambda}_L^\sigma - \boldsymbol{\lambda}_K^\sigma)}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}} \\ &+ (\boldsymbol{y} - \boldsymbol{y}_K) \cdot \boldsymbol{g}^\sigma. \end{aligned}$$

The point $\boldsymbol{y}_\sigma$ is then defined as the unique point $\boldsymbol{y} \in \sigma$ such that the preceding expression no longer depends on $\boldsymbol{g}^\sigma$, and the resulting expression for $u(\boldsymbol{y}_\sigma)$ follows. For all $v \in \mathcal{H}_\mathcal{D}$, $K \in \mathcal{T}$, $L \in \mathcal{N}_K$, we denote by $v_{K,L}$ the values,

$$v_{K,L} = \frac{\lambda_L d_{K,\sigma} u_L + \lambda_K d_{L,\sigma} u_K}{\lambda_L d_{K,\sigma} + \lambda_K d_{L,\sigma}}. \qquad \text{[IV.35]}$$

IV.5.2 Definition of the scheme

For all $K \in \mathcal{T}$, $L \in \mathcal{N}_K$, we denote by $y_{K,L}$ the harmonic averaging point given by (IV.5.1), by considering the two domains K and L and by $x_{K,L}$ the centre point of $\overline{K} \cap \overline{L}$. For all $K \in \mathcal{T}$, $L \in \mathcal{N}_K$, we denote by $h_{K,L}$ the point $y_{K,L}$ if $y_{K,L}$ belongs to the interior of $\overline{K} \cap \overline{L}$ else $h_{K,L} = x_{K,L}$. We require that for all $K \in \mathcal{T}$, $s \in \mathcal{V}_K$, $\sigma \in \mathcal{E}_s \cap \mathcal{E}_K$ with $\mathcal{T}_\sigma = \{K, L\}$, that $h_{K,L} \in \overline{\sigma}$. For all $K \in \mathcal{T}$, $\mathcal{E}_{\text{int},K} = \{]x_K, h_{K,L}[, L \in \mathcal{N}_K\} \cup \{]x_K, x_\sigma[, \sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}\}$ and the set of sub-cells of K, $\mathcal{S}_K$ is deduce from the set $\mathcal{E}_{\text{int},K} \cup \mathcal{E}_K$ according to Hypothesis (8)-$(H2)$ (see figure IV.3).

Remark 5 *We said that a mesh is not too distorted if for all $K \in \mathcal{T}$, $L \in \mathcal{N}_K$, $h_{K,L}$ is the harmonic averaging point given by (IV.5.1).*

For all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_{\text{int},K}$ such that $\sigma =]x_K, h_{K,L}[$ with $L \in \mathcal{N}_K$,

- if $h_{K,L}$ is the harmonic averaging point given by (IV.5.1), by considering the two domains K and L, then $\sigma \neq \mathcal{F}$ and for all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, $\Pi_\sigma v = \frac{1}{2}(v_K + v_{K,L})$
- else $\sigma \in \mathcal{F}$ which means that we have a face unknown.

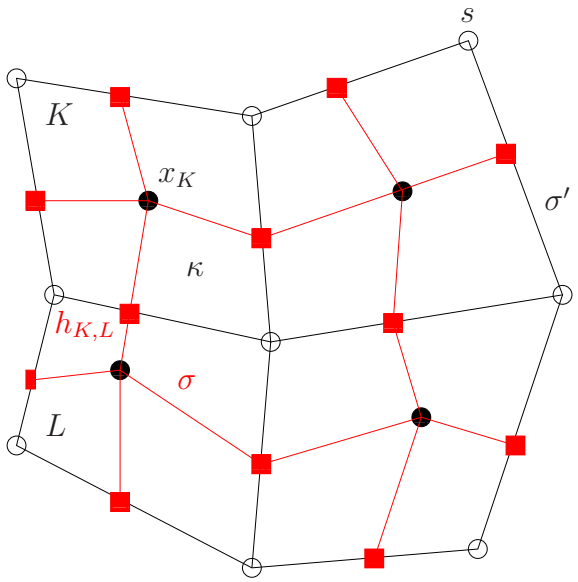


Figure IV.3: $\mathcal{E}_{\text{int},K}$: the red edges, $\mathcal{E}_K$: the black edges and $\kappa \in \mathcal{S}_K$, $K \in \mathcal{T}$.

For all $K \in \mathcal{T}$, $\sigma \in \mathcal{E}_{\text{int},K}$ such that $\sigma =]x_K, x_{\sigma'}[$ with $\sigma' \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}$, then $\sigma \neq \mathcal{F}$ and for all $v \in \mathcal{H}_{\mathcal{D},\mathcal{F}}$, $\Pi_\sigma v = \frac{1}{2} v_K$

Thanks to Lemma 17 and the properties of elements which belong to $\mathcal{Q}$, we can deduce that there exists $C > 0$ depending only on φ, ϱ_1, ϱ_2, and Λ such that for all $\varphi \in \mathcal{Q}$, $\sigma \in \mathcal{E}_{\text{int},K} \cap \overline{\mathcal{F}}$, $K \in \mathcal{T}$ such that $\sigma =]x_K, h_{K,L}[$ with $L \in \mathcal{N}_K$,

$$|\varphi(h_{K,L}) - (P_{\mathcal{F}}\varphi)_{K,L}| \leq C\ h_{\mathcal{D}}^2. \qquad \text{[IV.36]}$$

Owing to (IV.36), we get that for all $\sigma \in \mathcal{F}$, there exists $C > 0$ depending only on φ, ϱ_1, ϱ_2, and Λ such that for all $\varphi \in \mathcal{Q}$,

$$|\varphi(x_\sigma) - \Pi_\sigma(P_{\mathcal{F}}\varphi)| \leq C\ h_{\mathcal{D}}^2. \qquad \text{[IV.37]}$$

Chapitre V

Tests numériques 2D et 3D

Dans cette partie, nous avons consacré une rubrique traitant des tests numériques à chacun des schémas étudiés dans ce mémoire. Pour chaque section de ce chapitre, l'accent a été mis sur un schéma en particulier. Les tests numériques ont été menés dans le but d'étudier la convergence des schémas et leur robustesse sur des cas tests 2D, 3D synthétisant les difficultés que l'on peut rencontrer dans nos applications. nous avons étudié le comportement des schémas vis à vis :

- de l'hétérogénéité, l'anisotropie des tenseurs de diffusion,
- de la résolution des systèmes linéaires de façon itérative car nos codes industriels utilisent des solveurs itératifs,
- de la complexité des maillages rencontrés en modélisation de bassin, de réservoir et stockage du $Co2$.

Les résultats obtenus en terme de de convergence $L2$ de la solution discrète, des gradients ou des flux discrets (`cv`), de robustesse vis à vis , du maillage (`mesh`) de l'hétérogénéité (`het`), de l'anisotropie (`anis`) des tenseurs de diffusion et de la résolution itérative des systèmes linéaires (`resit`) sont résumés dans le tableau ci-dessous.

	`cv`	`mesh`	`resit`	`het`	`anis`
O scheme	oui	non	non	oui	non
G scheme	oui	oui	oui	oui	oui
VFSYM	oui	oui	non	oui	oui
CG method	oui	oui	oui	oui	oui
Dioptre	oui	oui	—	oui	oui

V.1 Implémentation

Pour les tests numériques 2D, on a développé un code prototype C++ se basant sur les librairies boost traitant de l'algèbre linéaire dense (www.boost.org) et de la librairie Petsc (www.mcs.anl.gov/petsc) pour résoudre nos systèmes linéaires. Au sein de ce code, on a développé un langage qui permet une implémentation formelle des schémas volumes finis sous sa formulation variationnelle. En ce qui concerne les tests numériques 3D, on a développé un code sous Arcane qui est une plateforme logicielle de développement pour des codes de simulation numérique à haute performance, pour des architectures de machines parallèles à mémoire distribuée.
Afin de permettre une implémentation formelle des schémas volumes finis, nous avons créé informatiquement une classe Expression composée d'un ensemble de couples d'objets définis par une valeur et un objet Inconnue. La classe Inconnue est définie par un mot clé et un entier. Le mot clé peut prendre deux valeurs ua pour préciser que c'est une inconnue de face et u pour préciser que c'est une inconnue de maille. L'entier donne l'indice de l'inconnue de face ou de maille suivant la valeur du mot clé. Pour la classe Expression, on a défini plusieurs méthodes, produit d'un objet Expression par un réel, addition, soustraction de deux objets Expression, addition, soustraction d'un objet Expression et d'un objet Inconnue, le résultat de ces méthodes donnant un nouvel objet Expression. On a aussi défini comme méthode le produit de deux objets Expression. Le produit de deux expressions $A = (\alpha_i, a_i)_{1\leq i\leq m}$, $B = (\beta_j, b_j)_{1\leq j\leq n}$ (a_i, b_j sont des objets Inconnue et α_i, β_j sont des réels) est une expression d'expressions. Pour $1 \leq j \leq n$, en écrivant b_j sous la forme $b_j = (\mu_j, l_j)$, où $\mu_j \in \{ua, u\}$ et $l_j \in \mathbb{N}^*$, le produit $A \cdot B$ est alors l'expression $\{(A * \beta_j, b_j)\}_{1\leq j\leq n}$, où les couples définissant celle-ci ont pour valeur, une expression, $A * \beta_j$ et pour objet Inconnue b_j. Le produit des expressions $A \cdot B$ peut être stocké dans une matrice, où chaque ligne correspond à une inconnue de face ou de maille, μ_j, de numéro l_j pour un j donné et cette ligne est égale à $A * \beta_j = \{\alpha_i\beta_j, a_i\}_{1\leq i\leq m}$, ainsi les colonnes correspondent aux objets Inconnue a_i. Prenons l'exemple d'un schéma volumes finis utilisant les inconnues de faces. Le schéma volumes finis sous formulation variationnelle est donné par :

$$\text{Trouver } u \in \mathcal{H}_\mathcal{D} \text{ s.t. } a_\mathcal{D}(u,v) = \int_\Omega f\ P_\mathcal{T} v \,\mathrm{d}x, \text{ Pour tout } v \in \mathcal{H}_\mathcal{D}. \qquad [V.1]$$

où la forme bilinéaire $a_\mathcal{D}$ est définie pour tout $(u, v) \in \mathcal{H}_\mathcal{D} \times \mathcal{H}_\mathcal{D}$ par,

$$a_\mathcal{D}(u,v) \;=\; \sum_{K\in\mathcal{T}} \sum_{\sigma\in\mathcal{E}_K} \underbrace{F_{K,\sigma}(u)}_{\text{Expr1}} \underbrace{(v_\sigma - v_K)}_{\text{Expr2}}.$$

Pour construire la matrice $\mathcal{A}$ provenant de la forme bilinéaire $a_\mathcal{D}$, on parcourt chaque maille $K \in \mathcal{T}$, chaque face de la maille K, $\sigma \in \mathcal{E}_K$, on construit formellement l'objet

Expression Expr1 du flux $F_{K,\sigma}(u)$ donné par la formule du schéma volumes finis considéré et l'objet Expression Expr2 du terme $v_\sigma - v_K$, puis en faisant le produit de ces deux objets, on obtient la contribution du terme $F_{K,\sigma}(u)(v_\sigma - v_K)$ dans la matrice $\mathcal{A}$.

V.2 3D numerical tests

Cell-Gradient method

In this section, we propose a comparison of the MPFA O, L, G and Cell-Gradient (referred as CG) methods on anisotropic heterogenous problems on challenging three-dimensional grids. Moving from two- to three-dimensional polyhedral meshes adds considerable difficulties. Indeed, all methods considered here can be proved to converge under suitable assumptions on both the permeability tensor and the mesh [10,12]. Such assumptions are more easily violated in three space dimensions. The aim of the comparison is to numerically evaluate the limitations of each scheme and to provide explications based on the theory developped in [10,12,15]. In order to illustrate the problems under investigation, some preliminary results are provided. To this purpose, the following indicators are used: `nunkw`, the number of unknowns; `erl2`, the L^2 error on the solution; `erinf`, the L^∞ error on the solution; `erFl2`, the L^2 error on the fluxes; `nit`, the number of (preconditioned) GMRes iterations required to solve the linear system.

Test 1 The first test is run on a sequence of squeezed grids. Each of them is obtained from a uniform hexahedral grid in a unit cube with step h by squeezing it 20 times along the z-axis and adding random perturbation in xy-plane of the order of $\frac{h}{2.5}$. A 2D slice of such a mesh is shown in Figure V.1. The following analytical solution (with suitable right hand side) is used in the convergence study:

$$u(x, y, z) = sin(\pi x)sin(\pi y)sin(\pi z).$$

Some results are presented in Figure V.2. While the Cell-Gradient schemes exhibit second order convergence, the L and G scheme also converge but to a lower order. The convergence of the O scheme seems more problematic.

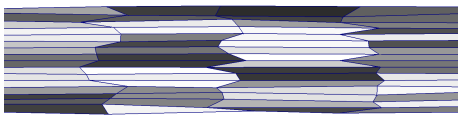


Figure V.1: Squeezed and randomly perturbed grid for Test 1

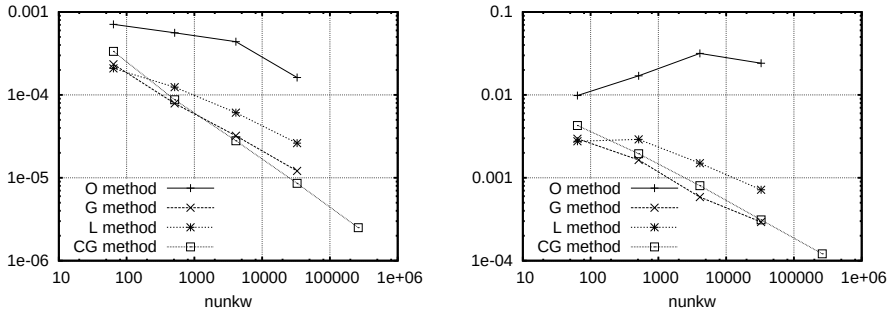


Figure V.2: `erl2` (left) and `erinf` for Test 1.

Test 2 The second test deals with a complex mixed mesh in the cube $[-500; 500]^3$ containing hexahedral, tetrahedral and pyramidal elements (see FigureV.3). The analitical solution for the convergence study (with suitable right hand side) is

$$u(x, y, z) = \sin\left(\frac{\pi}{1000}x\right) \sin\left(\frac{\pi}{1000}y\right) sin\left(\frac{\pi}{1000}z\right). \qquad [V.2]$$

In the first series of tests we solve the isotropic problem with $\Lambda = \mathrm{I}$. While a direct solver is used to assess the convergence of the scheme, the preconditioned GMRes method is used to estimate the solvability of the linear systems (see Figure V.4).

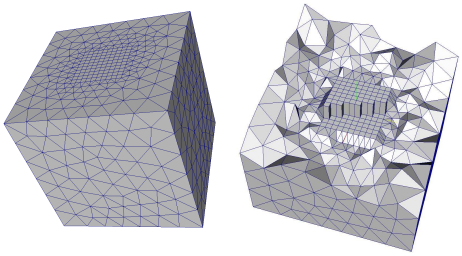


Figure V.3: Mixed grid for Tests 2 and 3.

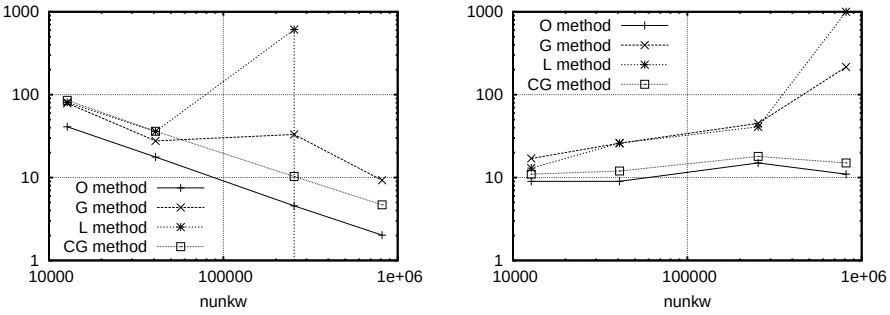


Figure V.4: `erl2` (left) and `Niter` (right) for Test 2

Test 3 In the second series of tests on the mesh family depicted in Figure V.3 we use again solution [V.2] with suitable right hand side and

$$\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0.1 \end{pmatrix}.$$

The results of Figure V.5 show convergence failure for the G scheme. Also, for the L scheme the linear solver systematically fails to reach convergence.

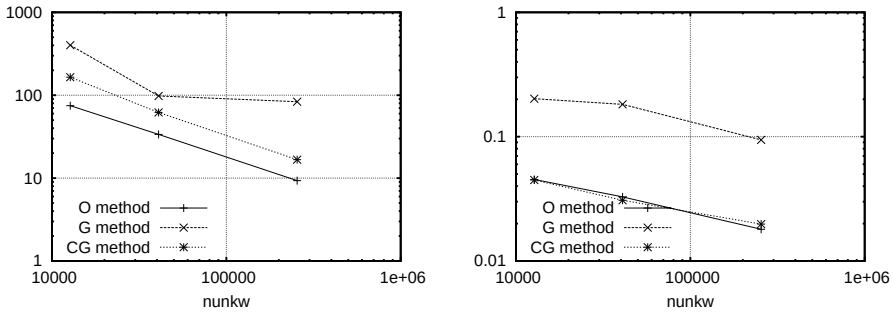


Figure V.5: `erl2` (left) and `erFl2` (right) for Test 3.

V.3 2D Numerical tests

V.3.1 MPFA O

There are many papers investigating the numerical convergence properties of the MPFA O scheme. For example, let us refer to [9] for quadrilateral grids in two and three dimensions, and to [50] in two dimensions with discontinuous diffusion coefficients. Also in [13], the MPFA O scheme is compared on challenging two dimensional anisotropic test cases with two unconditionally symmetric coercive finite volume schemes which exhibit a more robust convergence but at the expense of a much larger stencil.

Let us first discuss the coercivity condition (A.3) on a few particular remarkable cases. For all $\sigma \in \mathcal{E}_K \cap \mathcal{E}_s$ let us choose $m_\sigma = \frac{m_\sigma}{\mathrm{Card}(\mathcal{V}_\sigma)}$, x_K the isobarycenter of the vertices of the cell K and let x_σ be the center of gravity of the face σ. Then, for parallelogram and parallelepiped cells, the matrix B_K^s is equal to I. In such a case, the MPFA O scheme is symmetric and our sufficient condition of coercivity (A.3) is always satisfied. The same result holds for triangles with x_σ the barycenter with weights 2/3 at point s and 1/3 at the second end point of the edge σ. It holds again for tetrahedrons with x_σ the barycenter with weights 1/2 at point s and 1/4 at the two remaining end points of the face σ.

Let us now consider the case $d = 2$ with $\Lambda = I$, and let σ_1 and σ_2 be the two edges shared by a given vertex s of a given cell K. For $\sigma = \sigma_1, \sigma_2$, we assume that the continuity point x_σ is the center of gravity x_σ of the edge σ and that $m_\sigma = |x_\sigma - s|$. Then, the condition $\lambda_{\min}(B_K^s + (B_K^s)^t) \geq 2\theta$ is equivalent to $|x_{\sigma_1} - x_{\sigma_2}|\ |\overrightarrow{sx_{\sigma_1}} - \overrightarrow{x_{\sigma_2}x_K}| \leq 2(1-\theta)m_K^s$. For example, the trapezoidal mesh shown in Figure V.6

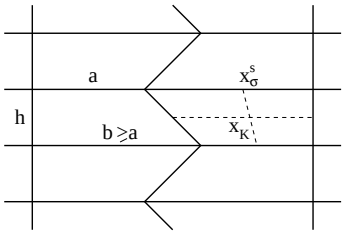


Figure V.6: Example of a trapezoidal mesh.

satisfies the coercivity condition (A.3) if and only if $\frac{b-a}{h} \leq (1-\theta)\frac{3a+b}{(b^2+h^2)^{1/2}}$ which exhibits the lack of robustness of the MPFA O scheme for distorted quadrangular meshes.

Next, let us discuss the sharpness of the coercivity criteria on a two dimensional example. We solve the anisotropic diffusion test case introduced in [28] on a family of skewed quadrangular meshes of the domain $\Omega = (0,1)^2$ of size $n_x \times n_x$ with $n_x = 20, 40, 80, 160$. The exact solution and the expression for the permeability coefficient are given below:

$$u = \sin(\pi x)\sin(\pi y), \quad K = \frac{1}{x^2+y^2}\begin{bmatrix} \delta x^2 + y^2 & (\delta-1)xy \\ (\delta-1)xy & x^2 + \delta y^2 \end{bmatrix}. \qquad \text{[V.3]}$$

We shall understand that Dirichlet boundary conditions are given on each boundary edge $\sigma \in \mathcal{E}_{\text{int}}$ by $u(x_\sigma)$, and that the forcing term is equal to $-\nabla \cdot (K\nabla u)$. The parameter δ is in fact the ratio between the minimum and the maximum eigenvalue of K.

The continuity points x_σ are the center of gravity of the edge σ and $m_\sigma = m_\sigma/2$ for all $s \in \mathcal{V}_\sigma$, $\sigma \in \mathcal{E}$, and the cell center is the isobarycenter of its four vertices.

The mesh $n_x = 20$ is plotted in Figure V.7 as well as the convergence of the MPFA O scheme for different values of δ, where *nunkw* denotes the number of cells n_x^2. We note that the convergence seems to be broken for $\delta = 0.001$.

In Table V.1 the sharpness of the two criteria of coercivity coercell$(\mathcal{D}, \Lambda)$, and coernode$(\mathcal{D}, \Lambda)$ are assessed. For that purpose we also compute the smallest eigenvalue of the symmetric part of the cell centered scheme matrix denoted by coerschurmesh$(\mathcal{D}, \Lambda)$, as well as coerschurnode$(\mathcal{D}, \Lambda)$, the non-zero smallest eigenvalue of the symmetric part of all the cell centered scheme submatrices around each vertex s of the mesh.

We note in Table V.1 that the positivity criteria coercell$(\mathcal{D}, \Lambda) \geq 0$ and coernode$(\mathcal{D}, \Lambda) \geq 0$ are more restrictive than the positivity of the cell centered scheme around each vertex coerschurnode$(\mathcal{D}, \Lambda) \geq 0$ which is a sufficient condition for the positivity of the cell centered finite volume scheme but not for the positivity of the hybrid finite volume scheme. From Table V.1 and Figure V.7, the convergence of the MPFA O scheme seems to be more closely related to the coercivity of the cell centered scheme. This is yet to be understood and goes beyond our framework based on the coercivity of the hybrid formulation.

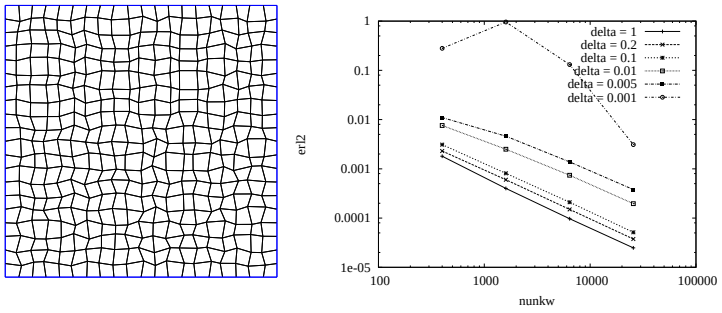


Figure V.7: Convergence of the L^2 error (erl2) for the MPFA O scheme.

criterion/mesh	$n_x = 10$	$n_x = 40$	$n_x = 80$	$n_x = 160$
coercell$(\mathcal{D}, \Lambda) \geq 0$	0.1	0.14	0.17	0.18
coernode$(\mathcal{D}, \Lambda) \geq 0$	0.06	0.09	0.09	0.11
coerschurnode$(\mathcal{D}, \Lambda) \geq 0$	0.012	0.014	0.016	0.02
coerschurmesh$(\mathcal{D}, \Lambda) \geq 0$	0.0055	0.0058	0.0068	0.014

Table V.1: Approximate smallest value of δ for which the coercivity criterion is positive for the different meshes and the various criteria.

V.3.2 Gscheme

The objective of this section is to assess the performance of the method described in Example 2 of the section III.2 on challenging, diffusion problems combining mild or strong anisotropy, heterogeneity and distorted or skewed meshes. For the sake of completeness, a comparison is provided against (i) the method of [45, §2.2] refered to as Success; (ii) the MPFA O method of [1] and (iii) the MPFA L method of [6, 7], also described in Example 1 of the section III.2. In the first test case, we consider the Dirichlet problem associated with the following exact solution featuring anisotropic permeability:

$$\bar{u} = \sin(\pi x)\sin(\pi y), \quad \Lambda = \mathrm{diag}(0.1, 1).$$

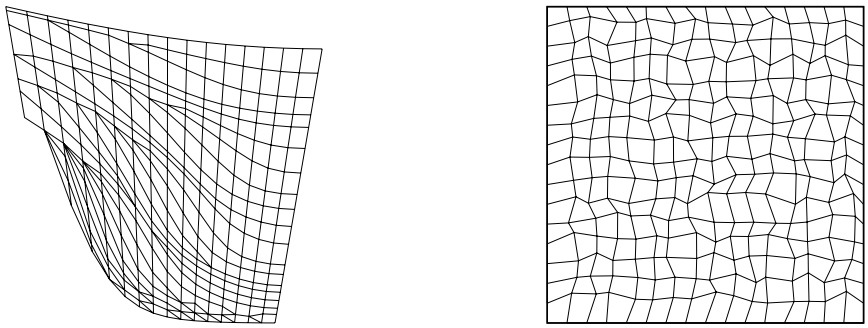
In the second test case, we consider the Dirichlet problem associated with the following exact solution featuring heterogeneous anisotropic permeability:

$$\bar{u} = \begin{cases} \sin(b\pi x)\sin(c\pi y) & \text{if } x \leq \delta, \\ \sin(b\pi\delta)\sin(c\pi y) + \pi b \dfrac{a_1}{a_2}\cos(b\pi\delta)\sin(c\pi y)(x-\delta) & \text{otherwise} \end{cases}$$

and

$$\Lambda = \begin{cases} \mathrm{diag}(a_1, b_1) & \text{if } x \leq \delta, \\ \mathrm{diag}(a_2, b_2) & \text{otherwise,} \end{cases}$$

where $b = \frac{1}{1.7}$, $c = 1.9$, $a_1 = 1$, $b_1 = 10$, $a_2 = 5, b_2 = 1, \delta = 0.5$. Both tests have been run on (i) a family of Corner Point Geometry basin meshes with erosion (see Figure V.8(a)) and (ii) a family of randomly distorted quadrangular meshes of $(0, 10) \times (0, 1)$ (see Figure V.8(b)).

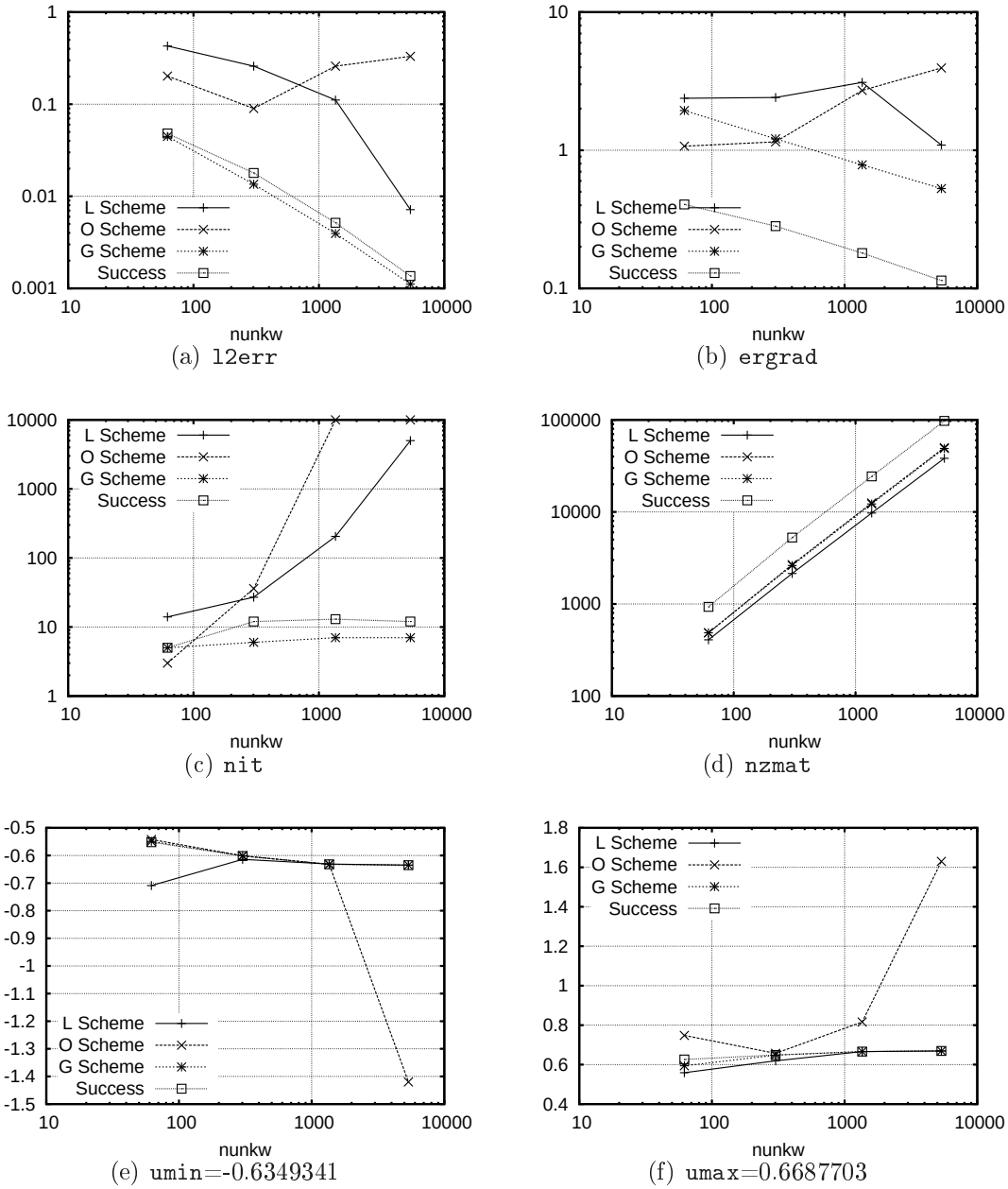


(a) Basin mesh. The actual aspect ratio is 10:1 (*x*:*y*)

(b) Randomly perturbed quadrangular mesh

Figure V.8: Mesh families.

The following indicators have been considered: `l2err`, the L^2 error; `ergrad`, the L^2 error on the gradient; `nit`, the number of preconditioned GMRes iterations; `nzmat`, the number of nonzero matrix entries; `umin`, the mininum of the discrete solution; `umax`, the maximum of the discrete solution. The number of degrees of freedom is denoted by `nunkw`. Blown up methods with respect to one indicator are not plotted to keep the scale readable. The linear systems have been solved with a direct solver for the indicators `l2err`, `umin`, `umax` and `ergrad`, whereas the GMRes algorithm from PETSc [18–20] with Hypre BoomerAMG preconditioner (see www.llnl.gov/CASC/hypre) has been used for `nit`. The stopping criterion required the preconditioned residual norm to be smaller than 10^{-7}. As expected, while sometimes displaying better accuracy, the Success scheme of [45, §2.2] has much denser matrices.

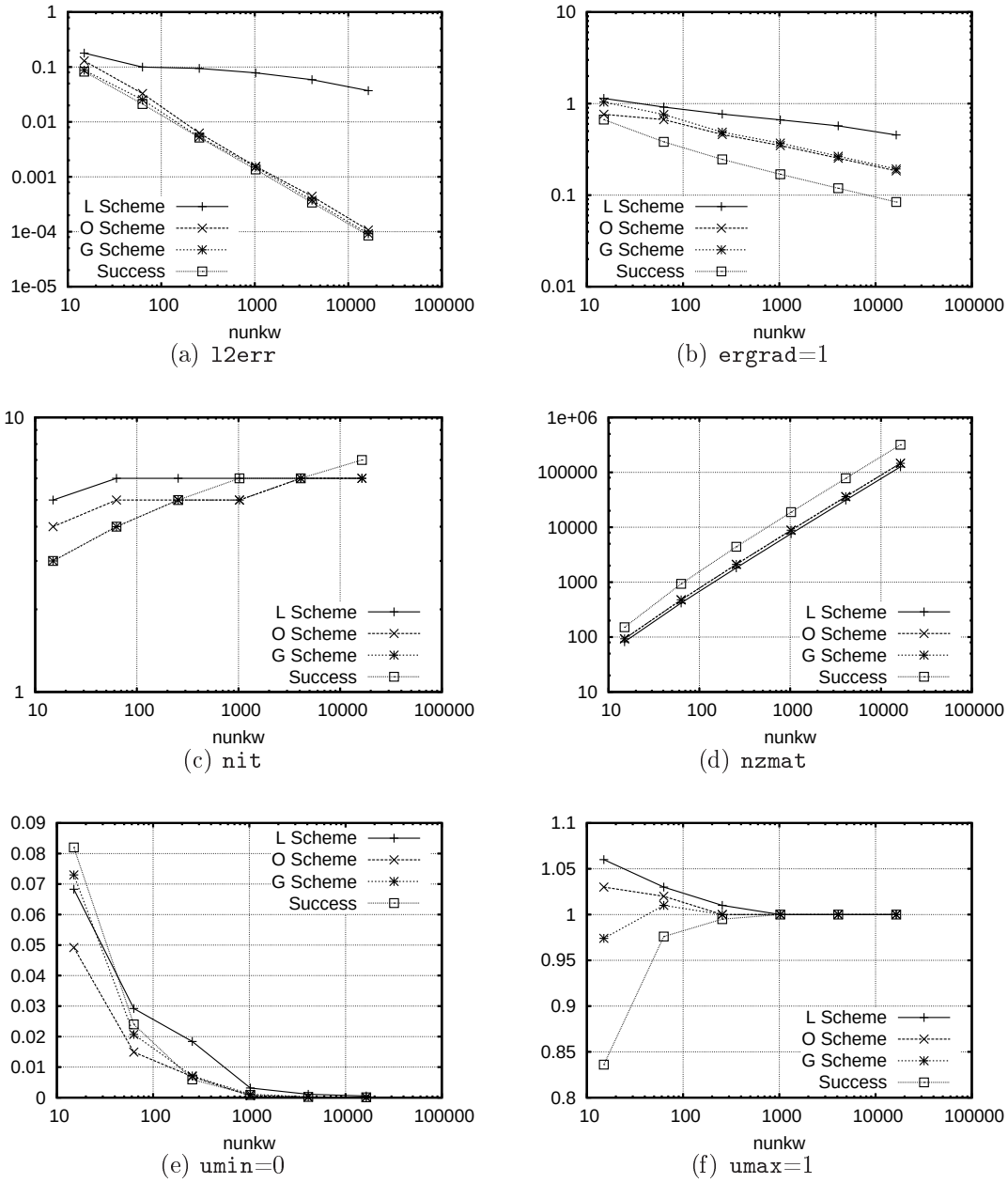


(a) `l2err` (b) `ergrad`

(c) `nit` (d) `nzmat`

(e) `umin`=-0.6349341 (f) `umax`=0.6687703

Figure V.9: Numerical results for test case 1 on the basin mesh family.

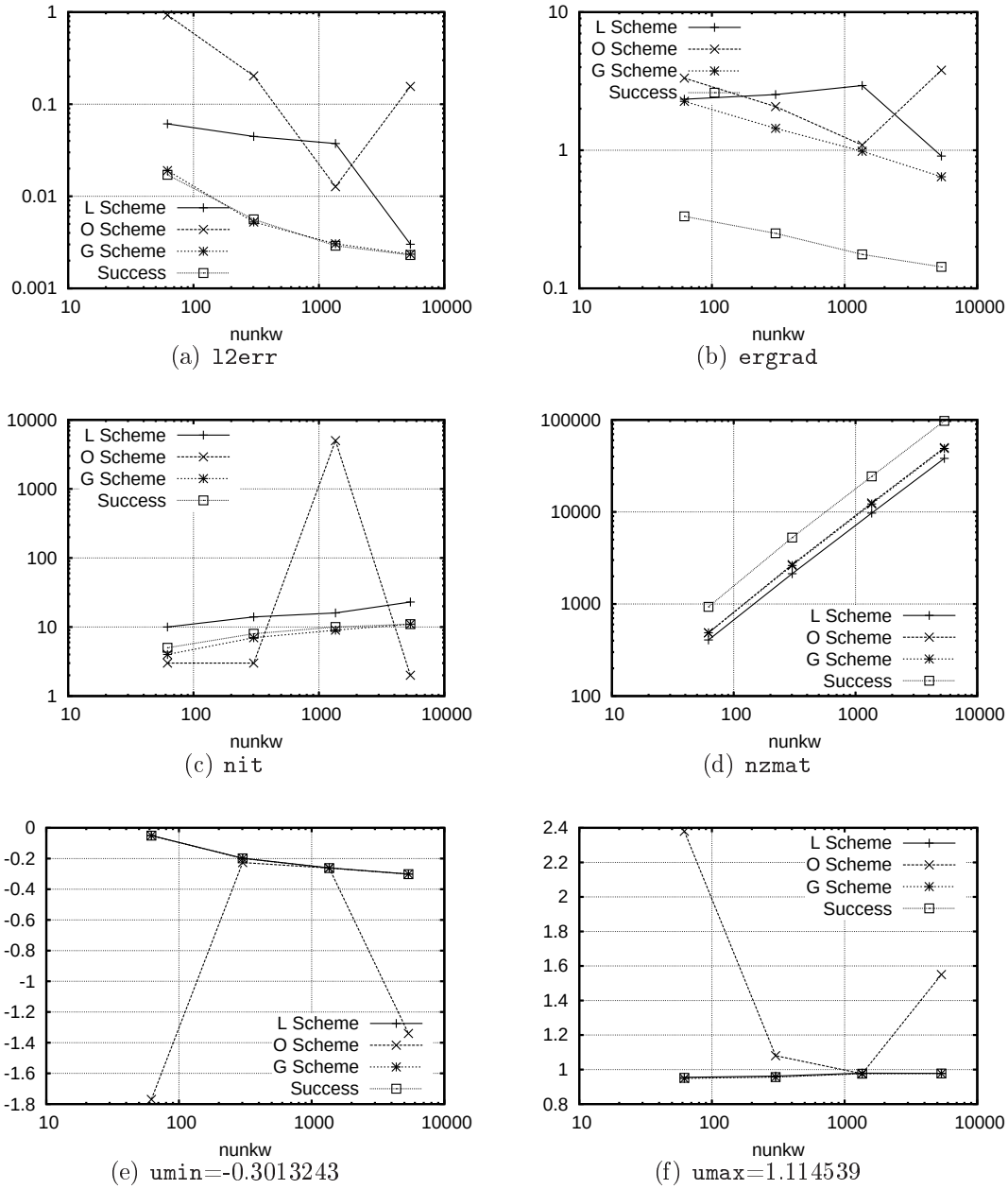


(a) `l2err` (b) `ergrad`=1

(c) `nit` (d) `nzmat`

(e) `umin`=0 (f) `umax`=1

Figure V.10: Numerical results for test case 1 on randomly perturbed mesh family



(a) `l2err` (b) `ergrad`

(c) `nit` (d) `nzmat`

(e) `umin`=-0.3013243 (f) `umax`=1.114539

Figure V.11: Numerical results for test case 2 on the basin mesh family.

V.3.3 VFSYM

The proposed method has been used to solve all the test cases described in [51]. The method has proved robust with respect to the skewdness of the mesh as well as with respect to the heterogeinity and anisotropy of the diffusion tensor. In a few cases, violations of the discrete maximum principle are observed, whose magnitude rapidly decreases when the mesh is refined. For the sake of simplicity, to calculate the indicator **sumflux**, we do not use the fluxes which derive from the scheme, but we use consistent fluxes, $F'_{K,\sigma}$, for all $K \in \mathcal{T}$ and $\sigma \in \mathcal{E}_K$ defined as follows

$$F'_{K,\sigma}(u) = \left(\frac{1}{\mathrm{m}_K} \sum_{\sigma \in \mathcal{E}_K} \mathrm{m}_{K_\sigma} (\nabla u)_{K_\sigma} \right) \cdot \mathrm{m}(\sigma) \Lambda_K \mathbf{n}_{K,\sigma}$$

This will explain the high values of **sumflux** in some tests.

The linear systems have been solved using PETSc (see `www.mcs.anl.gov/petsc`) and Hypre (see `www.llnl.gov/CASC/hypre`).

• **Test 1.1 Mild anisotropy,** $u(x,y) = 16x(1-x)y(1-y)$**,** min $= 0$, max $= 1$**, regular triangular mesh,** `mesh1`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
1_1	56	816	$-1.00e{+}00$	$1.51e{-}01$	–	$3.93e{-}01$	–
1_2	224	3736	$-3.37e{-}01$	$3.74e{-}02$	$2.01e{+}00$	$1.87e{-}01$	$1.07e{+}00$
1_3	896	15912	$-1.47e{-}01$	$9.22e{-}03$	$2.02e{+}00$	$9.23e{-}02$	$1.02e{+}00$
1_4	3584	65608	$-7.10e{-}02$	$2.28e{-}03$	$2.01e{+}00$	$4.58e{-}02$	$1.01e{+}00$
1_5	14336	266376	$-3.52e{-}02$	$5.68e{-}04$	$2.01e{+}00$	$2.28e{-}02$	$1.00e{+}00$

ocvl2=2.01, **ocvgradl2**=1.00

i	erflx0	erflx1	erfly0	erfly1	erflm
1_1	$5.21e{-}02$	$5.21e{-}02$	$5.21e{-}02$	$5.21e{-}02$	$2.42e{-}01$
1_2	$1.84e{-}02$	$1.84e{-}02$	$1.84e{-}02$	$1.84e{-}02$	$6.66e{-}02$
1_3	$8.54e{-}03$	$8.54e{-}03$	$8.54e{-}03$	$8.54e{-}03$	$1.82e{-}02$
1_4	$4.27e{-}03$	$4.27e{-}03$	$4.27e{-}03$	$4.27e{-}03$	$4.94e{-}03$
1_5	$2.16e{-}03$	$2.16e{-}03$	$2.16e{-}03$	$2.16e{-}03$	$1.28e{-}03$

i	umin	umax
1_1	$8.66e{-}02$	$1.03e{+}00$
1_2	$2.08e{-}02$	$1.01e{+}00$
1_3	$5.02e{-}03$	$1.00e{+}00$
1_4	$1.22e{-}03$	$1.00e{+}00$
1_5	$3.02e{-}04$	$1.00e{+}00$

• **Test 1.1 Mild anisotropy,** $u(x,y) = 16x(1-x)y(1-y)$**,** min $= 0$, max $= 1$**, skewed quadrangular mesh,,** `mesh4_1`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
4_1_1	289	5341	$-8.78e{-}01$	$3.22e{-}02$	–	$1.22e{-}01$	–
4_1_2	1156	22800	$-3.62e{-}01$	$5.37e{-}03$	$2.58e{+}00$	$3.49e{-}02$	$1.80e{+}00$
4_1_3	2601	52397	$-2.37e{-}01$	$2.48e{-}03$	$1.91e{+}00$	$1.48e{-}02$	$2.12e{+}00$
4_1_4	4624	94132	$-1.78e{-}01$	$1.49e{-}03$	$1.76e{+}00$	$9.83e{-}03$	$1.42e{+}00$
4_1_5	7225	148005	$-1.43e{-}01$	$1.01e{-}03$	$1.77e{+}00$	$6.55e{-}03$	$1.82e{+}00$

ocvl2=1.77, **ocvgradl2**=1.82

i	erflx0	erflx1	erfly0	erfly1	erflm
4_1_1	$5.29e{-}02$	$4.84e{-}02$	$4.16e{-}02$	$5.62e{-}02$	$9.70e{-}02$
4_1_2	$2.24e{-}02$	$2.08e{-}02$	$2.68e{-}02$	$1.54e{-}02$	$2.18e{-}02$
4_1_3	$1.45e{-}02$	$1.36e{-}02$	$1.88e{-}02$	$1.00e{-}02$	$9.59e{-}03$
4_1_4	$1.09e{-}02$	$1.03e{-}02$	$1.43e{-}02$	$7.74e{-}03$	$5.52e{-}03$
4_1_5	$8.72e{-}03$	$8.30e{-}03$	$1.16e{-}02$	$6.33e{-}03$	$3.45e{-}03$

i	umin	umax
4_1_1	$2.14e{-}03$	$9.84e{-}01$
4_1_2	$6.47e{-}04$	$9.98e{-}01$
4_1_3	$2.74e{-}04$	$1.00e{-}00$
4_1_4	$1.49e{-}04$	$1.00e{+}00$
4_1_5	$9.23e{-}05$	$1.00e{+}00$

• **Test 1.1 Mild anisotropy,** $u(x,y) = 16x(1-x)y(1-y)$**,** min $= 0$, max $= 1$**, skewed quadrangular mesh,,** `mesh4_2`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
4_2_1	1089	21437	$-3.59e{-}01$	$6.62e{-}03$	–	$3.13e{-}02$	–
4_2_2	4356	88592	$-1.83e{-}01$	$1.35e{-}03$	$2.29e{+}00$	$1.23e{-}02$	$1.34e{+}00$
4_2_3	9801	201485	$-1.23e{-}01$	$6.56e{-}04$	$1.78e{+}00$	$4.88e{-}03$	$2.28e{+}00$
4_2_4	17424	360116	$-9.23e{-}02$	$4.01e{-}04$	$1.71e{+}00$	$3.66e{-}03$	$1.00e{+}00$

ocvl2=1.71, **ocvgradl2**=1.00

i	erflx0	erflx1	erfly0	erfly1	erflm
4_2_1	$2.44e{-}02$	$2.46e{-}02$	$2.17e{-}02$	$1.37e{-}02$	$2.12e{-}02$
4_2_2	$1.14e{-}02$	$1.12e{-}02$	$1.27e{-}02$	$8.96e{-}03$	$4.92e{-}03$
4_2_3	$7.57e{-}03$	$7.40e{-}03$	$8.74e{-}03$	$6.34e{-}03$	$2.24e{-}03$
4_2_4	$5.68e{-}03$	$5.57e{-}03$	$6.63e{-}03$	$4.86e{-}03$	$1.42e{-}03$

i	umin	umax
4_2_1	$7.16e{-}04$	$9.93e{-}01$
4_2_2	$1.61e{-}04$	$9.99e{-}01$
4_2_3	$6.75e{-}05$	$1.00e{-}00$
4_2_4	$3.67e{-}05$	$1.00e{-}00$

• **Test 1.2 Mild anisotropy,** $u(x,y) = \sin((1-x)(1-y)) + (1-x)^3(1-y)^2$, $\min = 0$, $\max = 1 + \sin 1$, **regular triangular mesh,** `mesh1`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
1_1	56	816	$4.33e{-}01$	$3.78e{-}02$	–	$1.34e{-}01$	–
1_2	224	3736	$1.63e{-}01$	$1.05e{-}02$	$1.84e{+}00$	$6.88e{-}02$	$9.57e{-}01$
1_3	896	15912	$6.65e{-}02$	$2.72e{-}03$	$1.95e{+}00$	$3.45e{-}02$	$9.97e{-}01$
1_4	3584	65608	$2.92e{-}02$	$6.89e{-}04$	$1.98e{+}00$	$1.72e{-}02$	$1.00e{+}00$
1_5	14336	266376	$1.36e{-}02$	$1.73e{-}04$	$1.99e{+}00$	$8.57e{-}03$	$1.00e{+}00$

ocvl2=1.99, **ocvgradl2**=1.00

i	erflx0	erflx1	erfly0	erfly1	erflm
1_1	$7.23e{-}02$	$1.37e{-}02$	$8.17e{-}02$	$5.17e{-}02$	$1.11e{-}01$
1_2	$2.72e{-}02$	$1.94e{-}03$	$3.17e{-}02$	$1.75e{-}02$	$3.53e{-}02$
1_3	$1.10e{-}02$	$1.73e{-}04$	$1.32e{-}02$	$6.53e{-}03$	$9.94e{-}03$
1_4	$4.78e{-}03$	$4.66e{-}05$	$5.87e{-}03$	$2.69e{-}03$	$2.78e{-}03$
1_5	$2.21e{-}03$	$4.63e{-}05$	$2.75e{-}03$	$1.20e{-}03$	$7.37e{-}04$

i	umin	umax
1_1	$4.50e{-}03$	$1.37e{+}00$
1_2	$1.17e{-}03$	$1.59e{+}00$
1_3	$2.96e{-}04$	$1.71e{+}00$
1_4	$7.42e{-}05$	$1.78e{+}00$
1_5	$1.86e{-}05$	$1.81e{+}00$

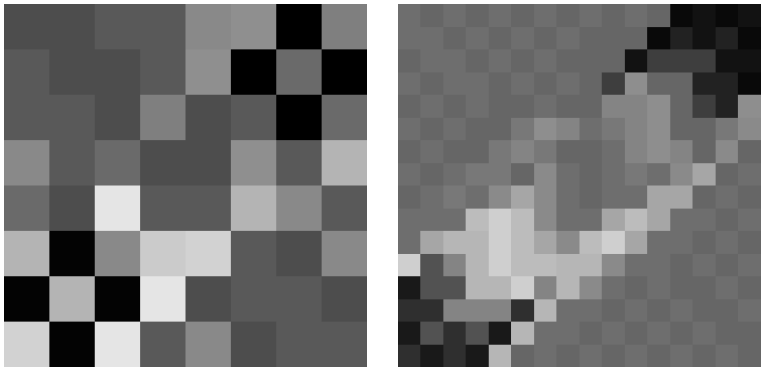
• **Test 1.2 Mild anisotropy,** $u(x,y) = \sin((1-x)(1-y)) + (1-x)^3(1-y)^2$, $\min = 0$, $\max = 1 + \sin 1$, **locally refined non-conforming rectangular mesh,** `mesh3`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
3_1	40	564	$2.03e{-}01$	$6.21e{-}02$	–	$6.72e{-}02$	–
3_2	160	2780	$9.86e{-}02$	$1.45e{-}02$	$2.10e{+}00$	$2.69e{-}02$	$1.32e{+}00$
3_3	640	12252	$4.64e{-}02$	$3.40e{-}03$	$2.09e{+}00$	$1.01e{-}02$	$1.42e{+}00$
3_4	2560	51356	$2.23e{-}02$	$8.20e{-}04$	$2.05e{+}00$	$3.67e{-}03$	$1.46e{+}00$
3_5	10240	210204	$1.09e{-}02$	$2.01e{-}04$	$2.03e{+}00$	$1.31e{-}03$	$1.48e{+}00$

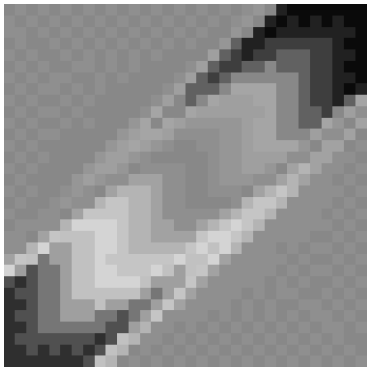
ocvl2=2.03, **ocvgradl2**=1.48

i	erflx0	erflx1	erfly0	erfly1	erflm
3_1	$4.74e{-}02$	$2.36e{-}02$	$4.81e{-}03$	$8.98e{-}02$	$1.10e{-}01$
3_2	$1.99e{-}02$	$1.44e{-}02$	$2.16e{-}03$	$5.57e{-}02$	$3.46e{-}02$
3_3	$8.77e{-}03$	$7.10e{-}03$	$2.26e{-}03$	$2.94e{-}02$	$9.70e{-}03$
3_4	$4.05e{-}03$	$3.51e{-}03$	$1.44e{-}03$	$1.51e{-}02$	$2.63e{-}03$
3_5	$1.93e{-}03$	$1.75e{-}03$	$7.95e{-}04$	$7.63e{-}03$	$7.07e{-}04$

i	umin	umax
3_1	$1.63e{-}03$	$1.65e{+}00$
3_2	$5.05e{-}04$	$1.75e{+}00$
3_3	$1.35e{-}04$	$1.79e{+}00$
3_4	$3.42e{-}05$	$1.82e{+}00$
3_5	$8.59e{-}06$	$1.83e{+}00$



(a) `mesh2_2` (b) `mesh2_3`



(c) `mesh2_4`

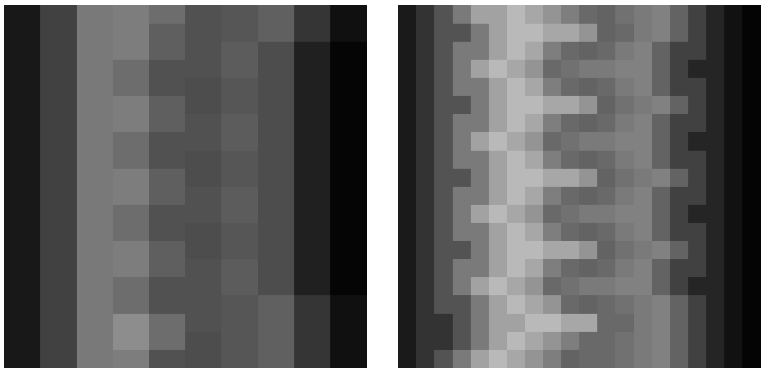
Figure V.12: Solution plots for test 3

• **Test 3 Oblique flow,** min = 0, max = 1**, uniform rectangular mesh,** `mesh2`

i	nunkw	nnmat	sumflux	fluxh0	fluxh1
2_1	16	180	$-7.22e{-16}$	$-2.47e{-01}$	$2.47e{-01}$
2_2	64	1012	$-1.11e{-16}$	$-2.21e{-01}$	$2.21e{-01}$
2_3	256	4692	$-2.43e{-15}$	$-2.08e{-01}$	$2.08e{-01}$
2_4	1024	20116	$-3.51e{-15}$	$-2.01e{-01}$	$2.01e{-01}$
2_5	4096	83220	$-4.49e{-14}$	$-1.97e{-01}$	$1.97e{-01}$
2_6	16384	338452	$-3.05e{-13}$	$-1.95e{-01}$	$1.95e{-01}$

i	fluyh0	fluyh1	umin	umax
2_1	$-1.86e{-01}$	$1.86e{-01}$	$-1.75e{-01}$	$1.17e{+00}$
2_2	$-1.33e{-01}$	$1.33e{-01}$	$-1.83e{-01}$	$1.18e{+00}$
2_3	$-1.18e{-01}$	$1.18e{-01}$	$-1.10e{-02}$	$1.01e{+00}$
2_4	$-1.08e{-01}$	$1.08e{-01}$	$-1.18e{-02}$	$1.01e{+00}$
2_5	$-1.03e{-01}$	$1.03e{-01}$	$7.51e{-04}$	$9.99e{-01}$
2_6	$-1.01e{-01}$	$1.01e{-01}$	$-1.00e{-03}$	$1.00e{+00}$

i	ener1	ener2
2_1	$2.13e{-01}$	$2.86e{-01}$
2_2	$2.39e{-01}$	$2.62e{-01}$
2_3	$2.38e{-01}$	$2.50e{-01}$
2_4	$2.42e{-01}$	$2.49e{-01}$
2_5	$2.42e{-01}$	$2.45e{-01}$
2_6	$2.42e{-01}$	$2.44e{-01}$



(a) `mesh5` (b) `mesh5_reg`

Figure V.13: Solution plots for test 4

• **Test 4 Vertical fault,** min = 0, max = 1**, non-conforming rectangular mesh,** `mesh5`

i	nunkw	nnmat	sumflux	fluxh0	fluxh1
5	105	1719	2.54e+00	−4.43e+01	4.63e+01
5_reg	400	7540	2.45e+00	−4.21e+01	4.44e+01

i	fluyh0	fluyh1	ener1	ener2	umin	umax
5	4.95e−01	1.55e−04	4.14e+01	4.41e+01	4.52e−02	9.59e−01
5_reg	1.71e−01	7.05e−04	4.14e+01	4.21e+01	2.12e−02	9.81e−01

• **Test 5 Heterogeneous rotating anisotropy,** min = 0, max = 1 **, uniform rectangular mesh,** `mesh2`

i	nunkw	nnmat	sumflux	erl2	ratioerl2	ergrad	ratioergrad
2_1	16	180	−4.19e−01	1.07e+01	–	2.28e+00	–
2_2	64	1012	−1.27e−01	1.50e+00	2.83e+00	1.46e+00	6.42e−01
2_3	256	4692	−1.05e−01	2.20e−01	2.77e+00	7.15e−01	1.03e+00
2_4	1024	20116	−7.03e−02	3.99e−02	2.47e+00	3.20e−01	1.16e+00
2_5	4096	83220	−3.55e−02	8.69e−03	2.20e+00	1.06e−01	1.59e+00
2_6	16384	338452	−1.33e−02	2.05e−03	2.09e+00	2.99e−02	1.83e+00

ocvl2=2.09, **ocvgradl2**=1.83

i	erflx0	erflx1	erfly0	erfly1	erflm
2_1	1.47e+00	4.08e−02	1.47e+00	4.08e−02	1.03e+00
2_2	1.49e+00	7.18e−03	1.49e+00	7.18e−03	2.94e−01
2_3	1.41e+00	2.03e−02	1.41e+00	2.03e−02	8.84e−02
2_4	1.19e+00	1.65e−02	1.19e+00	1.65e−02	2.42e−02
2_5	7.55e−01	8.58e−03	7.55e−01	8.58e−03	6.29e−03
2_6	3.16e−01	2.91e−03	3.16e−01	2.91e−03	1.60e−03

i	umin	umax
2_1	−1.92e+01	5.38e+00
2_2	−5.28e+00	1.34e+00
2_3	−1.39e+00	1.03e+00
2_4	−3.57e−01	1.00e+00
2_5	−9.06e−02	1.00e+00
2_6	−2.28e−02	1.00e+00

• **Test 6 Oblique drain,** min = −1.2, max = 0**, Coarse** `mesh6` **and Fine** `mesh7` **oblique meshes**

i	nunkw	nnmat	sumflux	erl2	ergrad
C	210	3748	−4.48e−14	8.18e−16	8.93e−15
F	230	3976	2.10e−12	3.41e−11	3.65e−09

i	erflx0	erflx1	erfly0	erfly1	erflm
C	$1.04e{-}15$	$9.55e{-}15$	$2.91e{-}15$	$4.16e{-}16$	$4.36e{-}14$
F	$4.40e{-}11$	$4.36e{-}11$	$2.05e{-}13$	$2.07e{-}13$	$5.65e{-}10$

i	umin	umax
C	$-1.15e{+}00$	$-5.43e{-}02$
F	$-1.15e{+}00$	$-5.43e{-}02$

• **Test 7 Oblique barrier,** min $= -5.575$, max $= 0.575$**, coarse** `mesh6` **and fine** `mesh7` **oblique meshes**

i	nunkw	nnmat	sumflux	erl2
6	210	3748	$5.98e{-}14$	$1.24e{-}15$

i	ergrad	erflx0	erflx1	erfly0	erfly1
6	$5.10e{-}14$	$6.93e{-}14$	$1.54e{-}14$	$2.33e{-}15$	$4.52e{-}14$

i	erflm	umin	umax
6	$1.04e{-}01$	$-5.54e{+}00$	$5.37e{-}01$

• **Test 8 Perturbed parallelograms,** min $= 0$**, perturbed parallelograms mesh** `mesh8`

i	nunkw	nnmat	umin	umax
8	121	2077	$-3.38e-02$	$1.12e-01$

• **Test 9 Anisotropy and wells,** min $= 0$**, Anisotropy and wells** `mesh9`

i	nunkw	nnmat	umin	umax
9	121	2037	$-3.69e-02$	$1.04e+00$

V.3.4 DIOPTRE

We tested the scheme for some of the cases described in the benchmark [44], in particular those with anisotropy and heterogeneity such as tests cases 5 and 6 (geological barrier and drain), and, as expected, the results are exact since the solution is piecewise affine in these cases. We have also run test case 5 (heterogeneous rotating anisotropy). An order 2 of convergence is then observed on the L^2-norm of the unknown. The finest mesh that we used for this test has 640×640 grid blocks, computed within a few minutes on a PC. A direct solver could be used, with numbering the unknowns by the classical domain decomposition strategy (this strategy holds for 9-point stencils).

We also consider a test case with a mesh inspired form those used in geological studies (see figure V.3.4). We take Λ =diag$(0.1, 1)$ and f such that the exact

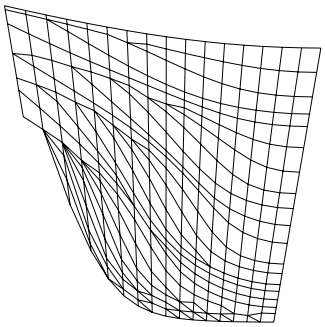


Figure V.14: Example of mesh

solution be given by $u(x, y) = \sin(\pi x)\sin(\pi y)$.

We get the following results when refining the mesh (mesh 1) depicted in Figure 2:

	mesh 1	mesh 2	mesh 3	mesh 4	mesh 5
$\#\mathcal{T}$	62	302	1357	5363	21031
# hybrid edges	1	3	6	10	17
L^2-error	$9.15\ 10^{-3}$	$3.07\ 10^{-3}$	$9.30\ 10^{-4}$	$2.66\ 10^{-4}$	$6.89\ 10^{-5}$

These results confirm the expected numerical convergence.

Chapitre VI

Conclusion et Perspectives

Dans ce travail de thèse, nous avons présenté plusieurs schémas volumes finis pour le problème de diffusion sur maillages généraux pour des tenseurs hétérogènes et anisotropes. Au vu de nos applications traitant de l'écoulement de fluides (huile, gaz) en milieux poreux hétérogènes, en simulation de réservoir, stockage du $Co2$ et bassin, les schémas de discrétisation devaient s'adapter à la complexité géologique des milieux poreux (canaux, érosions, failles). La description précise de ces milieux se traduit au niveau du maillage par des cellules qui peuvent présenter de grands contrastes de taille entre elles, des interfaces entre mailles pas nécessairement planes et au niveau des caractéristiques géologiques par des tenseurs hétérogènes, anisotropes. Il fallait donc des schémas de discrétisation pour le flux de pression, de température qui soient précis, robustes et peu couteux en temps calculs. Nous avons alors recherché des schémas volumes finis qui présentaient les caractéristiques suivantes : linéarité des flux en fonction des inconnues, stencil compact des flux (pour obtenir des gains en temps calculs), conservativité des flux (car on traite des problèmes de conservation de masse), consistance des flux (pour la précision du schéma), stabilité au sens où la forme bilinéaire dérivant du schéma volumes finis soit coercive (qui est un critère de robustesse du schéma). Pour certains des schémas présentés dans ce mémoire, la preuve de convergence a été donnée sous des hypothèses minimales : c'est à dire une hypothèse de consistance des flux vérifiée par les schémas volumes finis et une hypothèse de stabilité liée à la coercivité du schéma, permettant d'assurer l'existence et l'unicité de la solution du schéma et elle est essentielle dans la preuve de convergence. Bien que les schémas symétriques présentés dans ce mémoire satisfassent pratiquement à toutes ces propriétés sous uniquement des hypothèses de régularité du maillage, une propriété importante mais qui leur fait défaut en général est le stencil compact pour les flux, cela signifie que seules les proches cellules voisines et les faces de bords voisines de la face en question interviennent dans l'approximation du flux. En effet, les bonnes propriétés des schémas symétriques présentés dans ce mémoire sont soit au prix d'un large stencil comme

c'est le cas pour le schéma VFSYM, soit de sévères restrictions sur le maillage, par exemple sur des maillages triangulaires et tétraédriques, pour des points de continuité adéquates, le schéma MPFA O donne un schéma symétrique, consistant et coercive, avec en plus un stencil compact (cf V.3.1) et pour des maillages pas trop déformés au sens évoqué dans la remarque 5 où les cellules sont séparées par des plans, alors le schéma DIOPTRE qui est un schéma symétrique et coercive, donne un schéma avec un stencil compact. Cependant, la propriété de stencil compact du flux est très importante pour obtenir des gains calculs significatifs lors de la résolution sur ordinateurs à architectures parallèles. Notre choix s'est alors porté sur des schémas non symétriques qui ont été intégrés dans nos codes industriels et ceux prévus pour être industrialisés. Ces schémas présentent des stencils compacts pour les flux et satisfont à toutes les propriétés recherchées sous des hypothèses de régularité du maillage sauf celle de stabilité que l'on ne peut obtenir sous des hypothèses de régularité du maillage. Cette stabilité liée à la coercivité du schéma est souvent perdue pour des tenseurs très anisotropes. La convergence de ces schémas est alors obtenue sous l'hypothèse de coercivité que l'on ne peut vérifier que numériquement. Il en ressort que pour des maillages pas trop déformés ou pour des problèmes avec des tenseurs pas trop anisotropes comme c'est souvent le cas dans nos applications, les schémas non symétriques étudiés dans cette thèse sont alors adaptés. En ce qui concerne les schémas symétriques, le schéma DIOPTRE écrit en 2D, semble être le mieux adapté, cependant, sa mise en œuvre en 3D, reste une difficulté majeure. Les résultats originaux apportés dans cette thèse sont :

- l'introduction de nouveaux schémas
- l'introduction d'un cadre mathématique pour l'analyse des schémas volumes finis inspiré de celui introduit dans [42, 43],
- l'extension du résultat de convergence obtenu dans [45] aux schémas non symétriques,
- l'introduction d'un nouvel espace de fonctions tests dense dans H_0^1, permettant d'obtenir la vérification de l'hypothèse de consistance qui serait impossible à obtenir avec l'espace de fonctions tests usuel, C_c^∞ ou C_0^2,
- l'obtention d'une estimation d'erreur sans supposer une régularité supplémentaire de la solution du problème continu.

Dans le tableau ci-dessous, nous avons fait une synthèse des propriétés des différents schémas que nous avons proposés lors de cette thèse et qui figurent dans ce mémoire.

	2D	3D	stencil compact	coercivité	symétrique
MPFA O généralisé	oui	oui	oui	non	non
G scheme	oui	oui	oui	non	non
VFSYM	oui	oui	non	oui	oui
CG method	oui	oui	non	non	non
Dioptre	oui	non	oui	oui	oui

Les perspectives de ce travail seraient d'étudier les problèmes liés à la monotonie des schémas qui est une propriété recherchée dans les applications réservoirs car elle conditionne le changement de phase au niveau des puits, de voir comment ces schémas s'inscrivent dans un problème diphasique afin d'obtenir un schéma convergent pour ce problème.

Annexe A

Gradients discrets pour les schémas MPFA O

Two examples of construction of the gradient (III.6)

From Lemma 7, there is only one way to build a gradient (III.6) satisfying the consistency hypothesis 6 when the cardinal q_K^s of $\mathcal{E}_K \cap \mathcal{E}_s$ is equal to d. On the other hand, when $q_K^s > d$ there are many ways to build such gradient. Two examples are given in the two subsections below.

A.1 First construction

For all $K \in \mathcal{T}$ and $s \in \mathcal{V}_K$, let us define the square d-dimensional matrix B_K^s by

$$B_K^s = \frac{1}{\mathrm{m}_K^s} \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \mathrm{m}_\sigma \boldsymbol{n}_{K,\sigma} (x_\sigma - x_K)^t. \qquad [A.1]$$

The gradient (III.6) is defined by $B_K^s g_{K,\sigma}^s = \frac{\mathrm{m}_\sigma}{\mathrm{m}_K^s} \mathbf{n}_{K,\sigma}$, for all $\sigma \in \mathcal{E}_s \cap \mathcal{E}_K$, i.e.

$$(\overline{\nabla}_{\mathcal{D}} u)_K^s \quad = \quad (B_K^s)^{-1} (\widetilde{\nabla}_{\mathcal{D}} u)_K^s, \qquad [A.2]$$

assuming that the matrix B_K^s is non-singular. If q_K^s is equal to the space dimension d, and the set of vectors $(x_\sigma - x_K)_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ spans $\mathbb{R}^d$, the matrix B_K^s is non-singular iff the set of vectors $(\mathbf{n}_{K,\sigma})_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ spans also $\mathbb{R}^d$. For more general meshes, the non-singularity of B_K^s will be shown in subsection A.1 to result from a stronger assumption (A.3) ensuring also the coercivity of the scheme. Note however that if the set of vectors $(\mathbf{n}_{K,\sigma})_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ does not span $\mathbb{R}^d$, as it may be the case for non-matching meshes, the matrix B_K^s is singular and the present construction does not apply. This case will be taken into account in the second example.

Assuming that B_K^s is non-singular, we can easily check that the consistency hypothesis 6 is satisfied.

Coercivity of the finite volume scheme.

The main advantage of this construction is that a simple condition can be derived which ensures the non-singularity of the matrices B_K^s, the coercivity condition $\text{coernode}(\mathcal{D},\Lambda) \geq \theta_{\mathcal{D}}$ as well as an upper bound for the parameter ϱ_5 involved in the stability of the gradient function $\overline{\nabla}_{\mathcal{D}} u$.

This condition imposes the following non-negative lower bound

$$\text{coercell}(\mathcal{D},\Lambda) \geq \bar{\theta}_{\mathcal{D}} > 0, \qquad [A.3]$$

on the coercivity parameter defined by

$$\text{coercell}(\mathcal{D},\Lambda) = \min_{K\in\mathcal{T}, s\in\mathcal{V}_K} \lambda_{\min}\left(\frac{\Lambda_K B_K^s + (\Lambda_K B_K^s)^t}{2}\right). \qquad [A.4]$$

It can be easily computed for any given finite volume discretization $\mathcal{D}$ and diffusion tensor Λ.

The condition (A.3) ensures that the matrices B_K^s (A.1) defining the discrete gradients (A.2) are non-singular for all $s \in \mathcal{V}_K$, $K \in \mathcal{T}$ as stated in Lemma 19. To prove this result, we first need to state the following lemma.

Lemma 18 *Let $A \in \mathcal{M}_d(\mathbb{R})$ such that $\lambda_{\min}(A + A^t) > 0$, then A is a non-singular matrix and satisfies the estimate*

$$|A^{-1}| \leq \frac{8}{3}\,\frac{1}{\lambda_{\min}(A+A^t)}$$

Proof. We readily have $A \neq 0$. Let us consider the following estimates

$$\begin{aligned} |rA - I_d|^2 &= |(rA-I_d)^t(rA-I_d)| = |(I_d - r(A^t+A)) + r^2A^tA|, \\ &\leq |I_d - r(A^t+A)| + |r^2A^tA| = |I_d - r(A^t+A)| + r^2|A|^2. \end{aligned}$$

Choosing in the following $r = \dfrac{\lambda_{\min}(A+A^t)}{4|A|^2}$ ensures that all the eigenvalues of the symmetric matrix $I_d - r(A^t+A)$ are positive, and we have $|I_d - r(A^t+A)| = 1 - r\lambda_{\min}(A+A^t)$. Hence, we have proved the estimate

$$|rA - I_d|^2 \leq 1 - 3\left(\frac{\lambda_{\min}(A+A^t)}{4|A|}\right)^2.$$

It results that $|rA - I_d| < 1$. Then, setting $rA = I_d + (rA - I_d)$ we can obtain that rA is a non-singular matrix and that the following estimates hold

$$\begin{aligned} |(rA)^{-1}| &\leq \frac{1}{1-|rA-I_d|} = \frac{1+|rA-I_d|}{1-|rA-I_d|^2} \\ &\leq \frac{2}{1-|rA-I_d|^2} \\ &\leq \frac{2}{3}\left(\frac{4|A|}{\lambda_{\min}(A+A^t)}\right)^2, \end{aligned}$$

which concludes the proof.
□

Lemma 19 *Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1 such that there exists a real $\bar{\theta}_{\mathcal{D}} > 0$ with* coercell$(\mathcal{D}, \Lambda) \geq \bar{\theta}_{\mathcal{D}}$*, then for all $s \in \mathcal{V}_K$, $K \in \mathcal{T}$, the matrix B_K^s is non-singular, and its norm satisfies the following estimate*

$$|(B_K^s)^{-1}| \leq \frac{4\beta_0}{3\bar{\theta}_{\mathcal{D}}}. \qquad [A.5]$$

Proof. Using the assumption

$$2 \text{ coercell}(\mathcal{D}, \Lambda) = \lambda_{\min}(\Lambda_K B_K^s + (\Lambda_K B_K^s)^t) \geq 2\ \bar{\theta}_{\mathcal{D}} > 0$$

and Lemma 18, we deduce that the matrix $\Lambda_K B_K^s$ is non-singular as well as the matrix B_K^s. Still from Lemma 18, we have the estimate

$$|(\Lambda_K B_K^s)^{-1}| \leq \frac{4}{3\bar{\theta}_{\mathcal{D}}},$$

which concludes the proof from the bound $|\Lambda_K| \leq \beta_0$.
□

The following Lemmae 20 and 6 state respectively that the condition (A.3) provides an upper bound for the parameter ϱ_5 and that it ensures the coercivity condition coernode$(\mathcal{D}, \Lambda) \geq \theta_{\mathcal{D}}$.

Lemma 20 *Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1 such that there exists a real $\bar{\theta}_{\mathcal{D}} > 0$ with* coercell$(\mathcal{D}, \Lambda) \geq \bar{\theta}_{\mathcal{D}}$*. Then, we have the following estimate*

$$\max_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s, K \in \mathcal{T}_s, s \in \mathcal{V}} \frac{\mathrm{m}(\sigma)}{\mathrm{m}_K^s |g_{K,\sigma}^s|} \geq \frac{3\ \bar{\theta}_{\mathcal{D}}}{4\ \beta_0}.$$

Proof. The estimate derives from Lemma 19.
□

Proposition 6 *[coercivity of the scheme] Let $\mathcal{D}$ be an admissible discretization in the sense of Definition II.2.1 such that there exists a real $\bar{\theta}_{\mathcal{D}} > 0$ with* $\mathrm{coercell}(\mathcal{D}, \Lambda) \geq \bar{\theta}_{\mathcal{D}}$. *Then, setting* $\theta_{\mathcal{D}} = \frac{1}{2}\min\left(\mu_0, \frac{\varrho_1^2\ \bar{\theta}_{\mathcal{D}}}{\mathrm{CardFace}(\mathcal{D})}\right)$, *we have the lower bound* $\mathrm{coernode}(\mathcal{D}, \Lambda) \geq \theta_{\mathcal{D}}$ *and hence the coercivity of the bilinear form $a_{\mathcal{D}}$*

$$a_{\mathcal{D}}(u,u) \geq \theta_{\mathcal{D}} \|u\|_{\mathcal{D}}^2, \tag{A.6}$$

for all $u \in \mathcal{H}_{\mathcal{D}}$.

Proof. Let s be a given vertex of $\mathcal{V}$. From definition (III.18) of the bilinear form $a_{\mathcal{D}^s}$, we have for all $u \in \mathcal{H}_{\mathcal{D}^s}$

$$\begin{aligned} a_{\mathcal{D}^s}(u,u) = \sum_{K \in \mathcal{T}_s} \Bigg(& \mathrm{m}_K^s (\overline{\nabla}_{\mathcal{D}} u)_K^s \cdot \frac{\Lambda_K B_K^s + (B_K^s)^t \Lambda_K}{2} (\overline{\nabla}_{\mathcal{D}} u)_K^s \\ & + \alpha_K^s \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}} R_{K,\sigma}^s(u)^2 \Bigg). \end{aligned} \tag{A.7}$$

Using the following inequality

$$\mu(a-b)^2 \geq \frac{1}{2}\min(\mu,\lambda)a^2 - \lambda b^2, \ \text{ for all } (a,b,\mu,\lambda) \in (\mathbb{R}_+)^4,$$

with $\mu = \alpha_K^s,\ a = u_\sigma - u_K,\ b = (\overline{\nabla}_{\mathcal{D}} u)_K^s \cdot (x_\sigma - x_K)$ and $\lambda = \rho_K^s$, we obtain for all $\rho_K^s \geq 0$ the lower bound

$$\begin{aligned} & \alpha_K^s \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}} R_{K,\sigma}^s(u)^2 \geq \\ & \frac{1}{2}\min(\rho_K^s, \alpha_K^s) \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}} (u_\sigma - u_K)^2 - \rho_K^s (\overline{\nabla}_{\mathcal{D}} u)_K^s \cdot A_K^s (\overline{\nabla}_{\mathcal{D}} u)_K^s, \end{aligned} \tag{A.8}$$

where the square matrix A_K^s is defined by

$$A_K^s = \sum_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s} \frac{\mathrm{m}(\sigma)}{d_{K,\sigma}} (x_\sigma - x_K)(x_\sigma - x_K)^t, \tag{A.9}$$

and satisfies the bound

$$|A_K^s| \leq \frac{d \mathrm{m}_K^s}{\varrho_1^2}. \tag{A.10}$$

Let us choose ρ_K^s such that

$$\rho_K^s = \sup\left\{ \rho \in \mathbb{R},\ \mathrm{m}_K^s\ \frac{\Lambda_K B_K^s + (\Lambda_K B_K^s)^t}{2} - \rho A_K^s \geq 0 \right\}. \tag{A.11}$$

Using the upper bound (A.10), and the local coercivity assumption (A.3), (A.4), we can prove that ρ_K^s defined by (A.11) satisfies the lower bound

$$\rho_K^s \geq \frac{3}{4}\frac{\varrho_1^2\ \bar{\theta}_{\mathcal{D}}}{d}, \qquad [A.12]$$

for all $s \in \mathcal{V}_K$, $K \in \mathcal{T}$. Using (A.7), (A.8), (A.11), (A.12) we obtain the lower bound

$$a_{\mathcal{D}^s}(u,u) \geq \frac{1}{2}\min\left(\mu_0, \frac{3\varrho_1^2\ \bar{\theta}_{\mathcal{D}}}{4d}\right)\ \|u\|_{\mathcal{D}}^2, \qquad [A.13]$$

for all $u \in \mathcal{H}_{\mathcal{D}^s}$ which concludes the proof.
□

A.2 Second construction

This second finite volume scheme uses the construction of the gradient $(\overline{\nabla}_{\mathcal{D}}u)_K^s$ introduced in [30] for $d = 2$ and 3. Compared with the previous approach, its main advantage is to cover the case of non-matching or locally refined grids for which the set of vectors $(\boldsymbol{n}_{K,\sigma})_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_s}$ may not span $\mathbb{R}^d$.

For each $\sigma \in \mathcal{E}$, let us denote by $\mathcal{E}_{K,\sigma}^s$ the subset of $\mathcal{E}_s \cap \mathcal{E}_K$ of cardinality d defined as follows for $d = 2$ and $d = 3$. For $d = 2$, let us set $\mathcal{E}_{K,\sigma}^s = \mathcal{E}_s \cap \mathcal{E}_K$. For $d = 3$, let e_1 and e_2 be the two edges of the face σ intersecting the vertex s, and σ_1 and σ_2 be the two faces of $\mathcal{E}_s \cap \mathcal{E}_K$ sharing respectively the edge e_1 and e_2 with the face σ. Then, we set $\mathcal{E}_{K,\sigma}^s = \{\sigma, \sigma_1, \sigma_2\}$.

For all $K \in \mathcal{T}$ and $s \in \mathcal{V}_K$, the gradient $(\overline{\nabla}_{\mathcal{D}}u)_K^s$ is defined by

$$(\overline{\nabla}_{\mathcal{D}}u)_K^s = \sum_{\sigma\in\mathcal{E}_s\cap\mathcal{E}_K} \frac{\mathrm{m}_\sigma\ d_{K,\sigma}}{d\ \mathrm{m}_K^s} \sum_{\sigma'\in\mathcal{E}_{K,\sigma}^s} (u_{\sigma'} - u_K)\ g_{K,\sigma,\sigma'}^s,$$

where $\{g_{K,\sigma,\sigma'}^s,\ \sigma' \in \mathcal{E}_{K,\sigma}^s\}$ is the biorthogonal basis of $\{(x_{\sigma'}^s - x_K),\ \sigma' \in \mathcal{E}_{K,\sigma}^s\}$ such that

$$(x_{\sigma'}^s - x_K)\cdot g_{K,\sigma,\sigma''}^s = \delta_{\sigma',\sigma''}$$

for all $\sigma', \sigma'' \in \mathcal{E}_{K,\sigma}^s$, assuming that the set of vectors $(x_{\sigma'}^s - x_K)$, $\sigma' \in \mathcal{E}_{K,\sigma}^s$ is free. Note that by construction, $\sum_{\sigma'\in\mathcal{E}_{K,\sigma}^s} v\cdot(x_{\sigma'} - x_K)\ g_{K,\sigma,\sigma'}^s = v$ for any vector $v \in \mathbb{R}^d$. It results that the gradient $(\overline{\nabla}_{\mathcal{D}}u)_K^s$ is consistent in the sense of hypothesis 6.

The upper bound of the parameter ϱ_5 is controlled in two dimensions by the minimum angle between the two vectors $(x_{\sigma'} - x_K), \sigma' \in \mathcal{E}_s \cap \mathcal{E}_K$. In three dimensions it is controlled by the minimum angles between a vector of $\{(x_{\sigma'} - x_K), \sigma' \in \mathcal{E}_{K,\sigma}^s\}$ and the two remaining ones. These minimum angles should not tend to zero.

From Lemma 7, this second approach is equivalent to the MPFA O scheme described in [1] and [60] as soon as q_K^s is equal to the space dimension d for all cells K and all vertices s of the cell K. It is always the case in two dimensions $d = 2$. If in addition the set of vectors $(\boldsymbol{n}_{K,\sigma})_{\sigma \in \mathcal{E}_K \cap \mathcal{E}_s}$ spans $\mathbb{R}^d$, then both the first and second constructions are equivalent to the MPFA O scheme [1] and [60].

The coercivity condition coernode$(\mathcal{D}, \Lambda) \geq \theta_{\mathcal{D}}$ has to be checked numerically. The stronger but simpler condition coercell$(\mathcal{D}, \Lambda) \geq \bar{\theta}_{\mathcal{D}}$ can also be used when both constructions match.

Annexe B

Quelques lemmes techniques pour le schéma Gscheme

B.1 Proof of Lemma 3

The proof is trivial if $\Lambda \in C(\overline{\Omega})$ since, in this case, $C_c^\infty(\Omega)$ is contained in $\mathcal{Q}$. The difficulty comes from the possible discontinuities of Λ through the interfaces of P_Ω, in which case item (iii) of Lemma 3 is not easy to obtain and might impose discontinuity of $\nabla\varphi$ through these interfaces. The proof is made in several steps, following the idea of [34]: we first eliminate the singularities (vertices if $d = 2$, vertices and edges if $d = 3$, etc.) of the boundaries of the open sets $\{\Omega_i\}_{1\leq i\leq N_\Omega}$ by showing that we only need approximate functions which vanish around these singularities; then we reason on each $\overline{\Omega_i}$, approximating a given function by functions having the same value on the boundary and vanishing derivatives in the direction $\Lambda\boldsymbol{n}$; gluing these approximations together, we obtain a function in $\mathcal{Q}$ which is close to the initial given function.

(i) Elimination of the singularities of $\{\Omega_i\}_{1\leq i\leq N_\Omega}$. First of all we notice that, since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, the result of the lemma follows if we prove that functions in $\mathcal{Q}$ approximate, in $H_0^1(\Omega)$, any $\psi \in C_c^\infty(\Omega)$.
Let S be the set of singularities of $\cup_{i=1}^{N_\Omega}\partial\Omega_i$ (i.e. affine parts of dimension $d-2$ or less: the vertices in dimension $d = 2$, the vertices and edges if $d = 3$, etc.); it is known that S has a 2-capacity equal to 0 and we can therefore find a sequence of functions $\gamma_n \in C_c^\infty(\mathbb{R}^d;[0,1])$ s.t. $\gamma_n \to 0$ in $H^1(\mathbb{R}^n)$ as $n \to \infty$ and, for all $n \in \mathbb{N}$, $\gamma_n \equiv 1$ on a neighborhood of S. If $\psi \in C_c^\infty(\Omega)$ and $\psi_n = (1-\gamma_n)\psi \in C_c^\infty(\Omega)$, then $\psi_n \to \psi$ in $H_0^1(\Omega)$ and, for all n, $\psi_n \equiv 0$ on a neighborhood of S. Hence, denoting by $C_{c,S}^\infty(\Omega)$ the set of functions in $C_c^\infty(\Omega)$ which vanish on neighborhoods of S, the proof of the lemma is complete if we can approximate, in $H_0^1(\Omega)$, elements of $C_{c,S}^\infty(\Omega)$ by elements of $\mathcal{Q}$.

(ii) Reduction to a Ω_i. Let $\psi \in C^\infty_{c,S}(\Omega)$ and assume that, for all $1 \leq i \leq N_\Omega$, there exists a sequence $\varphi^i_n \in C^2(\overline{\Omega_i})$ which converges to ψ in $H^1(\Omega_i)$ as $n \to \infty$ and s.t., for all $n \in \mathbb{N}$, $\varphi^i_n = \psi$ and $(\Lambda \nabla \varphi^i_n)_{|\overline{\Omega_i}} \cdot \boldsymbol{n}_i = 0$ on $\partial\Omega_i$. Define then $\varphi_n : \overline{\Omega} \to \mathbb{R}$ as the function equal to φ^i_n on $\overline{\Omega_i}$ for all $i = 1, \ldots, N_\Omega$; since $\varphi^i_n = \varphi^j_n = \psi$ on $\partial\Omega_i \cap \partial\Omega_j$, φ_n is well defined and continuous on $\overline{\Omega}$, it is C^2 on each $\overline{\Omega_i}$, it vanishes on $\partial\Omega$ (on which $\psi = 0$) and the tangential derivatives of φ_n are continuous through the interfaces of P_Ω (for all $\mathbf{t}$ parallel to $\partial\Omega_i \cap \partial\Omega_j$, the values of $(\nabla\varphi_n)_{|\overline{\Omega_i}} \cdot \mathbf{t}$ and $(\nabla\varphi_n)_{|\overline{\Omega_j}} \cdot \mathbf{t}$ on $\partial\Omega_i \cap \partial\Omega_j$ can be computed using only the values of $\varphi^i_n = \varphi^j_n = \psi$ on $\partial\Omega_i \cap \partial\Omega_j$, and are therefore equal). The continuity of φ_n accross the boundary of Ω_i for each i moreover ensures that $\nabla\varphi_n$ has no singularity on these boundaries and it is therefore simply the function equal to $\nabla\varphi^i_n$ on Ω_i for all i; hence, $\varphi_n \to \psi$ in $H^1_0(\Omega)$. Finally, the fluxes $\Lambda\nabla\varphi_n \cdot \boldsymbol{n}$ are clearly continuous through the interfaces of P_Ω since they vanish on either side of each such interface $\partial\Omega_i \cap \partial\Omega_j$.
To conclude the proof, it remains to find the convenient approximations $\{\varphi^i_n\}_{n\geq 1}$ of $\psi \in C^\infty_{c,S}(\Omega)$ on $\overline{\Omega_i}$.

(iii) Approximation on $\overline{\Omega_i}$. Let $\psi \in C^\infty_{c,S}(\Omega)$ and let $\mathcal{O}$ be an open set containing S s.t. $\psi \equiv 0$ on a neighborhood of $\overline{\mathcal{O}}$. Let $(F_l)_{1\leq l\leq r}$ be the faces of Ω_i (i.e. the affine parts of $\partial\Omega_i$ of dimension $d-1$); for all $1 \leq l \leq r$, we denote by $\boldsymbol{n}_l$ the unit normal to F_l pointing inside Ω_i and we define the C^2 function $f_l : \mathbb{R} \times F_l \to \mathbb{R}^d$ by

$$\forall t \in \mathbb{R}\,,\ \forall y \in F_l\,,\ f_l(t,y) = y + t\Lambda(y)\boldsymbol{n}_l. \qquad \text{[B.1]}$$

If $(t,y) \in \mathbb{R}\times F_l$ and $(t',y') \in \mathbb{R}\times F_l$ are s.t. $f_l(t,y) = f_l(t',y')$ then, since $(y-y')\cdot\boldsymbol{n}_l = 0$, one has $t\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l = t'\Lambda(y')\boldsymbol{n}_l\cdot\boldsymbol{n}_l$ and thus

$$y - y' = t\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l \left(\frac{\Lambda(y')\boldsymbol{n}_l}{\Lambda(y')\boldsymbol{n}_l\cdot\boldsymbol{n}_l} - \frac{\Lambda(y)\boldsymbol{n}_l}{\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l} \right). \qquad \text{[B.2]}$$

Letting $\varepsilon > 0$ be smaller than the inverse of the Lipschitz constant of $y \to \beta_0 \frac{\Lambda(y)\boldsymbol{n}_l}{\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l}$ (which is well-defined since $\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l > \alpha_0$ for all y), [B.2] can happen with $y \neq y'$ only if $|t| \geq \varepsilon$. Hence, f_l is one-to-one on $(-\varepsilon,\varepsilon)\times F_l$. We also notice that $\Lambda(y)\boldsymbol{n}_l$ is uniformly transverse to the hyperplane H_l containing F_l (this is again $\Lambda(y)\boldsymbol{n}_l \cdot \boldsymbol{n}_l \geq \alpha_0$) and thus that, upon reducing ε, the Jacobian matrix of f_l at any $(t,y) \in (-\varepsilon,\varepsilon)\times F_l$ is invertible.
Let $\mathcal{V}_l$ be an open neighborhood of $\overline{F_l\backslash\mathcal{O}}$ in F_l s.t. $\mathrm{dist}(\mathcal{V}_l, S) > 0$; the preceding reasoning shows that f_l is a C^2-diffeomorphism from $(-\varepsilon,\varepsilon)\times\mathcal{V}_l$ to $f_l((-\varepsilon,\varepsilon)\times\mathcal{V}_l)$, an open set in $\mathbb{R}^d$ containing in particular $f_l(\{0\}\times\overline{F_l\backslash\mathcal{O}}) = \overline{F_l\backslash\mathcal{O}}$. Since $\Lambda(y)\boldsymbol{n}_l$ points inside Ω_i (one more time, $\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l > 0$) and $\mathrm{dist}(\mathcal{V}_l, S) > 0$, upon reducing again ε if needed, we also see that $\mathcal{U}_l \overset{\text{def}}{=} f_l([0,\varepsilon[\times\mathcal{V}_l)$ is contained in $\overline{\Omega_i}$ and is a neighborhood of $\overline{F_l\backslash\mathcal{O}}$ in $\overline{\Omega_i}$ (see Figure B.1 for a representation of some sets appearing in this proof). Moreover, for all $x \in \mathcal{U}_l$, if $x = f_l(t,y)$ for $(t,y) \in [0,\varepsilon[\times\mathcal{V}_l$

then $\mathrm{dist}(x, H_l) = (x-y)\cdot\boldsymbol{n}_l = t\Lambda(y)\boldsymbol{n}_l\cdot\boldsymbol{n}_l$ and thus $0 \le t \le \frac{1}{\alpha_0}\mathrm{dist}(x, H_l)$. This shows that

$$\forall x \in \mathcal{U}_l\,, \quad \text{if } (t,y) = (f_{l|[0,\varepsilon[\times\mathcal{V}_l})^{-1}(x) \text{ then } |x-y| \le \frac{\beta_0}{\alpha_0}\mathrm{dist}(x, H_l). \qquad [B.3]$$

Let us define ψ_l on $\mathcal{U}_l$ s.t.

$$\psi_l(f_l(t,y)) = \psi(y) \quad \text{for all } (t,y) \in [0,\varepsilon[\times\mathcal{V}_l. \qquad [B.4]$$

ψ_l belongs to $C^2(\mathcal{U}_l)$ and $\psi_l = \psi$ on $\mathcal{V}_l$ (because $f_l(0,y) = y$); derivating [B.4] with respect to t, taking $t = 0$ and using [B.1] we also have

$$0 = \frac{d}{dt}(\psi_l(f_l(t,y)))_{|t=0} = \nabla\psi_l(y)\cdot\Lambda(y)\boldsymbol{n}_l = \Lambda(y)\nabla\psi_l(y)\cdot\boldsymbol{n}_l \quad \text{for all } y \in \mathcal{V}_l. \qquad [B.5]$$

As ψ vanishes on a neighborhood of $\overline{\mathcal{O}}$, there exists a neighborhood $\mathcal{N}_l$ of $\mathcal{V}_l \cap \mathcal{O}$ in $\mathcal{V}_l$ s.t. $\psi = 0$ on $\mathcal{N}_l$; [B.4] then implies $\psi_l = 0$ on $f_l([0,\varepsilon[\times\mathcal{N}_l)$ which is, f_l being a diffeomorphism, a neighborhood in $\mathcal{U}_l$ of $f_l(\{0\}\times(\mathcal{V}_l\cap\mathcal{O})) = \mathcal{V}_l\cap\mathcal{O}$; to sum up,

$$\psi_l = 0 \text{ on a neighborhood of } \mathcal{V}_l \cap \mathcal{O} \text{ in } \mathcal{U}_l. \qquad [B.6]$$

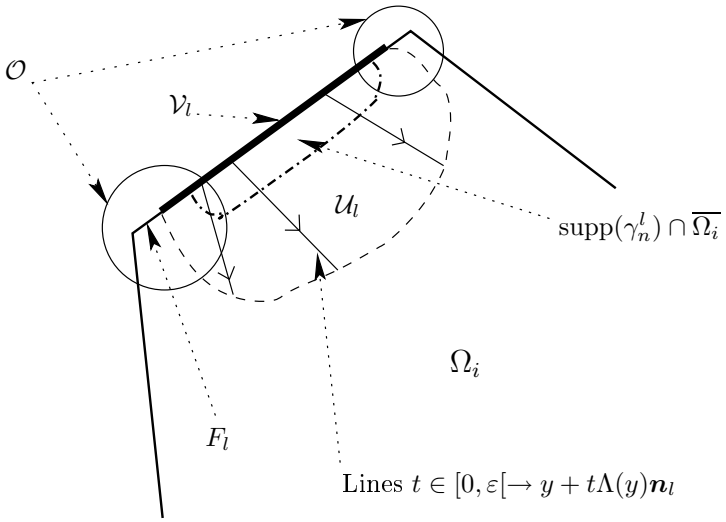


Figure B.1: Various sets appearing in the proof of Lemma 3.

For $1 \le l \le r$, we take a sequence $\gamma_n^l \in C_c^\infty(\mathbb{R}^d;[0,1])$ s.t., for all $n \in \mathbb{N}$, $\gamma_n^l \equiv 1$ on a neighborhood of $\overline{F_l\backslash\mathcal{O}}$ and

$$\gamma_n^l \equiv 0 \text{ on } \{x \in \mathbb{R}^d\,,\ \mathrm{dist}(x, F_l\backslash\mathcal{O}) \ge 1/n\} \quad \text{and} \quad ||\nabla\gamma_n^l||_{L^\infty(\mathbb{R}^d)} \le C_{13}n, \qquad [B.7]$$

with C_{13} not depending on n. If n is large, $\mathrm{supp}(\gamma_n^l)\cap\overline{\Omega_i}$ is a compact subset of $\mathcal{U}_l$ and $\gamma_n^l\psi_l$ can therefore be extended to $\overline{\Omega_i}$ by 0 outside $\mathcal{U}_l$ without loosing smoothness;

we then define $\varphi_n = \sum_{l=1}^r \gamma_n^l \psi_l + (1 - \sum_{l=1}^r \gamma_n^l)\psi \in C^2(\overline{\Omega_i})$. Since $\psi_l = \psi$ on $\mathcal{V}_l$ and, for n large enough, γ_n^l vanishes on $\partial\Omega_i$ outside $\mathcal{V}_l$, we have $\varphi_n = \psi$ on $\partial\Omega_i$ for such n. Still considering large n, on a neighborhood of $\overline{F_l \backslash \mathcal{O}}$ in $\overline{\Omega_i}$ we have $\gamma_n^l = 1$ and $\gamma_n^k = 0$ if $k \neq l$ and therefore, on such a neighborhood, $\varphi_n = \psi_l$; [B.5] thus shows that $\Lambda \nabla \varphi_n \cdot \boldsymbol{n} = 0$ on $\cup_{l=1}^r F_l \backslash \mathcal{O} = \partial\Omega_i \backslash \mathcal{O}$; since all the $\gamma_n^l \psi_l$ and ψ vanish on a neighborhood of $\partial\Omega_i \cap \mathcal{O}$ in $\overline{\Omega_i}$ (see [B.6]), we obviously also have $\Lambda \nabla \varphi_n \cdot \boldsymbol{n} = 0$ on $\partial\Omega_i \cap \mathcal{O}$, and thus on the whole boundary of Ω_i. It remains to prove that $\varphi_n \to \psi$ in $H^1(\Omega_i)$ as $n \to \infty$; in order to achieve this, we write $\varphi_n - \psi = \sum_{l=1}^r \gamma_n^l (\psi_l - \psi)$ and use [B.3], [B.4] and the smoothness of ψ to see that, if $\mathrm{dist}(x, F_l \backslash \mathcal{O}) \leq 1/n$, then $|\psi_l(x) - \psi(x)| \leq C_{14}/n$ with C_{14} not depending on n or x (because $x = f_l(t, y)$ with $y \in \mathcal{V}_l$ s.t. $|x - y| \leq \beta_0/(\alpha_0 n)$); we infer from [B.7] that

$$||\gamma_n^l(\psi_l - \psi)||_{L^2(\Omega_i)} \leq \frac{C_{14}}{n} \mathrm{meas}(\Omega_i)^{1/2}$$

and

$$||\nabla(\gamma_n^l(\psi_l - \psi))||_{L^2(\Omega_i)} \leq C_{13} C_{14} \mathrm{meas}(\Omega_i \cap \mathrm{supp}(\gamma_n^l))^{1/2} + ||\nabla(\psi_l - \psi)||_{L^2(\Omega_i \cap \mathrm{supp}(\gamma_n^l))}.$$

Since $\mathrm{meas}(\Omega_i \cap \mathrm{supp}(\gamma_n^l)) \to 0$ as $n \to \infty$, this concludes the proof that $\varphi_n \to \psi$ in $H^1(\Omega_i)$.

Remark 6 *The proof shows that Λ need not be C^2 on the whole of each $\overline{\Omega_i}$, only on the affine parts of $\partial\Omega_i$ (and the reader can check that the rest of the paper only requires the C^1 regularity of Λ on each $\overline{\Omega_i}$).*

B.2 Proof of Lemma 13

Proposition 7 *Let $\mathcal{D}$ be a generic element of a family of discretizations matching Definition II.2.1 satisfying hypothese (5)-(1). Let $\varphi \in \mathcal{Q}$, $\sigma \in \mathcal{E}_{\mathrm{int}}$ with $\mathcal{T}_\sigma = \{K, L\}$ and $y_\sigma \in \sigma$. Then $\nabla\varphi(x_K) - \nabla\varphi(x_L)$ can be decomposed as follows:*

$$\nabla\varphi(x_K) - \nabla\varphi(x_L) = \mu_\sigma \boldsymbol{n}_{K,\sigma} + \tau_\sigma \mathbf{t}_\sigma, \qquad \text{[B.8]}$$

where $|\mathbf{t}_\sigma| = 1$, $\mathbf{t}_\sigma \cdot \boldsymbol{n}_{K,\sigma} = 0$ and the reals μ_σ, τ_σ verify

$$|\tau_\sigma| \leq C_{15} \left[\mathrm{diam}L + \mathrm{diam}K\right], \qquad \text{[B.9]}$$

$$\mu_\sigma = -\frac{W_K(x_L)}{d_{L,\sigma}} + \tau_\sigma \frac{\mathbf{t}_\sigma \cdot (y_\sigma - x_L)}{d_{L,\sigma}} + \frac{W_K(y_\sigma) - W_L(y_\sigma)}{d_{L,\sigma}}, \qquad \text{[B.10]}$$

with

$$W_K(x) \stackrel{\mathrm{def}}{=} \varphi(x) - \varphi(x_K) - \nabla\varphi(x_K) \cdot (x - x_K) \qquad \text{[B.11]}$$

and $C_{15} \stackrel{\mathrm{def}}{=} \max(|\varphi''|_{L^\infty(K)}, |\varphi''|_{L^\infty(L)})$

PROOF. The vector $\mathbf{t}_\sigma$ is obviously the normed orthogonal projection of $\nabla\varphi(x_K) - \nabla\varphi(x_L)$ on the hyperplane parallel to σ, and the reals μ_σ, τ_σ are given by the formulæ

$$\mu_\sigma = (\nabla\varphi(x_K) - \nabla\varphi(x_L))\cdot\boldsymbol{n}_{K,\sigma}, \quad \tau_\sigma = (\nabla\varphi(x_K) - \nabla\varphi(x_L))\cdot\mathbf{t}_\sigma.$$

Since

$$\begin{aligned} -W_K(x_L) + W_K(y_\sigma) - W_L(y_\sigma) &= -\varphi(x_L) + \varphi(x_K) + \nabla\varphi(x_K)\cdot(x_L - x_K) \\ &\quad +\varphi(y_\sigma) - \varphi(x_K) - \nabla\varphi(x_K)\cdot(y_\sigma - x_K) \\ &\quad -\varphi(y_\sigma) + \varphi(x_L) + \nabla\varphi(x_L)\cdot(y_\sigma - x_L) \\ &= \nabla\varphi(x_K)\cdot(x_L - y_\sigma) + \nabla\varphi(x_L)\cdot(y_\sigma - x_L) \\ &= (\nabla\varphi(x_K) - \nabla\varphi(x_L))\cdot(x_L - y_\sigma), \end{aligned}$$

we can use [B.8] and the fact that $(x_L - y_\sigma)\cdot\boldsymbol{n}_{K,\sigma} = d_{L,\sigma}$ to re-write μ_σ under the form

$$\mu_\sigma = -\frac{W_K(x_L)}{d_{L,\sigma}} + \tau_\sigma\frac{\mathbf{t}_\sigma\cdot(y_\sigma - x_L)}{d_{L,\sigma}} + \frac{W_K(y_\sigma) - W_L(y_\sigma)}{d_{L,\sigma}}.$$

The face σ is completely contained either in one element of the partition P_Ω or in an interface of this partition; using then either the regularity of φ inside each element of the partition or the continuity of its tangential derivatives through the interfaces of P_Ω, we can re-write τ_σ under the form

$$\tau_\sigma = (\nabla\varphi(x_K) - \nabla\varphi(y_\sigma))\cdot\mathbf{t}_\sigma + (\nabla\varphi(y_\sigma) - \nabla\varphi(x_L))\cdot\mathbf{t}_\sigma$$

and the proof is complete since φ is C^2 on $\overline{K}$ and $\overline{L}$. □

Proposition 8 (Flux "quasi-continuity") *Let $\mathcal{D}$ be a generic element of a family of discretizations matching Definition II.2.1, satisfying hypotheses (5)-(1) and $\varphi \in \mathcal{Q}$. For all $G \in \mathcal{G}$, $\nabla\varphi(x_{K_G})$ is the solution of a linear system of equations of the form*

$$\mathcal{A}_G Y_G = \mathcal{B}_G(\varphi_\mathcal{T}) + \mathcal{C}_G(\varphi),$$

where $\varphi_\mathcal{T} \in H_\mathcal{T}(\Omega)$ is defined by the family $\{\varphi(x_K)\}_{K\in\mathcal{T}}$, $\mathcal{A}_G$ and $\mathcal{B}_G(\varphi_\mathcal{T})$ are the matrices defined in Lemma 11 and the vector $\mathcal{C}_G(\varphi)$ verifies

$$|\mathcal{C}_G(\varphi)| \leq C_1 \max_{K\in\mathcal{T}_G} \mathrm{diam}(K) \qquad [B.12]$$

with $C_1 > 0$ which only depends on ϱ_1, ϱ_2, Λ and φ.

PROOF. For a cell K, let W_K be the function defined by [B.11]. Since φ is C^2 regular on the closure of each element of P_Ω and since each cell is completely contained in one of these elements, there exists $C_16 > 0$ only depending on φ s.t., for all $K \in \mathcal{T}$,

$$|W_K(x)| \leq C_{16}\mathrm{diam}(K)^2 \text{ for all } x \in \overline{K}. \qquad [B.13]$$

For all $\sigma \in \mathcal{E}_{\rm int}$ with $\mathcal{T}_\sigma = \{K, L\}$ and $y_\sigma \in \sigma$, we apply Proposition 7 to decompose $\nabla\varphi(x_K) - \nabla\varphi(x_L)$ (note that the $W_K(x_L)$ appearing in [B.10] *is in general not* of order 2 with respect to the size of the mesh, since $x_L \notin \overline{K}$ and φ is not regular accross the boundary of some cells). Since $\varphi \in \mathcal{Q}$, we can also write

$$\frac{1}{\mathrm{m}_\sigma}\int_\sigma (\Lambda\nabla\varphi)_{|\overline{K}}(x)\cdot\boldsymbol{n}_{K,\sigma}\,\mathrm{d}x + \frac{1}{\mathrm{m}_\sigma}\int_\sigma (\Lambda\nabla\varphi)_{|\overline{L}}(x)\cdot\boldsymbol{n}_{L,\sigma}\,\mathrm{d}x = 0 \qquad [B.14]$$

and, $\nabla\varphi$ and Λ being C^1 on the closure of each control volume, we deduce from [B.14] that the real $\zeta_\sigma(\varphi) = \Lambda_K\nabla\varphi(x_K)\cdot\boldsymbol{n}_{K,\sigma} + \Lambda_L\nabla\varphi(x_L)\cdot\boldsymbol{n}_{L,\sigma}$ verifies

$$|\zeta_\sigma(\varphi)| \le C_{17}(\mathrm{diam}(K) + \mathrm{diam}(L)), \qquad [B.15]$$

where $C_{17} > 0$ depends only on φ, Λ.

Let us now consider $G \in \mathcal{G}$, $\sigma \in G \cap \mathcal{E}_{\rm int}$ and use these preliminary remarks with $K = K_G$ and L s.t. $\mathcal{T}_\sigma = \{K_G, L\}$. By definition of $\zeta_\sigma(\varphi)$ and [B.8],

$$\begin{aligned}(\Lambda_L\boldsymbol{n}_{L,\sigma} + \Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma})\cdot\nabla\varphi(x_{K_G}) &= \Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\nabla\varphi(x_{K_G}) - \Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\nabla\varphi(x_L) + \zeta_\sigma(\varphi)\\ &= -\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}\mu_\sigma + \tau_\sigma\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\mathbf{t}_\sigma + \zeta_\sigma(\varphi).\end{aligned}$$

Equation [B.10] and the definition of $W_{K_G}(x_L)$ then show

$$\begin{aligned}&(\Lambda_L\boldsymbol{n}_{L,\sigma} + \Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma})\cdot\nabla\varphi(x_{K_G})\\ &\quad= \frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}(\varphi(x_L) - \varphi(x_{K_G})) - \frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}\nabla\varphi(x_{K_G})\cdot(x_L - x_{K_G})\\ &\qquad+\zeta_\sigma(\varphi) - \frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}(\tau_\sigma\mathbf{t}_\sigma\cdot(y_\sigma - x_L) + W_{K_G}(y_\sigma) - W_L(y_\sigma)) + \tau_\sigma\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\mathbf{t}_\sigma\end{aligned}$$

and therefore

$$\begin{aligned}&\left(\Lambda_L\boldsymbol{n}_{L,\sigma} + \Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma} + \frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}(x_L - x_{K_G})\right)\cdot\nabla\varphi(x_{K_G}) =\\ &\qquad\qquad\frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}(\varphi(x_L) - \varphi(x_{K_G})) + c_\sigma(\varphi)\end{aligned}$$

with $c_\sigma(\varphi) = \zeta_\sigma(\varphi) - \frac{\Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\boldsymbol{n}_{L,\sigma}}{d_{L,\sigma}}(\tau_\sigma\mathbf{t}_\sigma\cdot(y_\sigma - x_L) + W_{K_G}(y_\sigma) - W_L(y_\sigma)) + \Lambda_L\boldsymbol{n}_{L,\sigma}\cdot\mathbf{t}_\sigma\tau_\sigma$.
If $\sigma \in G \cap \mathcal{E}_{\rm ext}$, using the definition of $W_{K_G}(x_\sigma)$, we have

$$\frac{\Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma}\cdot\boldsymbol{n}_{K_G,\sigma}}{d_{K_G,\sigma}}\nabla\varphi(x_{K_G})\cdot(x_\sigma - x_{K_G}) = \frac{\Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma}\cdot\boldsymbol{n}_{K_G,\sigma}}{d_{K_G,\sigma}}(-\varphi(x_{K_G})) + c_\sigma(\varphi)$$

with $c_\sigma(\varphi) = -\frac{\Lambda_{K_G}\boldsymbol{n}_{K_G,\sigma}\cdot\boldsymbol{n}_{K_G,\sigma}}{d_{K_G,\sigma}}W_{K_G}(x_\sigma)$.

We deduce that $\nabla\varphi(x_{K_G})$ is the solution of the linear system of equations

$$\mathcal{A}_G Y_G = \mathcal{B}_G(\varphi_\mathcal{T}) + \mathcal{C}_G(\varphi),$$

where $\mathcal{C}_G(\varphi)$ is the vector of $\mathbb{R}^d$ defined by $\{c_\sigma(\varphi)\}_{\sigma\in G}$. Thanks to (B.15), (B.9) and (B.13), there exists $C_{18} > 0$ which only depends on ϱ_1, ϱ_2, Λ and φ s.t., for all $\sigma \in G$ with $\mathcal{T}_\sigma = \{K_G, L\}$, $|c_\sigma(\varphi)| \le C_{18}(\mathrm{diam}(L) + \mathrm{diam}(K_G))$. The proof is complete. □

We are now in a position to prove Lemma 13. Let W_K the function defined by [B.11] and recall that [B.13] holds. Since $(\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G}$ is the solution of the linear system [III.29] with $v = \varphi_\mathcal{T}$, we can deduce from Proposition 8 that $\nabla\varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G}$ is the solution of the linear system $\mathcal{A}_G Z_G = \mathcal{C}_G(\varphi)$ where the vector $\mathcal{C}_G(\varphi)$ satisfies [B.12]. We obtain

$$|\nabla\varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G}| \le C_1 |\mathcal{A}_G^{-1}| \max_{K\in\mathcal{T}_G} \mathrm{diam}(K). \qquad [B.16]$$

For all $\sigma \in G \cap \mathcal{E}_{\mathrm{int}}$ with $\mathcal{T}_\sigma = \{K_G, L\}$, thanks to [III.30] with $v = \varphi_\mathcal{T}$, we have

$$(\nabla_\mathcal{D}\varphi_\mathcal{T})^{G,\sigma}_L = (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G} - \frac{R_{L,\sigma}(\varphi_\mathcal{T})}{d_{L,\sigma}}\boldsymbol{n}_{L,\sigma},$$

where $R_{L,\sigma}(\varphi_\mathcal{T}) = \varphi(x_L) - \varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G}\cdot(x_L - x_{K_G})$. Thanks to Proposition 7, we can deduce that

$$\begin{aligned}
\nabla\varphi(x_L) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^{G,\sigma}_L &= \nabla\varphi(x_{K_G}) + \mu_\sigma \boldsymbol{n}_{L,\sigma} - \tau_\sigma \mathbf{t}_\sigma - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G} + \frac{R_{L,\sigma}(\varphi_\mathcal{T})}{d_{L,\sigma}}\boldsymbol{n}_{L,\sigma} \\
&= \nabla\varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G} + \frac{R_{L,\sigma}(\varphi_\mathcal{T}) - W_{K_G}(x_L)}{d_{L,\sigma}}\boldsymbol{n}_{L,\sigma} \\
&\quad + \left(\tau_\sigma \frac{\mathbf{t}_\sigma\cdot(y_\sigma - x_L)}{d_{L,\sigma}} + \frac{W_{K_G}(y_\sigma) - W_L(y_\sigma)}{d_{L,\sigma}}\right)\boldsymbol{n}_{L,\sigma} - \tau_\sigma \mathbf{t}_\sigma \\
&= \nabla\varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G} + (\nabla\varphi(x_{K_G}) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^G_{K_G})\cdot\frac{(x_L - x_{K_G})}{d_{L,\sigma}}\boldsymbol{n}_{L,\sigma} \\
&\quad + \left(\tau_\sigma \frac{\mathbf{t}_\sigma\cdot(y_\sigma - x_L)}{d_{L,\sigma}} + \frac{W_{K_G}(y_\sigma) - W_L(y_\sigma)}{d_{L,\sigma}}\right)\boldsymbol{n}_{L,\sigma} - \tau_\sigma \mathbf{t}_\sigma.
\end{aligned}$$

Using then [B.16], [B.13] and [II.7], we can deduce that there exists a real $C_{19} > 0$ which only depends on ϱ_1, ϱ_2, Λ and φ s.t.

$$|\nabla\varphi(x_L) - (\nabla_\mathcal{D}\varphi_\mathcal{T})^{G,\sigma}_L| \le C_{19}(1 + |\mathcal{A}_G^{-1}|) \max_{K\in\mathcal{T}_G} \mathrm{diam}(K),$$

and the proof is complete.

B.3 Computation of the parameter γ_2, (example 2)

The alternative choice used in the numerical examples of § V.3.2 is designed so as to enhance the coercivity of the method. For each group $G \in \mathcal{G}$, define the space $\mathcal{H}_{\mathcal{T}_G} \stackrel{\text{def}}{=} \{u_K \in \mathbb{R},\ K \in \mathcal{T}_G\}$ endowed with the semi-norm

$$\|u\|^2_{\mathcal{T}_G} \stackrel{\text{def}}{=} \sum_{K\in\mathcal{T}_G} \sum_{\sigma\in\mathcal{E}_K\cap G} \frac{\mathrm{m}\sigma}{d_{K,\sigma}} (\gamma_\sigma u - u_K)^2.$$

For all $u,\ v \in \mathcal{H}_{\mathcal{T}_G}$ set $a_{\mathcal{T}_G}(u,v) = \sum_{K\in\mathcal{T}_G} \sum_{\sigma\in\mathcal{E}_K\cap G} F^G_{K,\sigma}(u)(\gamma_\sigma v - v_K)$. For each $G \in \mathcal{G}$ define

$$\gamma_2 \stackrel{\text{def}}{=} \inf_{\{u\in\mathcal{H}_{\mathcal{T}_G},\ \|u\|_{\mathcal{T}_G}=1\}} a_{\mathcal{T}_G}(u,u)$$

The computation of the parameter γ_2 requires to evaluate the eigenvalues of a local matrix of $\mathbb{R}^{d\times d}$ associated with the bilinear form $a_{\mathcal{T}_G}$, and its cost is negligible. Indeed, by conservativity of the subfluxes,

$$\begin{aligned} a_{\mathcal{T}_G}(u,u) &= \sum_{\sigma\in G} \sum_{K\in\mathcal{T}_\sigma} F^G_{K,\sigma}(u)(\gamma_\sigma u - u_K) \\ &= \sum_{\sigma\in G\cap\mathcal{E}_{\rm int},\mathcal{T}_\sigma=\{K_G,L\}} F^G_{K_G,\sigma}(u)(u_L - u_{K_G}) + \sum_{\sigma\in G\cap\mathcal{E}_{\rm ext},\mathcal{T}_\sigma=\{K_G\}} F^G_{K_G,\sigma}(u)(\gamma_\sigma u - u_{K_G}) \end{aligned}$$

Since for all $\sigma \in \mathcal{E}_{\rm int}$ with $\mathcal{T}_\sigma = \{K_G, L\}$, we have $u_L - u_{K_G} = \frac{d_{K_G,\sigma}+d_{L,\sigma}}{d_{K_G,\sigma}}(\gamma_\sigma u - u_{K_G})$, if we let $d_\sigma = d_{K_G,\sigma} + d_{L,\sigma}$ for all $\sigma \in \mathcal{E}_{\rm int}$ with $\mathcal{T}_\sigma = \{K_G, L\}$ and $d_\sigma = d_{K_G,\sigma}$ for all $\sigma \in \mathcal{E}_{\rm ext}$, then we can deduce that

$$a_{\mathcal{T}_G}(u,u) = \sum_{\sigma\in G} \frac{d_\sigma}{d_{K_G,\sigma}}\ F^G_{K_G,\sigma}(u)(\gamma_\sigma u - u_{K_G}).$$

Now, by [III.29], $F^G_{K_G,\sigma}(u)$ depends linearly on $\{u_L - u_{K_G}\}_{L\in\mathcal{T}_G\setminus\{K_G\}}$ (and u_{K_G} if $\sigma \in \mathcal{E}_{\rm ext}$), and it can therefore be written as

$$F^G_{K_G,\sigma}(u) = \sum_{\sigma'\in G} a^G_{\sigma,\sigma'} \frac{d_{\sigma'}}{d_{K_G,\sigma'}} (\gamma_{\sigma'} u - u_{K_G})$$

where $\{a^G_{\sigma,\sigma'}\}_{\sigma,\sigma'\in G\times G}$ is a family of reals. We obtain that

$$a_{\mathcal{T}_G}(u,u) \ = \sum_{(\sigma,\sigma')\ \in G\times G} \frac{d_\sigma}{d_{K_G,\sigma}} \frac{d_{\sigma'}}{d_{K_G,\sigma'}}\ a^G_{\sigma,\sigma'} (\gamma_{\sigma'} u - u_{K_G})(\gamma_\sigma u - u_{K_G}).$$

We denote by $X^G(u)$ the vector of size d defined by the family $\left\{\frac{\sqrt{d_\sigma \mathrm{m}\sigma}}{d_{K_G,\sigma}}(\gamma_\sigma u - u_{K_G})\right\}_{\sigma\in G}$ and by A^G the matrix of size d defined by the family of reals $\left\{\sqrt{\frac{d_\sigma d_{\sigma'}}{\mathrm{m}\sigma \mathrm{m}\sigma'}} a^G_{\sigma,\sigma'}\right\}_{\sigma,\sigma'\in G\times G}$.

Then, we can write $a_{\mathcal{T}_G}(u,u)$ under the form

$$a_{\mathcal{T}_G}(u,u) = (A^G X^G(u)) \cdot X^G(u)$$

or again,

$$a_{\mathcal{T}_G}(u,u) = \left(\frac{A^G + (A^G)^t}{2} X^G(u) \right) \cdot X^G(u) \qquad \text{[B.17]}$$

where $(A^G)^t$ is the transpose matrix of A^G. From (B.17), we deduce that γ_2 is the smallest eigenvalue of the matrix $\frac{A^G+(A^G)^t}{2}$ because the Euclidean norm of the vector $X^G(u)$ is exactly equal to $\|u\|_{\mathcal{T}_G}$. For a given $\epsilon > 0$, let

$$\begin{cases} g_\epsilon(x) = \frac{\epsilon^2}{\epsilon - x} & \text{if } x < 0, \\ g_\epsilon(x) = x + \epsilon & \text{otherwise,} \end{cases}$$

and, for all $G \in \mathcal{G}$, define $\beta^G = g_\epsilon(\gamma_2)$. The weights are defined as

$$\theta_\sigma^G = \frac{\beta^G}{\sum_{G' \in \mathcal{G}_\sigma} \beta^{G'}} \quad \forall G \in \mathcal{G},\ \forall \sigma \in G.$$

Therefore, for a given $G \in \mathcal{G}$, the larger γ_2, the more the subfluxes $\{F^G_{K,\sigma}\}_{K \in \mathcal{T}_G,\ \sigma \in \mathcal{E}_K \cap G}$ will contribute to the global fluxes $\{F_{K,\sigma}\}_{K \in \mathcal{T},\ \sigma \in \mathcal{E}_K \cap G}$. In the numerical tests of § V.3.2 we have taken $\epsilon = 0.1$.

Annexe C

Synthèse entre méthode Galerkin discontinu et volumes finis

Résumé

Dans ce chapitre, nous proposons un cadre unifié d'analyse regroupant un éventail de méthodes de discrétisations non conformes pour des opérateurs de diffusion hétérogènes anisotropiques sur maillages généraux. L'analyse repose sur deux outils d'analyse discrète d' espaces de fonctions polynômiaux par morceaux, à savoir une inégalité discrète de Sobolev-Poincaré et un théorème discret de Rellich.

Les conditions de convergence sont regroupées en sept hypothèses, chacune d'elle caractérisant un ingrédient essentiel de l'analyse. On montre dans cette partie que les schémas volumes finis aussi bien que les méthodes de Galerkin Discontinu les plus communes s'inscrivent dans cet analyse. Un nouveau schéma volume fini centré est également présenté.

Preliminary Several methods have been developed through the years to solve the single phase Darcy equation, often of non-conforming type. A crucial ingredient is a robust discretization of heterogeneous anisotropic diffusion operators. Indeed, strong anisotropy and heterogeneity are usually present in problems of practical interest, thus demanding an approach robust with respect to both. Moreover, even for simple domains, the low regularity of the diffusion coefficient may affect the regularity of the solution itself. It is thus important for a discretization method to ensure convergence to minimal regularity solutions, *i.e.* solutions belonging to the natural function spaces in which the weak formulation of the PDE is set. Furthermore, it is often desirable to handle general nonconforming meshes, both because end-users may have little or no control over the mesh and because local grid refinement could be required.

In this work, we have proposed a unified analysis framework encompassing a

wide range of non-conforming methods which respond to the above requirements. In particular, both Finite Volume (FV) and discontinuous Galerkin (dG) methods will be shown to fit in the framework. Although the analogies between these two families of discretization methods have often been highlighted, the present unified analysis is, to the best of our knowledge, new.

Finite Volume methods have been widely employed in industrial applications because of simplicity of implementation, closeness to physical intuition and reduced computational cost. In recent years, these methods have known an impetuous development thanks to both empirical and theoretical works. In particular, the convergence analysis of FV methods has been dealt with by Eymard, Gallouët, Herbin and co-authors (see e.g. [42, 45]), who have derived new discrete functional analysis tools allowing to prove the convergence to minimum regularity solutions. The discrete analysis framework above has been used for a variety of FV methods applied to linear or non-linear problems (see e.g. [15, 47]). Within the framework of Mimetic Finite Difference approximations, reduced-cost methods on general meshes have also been developed. These methods rely on different discrete analysis tools than the ones used here, and we refer to [24–26] for a unified analysis.

Discontinuous Galerkin methods were introduced over thirty years ago to approximate hyperbolic and elliptic PDEs (see e.g. [17, 38] for a historical perspective), and they have received extensive attention over the last decade. Up to now, convergence analysis has relied on classical Finite Element tools, yielding asimptotical order estimates but requiring regularity assumptions on the exact solution (see e.g. [17, 33, 38–40]). In a recent work [32], Di Pietro and Ern have extended the discrete analysis tools presented in [45] to piecewise polynomial function spaces on general meshes. By means of such tools, the convergence analysis of dG discretization of both linear and non-linear problems can be performed in the spirit of [45].

In this work we further extend the above results by proposing an abstract set of properties ensuring the convergence of a discretization method to minimal regularity solutions. The analysis framework proposed relies on the discrete functional analysis results of [32, 45], where the authors introduce discrete $W^{1,p}$ norms which satisfy discrete Sobolev inequalities and deduce a compactness result for bounded sequences in such norms using the Kolmogorov criterion (see, e.g., [23, Theorem IV.25]). In order to use the compactness results for sequences in piecewise polynomial spaces, we shall assume that, whatever the vector space V_h in which the solution is sought, a reconstruction operator on a suitable piecewise polynomial space is available. The key ideas of the analysis can be summarized as follows:

(i) V_h, is equipped with a norm $\|\cdot\|_{V_h}$ which, for all $v_h \in V_h$, controls the discrete H^1 norm of the piecewise polynomial reconstruction of v_h. As a consequence, bounded sequences in the $\|\cdot\|_{V_h}$ norm yield bounded sequences in the discrete H^1 norm;

(ii) an *a priori* estimate on the discrete solution is derived allowing to infer the strong convergence of a subsequence of (reconstruction of) discrete solutions to a function $u \in L^2(\Omega)$;

(iii) the construction of a discrete gradient weakly converging to ∇u in $[L^2(\Omega)]^d$ allows to prove that the limit u actually belongs to $H_0^1(\Omega)$;

(iv) the convergence of the scheme is finally proved testing against a discrete projection of a smooth function belonging to some convenient dense subspace, say $C_c^\infty(\Omega)$.

Since the exact solution is unique, the convergence of the whole sequence of discrete approximations is deduced. Moreover, stronger convergence results on the discrete gradient can be derived using the dissipative structure of the problem for both symmetric and non-symmetric schemes.

Besides providing a means to analyze existing methods and to develop new ones, the above framework ensures the convergence of arbitrary compositions of compliant methods. This can be particularly useful when one wishes to use a more accurate but expensive methods on a selected region of the domain along with a less accurate but faster method elsewhere.

This chapter is organized as follows. §C.1 introduces the abstract framework, including the assumptions on the mesh family as well as the properties required to prove convergence of a method. The latter are grouped into seven Hypotheses, each of them characterizing one salient ingredient of the analysis. The main result is Theorem C.1.2. §C.2 show some examples of methods which fit in the abstract analysis framework. In particular § III.2.1 presents a selection of dG methods robust with respect to the heterogeneity and anisotropy of the diffusion tensor; § III.2.2 deals with a new cell-centered finite volume method; § III.2.3 investigates a hybrid FV method using both cell- and face-unknowns. For all the methods, a precise definition possibly including further assumptions on the mesh family is followed by the verification of Hypotheses 10-16.

C.1 Abstract analysis framework

III.1.1 Model problem and setting

Let $\Omega \subset \mathbb{R}^d$, $\mathbb{N} \ni d \geq 1$, be a bounded polygonal domain and consider the following model problem:

$$\begin{cases} -\nabla\cdot(\nu\nabla u) = f, & \text{in } \Omega, \\ u = 0, & \text{on } \partial\Omega, \end{cases} \qquad \text{[C.1]}$$

where $\nu \in [L^\infty(\Omega)]^{d\times d}$ is s.t. (such that), for a.e. (almost every) $x \in \Omega$, $\nu(x)$ is symmetric and its spectrum $\{\lambda_i(x)\}_{i=1}^d$ is s.t. $0 < \underline{\lambda} \leq \lambda_i(x) \leq \overline{\lambda} < \infty$. In weak

formulation, problem [C.1] reads: Find $u \in H_0^1(\Omega)$ s.t.

$$a(u,v) = (f,v)_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega), \tag{C.2}$$

where $\mathcal{L}(H_0^1(\Omega) \times H_0^1(\Omega); \mathbb{R}) \ni a(u,v) \stackrel{\text{def}}{=} (\nu \nabla u, \nabla v)_{[L^2(\Omega)]^d}$. The well-posedness of problem [C.2] is classical.

Remark 7 *The analysis can be easily extended to $f \in L^r(\Omega)$ with $r \geq \frac{2d}{d+2}$ if $d \geq 3$ and $r > 1$ if $d = 2$; see [32] for the details in the case of dG methods. This requires more general Sobolev inequalities than the one of Hypothesis 21, which are proved in [32, 45]. Also, different boundary conditions can be handled with minor modifications, but we have decided to stick to the homogeneous Dirichlet problem for clarity of presentation.*

The following definition characterizes an admissible mesh family:

Définition C.1.1 (Admissible mesh family) *Let $\mathcal{H}$ be a countable set. The mesh family $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$, is said to be admissible if the following assumptions are satisfied for all $h \in \mathcal{H}$:*

(i) $\mathcal{T}_h$ is a finite family of non-empty connex (possibly non-convex) open disjoint sets T forming a partition of Ω and whose boundaries are a finite union of parts of hyperplanes. The d-dimensional Lebesgue measure and the diameter of the generic element $T \in \mathcal{T}_h$ will be denoted by $|T|$ and h_T respectively. The representative linear dimension of the discretization will be defined as $h \stackrel{\text{def}}{=} \max_{T\in\mathcal{T}_h} h_T$;

(ii) each $T \in \mathcal{T}_h$ is affine-equivalent to an element of a finite collection of reference elements;

(iii) there is a parameter N_∂ independent of h s.t., for all $h \in \mathcal{H}$, each $T \in \mathcal{T}_h$ has at most N_∂ faces. For all elements $T \in \mathcal{T}_h$, let $\mathcal{F}_h^T$ denote the set of faces of T. A set $F \in \mathcal{F}_h^T$ of non-zero $(d-1)$-dimensional Lebesgue measure $|F|$ is said to be a face of T if F is part of a hyperplane and if either F is located on the boundary of Ω (boundary face) or there is one and only one $T' \in \mathcal{T}_h$ s.t. $F = \mathcal{F}_h^T \cap \mathcal{F}_h^{T'}$ (interface). The diameter of the generic face $F \in \mathcal{F}_h$ will be denoted by h_F;

(iv) there is a parameter ϱ_1 independent of h s.t., for all $T \in \mathcal{T}_h$,

$$\sum_{F\in\mathcal{F}_h^T} h_F|F| \leq \varrho_1|T|;$$

The set of boundary faces will be denoted by $\mathcal{F}_h^b$, whereas the interfaces will be collected into the set $\mathcal{F}_h^i$. For every $F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}$ we let μ_F denote the outward normal to T_1; for all $T \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_h^T$, μ_F^T will denote the outward normal to T. For every $F \in \mathcal{F}_h^T \cap \mathcal{F}_h^b$, both μ_F and μ_F^T will denote the outward normal to Ω. Further assumptions on the mesh family may be required depending on the method considered, and will be specified in the corresponding section.

Remark 8 *According to Definition C.1.1, (i) the mesh elements are not supposed to be convex, and the mesh may possibly be nonconforming; (ii) in three space dimensions, general hexahedra can be treated by decomposing non-plane faces in a fixed number of plane sub-faces.*

Let $\mathcal{T}_h$ denote an element of an admissible mesh family and let $\mathcal{S}_h$ denote a submesh of $\mathcal{T}_h$ depending on the method at hand. We introduce the space of piecewise polynomial functions of total degree less than or equal to $\mathbb{N} \ni k \geq 0$,

$$P_h^k(\mathcal{X}_h) \stackrel{\text{def}}{=} \{v_h \in L^2(\Omega);\ v_{h|T} \in \mathbb{P}^k(T),\ \forall T \in \mathcal{X}_h\}, \quad \mathcal{X}_h \in \{\mathcal{T}_h, \mathcal{S}_h\}.$$

The symbols V_h and Σ_h denote two vector spaces associated with $\mathcal{T}_h$ and $\mathcal{S}_h$ respectively. We assume that $\Sigma_h = [P_h^{k_\Sigma}(\mathcal{S}_h)]^d$ for a fixed $\mathbb{N} \ni k_\Sigma \geq 0$ depending on the method considered. Also, in what follows, $r_h^V : V_h \to P_h^{k_V}(\mathcal{T}_h)$ will denote a reconstruction operator onto the piecewise polynomial space of degree k_V depending on the method at hand (see Hypothesis 10). In particular, for FV methods, $k_V = k_\Sigma = 0$ whereas $k_V \geq k_\Sigma \geq 0$, $k_V \geq 1$ for dG methods.

The symbols $\lesssim$ and $\gtrsim$ will be used in the present section for inequalities that hold up to a positive parameter independent of the mesh size h but possibly depending on the regularity parameters of the mesh family, on ν, k_V and k_Σ. More detailed expressions for these multiplicative constant will be given for each method in §C.2.

Hypothesis 10 (Piecewise polynomial reconstruction r_h^V) *For a fixed $\mathbb{N} \ni k_V \geq 0$ depending on the actual discretization method, there is a reconstruction operator $r_h^V : V_h \to P_h^{k_V}(\mathcal{T}_h)$ which maps every element $v_h \in V_h$ onto a piecewise polynomial function $r_h^V v_h \in P_h^{k_V}(\mathcal{T}_h)$.*

We define the following bilinear form

$$\mathcal{L}(V_h \times V_h; \mathbb{R}) \ni a_h(u_h, v_h) \stackrel{\text{def}}{=} (\nu G(u_h), \widetilde{G}(v_h))_{[L^2(\Omega)]^d} + j_h(u_h, v_h), \qquad [C.3]$$

where $G \in \mathcal{L}(V_h; \Sigma_h)$ and $\widetilde{G} \in \mathcal{L}(V_h; \Sigma_h)$ are linear gradient reconstructions whose properties will be detailed in Hypotheses 12, 13 and 16, whereas $j_h \in \mathcal{L}(V_h \times V_h; \mathbb{R})$ is a bilinear form meant to ensure the coercivity of a_h. We focus on the following family of approximations for problem [C.2]: Find $u_h \in V_h$ s.t.

$$a_h(u_h, v_h) = (f, r_h^V v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h. \qquad [C.4]$$

III.1.2 Discrete Rellich theorem

The piecewise polynomial space $P_h^{k_V}(\mathcal{T}_h)$, $k_V \geq 0$, must be equipped with a discrete H^1 norm $\|\cdot\|_{1,2,h}$ s.t. the following hypothesis is satisfied:

Hypothesis 11 (Compactness) *Let $\{p_h\}_{h\in\mathcal{H}}$ be a sequence in $P_h^{k_V}(\mathcal{T}_h)$, $k_V \geq 0$, bounded in the corresponding $\|\cdot\|_{1,2,h}$ norm. Then, the family $\{p_h\}_{h\in\mathcal{H}}$ is relatively compact in $L^2(\Omega)$ (and also in $L^2(\mathbb{R}^d)$ taking $p_h = 0$ outside Ω).*

Norms satisfying Hypothesis 11 will be defined in eqs. [C.20] and [C.26] below.

Lemma 21 (Discrete Sobolev-Poincaré inequality) *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a mesh family compliant with Definition C.1.1 and let us suppose that Hypothesis 11 holds. Then, for all $p_h \in P_h^{k_V}$, $k_V \geq 0$,*

$$\|p_h\|_{L^2(\Omega)} \lesssim \|p_h\|_{1,2,h}. \tag{C.5}$$

Proof. For the sake of simplicity, let $\mathcal{H} = \mathbb{N}$ and $h_n \to 0$ as $n \to \infty$. We proceed by contradiction. Let us admit that, for all $C > 0$, there is $n \in \mathbb{N}$ and $p_{h_n} \in \mathcal{T}_{h_n}$ s.t. $\|p_{h_n}\|_{L^2(\Omega)} > C\|p_{h_n}\|_{1,2,h_n}$. In particular, we can take $C = n$ and set $\widetilde{p}_{h_n} \stackrel{\text{def}}{=} p_{h_n}/\|p_{h_n}\|_{1,2,h}$, so that

$$\|\widetilde{p}_{h_n}\|_{L^2(\Omega)} > n, \quad \|\widetilde{p}_{h_n}\|_{1,2,h_n} = 1. \tag{C.6}$$

As n increases, the L^2 norm of $\widetilde{p}_{h_n}$ increases, whereas its $\|\cdot\|_{1,2,h}$ norm remains bounded. According to Hypothesis 11, $\{\widetilde{p}_{h_n}\}_{n\in\mathbb{N}}$ is thus relatively compact in $L^2(\Omega)$, and we can extract a subsequence $\{\widetilde{p}_{h_{\varphi(n)}}\}_{n\in\mathbb{N}}$ which converges to some $\overline{p}$ in $L^2(\Omega)$. As a consequence, $\|\widetilde{p}_{\varphi(n)}\|_{L^2(\Omega)} \to \|\overline{p}\|_{L^2(\Omega)}$ as $n \to \infty$, which is in contradiction with [C.6]. □ A direct proof of the Sobolev-Poincaré inequality on broken Sobolev spaces has been given in [16, 22, 45]; broken Sobolev embeddings have been derived by Lasis and Süli [56, 57] in the Hilbertian case; broken Sobolev embeddings in the non-Hilbertian case have been recently presented in [32]

Hypothesis 12 ($\|\cdot\|_{V_h}$ norm) *The vector space V_h is equipped with an inner product norm $\|\cdot\|_{V_h}$ s.t., for all $v_h \in V_h$,*

$$\|r_h^V v_h\|_{1,2,h} \lesssim \|v_h\|_{V_h}, \tag{C.7}$$

$$\|G(v_h)\|_{[L^2(\Omega)]^d} + \|\widetilde{G}(v_h)\|_{[L^2(\Omega)]^d} \lesssim \|v_h\|_{V_h}. \tag{C.8}$$

Inequality [C.7] will be used to derive an estimate for the piecewise polynomial reconstruction of the solution in terms of the discrete H^1 norm $\|\cdot\|_{1,2,h}$. This will, in turn, ensure the boundedness of the sequence of the reconstructed discrete solutions of [C.4] on the mesh family $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$, a key ingredient to infer a compactness result. Inequality [C.8] states that bounded sequences in the $\|\cdot\|_{V_h}$ norm yield bounded sequences of gradient approximations in the L^2 norm.

Hypothesis 13 (Weak convergence of $\widetilde{G}$) *Let $\{v_h\}_{h\in\mathcal{H}}$, be a sequence in V_h s.t. $\{r_h^V v_h\}_{h\in\mathcal{H}}$ converges to $v \in L^2(\Omega)$ in $L^2(\mathbb{R}^d)$ (prolonging $r_h^V v_h$ to zero outside Ω) and $\{\widetilde{G}(v_h)\}_{h\in\mathcal{H}}$ is bounded in the $[L^2(\mathbb{R}^d)]^d$ norm. Then, for all $\Phi \in [C_c^\infty(\mathbb{R}^d)]^d$,*

$$\lim_{h\to 0} \int_{\mathbb{R}^d} \widetilde{G}(v_h)\cdot\Phi = -\int_{\mathbb{R}^d} v\nabla\cdot\Phi.$$

Disposing of a weakly converging gradient allows to prove the following result concerning the regularity of the limit of a converging sequence in V_h:

Théorème C.1.1 (Discrete Rellich theorem) *Let $\{v_h\}_{h\in\mathcal{H}}$ be a sequence in V_h bounded in the $\|\cdot\|_{V_h}$ norm. Then, (i) $\{r_h^V v_h\}_{h\in\mathcal{H}}$ is relatively compact in $L^2(\Omega)$; (ii) if $r_h^V v_h \to v$ in $L^2(\Omega)$ as $h\to 0$, then $v\in H_0^1(\Omega)$.*

Proof. Owing to the assumptions of the theorem together with [C.7], there is $C\in\mathbb{R}_+$ s.t.

$$\|r_h^V v_h\|_{1,2,h} \le \|v_h\|_{V_h} \le C, \quad \forall h\in\mathcal{H}.$$

As a consequence, the sequence $\{r_h^V v_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{1,2,h}$ norm. Owing to Hypothesis 11, it is possible to extract a subsequence converging to some v in $L^2(\Omega)$ and also in $L^2(\mathbb{R}^d)$ provided we prolong $r_h^V v_h$ by zero outside Ω. Moreover, [C.8] yields, for all $h\in\mathcal{H}$,

$$\|\widetilde{G}(v_h)\|_{[L^2(\Omega)]^d} \lesssim \|v_h\|_{V_h} \le C.$$

We thus conclude that there exists a $\tau\in[L^2(\Omega)]^d$ to which the sequence $\{\widetilde{G}(v_h)\}_{h\in\mathcal{H}}$ converges in $[L^2(\Omega)]^d$ and also in $[L^2(\mathbb{R}^d)]^d$. On the other hand, the sequence $\{r_h^V v_h\}_{h\in\mathcal{H}}$ satisfies the assumptions of Hypothesis 13, so that $\tau=\nabla v$, which concludes the proof. □

III.1.3 Estimate on the solution

Let $\pi_h^V : C^0(\overline{\Omega})\to V_h$ denote an interpolator onto V_h whose properties will be detailed in Hypotheses 14 and 16. In what follows, π_h^V will be applied to functions of $C_c^\infty(\Omega)$, which is used as a pivot space.

Hypothesis 14 (Stabilization j_h) *The bilinear form j_h is symmetric, positive semidefinite and continuous with respect to the $\|\cdot\|_{V_h}$ norm, i.e. ,*

$$j_h(u_h,v_h) \lesssim \|u_h\|_{V_h}\|v_h\|_{V_h}, \quad \forall (u_h,v_h)\in[V_h]^2. \tag{C.9}$$

Furthermore, the following consistency property holds:

$$\lim_{h\to 0} j_h(\pi_h^V\varphi,\pi_h^V\varphi)=0, \quad \forall\varphi\in C_c^\infty(\Omega). \tag{C.10}$$

The following Cauchy-Schwarz type inequality is an immediate consequence of Hypothesis 14:

$$|j_h(u_h,v_h)| \lesssim [j_h(u_h,u_h)]^{1/2}\,[j_h(v_h,v_h)]^{1/2}. \tag{C.11}$$

Hypothesis 15 (Coercivity of a_h) *For all $v_h\in V_h$, $a_h(v_h,v_h)\gtrsim\|v_h\|_{V_h}^2$.*

The coercivity of the bilinear form a_h is an essential ingredient of the analysis, since it allows to obtain an estimate of the solution for use in the discrete Rellich Theorem C.1.1.

Lemma 22 (Well-posedness) *Problem* [C.4] *is well-posed. Furthermore, its solution satisfies the following* a priori *estimates:*

$$\|r_h^V u_h\|_{1,2,h} \lesssim \|u_h\|_{V_h} \lesssim \|f\|_{L^2(\Omega)}. \tag{C.12}$$

Proof. (i) To prove the well-posedness we use the Lax-Milgram lemma. Using [C.8] together with [C.9] we have, for all $(u_h, v_h) \in [V_h]^2$,

$$a_h(u_h, v_h) \lesssim \overline{\lambda}\|u_h\|_{V_h}\|v_h\|_{V_h} + \|u_h\|_{V_h}\|v_h\|_{V_h} \lesssim \|u_h\|_{V_h}\|v_h\|_{V_h},$$

i.e. , the bilinear form a_h is continuous in V_h. Cauchy-Schwarz inequality together with [C.5] and [C.7] yield, for all $v_h \in V_h$,

$$(f, r_h^V v_h)_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}\|r_h^V v_h\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}\|r_h^V v_h\|_{1,2,h} \leq \|f\|_{L^2(\Omega)}\|v_h\|_{V_h}.$$

We conclude using Hypothesis 15. (ii) If u_h is the null element of V_h, the estimate is trivially verified. If this is not the case, Hypothesis 15 together with Cauchy-Schwarz inequality, [C.5] and [C.7] yield

$$\|r_h^V u_h\|_{1,2,h}\|u_h\|_{V_h} \leq \|u_h\|_{V_h}^2 \lesssim a_h(u_h, u_h) \lesssim \|f\|_{L^2(\Omega)}\|u_h\|_{L^2(\Omega)} \lesssim \|f\|_{L^2(\Omega)}\|u_h\|_{V_h},$$

thus concluding the proof. □

III.1.4 Convergence

Hypothesis 16 (Consistency) *The following results hold:*

$$\|\pi_h^V \varphi\|_{V_h} \lesssim \sigma_\varphi, \quad \forall \varphi \in C_c^\infty(\Omega), \tag{C.13}$$

$$\lim_{h\to 0} \|(r_h^V \circ \pi_h^V)\varphi - \varphi\|_{L^2(\Omega)} = 0, \quad \forall \varphi \in C_c^\infty(\Omega), \tag{C.14}$$

$$\lim_{h\to 0} \|\nabla\varphi - G(\pi_h^V \varphi)\|_{[L^2(\Omega)]^d} = 0, \quad \forall \varphi \in C_c^\infty(\Omega), \tag{C.15}$$

where $\sigma_\varphi > 0$ is a parameter depending only on φ and on the mesh regularity parameters.

The above assumptions ensure that we can consistently approximate smooth functions and their gradients on the discrete spaces at hand. The consistency of G stated in [C.15] allows to prove the following

Lemma 23 (Convergence of G) *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of admissible meshes. Let u_h denote the unique solution of the discrete problem* [C.4] *on $\mathcal{T}_h$. Then, (i) there exists $\widetilde{u} \in H_0^1(\Omega)$ and a subsequence $\{r_h^V u_h\}_{h\in\mathcal{H}}$ converging to $\widetilde{u}$ in $L^2(\Omega)$ as $h \to 0$; (ii) $\{G(u_h)\}_{h\in\mathcal{H}}$ converges to $\nabla\widetilde{u}$ in $[L^2(\Omega)]^d$.*

PROOF. (i) Thanks to [C.12], the sequence $\{r_h^V u_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{1,2,h}$ norm. According to Theorem C.1.1, there is a subsequence of $\{r_h^V u_h\}_{h\in\mathcal{H}}$ (still denoted with the same symbol) and an element $\widetilde{u}\in H_0^1(\Omega)$ s.t. $\{r_h^V u_h\}_{h\in\mathcal{H}}$ converges to $\widetilde{u}$ in $L^2(\Omega)$ as $h\to 0$. (ii) Let $\varphi\in C_c^\infty$ and set $\varphi_h \stackrel{\text{def}}{=} \pi_h^V\varphi$. We have that

$$\|G(u_h)-\nabla\widetilde{u}\|_{[L^2(\Omega)]^d}^2 \le 3\left[\|G(u_h)-G(\varphi_h)\|_{[L^2(\Omega)]^d}^2+\|G(\varphi_h)-\nabla\varphi\|_{[L^2(\Omega)]^d}^2+\|\nabla\varphi-\nabla\widetilde{u}\|_{[L^2(\Omega)]^d}^2\right].$$

Let S_i, $i\in\{1\dots 3\}$ denote the terms in the right hand side. Thanks to Hypothesis 15 and to the linearity of a_h we have that

$$S_1 \lesssim a_h(u_h,u_h)-a_h(u_h,\varphi_h)-a_h(\varphi_h,u_h)+a_h(\varphi_h,\varphi_h).$$

Owing to [C.4], $a_h(u_h,u_h)=(f,r_h^V u_h)_{L^2(\Omega)}$ and $a_h(u_h,\varphi_h)=(f,r_h^V\varphi_h)_{L^2(\Omega)}$. As a consequence,

$$\lim_{h\to 0} a_h(u_h,u_h)=(f,\widetilde{u})_{L^2(\Omega)}.$$

Furthermore, using [C.14], we conclude that

$$0\le\limsup_{h\to 0}\left|a_h(u_h,\varphi_h)-(f,\varphi)_{L^2(\Omega)}\right|\le\limsup_{h\to 0}\|f\|_{L^2(\Omega)}\|r_h^V\varphi_h-\varphi\|_{L^2(\Omega)}=0,$$

that is, gathering the above results,

$$\forall\varphi\in C_c^\infty(\Omega),\quad \lim_{h\to 0}\left[a_h(u_h,u_h)-a_h(u_h,\varphi_h)\right]=(f,\widetilde{u}-\varphi)_{L^2(\Omega)}. \qquad [C.16]$$

To estimate the remaining terms, observe that

$$\begin{aligned}a_h(\varphi_h,\varphi_h)-a_h(\varphi_h,u_h) &= (\nu\nabla\varphi,\widetilde{G}(\varphi_h-u_h))_{[L^2(\Omega)]^d}\\ &\quad+(\nu(G(\varphi_h)-\nabla\varphi),\widetilde{G}(\varphi_h-u_h))_{[L^2(\Omega)]^d}+j_h(\varphi_h,\varphi_h-u_h).\end{aligned}$$

Owing to Hypothesis 13, the term in the first line tends to $(\nu\nabla\varphi,\nabla(\varphi-\widetilde{u}))_{[L^2(\Omega)]^d}$ as $h\to 0$. The term in the second line can be estimated as follows:

$$\begin{aligned}\left|(\nu(G(\varphi_h)-\nabla\varphi),\widetilde{G}(\varphi_h-u_h))_{[L^2(\Omega)]^d}\right| &\le \overline{\lambda}\|G(\varphi_h)-\nabla\varphi\|_{[L^2(\Omega)]^d}\|\widetilde{G}(\varphi_h-u_h)\|_{[L^2(\Omega)]^d}\\ &\lesssim \|G(\varphi_h)-\nabla\varphi\|_{[L^2(\Omega)]^d}\left(\|\varphi_h\|_{V_h}+\|u_h\|_{V_h}\right)\\ &\lesssim \|G(\varphi_h)-\nabla\varphi\|_{[L^2(\Omega)]^d}\left(\sigma_\varphi+\|u_h\|_{V_h}\right),\end{aligned}$$

where we have used Cauchy-Schwarz inequality followed by [C.8], [C.13] and [C.12]. Since $\|u_h\|_{V_h}$ is bounded, the right hand side of the above inequality tends to zero as $h\to 0$. On the other hand, [C.11], [C.13] and [C.12] yield

$$\begin{aligned}|j_h(\varphi_h,\varphi_h-u_h)| &\le [j_h(\varphi_h,\varphi_h)]^{1/2}\,[j_h(\varphi_h-u_h,\varphi_h-u_h)]^{1/2}\\ &\lesssim [j_h(\varphi_h,\varphi_h)]^{1/2}\left(\|\varphi_h\|_{V_h}+\|u_h\|_{V_h}\right)\\ &\lesssim [j_h(\varphi_h,\varphi_h)]^{1/2}\left(\sigma_\varphi+\|u_h\|_{V_h}\right),\end{aligned}$$

which, owing to [C.10] and to the boundedness of $\|u_h\|_{V_h}$, tends to zero as $h \to 0$. In conclusion,

$$\forall \varphi \in C_c^\infty(\Omega), \quad \lim_{h\to 0} [a_h(\varphi_h, \varphi_h) - a_h(\varphi_h, u_h)] = (\nu \nabla \varphi, \nabla(\varphi - \widetilde{u}))_{[L^2(\Omega)]^d}. \quad \text{[C.17]}$$

Equations [C.16] and [C.17] yield

$$\forall \varphi \in C_c^\infty(\Omega), \quad \lim_{h\to 0} S_1 = (\nu \nabla \varphi, \nabla(\varphi - \widetilde{u}))_{[L^2(\Omega)]^d} + (f, \widetilde{u} - \varphi)_{L^2(\Omega)}.$$

Using [C.15] we immediately conclude that, for all $\varphi \in C_c^\infty(\Omega)$, $\lim_{h\to 0} S_2 = 0$. Gathering the above results, for all $\varphi \in C_c^\infty(\Omega)$,

$$\limsup_{h\to 0} \|G(u_h) - \nabla \widetilde{u}\|^2_{[L^2(\Omega)]^d} \lesssim (\nu \nabla \varphi, \nabla(\varphi - \widetilde{u}))_{[L^2(\Omega)]^d} + (f, \widetilde{u} - \varphi)_{L^2(\Omega)} + \|\nabla \varphi - \nabla \widetilde{u}\|^2_{[L^2(\Omega)]^d}.$$

Let now $\{\varphi_m\}_{m\in\mathbb{N}}$ be a sequence converging to $\widetilde{u}$ in $H_0^1(\Omega)$ (the existence of such a sequence follows from the density of $C_c^\infty(\Omega)$ in $H_0^1(\Omega)$). Using the above bound, we conclude that

$$0 \le \liminf_{h\to 0} \|G(u_h) - \nabla \widetilde{u}\|^2_{[L^2(\Omega)]^d} \le \limsup_{h\to 0} \|G(u_h) - \nabla \widetilde{u}\|^2_{[L^2(\Omega)]^d} \le 0,$$

which proves the assert. □

Remark 9 *Observe that the passages to the limit for $h \to 0$ and for $m \to \infty$ cannot be exchanged in the proof. Indeed, the estimates from which* [C.14] *and* [C.15] *are obtained may depend on some norm of φ which does not remain bounded as $m \to \infty$, e.g. the H^2 norm.*

Théorème C.1.2 (Convergence of the method) *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of admissible meshes. Let u_h denote the unique solution of the discrete problem* [C.4] *on $\mathcal{T}_h$. Then, (i) the sequence $\{r_h^V u_h\}_{h\in\mathcal{H}}$ converges to the solution of* [C.2]*, say u, in $L^2(\Omega)$ as $h \to 0$; (ii) the sequence $\{G(u_h)\}_{h\in\mathcal{H}}$ converges to ∇u in $[L^2(\Omega)]^d$.*

Proof. Thanks to [C.12], the sequence $\{u_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{V_h}$ norm. Theorem C.1.1 states that we can extract a subsequence still denoted by $\{r_h^V u_h\}_{h\in\mathcal{H}}$ which converges to an element $\widetilde{u} \in H_0^1(\Omega)$ in $L^2(\Omega)$. Let us focus on the above subsequence. According to Lemma 23, $\{G(u_h)\}_{h\in\mathcal{H}}$ converges to $\nabla\widetilde{u}$ in $[L^2(\Omega)]^d$. In order to prove the convergence of the method, we have to prove that $\widetilde{u}$ solves [C.2]. Let, now, $\varphi \in C_c^\infty(\Omega)$ and set $\varphi_h \stackrel{\text{def}}{=} \pi_h^V \varphi$. We have that

$$a_h(u_h, \varphi_h) = (\nu G(u_h), \widetilde{G}(\varphi_h))_{[L^2(\Omega)]^d} + j_h(u_h, \varphi_h).$$

Using Hypothesis 13 together with Lemma 23 we conclude that

$$\forall \varphi \in C_c^\infty, \quad \lim_{h\to 0} (\nu G(u_h), \widetilde{G}(\varphi_h))_{[L^2(\Omega)]^d} = (\nu \nabla \widetilde{u}, \nabla \varphi)_{[L^2(\Omega)]^d} = a(\widetilde{u}, \varphi).$$

On the other hand, [C.11] together with [C.12] yield

$$|j_h(u_h,\varphi_h)| \le j_h(\varphi_h,\varphi_h)^{1/2} j_h(u_h,u_h)^{1/2} \le j_h(\varphi_h,\varphi_h)^{1/2}\|u_h\|_{V_h} \lesssim j_h(\varphi_h,\varphi_h)^{1/2}\|f\|_{L^2(\Omega)},$$

which tends to 0 as $h \to 0$ by virtue of [C.10]. Moreover,

$$(f, r_h^V\varphi_h)_{L^2(\Omega)} = (f,\varphi)_{L^2(\Omega)} + (f, \varphi - r_h^V\varphi_h)_{L^2(\Omega)},$$

and, using [C.14],

$$0 \le \limsup_{h\to 0} \left|(f, \varphi - r_h^V\varphi_h)_{L^2(\Omega)}\right| \lesssim \limsup_{h\to 0} \|f\|_{L^2(\Omega)}\|\varphi - r_h^V\varphi_h\|_{L^2(\Omega)} = 0,$$

so that, for all $\varphi \in C_c^\infty(\Omega)$, $(f, r_h^V\varphi_h)_{L^2(\Omega)} \to (f,\varphi)_{L^2(\Omega)}$ as $h \to 0$. Thanks to the above results, and since the u_h are solutions of the discrete problem [C.4], we have that

$$a(\widetilde{u},\varphi) = (f,\varphi)_{L^2(\Omega)}, \quad \forall\varphi \in C_c^\infty(\Omega).$$

Since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$, $\widetilde{u} = u$ for a.e. $x \in \Omega$. Furthermore, problem [C.2] has a unique solution, and so the convergence property extends to the whole sequence. The convergence of $\{G(u_h)\}_{h\in\mathcal{H}}$ to ∇u is an immediate consequence of Lemma 23 together with the uniqueness of the limit. □

III.1.5 Symmetric methods

In this section we show how the analysis can be simplified for symmetric methods. The following theorem replaces Lemma 23 and Theorem C.1.2:

Théorème C.1.3 (Convergence of symmetric methods) *Suppose that the bilinear form a_h is symmetric, i.e. $\widetilde{G} = G$ and let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of admissible meshes. Let u_h denote the unique solution of the discrete problem* [C.4] *on $\mathcal{T}_h$. Then, (i) the sequence $\{r_h^V u_h\}_{h\in\mathcal{H}}$ converges to the solution of* [C.2]*, say u, in $L^2(\Omega)$ as $h \to 0$; (ii) the sequence $\{G(u_h)\}_{h\in\mathcal{H}}$ converges to ∇u in $[L^2(\Omega)]^d$.*

Proof. Thanks to [C.12], the sequence $\{u_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{V_h}$ norm. Theorem C.1.1 states that we can extract a subsequence still denoted by $\{r_h^V u_h\}_{h\in\mathcal{H}}$ which converges to an element $\widetilde{u} \in H_0^1(\Omega)$ in $L^2(\Omega)$. Let us focus on the above sub-sequence. Owing to Hypothesis 13, $\widetilde{G}(u_h)$ weakly converges to $\nabla\widetilde{u}$ in L^2. Let $\varphi \in C_c^\infty(\Omega)$ and set $\varphi_h \stackrel{\text{def}}{=} \pi_h^V\varphi$. Observe that

$$a_h(u_h,\varphi_h) = (\nu G(u_h), \nabla\varphi)_{[L^2(\Omega)]^d} + (\nu G(u_h), G(u_h) - \nabla\varphi)_{[L^2(\Omega)]^d} + j_h(u_h,\varphi_h) \stackrel{\text{def}}{=} S_1 + S_2 + S_3.$$

Owing to the weak convergence of $G(u_h)$, $S_1 \to a(\widetilde{u},\varphi)$ as $h \to 0$. Using Cauchy-Schwarz inequality together with [C.11] we obtain

$$|S_2| \le \overline{\lambda}\|G(u_h)\|_{[L^2(\Omega)]^d}\|G(\varphi_h) - \nabla\varphi\|_{[L^2(\Omega)]^d} + [j_h(u_h,u_h)]^{1/2}[j_h(\varphi_h,\varphi_h)]^{1/2}.$$

Thanks to [C.8], [C.9] and [C.12], both $\|G(u_h)\|_{[L^2(\Omega)]^d}$ and $[j_h(u_h,u_h)]^{1/2}$ are bounded by $\|f\|_{L^2(\Omega)}$ up to a positive multiplicative constant. Equation [C.15] together with [C.10] then yield $|S_2| \to 0$ as $h \to 0$. Finally, $S_3 \to 0$ as $h \to 0$ by virtue of [C.10]. In conclusion,

$$(f,\varphi)_{L^2(\Omega)} \leftarrow (f,\varphi_h)_{L^2(\Omega)} = a_h(u_h,\varphi_h) \to a(\widetilde{u},\varphi),$$

i.e. , $\widetilde{u} = u$ for a.e. $x \in \Omega$ since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$. The strong convergence of $\{G(u_h)\}_{h\in\mathcal{H}}$ follows immediately. □

III.1.6 Adjoint methods

Let

$$a_h^*(u_h,v_h) \stackrel{\text{def}}{=} (\nu\widetilde{G}(u_h),G(v_h))_{[L^2(\Omega)]^d} + j_h(u_h,v_h).$$

In this section we investigate the convergence of the adjoint problem: Find $u_h \in V_h$ s.t.

$$a_h^*(u_h,v_h) = (f,r_h^V v_h)_{L^2(\Omega)}, \quad \forall v_h \in V_h. \tag{C.18}$$

Théorème C.1.4 (Convergence of adjoint methods) *Let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ be a family of admissible meshes. Let u_h^* denote the unique solution of the discrete problem* [C.18] *on $\mathcal{T}_h$. Then, the sequence $\{r_h^V u_h^*\}_{h\in\mathcal{H}}$ converges to the solution of* [C.2], *say u, in $L^2(\Omega)$ as $h \to 0$.*

Proof. Since also a_h^* is coercive, the sequence $\{u_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{V_h}$ norm. Theorem C.1.1 states that we can extract a subsequence still denoted by $\{r_h^V u_h\}_{h\in\mathcal{H}}$ which converges to an element $\widetilde{u} \in H_0^1(\Omega)$ in $L^2(\Omega)$. We shall focus our attention on the above sub-sequence. Let $\varphi \in C_c^\infty(\Omega)$ and set $\varphi \stackrel{\text{def}}{=} \pi_h^V\varphi$. We have

$$a_h^*(u_h^*,\varphi_h) = (\nu\widetilde{G}(u_h^*),\nabla\varphi)_{[L^2(\Omega)]^d} + (\nu\widetilde{G}(u_h^*),G(\varphi_h)-\nabla\varphi)_{[L^2(\Omega)]^d} + j_h(u_h^*,\varphi_h) \stackrel{\text{def}}{=} S_1+S_2+S_3.$$

Using Hypothesis 13 it is clear that $S_1 \to a(\widetilde{u},\varphi)$ as $h \to 0$. For the second term, using [C.12] we have

$$|S_2| \le \overline{\lambda}\|\widetilde{G}(u_h^*)\|_{[L^2(\Omega)]^d}\|G(\varphi_h)-\nabla\varphi\|_{[L^2(\Omega)]^d},$$

which, owing to [C.15], tends to zero as $h \to 0$. Similarly, using [C.10] together with [C.12], we can prove that $|S_3| \to 0$ as $h \to 0$. We thus have

$$(f,\varphi)_{L^2(\Omega)} \leftarrow (f,\varphi_h)_{L^2(\Omega)} = a_h(u_h^*,\varphi_h) \to a(\widetilde{u},\varphi),$$

i.e. , $\widetilde{u} = u$ for a.e. $x \in \Omega$ since $C_c^\infty(\Omega)$ is dense in $H_0^1(\Omega)$. This concludes the proof. □

C.2 Some examples

In this section we present some examples of conservative dG and FV methods which fit in the abstract framework above. Further examples which are not detailed here include the popular O-method (see, e.g., [1, 15]). Observe that the convergence results holds also for arbitrary compositions of the methods below.

III.2.1 Discontinuous Galerkin methods

In this section we shall present a number of dG methods which fit in the abstract analysis framework above. The weighted averaging techniques introduced in [27] and extended to dG methods in [33,41] will be used to ensure robust *a priori* estimates with respect to anisotropy and heterogeneity of the diffusion tensor in a suitable energy norm. The asimptotical convergence analysis can be performed following the guidelines of [33] and it is out of the scope of the present work. For all $F \in \mathcal{F}_h$ and for all φ s.t. a (possibly two-valued) trace is defined on F, we introduce the following jump operator:

$$[\![\varphi]\!] \stackrel{\text{def}}{=} \begin{cases} \varphi_{|T_1} - \varphi_{|T_2}, & \text{if } F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}, \\ \varphi_{|T}, & \text{if } F = \mathcal{F}_h^{T} \cap \mathcal{F}_h^{b}. \end{cases} \qquad \text{[C.19]}$$

The space $P_h^{k_V}(\mathcal{T}_h)$, $k_V \geq 1$, will be equipped with the following norm:

$$\|p_h\|_{1,2,h}^2 \stackrel{\text{def}}{=} \|\nabla_h p_h\|_{[L^2(\Omega)]^d}^2 + \sum_{F \in \mathcal{F}_h} \frac{1}{h_F} \|[\![p_h]\!]\|_{L^2(F)}^2, \quad \forall p_h \in P_h^{k_V}(\mathcal{T}_h), \qquad \text{[C.20]}$$

where ∇_h denotes the broken gradient. The proof of Hypothesis 11 can be found in [32, §6]. The following assumption need be added to those listed in Definition C.1.1:

Hypothesis 17 *Let $\mathcal{H}$ be a countable set and let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ denote a family of meshes matching Definition C.1.1. We require that the ratio of the diameter h_T, $T \in \mathcal{T}_h$, to the diameter of the largest ball inscribed in T be bounded from above by a parameter ϱ_3 independent of h.*

Remark 10 *Hypothesis 17 is not needed to prove Lemmata 11–21 for $k_V \geq 1$, so it is not listed in Definition C.1.1.*

For a given $k_V \geq 1$ we let $\mathcal{S}_h = \mathcal{T}_h$ and set

$$V_h \stackrel{\text{def}}{=} P_h^{k_V}(\mathcal{T}_h), \quad \Sigma_h \stackrel{\text{def}}{=} [P_h^{k_V}(\mathcal{T}_h)]^d.$$

We shall focus the piecewise constant case $\nu \in [P_h^0(\mathcal{T}_h)]^{d\times d}$. Let $\nu_{|T} = V_T D_T V_T^{-1}$ be the diagonalizaton of ν on $T \in \mathcal{T}_h$, *i.e.* , D_T is a diagonal matrix containing the

eigenvalues of ν. Denote with κ the element of $[P_h^0(\mathcal{T}_h)]^{d\times d}$ s.t. $\kappa_{|T} = V_T D_T^{1/2} V_T^{-1}$ for all $T \in \mathcal{T}_h$. The tensor field κ is symmetric, uniformly positive definite and s.t. $\nu = \kappa\kappa$ for a.e. $x \in \Omega$. Let, moreover, $\kappa^{-1} \in [P_h^0(\mathcal{T}_h)]^{d\times d}$ denote the inverse of κ, *i.e.* $\kappa\kappa^{-1} = I$ for a.e. $x \in \Omega$.

Remark 11 *The piecewise regular case $\nu \in [C_c^\infty(\mathcal{T}_h)]^{d\times d}$ requires only minor technical modifications in Lemma 24 below, which we omit for simplicity of exposition.*

Since V_h is a piecewise polynomial space, the reconstruction operator r_h^V can be taken equal to the identity on V_h. For all $F \in \mathcal{F}_h$ and for all φ s.t. a (possibly two-valued) trace is defined on F, we define the following weighted average operator: For a.e. $x \in F$,

$$\{\!\{\varphi\}\!\}_\omega \stackrel{\text{def}}{=} \begin{cases} \omega_2\varphi_{|T_1} + \omega_1\varphi_{|T_2}, & \text{if } F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}, \\ \varphi_{|T}, & \text{if } F = \mathcal{F}_h^T \cap \mathcal{F}_h^b, \end{cases}$$

where

$$\omega = (\omega_1, \omega_2) \stackrel{\text{def}}{=} \left(\tfrac{\lambda_1}{\lambda_1+\lambda_2}, \tfrac{\lambda_2}{\lambda_1+\lambda_2}\right), \quad \lambda_i \stackrel{\text{def}}{=} \sqrt{\nu_{|T_i}\mu_F \cdot \mu_F}, \quad i \in \{1,2\}.$$

Since $V_h = P_h^{k_V}$, we can take

$$\|v_h\|_{V_h} \stackrel{\text{def}}{=} \|v_h\|_{1,2,h},$$

with $\|\cdot\|_{1,2,h}$ defined as in [C.20]. The following lifting operators will play a crucial role in what follows: For all $F \in \mathcal{F}_h$ and for all $\varphi \in L^2(F)$, let $\mathbb{N} \ni l > 0$ and set

$$(r_{F,\kappa}^l(\varphi), \tau_h)_\Omega \stackrel{\text{def}}{=} (\varphi\mu_F, \{\!\{\kappa\tau_h\}\!\}_\omega)_{[L^2(F)]^d}, \quad \forall \tau_h \in [P_h^l(\mathcal{T}_h)]^d, \qquad [C.21]$$

and define $R_\kappa^l(\varphi) \stackrel{\text{def}}{=} \sum_{F\in\mathcal{F}_h} r_{F,\kappa}^l(\varphi)$. For $l = k_V$ the subscript will be omitted. For all $v_h \in V_h$, the weakly converging gradient is defined as

$$\widetilde{G}(v_h) \stackrel{\text{def}}{=} \nabla_h v_h - \kappa^{-1} R_\kappa(v_h),$$

where ∇_h denotes the broken gradient.

Remark 12 *To prove the convergence of the method, it is sufficient to work with the lifting operators r_F^0. However, if the exact solution u turns out to be more regular, optimal-order convergence rates can be established in the $\|\cdot\|_{V_h}$-norm when working with the lifting operators $r_F^{k_V-1}$ or $r_F^{k_V}$. The latter choice may be preferable for implementation purposes, especially if non-hierarchical, e.g. nodal-based, basis functions are used. For instance, if u belongs to the broken Sobolev space $H^{k+1}(\mathcal{T}_h)$, the usual* a priori *error analysis techniques can be used to infer a bound of the form $\|u-u_h\|_{V_h} \le C_u h^k$, with C_u a positive parameter depending on the norm of the exact solution u, on ϱ_i, $i \in \{1\dots3\}$, on k_V and on ν.*

Table C.1: Consistent gradient choices for dG methods. Symmetric methods are marked with a star.

Method	Ref.	$G(u_h)$
SIPG*	[16]	$\nabla_h u_h - \kappa^{-1} R_\kappa(u_h)$
NIPG	[62]	$\nabla_h u_h + \kappa^{-1} R_\kappa(u_h)$
IPG	[31]	$\nabla_h u_h$
BR*	[21]	$\nabla_h u_h - \kappa^{-1} R_\kappa(u_h)$
LDG*	[29]	$\nabla_h u_h - \kappa^{-1} R_\kappa(u_h)$

Several choices are possible for the consistent gradient G as well as for the bilinear form j_h. Some of the most common methods are presented in Tables C.1–C.2, where we have set

$$\lambda_{\min,F} \stackrel{\text{def}}{=} \begin{cases} \min(\lambda_1,\lambda_2), & \text{if } F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}, \\ \sqrt{\nu_{|T}\mu_F \cdot \mu_F}, & \text{if } F \in \mathcal{F}_h^T \cap \mathcal{F}_h^b, \end{cases} \qquad s_h(u_h,v_h) \stackrel{\text{def}}{=} (R_\kappa([\![u_h]\!]), R_\kappa([\![v_h]\!]))_{[L^2(\Omega)]^d}.$$

Remark 13 *The original formulation of the methods proposed in [16, 21, 29, 31, 62] has been modified using the averaging techniques introduced in [33]. Optimal asymptotic order estimates which are also robust with respect to anisotropy and heterogeneity can be obtained in the following norm:*

$$\|v_h\|_{\mathrm{DG},\nu}^2 \stackrel{\text{def}}{=} \|\kappa\nabla_h v_h\|_{[L^2(\Omega)]^d}^2 + |v_h|_J^2, \quad |v_h|_J^2 \stackrel{\text{def}}{=} \sum_{F\in\mathcal{F}_h} \frac{1}{h_F} \|\lambda_{\min,F}^{1/2} [\![v_h]\!]\|_{L^2(F)}^2.$$

The above norm is equivalent to $\|\cdot\|_{V_h}$ since, for all $v_h \in V_h$, $\underline{\lambda}^{1/2}\|v_h\|_{V_h} \le \|v_h\|_{\mathrm{DG},\nu} \le \overline{\lambda}^{1/2}\|v_h\|_{V_h}$.

The following result was proved in [32]:

Lemma 24 *Assume that Hypothesis 17 holds. Then, for all $F \in \mathcal{F}_h$, for all $v_h \in V_h$, there is $C_{\mathrm{IP}} > 0$ depending on ϱ_i, $i \in \{1\ldots3\}$, on k_V but not on h s.t.*

$$\|r_{F,\kappa}(v_h)\|_{[L^2(\Omega)]^d}^2 \le C_{\mathrm{IP}} |v_h|_J^2.$$

Furthermore, assume that there is a parameter ϱ_4 independent of h s.t.

$$h_F|F| \ge \varrho_4|T|, \quad \forall T \in \mathcal{T}_h,\ \forall F \in \mathcal{F}_h^T. \tag{C.22}$$

Then, for all $F \in \mathcal{F}_h$, for all $v_h \in V_h$, there is $c_{\mathrm{IP}} > 0$ depending on ϱ_i, $i \in \{1\ldots4\}$, on k_V but not on h s.t.

$$c_{\mathrm{IP}} |v_h|_J^2 \le \|r_{F,\kappa}(v_h)\|_{[L^2(\Omega)]^d}^2. \tag{C.23}$$

Remark 14 *Inequality* [C.23] *is only needed to prove the coercivity of the BR method (see Lemma 28 below), whereas it is not needed for the other methods listed in Tables C.1–C.2. In what follows we shall therefore tacitly require* [C.22] *only when dealing with the BR method.*

Lemma 25 (Proof of Hypothesis 12) *Let the assumptions of Lemma 24 hold true. Then, Hypothesis 12 holds for all the consistent gradients listed in Table C.1.*

PROOF. Property [C.7] is in fact verified with the equal sign. Let us prove [C.8] for $\widetilde{G}$ (the proof for the gradients listed in Table C.1 is similar and will be omitted). For all $v_h \in V_h$,

$$\|\widetilde{G}(v_h)\|^2_{[L^2(\Omega)]^d} \le 2\|\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + \frac{2}{\underline{\lambda}} \sum_{T\in\mathcal{T}_h} \|R_\kappa([\![v_h]\!])\|^2_{[L^2(T)]^d} \stackrel{\text{def}}{=} S_1 + S_2.$$

According to [C.21], for all $F \in \mathcal{F}_h$, $r_{F,\kappa}$ is solely supported by the elements which share F. We thus have that $R_\kappa([\![v_h]\!])_{|T} = \sum_{F\in\mathcal{F}_h^T} r_{F,\kappa}([\![v_h]\!])_{|T}$ and, owing to Lemma 24,

$$S_2 \le \frac{2N_\partial}{\underline{\lambda}} \sum_{F\in\mathcal{F}_h} \|r_{F,\kappa}(v_h)\|^2_{[L^2(\Omega)]^d} \le \frac{2C_{\rm IP}N_\partial}{\underline{\lambda}}|v_h|_J^2 \le \frac{2C_{\rm IP}N_\partial\overline{\lambda}}{\underline{\lambda}}\|v_h\|^2_{V_h},$$

which yields $\|\widetilde{G}(v_h)\|^2_{[L^2(\Omega)]^d} \le 2\left(1+\frac{2C_{\rm IP}N_\partial\overline{\lambda}}{\underline{\lambda}}\right)\|v_h\|^2_{V_h}$. □

Remark 15 *The L^2 projector π_h^1 onto the space $P_h^1(\mathcal{T}_h)$ enjoys the following property:*

$$\lim_{h\to\infty} \|\varphi - \pi_h^1\varphi\|_{V_h} = 0, \quad \forall\varphi \in C_c^\infty(\Omega). \tag{C.24}$$

Lemma 26 (Proof of Hypothesis 13) *Hypothesis 13 holds.*

PROOF. Let $\{v_h\}_{h\in\mathcal{H}}$ be a sequence in V_h satisfying the assumptions of Hypothesis 13. The sequence $\{\widetilde{G}(v_h)\}_{h\in\mathcal{H}}$ is bounded, and it converges (up to a subsequence) to some $\tau \in [L^2(\Omega)]^d$. It only remains to prove that $\tau = \nabla v$ for a.e. $x \in \mathbb{R}^d$. Let $\Phi \in [C_c^\infty(\mathbb{R}^d)]^d$, $v_h \in V_h$ and prolong v_h by zero outside Ω. Observe that

$$(\widetilde{G}(v_h), \pi_h^1\Phi)_{[L^2(\mathbb{R}^d)]^d} = -(v_h, \nabla_h{\cdot}\pi_h^1\Phi)_{L^2(\mathbb{R}^d)} + \sum_{F\in\mathcal{F}_h^i} (\{\!\{v_h\}\!\}_\omega, \mu_F{\cdot}[\![\pi_h^1\Phi]\!])_{L^2(F)},$$

where $\nabla_h\cdot$ denotes the broken divergence operator. Owing to the regularity of Φ, $[\![\Phi]\!] = 0$ for a.e. $x \in F$, $F \in \mathcal{F}_h$. The above identity then yields

$$\begin{aligned}&|(v_h, \nabla{\cdot}\Phi)_{L^2(\Omega)} + (\widetilde{G}(v_h), \pi_h^1\Phi)_{[L^2(\Omega)]^d}|\\ &= |(v_h, \nabla_h{\cdot}(\Phi-\pi_h^1\Phi))_{L^2(\Omega)} - \sum_{F\in\mathcal{F}_h^i} (\{\!\{v_h\}\!\}_\omega, \mu_F{\cdot}[\![\Phi-\pi_h^1\Phi]\!])_{L^2(\Omega)}| \le \|v_h\|_{V_h}\|\Phi-\pi_h^1\Phi\|_{V_h}.\end{aligned}$$

Passing to the limit and using [C.24] and the boundedness of $\{v_h\}_{h\in\mathcal{H}}$ in the $\|\cdot\|_{V_h}$ norm concludes the proof. □

Table C.2: Consistent stabilization choices for dG methods. Symmetric methods are marked with a star.

Method	$j_h(u_h, v_h)$
SIPG*	$\sum_{F\in\mathcal{F}_h} (\eta_{\mathrm{SIPG}} \frac{\lambda_{\min,F}}{h_F} [\![u_h]\!], [\![v_h]\!])_{L^2(F)} - s_h(u_h, v_h)$
NIPG	$\sum_{F\in\mathcal{F}_h} (\eta_{\mathrm{NIPG}} \frac{\lambda_{\min,F}}{h_F} [\![u_h]\!], [\![v_h]\!])_{L^2(F)} + s_h(u_h, v_h)$
IPG	$\sum_{F\in\mathcal{F}_h} (\eta_{\mathrm{IPG}} \frac{\lambda_{\min,F}}{h_F} [\![u_h]\!], [\![v_h]\!])_{L^2(F)}$
BR*	$\sum_{F\in\mathcal{F}_h} (\eta_{\mathrm{BR}} r_{F,\kappa}([\![u_h]\!]), r_{F,\kappa}([\![v_h]\!]))_{[L^2(F)]^d} - s_h(u_h, v_h)$
LDG*	$\sum_{F\in\mathcal{F}_h} (\eta_{\mathrm{LPG}} \frac{\lambda_{\min,F}}{h_F} [\![u_h]\!], [\![v_h]\!])_{L^2(F)}$

Lemma 27 (Proof of Hypothesis 14) *Let the stabilization parameters satisfy*

$$\eta_{\mathrm{SIPG}} > N_\partial C_{\mathrm{IP}}, \quad \eta_{\mathrm{NIPG}} > 0, \quad \eta_{\mathrm{IPG}} > N_\partial C_{\mathrm{IP}}/2, \quad \eta_{\mathrm{BR}} > N_\partial, \quad \eta_{\mathrm{LDG}} > 0.$$

Then, Hypothesis 14 holds for all the stabilizations of Table C.2.

Proof. The continuity of the stabilizations of Table C.2 stems from a simple application of Cauchy-Schwarz inequality. The IFP as well as the LDG stabilizations are clearly positive. Proceeding as in the proof of Lemma 25, we have that

$$s_h(v_h, v_h) \leq N_\partial \sum_{F\in\mathcal{F}_h} \|r_{F,\kappa}([\![v_h]\!])\|^2_{[L^2(\Omega)]^d} \leq C_{\mathrm{IP}} N_\partial |v_h|^2_J,$$

which yields the positivity of the SIPG, NIPG and BR stabilization. The term s_h is introduced to reduce the stencil of the above methods to neighbouring elements. C.2 immediately follows from the above remark provided the above assumptions on the stabilization parameters are matched. In order to prove consistency, let $\varphi \in C_c^\infty(\Omega)$. Since $[\![\varphi]\!] = 0$ for a.e. $x \in F$, $F \in \mathcal{F}_h$, the continuity of j_h gives

$$j_h(\pi_h^1\varphi, \pi_h^1\varphi) \lesssim |\varphi_h|^2_J = |\varphi_h - \varphi|^2_J \leq \overline{\lambda} \|\pi_h^1\varphi - \varphi\|^2_{V_h},$$

which, according to [C.24], tends to zero as $h \to 0$. □

Lemma 28 (Proof of Hypothesis 15) *Under the assumptions of Lemma 27, Hypothesis 15 holds true for all the methods of Tables C.1–C.2.*

PROOF. For the sake of brevity, the proof will be detailed for the BR and SIPG methods only. For all $v_h \in V_h$, Young inequality together with Lemma 24 yield

$$\begin{aligned} a_h^{\mathrm{BR}}(v_h, v_h) &= \|\kappa\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + 2(\kappa\nabla_h v_h, R_\kappa(v_h))_+ \eta_{\mathrm{BR}} \sum_{F\in\mathcal{F}_h} (r_{F,\kappa}([\![v_h]\!]), r_{F,\kappa}([\![v_h]\!]))_{[L^2(\Omega)]^d} \\ &\geq \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + (\eta_{\mathrm{BR}} - (1+\epsilon)N_\partial) \sum_{F\in\mathcal{F}_h} \|r_{F,\kappa}([\![v_h]\!])\|^2_{[L^2(\Omega)]^d} \\ &\geq \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + (\eta_{\mathrm{BR}} - (1+\epsilon)N_\partial) c_{\mathrm{IP}} |v_h|_J^2, \end{aligned}$$

for all $\epsilon > 0$. Coercivity then holds for $\eta_{\mathrm{BR}} > N_\partial$. Similarly,

$$\begin{aligned} a_h^{\mathrm{SIPG}}(v_h, v_h) &= \|\kappa\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + 2(\kappa\nabla_h v_h, R_\kappa(v_h))_+ \eta_{\mathrm{SIPG}} \sum_{F\in\mathcal{F}_h} (\frac{\lambda_{\min,F}}{h_F}[\![v_h]\!], [\![v_h]\!])_{L^2(F)} \\ &\geq \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|\nabla_h v_h\|^2_{[L^2(\Omega)]^d} + (\eta_{\mathrm{SIPG}} - (1+\epsilon)N_\partial C_{\mathrm{IP}}) \sum_{F\in\mathcal{F}_h} (\frac{\lambda_{\min,F}}{h_F}[\![v_h]\!], [\![v_h]\!])_{L^2(F)}, \end{aligned}$$

yielding coercivity for $\eta_{\mathrm{SIPG}} > N_\partial C_{\mathrm{IP}}$. □

Finally, Hypothesis [16] follows from [C.24]

III.2.2 A cell-based finite volume method

We consider hereafter a new finite volume method displaying all the ingredients introduced in §C.1. Throughout the present and the following section, the following assumption on the mesh need be added to those listed in Definition C.1.1:

Hypothesis 18 *Let $\mathcal{H}$ be a countable set and let $\{\mathcal{T}_h\}_{h\in\mathcal{H}}$ denote a family of meshes matching Definition C.1.1. Them*

(i) there is a positive parameter ϱ_5 independent of $h \in \mathcal{H}$ s.t.

$$\frac{|x_T - x_F|}{d_{T,F}} \leq \varrho_5, \quad \forall F \in \mathcal{F}_h^T,\ \forall T \in \mathcal{T}_h; \tag{C.25}$$

(ii) $\mathcal{P}_h$ is a family of points of Ω indexed by the elements of $\mathcal{T}_h$ and $\mathcal{P}_h = \{x_T\}_{T\in\mathcal{T}_h}$ is s.t., for all $T \in \mathcal{T}_h$, $x_T \in T$ and T is star-shaped with respect to x_T, i.e. , $[x_T, x] \subset T$ for all $x \in T$;

(iii) there is $\varrho_2 > 0$ s.t., for all $F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}$, $(T_1, T_2) \in [\mathcal{T}_h]^2$,

$$\varrho_2 \leq \frac{d_{T_1,F}}{d_{T_2,F}} \leq \frac{1}{\varrho_2},$$

where, for all $T \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_h^T$, we have set $d_{T,F} \stackrel{\mathrm{def}}{=} \mathrm{dist}(x_T, F) > 0$.

For all $T \in \mathcal{T}_h$ and for all $F \in \mathcal{F}_h^T$, we define

$$d_F \overset{\text{def}}{=} \begin{cases} d_{T_1,F} + d_{T_2,F}, & \text{if } F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}, \\ d_{T,F}, & \text{if } F = \mathcal{F}_h^T \cap \mathcal{F}_h^b. \end{cases}$$

In the present and in the following section, the space $P_h^0(\mathcal{T}_h)$ will be equipped with the the discrete H_0^1 norm:

$$\|p_h\|_{1,2,h}^2 \overset{\text{def}}{=} \sum_{F \in \mathcal{F}_h} \frac{1}{d_F} \|[\![p_h]\!]\|_{L^2(F)}^2, \quad \forall p_h \in P_h^0(\mathcal{T}_h), \tag{C.26}$$

where the jump operator has been defined in [C.19]. The proof that Hypothesis 11 holds for the norm [C.26] can be found in [45, §5]. Let

$$V_h \overset{\text{def}}{=} P_h^0(\mathcal{T}_h), \quad \Sigma \overset{\text{def}}{=} [P_h^0(\mathcal{T}_h)]^d.$$

Since V_h is a piecewise polynomial space, the reconstruction operator r_h^V can be taken equal to the identity on V_h. For all $F \in \mathcal{F}_h$ and for all $v_h \in V_h$ we define the following trace operator $\gamma_F : V_h \to \mathbb{P}^0(F)$:

$$\gamma_F(v_h) \overset{\text{def}}{=} \begin{cases} \omega_F^{T_2} v_{h|T_1} + \omega_F^{T_1} v_{h|T_2}, & \forall F = \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}, \\ 0, & \forall F = \mathcal{F}_h^T \cap \mathcal{F}_h^b, \end{cases}, \qquad \omega_F^T \overset{\text{def}}{=} \frac{d_{T,F}}{d_F} \leq 1.$$

For all $T \in \mathcal{T}_h$, for all $F \in \mathcal{F}_h^T$, let $\mathcal{I}_F^T : V_h \to \mathbb{P}^0(F)$ denote a linear interpolation operator s.t.

$$|(\mathcal{I}_F^T \circ \pi_h^0)\varphi - \varphi(x_F)| \leq C_\varphi h_F d_{T,F}, \quad \forall \varphi \in C_c^\infty(\Omega), \tag{C.27}$$

where $\pi_h^0 \equiv \pi_V$ denotes the L^2 projection onto V_h, x_F is the barycenter of F and C_φ denotes a positive parameter depending on some (bounded) norm of φ.

Remark 16 *A simple choice for the interpolator $\mathcal{I}_F^T$ is described hereafter. For the sake of simplicity, let $d = 2$. For all $F \in \mathcal{F}_h^T \cap \mathcal{F}_h^b$ we set $\mathcal{I}_F^T v_h = 0$. Let $F \in \mathcal{F}_h^{T_0} \cap \mathcal{F}_h^i$, $T_0 \in \mathcal{T}_h$, and let $T_1 \neq T_2$ be two elements of $\mathcal{T}_h \setminus \{T_0\}$ s.t. their barycenters are not aligned with that of T_0 (see Figure III.2.2 for an example). Denote by $\{\alpha_i\}_{i \in \{0 \dots d\}}$ the barycentric coordinates of x_F with respect to $\{x_{T_i}\}_{i \in \{0 \dots d\}}$. Then, for all $v_h \in V_h$, we set*

$$\mathcal{I}_F^{T_0} v_h \overset{\text{def}}{=} \sum_{i=0}^{d} \alpha_i v_{h|T_i}.$$

While the above choice ensures the convergence of the method, it does not yield strong consistency for piecewise linear exact solutions in the presence of heterogeneity. Other choices are possible, but their description lies out of the scope of the present paper. In particular, we refer to [11] for an alternative using the so called L interpolation introduced in [6].

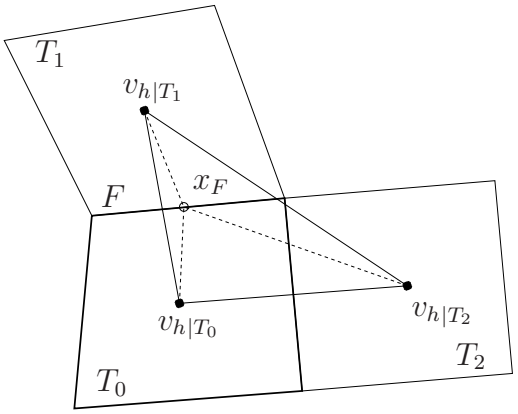


Figure C.1: Barycentric interpolation for $d = 2$.

For all $v_h \in V_h$, the gradient reconstructions are defined as follows: For all $T \in \mathcal{T}_h$,

$$\widetilde{G}(v_h)_{|T} \stackrel{\text{def}}{=} \frac{1}{|T|} \sum_{F \in \mathcal{F}_h^T} |F|(\gamma_F v_h - v_{h|T})\mu_F^T, \quad G(v_h)_{|T} \stackrel{\text{def}}{=} \frac{1}{|T|} \sum_{F \in \mathcal{F}_h^T} |F|(\mathcal{I}_F^T v_h - v_{h|T})\mu_F^T.$$

The space V_h will be equipped with the following norm:

$$\|v_h\|_{V_h}^2 \stackrel{\text{def}}{=} \sum_{T \in \mathcal{T}_h} \sum_{F \in \mathcal{F}_h^T} \frac{|F|}{d_{T,F}} (\mathcal{I}_F^T v_h - v_{h|T})^2.$$

Remark 17 *For all $h \in \mathcal{H}$ we have*

$$\sum_{F \in \mathcal{F}_h^T} \frac{|F| d_{T,F}}{|T|} = d, \quad \forall T \in \mathcal{T}_h. \tag{C.28}$$

Lemma 29 (Proof of Hypothesis 12) *Hypothesis 12 holds.*

Proof. Let v_h be a generic element of V_h. Cauchy-Schwarz inequality gives

$$\frac{[\![v_h]\!]^2}{d_F} \le \frac{(v_{h|T_1} - \mathcal{I}_F^{T_1} v_h)^2}{d_{T_1,F}} + \frac{(v_{h|T_2} - \mathcal{I}_F^{T_2} v_h)^2}{d_{T_2,F}}, \quad \forall F \in \mathcal{F}_h^{T_1} \cap \mathcal{F}_h^{T_2}.$$

Inequality [C.7] immediately follows. Cauchy-Schwarz inequality together with [C.28] yield

$$\begin{aligned} \|\widetilde{G}(v_h)\|_{[L^2(\Omega)]^d}^2 &= \sum_{T \in \mathcal{T}_h} \frac{1}{|T|} \Big| \sum_{F \in \mathcal{F}_h^T} |F| \omega_F^T [\![v_h]\!] \mu_F \Big|^2 \\ &\le \sum_{T \in \mathcal{T}_h} \left(\sum_{F \in \mathcal{F}_h^T} \frac{1}{d_{T,F}} \|[\![v_h]\!]\|_{L^2(F)}^2 \times \sum_{F \in \mathcal{F}_h^T} \frac{|F| d_{T,F}}{|T|} \right) \le d \|v_h\|_{1,2,h}^2 \le d \|v_h\|_{V_h}^2. \end{aligned} \tag{C.29}$$

Similarly,

$$\|G(v_h)\|_{[L^2(\Omega)]^d}^2 = \sum_{T\in\mathcal{T}_h}\frac{1}{|T|}\Big|\sum_{F\in\mathcal{F}_h^T}|F|(\mathcal{I}_F^T u_h - u_{h|T})\mu_F^T\Big|^2$$
$$\le \sum_{T\in\mathcal{T}_h}\left(\sum_{F\in\mathcal{F}_h^T}\frac{1}{d_{T,F}}\|\mathcal{I}_F^T u_h - u_{h|T}\|_{L^2(F)}^2 \times \sum_{F\in\mathcal{F}_h^T}\frac{|F|d_{T,F}}{|T|}\right) \le d\|v_h\|_{V_h}^2.$$

Observing that $\|v_h\|_{V_h}$ is bounded by assumption whereas the term in brackets tends to 0 as $h\to 0$ concludes the proof of [C.8]. □

Lemma 30 (Proof of Hypothesis 13) *Hypothesis 13 holds.*

Proof. Let $\{v_h\}_{h\in\mathcal{H}}$ be a sequence in V_h satisfying the assumptions of Hypothesis 13. The sequence $\{\widetilde{G}(v_h)\}_{h\in\mathcal{H}}$ is bounded, and it converges (up to a subsequence) to some $\tau\in[L^2(\Omega)]^d$. It only remains to prove that $\tau=\nabla v$ for a.e. $x\in\Omega$. Let $\Phi\in[C_c^\infty(\Omega)]^d$ and prolong v_h by zero outside Ω. Define $\Phi_h^T \stackrel{\text{def}}{=} \int_T\Phi/|T| = \pi_h^0\Phi_{|T}$ for all $T\in\mathcal{T}_h$ and $\Phi_h^F\stackrel{\text{def}}{=}\int_F\Phi/|F|$ for all $F\in\mathcal{F}_h$. Integration by parts yields

$$|(\widetilde{G}(v_h),\Phi)_{[L^2(\Omega)]^d} + (\nabla\cdot\Phi, r_h^V v_h)_{L^2(\Omega)}| = \Big|\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|F|(\gamma_F v_h - v_{h|T})(\Phi_h^F-\Phi_h^T)\cdot\mu_F^T\Big|$$
$$\le \|v_h\|_{V_h}\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|F|d_{T,F}(\Phi_h^F-\Phi_h^T)^2\right)^{\frac12},$$

which proves the assert. □

Define the stabilization term as follows:

$$j(u_h,v_h)\stackrel{\text{def}}{=}\sum_{T\in\mathcal{T}_h}\eta_{\mathrm{CVF}}^T\sum_{F\in\mathcal{F}_h^T}\frac{1}{d_{T,F}}(R_{T,F}(u_h),R_{T,F}(v_h))_{L^2(F)},$$

where, for all $v_h\in V_h$, we have set $R_{T,F}(v_h)\stackrel{\text{def}}{=}\mathcal{I}_F^T v_h - v_{h|T} - G(v_h)_{|T}\cdot(x_F-x_T)$, and, for all $T\in\mathcal{T}_h$, $0<\underline{\eta}\le\eta_{\mathrm{CVF}}^T<\overline{\eta}\le\infty$ denotes a positive stabilization parameter.

Lemma 31 (Proof of Hypothesis 14) *Hypothesis 14 holds.*

Proof. The proposed stabilization term is clearly symmetric and positive semi-definite. In order to prove the continuity, observe that, for all $v_h\in V_h$,

$$j_h(v_h,v_h)\le 2\sum_{T\in\mathcal{T}_h}\eta_{\mathrm{CVF}}^T\left(\sum_{F\in\mathcal{F}_h^T}\frac{1}{d_{T,F}}\|\mathcal{I}_F^T v_h - v_{h|T}\|_{L^2(F)}^2 + \sum_{F\in\mathcal{F}_h^T}\frac{|F|}{d_{T,F}}(G(v_h)\cdot(x_F-x_T))^2\right).$$

Let S_1^T, S_2^T the addends in brackets. Using [C.28] together with Hypothesis 18 and Lemma 29 we have that

$$\sum_{T\in\mathcal{T}_h} S_2^T \le \varrho_5 \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} \frac{|F|d_{T,F}}{|T|}|T||G(v_h)|^2 \le d\varrho_5\|G(v_h)\|^2_{[L^2(\Omega)]^d} \le d^2\varrho_5\|v_h\|^2_{V_h},$$

whence $j_h(v_h,v_h)\le 2\overline{\eta}(1+d^2\varrho_5)\|v_h\|^2_{V_h}$. Using the above result together with [C.11] we have

$$j_h(u_h,v_h)\le j_h(u_h,u_h)^{1/2}j_h(v_h,v_h)^{1/2}\le 2\overline{\eta}(1+d^2\varrho_5)\|u_h\|_{V_h}\|v_h\|_{V_h}.$$

It only remains to proof the consistency of j_h. In the rest of the proof, shall assume that [C.15] holds (a proof is given in Lemma 33 below). Let $\varphi\in C_c^\infty(\Omega)$ and set $\varphi_h \stackrel{\text{def}}{=} \pi_h^0\varphi$. Observe that

$$|R_{T,F}(v_h)| \le |\mathcal{I}_F^T\varphi_h - \varphi(x_F)| + |(\nabla\varphi(x_T)-G(\varphi_h))\cdot(x_F-x_T)| + c_\varphi|x_T-x_F|^2,$$

where c_φ denotes a positive parameter depending on a suitable (bounded) norm of φ. Substituting in the expression of j_h and using Hypothesis 18 we obtain

$$j_h(\varphi_h,\varphi_h)\le$$
$$4\overline{\eta}\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}\frac{|F|}{d_{T,F}}|\mathcal{I}_F^I\varphi_h-\varphi(x_F)|^2 + \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}\frac{|F|}{d_{T,F}}|\nabla\varphi(x_T)-G(\varphi_h)|^2 + |F|\varrho_5c_\varphi h_T^3\right).$$

Let S_i, $i\in\{1,2\}$ denote the first two addends in brackets. Using [C.27] together with [C.28] we have

$$S_1 \le C_\varphi\sum_{T\in\mathcal{T}_h}|T|\sum_{F\in\mathcal{F}_h^T}\frac{|F|d_{T,F}}{|T|}h_F^2 \le C_\varphi h^2 d|\Omega|,$$

i.e., $S_1\to 0$ as $h\to 0$. Using Hypothesis 18 and [C.28] we have

$$\begin{aligned}S_2 &\le \sum_{T\in\mathcal{T}_h}|T||\nabla\varphi(x_T)-G(\varphi_h)|^2\sum_{F\in\mathcal{F}_h^T}\frac{|F|d_{T,F}}{|T|}\frac{|x_F-x_T|^2}{d_{T,F}^2} \le d\varrho_5^2\sum_{T\in\mathcal{T}_h}\|\nabla\varphi(x_T)\|^2_{[L^2(T)]^d}\\ &\le 2d\varrho_5^2\sum_{T\in\mathcal{T}_h}\left(\|\nabla\varphi(x_T)-\nabla\varphi\|^2_{[L^2(T)]^d}+\|\nabla\varphi-G(\varphi_h)\|^2_{[L^2(T)]^d}\right),\end{aligned}$$

which, since [C.15] holds, shows that S_2 tends to zero as $h\to 0$. This concludes the proof. □

As the FV method proposed in this section is non-symmetric, it is conditionally coercive. In what follows, we shall provide a computable criterion to check coercivity

for a given mesh $\mathcal{T}_h$ and diffusion tensor ν. For the sake of simplicity we shall refer to the interpolator defined in Remark 16. For a given $T \in \mathcal{T}_h$ we introduce the bilinear form a_h^T defined as

$$a_h^T(u_h, v_h) = (\nu G(u_h)_{|T}, \widetilde{G}(u_h))_{[L^2(T)]^d} x + \eta_{\rm CVF}^T \sum_{F \in \mathcal{F}_h^T} \frac{1}{d_{T,F}} (R_{T,F}(u_h), R_{T,F}(v_h))_{L^2(F)}.$$

Let $\mathcal{T}_h^T \stackrel{\rm def}{=} \{T' \in \mathcal{T}_h, \mathcal{F}_h^{T'} \cap \mathcal{F}_h^T \neq \emptyset\}$ denote the set of elements sharing a face with T and set $m^T \stackrel{\rm def}{=} {\rm card}\mathcal{T}_h^T$. For brevity of notation, we shall note $\mathcal{T}_h^T = \{T_i\}_{1 \le i \le m^T}$ with T_i sharing the internal face F_i with T. Moreover, we define $m^F \stackrel{\rm def}{=} {\rm card}\mathcal{F}_h^{T_i} \cap \mathcal{F}_h^b$ and set $\{F_i\}_{m^T+1 \le i \le m^T+m^F} \stackrel{\rm def}{=} \mathcal{F}_h^T \cap \mathcal{F}_h^b$. Define the linear map $X^T : V_h \mapsto \mathbb{R}^{(m^T+m^F)}$ s.t., for all $v_h \in V_h$,

$$X^T(v_h) \stackrel{\rm def}{=} \{\{v_{h|T_i} - v_{h|T}\}_{1 \le i \le m^T}, \{\mathcal{I}_{F_i}^T(v_h) - v_{h|T}\}_{m^T+1 \le i \le m^T+m^F}\},$$

and recall that $\mathcal{I}_{F_i}^T(v_h) = 0$ for $m^T + 1 \le i \le m^T + m^F$ (since $\mathcal{I}_{F_i}^T(v_h)$ vanishes on boundary faces). It is a simple matter to verify that for all $T \in \mathcal{T}_h$, there exists a matrix $A_h^T \in \mathbb{R}^{(m^T+m^F)\times(m^T+m^F)}$ s.t., for all $(u_h, v_h) \in [V_h]^2$,

$$a_h^T(u_h, v_h) = (X^T(u_h))^t A^T X^T(v_h).$$

Notice also that, again because $\mathcal{I}_{F_i}^T(v_h) = 0$ for $m^T + 1 \le i \le m^T + m^F$, we can write

$$\mathcal{I}_{F_i}^T v_h = v_{h|T} + \sum_{j=1}^{m^T+m^F} \beta_{ij}^T X^T(v_h)_j, \quad 1 \le i \le m^T + m^F,$$

where the family of reals $\{\beta_{ij}^T\}_{1 \le j \le m^T+m^F}$ verifies $\sum_{j=1}^{m^T+m^F} \beta_{i,j}^T = 1$. Let $B^T \in \mathbb{R}^{(m^T+m^F)\times(m^T+m^F)}$ be the matrix of elements β_{ij}^T and define the norm $\|\cdot\|_T$ as follows: For all $x \in \mathbb{R}^{m^T+m^F}$,

$$\|x\|_T^2 \stackrel{\rm def}{=} \sum_{i=1}^{m^T+m^F} \frac{|F_i|}{d_{T,F_i}} (B^T x)_i^2. \qquad \text{[C.30]}$$

The following result provides a computable local criterion expressed in term of the local matrices $\{A^T\}_{T \in \mathcal{T}_h}$:

Lemma 32 (Proof of Hypothesis 15) *The bilinear form a_h is coercive if for all $T \in \mathcal{T}_h$, the matrix A^T is uniformly coercive for the norm $\|\cdot\|_T$, i.e. if there is $C > 0$ independent of h s.t., for all $x \in \mathbb{R}^{m^T+m^F}$, $x^t A^T x \ge C\|x\|_T^2$.*

Proof. For all $v_h \in V_h$,

$$a_h(v_h, v_h) = \sum_{T \in \mathcal{T}_h} a_h^T(u_h, u_h) = \sum_{T \in \mathcal{T}_h} (X^T(u_h))^t A^T X^T(u_h) \ge C \sum_{T \in \mathcal{T}_h} \|X^T(u_h)\|_T^2 = C\|u_h\|_{V_h}^2,$$

which concludes the proof. □

Lemma 33 (Proof of Hypothesis 16) *Hypothesis 16 holds.*

PROOF. Estimates [C.13]–[C.14] classically hold for $\pi_V = \pi_h^0$ (see, e.g., [37]). Let now $\varphi \in C_c^\infty(\Omega)$, set $\varphi_h \stackrel{\text{def}}{=} i_h^0\varphi$ and observe that, for all $T \in \mathcal{T}_h$,

$$(G(\varphi_h) - \nabla\varphi)_{|T} = \sum_{F\in\mathcal{F}_h^T} \frac{|F|}{|T|}(\mathcal{I}_F^T\varphi_h - \varphi(x_F))\mu_F^T + (\nabla\varphi(\hat{x}_T) - \nabla\varphi)$$
$$+ \left(\sum_{F\in\mathcal{F}_h^T} \frac{|F|}{|T|}(\varphi(x_F) - \varphi(\hat{x}_T))\mu_F^T - \nabla\varphi(\hat{x}_T)\right) \stackrel{\text{def}}{=} S_1^T + S_2^T + S_3^T,$$

where we have used the fact that, owing to assumption (iii) in Definition C.1.1, $\sum_{F\in\mathcal{F}_h^T}\mu_F^T = 0$ for all $T \in \mathcal{T}_h$ to replace $\varphi_{h|T}$ with $\varphi(\hat{x}_T)$ in S_3^T. Clearly, $\|G(\varphi_h) - \nabla\varphi\|^2_{[L^2(\Omega)]^d} \le 3\sum_{i=1}^3 \|S$
Estimate [C.27] together with [C.28] yields, for all $T \in \mathcal{T}_h$,

$$|S_1^T| \le \sum_{F\in\mathcal{F}_h^T} \frac{|F|d_{T,F}}{|T|}\frac{|\mathcal{I}_F^T\varphi_h - \varphi(x_F)|}{d_{T,F}} \le C'_\varphi dh_T,$$

so that $\|S_1^T\|_{[L^2(\Omega)]^d} \le C'_\varphi|\Omega|^{1/2}dh_T$. On the other hand, using classical estimates for π_h^0, we conclude that $\|S_2^T\|_{[L^2(\Omega)]^d} \le Ch_T\|\varphi\|_{[H^1(T)]^d}$. Finally, thanks to the regularity of φ, there is C''_φ depending on φ and on the mesh regularity s.t. $\|S_3^T\|^2_{[L^2(T)]^d} \le C''_\varphi|T|h_T^2$, *i.e.* , $\|S_3^T\|_{[L^2(\Omega)]^d} \le C''_\varphi|\Omega|^{1/2}h$. The above estimates yield the desired result. □ For the sake of completeness, the order of convergence of the new FV method presented in this section has been numerically evaluated by solving the Dirichlet problem for $d = 2$ with $u = \sin(\pi x)\sin(\pi y)$ ($u_{\min} = 0$, $u_{\max} = 1$), $f = -\triangle u$ and anisotropy ratios of 1 and 1000 on a family of randomly perturbed quadrangular meshes of $(0,1)^2$. The results are reported in Tables III.2.2 and III.2.2 and show second order convergence as well as robustness with respect to anisotropy and mesh skewdness. The following indicators are also listed: (i) `nunkw`, the number of unknowns; (ii) `nnmat`, the number of nonzero matrix entries; (iii) `erl2`, the L^2 error; (iv) `ocverl2`, the order of convergence for the L^2 error; (v) `umin` and `umax`, the minimum and maximum value of the discrete solution. A torough validation of the above method will be the subject of a future work. An asimptotic *a priori* analysis can be performed following the guidelines of [42], but it lies out of the scope of the present work.

III.2.3 A hybrid finite volume method

The goal of this section is to show that the hybrid finite volume method proposed in [43] fits in the frameword of §C.1. Hypothesis 18 is assumed to hold and $P_h^0(\mathcal{T}_h)$ is again equipped with the norm defined in [C.26]. To this purpose, for each $T \in \mathcal{T}_h$, for all $F \in \mathcal{F}_h^T$ we let $K_{T,F}$ denote the cone defined by F and x_T (see Figure III.2.3).

Table C.3: Convergence results for the FV method of § III.2.2 with anisotropy ratio of 1.

$1/h$	nunkw	nnmat	erl2	ocverl2	umin	umax
16	255	3001	$3.98e-03$	–	$7.54e-03$	$9.97e-01$
32	1023	12665	$1.00e-03$	$1.99e+00$	$1.02e-03$	$1.00e-00$
64	4095	51961	$2.71e-04$	$1.89e+00$	$2.79e-04$	$1.00e+00$
128	16383	210425	$6.58e-05$	$2.04e+00$	$9.84e-05$	$1.00e-00$

Table C.4: Convergence results for the FV method of § III.2.2 with anisotropy ratio of 1000.

$1/h$	nunkw	nnmat	erl2	ocverl2	umin	umax
16	255	3001	$2.82e-01$	–	$-2.93e-01$	$1.10e+00$
32	1023	12665	$7.98e-02$	$1.82e+00$	$-1.22e-01$	$1.01e+00$
64	4095	51961	$2.00e-02$	$2.00e+00$	$-1.04e-01$	$1.00e+00$
128	16383	210425	$3.94e-03$	$2.34e+00$	$-8.36e-03$	$1.00e-00$

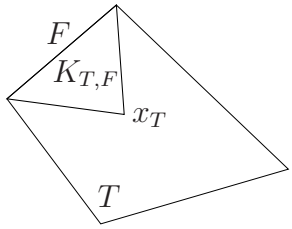


Figure C.2: A face based cone for $d = 2$.

Throughout this section, x_F will denote the barycenter of a face $F \in \mathcal{F}_h$. Thanks to Hypothesis 18, the cones are well-defined and they satisfy

$$|K_{T,F}| = \frac{|F| d_{T,F}}{d}. \tag{C.31}$$

Define the spaces of hybrid unknowns:

$$H_h \stackrel{\text{def}}{=} \mathbb{R}^{\text{card}\mathcal{T}_h \times \text{card}\mathcal{F}_h} = \{\{u_h^T\}_{T\in\mathcal{T}_h}, \{u_h^F\}_{F\in\mathcal{F}_h}\}, \quad H_h^0 \stackrel{\text{def}}{=} \{v_h \in H_h;\, v_h^F = 0,\, \forall F \in \mathcal{F}_h^b\}.$$

For all $h \in \mathcal{H}$, we let $\mathcal{S}_h \stackrel{\text{def}}{=} \{K_{T,F}\}_{(T\in\mathcal{T}_h, F\in\mathcal{F}_h^T)}$ and set

$$V_h \stackrel{\text{def}}{=} H_h^0, \quad \Sigma_h \stackrel{\text{def}}{=} [P_h^0(\mathcal{S}_h)]^d.$$

The space V_h is equipped with the following norm:

$$\|v_h\|_{V_h}^2 \stackrel{\text{def}}{=} \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_h^T} \frac{|F|}{d_{T,F}} (v_h^F - v_h^T)^2.$$

The gradient reconstructions are defined as follows: For all $v_h \in V_h$,

$$G(v_h)_{|K_{T,F}} = \widetilde{G}(v_h)_{|K_{T,F}} = G_T(v_h) + R_{T,F}(v_h)\mu_F^T, \quad \forall K_{T,F} \in \mathcal{S}_h,$$

where we have set

$$G_T(v_h) \stackrel{\text{def}}{=} \frac{1}{|T|} \sum_{F\in\mathcal{F}_h^T} |F|(v_h^F - v_h^T)\mu_F^T, \quad R_{T,F}(v_h) \stackrel{\text{def}}{=} \frac{d^{1/2}}{d_{T,F}} \left(v_h^F - v_h^T - G_T(v_h)\cdot(x_F - x_T)\right).$$

The reconstruction operator $r_h^V : V_h \to P_h^0(\mathcal{T}_h)$ is defined as follows: For all $v_h \in V_h$, $r_h^V v_h = p_h \in P_h^0(\mathcal{T}_h)$ with $p_{h|T} = v_h^T$, for all $T \in \mathcal{T}_h$. The interpolation operator onto V_h is defined as follows: For all $\varphi \in C_c^\infty(\Omega)$, $\pi_h^V \varphi = \varphi_h \in V_h$ with $\varphi_h^T = \varphi(x_T)$ for all $T \in \mathcal{T}_h$, $\varphi_h^F = \varphi(x_F)$. Observe that φ_h belongs to V_h since φ vanishes on the boundary of Ω.

Remark 18 *For all $T \in \mathcal{T}_h$ and for all $\hat{x} \in \mathbb{R}^d$, the following relation holds:*

$$\sum_{F\in\mathcal{F}_h^T} |F|(\mu_F^T)_i (x_F - \hat{x})_j = \delta_{ij}|T|, \tag{C.32}$$

where the ith component of a vector quantity was denoted $(\cdot)_i$ and δ_{ij} is the Kronecker symbol.

Proposition 9 *For all $v_h \in V_h$ and for all $\sigma_h \in [P_h^0(\mathcal{T}_h)]^d$,*

$$\sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_h^T} |K_{T,F}| \sigma_{h|T} \cdot \mu_F^T R_{T,F}(v_h) = 0.$$

PROOF. Using the definition of the residual, we obtain

$$\sum_{T\in\mathcal{T}_h} d^{1/2}\sigma_{h|T}\cdot\left(\sum_{F\in\mathcal{F}_h^T}\frac{|K_{T,F}|}{d_{T,F}}(v_h^F-v_h^T)\mu_F^T-\sum_{F\in\mathcal{F}_h^T}\frac{|K_{T,F}|}{d_{T,F}}G_T(v_h)\cdot(x_F-x_T)\mu_F^T\right).$$

Let S_1 and S_2 the addends in brackets. By definition, $S_1=|T|d^{-1}G_T(v_h)$. On the other hand, [C.31] together with [C.32] yield

$$S_2=-\frac{1}{d}(G_T(v_h))_i\sum_{F\in\mathcal{F}_h^T}|F|(x_F-x_T)_i\mu_F^T=-\frac{|T|}{d}G_T(v_h),$$

and the desired result follows. □

Lemma 34 (Proof of Hypothesis 12) *Hypothesis 12 holds.*

PROOF. The bound [C.7] can be proved as in Lemma 29. In order to prove [C.8], observe that, owing to Proposition 9,

$$\begin{aligned}\|G(v_h)\|_{[L^2(\Omega)]^d}^2&=\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}||G(v_h)|^2\\&=\sum_{T\in\mathcal{T}_h}|T||G_T(v_h)|^2+\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}||R_{T,F}(v_h)|^2\stackrel{\text{def}}{=}S_1+S_2.\end{aligned}$$

For the first term, using [C.31] together with [C.28] we have

$$S_1=\sum_{T\in\mathcal{T}_h}\frac{1}{|T|}\Big|\sum_{F\in\mathcal{F}_h^T}|F|(v_h^F-v_h^T)\mu_F^T\Big|^2\le\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}\frac{|F|d_{T,F}}{|T|}\times\sum_{F\in\mathcal{F}_h^T}\frac{|F|}{d_{T,F}}(v_h^F-v_h^T)^2\le d\|v_h\|_{V_h}^2. \tag{C.33}$$

Substituting the expression of $R_{T,F}$ in the second term yields

$$S_2\le 2\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}\frac{|F|}{d_{T,F}}(v_h^F-v_h^T)^2+\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}||G_T(v_h)|^2\frac{|x_F-x_T|^2}{d_{T,F}^2}\right)\le 2(1+\varrho_5 d)\|v_h\|_{V_h}^2,$$

which proves the assert. □

Lemma 35 (Proof of Hypothesis 13) *Hypothesis 13 holds.*

PROOF. Let $\{v_h\}_{h\in\mathcal{H}}$ be a sequence in V_h satisfying the assumptions of Hypothesis 13. The sequence $\{\widetilde{G}(v_h)\}_{h\in\mathcal{H}}$ is bounded, and it converges (up to a subsequence)

to some $\tau \in [L^2(\Omega)]^d$. It only remains to prove that $\tau = \nabla v$ for a.e. $x \in \mathbb{R}^d$. Let $\Phi \in [C_c^\infty(\mathbb{R}^d)]^d$ and let $\Phi_h = \pi_h^V \Phi$. We have

$$(\widetilde{G}(v_h), \Phi)_{[L^2(\mathbb{R}^d)]^d} = \sum_{T\in\mathcal{T}_h} (G_T(v_h), \Phi)_{[L^2(T)]^d} + \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} (R_{T,F}(v_h)\mu_F^T, \Phi)_{[L^2(K_{T,F})]^d} \stackrel{\text{def}}{=} S_1 + S_2.$$

Integrating by parts the second addend on the left hand side, simple algebraic manipulations yield

$$\begin{aligned}|(G_T(v_h), \Phi)_{[L^2(\Omega)]^d} + (\nabla\cdot\Phi, r_h^V v_h)_{L^2(\Omega)}| &= \Big|\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} |F|(v_h^F - v_h^T)(\Phi_h^F - \Phi_h^T)\cdot\mu_F^T\Big| \\ &\le \|v_h\|_{V_h}\left(\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} |F| d_{T,F}(\Phi_h^F - \Phi_h^T)^2\right)^{\frac12},\end{aligned}$$

which proves that $S_1 \to -(v, \nabla\cdot\Phi)_{L^2(\mathbb{R}^d)}$ as $h \to 0$ since the sequence $\{v_h\}_{h\in\mathcal{H}}$ is bounded in the $\|\cdot\|_{V_h}$ norm. Let us now consider the second term. Owing to the regularity of Φ, there exists $C_\Phi > 0$ only depending on Φ s.t. $|\int_{K_{T,F}}(\Phi - \Phi_h^T)| \le C_\Phi |K_{T,F}| h$. Using Proposition 9 with $\sigma_h = \Phi_h$ and [C.33], Cauchy-Schwarz inequality yields

$$\begin{aligned}S_2 &= \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} (R_{T,F}(v_h)\mu_F^T, \Phi - \Phi_h)_{[L^2(K_{T,F})]^d} \le \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} |R_{T,F}(v_h)|\Big|\int_{K_{T,F}}(\Phi - \Phi_h^T)\Big| \\ &\le \sqrt{2} C_\Phi h |\Omega|^{\frac12}\left(\|v_h\|_{V_h}^2 + d\varrho_5^2\|G_T(v_h)\|_{[L^2(\Omega)]^d}^2\right)^{\frac12} \le \sqrt{2} C_\Phi h |\Omega|^{\frac12}(1 + d\varrho_5)\|v_h\|_{V_h},\end{aligned}$$

which proves that $S_2 \to 0$ as $h \to 0$. □

Since residual terms are incorporated in the gradient reconstruction, the above method can be shown to be stable without further penalization. We thus take $j_h(u_h, v_h) = 0$, which trivially satisfies Hypothesis 14.

Let $\nu_h \in [P_h^0(\mathcal{T}_h)]^{d\times d}$ be s.t., for all $T \in \mathcal{T}_h$, $\nu_{h|T} = \int_T \nu/|T|$.

Lemma 36 (Proof of Hypothesis 15) *Hypothesis 15 holds.*

Proof. Let $v_h \in V_h$. Using Proposition 9 with σ_h s.t. $\sigma_{h|T} = G_T(v_h)$ for all $T \in \mathcal{T}_h$,

$$a_h(v_h, v_h) = \sum_{T\in\mathcal{T}_h} |T|\nu_{h|T} G_T(v_h)\cdot G_T(v_h) + \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T} |K_{T,F}|\nu_{h|T}\mu_F^T\cdot\mu_F^T R_{T,F}(v_h)^2 \stackrel{\text{def}}{=} S_1 + S_2.$$

Clearly, $S_1 \ge \underline{\lambda}\|G_T(v_h)\|_{[L^2(\Omega)]^d}^2$. Observe that, for $\epsilon > 0$ and for all $(a, b) \in \mathbb{R}^2$, $(a-b)^2 \ge \frac{\epsilon}{1+\epsilon}a^2 - \epsilon b^2$. Applying the above inequality with $a = v_h^F - v_h^T$ and $b =$

$G_T(v_h)\cdot(x_F - x_T)$ yields:

$$S_2 \geq \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|v_h\|_{V_h}^2 - \epsilon d\overline{\lambda}\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}||G_T(v_h)|^2\frac{|x_F - x_T|^2}{d_{T,F}^2}$$
$$\geq \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|v_h\|_{V_h}^2 - \epsilon d\varrho_5^2\overline{\lambda}\sum_{T\in\mathcal{T}_h}|T||G_T(v_h)|^2 = \frac{\epsilon\underline{\lambda}}{1+\epsilon}\|v_h\|_{V_h}^2 - \epsilon d\varrho_5^2\overline{\lambda}\|G_T(v_h)\|_{[L^2(\Omega)]^d}^2.$$

Coercivity thus holds for $\epsilon \leq \underline{\lambda}/(d\varrho_5^2\overline{\lambda})$. □

Remark 19 *A coercivity constant independent of the anisotropy ratio $\underline{\lambda}/\overline{\lambda}$ could be derived proceeding as in [45]. We have preferred this shorter proof for brevity of presentation.*

Lemma 37 (Proof of Hypothesis 16) *Hypothesis 16 holds.*

Proof. Let $\varphi \in C_c^\infty(\Omega)$ and set $\varphi_h \stackrel{\text{def}}{=} \pi_h^V\varphi$. Observe that

$$\|\varphi_h\|_{V_h}^2 = \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}|\frac{d}{d_{T,F}^2}(\varphi_h^F - \varphi_h^T)^2$$
$$\leq d\|\nabla\varphi\|_{[L^\infty(\Omega)]^d}^2\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_h^T}|K_{T,F}|\frac{|x_F - x_T|^2}{d_{T,F}^2} \leq d\varrho_5^2\|\nabla\varphi\|_{[L^\infty(\Omega)]^d}^2|\Omega|,$$

i.e. , [C.13] is verified with $\sigma_\varphi = (d|\Omega|)^{1/2}\varrho_5\|\nabla\varphi\|_{[L^\infty(\Omega)]^d}$. The proof of [C.14] is classical and will be omitted (see e.g. [37]). It has been proved in [45, Lemma 4.3] that $\|G(v_h) - \nabla\varphi\|_{[L^\infty(\Omega)]^d} \leq C_\varphi h$, where $C_\varphi > 0$ is a parameter depeding on φ, on d and on the mesh regularity parameters ϱ_i, $i \in \{1\ldots3, 5\}$. As a consequence, $\|G(v_h) - \nabla\varphi\|_{[L^2(\Omega)]^d} \leq |\Omega|^{1/2}\|G(v_h) - \nabla\varphi\|_{[L^\infty(\Omega)]^d}$ tends to zero as $h \to 0$, which concludes the proof. □

Bibliographie

[1] I. Aavatsmark. An introduction to multipoint flux approximations for quadrilateral grids. *Comput. Geosci.*, 6:405–432, 2002.

[2] I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretisation on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comput. Phys.*, 127(1):2–14, 1996.

[3] I. Aavatsmark, T. Barkve, O. Boe, and T. Mannseth. Discretisation on unstrutured grids for inhomogeneous, anisotropic media, part i: Derivation of the methods. *sisc*, 19:1700–1716, 1998.

[4] I. Aavatsmark, T. Barkve, Ω. B e, and T. Mannseth. Discretization on non-orthogonal, curvilinear grids for multi-phase flow. In *Prov. of the 4th European Conf. on the Mathematics of Oil Recovery*, volume D, Røros, Norway, 1994.

[5] I. Aavatsmark, G.T. Eigestad, R.A. Klausen, M.F. Wheeler, and I. Yotof. Convergence of a symmetric MPFA method on quadrilateral grids. *Comput. Geosci.*, 2007.

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

[7] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, J.M. Nordbotten, and E. Ω ian. A new finite volume approach to efficient discretization on challeging grids. In *Proc. SPE 106435, Houston*, 2005.

[8] I. Aavatsmark, G.T. Eigestad, B.T. Mallison, J.M. Nordbotten, and E. Ω ian. A compact multipoint flux approximation method with improved robustness. *Numer. Methods Partial Differential Equations*, 1(31), October 2007.

[9] Eigestad G.T. Aavatsmark I. and Klausen R.A. Numerical convergence of mpfa for general quadrilateral grids in two and three dimensions. *IMA*, 142:1–22, 2006.

[10] L. Agélas, D. A. Di Pietro, and J. Droniou. The G method for anisotropic heterogeneous diffusion. Preprint available at http://hal.archives-ouvertes.fr/hal-00342739/en/, November 2008. Submitted.

[11] L. Agélas and D.A. Di Pietro. A symmetric finite volume scheme for anisotropic heterogeneous second-order elliptic problems. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 705–716. John Wiley & Sons, 2008.

[12] L. Agélas, D.A. Di Pietro, R. Eymard, and R. Masson. An abstract analysis framework for nonconforming approximations of the single phase Darcy equation. Preprint available at http://hal.archives-ouvertes.fr/, June 2008. Submitted.

[13] L. Agélas, D.A. Di Pietro, and R. Masson. A symmetric and coercive finite volume scheme for multiphase porous media flow with applications in the oil industry. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 35–52. John Wiley & Sons, 2008.

[14] L. Agélas, R. Eymard, and R. Herbin. A nine-point finite volume scheme for the simulation of diffusion in heterogeneous media. *C. R. Acad. Sci. Paris, Sér. I*, (347):673–676, June 2009.

[15] L. Agélas and R. Masson. Convergence of the finite volume MPFA O scheme for heterogeneous anisotropic diffusion problems on general meshes. *C. R. Acad. Sci. Paris, Sér. I*, (346):1007–1012, October 2008.

[16] D.N. Arnold. An interior penalty finite element method with discontinuous elements. *SIAM J. Numer. Anal.*, 19:742–760, 1982.

[17] D.N. Arnold, F. Brezzi, B. Cockburn, and D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2002.

[18] Satish Balay, Kris Buschelman, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 2.1.5, Argonne National Laboratory, 2004.

[19] Satish Balay, Kris Buschelman, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang. PETSc Web page, 2001. www.mcs.anl.gov/petsc.

[20] Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

[21] F. Bassi, S. Rebay, G. Mariotti, S. Pedinotti, and M. Savini. A high-order accurate discontinuous finite element method for inviscid and viscous turbomachinery flows. In R. Decuypere and G. Dibelius, editors, *Proceedings of the* 2nd *European Conference on Turbomachinery Fluid Dynamics and Thermodynamics*, pages 99–109, 1997.

[22] S. Brenner. Poincaré-Friedrichs inequalities for piecewise H^1 functions. *SIAM J. Numer. Anal.*, 41(1):306–324, 2003.

[23] H. Brezis. *Analyse fonctionnelle : Théorie et applications*. Dunod, Paris, 2005 edition, 1983.

[24] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of mimetic finite difference methods for diffusion problems on polyhedral meshes. *SIAM J. Numer. Anal.*, 45:1872–1896, 2005.

[25] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of mimetic finite difference methods for diffusion problems on polyhedral meshes with curved faces. *Math. Mod. Meths. Appli. Sci. (M3AS)*, 26:275–298, 2006.

[26] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Math. Mod. Meths. Appli. Sci. (M3AS)*, 15:1533–1553, 2005.

[27] E. Burman and P. Zunino. A domain decomposition method based on weighted interior penalties for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, 44(2):1612–1638, 2006.

[28] Le Potier C. Finite volume scheme for highly anisotropic diffusion operators on unstructured meshes. *C. R. Math. Acad. Sci.*, 340, 2005.

[29] B. Cockburn and C.-W. Shu. The local discontinuous Galerkin finite element method for convection-diffusion systems. *SIAM J. Numer. Anal.*, 35:2440–2463, 1998.

[30] Gunasekera D., Herring J. Childs P., and Cox J. A multi-point flux discretization scheme for general polyhedral grids. In *SPE 48855,Proc. SPE 6th international Oil and Gas Conference and Exhibition*, volume D, China, 1998.

[31] C. Dawson, S. Sun, and M. F. Wheeler. Compatible algorithms for coupled flow and transport. *Comput. Methods Appl. Mech. Engrg.*, 193:2565–2580, 2004.

[32] D.A. Di Pietro and A. Ern. Discrete functional analysis tools for discontinuous Galerkin methods with application to the incompressible Navier-Stokes equations. Preprint available at http://hal.archives-ouvertes.fr/hal-00278925/fr/, May 2008. Submitted.

[33] D.A. Di Pietro, A. Ern, and J.L. Guermond. Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection. *SIAM J. Numer. Anal.*, 46(2):805–831, 2008.

[34] J. Droniou. A density result in Sobolev spaces. *J. Math. Pures Appl.*, 81(7):697–714, 2002.

[35] J. Droniou and R. Eymard. Study of the mixed finite volume method for stokes and navier-stokes equation. *Numer. Methods Partial Differential Equations*, 25(1):137–171, June 2008.

[36] M.G. Edwards and C.F. Rogers. A flux continuous scheme for the full tensor pressure equation. In *Prov. of the 4th European Conf. on the Mathematics of Oil Recovery*, volume D, Røros, Norway, 1994.

[37] A. Ern and J.L. Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, NY, 2004.

[38] A. Ern and J.L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006.

[39] A. Ern and J.L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, 44(6):2363–2388, 2006.

[40] A. Ern and J.L. Guermond. Discontinuous Galerkin methods for Friedrichs' systems. Part III. Multi-field theories with partial coercivity. *SIAM J. Numer. Anal.*, 46(2):776–804, 2008.

[41] A Ern, A. F. Stephansen, and P. Zunino. A discontinuous Galerkin method with weighted averages for advection-diffusion equations with locally small and anisotropic diffusivity. *IMA J. Num. Anal.*, 2008. doi:10.1093/imanum/drm050.

[42] R. Eymard, T. Gallouët, and R. Herbin. *The Finite Volume Method.* Ph. Charlet and J.L. Lions eds, North Holland, 2000.

[43] R. Eymard, T. Gallouët, and R. Herbin. A new finite volume scheme for anisotropic diffusion problems on general grids: convergence analysis. *C. R. Math. Acad. Sci.*, 344(6):403–406, 2007.

[44] R. Eymard, T. Gallouët, and R. Herbin. Benchmark on anisotropic problems, sushi: a scheme using stabilization and hybrid interfaces for anisotropic heterogeneous diffusion problems. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 801–814. John Wiley & Sons, 2008.

[45] R. Eymard, T. Gallouët, and R. Herbin. Discretization schemes for heterogeneous and anisotropic diffusion problems on general nonconforming meshes, suchi : a scheme using stabilisation and hybrid interfaces. Preprint available at , January 2008. to appear in IMAJNA.

[46] R. Eymard and R. Herbin. A new colocated finite volume scheme for the icompressible Navier-Stokes equations on general non matching grids. *C. R. Math. Acad. Sci.*, 344(10):659–662, 2007.

[47] R. Eymard, R. Herbin, and J.C. Latché. Convergence analysis of a colocated finite volume scheme for the incompressible Navier-Stokes equations on general 2D or 3D meshes. *SIAM J. Numer. Anal.*, 45(1):1–36, 2007.

[48] P. Grisvard. *Ordinary and Partial Differential Equations*, volume 564 of *Lecture note in Mathematics*, chapter Smoothness of the solution of a monotonic boundary value problem for a second order elliptic equation in a general domain, pages 135–151. Springer, Berlin, 1976.

[49] P. Grisvard. *Elliptic Problems in Nonsmooth Domains*. John Wiley Sons inc, 1986.

[50] Eigestad G.T. and Klausen R.A. On the convergence of the multi-point flux approximation o-method: Numerical experiments for discontinuous permeability. *Numer. Methods Partial Differential Equations*, 21(6):1079–1098, 2005.

[51] R. Herbin and F. Hubert. Finite volumes for complex applications. benchmark session: finite volume schemes on general grids for anisotropic and heterogeneous diffusion problems, June 2007. http://wwww.latp.univ-mrs.fr/fvca5/.

[52] R. Herbin and F. Hubert. Benchmark on duscretization schemes for anisotropic diffusion problems on general grids for anisotropic heterogeneous diffusion problems. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 659–692. John Wiley & Sons, 2008.

[53] Faille I. Jeannin L. and Gallouët T. How to model compressible two phase flows on hybrid grids. *Oil and Gas Science and Technology*, 55(3):269–279, 2000.

[54] R.A. Klausen and R. Winther. Robust convergence of multipoint flux approximation on rough grids. *Numer. Math.*, 104(3):317–337, 2006.

[55] Y. Kutznetsov and S. Repin. A new mixed finite element method on polygonal and polyhedral meshes. *J. Numer. Math. Mod.*, 18(3):261–278, 2003.

[56] A. Lasis and E. Süli. Poincaré-type inequalities for broken Sobolev spaces. Technical Report 03/10, Oxford University Computing Laboratory, Oxford, England, 2003.

[57] A. Lasis and E. Süli. *hp*-version discontinuous Galerkin finite element method for semilinear parabolic problems. *SIAM J. Numer. Anal.*, 45(4):1544–1569, 2007.

[58] Jenny P. Lee S.H. and Tchelepi H. A finite-volume method with hexahedral multiblock grids for modeling flow in porous media. *Comput. Geosci.*, 6(3):269–277, 2002.

[59] Yotov I. Lipnikov K., Shashkov M. Local flux mimetic finite difference methods. Technical Report Technical Report LA-UR-05-8364, Los Alamos National Laboratory, 2005.

[60] Edwards M.G. Unstructured control-volume distributed full tensor finite volume schemes with flow based grids. *Comput. Geosci.*, 6:433–452, 2002.

[61] Klausen R.A. and Winther R. Convergence of multi-point flux approximations on quadrilateral grids. *Numer. Methods Partial Differential Equations*, 22(6):1438–1454, 2006.

[62] B. Rivière, M.F. Wheeler, and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 8:337–360, 1999.