Arbitrary-order finite volume schemes preserving positivity for diffusion problems on deformed meshes

Julie Patela

Thèse de Doctorat

pour obtenir le grade de

Docteur d'Université Paris Cité

École Doctorale de Sciences Mathématiques de Paris-Centre (ED386)

Spécialité : « Mathématiques Appliquées »

Laboratoire Jacques-Louis Lions

présentée par

Julie PATELA

Sous la direction de **Xavier BLANC**
co-encadrée par **Emmanuel LABOURASSE** et **François HERMELINE**

Arbitrary-order finite volume schemes preserving positivity for diffusion problems on deformed meshes

---

Présentée et soutenue publiquement le 12 décembre 2023 devant un jury composé de :

M. Christophe Le Potier	Directeur de recherche	CEA	Rapporteur
M. Komla Domelevo	Professeur	University of Würzburg	Rapporteur
M. Laurent Desvillettes	Professeur	Université Paris Cité	Président du jury
Mme. Cindy Guichard	Maître de conférences	Sorbonne Université	Membre du jury
M. Franck Boyer	Professeur	Université Toulouse 3	Membre du jury
M. Xavier Blanc	Professeur	Université Paris Cité	Directeur de thèse
M. François Hermeline	Directeur de recherche	CEA	Encadrant de thèse
M. Emmanuel Labourasse	Directeur de recherche	CEA	Encadrant de thèse

Acknowledgements

Je souhaite tout d'abord remercier chaleureusement mon directeur de thèse Xavier Blanc ainsi que mes encadrants Emmanuel Labourasse et François Hermeline dont les conseils éclairés, les encouragements et les enseignements m'ont enrichie sur le plan scientifique, en particulier dans les domaines de l'analyse numérique et la programmation. Leur gentillesse et leur soutien m'ont été précieux tout au long de cette thèse.

Je remercie également Stéphane Del Pino qui a toujours été disponible pour m'aider à surmonter les problèmes de programmation qui se sont présentés.

Je remercie Christophe Le Potier et Komla Domelevo d'avoir accepté d'être rapporteurs de cette thèse et d'avoir relu ce manuscrit de manière approfondie. Mes remerciements vont également aux membres du jury, Cindy Guichard, Laurent Desvillettes et Franck Boyer, ainsi qu'au comité de suivi, Christophe Le Potier et Bruno Després, pour leur présence et leur contribution à l'évaluation de ce travail de recherche.

Je n'oublie pas de remercier toutes les personnes qui ont contribué à l'ambiance positive au sein du centre et qui ont toujours été disponibles en cas de besoin. Merci à Emmanuel, François, Gilles, Stéphane D-P., Jérôme, Xavier, Stéphane J., Nicolas, Pierre, Teddy, Benoît, Christophe, Isabelle, Patricia, Philippe et Pierre. Merci Céline, Agnès et Alain pour votre disponibilité et votre efficacité.

Un grand merci à mes collègues doctorants, Alexiane, Valentin, Axelle, Victor, Paul, Claire, Clément, Romain, Manuel, Simon, Florian, Eloïse, Christina, pour leur partage d'expérience et leur bonne humeur au sein de Teratec. Hélène, je te remercie pour ta gentillesse et ta disponibilité.

Enfin, je souhaite exprimer ma reconnaissance envers mes proches, mes parents et ma soeur, pour leur soutien constant tout au long de ce parcours.

Leurs contributions ont été essentielles à la réalisation de cette thèse. Je suis profondément reconnaissante envers chacun d'entre eux.

Un grand merci à tous !

Résumé

L'objectif de cette thèse est le développement et l'analyse de schémas volumes finis robustes et précis afin d'approcher la solution de l'équation de diffusion sur maillages quelconques avec un coefficient de diffusion qui peut être anisotrope et/ou discontinu. Afin de satisfaire ces propriétés, nos schémas devront préserver la positivité et être d'ordre élevé.

Dans ce manuscrit, nous proposons le premier schéma d'ordre arbitraire préservant la positivité pour la diffusion. Notre démarche est tout d'abord d'étudier le problème en 1D. Dans ce cas le problème de positivité n'apparaît qu'à partir de l'ordre 3. D'autre part, la dimension 1 nous permet de faire l'analyse mathématique de ce problème, notamment une preuve de convergence du schéma à un ordre arbitraire sous une hypothèse de stabilité. Ensuite, nous l'étendons en 2D à l'ordre 2, ce qui permet de nous appuyer sur des schémas connus. Nous avons étudié deux possibilités : un schéma type DDFV (Discrete Duality Finite Volume) que l'on compare à une méthode utilisant des reconstructions polynomiales. Enfin, cela nous permet de développer un schéma monotone d'ordre arbitraire sur maillage quelconque avec un coefficient de diffusion κ qui peut être discontinu et/ou anisotrope. La montée en ordre se fait grâce à une reconstruction polynomiale et la monotonicité s'obtient en se ramenant à une structure de M-matrice, ce qui nous donne des schémas non linéaires.

Chaque schéma est validé par des simulations numériques montrant l'ordre de convergence ainsi que la positivité de la solution obtenue.

Summary

The objective of this thesis is the development and the analysis of robust and accurate finite volume schemes for the approximation of the solution of the diffusion equation on deformed meshes with diffusion coefficient which can be anisotropic and/or discontinuous. To satisfy these properties, our schemes must preserve the positivity and achieve high-order accuracy.

In this manuscript, we propose the first positivity-preserving arbitrary-order scheme for diffusion. Our approach is first to study the problem in 1D. In such a case, the positivity problem only appears for order 3 and higher. The 1D setting allows us to perform the mathematical analysis of this problem, including a proof of convergence of the scheme to an arbitrary order under a stability assumption. We then extend it to 2D at order 2, relying on well-known schemes. We study two possibilities: a DDFV-type scheme (Discrete Duality Finite Volume), which we compare with a method using polynomial reconstruction. Finally, this allows us to develop a monotonic scheme of arbitrary order on any mesh with a κ diffusion coefficient that can be discontinuous and/or anisotropic. Improving the order is achieved through polynomial reconstruction, and monotonicity is obtained by reducing to a M-matrix structure, which gives nonlinear schemes.

Each scheme is validated by numerical simulations showing the order of convergence and the positivity of the solution obtained.

Résumé détaillé

Historiquement, calculer efficacement une approximation précise de la solution des équations de diffusion est d'un intérêt considérable dans divers domaines de la science et de l'ingénierie. L'opérateur de diffusion est d'une importance fondamentale parmi les opérateurs différentiels. Il apparaît dans de nombreuses équations aux dérivées partielles modélisant des modèles physiques tels que, par exemple, la conduction thermique, la diffusion radiative, l'élasticité, la diffusion dans les milieux poreux ou les équations de Navier Stokes. Discrétiser un tel opérateur de manière robuste et précise est un défi majeur.

L'objectif de cette thèse est de proposer une telle discrétisation de l'opérateur de diffusion sur maillages déformés. Des travaux récents ont montré qu'assurer la positivité de la solution numérique améliore la robustesse. Pour de nombreux modèles physiques, cette positivité est d'une grande importance. Pour l'équation de diffusion, cette propriété est appelée monotonicité. En ce qui concerne la précision, l'utilisation de méthodes d'ordre élevé semble appropriée.

Un schéma numérique est dit d'ordre k si l'erreur entre la solution numérique et la solution exacte est proportionnelle au pas du maillage h à la puissance k. L'opérateur de diffusion est bien adapté à la conception de méthodes d'ordre élevé, contrairement à l'opérateur hyperbolique pour lequel des chocs (c'est-à-dire des discontinuités de la solution) peuvent apparaître même pour des données lisses, alors que les solutions de l'opérateur de diffusion sont plus régulières. Cela permet une approximation polynomiale efficace, contrairement aux solutions avec des chocs. Les méthodes d'ordre élevé peuvent donner l'impression d'être moins efficaces que les méthodes d'ordre faible parce qu'elles sont plus gourmandes en temps de calcul que ces dernières pour un maillage donné. Cependant, les méthodes d'ordre inférieur nécessitent beaucoup plus de degrés de liberté (raffinement du maillage) pour atteindre une précision donnée que les méthodes d'ordre supérieur. Ainsi, l'efficacité d'une méthode ne se mesure pas en coût de calcul sur un maillage donné, mais en coût de calcul pour atteindre une erreur donnée, ou en examinant l'erreur obtenue pour un temps de calcul donné.

La monotonicité est l'une des propriétés les plus importantes à respecter pour un schéma numérique. L'absence de monotonicité peut entraîner de graves difficultés, car l'inconnue peut être la température, la densité des variables fluides ou la concentration des espèces chimiques, qui doivent rester positives. La présence de valeurs négatives peut faire échouer les simulations ou conduire à des résultats physiquement irréalistes, ce qui compromet la fiabilité et la précision de la simulation numérique. En outre, dans les applications visées, l'opérateur de diffusion doit, en général, être considéré comme une partie d'un système plus complexe d'équations aux dérivées partielles comprenant éventuellement des opérateurs hyperboliques et des termes sources. Un exemple simple est le système d'Euler avec conduction thermique (voir (1)) pour lequel la positivité de la température est requise.

Toutes les difficultés liées à l'obtention de la positivité sont déjà contenues dans le problème de diffusion que nous allons considérer dans la suite.

Nous nous intéressons aux maillages déformés car les applications que nous envisageons impliquent le couplage du problème de diffusion avec l'hydrodynamique lagrangienne (voir (1)). Dans ce cas, nous devrons résoudre l'équation de diffusion dans un grand code où le maillage est imposé. En outre, pour certaines applications, le maillage est imposé par des contraintes provenant de la physique, comme par exemple les couches sédimentaires en géologie.

De nombreuses études ont été consacrées à cet opérateur. Cependant, les méthodes numériques conventionnelles pour les équations de diffusion se heurtent souvent à la difficulté de préserver des

propriétés physiques importantes, telles que la positivité, tout en conservant une précision d'ordre élevé. Pour résoudre le problème de la monotonicité, on pourrait envisager de tronquer la solution discrète à zéro. Toutefois, cette solution ne serait pas satisfaisante, car la propriété de conservation, tout aussi importante, serait perdue. De nombreux travaux permettent d'obtenir un ordre élevé mais pas la monotonicité

- ▷ Les éléments finis (voir [27, 90]) sont les méthodes les plus populaires pour discrétiser les opérateurs elliptiques ou paraboliques, lorsque la conservation et la monotonicité ne sont pas essentielles. Ces méthodes offrent une base théorique solide, ce qui facilite la conception d'extensions d'ordre élevé. Malheureusement, ces méthodes ne sont généralement pas conservatives et ne préservent pas la monotonicité, à moins que des restrictions sévères sur le maillage ne soient supposées.

- ▷ Les méthodes hybrides d'ordre élevé (HHO, voir [33]) sont des approches numériques permettant de résoudre des équations sur des maillages polytopiques avec une grande précision. Ces méthodes combinent les avantages des méthodes des éléments finis et des volumes finis, en utilisant une formulation hybride. Les méthodes HHO visent à obtenir une convergence d'ordre élevé et une meilleure préservation des propriétés de la solution. Ces approches sont conservatives mais ne garantissent pas la positivité de la solution. L'ordre de convergence de ces méthodes dépend de plusieurs facteurs, tels que le degré des polynômes utilisés pour approximer la solution et les conditions de régularité de la solution.

- ▷ Les méthodes de Galerkin discontinues (voir [34]) utilisent des approximations continues par morceaux et des polynômes de haut degré pour obtenir une bonne précision. Ces méthodes ne préservent pas la monotonicité.

- ▷ Les méthodes des éléments virtuels (voir [6]) sont conçues pour traiter les maillages irréguliers et polygonaux à l'aide de fonctions virtuelles définies localement sur chaque élément. Elles peuvent atteindre un ordre de convergence élevé, en particulier pour les approximations polynomiales de degré élevé et les maillages réguliers. Ces méthodes ne préservent pas la positivité.

Les méthodes qui ne sont pas des méthodes de volumes finis peuvent poser des problèmes lorsqu'elles sont couplées à un code de volumes finis. En effet, il est difficile de coupler les méthodes de volumes finis pour l'hydrodynamique avec d'autres méthodes car, pour les premières, les valeurs sont considérées au centre des mailles alors que pour les autres, elles ne sont pas situées aux mêmes endroits. Pour améliorer l'ordre, les méthodes précédentes nécessitent plus de degrés de liberté par maille, ce qui les rend difficiles à coupler avec d'autres méthodes numériques. Par exemple, pour un ordre supérieur à 2, les méthodes d'éléments finis nécessitent plusieurs degrés de liberté par maille.

- ▷ Les méthodes de volumes finis (voir [48]) sont très populaires, notamment parce qu'elles sont conservatives et compatibles avec les méthodes classiques de discrétisation de la partie hyperbolique des équations.

  - ◇ L'article de Kershaw [64] présentent l'un des premiers schémas de volumes finis consistant pour les équations de diffusion sur des maillages déformés (voir également [86] pour un schéma apparenté). L'idée est d'écrire un schéma de volumes finis standard sur un maillage quadrilatéral de référence et de le transformer afin d'obtenir un schéma de volumes finis sur un maillage déformé. La matrice obtenue est symétrique, mais on peut prouver qu'elle n'est consistante que sur des maillages de parallélogrammes et qu'elle ne préserve pas la monotonicité.
  - ◇ Les schémas diamant (voir par exemple [30]) nous intéressent particulièrement. L'idée de ces schémas est d'utiliser des inconnues secondaires aux nœuds, qui sont calculées avec une méthode d'interpolation utilisant les valeurs aux mailles. Pour une interpolation suffisamment précise, le schéma est convergent d'ordre deux mais ne préserve pas la monotonicité.
  - ◇ La famille des méthodes appelées Schémas de Gradient (voir [38, 39]) comprend notamment les trois méthodes suivantes (DDFV, SUSHI, Mimetic):

- La méthode des volumes finis à dualité discrète (DDFV) a été proposée à l'origine par F. Hermeline (voir [57–62]). Cette méthode utilise également des inconnues secondaires aux nœuds. Contrairement aux schémas diamant, qui utilisent une méthode d'interpolation pour calculer ces inconnues secondaires à partir des inconnues aux mailles, le schéma DDFV les calcule en résolvant un problème de diffusion sur un maillage dual. Ainsi, deux problèmes de diffusion sont résolus, les inconnues secondaires de l'un étant les inconnues principales de l'autre. Le schéma est convergent d'ordre deux, même si le maillage est très déformé et/ou si le rapport d'anisotropie de κ est important. Cette approche donne une matrice symétrique, mais elle ne préserve pas la monotonicité. Nous étudierons une extension de ce schéma dans ce manuscrit.
- Les schémas mimétiques (voir [16,17,66,73,75]) sont conçus pour reproduire au niveau discret certaines des propriétés du système continu. Les flux étant considérés comme des inconnues supplémentaires, le nombre de degrés de liberté est plus important que pour les autres schémas. La matrice obtenue est symétrique. Ce schéma est convergent d'ordre deux mais ne préserve pas la monotonicité (voir [74]). La méthode SUSHI (Scheme Using Stabilization and Harmonic Interfaces) (voir [49]) et la méthode MFV (Mixed Finite Volume) (voir [37]) sont similaires à la méthode mimétique et partagent par conséquent les mêmes propriétés (voir [36]).

◇ L'approximation du flux multipoint (MPFA) (voir [1, 12, 42]) utilise des inconnues secondaires situées sur les faces du maillage. Ces inconnues secondaires sont utilisées pour calculer une approximation consistante du flux et sont éliminées en imposant la continuité du flux à travers chaque face. En fonction des variantes, le schéma peut ne pas converger sur maillages aléatoires ou aboutir à une formulation non coercive. La matrice obtenue est non symétrique. La méthode ne préserve pas la monotonicité (voir [43, 44, 50]).

La plupart des méthodes de volumes finis d'ordre 1 ou 2 vues ci-dessus ne sont consistantes et positives que sur des maillages admissibles (au sens de [48]) et pour des coefficients de diffusion scalaires. Ceci n'est pas suffisant dans notre contexte. De plus, si le coefficient de diffusion est tensoriel, cette propriété est perdue même sur les maillages admissibles. Dans le cas des maillages déformés, de nombreux travaux ont été consacrés à la conception de méthodes monotones depuis les articles fondateurs de [7, 40, 70, 72]. Certains d'entre eux traitent de la discrétisation par éléments finis comme [4, 5, 19, 102]. Elles ont au mieux une précision d'ordre deux. La plupart des méthodes monotones sont des méthodes volumes finis. L'approximation du flux à deux points (TPFA) (voir [47]) est linéaire et préserve la monotonicité, mais n'est pas consistante sur maillages déformés pour une dimension supérieure ou égale à 2. Dans [18], Christophe Buet et Stéphane Cordier montrent qu'un schéma linéaire avec un stencil fixé ne peut pas être monotone. Christophe Le Potier a montré dans [71] que si le stencil peut être arbitrairement grand, alors le schéma peut être linéaire et monotone (et l'ordre de convergence est compris entre un et deux). L'idée de son schéma est de créer des chemins pour trouver des combinaisons de mailles lui permettant d'être positif. Cependant, avoir un stencil de taille arbitraire est difficilement compatible avec le calcul parallèle. Le schéma étudié par Vincent Siess dans [97] est basé sur le maillage de Voronoï, sur lequel le schéma TPFA est appliqué. Ce schéma est consistant, linéaire et convergent d'ordre un (parce que la solution est projetée sur le maillage de Voronoï) et satisfait le principe du maximum. Cependant, le stencil peut être plus grand que pour une méthode standard. Toutes les méthodes que nous décrivons ci-dessous ont un stencil fixe mais sont non linéaires. Cette liste n'est pas exhaustive.

▷ Une extension monotone de la méthode DDFV a été proposée dans [21], mais elle n'est pas compatible avec les conditions aux limites de Neumann et n'est convergente qu'à l'ordre un pour les coefficients de diffusion tensoriels discontinus. Au chapitre 2, nous proposons une méthode DDFV monotone qui corrige ces inconvénients.

▷ Une nouvelle méthodologie d'adaptation, appelée M-Adaptation, qui renforce la monotonicité pour les méthodes de différences finies mimétiques a été décrite dans [55]. Cet article montre comment effectuer la M-adaptation pour la diffusion dans la forme primale et duale sur certaines formes d'éléments afin de garantir le principe du maximum discret.

- ▷ Une extension monotone de MPFA a été proposée dans [25]. Cette méthode utilise une approximation du flux multipoint avec un stencil en diamant et une stratégie de correction non linéaire pour garantir le principe du maximum discret. La formulation est basée sur le fait que le flux des méthodes MPFA peut être divisé en deux parties différentes, une composante d'approximation de flux à deux points (TPFA) et les termes de diffusion croisée. Le schéma est localement conservatif et convergent d'ordre deux.

- ▷ Le schéma de Droniou et Le Potier étudié dans [40] est basé sur l'idée de Bertolazzi et Manzini (voir [7]) qui est de construire un flux consistant à travers une face donnée pour chaque maille contenant cette face. Le schéma utilise des inconnues auxiliaires, dont les positions ne sont pas explicitement définies. Le flux est finalement obtenu par combinaison convexe, dont les coefficients dépendent des inconnues. Le schéma est non linéaire, convergent d'ordre deux et satisfait au principe du maximum. Un schéma similaire est proposé par Sheng et Yuan (voir [94]) où les inconnues auxiliaires sont explicitement situées au milieu des faces.

- ▷ Des schémas non linéaires monotones basés sur une approximation consistante du flux à deux points avec une combinaison convexe mais sans inconnues auxiliaires ont été décrits dans [31, 77–79, 85]. Ces schémas préservent la monotonicité et sont convergents d'ordre deux. Dans le cas d'un maillage fortement déformé, le stencil doit être étendu pour préserver la monotonicité.

- ▷ D'autres contributions proposent des schémas de volumes finis basés sur la même idée, c'est-à-dire définissant un schéma non linéaire pour préserver la monotonicité à l'aide d'une combinaison convexe [96, 104, 110]. Dans l'esprit de [76, 110], le schéma décrit dans [10] est plus simple car il n'utilise que deux flux ponctuels, il préserve la monotonicité et est convergent d'ordre deux.

- ▷ Une autre méthode pour obtenir la monotonicité est décrite dans [51, 52, 105, 106, 111]. Ces articles présentent des schémas non linéaires basés sur une approximation de flux à deux points. La monotonicité est obtenue en écrivant le flux comme un flux à deux points plus un reste et en distribuant la partie positive et négative de ce reste à chaque coefficient. Ces coefficients dépendent donc de l'inconnue, de sorte que le schéma est non linéaire mais préserve la monotonicité.

Comme nous l'avons vu, de nombreuses méthodes ont été proposées pour concevoir des schémas de volumes finis monotones. Cependant, deux d'entre elles, dont l'idée principale est de construire une M-matrice, semblent être les plus utilisées. Dans la première méthode on calcule les flux de part et d'autre de chaque élément. Ensuite, on fait une combinaison convexe de ces flux et on choisit astucieusement les coefficients de la combinaison pour imposer la monotonicité (voir par exemple [10, 96]). La seconde consiste à écrire le flux comme un flux à deux points plus un reste. Puis on écrit ce reste comme la différence entre ses parties positive et négative et on distribue ces deux parties aux coefficients devant chaque inconnue (voir par exemple [51, 52, 105, 106, 111]). Les deux méthodes conduisent à une matrice dont les coefficients dépendent de l'inconnue, ce qui rend le schéma non linéaire.

A notre connaissance, malgré tout le travail déjà effectué, il n'existe pas de méthode monotone d'ordre arbitraire. Cette thèse vise à contribuer à l'état de l'art des méthodes numériques pour les équations de diffusion en se concentrant sur le développement et l'analyse de méthodes monotones d'ordre élevé. L'objectif principal est de concevoir des schémas numériques qui peuvent capturer avec précision le comportement complexe des processus diffusifs tout en assurant la positivité de la solution.

Dans le premier chapitre, on commence par une étude en dimension 1 d'espace. Dans ce cas 1D, les problèmes de monotonicité n'apparaissent qu'à partir de l'ordre 3. En effet, le schéma utilisant des flux à deux points est positif et consistant d'ordre 2. Cependant, dès qu'on s'intéresse à un schéma d'ordre plus élevé, on perd la positivité. On propose une méthode permettant de définir un schéma monotone d'ordre quelconque. L'ordre est obtenu en utilisant une reconstruction polynomiale d'ordre suffisamment élevé pour l'estimation des flux numériques. Puis, on applique une méthode connue pour combiner les flux obtenus de part et d'autre d'un sommet du maillage. Ceci nous permet d'obtenir un flux consistant, qui s'écrit sous la forme d'une différence entre les valeurs aux mailles (flux à deux points), mais dont les coefficients dépendent de l'inconnue. Le schéma est donc non linéaire, et une

méthode de point fixe est nécessaire pour calculer sa solution. Une adaptation du schéma dans le cas d'un coefficient de diffusion discontinu est effectuée, à condition que la discontinuité coïncide avec un sommet du maillage. Le cadre 1D nous permet d'effectuer certaines preuves que nous ne pouvons pas étendre aux dimensions supérieures, en particulier, une preuve de convergence est proposée. Enfin, des résultats numériques sont présentés pour valider cette méthode. Ce travail a été publié dans *Computational and Applied Mathematics* (voir [8]).

Le deuxième chapitre applique notre méthode de monotonicité à deux schémas convergents d'ordre deux. Il donne une extension d'ordre deux de notre méthode 1D à la dimension 2. On détaille la construction de deux schémas monotones 2D de volumes finis, dont la différence réside dans la façon dont le calcul des valeurs aux nœuds est abordé. Nous proposons tout d'abord une extension du schéma DDFV qui respecte la positivité, en dimension 2. La méthode DDFV consiste à utiliser des inconnues intermédiaires aux sommets du maillage (en plus des inconnues aux centres des mailles), et de les calculer en résolvant un problème de diffusion sur un maillage dual. Ce schéma n'est pas positif. Pour corriger ce défaut, on adapte la méthode utilisée en dimension 1. Le deuxième schéma étudié utilise une méthode similaire (positive elle aussi) où les inconnues auxiliaires aux sommets sont calculées par interpolation. Ces schémas prennent en compte le cas d'un coefficient de diffusion tensoriel et/ou discontinu. Pour les coefficients de diffusion discontinus, les faces du maillage doivent suivre la discontinuité. Le schéma est ensuite testé pour vérifier la convergence (à l'ordre 2) et la positivité. Les résultats sont comparés, d'une part, entre les deux schémas, et d'autre part à ceux obtenus avec la méthode DDFV classique (qui, elle, n'est pas positive). Ce travail a été publié dans *Communications in Computational Physics* (voir [9]).

Enfin, le dernier chapitre concerne le problème central de la thèse : construire un schéma d'ordre arbitraire et positif en dimension 2. Ici la méthode proposée en dimension 1 est généralisée : montée en ordre en utilisant une interpolation polynomiale et positivité en combinant les flux consistants de façon à se réduire à des flux à deux points. Ceci permet d'avoir un schéma dont la matrice a une structure de M-matrice, donc un schéma monotone, au prix de la linéarité, et comme précédemment, un point fixe est nécessaire pour résoudre le système correspondant. Le cas d'un coefficient de diffusion discontinu et/ou tensoriel est également étudié, ainsi que le choix du stencil pour la reconstruction polynomiale que nous avons fait pour atteindre une précision d'ordre élevé. Pour les coefficients de diffusion discontinus, les faces du maillage doivent suivre la discontinuité. Nous utilisons des quadratures de Gauss pour calculer les intégrales de surface (flux) ou de volume (second membre). Ces quadratures sont conçues pour être exactes pour des polynômes d'ordre suffisamment élevé. Nous étudions deux schémas, dont la différence réside dans la manière dont le calcul des valeurs des nœuds est abordé. Le premier schéma est basé sur le schéma diamant qui interpole les valeurs des nœuds et le second est basé sur le schéma DDFV et considère les valeurs aux nœuds comme des inconnues. Le schéma de type diamant est le seul que nous ayons mis en œuvre. Les résultats numériques présentés permettent de valider la méthode du schéma diamant en proposant notamment une étude de convergence sur des maillages successivement raffinés. Une propriété remarquable de ce travail est que l'ordre est préservé même si le coefficient de diffusion est anisotrope. Ce travail a été soumis à *Journal of Computational Physics* (voir [11]).

Contents

Introduction

Historically, computing efficiently an accurate approximation of the solution of diffusion equations is of considerable interest in various fields of science and engineering. The diffusion operator is of fundamental importance among differential operators. It appears in many physical models that rely on the solution of partial differential equations, such as, for example, thermal conduction, radiative diffusion (see [84] p. 460), elasticity, diffusion in porous media or Navier Stokes equations. Discretizing the diffusion operator in a robust and accurate way is a major challenge.

The objective of this thesis is to propose a robust and accurate discretization of the diffusion operator on deformed meshes. Recent works have shown that ensuring the positivity of the numerical solution improves robustness. For a lot of physical models, positivity of the solution is of great significance. For the diffusion equation, this property is called monotonicity. Concerning accuracy, using high-order methods seems appropriate.

A numerical scheme is said to be of order k if the error between the numerical solution and the exact solution is proportional to the mesh size h to the power k. The diffusion operator is well suited for the design of high-order methods, in comparison with the hyperbolic operator. Indeed, for the latter, shocks (that is, discontinuities of the solution) may develop even for smooth data, while solutions to the former are more regular. This allows for efficient polynomial approximation, in contrast with shock solutions. High order methods may give the impression of being less efficient than low-order methods because they are more CPU time consuming than the latter for a given mesh. However, low-order methods require much more degrees of freedom (mesh refinement) to reach a given accuracy than high-order ones. Thus, the efficiency of a method is not measured in computational cost on a given mesh but in computational cost to achieve a given error, or by looking at the error obtained for a given computation time.

Monotonicity is one of the most important properties to satisfy for a numerical scheme. The lack of monotonicity can lead to serious difficulties since the unknown can be the temperature, the density of fluid variables or the concentration of chemical species that must remain nonnegative. The presence of negative values can make the simulations fail or lead to physically unrealistic results undermining the reliability and accuracy of the numerical simulation. Besides, in general, the diffusion operator has to be seen as a part of a more complex system of partial differential equations including possibly hyperbolic operators and source terms. A simple example is the Euler system with heat conduction.

$$\begin{cases} \partial_t \rho + \boldsymbol{\nabla} \cdot (\rho \mathbf{u}) = 0, \\ \partial_t (\rho \mathbf{u}) + \boldsymbol{\nabla} \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \boldsymbol{\nabla} p = 0, \\ \partial_t (\rho E) + \boldsymbol{\nabla} \cdot (\rho E \mathbf{u}) + \boldsymbol{\nabla} \cdot (p \mathbf{u}) = \boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} T, \end{cases} \tag{1}$$

where the unknowns are

- E : the total specific fluid energy,
- $e = E - \dfrac{\|\mathbf{u}\|^2}{2}$: the specific internal energy,
- $\rho > 0$: the density of the fluid,
- $p = p(\rho, e)$: the fluid pressure,
- $T = T(\rho,e)$: the temperature,
- $\kappa \geq 0$: the fluid's thermal conductivity coefficient,

▷ $\mathbf{u}$: the fluid velocity.

We are going to check the compatibility of Euler's equations with the second principle of thermodynamics. For this system, the Clausius-Duhem inequality implies

$$\frac{\kappa}{T}\left(\nabla T \cdot \nabla T\right) \geq 0. \tag{2}$$

Details of the computations are given in Appendix A.3.

Hence, since κ is non-negative, (2) is satisfied if and only if $T \geq 0$. In system (1) the positivity of T is also required in order to ensure the existence of p.

All the difficulties involved in achieving positivity are already contained in the following system, that we are going to consider in the sequel

$$\begin{cases} -\boldsymbol{\nabla} \cdot (\kappa \boldsymbol{\nabla} \bar{u}) + \lambda \bar{u} = f & \text{in} \quad \Omega, \\ \bar{u} = g_D & \text{on} \quad \Gamma_D, \\ \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} = g_N & \text{on} \quad \Gamma_N, \end{cases} \tag{3}$$

where Ω is a bounded open domain of $\mathbb{R}^2$ with $\partial\Omega = \Gamma_D \cup \Gamma_N$ ($\Gamma_D \cap \Gamma_N = \emptyset$), and $\mathbf{n} \in \mathbb{R}^2$ is the outgoing unit normal vector. The data are such that $f \in L^2(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in L^2(\Gamma_N)$, $\lambda \in \mathbb{R}^+\backslash\{0\}$, and $\kappa \in L^\infty(\Omega)$. The tensor-valued diffusion coefficient κ is bounded and satisfies the uniform ellipticity condition

$$\forall \mathbf{x} \in \Omega, \forall \xi \in \mathbb{R}^2, \quad \kappa_{\min}\|\xi\|^2 \leq \xi^t \kappa(\mathbf{x}) \xi,$$

where $\kappa_{\min}$ is a positive coefficient. Under the above conditions, one can prove (using Lax-Milgram Lemma in the spirit of [46], Chapter 6) that system (3) has a unique solution in $H^1(\Omega)$ that satisfies a positiveness principle, *i.e.* if $f \geq 0$ and $g \geq 0$, then $\bar{u} \geq 0$.

Let us state the maximum principle.

Theorem 1. *Let Ω be a bounded open domain of $\mathbb{R}^2$, $f \in L^2(\Omega)$, $g \in H^{1/2}(\Omega)$ and assume that $\bar{u} \in H^2(\Omega)$ is the solution of the following problem*

$$\begin{cases} -\boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} \bar{u} + \bar{u} = f & \textit{in} \quad \Omega, \\ \bar{u} = g & \textit{on} \quad \partial\Omega. \end{cases} \tag{4}$$

Then, we have

$$\operatorname{ess\,inf}\left(\operatorname*{ess\,inf}_{\partial\Omega}(g), \operatorname*{ess\,inf}_{\Omega}(f)\right) \leq \bar{u}(\mathbf{x}) \leq \operatorname{ess\,sup}\left(\operatorname*{ess\,sup}_{\partial\Omega}(g), \operatorname*{ess\,sup}_{\Omega}(f)\right), \quad \forall \mathbf{x} \in \Omega.$$

The proof is postponed in Appendix A.1 (see [15]).

For linear cases considered in this thesis, the monotonicity is equivalent to the discrete maximum principle at the continuous level. However, for nonlinear cases, the discrete maximum principle is more restrictive than monotonicity.

We are interested in deformed meshes, as the applications we have in mind involve the coupling of the diffusion problem with Lagrangian hydrodynamics (see (1)). In such a case, we will have to solve the diffusion equation within a large code where the mesh is imposed. Besides, for some applications, the mesh must follow the shape of fundamental features, as for instance the sedimentary layers in geology.

Numerous studies have focused on this operator. However, conventional numerical methods for diffusion equations often face challenges in preserving important physical properties, such as positivity,

while maintaining high-order accuracy. In order to address the problem of monotonicity, one could think of truncating the discrete solution to zero. However, this would not be satisfactory since the equally important property of conservation would be lost. Many works achieve high order but not monotonicity

- ▷ The Finite Elements (see [27,90]) are the most popular methods to discretize elliptic or parabolic operators, when conservation and monotonicity are not essential. These methods offer a strong mathematical background, making it easy to design high-order extensions. Unfortunatly, these methods are in general not conservative and do not preserve monotonicity, unless severe restrictions on the mesh are assumed.

- ▷ The Hybrid High-Order Methods (see [33]) are numerical approaches for solving equations on polytopal meshes with high accuracy. These methods combine the advantages of finite element and finite volume methods, using a hybrid formulation. HHO methods aim to achieve high-order convergence and better preservation of solution properties. Such approaches are conservative but do not guarantee the positivity of the solution. The order of convergence of these methods depends on several factors, such as the degree of the polynomials used to approximate the solution and the regularity conditions of the solution.

- ▷ The Discontinuous Galerkin methods (see [34]) use piecewise continuous approximations and high-degree polynomials to achieve good accuracy. These methods do not preserve monotonicity.

- ▷ The Virtual Elements methods (see [6]) are designed to deal with irregular and polygonal meshes using virtual functions defined locally on each element. They can reach a high order of convergence, especially for high-degree polynomial approximations and regular meshes. These methods do not preserve positivity.

Methods that are not Finite Volume methods can pose problems when coupled with a finite volume code. Indeed, it is difficult to couple finite volume methods for hydrodynamics with other methods because, for the first one, the values are considered at the center of the cells while for the others, they are not taken at the same points. To improve the order, more degrees of freedom per cell are required for the previous methods which makes them difficult to couple with other numerical methods. For example, for order larger than 2, finite element methods require several degrees of freedom per cell.

- ▷ The Finite Volume methods (see [48]) are very popular, especially because they are conservative and compatible with the classical discretization methods of the hyperbolic part of the equations.

  - ◇ The work of Kershaw [64] gives one of the first consistent finite volume scheme for diffusion equations on deformed meshes (see also [86] for a related scheme). The idea of this scheme is to write a standard finite volume scheme in a reference quadrilateral mesh and to transform it in order to obtain a finite volume scheme on a deformed mesh. The matrix obtained is symmetric but it can be proven to be consistent only on parallelogram meshes, and it does not preserve monotonicity.

  - ◇ Diamond schemes (see for example [30]) are of particular interest to us. The idea of such schemes is to use secondary unknowns at nodes, which are computed with an interpolation method involving cell values. For a sufficiently precise interpolation, the scheme is second-order convergent but does not preserve monotonicity.

  - ◇ The family of methods called Gradient Schemes (see [38, 39]) includes the following three methods (DDFV, SUSHI, Mimetic):

    - The Discrete Duality Finite Volume method (DDFV) was originally proposed by F. Hermeline (see [57–62]). This method also uses secondary unknowns at nodes. Unlike diamond schemes, which use an interpolation method to calculate these secondary unknowns from cells unknowns, the DDFV scheme computes them by solving a diffusion problem on a dual mesh. Thus, two diffusion problems are solved, the secondary unknowns of one being the main unknowns of the other. The scheme is second-order

convergent, even if the mesh is very distorted and/or the ratio of anisotropy of κ is important. This approach gives a symmetric matrix, but it does not preserve monotonicity. We will study an extension of this scheme in the present manuscript.

- The Mimetic schemes (see [16, 17, 66, 73, 75]) are designed to reproduce at the discete level some of the properties of the continuous system. Because the fluxes are considered as additional unknowns, the number of degrees of freedom is larger than for other schemes. The matrix obtained is symmetric. This scheme is second-order convergent but does not preserve monotonicity (see [74]). The Scheme Using Stabilization and Harmonic Interfaces (SUSHI) (see [49]) and the Mixed Finite Volume (MFV) methods (see [37]) are similar to the Mimetic method and consequently share the same properties (see [36]).

◇ The Multi-Point Flux Approximation (MPFA) (see [1, 12, 42]) uses secondary unknowns located on the faces of the mesh. These secondary unknowns are used to compute a consistent approximation of the flux and are eliminated by imposing the continuity of the flux accross each face. Depending on the variants, the scheme may not converge on random meshes or may result in a non-coercive formulation. The matrix obtained is non-symmetric. The method does not preserve monotonicity (see [43, 44, 50]).

Most of the previous finite volume methods of order 1 or 2 are consistent and positive only on admissible meshes (in the sense of [48]) and for scalar diffusion coefficient. This is not enough in our context. Besides, for tensor-valued coefficient κ, this property is lost even on admissible meshes. In the case of deformed meshes, a large amount of work has been devoted to design monotonic methods since the seminal papers of [7, 40, 70, 72]. Some of them deal with finite element discretization as [4, 5, 19, 102]. They are at best second-order accurate. Most of the monotonic methods are Finite Volume ones. The Two-Point Flux Approximation (TPFA) (see [47]) is linear and preserves the monotonicity but is not consistent on deformed meshes in dimension larger than 2. In [18], Christophe Buet and Stéphane Cordier show that a linear scheme with a fixed stencil can not be monotonic. Christophe Le Potier showed in [71] that if the stencil can be arbitrary large, then the scheme can be linear and monotonic (and the order of convergence is between one and two). The idea of his scheme is to make paths to find combinations of cells allowing it to be positive. However, having a stencil of arbitrary size is hardly compatible with parallel computing. The scheme studied by Siess in [97] is based on the Voronoï mesh, on which the TPFA scheme is applied. The scheme is consistent, linear and first-order convergent (because the solution is projected on the Voronoï mesh) and satisfies the maximum principle. However, the stencil may be larger than for a standard method. All the methods we describe below have a fixed stencil but are non-linear. This list is not meant to be exhaustive

▷ A monotonic extension of DDFV has been proposed in [21] but is not compatible with Neumann boundary conditions, and is only first-order convergent for discontinuous tensor coefficients κ. In Chapter 2, we propose a monotonic DDFV method which corrects these drawbacks.

▷ A new adaptation methodology, called the M-Adaptation, that enforces the monotonicity for the mimetic finite difference methods has been discribed in [55]. This article shows how to perform the M-adaptation for the diffusion in the primal and the dual form on some shapes of the elements to guarantee the discrete maximum principle.

▷ A monotonic extension of MPFA has been proposed in [25]. This method uses a Multipoint Flux Approximation with a Diamond Stencil and a Non-Linear defect correction strategy to guarantee the Discrete Maximum Principle. The formulation is based on the fact that the flux of MPFA methods can be split into two different parts, a Two Point Flux Approximation (TPFA) component and the Cross-Diffusion Terms. The scheme is locally conservative and second-order convergent.

▷ The scheme of Droniou and Le Potier studied in [40] is based on the idea of Bertolazzi and Manzini (see [7]) which is to construct a consistent flux through the face for each cell containing this face. The scheme uses auxilliary unknowns, the positions of which are not explicitly defined.

The flux is finally obtained by convex combination, the coefficients of which depend on the unknowns. The scheme is nonlinear, is second-order convergent and satisfies the maximum principle. A similar scheme is proposed by Sheng and Yuan (see [94]) where the auxiliary unknowns are explicitly located at the middle of the faces.

- ▷ Monotone nonlinear schemes based on a consistent two-point flux approximation with convex combination but without auxiliary unknowns were described in [31, 77–79, 85]. These schemes preserve monotonicity and are second-order convergent. In the case of a highly deformed mesh, the stencil must be extended to preserve monotonicity.

- ▷ Some others contributions propose finite volume schemes based on the same idea, that is, defining a nonlinear scheme to preserve monotonicity using a convex combination [96, 104, 110]. In the spirit of [76, 110], the scheme described in [10] is simpler because it uses only two point fluxes, it preserves monotonicity and is second-order convergent.

- ▷ Another method to obtain monotonicity is described in [51, 52, 105, 106, 111]. In these papers, non-linear schemes based on a two point flux approximation are presented. The monotonicity is obtained by writting the flux as a two points flux plus a remainder and distributing the positive and negative part of this remainder to each coefficient. These coefficients thus depend on the unknown so that the scheme is non-linear but preserves the monotonicity.

As we have seen, many methods have been proposed to design monotonic finite volume schemes. However, two of them seem to be most widely used and the main idea is to build a M-matrix. The idea of the first one is to calculate the fluxes on either side of each element. Then, one makes a convex combinaison of these fluxes and cleverly chooses the coefficients of the combination to impose monotonicity (see for example [10, 96]). The second one consists in writing the flux as a two-point flux plus a remainder. Then one writes this remainder as the difference between its positive and negative parts, distributes these two parts to the coefficients in front of each unknown (see for example [51, 52, 105, 106, 111]). Both methods lead to a matrix the coefficients of which depend on the unknown, making the scheme non-linear.

To our knowledge, despite all the work already done, there are no monotonic arbitrary order scheme. This thesis aims to contribute to the state-of-the-art in numerical methods for diffusion equations by focusing on the development and analysis of monotonic high order methods. The primary objective is to design numerical schemes that can accurately capture the intricate behavior of diffusive processes while ensuring nonnegativity of the solution.

This manuscript is organized as follows.

1. Chapter 1 deals with the 1D case. In 1D, monotonicity problems appear only at order 3 and higher. The 1D setting enables us to perform some proofs that we are unable to extend in higher dimensions. This chapter begins with an introduction giving a brief state of the art concerning this subject. Then, it details the construction of our monotonic arbitrary order 1D scheme using the finite-volume method. Next, a study of the properties of this scheme is proposed. An adaptation of the scheme in the case of a discontinuous diffusion coefficient is done, with the condition that mesh nodes follow the discontinuity. Finally, numerical results are presented to validate this method. This work has been published in *Computational and Applied Mathematics* (see [8]).

2. Chapter 2 applies our monotonicity method to two second-order convergent schemes. It gives a second-order extension of our 1D method to 2D. After an introduction giving a brief state of the art about 2D monotonic schemes, it details the construction of two monotonic 2D finite-volumes schemes, the difference of which resides in the way the computation of node values is addressed. These schemes take into account the case of a tensor-valued and/or discontinuous coefficient κ. For discontinuous diffusion coefficients, the faces of the mesh must follow the discontinuity. Then, the properties of these schemes are studied. Finally, the numerical results presented validate

these methods by comparing them to an existing scheme that does not preserve monotonicity. This work has been published in *Communications in Computational Physics* (see [9]).

3. Chapter 3 generalizes our 2D schemes of Chapter 2 to arbitrary order. The case of a discontinuous and/or tensor-valued coefficient κ is also studied, together with the choice of the stencil for the polynomial reconstruction that we made to achieve high order accuracy. For discontinuous diffusion coefficients, the faces of the mesh must follow the discontinuity. We use Gauss quadratures to calculate the surface (fluxes) or volume (right-hand side) integrals. These quadratures are designed to be exact for polynomials of sufficiently high order. We study two schemes, the difference of which resides in the way the computation of node values is addressed. The first scheme is based on the diamond scheme which interpolate the node values and the second one is based on the DDFV scheme and consider the node values as unknowns. The properties of both methods are then studied. The diamond type scheme is the only one we implemented. The numerical results presented allow to validate the diamond type method by proposing, in particular, a convergence study on successively refined meshes. A remarkable property of this work is that the order is preserved even if the diffusion coefficient is anisotropic. This work has been submitted to *Journal of Computational Physics* (see [11]).

Chapter 1

Arbitrary order monotonic finite-volume schemes for 1D elliptic problems

This chapter has been published by Springer as an article in Computational and Applied Mathematics (see [8]). Note that the definition of a M-matrix given in this article (see Definition 2.2 of [8]) is not completely rigorous, it has been fixed in the following chapter.

When solving numerically an elliptic problem, it is important in most applications that the scheme used preserves the positivity of the solution. When using finite volume schemes on deformed meshes, the question has been solved rather recently. Such schemes are usually (at most) second order convergent, and nonlinear. On the other hand, many high-order schemes have been proposed, that do not ensure positivity of the solution. In this chapter we propose a very high-order *monotonic* (that is, positivity preserving) numerical method for elliptic problems in 1D. We prove that this method converges to an arbitrary order (under reasonable assumptions on the mesh) and is indeed monotonic. We also show how to handle discontinuous sources or diffusion coefficients, while keeping the order of convergence. We assess the new scheme, on several test problems, with arbitrary (regular, distorted, random) meshes.

1.1 Introduction

In this chapter we consider the following elliptic problem with mixed boundary conditions

$$\begin{cases} -\boldsymbol{\nabla}\cdot(\kappa\boldsymbol{\nabla}\bar{u}) + \alpha\bar{u} = f & \text{in} \quad \Omega, \\ \beta\bar{u} + \gamma\kappa\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n} = g & \text{on} \quad \partial\Omega, \end{cases} \tag{1.1}$$

where Ω is a bounded open domain of $\mathbb{R}^d$ and $\mathbf{n} \in \mathbb{R}^d$ the external unit normal vector, with d the dimension. The data are such that $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$, $\alpha \in \mathbb{R}^+$ (if $\alpha = 0$, then $\beta \neq 0$), and $\kappa \in L^\infty(\Omega)$. The diffusion coefficient κ is bounded and satisfies the ellipticity condition

$$\forall x \in \Omega, \qquad \kappa(x) \geq \kappa_0 > 0. \tag{1.2}$$

Besides, β and γ are functions such that

$$\forall x \in \partial\Omega, \qquad \beta(x) \geq 0, \qquad \gamma(x) \geq 0$$

and they do not vanish at the same point. Under the above conditions, one can prove (see [46]) that system (1.1) has a unique solution in $H^1(\Omega)$. This solution satisfies a positivity principle, i.e. if $f \geq 0$ and $g \geq 0$, then $\bar{u} \geq 0$. For linear problems considered in this work, this property is equivalent to a maximum principle on $\bar{u}$, which can be stated as follows: if the data f_1,f_2 and g_1,g_2 are such that $f_1 \leq f_2$ and $g_1 \leq g_2$, then the associated solutions to (1.1), that we denote by $\bar{u}_1$ and $\bar{u}_2$ respectively, satisfy $\bar{u}_1 \leq \bar{u}_2$ almost everywhere in Ω.

Because system (1.1) is intended to model, for instance, concentration diffusion and thermal conduction, preservation of the positivity principle at the discrete level is highly desirable. An easy way to fix negative values is to truncate the solution to zero. However, it is not appropriate, since it breaks another very important property, which is the conservation. The standard finite volume two-point flux approximation (TPFA, see for example [47]) is positivity preserving (one also says monotonic) but is unfortunately inconsistent on deformed meshes, in dimension $d \geq 2$. For this reason, a great deal of work has been devoted to the design of positivity preserving schemes on general (namely non-κ-orthogonal) meshes over the past two decades. While elliptic problems are often solved using a finite element discretization, all the works we know of on monotonic methods on highly deformed meshes deal with finite volume schemes. Monotonic methods can be designed in the finite-element framework (see [26, 28, 63, 65, 99] among others), but rely on restrictive conditions on the mesh we cannot afford. The finite volume framework is well suited to achieve montonicity because it allows for an easy manipulation of the fluxes. The first works we know of are those of Le Potier [70] and Bertolazzi and Manzini [7]. In such methods, one uses a manipulation of the fluxes that leads to introduce a dependence on the discrete solution in the coefficients of the fluxes, making the scheme non-linear, although (1.1) is linear. Thus, mononicity is in general not equivalent to the maximum principle. In

such methods, one usually introduces secondary unknowns (for instance vertex-located or edge-located unknowns) in addition to the primal (cell-located) unknowns. Among others, important contributions to this field are [10, 76, 110], which propose efficient numerical schemes preserving the positivity of the primary unknowns. In [95], the requirement of positive secondary unknowns is relaxed. In [21], a non-linear solver based on an iterative resolution of two problems is described, the primary unknowns of one problem being the secondary unknowns of the other one. The works [77, 112] explain how to build monotonic schemes without relying on secondary unknowns. In [72, 79, 94], maximum principle preserving schemes are proposed. Cancès and Guichard obtained moreover an entropy diminishing property in [22], introducing the non-linearity directly at the continuous level with a change of variables. Some concepts and proofs about the existence of solutions for these types of scheme can be found in [32,40]. Recent advances in this field are [88,101,108]. All the works mentioned above concern 2D or 3D low-order (that is at most of order 2) numerical methods. Latterly, a third-order accurate monotonic method has been proposed in the Finite volume element (FVEM) context [106].

We are interested in designing a high-order positive scheme (that is at least of order 3). We start, in the present chapter, with the 1D case. Thus, for now on, the system we study is the 1D version of (1.1), that is,

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa\dfrac{d\bar{u}}{dx}\right) + \alpha\bar{u} = f & \text{in} \quad \Omega, \\ \beta\bar{u} + \gamma\kappa\dfrac{d\bar{u}}{dn} = g & \text{on} \quad \partial\Omega, \end{cases} \tag{1.3}$$

and we will suppose that $\Omega =]0,1[$ without loss of generality.

Although this setting is very specific, we believe it can be seen as a first step to tackle the question in higher dimensions. Let us be more precise about the 1D setting: in such a case, the TPFA scheme is actually consistent (and monotonic), contrary to dimensions $d \geq 2$. Thus, the relevant question here is to design a high-order scheme that satisfies the positivity principle. Of course, as one may expect, a naive extension to higher orders of the TPFA scheme gives non-positive schemes. In particular, none of the existing [6, 27, 33, 34] arbitrary high order methods for the problem (1.1) is monotonic. In [32] it is shown how to use Le Potier's trick [72] to obtain monotonic 1D schemes of order greater than 2. But as this method uses a finite difference discretization on Cartesian meshes, it seems hard to extend to general meshes even in 1D. In this chapter we propose a new numerical method that has the following properties:

- it has a provable arbitrarily high order of accuracy, under reasonable stability assumptions;
- it is monotonic;
- it is conservative, and
- it operates on general 1D meshes.

The organization of the chapter is as follows. In Section 1.2 we design a high-order Finite-Volume method by integrating the k-th order Taylor expansion of the unknown. The high-order derivatives of this series are approximated using a polynomial reconstruction of the solution while the degrees of freedom are the *integral mean values* of the solution on the cells. The monotonic behavior of the scheme is enforced using the trick described in [51, 52, 105, 111], which leads to a non-linear resolution. A symmetric version of the scheme is also proposed, allowing to obtain a Local Maximum Preserving (LMP) structure (see for instance [40] for a definition) for the fluxes. In Section 1.3, we prove the properties of the method: conservation, consistency of the fluxes at order k, monotonicity (or the LMP structure for the symmetric version) and convergence of the scheme. On this aspect, our analysis is not completely satisfactory. A first approach consists in applying the fairly general analysis performed in [92], using the assumption that the matrix of the scheme is coercive. This is what we do in Proposition 1.3.21 of Subsection 1.3.4.3, proving convergence at order k in L^2-norm. Unfortunately, we do not know how to prove that the matrix is coercive. Therefore, we propose a different approach, in which we replace such a coercivity assumption by a form of stability that is more general (see Assumption 1.3.17 of Subsection 1.3.4.1, and Proposition 1.3.18). We still do not know how to prove such an assumption, and Proposition 1.3.18 only gives convergence at order $k-1$ in L^1-norm. Finally

in Section 1.4 we verify the properties previously stated on 1D test problems, showing that the method is indeed monotonic and of order k in L^2-norm for the solution and the fluxes.

In all the chapter, C will denote an unspecified strictly positive constant independent of the mesh size.

1.2 High-order finite volume scheme

Consider a mesh of Ω whose cells are numbered from 1 to n. The center of cell i is denoted by x_i and its two vertices are $x_{i-\frac{1}{2}}$ and $x_{i+\frac{1}{2}}$. The length of cell i is h_i and the length between the centers x_i and x_{i+1} is $h_{i+\frac{1}{2}}$, see Figure 1.1. Without loss of generality, we will suppose that

$$x_i < x_{i+1}, \forall i \in [\![1,n-1]\!], \tag{1.4}$$

so that $\Omega =]x_{\frac{1}{2}} = 0, x_{n+\frac{1}{2}} = 1[$. We will also assume that the mesh is *quasi-uniform* that is there exists C such that

$$\max_{1\leq i\leq n}(h_i) < C \min_{1\leq i\leq n}(h_i). \tag{1.5}$$

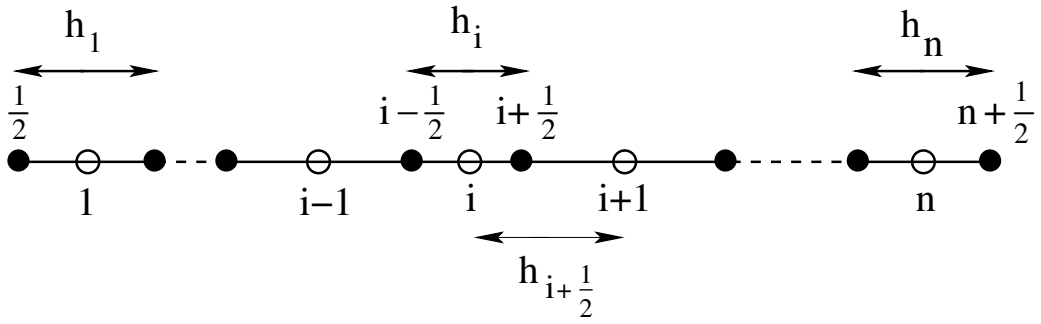


Fig. 1.1 – Definition of the mesh: i denotes the cells and $i+\frac{1}{2}$ the nodes.

We define $h = \max_{1\leq i\leq n}(h_i)$ and $\mathbf{u} = (u_i)_{1\leq i\leq n}$. The notation $\mathbf{u} > \mathbf{0}$ (resp. $\mathbf{u} \geq \mathbf{0}$) means that

$$u_i > 0, \text{ (resp. } u_i \geq 0) \ \forall i \in [\![1,n]\!].$$

Let us introduce some notations for the norms we are going to use. We first define the L^p norm, $p \in [1, +\infty[$

$$\begin{aligned} \|\cdot\|_{L^p} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \left(\sum_{i=1}^{n} h_i |u_i|^p\right)^{1/p} \end{aligned} \tag{1.6}$$

and the L^∞ norm

$$\begin{aligned} \|\cdot\|_{L^\infty} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \max_{1\leq i\leq n} |u_i|. \end{aligned} \tag{1.7}$$

Finally the H^1 norm

$$\begin{aligned} \|\cdot\|_{H^1} : \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{u} &\longmapsto \sqrt{\sum_{i=1}^{n-1} \frac{(u_{i+1}-u_i)^2}{h_{i+\frac{1}{2}}} + \sum_{i=1}^{n} h_i |u_i|^2}. \end{aligned} \tag{1.8}$$

Remark 1.2.1. *Note that* (1.6) *is an L^p-norm for grid functions. Defining* $u(x) = \sum_{i=1}^{n} u_i \mathbb{1}_{[i-\frac{1}{2}, i+\frac{1}{2}]}(x)$, *we have*

$$\|\mathbf{u}\|_{L^p} = \left(\int_\Omega |u(x)|^p dx \right)^{1/p}.$$

1.2.1 Finite volume formulation

In this section, $\kappa(x)$ is assumed to be a continuous function. The extension to discontinuous κ is explained in Section 1.3.5. From now on we note $\kappa_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}})$ and $\bar{\mathbf{u}} \in \mathbb{R}^n$ the vector defined by

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \bar{u}(x) dx. \tag{1.9}$$

Let $\bar{u} \in \mathcal{C}^{k+1}(\overline{\Omega})$. The first step to design a finite volume scheme consists in integrating (1.3) on cell i

$$-\left[\kappa_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx} \right)_{i+\frac{1}{2}} - \kappa_{i-\frac{1}{2}} \left(\frac{d\bar{u}}{dx} \right)_{i-\frac{1}{2}} \right] + \alpha h_i \bar{u}_i = h_i f_i,$$

with

$$f_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx. \tag{1.10}$$

Thus we need to define the fluxes

$$\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx} \right)_{i+\frac{1}{2}} \quad \text{and} \quad \bar{\mathcal{F}}_{i-\frac{1}{2}} = \kappa_{i-\frac{1}{2}} \left(\frac{d\bar{u}}{dx} \right)_{i-\frac{1}{2}}.$$

First of all, the Taylor expansion at order k in the neighborhood of $x_{i+\frac{1}{2}}$ gives

$$\forall x \in \overline{\Omega}, \quad \bar{u}(x) = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left((x - x_{i+\frac{1}{2}})^{k+1} \right). \tag{1.11}$$

In order to have mean values as degrees of freedom we integrate (1.11) from $x_{i+\frac{1}{2}}$ to $x_{i+\frac{3}{2}}$ and divide by h_{i+1}

$$\frac{1}{h_{i+1}} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \bar{u}(x) dx = \bar{u}(x_{i+\frac{1}{2}}) + \frac{1}{h_{i+1}} \sum_{\ell=1}^{k} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) dx + \mathcal{O}\left(h_{i+1}^{k+1} \right),$$

that is to say

$$\bar{u}_{i+1} = \bar{u}(x_{i+\frac{1}{2}}) + \frac{1}{h_{i+1}} \sum_{\ell=1}^{k} \left[\frac{(x - x_{i+\frac{1}{2}})^{\ell+1}}{(\ell+1)!} \right]_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_{i+1}^{k+1} \right),$$

namely

$$\bar{u}_{i+1} = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{h_{i+1}^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_{i+1}^{k+1} \right).$$

In a similar way, by integrating (1.11) from $x_{i-\frac{1}{2}}$ to $x_{i+\frac{1}{2}}$ we obtain

$$\bar{u}_i = \bar{u}(x_{i+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{(-1)^\ell h_i^\ell}{(\ell+1)!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h_i^{k+1} \right).$$

The difference between these last two equalities gives, using (1.5)

$$\bar{u}_{i+1} - \bar{u}_i = h_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \sum_{\ell=2}^{k} \frac{h_{i+1}^{\ell} - (-1)^{\ell} h_i^{\ell}}{(\ell+1)!} \frac{d^{\ell}\bar{u}}{dx^{\ell}}(x_{i+\frac{1}{2}}) + \mathcal{O}\left(h^{k+1}\right),$$

from which we obtain, using (1.5) again

$$\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) = \frac{1}{h_{i+\frac{1}{2}}}(\bar{u}_{i+1} - \bar{u}_i - \sum_{\ell=2}^{k} \frac{h_{i+1}^{\ell} + (-1)^{\ell+1} h_i^{\ell}}{(\ell+1)!} \frac{d^{\ell}\bar{u}}{dx^{\ell}}(x_{i+\frac{1}{2}})) + \mathcal{O}\left(h^{k}\right). \tag{1.12}$$

Let $\mathbf{u} = (u_i)_{1 \leq i \leq n}$ be the numerical solution. By mimicking the expression of the exact flux (1.12) the numerical flux is defined by

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right), \tag{1.13}$$

with

$$r_{i+\frac{1}{2}}(\mathbf{u}) = -\frac{1}{h_{i+\frac{1}{2}}} \sum_{\ell=2}^{k} \frac{h_{i+1}^{\ell} + (-1)^{\ell+1} h_i^{\ell}}{(\ell+1)!} \frac{d^{\ell}P}{dx^{\ell}}(x_{i+\frac{1}{2}}), \tag{1.14}$$

where P is a polynomial interpolation of u as we will see in the next section.

Remark 1.2.2. *For $k = 1$ (linear approximation of the fluxes), the remainder $r_{i+\frac{1}{2}}(\mathbf{u})$ vanishes, and the classical second-order accurate TPFA scheme is recovered.*

1.2.2 High-order reconstruction by interpolation

In the calculation of the fluxes, it is necessary to evaluate the derivatives of u in $x_{i+\frac{1}{2}}$. In this method, the neighboring cells of $x_{i+\frac{1}{2}}$ are used in order to compute the polynomial reconstruction of the solution by considering that the average of the polynomial in a cell is equal to the average of the solution in this cell.

For a polynomial of degree k, there are $k+1$ coefficients to calculate, so $k+1$ neighboring cells of $x_{i+\frac{1}{2}}$ will be necessary. When it is possible, the stencil will be centered in $x_{i+\frac{1}{2}}$, but the closer $x_{i+\frac{1}{2}}$ is to the boundary, the more the stencil will be shifted in order to stay in the interior of Ω.

The notation $u_0, ..., u_k$ denotes the $k+1$ values of $\mathbf{u}$ used for the calculation. With a small abuse of notation, we denote by $\mathcal{S}_{i+\frac{1}{2}} = \{x_0, ..., x_k\}$ the stencil of the node $x_{i+\frac{1}{2}}$. The polynomial will be of this form

$$P(x) = a_k(u_0, ...,u_k) \left(x - x_{i+\frac{1}{2}}\right)^k + ... + a_0(u_0, ..., u_k).$$

The coefficients of the polynomial $P(x)$ are approximated by

$$\frac{1}{h_j} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} P(x)dx = u_j, \quad \forall j \in [\![0,k]\!].$$

This leads to the following system

$$\underbrace{\begin{pmatrix} 1 & \frac{1}{x_{0+\frac{1}{2}} - x_{0-\frac{1}{2}}} \int_{x_{0-\frac{1}{2}}}^{x_{0+\frac{1}{2}}} x - x_{i+\frac{1}{2}} & \dots & \frac{1}{x_{0+\frac{1}{2}} - x_{0-\frac{1}{2}}} \int_{x_{0-\frac{1}{2}}}^{x_{0+\frac{1}{2}}} (x - x_{i+\frac{1}{2}})^k \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{1}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}} \int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} x - x_{i+\frac{1}{2}} & \dots & \frac{1}{x_{k+\frac{1}{2}} - x_{k-\frac{1}{2}}} \int_{x_{k-\frac{1}{2}}}^{x_{k+\frac{1}{2}}} (x - x_{i+\frac{1}{2}})^k \end{pmatrix}}_{=:M_k} \underbrace{\begin{pmatrix} a_0 \\ \vdots \\ a_k \end{pmatrix}}_{=:\mathbf{a}} = \begin{pmatrix} u_0 \\ \vdots \\ u_k \end{pmatrix}.$$

The matrix M_k can be rewritten

$$M_k = \begin{pmatrix} 1 & \frac{(x_{0+\frac{1}{2}}-x_{i+\frac{1}{2}})^2-(x_{0-\frac{1}{2}}-x_{i+\frac{1}{2}})^2}{2(x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}})} & \cdots & \frac{(x_{0+\frac{1}{2}}-x_{i+\frac{1}{2}})^{k+1}-(x_{0-\frac{1}{2}}-x_{i+\frac{1}{2}})^{k+1}}{(k+1)(x_{0+\frac{1}{2}}-x_{0-\frac{1}{2}})} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{(x_{k+\frac{1}{2}}-x_{i+\frac{1}{2}})^2-(x_{k-\frac{1}{2}}-x_{i+\frac{1}{2}})^2}{2(x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}})} & \cdots & \frac{(x_{k+\frac{1}{2}}-x_{i+\frac{1}{2}})^{k+1}-(x_{k-\frac{1}{2}}-x_{i+\frac{1}{2}})^{k+1}}{(k+1)(x_{k+\frac{1}{2}}-x_{k-\frac{1}{2}})} \end{pmatrix}. \tag{1.15}$$

Proposition 1.2.3. *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (1.4). Let $k\in\mathbb{N}^*$. The matrix M_k defined by (1.15) is invertible.*

Proof. $M_k\mathbf{a}=\mathbf{0}$ means that the integral of the polynomial $P(x)$ vanishes over $k+1$ distinct intervals. Therefore, this polynomial of degree k has at least $k+1$ roots. It is therefore zero, and all the coefficients $a_j, j\in[\![0,k]\!]$, vanish. Thus, this implies that $\mathbf{a}=\mathbf{0}$, so M_k is invertible. □

The exact derivatives can then be approximated by

$$\frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \approx \frac{d^\ell P}{dx^\ell}(x_{i+\frac{1}{2}}), \forall \ell \in [\![2,k]\!].$$

Remark 1.2.4. *A polynomial P is calculated for each node $x_{i+\frac{1}{2}}$. So, the polynomial $P=P_{i+\frac{1}{2}}$ can be different for each node but in order to simplify the notation, we will denote it by P.*

1.2.3 A method to obtain monotonicity

A method borrowed from [51,52,105,111] and developed in the framework of 2D diffusion on arbitrary meshes can be used to make the scheme monotonic. This method has been successfully applied in a recent work [106]. The flux (1.13) can be rewritten as follows

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left(\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r^+_{i+\frac{1}{2}}(\mathbf{u}) - r^-_{i+\frac{1}{2}}(\mathbf{u})\right),$$

with

$$r^+_{i+\frac{1}{2}}(\mathbf{u}) = \frac{|r_{i+\frac{1}{2}}(\mathbf{u})| + r_{i+\frac{1}{2}}(\mathbf{u})}{2} \geq 0 \quad \text{and} \quad r^-_{i+\frac{1}{2}}(\mathbf{u}) = \frac{|r_{i+\frac{1}{2}}(\mathbf{u})| - r_{i+\frac{1}{2}}(\mathbf{u})}{2} \geq 0.$$

Let us assume that $\mathbf{u}>\mathbf{0}$, the flux then reads as

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}}\right)u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i}\right)u_i\right], \tag{1.16}$$

and the coefficients of u_i, u_{i+1} are positive.

1.2.4 Symmetric version

Let us introduce a coefficient $s_{i+\frac{1}{2}}$ depending on $\mathbf{u}$ so that $\mathcal{F}_{i+\frac{1}{2}}$ can be rewritten as

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}}\left[\left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})+s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}}\right)u_{i+1} - \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})+s_{i+\frac{1}{2}}(\mathbf{u})}{u_i}\right)u_i\right]. \tag{1.17}$$

To make the scheme symmetric the coefficients of u_i and u_{i+1} must be equal

$$\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})+s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} = \frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})+s_{i+\frac{1}{2}}(\mathbf{u})}{u_i}, \tag{1.18}$$

which leads to

$$s_{i+\frac{1}{2}}(\mathbf{u}) = \frac{u_i r^+_{i+\frac{1}{2}}(\mathbf{u}) - u_{i+1} r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i}.$$

To preserve positivity, it is necessary to impose

$$\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} = \frac{1}{h_{i+\frac{1}{2}}} + \frac{r_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i} \geq 0,$$

that is to say

$$\frac{\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1} - u_i} \geq 0. \tag{1.19}$$

In other words, $u_{i+1} - u_i$ and $\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})$, defined by (1.13), must have the same sign which seems natural because if $\dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) \geq 0$ (resp. ≤ 0), then $\bar{u}$ is locally non-decreasing (resp. non-increasing) hence $\bar{u}_{i+1} \geq \bar{u}_i$ (resp. $\bar{u}_{i+1} \leq \bar{u}_i$).

In practice, if $\left(\frac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u})\right)(u_{i+1} - u_i) > 0$ we use the numerical flux (1.16), otherwise we use the first order approximation

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}\right). \tag{1.20}$$

1.2.5 Boundary conditions

1.2.5.1 Dirichlet boundary condition

In this section we only give the expression of the boundary conditions. Details are given in Appendix B.1. We consider problem (1.3) with $\beta = 1$, $\gamma = 0$. For the non-symmetric version of the scheme, application of the Dirichlet boundary condition on $x_{n+\frac{1}{2}}$ gives

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n\right], \tag{1.21}$$

and for $x_{\frac{1}{2}}$,

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left[\left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u})}{u_1}\right) u_1 - \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right) g(x_{\frac{1}{2}})\right],$$

For the symmetric version, we obtain

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})}\right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right) u_n\right], \tag{1.22}$$

and for the left boundary, similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left[\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1}\right) u_1 - \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})}\right) g(x_{\frac{1}{2}})\right]. \tag{1.23}$$

1.2.5.2 Neumann boundary condition

Consider problem (1.3) with $\beta = 0$, $\gamma = 1$. For the left ($i = 1$) boundary cell, the flux is

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{\frac{1}{2}} = -\kappa_{\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{\frac{1}{2}} = -g(x_{\frac{1}{2}}) \tag{1.24}$$

while for the right ($i = n$) boundary cell, the flux is

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dx}\right|_{n+\frac{1}{2}} = \kappa_{n+\frac{1}{2}} \left.\frac{d\bar{u}}{dn}\right|_{n+\frac{1}{2}} = g(x_{n+\frac{1}{2}}). \tag{1.25}$$

1.2.5.3 Mixed boundary condition

Consider finally problem (1.3) with $\beta(x) > 0, \gamma(x) > 0, \forall x \in \partial\Omega$. In this case we have for $i = 0$ or $i = n$

$$\bar{u}(x_{i+\frac{1}{2}}) = \frac{1}{\beta(x_{i+\frac{1}{2}})} \left(g(x_{i+\frac{1}{2}}) - \gamma(x_{i+\frac{1}{2}}) \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dn}(x_{i+\frac{1}{2}}) \right). \tag{1.26}$$

Consider first the right boundary of the domain. The adaptation for the left boundary is straightforward. We use the same method as for Dirichlet boundary condition in section 1.2.5.1. Replacing $u_{n+\frac{1}{2}}$ by its expression (1.26) in (1.18) (see also (B.1) in the Appendix) yields

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right) g(x_{n+\frac{1}{2}}) - \beta(x_{n+\frac{1}{2}}) \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) u_n}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}}) \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)}. \tag{1.27}$$

For the left boundary ($i = 0$) we obtain similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \frac{\beta(x_{\frac{1}{2}}) \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) u_1 - \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right) g(x_{\frac{1}{2}})}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}}) \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)}. \tag{1.28}$$

Remark 1.2.5. *In the expression of the fluxes* (1.28) *and* (1.27)*, if we take $\beta = 0$, $\gamma = 1$, we obtain the same fluxes as* (1.24) *and* (1.25)*. Likewise, if we take $\beta = 1, \gamma = 0$, we obtain the same fluxes as* (1.23) *and* (1.22)*.*

1.2.6 Summary of the method and matrix form

The scheme reads as

$$-(\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i u_i = h_i f_i, \tag{1.29}$$

that is, using (1.17),

$$\begin{aligned} &- \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} + \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i \\ &+ \kappa_{i-\frac{1}{2}} \left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^+_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i - \kappa_{i-\frac{1}{2}} \left(\frac{1}{h_{i-\frac{1}{2}}} + \frac{r^-_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_{i-1}} \right) u_{i-1} + \alpha h_i u_i = h_i f_i. \end{aligned}$$

With a more compact notation, we write this as $A\mathbf{u} = A(\mathbf{u})\mathbf{u} = \mathbf{b}(\mathbf{u}) = \mathbf{b}$, with

$$b_i = h_i f_i \quad \forall i \neq \{1,n\},$$

$$A_{ij} = \begin{cases} -\kappa_{i+\frac{1}{2}} \left(\dfrac{1}{h_{i+\frac{1}{2}}} + \dfrac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) & \text{if } j = i+1, \forall i \neq n, \\ \kappa_{i+\frac{1}{2}} \left(\dfrac{1}{h_{i+\frac{1}{2}}} + \dfrac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) & \\ \quad + \kappa_{i-\frac{1}{2}} \left(\dfrac{1}{h_{i-\frac{1}{2}}} + \dfrac{r^+_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_i} \right) + \alpha h_i & \text{if } j = i, \forall i \neq 1,n, \\ -\kappa_{i-\frac{1}{2}} \left(\dfrac{1}{h_{i-\frac{1}{2}}} + \dfrac{r^-_{i-\frac{1}{2}}(\mathbf{u}) + s_{i-\frac{1}{2}}(\mathbf{u})}{u_{i-1}} \right) & \text{if } j = i-1, \forall i \neq 1, \\ 0 & \text{else.} \end{cases} \tag{1.30}$$

The expression of the boundary terms depends on the type of boundary conditions. First, in the case of a Dirichlet boundary condition, we have

$$b_1 = h_1 f_1 + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})} \right) g(x_{\frac{1}{2}}), \tag{1.31}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \tag{1.32}$$

and

$$b_n = h_n f_n + \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})} \right) g(x_{n+\frac{1}{2}}), \tag{1.33}$$

$$A_{n,n} = \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n. \tag{1.34}$$

Next, in the case of a Neumann boundary condition, we have

$$b_1 = h_1 f_1 + g(x_{\frac{1}{2}}), \tag{1.35}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1, \tag{1.36}$$

and

$$b_n = h_n f_n + g(x_{n+\frac{1}{2}}), \tag{1.37}$$

$$A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n. \tag{1.38}$$

Finally, in the case of a mixed boundary condition, we have

$$b_1 = h_1 f_1 + \frac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r_{\frac{1}{2}}^-(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)}{\beta(x_{\frac{1}{2}}) + \gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r_{\frac{1}{2}}^-(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} g(x_{\frac{1}{2}}), \tag{1.39}$$

$$A_{1,1} = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r_{\frac{3}{2}}^-(\mathbf{u}) + s_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \frac{\kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r_{\frac{1}{2}}^+(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right)}{1 + \frac{\gamma(x_{\frac{1}{2}})\kappa_{\frac{1}{2}}}{\beta(x_{\frac{1}{2}})} \left(\frac{2}{h_1} + \frac{r_{\frac{1}{2}}^-(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_{\frac{1}{2}}} \right)} + \alpha h_1, \tag{1.40}$$

and

$$b_n = h_n f_n + \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r_{n+\frac{1}{2}}^+(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)}{\beta(x_{n+\frac{1}{2}}) + \gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r_{n+\frac{1}{2}}^+(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} g(x_{n+\frac{1}{2}}), \tag{1.41}$$

$$A_{n,n} = \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r_{n-\frac{1}{2}}^+(\mathbf{u}) + s_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \frac{\kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r_{n+\frac{1}{2}}^-(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right)}{1 + \frac{\gamma(x_{n+\frac{1}{2}})\kappa_{n+\frac{1}{2}}}{\beta(x_{n+\frac{1}{2}})} \left(\frac{2}{h_n} + \frac{r_{n+\frac{1}{2}}^+(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}} \right)} + \alpha h_n. \tag{1.42}$$

The matrix has been written for the symmetric version of the scheme. For the non-symmetric version, the matrix is the same with $s_{i+\frac{1}{2}}(\mathbf{u}) = s_{i-\frac{1}{2}}(\mathbf{u}) = 0, \forall i \in [\![1,n]\!]$.

Remark 1.2.6. *Assuming that $f \geq 0$ and $g \geq 0$, and that $\mathbf{u} > \mathbf{0}$, the right hand side $\mathbf{b}$ has all its components nonnegative, for any type of boundary conditions.*

Remark 1.2.7. *In the case of mixed boundary condition, the right hand side of the nonlinear system depends on $\mathbf{u}$.*

1.2.7 A fixed point method for handling nonlinearity

The system obtained is of the form $A\mathbf{u} = \mathbf{b}$, A being a matrix dependent on the solution. So, we use a fixed point algorithm (a Picard iteration method) to solve this system as, for instance, in [10,21,40,93]. We start with an initial guess $\mathbf{u}^0$, compute the matrix $A(\mathbf{u}^0)$ and solve $A(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b}$. Repeating this process, we build a sequence $\mathbf{u}^\nu$ that, if it converges, tends to the solution of the scheme. We perform this algorithm until the difference between the solution obtained between two iterations is small enough[a]. To summarize, the following loop is performed

$$\begin{aligned} &\nu = 0 \\ &A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\ &\text{While } \|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_{L_2} > \varepsilon \\ &\quad A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\ &\quad \nu = \nu + 1. \end{aligned} \tag{1.43}$$

Unfortunately, we have no proof of convergence of this algorithm. Nevertheless, the numerical tests we have performed did not provide any situation in which the above fix-point algorithm does not converge.

[a] In the numerical tests, we choose $\varepsilon = 10^{-12}$

Note that, in [40], the authors show that the nonlinear system has a solution. The proof is quite general and can be adapted to our case, but there is no proof of convergence of the fixed point algorithm. In some favorable cases, one can prove the convergence of the fixed point algorithm, *e.g.* if α is large enough (see [10]).

Remark 1.2.8. *We thus have two different schemes: the first one is linear and (expected to be) of high order, as we will see below. It is defined by the fluxes (1.13). Its definition does not require the unknown* $\mathbf{u}$ *to be positive, and its stencil is approximately of size* $k+1$*. The second scheme is nonlinear, and defined by the fluxes (1.16). We need* $\mathbf{u}$ *to be positive in order to define it, and its stencil is equal to* 2*. If it has a (positive) solution, then it is a solution of the linear scheme. Thus, two situations may occur:*

1. *the solution of the linear scheme is positive; then, it is also a solution to the nonlinear scheme;*

2. *the solution of the linear scheme has non-positive entries. Then, the nonlinear scheme* cannot *have a solution. Indeed, such a solution would be positive, hence be solution to the linear scheme. We nevertheless expect the above fix-point algorithm to converge to some* $\mathbf{u}$ *that is non-negative, but is* not *a solution to the nonlinear scheme (nor to the linear scheme).*

However, the solution of the continuous problem (1.3) *satisfies a local maximum principle. Hence, assuming that the solution* $\bar{u}$ *is positive and that the linear scheme converges in the* L^∞ *norm, its solution becomes a positive vector for small enough values of* h*. This situation corresponds to Item 1 above, and the solution of the nonlinear scheme coincides with the solution of the linear scheme. The case of Item 2 happens only for larger values of* h*. In such a case, the monotonicity correction allows to recover positive values of the solution, while giving up, to some extent, the equation defining the linear scheme, at least for points at which the solution to the linear scheme is non-positive. What we observe numerically (see Section 1.4 below) is that the fix-point algorithm always converges, to a "solution"* $\mathbf{u} \geq 0$ *that is an approximation of order* k *to the exact solution* $\bar{u}$*.*

1.2.8 Sketch of the method

We summarize the method as follows.

Initialization

- Initialize $\mathbf{u}^0 > 0$.
- Evaluate κ at the nodes: $\kappa_{i+\frac{1}{2}}, i \in [\![0,n]\!]$; and the mean value of f in each cell: $f_i, i \in [\![1,n]\!]$.

Picard iterations (ν):

Do

- Reconstruct polynomials $P_{i+\frac{1}{2}}, i \in [\![0,n]\!]$, of degree k, in each cells i using the method described in Section 1.2.2.
- Compute the remainder $\mathbf{r}_{i+\frac{1}{2}}(\mathbf{u}), i \in [\![0,n]\!]$ using equation (1.14).
- Distribute the remainder $\mathbf{r}_{i+\frac{1}{2}}(\mathbf{u})$ between cells i and $i+1$ to enforce monotonicity (see Section 1.2.3).
- Possibly, symmetrize the coefficients at each node, using the method of Section 1.2.4.
- Build the matrix $A(\mathbf{u}^\nu)$ and the right-hand side $\mathbf{b}^\nu$ (see Section 1.2.6).
- Solve $A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b}^\nu$.

While $\|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_{L_2} > \varepsilon$.

1.3 Properties

1.3.1 Conservation

Proposition 1.3.1. *Assume that* $\mathbf{u} > \mathbf{0}$ *and consider homogeneous Neumann boundary conditions, then the scheme defined by* (1.29) *is conservative. Indeed it satisfies the equality*

$$\alpha \sum_{i=1}^{n} h_i u_i = \sum_{i=1}^{n} h_i f_i,$$

that is to say

$$\sum_{i=1}^{n} (-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) = 0.$$

Proof. The sum is telescopic so only the boundary terms remain. The homogeneous Neumann boundary condition means that the boundary terms are zero, which leads to

$$\sum_{i=1}^{n} (-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) = 0,$$

that is to say

$$\alpha \sum_{i=1}^{n} h_i u_i = \sum_{i=1}^{n} h_i f_i.$$

The scheme is conservative. □

1.3.2 Monotonicity and Local Maximum Principle (LMP) structure

Consider the definition of an M-matrix (see for instance [87])

Definition 1.3.2. *An* $n \times n$ *matrix* $\mathbf{A}$ *that can be expressed in the forme* $\mathbf{A} = s\mathbf{I} - \mathbf{B}$*, where* $\mathbf{B} = (b_{ij})_{1 \leq i,j \leq n}$ *with* $b_{ij} \geq 0$*,* $1 \leq i,j \leq n$*, and* $s \geq \rho(\mathbf{B})$*, the maximum of the moduli of the eigenvalues of* $\boldsymbol{B}$*, is called an M-matrix.*

We use the following lemma

Lemma 1.3.3. *A matrix* $\mathbf{A} = (A_{ij})_{1 \leq i,j \leq n}$ *is an M-matrix if it satisfies the following inequalities*

$$\forall i \neq j, \quad A_{ij} \leq 0, \quad \textit{and} \quad \forall i, \quad \sum_{j=1}^{n} A_{ij} \geq 0.$$

Moreover, if the last inequality is strict, we say that $\mathbf{A}$ *is a strict M-matrix.*

1.3.2.1 Non-symmetric version: property of the matrix

Proposition 1.3.4. *Assume that* $\mathbf{u} > \mathbf{0}$*, the matrix* $A(\mathbf{u})$ *defined by (1.30) and (1.31) through (1.34), or (1.35) through (1.38), or (1.39) through (1.42) depending on the boundary conditions, with* $s_{i+\frac{1}{2}} = 0$*, is such that* $A^T(\mathbf{u})$ *is a strict M-matrix.*

Remark 1.3.5. *In the following proof we have considered Dirichlet boundary conditions, but the result also holds with other boundary conditions. For mixed boundary conditions, the sum of the first and the last column have also two positive terms. For Neumann boundary conditions, the sum of the first and the last column are also positive but the first term vanishes, that is to say* $\sum_i A_{i,1} = \alpha h_1 > 0$ *and* $\sum_i A_{i,n} = \alpha h_n > 0$.

Proof of Proposition 1.3.4. The matrix satisfies

$$\forall i \neq j, A_{ij}(\mathbf{u}) \leq 0 \quad \text{and} \quad \forall j, \sum_{i=1}^{n} A_{i,j}(\mathbf{u}) > 0.$$

Indeed, for the first column there are only two elements in the sum

$$\sum_i A_{i,1}(\mathbf{u}) = A_{1,1}(\mathbf{u}) + A_{2,1}(\mathbf{u}),$$

which leads to

$$\sum_i A_{i,1}(\mathbf{u}) = \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) - \kappa_{\frac{3}{2}} \left(\frac{1}{h_{\frac{3}{2}}} + \frac{r^-_{\frac{3}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1,$$

that is to say

$$\sum_i A_{i,1} = \kappa_{\frac{1}{2}} \left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) + \alpha h_1 > 0.$$

And for the last column,

$$\sum_i A_{i,n} = A_{n-1,n} + A_{n,n},$$

which leads to

$$\sum_i A_{i,n} = -\kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \kappa_{n-\frac{1}{2}} \left(\frac{1}{h_{n-\frac{1}{2}}} + \frac{r^+_{n-\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n,$$

that is to say

$$\sum_i A_{i,n} = \kappa_{n+\frac{1}{2}} \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) + \alpha h_n > 0.$$

Besides, for other columns

$$\sum_i A_{i,j} = A_{j-1,j} + A_{j,j} + A_{j+1,j},$$

which leads to

$$\begin{aligned} \sum_i A_{i,j} = -\kappa_{(j-1)+\frac{1}{2}} \left(\frac{1}{h_{(j-1)+\frac{1}{2}}} + \frac{r^+_{(j-1)+\frac{1}{2}}(\mathbf{u})}{u_{(j-1)+1}} \right) + \kappa_{j+\frac{1}{2}} \left(\frac{1}{h_{j+\frac{1}{2}}} + \frac{r^-_{j+\frac{1}{2}}(\mathbf{u})}{u_j} \right) + \alpha h_j \\ + \kappa_{j-\frac{1}{2}} \left(\frac{1}{h_{j-\frac{1}{2}}} + \frac{r^+_{j-\frac{1}{2}}(\mathbf{u})}{u_j} \right) - \kappa_{(j+1)-\frac{1}{2}} \left(\frac{1}{h_{(j+1)-\frac{1}{2}}} + \frac{r^-_{(j+1)-\frac{1}{2}}(\mathbf{u})}{u_{(j+1)-1}} \right), \end{aligned}$$

that is to say

$$\sum_i A_{i,j} = \alpha h_j > 0.$$

□

1.3.2.2 Strict monotonicity of the method

Proposition 1.3.6. *Assume that $f \geq 0$, $g \geq 0$, and either $\|f\|_{L^2(\Omega)} > 0$, $g(0) > 0$ or $g(1) > 0$. Assume moreover that $\mathbf{u}^0 > \mathbf{0}$. Then $\forall \nu, \mathbf{u}^\nu > \mathbf{0}$.*

To prove this property, we need to introduce the concept of irreducible matrix. We quote here [98, Definition 1.15].

Definition 1.3.7. *An $n \times n$ matrix A is **reducible** if there exits an $n \times n$ permutation matrix P such that*

$$PAP^T = \begin{bmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{bmatrix},$$

*where $A_{1,1}$ is an $r \times r$ submatrix and $A_{2,2}$ is an $(n-r) \times (n-r)$ submatrix, where $1 \leq r < n$. If no such permutation matrix exists, then A is **irreducible**.*

The matrix of the scheme can be proven to be irreducible in view of the following Lemma (see [98, Theorem 1.17]).

Lemma 1.3.8. *To any $n \times n$ matrix A we associate the graph of nodes 1,2,...,n and of directed edges connecting i to j if $A_{ij} \neq 0$. Then A is irreducible if and only if for any pair $i \neq j$ there exists a chain of edges that allows to go from i to j,*

$$A_{i,k1} \neq 0 \rightarrow A_{k1,k2} \neq 0 \rightarrow \cdots \rightarrow A_{km,j} \neq 0.$$

With these definitions we can make use of the following theorem (see [98], Corollary 3.20).

Theorem 1.3.9. *If A is an irreducible strict M-matrix, then it is invertible and $\forall i,j : (A^{-1})_{ij} > 0$.*

We are now in position to prove Proposition 1.3.6.

Proof of Proposition 1.3.6. We argue by induction on the index ν. We assume that $\mathbf{u}^\nu > 0$. Thus $A^T(\mathbf{u}^\nu)$ is a strict M-matrix (see Proposition 1.3.4). It is easy to check that $A^T(\mathbf{u}^\nu)$ is also irreducible. Thus all the entries of $A^{-T}(\mathbf{u}^\nu)$ are positive, using Theorem 1.3.9, and consequently all the entries of $A^{-1}(\mathbf{u}^\nu)$ are positive. Using Remark 1.2.6, we know that all components of $\mathbf{b}$ are non-negative. Moreover, because of the assumption that either $\|f\|_{L^2(\Omega)} > 0$, $g(0) > 0$ or $g(1) > 0$, at least one component of $\mathbf{b}$ is non zero. We thus have

$$\forall i \in [\![1,n]\!] : u_i^{\nu+1} = \sum_{j=1}^{n} A_{ij}^{-1} b_j > 0,$$

since all terms of this sum are non-negative, with one at least that is positive. □

Proposition 1.3.6 shows that the condition $\mathbf{u}^\nu > \mathbf{0}$ remains satisfied during the fixed point procedure, which allows to always define $A(\mathbf{u}^\nu)$. It shows moreover, than as long as hypothesis of the Proposition 1.3.6 are satisfied, all the properties requiring $\mathbf{u} > 0$ are verified for every fix point iteration.

1.3.2.3 Symmetric version: LMP structure

Proposition 1.3.10. *Assume that $\mathbf{u} > \mathbf{0}$, the matrix A defined by (1.30) and (1.31) through (1.34), or (1.35) through (1.38), or (1.39) through (1.42), depending on the boundary conditions, is symmetric.*

Proof. Let $x_{i+\frac{1}{2}}$, be an interior vertex of the mesh. If condition (1.19) is satisfied for this vertex, we use the definition of the flux (1.17), then symmetrization condition leads to $A_{i,i+1} = A_{i+1,i}$. Otherwise the flux is defined by (1.20), and once again $A_{i,i+1} = A_{i+1,i}$.

□

Proposition 1.3.11. *Assume that $\mathbf{u} > \mathbf{0}$, let A be defined by (1.30) and (1.31) through (1.34), or (1.35) through (1.38), or (1.39) through (1.42), depending on the boundary conditions, then the matrix A is a strict M-matrix.*

Proof. As for Proposition 1.3.4, it can be proved that the matrix A is the transpose of a strict M-matrix. Besides, A is symmetric, so A is itself a strict M-matrix. □

Definition 1.3.12. *This definition is taken from [40]. We say that a scheme for (1.3) has the local maximum principle structure (LMP structure for short) if it can be written in the form*

$$\forall i \in [\![1,n]\!] : \sum_{j=1}^{n} \lambda_{i,j}(\mathbf{u})(u_i - u_j) + \lambda_{i,\frac{1}{2}}(\mathbf{u})(u_i - u_{\frac{1}{2}}) + \lambda_{i,n+\frac{1}{2}}(\mathbf{u})(u_i - u_{n+\frac{1}{2}}) = f_i h_i, \tag{1.44}$$

for some functions $\lambda_{i,j} : \mathbb{R}^n \to \mathbb{R}^+$ satisfying,

$$\lambda_{1,\frac{1}{2}} > 0, \quad \lambda_{n,n+\frac{1}{2}} > 0, \quad \text{and} \quad \forall i \in [\![1,n-1]\!] : \lambda_{i,i\pm1} > 0. \tag{1.45}$$

Theorem 1.3.13. *Assume that $f \geq 0$, $g \geq 0$, and either $\|f\|_{L^2(\Omega)} > 0$, $g(0) > 0$ or $g(1) > 0$. Let A and $\mathbf{b}$ be defined by (1.30) and (1.31) through (1.34), or (1.35) through (1.38), or (1.39) through (1.42), depending on the boundary conditions. Assume that we have applied the symmetrization procedure defined in Section 1.2.4. Then $A^{-1}\mathbf{b} = \mathbf{u} \geq 0$. If moreover $\alpha = 0$, the scheme has the LMP structure.*

Proof. For interior vertices, we consider two cases:

- if condition (1.19) is satisfied, then the coefficients of the fluxes are defined by (1.18), and we have

$$\lambda_{i+\frac{1}{2}} := \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) = \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u}) + s_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right),$$

  which is positive because of (1.19).

- if condition (1.19) is not satisfied, then the coefficients of the fluxes are defined by (1.20), and

$$\lambda_{i+\frac{1}{2}} := \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}},$$

  which is positive.

Substituting $\lambda_{i+\frac{1}{2}}$ in equation (1.17) and using the definition of the scheme (1.29) with $\alpha = 0$ yields

$$\lambda_{i+\frac{1}{2}}(u_i - u_{i+1}) + \lambda_{i-\frac{1}{2}}(u_i - u_{i-1}) = h_i f_i.$$

In other words, we have (1.44), with $\lambda_{i,i\pm1} = \lambda_{i\pm\frac{1}{2}} > 0$, and $\lambda_{ij} = 0$ if $|i-j| > 1$. The proof is similar for boundary vertices, see equation (B.1). □

In addition to monotonicity, schemes with the LMP structure enjoy local stability properties as the nonoscillating property (Proposition 1.5 of [40]). In the present case, this reads as follows. Let $f = 0$ and $\mathbf{u}$ be a solution to the symmetric scheme; we have $\forall i \in [\![2,n-1]\!]$, $\min(u_{i-1}, u_{i+1}) \leq u_i \leq \max(u_{i-1}, u_{i+1})$, $\min\left(u_{\frac{1}{2}}, u_2\right) \leq u_1 \leq \max\left(u_{\frac{1}{2}}, u_2\right)$), and $\min(u_{n-1}, u_{n+\frac{1}{2}}) \leq u_n \leq \max\left(u_{n-1}, u_{n+\frac{1}{2}}\right)$. Another very interesting property, the preservation of initial bounds (Proposition 1.6 of [40]), holds for the parabolic version of the scheme.

1.3.3 Consistency of the fluxes

In order to state the following result (Proposition 1.3.15), we need to assume that the interpolation matrix M_k defined by (1.15) satisfies some regularity assumption in the limit $h \to 0$. Loosely speaking, we expect column j of M_k to be of order h^j. More precisely, we assume that

$$M_k = N_k \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & h & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & h^k \end{pmatrix}, \tag{1.46}$$

where the matrix N_k converges as $h \to 0$, the limit N_k^0 being invertible:

$$\lim_{h\to 0} N_k = N_k^0, \quad \det(N_k^0) \neq 0. \tag{1.47}$$

Remark 1.3.14. *Assumption (1.46)-(1.47) may be seen as a regularity assumption of the mesh. It is clearly satisfied by a regular mesh, for which an explicit computation gives (1.46), where the matrix N_k does not depend on h.*

We have the following result:

Proposition 1.3.15. *Let $k \in \mathbb{N}^*$ and $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (1.4), (1.5), (1.46) and (1.47) Let $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$. The fluxes defined by (1.13) are consistent of order k. More precisely, the vector $\bar{\mathbf{u}}$ being defined by (1.9), we have*

$$\left|\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})\right| \leq C_1 \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k,$$

where the constant C_1 depends only on k, on the constant C in (1.5) and on the norm of the matrix $(N_k^0)^{-1}$, where N_k^0 appears in (1.46)-(1.47). In particular it does not depend on $\bar{u}$ nor on i.

Proof. Since $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$, a Taylor expansion gives

$$\bar{u}(x) = \sum_{\ell=0}^{k} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!} + \rho(x) = Q(x) + \rho(x),$$

where Q is the k-th order polynomial

$$Q(x) = \sum_{\ell=0}^{k} \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) \frac{(x - x_{i+\frac{1}{2}})^\ell}{\ell!},$$

such that

$$\frac{d^\ell Q}{dx^\ell}(x_{i+\frac{1}{2}}) = \frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}), \quad \forall \ell \in [\![1, k]\!]. \tag{1.48}$$

The remainder ρ satisfies the estimate

$$|\rho(x)| \leq \frac{1}{(k+1)!} \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} \left|x - x_{i+\frac{1}{2}}\right|^{k+1}. \tag{1.49}$$

Applying our expression of the flux to $\bar{\mathbf{u}}$ gives

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{Q}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \kappa_{i+\frac{1}{2}} Q'(x_{i+\frac{1}{2}}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \kappa_{i+\frac{1}{2}} \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + \mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}),$$

where $\mathbf{Q}$ (resp. $\boldsymbol{\rho}$) is the vector defined as $\bar{\mathbf{u}}$ with the function Q (resp. ρ) instead of $\bar{u}$ (see (1.9)). Here, we have used first that the flux is linear, second that it is exact for polynomials of degree k (see Appendix D.3), and finally (1.48) with $\ell = 1$.

Proving the result thus amounts to show that $\left|\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho})\right| \leq Ch^k$. To this end, we write it as follows

$$\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \end{pmatrix} M_k^{-1} \boldsymbol{\rho},$$

and use (1.46)-(1.47)

$$\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho}) = \begin{pmatrix} 0 & h^{-1} & 0 & \dots & 0 \end{pmatrix} N_k^{-1} \boldsymbol{\rho}.$$

It is clear from estimate (1.49) that for each index ℓ, we have

$$|\rho_\ell| \leq C_k \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k+1},$$

where C_k depends only on k and on the constant appearing in (1.5). Hence,

$$\left|\mathcal{F}_{i+\frac{1}{2}}(\boldsymbol{\rho})\right| \leq C_k \left\|N_k^{-1}\right\| \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k.$$

Finally, property (1.47) allows to prove that $\left\|N_k^{-1}\right\|$ is bounded independently of h, at least for h small enough. This concludes the proof. □

Remark 1.3.16. *This proposition can be extended to the boundary fluxes. Indeed, for a Neumann boundary condition, the consistency is obvious and for Dirichlet or mixed boundary conditions, the proof is similar.*

1.3.4 Convergence

Consider again problem (1.3) with $\alpha > 0$, $\beta = 0$, $\gamma = 1$,

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) + \alpha \bar{u} = f & \text{in } \Omega, \\ \kappa \dfrac{d\bar{u}}{dn} = 0 & \text{on } \partial\Omega. \end{cases} \tag{1.50}$$

We will start by proving that the scheme is convergent at order $k-1$ in L^1 norm. Next, this will allow us to prove the convergence of the fluxes at order $k-1$ in L^2 norm.

1.3.4.1 Convergence at the order $k-1$

The scheme reads as

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \quad \forall i \in [\![1,n]\!], \tag{1.51}$$

with $\forall i \in [\![1, n-1]\!]$,

$$\begin{aligned} \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) &= \kappa_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right) \\ &= \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^+_{i+\frac{1}{2}}(\mathbf{u})}{u_{i+1}} \right) u_{i+1} - \kappa_{i+\frac{1}{2}} \left(\frac{1}{h_{i+\frac{1}{2}}} + \frac{r^-_{i+\frac{1}{2}}(\mathbf{u})}{u_i} \right) u_i, \end{aligned} \tag{1.52}$$

and

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = 0. \tag{1.53}$$

In order to state our convergence result, we need the following stability property:

Assumption 1.3.17. *If* $\mathbf{b} \geq 0$ *and* $A\mathbf{u} = \mathbf{b}$*, with* $b_i = h_i f_i, \forall i$*, then* $\forall i, u_i^- \leq C(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))$*, where* u_i^- *is the negative part of* u_i *and* $C > 0$ *a constant independent of* h*,* $\mathbf{b}$ *and* $\mathbf{u}$*.*

This assumption is a stability hypothesis similar to the one presented in Proposition 3.3 of [40].

Note that, if the scheme is convergent of order $\frac{1}{2}$, then Assumption 1.3.17 is satisfied. Let us be more precise: we assume that, denoting by $\bar{u}$ the exact solution and $\mathbf{u}$ the numerical one, we have

$$\|\mathbf{u} - \bar{\mathbf{u}}\|_{L^2} \leq C\sqrt{h}\,(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1)),$$

where the vector $\bar{\mathbf{u}}$ is defined by (1.9), the vector $\mathbf{f}$ is defined by (1.10), and C is a universal constant. Assuming that $f \geq 0$, we have $\bar{u} \geq 0$, and this estimate implies

$$\sum_{u_i<0} h_i \,(u_i - \bar{u}_i)^2 + \sum_{u_i \geq 0} h_i \,(u_i - \bar{u}_i)^2 \leq Ch(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))^2.$$

The second term in the right-hand side is non-negative, and, when $u_i < 0$, $(u_i - \overline{u}_i)^2 = \left(-u_i^- - \overline{u}_i\right)^2 \geq \left(u_i^-\right)^2$. Hence,

$$\sum_{i=1}^{n} h_i \left(u_i^-\right)^2 \leq C^2 h(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))^2.$$

Using (1.5), we infer that $u_i^- \leq C(\|\mathbf{f}\|_{L^2(\Omega)} + g(0) + g(1))$, that is, Assumption 1.3.17.

We now prove the following convergence result.

Proposition 1.3.18 (Convergence at order $k-1$ in L^1 norm)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$ be the exact solution of (1.50) and assume that $\bar{\mathbf{u}} \geq \mathbf{0}$. Let $\mathbf{e} = (\bar{u}_i - u_i)_{1 \leq i \leq n}$, where $\mathbf{u}$ is the solution of the scheme (1.51)-(1.52)-(1.53). Assume that Assumption 1.3.17 is satisfied. Then, we have*

$$\|\mathbf{e}\|_{L^1} \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1},$$

with $\|\cdot\|_{L^1}$ defined by (1.6), and C does not depend on h nor on $\bar{u}$, $\mathbf{u}$.

Proof. On the one hand the numerical flux defined by (1.52) satisfies (1.51) and on the other hand, the exact flux $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})$ satisfies

$$-\bar{\mathcal{F}}_{i+\frac{1}{2}} + \bar{\mathcal{F}}_{i-\frac{1}{2}} + \alpha h_i \bar{u}_i = h_i f_i, \quad \forall i \in [\![1,n]\!].$$

Subtracting (1.51) from this equation gives

$$-(\bar{\mathcal{F}}_{i+\frac{1}{2}} - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})) + (\bar{\mathcal{F}}_{i-\frac{1}{2}} - \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u})) + \alpha h_i(\bar{u}_i - u_i) = 0, \quad \forall i \in [\![1,n]\!].$$

Besides, the consistency of the fluxes gives that there exists a constant $C > 0$ such as

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \quad \forall i \in [\![1,n,]\!] \quad \text{with } |R_{i+\frac{1}{2}}| \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k, \text{ where } k \text{ is the order.} \tag{1.54}$$

These last two equations imply

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i e_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \quad \forall i \in [\![1,n]\!].$$

By choosing $\Delta = \frac{1}{\alpha} \max\limits_{1 \leq i \leq n} \left(\dfrac{R_{i+\frac{1}{2}} - R_{i-\frac{1}{2}}}{h_i} \right) \in \mathbb{R}^+$, that is to say $0 \leq \Delta \leq C \left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}$ such that

$$-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \quad \forall i \in [\![1,n]\!],$$

and adding it to e_i leads to

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i(e_i + \Delta) = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \quad \forall i \in [\![1,n]\!].$$

The flux is not modified since the remainder only involves derivatives (Δ being a constant, it no longer appears in the derivatives)

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e} + \boldsymbol{\Delta}) = \kappa_{i+\frac{1}{2}} \left(\frac{e_{i+1} + \Delta - e_i - \Delta}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{e}) \right) = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}), \quad \forall i \in [\![1,n]\!].$$

The corresponding matrix system writes

$$A(\mathbf{e} + \boldsymbol{\Delta}) = \mathbf{R} + \alpha \mathbf{h} \boldsymbol{\Delta},$$

with

$$(\mathbf{e} + \boldsymbol{\Delta})_i = e_i + \Delta, \quad (\mathbf{R} + \alpha \mathbf{h} \boldsymbol{\Delta})_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}} + \alpha h_i \Delta \geq 0, \quad \forall i \in [\![1,n]\!].$$

Using Assumption 1.3.17, we can deduce that

$$(e_i+\Delta)^- \le \left\| \frac{1}{h_i}\left(-R_{i+\frac{1}{2}}+R_{i-\frac{1}{2}}\right)+\alpha\Delta \right\|_{L^2} \le \left\| \frac{1}{h_i}\left(-R_{i+\frac{1}{2}}+R_{i-\frac{1}{2}}\right) \right\|_{L^2} + \alpha|\Delta| \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}. \tag{1.55}$$

Summing these inequalities over i, we obtain

$$\sum_{i=1}^{n} h_i\,(e_i+\Delta)^- \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}. \tag{1.56}$$

Next, we sum the equalities $-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e})+\mathcal{F}_{i-\frac{1}{2}}(\mathbf{e})+\alpha h_i(e_i+\Delta) = -R_{i+\frac{1}{2}}+R_{i-\frac{1}{2}}+\alpha h_i\Delta$, finding

$$\left|\alpha\sum_{i=1}^{n} h_i\,(e_i+\Delta)\right| \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}+\alpha\Delta \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1},$$

where we have used (1.54) and the above bound on Δ. Since $e_i+\Delta = (e_i+\Delta)^+ - (e_i+\Delta)^-$, this implies

$$\alpha\sum_{i=1}^{n} h_i\,(e_i+\Delta)^+ \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}+\alpha\sum_{i=1}^{n} h_i\,(e_i+\Delta)^-$$

Using (1.56), we conclude that

$$\sum_{i=1}^{n} h_i\,(e_i+\Delta)^+ \le C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^{k-1}. \tag{1.57}$$

Collecting (1.56) and (1.57), we conclude the proof.

□

1.3.4.2 Convergence of the fluxes

Let us denote by $H_M = \{(u_i)_{1\le i\le n}\}$ the set of cell values, $H_E = \{(f_{i+\frac{1}{2}})_{1\le i\le n-1}\}$ the set of node values and consider homogeneous Neumann boundary conditions, that is, for all $\mathbf{f}\in H_E$

$$f_{\frac{1}{2}} = f_{n+\frac{1}{2}} = 0. \tag{1.58}$$

Let us define the scalar products

$$\begin{cases} (\mathbf{u}|\mathbf{v})_{H_M} = \displaystyle\sum_{i=1}^{n} h_i u_i v_i, \\ (\mathbf{f}|\mathbf{g})_{H_E} = \displaystyle\sum_{i=1}^{n-1} h_{i+\frac{1}{2}} f_{i+\frac{1}{2}} g_{i+\frac{1}{2}}, \end{cases} \tag{1.59}$$

and the operators

$$\begin{cases} D : H_M \longrightarrow H_E \text{ defined by} & (D\mathbf{u})_{i+\frac{1}{2}} = \dfrac{u_{i+1}-u_i}{h_{i+\frac{1}{2}}}, \qquad 1\le i\le n-1, \\ D^* : H_E \longrightarrow H_M \text{ defined by} & (D^*\mathbf{f})_i = -\dfrac{f_{i+\frac{1}{2}}-f_{i-\frac{1}{2}}}{h_i}, \quad 1\le i\le n. \end{cases} \tag{1.60}$$

Proposition 1.3.19. *If condition (1.58) is satisfied the operators D and D^* are adjoints of each other, that is to say that $(D\mathbf{u}|\mathbf{f})_{H_E} = (\mathbf{u}|D^*\mathbf{f})_{H_M}$, $\forall \mathbf{u}\in H_M$, $\forall \mathbf{f}\in H_E$.*

Proof. The definition of the scalar product gives

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1} h_{i+\frac{1}{2}}(D\mathbf{u})_{i+\frac{1}{2}} f_{i+\frac{1}{2}},$$

which means

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1}(u_{i+1}-u_i)f_{i+\frac{1}{2}}.$$

The two sums can be separated

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=1}^{n-1} u_{i+1}f_{i+\frac{1}{2}} - \sum_{i=1}^{n-1} u_i f_{i+\frac{1}{2}}.$$

We shift the index of the first sum, which gives

$$(D\mathbf{u}|\mathbf{f})_{H_E} = \sum_{i=2}^{n} u_i f_{i-\frac{1}{2}} - \sum_{i=1}^{n-1} u_i f_{i+\frac{1}{2}}.$$

Then, the sums can be recombined as follows

$$(D\mathbf{u}|\mathbf{f})_{H_E} = u_n f_{n-\frac{1}{2}} - u_1 f_{\frac{3}{2}} - \sum_{i=2}^{n-1} u_i(f_{i+\frac{1}{2}} - f_{i-\frac{1}{2}}).$$

Condition (1.58) allows us to insert the boundary terms which are zero

$$(D\mathbf{u}|\mathbf{f})_{H_E} = u_n(f_{n-\frac{1}{2}}-f_{n+\frac{1}{2}})-u_1(f_{\frac{3}{2}}-f_{\frac{1}{2}})-\sum_{i=2}^{n-1} u_i(f_{i+\frac{1}{2}}-f_{i-\frac{1}{2}}) = -\sum_{i=1}^{n} u_i(f_{i+\frac{1}{2}}-f_{i-\frac{1}{2}}) = (\mathbf{u},D^*\mathbf{f})_{H_M}.$$

Thus, the operators D^* and D are adjoints of each other.

□

Proposition 1.3.20 (Convergence of the fluxes at order $k-1$)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (1.50) and assume that $\bar{u} \geq 0$. Let us denote $\mathbf{r}(\mathbf{e}) \in H_E$ the vector whose components are $r_{i+\frac{1}{2}}(\mathbf{e}), \forall i \in [\![0,n]\!]$ the remainders defined by (1.14) and the vector $\mathbf{e} \in H_M$ defined by $e_i = \bar{u}_i - u_i, \forall i \in [\![1,n]\!]$. Assume that $u_i > 0, \forall i \in [\![1,n]\!]$. Then we have*

$$\|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \bar{\boldsymbol{\mathcal{F}}}\|_{H_E} \leq Ch^{k-1},$$

where $\boldsymbol{\mathcal{F}}(\mathbf{u}) \in H_E$ is defined by $(\boldsymbol{\mathcal{F}}(\mathbf{u}))_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}), \forall i \in [\![0,n]\!]$, with $\mathcal{F}_{i+\frac{1}{2}}$ given by (1.52) and (1.53), and $\bar{\boldsymbol{\mathcal{F}}}$ is defined by $(\bar{\boldsymbol{\mathcal{F}}})_{i+\frac{1}{2}} = \bar{\mathcal{F}}_{i+\frac{1}{2}}$, with $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}), \forall i \in [\![0,n]\!]$.

Proof. The scheme

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{u}) + \alpha h_i u_i = h_i f_i, \quad \forall i \in [\![1,n]\!],$$

can be written as

$$D^*\boldsymbol{\kappa}(D\mathbf{u} + \mathbf{r}(\mathbf{u})) + \alpha\mathbf{u} = \mathbf{f}.$$

Besides, the exact flux $\bar{\mathcal{F}}_{i+\frac{1}{2}} = \kappa_{i+\frac{1}{2}} \dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}), \forall i \in [\![1,n]\!]$ also satisfies

$$-\bar{\mathcal{F}}_{i+\frac{1}{2}} + \bar{\mathcal{F}}_{i-\frac{1}{2}} + \alpha h_i \bar{u}_i = h_i f_i, \quad \forall i \in [\![1,n]\!].$$

Since the fluxes are consistent there exists C such that

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \bar{\mathcal{F}}_{i+\frac{1}{2}} + R_{i+\frac{1}{2}}, \quad \text{with } |R_{i+\frac{1}{2}}| \leq Ch^k, \quad \forall i \in [\![1,n]\!]. \tag{1.61}$$

Thus, we have

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i e_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \quad \forall i \in [\![1,n]\!],$$

that can be written

$$D^*\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))+\alpha\mathbf{e}=D^*\mathbf{R}.$$

Given $\mathbf{v}\in H_M$, we take the scalar product of this equation with $\mathbf{v}$

$$(D^*\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))|\mathbf{v})_{H_M}+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=(D^*\mathbf{R}|\mathbf{v})_{H_M},$$

that is to say

$$(D^*(\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))-\mathbf{R})|\mathbf{v})_{H_M}+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=0.$$

Besides $\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))-\mathbf{R}$ can be rewritten as

$$\kappa_{i+\frac{1}{2}}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))_{i+\frac{1}{2}}-R_{i+\frac{1}{2}}=-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})+\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}})-R_{i+\frac{1}{2}}=-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})+\bar{\mathcal{F}}_{i+\frac{1}{2}},\quad\forall i\in[\![1,n]\!],$$

and $\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u})$ and $\bar{\mathcal{F}}_{i+\frac{1}{2}}$ satisfy (1.58), so $\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))-\mathbf{R}$ satisfies (1.58) too.

Using Proposition 1.3.19 provides

$$(\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))|D\mathbf{v})_{H_E}+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=(\mathbf{R}|D\mathbf{v})_{H_E}.$$

We define $\mathbf{v}\in H_M$ by induction as follow

$$\begin{cases} v_1=0, \\ \\ v_{i+1}=h_{i+\frac{1}{2}}\kappa_{i+\frac{1}{2}}\left(\dfrac{e_{i+1}-e_i}{h_{i+\frac{1}{2}}}+r_{i+\frac{1}{2}}\right)+v_i \qquad \forall i\in[\![1,n-1]\!], \end{cases}$$

whence $D\mathbf{v}=\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))$. We thus have

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2+(\alpha\mathbf{e}|\mathbf{v})_{H_M}=(\mathbf{R}|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e})))_{H_E}.$$

The Cauchy-Schwarz inequality leads to

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2+(\alpha\mathbf{e}|\mathbf{v})_{H_M}\leq\|\mathbf{R}\|_{H_E}\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}. \tag{1.62}$$

Besides, we have

$$(\alpha\mathbf{e}|\mathbf{v})_{H_M}=\alpha\sum_{i=1}^{n}h_ie_iv_i.$$

Replacing v_i by its expression leads to

$$(\alpha\mathbf{e}|\mathbf{v})_{H_M}=\alpha\sum_{i=1}^{n}h_ie_i\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\kappa_{j+\frac{1}{2}}\left(\frac{e_{j+1}-e_j}{h_{j+\frac{1}{2}}}+r_{j+\frac{1}{2}}\right).$$

The Cauchy-Schwarz inequality gives

$$|(\alpha\mathbf{e}|\mathbf{v})_{H_M}|\leq\alpha\sum_{i=1}^{n}h_i|e_i|\left(\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\left(\kappa_{j+\frac{1}{2}}\left(\frac{e_{j+1}-e_j}{h_{j+\frac{1}{2}}}+r_{j+\frac{1}{2}}\right)\right)^2\right)^{1/2}\left(\sum_{j=1}^{i-1}h_{j+\frac{1}{2}}\right)^{1/2},$$

hence

$$|(\alpha\mathbf{e}|\mathbf{v})_{H_M}|\leq\alpha\left(\sum_{i=1}^{n}h_i|e_i|\right)\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}.$$

Inserting this estimate into (1.62), we have

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E}^2 \leq \alpha\left(\sum_{i=1}^n h_i|e_i|\right)\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} + \|\mathbf{R}\|_{H_E}\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E},$$

hence

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} \leq \|\mathbf{R}\|_{H_E} + \alpha\sum_{i=1}^n h_i|e_i|.$$

Equation (1.61) gives

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} \leq Ch^k + \alpha\sum_{i=1}^n h_i|e_i|.$$

Proposition 1.3.18 gives

$$\|\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e}))\|_{H_E} \leq Ch^k + \alpha Ch^{k-1}. \tag{1.63}$$

Recalling that

$$(\boldsymbol{\kappa}(D\mathbf{e}+\mathbf{r}(\mathbf{e})))_{i+\frac{1}{2}} = \mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) = \mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) - \mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}),$$

we infer

$$\|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \bar{\boldsymbol{\mathcal{F}}}\|_{H_E} = \|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \boldsymbol{\mathcal{F}}(\bar{\mathbf{u}}) + \mathbf{R}\|_{H_E} \leq \|\boldsymbol{\mathcal{F}}(\mathbf{u}) - \boldsymbol{\mathcal{F}}(\bar{\mathbf{u}})\|_{H_E} + \|\mathbf{R}\|_{H_E} \leq Ch^{k-1}.$$

So the fluxes are convergent at order $k-1$. □

1.3.4.3 Convergence at order k

Proposition 1.3.21 (Convergence at order k)**.** *Let $k \in \mathbb{N}^*$, $\bar{u} \in \mathcal{C}^{k+1}(\Omega)$ be the exact solution of (1.50) and assume that $\bar{\mathbf{u}} \geq \mathbf{0}$. Let $\mathbf{e} = (\bar{u}_i - u_i)_{1\leq i\leq n}$, where $\mathbf{u}$ is the solution of the scheme (1.52)-(1.53). Assume that the matrix A defining this scheme is uniformly coercive, that is, there exists a constant $C_c > 0$ independent of h such that*

$$\forall \mathbf{x} \in \mathbb{R}^n : \mathbf{x}^T A\mathbf{x} \geq C_c\|D\mathbf{x}\|_{L^2}^2,$$

where the operator D is defined by (1.60). Then, we have

$$\|\mathbf{e}\|_{L^2} \leq C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k,$$

where the constant C does not depend on $\bar{u}$, $\mathbf{u}$, h.

Proof. As in the proof of Proposition 1.3.18, we use the consistency of the flux to obtain that

$$-\mathcal{F}_{i+\frac{1}{2}}(\mathbf{e}) + \mathcal{F}_{i-\frac{1}{2}}(\mathbf{e}) + \alpha h_i e_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \quad \forall i \in [\![1,n]\!].$$

with $|R_{i+\frac{1}{2}}| \leq C\left\|\bar{u}^{(k+1)}\right\|_{L^\infty} h^k$. The corresponding matrix system writes

$$A\mathbf{e} = \mathbf{R},$$

with

$$(\mathbf{e})_i = e_i, \quad (\mathbf{R})_i = -R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}, \quad \forall i \in [\![1,n]\!].$$

Taking line i of the system $A\mathbf{e} = \mathbf{R}$, we multiply it by e_i and sum over i:

$$\mathbf{e}^T A\mathbf{e} = \sum_{i=1}^n \left(-R_{i+\frac{1}{2}} + R_{i-\frac{1}{2}}\right) e_i$$

Using a discrete integration by parts, then the Cauchy-Schwarz inequality, we have:

$$\mathbf{e}^T A\mathbf{e} = \sum_{i=0}^{n-1} R_{i+\frac{1}{2}} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}} \leq \left(\sum_{i=0}^{n-1} h_{i+\frac{1}{2}} R_{i+\frac{1}{2}}^2 \right)^{1/2} \left(\sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2 \right)^{1/2} \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k \|D\mathbf{e}\|_{L^2}.$$

The coercivity condition then gives

$$C_c \|D\mathbf{e}\|_{L^2} \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k.$$

A discrete mean Poincaré inequality, proved in Lemma 10.2 of [48], writes

$$\sum_{i=1}^{n} h_i e_i^2 \leq C \sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2 + \frac{1}{|\Omega|} \left(\sum_{i=1}^{n} h_i e_i \right)^2.$$

Owing to conservativity, we have $\sum_{i=1}^{n} h_i e_i = 0$, hence

$$\|\mathbf{e}\|_{L^2}^2 = \sum_{i=1}^{n} h_i e_i \leq C \sum_{i=0}^{n-1} h_{i+\frac{1}{2}} (D\mathbf{e})_{i+\frac{1}{2}}^2 = C \|D\mathbf{e}\|_{L^2}^2.$$

Thus, we have

$$\|\mathbf{e}\|_{L^2} \leq C \left\| \bar{u}^{(k+1)} \right\|_{L^\infty} h^k,$$

which concludes the proof. □

1.3.4.4 Asymptotic behavior of the symmetry condition

Lemma 1.3.22. *Let $\{x_i\}_{1\leq i\leq n}$ be a mesh satisfying (1.4) and (1.5). Let $k \in \mathbb{N}^*, k > 2$, $\bar{u} \in \mathcal{C}^k(\Omega)$ be the exact solution of (1.50) and assume that $\bar{u} \geq 0$. Let $\mathbf{u} \in \mathbb{R}^n$ be the solution of (1.51), (1.52) and (1.53) and assume that $u_i > 0, \forall i \in [\![1,n]\!]$. Assume moreover that $\frac{d\bar{u}}{dx} \neq 0$ on Ω, then the condition (1.19) is asymptotically fulfilled as $h \to 0$.*

Proof. Proposition 1.3.20 shows that

$$\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) = \frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + O(h^{k-1}),$$

and Proposition 1.3.18 that

$$u_{i+1} - u_i = \bar{u}_{i+1} - \bar{u}_i + O(h^{k-2}) = h_{i+\frac{1}{2}} \left(\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) + O(h) \right).$$

Then since $\frac{d\bar{u}}{dx}(x_{i+\frac{1}{2}}) \neq 0$, for h small enough these two quantities have the same sign. □

1.3.5 The case of discontinuous diffusion coefficient κ

In the case where κ is discontinuous at the node $x_{i+\frac{1}{2}}$, we compute two fluxes $\mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u})$ and $\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u})$. The first one is computed using a Taylor expansion in $[x_i, x_{i+\frac{1}{2}}]$ while the second one is computed via a Taylor expansion on $[x_{i+\frac{1}{2}}, x_{i+1}]$. Thus, we use two polynomial reconstructions, one on the left and the other on the right of $x_{i+\frac{1}{2}}$. For each node, we shift the stencil so that it does not cross the node where the discontinuity is located. Let us denote

$$\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \kappa^R_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_{i+\frac{1}{2}}}{\frac{h_{i+1}}{2}} + r^R_{i+\frac{1}{2}}(\mathbf{u}) \right) \quad \text{and} \quad \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u}) = \kappa^L_{i+\frac{1}{2}} \left(\frac{u_{i+\frac{1}{2}} - u_i}{\frac{h_i}{2}} + r^L_{i+\frac{1}{2}}(\mathbf{u}) \right),$$

with

$$\kappa^R_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}} + \epsilon) \quad \text{and} \quad \kappa^L_{i+\frac{1}{2}} = \kappa(x_{i+\frac{1}{2}} - \epsilon),$$

where $r^R_{i+\frac{1}{2}}(\mathbf{u})$ (resp. $r^L_{i+\frac{1}{2}}(\mathbf{u})$) denotes the remainder associated with the polynomial reconstruction of the solution using the cells located at the right (resp. left) of the node $x_{i+\frac{1}{2}}$.

Thus, the continuous problem imposing the equality of the fluxes (see also Figure 1.9 for an example), we also impose it at the discrete level, that is to say $\mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u})$ which leads to

$$\kappa^R_{i+\frac{1}{2}} \left(\frac{u_{i+1} - u_{i+\frac{1}{2}}}{\frac{h_{i+1}}{2}} + r^R_{i+\frac{1}{2}}(\mathbf{u}) \right) = \kappa^L_{i+\frac{1}{2}} \left(\frac{u_{i+\frac{1}{2}} - u_i}{\frac{h_i}{2}} + r^L_{i+\frac{1}{2}}(\mathbf{u}) \right),$$

which yields

$$u_{i+\frac{1}{2}} = \frac{h_i h_{i+1}}{2(h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i \kappa^R_{i+\frac{1}{2}})} \left[2 \left(\frac{\kappa^R_{i+\frac{1}{2}} u_{i+1}}{h_{i+1}} + \frac{\kappa^L_{i+\frac{1}{2}} u_i}{h_i} \right) + \kappa^R_{i+\frac{1}{2}} r^R_{i+\frac{1}{2}}(\mathbf{u}) - \kappa^L_{i+\frac{1}{2}} r^L_{i+\frac{1}{2}}(\mathbf{u}) \right].$$

Replacing $u_{i+\frac{1}{2}}$ by its expression in $\mathcal{F}^L_{i+\frac{1}{2}}$ or $\mathcal{F}^R_{i+\frac{1}{2}}$ gives

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^L_{i+\frac{1}{2}}(\mathbf{u}) = \mathcal{F}^R_{i+\frac{1}{2}}(\mathbf{u}) = \frac{2\kappa^L_{i+\frac{1}{2}} \kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i \kappa^R_{i+\frac{1}{2}}} \left[(u_{i+1} - u_i) + \frac{1}{2} \left(h_{i+1} r^R_{i+\frac{1}{2}}(\mathbf{u}) + h_i r^L_{i+\frac{1}{2}}(\mathbf{u}) \right) \right],$$

that is

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \tilde{\alpha}_{i+\frac{1}{2}}(u_{i+1} - u_i) + \tilde{r}_{i+\frac{1}{2}}(\mathbf{u}) \tag{1.64}$$

with

$$\tilde{\alpha}_{i+\frac{1}{2}} = \frac{2\kappa^L_{i+\frac{1}{2}} \kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i \kappa^R_{i+\frac{1}{2}}}, \quad \tilde{r}_{i+\frac{1}{2}}(\mathbf{u}) = \frac{h_{i+1}\kappa^L_{i+\frac{1}{2}} \kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i \kappa^R_{i+\frac{1}{2}}} r^R_{i+\frac{1}{2}}(\mathbf{u}) + \frac{h_i \kappa^L_{i+\frac{1}{2}} \kappa^R_{i+\frac{1}{2}}}{h_{i+1}\kappa^L_{i+\frac{1}{2}} + h_i \kappa^R_{i+\frac{1}{2}}} r^L_{i+\frac{1}{2}}(\mathbf{u}).$$

The coefficient $\tilde{\alpha}_{i+\frac{1}{2}}$ being positive, we can achieve monotonicity as in Section 1.2.3 and the symmetrization can be done again for this scheme. Besides, the previous analysis applies to this case. In the case where the condition of symmetrization is not satisfied, the flux (1.64) is replaced by the first-order approximation

$$\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) = \tilde{\alpha}_{i+\frac{1}{2}}(u_{i+1} - u_i).$$

Remark 1.3.23. *For $k = 1$, the remainders $r^L_{i+\frac{1}{2}}(\mathbf{u})$ and $r^R_{i+\frac{1}{2}}(\mathbf{u})$ vanish, and we obtain the classical harmonic mean for the equivalent diffusion coefficient.*

Remark 1.3.24. *In the case of a discontinuous right hand side f, we use the same type of strategy. In such a case, the second derivative of the solution $\bar{u}$ is discontinuous. Thus, the reconstruction is made on each side of the discontinuity.*

1.4 Numerical experiments

Before giving numerical results, we explain how we deal with possibly vanishing Dirichlet boundary conditions. The definition of the nonlinear scheme requires $\mathbf{u} > 0$ (which is enforced by construction, see Prop.1.3.6), and $g(x_{\frac{1}{2}}) > 0$ and $g(x_{\frac{1}{2}}) > 0$ for Dirichlet boundary conditions (see Section 1.2.5.1). However, we want to be able to deal with homogeneous Dirichlet boundary conditions. In order to

circumvent this difficulty, it is possible to add a term proportional to h^k to the denominator in the flux. Let $\epsilon > 0$, the flux (1.21) is given by[b]

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}}) + \epsilon h^k} \right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) u_n \right],$$

Same modification is made if needed for $\mathcal{F}_{\frac{1}{2}}$. We use also a correction to prevent the denominator of (1.19) to be zero. The condition (1.19) is replaced with

$$\left(\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + r_{i+\frac{1}{2}}(\mathbf{u}) \right) (u_{i+1} - u_i) \geq 0.$$

The L^2 norm of the error is computed as

$$e_{L^2} = \left(\sum_{i=1}^{n} h_i |u_i - \bar{u}_i|^2 \right)^{1/2}$$

for the solution, and

$$f_{L^2} = \left(e^2_{L^2} + \sum_{i=0}^{n} h_{i+\frac{1}{2}} |\mathcal{F}_{i+\frac{1}{2}}(\mathbf{u}) - \bar{\mathcal{F}}(x_{i+\frac{1}{2}})|^2 \right)^{1/2} \tag{1.65}$$

for the flux.

Given $\Omega =]0,1[$, κ a diffusion coefficient and g a function defined on $\partial\Omega$, we consider problem (1.3) with $\alpha = 0$, $\beta = 1$, $\gamma = 0$

$$\begin{cases} -\dfrac{d}{dx}\left(\kappa \dfrac{d\bar{u}}{dx}\right) = f & \text{in } \Omega, \\ \bar{u} = g & \text{on } \partial\Omega. \end{cases} \tag{1.66}$$

We will use three types of meshes:

1. Cartesian meshes,
2. deformed meshes, the deformation of which is given by: $x \to x + 0.65x(1-x)(0.5-x)\sin(0.8\pi)$,
3. random meshes, the deformation of which is given by: $x \to x + \dfrac{\eta}{n}$, with $\eta \in [-0.45, 0.45]$, and n the number of cells. Thus, $C = 19$ for inequality (1.5). An example of which with 8 cells being given in Figure 1.2.

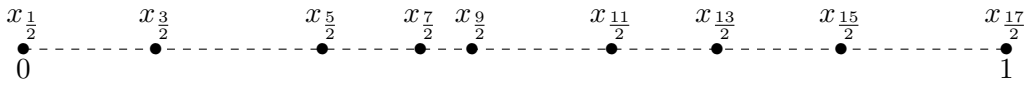


Fig. 1.2 – An example of a random mesh with 8 cells.

Figure 1.3 gives an example of the repartition of the cell volumes for a random mesh with 64 cells.

For all the tests, the ε and $\mathbf{u}^0$ of the fixed-point algorithm (1.43) are $\varepsilon = 10^{-12}$ and $u^0_i = 1, \forall i$. We use the linear solver GMRES with the preconditioner ILU (see [83], Chapter 7.4) and the convergence criterion is 10^{-14}.

[b] In the benchmarks we have chosen $\epsilon = 10^{-11}$.

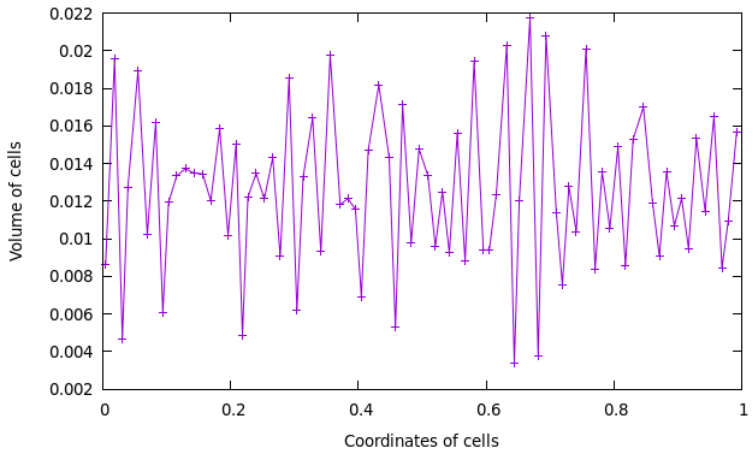


Fig. 1.3 – Example of a repartition of the volume for a random mesh with 64 cells.

1.4.1 L^2 convergence for polynomial solutions

Given $\kappa = 1$, $f(x) = -6x$ (resp. $f(x) = -72x^7$), $g(0) = 1$ and $g(1) = 2$, the function $\bar{u}(x) = x^3 + 1$ (resp. $\bar{u}(x) = x^9 + 1$) is solution to (1.66). We perform a spectral convergence study for these problems on a deformed mesh with 64 cells. The L^2-error between the exact $\bar{u}$ and approximated u solutions are reported in the Table 1.1.

Order	$\bar{u}(x) = x^3 + 1$	$\bar{u}(x) = x^9 + 1$
1	1.64e-04	1.56e-03
2	3.46e-06	7.00e-04
3	4.53e-15	2.70e-04
4	3.79e-15	1.39e-06
5	8.15e-15	7.43e-07
6	2.57e-14	7.07e-09
7	4.21e-15	5.24e-10
8	5.02e-15	6.58e-13
9	7.86e-15	8.17e-15

Tab. 1.1 – The L^2-error between the exact $\bar{u}$ and approximated u solutions.

The proof of exactness for polynomial of degree k (see appendix D.3) shows that the numerical solution must be exact for an order greater than 3 (resp. 9). The Table of convergence 1.1 agrees with the theory since the error is zero, to machine precision, for the order greater than 3 (resp. 9).

1.4.2 L^2 convergence for a smooth diffusion coefficient

Given $\kappa = \exp(x)$, $f(x) = 4\exp(x) + 4x\exp(x) - \pi\cos(\pi x)\exp(x) + \pi^2\exp(x)\sin(\pi x)$ (note that f is positive), $g(0) = 4$ and $g(1) = 2$, the function $\bar{u}(x) = \sin(\pi x) - 2x^2 + 4$ is solution to (1.66). We perform a convergence study for this problem with the non-symmetric and symmetric schemes on the deformed mesh. The L^2-error between the exact $\bar{u}$ and approximated u solutions and f_{L^2} (refer to Equation (1.65)) are reported in Figures 1.4.

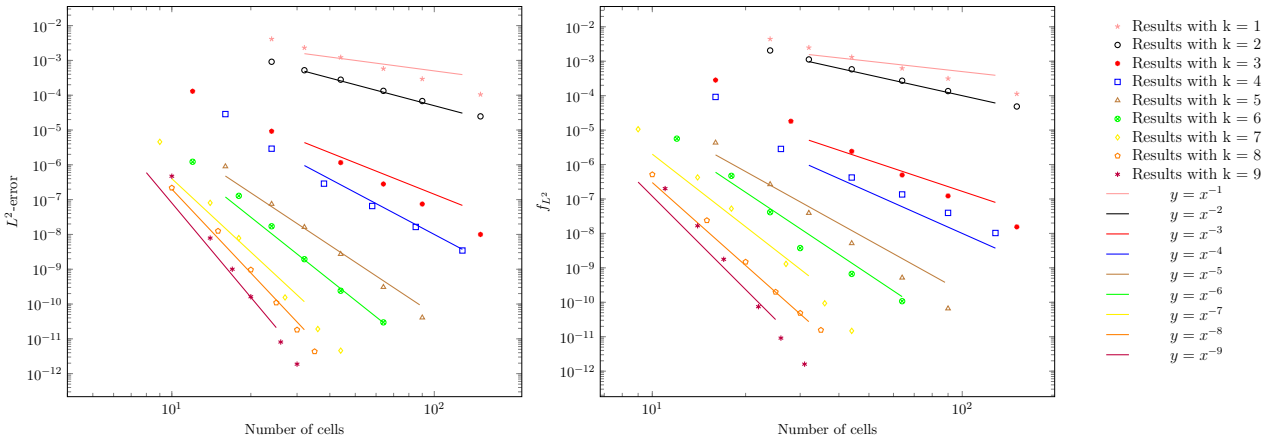


Fig. 1.4 – L^2-error, at the left, and f_{L^2} (refer to Equation (1.65)), at the right, with the non-symmetric scheme for problem of Section 1.4.2.

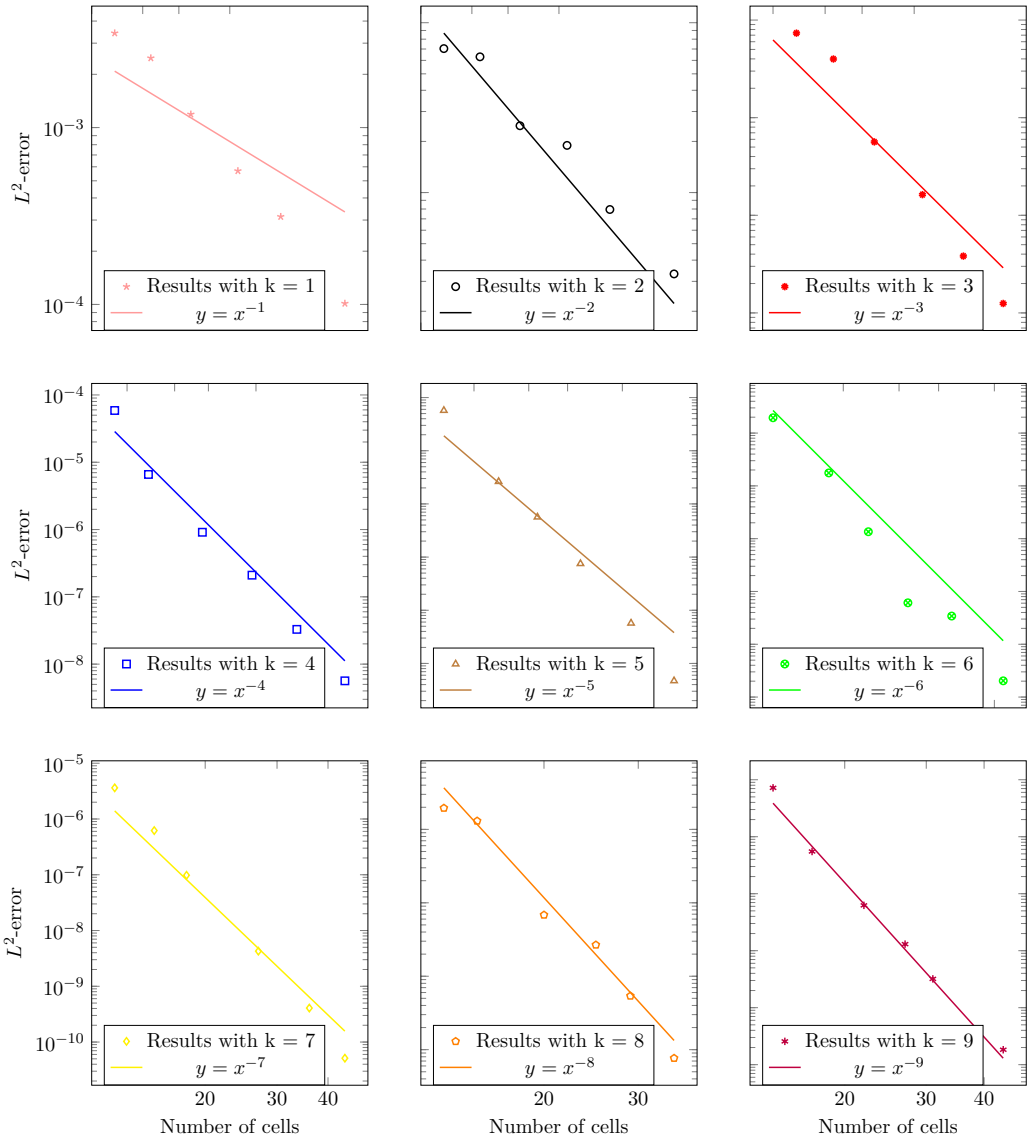


Fig. 1.5 – L^2-error with non-symmetric scheme and random mesh for problem of Section 1.4.2.

The results show that the numerical convergence order is at worst equal to the theoretical order k (for the theoretical order 4 one obtains convergence at order 4) or better (for the theoretical order 3 one obtains the order 4). Besides, the results are qualitatively the same for the symmetric case and

for the non-symmetric case (the results are only given for the non symmetric case because the figures are similar). We observe similar convergence orders for e_{L^2} and f_{L^2}.

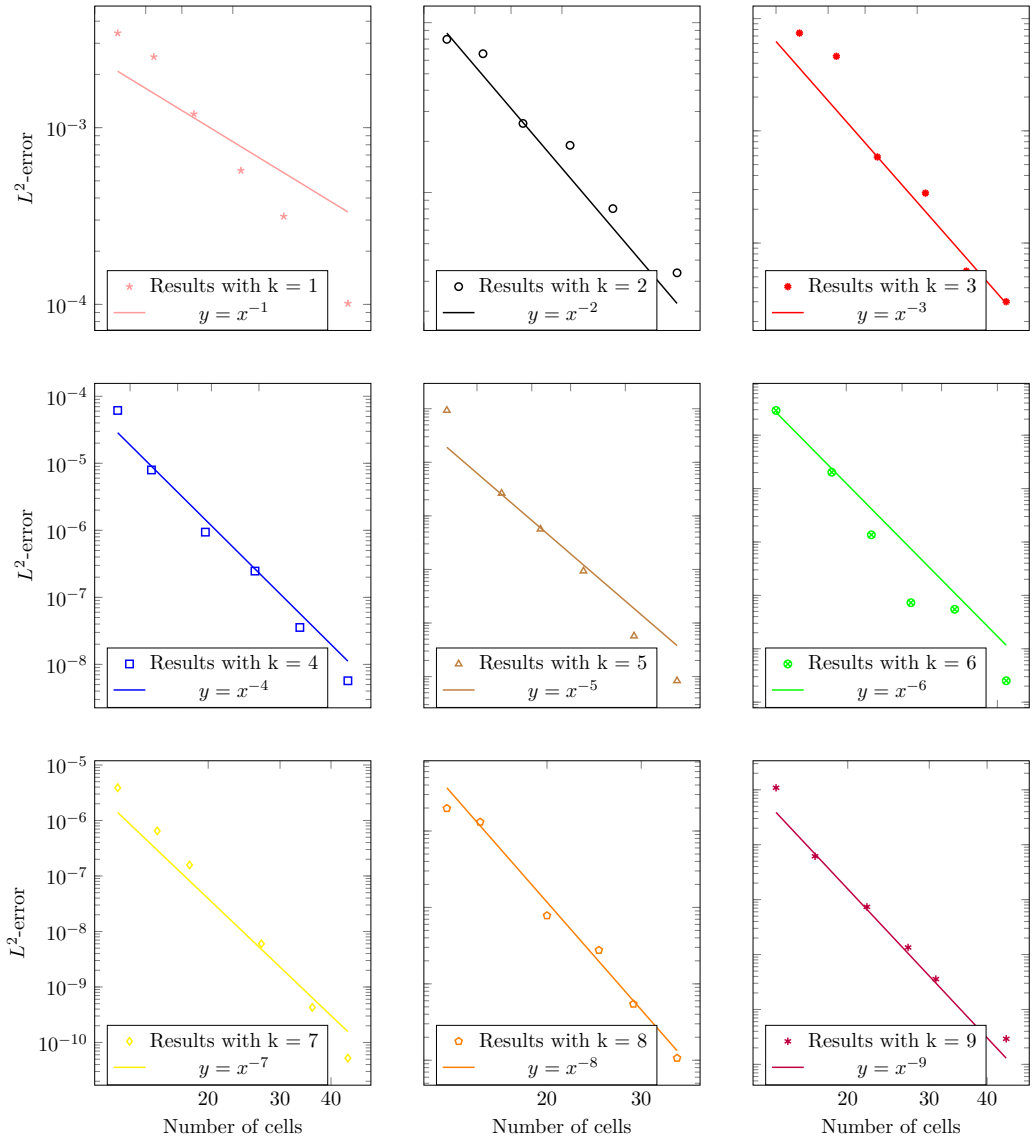


Fig. 1.6 – f_{L^2}-error with non-symmetric scheme and random mesh for problem of Section 1.4.2.

We also perform a convergence study for the same problem on the random mesh: see Figures 1.5 and 1.6. As for the deformed mesh, the results show that the numerical convergence order is at worst equal to the theoretical order k (for the theoretical order 4 one obtains convergence at order 4) or better (for the theoretical order 3 one obtains convergence at order 4). The results are similar for the symmetric case and for the non-symmetric case (the results are only given for the non symmetric case). We observe similar convergence orders for e_{L^2} and f_{L^2}. However, the curves are slightly translated: for a given mesh size, the error is larger when the mesh is deformed. This is illustrated on the Figure 1.7 for the fourth-order non-symmetric scheme.

Figure 1.8 show that the number of iterations of the fixed-point algorithm depends weakly on the number of cells. This is especially visible in the case of a deformed mesh. Besides, for a random mesh, the number of iterations (for the fixed-point algorithm to reach stagnation) is significantly larger than for a deformed mesh.

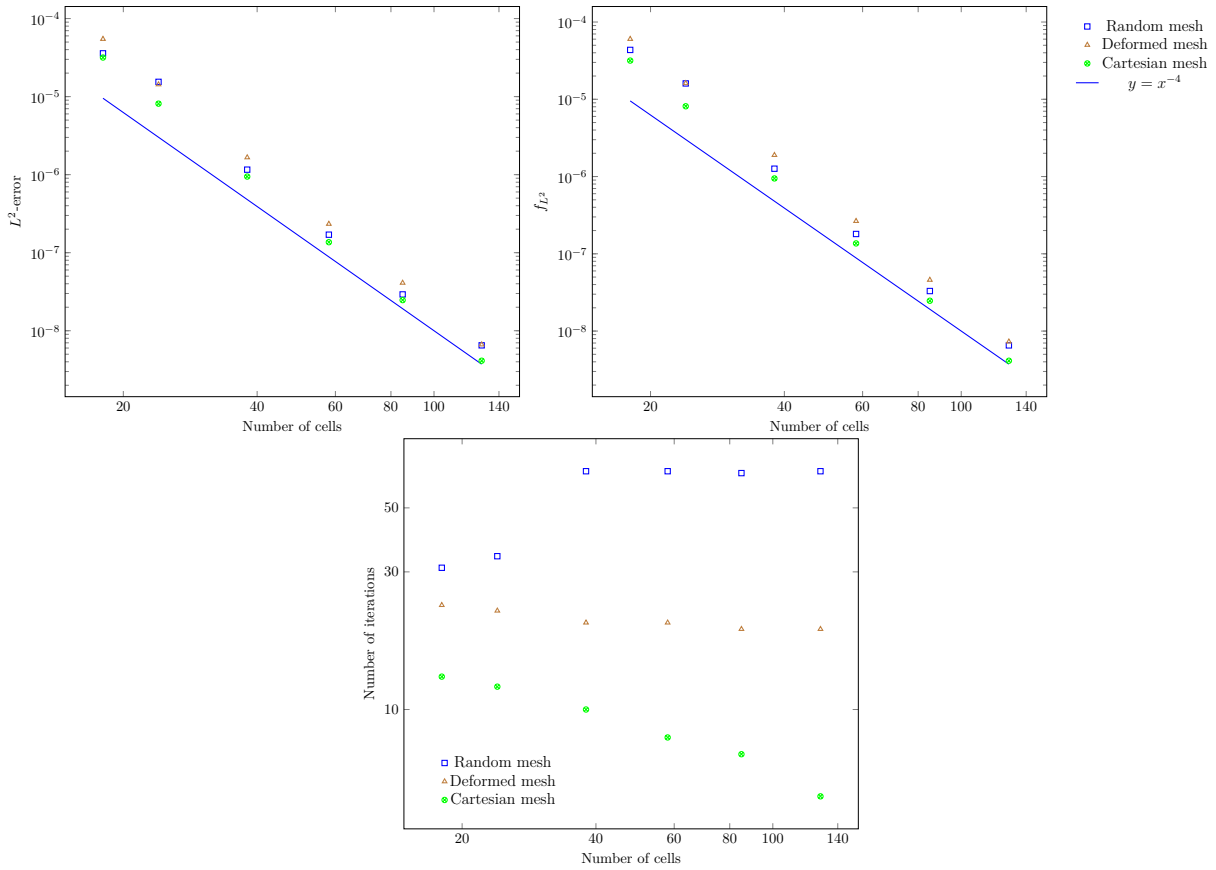


Fig. 1.7 – L^2-error, at the top left, and f_{L^2} (refer to Equation (1.65)), at the top right, and number of iterations of the fixed point (bottom) with the non-symmetric scheme at order $k = 4$ for problem of Section 1.4.2. It shows that the mesh deformation impacts only slightly the error, but strongly the number of fixed point iterations to achieve convergence.

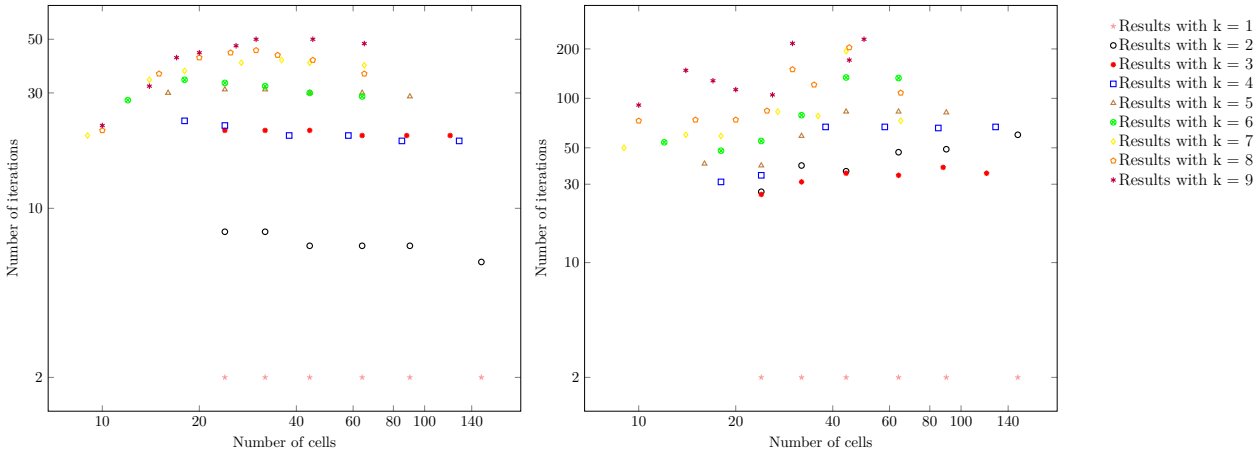


Fig. 1.8 – Number of iterations of the fixed point algorithm with the non-symmetric scheme for problem of Section 1.4.2 for a deformed mesh at the left and for a random mesh at the right. The number of iteration of the fixed point algorithm increases with the order of convergence k, but is weakly affected by the mesh refinement.

1.4.3 Comparison with a non-monotonic scheme

To show the effect of the monotonicity correction, we compare our scheme with a non-monotonicity preserving scheme.

Given $\kappa = 1$, $f = \pi^2 \sin(\pi x)$, $g(0) = g(1) = 0$, the function $\bar{u}(x) = \sin(\pi x)$ is solution to (1.66). We perform a monotonicity study for this problem on a Cartesian mesh with the third-order version

for different grid sizes. Results are summarized in Table 1.2. Note that the non-monotonic scheme does not exhibit negative entries for all the grid resolutions, but when it happens, it is corrected with the monotonic version.

Number of cells	High order monotonic scheme	High order non-monotonic scheme
8	0	1
16	0	0
32	0	0
64	0	1
128	0	0

Tab. 1.2 – Negative entries for the non-monotonic and the monotonic schemes for problem of Section 1.4.3.

1.4.4 Discontinuous diffusion coefficient κ

Given κ such that

$$\kappa(x) = \begin{cases} 1 & \text{if} \quad x \leq \dfrac{1}{2}, \\ 2 & \text{if} \quad x > \dfrac{1}{2}, \end{cases}$$

and $f(x) = \pi^2 \sin(\pi x)$, the function

$$\bar{u}(x) = (\sin(\pi x) + 2x)\, \mathbb{1}_{\{x \leq \frac{1}{2}\}}(x) + \left(\frac{1}{2}\sin(\pi x) + x + 1\right) \mathbb{1}_{\{x > \frac{1}{2}\}}(x),$$

is solution to (1.66). The solution of this problem is displayed on Figure 1.9. We perform a convergence study for this problem, using the method described in Section 1.3.5, on a Cartesian mesh for order 1 to 9.

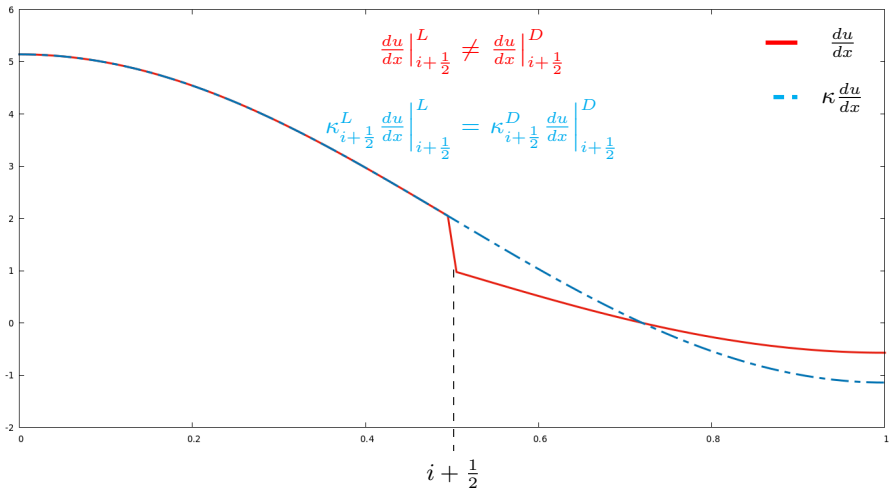


Fig. 1.9 – Illustration of problem of Section 1.4.4. The diffusion being discontinuous, so is the gradient, but the flux remains continuous.

An even number of cells is required to have a node coinciding with the discontinuity of κ ($x = \frac{1}{2}$). Results are summarized in Figure 1.10. These graphs show that we achieve the expected convergence rate, even with discontinuous κ.

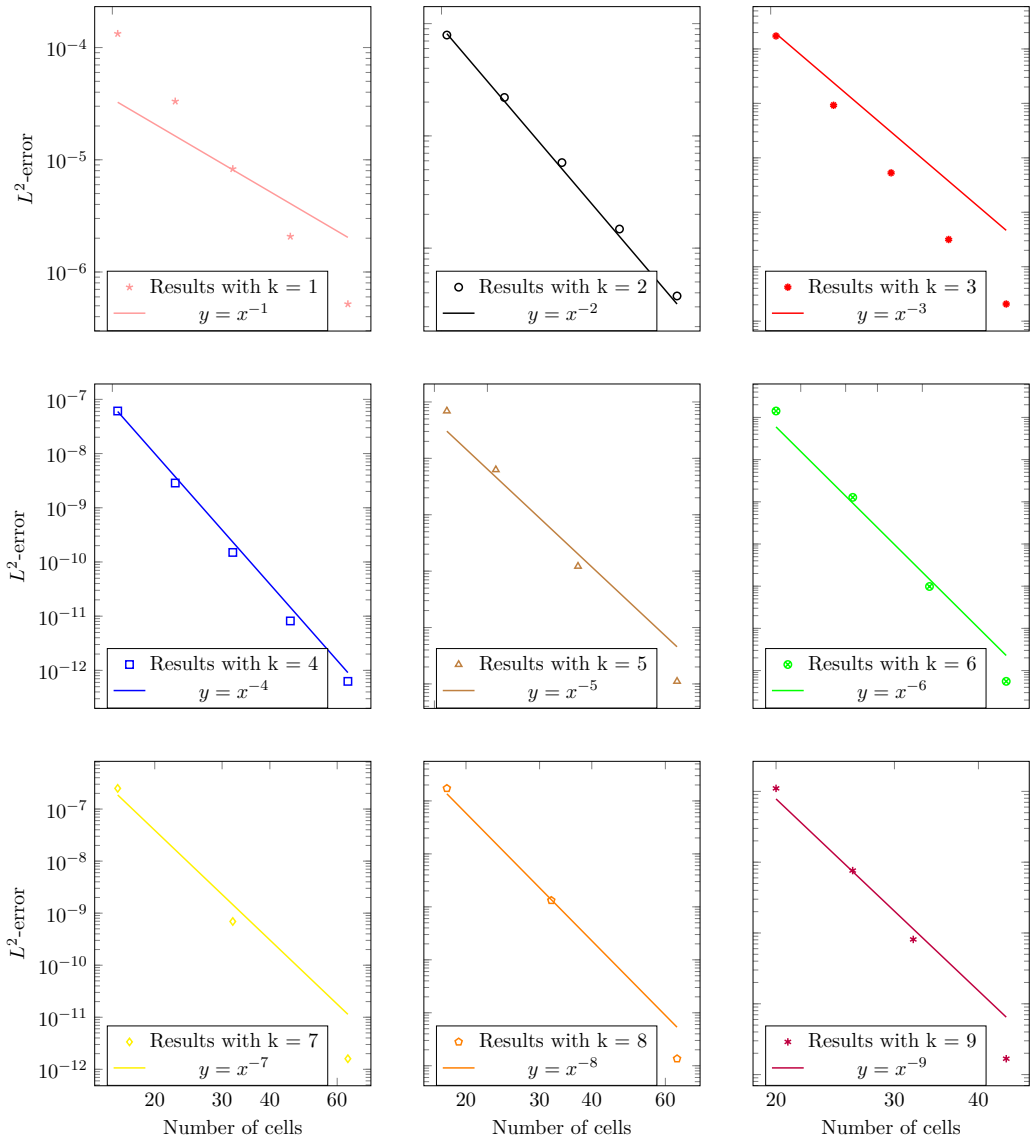


Fig. 1.10 – L^2-error with symmetric scheme and discontinuous κ for problem of Section 1.4.4.

1.5 Concluding remarks

In this chapter we have proposed an arbitrary-order monotonic scheme for the elliptic problem (1.3), on arbitrary 1D meshes. The properties of convergence at a given order, and the preservation of the positivity of the discrete solution have been proven with reasonable assumptions on the mesh. We also proposed a symmetric version of the method. We have shown how to extend these schemes to the case of a discontinuous diffusion coefficient. These properties have been illustrated numerically up to the order 9. In future works, we aim to extend these schemes to higher spatial dimensions and to parabolic problems. We are quite confident in the fact that our scheme can be extended to 2D because we used the same method to enforce monotonicity than in [51,52,105,111], who have applied it in the context of 2D diffusion on arbitrary meshes. To extend this method in 2D, we will need secondary unknowns. In order to compute them, several strategies are possible. Among others, one may use interpolation (see [30]), or a dual partition, in the spirit of the DDFV method (see [58]).

In the following chapter, we study the 2D case. We apply the same method to enforce monotonicity and compare the two possibilities to compute secondary unknowns.

Chapter 2

Monotonic diamond and DDFV type finite-volume schemes for 2D elliptic problems

This chapter has been published as an article in Communications in Computational Physics (see [9]). In this chapter, the notations have been modified to be consistent with the rest of the manuscript. For all the tests, we use the linear solver GMRES with the preconditioner ILU (see [83], Chapter 7.4) and the convergence criterion is 10^{-14}. We add a numerical test in Section 2.7.2.3.

The DDFV (Discrete Duality Finite Volume) method is a finite volume scheme mainly dedicated to diffusion problems, with some outstanding properties. This scheme has been found to be one of the most accurate finite volume methods for diffusion problems. In this chapter, we propose a new monotonic extension of DDFV, which can handle discontinuous tensor-valued diffusion coefficient. Moreover, we compare its performance to a diamond type method with an original interpolation method relying on polynomial reconstructions. Monotonicity is achieved by adapting the method of [51, 52, 105, 111]. Such a technique does not require the positiveness of the vertex unknowns. We show that the two new methods are second-order accurate and are indeed monotonic on some challenging benchmarks as, for instance, a Fokker-Planck problem.

2.1 Introduction

Consider the model stationary diffusion problem

$$\begin{cases} -\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}\bar{u}) + \lambda\bar{u} = f & \text{in} \quad \Omega, \\ \bar{u} = g_D & \text{on} \quad \Gamma_D, \\ \boldsymbol{\kappa}\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n} = g_N & \text{on} \quad \Gamma_N, \end{cases} \tag{2.1}$$

where Ω is a bounded open domain of $\mathbb{R}^2$ with $\partial\Omega = \Gamma_D \cup \Gamma_N$ ($\Gamma_D \cap \Gamma_N = \emptyset$) and $\mathbf{n} \in \mathbb{R}^2$ the outgoing unit normal vector. The data are such that f,$\lambda \in L^2(\Omega)$, with $\lambda \geq 0$ (if $\lambda = 0$, then $|\Gamma_D| > 0$), and $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in L^2(\Gamma_N)$. The tensor-valued diffusion coefficient $\boldsymbol{\kappa}$ is supposed to be bounded and to satisfy the uniform ellipticity condition

$$\forall \mathbf{x} \in \Omega, \quad \forall \mathbf{y} \in \mathbb{R}^2, \quad \alpha_{min}\|\mathbf{y}\|^2 \leq \mathbf{y}^T\boldsymbol{\kappa}(\mathbf{x})\mathbf{y} \leq \alpha_{max}\|\mathbf{y}\|^2,$$

where α_{min}, α_{max} are positive coefficients. Under the above conditions, and if either $\lambda > 0$ or Γ_D is of positive length, it is well known that system (2.1) has a unique solution in $H^1(\Omega)$. Such a solution satisfies a positiveness principle, i.e. if $f \geq 0$ and $g \geq 0$, then $\bar{u} \geq 0$ (using Lax-Milgram Lemma in the spirit of [46], Chapter 6).

Standard methods may be applied to the discretization of such diffusion equations with possibly discontinuous κ on arbitrary meshes. This proves to be an efficient strategy, as far as accuracy (or convergence) is concerned. However, it is well known that positiveness of the discrete solution does not hold. This lack of positiveness (also called monotonicity) can lead to serious difficulties, since $\bar{u}$ can account for a temperature or a concentration. A first attempt to solve the issue of monotonicity would be to truncate the discrete solution to zero. This is not satisfactory because conservation is lost in such a process, and conservation is an important property of the scheme. Some algorithms based on the repair technique introduced in [81] are employed to fix the conservation issue [23,80,103,107]. However, these algorithms are only *globally* (and not locally) conservative, and the consistency is unclear. Some monotonic methods have been designed in the finite-element framework (see [26, 28, 63, 65, 99] among others), but they rely on restrictive conditions on the mesh, that we cannot afford.
For fifteen years many original finite volume methods have been proposed to address the issue of monotonicity, while preserving conservation. Most of these schemes are nonlinear or have a larger stencil than standard methods. The finite volume framework is well suited to achieve monotonicity because it allows for an easy manipulation of the fluxes. The first works we know of are those of Le Potier [70] and Bertolazzi and Manzini [7]. In such methods, one uses a manipulation of the fluxes that leads to introduce a dependence on the discrete solution in the coefficients of the fluxes, making the scheme nonlinear, although (2.1) is linear. To this end, one usually introduces secondary unknowns (for instance vertex-located or face-located unknowns) in addition to the primary (cell-located) unknowns.

Among others, important contributions to this field are [10, 51, 76, 96, 110], which propose efficient numerical schemes preserving the positiveness of the primary unknowns. In [95] the requirement of positive secondary unknowns is relaxed. The works [77, 112] explain how to build monotonic schemes without relying on secondary unknowns. In [72,79,94], maximum principle preserving schemes are proposed. Cancès and Guichard obtained moreover an entropy diminishing property in [22], introducing the nonlinearity directly at the continuous level via a change of variables. Some concepts and proofs about the existence of solutions for these types of scheme can be found in [32, 40, 92]. See also [101, 108] for recent advances in this field.
The DDFV (Discrete Duality Finite Volume [58], [35]) scheme relies on secondary (nodal) unknowns. However, in contrast with most above-mentioned methods, one considers an additional diffusion problem on a so-called *dual* mesh to calculate them. This scheme has been found to be one of the most accurate finite volume methods for diffusion problems [56], at the price of doubling the number of degrees of freedom compared for instance to the linear or bilinear finite element method or to cell centered methods such as MPFA (Multi Point Flux Approximation [1]) or SUSHI (Scheme Using Stabilization and Hybrid Interfaces [49]). However, none of latter methods are monotonic.
A monotonic extension of DDFV has been proposed in [21], but was not compatible with Neumann boundary conditions, and only first-order convergent for discontinuous tensor coefficients $\boldsymbol{\kappa}$. In the present chapter, we propose a new monotonic extension of DDFV that remedies these flaws. Moreover, we compare its performance to a diamond type method with an original interpolation method relying on polynomial reconstructions. Monotonicity is achieved by adapting the method of [51, 52, 105, 111] to our schemes. Such a technique does not require the positiveness of the secondary unknowns.
The main steps of the proposed methods may be briefly summarized as follows.

1. Integration of the equation over each cell of the user's mesh that we will call *primal*.
2. Transformation of this surface integral into a sum of fluxes using the divergence theorem.
3. Approximation of the fluxes using the midpoint quadrature rule on each face of the cell.
4. Taylor expansion of the solution $\bar{u}$ in the neighborhood of the midpoint of each face along *two* independent privileged directions in order to obtain an approximation of $\boldsymbol{\nabla}\bar{u}$ involving the values of $\bar{u}$ and its derivatives at certain suitably chosen points, in this case the center and vertices of the cell.
5. Thanks to this Taylor expansion, estimation of $(\boldsymbol{\kappa}\boldsymbol{\nabla}\bar{u})\cdot\mathbf{n} = (\boldsymbol{\nabla}\bar{u})\cdot(\boldsymbol{\kappa}^T\mathbf{n})$.
6. Calculation of the values of $\bar{u}$ at vertices either by a polynomial interpolation formula in the neighborhood of the midpoint of each primal cell face or by integration of the equation over each cell of the dual mesh.
7. Calculation of the values of derivatives of $\bar{u}$ at centers and vertices of the neighboring cells by differentiating this polynomial interpolation.
8. Transformation of the scheme into a monotonic nonlinear two point flux approximation (or four point flux approximation if a DDFV type method is used).
9. Resolution of the nonlinear system by the Picard iteration method.

The integration over the primal mesh is common to the two monotonic schemes proposed here and is described in Section 2.3. The treatment of the vertex unknowns depends on the scheme and is addressed in Section 2.4. Monotonicity of both schemes is based on the same strategy, which is described in Section 2.5. It leads to a two point flux the coefficients of which depend on the unknown. The Picard iteration method to handle the nonlinearity is also described. The properties of the new DDFV schemes are listed in Section 2.6. Finally, both schemes are assessed in term of accuracy, monotonicity and computational efficiency, and compared with the non monotonic DDFV scheme in Section 2.7. It is shown that the interpolation-based scheme is more efficient for a given L^2 accuracy, but that the DDFV-based scheme achieves second-order accuracy in H^1 norm for the tests we ran.

This outstanding feature has been already observed in [56, 61]. Our final test problem is a solution of a simplified Fokker-Planck equation. We show that our monotonic DDFV scheme is able to compute a correct monotonic solution while achieving the energy conservation.

In all that follows vectors and matrices will be denoted with bold letters. Moreover, $\mathbf{x} = (x,y)$ and $\mathbf{I}$ will stand for the position and 2×2 identity matrix, respectively.

2.2 Definitions and notations

In this section we gather most of the notations that will be used later.

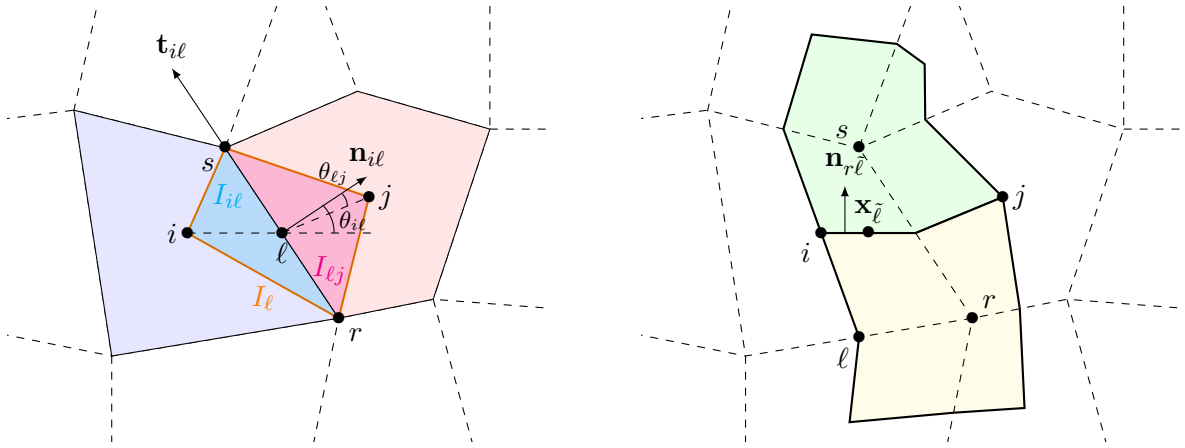


Fig. 2.1 – Primal mesh (at the left) and dual mesh (at the right)

Consider an arbitrary primal mesh made of (possibly distorted, non-conformal, non convex...) *polygonal* cells that are numbered from 1 to n. The primal cells are denoted i or j. The center of a cell i is denoted by $\mathbf{x}_i$ (in general $\mathbf{x}_i$ is the *mass* center of i but other interior points for which i is star-shaped could be chosen), its faces are ℓ (which length is $|\ell|$) and its vertices $\mathbf{x}_r$ and $\mathbf{x}_s$. The position of the center of the face ℓ is $\mathbf{x}_\ell$ and the position of a vertex r is $\mathbf{x}_r$. The volume of a cell i is V_i. The normal vector $\mathbf{n}_{i\ell}$ is the unit vector which is orthogonal to the edge ℓ and outgoing for the cell i, and $\mathbf{N}_{i\ell} = |\ell|\mathbf{n}_{i\ell}$. The diamond cell $\mathbf{x}_i\mathbf{x}_r\mathbf{x}_j\mathbf{x}_s$ is denoted by I_ℓ and its volume is V_ℓ. The diamond subcell $\mathbf{x}_i\mathbf{x}_r\mathbf{x}_\ell\mathbf{x}_s$ (resp. $\mathbf{x}_j\mathbf{x}_s\mathbf{x}_\ell\mathbf{x}_r$) is denoted by $I_{i\ell}$ (resp. $I_{\ell j}$). Let $\theta_{i\ell}$ (resp. $\theta_{\ell j}$) be the angle between $\mathbf{x}_\ell - \mathbf{x}_i$ (resp. $\mathbf{x}_j - \mathbf{x}_\ell$) and $\mathbf{n}_{i\ell}$.

In order to define DDFV type schemes we also need to define a dual mesh (often named barycentric or Donald dual mesh). The dual mesh is obtained from the primal mesh by joining the primal cell centers to the primal face centers. The dual cells are numbered from 1 to m. The cells of the dual mesh are denoted by r or s, its faces are $\tilde{\ell}$ and its vertices are i, j and ℓ. The volume of a cell r is V_r. The normal vector $\mathbf{n}_{r\tilde{\ell}}$ is the unit vector which is orthogonal to the edge $\tilde{\ell}$ and outgoing for the cell r, and $\mathbf{N}_{r\tilde{\ell}} = |\tilde{\ell}|\mathbf{n}_{r\tilde{\ell}}$. These notations are summarized on Figure 3.1.

Remark 2.2.1. *The dual mesh is useful for the DDFV scheme only (Section 2.4.2). When interpolation is used to define vertex values (Section 2.4.1), the definition of the dual mesh is useless.*

We consider conform meshes, which cover the whole domain and in which there is no overlap. We define

$$h = \max_{\ell,\tilde{\ell}} \left(|\ell|, |\tilde{\ell}| \right),$$

we will assume that the primal and dual meshes satisfy the following assumptions.

- (**H1**) There exists a constant θ_0 independent of h such that, for all ℓ,

$$|\theta_0| < \frac{\pi}{2}, \quad \cos(\theta_0) < \cos(\theta_{i\ell}), \quad \cos(\theta_0) < \cos(\theta_{\ell j}).$$

- (**H2**) Given N_i (resp. N_r) the number of faces of the primal (resp. dual) cell i (resp. r), there exists a constant $N_{\max}$ independent of h such that

$$\max(\max_i N_i, \max_r N_r) < N_{\max}.$$

- (**H3**) There exists a constant ξ independent of h such that, for all ℓ,

$$V_\ell \leq \xi \min(V_i, V_j, V_r, V_s).$$

Given $\mathbf{v} = (v_i)$ a vector in $\mathbb{R}^n$ we will denote respectively its Euclidian, L^2 and L^∞ norms by

$$\|\mathbf{v}\| = \left(\sum_{i=1}^n v_i^2\right)^{1/2}, \quad \|\mathbf{v}\|_2 = \left(\sum_{i=1}^n V_i v_i^2\right)^{1/2}, \quad \|\mathbf{v}\|_\infty = \max_{1\leq i\leq n} |v_i|,$$

and we use the following notations

$$\mathbf{v} \geq 0 \quad \text{if} \quad \forall i,\ v_i \geq 0,$$
$$\mathbf{v} > 0 \quad \text{if} \quad \forall i,\ v_i > 0.$$

Given $\mathbf{x}_k$ any point and ϕ any function we will often note $\phi_k = \phi(\mathbf{x}_k)$.

2.3 Finite volume formulation on the primal mesh

2.3.1 Computation of the flux

We will assume that $\boldsymbol{\kappa}$ is continuous inside each cell but can be discontinuous along some primal faces ℓ. To simplify the presentation it will be assumed for now on that $\boldsymbol{\kappa}$ is scalar-valued, that is, $\boldsymbol{\kappa} = \kappa \boldsymbol{I}$ with $\alpha_{min} \leq \kappa \leq \alpha_{max}$. For a tensor-valued coefficient $\boldsymbol{\kappa} \in \mathbb{R}^{2,2}$ it is enough to replace $\kappa\mathbf{n}$ by $\boldsymbol{\kappa}^T\mathbf{n}$ in the following calculations.

The first step to design a finite volume scheme consists in integrating (2.1) on cell i

$$-\int_i \boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} \bar{u} + \int_i \lambda \bar{u} = \int_i f.$$

We can make use of the divergence formula to obtain

$$-\sum_{\ell \in i} \int_\ell \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} + \int_i \lambda \bar{u} = \int_i f. \tag{2.2}$$

We need to approximate the flux

$$\bar{\mathcal{F}}_\ell = \int_\ell \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n}.$$

With a second-order approximation, we have

$$-\sum_{\ell \in i} |\ell| \left(\kappa \boldsymbol{\nabla} \bar{u}\right)_\ell \cdot \mathbf{n}_{i\ell} + \int_i \lambda \bar{u} = \int_i f + \mathcal{O}\left(h^3\right).$$

Thus we need to approximate

$$\left(\kappa \boldsymbol{\nabla} \bar{u}\right)_\ell \cdot \mathbf{n}_{i\ell}.$$

Suppose that $\bar{u} \in W^{1,\infty}(\Omega)$. A Taylor expansion in the neighborhood of $\mathbf{x}_\ell$ gives

$$\bar{u}(\mathbf{x}) = \bar{u}(\mathbf{x}_\ell) + (\mathbf{x} - \mathbf{x}_\ell) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(\|\mathbf{x} - \mathbf{x}_\ell\|^2\right).$$

Replacing $\mathbf{x}$ respectively by $\mathbf{x}_i$, $\mathbf{x}_j$, $\mathbf{x}_r$, $\mathbf{x}_s$, we obtain

$$(\mathbf{x}_j - \mathbf{x}_\ell) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) = \bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right),$$

$$(\mathbf{x}_i - \mathbf{x}_\ell) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) = \bar{u}(\mathbf{x}_i) - \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right),$$

$$\bar{u}(\mathbf{x}_r) = \bar{u}(\mathbf{x}_\ell) + (\mathbf{x}_r - \mathbf{x}_\ell) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right),$$

$$\bar{u}(\mathbf{x}_s) = \bar{u}(\mathbf{x}_\ell) + (\mathbf{x}_s - \mathbf{x}_\ell) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right).$$

Subtracting the last two equations we have

$$(\mathbf{x}_s - \mathbf{x}_r) \cdot \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) = \bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right).$$

Thus, we have the system

$$\begin{cases} \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_j - \mathbf{x}_\ell) = \bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right), \\ \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_\ell - \mathbf{x}_i) = \bar{u}(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i) + \mathcal{O}\left(h^2\right), \\ \boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_s - \mathbf{x}_r) = \bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right). \end{cases} \tag{2.3}$$

We can decompose the normal vector $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_j - \mathbf{x}_\ell),(\mathbf{x}_s - \mathbf{x}_r))$ or $((\mathbf{x}_\ell - \mathbf{x}_i),(\mathbf{x}_s - \mathbf{x}_r))$

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,j} \frac{\mathbf{x}_j - \mathbf{x}_\ell}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j} \frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|} = \alpha_{i\ell,i} \frac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i} \frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

with

$$\alpha_{i\ell,j} = \frac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{(\mathbf{x}_j - \mathbf{x}_\ell) \cdot \mathbf{n}_{i\ell}}, \qquad \beta_{i\ell,j} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\| \mathbf{n}_{i\ell} \cdot (\mathbf{x}_j - \mathbf{x}_\ell)^\perp}{(\mathbf{x}_s - \mathbf{x}_r) \cdot (\mathbf{x}_j - \mathbf{x}_\ell)^\perp}. \tag{2.4}$$

and

$$\alpha_{i\ell,i} = \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{(\mathbf{x}_\ell - \mathbf{x}_i) \cdot \mathbf{n}_{i\ell}}, \qquad \beta_{i\ell,i} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\| \mathbf{n}_{i\ell} \cdot (\mathbf{x}_\ell - \mathbf{x}_i)^\perp}{(\mathbf{x}_s - \mathbf{x}_r) \cdot (\mathbf{x}_\ell - \mathbf{x}_i)^\perp}, \tag{2.5}$$

The details of these computations are given in Appendix C.1. That is, in view of Figure C.1

$$\alpha_{i\ell,i} = \frac{1}{\cos(\theta_{i\ell})}, \qquad \beta_{i\ell,i} = \frac{\sin(\theta_{i\ell})}{\cos(\theta_{i\ell})}, \qquad \alpha_{i\ell,j} = \frac{1}{\cos(\theta_{j\ell})}, \qquad \beta_{i\ell,j} = \frac{\sin(\theta_{j\ell})}{\cos(\theta_{j\ell})}. \tag{2.6}$$

According to assumption **H1** these values are well defined. Note that $\alpha_{i\ell,j} > 0$ as soon as the centers $\mathbf{x}_i$ and $\mathbf{x}_j$ of the primal cells i and j are separated by the line corresponding to their face $\ell = i \cap j$. It may happen that $\mathbf{x}_i$ and $\mathbf{x}_j$ are *not* separated by the face ℓ. This is the case for a non-convex cell i if its mass center $\mathbf{x}_i$ is not inside i (see the left-hand side of Figure 2.2). In such a case we replace $\mathbf{x}_i$ by the midpoint of an inner diagonal of i or by any interior point for which i is star-shaped (right-hand side of Figure 2.2). Doing so, the inequalities $\alpha_{i\ell,j} > 0$, which are mandatory to enforce the positiveness of the scheme (see Section 2.5), are always satisfied.
Thus, the gradient in the direction $\mathbf{n}_{sr}$ in the cell j, denoted by $\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell}$ the writes

$$\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,j} \frac{\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_j - \mathbf{x}_\ell)}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j} \frac{\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

and in the cell i, denoted by $\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell}$ writes

$$\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,i} \frac{\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_\ell - \mathbf{x}_i)}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i} \frac{\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

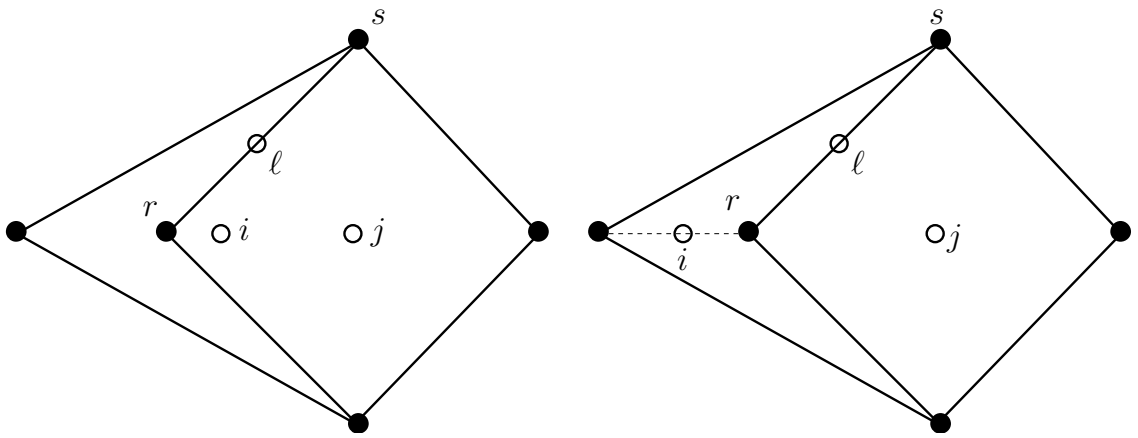


Fig. 2.2 – A non convex cell i and a convex cell j such that $\mathbf{x}_i$ and $\mathbf{x}_j$ are not separated by the line defined by face ℓ.

that is to say, using (2.3)

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,j} \dfrac{\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j} \dfrac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,i} \dfrac{\bar{u}(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i} \dfrac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|}. \end{cases} \tag{2.7}$$

Note that these approximations can also be obtained by using the Green-Gauss formula applied to $\boldsymbol{\nabla}\bar{u}$ in diamond sub-cells $I_{i\ell}$ and $I_{\ell j}$

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i = \dfrac{1}{|I_{i\ell}|} \displaystyle\int_{I_{i\ell}} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell) + \mathcal{O}(h) = \frac{1}{2}\frac{1}{|I_{i\ell}|} \left((\bar{u}(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i))\mathbf{N}_{i\ell} + (\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell},i} \right) + \mathcal{O}(h), \\ \\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j = \dfrac{1}{|I_{\ell j}|} \displaystyle\int_{I_{\ell j}} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell) + \mathcal{O}(h) = \frac{1}{2}\frac{1}{|I_{\ell j}|} \left((\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_\ell))\mathbf{N}_{i\ell} + (\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell},j} \right) + \mathcal{O}(h). \end{cases} \tag{2.8}$$

The fluxes can be *indifferently* estimated using one or the other of formulas (2.7), (2.8).

Let us now recall that the properties of (2.1) imply that the normal component of the flux is continuous across the primal face F_ℓ. We therefore impose

$$\kappa_i \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \kappa_j \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell}, \tag{2.9}$$

which leads to

$$\begin{aligned} \bar{u}(\mathbf{x}_\ell) = & \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\bar{u}(\mathbf{x}_j) + \|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i}\bar{u}(\mathbf{x}_i)}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \\ & + \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|\|\mathbf{x}_j - \mathbf{x}_\ell\|(\kappa_j\beta_{i\ell,j} - \kappa_i\beta_{i\ell,i})}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)}(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) + \mathcal{O}\left(h^2\right). \end{aligned} \tag{2.10}$$

Inserting (2.10) into one of the two equations of (2.7) results in

$$\begin{aligned} \kappa_j \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell} = \kappa_i \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = & \left(\frac{\kappa_i\kappa_j\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \right)(\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_i)) \\ & + \left(\frac{\kappa_i\kappa_j\left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)} \right)(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) + \mathcal{O}\left(h\right). \end{aligned} \tag{2.11}$$

Let $\mathbf{u}^{\text{primal}} = (u_i)_{1\leq i\leq n}$ be the numerical solution on the primal mesh. By mimicking the expression of the exact flux (2.11), the numerical flux through the primal face ℓ is defined by

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| \left[\left(\frac{\kappa_i\kappa_j\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \right) (u_j - u_i) \right.$$
$$\left. + \left(\frac{\kappa_i\kappa_j \left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|\right)}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)} \right) (u_s - u_r) \right].$$

In other words

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell(u_j - u_i) + r_\ell(\mathbf{u}), \tag{2.12}$$

with

$$\begin{cases} r_\ell(\mathbf{u}) = |\ell| \left(\dfrac{\kappa_i\kappa_j \left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|\right)}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)} \right) (u_s - u_r), \\ \gamma_\ell = |\ell| \left(\dfrac{\kappa_i\kappa_j\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \right) \geq 0. \end{cases}$$

This decomposition will be used hereafter to enforce the positiveness of the scheme (see Section 2.5).

2.3.2 Boundary conditions

2.3.2.1 Neumann boundary condition

In the case of a Neumann boundary condition on a primal boundary face $\ell \subset \Gamma_N$, we have

$$\int_\ell \kappa \boldsymbol{\nabla}\bar{u} \cdot \mathbf{n}_{i\ell} = \int_\ell g_N,$$

that is to say

$$\bar{\mathcal{F}}_\ell = |\ell| g_N(\mathbf{x}_\ell) + \mathcal{O}(h^2),$$

we thus impose this equation on the numerical flux

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| g_N(\mathbf{x}_\ell).$$

2.3.2.2 Dirichlet boundary condition

In the case of Dirichlet boundary condition, we have $\bar{u}_\ell = g_D(\mathbf{x}_\ell)$ as soon as $\mathbf{x}_\ell \in \Gamma_D$. From (2.7) we then obtain

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \frac{\alpha_{i\ell,i}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} \left(g_D(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i)\right) + \frac{\beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)\right) + \mathcal{O}\left(h\right),$$

so that the Dirichlet boundary flux is defined by

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell|\kappa_\ell \left(\frac{\alpha_{i\ell,i}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} \left(g_D(\mathbf{x}_\ell) - u_i\right) + \frac{\beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left(u_s - u_r\right) \right).$$

In other words

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell(g_D(\mathbf{x}_\ell) - u_i) + r_\ell(\mathbf{u}), \tag{2.13}$$

with

$$\gamma_\ell = \left(\frac{\kappa_\ell\alpha_{i\ell,i}|\ell|}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} \right) \geq 0, \qquad r_\ell(\mathbf{u}) = |\ell| \frac{\kappa_\ell\beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(u_s - u_r).$$

2.4 Dealing with vertex unknows

In order to evaluate the numerical fluxes $\mathcal{F}_\ell(\mathbf{u})$, Equations (2.12) and (2.13) require the knowledge of the values of u at the vertices $\mathbf{x}_r$ of the primal mesh. To compute these values, we propose to use two different methods. For the first one, described in Section 2.4.1, vertex values are calculated by interpolation while for the second one, described in Section 2.4.2, they are calculated as the solution to the same diffusion problem (2.1) discretized on the dual mesh.

2.4.1 Diamond type scheme

The first way to calculate the vertex values u_r is to use a polynomial approximation calculated using the cell-centered values u_i.
For a polynomial of degree 1, we have 3 coefficients to calculate, so at least 6 (3×dimension) neighboring cells of the considered cell are required for stability reason: see [41,67]. When it is possible, the stencil will be centered on the cell, but the closer the cell is to the boundary of the domain or to the discontinuity of κ, the more the stencil will be shifted in order not to cross the discontinuity.

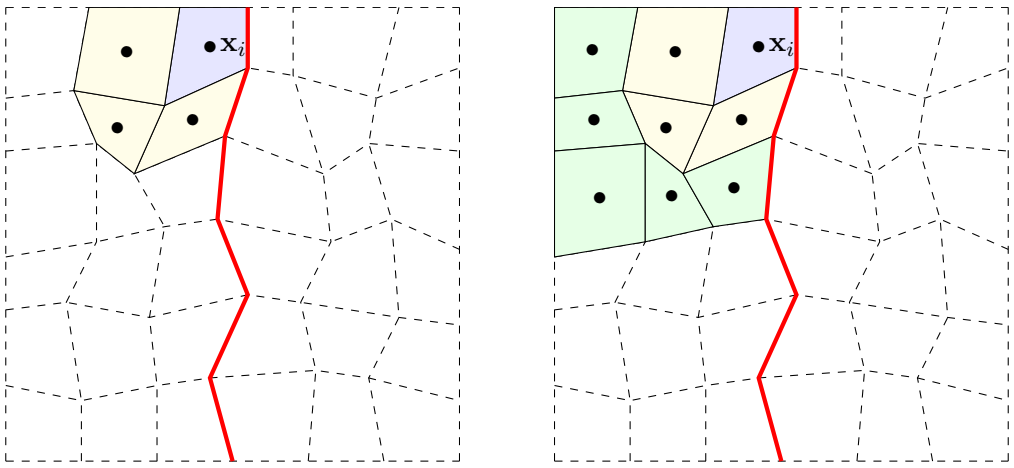


Fig. 2.3 – Construction of the stencil for the cell i with a discontinuity (in red).

To be more precise, the construction of the stencil of a cell i is illustrated on Figure 3.2. We denote this stencil by $\mathcal{S}_i = \{0,...,k\}$. For the sake of simplicity, we have assumed that the cells involved in the stencil have been renumbered. First the cell i itself (in blue) is added to the stencil and then we add the cells that share, at least, a vertex with the cell i (in yellow). If the number of cells we have already selected is not sufficient (in our case, 6 cells for a polynomial of order 1), we add the cells that have, at least, a vertex linked to the cells that we have just been added to the stencil (in green) and so on until we have enough cells. In all the above process, we impose that the stencil does not cross any discontinuity of κ (see Figure 3.2).

Let $u_0, ..., u_k$ denote the $k+1$ values used for the calculation ($k \geq 5$). The polynomial is of the form

$$\mathcal{P}_i(\mathbf{x}) = a_{00}(u_0, ..., u_k) + a_{10}(u_0, ..., u_k)(x - x_i) + a_{01}(u_0, ..., u_k)(y - y_i),$$

and its coefficients a_{00}, a_{10}, a_{01} are chosen such that

$$\mathcal{P}_i(\mathbf{x}_0) = u_0, ..., \mathcal{P}_i(\mathbf{x}_k) = u_k.$$

This leads to the following system

$$\underbrace{\begin{pmatrix} 1 & x_0 - x_i & y_0 - y_i \\ \vdots & \vdots & \vdots \\ 1 & x_k - x_i & y_k - y_i \end{pmatrix}}_{=:\mathbf{M}} \underbrace{\begin{pmatrix} a_{00} \\ a_{10} \\ a_{01} \end{pmatrix}}_{=:\mathbf{a}} = \underbrace{\begin{pmatrix} u_0 \\ \vdots \\ u_k \end{pmatrix}}_{=:\mathbf{b}}.$$

Since matrix $\mathbf{M}$ has more rows than columns we have to use the least square method so that vector $\mathbf{a}$ is computed as a solution to the linear system: $\mathbf{M}^T\mathbf{M}\mathbf{a} = \mathbf{M}^T\mathbf{b}$.

In this process note that we do *not* enforce the continuity of u at the vertices. Indeed, a priori, $\mathcal{P}_i(\mathbf{x}_r) \neq \mathcal{P}_j(\mathbf{x}_r)$ for $i \neq j$.

We thus obtain expressions of the gradients in the direction $\mathbf{n}_{i\ell}$ in the cells i and j similar to (2.7)

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,j}\dfrac{\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_\ell) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j}\dfrac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,i}\dfrac{\bar{u}(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i}\dfrac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \mathcal{O}\left(h^2\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|}. \end{cases} \tag{2.14}$$

Assuming the continuity of the flux $\mathcal{F}_\ell$ through the primal face F_ℓ

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell} = \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_j \cdot \mathbf{n}_{i\ell},$$

provides

$$\begin{aligned} \bar{u}(\mathbf{x}_\ell) = &\frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\bar{u}(\mathbf{x}_j) + \|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i}\bar{u}(\mathbf{x}_i)}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \\ &+ \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|\|\mathbf{x}_j - \mathbf{x}_\ell\|\left(\kappa_i\beta_{i\ell,i}(P_j(\mathbf{x}_s) - P_j(\mathbf{x}_r)) - \kappa_j\beta_{i\ell,j}(P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r))\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)} + \mathcal{O}\left(h^2\right). \end{aligned}$$

Let $\mathbf{u} = \mathbf{u}^{\text{primal}} = (u_i)_{1\leq i\leq n}$ be the numerical solution for this method. Replacing $\bar{u}(\mathbf{x}_j)$ by u_j, $\bar{u}(\mathbf{x}_i)$ by u_i in (2.14) and face and nodal values by interpolated approximations, the numerical flux $\mathcal{F}_\ell$ through the primal face ℓ results in

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell(u_j - u_i) + r_\ell(\mathbf{u}), \tag{2.15}$$

with

$$\begin{cases} r_\ell(\mathbf{u}) = |\ell|\left(\dfrac{\kappa_i\kappa_j\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)}(P_j(\mathbf{x}_s) - P_j(\mathbf{x}_r))\right. \\ \qquad \left. + \dfrac{\kappa_i\kappa_j\alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)}(P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r))\right), \\ \gamma_\ell = |\ell|\left(\dfrac{\kappa_i\kappa_j\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}}\right) \geq 0, \end{cases}$$

where P_j is a polynomial local to the cell j. The choice of cell-based polynomials is consistent with the fact that the diffusion coefficient is continuous inside each cell.

So that the diamond scheme writes

$$\begin{cases} -\displaystyle\sum_{\ell\in i,\ell\notin\partial\Omega}(\gamma_\ell(u_j - u_i) + \delta_{i\ell}(\mathcal{P}_i(\mathbf{x}_s) - \mathcal{P}_i(\mathbf{x}_r)) + \delta_{\ell j}(\mathcal{P}_j(\mathbf{x}_s) - \mathcal{P}_j(\mathbf{x}_r))) \\ -\displaystyle\sum_{\ell\in i,\ell\in\partial\Omega}(\gamma_\ell(u_\ell - u_i) + \delta_\ell(\mathcal{P}_i(\mathbf{x}_s) - \mathcal{P}_i(\mathbf{x}_r))) + V_i\lambda_i u_i = V_i f_i, \\ u_\ell = g_D(\mathbf{x}_\ell) \qquad \mathbf{x}_\ell \in \Gamma_D, \\ \gamma_\ell(u_\ell - u_i) + \delta_\ell(\mathcal{P}_i(\mathbf{x}_s) - \mathcal{P}_i(\mathbf{x}_r)) = |\ell| g_N(\mathbf{x}_\ell) \qquad \mathbf{x}_\ell \in \Gamma_N, \end{cases} \tag{2.16}$$

with

$$\gamma_\ell = \frac{|\ell|\kappa_i\kappa_j\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j-\mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i}+\|\mathbf{x}_\ell-\mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}} \quad \text{if } \ell \notin \partial\Omega, \qquad \gamma_\ell = \frac{\kappa_\ell\alpha_{i\ell,i}|\ell|}{\|\mathbf{x}_\ell-\mathbf{x}_i\|} \quad \text{if } \ell \in \gamma_D,$$

$$\delta_{i\ell} = \frac{|\ell|\kappa_i\kappa_j\alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell-\mathbf{x}_i\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i}+\|\mathbf{x}_\ell-\mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)}, \qquad \delta_\ell = |\ell|\frac{\kappa_\ell\beta_{i\ell,i}}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

$$\delta_{\ell j} = \frac{|\ell|\kappa_i\kappa_j\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j-\mathbf{x}_\ell\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_\ell\|\kappa_i\alpha_{i\ell,i}+\|\mathbf{x}_\ell-\mathbf{x}_i\|\kappa_j\alpha_{i\ell,j}\right)}.$$

2.4.2 DDFV scheme

The second way to calculate the vertex values u_r is to consider them as additional unknowns that are solutions to problem (2.1) integrated on each cell of the dual mesh, thus following [58]. We have

$$-\int_r \boldsymbol{\nabla}\cdot\kappa\boldsymbol{\nabla}\bar{u}+\int_r\lambda\bar{u}=\int_r f.$$

that is, thanks to the divergence theorem

$$-\sum_{\tilde{\ell}\in r}\int_{\tilde{\ell}}\kappa\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n}+\int_r\lambda\bar{u}=\int_r f,$$

that is to say, up to second-order terms,

$$-\sum_{\tilde{\ell}\in r}|\tilde{\ell}|\kappa_{\tilde{\ell}}\left(\boldsymbol{\nabla}\bar{u}\right)_{\tilde{\ell}}\cdot\mathbf{n}_{i\tilde{\ell}}+\int_r\lambda\bar{u}=\int_r f+\mathcal{O}\left(h^2\right).$$

where $\kappa_{\tilde{\ell}}$ is an evaluation of κ at the center of the edge $\tilde{\ell}$. Thus we need to define the fluxes

$$\bar{\mathcal{F}}_{\tilde{\ell}}=\kappa_{\tilde{\ell}}\left(\boldsymbol{\nabla}\bar{u}\right)_{\tilde{\ell}}\cdot\mathbf{n}_{i\tilde{\ell}}.$$

Let us consider a dual face $\tilde{\ell}$ located in a primal cell i. A Taylor second-order expansion in the neighborhood of the face $\tilde{\ell}$ gives

$$\bar{u}(\mathbf{x})=\bar{u}(\mathbf{x}_{\tilde{\ell}})+(\mathbf{x}-\mathbf{x}_{\tilde{\ell}})\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})+\mathcal{O}\left(\|\mathbf{x}-\mathbf{x}_{\tilde{\ell}}\|^2\right). \tag{2.17}$$

Replacing $\mathbf{x}$ respectively by $\mathbf{x}_i$ and $\mathbf{x}_\ell$ in the Taylor expansion (2.17) we have

$$\bar{u}(\mathbf{x}_i)=\bar{u}(\mathbf{x}_{\tilde{\ell}})+(\mathbf{x}_i-\mathbf{x}_{\tilde{\ell}})\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\mathcal{O}\left(h^2\right),$$

$$\bar{u}(\mathbf{x}_\ell)=\bar{u}(\mathbf{x}_{\tilde{\ell}})+(\mathbf{x}_\ell-\mathbf{x}_{\tilde{\ell}})\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})+\mathcal{O}\left(h^2\right).$$

The difference of the two previous equations gives

$$(\mathbf{x}_\ell-\mathbf{x}_i)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})=\bar{u}(\mathbf{x}_\ell)-\bar{u}(\mathbf{x}_i)+\mathcal{O}\left(h^2\right).$$

Then, using $\mathbf{x}=\mathbf{x}_r$ and $\mathbf{x}=\mathbf{x}_s$ in (2.17) gives

$$\bar{u}(\mathbf{x}_r)=\bar{u}(\mathbf{x}_{\tilde{\ell}})+(\mathbf{x}_r-\mathbf{x}_{\tilde{\ell}})\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})+\mathcal{O}\left(h^2\right),$$

$$\bar{u}(\mathbf{x}_s)=\bar{u}(\mathbf{x}_{\tilde{\ell}})+(\mathbf{x}_s-\mathbf{x}_{\tilde{\ell}})\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})+\mathcal{O}\left(h^2\right).$$

The difference of the two previous equations gives

$$(\mathbf{x}_s-\mathbf{x}_r)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})=\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)+\mathcal{O}\left(h^2\right).$$

Thus, we have the system

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot(\mathbf{x}_\ell-\mathbf{x}_i)=\bar{u}(\mathbf{x}_\ell)-\bar{u}(\mathbf{x}_i)+\mathcal{O}\left(h^2\right), \\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot(\mathbf{x}_s-\mathbf{x}_r)=\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)+\mathcal{O}\left(h^2\right). \end{cases}$$

We can decompose the normal vector $\mathbf{n}_{r\tilde{\ell}}$ in the basis $((\mathbf{x}_\ell-\mathbf{x}_i),(\mathbf{x}_s-\mathbf{x}_r))$

$$\mathbf{n}_{r\tilde{\ell}}=\alpha_{r\tilde{\ell}}\frac{\mathbf{x}_\ell-\mathbf{x}_i}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}+\beta_{r\tilde{\ell}}\frac{\mathbf{x}_s-\mathbf{x}_r}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

with

$$\beta_{r\tilde{\ell}}=\frac{\|\mathbf{x}_s-\mathbf{x}_r\|}{(\mathbf{x}_s-\mathbf{x}_r)\cdot\mathbf{n}_{r\tilde{\ell}}},\qquad \alpha_{r\tilde{\ell}}=\frac{\|\mathbf{x}_\ell-\mathbf{x}_i\|\mathbf{n}_{r\tilde{\ell}}\cdot(\mathbf{x}_s-\mathbf{x}_r)^\perp}{(\mathbf{x}_\ell-\mathbf{x}_i)\cdot(\mathbf{x}_s-\mathbf{x}_r)^\perp}, \tag{2.18}$$

the details of the computations are given in Appendix C.2.

Thus, we have

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{n}_{r\tilde{\ell}}=\alpha_{r\tilde{\ell}}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot(\mathbf{x}_\ell-\mathbf{x}_i)}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}+\beta_{r\tilde{\ell}}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot(\mathbf{x}_s-\mathbf{x}_r)}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

that is to say

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{n}_{r\tilde{\ell}}=\alpha_{r\tilde{\ell}}\frac{\bar{u}(\mathbf{x}_\ell)-\bar{u}(\mathbf{x}_i)+\mathcal{O}\left(h^2\right)}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}+\beta_{r\tilde{\ell}}\frac{\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)+\mathcal{O}\left(h^2\right)}{\|\mathbf{x}_s-\mathbf{x}_r\|}. \tag{2.19}$$

Let $\mathbf{u}=\begin{pmatrix}\mathbf{u}^{\text{primal}}\\ \mathbf{u}^{\text{dual}}\end{pmatrix}$ be the numerical solution, where $\mathbf{u}^{\text{dual}}=(u_r)_{1\le r\le m}$ is the numerical solution on the dual mesh and $\mathbf{u}^{\text{primal}}=(u_i)_{1\le i\le n}$ is the numerical solution on the primal mesh. By mimicking the expression of the exact flux (2.19), the numerical flux is defined by

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u})=|\tilde{\ell}|\kappa_{\tilde{\ell}}\left[\alpha_{r\tilde{\ell}}\frac{u_\ell-u_i}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}+\beta_{r\tilde{\ell}}\frac{u_s-u_r}{\|\mathbf{x}_s-\mathbf{x}_r\|}\right].$$

In other words

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u})=\gamma_{\tilde{\ell}}(u_s-u_r)+r_{\tilde{\ell}}(\mathbf{u}), \tag{2.20}$$

with

$$\begin{cases} r_{\tilde{\ell}}(\mathbf{u})=\dfrac{|\tilde{\ell}|\kappa_{\tilde{\ell}}\alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}(u_\ell-u_i), \\ \gamma_{\tilde{\ell}}=\dfrac{|\tilde{\ell}|\kappa_{\tilde{\ell}}\beta_{r\tilde{\ell}}}{\|\mathbf{x}_s-\mathbf{x}_r\|}\ge 0, \end{cases}$$

where u_ℓ is obtained by mimicking the expression given by (2.10) if ℓ is not a boundary face. If ℓ is a boundary face, there are two cases.

First, if ℓ is a Dirichlet boundary face, we have

$$\bar{u}(\mathbf{x}_\ell)=g_D(\mathbf{x}_\ell), \tag{2.21}$$

Second, if ℓ is a Neumann boundary face, belonging to a cell i. We have on the one hand

$$\kappa_\ell\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell}=g_N(\mathbf{x}_\ell), \tag{2.22}$$

and on the other hand

$$\kappa_\ell\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell}=\kappa_\ell\left[\alpha_{i\ell}(\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot(\mathbf{x}_\ell-\mathbf{x}_i))+\beta_{i\ell}(\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot(\mathbf{x}_s-\mathbf{x}_r))\right],$$

that is to say

$$\kappa_\ell\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell}=\kappa_\ell\left[\alpha_{i\ell}\left(\bar{u}(\mathbf{x}_\ell)-\bar{u}(\mathbf{x}_i)\right)+\beta_{i\ell}\left(\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)\right)\right]+\mathcal{O}(h^2). \tag{2.23}$$

Equations (2.22) and (2.23) lead to

$$\bar{u}(\mathbf{x}_\ell) = \frac{g_N(\mathbf{x}_\ell)}{\kappa_\ell \alpha_{i\ell}} + \bar{u}(\mathbf{x}_i) - \frac{\beta_{i\ell}}{\alpha_{i\ell}} \left(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)\right) + \mathcal{O}(h^2). \tag{2.24}$$

So, if ℓ is a boundary face, u_ℓ is obtained by mimicking the expression given by (2.24) or (2.21).

Let us see now how to deal with the dual boundary conditions. Let $\tilde{\ell} \subset \gamma_N$ be a Neumann boundary face of the dual mesh. Applying the same process as for the primal mesh, we have

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = |\tilde{\ell}| g_N(\mathbf{x}_{\tilde{\ell}}).$$

On the dual mesh, we penalize the diagonal entries of the matrix and the right-hand side to impose the Dirichlet boundary condition.

The DDFV scheme thus writes

$$\begin{cases} -\displaystyle\sum_{\ell \in i, \ell \notin \partial\Omega} \gamma_\ell (u_j - u_i) - \sum_{\ell \in i, \ell \in \partial\Omega} \gamma_\ell (u_\ell - u_i) - \sum_{\ell \in i} \delta_\ell (u_s - u_r) + V_i \lambda_i u_i = V_i f_i & \\ -\displaystyle\sum_{\tilde{\ell} \in r} \left(\delta_{\tilde{\ell}} (u_\ell - u_i) + \gamma_{\tilde{\ell}} (u_s - u_r)\right) + V_r \lambda_r u_r = V_r f_r & \mathbf{x}_r \notin \Gamma_D, \\ u_\ell = g_D(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_D, \\ u_r = g_D(\mathbf{x}_r) & \mathbf{x}_r \in \Gamma_D, \\ \gamma_\ell (u_\ell - u_i) + \delta_\ell (u_s - u_r) = |\ell| g_N(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_N, \\ \delta_{\tilde{\ell}} (u_\ell - u_i) + \gamma_{\tilde{\ell}} (u_s - u_r) = |\tilde{\ell}| g_N(\mathbf{x}_{\tilde{\ell}}) & \mathbf{x}_{\tilde{\ell}} \in \Gamma_N, \end{cases} \tag{2.25}$$

with

$$\gamma_\ell = \frac{|\ell| \kappa_i \kappa_j \alpha_{i\ell,j} \alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\| \kappa_i \alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\| \kappa_j \alpha_{i\ell,j}} \quad \text{if } \ell \notin \partial\Omega, \qquad \gamma_\ell = \frac{\kappa_\ell \alpha_{i\ell,i} |\ell|}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} \quad \text{if } \ell \in \gamma_D,$$

$$\delta_\ell = \frac{|\ell| \kappa_i \kappa_j \left(\alpha_{i\ell,i} \beta_{i\ell,j} \|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j} \beta_{i\ell,i} \|\mathbf{x}_\ell - \mathbf{x}_i\|\right)}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\| \kappa_i \alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\| \kappa_j \alpha_{i\ell,j}\right)} \quad \text{if } \ell \notin \partial\Omega, \quad \delta_\ell = \frac{|\ell| \kappa_\ell \beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \quad \text{if } \ell \in \gamma_D,$$

$$\gamma_{\tilde{\ell}} = \frac{|\tilde{\ell}| \kappa_{\tilde{\ell}} \beta_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

$$\delta_{\tilde{\ell}} = \frac{|\tilde{\ell}| \kappa_{\tilde{\ell}} \alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}.$$

2.5 Monotonicity

In this section we propose to find a method for the previous methods to be made monotonic. A method borrowed from [51, 52, 105, 111] and developed in the framework of 2D diffusion on arbitrary meshes can be used. For any value $r_\ell(\mathbf{u})$ we will use the common notation $r_\ell(\mathbf{u}) = r_\ell(\mathbf{u})^+ - r_\ell(\mathbf{u})^-$ with

$$r_\ell(\mathbf{u})^+ = \frac{|r_\ell(\mathbf{u})| + r_\ell(\mathbf{u})}{2} \geq 0 \quad \text{and} \quad r_\ell(\mathbf{u})^- = \frac{|r_\ell(\mathbf{u})| - r_\ell(\mathbf{u})}{2} \geq 0.$$

The primal flux (2.12) can be rewritten as follows

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell (u_j - u_i) + r_\ell(\mathbf{u})^+ - r_\ell(\mathbf{u})^-,$$

Let us assume that $\mathbf{u} > \mathbf{0}$, the flux then reads as

$$\mathcal{F}_\ell(\mathbf{u}) = \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j}\right) u_j - \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i}\right) u_i.$$

and the coefficients $\left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j}\right)$ and $\left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i}\right)$ are positive.

The same method can be applied to the dual flux

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}(\mathbf{u})^+}{u_s}\right) u_s - \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}(\mathbf{u})^-}{u_r}\right) u_r.$$

As $\gamma_{\tilde{\ell}} > 0$ and $\gamma_{\tilde{\ell}} > 0$ we end up with two points primal and dual flux approximations with *positive* coefficients, which is very favorable for the resolution of the linear system.

The diamond scheme (2.16) then rewrites

$$\begin{cases} -\displaystyle\sum_{\ell\in i, \ell\notin\partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_j}\right) u_j - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i\right) & \\ -\displaystyle\sum_{\ell\in i, \ell\in\partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_\ell}\right) u_\ell - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i\right) + V_i\lambda_i u_i = V_i f_i, & \\ u_\ell = g_D(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_D, \\ \left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_\ell}\right) u_\ell - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i = |\ell| g_N(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_N, \end{cases}$$

while the DDFV scheme (2.25) rewrites

$$\begin{cases} -\displaystyle\sum_{\ell\in i, \ell\notin\partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_j}\right) u_j - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i\right) & \\ -\displaystyle\sum_{\ell\in i, \ell\in\partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_\ell}\right) u_\ell - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i\right) + V_i\lambda_i u_i = V_i f_i, & \\ -\displaystyle\sum_{\tilde{\ell}\in r} \left(\left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^+(\mathbf{u})}{u_s}\right) u_s - \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^-(\mathbf{u})}{u_r}\right) u_r\right) + V_r\lambda_r u_r = V_r f_r, & \mathbf{x}_r \notin \Gamma_D, \\ u_\ell = g_D(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_D, \\ u_r = g_D(\mathbf{x}_r) & \mathbf{x}_r \in \Gamma_D, \\ \left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u})}{u_\ell}\right) u_\ell - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u})}{u_i}\right) u_i = |\ell| g_N(\mathbf{x}_\ell) & \mathbf{x}_\ell \in \Gamma_N, \\ \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^+(\mathbf{u})}{u_s}\right) u_s - \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^-(\mathbf{u})}{u_r}\right) u_r = |\tilde{\ell}| g_N(\mathbf{x}_{\tilde{\ell}}) & \mathbf{x}_{\tilde{\ell}} \in \Gamma_N. \end{cases} \tag{2.26}$$

The matrices associated with these systems are not symmetric and depend respectively on u_i, u_ℓ ($\ell \in \partial\Omega$) and u_r. More details about this are given in the following section.

2.5.1 Matrix form

The scheme reads as

$$-\sum_{\ell\in i}\mathcal{F}_\ell(\mathbf{u})+\lambda_i V_i u_i = V_i f_i \quad \text{and} \quad -\sum_{\tilde{\ell}\in r}\mathcal{F}_{\tilde{\ell}}(\mathbf{u})+\lambda_r V_r u_r = V_r f_r. \tag{2.27}$$

Consider a mesh the cells of which are numbered from 1 to n and the vertices of which are numbered from 1 to m. Denoting

$$\begin{aligned} &\mathbf{u}^{\text{primal}} = (u_i)_{1\le i\le n}, \quad &&\mathbf{u}^{\text{dual}} = (u_r)_{1\le r\le m}, \quad &&\mathbf{u} = (\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}),\\ &\mathbf{b}^{\text{primal}} = (b_i)_{1\le i\le n}, \quad &&\mathbf{b}^{\text{dual}} = (b_r)_{1\le r\le m}, \quad &&\mathbf{b} = (\mathbf{b}^{\text{primal}},\mathbf{b}^{\text{dual}}), \end{aligned} \tag{2.28}$$

and

$$\begin{cases} \mathbf{b}_i^{\text{primal}} &= V_i f_i + \displaystyle\sum_{\ell\in i,\ell\in\Gamma_N}|\ell| g_N(\mathbf{x}_\ell) + \sum_{\ell\in i,\ell\in\Gamma_D}\left(r_\ell(\mathbf{u}^{\text{dual}})^+ + \gamma_\ell g_D(\mathbf{x}_\ell)\right), \forall i\in[\![1,n]\!],\\ \mathbf{b}_r^{\text{dual}} &= V_r f_r + \displaystyle\sum_{\tilde{\ell}\in r,\ell\in\Gamma_N}|\tilde{\ell}| g_N(\mathbf{x}_{\tilde{\ell}}) + \sum_{\tilde{\ell}\in r,\ell\in\Gamma_D}\zeta g_D(\mathbf{x}_{\tilde{\ell}}), \forall r\in[\![1,m]\!], \end{cases} \tag{2.29}$$

where ζ is a large value dedicated to taking into account of the Dirichlet boundary conditions by penalization (for example $\zeta = 10^{12}$).

We can then write this as the matrix-vector product

$$\mathbf{A}(\mathbf{u})\mathbf{u} = \begin{pmatrix} \mathbf{A}^{\text{primal}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) & \mathbf{0}\\ \mathbf{0} & \mathbf{A}^{\text{dual}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) \end{pmatrix}\begin{pmatrix} \mathbf{u}^{\text{primal}}\\ \mathbf{u}^{\text{dual}}\end{pmatrix} = \begin{pmatrix} \mathbf{b}^{\text{primal}}\\ \mathbf{b}^{\text{dual}}\end{pmatrix} = \mathbf{b}, \tag{2.30}$$

with

$$\begin{cases} A_{ii}^{\text{primal}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) = \lambda_i V_i + \displaystyle\sum_{\ell\in i,\ell\notin\Gamma_N}\left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{dual}})^-}{u_i}\right),\\ A_{ij}^{\text{primal}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) = -\displaystyle\sum_{\ell\in i\cap j}\left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{dual}})^+}{u_j}\right), & \forall i\neq j,\\ A_{rr}^{\text{dual}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) = \lambda_r V_r + \displaystyle\sum_{\tilde{\ell}\in r,\tilde{\ell}\notin\partial\Omega}\left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}(\mathbf{u}^{\text{primal}})^-}{u_r}\right) + \sum_{\tilde{\ell}\in r,\tilde{\ell}\in\gamma_D}\zeta,\\ A_{rs}^{\text{dual}}(\mathbf{u}^{\text{primal}},\mathbf{u}^{\text{dual}}) = -\displaystyle\sum_{\tilde{\ell}\in r\cap s,\tilde{\ell}\notin\partial\Omega}\left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}(\mathbf{u}^{\text{primal}})^+}{u_s}\right), & \forall r\neq s. \end{cases} \tag{2.31}$$

Thus the monotonicity enforcing procedure leads to two decoupled sparse matrices of size $m\times m$ and $n\times n$ depending on $\mathbf{u}$. This is a significant difference with the usual DDFV scheme for which all degrees of freedom are coupled, leading to a single $(m+n)\times(m+n)$ matrix independent of $\mathbf{u}$.
In the case of the monotonic diamond method, we obtain a system

$$\mathbf{A}^{\text{diamond}}(\mathbf{u}^{\text{primal}})\mathbf{u}^{\text{primal}} = \mathbf{b}^{\text{diamond}}, \tag{2.32}$$

with

$$\begin{cases} A_{ii}^{\text{diamond}}(\mathbf{u}^{\text{primal}}) = \displaystyle\sum_{\ell\in i,\ell\notin\Gamma_N}\left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{primal}})^-}{u_i}\right) + V_i\lambda_i,\\ A_{ij}^{\text{diamond}}(\mathbf{u}^{\text{primal}}) = -\displaystyle\sum_{\ell\in i\cap j}\left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{primal}})^+}{u_j}\right) & \forall i\neq j, \end{cases} \tag{2.33}$$

and

$$\mathbf{b}_i^{\text{diamond}} = V_i f_i + \sum_{\ell \in i, \ell \in \Gamma_D} \Big(r_\ell(\mathbf{u}^{\text{primal}})^+ + \gamma_\ell g_D(\mathbf{x}_\ell) \Big) + \sum_{\ell \in i, \ell \in \Gamma_N} |\ell| g_N(\mathbf{x}_\ell). \tag{2.34}$$

Remark 2.5.1. *Assuming that $f \geq 0$ and $g \geq 0$, all the components of the right hand side $\mathbf{b}$ are non-negative. Assuming moreover that f and g are not zero, then at least one component of $\mathbf{b}$ is positive.*

2.5.2 Picard iteration method

Both systems (2.30) and (2.32) are of the form $\mathbf{A}(\mathbf{u})\mathbf{u} = \mathbf{b}$. In order to solve them, we use a Picard iteration method. We start with an initial guess $\mathbf{u}^0 > 0$, compute the matrix $\mathbf{A}(\mathbf{u}^0)$ and solve $\mathbf{A}(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b}$. Repeating this process, we build a sequence $(\mathbf{u}^\nu)$ that, if it converges to a positive vector, tends to a solution of the scheme. We stop the algorithm when the difference $\mathbf{u}^{\nu+1} - \mathbf{u}^\nu$ between two successive iterates is small enough. To summarize, the following algorithm is used

$$\begin{aligned}
&\nu = 0 \\
&A(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b} \\
&\text{While } \frac{\|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_2}{\|\mathbf{u}^\nu\|_2} > \mu \\
&\quad A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\
&\quad \nu = \nu + 1.
\end{aligned}$$

For the monotonic DDFV scheme (2.26), for example, the linear system $\mathbf{A}(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b}$ writes

$$\left\{
\begin{aligned}
&- \sum_{\ell \in i, \ell \notin \partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u}^\nu)}{u_j^\nu} \right) u_j^{\nu+1} - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u}^\nu)}{u_i^\nu} \right) u_i^{\nu+1} \right) \\
&- \sum_{\ell \in i, \ell \in \partial\Omega} \left(\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u}^\nu)}{u_\ell^\nu} \right) u_\ell^{\nu+1} - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u}^\nu)}{u_i^\nu} \right) u_i^{\nu+1} \right) + V_i \lambda_i u_i^{\nu+1} = V_i f_i, \\
&- \sum_{\tilde{\ell} \in r} \left(\left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^+(\mathbf{u}^\nu)}{u_s^\nu} \right) u_s^{\nu+1} - \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^-(\mathbf{u}^\nu)}{u_r^\nu} \right) u_r^{\nu+1} \right) + V_r \lambda_r u_r^{\nu+1} = V_r f_r && \mathbf{x}_r \notin \Gamma_D, \\
&u_\ell^{\nu+1} = g_D(\mathbf{x}_\ell) && \mathbf{x}_\ell \in \Gamma_D, \\
&u_r^{\nu+1} = g_D(\mathbf{x}_r) && \mathbf{x}_r \in \Gamma_D, \\
&\left(\gamma_\ell + \frac{r_\ell^+(\mathbf{u}^\nu)}{u_\ell^\nu} \right) u_\ell^{\nu+1} - \left(\gamma_\ell + \frac{r_\ell^-(\mathbf{u}^\nu)}{u_i^\nu} \right) u_i^{\nu+1} = |\ell| g_N(\mathbf{x}_\ell) && \mathbf{x}_\ell \in \Gamma_N, \\
&\left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^+(\mathbf{u}^\nu)}{u_s^\nu} \right) u_s^{\nu+1} - \left(\gamma_{\tilde{\ell}} + \frac{r_{\tilde{\ell}}^-(\mathbf{u}^\nu)}{u_r^\nu} \right) u_r^{\nu+1} = |\tilde{\ell}| g_N(\mathbf{x}_{\tilde{\ell}}) && \mathbf{x}_{\tilde{\ell}} \in \Gamma_N.
\end{aligned}
\right. \tag{2.35}$$

Unfortunately, we are unable to prove that the above Picard algorithm converges. Nevertheless, we prove in Section 2.6.3 below that the scheme is well defined at each iteration of the algorithm, as soon as the initial guess $\mathbf{u}^0$ is positive.

2.6 Properties

2.6.1 Conservation

Proposition 2.6.1. *Assume that* $\mathbf{u} > \mathbf{0}$ *and consider homogeneous Neumann boundary conditions, then the scheme defined by* (2.27) *is conservative. Indeed it satisfies the equality*

$$\sum_{i=1}^{n} V_i \lambda_i u_i = \sum_{i=1}^{n} V_i f_i,$$

that is to say

$$\sum_{i=1}^{n} \left(-\sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) \right) = 0.$$

The proof is given in Appendix C.4.

2.6.2 Monotonicity

Consider the definition of an M-matrix (see for instance [87])

Definition 2.6.2. *An* $n \times n$ *matrix* $\mathbf{A}$ *that can be expressed in the forme* $\mathbf{A} = s\mathbf{I} - \mathbf{B}$*, where* $\mathbf{B} = (b_{ij})_{1 \le i,j \le n}$ *with* $b_{ij} \ge 0$*,* $1 \le i,j \le n$*, and* $s \ge \rho(\mathbf{B})$*, the maximum of the moduli of the eigenvalues of* $\boldsymbol{B}$*, is called an M-matrix.*

We use the following lemma

Lemma 2.6.3. *A matrix* $\mathbf{A} = (A_{ij})_{1 \le i,j \le n}$ *is an M-matrix if it satisfies the following inequalities*

$$\forall i \neq j, \quad A_{ij} \le 0, \quad \textit{and} \quad \forall i, \quad \sum_{j=1}^{n} A_{ij} \ge 0.$$

Moreover, if the last inequality is strict, we say that $\mathbf{A}$ *is a strict M-matrix.*

Proposition 2.6.4. *Assume that* $\mathbf{u} > \mathbf{0}$*. Then the matrices* $\mathbf{A}^{primal}$ *and* $\mathbf{A}^{dual}$ *defined by (2.31) and the matrix* $\mathbf{A}^{diamond}$ *defined by* (2.33) *are such that* $(\mathbf{A}^{primal})^T$*,* $(\mathbf{A}^{dual})^T$ *and* $(\mathbf{A}^{diamond})^T$ *are strict M-matrices.*

Proof. The matrix $\mathbf{A}^{\text{primal}}$ satisfies

$$\forall i \neq j, \quad A_{ij}^{\text{primal}} \le 0 \quad \text{and} \quad \forall j, \quad \sum_{i=1}^{n} A_{ij}^{\text{primal}} > 0.$$

Indeed we have, for all j

$$\sum_{i=1}^{n} A_{ij}^{\text{primal}} = \sum_{i=1}^{n} \left(\sum_{\ell \in i, \ell \notin \Gamma_N} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{dual}})^-}{u_i} \right) - \sum_{\ell \in i \cap j} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{dual}})^+}{u_j} \right) \right) + \lambda_j V_j.$$

Thanks to Proposition 2.6.1, only the boundary terms and the mass term remain, for all j

$$\sum_{i=1}^{n} A_{ij}^{\text{primal}} = \sum_{i=1}^{n} \sum_{\ell \in (i \cap \Gamma_D)} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u}^{\text{dual}})^-}{u_i} \right) + \lambda_j V_j > 0.$$

The above argument has been carried out on $\mathbf{A}^{\text{primal}}$ but the proof applies *mutatis mutandis* for $\mathbf{A}^{\text{dual}}$ or $\mathbf{A}^{\text{diamond}}$. □

Remark 2.6.5. *According to* (2.30)*, it is sufficient to prove that* $\mathbf{A}^{primal}$ *and* $\mathbf{A}^{dual}$ *are both strict M-matrices to prove that* $\mathbf{A}$ *is a strict M-matrix.*

Theorem 2.6.6. *Assume that $f > 0$ and $g > 0$. Let $\mathbf{A}$ and $\mathbf{b}$ be defined by (2.29)-(2.31) or (2.33)-(2.34). Then $\mathbf{A}^{-1}\mathbf{b} = \mathbf{u} \geq \mathbf{0}$.*

Proof. As $\mathbf{A}^T$ is a strict M-matrix $\mathbf{A}$ is invertible and its inverse has only non-negative entries (see for example [98], Corollary 3.20). In view of Remark 2.5.1, the right hand side is non-negative, hence $\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} \geq \mathbf{0}$. □

2.6.3 Well-posedness of the Picard iteration method

Proposition 2.6.7. *Assume that $f \geq 0$, $g \geq 0$, and either $\|f\|_{L^2(\Omega)} > 0$ or $\|g\|_{L^2(\partial\Omega)} > 0$. Assume moreover that $\mathbf{u}^0 > \mathbf{0}$. Then for all ν, $\mathbf{u}^\nu > \mathbf{0}$.*

To prove this property, we need to introduce the concept of irreducible matrix. We quote here [98, Definition 1.15].

Definition 2.6.8. *An $n \times n$ matrix $\mathbf{A}$ is **reducible** if there exits an $n \times n$ permutation matrix $\mathbf{P}$ such that*

$$\mathbf{PAP}^T = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix},$$

*where $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{22}$ are respectively $r \times r$, $r \times (n-r)$ and $(n-r) \times (n-r)$ sub-matrices with $1 \leq r < n$. If no such permutation matrix exists, then $\mathbf{A}$ is **irreducible**.*

The matrices $\mathbf{A}^{\text{primal}}$, $\mathbf{A}^{\text{dual}}$ defined by (2.31) and the matrix $\mathbf{A}^{\text{diamond}}$ defined by (2.33) are irreducible thanks to the following Lemma (see [98, Theorem 1.17]).

Lemma 2.6.9. *To any $n \times n$ matrix $\mathbf{A}$ we associate the graph of nodes 1,2,...,n and of directed edges connecting $\mathbf{x}_i$ to $\mathbf{x}_j$ if $A_{ij} \neq 0$. Then $\mathbf{A}$ is irreducible if and only if for any pair $i \neq j$ there exists a chain of edges that allows to go from $\mathbf{x}_i$ to $\mathbf{x}_j$,*

$$A_{i,k1} \neq 0 \rightarrow A_{k1,k2} \neq 0 \rightarrow \cdots \rightarrow A_{km,j} \neq 0.$$

With these definitions we can make use of the following theorem (see [98], Corollary 3.20).

Theorem 2.6.10. *If $\mathbf{A}$ is an irreducible strict M-matrix, then it is invertible and, for all i, j ($1 \leq i,j \leq n$), $(A^{-1})_{ij} > 0$.*

We are now in position to prove Proposition 2.6.7.

Proof of Proposition 2.6.7. We argue by induction on the index ν. We assume that $\mathbf{u}^\nu > 0$. Hence $(\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu))^T$ is a strict M-matrix (see Proposition 2.6.4). It is easy to check that $(\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu))^T$ is also irreducible. Thus, applying Theorem 2.6.10, all the entries of $(\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu))^{-T}$ are positive. Consequently, all the entries of $(\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu))^{-1}$ are positive. Using Remark 2.5.1, we know that all components of $\mathbf{b}$ are non-negative. Moreover, because of the assumption that either $\|f\|_{L^2(\Omega)} > 0$ or $\|g\|_{L^2(\partial\Omega)} > 0$, at least one component of $\mathbf{b}$ is positive. We thus have, for all i ($1 \leq i \leq n$)

$$u_i^{\nu+1} = \sum_{j=1}^{n} (\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu))^{-1}_{ij} b_j > 0,$$

since all terms of this sum are non-negative, with one at least that does not vanish.

The above argument has been carried out on $\mathbf{A}^{\text{primal}}$ but the proof applies *mutatis mutandis* for $\mathbf{A}^{\text{dual}}$ or $\mathbf{A}^{\text{diamond}}$. □

Proposition 2.6.7 shows that the condition $\mathbf{u}^\nu > \mathbf{0}$ remains satisfied during the Picard iteration method, which allows to define $\mathbf{A}^{\text{primal}}(\mathbf{u}^\nu)$ for all $\nu \geq 0$.

2.6.4 About the convergence of the fixed-point for the monotonic DDFV scheme

Recall that

- $\bar{\mathbf{u}} = ((\bar{u}_i)_{1\le i\le n},(\bar{u}_r)_{1\le r\le m})$ is the *exact* solution of (2.1),
- $\mathbf{u} = ((u_i)_{1\le i\le n},(u_r)_{1\le r\le m})$ is the *DDFV* solution defined by (2.25),
- $\mathbf{u}^\nu = ((u_i^\nu)_{1\le i\le n},(u_r^\nu)_{1\le r\le m}))$ is the ν-th iterate associated with the *monotonic DDFV* scheme, that is, the solution to (2.35).

For simplicity we will restrict ourselves to the case $\Gamma_N = \emptyset$ in (2.1), that is,

$$\begin{cases} -\boldsymbol{\nabla}\cdot\kappa\,(\boldsymbol{\nabla}\bar{u}) + \lambda\bar{u} = f & \text{in} \quad \Omega, \\ \bar{u} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{2.36}$$

We will make use of the following theorem, the proof of which is postponed to Appendix C.5.

Theorem 2.6.11. *Under assumptions* **H1**, **H2**, **H3** *the DDFV scheme defined by* (2.25) *is first-order accurate in the discrete L^2 norm, that is, there exists a constant C_1 independent of h such that*

$$\|\bar{\mathbf{u}} - \mathbf{u}\|_2 = \left(\sum_i V_i(\bar{u}(\mathbf{x}_i) - u_i)^2 + \sum_r V_r(\bar{u}(\mathbf{x}_r) - u_r)^2\right)^{1/2} \le C_1 h.$$

We will need the following lemma to prove Theorem 2.6.13.

Lemma 2.6.12. *Assume that there exists $\nu > 0$ and $\epsilon > 0$ such that*

$$\max\left(\max_i\left|\frac{u_i^{\nu+1} - u_i^\nu}{u_i^\nu}\right|, \max_r\left|\frac{u_r^{\nu+1} - u_r^\nu}{u_r^\nu}\right|\right) \le \epsilon. \tag{2.37}$$

Then the monotonic DDFV scheme (2.35) *writes*

$$\begin{cases} -\displaystyle\sum_{\ell\in i,\ell\notin\partial\Omega}\left(\gamma_\ell\left(u_j^{\nu+1} - u_i^{\nu+1}\right) + \delta_\ell\left(u_s^{\nu+1} - u_r^{\nu+1}\right)\right) + V_i\lambda_i u_i^{\nu+1} = V_i f_i + \rho_i^\nu, \\ -\displaystyle\sum_{\tilde{\ell}\in r,\tilde{\ell}\cap\partial\Omega=\emptyset}\left(\delta_{\tilde{\ell}}\left(u_\ell^{\nu+1} - u_i^{\nu+1}\right) + \gamma_{\tilde{\ell}}\left(u_s^{\nu+1} - u_r^{\nu+1}\right)\right) + V_r\lambda_r u_r^{\nu+1} = V_r f_r + \rho_r^\nu, \end{cases} \tag{2.38}$$

with

$$|\rho_i^\nu| \le C\epsilon, \quad |\rho_r^\nu| \le C\epsilon, \tag{2.39}$$

where C is a constant independant of h and ϵ.

Proof. Recall that, for all i, r, ν, $u_i^\nu > 0$ and $u_r^\nu > 0$. Suppose, for example, that

$$r_\ell(\mathbf{u}^\nu) = \delta_\ell(u_s^\nu - u_r^\nu) \ge 0, \quad r_{\tilde{\ell}}(\mathbf{u}^\nu) = \delta_{\tilde{\ell}}(u_j^\nu - u_i^\nu) \ge 0,$$

then $r_\ell^-(\mathbf{u}^\nu) = r_{\tilde{\ell}}^-(\mathbf{u}^\nu) = 0$ and the scheme (2.35) rewrites

$$\begin{cases} -\displaystyle\sum_{\ell\in i,\ell\notin\partial\Omega}\left(\gamma_\ell\left(u_j^{\nu+1} - u_i^{\nu+1}\right) + \delta_\ell(u_s^\nu - u_r^\nu)\frac{u_j^{\nu+1}}{u_j^\nu}\right) + V_i\lambda_i u_i^{\nu+1} = V_i f_i, \\ -\displaystyle\sum_{\tilde{\ell}\in r,\tilde{\ell}\cap\partial\Omega=\emptyset}\left(\delta_{\tilde{\ell}}(u_\ell^\nu - u_i^\nu)\frac{u_s^{\nu+1}}{u_s^\nu} + \gamma_{\tilde{\ell}}\left(u_s^{\nu+1} - u_r^{\nu+1}\right)\right) + V_r\lambda_r u_r^{\nu+1} = V_r f_r. \end{cases} \tag{2.40}$$

From assumption (2.37) we deduce that, for all i, r, there exists ϵ_i ($|\epsilon_i| \le \epsilon$) and ϵ_r ($|\epsilon_r| \le \epsilon$) such that

$$u_i^{\nu+1} = u_i^\nu + \epsilon_i u_i^\nu, \quad u_r^{\nu+1} = u_r^\nu + \epsilon_r u_r^\nu.$$

Inserting these values into (2.40) gives (2.38) with

$$\rho_i^\nu = \sum_{\ell \in i, \ell \notin \partial\Omega} \delta_\ell \left(\epsilon_r u_r^\nu - \epsilon_s u_s^\nu - \epsilon_j u_r^\nu + \epsilon_j u_s^\nu \right), \quad \rho_r^\nu = \sum_{\tilde{\ell} \in r, \tilde{\ell} \cap \partial\Omega = \emptyset} \delta_{\tilde{\ell}} \left(\epsilon_i u_i^\nu - \epsilon_s u_i^\nu + \epsilon_s u_\ell^\nu \right).$$

As a consequence,

$$|\rho_i^\nu| \leq 4 N_{max} \left(\max_\ell |\delta_\ell| \right) \left(\max_r u_r^\nu \right) \epsilon, \quad |\rho_r^\nu| \leq 3 N_{max} \left(\max_{\tilde{\ell}} |\delta_{\tilde{\ell}}| \right) \left(\max \left(\max_i u_i^\nu, \max_\ell u_\ell^\nu \right) \right) \epsilon,$$

where we recall that N_{max} is the maximum number of faces of primal and dual cells.

Considering Dirichlet boundary conditions, we have

$$u_\ell^\nu = g_D(\mathbf{x}_\ell) = u_\ell^{\nu+1}.$$

Thus, we have

$$\begin{cases} -\sum_{\ell \in i, \ell \notin \partial\Omega} \left(\gamma_\ell \left(u_j^{\nu+1} - u_i^{\nu+1} \right) + \delta_\ell \left(u_s^{\nu+1} - u_r^{\nu+1} \right) \right) + V_i \lambda_i u_i^{\nu+1} = V_i f_i + \rho_i^\nu, \\ -\sum_{\tilde{\ell} \in r, \tilde{\ell} \cap \partial\Omega = \emptyset} \left(\delta_{\tilde{\ell}} \left(u_\ell^{\nu+1} - u_i^{\nu+1} \right) + \gamma_{\tilde{\ell}} \left(u_s^{\nu+1} - u_r^{\nu+1} \right) \right) + V_r \lambda_r u_r^{\nu+1} = V_r f_r + \rho_r^\nu, \end{cases}$$

This concludes the proof. □

Theorem 2.6.13. *Assume that* **H1**, **H2**, **H3** *hold, and that the assumptions of Lemma 2.6.12 are satisfied. Then, there exists a constant C_4, independent of h and ϵ, such that*

$$\|\bar{\mathbf{u}} - \mathbf{u}^{\nu+1}\|_2 \leq C_1 h + C_4 \epsilon,$$

with C_1 the constant defined by Theorem 2.6.11.

Proof. System (2.25) writes

$$\mathbf{A}\mathbf{u} = \mathbf{f}$$

with

$$\mathbf{f} = \left((V_i f_i)_{1 \leq i \leq n}, (V_r f_r)_{1 \leq r \leq m} \right),$$

while system (2.38) writes

$$\mathbf{A}\mathbf{u}^{\nu+1} = \mathbf{f} + \mathbf{f}_\epsilon$$

with

$$\mathbf{f}_\epsilon = \left((\rho_i^\nu)_{1 \leq i \leq n}, (\rho_r^\nu)_{1 \leq r \leq m} \right).$$

By difference and thanks to the stability Lemma C.5.5, there exists a constant C_2 such that

$$\|\mathbf{u} - \mathbf{u}^{\nu+1}\|_2 \leq C_2 \|\mathbf{f}_\epsilon\|_2.$$

Thanks to Lemma 2.6.12 there exists a constant C_3 such that

$$\|\mathbf{f}_\epsilon\|_2 \leq C_3 \epsilon.$$

Then choosing $C_4 = C_2 C_3$ and applying the triangle inequality and Theorem 2.6.11 we obtain

$$\|\bar{\mathbf{u}} - \mathbf{u}^{\nu+1}\|_2 \leq \|\bar{\mathbf{u}} - \mathbf{u}\|_2 + \|\mathbf{u} - \mathbf{u}^{\nu+1}\|_2 \leq C_1 h + C_4 \epsilon,$$

which concludes the proof. □

Note that Theorem 2.6.13 is *not* a convergence theorem. Indeed if we make both h and ϵ tend to zero, the positive solution $\mathbf{u}^{\nu+1}$ tends to the DDFV numerical solution $\mathbf{u}$ which is only possible if $\mathbf{u}$ *itself* is non negative. Roughly speaking one can say that the (positive) numerical solution $\mathbf{u}^{\nu+1}$ obtained at the end of the iterative process is *close* to the (non necessarily positive) DDFV numerical solution $\mathbf{u}$ that itself is close to the exact solution $\bar{\mathbf{u}}$.

Note also that condition (2.37) is restrictive: in practice we rather use the condition $\|\mathbf{u}^{\nu+1} - \mathbf{u}^{\nu}\|_\infty \leq \epsilon\|\mathbf{u}^{\nu}\|_\infty$ or $\|\mathbf{u}^{\nu+1} - \mathbf{u}^{\nu}\|_2 \leq \epsilon\|\mathbf{u}^{\nu}\|_2$ as a stopping criterion.

2.7 Numerical experiments

Given $\Omega =]0,1[^2$, κ a diffusion coefficient and g a function defined on $\partial\Omega$, consider Problem (2.1) with $\lambda = 0$ and $\Gamma_N = \emptyset$

$$\begin{cases} -\boldsymbol{\nabla} \cdot (\kappa \boldsymbol{\nabla} \bar{u}) = f & \text{in} \quad \Omega, \\ \bar{u} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{2.41}$$

In addition to Cartesian meshes we will use the two following types of meshes (see Figure 2.4):

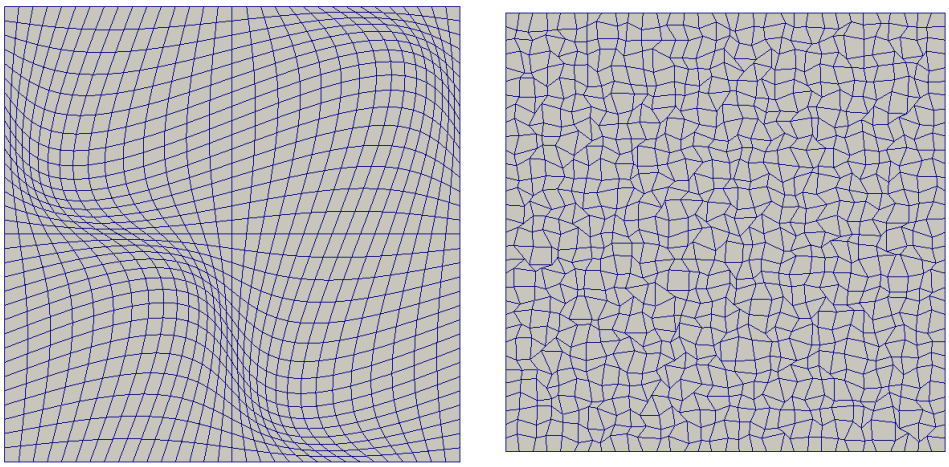
1. deformed meshes, the deformation of which from the Cartesian mesh is given by

$$(x,y) \to (x + 0.1\sin(2\pi x)\sin(2\pi y), y + 0.1\sin(2\pi x)\sin(2\pi y)),$$

2. randomly deformed meshes, the deformation of which from the unit Cartesian mesh with cells of size Δx is given by

$$(x,y) \to 0.1(x,y) + 0.9(x + 0.45a\Delta x, y + 0.45b\Delta x),$$

where a, b are random numbers distributed according to the uniform law on $[-1,1]$.



(a) A deformed mesh (b) A random mesh

Fig. 2.4 – Examples of deformed meshes.

The L^2 and H^1-errors used in the following tests are respectively given by

$$\frac{\|\mathbf{u} - \bar{\mathbf{u}}\|_2}{\|\bar{\mathbf{u}}\|_2} \quad \text{and} \quad \frac{\|\boldsymbol{\nabla}_h \mathbf{u} - \boldsymbol{\nabla}\bar{u}\|_2}{\|\boldsymbol{\nabla}\bar{u}\|_2},$$

where

$$\|\boldsymbol{\nabla}\bar{u}\|_2 = \left(\sum_\ell V_\ell \|\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\|^2\right)^{1/2},$$

$$\|\boldsymbol{\nabla}_h\mathbf{u} - \boldsymbol{\nabla}\bar{u}\|_2 = \left(\sum_\ell V_\ell \left\|\frac{1}{2}\frac{1}{V_\ell}\left((u_j-u_i)\mathbf{x}_s\mathbf{x}_r^\perp + (u_s-u_r)\mathbf{x}_i\mathbf{x}_j^\perp\right) - \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\right\|^2\right)^{1/2},$$

V_ℓ being the surface of the diamond cell ℓ.

For DDFV type schemes we plot on figures 2.7, 2.8, 2.10, 2.9, the *primal* numerical values while on tables 2.2, 2.4, 2.3, the maxima and minima are computed over *both* primal *and* dual values.

2.7.1 Accuracy

Three simple benchmarks are proposed to assess the accuracy of our monotonic schemes in comparison with the usual (non monotonic) DDFV scheme. For these three tests, we choose $\epsilon = 10^{-12}$ as the stopping criterion of the fixed point algorithm.

2.7.1.1 Checking the preservation of linear solutions

Given $\kappa(\mathbf{x}) = 1$ $f(\mathbf{x}) = 0$ and $g(\mathbf{x}) = -x - y + 2$, the positive linear function $\bar{u}(\mathbf{x}) = -x - y + 2$ is solution to (2.41). We perform a study of this problem on the deformed mesh (see Figure 2.4a) with 32×32 cells for each of the three schemes. The L^2 and H^1 errors between the exact solution $\bar{\mathbf{u}}$ and the approximated one $\mathbf{u}$ are reported in Table 2.1. This agrees with the theory (see Appendix D.3) since the error is zero, to machine precision, when $\bar{u}$ is a polynomial of degree 1.

Scheme	L^2-error	H^1-error
DDFV	$2.58e-15$	$4.46e-14$
Monotonic DDFV	$9.42e-15$	$6.30e-13$
Monotonic diamond (degree 1)	$1.05e-14$	$1.02e-13$

Tab. 2.1 – Comparison between the different schemes for the positive linear solution to problem of Section 2.7.1.1.

2.7.1.2 Anisotropic diffusion coefficient

Given

$$\boldsymbol{\kappa}(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad f(\mathbf{x}) = 3\pi^2 \sin(\pi x)\sin(\pi y), \quad g(\mathbf{x}) = 0,$$

the function $\bar{u}(\mathbf{x}) = \sin(\pi x)\sin(\pi y)$ is solution to (2.41). We perform a convergence study for this problem with a sequence of successively refined deformed meshes like the one of Figure 2.4a.

Results are summarized in Figure 2.5 which shows that all schemes are second-order accurate in the L^2 norm. Of course, similar results may be obtained for a scalar-valued diffusion coefficient κ. We see that the error in H^1-norm is second-order convergent for DDFV methods while the diamond scheme is only first-order accurate in the H^1 norm. However, if we perform a second-order reconstruction of the gradient, we also obtain second-order accuracy of the diamond scheme for the H^1 norm.

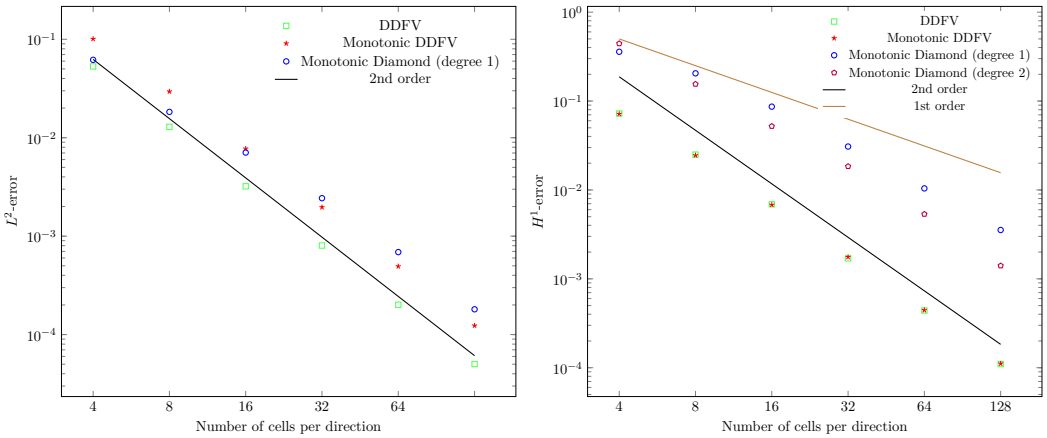


Fig. 2.5 – L^2 (on the left) and H^1 (on the right) errors for problem of Section 2.7.1.2.

2.7.1.3 Discontinuous diffusion coefficient

Recall that we have assumed that possible discontinuities of the diffusion coefficient κ occur only along the primal cell faces. Given

$$\kappa(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad x \leq \dfrac{1}{2} \\ 2 & \text{if} \quad x > \dfrac{1}{2} \end{cases}, \quad f(\mathbf{x}) = 2\pi^2 \cos(\pi x)\cos(\pi y) + 20, \quad g(\mathbf{x}) = 0,$$

the function

$$\bar{u}(\mathbf{x}) = \begin{cases} \cos(\pi x)\cos(\pi y) - 10x^2 + 12 & \text{if} \quad x \leq \dfrac{1}{2}, \\ \dfrac{1}{2}\cos(\pi x)\cos(\pi y) - 5x^2 + \dfrac{43}{4} & \text{if} \quad x > \dfrac{1}{2}, \end{cases}$$

is solution to (2.41). We perform a convergence study for this problem with a sequence of successively refined deformed meshes as shown in Figure 2.4a.

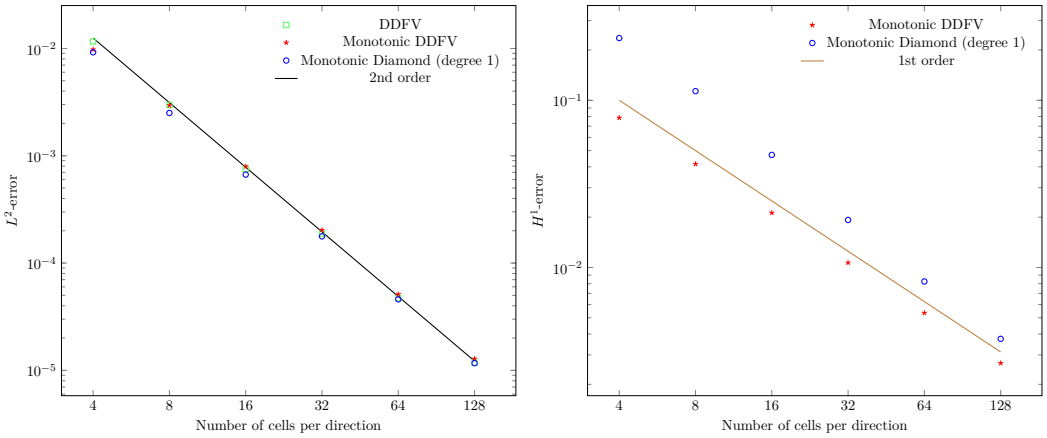


Fig. 2.6 – L^2 (on the left) and H^1 (on the right) errors for problem of Section 2.7.1.3.

Figure 2.6 shows that, in the present case of a discontinuous κ, the results are similar to those of the continuous case, that is to say, the scheme is second-order accurate. However, both schemes are only first-order accurate in H^1 norm in this case.

2.7.2 Monotonicity test problems

We propose two benchmarks to compare the usual DDFV scheme, which can give nonpositive solutions, with our monotonic diamond and DDFV schemes which always give nonnegative solutions.

2.7.2.1 Tensor-valued coefficient κ and square domain with a square hole

Consider the square domain with a square hole $\Omega =]0,1[^2 \setminus \left[\frac{4}{9},\frac{5}{9}\right]^2$, $f(\mathbf{x}) = 0$ in Ω and $g(\mathbf{x}) = 0$ (resp. $g(\mathbf{x}) = 2$) on the external (resp. internal) boundary. We have chosen

$$\boldsymbol{\kappa} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 10^4 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{\pi}{6}.$$

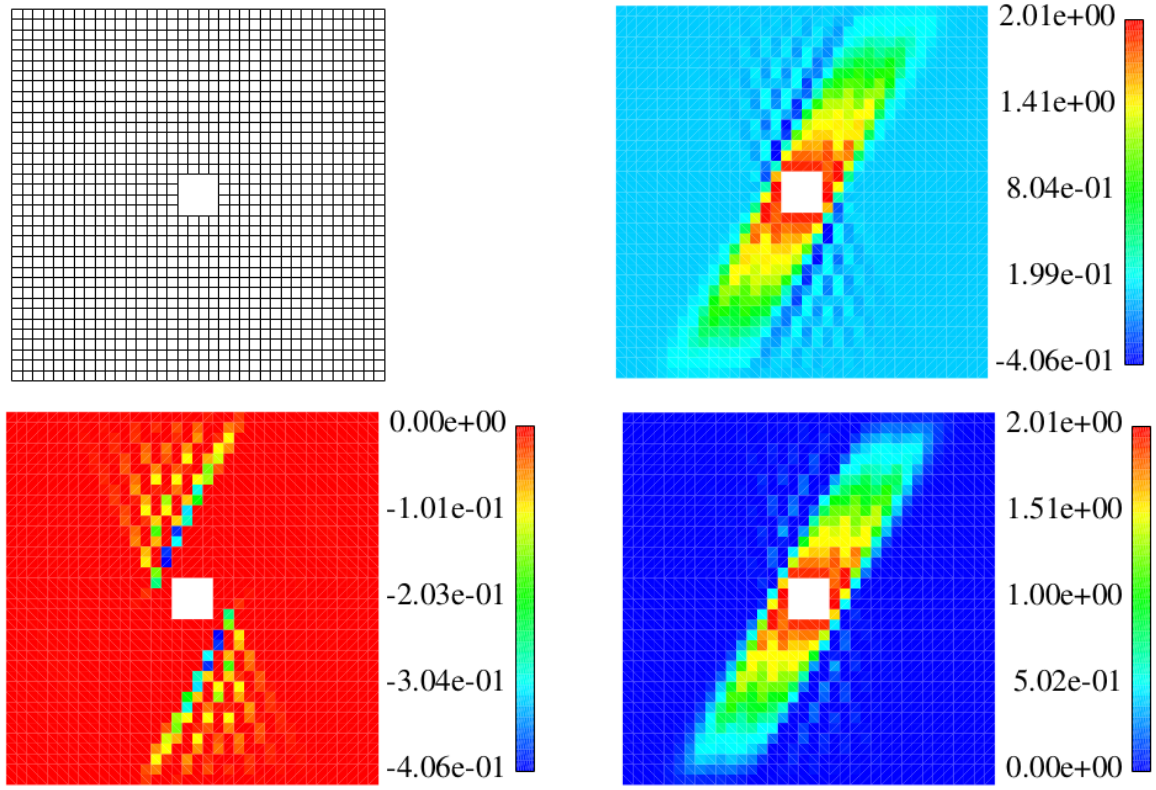


Fig. 2.7 – Mesh (top, left), DDFV solution to problem of section 2.7.2.1 (top, right) and its negative (bottom, left) and positive (bottom, right) parts.

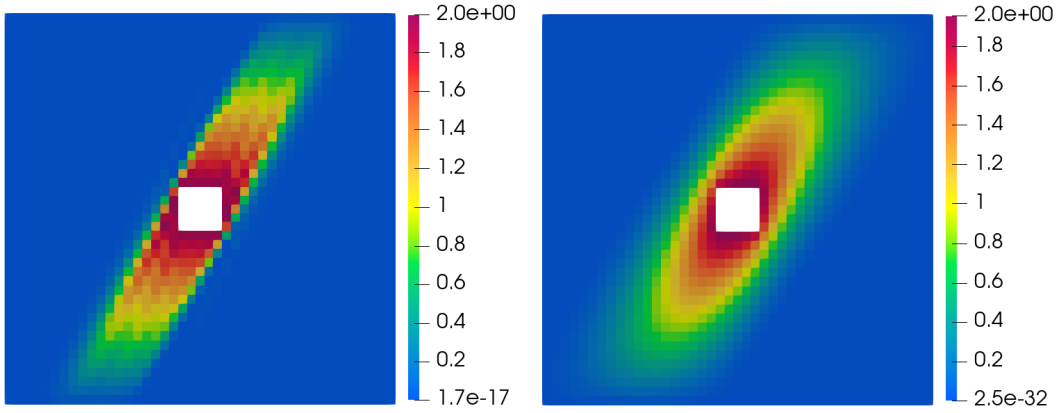


Fig. 2.8 – Monotonic DDFV (on the left) and diamond (degree 1, on the right) solutions to problem of section 2.7.2.1.

We compare the results obtained with the monotonic diamond and DDFV schemes on a Cartesian mesh with 36 cells per direction. The stopping criterion of the fixed point algorithm is $\epsilon = 10^{-12}$.

Figure 2.7 shows the mesh, the DDFV solution and its negative and positive parts. Figure 2.8 displays the monotonic DDFV and diamond solutions while Table 2.2 gives the minimum and the maximum of each solution.

Scheme	Minimum of the solution	Maximum of the solution
DDFV	-4.59×10^{-1}	2.05
Monotonic DDFV	1.65×10^{-17}	2.01
Monotonic diamond (degree 1)	2.46×10^{-32}	1.95

Tab. 2.2 – Minimum and maximum of the numerical solution to the problem of section 2.7.2.1 for the Cartesian mesh with 36 cells per direction.

While the solution obtained with the usual DDFV scheme has a negative minimum we can see that the solutions obtained with the monotonic methods are always positive, as expected.

2.7.2.2 Fokker-Planck type diffusion equation

This benchmark is a simplified version of the one from [69]. Given $\Omega =]-50,50[^2$, $T = 250$, $\mathbf{v} = (v_x,v_y)$ the velocity variable and $\mathbf{V} = (-20,20)$ the averaged velocity, we are looking for the distribution function $\bar{u} = \bar{u}(\mathbf{v},t)$, solution to the simplified Fokker-Planck equation

$$\begin{cases} \dfrac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla}_{\mathbf{v}} \left(\boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u}\right) = 0 & \text{in} \quad \Omega \times [0,T], \\ \boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u} \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times [0,T], \\ \bar{u}(0) = \bar{u}^0 & \text{in} \quad \Omega, \end{cases} \tag{2.42}$$

where the diffusion coefficient $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathbf{v})$ and the initial condition $\bar{u}^0$ are given by

$$\boldsymbol{\kappa}(\mathbf{v}) = \mathbf{I} - \frac{1}{\|\mathbf{v}\|^2} \mathbf{v} \otimes \mathbf{v}, \quad \bar{u}^0(\mathbf{v}) = \frac{1}{\pi} \exp(-\|\mathbf{v} - \mathbf{V}\|^2).$$

Note that the full Fokker-Planck equation would read as

$$\frac{\partial \bar{u}}{\partial t} + \boldsymbol{\nabla}_{\mathbf{v}} \cdot (\mathbf{v} \bar{u}) - \boldsymbol{\nabla}_{\mathbf{v}} \left(\boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u}\right) = 0.$$

It is well known that the n-order moments of $\bar{u}$ $(0 \leq n \leq 2)$ are preserved over the time

$$\frac{d}{dt}\left(\int_\Omega \bar{u}\right) = 0, \quad \frac{d}{dt}\left(\int_\Omega \mathbf{v} \bar{u}\right) = \mathbf{0}, \quad \frac{d}{dt}\left(\int_\Omega \|\mathbf{v}\|^2 \bar{u}\right) = 0.$$

The backward Euler scheme is used for time discretization.

To limit the calculation time, the stopping criterion of the fixed point algorithm is $\epsilon = 10^{-5}$. Figure 2.10 (resp. 2.9) displays the DDFV (resp. monotonic DDFV and diamond) numerical solutions obtained with the Cartesian, deformed and random meshes of 200^2 cells. Table 2.4 gives the minima and maxima of the DDFV solution for a sequence of refined Cartesian meshes and Table 2.3 gives the minima and the maxima of the numerical solution obtained with the DDFV, monotonic DDFV and diamond schemes. We observe that the minima of the DDFV solution are negative but converge to zero as h tends to zero while the minima of the solutions to monotonic schemes always remain non negative, as expected. Compared to both the non monotonic and monotonic DDFV schemes the

monotonic diamond scheme is more diffusive (in the radial direction). This could be explained by the use of a larger stencil required for polynomial reconstruction.

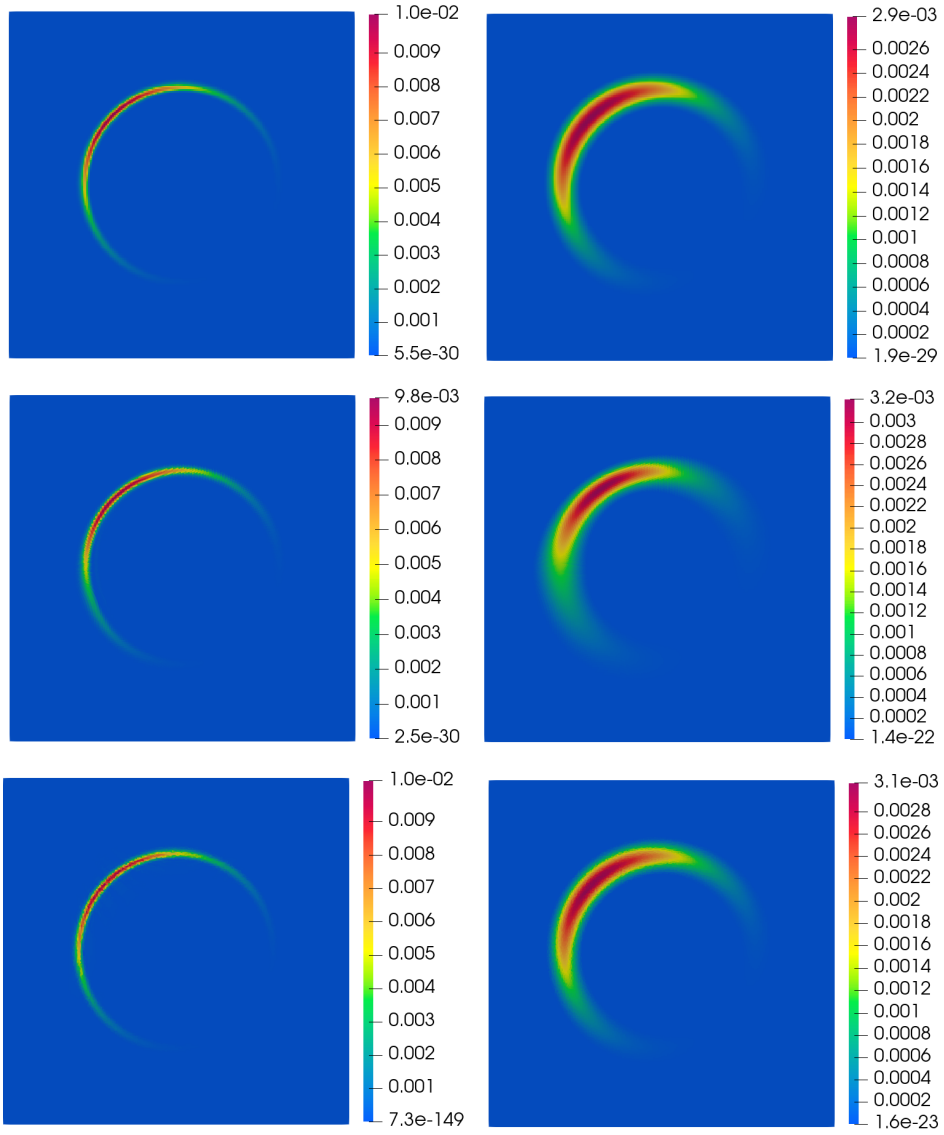


Fig. 2.9 – Monotonic DDFV (on the left) and diamond (degree 1, on the right) solutions to (2.42) at time $T = 250$ on the Cartesian (top), deformed (middle) and random (bottom) mesh of 200×200 cells.

	Scheme	Cartesian mesh	Deformed mesh	Random mesh
Minima	DDFV	-2.48×10^{-4}	-1.25×10^{-3}	-2.41×10^{-3}
	Monotonic DDFV	5.46×10^{-30}	2.53×10^{-30}	4.63×10^{-40}
	Monotonic diamond (degree 1)	1.86×10^{-29}	1.42×10^{-22}	1.58×10^{-23}
Maxima	DDFV	1.04×10^{-2}	1.04×10^{-2}	1.14×10^{-2}
	Monotonic DDFV	1.04×10^{-2}	0.97×10^{-2}	1.02×10^{-2}
	Monotonic diamond (degree 1)	0.29×10^{-2}	0.32×10^{-2}	0.31×10^{-2}

Tab. 2.3 – Minima and maxima of the numerical solutions to (2.42) at time $T = 250$ on the three types of 200×200 cells meshes.

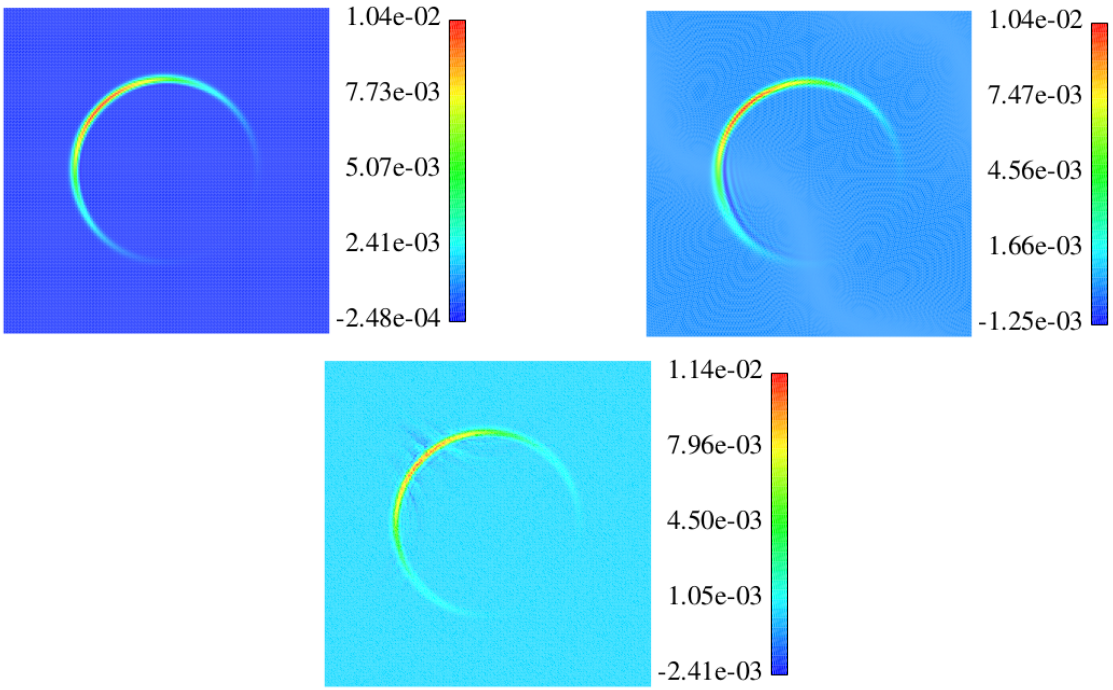


Fig. 2.10 – DDFV solution to (2.42) at time $T = 250$ on the Cartesian (top left), deformed (top right) and random (bottom) mesh of 200×200 cells.

	Number of cells	Cartesian mesh	Deformed mesh	Random mesh
Minima	100×100	-1.89×10^{-3}	-2.38×10^{-3}	-3.15×10^{-3}
	200×200	-2.48×10^{-4}	-1.25×10^{-3}	-2.41×10^{-3}
	400×400	-6.32×10^{-13}	-2.14×10^{-4}	-9.92×10^{-4}
	800×800	-1.66×10^{-13}	-7.95×10^{-7}	-7.63×10^{-4}
	1600×1600	-8.53×10^{-14}	-1.97×10^{-7}	-4.58×10^{-4}
Maxima	100×100	1.19×10^{-2}	1.16×10^{-2}	1.65×10^{-2}
	200×200	1.04×10^{-2}	1.04×10^{-2}	1.14×10^{-2}
	400×400	1.01×10^{-2}	1.01×10^{-2}	1.09×10^{-2}
	800×800	1.01×10^{-2}	1.01×10^{-2}	1.09×10^{-2}
	1600×1600	1.01×10^{-2}	1.01×10^{-2}	1.08×10^{-2}

Tab. 2.4 – Minima and maxima of the DDFV solution of (2.42) at time $T = 250$ on refined Cartesian meshes.

The conservation of the zero-order moment of $\bar{u}$ at the discrete level is a property of our schemes. It is more challenging to obtain a conservation of a discrete equivalent of the second-order moment. Thanks to the identity

$$\mathbf{v} = \frac{1}{2}\boldsymbol{\nabla}_{\mathbf{v}}(\|\mathbf{v}\|^2),$$

one can introduce an approximation $\mathbf{v}_\ell$ of $\mathbf{v}$ in the diamond cell I_ℓ by using the Green-Gauss formula

$$\begin{cases} \mathbf{v}_\ell = \dfrac{1}{4}\dfrac{1}{V_\ell}\left((\|\mathbf{v}_j\|^2 - \|\mathbf{v}_i\|^2)\mathbf{N}_{i\ell} + (\|\mathbf{v}_s\|^2 - \|\mathbf{v}_r\|^2)\mathbf{N}_{r\tilde{\ell}}\right) & \ell \notin \partial\Omega, \\ \mathbf{v}_\ell = \dfrac{1}{4}\dfrac{1}{V_\ell}\left((\|\mathbf{v}_\ell\|^2 - \|\mathbf{v}_i\|^2)\mathbf{N}_{i\ell} + (\|\mathbf{v}_s\|^2 - \|\mathbf{v}_r\|^2)\mathbf{N}_{r\tilde{\ell}}\right) & \ell \in \partial\Omega. \end{cases} \tag{2.43}$$

We then prove the following proposition.

Proposition 2.7.1. *Consider the DDFV solution to* (2.42)*, that is,*

$$\begin{cases} V_i \dfrac{u_i^{n+1}-u_i^n}{\Delta t} - \dfrac{1}{2}\displaystyle\sum_{\ell\in i,\ell\notin\partial\Omega} \dfrac{1}{V_\ell}\Big((u_j^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell}\boldsymbol{\kappa}_\ell\mathbf{N}_{i\ell} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\boldsymbol{\kappa}_\ell\mathbf{N}_{i\ell}\Big) \\ \qquad\qquad -\dfrac{1}{2}\displaystyle\sum_{\ell\in i,\ell\in\partial\Omega} \dfrac{1}{V_\ell}\Big((u_\ell^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell}\boldsymbol{\kappa}_\ell\mathbf{N}_{i\ell} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\boldsymbol{\kappa}_\ell\mathbf{N}_{i\ell}\Big) = 0, \\ V_r \dfrac{u_r^{n+1}-u_r^n}{\Delta t} - \dfrac{1}{2}\displaystyle\sum_{\ell\in r,\ell\notin\partial\Omega} \dfrac{1}{V_\ell}\Big((u_j^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}}\Big) \\ \qquad\qquad -\dfrac{1}{2}\displaystyle\sum_{\ell\in r,\ell\in\partial\Omega} \dfrac{1}{V_\ell}\Big((u_\ell^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}}\Big) = 0, \\ \dfrac{1}{2}\dfrac{1}{V_\ell}\Big((u_\ell^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\boldsymbol{\kappa}_\ell\mathbf{N}_{r\tilde{\ell}}\Big) = 0 \qquad \mathbf{x}_\ell\in\partial\Omega, \end{cases} \tag{2.44}$$

with the following approximations of $\boldsymbol{\kappa}$ *in a diamond cell* I_ℓ *such that* $\mathbf{v}_\ell \notin \partial\Omega$

$$\boldsymbol{\kappa}_\ell = \mathbf{I} - \frac{1}{\|\mathbf{v}_\ell\|^2}\mathbf{v}_\ell\otimes\mathbf{v}_\ell,$$

with $\mathbf{v}_\ell$ *calculated by* (2.43)*.*

Let E^n *be the following discrete equivalent of the second-order moment*

$$E^n = \frac{1}{2}\left(\sum_i V_i\|\mathbf{v}_i\|^2 u_i^n + \sum_r V_r\|\mathbf{v}_r\|^2 u_r^n\right).$$

Then, for all $n\geq 0$*,* $E^n = E^0$*.*

Proof. We multiply the first (resp. second) equation of (2.44) by $\|\mathbf{v}_i\|^2$ (resp. $\|\mathbf{v}_r\|^2$) and sum over primal (resp. dual) cells i (resp. r). Adding these two sums we get

$$\frac{1}{\Delta t}\left(\sum_i V_i\|\mathbf{v}_i\|^2 u_i^{n+1} + \sum_r V_r\|\mathbf{v}_r\|^2 u_r^{n+1} - \sum_i V_i\|\mathbf{v}_i\|^2 u_i^n - \sum_r V_r\|\mathbf{v}_r\|^2 u_r^n\right)$$

$$-\frac{1}{2}\sum_\ell \frac{1}{V_\ell}\Big((\|\mathbf{v}_i\|^2-\|\mathbf{v}_j\|^2)\mathbf{N}_{i\ell} + (\|\mathbf{v}_r\|^2-\|\mathbf{v}_s\|^2)\mathbf{N}_{r\tilde{\ell}}\Big)\boldsymbol{\kappa}_\ell\Big((u_j^{n+1}-u_i^{n+1})\mathbf{N}_{i\ell} + (u_s^{n+1}-u_r^{n+1})\mathbf{N}_{r\tilde{\ell}}\Big) = 0.$$

Then, noting that $\boldsymbol{\kappa}_\ell\mathbf{v}_\ell = \mathbf{0}$, we obtain thanks to (2.43)

$$\sum_i V_i\|\mathbf{v}_i\|^2 u_i^{n+1} + \sum_r V_r\|\mathbf{v}_r\|^2 u_r^{n+1} = \sum_i V_i\|\mathbf{v}_i\|^2 u_i^n + \sum_r V_r\|\mathbf{v}_r\|^2 u_r^n,$$

that is, $E^{n+1} = E^n$. □

The numerical results displayed in Figure 2.11 show that the second order moment is conserved over time for the non-monotonic DDFV scheme, as it has been proved. However, it is not exactly conserved with monotonic DDFV scheme because we do not exactly solve the DDFV system. Nevertheless, the conservation error is far lower than for the positive diamond scheme.

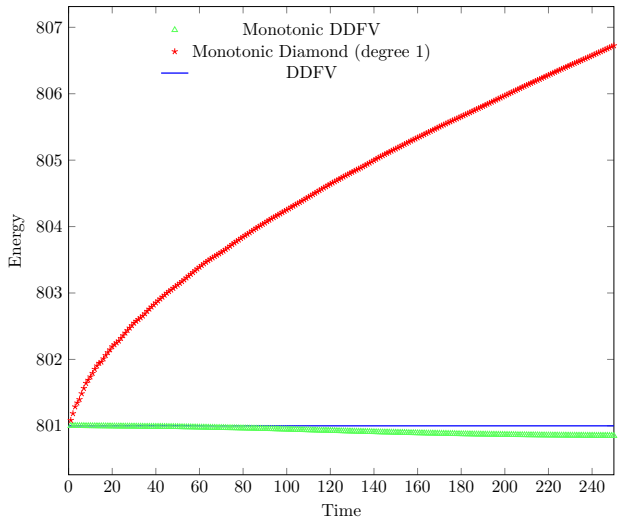


Fig. 2.11 – Variation of energy over time for the 3 schemes on cartesian mesh of 200×200 cells.

2.7.2.3 Convection-diffusion type diffusion equation

This test was added after the publication of the article [9].

Given $\Omega =]-1,1[^2$, $T = 1$, consider the problem

$$\begin{cases} \dfrac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot (\boldsymbol{\kappa} \boldsymbol{\nabla} \bar{u}) = 0 & \text{in} \quad \Omega \times [0,T], \\ \boldsymbol{\kappa} \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times [0,T], \\ \bar{u}(0) = \bar{u}^0 & \text{in} \quad \Omega, \end{cases} \tag{2.45}$$

where the diffusion coefficient $\boldsymbol{\kappa}$ is given by

$$\boldsymbol{\kappa} = \begin{pmatrix} 10^{-8} & \pi(x^2+y^2) \\ -\pi(x^2+y^2) & 10^{-8} \end{pmatrix}, \tag{2.46}$$

and the initial condition $\bar{u}^0$ is given by

$$\bar{u}^0 = \exp\left(-\frac{(x-x_0)^2 + (y-y_0)^2}{R^2} \right),$$

with $R = 10^{-1}$, $x_0 = 0$, $y_0 = 5 \times 10^{-1}$.

The solution $\bar{u}$ of (2.45) should remain positive, and the non-monotonic DDFV scheme produces non-physical negative values. As we will see, our monotonic scheme gives a positive solution.

Proposition 2.7.2. *Let Ω be a bounded open connected domain of $\mathbb{R}^2$ and let $\alpha \in C^1(\Omega)$. Consider the equation*

$$\frac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot (\bar{u}\mathbf{v}) = 0, \tag{2.47}$$

with $\mathbf{v} = \boldsymbol{\nabla} \times \alpha$, noting $\boldsymbol{\nabla}\times$ the rotational operator, with the convention $\boldsymbol{\nabla} \times \alpha = \begin{pmatrix} \frac{\partial \alpha}{\partial y} \\ -\frac{\partial \alpha}{\partial x} \end{pmatrix}$. Then, any $\bar{u} \in C^1([0,T],H^2(\Omega))$ solution to (2.47) is also solution to

$$\frac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot (\boldsymbol{\kappa} \boldsymbol{\nabla} \bar{u}) = 0, \tag{2.48}$$

with

$$\boldsymbol{\kappa} = \begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix}.$$

Conversely, any $\bar{u} \in C^1([0,T],H^2(\Omega))$ *solution to* (2.48) *is solution to* (2.47).

Proof. First, the diffusion equation

$$\frac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot (\boldsymbol{\kappa} \boldsymbol{\nabla} \bar{u}) = 0,$$

with

$$\boldsymbol{\kappa} = \begin{pmatrix} 0 & \alpha \\ -\alpha & 0 \end{pmatrix},$$

gives

$$\frac{\partial \bar{u}}{\partial t} - \frac{\partial \alpha}{\partial x}\frac{\partial \bar{u}}{\partial y} + \frac{\partial \bar{u}}{\partial x}\frac{\partial \alpha}{\partial y} = 0.$$

Second, the convection equation

$$\frac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot (\bar{u}\mathbf{v}) = 0,$$

with $\mathbf{v} = \boldsymbol{\nabla} \times \alpha$, gives

$$\frac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla} \cdot \left(\bar{u} \begin{pmatrix} \frac{\partial \alpha}{\partial y} \\ -\frac{\partial \alpha}{\partial x} \end{pmatrix}, \right) = 0,$$

that is to say

$$\frac{\partial \bar{u}}{\partial t} - \frac{\partial \alpha}{\partial x}\frac{\partial \bar{u}}{\partial y} + \frac{\partial \bar{u}}{\partial x}\frac{\partial \alpha}{\partial y} = 0.$$

Thus, the convection and the diffusion equations are equivalent. □

Cartesian mesh refinement	Time $t = 0.25$	Time $t = 0.50$	Time $t = 0.75$	Time $t = 1$
60×60	-0.11	-0.19	-0.22	-0.23
120×120	-2.9×10^{-3}	-3.8×10^{-2}	-8.5×10^{-2}	-1.23×10^{-1}
240×240	-6.4×10^{-11}	-5.8×10^{-6}	-4.7×10^{-4}	-3.18×10^{-3}
480×480	-1.87×10^{-10}	-2.9×10^{-10}	-4.14×10^{-10}	-4.2×10^{-10}

Tab. 2.5 – Minima of the numerical solutions to (2.45) with the DDFV scheme at different times on cartesian meshes.

Figure 2.12 (resp. Figure 2.13 and Figure 2.14) shows the solution obtained with the monotonic DDFV (resp. DDFV and Upwind) scheme on a Cartesian mesh with 60 cells per direction. We note that the solution obtained with the monotonic DDFV scheme is positive unlike the one obtained with the DDFV scheme. The monotonic DDFV solution is much more diffusive (in the radial direction) than the DDFV solution but the latter exhibits oscillations unlike the solution to the monotonic scheme. The Upwind solution is highly inaccurate because it is too diffuse (much more than the monotonic DDFV scheme). We notice that the more the mesh is refined, the less the DDFV scheme solution is negative (see Table 2.5) and the less oscillations there are.

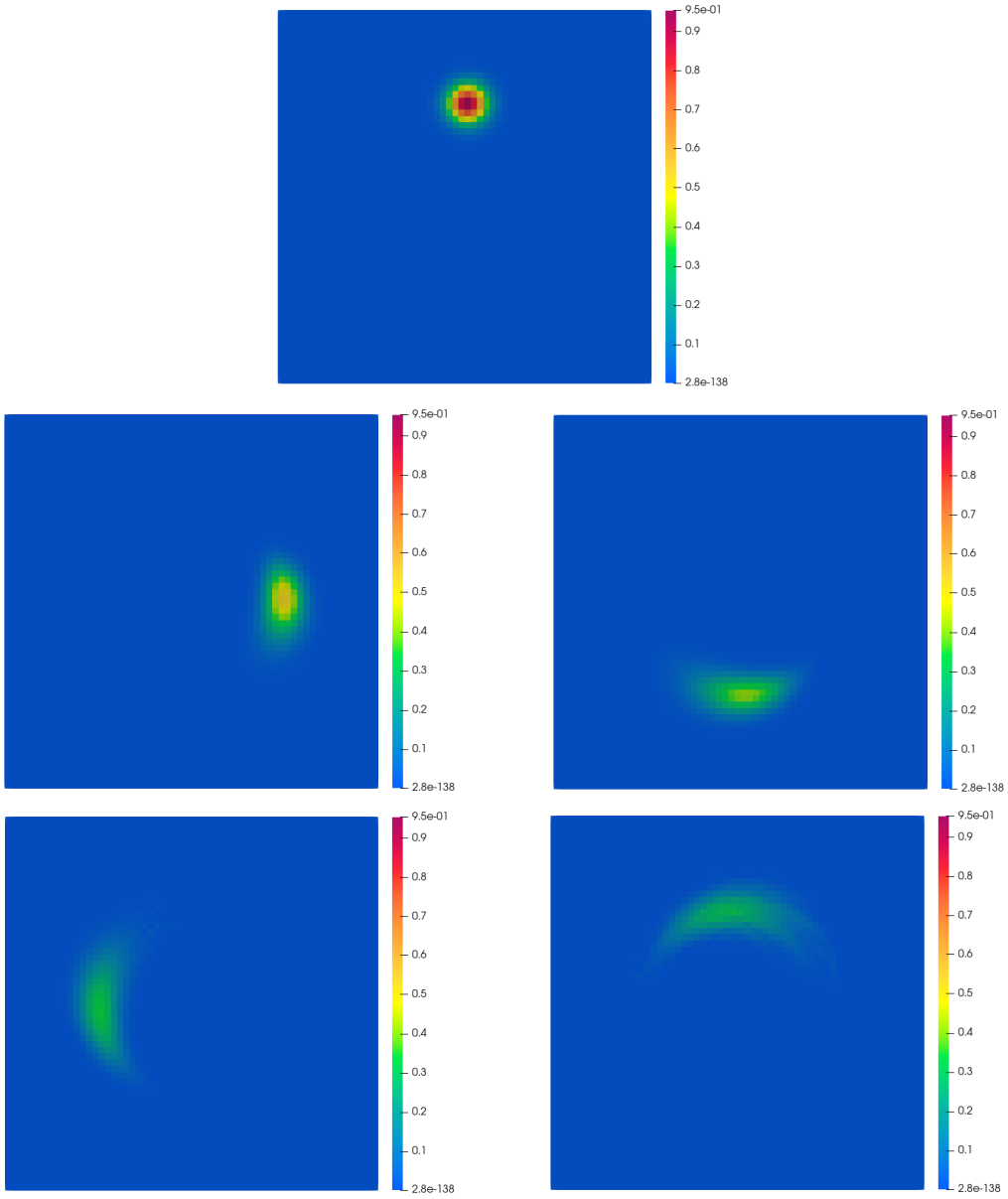


Fig. 2.12 – Solutions at time $t = 0$ (top), $t = 0.25$ (middle left), $t = 0.5$ (middle right), $t = 0.75$ (bottom left) and final time (bottom right) obtained with the monotonic DDFV scheme with a cartesian mesh of 60×60 cells.

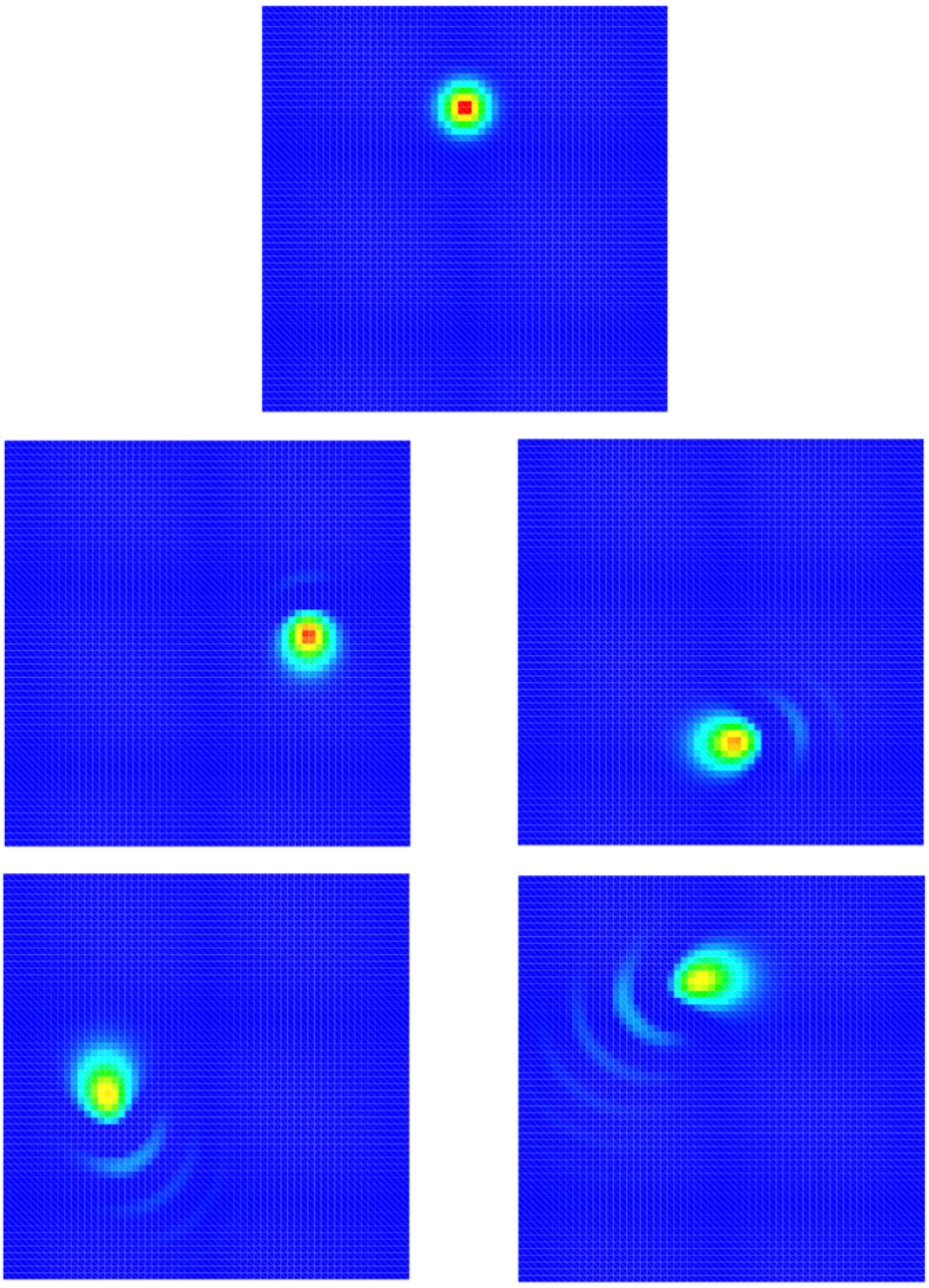


Fig. 2.13 – Solutions at time $t = 0$ (top), $t = 0.25$ (middle left), $t = 0.5$ (middle right), $t = 0.75$ (bottom left) and final time (bottom right) obtained with the DDFV scheme with a cartesian mesh of 60×60 cells.

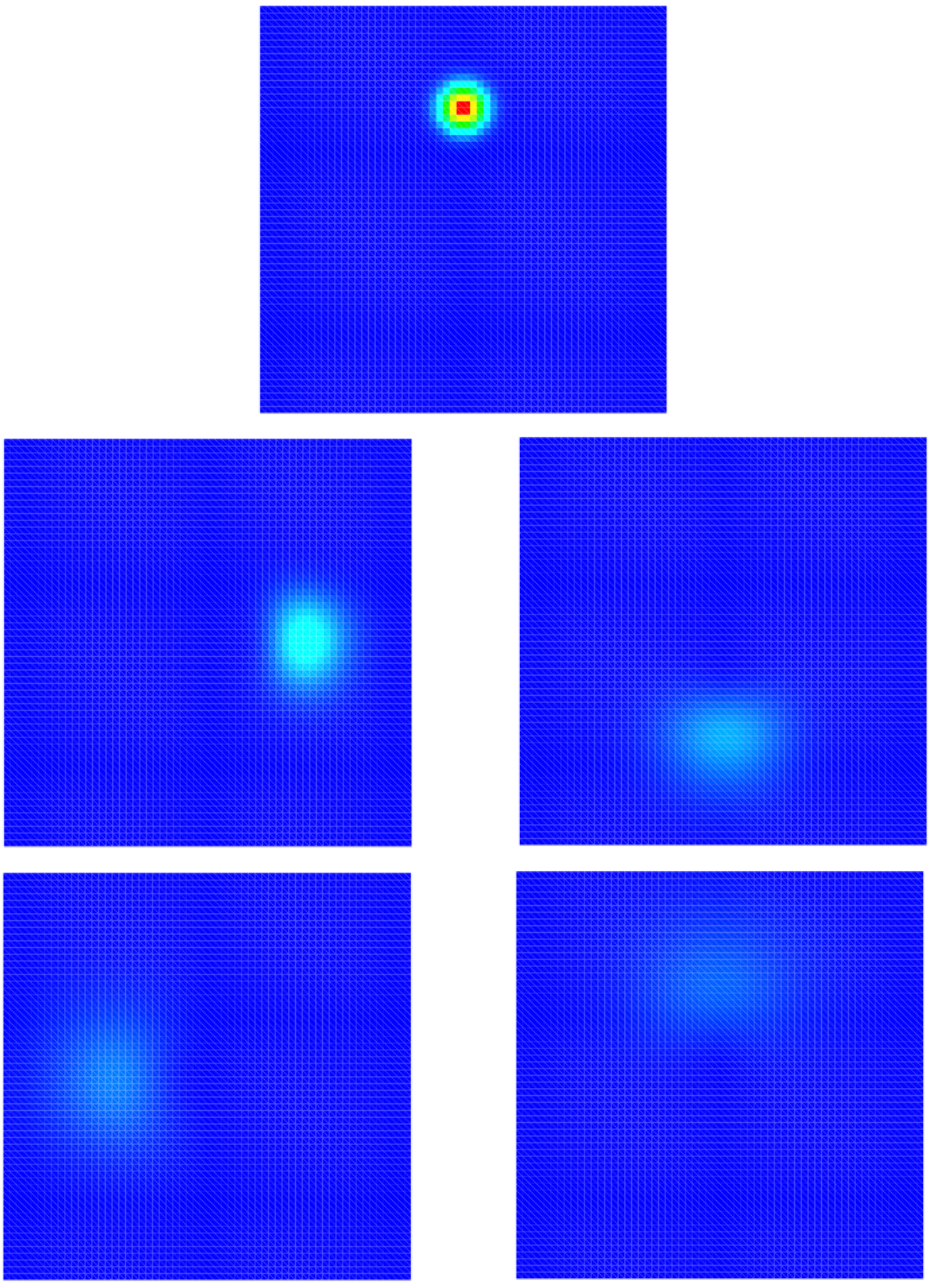


Fig. 2.14 – Solutions at time $t = 0$ (top), $t = 0.25$ (middle left), $t = 0.5$ (middle right), $t = 0.75$ (bottom left) and final time (bottom right) obtained with the Upwind scheme with a cartesian mesh of 60×60 cells.

2.8 Concluding remarks

In this chapter, we propose two new monotonic schemes for the diffusion equation, which are based on the same cell-centered discretization. This first step is called primal scheme, and the consistency of the primal fluxes relies on a correct evaluation of dual (node-centered) unknowns. The difference between the two schemes lies in the evaluation of these dual quantities. For the first one, which is called diamond type, the dual unknowns are evaluated, using a polynomial reconstruction involving values in neighbouring (primal) cells. For the second one, called DDFV type, the evaluation of the dual unknown is obtained by solving a diffusion problem discretized on the dual mesh. This second scheme is an improvement with respect to the nonlinear monotonic DDFV method of [21]. Indeed, the new nonlinear method we have proposed here makes it possible to deal with all types of boundary conditions (Dirichlet, Neumann) and is second-order convergent even for discontinuous diffusion coefficients. For both methods, we adapt the same non-linear process borrowed from [51, 52, 105, 111], we assess their monotonicity and accuracy on several test cases and compare the results with the classical (non-monotonic) DDFV scheme. Moreover, the DDFV type monotonic scheme takes advantage of very nice features of the DDFV scheme, such as second-order accuracy in H^1 norm, while providing non-negative solutions.

In the next chapter, we will extend theses schemes to arbitrary order, using the techniques developed in the 1D setting in Chapter 1 (see also [8]).

Chapter 3

Arbitrary order monotonic finite-volume schemes for 2D elliptic problems

This chapter has been submitted as an article in Journal of Computational Physics (see [11]). We add to the content of the article the formulation of a monotonic DDFV scheme of arbitrary order, which has not been coded.

Monotonicity is very important in most applications solving elliptic problems. Many schemes preserving positivity has been proposed but are at most second-order convergent. Besides, in general, high-order schemes do not preserve positivity. In the present chapter, we propose an arbitrary-order monotonic method for elliptic problems in 2D. We show how to adapt our method to the case of a discontinuous and/or tensor-valued diffusion coefficient, while keeping the order of convergence. We assess the new scheme on several test problems.

3.1 Introduction

This chapter describes a follow-up of two recently published works [8, 9]. In the former work, we designed a monotonic and arbitrary-order numerical method for an elliptic equation in 1D. In the latter one, we showed that the approach used in 1D extends to second-order accurate methods in 2D. Our goal in this paper is to propose the first arbitrary-order monotonic method for elliptic problems in 2D.

The model we consider is

$$\begin{cases} -\operatorname{div}(\kappa\nabla\bar{u}) + \lambda\bar{u} = f & \text{in} \quad \Omega, \\ \bar{u} = g_D & \text{on} \quad \Gamma_D, \\ \kappa\nabla\bar{u}\cdot\mathbf{n} = g_N & \text{on} \quad \Gamma_N, \end{cases} \tag{3.1}$$

where Ω is a bounded open domain of $\mathbb{R}^2$ with $\partial\Omega = \Gamma_D \cup \Gamma_N$ ($\Gamma_D \cap \Gamma_N = \emptyset$), and $\mathbf{n} \in \mathbb{R}^2$ is the outgoing unit normal vector. The data are such that $f \in L^2(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in L^2(\Gamma_N)$, $\lambda \in \mathbb{R}^+$ (if $\lambda = 0$, then $|\Gamma_D| > 0$), and $\kappa \in L^\infty(\Omega)$. The tensor-valued diffusion coefficient κ satisfies the uniform ellipticity condition:

$$\forall \mathbf{x} \in \Omega, \forall \xi \in \mathbb{R}^2, \quad \kappa_{\min}\|\xi\|^2 \leq \xi^t \kappa(\mathbf{x})\xi. \tag{3.2}$$

where $\kappa_{\min}$ is a strictly positive coefficient. Under the above conditions, one can prove (using Lax-Milgram Lemma in the spirit of [46], Chapter 6) that system (3.1) has a unique solution in $H^1(\Omega)$ which satisfies a positiveness principle, i.e. if $f \geq 0$ and $g \geq 0$, then $\bar{u} \geq 0$. One often refers to monotonicity in the literature for this principle.

For the applications we have in mind, such as inertial confinement fusion simulation, we need to be able to slve problem (3.1) on (almost) arbitrary meshes. The reason for this is twofold. First, the domain Ω can be very distorted. Second, problem (3.1) is coupled to the incompressible Euler system, which is descretized using a Lagrangian finite volume scheme (see [24,68,82]). We thus have no control on the quality of the mesh. Further, a fundamental property of the hydrodynamics scheme is to be conservative, in order to reproduce as precisely as possible singular solutions, such as shocks. Thus, the diffusion scheme applied to (3.1) should be conservative too, in order to preserve this property. As a consequence, monotonicity cannot be recovered by merely truncating negative values: such a strategy is incompatible with conservativity.

This is why a large amount of work has been devoted to the design of monotone schemes since the seminal works of [7,70]. Among other publications, let us cite recent works [21,22,88,95,101,105,106, 112] and citations therein about this topic. However, none of these methods is arbitrarily high-order accurate. The most advanced work in this direction is [106], which achieved third-order accuracy.

Some methods are particularly well-suited for achieving arbitrary high-order for elliptic problems. Let us cite for instance the finite-element method [27], the Virtual Element method [6], the Discontinuous Galerkin method [29], and the Hybrid High-Order method [33]. However, very few

(see [4,5,20,102] and references therein) can enforce the positiveness of the unknown without imposing severe constraints on the mesh, and none of them achieve a convergence order higher than two. Another reason for not using these methods in our context, is that their coupling with other models can be problematic since the degrees of freedom of the different discrete operators approximations do not match.

This work proposes the first arbitrary-order monotonic scheme for the elliptic equation (3.1). The diffusion coefficient can be tensor-valued and/or discontinuous. We show that we preserve the arbitrary high-order accuracy even with a discontinuous diffusion coefficient as long as discontinuities are known and coincide edges of the mesh. We recall the main steps of the proposed method (see also [9]):

1. Integration of the equation over each cell of the initial mesh that we will call *primal*.
2. Transformation of this surface integral into a sum of fluxes using the divergence theorem.
3. Approximation of the fluxes using a Gauss quadrature rule on each face of the cell.
4. Taylor expansion of the solution $\bar{u}$ in the neighborhood of each Gauss quadrature point of each face along *two* independent privileged directions in order to obtain an approximation of $\boldsymbol{\nabla}\bar{u}$ involving the values of $\bar{u}$ and its derivatives at certain suitably chosen points, in this case the center and vertices of the cell.
5. Using this Taylor expansion, estimation of $(\boldsymbol{\kappa}\boldsymbol{\nabla}\bar{u})\cdot\mathbf{n} = (\boldsymbol{\nabla}\bar{u})\cdot(\boldsymbol{\kappa}^t\mathbf{n})$.
6. Calculation of the values of $\bar{u}$ at vertices by a polynomial interpolation formula in the neighborhood of the Gauss quadrature points of each primal cell face.
7. Calculation of the values of derivatives of $\bar{u}$ at centers and vertices of the neighboring cells by differentiating this polynomial interpolation.
8. Transformation of the scheme into a monotonic nonlinear two point flux approximation.
9. Resolution of the nonlinear system by the Picard iteration method.

The paper is structured as follows. Definitions and notations are given in Section 3.2. The proposed arbitrarily high-order Finite-Volume method is described in Section 3.3. Then, we explain how the scheme is modified to enforce the monotonicity in Section 3.4. In Section 3.5, we prove some nice properties of the method. Finally the arbitrary high-order accuracy and the monotonicity of the method are assessed in Section 3.6 on classical benchmarks including test cases with anisotropic and discontinuous diffusion coefficients.

3.2 Definitions and notations

Given an arbitrary mesh the cells of which are numbered from 1 to n, consider a cell denoted i and its neighbor j (see Figure 3.1). The center of mass of i (resp. j) is denoted by $\mathbf{x}_i$ (resp. $\mathbf{x}_j$), their common face is ℓ and the vertices of ℓ are r and s. The position of the center of the face ℓ is $\mathbf{x}_\ell$, and the positions of its vertices are $\mathbf{x}_r$ and $\mathbf{x}_s$. We denote by $\mathbf{x}_g$ a Gauss quadrature point located on the face ℓ. The length of ℓ is $|\ell|$ and the volume of a cell i is V_i. The normal vector $\mathbf{n}_{i\ell}$ is the unit vector which is orthogonal to the edge ℓ and outgoing for the cell i. We define $h = \min_{\ell} |\ell|$.

Given $\mathbf{v} = (v_i)$ a vector in $\mathbb{R}^n$ we will denote respectively its Euclidian, L^2 and L^∞ norms by

$$\|\mathbf{v}\| = \left(\sum_{i=1}^{n} v_i^2\right)^{1/2}, \quad \|\mathbf{v}\|_2 = \left(\sum_{i=1}^{n} V_i v_i^2\right)^{1/2}, \quad \|\mathbf{v}\|_\infty = \max_{1\le i\le n} |v_i|,$$

and we use the notation $\mathbf{v} > \mathbf{0}$ (resp. $\mathbf{v} \ge \mathbf{0}$) if, for all i, $v_i > 0$ (resp. $v_i \ge 0$).

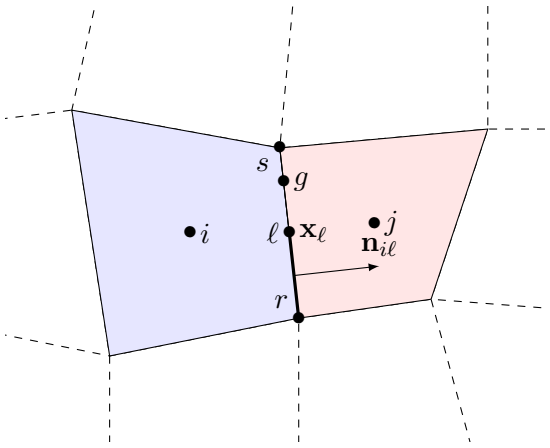


Fig. 3.1 – Example of a mesh with our notations

3.3 Finite volume formulation

To simplify the presentation we suppose that $\boldsymbol{\kappa}$ is isotropic : $\boldsymbol{\kappa} = \kappa \boldsymbol{I}$, with $\kappa > \kappa_{\min}$. It is worth noting that the full anisotropic case can be immediately dealt with by remarking that $(\boldsymbol{\kappa}\boldsymbol{\nabla}\bar{u})\cdot\mathbf{n} = (\boldsymbol{\nabla}\bar{u})\cdot(\boldsymbol{\kappa}^t\mathbf{n})$ and by replacing $\mathbf{n}$ by $\boldsymbol{\kappa}^t\mathbf{n}$ in what follows. Moreover we assume that the discontinuities of κ coincide with faces of the mesh and therefore that κ is a continuous function inside each cell.

The first step to design a finite volume scheme consists in integrating (3.1) on cell i

$$-\int_i \boldsymbol{\nabla}\cdot\kappa\boldsymbol{\nabla}\bar{u} + \int_i \lambda\bar{u} = \int_i f.$$

The properties of the continuous problem (3.1) impose that the normal component of the flux is continuous across the faces. Then, we can make use of the divergence formula to obtain

$$-\sum_{\ell\in i}\int_\ell \kappa\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n} + \int_i \lambda\bar{u} = \int_i f. \tag{3.3}$$

Using a k-th order accurate Gauss's quadrature formula for approximating the flux through the edge ℓ

$$\bar{\mathcal{F}}_\ell = \int_\ell \kappa\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n}$$

we have

$$-\sum_{\ell\in i}|\ell|\sum_{g\in\ell}\omega_g\kappa_g\left(\boldsymbol{\nabla}\bar{u}\right)_g\cdot\mathbf{n}_{i\ell} + \int_i\lambda\bar{u} = \int_i f + \mathcal{O}(h^k),$$

where ω_g and $\mathbf{x}_g$ are respectively the weights and the points of the quadrature. Thus we have to approximate

$$\kappa_g\left(\boldsymbol{\nabla}\bar{u}\right)_g\cdot\mathbf{n}_{i\ell}.$$

3.3.1 Approximation of the interior fluxes with the diamond method

Consider the case where the diffusion coefficient κ can be discontinuous on a face ℓ of the mesh and suppose that $\bar{u}\in W^{1,\infty}(\Omega)$. A Taylor expansion at order k in the neighborhood of $\mathbf{x}_g$ gives

$$\bar{u}(\mathbf{x}) = \bar{u}(\mathbf{x}_g)+(\mathbf{x}-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)(x-x_g)^q(y-y_g)^{(p-q)}+\mathcal{O}\left(\|\mathbf{x}-\mathbf{x}_g\|^{k+1}\right). \tag{3.4}$$

Denote by $\bar{u}_i$ the mean value of u in cell i

$$\bar{u}_i = \frac{1}{V_i}\int_i \bar{u}(\mathbf{x})dx.$$

In order to have mean values as degrees of freedom, we integrate (3.4) on the cell j and divide by its volume V_j

$$\frac{1}{V_j}\int_j \bar{u}(\mathbf{x})dx = \bar{u}(\mathbf{x}_g) + \frac{1}{V_j}\int_j (\mathbf{x}-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)dx + \frac{1}{V_j}\int_j \sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)(x-x_g)^q(y-y_g)^{(p-q)}dx + \mathcal{O}\left(h^{k+1}\right),$$

that is to say

$$\bar{u}_j = \bar{u}(\mathbf{x}_g)+(\mathbf{x}_j-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\sum_{p=2}^{k}\frac{1}{V_j p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\int_j (x-x_g)^q(y-y_g)^{(p-q)}dx+\mathcal{O}\left(h^{k+1}\right). \tag{3.5}$$

In a similar way, by integrating on the cell i, we obtain

$$\bar{u}_i = \bar{u}(\mathbf{x}_g)+(\mathbf{x}_i-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\sum_{p=2}^{k}\frac{1}{V_i p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\int_i (x-x_g)^q(y-y_g)^{(p-q)}dx+\mathcal{O}\left(h^{k+1}\right). \tag{3.6}$$

Equalities (3.5) and (3.6) give

$$(\mathbf{x}_j-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) = \bar{u}_j-\bar{u}(\mathbf{x}_g)\underbrace{-\sum_{p=2}^{k}\frac{1}{V_j p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\int_j (x-x_g)^q(y-y_g)^{(p-q)}dx + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{gj}}, \tag{3.7}$$

and

$$(\mathbf{x}_g-\mathbf{x}_i)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) = \bar{u}(\mathbf{x}_g)-\bar{u}_i+\underbrace{\sum_{p=2}^{k}\frac{1}{V_i p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\int_i (x-x_g)^q(y-y_g)^{(p-q)}dx + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{ig}}. \tag{3.8}$$

Using respectively $\mathbf{x}=\mathbf{x}_r$ and $\mathbf{x}=\mathbf{x}_s$ in the Taylor expansion (3.4), we obtain

$$\bar{u}(\mathbf{x}_r) = \bar{u}(\mathbf{x}_g)+(\mathbf{x}_r-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)(x_r-x_g)^q(y_r-y_g)^{(p-q)}+\mathcal{O}\left(h^{k+1}\right), \tag{3.9}$$

and

$$\bar{u}(\mathbf{x}_s) = \bar{u}(\mathbf{x}_g)+(\mathbf{x}_s-\mathbf{x}_g)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)(x_s-x_g)^q(y_s-y_g)^{(p-q)}+\mathcal{O}\left(h^{k+1}\right). \tag{3.10}$$

Subtracting the two last equalities gives

$$(\mathbf{x}_s-\mathbf{x}_r)\cdot\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) = \bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r) \underbrace{-\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)}-(x_r-x_g)^q(y_r-y_g)^{(p-q)}\right)+\mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{rs}}. \tag{3.11}$$

Thus, we have the system

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_j-\mathbf{x}_g)=\bar{u}_j-\bar{u}(\mathbf{x}_g)+\bar{r}_{gj},\\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_g-\mathbf{x}_i)=\bar{u}(\mathbf{x}_g)-\bar{u}_i+\bar{r}_{ig},\\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s-\mathbf{x}_r)=\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)+\bar{r}_{rs}. \end{cases} \tag{3.12}$$

We can decompose the normal vector $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_j-\mathbf{x}_g),(\mathbf{x}_s-\mathbf{x}_r))$

$$\mathbf{n}_{i\ell}=\alpha_{i\ell,jg}\frac{\mathbf{x}_j-\mathbf{x}_g}{\|\mathbf{x}_j-\mathbf{x}_g\|}+\beta_{i\ell,jg}\frac{\mathbf{x}_s-\mathbf{x}_r}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

with

$$\alpha_{i\ell,jg}=\frac{\|\mathbf{x}_j-\mathbf{x}_g\|}{(\mathbf{x}_j-\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}}\geq 0, \tag{3.13}$$

and

$$\beta_{i\ell,jg}=\frac{\|\mathbf{x}_s-\mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_j-\mathbf{x}_g)^{\perp}}{(\mathbf{x}_s-\mathbf{x}_r)\cdot(\mathbf{x}_j-\mathbf{x}_g)^{\perp}}. \tag{3.14}$$

The details of these computations are given in Appendix D.1.

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell j, denoted by $\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_j\cdot\mathbf{n}_{i\ell}$

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_j\cdot\mathbf{n}_{i\ell}=\alpha_{i\ell,jg}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_j-\mathbf{x}_g)}{\|\mathbf{x}_j-\mathbf{x}_g\|}+\beta_{i\ell,jg}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s-\mathbf{x}_r)}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

that is to say, using (3.12)

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_j\cdot\mathbf{n}_{i\ell}=\alpha_{i\ell,jg}\frac{\bar{u}_j-\bar{u}(\mathbf{x}_g)+\bar{r}_{gj}}{\|\mathbf{x}_j-\mathbf{x}_g\|}+\beta_{i\ell,jg}\frac{\bar{u}_s-\bar{u}_r+\bar{r}_{rs}}{\|\mathbf{x}_s-\mathbf{x}_r\|}, \tag{3.15}$$

Then, we can decompose $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_g-\mathbf{x}_i),(\mathbf{x}_s-\mathbf{x}_r))$

$$\mathbf{n}_{i\ell}=\alpha_{i\ell,ig}\frac{\mathbf{x}_g-\mathbf{x}_i}{\|\mathbf{x}_g-\mathbf{x}_i\|}+\beta_{i\ell,ig}\frac{\mathbf{x}_s-\mathbf{x}_r}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

with

$$\alpha_{i\ell,ig}=\frac{\|\mathbf{x}_g-\mathbf{x}_i\|}{(\mathbf{x}_g-\mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}\geq 0, \tag{3.16}$$

and

$$\beta_{i\ell,ig}=\frac{\|\mathbf{x}_s-\mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_g-\mathbf{x}_i)^{\perp}}{(\mathbf{x}_s-\mathbf{x}_r)\cdot(\mathbf{x}_g-\mathbf{x}_i)^{\perp}}. \tag{3.17}$$

The details of these computations are given in Appendix D.1.

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell i, denoted by $\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i\cdot\mathbf{n}_{i\ell}$

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_i\cdot\mathbf{n}_{i\ell}=\alpha_{i\ell,ig}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_g-\mathbf{x}_i)}{\|\mathbf{x}_g-\mathbf{x}_i\|}+\beta_{i\ell,ig}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s-\mathbf{x}_r)}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

that is to say, using (3.12)

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_i\cdot\mathbf{n}_{i\ell}=\alpha_{i\ell,ig}\frac{\bar{u}(\mathbf{x}_g)-\bar{u}_i+\bar{r}_{ig}}{\|\mathbf{x}_g-\mathbf{x}_i\|}+\beta_{i\ell,ig}\frac{\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)+\bar{r}_{rs}}{\|\mathbf{x}_s-\mathbf{x}_r\|},$$

If κ is continuous on a Gauss point $\mathbf{x}_g$ of a face ℓ we define

$$\kappa_{g,i} = \kappa_{g,j} = \kappa(\mathbf{x}_g)$$

while if it is not we define

$$\kappa_{g,i} = \lim_{\mathbf{x}\in i\to\mathbf{x}_g} \kappa(\mathbf{x}), \quad \kappa_{g,j} = \lim_{\mathbf{x}\in j\to\mathbf{x}_g} \kappa(\mathbf{x}).$$

Thanks to the continuity of the flux

$$\kappa_{g,i}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_i \cdot \mathbf{n}_{i\ell} = \kappa_{g,j}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell}, \tag{3.18}$$

we obtain

$$\begin{aligned}\bar{u}(\mathbf{x}_g) = \frac{1}{\frac{\kappa_{g,i}\alpha_{i\ell,ig}}{\|\mathbf{x}_g-\mathbf{x}_i\|} + \frac{\kappa_{g,j}\alpha_{i\ell,jg}}{\|\mathbf{x}_j-\mathbf{x}_g\|}} \Bigg(&\frac{\kappa_{g,j}\alpha_{i\ell,jg}}{\|\mathbf{x}_j-\mathbf{x}_g\|}(\bar{u}_j + \bar{r}_{gj}) + \frac{\kappa_{g,i}\alpha_{i\ell,ig}}{\|\mathbf{x}_g-\mathbf{x}_i\|}(\bar{u}_i - \bar{r}_{ig}) \\ &+\frac{\kappa_{g,j}\beta_{i\ell,jg}}{\|\mathbf{x}_s-\mathbf{x}_r\|}(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs}) - \frac{\kappa_{g,i}\beta_{i\ell,ig}}{\|\mathbf{x}_s-\mathbf{x}_r\|}(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs})\Bigg).\end{aligned} \tag{3.19}$$

Inserting (3.19) into (3.15) results in

$$\begin{aligned}\kappa_{g,j}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell} = &\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}}\right)(\bar{u}_j - \bar{u}_i + \bar{r}_{gj} + \bar{r}_{ig}) \\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j-\mathbf{x}_g\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs}) \\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_g-\mathbf{x}_i\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs}).\end{aligned} \tag{3.20}$$

Let us assume that we have at our disposal an approximation $\mathbf{u} = (u_i)_{1\le i\le n}$ of $\bar{\mathbf{u}} = (\bar{u}_i)_{1\le i\le n}$. From $\mathbf{u}$ we can find a high-order polynomial approximation $P_i(\mathbf{x})$ of $\bar{u}$ in each cell i while respecting the discontinuity lines of the diffusion coefficient κ (see Section 3.3.6). So, the numerical flux $\mathcal{F}_\ell(\mathbf{u})$ is defined by

$$\begin{aligned}\mathcal{F}_\ell(\mathbf{u}) = |\ell|\sum_{g\in\ell}\omega_g\Bigg[&\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}}\right)(u_j - u_i + r_{gj}(\mathbf{u}) + r_{ig}(\mathbf{u})) \\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j-\mathbf{x}_g\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(P_j(\mathbf{x}_s) - P_j(\mathbf{x}_r) + r_{rs,j}(\mathbf{u})) \\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_g-\mathbf{x}_i\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r) + r_{rs,i}(\mathbf{u}))\Bigg],\end{aligned}$$

with

$$\begin{cases} r_{ig}(\mathbf{u}) = \displaystyle\sum_{p=2}^{k}\frac{1}{V_i p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_i}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\int_i (x-x_g)^q(y-y_g)^{(p-q)}dx, \\ r_{gj}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k}\frac{1}{V_j p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_j}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\int_j (x-x_g)^q(y-y_g)^{(p-q)}dx, \\ r_{rs,i}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_i}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)} - (x_r-x_g)^q(y_r-y_g)^{(p-q)}\right), \\ r_{rs,j}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_j}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)} - (x_r-x_g)^q(y_r-y_g)^{(p-q)}\right), \end{cases}$$

where P_j is a polynomial local to the cell j. The choice of cell-based polynomials is consistent with the fact that the diffusion coefficient is continuous inside each cell.

Finally we obtain in a more compact form the following approximation of the flux through the face ℓ

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell(u_j - u_i) + r_\ell(\mathbf{u}), \tag{3.21}$$

with

$$\begin{cases} \gamma_\ell = |\ell| \sum_{g\in\ell} \omega_g \left(\dfrac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}} \right) \geq 0, \\ r_\ell(\mathbf{u}) = |\ell| \sum_{g\in\ell} \omega_g \left[\left(\dfrac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}} \right) (r_{gj}(\mathbf{u}) + r_{ig}(\mathbf{u})) \right. \\ \qquad + \left(\dfrac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j - \mathbf{x}_g\|}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)} \right) (P_j(\mathbf{x}_s) - P_j(\mathbf{x}_r) + r_{rs,j}(\mathbf{u})) \\ \qquad \left. + \left(\dfrac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_g - \mathbf{x}_i\|}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)} \right) (P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r) + r_{rs,i}(\mathbf{u})) \right]. \end{cases}$$

This decomposition will be used hereafter to enforce the positiveness of the scheme (see Section 3.4).

3.3.2 Approximation of the boundary fluxes with the diamond method

In this section, we use the boundary conditions to estimate the boundary fluxes.

3.3.2.1 Neumann boundary condition

Consider the problem (3.1) with $\Gamma_D = \emptyset$. Let ℓ be a boundary face of the mesh on Γ_N. Integrating the Neumann boundary condition on the face $\ell \subset \Gamma_N$, we have

$$\int_\ell \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} = \int_\ell g_N,$$

that is to say

$$\bar{\mathcal{F}}_\ell = |\ell| \sum_{g\in\ell} \omega_g g_N(\mathbf{x}_g) + \mathcal{O}(h^k),$$

we thus impose this equation on the numerical flux

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| \sum_{g\in\ell} \omega_g g_N(\mathbf{x}_g).$$

3.3.2.2 Dirichlet boundary condition

Consider the problem (3.1) with $\Gamma_N = \emptyset$. Let ℓ be a boundary face of the cell i on Γ_D. Taking into account the Dirichlet boundary condition $\bar{u}(\mathbf{x}_g) = g_D(\mathbf{x}_g)$ in (3.8) for $g \in \ell \subset \Gamma_D$, equalities (3.8) and (3.11) give the system

$$\begin{cases} \boldsymbol{\nabla} \bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_g - \mathbf{x}_i) = g_D(\mathbf{x}_g) - \bar{u}_i + \bar{r}_{ig}, \\ \boldsymbol{\nabla} \bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_s - \mathbf{x}_r) = \bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs}. \end{cases} \tag{3.22}$$

Then, we can decompose $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_g - \mathbf{x}_i),(\mathbf{x}_s - \mathbf{x}_r))$

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,ig} \frac{\mathbf{x}_g - \mathbf{x}_i}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig} \frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

with

$$\alpha_{i\ell,ig} = \frac{\|\mathbf{x}_g - \mathbf{x}_i\|}{(\mathbf{x}_g - \mathbf{x}_i)\cdot \mathbf{n}_{i\ell}} \geq 0,$$

and

$$\beta_{i\ell,ig} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\| \mathbf{n}_{i\ell}\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp}{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp}.$$

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell i, denoted by $\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell}$

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell} = \alpha_{i\ell,ig}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_g - \mathbf{x}_i)}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig}\frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say, using (3.22)

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell} = \alpha_{i\ell,ig}\frac{g_D(\mathbf{x}_g) - \bar{u}_i + \bar{r}_{ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig}\frac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + \bar{r}_{rs}}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \tag{3.23}$$

Let $\mathbf{u} = (u_i)_{1\leq i\leq n}$ be the numerical solution on the mesh. By mimicking the expression of the exact flux (3.23), the numerical flux is defined by

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell|\sum_{g\in\ell}\omega_g\kappa_g\left(\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|}\left(g_D(\mathbf{x_g}) - u_i + r_{ig}(\mathbf{u})\right) + \frac{\beta_{i\ell,ig}}{\|\mathbf{x}_s - \mathbf{x}_r\|}\left(P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r) + r_{rs}(\mathbf{u})\right)\right),$$

with

$$\begin{cases} r_{ig}(\mathbf{u}) = \displaystyle\sum_{p=2}^{k}\frac{1}{V_i p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_i}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\int_i (x-x_g)^q (y-y_g)^{(p-q)}dx, \\ r_{rs}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p P_i}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)} - (x_r-x_g)^q(y_r-y_g)^{(p-q)}\right). \end{cases}$$

In a more compact form, we have

$$\mathcal{F}_\ell(\mathbf{u}) = -\gamma_\ell u_i + \sum_{g\in\ell}\left(\frac{\omega_g\kappa_g\alpha_{i\ell,ig}|\ell|}{\|\mathbf{x}_g - \mathbf{x}_i\|}g_D(\mathbf{x_g})\right) + r_\ell(\mathbf{u}),$$

with

$$\begin{cases} \gamma_\ell = \displaystyle\sum_{g\in\ell}\left(\frac{\omega_g\kappa_g\alpha_{i\ell,ig}|\ell|}{\|\mathbf{x}_g - \mathbf{x}_i\|}\right) \geq 0, \\ r_\ell(\mathbf{u}) = |\ell|\displaystyle\sum_{g\in\ell}\omega_g\kappa_g\left(\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|}r_{ig}(\mathbf{u}) + \frac{\beta_{i\ell,ig}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(P_i(\mathbf{x}_s) - P_i(\mathbf{x}_r) + r_{rs}(\mathbf{u}))\right). \end{cases}$$

3.3.3 Approximation of the interior fluxes with the DDFV method

3.3.3.1 Primal flux

Consider the case where the diffusion coefficient κ can be discontinuous on a face ℓ of the mesh and suppose that $\bar{u} \in W^{1,\infty}(\Omega)$. A Taylor expansion at order k in the neighborhood of $\mathbf{x}_g$ gives (3.5) and (3.6), that is to say

$$(\mathbf{x}_j - \mathbf{x}_g)\cdot\nabla\bar{u}(\mathbf{x}_g) = \bar{u}_j - \bar{u}(\mathbf{x}_g) \underbrace{- \sum_{p=2}^{k}\frac{1}{V_j p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\int_j (x-x_g)^q(y-y_g)^{(p-q)}dx + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{gj}},$$

and

$$(\mathbf{x}_g-\mathbf{x}_i)\cdot\nabla\bar{u}(\mathbf{x}_g)=\bar{u}(\mathbf{x}_g)-\bar{u}_i+\underbrace{\sum_{p=2}^{k}\frac{1}{V_i p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\int_i(x-x_g)^q(y-y_g)^{(p-q)}dx+\mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{ig}}.$$

Using respectively $\mathbf{x}=\mathbf{x}_r$ and $\mathbf{x}=\mathbf{x}_s$ in the Taylor expansion (3.4), we obtain (3.9) and (3.10) as previously and the difference between these two equalities gives (3.11), that is

$$\begin{aligned}&(\mathbf{x}_s-\mathbf{x}_r)\cdot\nabla\bar{u}(\mathbf{x}_g)=\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)\\&-\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)}-(x_r-x_g)^q(y_r-y_g)^{(p-q)}\right)+\mathcal{O}\left(h^{k+1}\right).\end{aligned}$$

Noting s_i as the intersection between the primal mesh i and the dual mesh s, we have

$$u_s=\frac{1}{V_s}\int_s u(\mathbf{x})dx=\frac{1}{V_s}\sum_{s_i\in s}\int_{s_i}u(\mathbf{x})dx.$$

A Taylor expansion with respect to $\mathbf{x}_s$ at a point $\mathbf{x}\in s_i$ gives us

$$\bar{u}(\mathbf{x})=\bar{u}(\mathbf{x}_s)+(\mathbf{x}-\mathbf{x}_s)\cdot\nabla\bar{u}(\mathbf{x}_{s_i})+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_s)(x-x_s)^q(y-y_s)^{(p-q)}+\mathcal{O}\left(h^{k+1}\right),$$

then, by integrating it over s_i, this gives

$$\int_{s_i}\bar{u}(\mathbf{x})=\bar{u}(\mathbf{x}_s)V_{s_i}+\int_{s_i}(\mathbf{x}-\mathbf{x}_s)\cdot\nabla\bar{u}(\mathbf{x}_{s_i})+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_s)\int_{s_i}(x-x_s)^q(y-y_s)^{(p-q)}dx+\mathcal{O}\left(h^{k+1}\right).$$

Thus, we have

$$\bar{u}_s=\bar{u}(\mathbf{x}_s)+\frac{1}{V_s}\sum_{s_i\in s}\int_{s_i}(\mathbf{x}-\mathbf{x}_s)\cdot\nabla\bar{u}(\mathbf{x}_{s_i})+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_s)\int_{s_i}(x-x_s)^q(y-y_s)^{(p-q)}dx+\mathcal{O}\left(h^{k+1}\right).\tag{3.24}$$

Using the same principle for the node r and replacing $\bar{u}(\mathbf{x}_s)$ and $\bar{u}(\mathbf{x}_r)$ by their expressions in (3.11), we obtain

$$\begin{aligned}(\mathbf{x}_s-\mathbf{x}_r)\cdot\nabla\bar{u}(\mathbf{x}_g)=\bar{u}_s-\bar{u}_r&\underbrace{-\frac{1}{V_s}\sum_{s_i\in s}\int_{s_i}(\mathbf{x}-\mathbf{x}_s)\cdot\nabla\bar{u}(\mathbf{x}_{s_i})}\\&\underbrace{-\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_s)\int_{s_i}(x-x_s)^q(y-y_s)^{(p-q)}dx+\frac{1}{V_r}\sum_{r_i\in r}\int_{r_i}(\mathbf{x}-\mathbf{x}_r)\cdot\nabla\bar{u}(\mathbf{x}_{r_i})}\\&\underbrace{+\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_r)\int_{r_i}(x-x_r)^q(y-y_r)^{(p-q)}dx}\\&\underbrace{-\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)\left((x_s-x_g)^q(y_s-y_g)^{(p-q)}-(x_r-x_g)^q(y_r-y_g)^{(p-q)}\right)+\mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{rs}}.\end{aligned}\tag{3.25}$$

Thus, we have the system

$$\begin{cases}\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_j-\mathbf{x}_g)=\bar{u}_j-\bar{u}(\mathbf{x}_g)+\bar{r}_{gj},\\\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_g-\mathbf{x}_i)=\bar{u}(\mathbf{x}_g)-\bar{u}_i+\bar{r}_{ig},\\\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s-\mathbf{x}_r)=\bar{u}_s-\bar{u}_r+\bar{r}_{rs}.\end{cases}\tag{3.26}$$

We can decompose the normal $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_j - \mathbf{x}_g),(\mathbf{x}_s - \mathbf{x}_r))$

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,j}\frac{\mathbf{x}_j - \mathbf{x}_g}{\|\mathbf{x}_j - \mathbf{x}_g\|} + \beta_{i\ell,j}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

with $\alpha_{i\ell,j} \geq 0$.

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell j, denoted by $\nabla\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell}$

$$\nabla\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,j}\frac{\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_j - \mathbf{x}_g)}{\|\mathbf{x}_j - \mathbf{x}_g\|} + \beta_{i\ell,j}\frac{\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say, using (3.26)

$$\nabla\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,j}\frac{\bar{u}_j - \bar{u}(\mathbf{x}_g) + \bar{r}_{gj}}{\|\mathbf{x}_j - \mathbf{x}_g\|} + \beta_{i\ell,j}\frac{\bar{u}_s - \bar{u}_r + \bar{r}_{rs}}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \tag{3.27}$$

Then, we can decompose $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_g - \mathbf{x}_i),(\mathbf{x}_s - \mathbf{x}_r))$

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,i}\frac{\mathbf{x}_g - \mathbf{x}_i}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,i}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

with $\alpha_{i\ell,i} \geq 0$.

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell i, denoted by $\nabla\bar{u}(\mathbf{x}_g)_i \cdot \mathbf{n}_{i\ell}$

$$\nabla\bar{u}(\mathbf{x}_g)_i \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,i}\frac{\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_g - \mathbf{x}_i)}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,i}\frac{\nabla\bar{u}(\mathbf{x}_g)\cdot(\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say, using (3.26)

$$\nabla\bar{u}(\mathbf{x}_g)_i \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,i}\frac{\bar{u}(\mathbf{x}_g) - \bar{u}_i + \bar{r}_{ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,i}\frac{\bar{u}_s - \bar{u}_r + \bar{r}_{rs}}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

The continuity of the flux imposes

$$\kappa_{g,i}\nabla\bar{u}(\mathbf{x}_g)_i \cdot \mathbf{n}_{i\ell} = \kappa_{g,j}\nabla\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell},$$

which leads to

$$\bar{u}(\mathbf{x}_g) = \frac{1}{\frac{\kappa_{g,i}\alpha_{i\ell,i}}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \frac{\kappa_{g,j}\alpha_{i\ell,j}}{\|\mathbf{x}_j - \mathbf{x}_g\|}}\left(\frac{\kappa_{g,j}\alpha_{i\ell,j}}{\|\mathbf{x}_j - \mathbf{x}_g\|}(\bar{u}_j + \bar{r}_{gj}) + \frac{\kappa_{g,i}\alpha_{i\ell,i}}{\|\mathbf{x}_g - \mathbf{x}_i\|}(\bar{u}_i - \bar{r}_{ig}) + \frac{\kappa_{g,j}\beta_{i\ell,j} - \kappa_{g,i}\beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(\bar{u}_s - \bar{u}_r + \bar{r}_{rs})\right). \tag{3.28}$$

Inserting (3.28) into (3.27), results in

$$\begin{aligned}\kappa_{g,j}\nabla\bar{u}(\mathbf{x}_g)_j \cdot \mathbf{n}_{i\ell} = &\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}}\right)(\bar{u}_j - \bar{u}_i + \bar{r}_{gj} + \bar{r}_{ig}) \\ &+ \left(\frac{\kappa_{g,i}\kappa_{g,j}\left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_g\| - \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_g - \mathbf{x}_i\|\right)}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g - \mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}\right)}\right)(\bar{u}_s - \bar{u}_r + \bar{r}_{rs}).\end{aligned} \tag{3.29}$$

Let $\mathbf{u} = \begin{pmatrix} \mathbf{u}^{\text{primal}} \\ \mathbf{u}^{\text{dual}} \end{pmatrix}$ be the numerical solution, where $\mathbf{u}^{\text{dual}} = (u_r)_{1\leq r\leq m}$ is the numerical solution on the dual mesh and $\mathbf{u}^{\text{primal}} = (u_i)_{1\leq i\leq n}$ is the numerical solution on the primal mesh. We approximate

the derivatives of u at the Gauss points with a derived polynomial at this Gauss point. By mimicking the expression of the exact flux (3.29), the numerical flux is defined by

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| \sum_{g\in\ell} \omega_g \left[\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}} \right) (u_j - u_i + r_{gj}(\mathbf{u}) + r_{ig}(\mathbf{u})) \right.$$
$$\left. + \left(\frac{\kappa_{g,i}\kappa_{g,j}\left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j-\mathbf{x}_g\| - \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_g-\mathbf{x}_i\|\right)}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}\right)} \right) (u_s - u_r + r_{rs}(\mathbf{u})) \right],$$

with

$$\begin{cases}
r_{ig}(\mathbf{u}) = \displaystyle\sum_{p=2}^{k} \frac{1}{V_i p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \int_i (x-x_g)^q (y-y_g)^{(p-q)} dx, \\
r_{gj}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k} \frac{1}{V_j p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \int_j (x-x_g)^q (y-y_g)^{(p-q)} dx, \\
r_{rs}(\mathbf{u}) = -\dfrac{1}{V_s} \displaystyle\sum_{s_i\in s} \int_{s_i} (\mathbf{x}-\mathbf{x}_s)\cdot\nabla P(\mathbf{x}_{s_i}) + \dfrac{1}{V_r} \sum_{r_i\in r} \int_{r_i} (\mathbf{x}-\mathbf{x}_r)\cdot\nabla P(\mathbf{x}_{r_i}) \\
\qquad - \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x-x_s)^q (y-y_s)^{(p-q)} dx \\
\qquad + \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_r) \int_{r_i} (x-x_r)^q (y-y_r)^{(p-q)} dx \\
\qquad - \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \left((x_s-x_g)^q (y_s-y_g)^{(p-q)} - (x_r-x_g)^q (y_r-y_g)^{(p-q)} \right).
\end{cases}$$

In other words

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell (u_j - u_i) + r_\ell(\mathbf{u}),$$

with

$$\begin{cases}
\gamma_\ell = |\ell| \displaystyle\sum_{g\in\ell} \omega_g \left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}} \right) \geq 0, \\
r_\ell(\mathbf{u}) = |\ell| \displaystyle\sum_{g\in\ell} \omega_g \left[\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}} \right) (r_{gj}(\mathbf{u}) + r_{ig}(\mathbf{u})) \right. \\
\qquad \left. + \left(\dfrac{\kappa_{g,i}\kappa_{g,j}\left(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j-\mathbf{x}_g\| - \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_g-\mathbf{x}_i\|\right)}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,i} + \|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,j}\right)} \right) (u_s - u_r + r_{rs}(\mathbf{u})) \right]
\end{cases}$$

3.3.3.2 Computation of u_ℓ on the primal mesh

Let us consider a face ℓ inside the domain, then u_ℓ is obtained by the same process as the one to obtain u_g in (3.28)

$$\bar{u}(\mathbf{x}_\ell) = \frac{1}{\frac{\kappa_{\ell,i}\alpha_{i\ell,i}}{\|\mathbf{x}_\ell-\mathbf{x}_i\|} + \frac{\kappa_{\ell,j}\alpha_{i\ell,j}}{\|\mathbf{x}_j-\mathbf{x}_\ell\|}} \left(\frac{\kappa_{\ell,j}\alpha_{i\ell,j}}{\|\mathbf{x}_j-\mathbf{x}_\ell\|}(\bar{u}_j + \bar{r}_{lj}) + \frac{\kappa_{\ell,i}\alpha_{i\ell,i}}{\|\mathbf{x}_\ell-\mathbf{x}_i\|}(\bar{u}_i - \bar{r}_{i\ell}) + \frac{\kappa_{\ell,j}\beta_{i\ell,j} - \kappa_{\ell,i}\beta_{i\ell,i}}{\|\mathbf{x}_s-\mathbf{x}_r\|}(\bar{u}_s - \bar{u}_r + \bar{r}_{rs}) \right). \tag{3.30}$$

Let us consider a face ℓ as a Dirichlet boundary face, then we have

$$\bar{u}(\mathbf{x}_\ell) = g(\mathbf{x}_\ell).$$

Let us consider ℓ a Neumann boundary face, belonging to a cell i. We have

$$\kappa_\ell \nabla\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell} = g(\mathbf{x}_\ell), \tag{3.31}$$

and

$$\kappa_\ell \nabla \bar{u}(\mathbf{x}_\ell) \cdot \mathbf{n}_{i\ell} = \kappa_\ell \left[\alpha_{i\ell} (\nabla \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_\ell - \mathbf{x}_i)) + \beta_{i\ell} (\nabla \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_s - \mathbf{x}_r)) \right],$$

that is to say

$$\kappa_\ell \nabla \bar{u}(\mathbf{x}_\ell) \cdot \mathbf{n}_{i\ell} = \kappa_\ell \left[\alpha_{i\ell} \left(\bar{u}(\mathbf{x}_\ell) - \bar{u}_i + \bar{r}_{i\ell} \right) + \beta_{i\ell} \left(\bar{u}_s - \bar{u}_r + \bar{r}_{rs} \right) \right]. \tag{3.32}$$

Equations (3.31) and (3.32) give

$$\bar{u}(\mathbf{x}_\ell) = \frac{g(\mathbf{x}_\ell)}{\kappa_\ell \alpha_{i\ell}} + \bar{u}_i - \bar{r}_{i\ell} - \frac{\beta_{i\ell}}{\alpha il} \left(\bar{u}_s - \bar{u}_r + \bar{r}_{rs} \right). \tag{3.33}$$

3.3.3.3 Dual flux

The second way to calculate the vertex values u_r is to consider them as additional unknowns that are solutions to problem (3.1) integrated on each cell of the dual mesh, thus following [58]. We have

$$-\int_r \boldsymbol{\nabla} \cdot \kappa \boldsymbol{\nabla} \bar{u} + \int_r \lambda \bar{u} = \int_r f.$$

that is, thanks to the divergence theorem

$$-\sum_{\tilde{\ell} \in r} \int_{\tilde{\ell}} \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} + \int_r \lambda \bar{u} = \int_r f,$$

With a k-order approximation, we have

$$-\sum_{\tilde{\ell} \in r} |\tilde{\ell}| \sum_{\tilde{g} \in \tilde{\ell}} \omega_{\tilde{g}} \kappa_{\tilde{g}} \left(\boldsymbol{\nabla} \bar{u} \right)_{\tilde{g}} \cdot \mathbf{n}_{i\tilde{\ell}} + \int_i \lambda \bar{u} = \int_i f + \mathcal{O}(h^k),$$

with ω_g the weight of the Gauss quadrature.

Thus we need to approximate

$$\kappa_{\tilde{g}} \left(\boldsymbol{\nabla} \bar{u} \right)_{\tilde{g}} \cdot \mathbf{n}_{i\tilde{\ell}}.$$

Let us consider a Gauss point $\tilde{g}$ on a dual face $\tilde{\ell}$ located in a primal cell i. Using the Taylor expansion (3.6), we have

$$\bar{u}_i = \bar{u}(\mathbf{x}_{\tilde{g}}) + (\mathbf{x}_i - \mathbf{x}_{\tilde{g}}) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) + \sum_{p=2}^{k} \frac{1}{V_i p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) \int_i (x - x_{\tilde{g}})^q (y - y_{\tilde{g}})^{(p-q)} dx + \mathcal{O}\left(h^{k+1}\right), \tag{3.34}$$

and the ponctual Taylor expansion (3.4) applied to the point $\mathbf{x}_\ell$ gives

$$\bar{u}(\mathbf{x}_\ell) = \bar{u}(\mathbf{x}_{\tilde{g}}) + (\mathbf{x}_\ell - \mathbf{x}_{\tilde{g}}) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) + \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) (x_\ell - x_{\tilde{g}})^q (y_\ell - y_{\tilde{g}})^{(p-q)} + \mathcal{O}\left(h^{k+1}\right).$$

The difference between the two previous equations gives us

$$\begin{aligned} &(\mathbf{x}_\ell - \mathbf{x}_i) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) = \bar{u}(\mathbf{x}_\ell) - \bar{u}_i \\ &\underbrace{-\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) \left((x_\ell - x_{\tilde{g}})^q (y_\ell - y_{\tilde{g}})^{(p-q)} - \frac{1}{V_i} \int_i (x - x_{\tilde{g}})^q (y - y_{\tilde{g}})^{(p-q)} dx \right) + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{i\ell}}. \end{aligned}$$

Then, we have (3.9)

$$\bar{u}(\mathbf{x}_r) = \bar{u}(\mathbf{x}_{\tilde{g}}) + (\mathbf{x}_r - \mathbf{x}_{\tilde{g}}) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) + \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) (x_r - x_{\tilde{g}})^q (y_r - y_{\tilde{g}})^{(p-q)} + \mathcal{O}\left(h^{k+1}\right),$$

and (3.10)

$$\bar{u}(\mathbf{x}_s) = \bar{u}(\mathbf{x}_{\tilde{g}}) + (\mathbf{x}_s - \mathbf{x}_{\tilde{g}}) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) + \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}})(x_s - x_{\tilde{g}})^q (y_s - y_{\tilde{g}})^{(p-q)} + \mathcal{O}\left(h^{k+1}\right).$$

Using the relation (3.24) applied to the nodes r and s, we have

$$(\mathbf{x}_{\tilde{g}} - \mathbf{x}_r) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) = \bar{u}(\mathbf{x}_{\tilde{g}}) - \bar{u}_r + \underbrace{\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}})(x_r - x_{\tilde{g}})^q (y_r - y_{\tilde{g}})^{(p-q)}}$$

$$\underbrace{+\frac{1}{V_r} \sum_{r_i \in r} \int_{r_i} (\mathbf{x} - \mathbf{x}_r) \cdot \nabla \bar{u}(\mathbf{x}_{r_i}) + \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x - x_s)^q (y - y_s)^{(p-q)} dx + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{r\tilde{g}}},$$

and

$$(\mathbf{x}_s - \mathbf{x}_{\tilde{g}}) \cdot \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) = \bar{u}_s - \bar{u}(\mathbf{x}_{\tilde{g}}) - \underbrace{\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}})(x_s - x_{\tilde{g}})^q (y_s - y_{\tilde{g}})^{(p-q)}}$$

$$\underbrace{-\frac{1}{V_s} \sum_{s_i \in s} \int_{s_i} (\mathbf{x} - \mathbf{x}_s) \cdot \nabla \bar{u}(\mathbf{x}_{s_i}) - \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x - x_s)^q (y - y_s)^{(p-q)} dx + \mathcal{O}\left(h^{k+1}\right)}_{\bar{r}_{\tilde{g}s}}.$$

Thus, we have the system

$$\begin{cases} \nabla \bar{u}(\mathbf{x}_\ell) \cdot (\mathbf{x}_\ell - \mathbf{x}_i) = \bar{u}(\mathbf{x}_\ell) - \bar{u}_i + \bar{r}_{i\ell}, \\ \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_{\tilde{g}} - \mathbf{x}_r) = \bar{u}(\mathbf{x}_{\tilde{g}}) - \bar{u}_r + \bar{r}_{r\tilde{g}}, \\ \nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_s - \mathbf{x}_{\tilde{g}}) = \bar{u}_s - \bar{u}(\mathbf{x}_{\tilde{g}}) + \bar{r}_{\tilde{g}s}. \end{cases}$$

We can decompose the normal $\mathbf{n}_{r\tilde{\ell}}$ in the basis $((\mathbf{x}_\ell - \mathbf{x}_i),(\mathbf{x}_{\tilde{g}} - \mathbf{x}_r))$

$$\mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},r} \frac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},r} \frac{\mathbf{x}_{\tilde{g}} - \mathbf{x}_r}{\|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|},$$

with $\beta_{r\tilde{\ell},r} \geq 0$.

Thus we have

$$\nabla \bar{u}(\mathbf{x}_{\tilde{g}})_r \cdot \mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},r} \frac{\nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_\ell - \mathbf{x}_i)}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},r} \frac{\nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_{\tilde{g}} - \mathbf{x}_r)}{\|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|},$$

that is to say

$$\nabla \bar{u}(\mathbf{x}_{\tilde{g}})_r \cdot \mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},r} \frac{\bar{u}(\mathbf{x}_\ell) - \bar{u}_i + \bar{r}_{i\ell}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},r} \frac{\bar{u}_{\tilde{g}} - \bar{u}_r + \bar{r}_{r\tilde{g}}}{\|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|}. \tag{3.35}$$

Then, we can decompose the normal in the basis $((\mathbf{x}_\ell - \mathbf{x}_i),(\mathbf{x}_s - \mathbf{x}_{\tilde{g}}))$

$$\mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},s} \frac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},s} \frac{\mathbf{x}_s - \mathbf{x}_{\tilde{g}}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|},$$

with $\beta_{r\tilde{\ell},s} \geq 0$.

Thus we have

$$\nabla \bar{u}(\mathbf{x}_{\tilde{g}})_s \cdot \mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},s} \frac{\nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_\ell - \mathbf{x}_i)}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},s} \frac{\nabla \bar{u}(\mathbf{x}_{\tilde{g}}) \cdot (\mathbf{x}_s - \mathbf{x}_{\tilde{g}})}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|},$$

that is to say

$$\nabla \bar{u}(\mathbf{x}_{\tilde{g}})_s \cdot \mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell},s} \frac{\bar{u}(\mathbf{x}_\ell) - \bar{u}_i + \bar{r}_{i\ell}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell},s} \frac{\bar{u}_s - \bar{u}_{\tilde{g}} + \bar{r}_{\tilde{g}s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|},$$

The continuity of the flux imposes

$$\nabla \bar{u}(\mathbf{x}_{\tilde{g}})_r \cdot \mathbf{n}_{r\tilde{\ell}} = \nabla \bar{u}(\mathbf{x}_{\tilde{g}})_s \cdot \mathbf{n}_{r\tilde{\ell}},$$

that is to say

$$\bar{u}(\mathbf{x}_{\tilde{g}}) = \frac{1}{\frac{\beta_{r\tilde{\ell},r}}{\|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|} + \frac{\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|}} \left(\frac{\alpha_{r\tilde{\ell},s} - \alpha_{r\tilde{\ell},r}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} (\bar{u}(\mathbf{x}_\ell) - \bar{u}_i - \bar{r}_{i\ell}) + \frac{\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|} (\bar{u}_s + \bar{r}_{\tilde{g}s}) + \frac{\beta_{r\tilde{\ell},r}}{\|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|} (\bar{u}_r - \bar{r}_{r\tilde{g}}) \right), \tag{3.36}$$

with $\bar{u}(\mathbf{x}_\ell)$ defined by (3.30) with $\bar{u}_r$, $\bar{u}_s$ taken at the previous iteration of the fixed point algorithm.

Inserting (3.36) into (3.35), we obtain

$$\begin{aligned} \nabla \bar{u}(\mathbf{x}_{\tilde{g}})_r \cdot \mathbf{n}_{r\tilde{\ell}} = {} & \frac{\beta_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s}} (\bar{u}_s - \bar{u}_r + \bar{r}_{r\tilde{g}} + \bar{r}_{\tilde{g}s}) \\ & + \frac{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\alpha_{r\tilde{\ell},s}\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\alpha_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|(\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s})} (\bar{u}(\mathbf{x}_\ell) - \bar{u}_i + \bar{r}_{i\ell}). \end{aligned} \tag{3.37}$$

We approximate the derivatives of u at the Gauss points with a derived polynomial at this Gauss point. By mimicking the expression of the exact flux (3.37), the numerical flux is defined by

$$\begin{aligned} \mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = |\tilde{\ell}| \sum_{\tilde{g} \in \tilde{\ell}} \omega_{\tilde{g}} \kappa_{\tilde{g}} \Bigg[& \frac{\beta_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s}} (u_s - u_r + r_{r\tilde{g}}(\mathbf{u}) + r_{\tilde{g}s}(\mathbf{u})) \\ & + \frac{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\alpha_{r\tilde{\ell},s}\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\alpha_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|(\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s})} (u(\mathbf{x}_\ell) - u_i + r_{i\ell}(\mathbf{u})) \Bigg], \end{aligned}$$

with

$$\begin{cases} r_{i\ell}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) \left(\frac{1}{V_\ell} \int_\ell (x - x_{\tilde{g}})^q (y - y_{\tilde{g}})^{(p-q)} dx \right. \\ \qquad \left. - \dfrac{1}{V_i} \displaystyle\int_i (x - x_{\tilde{g}})^q (y - y_{\tilde{g}})^{(p-q)} dx \right), \\ r_{r\tilde{g}}(\mathbf{u}) = \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) (x_r - x_{\tilde{g}})^q (y_r - y_{\tilde{g}})^{(p-q)} + \frac{1}{V_r} \displaystyle\sum_{r_i \in r} \int_{r_i} (\mathbf{x} - \mathbf{x}_r) \cdot \nabla P(\mathbf{x}_{r_i}) \\ \qquad + \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x - x_s)^q (y - y_s)^{(p-q)} dx, \\ r_{\tilde{g}s}(\mathbf{u}) = -\displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_{\tilde{g}}) (x_s - x_{\tilde{g}})^q (y_s - y_{\tilde{g}})^{(p-q)} - \frac{1}{V_s} \displaystyle\sum_{s_i \in s} \int_{s_i} (\mathbf{x} - \mathbf{x}_s) \cdot \nabla P(\mathbf{x}_{s_i}) \\ \qquad - \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x - x_s)^q (y - y_s)^{(p-q)} dx \end{cases}$$

In other words

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = \gamma_{\tilde{\ell}}(u_s - u_r) + r_{\tilde{\ell}}(\mathbf{u}),$$

with

$$\begin{cases} \gamma_{\tilde{\ell}} = |\tilde{\ell}| \sum_{\tilde{g}\in\tilde{\ell}} \omega_{\tilde{g}}\kappa_{\tilde{g}} \left(\dfrac{\beta_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s}} \right) \geq 0, \\ r_{\tilde{\ell}}(\mathbf{u}) = |\tilde{\ell}| \sum_{\tilde{g}\in\tilde{\ell}} \omega_{\tilde{g}}\kappa_{\tilde{g}} \left[\dfrac{\beta_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s}} (r_{r\tilde{g}}(\mathbf{u}) + r_{\tilde{g}s}(\mathbf{u})) \right. \\ \qquad \left. + \dfrac{\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\alpha_{r\tilde{\ell},s}\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\alpha_{r\tilde{\ell},r}\beta_{r\tilde{\ell},s}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|(\|\mathbf{x}_s - \mathbf{x}_{\tilde{g}}\|\beta_{r\tilde{\ell},r} + \|\mathbf{x}_{\tilde{g}} - \mathbf{x}_r\|\beta_{r\tilde{\ell},s})} (u(\mathbf{x}_\ell) - u_i + r_{i\ell}(\mathbf{u})) \right]. \end{cases}$$

3.3.4 Approximation of the primal boundary fluxes with the DDFV method

3.3.4.1 Neumann boundary condition

Consider the problem (3.1) with $\Gamma_D = \emptyset$. Let ℓ be a boundary face of the mesh on Γ_N. Integrating the boundary condition on the face ℓ, we have

$$\int_\ell \kappa \boldsymbol{\nabla}\bar{u} \cdot \mathbf{n} = \int_\ell g_N,$$

that is to say

$$\bar{\mathcal{F}}_\ell = |\ell| \sum_{g\in\ell} \omega_g g_N(\mathbf{x}_g) + \mathcal{O}(h^k),$$

we thus impose this equation on the numerical flux

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| \sum_{g\in\ell} \omega_g g_N(\mathbf{x}_g).$$

3.3.4.2 Dirichlet boundary condition

Consider the problem (3.1) with $\Gamma_N = \emptyset$. Let ℓ be a boundary face of the cell i on Γ_D. Using the Dirichlet boundary condition on g in (3.8), equalities (3.8) and (3.25) give the system

$$\begin{cases} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_g - \mathbf{x}_i) = g_D(\mathbf{x}_g) - \bar{u}_i + \bar{r}_{ig}, \\ \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_s - \mathbf{x}_r) = \bar{u}_s - \bar{u}_r + \bar{r}_{rs}. \end{cases} \tag{3.38}$$

Then, we can decompose $\mathbf{n}_{i\ell}$ in the basis $((\mathbf{x}_g - \mathbf{x}_i),(\mathbf{x}_s - \mathbf{x}_r))$

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,ig} \frac{\mathbf{x}_g - \mathbf{x}_i}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig} \frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

with

$$\alpha_{i\ell,ig} = \frac{\|\mathbf{x}_g - \mathbf{x}_i\|}{(\mathbf{x}_g - \mathbf{x}_i) \cdot \mathbf{n}_{i\ell}} \geq 0,$$

and

$$\beta_{i\ell,ig} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\| \mathbf{n}_{i\ell} \cdot (\mathbf{x}_g - \mathbf{x}_i)^\perp}{(\mathbf{x}_s - \mathbf{x}_r) \cdot (\mathbf{x}_g - \mathbf{x}_i)^\perp}.$$

Thus, we have the expression of the gradient in the direction of the normal vector seen by the cell i, denoted by $\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)_i \cdot \mathbf{n}_{i\ell}$

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,ig} \frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_g - \mathbf{x}_i)}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig} \frac{\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot (\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say, using (3.38)

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = \alpha_{i\ell,ig} \frac{g_D(\mathbf{x}_g) - \bar{u}_i + \bar{r}_{ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig} \frac{\bar{u}_s - \bar{u}_r + \bar{r}_{rs}}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \tag{3.39}$$

Let $\mathbf{u} = (u_i)_{1\leq i\leq n}$ be the numerical solution on the mesh. By mimicking the expression of the exact flux (3.39), the numerical flux is defined by

$$\mathcal{F}_\ell(\mathbf{u}) = |\ell| \sum_{g\in\ell} \omega_g \kappa_g \left(\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|} \left(g_D(\mathbf{x_g}) - u_i + r_{ig}(\mathbf{u})\right) + \frac{\beta_{i\ell,ig}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left(u_s - u_r + r_{rs}(\mathbf{u})\right) \right),$$

with

$$\begin{cases} r_{ig}(\mathbf{u}) = \displaystyle\sum_{p=2}^{k} \frac{1}{V_i p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P_i}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \int_i (x-x_g)^q (y-y_g)^{(p-q)} dx, \\ r_{rs}(\mathbf{u}) = -\dfrac{1}{V_s} \displaystyle\sum_{s_i\in s} \int_{s_i} (\mathbf{x}-\mathbf{x}_s)\cdot\nabla P(\mathbf{x}_{s_i}) \\ \qquad - \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_s) \int_{s_i} (x-x_s)^q (y-y_s)^{(p-q)} dx + \dfrac{1}{V_r} \displaystyle\sum_{r_i\in r} \int_{r_i} (\mathbf{x}-\mathbf{x}_r)\cdot\nabla P(\mathbf{x}_{r_i}) \\ \qquad + \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_r) \int_{r_i} (x-x_r)^q (y-y_r)^{(p-q)} dx \\ \qquad - \displaystyle\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p P}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \left((x_s-x_g)^q (y_s-y_g)^{(p-q)} - (x_r-x_g)^q (y_r-y_g)^{(p-q)} \right) \end{cases}$$

In other words

$$\mathcal{F}_\ell(\mathbf{u}) = -\gamma_\ell u_i + \sum_{g\in\ell} \left(\frac{\omega_g \kappa_g \alpha_{i\ell,ig} |\ell|}{\|\mathbf{x}_g - \mathbf{x}_i\|} g_D(\mathbf{x_g}) \right) + r_\ell(\mathbf{u}),$$

with

$$\begin{cases} \gamma_\ell = \displaystyle\sum_{g\in\ell} \left(\frac{\omega_g \kappa_g \alpha_{i\ell,ig} |\ell|}{\|\mathbf{x}_g - \mathbf{x}_i\|} \right) \geq 0, \\ r_\ell(\mathbf{u}) = |\ell| \displaystyle\sum_{g\in\ell} \omega_g \kappa_g \left(\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_g - \mathbf{x}_i\|} r_{ig}(\mathbf{u}) + \frac{\beta_{i\ell,ig}}{\|\mathbf{x}_s - \mathbf{x}_r\|} (u_s - u_r + r_{rs}(\mathbf{u})) \right). \end{cases}$$

3.3.5 Approximation of the dual boundary fluxes with the DDFV method

3.3.5.1 Neumann boundary condition

Consider the problem (3.1) with $\Gamma_D = \emptyset$. Let $\tilde{\ell}$ be a boundary face of the mesh on Γ_N. Integrating the boundary condition on the face $\tilde{\ell}$, we have

$$\int_{\tilde{\ell}} \kappa \boldsymbol{\nabla} \bar{u} \cdot \mathbf{n} = \int_{\tilde{\ell}} g_N,$$

that is to say

$$\bar{\mathcal{F}}_{\tilde{\ell}} = |\tilde{\ell}| \sum_{g\in\tilde{\ell}} \omega_g g_N(\mathbf{x}_g) + \mathcal{O}(h^k),$$

we thus impose this equation on the numerical flux

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = |\tilde{\ell}| \sum_{g\in\tilde{\ell}} \omega_g g_N(\mathbf{x}_g).$$

3.3.5.2 Dirichlet boundary condition

On the dual mesh, we penalize the diagonal entries of the matrix and the right-hand side to impose the Dirichlet boundary condition.

3.3.6 Reconstruction of high order by interpolation

For a polynomial of degree k, we have $\frac{(k+1)(k+2)}{2}$ coefficients to calculate, so at least $(k+1)(k+2)$ ($\frac{(k+1)(k+2)}{2}\times$dimension) neighboring cells of the cell are required for stability purpose [41, 67]. When it is possible, the stencil will be centered on the cell, but the closer the cell is to the boundary or the discontinuity of κ, the more the stencil will be shifted in order to not to cross the discontinuity.

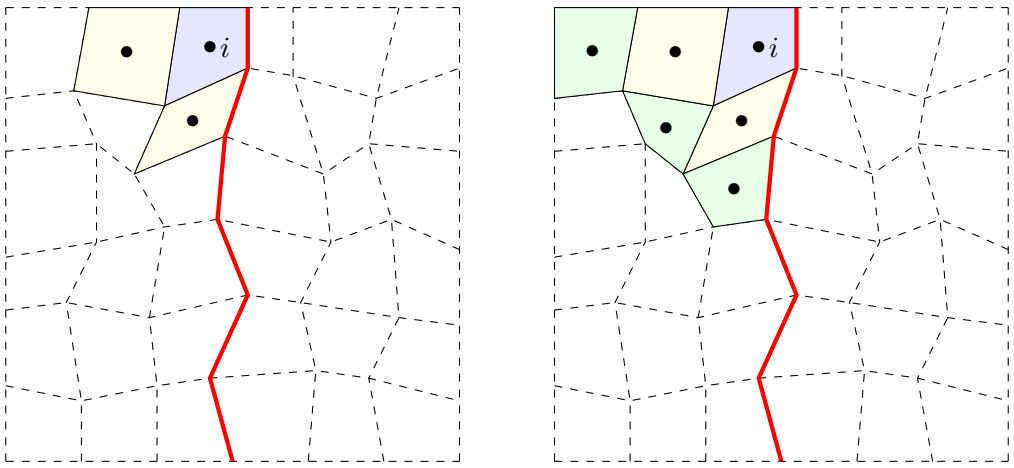


Fig. 3.2 – Construction of the stencil for the cell i with a discontinuity (in red)

To be more precise, the construction of the stencil of a cell i is illustrated on Figure 3.2. We denote this stencil by $\mathcal{S}_i = \{0,...,p\}$. For the sake of simplicity, we have assumed that the cells involved in the stencil have been renumbered. First the cell i itself (in blue) is added to the stencil and then we add the cells that share, at least, a face with the cell i (in yellow). If the number of cells we have already selected is not sufficient (in our case, $(k+1)(k+2)$ cells for a polynomial of order k), we add the cells that have, at least, a face linked to the cells that we have just been added to the stencil (in green) and so on until we have enough cells. In all the above process, we impose that the stencil does not cross any discontinuity of κ (see Figure 3.2).

Let $u_0, ..., u_p$ denote the $p+1$ values of $\mathbf{u}$ used for the calculation, with $p \geq 2$. The polynomial is of the form

$$P(\mathbf{x}) = \sum_{m=0}^{k} \sum_{n=0}^{k-m} a_{m,n}(\bar{u})(x - x_i)^m (y - y_i)^n.$$

The coefficients of the polynomial $P(\mathbf{x})$ are assumed to satisfy

$$\frac{1}{V_j} \int_j P(\mathbf{x}) dx = \bar{u}_j, \forall j \in \mathcal{S}_i.$$

This leads to the following system

$$\underbrace{\begin{pmatrix} \frac{1}{V_0}\int_0 1 & \frac{1}{V_0}\int_0 x - x_i & \frac{1}{V_0}\int_0 y - y_i & \dots & \frac{1}{V_0}\int_0 (y-y_i)^k \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{V_p}\int_p 1 & \frac{1}{V_p}\int_p x - x_i & \frac{1}{V_p}\int_p y - y_i & \dots & \frac{1}{V_p}\int_p (y-y_i)^k \end{pmatrix}}_{=:M} \underbrace{\begin{pmatrix} a_{0,0} \\ a_{1,0} \\ a_{0,1} \\ \vdots \\ a_{0,k} \end{pmatrix}}_{=:\mathbf{a}} = \underbrace{\begin{pmatrix} u_0 \\ \vdots \\ u_p \end{pmatrix}}_{=:\mathbf{d}}.$$

Since the matrix $\mathbf{M}$ has more rows than columns we have to use the least square method so that the vector $\mathbf{a}$ is computed as a solution to the linear system: $\mathbf{M}^t\mathbf{M}\mathbf{a} = \mathbf{M}^t\mathbf{d}$. We use the Givens method (see [53] p.206 and following) to solve the least-square problem.

In this process, we do not enforce the continuity of u at the vertices. Indeed, a priori, $P_j(\mathbf{x}_s) \neq P_i(\mathbf{x}_s)$ for $i \neq j$.

3.4 Monotonicity

A method borrowed from [51,52,105,111] and developed in the framework of 2D diffusion on arbitrary meshes can be used to make the scheme monotonic. The flux (3.21) can be rewritten as follows

$$\mathcal{F}_\ell(\mathbf{u}) = \gamma_\ell(u_j - u_i) + r_\ell(\mathbf{u})^+ - r_\ell(\mathbf{u})^-,$$

with

$$r_\ell(\mathbf{u})^+ = \frac{|r_\ell(\mathbf{u})| + r_\ell(\mathbf{u})}{2} \geq 0 \quad \text{and} \quad r_\ell(\mathbf{u})^- = \frac{|r_\ell(\mathbf{u})| - r_\ell(\mathbf{u})}{2} \geq 0.$$

Let us assume that $\mathbf{u} > \mathbf{0}$, the flux then reads as

$$\mathcal{F}_\ell(\mathbf{u}) = \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j}\right) u_j - \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i}\right) u_i.$$

and the coefficients $\left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j}\right)$ and $\left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i}\right)$ are positive. We end up with a two point flux, which is very favorable for the resolution of the system. However note that this system is non-symmetric and non-linear since the coefficients depend on the unknown vector $\mathbf{u}$.

3.4.1 Matrix form

The scheme reads as

$$-\sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) + \lambda_i V_i u_i = V_i f_i. \tag{3.40}$$

Consider a mesh the cells of which are numbered from 1 to n. Denoting

$$\begin{cases} \mathbf{u} = (u_i)_{1 \leq i \leq n}, \\ \mathbf{b} = (b_i)_{1 \leq i \leq n}, \\ \mathbf{A} = (A_{ij})_{1 \leq i,j \leq n}, \end{cases} \tag{3.41}$$

we can write this as the matrix-vector product

$$\mathbf{A}(\mathbf{u})\mathbf{u} = \mathbf{b}, \tag{3.42}$$

with

$$\begin{cases} A_{ii}(\mathbf{u}) = \displaystyle\sum_{\ell \in i, \ell \notin \Gamma_N} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i}\right) + V_i \lambda_i, \\ \\ A_{ij}(\mathbf{u}) = -\displaystyle\sum_{\ell \in i \cap j} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j}\right) \qquad \forall i \neq j, \end{cases} \tag{3.43}$$

and

$$\mathbf{b}_i = V_i f_i + \sum_{\ell \in i, \ell \in \Gamma_D} \left(r_\ell(\mathbf{u})^+ + \sum_{g \in \ell} \left(\frac{\omega_g \kappa_g \alpha_{i\ell,ig} |\ell|}{\|\mathbf{x}_g - \mathbf{x}_i\|} \right) g_D(\mathbf{x_g}) \right) + \sum_{\ell \in i, \ell \in \Gamma_N} |\ell| \sum_{g \in \ell} \omega_g g_N(\mathbf{x}_g). \tag{3.44}$$

Remark 3.4.1. *Assuming that $f \geq 0$ and $g \geq 0$, all the components of the right hand side $\mathbf{b}$ are non-negative. Assuming moreover that f and g are not both identically zero, then at least one component of $\mathbf{b}$ is positive.*

3.4.2 Picard iteration method

The system (3.43) is of the form $\mathbf{A}(\mathbf{u})\mathbf{u} = \mathbf{b}$. In order to solve them, we use a Picard iteration method. We start with an initial guess $\mathbf{u}^0 > 0$, compute the matrix $\mathbf{A}(\mathbf{u}^0)$ and solve $\mathbf{A}(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b}$. Repeating this process, we build a sequence $(\mathbf{u}^\nu)$ that, if it converges to a positive vector, tends to a solution of the scheme. We stop the algorithm when the difference $\mathbf{u}^{\nu+1} - \mathbf{u}^\nu$ between two successive iterates is small enough. To summarize, the following algorithm is used

$$\begin{aligned} &\nu = 0 \\ &A(\mathbf{u}^0)\mathbf{u}^1 = \mathbf{b} \\ &\text{While } \frac{\|\mathbf{u}^{\nu+1} - \mathbf{u}^\nu\|_2}{\|\mathbf{u}^\nu\|_2} > \mu \\ &\quad A(\mathbf{u}^\nu)\mathbf{u}^{\nu+1} = \mathbf{b} \\ &\quad \nu = \nu + 1. \end{aligned} \tag{3.45}$$

Unfortunately, we are unable to prove that the above algorithm converges. Nevertheless, we prove in Section 3.5.3 below that the scheme is well defined at each iteration of the algorithm, as soon as the initial guess $\mathbf{u}^0$ is positive.

3.5 Properties

3.5.1 Conservation

Proposition 3.5.1. *Assume that $\mathbf{u} > \mathbf{0}$ and consider homogeneous Neumann boundary conditions, then the scheme defined by (3.40) is conservative, that is to say*

$$\sum_{i=1}^{n} V_i \lambda_i u_i = \sum_{i=1}^{n} V_i f_i,$$

Indeed it satisfies the equality

$$\sum_{i=1}^{n} \left(-\sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) \right) = 0.$$

The proof is done in Appendix D.2.

3.5.2 Monotonicity

Consider the definition of an M-matrix (see for instance [87])

Definition 3.5.2. *An $n \times n$ matrix $\mathbf{A}$ that can be expressed in the forme $\mathbf{A} = s\mathbf{I} - \mathbf{B}$, where $\mathbf{B} = (b_{ij})_{1 \le i,j \le n}$ with $b_{ij} \ge 0$, $1 \le i,j \le n$, and $s \ge \rho(\mathbf{B})$, the maximum of the moduli of the eigenvalues of $\boldsymbol{B}$, is called an M-matrix.*

We use the following lemma

Lemma 3.5.3. *A matrix $\mathbf{A} = (A_{ij})_{1 \le i,j \le n}$ is an M-matrix if it satisfies the following inequalities*

$$\forall i \neq j, \quad A_{ij} \le 0, \quad \text{and} \quad \forall i, \quad \sum_{j=1}^{n} A_{ij} \ge 0.$$

Moreover, if the last inequality is strict, we say that $\mathbf{A}$ is a strict M-matrix.

Proposition 3.5.4. *Assume that $\mathbf{u} > \mathbf{0}$. Then the matrice $\mathbf{A}$ defined by (3.43) is such that $\mathbf{A}^t$ is a strict M-matrice.*

Proof. The matrix $\mathbf{A}$ satisfies

$$\forall i \neq j, \quad A_{ij} \leq 0 \quad \text{and} \quad \forall j, \quad \sum_{i=1}^{n} A_{ij} > 0.$$

Indeed we have, for all j

$$\sum_{i=1}^{n} A_{ij} = \sum_{i=1}^{n} \left(\sum_{\ell \in i, \ell \notin \Gamma_N} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i} \right) - \sum_{\ell \in i \cap j} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^+}{u_j} \right) \right) + \lambda_j V_j.$$

Thanks to Proposition 3.5.1, only the boundary terms and the mass term remain, for all j

$$\sum_{i=1}^{n} A_{ij} = \sum_{i=1}^{n} \sum_{\ell \in (i \cap \Gamma_D)} \left(\gamma_\ell + \frac{r_\ell(\mathbf{u})^-}{u_i} \right) + \lambda_j V_j > 0.$$

□

Theorem 3.5.5. *Assume that $f > 0$ and $g > 0$. Let $\mathbf{A}$ and $\mathbf{b}$ be defined by* (3.43)-(3.44)*. Then $\mathbf{A}^{-1}\mathbf{b} = \mathbf{u} \geq \mathbf{0}$.*

Proof. As $\mathbf{A}^t$ is a strict M-matrix $\mathbf{A}$ is invertible and its inverse has only non-negative entries (see for example [98], Corollary 3.20). In view of Remark 2.5.1, the right hand side is non-negative, hence $\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} \geq \mathbf{0}$. □

Remark 3.5.6. *The scheme preserves positivity if the inversion of the linear system is exact. The above proof assumes that the matrix M^{-1} is calculated exactly. Obviously, in practice, this is not the case. In the tests we have carried out, the error is small enough not to affect the calculations. However, in rare cases, the inversion of the matrix led to a solution with negative components, causing the calculation to stop. This error can be reduced by working on the condition number of the matrix or on methods for solving linear systems, which is a perspective.*

3.5.3 Well-posedness of the Picard iteration method

Proposition 3.5.7. *Assume that $f \geq 0$, $g \geq 0$, and either $\|f\|_{L^2(\Omega)} > 0$ or $\|g\|_{L^2(\partial\Omega)} > 0$. Assume moreover that $\mathbf{u}^0 > \mathbf{0}$. Then, the algorithm* (3.45) *defines a sequence $(\mathbf{u}^\nu)_{\nu \geq 0}$ such, that for all ν, $\mathbf{u}^\nu > \mathbf{0}$.*

To prove this property, we need to introduce the concept of irreducible matrix. We quote here [98, Definition 1.15].

Definition 3.5.8. *An $n \times n$ matrix $\mathbf{A}$ is **reducible** if there exits an $n \times n$ permutation matrix $\mathbf{P}$ such that*

$$\mathbf{P}\mathbf{A}\mathbf{P}^t = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{bmatrix},$$

*where $\mathbf{A}_{11}$, $\mathbf{A}_{12}$, $\mathbf{A}_{22}$ are respectively $r \times r$, $r \times (n-r)$ and $(n-r) \times (n-r)$ sub-matrices with $1 \leq r < n$. If no such permutation matrix exists, then $\mathbf{A}$ is **irreducible**.*

The matrice $\mathbf{A}$ defined by (3.43) is irreducible thanks to the following Lemma (see [98, Theorem 1.17]).

Lemma 3.5.9. *To any $n \times n$ matrix $\mathbf{A}$ we associate the graph of nodes 1,2,...,n and of directed edges connecting $\mathbf{x}_i$ to $\mathbf{x}_j$ if $A_{ij} \neq 0$. Then $\mathbf{A}$ is irreducible if and only if for any pair $i \neq j$ there exists a chain of edges that allows to go from $\mathbf{x}_i$ to $\mathbf{x}_j$,*

$$A_{i,k1} \neq 0 \rightarrow A_{k1,k2} \neq 0 \rightarrow \cdots \rightarrow A_{km,j} \neq 0.$$

With these definitions we can make use of the following theorem (see [98], Corollary 3.20).

Theorem 3.5.10. *If* $\mathbf{A}$ *is an irreducible strict M-matrix, then it is invertible and, for all* i, j *(*$1 \leq i,j \leq n$*),* $(A^{-1})_{ij} > 0$.

We are now in position to prove Proposition 3.5.7.

Proof of Proposition 3.5.7. We argue by induction on the index ν. We assume that $\mathbf{u}^\nu > 0$. Hence $(\mathbf{A}(\mathbf{u}^\nu))^t$ is a strict M-matrix (see Proposition 3.5.4). It is easy to check that $(\mathbf{A}(\mathbf{u}^\nu))^t$ is also irreducible. Thus, applying Theorem 3.5.10, $(\mathbf{A}(\mathbf{u}^\nu))^t$ is invertible and all the entries of $(\mathbf{A}(\mathbf{u}^\nu))^{-t}$ are positive. Consequently, all the entries of $(\mathbf{A}(\mathbf{u}^\nu))^{-1}$ are positive. Using Remark 2.5.1, we know that all components of $\mathbf{b}$ are non-negative. Moreover, because of the assumption that either $\|f\|_{L^2(\Omega)} > 0$ or $\|g\|_{L^2(\partial\Omega)} > 0$, at least one component of $\mathbf{b}$ is positive. We thus have, for all i $(1 \leq i \leq n)$

$$u_i^{\nu+1} = \sum_{j=1}^{n} (\mathbf{A}(\mathbf{u}^\nu))_{ij}^{-1} b_j > 0,$$

since all terms of this sum are non-negative, with one at least that does not vanish. □

Proposition 3.5.7 shows that the condition $\mathbf{u}^\nu > \mathbf{0}$ remains satisfied during the Picard iteration method, which allows to define $\mathbf{A}(\mathbf{u}^\nu)$ for all $\nu \geq 0$.

3.6 Numerical experiments

Given $\Omega =]0,1[^2$, κ a diffusion coefficient and g a function defined on $\partial\Omega$, consider Problem (3.1) with $\lambda = 0$ and $\Gamma_N = \emptyset$

$$\begin{cases} -\boldsymbol{\nabla} \cdot (\kappa \boldsymbol{\nabla} \bar{u}) = f & \text{in} \quad \Omega, \\ \bar{u} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{3.46}$$

In addition to Cartesian meshes we will use the two following types of meshes (see Figure 3.3):

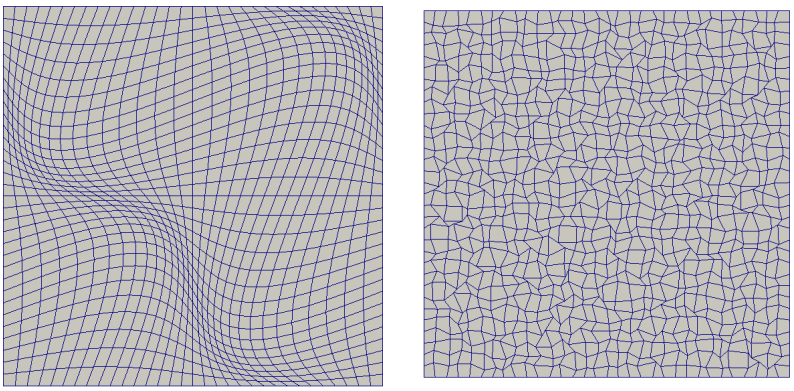
1. deformed meshes, the deformation of which from the Cartesian mesh is given by

$$(x,y) \to (x + 0.1\sin(2\pi x)\sin(2\pi y), y + 0.1\sin(2\pi x)\sin(2\pi y)),$$

2. randomly deformed meshes, the deformation of which from the unit Cartesian mesh with cells of size Δx is given by

$$(x,y) \to 0.1(x,y) + 0.9(x + 0.45a\Delta x, y + 0.45b\Delta x),$$

where a, b are random numbers distributed according to the uniform law on $[-1,1]$.



(a) A deformed mesh (b) A random mesh

Fig. 3.3 – Examples of deformed meshes.

The L^2-error and L^2-error on the fluxes used in the following tests are respectively given by

$$\frac{\|\mathbf{u}-\bar{\mathbf{u}}\|_2}{\|\bar{\mathbf{u}}\|_2} \quad \text{and} \quad \frac{\left(\sum_\ell \left(\mathcal{F}_\ell(\mathbf{u}) - |\ell| \sum_{g\in\ell} \omega_g \kappa_g \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}\right)^2\right)^{1/2}}{\left(\sum_\ell \left(|\ell| \sum_{g\in\ell} \omega_g \kappa_g \boldsymbol{\nabla}\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}\right)^2\right)^{1/2}},$$

We also use the H^1 semi-norm error defined by

$$\frac{\|\boldsymbol{\nabla}_h\mathbf{u} - \boldsymbol{\nabla}\bar{u}\|_2}{\|\boldsymbol{\nabla}\bar{u}\|_2},$$

where

$$\|\boldsymbol{\nabla}\bar{u}\|_2 = \left(\sum_i V_i \|\boldsymbol{\nabla}\bar{u}(\mathbf{x}_i)\|^2\right)^{1/2} \quad \text{and} \quad \|\boldsymbol{\nabla}_h\mathbf{u} - \boldsymbol{\nabla}\bar{u}\|_2 = \left(\sum_i V_i \|\boldsymbol{\nabla}P_i(\mathbf{x}_i) - \boldsymbol{\nabla}\bar{u}(\mathbf{x}_i)\|^2\right)^{1/2},$$

P_i being the polynomial obtained by reconstruction with the values of the solution $\mathbf{u}$.

For all the tests, the stopping criterion μ and the initial guess $\mathbf{u}^0$ of the fixed-point algorithm (3.45) are $\mu = 10^{-12}$ and $u_i^0 = 1, \forall i$. We use the linear solver GMRES with the preconditioner ILU (see [83], Chapter 7.4) with the convergence criterion is 10^{-14}.

3.6.1 Numerical accuracy assessment

In this section we present numerical results for diffusion problems of type (3.46) with analytical solutions. The first (resp. second) case involves a discontinuous (resp. anisotropic) diffusion coefficient. Numerical convergence rates are evaluated using the L^2 norm of the solution as well the L^2 norm of the fluxes and the H^1 semi-norm. We perform a convergence study for these problems with a sequence of successively refined deformed meshes as that shown in Figure 3.3a. For the sake of brevity we present only the results on this type of mesh. We obtain similar results on randomly deformed meshes as that shown on Figure 3.3b. We will also skip the case of continuous scalar diffusion coefficient, as it is simpler than the discontinuous and anisotropic cases.

3.6.1.1 Discontinuous diffusion coefficient

Recall that we have assumed the possible discontinuities of the diffusion coefficient κ coincide with edges of the mesh. Given

$$\kappa(\mathbf{x}) = \begin{cases} 1 & \text{if} \quad x \leq \dfrac{1}{2} \\ 2 & \text{if} \quad x > \dfrac{1}{2} \end{cases}, \quad f(\mathbf{x}) = 2\pi^2 \cos(\pi x)\cos(\pi y) + 20, \quad g(\mathbf{x}) = 0,$$

the function

$$\bar{u}(\mathbf{x}) = \begin{cases} \cos(\pi x)\cos(\pi y) - 10x^2 + 12 & \text{if} \quad x \leq \dfrac{1}{2}, \\ \dfrac{1}{2}\cos(\pi x)\cos(\pi y) - 5x^2 + \dfrac{43}{4} & \text{if} \quad x > \dfrac{1}{2}, \end{cases}$$

is solution to (2.41). Results are summarized in Figure 3.4 which shows that all schemes are k-th-order accurate in the L^2 norm, the L^2 norm of the fluxes and the H^1 seminorm. We can note that there is a superconvergence for odd orders.

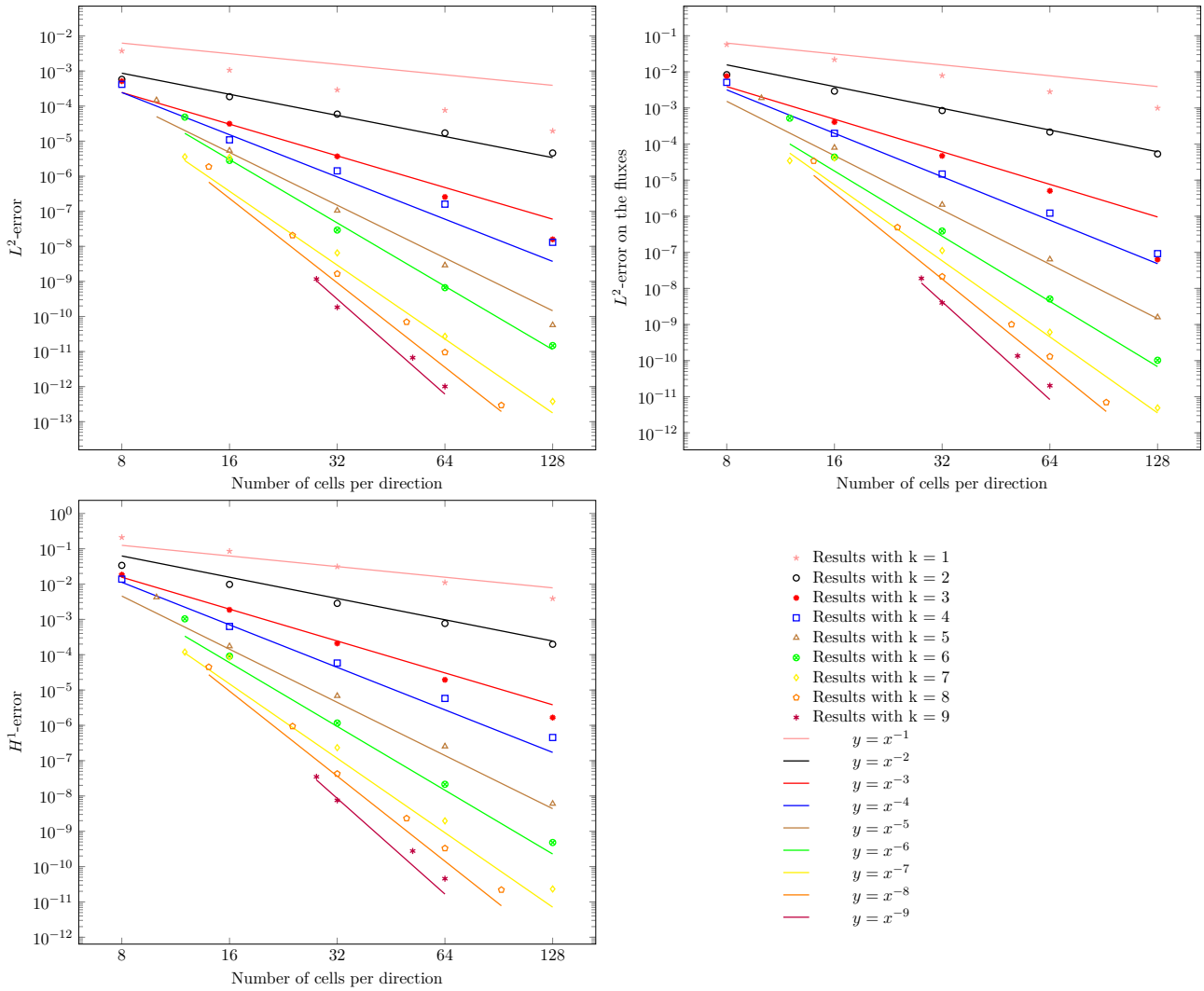


Fig. 3.4 – L^2-error (on the top left), L^2-error on the fluxes (on the top right) and error in the H^1 seminorm (on the bottom left) for problem of Section 3.6.1.1.

We see that, even if $\nabla\bar{u}$ is discontinuous in this problem, we are able to achieve an arbitrary order of accuracy. The by point for this is to design a stencil that do not cross discontinuities of κ, as explained in Section 3.3.6.

3.6.1.2 Anisotropic diffusion coefficient

Given

$$\boldsymbol{\kappa}(\mathbf{x}) = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix},$$

and

$$f(\mathbf{x}) = 3\pi^2 \sin(\pi x)\sin(\pi y), \quad g(\mathbf{x}) = 0,$$

the function

$$\bar{u}(\mathbf{x}) = \sin(\pi x)\sin(\pi y)$$

is solution to (2.41). Results are summarized in Figure 3.5 which shows that all schemes are k-th-order accurate in the L^2 norm, the L^2 norm of the fluxes and the H^1 seminorm. We can note that there is a superconvergence for odd orders. Of course, similar results may be obtained for a scalar-valued diffusion coefficient κ.

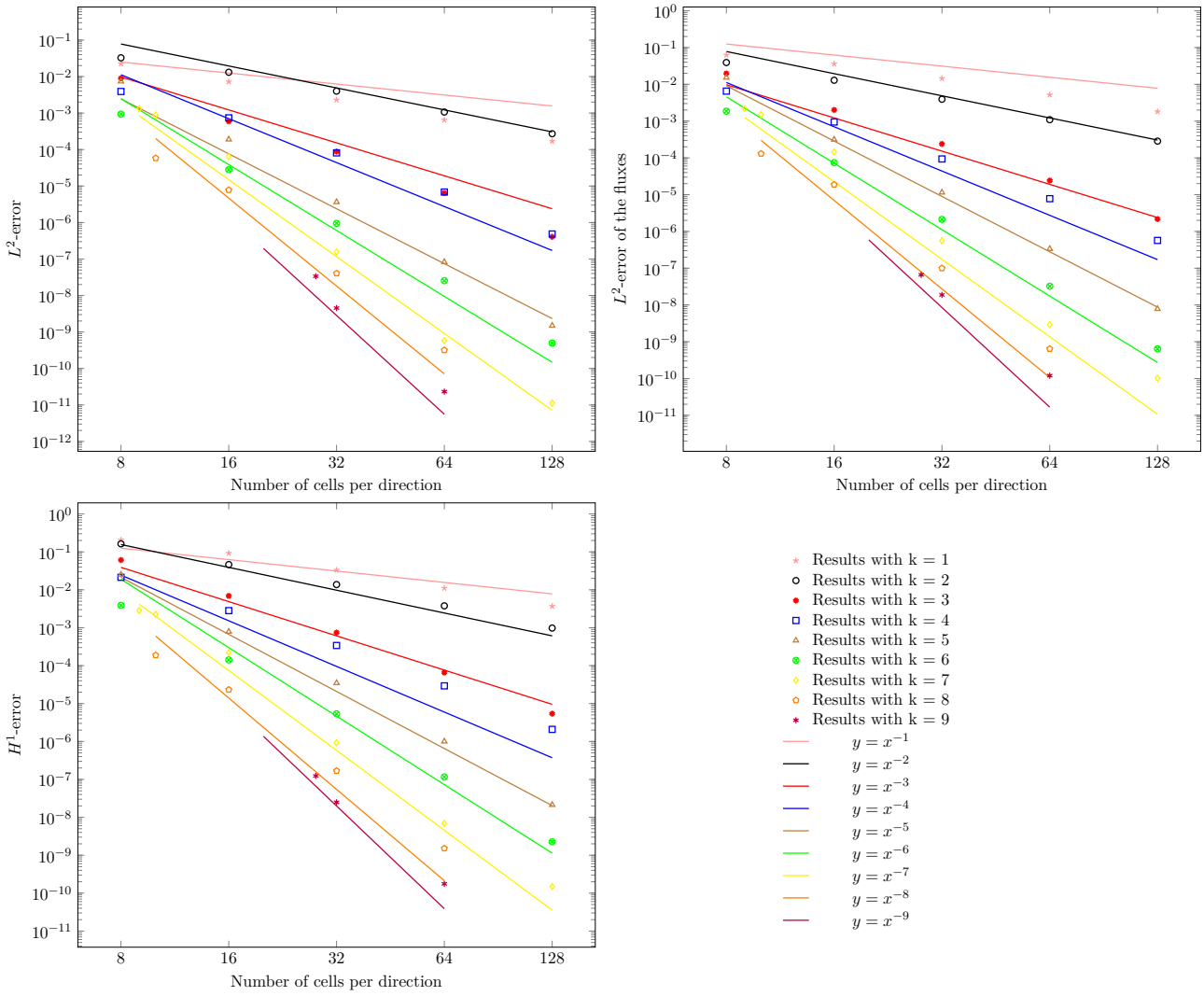


Fig. 3.5 – L^2-error (on the top left), L^2-error on the fluxes (on the top right) and error in the H^1 seminorm (on the bottom left) for problem of Section 3.6.1.2.

Scheme	Number of cells per direction	Number of iterations	Execution time (ratio)
Order 1	168	172	1
Order 2	212	180	2.33
Order 3	31	132	0.10
Order 4	31	120	0.20
Order 5	19	103	0.20
Order 6	14	124	0.26
Order 7	16	143	1.08
Order 8	10	154	0.78

Tab. 3.1 – Minimum number of cells to reach a precision on the L^2-error of 10^{-5} with the time of execution and the number of iterations of the fixed point algorithm for order 1 to 8 for problem of Section 3.6.1.2.

Scheme	Number of cells per direction	Number of iterations	Execution time (ratio)
Order 3	323	135	1
Order 4	343	135	2.49
Order 5	93	122	0.56
Order 6	76	134	0.73
Order 7	46	90	0.52
Order 8	40	76	0.62
Order 9	30	75	0.75

Tab. 3.2 – Minimum number of cells to reach a precision on the L^2-error of 10^{-9} with the time of execution and the number of iterations of the fixed point algorithm for order 3 to 9 for problem of Section 3.6.1.2.

Table 3.1 (resp. Table 3.2) gives the minimum number of cells per direction required to achieve an accuracy of 10^{-5} (resp. 10^{-9}) on the L^2-error, with the number of iterations of the fixed point algorithm and the time of execution. As expected, the number of cells needed to achieve the desired precision (first column) is a decreasing function of the order. The second column gives the number of fixed point iterations required to satisfy the stagnation criterion. This number is either constant or decreasing with the order, which is not intuitive and is a good point. The more interesting column is the last one giving the total computational cost of the method. This computational time is a trade-off between the algorithmic complexity and the precision of the method, which both increase with the order. We notice that, in general, execution time decreases as the order increases. For a large error setpoint value (10^{-5}), the optimal choice of scheme is the third-order one. However, when decreasing the error setpoint value (10^{-9}) higher-order schemes perform better, and the optimal order becomes seven. We anticipate that small values of the error setpoint will favor the highest orders. We obtain speed-ups of factors up to ten in term of computational time to reach the desired precision. We also observed that odd orders perform better than even orders. This confirms what we notice on Figures 3.4 and 3.5: a super-convergence is achieved for odd orders. We also observe a somewhat spectral convergence: for a fixed mesh size, the error decreases as k grows.

3.6.2 Monotonicity assessment

We propose a challenging benchmark borrowed from [110] to compare a non-monotonic scheme, which can give nonpositive solutions (in this case the usual DDFV scheme), with our monotonic high-order scheme which always gives nonnegative solutions. For this test we have used Cartesian meshes.

3.6.2.1 Tensor-valued coefficient κ and square domain with a square hole

Consider the square domain with a square hole $\Omega =]0,1[^2 \setminus \left[\frac{4}{9},\frac{5}{9}\right]^2$, $f(\mathbf{x}) = 0$ in Ω and $g(\mathbf{x}) = 0$ (resp. $g(\mathbf{x}) = 2$) on the external (resp. internal) boundary. We choose

$$\boldsymbol{\kappa} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 10^4 \end{pmatrix} \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}, \quad \theta = \frac{\pi}{6}.$$

We compare the results obtained with the monotonic high order schemes and DDFV schemes on a Cartesian mesh with 2000 cells of size 1/45. The stopping criterion of the fixed point algorithm is $\mu = 10^{-12}$, except for order 6 for which $\mu = 10^{-10}$ and for order 7 and 8 for which $\mu = 10^{-6}$ to reduce the computing time. Figure 2.7 shows the mesh, the DDFV solution and its negative and positive parts. Figure 3.7 displays the monotonics high order solutions while Table 3.3 gives the minimum and the maximum of each solution.

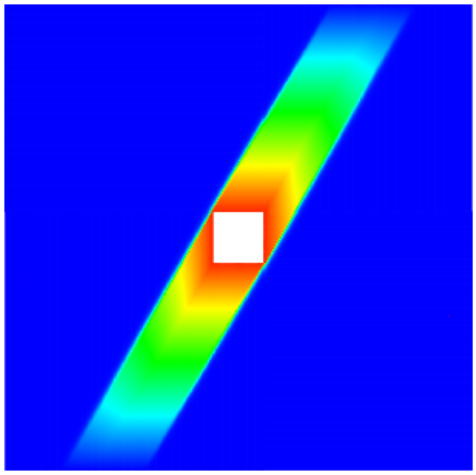


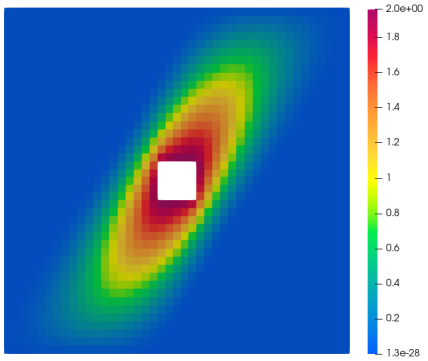
Fig. 3.6 – Numerical solution obtained with the DDFV scheme on a highly refined mesh (1310720 cells of size $\Delta x = 1/1152$).

As explained in Remark 3.5.6, the precision of the inversion of the linear system sometimes leads to negative entries in the solution vector $\mathbf{u}$. In general, this can be fixed by using the result of a low-order calculation as the initial guess of the high-order calculation. This procedure is also favorable regarding the computation time. It significantly reduces the overall cost of the simulation. However, we encountered one case for which this fix was not sufficient. For the test of order 5, for a Cartesian mesh with 86 cells per direction, we did not manage to run the simulation. We think that this is a severe issue for this kind of methods which is in general not addressed in the papers. In the near future, we intend to work on the linear system inversion.

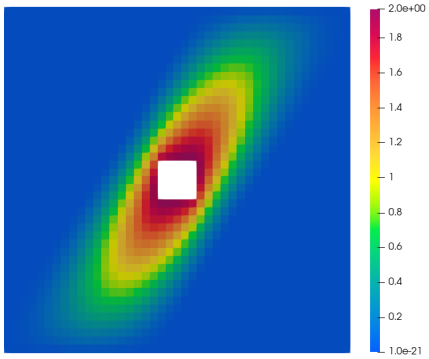
Scheme	Minimum of the solution	Maximum of the solution
DDFV	-4.59×10^{-1}	2.05
Monotonic scheme of order 1	$1.3e-28$	1.96
Monotonic scheme of order 2	$1e-21$	1.96
Monotonic scheme of order 3	$1.7e-27$	1.98
Monotonic scheme of order 4	$3.9e-30$	1.97
Monotonic scheme of order 5	$1.1e-27$	1.97
Monotonic scheme of order 6	$4.3e-27$	1.98
Monotonic scheme of order 7	$7.9e-25$	1.98
Monotonic scheme of order 8	$5.4e-21$	1.98

Tab. 3.3 – Minimum and maximum of the numerical solution to the problem of section 3.6.2.1 for the Cartesian mesh with 2000 cells of size 1/45.

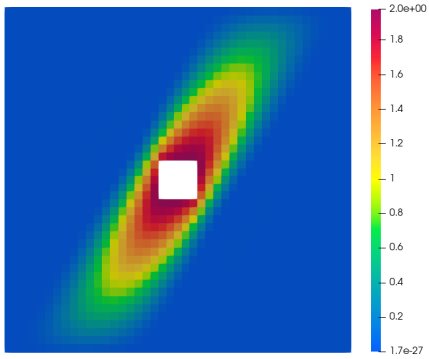
Even for a highly refined mesh (1310720 squares of size $\Delta x = 1/1152$) the solution obtained with the usual (non-monotonic) DDFV scheme (see Figure 3.6) has negative values up to -2.11×10^{-3}. On the other hand the high-order solutions obtained with the monotonic scheme remain always positive whatever the order: see Figure 3.7 and Table 3.3 which gives the minimum and the maximum of each solution calculated with a Cartesian mesh (2000 cells of size 1/45), up to order 6. We also observe on Figure 3.7 that the solution for $k = 3$ is closer to the converged solution (see 3.6) than the solution for $k = 1$. This is reminiscent of the spectral convergence we pointed out in Section 3.6.1.



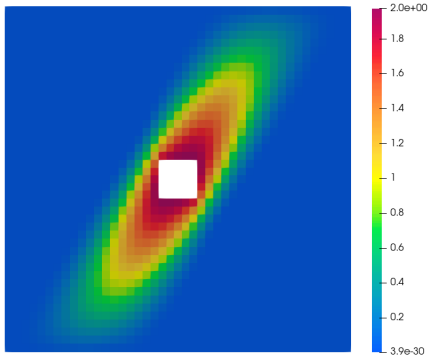
(a) Solution obtained with the monotonic scheme of order 1



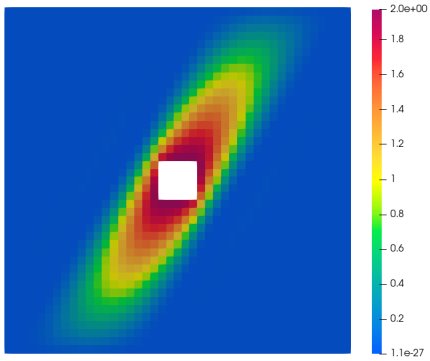
(b) Solution obtained with the monotonic scheme of order 2



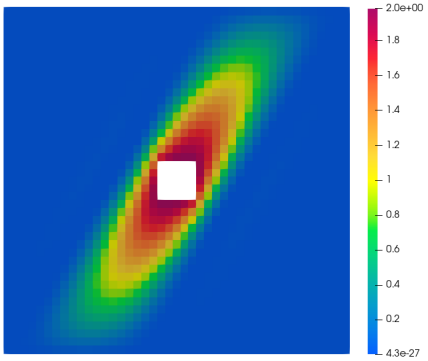
(c) Solution obtained with the monotonic scheme of order 3



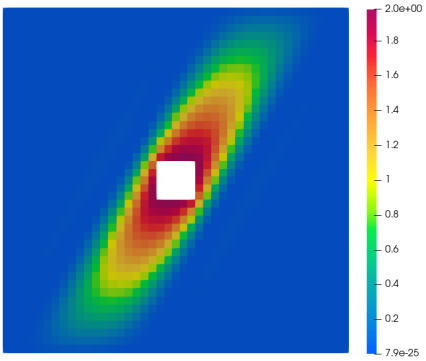
(d) Solution obtained with the monotonic scheme of order 4



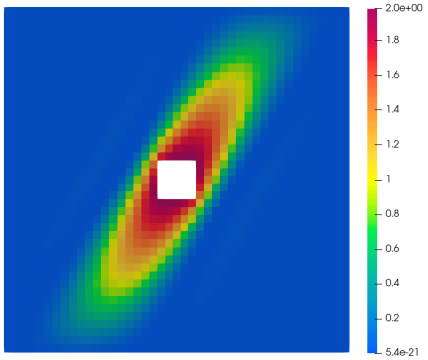
(e) Solution obtained with the monotonic scheme of order 5



(f) Solution obtained with the monotonic scheme of order 6

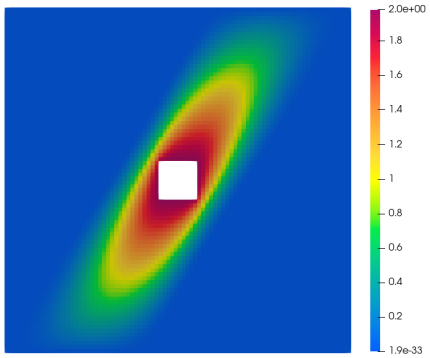


(g) Solution obtained with the monotonic scheme of order 7

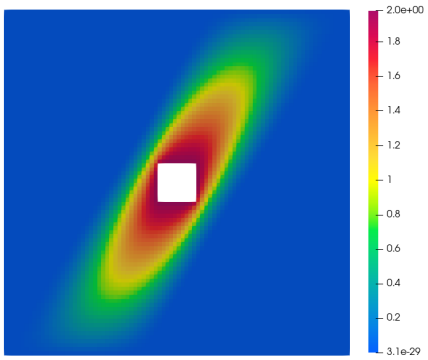


(h) Solution obtained with the monotonic scheme of order 8

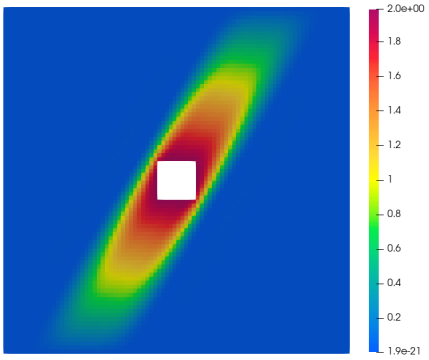
Fig. 3.7 – Numerical solutions obtained with monotonic schemes of order 1 to 8 for a cartesian mesh (2000 cells of size 1/45)



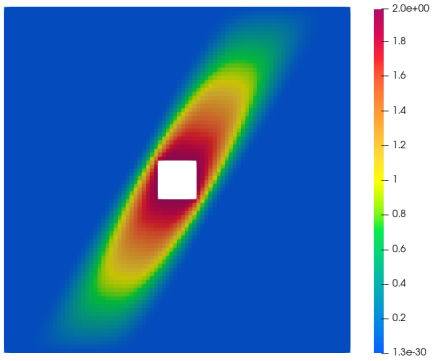
(a) Solution obtained with the monotonic scheme of order 1



(b) Solution obtained with the monotonic scheme of order 2



(c) Solution obtained with the monotonic scheme of order 3



(d) Solution obtained with the monotonic scheme of order 4

Fig. 3.8 – Numerical solutions obtained with monotonic schemes of order 1 to 4 for a cartesian mesh with 90 cells per direction

3.6.2.2 Fokker-Planck type diffusion equation

This benchmark is a simplified version of the one from [69]. Given $\Omega =]-50,50[^2$, $T = 250$, $\mathbf{v} = (v_x, v_y)$ the velocity variable and $\mathbf{V} = (-20,20)$ the averaged velocity, we are looking for the distribution function $\bar{u} = \bar{u}(\mathbf{v},t)$, solution to the simplified Fokker-Planck equation

$$\begin{cases} \dfrac{\partial \bar{u}}{\partial t} - \boldsymbol{\nabla}_{\mathbf{v}} \left(\boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u} \right) = 0 & \text{in} \quad \Omega \times [0,T], \\ \boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u} \cdot \mathbf{n} = 0 & \text{on} \quad \partial\Omega \times [0,T], \\ \bar{u}(0) = \bar{u}^0 & \text{in} \quad \Omega, \end{cases} \tag{3.47}$$

where the diffusion coefficient $\boldsymbol{\kappa} = \boldsymbol{\kappa}(\mathbf{v})$ and the initial condition $\bar{u}^0$ are given by

$$\boldsymbol{\kappa}(\mathbf{v}) = \mathbf{I} - \frac{1}{\|\mathbf{v}\|^2} \mathbf{v} \otimes \mathbf{v}, \quad \bar{u}^0(\mathbf{v}) = \frac{1}{\pi} \exp(-\|\mathbf{v} - \mathbf{V}\|^2). \tag{3.48}$$

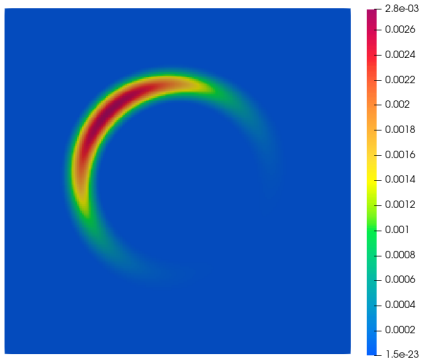
Note that the full Fokker-Planck equation would read as

$$\frac{\partial \bar{u}}{\partial t} + \boldsymbol{\nabla}_{\mathbf{v}} \cdot (\mathbf{v} \bar{u}) - \boldsymbol{\nabla}_{\mathbf{v}} \left(\boldsymbol{\kappa} \boldsymbol{\nabla}_{\mathbf{v}} \bar{u} \right) = 0.$$

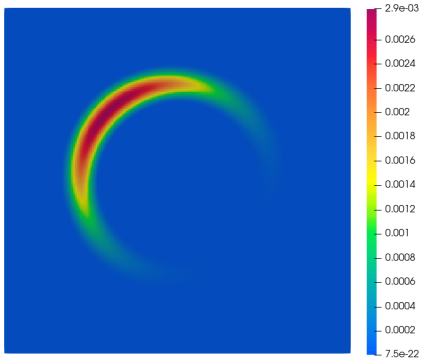
The diffusion coefficient κ defined by (3.48) is degenerated: it does not satisfy (3.2), hence the theoritical results of the preceding Sections do not apply to the present case. It follows in particular that the well-posedness of the fixed-point algorithm (see Section 3.5.3) is no longer ensured. However, $\bar{u}$ should remain positive, and the non-monotonic DDFV scheme produces non-physical negative values. We will see that our monotonic scheme fixes it. This diffusion tensor correspond to a degenerate

diffusion problem along the circle of radius $\|\mathbf{v}\|$.

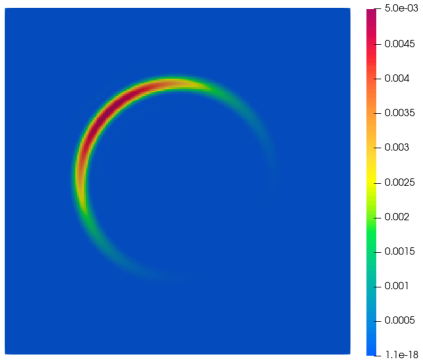
The backward Euler scheme is used for time integration.



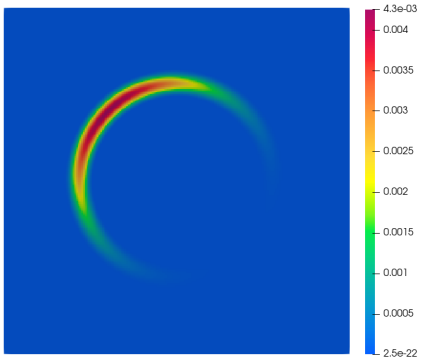
(a) Solution obtained with the monotonic scheme of order 1



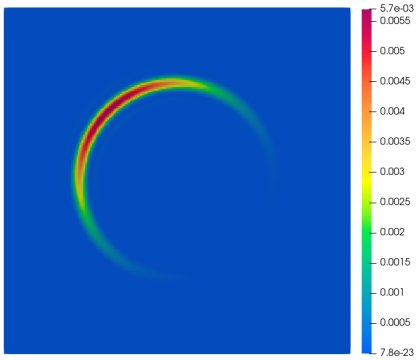
(b) Solution obtained with the monotonic scheme of order 2



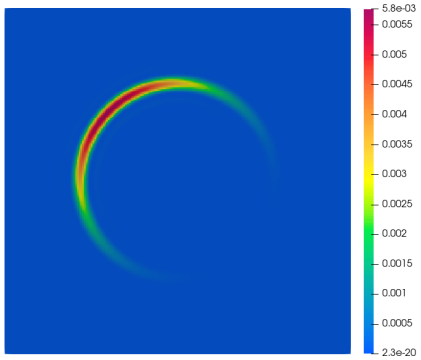
(c) Solution obtained with the monotonic scheme of order 3



(d) Solution obtained with the monotonic scheme of order 4



(e) Solution obtained with the monotonic scheme of order 5



(f) Solution obtained with the monotonic scheme of order 6

Fig. 3.9 – Numerical solutions obtained with monotonic schemes of order 1 to 6 for a cartesian mesh with 200 cells per direction for problem of Section 3.6.2.2

To limit the calculation time, the stopping criterion of the fixed point algorithm is $\mu = 10^{-5}$. Figure 3.9 displays the numerical solutions obtained with the Cartesian mesh of 200^2 cells. Table 2.4 gives the minima and maxima of the DDFV solution for a sequence of refined Cartesian meshes and Table 3.4 gives the minima and the maxima of the numerical solution obtained with the monotonic schemes up to order 6. We observe that the minima of the solutions to monotonic schemes always remain non negative, as expected. Compared to the solutions obtained with the DDFV scheme, given by Figure 2.10 and the solutions obtained by the monotonic DDFV schemes, given by Figure 2.9, the monotonic arbitrary order schemes are more diffusive (in the radial direction). However, we can note that the higher is the order, the less diffusive (in the radial direction) is the scheme.

Scheme	Minimum of the solution	Maximum of the solution
Monotonic scheme of order 1	$1.5e-23$	$2.8e-3$
Monotonic scheme of order 2	$7.5e-22$	$2.9e-3$
Monotonic scheme of order 3	$1.1e-18$	$5.0e-3$
Monotonic scheme of order 4	$2.5e-22$	$4.3e-3$
Monotonic scheme of order 5	$7.8e-23$	$5.7e-3$
Monotonic scheme of order 6	$2.3e-20$	$5.8e-3$

Tab. 3.4 – Minimum and maximum of the numerical solution to the problem of section 3.6.2.2 for the Cartesian mesh with 200 cells per direction.

3.7 Concluding remarks

This chapter proposes an arbitrary-order monotonic Finite Volume scheme for the elliptic problem (3.1) on general 2D meshes. The new non-linear method we have detailed here is arbitrary-order convergent even for anisotropic and/or discontinuous diffusion coefficients on deformed meshes. Furthermore it allows to deal with all boundary conditions (Dirichlet, Neumann). This scheme uses a polynomial reconstruction involving values in neighboring cells to evaluate the secondary unknowns at the Gauss quadrature points. We have adapted the non-linear process from [105] to enforce monotonicity. We have assessed numerically both its accuracy and monotonicity.

Numerical performance could be improved. Indeed, the convergence of the fixed-point algorithm is not guaranteed and may be very slow. This is observed in particular in test cases where the classical DDFV scheme gives negative solutions. Techniques for accelerating this fixed point could be explored, such as Anderson acceleration (see [2, 91]) or the ϵ-algorithm (see [13, 14]).

The next step is to extend the method to non-linear diffusion (with a diffusion coefficient depending on the unknown) and to arbitrary order unsteady diffusion, taking inspiration from [45] for example. The extension of the scheme to the three-dimensional case, based on the same ideas, is the subject of ongoing works.

Conclusions and perspectives

In this thesis, we developed positive high-order finite volume schemes for diffusion on deformed meshes. We started by proposing an arbitrary-order positivity-preserving scheme for diffusion on 1D deformed meshes in Chapter 1. We dealt with any boundary conditions (Neumann, Dirichlet, mixed). We have also proposed a symmetrical version of this scheme to guarantee the maximum principle. Studying the 1D case, we have shown that the proposed schemes are conservative, that the fluxes are consistent, and prove the convergence of the scheme at expected order under a reasonable assumption on the mesh. We have carried out numerical tests to illustrate these properties up to order 9. Then, we have presented two positivity-preserving schemes for diffusion on 2D deformed meshes in Chapter 2. They are based on the same principle, but differ in the way they handle secondary unknowns (node values). The first scheme, called monotonic diamond scheme, uses polynomial reconstruction to evaluate node values. The second scheme, called monotonic DDFV scheme, treats these node values as unknowns and compute them as solution to a diffusion problem on a dual mesh. We consider both Dirichlet and Neumann boundary conditions. We proposed numerical tests showing the second-order accuracy and highlighted the monotonicity by comparing them with the classical DDFV scheme (see [58]). Finally, we presented in Chapter 3 an extension of the schemes proposed in Chapter 2 to arbitrary order, which is a natural outcome of Chapter 1. We therefore have two positivity-preserving arbitrary-order schemes on 2D deformed meshes, based on the same principle but differing in the way they handle node values. For both methods, we have addressed the case of Dirichlet and Neumann boundary conditions. However, we have only implemented the positive diamond scheme of arbitrary order. Numerical tests have been proposed to confirm the order of convergence and monotonicity. We observe a superconvergence for odd order. For both methods, high order is obtained using Taylor expansion at the desired order. For the methods that require a polynomial reconstruction to evaluate some quantites, the choice of the stencil has been made with the aim of accurately reconstructing polynomials of sufficiently large degrees. The higher the order, the larger the stencil needs to be. All these methods have been adapted to the case of a discontinuous and/or tensor-valued diffusion coefficient. It is important to note that high order is achieved even in the case of a discontinuous diffusion coefficient, provided that the faces coincide with the discontinuities. In this case, a flux is considered on each side of the face, taking care not to cross the discontinuity, especially when building the stencil. This result could be of particular interest in the context of coupling with Lagrangian hydrodynamics.

We now list some natural perspectives to this work :

- ▷ First, concerning the theory
  - ◇ It would be interesting to extend our 1D convergence proofs to Multi-D.
- ▷ Second, improving the effectiveness of the scheme could be considered
  - ◇ It seems that the scheme performs better for odd orders. We should try to explain why.
  - ◇ We could investigate the effect of a weighted reconstruction for polynomials (see [89]).
  - ◇ We could work on the condition number of the matrix or on the solver in order to improve the resolution of the linear system.
  - ◇ Numerical performance could be improved. Indeed, the convergence of the fixed-point algorithm is not guaranteed and may be very slow in 2D. This is exibited in particular

in test cases where the classical DDFV scheme obtains negative solutions. Techniques for accelerating this fixed point could be explored, such as Anderson acceleration (see [2, 91]) or the ϵ-algorithm (see [13, 14]). This particular point has been studied by Clément Vincent during his master's intership (see [100]). He obtained an acceleration of the convergence up to 46%.

- Last, some possible extensions of the schemes may be addressed
  - It would be interesting to implement and test the monotonic DDFV scheme of arbitrary order in order to compare it with the monotonic diamond scheme. As one has seen in numerical tests in Chapter 2, the monotonic diamond scheme seems more diffusive (in the radial direction) than the monotonic DDFV scheme. This could be explained by the stencil used for the polynomial reconstruction required for the monotonic diamond scheme in constrast with monotonic DDFV. It would be interesting to see whether this is also the case for arbitrary order since, in such a case, polynomial reconstruction is required for both schemes.
  - Another natural follow-up to this work would be to consider the non-stationary case and perform a high-order time discretization. In particular, we could consider some positivity-preserving variant of Runge Kutta methods (see [45, 54]).
  - Moreover, some applications could involve a diffusion coefficient depending on the unknown. It would therefore be appropriate to adapt these schemes and implement them in the non-linear case.
  - Besides, these methods could be coupled into a high-order hydrodynamics code.
  - Finally, this work could be extended to 3D. Extension to 3D tetrahedral meshes does not seem to rise any difficulties. However, extending this method to any mesh, for example hexahedral meshes, which could have non-planar faces, is far more complicated, in particular for integration. Indeed, integrating a polynomial exactly on a non-planar face is not straightforward.

Appendices

Appendix A

Appendix of the Introduction

A.1 Proof of the Theorem 1

Proof. A variational formulation of (4) gives

$$\int_\Omega \kappa \boldsymbol{\nabla} u \boldsymbol{\nabla} v + \int_\Omega uv = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega). \tag{A.1}$$

Let us pose $G \in C^1(\mathbb{R})$ such that

$$\begin{cases} G' > 0 \ \text{ on } \mathbb{R}^+, \\ G = 0 \ \text{ on } \mathbb{R}^-, \end{cases} \tag{A.2}$$

and $K = \max\left(\max_\Omega(g), \max_\Omega(f)\right)$, we suppose $K < \infty$.

We can choose $v = G(u - K) \in H_0^1(\Omega)$ because

$$v(\mathbf{x}) = G(u(\mathbf{x}) - K) = G(g(\mathbf{x}) - K), \forall \mathbf{x} \in \partial\Omega, \tag{A.3}$$

and

$$g(\mathbf{x}) - K \leq 0, \forall \mathbf{x} \in \partial\Omega,$$

so $v(\mathbf{x}) = 0$, $\forall \mathbf{x} \in \partial\Omega$.

Then (A.1) gives

$$\int_\Omega \kappa \boldsymbol{\nabla} u \left(G(u - K)\right)' + \int_\Omega uG(u - K) = \int_\Omega fG(u - K),$$

that is to say

$$\int_\Omega \kappa \left(\boldsymbol{\nabla} u \cdot \boldsymbol{\nabla} u\right) G'(u - K) + \int_\Omega (u - K)G(u - K) = \int_\Omega (f - K)G(u - K),$$

Since $G(u - K) \leq 0$ and $f - K \leq 0$, the right hand side is non positive. The first term of the left hand side is non negative, so $(u - K)G(u - K) \leq 0$. But we also have $(u - K)G(u - K) \leq 0$ since if $u - K \in \mathbb{R}^-$, then $(u - K)G(u - K) = 0$ and if $u - K \in \mathbb{R}^+$, then $G(u - K) \leq 0$. Thus, $(u - K)G(u - K) = 0$.

So, we have $u = K$ or $G(u - K) = 0$, that is to say $u \leq K$. □

A.2 Formulation with the particle derivative of the Euler equations with thermal conduction

While the Eulerian description describes the velocity field at a given instant, the formulation with the particle derivative describes the trajectories followed by the particles over time. This formulation therefore allows the material to be tracked as it moves.

To do this, we will perform a few algebraic manipulations to obtain the formulation with the particle derivative of Euler's equations with heat conduction.

Let us introduce the specific volume $\tau = \dfrac{1}{\rho} = \rho^{-1}$. Thus, we have $\tau\rho = 1$.

▷ Let us start with the conservation of mass equation.

We start with the following equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0. \tag{A.4}$$

By multiplying equation (A.4) by τ, we obtain

$$\tau \frac{\partial \rho}{\partial t} + \tau \nabla \cdot (\rho \mathbf{u}) = 0.$$

We know that $\tau\rho = 1$, which implies, in the case of ρ (and so τ) is regular (derivable), that $\dfrac{\partial}{\partial t}(\tau\rho) = 0$.

Besides, we have the following relation

$$\frac{\partial}{\partial t}(\tau\rho) = \tau \frac{\partial \rho}{\partial t} + \rho \frac{\partial \tau}{\partial t},$$

which gives

$$\tau \frac{\partial \rho}{\partial t} + \rho \frac{\partial \tau}{\partial t} = 0 \quad \Longleftrightarrow \quad \tau \frac{\partial \rho}{\partial t} = -\rho \frac{\partial \tau}{\partial t}.$$

Thus, the equation becomes

$$-\rho \frac{\partial \tau}{\partial t} + \tau \nabla \cdot (\rho \mathbf{u}) = 0.$$

In the case of regular solutions, we have the following relation

$$\nabla \cdot (\rho \mathbf{u}) = \rho \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \nabla \rho.$$

By applying this relation, we have

$$-\rho \frac{\partial \tau}{\partial t} + \tau\rho \nabla \cdot \mathbf{u} + \mathbf{u} \cdot \tau \nabla \rho = 0.$$

Using the fact that $\tau\rho = 1$, this implies, in the regular case, that $\nabla(\rho\tau) = 0$, that is to say

$$\tau \nabla \rho + \rho \nabla \tau = 0 \quad \Longleftrightarrow \quad \tau \nabla \rho = -\rho \nabla \tau.$$

So, we have

$$-\rho \frac{\partial \tau}{\partial t} + \tau\rho \nabla \cdot \mathbf{u} - \mathbf{u} \cdot \rho \nabla \tau = 0.$$

After factorization

$$-\rho \left(\frac{\partial \tau}{\partial t} + \mathbf{u} \cdot \nabla \tau \right) + \tau\rho \nabla \cdot \mathbf{u} = 0.$$

Using $\tau\rho = 1$, this gives

$$\rho \left(\frac{\partial \tau}{\partial t} + \mathbf{u} \cdot \nabla \tau \right) - \nabla \cdot \mathbf{u} = 0.$$

Let us introduce the operator "particle derivative" or "total derivative" : $D_t = \partial_t + \mathbf{u} \cdot \nabla$, which gives

$$\rho D_t \tau - \nabla \cdot \mathbf{u} = 0.$$

▷ Let us now turn to the conservation equation for momentum.

We start with the following equation

$$\frac{\partial}{\partial t}(\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) + \nabla p = 0.$$

Using the formula for the derivative of a product in the regular case

$$\frac{\partial}{\partial t}(\rho \mathbf{u}) = \rho \frac{\partial \mathbf{u}}{\partial t} + \mathbf{u} \frac{\partial \rho}{\partial t},$$

gives

$$\rho\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u}) + \nabla p = 0.$$

Then, using the formula for the divergence of a tensor product

$$\nabla\cdot(\rho\mathbf{u}\otimes\mathbf{u}) = (\rho\mathbf{u})\nabla\mathbf{u} + \nabla\cdot(\rho\mathbf{u})\,\mathbf{u},$$

we obtain

$$\rho\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\frac{\partial \rho}{\partial t} + \rho\mathbf{u}\nabla\mathbf{u} + \nabla\cdot(\rho\mathbf{u})\mathbf{u} + \nabla p = \mathbf{0}.$$

Rearranging terms to bring out known formules, we have

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\nabla\mathbf{u}\right) + \left(\mathbf{u}\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\mathbf{u}\right) + \nabla p = \mathbf{0}.$$

The second term reveals the formula for conservation of mass, which is zero

$$\mathbf{u}\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\mathbf{u} = \mathbf{u}\left(\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\right) = \mathbf{0}.$$

The first term allows us to display the D_t operator

$$\rho\left(\frac{\partial \mathbf{u}}{\partial t} + \mathbf{u}\nabla\mathbf{u}\right) = \rho D_t\mathbf{u}.$$

The equation becomes

$$\rho D_t\mathbf{u} + \nabla p = \mathbf{0}.$$

▷ Let us finish with the conservation equation for total energy with heat conduction.

We start with the following equation

$$\frac{\partial}{\partial t}(\rho E) + \nabla\cdot(\rho E\mathbf{u}) + \nabla\cdot(p\mathbf{u}) = \nabla\cdot\kappa\nabla T.$$

Using the formula for the derivative of a product, for regular solutions

$$\frac{\partial}{\partial t}(\rho E) = \rho\frac{\partial E}{\partial t} + E\frac{\partial \rho}{\partial t},$$

we have

$$\rho\frac{\partial E}{\partial t} + E\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho E\mathbf{u}) + \nabla\cdot(p\mathbf{u}) = \nabla\cdot\kappa\nabla T.$$

Then, using the formula for the divergence of a product

$$\nabla\cdot(\rho E\mathbf{u}) = E\nabla\cdot(\rho\mathbf{u}) + \rho\mathbf{u}\cdot\nabla E,$$

we obtain

$$\rho\frac{\partial E}{\partial t} + E\frac{\partial \rho}{\partial t} + E\nabla\cdot(\rho\mathbf{u}) + \rho\mathbf{u}\cdot\nabla E + \nabla\cdot(p\mathbf{u}) = \nabla\cdot\kappa\nabla T.$$

Rearranging the terms, we have

$$\rho\left(\frac{\partial E}{\partial t} + \mathbf{u}\cdot\nabla E\right) + E\left(\frac{\partial \rho}{\partial t} + \nabla\cdot(\rho\mathbf{u})\right) + \nabla\cdot(p\mathbf{u}) = \nabla\cdot\kappa\nabla T.$$

The first term allows us to display the D_t operator

$$\rho\left(\frac{\partial E}{\partial t} + \mathbf{u}\cdot\nabla E\right) = \rho D_t E.$$

The second term reveals the formula for conservation of mass

$$E\left(\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u})\right) = E \times 0 = 0.$$

The equation becomes

$$\rho D_t E + \nabla \cdot (p\mathbf{u}) = \nabla \cdot \kappa \nabla T.$$

So, we obtain the formulation with the particle derivative for the Euler equations

$$\begin{cases} \rho D_t \tau - \nabla \cdot \mathbf{u} = & 0, \\ \rho D_t \mathbf{u} + \nabla p = & \mathbf{0}, \\ \rho D_t E + \nabla \cdot (p\mathbf{u}) = & \nabla \cdot \kappa \nabla T. \end{cases} \tag{A.5}$$

In the case of regular solutions, this formulation is equivalent to the system (1)

A.3 Details of computations for the Equation (2)

The entropy η can be defined by the Gibbs law

$$T d\eta = p d\tau + de,$$

where $\tau = \dfrac{1}{\rho}$ is the specific volume. This implies

$$T \frac{d}{dt}\eta = p \frac{d}{dt}\tau + \frac{d}{dt}e. \tag{A.6}$$

Introducing the notation $D_t = \dfrac{d}{dt} = \partial_t + \mathbf{u} \cdot \nabla$, we have

$$T D_t \eta = p D_t \tau + D_t e,$$

that is to say, multiplying by ρ,

$$\rho T D_t \eta = \rho p D_t \tau + \rho D_t e.$$

The Clausius Duhem inequality writes

$$\rho T D_t \eta - T \nabla \cdot \left(\frac{\kappa \nabla T}{T}\right) \geq 0. \tag{A.7}$$

Equation (A.6) gives

$$\rho T D_t \eta = \rho p D_t \tau + \rho D_t e.$$

Replacing e by its expression $E - \dfrac{\|\mathbf{u}\|^2}{2}$, we obtain

$$\rho T D_t \eta = \rho p D_t \tau + \rho D_t (E) - \rho D_t \left(\frac{\|\mathbf{u}\|^2}{2}\right).$$

The first equation of (A.5) gives

$$\rho D_t \tau = \nabla \cdot \mathbf{u}.$$

which implies

$$\rho p D_t \tau = p \nabla \cdot \mathbf{u}.$$

The third equation of (A.5) gives

$$\rho D_t E = -\nabla \cdot (p\mathbf{u}) + \nabla \cdot \kappa \nabla T.$$

For the last term, we use the following expression

$$D_t\left(\frac{\|\mathbf{u}\|^2}{2}\right) = \partial_t\left(\frac{\|\mathbf{u}\|^2}{2}\right) + \mathbf{u}\cdot\nabla\left(\frac{\|\mathbf{u}\|^2}{2}\right) = \frac{1}{2}2\mathbf{u}\partial_t(\mathbf{u}) + \left(\frac{\mathbf{u}}{2}\right)2\mathbf{u}\nabla\mathbf{u} = \mathbf{u}\partial_t\mathbf{u} + \|\mathbf{u}\|^2\nabla\mathbf{u} = \mathbf{u}D_t(\mathbf{u}).$$

Thus, we have

$$-\rho D_t\left(\frac{\|\mathbf{u}\|^2}{2}\right) = -\rho\mathbf{u}D_t(\mathbf{u}) = -\mathbf{u}(\rho D_t(\mathbf{u})).$$

Then, the second equation of (A.5) gives

$$-\mathbf{u}(\rho D_t(\mathbf{u})) = \mathbf{u}\nabla p.$$

Finally, we obtain

$$\rho T D_t(\eta) = p\nabla\cdot\mathbf{u} - \nabla\cdot(p\mathbf{u}) + \nabla\cdot\kappa\nabla T + \mathbf{u}\nabla p.$$

Besides, in a regular case,

$$\nabla\cdot(p\mathbf{u}) = p\nabla\cdot\mathbf{u} + \mathbf{u}\nabla p.$$

Then, we have

$$\rho T D_t\eta = \nabla\cdot\kappa\nabla T.$$

Inequality (A.7) thus becomes

$$\nabla\cdot\kappa\nabla T - T\nabla\cdot\left(\frac{\kappa\nabla T}{T}\right) \geq 0,$$

Using the formula for the divergence of a product on both terms, we have

$$\kappa\nabla\cdot(\nabla T) + \nabla T\nabla\kappa - \kappa\nabla\cdot(\nabla T) - T\nabla T\nabla\left(\frac{\kappa}{T}\right) \geq 0,$$

Reusing the same formula on the last term, we obtain

$$\nabla T\nabla\kappa - T\nabla T\left[\frac{1}{T}\nabla\kappa + \kappa\nabla\left(\frac{1}{T}\right)\right] \geq 0,$$

that is to say

$$\nabla T\nabla\kappa - \nabla T\nabla\kappa - T\kappa\nabla T\nabla\left(\frac{1}{T}\right) \geq 0,$$

Using the formula of the derivative of an inverse, this gives

$$-T\kappa\nabla T\left(-\frac{\nabla T}{T^2}\right) \geq 0,$$

that is to say

$$\frac{\kappa}{T}(\nabla T\cdot\nabla T) \geq 0.$$

Appendix B

Appendix of the Chapter 1

B.1 Dirichlet boundary conditions

In this appendix we give the details of the computations of the fluxes in the case of Dirichlet boundary condition given in Section 1.2.5.1, that is to say considering problem (1.3) with $\beta = 1$, $\gamma = 0$.

Consider first the right boundary of the domain. The adaptation to the left boundary is straightforward. The k-th order Taylor expansion in the neighborhood of $x_{n+\frac{1}{2}}$ gives

$$\forall x, \quad \bar{u}(x) = \bar{u}(x_{n+\frac{1}{2}}) + \sum_{\ell=1}^{k} \frac{(x - x_{n+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left(\left(x - x_{n+\frac{1}{2}}\right)^{k+1}\right).$$

Here again, we integrate this expression in order to use mean values. This gives

$$\frac{1}{h_n}\int_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} \bar{u}(x)dx = \bar{u}(x_{n+\frac{1}{2}}) + \frac{1}{h_n}\sum_{\ell=1}^{k}\int_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} \frac{(x - x_{n+\frac{1}{2}})^\ell}{\ell!} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}})dx + \mathcal{O}\left(h_n^{k+1}\right),$$

that is to say

$$\bar{u}_n = \bar{u}(x_{n+\frac{1}{2}}) + \frac{1}{h_n}\sum_{\ell=1}^{k}\left[\frac{(x - x_{n+\frac{1}{2}})^{\ell+1}}{(\ell+1)!}\right]_{x_{n-\frac{1}{2}}}^{x_{n+\frac{1}{2}}} \frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left(h_n^{k+1}\right),$$

from which we obtain

$$\frac{d\bar{u}}{dx}(x_{n+\frac{1}{2}}) = \frac{2}{h_n}\left(\bar{u}(x_{n+\frac{1}{2}}) - \bar{u}_n\right) + 2\sum_{\ell=2}^{k}\frac{(-1)^\ell h_n^{\ell-1}}{(\ell+1)!}\frac{d^\ell \bar{u}}{dx^\ell}(x_{n+\frac{1}{2}}) + \mathcal{O}\left(h_n^k\right).$$

The numerical flux is obtained by approximating the derivatives of $\bar{u}$ at $x_{n+\frac{1}{2}}$ using a polynomial reconstruction of the solution

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left(\frac{2}{h_n}\left(u_{n+\frac{1}{2}} - u_n\right) + r_{n+\frac{1}{2}}(\mathbf{u})\right).$$

The trick of Section 1.2.3 can be applied to ensure monotonicity, that is, in the non-symmetric version

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right)u_{n+\frac{1}{2}} - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right)u_n\right],$$

and, in the symmetric version

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}}\left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}}}\right)u_{n+\frac{1}{2}} - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n}\right)u_n\right], \quad \text{(B.1)}$$

with

$$s_{n+\frac{1}{2}}(\mathbf{u}) = \frac{u_n r^+_{n+\frac{1}{2}}(\mathbf{u}) - u_{n+\frac{1}{2}} r^-_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}} - u_n}.$$

In order to preserve positivity, a condition similar to (1.19) must be satisfied for the symmetric version of the scheme

$$\frac{\frac{2}{h_n}\left(u_{n+\frac{1}{2}} - u_n\right) + r_{n+\frac{1}{2}}(\mathbf{u})}{u_{n+\frac{1}{2}} - u_n} \geq 0,$$

that is to say that $u_{n+\frac{1}{2}} - u_n$ and $\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u})$ must have the same sign. As in Section 1.2.4, this condition seems natural because if $\dfrac{d\bar{u}}{dx}\left(x_{n+\frac{1}{2}}\right) \geq 0$ (resp. ≤ 0), then $\bar{u}$ is locally increasing (resp. decreasing) so

$\bar{u}_{n+\frac{1}{2}} \geq \bar{u}_n$ (resp. $\bar{u}_{n+\frac{1}{2}} \leq \bar{u}_n$).

Applying the boundary condition, (B.1) becomes

$$\mathcal{F}_{n+\frac{1}{2}}(\mathbf{u}) = \kappa_{n+\frac{1}{2}} \left[\left(\frac{2}{h_n} + \frac{r^+_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{g(x_{n+\frac{1}{2}})} \right) g(x_{n+\frac{1}{2}}) - \left(\frac{2}{h_n} + \frac{r^-_{n+\frac{1}{2}}(\mathbf{u}) + s_{n+\frac{1}{2}}(\mathbf{u})}{u_n} \right) u_n \right]. \quad \text{(B.2)}$$

For the left boundary we obtain similarly

$$\mathcal{F}_{\frac{1}{2}}(\mathbf{u}) = \kappa_{\frac{1}{2}} \left[\left(\frac{2}{h_1} + \frac{r^+_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{u_1} \right) u_1 - \left(\frac{2}{h_1} + \frac{r^-_{\frac{1}{2}}(\mathbf{u}) + s_{\frac{1}{2}}(\mathbf{u})}{g(x_{\frac{1}{2}})} \right) g(x_{\frac{1}{2}}) \right]. \quad \text{(B.3)}$$

B.2 Exactness for polynomials of degree k

In this appendix, we give the proof that our flux is exact for polynomials of degree k.

To simplify the calculation let us consider a polynomial of degree k centered on $x_{i+\frac{1}{2}}$ as an exact solution in order to demonstrate that the approximation of $\dfrac{d\bar{u}}{dx}(x_{i+\frac{1}{2}})$ is exact for polynomials of degree k. For

$$\bar{u}(x) = \sum_{p=0}^{k} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^p,$$

we obtain

$$\frac{d^\ell \bar{u}}{dx^\ell}(x) = \sum_{p=\ell}^{k} \frac{p!}{(p-\ell)!} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^{p-\ell},$$

hence

$$\frac{d^\ell \bar{u}}{dx^\ell}(x_{i+\frac{1}{2}}) = \ell! a_{i+\frac{1}{2},\ell}.$$

Besides, mean values were used to estimate the values of u at the centers of the cells, so

$$\bar{u}_{i+1} = \frac{1}{h_{i+1}} \int_{x_{i+\frac{1}{2}}}^{x_{i+\frac{3}{2}}} \sum_{p=0}^{k} a_{i+\frac{1}{2},p}(x - x_{i+\frac{1}{2}})^p = \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p}{p+1},$$

and

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} \sum_{p=0}^{k} a_{i+\frac{1}{2},n}(x - x_{i+\frac{1}{2}})^p = \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{(-1)^p h_i^p}{p+1}.$$

The flux is

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}} \left[\bar{u}_{i+1} - \bar{u}_i - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{(p+1)!} \frac{d^p P}{dx^p}(x_{i+\frac{1}{2}}) \right],$$

where P is an interpolation polynomial of $\bar{u}$. Besides, $P = \bar{u}$ in that case since $\bar{u}$ is a polynomial of degree k and polynomials of degree k are invariant under the polynomial reconstruction. The flux becomes

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \frac{\kappa_{i+\frac{1}{2}}}{h_{i+\frac{1}{2}}} \left(\left[\sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p}{p+1} - \sum_{p=0}^{k} a_{i+\frac{1}{2},p} \frac{(-1)^p h_i^p}{p+1} \right] - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{(p+1)!} p! a_{i+\frac{1}{2},p} \right),$$

that is to say

$$\mathcal{F}_{i+\frac{1}{2}}(\bar{\mathbf{u}}) = \kappa_{i+\frac{1}{2}} \left(a_{i+\frac{1}{2},1} + \sum_{p=2}^{k} a_{i+\frac{1}{2},p} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{h_{i+\frac{1}{2}}(p+1)} - \sum_{p=2}^{k} \frac{h_{i+1}^p + (-1)^{p+1} h_i^p}{h_{i+\frac{1}{2}}(p+1)} a_{i+\frac{1}{2},p} \right) = \kappa_{i+\frac{1}{2}} a_{i+\frac{1}{2},1}.$$

The flux is exact for polynomials of degree k.

Appendix C

Appendix of the Chapter 2

C.1 Computation of the coefficients $\alpha_{i\ell,i}$, $\alpha_{i\ell,j}$, $\beta_{i\ell,i}$ and $\beta_{i\ell,j}$

In this appendix, we give the details of the computations of the primal coefficents $\alpha_{i\ell,i}$, $\alpha_{i\ell,j}$, $\beta_{i\ell,i}$ and $\beta_{i\ell,j}$ given by Equations (2.4) and (2.5).

First, we have

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,i}\frac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}.$$

Since $\mathbf{n}_{i\ell}$ is orthogonal to the edge ℓ the vertices of which are r and s, by taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain on the one hand

$$1 = \alpha_{i\ell,i}\frac{(\mathbf{x}_\ell - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|},$$

that is to say

$$\alpha_{i\ell,i} = \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{(\mathbf{x}_\ell - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}.$$

On the other hand, we also have

$$\mathbf{n}_{i\ell}\cdot(\mathbf{x}_\ell - \mathbf{x}_i)^\perp = \beta_{i\ell,i}\frac{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_\ell - \mathbf{x}_i)^\perp}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say

$$\beta_{i\ell,i} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_\ell - \mathbf{x}_i)^\perp}{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_\ell - \mathbf{x}_i)^\perp}.$$

Second, we have

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,j}\frac{\mathbf{x}_j - \mathbf{x}_\ell}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}.$$

Since $\mathbf{n}_{i\ell}$ is orthogonal to the edge ℓ the vertices of which are r and s, by taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain on the one hand

$$1 = \alpha_{i\ell,j}\frac{(\mathbf{x}_j - \mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|},$$

that is to say

$$\alpha_{i\ell,j} = \frac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{(\mathbf{x}_j - \mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell}}.$$

On the other hand, we also have

$$\mathbf{n}_{i\ell}\cdot(\mathbf{x}_j - \mathbf{x}_\ell)^\perp = \beta_{i\ell,j}\frac{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_j - \mathbf{x}_\ell)^\perp}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say

$$\beta_{i\ell,j} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_j - \mathbf{x}_\ell)^\perp}{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_j - \mathbf{x}_\ell)^\perp}.$$

C.2 Computation of the coefficients $\alpha_{r\tilde{\ell}}$ and $\beta_{r\tilde{\ell}}$

In this appendix, we give the details of the computations of the dual coefficents $\alpha_{r\tilde{\ell}}$ and $\beta_{r\tilde{\ell}}$ given by Equation (2.18).

We have

$$\mathbf{n}_{r\tilde{\ell}} = \alpha_{r\tilde{\ell}} \frac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{r\tilde{\ell}} \frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}.$$

Since $\mathbf{n}_{r\tilde{\ell}}$ is orthogonal to the edge $\tilde{\ell}$ the two vertices of which are i and ℓ, by taking the scalar product with $\mathbf{n}_{r\tilde{\ell}}$, we obtain on the one hand

$$1 = \beta_{r\tilde{\ell}} \frac{(\mathbf{x}_s - \mathbf{x}_r) \cdot \mathbf{n}_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say

$$\beta_{r\tilde{\ell}} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\|}{(\mathbf{x}_s - \mathbf{x}_r) \cdot \mathbf{n}_{r\tilde{\ell}}}.$$

On the other hand, we also have

$$\mathbf{n}_{r\tilde{\ell}} \cdot (\mathbf{x}_s - \mathbf{x}_r)^\perp = \alpha_{r\tilde{\ell}} \frac{(\mathbf{x}_\ell - \mathbf{x}_i) \cdot (\mathbf{x}_s - \mathbf{x}_r)^\perp}{\|\mathbf{x}_\ell - \mathbf{x}_i\|},$$

that is to say

$$\alpha_{r\tilde{\ell}} = \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\| \mathbf{n}_{r\tilde{\ell}} \cdot (\mathbf{x}_s - \mathbf{x}_r)^\perp}{(\mathbf{x}_\ell - \mathbf{x}_i) \cdot (\mathbf{x}_s - \mathbf{x}_r)^\perp}.$$

C.3 Exactness for polynomials of degree 1

In this appendix, we give the proof that our fluxes are exact for polynomials of degree 1. The proof is first given for the primal flux and then for the dual flux.

C.3.1 Primal flux

We will show that our approximation of $\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell)$ is exact for polynomials of degree 1.

First, let us assume that

$$\bar{u}(\mathbf{x}) = a_{0,0} + a_{1,0} x + a_{0,1} y.$$

Then, we have

$$\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot \mathbf{n}_{i\ell} = \frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \left(a_{1,0}(x_j - x_i) + a_{0,1}(y_j - y_i) \right) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left(a_{1,0}(x_s - x_r) + a_{0,1}(y_s - y_r) \right).$$

For the primal mesh, the flux is defined by

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell| \left[\left(\frac{\kappa_i \kappa_j \alpha_{i\ell,j} \alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\| \kappa_i \alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\| \kappa_j \alpha_{i\ell,j}} \right) (\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_i)) \right.$$
$$\left. + \left(\frac{\kappa_i \kappa_j \left(\alpha_{i\ell,i} \beta_{i\ell,j} \|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j} \beta_{i\ell,i} \|\mathbf{x}_\ell - \mathbf{x}_i\| \right)}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\| \kappa_i \alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\| \kappa_j \alpha_{i\ell,j} \right)} \right) (\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) \right].$$

Considering κ continuous, the flux becomes

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\kappa_\ell \left[\left(\frac{\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,j}} \right) (\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_i)) \right.$$
$$\left. + \left(\frac{(\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|)}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,j}\right)} \right) (\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) \right].$$

Besides, we have

$$\begin{cases} \mathbf{n}_{i\ell} = \alpha_{i\ell}\dfrac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|} + \beta_{i\ell}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \mathbf{n}_{i\ell} = \alpha_{i\ell,j}\dfrac{\mathbf{x}_j - \mathbf{x}_\ell}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,j}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \mathbf{n}_{i\ell} = \alpha_{i\ell,i}\dfrac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,i}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}. \end{cases} \tag{C.1}$$

By taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain

$$\begin{cases} \dfrac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} = \dfrac{1}{(\mathbf{x}_j - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}, \\ (\mathbf{x}_j - \mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell} = \dfrac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,j}}, \\ (\mathbf{x}_\ell - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell} = \dfrac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,i}}, \end{cases} \tag{C.2}$$

that is to say

$$\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} = \frac{1}{\frac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,j}} + \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,i}}} = \frac{\alpha_{i\ell,j}\alpha_{i\ell,i}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,j}}.$$

Then, by taking the scalar product with $\mathbf{x}_s - \mathbf{x}_r$ on (C.1), we have

$$\begin{cases} \beta_{i\ell} = -\alpha_{i\ell}\dfrac{(\mathbf{x}_j - \mathbf{x}_i)\cdot(\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_j - \mathbf{x}_i\|}, \\ (\mathbf{x}_j - \mathbf{x}_\ell)\cdot(\mathbf{x}_s - \mathbf{x}_r) = -\dfrac{\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,j}}, \\ (\mathbf{x}_\ell - \mathbf{x}_i)\cdot(\mathbf{x}_s - \mathbf{x}_r) = -\dfrac{\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,i}}, \end{cases}$$

that is to say

$$\frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} = \frac{\frac{\alpha_{i\ell,i}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}\frac{\beta_{i\ell,j}}{\|\mathbf{x}_s - \mathbf{x}_r\|} + \frac{\alpha_{i\ell,j}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|}\frac{\beta_{i\ell,i}}{\|\mathbf{x}_s - \mathbf{x}_r\|}}{\frac{\alpha_{i\ell,i}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \frac{\alpha_{i\ell,j}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|}} = \frac{\alpha_{i\ell,i}\beta_{i\ell,j}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,j}\beta_{i\ell,i}\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\|\mathbf{x}_s - \mathbf{x}_r\| \left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,i} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,j}\right)}.$$

In the case of a contiuous κ, the flux is

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\kappa_\ell \left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}(\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_i)) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) \right].$$

Consider first the interpolation method. Since $\bar{u}$ is an affine function, the interpolation polynomial P is exactly equal to $\bar{u}$. Therefore, the node values $P(\mathbf{x}_r)$ are equal to $\bar{u}(\mathbf{x}_r)$. Second, in the DDFV method, the vertex values are degrees of freedom. In both cases, using the definition of $\bar{u}$, the flux becomes

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\kappa_\ell \left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}(a_{0,0} + a_{1,0}x_j + a_{0,1}y_j) - (a_{0,0} + a_{1,0}x_i + a_{0,1}y_i)) \right.$$
$$\left. + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(a_{0,0} + a_{1,0}x_s + a_{0,1}y_s - (a_{0,0} + a_{1,0}x_r + a_{0,1}y_r)) \right],$$

that is to say

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\kappa_\ell \left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}(a_{1,0}(x_j - x_i) + a_{0,1}(y_j - y_i)) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(a_{1,0}(x_s - x_r) + a_{0,1}(y_s - y_r)) \right],$$

As a conclusion, the primal flux is exact for polynomials of degree 1.

C.3.2 Dual flux

We will show that our approximation of $\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})$ is exact for polynomials of degree 1.

First, let us assume that

$$\bar{u}(\mathbf{x}) = a_{0,0} + a_{1,0}x + a_{0,1}y.$$

Then, we have

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}}) \cdot \mathbf{n}_{r\tilde{\ell}} = \frac{\alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}\left(a_{1,0}(x_\ell - x_i) + a_{0,1}(y_\ell - y_i)\right) + \frac{\beta_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|}\left(a_{1,0}(x_s - x_r) + a_{0,1}(y_s - y_r)\right).$$

For the dual mesh, the flux is defined by

$$\mathcal{F}_{\tilde{\ell}}(\bar{\mathbf{u}}) = |\tilde{\ell}|\kappa_{\tilde{\ell}}\left(\frac{\beta_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)) + \frac{\alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}(\bar{u}(\mathbf{x}_\ell) - \bar{u}(\mathbf{x}_i)) \right).$$

Using the definition of $\bar{u}$, the flux becomes

$$\mathcal{F}_{\tilde{\ell}}(\bar{\mathbf{u}}) = |\tilde{\ell}|\kappa_{\tilde{\ell}}\left(\frac{\beta_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(a_{0,0} + a_{1,0}x_s + a_{0,1}y_s - (a_{0,0} + a_{1,0}x_r + a_{0,1}y_r) \right.$$
$$\left. + \frac{\alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}(a_{0,0} + a_{1,0}x_\ell + a_{0,1}y_\ell) - (a_{0,0} + a_{1,0}x_i + a_{0,1}y_i)) \right).$$

that is to say

$$\mathcal{F}_{\tilde{\ell}}(\bar{\mathbf{u}}) = |\tilde{\ell}|\kappa_{\tilde{\ell}}\left(\frac{\beta_{r\tilde{\ell}}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(a_{1,0}(x_s - x_r) + a_{0,1}(y_s - y_r)) + \frac{\alpha_{r\tilde{\ell}}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}(a_{1,0}(x_\ell - x_i) + a_{0,1}(y_\ell - y_i)) \right).$$

As a conclusion, the flux is exact for polynomials of degree 1.

C.4 Proof of Proposition 2.6.1

Proof. The sum can be rewritten by interverting the sum over the cells and the sum over the faces. Besides, the sum can be separated into boundary terms and non-boundary-terms

$$\sum_{i=1}^{n}\left(-\sum_{\ell\in i}\mathcal{F}_\ell(\mathbf{u}) \right) = -\sum_{\ell\in\Gamma}\mathcal{F}_\ell(\mathbf{u}) - \sum_{\ell\in\mathring{\Omega}}\left(\mathcal{F}_{\ell,i}(\mathbf{u}) + \mathcal{F}_{\ell,j}(\mathbf{u})\right),$$

where ℓ is the face shared by the cells i and j, and with

$$\begin{cases} \mathcal{F}_{\ell,i}(\mathbf{u}) = \gamma_\ell(u_j - u_i) + r_{\ell,i}(\mathbf{u}), \\ \mathcal{F}_{\ell,j}(\mathbf{u}) = \gamma_\ell(u_i - u_j) + r_{\ell,j}(\mathbf{u}), \end{cases}$$

with $r_{\ell,i}(\mathbf{u}) = -r_{\ell,j}(\mathbf{u})$. Then,

$$\mathcal{F}_{\ell,i}(\mathbf{u}) + \mathcal{F}_{\ell,j}(\mathbf{u}) = 0.$$

The homogeneous Neumann boundary condition means that the boundary terms are zero, which leads to

$$\sum_{i=1}^{n} \left(-\sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) \right) = 0,$$

that is to say

$$\sum_{i=1}^{n} V_i \lambda_i u_i = \sum_{i=1}^{n} V_i f_i.$$

The scheme is conservative. □

C.5 Proof of convergence for DDFV scheme

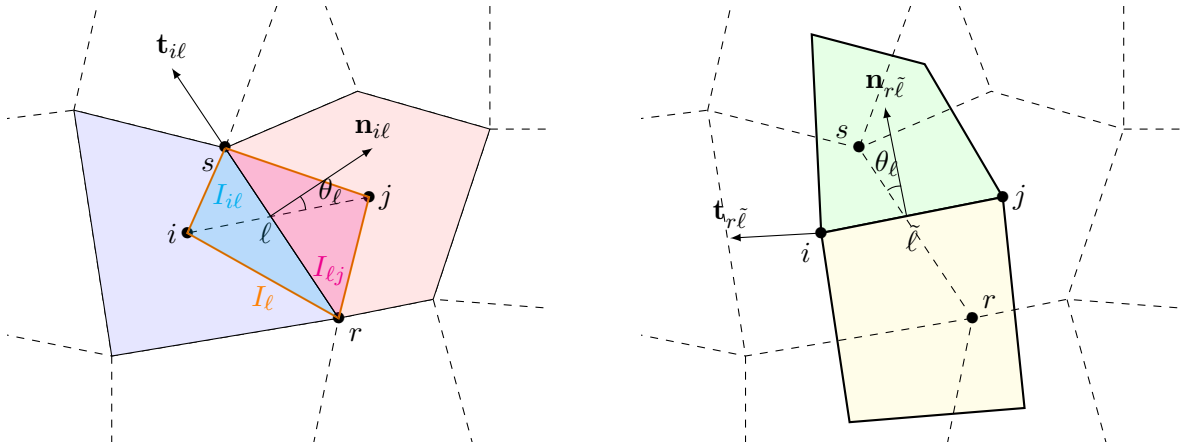


Fig. C.1 – Primal mesh (at the left) and dual mesh (at the right)

For simplicity we will restrict ourselves to the case $\kappa = 1$, $\lambda = 0$, $g = 0$ and $\Gamma_N = \emptyset$ in (2.1), that is,

$$\begin{cases} -\boldsymbol{\nabla} \cdot (\boldsymbol{\nabla} \bar{u}) = f & \text{in} \quad \Omega, \\ \bar{u} = g & \text{on} \quad \partial\Omega. \end{cases} \tag{C.3}$$

Suppose further that the dual mesh is made of cells obtained by joining the center of each primal cell with the center of each of its neightbors and with the middle of its boundary faces. In this case we observe that the dual boundary $\tilde{\ell} = r \cap s$ coincides with the segment $\mathbf{x}_i\mathbf{x}_j$. Denote by $\mathbf{n}_{r\tilde{\ell}}$ the unit vector orthogonal to $\tilde{\ell}$ directed from the dual cell r to s, $\mathbf{N}_{r\tilde{\ell}} = \|\mathbf{x}_i - \mathbf{x}_j\| \mathbf{n}_{r\tilde{\ell}}$, and by θ_ℓ the angle between vectors $-\mathbf{n}_{r\tilde{\ell}}^{\perp}$ (that is to say $\mathbf{x}_i\mathbf{x}_j$) and $\mathbf{n}_{i\ell}$ (see Figure C.1).

We define

$$h = \max_{\ell}(|\ell|, |\tilde{\ell}|).$$

Applying the method used in Sections 2.3 and 2.4.2, we have

$$\boldsymbol{\nabla} \bar{u}(\mathbf{x}_\ell) \cdot \mathbf{n}_{i\ell} = \frac{1}{\cos(\theta_\ell)} \frac{\bar{u}(\mathbf{x}_j) - \bar{u}(\mathbf{x}_i)}{\|\mathbf{x}_j - \mathbf{x}_i\|} + \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)} \frac{\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r)}{\|\mathbf{x}_s - \mathbf{x}_r\|} + \mathcal{O}(h),$$

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{n}_{r\tilde{\ell}} = \frac{1}{\cos(\theta_\ell)}\frac{\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)}{\|\mathbf{x}_s-\mathbf{x}_r\|} + \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}\frac{\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i)}{\|\mathbf{x}_j-\mathbf{x}_i\|} + \mathcal{O}(h).$$

This is equivalent to say that $\boldsymbol{\nabla}\bar{u}$ is approximated in the diamond cell I_ℓ using the Green-Gauss formula

$$\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell) = \frac{1}{V_\ell}\int_{I_\ell}\boldsymbol{\nabla}\bar{u} + \mathcal{O}(h) = \frac{1}{2}\frac{1}{V_\ell}\left(\mathbf{N}_{i\ell}(u_j-u_i) + \mathbf{N}_{r\tilde{\ell}}(u_s-u_r)\right) + \mathcal{O}(h).$$

The discretization of (C.3) with the DDFV scheme then writes

$$\begin{cases} -\frac{1}{2}\sum\limits_{\ell\in i,\ell\notin\partial\Omega}\frac{1}{V_\ell}\left(|\ell|^2(u_i-u_j)+|\ell||\tilde{\ell}|\mathbf{n}_{i\ell}\cdot\mathbf{n}_{r\tilde{\ell}}(u_r-u_s)\right)+ \\ -\frac{1}{2}\sum\limits_{\ell\in i,\ell\in\partial\Omega}\frac{1}{V_\ell}\left(|\ell|^2(u_i-u_\ell)+|\ell||\tilde{\ell}|\mathbf{n}_{i\ell}\cdot\mathbf{n}_{r\tilde{\ell}}(u_r-u_s)\right) = V_j f_j, \\ -\frac{1}{2}\sum\limits_{\ell\in r}\frac{1}{V_\ell}\left(|\ell||\tilde{\ell}|\mathbf{n}_{i\ell}\cdot\mathbf{n}_{r\tilde{\ell}}(u_i-u_j)+|\tilde{\ell}|^2(u_r-u_s)\right) = V_r f_r \qquad \mathbf{x}_r\notin\partial\Omega, \\ u_\ell = g(\mathbf{x}_\ell) \qquad \mathbf{x}_\ell\in\partial\Omega, \\ u_r = g(\mathbf{x}_r) \qquad \mathbf{x}_r\in\partial\Omega. \end{cases} \tag{C.4}$$

where I_ℓ is the diamond cell $\mathbf{x}_i\mathbf{x}_s\mathbf{x}_j\mathbf{x}_r$, V_ℓ is the surface of the diamond cell associated with the face ℓ. Recall that the dual edge $\mathbf{x}_i\mathbf{x}_j$ will be denoted by $\tilde{\ell}$.

The following proofs are inspired from the arguments of [48] for *admissible* meshes and from [3] for general meshes (see also [35], [109]). In the sequel we will assume that the exact solution $\bar{u}$ satisfies $\bar{u}\in W^{1,\infty}(\Omega)$.

C.5.1 Consistency of the fluxes

Let us denote by

1. $\bar{\mathcal{F}}_\ell$, $\bar{\mathcal{F}}_{\tilde{\ell}}$ the *exact* primal and dual fluxes

$$\bar{\mathcal{F}}_\ell = \int_\ell \boldsymbol{\nabla}\bar{u}\cdot\mathbf{n}_{i\ell}, \quad \bar{\mathcal{F}}_{\tilde{\ell}} = \int_{\tilde{\ell}}\boldsymbol{\nabla}\bar{u}\cdot\mathbf{n}_{r\tilde{\ell}},$$

2. $\mathcal{F}_\ell(\mathbf{u})$, $\mathcal{F}_{\tilde{\ell}}(\mathbf{u})$ the *approximated* primal and dual fluxes

$$\mathcal{F}_\ell(\mathbf{u}) = \frac{1}{2}\frac{1}{V_\ell}\left((u_j-u_i)\mathbf{N}_{i\ell}+(u_s-u_r)\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{i\ell},$$

$$\mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = \frac{1}{2}\frac{1}{V_\ell}\left((u_j-u_i)\mathbf{N}_{i\ell}+(u_s-u_r)\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{r\tilde{\ell}},$$

3. $\mathcal{F}_\ell(\bar{\mathbf{u}})$, $\mathcal{F}_{\tilde{\ell}}(\bar{\mathbf{u}})$ what we can call the *semi-approximated* primal and dual fluxes

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = \frac{1}{2}\frac{1}{V_\ell}\left((\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i))\mathbf{N}_{i\ell}+(\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{i\ell},$$

$$\mathcal{F}_{\tilde{\ell}}(\bar{\mathbf{u}}) = \frac{1}{2}\frac{1}{V_\ell}\left((\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i))\mathbf{N}_{i\ell}+(\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{r\tilde{\ell}}.$$

Proposition C.5.1 (Consistency of the fluxes for the DDFV scheme)**.** *Let $\bar{u} \in W^{1,\infty}(\Omega)$ be the exact solution of* (C.3) *and θ_ℓ be the angle between $\mathbf{x}_i\mathbf{x}_j$ and $\mathbf{n}_{il}$ (see Figure C.1). Assume that* **H1** *is satisfied. Then we have*

$$\begin{cases} \left|\bar{\mathcal{F}}_\ell - \mathcal{F}_\ell(\bar{u})\right| \leq \dfrac{C_\ell}{\cos(\theta_\ell)}|\ell|\left((1+|\sin(\theta_\ell)|)|\ell| + |\tilde{\ell}|\right) \leq \dfrac{2C}{\cos(\theta_\ell)}h^2, \\ \left|\bar{\mathcal{F}}_{\tilde{\ell}} - \mathcal{F}_{\tilde{\ell}}(\bar{u})\right| \leq \dfrac{C_\ell}{\cos(\theta_\ell)}|\tilde{\ell}|\left((1+|\sin(\theta_\ell)|)|\tilde{\ell}| + |\ell|\right) \leq \dfrac{2C}{\cos(\theta_\ell)}h^2, \end{cases} \tag{C.5}$$

where $C_\ell \leq C_0\|D^2\bar{u}\|_{L^\infty}$, where C_0 is a universal constant, and $C = \max_\ell C_\ell$.

Proof. Using the midpoint integration formula we have

$$\bar{\mathcal{F}}_\ell = \int_\ell \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell} + \mathcal{O}(|\ell|^2), \quad \text{and} \quad \bar{\mathcal{F}}_{\tilde{\ell}} = \int_{\tilde{\ell}} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{n}_{r\tilde{\ell}} + \mathcal{O}(|\tilde{\ell}|^2),$$

hence

$$\begin{aligned} \bar{\mathcal{F}}_\ell - \mathcal{F}_\ell(\bar{u}) &= \int_\ell \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell} - \frac{1}{2}\frac{1}{V_\ell}\left((\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i))\mathbf{N}_{i\ell} + (\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{i\ell} + \mathcal{O}(|\ell|^2), \\ \bar{\mathcal{F}}_{\tilde{\ell}} - \mathcal{F}_{\tilde{\ell}}(\bar{u}) &= \int_{\tilde{\ell}} \boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{n}_{r\tilde{\ell}} - \frac{1}{2}\frac{1}{V_\ell}\left((\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i))\mathbf{N}_{i\ell} + (\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r))\mathbf{N}_{r\tilde{\ell}}\right)\cdot\mathbf{N}_{r\tilde{\ell}} + \mathcal{O}(|\tilde{\ell}|^2). \end{aligned}$$

We have

$$V_\ell = \frac{1}{2}\cos(\theta_\ell)|\ell||\tilde{\ell}|, \quad \mathbf{N}_{i\ell} = \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}\mathbf{N}_{i\tilde{\ell}}^\perp - \frac{1}{\cos(\theta_\ell)}\frac{|\ell|}{|\tilde{\ell}|}\mathbf{N}_{r\tilde{\ell}}^\perp, \quad \mathbf{N}_{r\tilde{\ell}} = \frac{1}{\cos(\theta_\ell)}\frac{|\tilde{\ell}|}{|\ell|}\mathbf{N}_{i\ell}^\perp - \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}\mathbf{N}_{r\tilde{\ell}}^\perp.$$

where θ_ℓ is the angle between $\mathbf{x}_i\mathbf{x}_j$ and $\mathbf{n}_{il}$ (see Figure C.1)[a]. Using the relation $\mathbf{a}\cdot\mathbf{b} = \|\mathbf{a}\|\|\mathbf{b}\|\cos(\mathbf{a},\mathbf{b}), \forall \mathbf{a},\mathbf{b} \in \mathbb{R}^2$, we obtain

$$\begin{aligned} \bar{\mathcal{F}}_\ell - \mathcal{F}_\ell(\bar{u}) &= \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{N}_{i\tilde{\ell}}^\perp - \frac{1}{\cos(\theta_\ell)}\frac{|\ell|}{|\tilde{\ell}|}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{N}_{r\tilde{\ell}}^\perp \\ &\quad - \frac{1}{\cos(\theta_\ell)}\frac{|\ell|}{|\tilde{\ell}|}(\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i)) - \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}(\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)) + \mathcal{O}(|\ell|^2), \\ \bar{\mathcal{F}}_{\tilde{\ell}} - \mathcal{F}_{\tilde{\ell}}(\bar{u}) &= \frac{1}{\cos(\theta_\ell)}\frac{|\tilde{\ell}|}{|\ell|}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{N}_{i\tilde{\ell}}^\perp - \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}\boldsymbol{\nabla}\bar{u}(\mathbf{x}_{\tilde{\ell}})\cdot\mathbf{N}_{r\tilde{\ell}}^\perp \\ &\quad - \frac{\sin(\theta_\ell)}{\cos(\theta_\ell)}(\bar{u}(\mathbf{x}_j)-\bar{u}(\mathbf{x}_i)) - \frac{1}{\cos(\theta_\ell)}\frac{|\tilde{\ell}|}{|\ell|}(\bar{u}(\mathbf{x}_s)-\bar{u}(\mathbf{x}_r)) + \mathcal{O}(|\tilde{\ell}|^2). \end{aligned}$$

Using Taylor expansions in the neighborhood of $\mathbf{x}_\ell$

$$\bar{u}(\mathbf{x}_j) = \bar{u}(\mathbf{x}_i) - \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{N}_{r\tilde{\ell}}^\perp + \mathcal{O}(|\tilde{\ell}|^2), \quad \bar{u}(\mathbf{x}_s) = \bar{u}(\mathbf{x}_r) + \boldsymbol{\nabla}\bar{u}(\mathbf{x}_\ell)\cdot\mathbf{N}_{i\tilde{\ell}}^\perp + \mathcal{O}(|\ell|^2).$$

we deduce (C.5). □

C.5.2 Discrete Poincaré inequality

Lemma C.5.2 (Discrete Poincaré inequality)**.** *Assume that* **H2** *and* **H3** *are satisfied. Consider $\mathbf{e} = \left(\mathbf{e}^{primal}, \mathbf{e}^{dual}\right) \in \mathbb{R}^{n+m}$, where $\mathbf{e}^{primal} = (e_i)_{1\leq i\leq n}$ and $\mathbf{e}^{dual} = (e_r)_{1\leq r\leq m}$. Assume moreover that we have homogeneous Dirichlet boundary condition, that is to say*

$$\forall r \in \partial\Omega, \quad e_r = 0. \tag{C.6}$$

[a] Note that $\cos(\theta_\ell)$ is always positive.

Then we have

$$\left(\sum_i V_i e_i^2 + \sum_r V_r e_r^2\right)^{1/2} \leq 2\sqrt{2}\,\mathrm{diam}(\Omega)\frac{\sqrt{N_{\max}\xi}}{\cos(\theta_0)}\left(\sum_\ell V_\ell\left(\left(\frac{e_j - e_i}{|\tilde{\ell}|}\right)^2 + \left(\frac{e_s - e_r}{|\ell|}\right)^2\right)\right)^{1/2},$$

with $N_{\max}, \xi, \theta_0$ *the constants definied by* **H2** *and* **H3**, *and where we use the convention that, if* $\ell \subset \partial\Omega$, *then* $e_i - e_j = e_i$.

Proof. Given a point $\mathbf{x} \in \Omega$, let $\boldsymbol{\eta}(\mathbf{x})$ be the (first) point of intersection between the horizontal half line (for example) passing through $\mathbf{x}$ and the boundary $\partial\Omega$ (see Figure (C.2)). For all primal face ℓ, let $\chi_\ell : \Omega \longrightarrow \{0,1\}$ be defined by

$$\chi_\ell(\mathbf{x}) = \begin{cases} 1 & \text{if } \ell \cap [\mathbf{x},\boldsymbol{\eta}(\mathbf{x})] \neq \emptyset, \\ 0 & \text{otherwise.} \end{cases}$$

We note that

$$\int_\Omega \chi_\ell \leq \mathrm{diam}(\Omega)|\ell|, \tag{C.7}$$

where $\mathrm{diam}(\Omega) = \max_{\mathbf{x},\mathbf{y}\in\Omega} \|\mathbf{x} - \mathbf{y}\|$ is the diameter of Ω.

Fixing $\mathbf{x} \in i$, we write e_i^2 as a telescopic sum along the segment $[\mathbf{x}, \mathbf{y}(\mathbf{x})]$, that is,

$$e_i^2 = e_i^2 - e_j^2 + ... + e_k^2 - e_r^2,$$

where we assume that e_r, with $\mathbf{x}_r$ a node of the boundary of the domain, is zero thanks to the homogeneous Dirichlet boundary condition. Using the triangle inequality, we deduce that

$$|e_i^2| \leq |e_i^2 - e_j^2 + ... + e_k^2 - e_r^2| \leq \sum_\ell |e_i^2 - e_j^2|.$$

with the convention that, in the right hand side, if $\ell \subset \partial\Omega$, then $e_j = 0$.

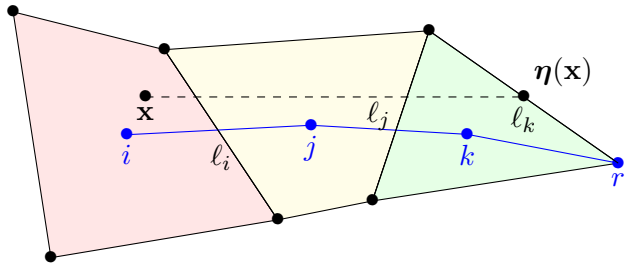


Fig. C.2 – An example of three adjacent primal cells and a horizontal half line (dashed lines) from the point $\mathbf{x} \in i$ and intersecting the two interior sides ℓ_i, ℓ_j and the border side ℓ_k at point $\boldsymbol{\eta}(\mathbf{x})$.

The definition of χ_ℓ allows to write this as follow

$$e_i^2 \leq \sum_\ell |e_j^2 - e_i^2|\chi_\ell(\mathbf{x}),$$

where the sum runs over all faces ℓ such that $\ell \cap [\mathbf{x}, \boldsymbol{\eta}(\mathbf{x})]$. Integrating this inequality over i with respect to $\mathbf{x}$, we have

$$\int_i e_i^2 = V_i e_i^2 \leq \sum_\ell |e_j^2 - e_i^2| \int_i \chi_\ell.$$

Using (C.7), we deduce that

$$\sum_i V_i e_i^2 \leq \sum_i \left(\sum_\ell |e_j^2 - e_i^2| \int_i \chi_\ell \right) = \sum_\ell |e_j^2 - e_i^2| \int_\Omega \chi_\ell \leq \mathrm{diam}(\Omega) \sum_\ell |\ell| |e_j^2 - e_i^2|,$$

that is to say

$$\sum_i V_i e_i^2 \leq \mathrm{diam}(\Omega) \sum_\ell |\ell| |e_j^2 - e_i^2|. \tag{C.8}$$

Noting that

$$\sum_\ell |\ell| |e_j^2 - e_i^2| = \sum_\ell \frac{1}{\cos(\theta_\ell)} \left(\cos(\theta_\ell) |\ell| |\tilde{\ell}| \right)^{1/2} \frac{|e_j - e_i|}{|\tilde{\ell}|} \left(\cos(\theta_\ell) |\ell| |\tilde{\ell}| \right)^{1/2} |e_j + e_i|,$$

and using assumption **H1**, we obtain

$$\sum_\ell |\ell| |e_j^2 - e_i^2| \leq \sum_\ell \frac{1}{cos(\theta_0)} \left(\cos(\theta_\ell) |\ell| |\tilde{\ell}| \right)^{1/2} \frac{|e_j - e_i|}{|\tilde{\ell}|} \left(\cos(\theta_\ell) |\ell| |\tilde{\ell}| \right)^{1/2} (|e_j| + |e_i|).$$

Hence, using the Cauchy-Schwarz inequality and the fact that $V_\ell = \frac{1}{2} \cos(\theta_\ell) |\ell| \ |\tilde{\ell}|$, we infer

$$\sum_\ell |\ell| |e_j^2 - e_i^2| \leq \frac{2}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 \right)^{1/2} \left(\sum_\ell V_\ell (|e_j| + |e_i|)^2 \right)^{1/2}.$$

Since

$$(|e_j| + |e_i|)^2 \leq 2(|e_j|^2 + |e_i|^2),$$

this gives

$$\sum_\ell |\ell| |e_j^2 - e_i^2| \leq \frac{2\sqrt{2}}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 \right)^{1/2} \left(\sum_\ell V_\ell (|e_j|^2 + |e_i|^2) \right)^{1/2}. \tag{C.9}$$

Taking into account assumptions **H2** and **H3** we have

$$\sum_\ell V_\ell (e_i^2 + e_j^2) \leq \xi \sum_\ell (V_i e_i^2 + V_j e_j^2) \leq N_{\max} \xi \sum_i V_i e_i^2,$$

Inserting this estimate into (C.9), we deduce that

$$\sum_\ell |\ell| |e_j^2 - e_i^2| \leq 2\sqrt{2} \frac{\sqrt{N_{\max} \xi}}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 \right)^{1/2} \left(\sum_i V_i e_i^2 \right)^{1/2}.$$

Using Equation (C.8) gives

$$\left(\sum_i V_i e_i^2 \right)^{1/2} \leq 2\sqrt{2} \ \mathrm{diam}(\Omega) \frac{\sqrt{N_{\max} \xi}}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 \right)^{1/2}. \tag{C.10}$$

Applying the same argument to the dual mesh, we also have

$$\left(\sum_r V_r e_r^2 \right)^{1/2} \leq 2\sqrt{2} \ \mathrm{diam}(\Omega) \frac{\sqrt{N_{\max} \xi}}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right)^{1/2}. \tag{C.11}$$

Collecting (C.10) and (C.11), we obtain

$$\left(\sum_i V_i e_i^2 + \sum_r V_r e_r^2 \right)^{1/2} \leq 2\sqrt{2} \ \mathrm{diam}(\Omega) \frac{\sqrt{N_{\max} \xi}}{\cos(\theta_0)} \left(\sum_\ell V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right) \right)^{1/2},$$

which concludes the proof. □

C.5.3 Convergence

Proposition C.5.3 (Convergence of the DDFV scheme)**.** *Let $\bar{u} \in W^{1,\infty}(\Omega)$ be the exact solution of* (C.3)*. Let $e_i = \bar{u}(\mathbf{x}_i) - u_i$, $\forall i \in [\![1,n]\!]$ and $e_r = \bar{u}(\mathbf{x}_r) - u_r$, $\forall r \in [\![1,m]\!]$, where $\mathbf{u}$ is the solution of the scheme* (C.4)*. Assume that* **H1**, **H2**, **H3** *are satisfied. Then we have*

$$\left(\sum_i V_i e_i^2 + \sum_r V_r e_r^2 \right)^{1/2} \leq C_1 h,$$

with C_1 a constant independent of h.

Proof. The fluxes $\bar{\mathcal{F}}_\ell$, $\mathcal{F}_\ell(\mathbf{u})$, $\bar{\mathcal{F}}_{\tilde{\ell}}$, $\mathcal{F}_{\tilde{\ell}}(\bar{u})$ are such that

$$-\sum_{\ell \in i} \bar{\mathcal{F}}_\ell = -\sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) = \int_i f \quad \text{and} \quad -\sum_{\tilde{\ell} \in r} \bar{\mathcal{F}}_{\tilde{\ell}} = -\sum_{\tilde{\ell} \in r} \mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = \int_r f.$$

Therefore

$$\sum_{\ell \in i} \bar{\mathcal{F}}_\ell = \sum_{\ell \in i} \mathcal{F}_\ell(\mathbf{u}) \quad \text{and} \quad \sum_{\tilde{\ell} \in r} \bar{\mathcal{F}}_{\tilde{\ell}} = \sum_{\tilde{\ell} \in r} \mathcal{F}_{\tilde{\ell}}(\mathbf{u}).$$

Given $e_i = \bar{u}(\mathbf{x}_i) - \mathbf{u}_i$ and $e_r = \bar{u}(\mathbf{x}_r) - \mathbf{u}_r$ we deduce that

$$\sum_{\ell \in i} \mathcal{F}_\ell(\bar{u}) - \mathcal{F}_\ell(\mathbf{u}) = \sum_{\ell \in i} \mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell = \frac{1}{2} \frac{1}{V_\ell} \left((e_j - e_i) \mathbf{N}_{i\ell} + (e_s - e_r) \mathbf{N}_{r\tilde{\ell}} \right) \cdot \mathbf{N}_{i\ell},$$
$$\sum_{\tilde{\ell} \in r} \mathcal{F}_{\tilde{\ell}}(\bar{u}) - \mathcal{F}_{\tilde{\ell}}(\mathbf{u}) = \sum_{\tilde{\ell} \in r} \mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} = \frac{1}{2} \frac{1}{V_\ell} \left((e_j - e_i) \mathbf{N}_{i\ell} + (e_s - e_r) \mathbf{N}_{r\tilde{\ell}} \right) \cdot \mathbf{N}_{r\tilde{\ell}}.$$

Multiplying these relations respectively by e_i and e_r and summing over the primal cells i and dual cells r, we obtain

$$\sum_i e_i \sum_{\ell \in i} (\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell) + \sum_r e_r \sum_{\tilde{\ell} \in r} (\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}}) =$$
$$\frac{1}{2} \sum_i \sum_{\ell \in i} \left(e_i \frac{1}{V_\ell} \left((e_j - e_i) \mathbf{N}_{i\ell} + (e_s - e_r) \mathbf{N}_{r\tilde{\ell}} \right) \cdot \mathbf{N}_{i\ell} \right) + \frac{1}{2} \sum_r \sum_{\tilde{\ell} \in r} \left(e_r \frac{1}{V_\ell} \left((e_j - e_i) \mathbf{N}_{i\ell} + (e_s - e_r) \mathbf{N}_{r\tilde{\ell}} \right) \cdot \mathbf{N}_{r\tilde{\ell}} \right).$$

Exchanging the sums, and grouping the terms by diamond cells, this reads as

$$\sum_\ell \left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right) (e_j - e_i) + \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right) (e_s - e_r) =$$
$$\frac{1}{2} \sum_\ell \frac{1}{V_\ell} \left((e_j - e_i)^2 \mathbf{N}_{i\ell} \cdot \mathbf{N}_{i\ell} + (e_s - e_r)^2 \mathbf{N}_{r\tilde{\ell}} \cdot \mathbf{N}_{r\tilde{\ell}} + 2(e_j - e_i)(e_s - e_r) \mathbf{N}_{i\ell} \cdot \mathbf{N}_{r\tilde{\ell}} \right) =$$
$$2 \sum_\ell \frac{1}{\cos(\theta_\ell)^2} V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 + 2 \sin(\theta_\ell) \frac{e_j - e_i}{|\tilde{\ell}|} \frac{e_s - e_r}{|\ell|} \right). \quad \text{(C.12)}$$

We next use the following inequality, which holds for all $X, Y \in \mathbb{R}^n$

$$X^2 + Y^2 \leq \frac{1}{1 - |\sin(\theta_\ell)|} \left(X^2 + Y^2 + 2 \sin(\theta_\ell) XY \right) = \frac{1 + |\sin(\theta_\ell)|}{\cos(\theta_\ell)^2} \left(X^2 + Y^2 + 2 \sin(\theta_\ell) XY \right). \quad \text{(C.13)}$$

Estimate (C.13) and equality (C.12) imply

$$\sum_\ell V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right)$$

$$\leq \sum_\ell \frac{1+|\sin(\theta_\ell)|}{\cos(\theta_\ell)^2} V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 + 2\sin(\theta_\ell) \frac{e_j - e_i}{|\tilde{\ell}|} \frac{e_s - e_r}{|\ell|} \right)$$

$$\leq 2 \sum_\ell \frac{1}{\cos(\theta_\ell)^2} V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 + 2\sin(\theta_\ell) \frac{e_j - e_i}{|\tilde{\ell}|} \frac{e_s - e_r}{|\ell|} \right)$$

$$= \sum_\ell \left(\left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right) (e_j - e_i) + \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right) (e_s - e_r) \right).$$

Using the Cauchy-Schwarz inequality we obtain

$$\sum_\ell \left(\left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right) (e_j - e_i) + \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right) (e_s - e_r) \right)$$

$$= \sum_\ell \left(\frac{|\tilde{\ell}|}{V_\ell^{1/2}} \left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right) \frac{V_\ell^{1/2}}{|\tilde{\ell}|} (e_j - e_i) + \frac{|\ell|}{V_\ell^{1/2}} \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right) \frac{V_\ell^{1/2}}{|\ell|} (e_s - e_r) \right)$$

$$\leq \left(\sum_\ell \frac{|\tilde{\ell}|^2}{V_\ell} \left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right)^2 + \frac{|\ell|^2}{V_\ell} \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right)^2 \right)^{1/2} \left(\sum_\ell V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right) \right)^{1/2},$$

hence

$$\left(\sum_\ell V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right) \right)^{1/2} \leq \left(\sum_\ell \frac{|\tilde{\ell}|^2}{V_\ell} \left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right)^2 + \frac{|\ell|^2}{V_\ell} \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right)^2 \right)^{1/2}. \tag{C.14}$$

Applying the consistency of fluxes (C.5) we have

$$\sum_\ell \frac{|\tilde{\ell}|^2}{V_\ell} \left(\mathcal{F}_\ell(\bar{u}) - \bar{\mathcal{F}}_\ell \right)^2 + \frac{|\ell|^2}{V_\ell} \left(\mathcal{F}_{\tilde{\ell}}(\bar{u}) - \bar{\mathcal{F}}_{\tilde{\ell}} \right)^2$$
$$\leq 4 \frac{C_\ell^2}{\cos(\theta_\ell)^4} \left(\left((1+|\sin(\theta_\ell)|)|\ell| + |\tilde{\ell}| \right)^2 + \left((1+|\sin(\theta_\ell)|)|\tilde{\ell}| + |\ell| \right)^2 \right) \leq 8 \frac{C^2}{\cos(\theta_0)^4} |\Omega| (2+\sigma)^2 h^2, \tag{C.15}$$

with

$$C = \max_\ell C_\ell, \quad \sigma = \max_\ell |\sin(\theta_\ell)|.$$

Inserting (C.15) into (C.14), we deduce

$$\left(\sum_\ell V_\ell \left(\left(\frac{e_j - e_i}{|\tilde{\ell}|} \right)^2 + \left(\frac{e_s - e_r}{|\ell|} \right)^2 \right) \right)^{1/2} \leq \sqrt{8} \frac{C}{\cos(\theta_0)^2} |\Omega|^{\frac{1}{2}} (2+\sigma) h. \tag{C.16}$$

Applying Lemma C.5.2 to the left-hand side of Equation (C.16), we conclude that

$$\left(\sum_i V_i e_i^2 + \sum_r V_r e_r^2 \right)^{1/2} \leq C_1 h,$$

with

$$C_1 = 8\sqrt{2}\, \mathrm{diam}(\Omega) |\Omega|^{1/2} \frac{C}{\cos(\theta_0)^3} (2+\sigma) \sqrt{N_{\max} \xi},$$

hence the method is (at least) first-order convergent. □

C.5.4 Coercivity

Lemma C.5.4 (Coercivity)**.** *Let* $\mathbf{A}$ *be the matrix associated with the DDFV discretization* (C.4) *of equation* (C.3)*. There exists a constant* C_2 *independent of* h *such that*

$$\forall \mathbf{u} \in \mathbb{R}^n, \quad \|\mathbf{u}\|_2^2 \leq C_2 \mathbf{u}^T \mathbf{A}\mathbf{u}.$$

Proof. Owing to the identity

$$V_\ell = \frac{1}{2}\cos(\theta_\ell)|\ell||\tilde{\ell}|,$$

we have

$$\begin{aligned}
\mathbf{u}^T\mathbf{A}\mathbf{u} &= \frac{1}{2}\sum_{\ell\notin\partial\Omega}\frac{1}{V_\ell}\|\mathbf{N}_{i\ell}(u_j-u_i)+\mathbf{N}_{r\tilde{\ell}}(u_s-u_r)\|^2 + \frac{1}{2}\sum_{\ell\in\partial\Omega}\left(\frac{1}{V_\ell}\|\mathbf{N}_{i\ell}(u_\ell-u_i)+\mathbf{N}_{r\tilde{\ell}}(u_s-u_r)\|^2\right)\\
&= 2\sum_{\ell\notin\partial\Omega}\frac{1}{\cos(\theta_\ell)^2}V_\ell\left(\left(\frac{u_j-u_i}{|\tilde{\ell}|}\right)^2+\left(\frac{u_s-u_r}{|\ell|}\right)^2+2\sin(\theta_\ell)\frac{u_j-u_i}{|\tilde{\ell}|}\frac{u_s-u_r}{|\ell|}\right)\\
&+2\sum_{\ell\in\partial\Omega}\frac{1}{\cos(\theta_\ell)^2}V_\ell\left(\left(\frac{u_\ell-u_i}{|\tilde{\ell}|}\right)^2+\left(\frac{u_s-u_r}{|\ell|}\right)^2+2\sin(\theta_\ell)\frac{u_\ell-u_i}{|\tilde{\ell}|}\frac{u_s-u_r}{|\ell|}\right).
\end{aligned}$$

As we have assumed that $u = g = 0$ on $\partial\Omega$ we can apply Lemma C.5.2 to $\mathbf{u} = ((u_i)_{1\leq i\leq n},(u_r)_{1\leq r\leq m})$ instead of $\mathbf{e} = ((e_i)_{1\leq i\leq n},(e_r)_{1\leq r\leq m})$. Therefore there exists a constant C_2 independent of h such that

$$\left(\sum_i V_i u_i^2 + \sum_r V_r u_r^2\right)^{1/2} \leq C_2\left(\sum_\ell V_\ell\left(\left(\frac{u_j-u_i}{|\tilde{\ell}|}\right)^2+\left(\frac{u_s-u_r}{|\ell|}\right)^2\right)\right)^{1/2}.$$

Using inequality (C.13), we have

$$\begin{aligned}
&\sum_\ell V_\ell\left(\left(\frac{u_j-u_i}{|\tilde{\ell}|}\right)^2+\left(\frac{u_s-u_r}{|\ell|}\right)^2\right)\\
&\leq 2\sum_\ell\frac{1}{\cos(\theta_\ell)^2}V_\ell\left(\left(\frac{u_j-u_i}{|\tilde{\ell}|}\right)^2+\left(\frac{u_s-u_r}{|\ell|}\right)^2+2\sin(\theta_\ell)\frac{u_j-u_i}{|\tilde{\ell}|}\frac{u_s-u_r}{|\ell|}\right),
\end{aligned}$$

which allows to conclude the proof. □

C.5.5 Stability

Lemma C.5.5 (Stability)**.** *Let* $\mathbf{u}$ *be the solution to* (C.4)*. We have*

$$\|\mathbf{u}\|_2 \leq C_2\|\mathbf{f}\|_2,$$

where C_2 *does not depend on* $\mathbf{u}$*,* $\mathbf{f}$ *and* h*.*

Proof. We have

$$\mathbf{u}^t\mathbf{A}\mathbf{u} = \sum_i V_i f_i u_i + \sum_r V_r f_r u_r,$$

hence, owing to the Cauchy-Schwarz inequality

$$\mathbf{u}^t\mathbf{A}\mathbf{u} \leq \left(\sum_i V_i f_i^2 + \sum_r V_r f_r^2\right)^{1/2}\left(\sum_i V_i u_i^2 + \sum_r V_r u_r^2\right)^{1/2} = \|\mathbf{f}\|_2\|\mathbf{u}\|_2.$$

Now, thanks to Lemma C.5.4, we obtain

$$\|\mathbf{u}\|_2^2 \leq C_2 \mathbf{u}^t \mathbf{A}\mathbf{u} \leq C_2 \|\mathbf{f}\|_2 \|\mathbf{u}\|_2,$$

which allows to conclude. □

Appendix D

Appendix of the Chapter 3

D.1 Computation of the coefficients $\alpha_{i\ell,ig}$, $\alpha_{i\ell,jg}$, $\beta_{i\ell,ig}$ and $\beta_{i\ell,jg}$

In this appendix, we give the details of the computations of the coefficents $\alpha_{i\ell,ig}$, $\alpha_{i\ell,jg}$, $\beta_{i\ell,ig}$ and $\beta_{i\ell,jg}$ given by Equations (3.13), (3.14), (3.16) and (3.17).

First, we have

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,ig}\frac{\mathbf{x}_g - \mathbf{x}_i}{\|\mathbf{x}_g - \mathbf{x}_i\|} + \beta_{i\ell,ig}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}.$$

Since $\mathbf{n}_{i\ell}$ is orthogonal to the edge ℓ the vertices of which are r and s, by taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain on the one hand

$$1 = \alpha_{i\ell,ig}\frac{(\mathbf{x}_g - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}{\|\mathbf{x}_g - \mathbf{x}_i\|},$$

that is to say

$$\alpha_{i\ell,ig} = \frac{\|\mathbf{x}_g - \mathbf{x}_i\|}{(\mathbf{x}_g - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}.$$

On the other hand, we also have

$$\mathbf{n}_{i\ell}\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp = \beta_{i\ell,ig}\frac{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say

$$\beta_{i\ell,ig} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp}{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_g - \mathbf{x}_i)^\perp}.$$

Secondly, we have

$$\mathbf{n}_{i\ell} = \alpha_{i\ell,jg}\frac{\mathbf{x}_j - \mathbf{x}_g}{\|\mathbf{x}_j - \mathbf{x}_g\|} + \beta_{i\ell,jg}\frac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}.$$

Since $\mathbf{n}_{i\ell}$ is orthogonal to the edge ℓ the vertices of which are r and s, by taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain on the one hand

$$1 = \alpha_{i\ell,jg}\frac{(\mathbf{x}_j - \mathbf{x}_g)\cdot\mathbf{n}_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_g\|},$$

that is to say

$$\alpha_{i\ell,jg} = \frac{\|\mathbf{x}_j - \mathbf{x}_g\|}{(\mathbf{x}_j - \mathbf{x}_g)\cdot\mathbf{n}_{i\ell}}.$$

On the other hand, we also have

$$\mathbf{n}_{i\ell}\cdot(\mathbf{x}_j - \mathbf{x}_g)^\perp = \beta_{i\ell,jg}\frac{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_j - \mathbf{x}_g)^\perp}{\|\mathbf{x}_s - \mathbf{x}_r\|},$$

that is to say

$$\beta_{i\ell,jg} = \frac{\|\mathbf{x}_s - \mathbf{x}_r\|\mathbf{n}_{i\ell}\cdot(\mathbf{x}_j - \mathbf{x}_g)^\perp}{(\mathbf{x}_s - \mathbf{x}_r)\cdot(\mathbf{x}_j - \mathbf{x}_g)^\perp}.$$

D.2 Proof of Proposition 3.5.1

Proof. The sum can be rewritten by inverting the sum on the cells and on the faces. Besides, the sum can be separated into boundary terms and non-boundary-terms

$$\sum_{i=1}^{n}\left(-\sum_{\ell\in i}\mathcal{F}_\ell(\mathbf{u})\right) = -\sum_{\ell\in\Gamma}\mathcal{F}_\ell(\mathbf{u}) - \sum_{\ell\in\mathring{\Omega}}\left(\mathcal{F}_{\ell,i}(\mathbf{u})+\mathcal{F}_{\ell,j}(\mathbf{u})\right),$$

where ℓ is the face shared by the cells i and j, and with

$$\begin{cases}\mathcal{F}_{\ell,i}(\mathbf{u}) = \gamma_\ell(u_j-u_i)+r_{\ell,i}(\mathbf{u}),\\ \mathcal{F}_{\ell,j}(\mathbf{u}) = \gamma_\ell(u_i-u_j)+r_{\ell,j}(\mathbf{u}),\end{cases}$$

with $r_{\ell,i}(\mathbf{u}) = -r_{\ell,j}(\mathbf{u})$. Then,

$$\mathcal{F}_{\ell,i}(\mathbf{u})+\mathcal{F}_{\ell,j}(\mathbf{u}) = 0.$$

The homogeneous Neumann boundary condition means that the boundary terms are zero, which leads to

$$\sum_{i=1}^{n}\left(-\sum_{\ell\in i}\mathcal{F}_\ell(\mathbf{u})\right) = 0,$$

that is to say

$$\sum_{i=1}^{n} V_i\lambda_i u_i = \sum_{i=1}^{n} V_i f_i.$$

The scheme is conservative. □

D.3 Exactness for polynomials of degree k

In this appendix, we give the proof that our approximation of the flux is exact for polynomials of degree k.

The flux is defined by

$$\begin{aligned}\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\sum_{g\in\ell}\omega_g\Bigg[&\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}}\right)(\bar{u}_j-\bar{u}_i+r_{gj}(\bar{\mathbf{u}})+r_{ig}(\bar{\mathbf{u}}))\\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j-\mathbf{x}_g\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(P_j(\mathbf{x}_s)-P_j(\mathbf{x}_r)+r_{rs,j}(\bar{\mathbf{u}}))\\ &+\left(\frac{\kappa_{g,i}\kappa_{g,j}\alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_g-\mathbf{x}_i\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\kappa_{g,i}\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\kappa_{g,j}\alpha_{i\ell,jg}\right)}\right)(P_i(\mathbf{x}_s)-P_i(\mathbf{x}_r)+r_{rs,i}(\bar{\mathbf{u}}))\Bigg],\end{aligned}$$

Considering κ continuous, the flux becomes

$$\begin{aligned}\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\sum_{g\in\ell}\omega_g\kappa_g\Bigg[&\left(\frac{\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j-\mathbf{x}_g\|\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\alpha_{i\ell,jg}}\right)(\bar{u}_j-\bar{u}_i+r_{gj}(\bar{\mathbf{u}})+r_{ig}(\bar{\mathbf{u}}))\\ &+\left(\frac{\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j-\mathbf{x}_g\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\alpha_{i\ell,jg}\right)}\right)(P_j(\mathbf{x}_s)-P_j(\mathbf{x}_r)+r_{rs,j}(\bar{\mathbf{u}}))\\ &+\left(\frac{\alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_g-\mathbf{x}_i\|}{\|\mathbf{x}_s-\mathbf{x}_r\|\left(\|\mathbf{x}_j-\mathbf{x}_g\|\alpha_{i\ell,ig}+\|\mathbf{x}_g-\mathbf{x}_i\|\alpha_{i\ell,jg}\right)}\right)(P_i(\mathbf{x}_s)-P_i(\mathbf{x}_r)+r_{rs,i}(\bar{\mathbf{u}}))\Bigg],\end{aligned}$$

Besides, we have

$$\begin{cases} \mathbf{n}_{i\ell} = \alpha_{i\ell}\dfrac{\mathbf{x}_j - \mathbf{x}_i}{\|\mathbf{x}_j - \mathbf{x}_i\|} + \beta_{i\ell}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \mathbf{n}_{i\ell} = \alpha_{i\ell,jg}\dfrac{\mathbf{x}_j - \mathbf{x}_\ell}{\|\mathbf{x}_j - \mathbf{x}_\ell\|} + \beta_{i\ell,jg}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}, \\ \mathbf{n}_{i\ell} = \alpha_{i\ell,ig}\dfrac{\mathbf{x}_\ell - \mathbf{x}_i}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \beta_{i\ell,ig}\dfrac{\mathbf{x}_s - \mathbf{x}_r}{\|\mathbf{x}_s - \mathbf{x}_r\|}. \end{cases} \tag{D.1}$$

By taking the scalar product with $\mathbf{n}_{i\ell}$, we obtain

$$\begin{cases} \dfrac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} = \dfrac{1}{(\mathbf{x}_j - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell}}, \\ (\mathbf{x}_j - \mathbf{x}_\ell)\cdot\mathbf{n}_{i\ell} = \dfrac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,jg}}, \\ (\mathbf{x}_\ell - \mathbf{x}_i)\cdot\mathbf{n}_{i\ell} = \dfrac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,ig}}, \end{cases} \tag{D.2}$$

that is to say

$$\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} = \frac{1}{\frac{\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,jg}} + \frac{\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,ig}}} = \frac{\alpha_{i\ell,jg}\alpha_{i\ell,ig}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,ig} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,jg}}.$$

Then, by taking the scalar product with $\mathbf{x}_s - \mathbf{x}_r$ on (D.1), we have

$$\begin{cases} \beta_{i\ell} = -\alpha_{i\ell}\dfrac{(\mathbf{x}_j - \mathbf{x}_i)\cdot(\mathbf{x}_s - \mathbf{x}_r)}{\|\mathbf{x}_j - \mathbf{x}_i\|}, \\ (\mathbf{x}_j - \mathbf{x}_\ell)\cdot(\mathbf{x}_s - \mathbf{x}_r) = -\dfrac{\beta_{i\ell,jg}\|\mathbf{x}_j - \mathbf{x}_\ell\|}{\alpha_{i\ell,j}}, \\ (\mathbf{x}_\ell - \mathbf{x}_i)\cdot(\mathbf{x}_s - \mathbf{x}_r) = -\dfrac{\beta_{i\ell,ig}\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\alpha_{i\ell,i}}, \end{cases}$$

that is to say

$$\frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} = \frac{\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|}\frac{\beta_{i\ell,jg}}{\|\mathbf{x}_s - \mathbf{x}_r\|} + \frac{\alpha_{i\ell,jg}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|}\frac{\beta_{i\ell,ig}}{\|\mathbf{x}_s - \mathbf{x}_r\|}}{\frac{\alpha_{i\ell,ig}}{\|\mathbf{x}_\ell - \mathbf{x}_i\|} + \frac{\alpha_{i\ell,jg}}{\|\mathbf{x}_j - \mathbf{x}_\ell\|}} = \frac{\alpha_{i\ell,ig}\beta_{i\ell,jg}\|\mathbf{x}_j - \mathbf{x}_\ell\| + \alpha_{i\ell,jg}\beta_{i\ell,ig}\|\mathbf{x}_\ell - \mathbf{x}_i\|}{\|\mathbf{x}_s - \mathbf{x}_r\|\left(\|\mathbf{x}_j - \mathbf{x}_\ell\|\alpha_{i\ell,ig} + \|\mathbf{x}_\ell - \mathbf{x}_i\|\alpha_{i\ell,jg}\right)}.$$

Since $\bar{u}$ is a polynomial function of degree k, the interpolation polynomial P is exactly equal to $\bar{u}$. Therefore, the node values $P_j(\mathbf{x}_r)$ are equal to $\bar{u}(\mathbf{x}_r)$. In the case of a contiuous κ, the flux is

$$\mathcal{F}_\ell(\bar{\mathbf{u}}) = |\ell|\sum_{g\in\ell}\omega_g\kappa_g\left[\left(\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}\right)(\bar{u}_j - \bar{u}_i + r_{gj}(\bar{\mathbf{u}}) + r_{ig}(\bar{\mathbf{u}})) + \left(\frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}\right)(\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + r_{rs}(\bar{\mathbf{u}}))\right]. \tag{D.3}$$

We will prove that our approximation of $\nabla\bar{u}(\mathbf{x}_g)$ is exact for polynomials of degree k, for each Gauss point g. To show that we will consider that our exact solution $\bar{u}$ is a polynomial of degree k centered in $\mathbf{x}_g$ in order to simplify the calculations. Moreover, to prove this exactness for a polynomial of degree k, we will prove it for each monomial of the basis centered in $\mathbf{x}_g$ in which we decompose our polynomial, that is to say $\{1,x - x_g,y - y_g,(x - x_g)^m, m \in [\![1,k]\!],(y - y_g)^n, n \in [\![1,k]\!],(x - x_g)(y - y_g),(x - x_g)^m(y - y_g)^n, m \in [\![1,k]\!],n \in [\![1,k]\!]$ tels que $3 \leq m + n \leq k\}$.

First, let's take

$$\bar{u}(\mathbf{x}) = (x - x_g)^m(y - y_g)^n, \quad m \in [\![1,k]\!], n \in [\![1,k]\!] \text{ tels que } 3 \leq m + n \leq k.$$

Then, we have

$$\nabla \bar{u}(\mathbf{x}) = \begin{pmatrix} m(x-x_g)^{m-1}(y-y_g)^n \\ n(x-x_g)^m(y-y_g)^{n-1} \end{pmatrix},$$

and so $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = 0$.

Then,

$$\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}) = \frac{m!n!}{(m-q)!(n-(p-q))!}(x-x_g)^{m-q}(y-y_g)^{n-(p-q)},$$

and so, in $\mathbf{x}_g$, the only non-zero terms are those for $m = q$ and $n = p - q$

$$\frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) = m!n!,$$

our sums therefore contain only one term for $p = m + n$ et $q = m$

$$\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) = \binom{p}{p-q} \frac{m!n!}{p!} = \frac{p!}{q!(p-q)!} \frac{q!(p-q)!}{p!} = 1.$$

In addition, we used integral values to estimate the values of u at cell centers, so we have

$$\bar{u}_j = \frac{1}{V_j} \int_j \bar{u}(\mathbf{x}) dx = \frac{1}{V_j} \int_j (x-x_g)^m (y-y_g)^n.$$

The approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ used in the flux (D.3)

$$\left[\left(\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \right) (\bar{u}_j - \bar{u}_i + r_{gj}(\bar{\mathbf{u}}) + r_{ig}(\bar{\mathbf{u}})) \; + \left(\frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \right) (\bar{u}(\mathbf{x}_s) - \bar{u}(\mathbf{x}_r) + r_{rs}(\bar{\mathbf{u}})) \right],$$

becomes

$$\begin{aligned}
&\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \left(\frac{1}{V_j} \int_j (x-x_g)^m (y-y_g)^n - \frac{1}{V_i} \int_i (x-x_g)^m (y-y_g)^n \right.\right. \\
&- \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \left(\frac{1}{V_j} \int_j (x-x_g)^q (y-y_g)^{(p-q)} dx - \frac{1}{V_i} \int_i (x-x_g)^q (y-y_g)^{(p-q)} dx \right) \Bigg) \\
&\qquad + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left((x_s-x_g)^m (y_s-y_g)^n - (x_s-x_g)^m (y_s-y_g)^n \right. \\
&\left.\left. - \sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) \left((x_s-x_g)^q (y_s-y_g)^{(p-q)} dx - (x_s-x_g)^q (y_s-y_g)^{(p-q)} dx \right) \right) \right],
\end{aligned}$$

that is to say

$$\begin{aligned}
&\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \left(\frac{1}{V_j} \int_j (x-x_g)^m (y-y_g)^n - \frac{1}{V_i} \int_i (x-x_g)^m (y-y_g)^n \right.\right. \\
&\qquad - \left(\frac{1}{V_j} \int_j (x-x_g)^m (y-y_g)^n dx - \frac{1}{V_i} \int_i (x-x_g)^m (y-y_g)^n dx \right) \Bigg) \\
&\qquad + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left((x_s-x_g)^m (y_s-y_g)^n - (x_s-x_g)^m (y_s-y_g)^n \right. \\
&\qquad\qquad \left.\left. - ((x_s-x_g)^m (y_s-y_g)^n dx - (x_s-x_g)^m (y_s-y_g)^n dx)) \right] = 0,
\end{aligned}$$

Therefore, the approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ is exact for monomials of type $(x-x_g)^m (y-y_g)^n, m \in [\![1,k]\!], n \in [\![1,k]\!]$ such that $3 \leq m + n \leq k$.

Next, let's take

$$\bar{u}(\mathbf{x}) = (x - x_g)(y - y_g).$$

Then, we have

$$\nabla \bar{u}(\mathbf{x}) = \begin{pmatrix} y - y_g \\ x - x_g \end{pmatrix},$$

and so $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = 0$.

Then, the only terms of our double sums that are non-zero are those obtained by taking $p = 2$ and $q = 1$

$$\frac{\partial^2 \bar{u}}{\partial x \partial y}(\mathbf{x}) = 1,$$

our sums therefore contain only one term for $p = 2$ et $q = 1$

$$\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) = \frac{1}{2!}\binom{2}{1} 1 = 1.$$

The approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\begin{aligned}\Bigg[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \Bigg(&\frac{1}{V_j}\int_j (x - x_g)(y - y_g) - \frac{1}{V_i}\int_i (x - x_g)(y - y_g) \\ &- \left(\frac{1}{V_j}\int_j (x - x_g)(y - y_g) - \frac{1}{V_i}\int_i (x - x_g)(y - y_g)\right)\Bigg) \\ &+ \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}\left((x_s - x_g)(y_s - y_g) - (x_r - x_g)(y_r - y_g)\right. \\ &\left. - \left((x_s - x_g)(y_s - y_g) - (x_r - x_g)(y_r - y_g)\right)\right)\Bigg] = 0.\end{aligned}$$

Thereore, the approxiation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ is exact for monomials of type $(x - x_g)(y - y_g)$.

Then

$$\bar{u}(\mathbf{x}) = x - x_g.$$

Thus, we have

$$\nabla \bar{u}(\mathbf{x}) = \begin{pmatrix} 1 \\ 0 \end{pmatrix},$$

and then $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = \frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}(x_j - x_i) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(x_s - x_r)$.

The approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}\left(\frac{1}{V_j}\int_j (x - x_g) - \frac{1}{V_i}\int_i (x - x_g)\right) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}\left((x_s - x_g) - (x_r - x_g)\right)\right],$$

that is to say

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|}(x_j - x_i) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|}(x_s - x_r)\right].$$

Therefore, the approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ is exact for monomials of type $x - x_g$.

And

$$\bar{u}(\mathbf{x}) = (x - x_g)^m, \quad m \in [\![1,k]\!].$$

Thus, we have

$$\nabla \bar{u}(\mathbf{x}) = \begin{pmatrix} m(x - x_g)^{m-1} \\ 0 \end{pmatrix},$$

and so $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}=0$.

Then, since our function does not depend on y, the sum $\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x})$ is zero for all $p\neq q$. We have

$$\frac{\partial^p\bar{u}}{\partial x^p}(\mathbf{x})=\frac{m!}{(m-p)!}(x-x_g)^{m-p},$$

and thus, in $\mathbf{x}_g$, the only non-zero terms are those for $m=p$

$$\frac{\partial^p\bar{u}}{\partial x^q}(\mathbf{x}_g)=p!,$$

our sums therefore contain only one term for $q=p=m$

$$\sum_{p=2}^{k}\frac{1}{p!}\sum_{q=0}^{p}\binom{p}{p-q}\frac{\partial^p\bar{u}}{\partial x^q\partial y^{(p-q)}}(\mathbf{x}_g)=\frac{m!}{m!}=1.$$

The approximation of $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j-\mathbf{x}_i\|}\left(\frac{1}{V_j}\int_j(x-x_g)^m-\frac{1}{V_i}\int_i(x-x_g)^m-\left(\frac{1}{V_j}\int_j(x-x_g)^m-\frac{1}{V_i}\int_i(x-x_g)^m\right)\right)\right.$$
$$\left.+\frac{\beta_{i\ell}}{\|\mathbf{x}_s-\mathbf{x}_r\|}\left((x_s-x_g)^m-(x_r-x_g)^m-((x_s-x_g)^m-(x_r-x_g)^m)\right)\right]=0.$$

Therefore, the approximation of $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}$ is exact for monomials of type $(x-x_g)^m, m\in[\![1,k]\!]$.

And

$$\bar{u}(\mathbf{x})=y-y_g.$$

Then, we have

$$\nabla\bar{u}(\mathbf{x})=\begin{pmatrix}0\\1\end{pmatrix},$$

and so $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}=\frac{\alpha_{i\ell}}{\|\mathbf{x}_j-\mathbf{x}_i\|}(y_j-y_i)+\frac{\beta_{i\ell}}{\|\mathbf{x}_s-\mathbf{x}_r\|}(y_s-y_r)$.

The approximation of $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j-\mathbf{x}_i\|}\left(\frac{1}{V_j}\int_j(y-y_g)-\frac{1}{V_i}\int_i(y-y_g)\right)+\frac{\beta_{i\ell}}{\|\mathbf{x}_s-\mathbf{x}_r\|}\left((y_s-y_g)-(y_r-y_g)\right)\right],$$

that is to say

$$\mathcal{F}_\ell(\bar{\mathbf{u}})=|\ell|\sum_{g\in\ell}\omega_g\kappa_g\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j-\mathbf{x}_i\|}(y_j-y_i)+\frac{\beta_{i\ell}}{\|\mathbf{x}_s-\mathbf{x}_r\|}(y_s-y_r)\right].$$

Therefore, the approximation of $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}$ is exact for monomials of type $y-y_g$.

And

$$\bar{u}(\mathbf{x})=(y-y_g)^n,\quad m\in[\![1,k]\!].$$

Then, we have

$$\nabla\bar{u}(\mathbf{x})=\begin{pmatrix}0\\n(y-y_g)^{n-1}\end{pmatrix},$$

and so $\nabla\bar{u}(\mathbf{x}_g)\cdot\mathbf{n}_{i\ell}=0$.

Then, since our function does not depend on x, the sum $\sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x})$ is zero for all $q \neq 0$. We have

$$\frac{\partial^p \bar{u}}{\partial y^p}(\mathbf{x}) = \frac{n!}{(n-p)!}(y-y_g)^{n-p},$$

and so, in $\mathbf{x}_g$, the only non-zero terms are those for $n = p$

$$\frac{\partial^p \bar{u}}{\partial y^p}(\mathbf{x}_g) = p!,$$

our sums therefore contain only one term for $q = 0$ and $p = n$

$$\sum_{p=2}^{k} \frac{1}{p!} \sum_{q=0}^{p} \binom{p}{p-q} \frac{\partial^p \bar{u}}{\partial x^q \partial y^{(p-q)}}(\mathbf{x}_g) = \frac{n!}{n!} = 1.$$

The approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \left(\frac{1}{V_j} \int_j (y-y_g)^n - \frac{1}{V_i} \int_i (y-y_g)^n - \left(\frac{1}{V_j} \int_j (y-y_g)^n - \frac{1}{V_i} \int_i (y-y_g)^n \right) \right) \right.$$
$$\left. + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left((y_s-y_g)^n - (y_r-y_g)^n - ((y_s-y_g)^n - (y_r-y_g)^n) \right) \right] = 0.$$

Therefore, the approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ is exact for monomials of type $(y-y_g)^n, n \in [\![1,k]\!]$.

Finally,

$$\bar{u}(\mathbf{x}) = 1.$$

Then, we have $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell} = 0$, and the approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ used in the flux (D.3) becomes

$$\left[\frac{\alpha_{i\ell}}{\|\mathbf{x}_j - \mathbf{x}_i\|} \left(\frac{1}{V_j} \int_j 1 - \frac{1}{V_i} \int_i 1 \right) + \frac{\beta_{i\ell}}{\|\mathbf{x}_s - \mathbf{x}_r\|} \left(\frac{1}{V_s} \int_s 1 - \frac{1}{V_r} \int_r 1 \right) \right] = 0,$$

Therefore, the approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ are exact for constants.

The approximation of $\nabla \bar{u}(\mathbf{x}_g) \cdot \mathbf{n}_{i\ell}$ is therefore exact for each monomials of the basis centered on $\mathbf{x}_g$: $\{1, x-x_g, y-y_g, (x-x_g)^m, m \in [\![1,k]\!], (y-y_g)^n, n \in [\![1,k]\!], (x-x_g)(y-y_g), (x-x_g)^m(y-y_g)^n, m \in [\![1,k]\!], n \in [\![1,k]\!]$ tels que $3 \leq m+n \leq k\}$, hence they are exact for any poynomial of degree k centered in $\mathbf{x}_g$, $\sum_{m=0}^{k} \sum_{n=0}^{k-m} a_{m,n}(x-x_g)^m(y-y_g)^n$. Besides, the Gauss quadrature formula of order k is also exact for polynomials of degree k, thus the fluxes are exact for polynomials of degree k.

Bibliography

[1] I. Aavatsmark, G.T. Eigestad, R.A. Klausen, M.F. Wheeler, and I. Yotov. Convergence of a symmetric MPFA method on quadrilateral grids. *Computational Geosciences*, 11(4):333–345, 2007. — Cited on pages vii, 4, and 41.

[2] D. G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12:547–560, 1965. — Cited on pages 103 and 106.

[3] B. Andreianov, F. Boyer, and F. Hubert. Discrete duality finite volume schemes for Leray-Lions type elliptic problems on general 2D meshes. *Numerical Methods for Partial Differential Equations*, 23:pp 145–195, 2007. — Cited on page 125.

[4] G. Barrenechea, V. John, and P. Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Mathematical Models and Methods in Applied Sciences*, 27(03):525–548, 2017. — Cited on pages vii, 4, and 75.

[5] G. Barrenechea, V. John, and P. Knobloch. Finite element methods respecting the discrete maximum principle for convection-diffusion equations, 2023. — Cited on pages vii, 4, and 75.

[6] L. Beirão da Veiga, F. Brezzi, L. D. Marini, and A. Russo. Virtual element method for general second-order elliptic problems on polygonal meshes. *Mathematical Models and Methods in Applied Sciences*, 26(04):729–750, 2016. — Cited on pages vi, 3, 9, and 74.

[7] E. Bertolazzi and G. Manzini. A second-order maximum principle preserving finite volume method for steady convection-diffusion problems. *SIAM Journal on Numerical Analysis*, 43(5):2172–2199, 2005. — Cited on pages vii, viii, 4, 8, 40, and 74.

[8] X. Blanc, F. Hermeline, E. Labourasse, and J. Patela. Arbitrary-order monotonic finite-volume schemes for 1D elliptic problems. *Computational and Applied Mathematics*, 42(4):195, June 2023. — Cited on pages ix, 5, 8, 72, and 74.

[9] X. Blanc, F. Hermeline, E. Labourasse, and J. Patela. Monotonic diamond and DDFV type finite-volume schemes for 2D elliptic problems. *Communications in Computational Physics*, 34(2):456–502, June 2023. — Cited on pages ix, 6, 40, 67, 74, and 75.

[10] X. Blanc and E. Labourasse. A positive scheme for diffusion problems on deformed meshes. *ZAMM - Journal of Applied Mathematics and Mechanics*, 96(6):660–680, 2016. — Cited on pages viii, 5, 9, 17, 18, and 41.

[11] Xavier Blanc, Francois Hermeline, Emmanuel Labourasse, and Julie Patela. Arbitrary order monotonic finite-volume schemes for 2D elliptic problems. working paper or preprint, September 2023. — Cited on pages ix, 6, and 74.

[12] J. Breil and P.-H. Maire. A cell-centered diffusion scheme on two-dimensional unstructured meshes. *Journal of Computational Physics*, 224(2):785–823, 2007. — Cited on pages vii and 4.

[13] C. Brezinski. *Accélération de la convergence en analyse numérique*. Lecture notes in mathematics. Springer-Verlag, 1977. — Cited on pages 103 and 106.

[14] C. Brezinski. *Algorithmes d'accélération de la convergence: étude numérique*. Collection Langages et algorithmes de l'informatique. Technip, 1978. — Cited on pages 103 and 106.

[15] H. Brézis. *Analyse fonctionnelle: théorie et applications*. Collection Mathématiques appliquées pour la maîtrise. Masson, 1983. — Cited on page 2.

[16] F. Brezzi, A. Buffa, and K. Lipnikov. Mimetic finite differences for elliptic problems. *ESAIM, Mathematical Modelling and Numerical Analysis*, 43(2):277–295, 2009. — Cited on pages vii and 4.

[17] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM Journal on Numerical Analysis*, 43(5):1872–1896, 2005. — Cited on pages vii and 4.

[18] C. Buet and S. Cordier. On the non existence of monotone linear schema for some linear parabolic equations. *Comptes Rendus Mathematique*, 340(5):399–404, 2005. — Cited on pages vii and 4.

[19] E. Burman and A. Ern. Discrete maximum principle for galerkin approximations of the laplace operator on arbitrary meshes. *Comptes Rendus Mathematique*, 338(8):641–646, 2004. — Cited on pages vii and 4.

[20] E. Burman and A. Ern. Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. *Comptes Rendus Mathematique*, 338(8):641–646, 2004. — Cited on page 75.

[21] J.-S. Camier and F. Hermeline. A monotone nonlinear finite volume method for approximating diffusion operators on general meshes. *International Journal for Numerical Methods in Engineering*, 107:496–519, 2016. — Cited on pages vii, 4, 9, 17, 41, 72, and 74.

[22] C. Cancès and C. Guichard. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Foundations of Computational Mathematics*, 17:1525–1584, 2017. — Cited on pages 9, 41, and 74.

[23] F. Cao, Y. Yao, Y. Yu, and G. Yuan. A conservative enforcing positivity-preserving algorithm for diffusion scheme on general meshes. *International Journal of Numerical Analysis and Modeling*, 13(5):739–752, 2016. — Cited on page 40.

[24] G. Carré, S. Del Pino, B. Després, and E. Labourasse. A cell-centered Lagrangian hydrodynamics scheme on general unstructured meshes in arbitrary dimension. *Journal of Computational Physics*, 228(14):5160–5183, 2009. — Cited on page 74.

[25] T. Cavalcante, R. Filho, A. Souza, D. K. de Carvalho, and P. Lyra. A multipoint flux approximation with a diamond stencil and a non-linear defect correction strategy for the numerical solution of steady state diffusion problems in heterogeneous and anisotropic media satisfying the discrete maximum principle. *Journal of Scientific Computing*, 93, 09 2022. — Cited on pages viii and 4.

[26] P. Ciarlet. Discrete maximum principle for finite-difference operators. *Aequationes Mathematicae*, 4:338–352, 1970. — Cited on pages 8 and 40.

[27] P. Ciarlet. *The Finite Element Method for elliptic problems*, volume 40. SIAM, Philadelphia, 2002. — Cited on pages vi, 3, 9, and 74.

[28] P. Ciarlet and P.-A. Raviart. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Foundations of Computational Mathematics*, 2:17–31, 1973. — Cited on pages 8 and 40.

[29] B. Cockburn, G. E. Karniadakis, and C.-W. Shu. *Discontinuous Galerkin methods: theory, computation and applications*, volume 11. Springer Science & Business Media, 2012. — Cited on page 74.

[30] Y. Coudière, J.-P. Vila, and P. Villedieu. Convergence rate of a finite volume scheme for a two dimensional convection-diffusion problem. *ESAIM, Mathematical Modelling and Numerical Analysis*, 33(3):493–516, 1999. — Cited on pages vi, 3, and 38.

[31] A. Danilov and Y. Vassilevski. A monotone nonlinear finite volume method for diffusion equations on conformal polyhedral meshes. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 24(3):207–227, 2009. — Cited on pages viii and 5.

[32] B. Després. Non linear schemes for the heat equation in 1D. *ESAIM, Mathematical Modelling and Numerical Analysis*, 48(1):107–134, 2014. — Cited on pages 9 and 41.

[33] D. A. Di Pietro and J. Droniou. *The Hybrid High-Order method for polytopal meshes*, volume 19. Springer, 2020. — Cited on pages vi, 3, 9, and 74.

[34] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69. Springer, 2012. — Cited on pages vi, 3, and 9.

[35] K. Domelevo and P. Omnes. A finite volume method for the laplace equation on almost arbitrary two-dimensional grids. *ESAIM, Mathematical Modelling and Numerical Analysis*, 39(6):1203–1249, 2005. — Cited on pages 41 and 125.

[36] J. Droniou. Finite volume schemes for diffusion equations: introduction to and review of modern methods. *Mathematical Models and Methods in Applied Sciences*, 24(08):1575–1619, 2014. — Cited on pages vii and 4.

[37] J. Droniou and R. Eymard. Study of the mixed finite volume method for stokes and navier-stokes equations. *Numerical Methods for Partial Differential Equations*, 25(1):137–171, 2009. — Cited on pages vii and 4.

[38] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. *The gradient discretisation method*, volume 82. Springer, 2018. — Cited on pages vi and 3.

[39] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Math. Mod. Meth. Appl. Sci.*, 20(2):265–295, 2010. — Cited on pages vi and 3.

[40] J. Droniou and C. Le Potier. Construction and convergence study of schemes preserving the elliptic local maximum principle. *SIAM Journal on Numerical Analysis*, 49(2):459–490, 2011. — Cited on pages vii, viii, 4, 9, 17, 18, 22, 24, and 41.

[41] M. Dumbser, W. Boscheri, M. Semplice, and G. Russo. Central weighted ENO schemes for hyperbolic conservation laws on fixed and moving unstructured meshes. *SIAM Journal on Scientific Computing*, 39(6):A2564–A2591, 2017. — Cited on pages 47 and 90.

[42] M. Edwards and C. Rogers. Finite volume discretization with imposed flux continuity for the general tensor pressure equation. *Computational Geosciences*, 2:259–290, 1998. — Cited on pages vii and 4.

[43] M. G. Edwards and H. Zheng. A quasi-positive family of continuous Darcy-flux finite-volume schemes with full pressure support. *Journal of Computational Physics*, 227(22):9333–9364, 2008. — Cited on pages vii and 4.

[44] G. T. Eigestad, I. Aavatsmark, and M. Espedal. Symmetry and M-matrix issues for the O-method on an unstructured grid. *Computational Geosciences*, 6(3-4):381–404, 2002. Locally conservative numerical methods for flow in porous media. — Cited on pages vii and 4.

[45] A. Ern and J.-L. Guermond. Invariant-domain preserving high-order time stepping: II. IMEX schemes. *hal-03703035*, v1, 2022. — Cited on pages 103 and 106.

[46] L. Evans. Application of nonlinear semigroup theory to certain partial differential equations. In Michael G. Crandall, editor, *Nonlinear Evolution Equations*, pages 163–188. Academic Press, 1978. — Cited on pages 2, 8, 40, and 74.

[47] R. Eymard, T. Gallouët, C. Guichard, R. Herbin, and R. Masson. TP or not TP, that is the question. *Computational Geosciences*, 18(3-4):285–296, 2014. — Cited on pages vii, 4, and 8.

[48] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In Ph. G. Ciarlet and J.-L. Lions, editors, *Handbook of numerical analysis*, volume VII. North-Holland, Amsterdam, 2000. — Cited on pages vi, vii, 3, 4, 30, and 125.

[49] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: A scheme using stabilization and hybrid interfaces. *IMA Journal of Numerical Analysis*, 30(4):1009–1043, 2010. — Cited on pages vii, 4, and 41.

[50] H. A. Friis and M. G. Edwards. A family of MPFA finite-volume schemes with full pressure support for the general tensor pressure equation on cell-centered triangular grids. *Journal of Computational Physics*, 230(1):205–231, 2011. — Cited on pages vii and 4.

[51] Y. Gao, G. Yuan, S. Wang, and X. Hang. A finite volume element scheme with a monotonicity correction for anisotropic diffusion problems on general quadrilateral meshes. *Journal of Computational Physics*, 407:109143, 2020. — Cited on pages viii, 5, 9, 13, 38, 40, 41, 51, 72, and 91.

[52] Z. Gao and J. Wu. A second-order positivity-preserving finite volume scheme for diffusion equations on general meshes. *SIAM Journal on Scientific Computing*, 37(1):A420–A438, 2015. — Cited on pages viii, 5, 9, 13, 38, 40, 41, 51, 72, and 91.

[53] G. H. Golub and C. F. Van Loan. *Matrix Computations (3rd Ed.)*. Johns Hopkins University Press, USA, 1996. — Cited on page 90.

[54] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM review*, 43(1):89–112, 2001. — Cited on page 106.

[55] V. Gyrya, K. Lipnikov, G. Manzini, and D. Svyatskiy. M-adaptation in the mimetic finite difference method. *Mathematical Models and Methods in Applied Sciences*, 24(08):1621–1663, 2014. — Cited on pages vii and 4.

[56] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In R. Eymard and J.-M. Herard, editors, *Finite volume for complex applications, problems and perspectives V*. Wiley, 2008. — Cited on pages 41 and 42.

[57] F. Hermeline. A finite volume method for second-order elliptic equations. (Une méthode de volumes finis pour les équations elliptiques du second ordre.). *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics*, 1998. — Cited on pages vii and 3.

[58] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *Journal of Computational Physics*, 160(2):481–499, 2000. — Cited on pages vii, 3, 38, 41, 49, 85, and 105.

[59] F. Hermeline. Approximation of diffusion operators with discontinuous tensor coefficients on distorted meshes. *Computer Methods in Applied Mechanics and Engineering*, 192(16-18):1939–1959, 2003. — Cited on pages vii and 3.

[60] F. Hermeline. Approximation of 2D and 3D diffusion operators with variable full tensor coefficients on arbitrary meshes. *Computer Methods in Applied Mechanics and Engineering*, 196(21-24):2497–2526, 2007. — Cited on pages vii and 3.

[61] F. Hermeline. *Nouvelles méthodes de volumes finis pour approcher des équations aux dérivées partielles sur des maillages quelconques.* Habilitation à diriger des recherches, CEA/DAM Ile de France, 2008. — Cited on pages vii, 3, and 42.

[62] F. Hermeline. A finite volume method for approximating 3D diffusion operators on general meshes. *Journal of Computational Physics*, 228(16):5763–5786, 2009. — Cited on pages vii and 3.

[63] J. Karátson, S. Korotov, and M. Křížek. On discrete maximum principles for nonlinear elliptic problems. *Mathematics and Computers in Simulation*, 76(1):99–108, 2007. — Cited on pages 8 and 40.

[64] D. S. Kershaw. Differencing of the diffusion equation in Lagrangian hydrodynamic codes. *Journal of Computational Physics*, 39:375–395, 1981. — Cited on pages vi and 3.

[65] S. Korotov, M. Křížek, and P. Neittaanmäki. Weakened acute type condition for tetrahedral triangulations and the discrete maximum principle. *Mathematics of Computation*, 70(233):107–119, 2000. — Cited on pages 8 and 40.

[66] Y. Kuznetsov, K. Lipnikov, and M. Shashkov. The mimetic finite difference method on polygonal meshes for diffusion-type problems. *Computational Geosciences*, 8(4):301–324, 2005. — Cited on pages vii and 4.

[67] M. Käser and A. Iske. Ader schemes on adaptive triangular meshes for scalar conservation laws. *Journal of Computational Physics*, 205(2):486–508, 2005. — Cited on pages 47 and 90.

[68] E. Labourasse. *Contribution to the numerical simulation of radiative hydrodynamics.* Habilitation à diriger des recherches, sorbonne university, December 2021. — Cited on page 74.

[69] O. Larroche. An efficient explicit numerical scheme for diffusion-type equations with a highly inhomogeneous and highly anisotropic diffusion tensor. *Journal of Computational Physics*, 223:436–450, 2007. — Cited on pages 63 and 101.

[70] C. Le Potier. Schéma volumes finis monotone pour des opérateurs de diffusion fortement anisotropes sur des maillages de triangles non structurés. *Comptes Rendus Mathematique*, 341(12):787–792, 2005. — Cited on pages vii, 4, 8, 40, and 74.

[71] C. Le Potier. A linear scheme satisfying a maximum principle for anisotropic diffusion operators on distorted grids. (Un schéma linéaire vérifiant le principe du maximum pour des opérateurs de diffusion très anisotropes sur des maillages déformés.). *Comptes Rendus. Mathématique. Académie des Sciences, Paris*, 347(1-2):105–110, 2009. — Cited on pages vii and 4.

[72] C. Le Potier. Correction non linéaire et principe du maximum pour la discrétisation d'opérateurs de diffusion avec des schémas volumes finis centrés sur les mailles. *Comptes Rendus Mathematique*, 348(11-12):691–695, 2010. — Cited on pages vii, 4, 9, and 41.

[73] K. Lipnikov, G. Manzini, and M. Shashkov. Mimetic finite difference method. *Journal of Computational Physics*, 257, Part B(0):1163 – 1227, 2014. Physics-compatible numerical methods. — Cited on pages vii and 4.

[74] K. Lipnikov, G. Manzini, and D. Svyatskiy. Monotonicity conditions in the mimetic finite difference method. In *Finite volumes for complex applications. VI. Problems & perspectives. Volume 1, 2*, volume 4 of *Springer Proceedings in Mathematics and Statistics*, pages 653–661. Springer, Heidelberg, 2011. — Cited on pages vii and 4.

[75] K. Lipnikov, M. Shashkov, and D. Svyatskiy. The mimetic finite difference discretization of diffusion problem on unstructured polyhedral meshes. *Journal of Computational Physics*, 211(2):473–491, 2006. — Cited on pages vii and 4.

[76] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Yu. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227(1):492–512, 2007. — Cited on pages viii, 5, 9, and 41.

[77] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Interpolation-free monotone finite volume method for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 228(3):703–716, 2009. — Cited on pages viii, 5, 9, and 41.

[78] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. A monotone finite volume method for advection-diffusion equations on unstructured polygon meshes. *Journal of Computational Physics*, 229(11):4017–4032, 2010. — Cited on pages viii and 5.

[79] K. Lipnikov, D. Svyatskiy, and Y. Vassilevski. Minimal stencil finite volume scheme with the discrete maximum principle. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 27(4):369–385, 2012. — Cited on pages viii, 5, 9, and 41.

[80] R. Liska and M. Shashkov. Enforcing the discrete maximum principle for linear finite element solutions of second-order elliptic problems. *Communications in Computational Physics*, 3(4):852–877, 2008. — Cited on page 40.

[81] R. Loubère, M. Staley, and B. Wendroff. The repair paradigm: New algorithms and applications to compressible flow. *Journal of Computational Physics*, 211(2):385–404, 2006. — Cited on page 40.

[82] P.-H. Maire. A high-order cell-centered Lagrangian scheme for two-dimensional compressible fluid flows on unstructured meshes. *Journal of Computational Physics*, 228(7):2391–2425, 2009. — Cited on page 74.

[83] G. Meurant. *Computer solution of large linear systems*. Elsevier, 1999. — Cited on pages 32, 40, and 95.

[84] D. Mihalas and B. Weibel-Mihalas. *Foundations of radiation hydrodynamics*. Dover Books on Physics. Dover Publications, New York, NY, USA, 1999. — Cited on page 1.

[85] K. Nikitin and Y. Vassilevski. A monotone nonlinear finite volume method for advection-diffusion equations on unstructured polyhedral meshes in 3D. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 25(4):335–358, 2010. — Cited on pages viii and 5.

[86] G. J. Pert. Physical constraints in numerical calculations of diffusion. *Journal of Computational Physics*, 42(1):20–52, 1981. — Cited on pages vi and 3.

[87] R.J. Plemmons. M-matrix characterizations.I – nonsingular M-matrices. *Linear Algebra and its Applications*, 18(2):175 – 188, 1977. — Cited on pages 19, 55, and 92.

[88] E. H. Quenjel. Enhanced positive vertex-centered finite volume scheme for anisotropic convection-diffusion equations. *ESAIM, Mathematical Modelling and Numerical Analysis*, 54(2):591–618, 2020. — Cited on pages 9 and 74.

[89] L. Ramírez, L. Edreira, I. Couceiro, P. Ouro, X. Nogueira, and I. Colominas. A new mean preserving moving least squares method for arbitrary order finite volume schemes. *Applied Mathematics and Computation*, 443:127768, 2023. — Cited on page 105.

[90] P.-A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983. — Cited on pages vi and 3.

[91] L. Rebholz and M. Xiao. The effect of anderson acceleration on superlinear and sublinear convergence. *Journal of Scientific Computing*, 96, 06 2023. — Cited on pages 103 and 106.

[92] M. Schneider, L. Agélas, G. Enchéry, and B. Flemisch. Convergence of nonlinear finite volume schemes for heterogeneous anisotropic diffusion on general meshes. *Journal of Computational Physics*, 351:80–107, 2017. — Cited on pages 9 and 41.

[93] Z. Sheng and G. Yuan. A finite volume scheme for diffusion equations on distorted quadrilateral meshes. *Transport Theory and Statistical Physics*, 37(2-4):171–207, 2008. — Cited on page 17.

[94] Z. Sheng and G. Yuan. The finite volume scheme preserving extremum principle for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 230(7):2588–2604, 2011. — Cited on pages viii, 5, 9, and 41.

[95] Z. Sheng and G. Yuan. A new nonlinear finite volume scheme preserving positivity for diffusion equations. *Journal of Computational Physics*, 315:182–193, 2016. — Cited on pages 9, 41, and 74.

[96] Z. Sheng, J. Yue, and G. Yuan. Monotone finite volume schemes of nonequilibrium radiation diffusion equations on distorted meshes. *SIAM Journal on Scientific Computing*, 31(4):2915–2934, 2009. — Cited on pages viii, 5, and 41.

[97] V. Siess. A linear and accurate diffusion scheme respecting the maximum principle on distorted meshes. *Comptes Rendus de l'Académie des Sciences, Paris - Series I - Mathématique*, 347:1317–1320, 2009. — Cited on pages vii and 4.

[98] R. S. Varga. *Matrix iterative analysis*, volume 1. Prentice Hall, 1962. — Cited on pages 21, 56, and 93.

[99] T. Vejchodskỳ and P. Šolín. Discrete maximum principle for higher-order finite elements in 1D. *Mathematics of Computation*, 76(260):1833–1846, 2007. — Cited on pages 8 and 40.

[100] Clément Vincent and Mohamed Khelifi. Acceleration of convergence for numerical sequences. Technical report, CEA - Commissariat à l'énergie atomique et aux énergies alternatives, September 2023. — Cited on page 106.

[101] J. Wang, Z. Sheng, and G. Yuan. A finite volume scheme preserving maximum principle with cell-centered and vertex unknowns for diffusion equations on distorted meshes. *Applied mathematics and computation*, 398(1):1–21, 2021. — Cited on pages 9, 41, and 74.

[102] S. Wang and G. Yuan. Discrete strong extremum principles for finite element solutions of diffusion problems with nonlinear corrections. *Applied Numerical Mathematics*, 174:1–16, 2022. — Cited on pages vii, 4, and 75.

[103] S. Wang, G. Yuan, Y. Li, and Z. Sheng. Discrete maximum principle based on repair technique for diamond type scheme of diffusion problems. *International journal for numerical methods in fluids*, 70(9):1188–1205, 2012. — Cited on page 40.

[104] S. Wang, G. Yuan, Y. Li, and Z. Sheng. A monotone finite volume scheme for advection-diffusion equations on distorted meshes. *International Journal for Numerical Methods in Fluids*, 69(7):1283–1298, 2012. — Cited on pages viii and 5.

[105] J. Wu and Z. Gao. Interpolation-based second-order monotone finite volume schemes for anisotropic diffusion equations on general grids. *Journal of Computational Physics*, 275:569–588, 2014. — Cited on pages viii, 5, 9, 13, 38, 40, 41, 51, 72, 74, 91, and 103.

[106] H. Yang, B. Yu, Y. Li, and G. Yuan. Monotonicity correction for second order element finite volume methods of anisotropic diffusion problems. *Journal of Computational Physics*, 449:110759, 2022. — Cited on pages viii, 5, 9, 13, and 74.

[107] Y. Yao and G. Yuan. Enforcing positivity with conservation for nine-point scheme of nonlinear diffusion equations. *Computer methods in applied mechanics and engineering*, 223:161–172, 2012. — Cited on page 40.

[108] Y. Yu, X. Chen, and G. Yuan. A finite volume scheme preserving maximum principle for the system of radiation diffusion equation with three temperatures. *SIAM Journal on Scientific Computing*, 41(1):93–113, 2019. — Cited on pages 9 and 41.

[109] G. Yuan and Z. Sheng. Analysis of accuracy of a finite volume scheme for diffusion equations on distorted meshes. *Journal of Computational Physics*, 224:1170, 2007. — Cited on page 125.

[110] G. Yuan and Z. Sheng. Monotone finite volume schemes for diffusion equations on polygonal meshes. *Journal of Computational Physics*, 227(12):6288–6312, 2008. — Cited on pages viii, 5, 9, 41, and 98.

[111] X. Zhang, S. Su, and J. Wu. A vertex-centered and positivity-preserving scheme for anisotropic diffusion problems on arbitrary polygonal grids. *Journal of Computational Physics*, 344:419–436, 2017. — Cited on pages viii, 5, 9, 13, 38, 40, 41, 51, 72, and 91.

[112] F. Zhao, Z. Sheng, and G. Yuan. A monotone combination scheme of diffusion equations on polygonal meshes. *ZAMM - Journal of Applied Mathematics and Mechanics*, 100(5):1–25, 2020. — Cited on pages 9, 41, and 74.

Schémas volumes finis d'ordre arbitraire positifs pour la diffusion sur maillage quelconque

Résumé :

L'objectif de cette thèse est le développement et l'analyse de schémas volumes finis robustes et précis afin d'approcher la solution de l'équation de diffusion sur maillages quelconques avec un coefficient de diffusion qui peut être anisotrope et/ou discontinu. Afin de satisfaire ces propriétés, nos schémas devront préserver la positivité et être d'ordre élevé.

Dans ce manuscrit, nous proposons le premier schéma d'ordre arbitraire préservant la positivité pour la diffusion. Notre démarche est tout d'abord d'étudier le problème en 1D. Dans ce cas le problème de positivité n'apparaît qu'à partir de l'ordre 3. D'autre part, la dimension 1 nous permet de faire l'analyse mathématique de ce problème, notamment une preuve de convergence du schéma à un ordre arbitraire sous une hypothèse de stabilité. Ensuite, nous l'étendons en 2D à l'ordre 2, ce qui permet de nous appuyer sur des schémas connus. Nous avons étudié deux possibilités : un schéma type DDFV (Discrete Duality Finite Volume) que l'on compare à une méthode utilisant des reconstructions polynomiales. Enfin, cela nous permet de développer un schéma monotone d'ordre arbitraire sur maillage quelconque avec un coefficient de diffusion κ qui peut être discontinu et/ou anisotrope. La montée en ordre se fait grâce à une reconstruction polynomiale et la monotonicité s'obtient en se ramenant à une structure de M-matrice, ce qui nous donne des schémas non linéaires.

Chaque schéma est validé par des simulations numériques montrant l'ordre de convergence ainsi que la positivité de la solution obtenue.

Mots-clés : Méthode volumes-finis, ordre élevé, diffusion anisotrope, positivité, schéma DDFV.

Arbitrary-order finite volume schemes preserving positivity for diffusion problems on deformed meshes

Abstract:

The objective of this thesis is the development and the analysis of robust and accurate finite volume schemes for the approximation of the solution of the diffusion equation on deformed meshes with diffusion coefficient which can be anisotropic and/or discontinuous. To satisfy these properties, our schemes must preserve the positivity and achieve high-order accuracy.

In this manuscript, we propose the first positivity-preserving arbitrary-order scheme for diffusion. Our approach is first to study the problem in 1D. In such a case, the positivity problem only appears for order 3 and higher. The 1D setting allows us to perform the mathematical analysis of this problem, including a proof of convergence of the scheme to an arbitrary order under a stability assumption. We then extend it to 2D at order 2, relying on well-known schemes. We study two possibilities: a DDFV-type scheme (Discrete Duality Finite Volume), which we compare with a method using polynomial reconstruction. Finally, this allows us to develop a monotonic scheme of arbitrary order on any mesh with a κ diffusion coefficient that can be discontinuous and/or anisotropic. Improving the order is achieved through polynomial reconstruction, and monotonicity is obtained by reducing to a M-matrix structure, which gives nonlinear schemes.

Each scheme is validated by numerical simulations showing the order of convergence and the positivity of the solution obtained.

Keywords: Finite volume method, high-order, anisotropic diffusion, monotonic method, DDFV scheme.

Thèse de doctorat, Julie PATELA

Établissements de recherche :
CEA, DAM, DIF, F-91297 Arpajon, France,
Université Paris Cité, Sorbonne Université, CNRS, Laboratoire Jacques-Louis Lions, F-75013, Paris, France.