



Approximation of scalar and vector transport problems on polyhedral meshes

Pierre Cantin

► To cite this version:

Pierre Cantin. Approximation of scalar and vector transport problems on polyhedral meshes. General Mathematics [math.GM]. Université Paris-Est, 2016. English. NNT : 2016PESC1028 . tel-01419312

HAL Id: tel-01419312

<https://pastel.hal.science/tel-01419312>

Submitted on 19 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



THÈSE

présentée pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS-EST

École doctorale MSTIC, mention MATHÉMATIQUES APPLIQUÉES

par Pierre CANTIN

Approximation of scalar and vector transport problems on
polyhedral meshes

Approximation des problèmes de transport scalaire et vectoriel
sur maillages polyédriques

Soutenue publiquement le 14 novembre 2016 devant le jury composé de

Rapporteurs	Pr. Raphaële HERBIN	Université de Provence
	Pr. Eric SONNENDRÜCKER	Max-Planck Institut (Allemagne)
Examineurs	Pr. Boris ANDREIANOV	Université de Tours
	Pr. Lourenco BEIRÃO DA VEIGA	Università di Milano-Bicocca (Italie)
	Pr. Bruno DESPRÉS	Université Pierre et Marie Curie
	Pr. Roland MASSON	Université de Nice Sophia Antipolis
Directeur de thèse	Pr. Alexandre ERN	Université Paris-Est
Encadrant industriel	Dr. Jérôme BONELLE	EDF R&D

À mes parents, à mon frère,

"Il faut sans cesse se jeter du haut d'une falaise
et se fabriquer des ailes durant la chute"

Ray Bradbury

Remerciements

Les premiers remerciements qui me viennent à l'esprit sont bien évidemment destinés à mon directeur de thèse, Alexandre Ern. Avoir travaillé sous son encadrement durant ces trois années reste pour moi un privilège. Je souhaite le remercier très sincèrement pour sa disponibilité, ses conseils scientifiques (et autres), et surtout pour m'avoir initié à son approche et sa vision des mathématiques.

Mes remerciements vont ensuite à Jérôme Bonelle, mon encadrant industriel à EDF R&D, qui a sans cesse œuvré à ce que cette thèse soit une réussite tant sur le plan académique, que sur le plan industriel. Son encadrement calme et juste m'a beaucoup servi durant cette thèse, et j'en suis sûr, le restera pour la suite de mon parcours. Au même titre, je souhaite également remercier EDF R&D, et plus particulièrement le projet et l'équipe de *Code_Saturne*, qui m'a permis d'effectuer cette thèse dans les meilleures conditions.

Je souhaite maintenant remercier Raphaële Herbin et Eric Sonnendrücker, pour m'avoir fait l'honneur de rapporter ma thèse. Je suis également reconnaissant à Boris Andreianov, Lourenco Beirão da Veiga, Bruno Després et Roland Masson pour leur présence dans mon jury de thèse. Je souhaite également remercier Erik Burman pour avoir accepté une collaboration au cours cette thèse et pour son accueil à Londres.

Mes remerciements vont maintenant à toutes les personnes que j'ai côtoyé durant cette thèse et qui m'ont témoigné leur soutien. En particulier à EDF, je souhaite remercier Jean-Marc Hérard, pour ses conseils et son attention portée au bon déroulement de cette thèse, Sofiane Benhamadouche et Martin Ferrand pour leurs regards critiques et leur présence à mes comités de thèse, Romain Camy pour l'utilisation de ses maillages, David Monfort et Marine Le Coq pour tous les à-côtés administratifs et bien évidemment l'ex et le toujours thésard, Jean-François et Benjamin. Au CERMICS, je souhaite également remercier Laurent Monasse pour l'intérêt porté à mes travaux, Isabelle Simunic pour son aide administrative et son efficacité, ainsi que les thésards et Post-Doc, Yannick, Boris, Karol, Amina et Thomas.

Puisque si je suis arrivé jusqu'ici, c'est aussi grâce à eux, je remercie avec un certain engouement mes potes nantais de toujours, Alex, Audrey et Charles, ainsi que les éminents membres de la Secte, Alexandre, Anne-Sophie, Charles, Mathieu, Matthieu, Noémie, Odile, Sylvain, ainsi que Maylis et Romain, mes colocos de ces une ou trois dernières années.

Enfin, je termine en témoignant ma profonde gratitude envers ma famille, mes parents et mon frère, pour leurs encouragements et leurs soutiens, sans lesquels je ne serais pas arrivé jusqu'ici.

Table des matières

1	Introduction	1
1.1	Contexte et motivations industrielles	1
1.2	État de l'art et contributions	6
1.3	Organisation de la thèse	10
2	On the Friedrichs positivity assumption for advection-reaction problems in Banach spaces	15
2.1	Continuous settings	16
2.2	Scalar advection-reaction problem	17
2.3	Vector advection-reaction problem	22
3	Vertex-based scheme for advection-reaction	25
3.1	Discrete setting	26
3.2	Advection-reaction problem	30
3.3	Numerical results	41
4	Péclet-robust vertex-based scheme for diffusion-advection-reaction	47
4.1	Discrete setting	48
4.2	Pure Diffusion problem	50
4.3	Diffusion-advection-reaction problem	56
4.4	Numerical results	62
5	Improved vertex-based scheme for scalar transport	67
5.1	Discrete setting	68
5.2	Advection-reaction problem	69
5.3	Diffusion-advection-reaction problem	77
5.4	Numerical results	79
6	Edge-based scheme for vector advection-reaction	87
6.1	Discrete setting	88
6.2	Advection-reaction problem	89
6.3	Numerical results	101
7	Analysis on polyhedral meshes	105
7.1	Polyhedral meshes	105
7.2	Mesh partitions	109
7.3	Functional inequalities in polyhedral meshes	113
7.4	Reconstruction and approximation	116
8	Conclusions et perspectives	127
A	Additional numerical results on Kershaw mesh sequences	131

Chapitre 1

Introduction

Contents

1.1	Contexte et motivations industrielles	1
1.1.1	Utilisation de maillages polyédriques	2
1.1.2	Des équations de Navier-Stokes aux problèmes de transport	3
1.2	État de l’art et contributions	6
1.2.1	Analyse des problèmes continus	7
1.2.2	Résolution du problème d’advection-réaction scalaire	8
1.2.3	Résolution du problème de diffusion-advection-réaction scalaire	9
1.2.4	Résolution du problème de transport vectoriel	10
1.3	Organisation de la thèse	10

1.1 Contexte et motivations industrielles

La mécanique des fluides est au cœur des activités de l’industrie de l’énergie et de l’aéronautique. C’est le cas notamment en ce qui concerne les activités de EDF R&D, et plus particulièrement celles du département Mécanique des Fluides, Énergie et Environnement (MFEE), qui étudie le renforcement de la sûreté et l’optimisation du fonctionnement des moyens de production d’énergie. À ces fins, des codes de simulation numérique sont développés au sein de ce département, notamment *Code_Saturne*¹ pour les problèmes issus de la thermohydraulique à l’échelle locale.

Le cadre général de cette thèse est de proposer de nouveaux schémas numériques pour le code de référence *Code_Saturne* développé depuis la fin des années 90 (voir e.g., Archambeau *et al.* (2004)). Ce code, open-source depuis 2007, propose différentes approches numériques utilisant la méthode des volumes finis à variables co-localisées, afin de résoudre les problèmes modélisant les écoulements monophasiques. Parmi les nombreuses applications, ce code est notamment utilisé pour simuler les écoulements au sein des turbo-machines et des réacteurs des centrales du parc nucléaire français, pour simuler les écoulements souterrains, ou bien encore pour l’étude du productible éolien ou pour la modélisation des incendies.

Cette thèse s’inscrit dans la continuité des travaux de thèse de Bonelle (2014), portant sur les méthodes numériques utilisant l’approche *Compatible Discrete Operator* (CDO). L’objectif de ces méthodes est de préserver lors du passage du niveau continu au niveau discret, certaines propriétés du problème, comme le noyau des opérateurs mis en jeu et leurs relations d’adjonction. De ce fait, ces méthodes appartiennent à la classe des méthodes *mimétiques* (ou structure-preserving). Utilisant les travaux de Tonti (1975, 2013), de Bossavit (2000) et de Matthusi (2002), les méthodes CDO consistent à dégager une structure mathématique générale des

1. <http://www.code-saturne.org>

principaux phénomènes physiques à partir de concepts issus de la géométrie différentielle et de la topologie algébrique. Ces méthodes ont pour l'instant été analysées dans le cas du problème de la diffusion pure par Bonelle & Ern (2014a) ainsi que dans le cas du problème de Stokes par Bonelle & Ern (2014b). L'objectif de cette thèse est d'étendre ces travaux afin de traiter plus généralement les phénomènes de transport sur maillages polyédriques, et notamment ceux liés à l'*advection*.

1.1.1 Utilisation de maillages polyédriques

Comme son nom l'indique, un maillage polyédrique d'un domaine connexe borné polyédrique Ω de dimension 3 est un pavage de Ω formé de polyèdres de $\mathbb{R}^3$, constitué de mailles délimitées par un nombre fini de faces planes polygonales. Contrairement à un maillage formé de simplexes, les éléments d'un maillage polyédrique ont des caractéristiques géométriques différentes. Un maillage polyédrique peut par exemple contenir à la fois des mailles convexes et non convexes, ainsi que des mailles possédant un nombre variable de faces. D'un point de vue pratique, l'utilisation de maillages polyédriques est plus délicate que celle de maillages classiques, généralement de type structuré ou bien formé d'un seul type d'élément simple (tétraèdre, hexaèdre, prisme à base triangulaire). D'un point de vue théorique, il est alors nécessaire d'introduire un formalisme adapté afin de prendre en compte ces généralités. En comparaison, les schémas de discrétisation classiques considèrent généralement des maillages spécifiques, comme par exemple ceux de type structurés dans le cas des méthodes aux différences finies, ou bien ceux formés d'un seul type d'élément simple pour la méthode des éléments finis.

Plusieurs raisons motivent l'utilisation de maillages polyédriques dans l'industrie. La première concerne les aspects "coût mémoire" liés à l'utilisation de gros maillages (contenant généralement plusieurs dizaines voire centaines de millions de mailles) lors des simulations sur des géométries de grande taille avec des mailles très fines. Grâce à l'utilisation de polyèdres, il est possible de diminuer le nombre nécessaire de mailles en utilisant plusieurs types d'éléments. L'utilisation simultanée d'hexaèdres, de pyramides à base carrée et de tétraèdres dans un même maillage permet par exemple de diviser localement par 3 le nombre de mailles, comme on peut l'observer sur la Figure 1.1. Ces maillages multi-éléments sont également utilisés en pratique pour mailler finement les couches limites.

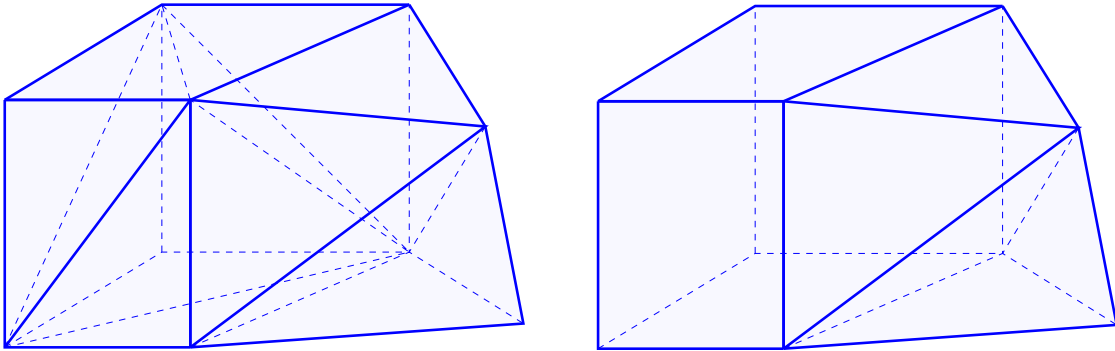


FIGURE 1.1 – Un pavage de $\mathbb{R}^3$ formé de 9 tétraèdres à gauche et de 3 polyèdres à droite.

La deuxième raison motivant l'utilisation de maillages polyédriques est la possibilité de considérer des recollements non-conformes de sous-maillages. Dans ce cas, le maillage polyédrique est obtenu en deux étapes. La première étape consiste à décomposer le domaine initial Ω en sous-domaines $\Omega = \cup_i \Omega_i$ d'intersection vide deux à deux et à mailler séparément ces sous-domaines Ω_i . La deuxième étape consiste ensuite à agglomérer les maillages de tous ces sous-domaines Ω_i afin d'obtenir un maillage du domaine Ω . Les non-conformités qui résultent de ces recollements conduisent alors à des mailles comportant plusieurs faces dans un même hyperplan. Les mailles en question sont alors vues comme des polyèdres contenant plus de faces. Dans le cas par exemple du recollement non-conforme représenté sur la Figure 1.2, la

non-conformité de recollement modifie l'élément de gauche de telle sorte qu'on ne le considère plus comme un hexaèdre, mais comme un nonaèdre. En pratique, les maillages obtenus par

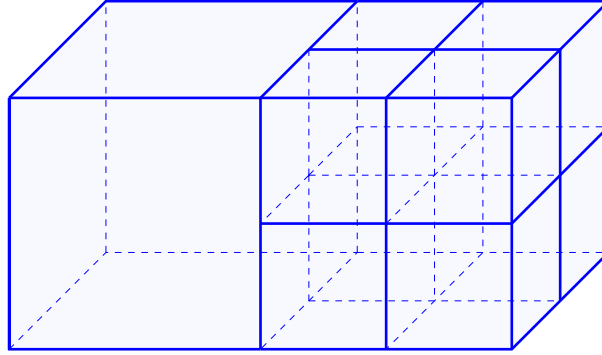


FIGURE 1.2 – Recollement non-conforme de deux maillages hexaédriques.

ce procédé correspondent à ceux utilisés dans l'industrie. Ces maillages ont l'avantage d'une part de pouvoir être générés successivement à partir de mailleurs classiques et d'autre part d'offrir la possibilité de pouvoir être raffinés localement en fonction de la physique étudiée. La Figure 1.3 illustre deux situations considérées dans le cadre des études menées à EDF R&D, où l'on observe notamment les différents recollements non-conformes entre les sous-domaines constitutifs de la géométrie globale.

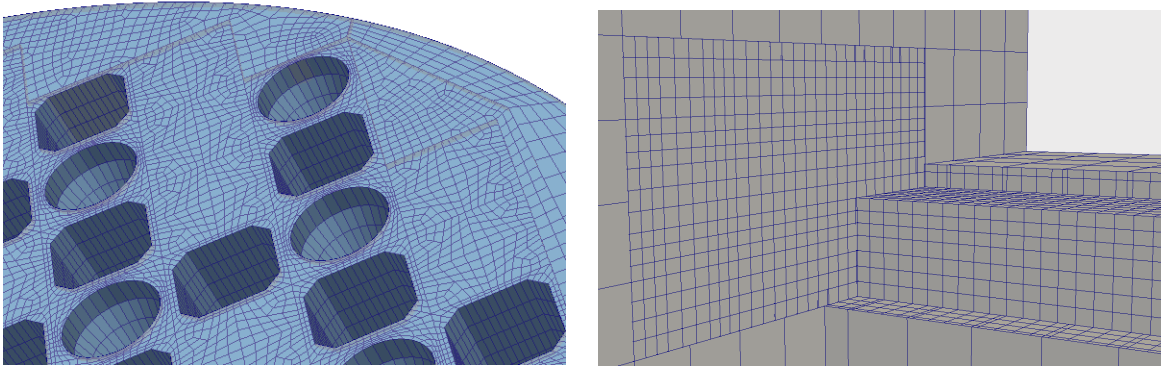


FIGURE 1.3 – Exemples de recollements non-conformes pour des applications industrielles à EDF R&D.

1.1.2 Des équations de Navier-Stokes aux problèmes de transport

Les équations de Navier-Stokes modélisent l'écoulement des fluides à l'échelle macroscopique (voir e.g., Darrozes & François (1982)). Dans le cas de l'écoulement incompressible dans un domaine Ω de $\mathbb{R}^3$ (fixe au cours du temps) d'un fluide de masse volumique constante ρ , de viscosité constante λ et soumis à une densité de champ de force $\mathbf{s}$ (typiquement la gravité), le problème de Navier-Stokes traduit la conservation de la masse et de la quantité de mouvement et consiste à déterminer la pression $p : \Omega \rightarrow \mathbb{R}$ et la vitesse $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ vérifiant

$$\rho \frac{\partial \mathbf{u}}{\partial t} + \rho(\mathbf{u} \cdot \nabla) \mathbf{u} - \lambda \Delta \mathbf{u} + \nabla p = \rho \mathbf{s} \quad \text{dans } \Omega, \quad (1.1a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{dans } \Omega, \quad (1.1b)$$

ainsi que des conditions à la limite (que nous omettons ici pour simplifier la discussion). Lorsque l'écoulement est supposé de surcroît stationnaire et en renormalisant la viscosité et la pression

par $\rho \neq 0$, le problème (1.1) se simplifie pour devenir

$$(\mathbf{u} \cdot \nabla) \mathbf{u} - \lambda \Delta \mathbf{u} + \nabla p = \mathbf{s} \quad \text{dans } \Omega, \quad (1.2a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{dans } \Omega. \quad (1.2b)$$

Malgré ces simplifications, l'analyse mathématique du problème (1.2) est délicate dans son ensemble (voir e.g., Temam (1977) ou Galdi (1994)), en raison notamment de la présence du terme d'advection non-linéaire (ou de convection) $(\mathbf{u} \cdot \nabla) \mathbf{u}$. En pratique, l'approximation numérique de la solution $(\mathbf{u}, p)$ du problème (1.2) peut être obtenue par l'intermédiaire d'un problème linéarisant le terme de convection dans l'équation (1.2a). Pour un champ vectoriel $\beta : \Omega \rightarrow \mathbb{R}^3$ donné, cette linéarisation conduit au problème de Oseen :

$$(\beta \cdot \nabla) \mathbf{u} - \lambda \Delta \mathbf{u} + \nabla p = \mathbf{s} \quad \text{dans } \Omega, \quad (1.3a)$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{dans } \Omega. \quad (1.3b)$$

La solution du problème (1.2) est alors obtenue par itération afin de déterminer un point fixe de la fonctionnelle qui à β associe $\mathbf{u}$.

Une méthode permettant de résoudre numériquement le problème (1.3) consiste à découpler les inconnues vitesse et pression. Dans cette approche, on est alors amené à considérer le cas du problème de *transport vectoriel* de la vitesse. En plus de ce problème, il est également intéressant d'introduire le problème de *transport scalaire*, notamment dans le cas où l'on adopte une vision composante par composante du problème de transport vectoriel. L'étude du problème de transport scalaire est également pertinente dans le cas de la modélisation du transport de la concentration d'espèces chimiques ou de grandeurs turbulentes.

Le problème de transport scalaire. Dans un domaine Ω de $\mathbb{R}^3$ de frontière $\partial\Omega$, considérons $\mu : \Omega \rightarrow \mathbb{R}$ un coefficient de réaction, $\beta : \Omega \rightarrow \mathbb{R}^3$ un champ d'advection et $\lambda : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ un tenseur de diffusion à valeurs symétriques. Le problème de transport scalaire considéré dans cette thèse consiste à déterminer la fonction $u : \Omega \rightarrow \mathbb{R}$ résolvant le problème

$$-\nabla \cdot (\lambda \nabla u) + \beta \cdot \nabla u + \mu u = s \quad \text{dans } \Omega, \quad (1.4a)$$

$$u = u_D \quad \text{sur } \partial\Omega, \quad (1.4b)$$

où $s : \Omega \rightarrow \mathbb{R}$ est le terme source et $u_D : \partial\Omega \rightarrow \mathbb{R}$ est la donnée au bord de Dirichlet (d'autres conditions à la limite peuvent être considérées, comme par exemple celles de type Neumann ou de type Robin).

Selon les valeurs que prennent les différents coefficients physiques du problème (1.4), on distingue plusieurs régimes. Pour simplifier la discussion, nous ne considérons pas le cas de la réaction dominante et nous supposons que les effets réactifs sont au plus du même ordre de grandeur que les effets advectifs. L'écoulement associé au problème (1.4) est alors caractérisé par un seul nombre adimensionnel, le *nombre de Péclet*. Ce nombre traduit le rapport des effets advectifs sur les effets diffusifs. Il est défini sur le domaine Ω par

$$\Pi_\Omega = \frac{L_\Omega \beta_\Omega}{\lambda_\Omega}, \quad (1.5)$$

où L_Ω correspond à une longueur caractéristique du domaine Ω , β_Ω à une échelle de vitesse du champ β et λ_Ω à une échelle de diffusion du tenseur λ . Lorsque les effets diffusifs dominent, i.e., $\Pi_\Omega \rightarrow 0$, le problème (1.4) tend formellement vers le problème elliptique de diffusion pure consistant à déterminer la fonction $u : \Omega \rightarrow \mathbb{R}$ solution du problème

$$-\nabla \cdot (\lambda \nabla u) = s \quad \text{dans } \Omega, \quad (1.6a)$$

$$u = u_D \quad \text{sur } \partial\Omega. \quad (1.6b)$$

À l'inverse, lorsque les effets advectifs dominent, i.e., $\Pi_\Omega \rightarrow +\infty$, le problème (1.4) tend formellement vers le problème du premier ordre d'advection-réaction où l'on cherche la fonction $u : \Omega \rightarrow \mathbb{R}$ solution du problème

$$\beta \cdot \nabla u + \mu u = s \quad \text{dans } \Omega, \quad (1.7a)$$

$$u = u_D \quad \text{sur } \partial\Omega^-, \quad (1.7b)$$

où la condition à la limite de Dirichlet est uniquement posée sur la frontière *entrante* (ou *inflow*) $\partial\Omega^- = \{\mathbf{x} \in \partial\Omega \mid \mathbf{n}(\mathbf{x}) \cdot \beta(\mathbf{x}) < 0\}$, avec $\mathbf{n}$ la normale sortante à Ω . La formulation du terme d'advection dans le problème (1.7) est dite *non-conservative*, tandis que l'on obtient la formulation *conservative* (1.8) grâce à la formule de Leibniz $\nabla \cdot (\beta u) = \beta \cdot \nabla u + (\nabla \cdot \beta)u$ en choisissant $\mu = \nabla \cdot \beta$:

$$\nabla \cdot (\beta u) = s \quad \text{dans } \Omega, \quad (1.8a)$$

$$u = u_D \quad \text{sur } \partial\Omega^-. \quad (1.8b)$$

Signalons que l'étude des problèmes (1.7) et (1.8) peut notamment être appliquée à l'équation de conservation de la masse lorsque l'écoulement modélisé par le problème de Navier-Stokes (1.1) est non-homogène (i.e., $\nabla \rho \neq 0$).

Le problème de transport vectoriel. Dans un domaine Ω de $\mathbb{R}^3$ de frontière $\partial\Omega$, le problème de transport vectoriel consiste à déterminer la fonction $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ solution du système

$$\nabla \times (\lambda \nabla \times \mathbf{u}) + \nabla(\beta \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \beta + \mu \mathbf{u} = \mathbf{s} \quad \text{dans } \Omega, \quad (1.9a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-, \quad (1.9b)$$

$$\mathbf{u} \times \mathbf{n} = \mathbf{u}_D \times \mathbf{n} \quad \text{sur } \partial\Omega \setminus \partial\Omega^-, \quad (1.9c)$$

avec $\mu : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ un tenseur de réaction, $\beta : \Omega \rightarrow \mathbb{R}^3$ un champ advectif, $\lambda : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ un tenseur de diffusion à valeurs symétriques, $\mathbf{s} : \Omega \rightarrow \mathbb{R}^3$ le terme source et $\mathbf{u}_D : \partial\Omega \rightarrow \mathbb{R}^3$ la donnée au bord de Dirichlet. Plusieurs formulations du problème de transport vectoriel (1.9) sont disponibles dans la littérature, notamment en ce qui concerne l'écriture du terme de diffusion. Ici, nous avons retenu la formulation dite *curl-curl*, en cohérence avec le problème de Stokes étudié par Bonelle & Ern (2014b), et initialement analysée par Nédélec (1982). Dans le cas d'un tenseur de diffusion constant et isotrope $\lambda = \lambda \text{Id}$, observons que le terme $\nabla \times (\lambda \nabla \times \mathbf{u})$ se réécrit $-\lambda \Delta \mathbf{u} + \lambda \nabla(\nabla \cdot \mathbf{u})$, permettant de retrouver le terme de diffusion du problème de Oseen (1.3) lorsque la vitesse $\mathbf{u}$ satisfait la contrainte d'isovolume $\nabla \cdot \mathbf{u} = 0$.

En ce qui concerne le terme d'advection, il est composé de deux termes : $\nabla(\beta \cdot \mathbf{u})$ et $(\nabla \times \mathbf{u}) \times \beta$. Le premier terme correspond aux variations de la fonction $\mathbf{u}$ le long des lignes de courant du champ β , tandis que le second terme décrit la rotation de la fonction $\mathbf{u}$ perpendiculairement à ces lignes de courant. Cette formulation du terme d'advection est pertinente dans de nombreuses situations, et notamment dans le cas du problème de Oseen. En effet, en supposant les champs $\mathbf{u}$ et β suffisamment réguliers, on rappelle que l'on a :

$$\nabla(\beta \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \beta = (\beta \cdot \nabla) \mathbf{u} + (\mathbf{u} \cdot \nabla) \beta - (\nabla \times \beta) \times \mathbf{u}, \quad (1.10a)$$

$$(\mathbf{u} \cdot \nabla) \beta = (\nabla \times \beta) \times \mathbf{u} + (\nabla \beta)^T \mathbf{u}. \quad (1.10b)$$

En combinant ces expressions, on obtient alors $\nabla(\beta \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \beta = (\beta \cdot \nabla) \mathbf{u} + (\nabla \beta)^T \mathbf{u}$. Le terme de droite est ainsi composé du terme d'advection composante par composante $(\beta \cdot \nabla) \mathbf{u}$, considéré dans le problème de Oseen (1.3), et d'un opérateur d'ordre zéro $(\nabla \beta)^T \mathbf{u}$. En particulier, le choix $\mu = -\nabla \beta^T$ dans le problème (1.9) permet de considérer le problème suivant

$$\nabla \times (\lambda \nabla \times \mathbf{u}) + (\beta \cdot \nabla) \mathbf{u} = \mathbf{s} \quad \text{dans } \Omega, \quad (1.11a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-, \quad (1.11b)$$

$$\mathbf{u} \times \mathbf{n} = \mathbf{u}_D \times \mathbf{n} \quad \text{sur } \partial\Omega \setminus \partial\Omega^-, \quad (1.11c)$$

où le terme d'advection est écrit en formulation *non-conservative* tandis que le choix $\boldsymbol{\mu} = -\nabla\boldsymbol{\beta}^\top + (\nabla\cdot\boldsymbol{\beta})\text{Id}$ fournit

$$\nabla\times(\boldsymbol{\lambda}\nabla\times\mathbf{u}) + \nabla\cdot(\mathbf{u}\otimes\boldsymbol{\beta}) = \mathbf{s} \quad \text{dans } \Omega, \quad (1.12a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-, \quad (1.12b)$$

$$\mathbf{u}\times\mathbf{n} = \mathbf{u}_D\times\mathbf{n} \quad \text{sur } \partial\Omega\setminus\partial\Omega^-, \quad (1.12c)$$

correspondant à la formulation *conservative* du terme d'advection. Comme dans le cas du problème de transport scalaire, et en supposant que les effets réactifs sont au plus de l'ordre de grandeur des effets advectifs, l'écoulement modélisé par le problème de transport vectoriel (1.9) est caractérisé par le nombre de Péclet. Dans le régime advectif, le problème (1.9) tend alors formellement vers le problème du premier ordre d'advection-réaction vectoriel

$$\nabla(\boldsymbol{\beta}\cdot\mathbf{u}) + (\nabla\times\mathbf{u})\times\boldsymbol{\beta} + \boldsymbol{\mu}\mathbf{u} = \mathbf{s} \quad \text{dans } \Omega, \quad (1.13a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-. \quad (1.13b)$$

Le problème de transport en géométrie différentielle. Plus généralement, les problèmes de transport (1.4) et (1.9) peuvent être reformulés dans le cadre de la géométrie différentielle à l'aide de trois opérateurs : la dérivée extérieure, l'opérateur de Hodge et l'opérateur de contraction. Considérant Ω une variété de dimension d , l'espace des k -formes différentielles pour $k \in \llbracket 0, d \rrbracket$ est noté $\Lambda^k(\Omega)$. En dimension $d = 3$, une 0-forme est appelée *potentiel*, une 1-forme est appelée *circulation*, une 2-forme est appelée *flux* et une 3-forme est appelée *densité*. Dans le formalisme de l'analyse vectorielle, l'équivalent (correspondant au proxy) d'une 0-forme et d'une 3-forme est une fonction $\Omega \rightarrow \mathbb{R}$, tandis que le proxy d'une 1-forme et d'une 2-forme est une fonction $\Omega \rightarrow \mathbb{R}^3$. La dérivée extérieure est notée $d : \Lambda^k(\Omega) \rightarrow \Lambda^{k+1}(\Omega)$ pour $k \in \llbracket 0, d-1 \rrbracket$ et admet pour proxy en dimension $d = 3$ les opérateurs différentiels $v \mapsto \nabla v$ pour $k = 0$, $\mathbf{v} \mapsto \nabla\times\mathbf{v}$ pour $k = 1$ et $\mathbf{v} \mapsto \nabla\cdot\mathbf{v}$ pour $k = 2$. L'opérateur de Hodge associé au paramètre $\boldsymbol{\lambda}$ est quant à lui noté $\star_{\boldsymbol{\lambda}} : \Lambda^k(\Omega) \rightarrow \Lambda^{d-k}(\Omega)$ et admet pour proxy en dimension $d = 3$ l'application $\mathbf{u} \mapsto \boldsymbol{\lambda}\mathbf{u}$ pour $k = 1$ et $k = 2$. Les opérateurs de diffusion $u \mapsto -\nabla\cdot(\boldsymbol{\lambda}\nabla u)$ et $\mathbf{u} \mapsto \nabla\times(\boldsymbol{\lambda}\nabla\times\mathbf{u})$ se représentent ainsi de manière abstraite par l'application

$$\Lambda^k(\Omega) \ni \omega^k \mapsto (-1)^{k+1} d \star_{\boldsymbol{\lambda}} d \omega^k \in \Lambda^{d-k}(\Omega), \quad (1.14)$$

pour $k = 0$ et $k = 1$, respectivement. Dans ce même formalisme, on représente les termes advectifs à l'aide de l'opérateur de contraction noté $\iota_{\boldsymbol{\beta}} : \Lambda^k(\Omega) \rightarrow \Lambda^{k-1}(\Omega)$ pour $k \in \llbracket 1, d \rrbracket$, où $\boldsymbol{\beta} : \Omega \rightarrow T\Omega$ est un champ vectoriel donné, défini comme une section du fibré tangent $T\Omega$ de Ω (voir Abraham *et al.* (1988)) isomorphe à l'espace des 1-formes différentielles $\Lambda^1(\Omega)$. L'opérateur de contraction $\iota_{\boldsymbol{\beta}}$ admet pour proxy en dimension $d = 3$ les applications $\mathbf{v} \mapsto \mathbf{v}\cdot\boldsymbol{\beta}$ pour $k = 1$, $\mathbf{v} \mapsto \mathbf{v}\times\boldsymbol{\beta}$ pour $k = 2$ et $v \mapsto \boldsymbol{\beta}v$ pour $k = 3$. Les opérateurs d'advection linéaires $v \mapsto \boldsymbol{\beta}\cdot\nabla v$ et $\mathbf{v} \mapsto \nabla(\mathbf{v}\cdot\boldsymbol{\beta}) + (\nabla\times\mathbf{v})\times\boldsymbol{\beta}$ se représentent ainsi de manière abstraite par l'application

$$\Lambda^k(\Omega) \ni \omega^k \mapsto (d\iota_{\boldsymbol{\beta}} + \iota_{\boldsymbol{\beta}}d) \omega^k \in \Lambda^k(\Omega), \quad (1.15)$$

pour $k = 0$ et $k = 1$, respectivement, avec la convention $\iota_{\boldsymbol{\beta}}\omega^0 = 0$ pour tout $\omega^0 \in \Lambda^0(\Omega)$. L'opérateur défini par (1.15) est plus communément appelé *dérivée de Lie*. À la manière de la dérivée extérieure (cf. Palha *et al.* (2010) ou Abraham *et al.* (1988)), on montre que l'opérateur de contraction est nilpotent d'ordre 2 i.e., qu'il satisfait $\iota_{\boldsymbol{\beta}}\circ\iota_{\boldsymbol{\beta}} = 0$. Le proxy de cette propriété en dimension $d = 3$ n'est rien d'autre qu'une conséquence de la définition du produit scalaire, du produit vectoriel et du produit de colinéarité : pour $k = 2$, on a $\boldsymbol{\beta}\cdot(\mathbf{v}\times\boldsymbol{\beta}) = 0$ pour tout flux $\mathbf{v}$ et tout champ $\boldsymbol{\beta}$ et pour $k = 3$, on a $(\boldsymbol{\beta}v)\times\boldsymbol{\beta} = \mathbf{0}$ pour toute densité v et tout champ $\boldsymbol{\beta}$.

1.2 État de l'art et contributions

L'objectif de cette thèse est l'analyse et la résolution sur maillages polyédriques tridimensionnels des problèmes de transport scalaire et vectoriel lorsque les effets advectifs prédominent

sur les effets diffusifs. On s'intéressera en particulier (mais pas uniquement) aux équations d'advection-réaction d'un scalaire et d'un vecteur en négligeant les termes de diffusion. Cette étude est essentielle car la possibilité de traiter de manière robuste le problème complet avec une diffusion faible requiert la maîtrise du cas limite où la diffusion est nulle.

1.2.1 Analyse des problèmes continus

Un cadre adapté à l'étude mathématique des problèmes linéaires d'advection-réaction scalaire (1.7) et vectoriel (1.13) est celui des systèmes de Friedrichs (1958). Ces systèmes englobent un certain nombre de problèmes physiques, qu'ils soient de nature elliptique (e.g., le problème de diffusion (1.6) ou le problème de Maxwell dans le régime diffusif), hyperbolique (e.g., le problèmes d'advection-réaction scalaire (1.7) ou vectoriel (1.13)) ou bien comportant à la fois des termes elliptiques et hyperboliques (e.g., le problème de diffusion-advection-réaction scalaire (1.4) ou vectoriel (1.9)).

Les travaux récents de Ern & Guermond (2006a,b), de Ern *et al.* (2007) et de Ern & Guermond (2008) ont établi l'existence et l'unicité de la solution faible dans les espaces du graphe associé à $L^2(\Omega)$. Outre le fait de caractériser les bonnes conditions à la limite à imposer, cette analyse met en évidence l'existence d'un tenseur, que nous appelons *tenseur de Friedrichs*, dont la positivité est une condition suffisante pour établir la bonne position du problème. Dans le cas des problèmes elliptiques, cette condition de positivité est généralement trivialement satisfaite pour des coefficients associés à une physique représentative (e.g., des coefficients de diffusion, de perméabilité ou de conductivité positifs). En ce qui concerne les problèmes d'advection-réaction (1.7) et (1.13), cette condition de positivité n'est généralement pas satisfaite pour des champs advectifs et réactifs quelconques. Pour le problème scalaire (1.7), le tenseur de Friedrichs dans $L^2(\Omega)$ est à valeurs scalaires et s'écrit

$$\sigma_{\beta,\mu} = \mu - \frac{1}{2} \nabla \cdot \beta : \Omega \rightarrow \mathbb{R}. \quad (1.16)$$

La condition de positivité correspond alors à l'existence d'un réel $\sigma_0 > 0$, homogène à l'inverse d'une échelle de temps caractéristique des phénomènes d'advection-réaction, tel que $\sigma_{\beta,\mu} \geq \sigma_0$ presque partout dans Ω . De manière analogue, le tenseur de Friedrichs associé au problème vectoriel (1.13) dans $L^2(\Omega; \mathbb{R}^3)$ est à valeurs tensorielles et s'écrit

$$\sigma_{\beta,\mu} = \frac{\nabla \beta + \nabla \beta^T}{2} + \frac{\mu + \mu^T}{2} - \frac{1}{2} (\nabla \cdot \beta) \text{Id} : \Omega \rightarrow \mathbb{R}^{3 \times 3}. \quad (1.17)$$

La condition de positivité correspond alors à l'existence d'un réel $\aleph_0 > 0$ (également homogène à l'inverse d'un temps) tel que $\aleph \geq \aleph_0$ presque partout dans Ω , où $\aleph$ correspond à la plus petite valeur propre du tenseur symétrique $\sigma_{\beta,\mu}$.

Les conditions de positivité ci-dessus limitent la portée de l'analyse mathématique car elles ne sont pas toujours satisfaites. Dans le cas du problème (1.7) et en l'absence de terme réactif μ , la condition de positivité impose notamment que le champ advectif soit strictement contractant, i.e., qu'il satisfasse $-\nabla \cdot \beta \geq \sigma_0 > 0$ presque partout dans Ω . Il ne peut alors par exemple être constant. De manière analogue, pour le problème d'advection vectorielle

$$(\beta \cdot \nabla) \mathbf{u} = \mathbf{s} \quad \text{dans } \Omega, \quad (1.18a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-, \quad (1.18b)$$

obtenu à partir du problème (1.11) pour un tenseur de diffusion nul et en choisissant $\mu = -\nabla \beta^T$ si bien que $\sigma_{\beta,\mu} = -\frac{1}{2} (\nabla \cdot \beta) \text{Id}$, la condition de positivité contraint également le champ β à être strictement contractant.

Contributions. Dans cette thèse, nous identifions les tenseurs de Friedrichs relatifs aux problèmes (1.7) et (1.13) dans les espaces du graphe associés aux espaces de Lebesgue $L^p(\Omega)$ pour $p \in (1, \infty]$. Dans le cas où la plus petite valeur propre des tenseurs (1.16) et (1.17) prend des valeurs nulles ou raisonnablement négatives, nous introduisons à la place les tenseurs

$$\sigma'_{\beta,\mu} = \zeta \sigma_{\beta,\mu} - \frac{1}{2} \beta \cdot \nabla \zeta : \Omega \rightarrow \mathbb{R} \quad \text{et} \quad \sigma'_{\beta,\mu} = \zeta \sigma_{\beta,\mu} - \frac{1}{2} \beta \cdot \nabla \zeta \text{Id}_{\mathbb{R}^3} : \Omega \rightarrow \mathbb{R}^{3 \times 3}, \quad (1.19)$$

pour $p = 2$, fonction d'un potentiel $\zeta : \Omega \rightarrow \mathbb{R}$. Dans ce cas, nous établissons les conditions suffisantes permettant d'établir l'unicité de la solution faible pour $p \in (1, \infty]$ et nous montrons l'existence de la solution faible pour $p \in (1, 2]$.

1.2.2 Résolution du problème d'advection-réaction scalaire

Dans un domaine Ω de $\mathbb{R}^3$, le problème d'advection-réaction scalaire consiste à déterminer la fonction $u : \Omega \rightarrow \mathbb{R}$ solution du système

$$\beta \cdot \nabla u + \mu u = s \quad \text{dans } \Omega, \quad (1.20a)$$

$$u = u_D \quad \text{sur } \partial\Omega^-. \quad (1.20b)$$

Pour des maillages généraux, la méthode des volumes finis permet de résoudre ce problème en considérant un degré de liberté sur chaque cellule du maillage et en approchant les termes de flux sur les faces du maillage par des méthodes upwind. Ces méthodes, très utilisées dans l'industrie, ont un coût relativement faible, sont faciles à implémenter mais sont peu précises, avec typiquement une convergence de l'erreur en norme L^2 à l'ordre $\frac{1}{2}$. Les méthodes *discontinuous Galerkin* (dG), introduites pour les équations de la neutronique par Reed & Hill (1973) et analysées pour la première fois par Lesaint & Raviart (1974) (voir également Johnson & Pitkäranta (1986) et Ern & Guermond (2006a,b)), étendent les méthodes volumes finis en considérant des polynômes de degré $k \geq 0$ au sein de chaque cellule, i.e., en considérant $\frac{(k+d)!}{k!d!}$ degrés de liberté par cellule en dimension d . Pour des solutions suffisamment régulières, ces méthodes fournissent typiquement une estimation de l'erreur en norme L^2 à l'ordre $k + \frac{1}{2}$ et une estimation de l'erreur sur la dérivée advective en norme L^2 à l'ordre k . Néanmoins, ces méthodes ont souvent un coût prohibitif pour des applications industrielles.

Une alternative aux méthodes dG permettant d'approcher à l'ordre élevé la solution du problème d'advection-réaction (1.7) est la méthode des éléments finis H^1 -conformes d'ordre k sur maillages formés de simplexes. Ces méthodes sont basées sur la formulation de Galerkin standard (où l'espace des solutions coïncide avec celui des fonctions tests) à laquelle on rajoute un terme de stabilisation symétrique, positif et consistant de manière exacte ou asymptotique. Parmi les méthodes les plus récentes, citons la stabilisation par viscosité de sous-maille proposée par Guermond (1999), la méthode de projection locale LPS proposée par Becker & Braack (2001) (voir également Matthies *et al.* (2007, 2008)), ainsi que la méthode CIP introduite par Burman & Hansbo (2004) et Burman (2005), consistant à pénaliser sur les faces du maillage les sauts de la dérivée advective (voir également Burman & Ern (2007) pour l'extension aux systèmes de Friedrichs). Pour des solutions suffisamment régulières, toutes ces méthodes fournissent des résultats de convergence similaires aux méthodes dG.

Contributions. Dans cette thèse, nous considérons dans un premier temps un schéma équivalent aux méthodes volumes finis sur maillages polyédriques considérant les cellules duales associées aux sommets du maillage comme volumes de contrôle. Les degrés de liberté sont associés aux sommets du maillage et l'on obtient, pour des solutions suffisamment régulières, une estimation de l'erreur en norme L^2 à l'ordre $\frac{1}{2}$ lorsque le tenseur de Friedrichs satisfait la condition de positivité. Une contribution originale (voir aussi les travaux de Deuring *et al.* (2015) et de Ayuso & Marini (2009)) est l'analyse dans le cas où le tenseur de Friedrichs prend des valeurs nulles, si bien qu'il ne satisfait plus la condition de positivité. Dans un second temps,

nous proposons un nouveau schéma sur maillages polyédriques issu de l'approche par éléments finis stabilisés H^1 -conformes. Les degrés de libertés sont toujours associés aux sommets du maillage et l'on obtient pour des solutions suffisamment régulières, une estimation de l'erreur en norme L^2 à l'ordre $\frac{3}{2}$ et de l'erreur sur la dérivée advective en norme L^2 à l'ordre 1.

1.2.3 Résolution du problème de diffusion-advection-réaction scalaire

Dans un domaine Ω de $\mathbb{R}^3$, le problème de diffusion-advection-réaction scalaire consiste à déterminer la fonction $u : \Omega \rightarrow \mathbb{R}$ solution du système

$$-\nabla \cdot (\lambda \nabla u) + \beta \cdot \nabla u + \mu u = s \quad \text{dans } \Omega, \quad (1.21a)$$

$$u = u_D \quad \text{sur } \partial\Omega, \quad (1.21b)$$

quel que soit la valeur du nombre de Péclet (1.5) (et lorsque les effets réactifs sont au plus du même ordre de grandeur que les effets advectifs). Bien que le cas de la réaction dominante ne soit pas étudié dans cette thèse, signalons néanmoins les travaux de Apel & Lube (1998), de Xenophontos & Fulton (2003) et de Heuer & Karkulik (2015) étudiant entre autres la transition entre la diffusion et la réaction.

Une idée afin d'approcher la solution du problème (1.21) consiste à coupler l'une des approches citée dans la Section 1.2.2 pour les termes d'advection-réaction avec une approximation par éléments finis du terme de diffusion. Parmi les travaux traitant ce sujet sur maillages formés de simplexes, citons par exemple ceux de Ohmori & Ushijima (1984) et de Angot *et al.* (1998) utilisant les volumes duaux associés aux arêtes du maillage en dimension 2 ainsi que les travaux de Baba & Tabata (1981) et de Risch (1990) utilisant les volumes duaux associés aux sommets du maillages. De manière générale, Roos *et al.* (2008) proposent un bon aperçu des différentes techniques utilisant la méthode des différences finies et des éléments finis, disponibles dans la littérature afin de résoudre sur maillages classiques le problème de diffusion-advection (1.21).

Pour des maillages généraux, un certain nombre de méthodes numériques sont également disponibles. Appartenant à la famille des *gradient schemes* (cf. Droniou *et al.* (2016) pour les problèmes de diffusion), les schémas *Vertex Approximate Gradient* (VAG) proposés par Guichard (2011) et par Eymard *et al.* (2012a,b), permettent de résoudre à l'ordre bas le problème (1.21) en considérant des degrés de liberté associés à la fois aux cellules et aux sommets du maillage. Dans une approche différente, les méthodes hybrides considèrent des degrés de libertés associés aux faces et aux cellules du maillage. À l'ordre bas, citons les méthodes *Hybrid Mimetic Mixed* (HMM) introduites par Droniou *et al.* (2010), Droniou (2010) et par Beirão da Veiga *et al.* (2011), combinant les approches *Mimetic Finite Difference* (MFD) proposée par Hyman *et al.* (2002), *Hybrid Finite Volume* (HFV) proposée par Eymard *et al.* (2010) et *Mixed Finite Volume* (MFV) proposée par Droniou & Eymard (2006). Pour des méthodes d'ordre $k \geq 1$, citons les méthodes *Hybridizable discontinuous Galerkin* (HdG) introduites par Cockburn *et al.* (2009) ainsi que les méthodes *Hybrid High-Order* proposées récemment par Di Pietro & Ern (2015), où les inconnues cellules sont statiquement condensées, afin de ne considérer que $\frac{(k+d-1)!}{k!(d-1)!}$ degrés de liberté par face en dimension d . Une autre méthode d'ordre $k \geq 1$ est la méthode des *Virtual Elements* (VEM) proposée par Beirão da Veiga *et al.* (2013) pour le problème de diffusion et par Beirão da Veiga *et al.* (2016b,a) pour le problème de diffusion-advection dans le régime diffusif.

Contributions. Dans cette thèse, nous proposons un nouveau schéma à partir des travaux de Bonelle & Ern (2014a) traitant le terme diffusif, afin d'approcher la solution du problème de transport (1.21) en fonction d'un nombre de Péclet local Π_h associé aux échelles de maille, de manière à retrouver dans les cas limites, le schéma d'ordre $\frac{1}{2}$ présenté dans la Section 1.2.2 pour le problème d'advection-réaction (1.20) lorsque $\Pi_h = +\infty$ et le schéma associé au problème de diffusion de Bonelle & Ern (2014a) lorsque $\Pi_h = 0$ avec des conditions à la limite à la Nitsche (1971). Les degrés de liberté sont associés aux sommets du maillage polyédrique et

nous obtenons, pour des solutions suffisamment régulières et pour des tenseurs de Friedrichs à valeurs positives ou nulles, une estimation de l'erreur en norme L^2 convergeant à l'ordre $\frac{1}{2}$ si les effets advectifs dominent et à l'ordre 1 si les effets diffusifs dominent.

1.2.4 Résolution du problème de transport vectoriel

Dans un domaine Ω de $\mathbb{R}^3$, le problème de transport vectoriel consiste à déterminer la fonction $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ solution du système

$$\nabla \times (\lambda \nabla \times \mathbf{u}) + \nabla(\beta \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \beta + \mu \mathbf{u} = \mathbf{s} \quad \text{dans } \Omega, \quad (1.22a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{sur } \partial\Omega^-, \quad (1.22b)$$

$$\mathbf{u} \times \mathbf{n} = \mathbf{u}_D \times \mathbf{n} \quad \text{sur } \partial\Omega \setminus \partial\Omega^-. \quad (1.22c)$$

Nous nous concentrons ici sur l'étude de ce problème dans le régime advectif, donné par (1.13) dans le cas limite où la diffusion λ s'annule uniformément. Un traitement du problème dans le régime de diffusion pure, lorsque $\beta = \mathbf{0}$ et $\mu = \mathbf{0}$, peut être envisagé à partir des travaux de Bonelle & Ern (2014b) concernant le problème de Stokes.

Les méthodes d'approximation du problème modèle (1.22) sont peu nombreuses dans la littérature. Le plus souvent, seul un des deux termes de l'opérateur d'advection vectoriel est considéré en présence d'un terme de diffusion. C'est le cas notamment du problème modélisant l'induction électromagnétique dans un plasma de vitesse β , où l'on souhaite déterminer le champ électromagnétique $(\mathbf{E}, \mathbf{B})$ satisfaisant le système

$$\nabla \times \mathbf{E} = \mathbf{f} \quad \text{dans } \Omega, \quad (1.23a)$$

$$\mathbf{E} - \lambda^{-1} \nabla \times \mathbf{B} - \mathbf{B} \times \beta = \mathbf{g} \quad \text{dans } \Omega, \quad (1.23b)$$

avec λ le coefficient de permittivité du milieu. Sur des maillages tétraédriques, Guermond & Mineev (2003) proposent une alternative à l'utilisation classique des éléments finis d'arête afin de résoudre ce problème en utilisant des éléments finis H^1 -conformes d'ordre bas.

Également motivé par des problèmes issus de l'électromagnétisme, Heumann (2011) propose et analyse dans le cas instationnaire des méthodes Eulériennes et semi-Lagrangiennes résolvant le problème d'advection vectorielle (1.13) à partir d'éléments $H(\mathbf{curl})$ -conformes sur maillages formés de simplexes. Les schémas proposés dans ces travaux font notamment écho aux travaux de Bossavit (2003) consacrés à l'approximation de la dérivée de Lie d'une forme différentielle par extrusion des éléments du maillage. Cette approche est également considérée dans les travaux de Mullen *et al.* (2011), où la solution du problème d'advection vectorielle (1.13) est approchée par une méthode d'extrusion des arêtes d'un maillage formé de quadrangles. Signalons finalement les travaux de Palha (2013) abordant également l'approximation de ce problème sur des maillages formés de quadrangles en approchant l'opérateur de contraction à partir de son opérateur adjoint, appelé produit extérieur, dans le cadre des méthodes *Mimetic Spectral Elements* (MSE).

Contributions. Dans cette thèse, nous proposons un nouveau schéma approchant la solution du problème (1.13) en considérant un seul degré de liberté scalaire sur chacune des arêtes d'un maillage polyédrique. Le choix de ce positionnement des degrés de liberté permet non seulement d'être cohérent avec le positionnement de la vitesse dans le schéma CDO de Bonelle & Ern (2014b) pour le problème Stokes, mais également avec la nature physique de la circulation $\mathbf{u}$, correspondant au proxy d'une 1-forme et donc discrétisée sur les 1-co-chaines (i.e., sur les arêtes dans le formalisme de la topologie algébrique). Une estimation de l'erreur est obtenue en norme L^2 à l'ordre $\frac{1}{2}$ pour des solutions suffisamment régulières et pour des tenseurs de Friedrichs prenant des valeurs positives, nulles ou négatives. On notera la différence conceptuelle de notre schéma avec le schéma volumes finis sur le maillage dual associé aux arêtes du maillage, considérant comme degrés de liberté les trois composantes de l'inconnue sur chaque volume dual.

1.3 Organisation de la thèse

Cette thèse est composée de 7 chapitres. Les Chapitres de 2 à 6, résumés ci-dessous, sont globalement indépendants afin de permettre une lecture non-linéaire et par morceaux. Le Chapitre 7 collecte les différents outils et résultats relatifs à l'utilisation des maillages polyédriques et à l'analyse de nos schémas. Pour faciliter la lecture, les principaux résultats obtenus dans ce dernier chapitre sont succinctement rappelés au cours des chapitres de 2 à 6.

Chapitre 2) Analyse des problèmes continus. À partir de l'analyse Hilbertienne proposée par Ern & Guermond (2006a) dans l'espace du graphe de $L^2(\Omega)$, on étend l'analyse de l'unicité de la solution des problèmes (1.7) et (1.13) dans l'espace du graphe associé à $L^p(\Omega)$ et $\mathbf{L}^p(\Omega)$, respectivement, pour $p \in (1, +\infty)$. L'analyse relative à l'existence d'une solution est quant à elle menée pour des exposants $p \in (1, 2]$. La bonne position de ces problèmes pour $p \in (1, 2]$ permet par exemple de considérer des termes sources peu réguliers dans $L^p(\Omega)$ et $\mathbf{L}^p(\Omega)$, respectivement. Cette analyse est effectuée lorsque la plus petite valeur propre du tenseur de Friedrichs dans $L^p(\Omega)$ et $\mathbf{L}^p(\Omega)$, respectivement, prend des valeurs positives, nulles ou raisonnablement négatives. L'extension de l'analyse dans ces deux derniers cas est notamment envisageable à partir des travaux de Devinatz *et al.* (1974) et de Azérad & Pousin (1996). Dans le cas où pour tout point du domaine Ω , il existe une caractéristique *backward* partant de ce point et atteignant $\partial\Omega^-$ en un temps fini, ces travaux introduisent une fonction scalaire $\zeta : \Omega \rightarrow \mathbb{R}$ permettant de construire un nouveau tenseur satisfaisant une condition de positivité.

Chapitre 3) Schéma d'ordre bas pour le problème d'advection-réaction scalaire. Un schéma approchant la solution du problème modèle (1.7) est proposé sur des maillages polyédriques avec des degrés de liberté scalaires associés aux sommets. Ce schéma utilise le formalisme des schémas CDO introduit par Bonelle (2014) et discrétise la dérivée advective $v \mapsto \beta \cdot \nabla v$ en dissociant l'opérateur différentiel $v \mapsto \nabla v$ et l'opérateur de contraction $\mathbf{v} \mapsto \beta \cdot \mathbf{v}$. Nous introduisons ainsi la notion d'*opérateur de contraction discret*, construit de manière à satisfaire au niveau discret la formule de Leibniz et la formule d'intégration par parties. Les conditions à la limite sont imposées faiblement sur le bord entrant du domaine et l'analyse discrète du schéma s'inspire de l'analyse des systèmes de Friedrichs. Cette approche permet de retrouver en particulier le schéma volumes finis upwind où les volumes de contrôle correspondent aux cellules duales associées aux sommets du maillage. Lorsque le tenseur de Friedrichs satisfait la condition de positivité, le schéma est stable par coercivité. Lorsque ce tenseur prend des valeurs nulles, le schéma est inf-sup stable pour un maillage suffisamment raffiné et inconditionnellement stable si le champ advectif est à divergence nulle. Une estimation de l'erreur discrète quasi-optimale à l'ordre $\frac{1}{2}$ est obtenue en norme de stabilité. Les résultats théoriques sont illustrés par deux cas tests sur des maillages polyédriques.

Chapitre 4) Schéma d'ordre bas pour le problème de diffusion-advection-réaction scalaire. À partir du schéma CDO obtenu pour le problème (1.7) dans le Chapitre 3 et du schéma CDO proposé par Bonelle & Ern (2014a) pour le problème de diffusion (1.6), un nouveau schéma robuste en fonction du nombre de Péclet est obtenu afin d'approcher la solution du problème de diffusion-advection-réaction (1.4) sur maillages polyédriques. Ce schéma est présenté dans le formalisme des schémas CDO et utilise des degrés de liberté associés aux sommets du maillage. Afin de traiter de manière cohérente les conditions à la limite quel que soit le nombre de Péclet, le schéma proposé par Bonelle & Ern (2014a) est étendu afin d'imposer faiblement la condition de Dirichlet par une méthode de pénalisation à la Nitsche (1971). Ce schéma pour la diffusion est alors couplé au schéma associé au problème d'advection-réaction obtenu dans le Chapitre 3 en utilisant une méthode de décentrement en fonction du nombre de Péclet local Π_h . En l'absence d'advection et de réaction, i.e., lorsque $\Pi_h = 0$, on retrouve le schéma associé au problème de diffusion pure avec des conditions à la limite faible, tandis

que lorsque $\Pi_h = +\infty$, on retrouve le schéma d'advection-réaction proposé dans le Chapitre 3. La stabilité du schéma est prouvée par coercivité lorsque le tenseur de Friedrichs est positif. Lorsque celui-ci peut prendre des valeurs nulles, la stabilité uniforme en fonction du nombre de Péclet est établie sous la forme d'une condition inf-sup, satisfaite si le maillage est suffisamment raffiné. Une estimation de l'erreur discrète dépendante du nombre de Péclet est obtenue, de telle sorte que le schéma converge à l'ordre 1 en norme de stabilité dans le régime diffusif et à l'ordre $\frac{1}{2}$ dans le régime advectif. Des simulations numériques sur maillages polyédriques permettent d'illustrer les résultats théoriques.

Chapitre 5) Extension des schémas pour le problème de transport scalaire. Un nouveau schéma pour le problème de advection-réaction (1.7) est proposé sur maillages polyédriques avec des degrés de liberté attachés aux sommets du maillage. Ce schéma est plus précis que celui proposé dans le Chapitre 3. L'idée est d'introduire en plus des degrés de liberté sommets, des degrés de liberté scalaires associés aux cellules du maillage. Ces degrés de liberté supplémentaires ne sont pas couplés entre eux, ce qui permet de les éliminer à moindre coût par une méthode de condensation statique (également appelée complément de Schur). Le schéma est stabilisé par une approche inspirée des méthodes *continuous interior penalty* (CIP), consistant à pénaliser les sauts de la dérivée advective d'une certaine fonction reconstruite à partir des degrés de liberté sommets et cellules, à travers les sous-faces contenues dans les cellules du maillage. Le schéma est alors inconditionnellement inf-sup stable si la condition de positivité sur le tenseur de Friedrichs est satisfaite, et l'on contrôle en plus de la norme de coercivité, la norme L^2 de la dérivée advective. Une estimation de l'erreur en norme L^2 de l'erreur est obtenue à l'ordre $\frac{3}{2}$ tandis que l'erreur en norme L^2 sur la dérivée advective converge à l'ordre 1. Ce schéma est finalement couplé à une discretisation du terme diffusif afin de considérer une approximation du problème de transport complet avec un traitement des conditions à la limite à la Nitsche (1971). Des résultats numériques illustrent les résultats théoriques et sont comparés aux résultats des simulations précédemment obtenus dans les Chapitres 3 et 4.

Chapitre 6) Schéma d'ordre bas pour le problème d'advection-réaction vectoriel. Motivés par la discrétisation du problème de Stokes sur maillages polyédriques considérée par Bonelle & Ern (2014b), nous proposons un nouveau schéma permettant d'approcher la solution du problème d'advection-réaction vectoriel (1.13) à l'aide de degrés de liberté scalaires associés aux arêtes du maillage. Le schéma est basé sur la formulation de Galerkin standard du problème. À l'aide d'un opérateur permettant de reconstruire un champ vectoriel à partir des degrés de liberté scalaires associés aux arêtes, le schéma est stabilisé par une méthode de pénalisation des sauts au travers des sous-faces induites par les volumes diamants entourant les arêtes du maillage. Le schéma est stable par coercivité dans la norme de stabilité si la plus petite valeur propre du tenseur de Friedrichs (1.17) est strictement positive et inf-sup stable pour des maillages suffisamment raffinés dans le cas où cette valeur propre prend des valeurs nulles ou raisonnablement négatives. Dans le cas du problème (1.18), cette restriction sur le pas du maillage disparaît. Une estimation *a priori* de l'erreur discrète quasi-optimale à l'ordre p est obtenue pour des solutions (peu régulières) appartenant à $\mathbf{W}^{1,p+\frac{3}{2}}(\Omega)$ avec $p \in (0, \frac{1}{2}]$. Des résultats numériques sur maillages polyédriques sont présentés afin d'illustrer les résultats théoriques.

Publications. Les travaux de cette thèse ont fait l'objet de trois publications soumises à des revues internationales.

- 1) P. Cantin & A. Ern. "*Vertex-based compatible Discrete Operator Schemes on Polyhedral meshes for Advection-Diffusion Equations*" publié dans Computational Methods in Applied Mathematics. 2016.
- 2) P. Cantin, J. Bonelle, E. Burman & A. Ern. "*A vertex-based scheme on polyhedral meshes for advection-reaction equations with sub-mesh stabilization*" à paraître dans Computers and Mathematics with Applications. 2016.
- 3) P. Cantin & A. Ern. "*Edge-based low-order scheme on polyhedral meshes for vector advection-reaction equations*", soumis à M2AN. 2016.

Chapter 2

On the Friedrichs positivity assumption for advection-reaction problems in Banach spaces

Contents

2.1	Continuous settings	16
2.1.1	Functional spaces	16
2.1.2	Banach-Nečas-Babuška (BNB) theorem	17
2.2	Scalar advection-reaction problem	17
2.2.1	The graph space	17
2.2.2	Well-posedness	19
2.3	Vector advection-reaction problem	22
2.3.1	The graph space	22
2.3.2	Well-posedness	23

Let Ω be a domain of $\mathbb{R}^3$ with Lipschitz-continuous boundary $\partial\Omega$. We are looking for $u : \Omega \rightarrow \mathbb{R}$ and $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ solving the following first-order boundary-value problems:

$$\boldsymbol{\beta} \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (2.1a)$$

$$u = 0 \quad \text{a.e. on } \partial\Omega^-, \quad (2.1b)$$

and

$$\nabla(\boldsymbol{\beta} \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \boldsymbol{\beta} + \boldsymbol{\mu} \mathbf{u} = \mathbf{s} \quad \text{a.e. in } \Omega, \quad (2.2a)$$

$$\mathbf{u} = \mathbf{0} \quad \text{a.e. on } \partial\Omega^-. \quad (2.2b)$$

Here, $\boldsymbol{\beta}$ denotes a smooth $\mathbb{R}^3$ -valued vector field on Ω , say Lipschitz-continuous, and μ and $\boldsymbol{\mu}$ are two bounded reaction terms taking values in $\mathbb{R}$ and $\mathbb{R}^{3 \times 3}$, respectively. The inflow boundary $\partial\Omega^-$ is defined by $\partial\Omega^- = \{\mathbf{x} \in \partial\Omega \mid \boldsymbol{\beta}(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) < 0\}$, with $\mathbf{n}$ the outward unit normal to $\partial\Omega$. These problems (or variants) have been studied many times in the literature. We mention the pioneering work of Bardos (1970) and of Beirão da Veiga (1988) for the analysis of the well-posedness in smooth domains with smooth model parameters and the work of DiPerna & Lions (1989) when the problem is expressed in the whole space with irregular model parameters. More recently, Girault & Tartar (2010) proved (using a viscous and a Yosida regularization technique), for all $p \geq 2$, the well-posedness of these problems in $L^p(\Omega)$ if $\boldsymbol{\beta} \in \mathbf{W}^{1,2}(\Omega)$ and also the $W^{1,p}(\Omega)$ -regularity of the solution of (2.1) if $s \in W^{1,p}(\Omega)$ and if $\boldsymbol{\beta} \in \mathbf{W}^{1,\infty}(\Omega)$ is sufficiently small. Approximations of these problems in Banach spaces are extremely few; let

us mention the work of Guermond (2005) presenting a finite element approximation in L^1 and also the reweighted least-squares method of Jiang (1993).

This chapter analyzes the well-posedness of problems (2.1) and (2.2) when a weak solution is sought in Lebesgue graph spaces of exponent $p \in (1, \infty)$. Observing that these problems define two Friedrichs' systems, it is well-known from Ern & Guermond (2006a) or from Ern *et al.* (2007) that, for $p = 2$, the well-posedness for the problem (2.1) is a consequence of the positivity of the $\mathbb{R}$ -valued Friedrichs tensor

$$\sigma_{\beta, \mu; p} := \mu - \frac{1}{p} \nabla \cdot \beta, \quad (2.3)$$

and for the problem (2.2) of the positivity of the lowest eigenvalue of the $\mathbb{R}^{3 \times 3}$ -valued Friedrichs tensor

$$\sigma_{\beta, \mu; p} := \frac{\mu + \mu^T}{2} + \frac{\nabla \beta + \nabla \beta^T}{2} - \frac{1}{p} (\nabla \cdot \beta) \text{Id}. \quad (2.4)$$

The first contribution of this chapter concerns the well-posedness of these two problems in Banach graph spaces. Following the analysis of Friedrichs' systems in Hilbert spaces and assuming that the above tensors (2.3) and (2.4) are positive, we prove the uniqueness of the weak solution of these two problems for $p \in (1, \infty)$ while the existence is obtained if $p \in (1, 2]$. The second part of this work is devoted to the analysis when these positivity assumptions are not satisfied. Assuming the existence of a so-called potential (whose existence follows from properties of the vector field β in Ω , see Devinatz *et al.* (1974)), we prove that one may generalize the Friedrichs positivity assumptions so as to consider tensors whose lowest eigenvalue is null or even reasonably negative.

This chapter is organized as follows. First, we introduce some notation and we recall the classical statement of the Banach-Nečas-Babuška (BNB) theorem. Section 2.2 is concerned with the scalar-valued problem (2.1) and Section 2.3 with the vector-valued problem (2.2).

2.1 Continuous settings

In this section, we introduce some notation, in particular concerning functional spaces. In order to stay general, we will write $d = 3$ the dimension of Ω .

2.1.1 Functional spaces

For all $p \in [1, \infty]$, $p' \in [1, \infty]$ denotes its conjugate real number such that $\frac{1}{p} + \frac{1}{p'} = 1$. $(x_1, \dots, x_d)$ denotes the classical Cartesian basis of the Euclidean space $\mathbb{R}^d$ and ∂_i is the weak derivative in the direction i , for all $i \in \llbracket 1, d \rrbracket$. Classical fonts are used for $\mathbb{R}$ -valued quantities and boldface fonts are used for $\mathbb{R}^d$ or $\mathbb{R}^{d \times d}$ -valued quantities. The inner product in $\mathbb{R}^d$ is denoted by $\cdot$, the cross product by $\times$ and the tensorial product by $\otimes$. To alleviate the notation, $|\cdot|$ denotes either the Lebesgue measure of a set, the absolute value of a real number, the Euclidean norm of a vector or the Frobenius norm of a matrix.

For all $q \in \{1, d\}$ and for all $\omega \subseteq \Omega$, the Banach space $L^p(\omega; \mathbb{R}^q)$ collects all Lebesgue measurable functions $v : \omega \rightarrow \mathbb{R}^q$ whose the Euclidean norm raised to the power p is Lebesgue integrable. We define the norm

$$\|v\|_{L^p(\omega; \mathbb{R}^q)} = \left(\int_{\omega} |v|^p \right)^{\frac{1}{p}}, \quad \text{for all } 1 \leq p < \infty, \quad (2.5)$$

and $\|v\|_{L^\infty(\omega; \mathbb{R}^q)} = \text{ess sup}_{x \in \omega} |v(x)|$. Similarly, for all $s \in \mathbb{N}$, $W^{s,p}(\omega; \mathbb{R}^q)$ denotes the Banach space collecting all Lebesgue measurable functions $v : \omega \rightarrow \mathbb{R}^q$ whose derivatives of global order up to s belong to $L^p(\omega; \mathbb{R}^q)$. We define the semi-norm

$$|v|_{W^{s,p}(\omega; \mathbb{R}^q)} = \left(\sum_{\alpha_1 + \dots + \alpha_d = s} \|\partial^\alpha v\|_{L^p(\omega; \mathbb{R}^q)}^p \right)^{\frac{1}{p}}, \quad \text{for all } 1 \leq p < \infty, \quad (2.6)$$

and $|v|_{W^{s,\infty}(\omega;\mathbb{R}^q)} = \max_{\alpha_1+\dots+\alpha_d=s} \|\partial^\alpha v\|_{L^\infty(\omega;\mathbb{R}^q)}$ with the d -uplet $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}^d$ and the weak derivative of order α defined as $\partial^\alpha = \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}$. In essence, we have $v \in W^{s,p}(\omega;\mathbb{R}^q)$ if and only if $|v|_{W^{t,p}(\omega;\mathbb{R}^q)} < \infty$ for all $t \leq s$. We use the following notations:

$$L^p(\omega) \equiv L^p(\omega;\mathbb{R}) \quad \text{and} \quad \mathbf{L}^p(\omega) \equiv L^p(\omega;\mathbb{R}^d), \quad \text{for all } p \in [1, \infty],$$

$$W^{s,p}(\omega) \equiv W^{s,p}(\omega;\mathbb{R}) \quad \text{and} \quad \mathbf{W}^{s,p}(\omega) \equiv W^{s,p}(\omega;\mathbb{R}^d), \quad \text{for all } p \in [1, \infty] \text{ and for all } s \in \mathbb{N}.$$

Conventionally, we set $W^{0,p}(\omega;\mathbb{R}^q) = L^p(\omega;\mathbb{R}^q)$. The space of k -continuously differentiable functions on ω are denoted by $\mathcal{C}^k(\omega;\mathbb{R}^q)$ and $\mathcal{C}_0^k(\omega;\mathbb{R}^q)$ those functions that are compactly supported in ω . The space of Lipschitz continuous functions is denoted by $\text{Lip}(\omega;\mathbb{R}^q)$ and we recall that we always have $\text{Lip}(\omega;\mathbb{R}^q) \subset W^{1,\infty}(\omega;\mathbb{R}^q)$, while the reverse inclusion holds for domains ω with smooth enough boundary, see Grisvard (2011).

Generic constants are denoted by $\mathbf{C}_\bullet$ and are always assumed to be independent of any physical quantity, such as the model parameters or the size of the mesh for discrete problems. The notation $A \lesssim B$ means that $A \leq \mathbf{C}_\bullet B$, where $\mathbf{C}_\bullet$ may change from one inequality to another.

2.1.2 Banach-Nečas-Babuška (BNB) theorem

Hereafter, we recall the classical statement of the Banach-Nečas-Babuška Theorem. Consider the following abstract variational problem:

$$\text{Find } u \in U \quad \text{s.t.} \quad a(u, v) = \langle f, v \rangle_{V', V}, \quad \forall v \in V, \quad (2.7)$$

where U and V are two Banach spaces equipped with norms $\|\cdot\|_U$ and $\|\cdot\|_V$, respectively, V is reflexive, $a \in \mathcal{L}(U \times V; \mathbb{R})$, $f \in V'$ and $\langle \cdot, \cdot \rangle_{V', V}$ is the duality pairing between $V' \equiv \mathcal{L}(V; \mathbb{R})$ and V . A necessary and sufficient condition for (2.7) to be well-posed is given by the BNB theorem, see e.g., Ern & Guermond (2004).

Theorem 2.1 (Banach-Nečas-Babuška). *The problem (2.7) is well-posed if and only if:*
(BNB1) *There exists $\mathbf{C}_{\text{BNB}} > 0$ such that*

$$\mathbf{C}_{\text{BNB}} \|v\|_U \leq \sup_{w \in V \setminus \{0\}} \frac{a(v, w)}{\|w\|_V}, \quad \forall v \in U. \quad (2.8)$$

(BNB2) *For all $w \in V$, $(\forall v \in U, a(v, w) = 0) \implies (w = 0)$.*

Remark 2.2 (Finite-dimensional case). *In this thesis, the well-posedness of our numerical schemes is obtained owing to (BNB1) since the two statements of Theorem 2.1 are equivalent if U and V are two finite-dimensional spaces satisfying $\dim U = \dim V$.*

2.2 Scalar advection-reaction problem

This section analyzes the well-posedness of the continuous problem (2.1) in Lebesgue graph spaces and generalized the sign condition on the Friedrichs tensor $\sigma_{\beta, \mu; p}$ defined by (2.3).

2.2.1 The graph space

Let $p \in (1, \infty)$. The graph space associated with (2.1) is defined by

$$V_{\beta; p}(\Omega) := \{v \in L^p(\Omega) \mid \beta \cdot \nabla v \in L^p(\Omega)\}, \quad (2.9)$$

and is equipped with the norm $\|v\|_{V_{\beta; p}(\Omega)} := (\|v\|_{L^p(\Omega)}^p + \|\beta \cdot \nabla v\|_{L^p(\Omega)}^p)^{\frac{1}{p}}$ for all $v \in V_{\beta; p}(\Omega)$. This space defines a reflexive Banach space owing to the first and the second Clarkson inequalities (see Brezis (2010)) where for all $v \in V_{\beta; p}(\Omega)$, $\beta \cdot \nabla v \in L^p(\Omega)$ means that the linear form

$$\mathcal{C}_0^\infty(\Omega) \ni \varphi \mapsto - \int_{\Omega} v \nabla \cdot (\beta \varphi), \quad (2.10)$$

is bounded in $L^{p'}(\Omega)$, so that $\beta \cdot \nabla v$ is the Riesz representative of (2.10) in $L^p(\Omega)$. To specify the meaning of the trace of a function in $V_{\beta;p}(\Omega)$, we introduce the space $L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$ given by

$$L^p(|\beta \cdot \mathbf{n}|; \partial\Omega) := \left\{ v : \partial\Omega \rightarrow \mathbb{R} \text{ is Lebesgue measurable on } \partial\Omega \mid \int_{\partial\Omega} |\beta \cdot \mathbf{n}| |v|^p < \infty \right\}, \quad (2.11)$$

which is a Banach space when equipped with the norm $\|v\|_{L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)} = (\int_{\partial\Omega} |\beta \cdot \mathbf{n}| |v|^p)^{\frac{1}{p}}$ for all $v \in L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$. As observed by Ern & Guermond (2006a), the existence of traces in $L^2(|\beta \cdot \mathbf{n}|; \partial\Omega)$ for function in $V_{\beta;2}(\Omega)$ is not always guaranteed. In particular, the boundary $\partial\Omega$ must be well-separated with respect to the vector field β in the sense that

$$\text{dist}(\partial\Omega^-, \partial\Omega^+) > 0 \text{ with } \partial\Omega^\pm = \{x \in \partial\Omega \mid \pm \beta(x) \cdot \mathbf{n}(x) > 0\}. \quad (2.12)$$

In this thesis, this condition will always be assumed to be satisfied. Let us adapt the proof of (Ern & Guermond, 2006a, Lemma 3.1) to the general case $p \in (1, \infty)$ to prove the existence of such traces.

Lemma 2.3 (Trace in $L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$). *Let $p \in (1, \infty)$. The trace map $\gamma : C^\infty(\bar{\Omega}) \rightarrow L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$ with $\gamma(\varphi) = \varphi|_{\partial\Omega}$ for all $\varphi \in C^\infty(\bar{\Omega})$, extends continuously to $V_{\beta;p}(\Omega)$, i.e., there exists $\mathcal{C}_\gamma > 0$ such that*

$$\|\gamma(v)\|_{L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)} \leq \mathcal{C}_\gamma \|v\|_{V_{\beta;p}(\Omega)}, \quad \forall v \in V_{\beta;p}(\Omega).$$

Proof. Owing to the separation of the boundary from assumption (2.12), there exist $\psi^+, \psi^- \in C^\infty(\bar{\Omega})$ such that $\psi^+ + \psi^- \equiv 1$ on $\partial\Omega$, $\psi^\pm \geq 0$, $\psi^+|_{\partial\Omega^-} \equiv 0$ and $\psi^-|_{\partial\Omega^+} \equiv 0$. Proceeding as in Ern & Guermond (2006a), we infer that

$$\int_{\partial\Omega} |\beta \cdot \mathbf{n}| |\varphi|^p = \int_{\partial\Omega} \beta \cdot \mathbf{n} |\varphi|^p (\psi^+ - \psi^-) = \int_{\Omega} \nabla \cdot (\beta |\varphi|^p (\psi^+ - \psi^-)), \quad \forall \varphi \in C^\infty(\bar{\Omega}),$$

where we have used the partition of the unity on the boundary and the Stokes formula. Applying now the Leibniz product rule and recalling that $\nabla |\varphi|^p = p|\varphi|^{p-2} \nabla \varphi$, we obtain

$$\int_{\partial\Omega} |\beta \cdot \mathbf{n}| |\varphi|^p = p \int_{\Omega} (\psi^+ - \psi^-) (\beta \cdot \nabla \varphi) |\varphi|^{p-2} + \int_{\Omega} |\varphi|^p \nabla \cdot (\beta (\psi^+ - \psi^-)).$$

Next, Hölder's and Young's inequalities together with the identity $\|\varphi|\varphi|^{p-2}\|_{L^{p'}(\Omega)}^{p'} = \|\varphi\|_{L^p(\Omega)}^p$ yield

$$\int_{\Omega} |(\beta \cdot \nabla \varphi) |\varphi|^{p-2}| \leq \|\beta \cdot \nabla \varphi\|_{L^p(\Omega)} \|\varphi\|_{L^p(\Omega)}^{p/p'} \leq \frac{1}{p} \|\beta \cdot \nabla \varphi\|_{L^p(\Omega)}^p + \frac{1}{p'} \|\varphi\|_{L^p(\Omega)}^p.$$

It follows that $\|\varphi\|_{L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)} \leq \mathcal{C}' (\|\beta \cdot \nabla \varphi\|_{L^p(\Omega)}^p + p \|\varphi\|_{L^p(\Omega)}^p)^{\frac{1}{p}}$ with the constant $\mathcal{C}' = 2^{\frac{1}{p}} (\|\psi^+ - \psi^-\|_{L^\infty(\Omega)} + \|\nabla \cdot (\beta (\psi^+ - \psi^-))\|_{L^\infty(\Omega)})^{\frac{1}{p}}$. Then, observing that $p^{\frac{1}{p}} \leq e^{\frac{p-1}{p}} \leq e$, we obtain

$$\|\varphi\|_{L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)} \leq \mathcal{C}_\gamma \|\varphi\|_{V_{\beta;p}(\Omega)}, \quad \forall \varphi \in C^\infty(\bar{\Omega}),$$

with $\mathcal{C}_\gamma = e\mathcal{C}'$. Finally, recalling that $C^\infty(\bar{\Omega})$ is dense in $V_{\beta;p}(\Omega)$ for all $p \in (1, \infty)$ (see (Jensen, 2004, Theorem 2)), this inequality holds as well for any function in $V_{\beta;p}(\Omega)$. $\square$

Owing to the existence of traces in $L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$, the following integration by parts formulae hold.

Lemma 2.4 (Integration by parts). *Let $p \in (1, \infty)$. Then, for all $v \in V_{\beta;p}(\Omega)$ and for all $w \in V_{\beta;p'}(\Omega)$,*

$$\int_{\Omega} (\beta \cdot \nabla v) w + \int_{\Omega} (\beta \cdot \nabla w) v + \int_{\Omega} (\nabla \cdot \beta) vw = \int_{\partial\Omega} (\beta \cdot \mathbf{n}) vw. \quad (2.13a)$$

In addition, for all $v \in V_{\beta;p}(\Omega)$ and for all $z \in W^{1,\infty}(\Omega)$,

$$\int_{\Omega} (\beta \cdot \nabla v) v |v|^{p-2} z + \frac{1}{p} \int_{\Omega} (\nabla \cdot \beta) |v|^p z + \frac{1}{p} \int_{\Omega} \beta \cdot \nabla z |v|^p = \frac{1}{p} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) |v|^p z. \quad (2.13b)$$

Proof. These formulae follow from the density of $\mathcal{C}^\infty(\overline{\Omega})$ in $V_{\beta;p}(\Omega)$ for all $p \in (1, \infty)$. The first one results from the Leibniz product rule while the second one is a consequence of the identity

$$\beta \cdot \nabla(\varphi|\varphi|^{p-2}z) = \varphi|\varphi|^{p-2}\beta \cdot \nabla z + (p-1)|\varphi|^{p-2}z\beta \cdot \nabla \varphi, \quad (2.14)$$

for all $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$ and for all $z \in W^{1,\infty}(\Omega)$. $\square$

The following proposition identifies some elements in the graph space $V_{\beta;p'}(\Omega)$ if $p \in [2, \infty)$.

Proposition 2.5 (Elements of $V_{\beta;p'}(\Omega)$). *Let $p \in [2, \infty)$. Then, for all $v \in V_{\beta;p}(\Omega)$, $v|v|^{p-2} \in V_{\beta;p'}(\Omega)$.*

Proof. Let $p \in [2, \infty)$ and let $v \in V_{\beta;p}(\Omega)$. Clearly, $v|v|^{p-2} \in L^{p'}(\Omega)$ so that it remains to prove that $\beta \cdot \nabla(v|v|^{p-2}) \in L^{p'}(\Omega)$. Owing to the chain rule, we have $\beta \cdot \nabla(v|v|^{p-2}) = (p-1)|v|^{p-2}\beta \cdot \nabla v$. Hence, using Hölder's inequality with $\frac{1}{p} + \frac{1}{r} = \frac{1}{p'}$ since $p \geq 2$ (so that $r = \frac{p}{p-2} > 1$), we infer that

$$\|\beta \cdot \nabla(v|v|^{p-2})\|_{L^{p'}(\Omega)} \leq (p-1)\|\beta \cdot \nabla v\|_{L^p(\Omega)}\|v|v|^{p-2}\|_{L^r(\Omega)}.$$

Hence, since $\|v|v|^{p-2}\|_{L^r(\Omega)} = \|v\|_{L^p(\Omega)}^{p-2}$, we deduce that $v|v|^{p-2} \in V_{\beta;p'}(\Omega)$. $\square$

2.2.2 Well-posedness

To examine the well-posedness of (2.1), we introduce the bilinear form $a_{\beta,\mu;p} \in \mathcal{L}(V_{\beta;p}^0(\Omega) \times L^{p'}(\Omega); \mathbb{R})$, where $V_{\beta;p}^0(\Omega) := \{w \in V_{\beta;p}(\Omega) \mid w|_{\partial\Omega^-} = 0\}$, and such that for all $v \in V_{\beta;p}^0(\Omega)$ and all $w \in L^{p'}(\Omega)$,

$$a_{\beta,\mu;p}(v, w) := \int_{\Omega} (\beta \cdot \nabla v) w + \int_{\Omega} \mu v w. \quad (2.15)$$

Observe that, for all $p \in (1, \infty)$, $V_{\beta;p}^0(\Omega)$ is a closed subspace of $V_{\beta;p}(\Omega)$ owing to Lemma 2.3. Assuming that $s \in L^p(\Omega)$, the weak formulation of (2.1) in the graph space $V_{\beta;p}^0(\Omega)$ is:

$$\text{Find } u \in V_{\beta;p}^0(\Omega) \text{ s.t. } a_{\beta,\mu;p}(u, v) = \int_{\Omega} s v, \quad \forall v \in L^{p'}(\Omega). \quad (2.16)$$

It is readily seen that if $u \in V_{\beta;p}^0(\Omega)$ solves (2.16), the PDE (2.1a) holds in $L^p(\Omega)$ and the boundary condition (2.1b) holds in $L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$. Note that the boundary conditions are *strongly* enforced in (2.16). To prove the uniqueness of the solution of (2.16) for all $p \in (1, \infty)$, we recall the $\mathbb{R}$ -valued Friedrichs tensor

$$\sigma_{\beta,\mu;p} := \mu - \frac{1}{p} \nabla \cdot \beta, \quad (2.17)$$

and we assume that it satisfies the so-called Friedrichs positivity assumption $(\mathcal{H}_p)$:

$(\mathcal{H}_p)$ $\text{ess inf}_{\Omega} \sigma_{\beta,\mu;p} > 0$. We define the reference time $\tau = (\text{ess inf}_{\Omega} \sigma_{\beta,\mu;p})^{-1}$.

Proposition 2.6 (Uniqueness under $(\mathcal{H}_p)$). *Let $p \in (1, \infty)$ and assume that $(\mathcal{H}_p)$ holds. Then,*

$$a_{\beta,\mu;p}(v, v|v|^{p-2}) \geq \tau^{-1}\|v\|_{L^p(\Omega)}^p, \quad \forall v \in V_{\beta;p}^0(\Omega). \quad (2.18)$$

Proof. Let $v \in V_{\beta;p}^0(\Omega)$. Observing that $v|v|^{p-2} \in L^{p'}(\Omega)$, the quantity $a_{\beta,\mu;p}(v, v|v|^{p-2})$ is well-defined. Owing to the integration by parts formula (2.13b) with $z \equiv 1$ on Ω (so that $\beta \cdot \nabla z \equiv 0$), we infer that

$$a_{\beta,\mu;p}(v, v|v|^{p-2}) = \int_{\Omega} \left(\mu - \frac{1}{p} \nabla \cdot \beta \right) |v|^p + \frac{1}{p} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) |v|^p,$$

whence, using the definition (2.17) of the Friedrichs tensor $\sigma_{\beta,\mu;p}$ and the fact that $v|_{\partial\Omega^-} = 0$, we obtain

$$a_{\beta,\mu;p}(v, v|v|^{p-2}) = \int_{\Omega} \sigma_{\beta,\mu;p} |v|^p + \frac{1}{p} \int_{\partial\Omega^+} (\beta \cdot \mathbf{n}) |v|^p.$$

The desired bound then follows from $(\mathcal{H}_p)$ and the definition of $\partial\Omega^+$. $\square$

Let us now prove the well-posedness of (2.16) under assumption $(\mathcal{H}_p)$ for $p \in (1, 2]$.

Theorem 2.7 (Well-posedness). *Let $p \in (1, 2]$ and assume that $(\mathcal{H}_p)$ holds. Then the problem (2.16) is well-posed.*

Proof. We apply the BNB Theorem 2.1. Adapting the proof of (Ern & Guermond, 2004, Theorem 5.7), the condition (BNB1) follows from Proposition 2.6 with the constant $\mathbf{c}_{\text{BNB}} = (\tau^p + (1 + \|\mu\|_{L^\infty(\Omega)} \tau)^p)^{\frac{1}{p}}$. Let us prove the second condition (BNB2). Consider $w \in L^{p'}(\Omega)$ such that $a_{\beta,\mu;p}(v, w) = 0$ for all $v \in V_{\beta;p}^0(\Omega)$. Owing to the inclusion $\mathcal{C}_0^\infty(\Omega) \subset V_{\beta;p}^0(\Omega)$, it follows that $\mu w - \nabla \cdot (\beta w) = 0$ a.e. in Ω , so that the dense inclusion $\mathcal{C}_0^\infty(\Omega) \subset L^{p'}(\Omega)$ implies that $\beta \cdot \nabla w = (\mu - \nabla \cdot \beta) w \in L^{p'}(\Omega)$, whence $w \in V_{\beta;p'}(\Omega)$. Applying now the integration by part formula (2.13a), we observe that

$$\int_{\partial\Omega} (\beta \cdot \mathbf{n}) v w = a_{\beta,\mu;p}(v, w) - \int_{\Omega} (\mu - \nabla \cdot \beta) v w + \int_{\Omega} (\beta \cdot \nabla w) v = 0,$$

for all $v \in V_{\beta;p}^0(\Omega)$. Then, choosing $v = \psi^+ |w|^{p'-2}$ belonging to $V_{\beta;p}^0(\Omega)$ owing to Proposition 2.5 since $p' \geq 2$ by assumption, we infer that

$$0 = \int_{\partial\Omega} (\beta \cdot \mathbf{n}) \psi^+ |w|^{p'} = \int_{\partial\Omega^+} (\beta \cdot \mathbf{n}) |w|^{p'},$$

so that $w|_{\partial\Omega^+} = 0$. Now, we test the identity $\mu w - \nabla \cdot (\beta w) = 0$ by an arbitrary $y \in L^p(\Omega)$ and we use the chain rule $\nabla \cdot (\beta w) = \beta \cdot \nabla w + (\nabla \cdot \beta) w$ to infer that

$$0 = \int_{\Omega} (\mu w - \nabla \cdot (\beta w)) y = \int_{\Omega} (\mu - \nabla \cdot \beta) w y - \int_{\Omega} \beta \cdot \nabla w y.$$

Hence, the particular choice $y = w |w|^{p'-2}$ along with the identity (2.13b) with p replaced by p' and with $z \equiv 1$ yields

$$\begin{aligned} 0 &= \int_{\Omega} (\mu - \nabla \cdot \beta) |w|^{p'} + \frac{1}{p'} \int_{\Omega} (\nabla \cdot \beta) |w|^{p'} - \frac{1}{p'} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) |w|^{p'} = \int_{\Omega} \sigma_{\beta,\mu;p} |w|^{p'} - \frac{1}{p'} \int_{\partial\Omega} \beta \cdot \mathbf{n} |w|^{p'} \\ &\geq \tau^{-1} \|w\|_{L^{p'}(\Omega)}^{p'}, \end{aligned}$$

where we have used that $w|_{\partial\Omega^+} = 0$ and the assumption $(\mathcal{H}_p)$. As a result, $w = 0$ a.e. in Ω , so that the condition (BNB2) is satisfied. Hence, there exists a unique solution solving the problem (2.16). $\square$

Summarizing the results obtained so far, we have proved under assumption $(\mathcal{H}_p)$ the uniqueness of the solution for all $p \in (1, \infty)$ while the existence has been obtained for all $p \in (1, 2]$. These results are now generalized under the following new assumption $(\mathcal{H}'_p)$, for all $p \in (1, \infty)$, so as to include the situation where the infimum of the Friedrichs' tensor $\sigma_{\beta,\mu;p}$ takes null or even slightly negative values:

$(\mathcal{H}'_p)$ $\text{ess inf}_{\Omega} \sigma_{\beta,\mu;p} \leq 0$ and there exists a non dimensional function $\zeta \in \text{Lip}(\Omega)$ such that $\zeta \geq \zeta_0$ in Ω with $\zeta_0 > 0$ and

$$\text{ess inf}_{\Omega} \left(\zeta \sigma_{\beta,\mu;p} - \frac{1}{p} \beta \cdot \nabla \zeta \right) > 0. \quad (2.19)$$

We define the reference time $\tau = \left(\text{ess inf}_{\Omega} \left(\zeta \sigma_{\beta,\mu;p} - \frac{1}{p} \beta \cdot \nabla \zeta \right) \right)^{-1}$.

Proposition 2.8 (Uniqueness under $(\mathcal{H}'_p)$). *Let $p \in (1, \infty)$ and assume that $(\mathcal{H}'_p)$ holds. Then,*

$$a_{\beta, \mu; p}(v, \zeta v |v|^{p-2}) \geq \tau^{-1} \|v\|_{L^p(\Omega)}^p, \quad \forall v \in V_{\beta; p}^0(\Omega). \quad (2.20)$$

Proof. Let $p \in (1, \infty)$ and let $v \in V_{\beta; p}^0(\Omega)$. Owing to the integration by part formula (2.13b) with $z = \zeta$, we have

$$\int_{\Omega} (\beta \cdot \nabla v) \zeta v |v|^{p-2} = \frac{1}{p} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) \zeta |v|^p - \frac{1}{p} \int_{\Omega} (\nabla \cdot \beta) \zeta |v|^p - \frac{1}{p} \int_{\Omega} \beta \cdot \nabla \zeta |v|^p.$$

Hence, we obtain

$$\begin{aligned} a_{\beta, \mu; p}(v, \zeta v |v|^{p-2}) &= \int_{\Omega} \zeta \left(\mu - \frac{1}{p} \nabla \cdot \beta \right) |v|^p - \frac{1}{p} \int_{\Omega} \beta \cdot \nabla \zeta |v|^p + \frac{1}{p} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) \zeta |v|^p \\ &\geq \int_{\Omega} \left(\zeta \sigma_{\beta, \mu; p} - \frac{1}{p} \beta \cdot \nabla \zeta \right) |v|^p, \end{aligned}$$

where we have used that $v|_{\partial\Omega^-} = 0$ and $\zeta > 0$. Hence, the expected result follows from $(\mathcal{H}'_p)$. $\square$

Example 2.9. *Assumption $(\mathcal{H}'_p)$ indeed generalizes the assumption $(\mathcal{H}_p)$ since it is now possible to consider situations not handled by $(\mathcal{H}_p)$. For example, in the domain $\Omega = [0, 1]^3$, define $\mu \equiv 0$ and $\beta = ((1 + \alpha p)(1 + x), -y, 0)$ for any $\alpha \geq 0$, in the Cartesian coordinates. Observing that $\sigma_{\beta, \mu; p} = -\alpha \leq 0$, the assumption $(\mathcal{H}_p)$ is not satisfied whereas $(\mathcal{H}'_p)$ is satisfied for example with the potential $\zeta(\mathbf{x}) = (x - 2)^2$ on Ω . Indeed, a short calculation shows that*

$$\zeta \sigma_{\beta, \mu; p} - \frac{1}{p} \beta \cdot \nabla \zeta = \frac{1}{p} (2 - x) ((3\alpha p + 2)x + 2),$$

which is strictly positive on $[0, 1]$.

Remark 2.10 (Existence of ζ). *Following Devinatz et al. (1974) and considering a continuously differentiable field $\beta \in C^1(\mathbb{R}^d)$, the existence of the potential ζ relies on the assumption that every solution of the Cauchy problem $d_t \mathbf{x}(t) = \beta(\mathbf{x}(t))$, $\mathbf{x}(0) = \mathbf{x}_0 \in \Omega$ remains in the domain Ω for a finite time only. Observing that the proof is based on the flow box theorem, the extension to a less regular field (e.g. $\beta \in \mathbf{Lip}(\Omega)$) is a priori not obvious.*

We are now in a position to state the well-posedness of (2.1) under assumption $(\mathcal{H}'_p)$ for all $p \in (1, 2]$.

Theorem 2.11 (Well-posedness). *Let $p \in (1, 2]$ and assume that $(\mathcal{H}'_p)$ holds. Then the problem (2.16) is well-posed.*

Proof. We follow the same ideas as in the proof of Theorem 2.7. The condition (BNB1) is inferred from Proposition 2.8 with $\mathbf{C}_{\text{BNB}} = (\|\zeta\|_{L^\infty(\Omega)}^p \tau^p + (1 + \|\mu\|_{L^\infty(\Omega)} \tau \|\zeta\|_{L^\infty(\Omega)})^p)^{\frac{1}{p}}$. Turning to the second condition (BNB2), we consider $w \in L^{p'}(\Omega)$ such that $a_{\beta, \mu; p}^0(v, w) = 0$ for all $v \in V_{\beta; p}^0(\Omega)$. Proceeding as in the proof of Theorem 2.7, this implies that w belongs to the graph space $V_{\beta; p'}(\Omega)$ and that it satisfies $\mu w - \nabla \cdot (\beta w) = 0$ a.e. in Ω and $w|_{\partial\Omega^+} = 0$. Let us prove that $w \equiv 0$. First, we observe that replacing p by p' and choosing $z = \zeta^{1-p'}$ in (2.13b) yields

$$\begin{aligned} \int_{\Omega} (\beta \cdot \nabla w) w |w|^{p'-2} \zeta^{1-p'} &= \frac{1}{p'} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) |w|^{p'} \zeta^{1-p'} - \frac{1}{p'} \int_{\Omega} (\nabla \cdot \beta) |w|^{p'} \zeta^{1-p'} \\ &\quad - \frac{1-p'}{p'} \int_{\Omega} (\beta \cdot \nabla \zeta) |w|^{p'} \zeta^{-p'}. \end{aligned}$$

Hence, testing the identity $\mu w - \nabla \cdot (\beta w) = 0$ with the function $w |w|^{p'-2} \zeta^{1-p'}$, we infer that

$$\begin{aligned} 0 &= \int_{\Omega} (\mu - \nabla \cdot \beta) |w|^{p'} \zeta^{1-p'} - \int_{\Omega} (\beta \cdot \nabla w) w |w|^{p'-2} \zeta^{1-p'} \\ &= \int_{\Omega} (\mu - \nabla \cdot \beta) |w|^{p'} \zeta^{1-p'} + \frac{1}{p'} \int_{\Omega} (\nabla \cdot \beta) |w|^{p'} \zeta^{1-p'} - \frac{1}{p} \int_{\Omega} (\beta \cdot \nabla \zeta) |w|^{p'} \zeta^{-p'} \\ &\quad - \frac{1}{p'} \int_{\partial \Omega} (\beta \cdot \mathbf{n}) |w|^{p'} \zeta^{1-p'}. \end{aligned}$$

Then, collecting these terms and using the fact that $w|_{\partial \Omega^+} = 0$, we obtain

$$0 = \int_{\Omega} \left(\zeta \sigma_{\beta, \mu; p} - \frac{1}{p} \beta \cdot \nabla \zeta \right) |w|^{p'} \zeta^{-p'} - \frac{1}{p'} \int_{\partial \Omega} (\beta \cdot \mathbf{n}) |w|^{p'} \zeta^{1-p'} \geq \int_{\Omega} \left(\zeta \sigma_{\beta, \mu; p} - \frac{1}{p} \beta \cdot \nabla \zeta \right) |w|^{p'} \zeta^{-p'}.$$

As a result, owing to $(\mathcal{H}_{p'})$ and the fact that $\zeta \in L^\infty(\Omega)$ and $\zeta > 0$, it follows that $w \equiv 0$ a.e. in Ω . We conclude that (2.16) is well-posed under assumption $(\mathcal{H}_{p'})$ for all $p \in (1, 2]$. $\square$

2.3 Vector advection-reaction problem

In this section, we apply similar ideas to analyze the well-posedness of the vector-valued problem (2.2) in Lebesgue graph spaces where we generalize the assumption of the sign of the Friedrichs tensor $\sigma_{\beta, \mu; p}$ defined by (2.4). For the sake of brevity, the proofs are omitted if they are straightforwardly adapted from those of the scalar case in Section 2.2.

2.3.1 The graph space

Let us introduce the graph space

$$\mathbf{V}_{\beta; p}(\Omega) := \{ \mathbf{v} \in \mathbf{L}^p(\Omega) \mid (\beta \cdot \nabla) \mathbf{v} \in \mathbf{L}^p(\Omega) \}, \quad (2.21)$$

where the i -th component in the Cartesian basis of $(\beta \cdot \nabla) \mathbf{v}$ is given by $\beta_j \partial_j v_i$ (where repeated indices are summed) and where $(\beta \cdot \nabla) \mathbf{v} \in \mathbf{L}^p(\Omega)$ means that the linear form

$$\mathcal{C}_0^\infty(\Omega) \ni \varphi \mapsto - \int_{\Omega} \nabla \cdot (\beta \otimes \varphi) \cdot \mathbf{v}, \quad (2.22)$$

is bounded in $\mathbf{L}^{p'}(\Omega)$, so that $(\beta \cdot \nabla) \mathbf{v}$ is the Riesz representative of (2.22) in $\mathbf{L}^p(\Omega)$. Equipped with the norm $\|\mathbf{v}\|_{\mathbf{V}_{\beta; p}(\Omega)} := (\|\mathbf{v}\|_{\mathbf{L}^p(\Omega)}^p + \|(\beta \cdot \nabla) \mathbf{v}\|_{\mathbf{L}^p(\Omega)}^p)^{\frac{1}{p}}$ for all $\mathbf{v} \in \mathbf{V}_{\beta; p}(\Omega)$, this space defines a reflexive Banach space. Owing to the following proposition, the problem (2.2) is well-defined in the graph space $\mathbf{V}_{\beta; p}(\Omega)$.

Proposition 2.12 (Equivalent definition of $\mathbf{V}_{\beta; p}(\Omega)$). *For all $p \in (1, \infty)$, we have*

$$\mathbf{V}_{\beta; p}(\Omega) = \{ \mathbf{v} \in \mathbf{L}^p(\Omega) \mid \nabla(\beta \cdot \mathbf{v}) + (\nabla \times \mathbf{v}) \times \beta \in \mathbf{L}^p(\Omega) \},$$

where $\nabla(\beta \cdot \mathbf{v}) + (\nabla \times \mathbf{v}) \times \beta \in \mathbf{L}^p(\Omega)$ means that the linear form

$$\mathcal{C}_0^\infty(\Omega) \ni \varphi \mapsto - \int_{\Omega} (\beta \cdot \nabla \varphi + \nabla \times (\varphi \times \beta)) \cdot \mathbf{v}, \quad (2.23)$$

is bounded in $\mathbf{L}^{p'}(\Omega)$.

Proof. Let $\mathbf{v} \in \mathbf{V}_{\beta; p}(\Omega)$. By definition, we have

$$\int_{\Omega} (\beta \cdot \nabla) \mathbf{v} \cdot \varphi = - \int_{\Omega} \nabla \cdot (\beta \otimes \varphi) \cdot \mathbf{v}, \quad \forall \varphi \in \mathcal{C}_0^\infty(\Omega).$$

Now, recalling the identity $\nabla \cdot (\beta \otimes \varphi) = \beta \nabla \cdot \varphi + \nabla \times (\varphi \times \beta) + (\varphi \cdot \nabla) \beta$ for all $\beta \in \mathbf{Lip}(\Omega)$ and for all $\varphi \in \mathcal{C}_0^\infty(\Omega)$, it follows that

$$\int_{\Omega} (\beta \cdot \nabla) \mathbf{v} \cdot \varphi = - \int_{\Omega} (\beta \nabla \cdot \varphi + \nabla \times (\varphi \times \beta)) \cdot \mathbf{v} - \int_{\Omega} ((\varphi \cdot \nabla) \beta) \cdot \mathbf{v}, \quad \forall \varphi \in \mathcal{C}_0^\infty(\Omega).$$

Hence, observing that $((\varphi \cdot \nabla) \beta) \cdot \mathbf{v} = ((\nabla \beta) \varphi) \cdot \mathbf{v}$, we obtain

$$\int_{\Omega} ((\beta \cdot \nabla) \mathbf{v} + (\nabla \beta)^\top \mathbf{v}) \cdot \varphi = - \int_{\Omega} (\beta \nabla \cdot \varphi + \nabla \times (\varphi \times \beta)) \cdot \mathbf{v}, \quad \forall \varphi \in \mathcal{C}_0^\infty(\Omega).$$

Hence, the linear form (2.23) is bounded yielding $\nabla(\beta \cdot \mathbf{v}) + (\nabla \times \mathbf{v}) \times \beta \in \mathbf{L}^p(\Omega)$, so that the inclusion holds. Note that we have the identity $(\beta \cdot \nabla) \mathbf{v} + (\nabla \beta)^\top \mathbf{v} = \nabla(\beta \cdot \mathbf{v}) + (\nabla \times \mathbf{v}) \times \beta$. Since the proof of the converse inclusion is similar, the proof is completed. $\square$

Recalling now that the boundary $\partial\Omega$ is well-separated in the sense of (2.12), functions in the graph space $\mathbf{V}_{\beta;p}(\Omega)$ have a trace in the space

$$\mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega) := \{\mathbf{v} : \partial\Omega \rightarrow \mathbb{R}^3 \text{ is Lebesgue measurable} \mid \int_{\partial\Omega} |\beta \cdot \mathbf{n}| |\mathbf{v}|^p < \infty\}. \quad (2.24)$$

Equipped with the norm $\|\mathbf{v}\|_{\mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)}^p := \int_{\partial\Omega} |\beta \cdot \mathbf{n}| |\mathbf{v}|^p$ for all $\mathbf{v} \in \mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$, this space defines a Banach space.

Lemma 2.13 (Trace in $\mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$). *Let $p \in (1, \infty)$. The trace map $\gamma : \mathcal{C}^\infty(\overline{\Omega}) \rightarrow \mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$ with $\gamma(\varphi) = \varphi|_{\partial\Omega}$ for all $\varphi \in \mathcal{C}^\infty(\overline{\Omega})$ extends continuously to $\mathbf{V}_{\beta;p}(\Omega)$, i.e., there exists $\mathcal{C}_\gamma > 0$ such that*

$$\|\gamma(\mathbf{v})\|_{\mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)} \leq \mathcal{C}_\gamma \|\mathbf{v}\|_{\mathbf{V}_{\beta;p}(\Omega)}, \quad \forall \mathbf{v} \in \mathbf{V}_{\beta;p}(\Omega).$$

Proposition 2.14 (Integration by parts). *For all $\mathbf{v} \in \mathbf{V}_{\beta;p}(\Omega)$ and for all $\mathbf{w} \in \mathbf{V}_{\beta;p'}(\Omega)$,*

$$\int_{\Omega} \mathbf{w} \cdot (\beta \cdot \nabla) \mathbf{v} + \int_{\Omega} \mathbf{v} \cdot (\beta \cdot \nabla) \mathbf{w} + \int_{\Omega} (\nabla \cdot \beta) \mathbf{v} \cdot \mathbf{w} = \int_{\partial\Omega} (\beta \cdot \mathbf{n}) \mathbf{v} \cdot \mathbf{w}. \quad (2.25a)$$

In addition, for all $\mathbf{v} \in \mathbf{V}_{\beta;p}(\Omega)$ and for all $z \in W^{1,\infty}(\Omega)$,

$$\int_{\Omega} |\mathbf{v}|^{p-2} z \mathbf{v} \cdot (\beta \cdot \nabla) \mathbf{v} + \frac{1}{p} \int_{\Omega} (\nabla \cdot \beta) |\mathbf{v}|^p z + \frac{1}{p} \int_{\Omega} \beta \cdot \nabla z |\mathbf{v}|^p = \frac{1}{p} \int_{\partial\Omega} (\beta \cdot \mathbf{n}) |\mathbf{v}|^p. \quad (2.25b)$$

2.3.2 Well-posedness

Similarly to Section 2.2, we introduce the bilinear form $\mathbf{a}_{\beta,\mu;p} \in \mathcal{L}(\mathbf{V}_{\beta;p}^0(\Omega) \times \mathbf{L}^{p'}(\Omega); \mathbb{R})$ with $\mathbf{V}_{\beta;p}^0(\Omega) = \{\mathbf{w} \in \mathbf{V}_{\beta;p}(\Omega) \mid \mathbf{w}|_{\partial\Omega^-} = \mathbf{0}\}$ such that for all $\mathbf{v} \in \mathbf{V}_{\beta;p}(\Omega)$ and for all $\mathbf{w} \in \mathbf{L}^{p'}(\Omega)$,

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{w}) := \int_{\Omega} (\nabla(\beta \cdot \mathbf{v}) + (\nabla \times \mathbf{v}) \times \beta) \cdot \mathbf{w} + \int_{\Omega} \mu \mathbf{v} \cdot \mathbf{w}, \quad (2.26)$$

As in the proof of Proposition 2.12, we observe that the bilinear form (2.26) can be reformulated as

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{w}) = \int_{\Omega} \mathbf{w} \cdot (\beta \cdot \nabla) \mathbf{v} + \int_{\Omega} \mathbf{w} \cdot (\nabla \beta^\top + \mu) \mathbf{v}, \quad (2.27)$$

and the writing $\mu' = \nabla \beta^\top + \mu$ yields $\mu' \in \mathbf{L}^\infty(\Omega)$ and

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{w}) := \int_{\Omega} (\beta \cdot \nabla) \mathbf{v} \cdot \mathbf{w} + \int_{\Omega} \mu' \mathbf{v} \cdot \mathbf{w}. \quad (2.28)$$

Assuming that $\mathbf{s} \in \mathbf{L}^p(\Omega)$, the weak formulation of (2.2) in the graph space $\mathbf{V}_{\beta;p}^0(\Omega)$ is:

$$\text{Find } \mathbf{u} \in \mathbf{V}_{\beta;p}(\Omega) \quad \text{s.t.} \quad \mathbf{a}_{\beta,\mu;p}(\mathbf{u}, \mathbf{v}) = \int_{\Omega} \mathbf{s} \cdot \mathbf{v}, \quad \forall \mathbf{v} \in \mathbf{L}^{p'}(\Omega). \quad (2.29)$$

We readily see that if $u \in \mathbf{V}_{\beta;p}^0(\Omega)$ solves (2.29), the PDE (2.2a) holds in $\mathbf{L}^p(\Omega)$ and the boundary condition (2.2b) holds in $\mathbf{L}^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$. The uniqueness of the solution of problem (2.29) relies on the sign of the lowest eigenvalue of the $\mathbb{R}^{3 \times 3}$ -valued Friedrichs tensor

$$\sigma_{\beta,\mu;p} := \frac{\mu + \mu^\top}{2} + \frac{\nabla\beta + \nabla\beta^\top}{2} - \frac{1}{p}(\nabla \cdot \beta) \text{Id}. \quad (2.30)$$

This lowest eigenvalue is denoted by

$$\aleph_p(\mathbf{x}) = \min_{\mathbf{y} \in \mathbb{R}^3 \setminus \{0\}} \frac{(\sigma_{\beta,\mu;p}(\mathbf{x})\mathbf{y}, \mathbf{y})}{|\mathbf{y}|^2}, \quad \forall \mathbf{x} \in \Omega,$$

with $(\cdot, \cdot)$ the classical Euclidean inner product in $\mathbb{R}^3$. The uniqueness of the solution of (2.29) is then a consequence of the following Friedrichs positivity assumption.

$(\mathcal{H}_p)$ $\text{ess inf}_\Omega \aleph_p > 0$. We define $\tau = (\text{ess inf}_\Omega \aleph_p)^{-1}$.

Proposition 2.15 (Uniqueness under $(\mathcal{H}_p)$). *Let $p \in (1, \infty)$ and assume that $(\mathcal{H}_p)$ holds. Then,*

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{v} |\mathbf{v}|^{p-2}) \geq \tau^{-1} \|\mathbf{v}\|_{\mathbf{L}^p(\Omega)}^p, \quad \forall \mathbf{v} \in \mathbf{V}_{\beta;p}^0(\Omega).$$

Proof. Let $\mathbf{v} \in \mathbf{V}_{\beta;p}^0(\Omega)$. First, observe that $\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{v} |\mathbf{v}|^{p-2})$ is well-defined since $\mathbf{v} |\mathbf{v}|^{p-2} \in \mathbf{L}^{p'}(\Omega)$. Owing to the identity (2.27), we infer that

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{v} |\mathbf{v}|^{p-2}) = \int_\Omega |\mathbf{v}|^{p-2} \mathbf{v} \cdot (\beta \cdot \nabla) \mathbf{v} + \int_\Omega |\mathbf{v}|^{p-2} \mathbf{v} \cdot (\nabla \beta^\top + \mu) \mathbf{v}.$$

Using now the integration by parts formula (2.25b) with $z \equiv 1$, we obtain

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \mathbf{v} |\mathbf{v}|^{p-2}) = \int_\Omega |\mathbf{v}|^{p-2} \mathbf{v} \cdot \sigma_{\beta,\mu;p} \mathbf{v} + \frac{1}{p} \int_{\partial\Omega} |\beta \cdot \mathbf{n}| |\mathbf{v}|^p,$$

whence the result follows using $(\mathcal{H}_p)$ and recalling that $\mathbf{v}|_{\partial\Omega^-} = 0$. $\square$

To take into account situations where the smallest eigenvalue $\aleph_p$ takes null or slightly negative values in Ω , we now consider the following assumption:

$(\mathcal{H}'_p)$ $\text{ess inf}_\Omega \aleph_p \leq 0$ and there exists a non dimensional function $\zeta \in \text{Lip}(\Omega)$ such that $\zeta \geq \zeta_0$ with $\zeta_0 > 0$ and

$$\text{ess inf}_\Omega \left(\zeta \aleph_p - \frac{1}{p} \beta \cdot \nabla \zeta \right) > 0.$$

We define $\tau^{-1} = \text{ess inf}_\Omega \left(\zeta \aleph_p - \frac{1}{p} \beta \cdot \nabla \zeta \right)$.

Observe the similar structure of the generalized positivity assumption $(\mathcal{H}'_p)$ and $(\mathcal{H}_p)$ for the two problems (2.1) and (2.2).

Proposition 2.16 (Uniqueness under Hypothesis $(\mathcal{H}'_p)$). *Let $p \in (1, \infty)$ and assume that $(\mathcal{H}'_p)$ holds. Then,*

$$\mathbf{a}_{\beta,\mu;p}(\mathbf{v}, \zeta \mathbf{v} |\mathbf{v}|^{p-2}) \geq \tau^{-1} \|\mathbf{v}\|_{\mathbf{L}^p(\Omega)}^p, \quad \forall \mathbf{v} \in \mathbf{V}_{\beta;p}^0(\Omega).$$

Finally, the well-posedness of (2.2) holds under assumption $(\mathcal{H}_p)$ or $(\mathcal{H}'_p)$ for all $p \in (1, 2]$. The proof follows the same ideas used to prove Theorems 2.7 and 2.11, this time using Propositions 2.15 and 2.16, respectively.

Theorem 2.17 (Well-posedness). *Let $p \in (1, 2]$ and assume that $(\mathcal{H}_p)$ or $(\mathcal{H}'_p)$ holds. Then the problem (2.29) is well-posed.*

Chapter 3

Vertex-based scheme for advection-reaction

Contents

3.1	Discrete setting	26
3.1.1	Primal and dual meshes	26
3.1.2	Degrees of freedom	28
3.1.3	Compatible Discrete Operator tools	28
3.2	Advection-reaction problem	30
3.2.1	Discrete Friedrichs problem	31
3.2.2	Analysis	35
3.3	Numerical results	41
3.3.1	Computational setting	42
3.3.2	Test case 1. Rotating advective field	43
3.3.3	Test case 2. Sharp internal layer	44

This chapter devises and analyzes vertex-based schemes to approximate on polyhedral meshes the solution of the scalar-valued advection-reaction problem

$$\beta \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (3.1)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega^-. \quad (3.2)$$

A salient feature of this chapter is that our schemes are formulated using the algebraic Compatible Discrete Operator (CDO) framework proposed by Bonelle (2014). This framework takes inspiration from the seminal works of Tarhassaari *et al.* (1999); Bossavit (1998, 2000) and from the Discrete Hodge operator setting proposed by Hiptmair (2001, 2002).

This chapter contains two main contributions. The first one is to devise a CDO scheme for advection-reaction. The key idea is to build a *discrete contraction operator* that is the discrete counterpart of the map $\mathbf{v} \mapsto \beta \cdot \mathbf{v}$. This operator maps degrees of freedom (dofs) attached to mesh edges to dofs attached to mesh vertices, and is also called interior product (see Abraham *et al.* (1988) or Gerritsma (2012)). The CDO framework allows one to discretize the advective derivative $\beta \cdot \nabla \mathbf{v}$ by two distinct operators: a well-known topological discrete gradient operator mapping dofs attached to mesh vertices to dofs attached to mesh edges and the above discrete contraction operator. The second contribution is to extend at the discrete level the classical positivity assumption (denoted by $\mathcal{H}_p$ in the Chapter 2) so as to consider a scalar-valued Friedrichs tensor taking possibly null values. This extension is rarely addressed in the literature (see Deuring *et al.* (2015) or Ayuso & Marini (2009) for some examples). In particular, the stability of the scheme now hinges on an inf-sup condition under a mesh-size restriction, which becomes void if the advective field is divergence-free.

Putting our schemes into perspective with existing schemes, we observe that our scheme is essentially an upwind finite volume (or lowest-order dG) scheme on the dual mesh with vertex-centered dual cells as control volumes. Taking inspiration from the analysis of Friedrichs systems by Ern & Guermond (2006a,b) for dG methods, our analysis uses similar techniques (see also the seminal work of Johnson & Pitkäranta (1986)). In particular, the present algebraic point of view sheds new light on the theory of Friedrichs' systems at the discrete level. We also mention other approaches to discretize similar problems in the context of the differential geometry. Using the notion of extrusion defined by Bossavit (2003), Heumann & Hiptmair (2008) proposed a discretization of the aforementioned contraction operator on triangular and Cartesian square meshes, respectively. Palha (2013) also investigates this problem using the notion of wedge product, seen as the adjoint of the contraction operator.

This chapter is organized as follows. First, we introduce the discrete setting, namely the polyhedral mesh, the dofs and some of the CDO tools from Bonelle (2014), i.e., duality products, discrete differential operators and reduction maps. Next, we devise and analyze our schemes. In Section 3.3, numerical results are presented on three-dimensional polyhedral meshes.

The content of this chapter is an extended version of some parts of the paper "*Vertex-based Compatible Discrete Operator Schemes on Polyhedral meshes for Advection-Diffusion Equations*" by P. Cantin & A. Ern, published in Computational Methods in Applied Mathematics, 2016.

3.1 Discrete setting

This section introduces the main ingredients underlying the discrete setting: mesh entities, degrees of freedom, reduction maps and discrete differential operators. For brevity, we only present the concepts needed in this Chapter. A broader presentation can be found in Chapter 7.

3.1.1 Primal and dual meshes

The computational domain is denoted by Ω and is assumed to be an open, bounded, connected, polyhedral subset of $\mathbb{R}^3$. Its boundary is denoted by $\partial\Omega$ with $\mathbf{n}$ its outward normal component.

Primal mesh. The primal mesh of the domain Ω is denoted by $M := \{V, E, F, C\}$, where V collects the mesh vertices generically denoted by v (0-cell), E collects the mesh edges denoted by e (1-cell), F collects the mesh faces denoted by f (2-cell), and C collects the mesh cells denoted by c (3-cell). For all $k \in \llbracket 0, 3 \rrbracket$, these k -cell are homeomorphic by a bi-Lipschitz map to the open unit k -ball of $\mathbb{R}^k$. Considering an arbitrary geometric set $X \in \{E, F, C\}$ with $x \in X$, we denote $\bar{x}$ the closure of x and ∂x the boundary of x . We orient mesh edges by assigning an arbitrary tangential unit vector $\mathbf{t}_e$ for all $e \in E$ and we orient all mesh faces by assigning an arbitrary normal unit vector $\mathbf{n}_f$ for all $f \in F$. An example of a simple polyhedral mesh is depicted in Figure 3.1.

Mesh regularity. Hereafter, we consider families of polyhedral meshes M satisfying the following requirements:

- (C) The mesh M defines a *cellular complex* of Ω : $\bar{\Omega} = \cup\{\bar{c} \mid c \in C\}$, edges are straight and faces are planar, distinct elements have empty intersection and the boundary of all the cells and all the faces is composed of a finite number of faces and edges, respectively.
- (St) The mesh M is *star-shaped*: all the mesh cells and all the mesh faces are star-shaped with respect to their barycenter.
- (Sh) The mesh M is *shape-regular*: there exists a simplicial sub-mesh of M that is shape-regular in the sense of Ciarlet (1978).

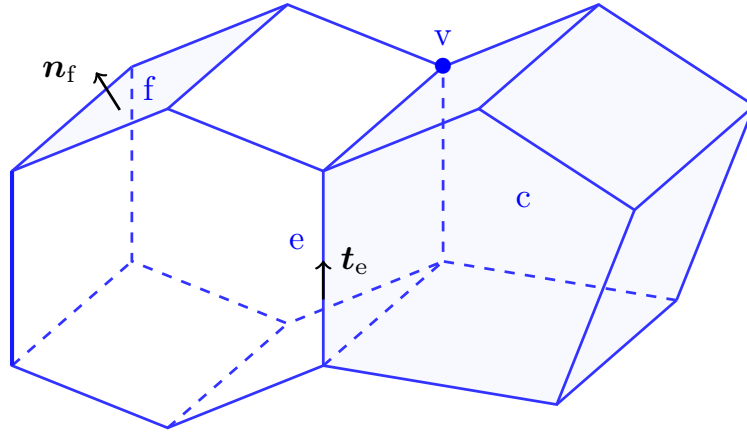


Figure 3.1 – A primal polyhedral mesh containing the cell $c \in \mathcal{C}$, the face $f \in \mathcal{F}$, the edge $e \in \mathcal{E}$ and the vertex $v \in \mathcal{V}$

Definition 3.1 (Geometric subsets). *Let $X, Y \in \mathcal{M}$ and let $y \in Y$. Then, we define*

$$X_y := \{x \in X \mid y \subset \partial x\} \quad \text{or} \quad X_y := \{x \in X \mid x \subset \partial y\}, \quad (3.3)$$

*depending if the dimension of y is smaller or not than that of element of X . For instance, $E_c := \{e \in \mathcal{E} \mid e \subset \partial c\}$ and $C_e := \{c \in \mathcal{C} \mid e \subset \partial c\}$ and so on. Owing to assumption **(C)**, the cardinal $\#X_y$ is uniformly bounded.*

Dual mesh. In addition to the primal mesh $\mathcal{M}$, Compatible Discrete Operator (CDO) schemes are formulated using also a dual mesh. The dual mesh is denoted by $\tilde{\mathcal{M}} := \{\tilde{\mathcal{V}}, \tilde{\mathcal{E}}, \tilde{\mathcal{F}}, \tilde{\mathcal{C}}\}$ and is composed of dual entities such that there is a one-to-one pairing between a primal k -cell and a dual $(3 - k)$ -cell. In essence, a primal vertex is associated with a dual cell, a primal edge is associated with a dual face, and so on. Figure 3.2 represents dual elements of the primal mesh depicted in Figure 3.1. To enhance the link between primal and dual entities, we write

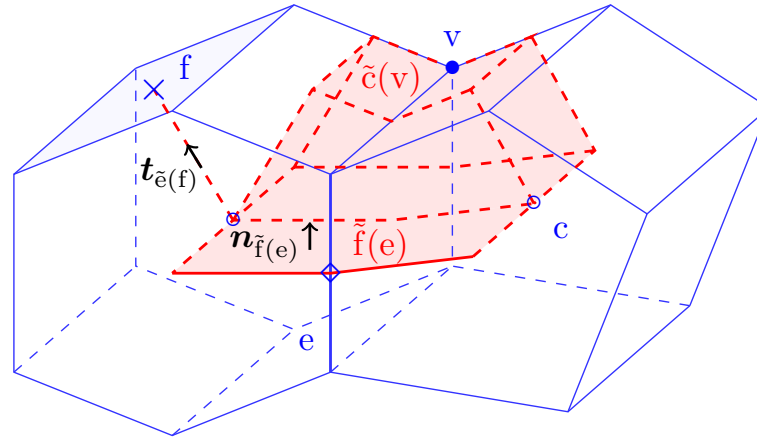


Figure 3.2 – Dual elements of $\mathcal{M}$: the dual face $\tilde{f}(e) \in \tilde{\mathcal{F}}$ and the dual cell $\tilde{c}(v) \in \tilde{\mathcal{C}}$.

$\tilde{f}(e)$ the dual face associated with the primal edge $e \in \mathcal{E}$ and $\tilde{c}(v)$ the dual cell associated with the primal vertex $v \in \mathcal{V}$. We stress that the dual mesh is not seen by the final-user and that its definition is not unique. In particular, we only consider in this thesis *fully barycentric* dual meshes, built from the barycenters of the primal mesh entities owing to assumption **(St)**. It is also useful to notice that there exists a common simplicial sub-mesh of the primal mesh $\mathcal{M}$ and the dual mesh $\tilde{\mathcal{M}}$.

In general, dual edges are not straight and dual faces are not planar, as can be seen from Figure 3.2. For this reason, the normal component $\mathbf{n}_{\tilde{f}(e)}$ attached to a dual face $\tilde{f}(e)$ is only piece-wise constant on each simplex composing this dual face. The normal vector $\mathbf{n}_{\tilde{f}(e)}$ is oriented in the same direction as $\mathbf{t}_e$, i.e., $\mathbf{n}_{\tilde{f}(e)} \cdot \mathbf{t}_e > 0$ holds almost everywhere.

Boundary meshes. Since boundary conditions are weakly enforced in our schemes, mesh entities at the boundary play an important role. The trace of the primal mesh M at the boundary $\partial\Omega$ is denoted by $M^\partial := \{V^\partial, E^\partial, F^\partial\}$, where V^∂ collects all the primal vertices lying at the boundary, and so on. Let $\tilde{c}(v)$ be the dual cell attached to a boundary vertices $v \in V^\partial$. An important observation is that its boundary $\partial\tilde{c}(v)$ is not completely covered by dual faces. For this reason, we additionally introduce the set $\tilde{F}^\partial$ collecting the boundary dual faces defined by $\tilde{f}^\partial(v) = \partial\tilde{c}(v) \cap \partial\Omega$ for all $v \in V^\partial$; see Figure 3.3.

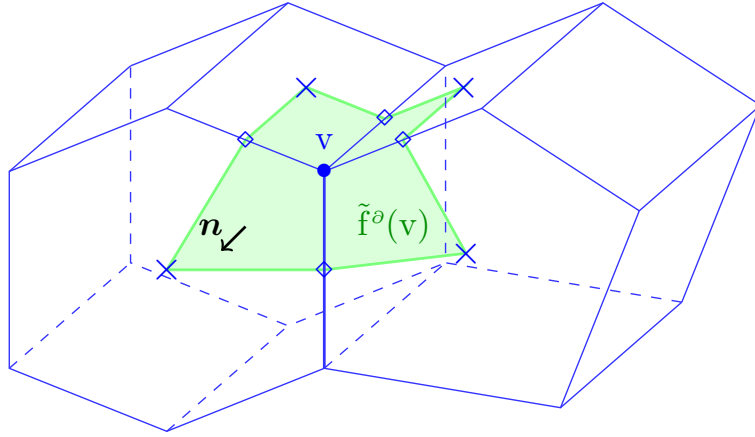


Figure 3.3 – The dual boundary face $\tilde{f}^\partial(v)$ associated with the boundary vertex $v \in V^\partial$.

3.1.2 Degrees of freedom

CDO schemes consider degrees of freedom (dofs) attached to mesh entities according to the physical nature of the discrete field. For instance, dofs of a discrete potential field (also called 0-cochain in algebraic topology) are attached to mesh vertices, either primal or dual ones (also called 0-chains in algebraic topology). In this chapter, we focus on vertex-based CDO schemes where dofs are attached to primal vertices. For a discrete potential u , we use the notation $u \in \mathcal{V} \equiv \mathbb{R}^{\#V}$, where $\mathcal{V}$ denotes the finite-dimensional space collecting dofs attached to the mesh vertices and where $\#V$ denotes the cardinal of the set V . u_v corresponds to the entry of u attached to $v \in V$. We also consider a discrete circulation field (also called 1-cochain), attached to primal mesh edges (also called 1-chains) and collected in the set $\mathcal{E}$. For all $u \in \mathcal{E}$, u_e denotes the entry of u attached to $e \in E$. These dofs spaces are respectively equipped with the following Euclidean inner products:

$$\langle\langle v, w \rangle\rangle_{\mathcal{V}} = \sum_{v \in V} v_v w_v, \quad \forall v, w \in \mathcal{V} \quad \text{and} \quad \langle\langle v, w \rangle\rangle_{\mathcal{E}} = \sum_{e \in E} v_e w_e, \quad \forall v, w \in \mathcal{E}. \quad (3.4)$$

3.1.3 Compatible Discrete Operator tools

Discrete differential operators. Recalling that mesh edges $e \in E$ are oriented by an unit tangential vector $\mathbf{t}_e$, $\iota_{v,e}$ denotes the incidence number equal to 1 if $\mathbf{t}_e$ points toward v and -1 otherwise. Similarly, $\iota_{\tilde{f}(e), \tilde{c}(v)}$ is the dual incidence number of $\iota_{v,e}$, equal to 1 if $\mathbf{n}_{\tilde{f}(e)}$ points outward $\tilde{c}(v)$ and -1 otherwise. One can verify that $\iota_{\tilde{f}(e), \tilde{c}(v)} = -\iota_{v,e}$. Key ingredients to devise CDO schemes are discrete differential operators. Hereafter, we only introduce the

discrete primal gradient and the discrete dual divergence, which are defined, respectively, as

$$\text{GRAD} : \mathcal{V} \rightarrow \mathcal{E} \text{ such that } \forall e \in E, \quad \text{GRAD}(\mathbf{w})|_e := \sum_{v \in V_e} \iota_{v,e} \mathbf{w}_v, \quad \forall \mathbf{w} \in \mathcal{V}, \quad (3.5a)$$

$$\widetilde{\text{DIV}} : \mathcal{E} \rightarrow \mathcal{V} \text{ such that } \forall v \in V, \quad \widetilde{\text{DIV}}(\mathbf{w})|_v := \sum_{e \in E_v} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{w}_e, \quad \forall \mathbf{w} \in \mathcal{E}, \quad (3.5b)$$

with $V_e := \{v \in V \mid v \subset \partial e\}$ for all $e \in E$ and $E_v := \{e \in E \mid v \subset \partial e\}$ for all $v \in V$, owing to Definition 3.1.

Remark 3.2 (Topological operators). *Operators defined by (3.5a) and (3.5b) are topological or metric-free since only the connectivity vertices-edges plays a role in their definition.*

The discrete primal gradient and the discrete dual divergence are anti-adjoints with respect to the Euclidean inner products (3.4).

Proposition 3.3 (Adjunction). *For all $v \in \mathcal{V}$ and for all $\mathbf{w} \in \mathcal{E}$, we have*

$$\langle \mathbf{w}, \text{GRAD}(v) \rangle_{\mathcal{E}} = -\langle v, \widetilde{\text{DIV}}(\mathbf{w}) \rangle_{\mathcal{V}}. \quad (3.6)$$

Proof. This identity is a direct consequence of definitions (3.4) together with the definitions (3.5a) and (3.5b) of GRAD and $\widetilde{\text{DIV}}$, respectively, and the relation $\iota_{v,e} = -\iota_{\tilde{f}(e), \tilde{c}(v)}$. $\square$

Reduction maps. To measure the approximation error resulting from CDO schemes and to discretize continuous fields (such as the source terms or the boundary datum), we introduce the notion of reduction map. The reduction map $\hat{R}_{\mathcal{V}} : L^1(\Omega) \rightarrow \mathcal{V}$ acts as follows:

$$\forall v \in V, \quad \hat{R}_{\mathcal{V}}(w)|_v = \frac{1}{|\tilde{c}(v)|} \int_{\tilde{c}(v)} w, \quad \forall w \in L^1(\Omega), \quad (3.7)$$

where $|\tilde{c}(v)|$ denotes the 3-dimensional Lebesgue measure of the dual cell $\tilde{c}(v)$.

Remark 3.4 (Comparison with the de Rham map). *Another reduction map mapping into $\mathcal{V}$ is the classical de Rham map $R_{\mathcal{V}} : H^{\frac{3}{2}+\epsilon}(\Omega) \rightarrow \mathcal{V}$ with $\epsilon > 0$, such that for all $v \in V$ and for all $w \in H^{\frac{3}{2}+\epsilon}(\Omega)$, $R_{\mathcal{V}}(w)|_v = w(\mathbf{x}_v)$, where $\mathbf{x}_v$ denotes the coordinates of v . Compared with $\hat{R}_{\mathcal{V}}$, the operator $R_{\mathcal{V}}$ requires more regularity. This is the reason why we do not use this operator in the context of low-order approximation of first-order problems since typically, only the $H^1(\Omega)$ regularity of the exact solution is needed to achieve a quasi-optimal convergence rate.*

Dual reduction maps attached to dual cells and dual faces are also used in the analysis. These maps are denoted by $\tilde{R}_{\mathcal{V}} : L^1(\Omega) \rightarrow \mathcal{V}$ and by $\tilde{R}_{\mathcal{E}} : \mathbf{W}^{s,p}(\Omega) \rightarrow \mathcal{E}$ with $sp > 1$, respectively, and act as follows:

$$\forall v \in V, \quad \tilde{R}_{\mathcal{V}}(w)|_v = \int_{\tilde{c}(v)} w, \quad \forall w \in L^1(\Omega), \quad (3.8a)$$

$$\forall e \in E, \quad \tilde{R}_{\mathcal{E}}(\mathbf{w})|_e = \int_{\tilde{f}(e)} \mathbf{w} \cdot \mathbf{n}_{\tilde{f}(e)}, \quad \forall \mathbf{w} \in \mathbf{W}^{s,p}(\Omega). \quad (3.8b)$$

Finally, since the boundary conditions are weakly enforced, we also consider a dual reduction map attached to boundary dual faces, denoted by $\tilde{R}_{\mathcal{V}}^{\partial} : L^1(\partial\Omega) \rightarrow \mathcal{V}$ and acting as follows:

$$\forall v \in V, \quad \tilde{R}_{\mathcal{V}}^{\partial}(w)|_v = \int_{\tilde{f}^{\partial}(v)} w, \quad \forall w \in L^1(\partial\Omega). \quad (3.8c)$$

By convention, since $\tilde{f}^{\partial}(v) = \emptyset$ for all $v \in V^{\circ}$ with $V^{\circ} = V \setminus V^{\partial}$, we have $\tilde{R}_{\mathcal{V}}^{\partial}(\cdot)|_v = 0$ for all $v \in V^{\circ}$. The following Proposition 3.5 states that the differential operator $\widetilde{\text{DIV}}$ is indeed the discrete counterpart of the classical divergence operator $\nabla \cdot$.

Proposition 3.5 (Dual commutation). *For all $\mathbf{v} \in \mathbf{W}^{s,p}(\Omega)$ with $sp > 1$, we have*

$$\widetilde{\mathbf{R}}_{\mathcal{V}}(\nabla \cdot \mathbf{v}) = \widetilde{\mathbf{DIV}} \widetilde{\mathbf{R}}_{\mathcal{E}}(\mathbf{v}) + \widetilde{\mathbf{R}}_{\mathcal{V}}^{\partial}(\mathbf{v} \cdot \mathbf{n}). \quad (3.9)$$

Proof. This identity follows from the classical divergence formula. $\square$

Remark 3.6 (Boundary contribution in dual operators). *We have conserved the definition of $\widetilde{\mathbf{DIV}}$ proposed by Bonelle (2014), since it is naturally associated with the dual mesh, where dual cells attached to boundary vertices are not covered at the boundary by dual faces in $\widetilde{\mathbf{F}}$. An alternative choice is to modify the definition of this operator so as to include dual boundary faces attached to primal boundary vertices; by doing so, the boundary term appears in (3.6) and no longer in (3.9).*

Restriction to mesh cells. It is convenient to localize discrete objects to the primal mesh cells. For all $c \in \mathcal{C}$, $\mathcal{V}_c$ and $\mathcal{E}_c$ denote the dofs subspaces of $\mathcal{V}$ and $\mathcal{E}$, respectively, collecting dofs attached to V_c and E_c . Using straightforward notation, the local Euclidean inner products on the dofs spaces $\mathcal{V}_c$ and $\mathcal{E}_c$ are denoted by $\langle\langle \cdot, \cdot \rangle\rangle_{\mathcal{V}_c}$ and $\langle\langle \cdot, \cdot \rangle\rangle_{\mathcal{E}_c}$. Slightly abusing the notation, $\mathbf{GRAD}$ and $\widetilde{\mathbf{DIV}}$ also denote the discrete differential operators (3.5a) and (3.5b) restricted to the local dofs spaces $\mathcal{V}_c$ and $\mathcal{E}_c$, respectively. We have the following local version of Proposition 3.3.

Proposition 3.7 (Local adjunction). *For all $\mathbf{v} \in \mathcal{V}_c$ and for all $\mathbf{w} \in \mathcal{E}_c$, we have*

$$\langle\langle \mathbf{w}, \mathbf{GRAD}(\mathbf{v}) \rangle\rangle_{\mathcal{E}_c} = -\langle\langle \mathbf{v}, \widetilde{\mathbf{DIV}}(\mathbf{w}) \rangle\rangle_{\mathcal{V}_c}.$$

We also localize the primal and the dual reduction maps (3.7) and (3.8). The local primal reduction map attached to V_c is denoted by $\widehat{\mathbf{R}}_{\mathcal{V}_c} : L^1(\widehat{c}) \rightarrow \mathcal{V}_c$ with the patch cell $\widehat{c} := \cup\{\widehat{c}(\mathbf{v}) \mid \mathbf{v} \in V_c\}$ and acts similarly to the global map $\mathbf{R}_{\mathcal{V}}$ as follows:

$$\forall \mathbf{v} \in V_c, \quad \widehat{\mathbf{R}}_{\mathcal{V}_c}(\mathbf{w})|_{\mathbf{v}} = \frac{1}{|\widehat{c}(\mathbf{v})|} \int_{\widehat{c}(\mathbf{v})} \mathbf{w}, \quad \forall \mathbf{w} \in L^1(\widehat{c}). \quad (3.10)$$

Remark 3.8 (Comparison with the de Rham map). *From Remark 3.4, the local de Rham map defining elements of $\mathcal{V}_c$ is defined as $\mathbf{R}_{\mathcal{V}_c} : H^{\frac{3}{2}+\epsilon}(c) \rightarrow \mathcal{V}_c$ and acts similarly to the global map $\mathbf{R}_{\mathcal{V}}$. Observe that the domain of $\mathbf{R}_{\mathcal{V}_c}$ considers functions only defined on the cell c whereas the domain of $\widehat{\mathbf{R}}_{\mathcal{V}_c}$ considers functions defined on the slightly larger patch cell $\widehat{c}$.*

The local dual reduction maps on dual cells and on dual faces are denoted by $\widetilde{\mathbf{R}}_{\mathcal{V}_c} : L^1(c) \rightarrow \mathcal{V}_c$ and by $\widetilde{\mathbf{R}}_{\mathcal{E}_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathcal{E}_c$ with $sp > 1$, respectively. They are defined by

$$\forall \mathbf{v} \in V_c, \quad \widetilde{\mathbf{R}}_{\mathcal{V}_c}(\mathbf{w})|_{\mathbf{v}} = \int_{\widehat{c}(\mathbf{v}) \cap c} \mathbf{w}, \quad \forall \mathbf{w} \in L^1(c), \quad (3.11a)$$

$$\forall \mathbf{e} \in E_c, \quad \widetilde{\mathbf{R}}_{\mathcal{E}_c}(\mathbf{w})|_{\mathbf{e}} = \int_{\widetilde{\mathbf{f}}_c(\mathbf{e})} \mathbf{w} \cdot \mathbf{n}_{\widetilde{\mathbf{f}}_c(\mathbf{e})}, \quad \forall \mathbf{w} \in \mathbf{W}^{s,p}(c). \quad (3.11b)$$

Here, $\widetilde{\mathbf{f}}_c(\mathbf{e})$ denotes the restriction of the dual face $\widetilde{\mathbf{f}}(\mathbf{e})$ to c , defined as $\widetilde{\mathbf{f}}_c(\mathbf{e}) = \widetilde{\mathbf{f}}(\mathbf{e}) \cap c$ and with $\mathbf{n}_{\widetilde{\mathbf{f}}_c(\mathbf{e})}$ its normal vector oriented by $\mathbf{t}_e$.

For any boundary face $\mathbf{f} \in \mathbf{F}^{\partial}$, $c(\mathbf{f})$ denotes the unique cell containing $\mathbf{f}$. The local reduction map attached to dual boundary faces is denoted by $\widetilde{\mathbf{R}}_{\mathcal{V}_{\mathbf{f}}}^{\partial} : L^1(\mathbf{f}) \rightarrow \mathcal{V}_{c(\mathbf{f})}$ and is such that

$$\forall \mathbf{v} \in V_{c(\mathbf{f})}, \quad \widetilde{\mathbf{R}}_{\mathcal{V}_{\mathbf{f}}}^{\partial}(\mathbf{w})|_{\mathbf{v}} = \int_{\widetilde{\mathbf{f}}^{\partial}(\mathbf{v}) \cap \mathbf{f}} \mathbf{w}, \quad \forall \mathbf{w} \in L^1(\mathbf{f}). \quad (3.11c)$$

Observe that $\widetilde{\mathbf{R}}_{\mathcal{V}_{\mathbf{f}}}^{\partial}(\cdot)|_{\mathbf{v}} = 0$ for all $\mathbf{v} \notin V_{c(\mathbf{f})} \cap V^{\partial}$.

3.2 Advection-reaction problem

This section is concerned with the derivation and the analysis of a vertex-based CDO scheme to approximate the solution of the advection-reaction problem

$$\beta \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (3.12a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega^-, \quad (3.12b)$$

with the advective field $\beta \in \mathbf{Lip}(\Omega)$ and the reaction coefficient $\mu \in L^\infty(\Omega)$. We assume that the data are such that $s \in L^2(\Omega)$ and $u_D \in L^2(|\beta \cdot \mathbf{n}|; \partial\Omega^-)$. We also assume that the boundary $\partial\Omega$ is well-separated with respect to β , i.e., the inflow part $\partial\Omega^-$ and the outflow part $\partial\Omega^+$, defined as $\partial\Omega^\pm = \{\mathbf{x} \in \partial\Omega \mid \pm \beta(\mathbf{x}) \cdot \mathbf{n}(\mathbf{x}) > 0\}$, are such that $\text{dist}(\partial\Omega^-, \partial\Omega^+) > 0$. In addition, we assume that the Friedrichs tensor

$$\sigma_{\beta, \mu} := \mu - \frac{1}{2} \nabla \cdot \beta, \quad (3.13)$$

satisfies one of the following assumptions:

($\mathcal{H}$) $\text{ess inf}_\Omega \sigma_{\beta, \mu} > 0$. We define the reference time $\tau = (\text{ess inf}_\Omega \sigma_{\beta, \mu})^{-1}$.

($\mathcal{H}'$) $\text{ess inf}_\Omega \sigma_{\beta, \mu} = 0$ and there exists a non dimensional function $\zeta \in \text{Lip}(\Omega)$ with $\zeta \geq 1$ in Ω such that

$$\text{ess inf}_\Omega -\frac{1}{2} \beta \cdot \nabla \zeta > 0. \quad (3.14)$$

We define the reference time $\tau = \left(\text{ess inf}_\Omega -\frac{1}{2} \beta \cdot \nabla \zeta \right)^{-1}$.

Recalling the continuous analysis of Chapter 2, we observe that assumption ($\mathcal{H}$) stands for the assumption ($\mathcal{H}_2$) whereas ($\mathcal{H}'$) slightly differs from ($\mathcal{H}'_2$) since we only assume here that $\text{ess inf}_\Omega \sigma_{\beta, \mu} = 0$. However, we notice that ($\mathcal{H}'$) implies ($\mathcal{H}'_2$) since

$$\text{ess inf}_\Omega \left(\zeta \sigma_{\beta, \mu} - \frac{1}{2} \beta \cdot \nabla \zeta \right) \geq \text{ess inf}_\Omega \left(-\frac{1}{2} \beta \cdot \nabla \zeta \right) > 0 \quad (3.15)$$

whenever $\text{ess inf}_\Omega \sigma_{\beta, \mu} = 0$ with $\zeta \geq 0$. In addition, note that we have assumed in ($\mathcal{H}'$) that the potential satisfies $\zeta \geq 1$ without loss of generality since (3.14) is invariant by adding any constant to ζ . Under either of the above assumptions, the problem (3.12) is well-posed owing to Theorems 2.7 and 2.11.

Remark 3.9 (Negative Friedrichs tensor). *We actually consider the case where the Friedrichs tensor takes negative values in Chapter 6 concerning the vector advection-reaction problem. Hence, we expect that the present analysis can be extended to the case where $\text{ess inf}_\Omega \sigma_{\beta, \mu} < 0$.*

3.2.1 Discrete Friedrichs problem

Besides the classical CDO discrete operators presented in Section 3.1.3, vertex-based CDO schemes for the advection-reaction problem are built using three additional discrete operators: the discrete contraction operator $J_\beta^\varepsilon : \mathcal{E} \rightarrow \mathcal{V}$ which is the discrete counterpart of the map $\mathbf{w} \mapsto \beta \cdot \mathbf{w}$ and two weighted mass operators, $H_\alpha^\nu : \mathcal{V} \rightarrow \mathcal{V}$ and $H_\alpha^{\nu, \partial} : \mathcal{V} \rightarrow \mathcal{V}$ (with a generic parameter α) that are the discrete counterpart of the map $u \mapsto \alpha u$ in Ω and at the boundary, respectively. Note that the operator $H_\alpha^{\nu, \partial}$ is defined from $\mathcal{V}$ since the information carried on by the internal vertices is typically used to define the value at the boundary, while this operator maps into $\mathcal{V}$ in order to alleviate the notation and we always have $H_\alpha^{\nu, \partial}(\cdot)|_{\mathcal{V}} = 0$ for all $\mathbf{v} \in \mathcal{V}^\circ$.

Bilinear forms. The discrete problem hinges on the bilinear form $\mathbb{A}_{\beta,\mu}^\nu : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$, defined as

$$\mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \mathbf{w}) = \langle\langle \mathbf{J}_\beta^\mathcal{E} \text{GRAD}(\mathbf{v}), \mathbf{w} \rangle\rangle_\mathcal{V} + \langle\langle \mathbf{H}_\mu^\nu(\mathbf{v}), \mathbf{w} \rangle\rangle_\mathcal{V} + \langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus}^{\nu,\partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_\mathcal{V}, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}. \quad (3.16)$$

The first two terms approximate (3.12a), whereas the third term weakly enforces the boundary condition on $\partial\Omega^-$. The discrete gradient $\text{GRAD} : \mathcal{V} \rightarrow \mathcal{E}$ is defined by (3.5a) and the discrete contraction operator $\mathbf{J}_\beta^\mathcal{E} : \mathcal{E} \rightarrow \mathcal{V}$ is defined below. The reactive mass operator $\mathbf{H}_\mu^\nu : \mathcal{V} \rightarrow \mathcal{V}$ is defined as

$$\langle\langle \mathbf{H}_\mu^\nu(\mathbf{v}), \mathbf{w} \rangle\rangle_\mathcal{V} = \int_\Omega \mu \mathbf{L}_\mathcal{V}(\mathbf{v}) \mathbf{L}_\mathcal{V}(\mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \quad (3.17)$$

where we have introduced the reconstruction map $\mathbf{L}_\mathcal{V} : \mathcal{V} \rightarrow \mathbb{P}_0(\tilde{\mathcal{C}})$ defined by

$$\forall \mathbf{v} \in \mathbf{V}, \quad \mathbf{L}_\mathcal{V}(\mathbf{w})|_{\tilde{\mathcal{C}}(\mathbf{v})} = \mathbf{w}_\mathbf{v}, \quad \forall \mathbf{w} \in \mathcal{V}. \quad (3.18)$$

Note that if $\text{ess inf}_\Omega \mu > 0$, $\mathbf{H}_\alpha^\nu$ is positive and defines a Hodge operator in the sense of statement **(H)** in (Bonelle, 2014, Section 3.4.2). Boundary conditions are weakly enforced on the inflow boundary using the operator $\mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus}^{\nu,\partial} : \mathcal{V} \rightarrow \mathcal{V}$, defined by

$$\langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus}^{\nu,\partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_\mathcal{V} = \int_{\partial\Omega} (\beta \cdot \mathbf{n})^\ominus \mathbf{L}_\mathcal{V}^\partial(\mathbf{v}) \mathbf{L}_\mathcal{V}^\partial(\mathbf{w}), \quad (3.19)$$

where we have introduced the reconstruction map $\mathbf{L}_\mathcal{V}^\partial : \mathcal{V} \rightarrow \mathbb{P}_0(\tilde{\mathcal{F}}^\partial)$ on the boundary, such that

$$\forall \mathbf{v} \in \mathbf{V}^\partial, \quad \mathbf{L}_\mathcal{V}^\partial(\mathbf{w})|_{\tilde{\mathcal{F}}^\partial(\mathbf{v})} = \mathbf{w}_\mathbf{v}, \quad \forall \mathbf{w} \in \mathcal{V}. \quad (3.20)$$

Observe that for all interior mesh vertices $\mathbf{v} \in \mathbf{V}^\circ$, $\mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus}^{\nu,\partial}(\mathbf{w})|_\mathbf{v} = 0$ since then $\tilde{\mathcal{F}}^\partial(\mathbf{v}) = \emptyset$.

Discrete problem. The scheme approximating (3.12) reads:

$$\text{Find } \mathbf{u} \in \mathcal{V} \text{ s.t. } \mathbb{A}_{\beta,\mu}^\nu(\mathbf{u}, \mathbf{v}) = \mathbb{S}(s, u_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{V}, \quad (3.21)$$

with the source linear form $\mathbb{S}(s, u_D; \cdot) : \mathcal{V} \rightarrow \mathbb{R}$ defined by

$$\mathbb{S}(s, u_D; \mathbf{v}) := \int_\Omega s \mathbf{L}_\mathcal{V}(\mathbf{v}) + \int_{\partial\Omega} (\beta \cdot \mathbf{n})^\ominus u_D \mathbf{L}_\mathcal{V}^\partial(\mathbf{v}) = \langle\langle \tilde{\mathbf{R}}_\mathcal{V}(s), \mathbf{v} \rangle\rangle_\mathcal{V} + \langle\langle \tilde{\mathbf{R}}_\mathcal{V}^\partial((\beta \cdot \mathbf{n})^\ominus u_D), \mathbf{v} \rangle\rangle_\mathcal{V}, \quad (3.22)$$

with the dual reduction maps $\tilde{\mathbf{R}}_\mathcal{V}$ and $\tilde{\mathbf{R}}_\mathcal{V}^\partial$ defined by (3.8a) and (3.8c), respectively, and the reconstruction map $\mathbf{L}_\mathcal{V}$ defined by (3.18) and (3.20).

Discrete contraction operators. The discrete contraction map $\mathbf{J}_\beta^\mathcal{E} : \mathcal{E} \rightarrow \mathcal{V}$ is built using the algebraic parameters $\{\beta_e\}_{e \in \mathbf{E}}$ and $\{\kappa_{\mathbf{ve}}\}_{(\mathbf{v}, e) \in \mathbf{VE}}$ with $\mathbf{VE} := \{(\mathbf{v}, e) \in \mathbf{V} \times \mathbf{E} \mid \mathbf{v} \subset \partial e\}$. For all $e \in \mathbf{E}$, β_e denotes the reduction of the vector field β on the dual face $\tilde{\mathcal{F}}(e)$:

$$\beta_e := \tilde{\mathbf{R}}_\mathcal{E}(\beta)|_e = \int_{\tilde{\mathcal{F}}(e)} \beta \cdot \mathbf{n}_{\tilde{\mathcal{F}}(e)}, \quad \forall e \in \mathbf{E}. \quad (3.23)$$

For all $\mathbf{v} \in \mathbf{V}$ and for all $e \in \mathbf{E}_\mathbf{v}$, we define $\kappa_{\mathbf{ve}} \in [-1, 1]$ satisfying the two following assumptions:

- (i_κ) $\sum_{\mathbf{v} \in \mathbf{V}_e} \kappa_{\mathbf{ve}} = 0$ and setting $\kappa_e := \frac{1}{2} \sum_{\mathbf{v} \in \mathbf{V}_e} \iota_{\tilde{\mathcal{F}}(e), \tilde{\mathcal{C}}(\mathbf{v})} \kappa_{\mathbf{ve}}$, $\beta_e \kappa_e \geq 0$ holds.
- (ii_κ) There exists $\mathbf{C}_\kappa > 0$, uniform with respect to the mesh size and the model parameters, such that $\beta_e \kappa_e \geq \mathbf{C}_\kappa |\beta_e|$.

Remark 3.10 (Upwind schemes). *The reason to distinguish the property $\beta_e \kappa_e \geq 0$ in (i_κ) from the property $\beta_e \kappa_e \geq \mathbf{C}_\kappa |\beta_e|$ in (ii_κ), is that the former is satisfied by the so-called centered scheme corresponding to $\kappa_{\mathbf{ve}} = 0$ for all $\mathbf{v} \in \mathbf{V}_e$, and the latter by an upwind scheme, e.g., corresponding to $\kappa_{\mathbf{ve}} = \text{sign}(\iota_{\tilde{\mathcal{F}}(e), \tilde{\mathcal{C}}(\mathbf{v})} \beta_e)$, with the sign function defined by $\text{sign}(t) = -1$ if $t \in \mathbb{R}_{<0}$, $\text{sign}(0) = 0$, and $\text{sign}(t) = 1$ if $t \in \mathbb{R}_{>0}$. For that particular choice, (ii_κ) holds with $\mathbf{C}_\kappa = 1$.*

Proposition 3.11 (Jump across dual faces). *Let $\mathbf{w} \in \mathcal{V}$ and let $e \in E$. Define the jump of $\mathbf{w}$ across the dual face $\tilde{f}(e)$ as $[\mathbf{w}]_e = \sum_{v \in V_e} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{w}_v$. Then,*

$$[\mathbf{w}]_e \kappa_e = \sum_{v \in V_e} \mathbf{w}_v \kappa_{ve}. \quad (3.24)$$

Proof. Let $e \in E$ and denote $v, v' \in V_e$ the two vertices that are the end points of the edge e . Owing to the definition of κ_e and of the jump $[\mathbf{w}]_e$, we have

$$[\mathbf{w}]_e \kappa_e = \frac{1}{2} \left(\iota_{\tilde{f}(e), \tilde{c}(v)} \kappa_{ve} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{w}_v + \iota_{\tilde{f}(e), \tilde{c}(v')} \kappa_{v'e} \iota_{\tilde{f}(e), \tilde{c}(v')} \mathbf{w}_{v'} \right. \\ \left. + \iota_{\tilde{f}(e), \tilde{c}(v)} \kappa_{ve} \iota_{\tilde{f}(e), \tilde{c}(v')} \mathbf{w}_{v'} + \iota_{\tilde{f}(e), \tilde{c}(v')} \kappa_{v'e} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{w}_v \right).$$

Hence, observing that $\iota_{\tilde{f}(e), \tilde{c}(v)} \iota_{\tilde{f}(e), \tilde{c}(v)} = 1$ and $\iota_{\tilde{f}(e), \tilde{c}(v)} \iota_{\tilde{f}(e), \tilde{c}(v')} = -1$, it follows that

$$[\mathbf{w}]_e \kappa_e = \frac{1}{2} \sum_{v \in V_e} \mathbf{w}_v \kappa_{ve} - \frac{1}{2} (\mathbf{w}_{v'} \kappa_{ve} + \mathbf{w}_v \kappa_{v'e}),$$

whence the expected result follows from the identity $\sum_{v \in V_e} \kappa_{ve} = 0$ (see assumption (i_κ)). $\square$

In the spirit of the analysis of Friedrichs' systems (see Ern & Guermond (2006a,b)), we introduce at the same time the contraction operator $J_\beta^\varepsilon : \mathcal{E} \rightarrow \mathcal{V}$ and its companion $J_\beta^\nu : \mathcal{V} \rightarrow \mathcal{E}$, which are somehow anti-adjoint, up to the stabilization term (see Lemma 3.14).

Definition 3.12 (Contraction operators). *The contraction operator $J_\beta^\varepsilon : \mathcal{E} \rightarrow \mathcal{V}$ is algebraically defined by*

$$\forall v \in V, \quad J_\beta^\varepsilon(\mathbf{w})|_v := \frac{1}{2} \sum_{e \in E_v} \mathbf{w}_e \beta_e (1 - \kappa_{ve}), \quad \forall \mathbf{w} \in \mathcal{E}. \quad (3.25)$$

The companion operator of $J_\beta^\varepsilon : \mathcal{V} \rightarrow \mathcal{E}$ is defined by

$$\forall e \in E, \quad J_\beta^\nu(\mathbf{w})|_e := \frac{1}{2} \sum_{v \in V_e} \mathbf{w}_v \beta_e (1 + \kappa_{ve}), \quad \forall \mathbf{w} \in \mathcal{V}. \quad (3.26)$$

These operators are jointly analyzed and satisfy the following two Lemmata 3.13 and 3.14.

Lemma 3.13 (Discrete Leibniz rule). *Let J_β^ε and J_β^ν be defined by (3.25) and (3.26), respectively. Assume that (i_κ) holds. Then,*

$$\langle\langle H_{\nabla \cdot \beta}^\nu(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu + \langle\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu - \langle\langle \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu - \langle\langle H_{(\beta \cdot \mathbf{n})}^{\nu, \partial}(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu = 0, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \quad (3.27)$$

where $H_{\nabla \cdot \beta}^\nu$ and $H_{\beta \cdot \mathbf{n}}^{\nu, \partial}$ are defined by (3.17) with $\nabla \cdot \beta$ instead of μ and by (3.19) with $(\beta \cdot \mathbf{n})$ instead of $(\beta \cdot \mathbf{n})^\ominus$, respectively.

Proof. Owing to the definitions (3.25) and (3.26) of J_β^ε and J_β^ν , we infer that

$$\langle\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu = \frac{1}{2} \sum_{v \in V} \mathbf{v}_v \sum_{e \in E_v} \sum_{v' \in V_e} \iota_{v', e} \mathbf{w}_{v'} \beta_e (1 - \kappa_{ve}), \\ \langle\langle \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu = \frac{1}{2} \sum_{v \in V} \mathbf{v}_v \sum_{e \in E_v} \sum_{v' \in V_e} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{w}_{v'} \beta_e (1 + \kappa_{v'e}).$$

Recalling that $\iota_{v, e} = -\iota_{\tilde{f}(e), \tilde{c}(v)}$, it follows that

$$\langle\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu - \langle\langle \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}), \mathbf{v} \rangle\rangle_\nu = \frac{1}{2} \sum_{v \in V} \mathbf{v}_v \sum_{e \in E_v} \sum_{v' \in V_e} \mathbf{w}_{v'} \beta_e (\iota_{v', e} (1 - \kappa_{ve}) + \iota_{v, e} (1 + \kappa_{v'e})).$$

Hence, using (i_κ) and the identity $\sum_{v \in V_e} \iota_{v,e} = 0$, we infer that

$$\sum_{v' \in V_e} \mathbf{w}_{v'} \beta_e (\iota_{v',e} (1 - \kappa_{ve}) + \iota_{v,e} (1 + \kappa_{v'e})) = \mathbf{w}_v (\iota_{v,e} (1 - \kappa_{ve}) + \iota_{v,e} (1 + \kappa_{ve})) = 2\iota_{v,e} \mathbf{w}_v,$$

so that the definition of β_e yields

$$\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{w}), \mathbf{v} \rangle_v - \langle \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}), \mathbf{v} \rangle_v = \sum_{v \in V} \mathbf{v}_v \mathbf{w}_v \sum_{e \in E_v} \iota_{v,e} \beta_e = - \sum_{v \in V} \mathbf{v}_v \mathbf{w}_v \sum_{e \in E_v} \iota_{\tilde{f}(e), \tilde{c}(v)} \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}.$$

To conclude, owing to the divergence theorem, observe that if $v \in V^\circ$,

$$\sum_{e \in E_v} \iota_{\tilde{f}(e), \tilde{c}(v)} \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)} = \int_{\partial \tilde{c}(v)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{c}(v)} = \int_{\tilde{c}(v)} \nabla \cdot \boldsymbol{\beta},$$

with $\mathbf{n}_{\tilde{c}(v)}$ the unit normal at the boundary of $\tilde{c}(v)$ and pointing outward, while for the boundary vertices $v \in V^\partial$, we use the definition of the boundary Hodge operator to infer that

$$\sum_{v \in V^\partial} \mathbf{v}_v \mathbf{w}_v \sum_{e \in E_v} \iota_{\tilde{f}(e), \tilde{c}(v)} \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)} = \sum_{v \in V^\partial} \mathbf{v}_v \mathbf{w}_v \int_{\partial \tilde{c}(v)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{c}(v)} - \langle H_{\boldsymbol{\beta} \cdot \mathbf{n}}^{\nu, \partial}(\mathbf{w}), \mathbf{v} \rangle_v,$$

whence the expected result. $\square$

Lemma 3.14 (Integration by parts). *Let J_β^ε and J_β^ν be defined by (3.25) and (3.26), respectively. Assume that (i_κ) holds. Then, the bilinear form defined on $\mathcal{V} \times \mathcal{V}$ by*

$$s_\beta^\nu(\mathbf{v}, \mathbf{w}) := \langle J_\beta^\varepsilon \text{GRAD}(\mathbf{v}), \mathbf{w} \rangle_v + \langle \mathbf{v}, \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}) \rangle_v, \quad (3.28)$$

defines a semi-inner product and we have

$$s_\beta^\nu(\mathbf{v}, \mathbf{w}) = \sum_{e \in E} [\mathbf{v}]_e [\mathbf{w}]_e \kappa_e \beta_e, \quad (3.29)$$

with $[\mathbf{w}]_e$ the jump of $\mathbf{w}$ across the dual face $\tilde{f}(e)$, defined in Proposition 3.11.

Proof. Owing to the definition of J_β^ε and J_β^ν and recalling that $\iota_{v,e} = -\iota_{\tilde{f}(e), \tilde{c}(v)}$, we infer that

$$\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{v}), \mathbf{w} \rangle_v = \frac{1}{2} \sum_{v \in V} \sum_{e \in E_v} \sum_{v' \in V_e} \iota_{v',e} \mathbf{w}_{v'} \mathbf{v}_v \beta_e (1 - \kappa_{ve}) = \frac{1}{2} \sum_{v \in V} \sum_{e \in E_v} \mathbf{w}_v [\mathbf{v}]_e \beta_e (\kappa_{ve} - 1), \quad (3.30)$$

and

$$\langle \mathbf{v}, \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}) \rangle_v = \frac{1}{2} \sum_{v \in V} \sum_{e \in E_v} \sum_{v' \in V_e} \iota_{\tilde{f}(e), \tilde{c}(v)} \mathbf{v}_v \mathbf{w}_{v'} \beta_e (1 + \kappa_{v'e}) = \frac{1}{2} \sum_{e \in E} \sum_{v \in V_e} \mathbf{w}_v [\mathbf{v}]_e \beta_e (1 + \kappa_{ve}). \quad (3.31)$$

Exchanging the summations in the first line leads to

$$\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{v}), \mathbf{w} \rangle_v + \langle \mathbf{v}, \widetilde{\text{DIV}} J_\beta^\nu(\mathbf{w}) \rangle_v = \sum_{e \in E} \sum_{v \in V_e} \mathbf{w}_v [\mathbf{v}]_e \kappa_{ve} \beta_e.$$

Then (3.29) follows from the identity $\sum_{v \in V_e} \mathbf{w}_v \kappa_{ve} = [\mathbf{w}]_e \kappa_e$ from Proposition 3.11. $\square$

Remark 3.15 (Interpretations). *Recalling from Remark 3.6 that the discrete dual divergence operator $\widetilde{\text{DIV}}$ does not involve faces on the boundary $\partial\Omega$, Lemma 3.13 is the discrete counterpart of the Leibniz formula $\int_\Omega \nabla \cdot \boldsymbol{\beta} v w + \int_\Omega (\boldsymbol{\beta} \cdot \nabla w) v - \int_\Omega \nabla \cdot (\boldsymbol{\beta} w) v = 0$, where the two rightmost terms in (3.27) form together the discrete counterpart of $\int_\Omega \nabla \cdot (\boldsymbol{\beta} w) v$. Furthermore, Lemma 3.14 is the discrete counterpart of the integration by parts formula $\int_\Omega (\boldsymbol{\beta} \cdot \nabla w) v + \int_\Omega \nabla \cdot (\boldsymbol{\beta} v) w - \int_{\partial\Omega} w (\boldsymbol{\beta} \cdot \mathbf{n}) v = 0$. At the discrete level, this quantity is not equal to zero owing to the use of stabilization. We also notice that the symmetry of the map $s_\beta^\nu(\mathbf{v}, \mathbf{w})$ results from $s_\beta^\nu(\mathbf{v}, \mathbf{w}) - s_\beta^\nu(\mathbf{w}, \mathbf{v}) = \langle H_{\nabla \cdot \boldsymbol{\beta}}^\nu(\mathbf{v}), \mathbf{w} \rangle_v - \langle \mathbf{v}, H_{\nabla \cdot \boldsymbol{\beta}}^\nu(\mathbf{w}) \rangle_v = 0$ where we have used the self-adjointness of $H_{\nabla \cdot \boldsymbol{\beta}}^\nu$ and of $H_{\boldsymbol{\beta} \cdot \mathbf{n}}^{\nu, \partial}$. Finally, we observe in general that neither J_β^ε nor J_β^ν depend linearly on its argument $\boldsymbol{\beta}$ owing to the use of stabilization.*

Remark 3.16 (Conservative formulation). *A possible variant of the continuous problem (3.12) is the conservative form of the advective derivative. The equation (3.12a) becomes $\nabla \cdot (\beta u) + \mu' u = s$ a.e. in Ω with $\mu' = \mu - \nabla \cdot \beta$ and the Dirichlet condition is still enforced on the inflow boundary. Then, the stability hinges on the modified assumption $(\mathcal{H})$ by replacing $\sigma_{\beta, \mu}$ by $\mu + \frac{1}{2} \nabla \cdot \beta$. The discrete bilinear form $\mathbb{A}_{\beta, \mu}^\nu$ in (3.21) then becomes*

$$\mathbb{A}_{\beta, \mu}^\nu(u, v) = \langle \widetilde{\text{DIV}} J_\beta^\nu(u), v \rangle_\nu + \langle H_{\mu'}^\nu(u), v \rangle_\nu + \langle H_{(\beta \cdot \mathbf{n})^\oplus}^{\nu, \partial}(u), v \rangle_\nu.$$

The source term remains unchanged.

Link with other advection schemes. For some particular choices of the algebraic parameters κ_{ve} , the scheme (3.21) is equivalent to the lowest order dG scheme on dual cells attached to primal vertices. Indeed, recalling the expression of the dG bilinear form a_h^{upw} (with no reactive term) from (Di Pietro & Ern, 2012, Section 2.3.1), for all $v_h, w_h \in \mathbb{P}_k(\tilde{\mathcal{C}})$ with $k \geq 0$, we have

$$\begin{aligned} a_h^{\text{upw}}(v_h, w_h) = & \sum_{v \in V} \int_{\tilde{\mathcal{C}}(v)} \beta \cdot \nabla v_h w_h + \sum_{e \in E} \int_{\tilde{\mathbf{f}}(e)} \llbracket v_h \rrbracket \left(\frac{1}{2} |\beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}| \llbracket w_h \rrbracket - \beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)} \{ \{ w_h \} \} \right) \\ & + \sum_{v \in V^\partial} \int_{\tilde{\mathbf{f}}^\partial(v)} (\beta \cdot \mathbf{n})^\ominus v_h w_h, \end{aligned}$$

with $\llbracket \cdot \rrbracket$ and $\{ \{ \cdot \} \}$ denoting the jump and the average operators, respectively. Then, choosing $v_h = L_\nu(v)$ and $w_h = L_\nu(w)$, the first term on the right-hand side vanishes, and rewriting the second term, we obtain

$$a_h^{\text{upw}}(L_\nu(v), L_\nu(w)) = \frac{1}{2} \sum_{e \in E} [v]_e \left([w]_e \int_{\tilde{\mathbf{f}}(e)} |\beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}| - \beta_e \sum_{v \in V_e} w_v \right) + \sum_{v \in V^\partial} \int_{\tilde{\mathbf{f}}^\partial(v)} (\beta \cdot \mathbf{n})^\ominus L_\nu^\partial(v) L_\nu^\partial(w). \quad (3.32)$$

Comparing (3.32) with (3.30), we infer that

$$a_h^{\text{upw}}(L_\nu(v), L_\nu(w)) = \langle J_\beta^\varepsilon \text{GRAD}(v), w \rangle_\nu + \langle H_{(\beta \cdot \mathbf{n})^\ominus}^{\nu, \partial}(w), v \rangle_\nu = \mathbb{A}_{\beta, \mu}^\nu(v, w),$$

where, for all $e \in E$ and for all $v \in V_e$, we have $\beta_e(\kappa_{ve} - 1) = \iota_{\tilde{\mathbf{f}}(e), \tilde{\mathcal{C}}(v)} \int_{\tilde{\mathbf{f}}(e)} |\beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}| - \beta_e$, i.e., $\beta_e \kappa_{ve} = \iota_{\tilde{\mathbf{f}}(e), \tilde{\mathcal{C}}(v)} \int_{\tilde{\mathbf{f}}(e)} |\beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}|$. With this choice, assumptions (i_κ) and (ii_κ) are satisfied with $\mathbf{C}_\kappa = 1$. Proceeding similarly, we prove as well that the scheme (3.21) is equivalent to the upwind Finite Volume method (see Eymard *et al.* (2000)), with $\beta_e \kappa_{ve} = \iota_{\tilde{\mathbf{f}}(e), \tilde{\mathcal{C}}(v)} \int_{\tilde{\mathbf{f}}(e)} \beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}$.

Remark 3.17 (Upwinding). *Generally speaking, there are several possible variations in the geometric quantities considered for upwinding. Instead of considering the full dual face $\tilde{\mathbf{f}}(e)$ in (3.23), one possibility is to consider the average of the normal advection velocity on the dual sub-faces $\tilde{\mathbf{f}}_c(e) = \tilde{\mathbf{f}}(e) \cap c$ and to design the upwinding parameters based on the sign of these quantities. In general, the smaller the underlying geometric objects, the larger the dissipation introduced by upwinding. The advantage of considering the dual sub-faces $\tilde{\mathbf{f}}_c(e)$ is that upwinding is then compatible with the assembly of the scheme on primal cells.*

3.2.2 Analysis

Let us now analyze the scheme (3.21). First, we define the discrete 2-norm on the space $\mathcal{V}$ as

$$\|w\|_{\mathcal{V}, 2}^2 := \sum_{v \in V} |\tilde{\mathcal{C}}(v)| |w_v|^2, \quad \forall w \in \mathcal{V}. \quad (3.33)$$

Observe that this norm can be localized on primal mesh cells as $\|w\|_{\mathcal{V}, 2}^2 = \sum_{c \in \mathcal{C}} \|w\|_{\mathcal{V}_c, 2}^2$, where the local discrete 2-norm on $\mathcal{V}_c$ is defined by $\|w\|_{\mathcal{V}_c, 2}^2 := \sum_{v \in V_c} |\tilde{\mathcal{C}}(v) \cap c| |w_v|^2$.

Stability. For all $\mathbf{v} \in \mathcal{V}$, the stability norm used to analyze (3.21) is defined by

$$\|\mathbf{v}\|_{\mathcal{V},\mathbf{a}}^2 := \tau^{-1} \|\mathbf{v}\|_{\mathcal{V},2}^2 + \mathbf{s}_{\beta}^{\mathcal{V}}(\mathbf{v}, \mathbf{v}) + \langle \mathbf{H}_{|\beta \cdot \mathbf{n}|}^{\mathcal{V},\partial}(\mathbf{v}), \mathbf{v} \rangle_{\mathcal{V}}, \quad (3.34)$$

where the parameter $\tau > 0$ is defined by assumption $(\mathcal{H})$ or $(\mathcal{H}')$, $\|\cdot\|_{\mathcal{V},2}$ is defined by (3.33), the bilinear form $\mathbf{s}_{\beta}^{\mathcal{V}}(\cdot, \cdot)$ by (3.29) and the operator $\mathbf{H}_{|\beta \cdot \mathbf{n}|}^{\mathcal{V},\partial}$ by (3.19). To alleviate the reading, the proofs are postponed to the end of this Section.

Lemma 3.18 (Coercivity). *Assume that assumption $(\mathcal{H})$ holds. Then,*

$$\mathbf{A}_{\beta,\mu}^{\mathcal{V}}(\mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V},\mathbf{a}}^2, \quad \forall \mathbf{v} \in \mathcal{V}.$$

Consequently, the problem (3.21) is well-posed.

The stability of the bilinear form $\mathbf{A}_{\beta,\mu}^{\mathcal{V}}$ is now studied under the assumption $(\mathcal{H}')$. We denote $L_{\zeta} = |\zeta|_{W^{1,\infty}(\Omega)}$ and $L_{\beta} = |\beta|_{W^{1,\infty}(\Omega)}$ the Lipschitz constant of ζ and β . To simplify the tracking in the analysis of the dependency on the model parameter, we assume that there exists a constant $\mathbf{C}_{\mathcal{V},\mathbf{a}} > 0$, independent of the mesh-size and the physical parameters, such that

$$\tau \max \left(L_{\beta}, L_{\zeta} h \|\mu\|_{L^{\infty}(\Omega)}, L_{\zeta}^2 h \|\beta\|_{L^{\infty}(\Omega)} \right) \leq \mathbf{C}_{\mathcal{V},\mathbf{a}}, \quad (3.35)$$

where $h = \max_{c \in \mathcal{C}} h_c$ denotes the size of mesh and h_c denotes the diameter of c . For instance, $\mathbf{C}_{\mathcal{V},\mathbf{a}} \leq 1$ yields $\tau L_{\zeta} h \|\mu\|_{L^{\infty}(\Omega)} \leq 1$, meaning that we do not consider strongly reactive regimes. In addition, we introduce the reference length

$$h_0 = \left(L_{\zeta} \tau \|\nabla \cdot \beta\|_{L^{\infty}(\Omega)} \right)^{-1} \quad (3.36)$$

where by convention, $h_0 = +\infty$ if β is divergence-free.

Lemma 3.19 (Inf-sup stability). *Assume that assumption $(\mathcal{H}')$ holds. Assume that (3.35) holds and that the mesh-size satisfies $h < h_0$. Then, there is $\varrho > 0$ such that*

$$\sup_{\mathbf{w} \in \mathcal{V}; \|\mathbf{w}\|_{\mathcal{V},\mathbf{a}}=1} \mathbf{A}_{\beta,\mu}^{\mathcal{V}}(\mathbf{v}, \mathbf{w}) \geq \varrho \|\mathbf{v}\|_{\mathcal{V},\mathbf{a}}, \quad \forall \mathbf{v} \in \mathcal{V}.$$

Consequently, the problem (3.21) is well-posed.

Remark 3.20 (Comparison with Cantin & Ern (2016b)). *Lemma 3.19 extends Lemma 5.1 from Cantin & Ern (2016b) where the simpler case $\nabla \cdot \beta = 0$ and $\mu = 0$ a.e. in Ω is considered with no mesh-size restriction.*

The following Table 3.1 recapitulates the different conditions from Lemmata 3.18 and 3.19 for which the discrete problem (3.21) is well-posed.

Case	$\text{ess inf}_{\Omega} \sigma_{\beta,\mu} > 0$	$\text{ess inf}_{\Omega} \sigma_{\beta,\mu} = 0$	
Assumption	$(\mathcal{H})$	$(\mathcal{H}')$	
Advective field		$\nabla \cdot \beta = 0$	$\nabla \cdot \beta \neq 0$
Mesh-size	$0 < h$	$0 < h$	$0 < h < h_0$

Table 3.1 – Stability conditions for the discrete problem (3.21) with h_0 defined by (3.36).

Proof of Lemma 3.18. Using Lemmata 3.13 and 3.14, we observe that

$$\langle\langle J_\beta^\varepsilon \text{GRAD}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}} = -\frac{1}{2} \langle\langle \mathbf{H}_{\nabla \cdot \beta}^\nu(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}} + \frac{1}{2} s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2} \langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})}^{\nu, \partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}},$$

so that, using the linear dependence of $\mathbf{H}_\alpha^\nu$ with α , we rewrite the quantity $\mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \mathbf{v})$ as

$$\mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \mathbf{v}) = \langle\langle \mathbf{H}_{\sigma_{\beta, \mu}}^\nu(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}} + \frac{1}{2} s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2} \langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})}^{\nu, \partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}} + \frac{1}{2} \langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus}^{\nu, \partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}}.$$

Hence, since $\mathbf{H}_\alpha^{\nu, \partial}$ for all $\alpha \in L^\infty(\partial\Omega)$ linearly depends as well on α , we infer that

$$\mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \mathbf{v}) = \langle\langle \mathbf{H}_{\sigma_{\beta, \mu}}^\nu(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}} + \frac{1}{2} s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2} \langle\langle \mathbf{H}_{|\beta \cdot \mathbf{n}|}^{\nu, \partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}}. \quad (3.37)$$

The expected result then follows from the definition (3.17) of $\mathbf{H}_\alpha^\nu$ and from the assumption $(\mathcal{H})$. $\square$

To alleviate the proof of Lemma 3.19, we first consider the following proposition.

Proposition 3.21 (Multiplicative stability). *Let $\mathbf{v} \in \mathcal{V}$ and define $\zeta \mathbf{v} \in \mathcal{V}$ such that $(\zeta \mathbf{v})_{\mathbf{v}} = \zeta(\mathbf{x}_{\mathbf{v}}) \mathbf{v}_{\mathbf{v}}$. Assume that (3.35) holds. Then,*

$$\|\zeta \mathbf{v}\|_{\mathcal{V}, \mathbf{a}} \leq \mathcal{C}_\zeta \|\mathbf{v}\|_{\mathcal{V}, \mathbf{a}}, \quad \forall \mathbf{v} \in \mathcal{V}. \quad (3.38)$$

where $\mathcal{C}_\zeta$ linearly depends on $(\|\zeta\|_{L^\infty}^2 + \mathcal{C}_{\mathcal{V}, \mathbf{a}})^{\frac{1}{2}}$.

Proof. Let $\mathbf{v} \in \mathcal{V}$ and let $\zeta \mathbf{v} \in \mathcal{V}$. Owing to the definition of the norm $\|\cdot\|_{\mathcal{V}, 2}$, we readily infer that $\|\zeta \mathbf{v}\|_{\mathcal{V}, 2} \leq \|\zeta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{\mathcal{V}, 2}$. Similarly, $\langle\langle \mathbf{H}_{|\beta \cdot \mathbf{n}|}^{\nu, \partial}(\zeta \mathbf{v}), \zeta \mathbf{v} \rangle\rangle_{\mathcal{V}} \leq \|\zeta\|_{L^\infty(\Omega)}^2 \langle\langle \mathbf{H}_{|\beta \cdot \mathbf{n}|}^{\nu, \partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{V}}$ so that it remains to bound $s_\beta^\nu(\zeta \mathbf{v}, \zeta \mathbf{v})$. Owing to Lemma 3.14 and recalling from assumption (i_κ) that $\beta_e \kappa_e \geq 0$, we observe that

$$s_\beta^\nu(\zeta \mathbf{v}, \zeta \mathbf{v}) = \sum_{e \in \mathbf{E}} [\zeta \mathbf{v}]_e^2 \beta_e \kappa_e \leq 2 \sum_{e \in \mathbf{E}} \left(\{\zeta\}_e^2 [\mathbf{v}]_e^2 + \{\mathbf{v}\}_e^2 [\zeta]_e^2 \right) \beta_e \kappa_e,$$

where, for all $e \in \mathbf{E}$, we have defined $\{\zeta\}_e = \frac{1}{2} \sum_{\mathbf{v} \in \mathbf{V}_e} \zeta(\mathbf{x}_{\mathbf{v}})$, $[\zeta]_e = \sum_{\mathbf{v} \in \mathbf{V}_e} \iota_{\tilde{\mathbf{f}}(e), \tilde{\mathbf{c}}(\mathbf{v})} \zeta(\mathbf{x}_{\mathbf{v}})$, $\{\mathbf{v}\}_e = \frac{1}{2} \sum_{\mathbf{v} \in \mathbf{V}_e} \mathbf{v}_{\mathbf{v}}$ and $[\mathbf{v}]_e = \sum_{\mathbf{v} \in \mathbf{V}_e} \iota_{\tilde{\mathbf{f}}(e), \tilde{\mathbf{c}}(\mathbf{v})} \mathbf{v}_{\mathbf{v}}$ (already defined in Lemma 3.14). Noting that $\{\zeta\}_e \leq \|\zeta\|_{L^\infty(\Omega)}$ and that $[\zeta]_e \leq L_\zeta h_e$ since ζ is Lipschitz, we infer that

$$s_\beta^\nu(\zeta \mathbf{v}, \zeta \mathbf{v}) \leq 2 \|\zeta\|_{L^\infty(\Omega)}^2 s_\beta^\nu(\mathbf{v}, \mathbf{v}) + 2 \sum_{e \in \mathbf{E}} \{\mathbf{v}\}_e^2 L_\zeta^2 h_e^2 \beta_e \kappa_e.$$

Hence, since $\{\mathbf{v}\}_e^2 \leq \frac{1}{2} \sum_{\mathbf{v} \in \mathbf{V}_e} \mathbf{v}_{\mathbf{v}}^2$ and observing that $0 \leq \beta_e \kappa_e \leq \|\beta\|_{L^\infty(\Omega)} |\tilde{\mathbf{f}}(e)|$, it follows from mesh regularity **(Sh)** that

$$s_\beta^\nu(\zeta \mathbf{v}, \zeta \mathbf{v}) \lesssim \|\zeta\|_{L^\infty(\Omega)}^2 s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \left(L_\zeta^2 h \|\beta\|_{L^\infty(\Omega)} \max_{\mathbf{v} \in \mathbf{V}} \#E_{\mathbf{v}} \right) \sum_{\mathbf{v} \in \mathbf{V}} |\tilde{\mathbf{c}}(\mathbf{v})| |\mathbf{v}|_{\mathbf{v}}^2,$$

whence $s_\beta^\nu(\zeta \mathbf{v}, \zeta \mathbf{v}) \lesssim \|\zeta\|_{L^\infty(\Omega)}^2 s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \mathcal{C}_{\mathcal{V}, \mathbf{a}} \tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2$, using the mesh regularity **(C)** to bound $\max_{\mathbf{v} \in \mathbf{V}} \#E_{\mathbf{v}}$ and recalling assumption (3.35). Collecting these estimates finally yields the desired result. $\square$

Proof of Lemma 3.19. Let $\mathbf{v} \in \mathcal{V}$ and let $\zeta \mathbf{v} \in \mathcal{V}$. Owing to Proposition 3.21, there is $\mathcal{C}_{\zeta, 1} > 0$ such that $\|\zeta \mathbf{v}\|_{\mathcal{V}, \mathbf{a}} \leq \mathcal{C}_{\zeta, 1} \|\mathbf{v}\|_{\mathcal{V}, \mathbf{a}}$. Next, let us find $\theta_a > 0$ and $\mathcal{C}_{\zeta, a} > 0$ such that $\mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \zeta \mathbf{v} + \theta_a \mathbf{v}) \geq \mathcal{C}_{\zeta, a} \|\mathbf{v}\|_{\mathcal{V}, \mathbf{a}}^2$. First, we rewrite $\mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \zeta \mathbf{v})$ as

$$\begin{aligned} \mathbb{A}_{\beta, \mu}^\nu(\mathbf{v}, \zeta \mathbf{v}) &= \mathbb{A}_{\beta, \frac{1}{2} \nabla \cdot \beta}^\nu(\mathbf{v}, \zeta \mathbf{v}) + \langle\langle \mathbf{H}_{\sigma_{\beta, \mu}}^\nu(\mathbf{v}), \zeta \mathbf{v} \rangle\rangle_{\mathcal{V}} = \mathbb{A}_{\zeta \beta, \frac{1}{2} \zeta \nabla \cdot \beta}^\nu(\mathbf{v}, \mathbf{v}) \\ &\quad + \mathbb{A}_{\beta, \frac{1}{2} \nabla \cdot \beta}^\nu(\mathbf{v}, \zeta \mathbf{v}) - \mathbb{A}_{\zeta \beta, \frac{1}{2} \zeta \nabla \cdot \beta}^\nu(\mathbf{v}, \mathbf{v}) \\ &\quad + \langle\langle \mathbf{H}_{\sigma_{\beta, \mu}}^\nu(\mathbf{v}), \zeta \mathbf{v} \rangle\rangle_{\mathcal{V}} = T_1 + T_2 + T_3, \end{aligned}$$

with $\sigma_{\beta,\mu} = \mu - \frac{1}{2}\nabla \cdot \beta$ and where we have used the linear dependency of H_α^ν with α . Proceeding as in the proof of Lemma 3.18, we observe that

$$T_1 = \langle H_{\sigma_{\beta,\frac{1}{2}\zeta\nabla\cdot\beta}}^\nu(\mathbf{v}), \mathbf{v} \rangle_\nu + \frac{1}{2}s_{\zeta\beta}^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2}\langle H_{|\zeta\beta\cdot\mathbf{n}|}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu,$$

where $\sigma_{\beta,\frac{1}{2}\zeta\nabla\cdot\beta} = -\frac{1}{2}\beta\cdot\nabla\zeta$, so that, using $\zeta \geq 1$, it follows that

$$T_1 \geq \langle H_{-\frac{1}{2}\beta\cdot\nabla\zeta}^\nu(\mathbf{v}), \mathbf{v} \rangle_\nu + \frac{1}{2}s_{\zeta\beta}^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2}\langle H_{|\beta\cdot\mathbf{n}|}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu.$$

Now, let's turn to the term T_2 . By definition,

$$T_2 = \left(\langle J_\beta^\varepsilon(\text{GRAD}(\mathbf{v})), \zeta\mathbf{v} \rangle_\nu - \langle J_{\zeta\beta}^\varepsilon(\text{GRAD}(\mathbf{v})), \mathbf{v} \rangle_\nu \right) + \left(\langle H_{\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \zeta\mathbf{v}) \rangle_\nu - \langle H_{\frac{1}{2}\zeta\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v}) \rangle_\nu \right) \\ \left(\langle H_{(\beta\cdot\mathbf{n})^\ominus}^{\nu,\partial}(\mathbf{v}), \zeta\mathbf{v} \rangle_\nu - \langle H_{(\zeta\beta\cdot\mathbf{n})^\ominus}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu \right) = T_{2,1} + T_{2,2} + T_{2,3},$$

with

$$T_{2,1} = s_\beta^\nu(\mathbf{v}, \zeta\mathbf{v}) - s_{\zeta\beta}^\nu(\mathbf{v}, \mathbf{v}) + \langle \mathbf{v}, \widetilde{\text{DIV}}(J_\beta^\nu(\zeta\mathbf{v})) \rangle_\nu - \langle \mathbf{v}, \widetilde{\text{DIV}}(J_{\zeta\beta}^\nu(\mathbf{v})) \rangle_\nu,$$

owing to Lemma (3.14). Hence, using the identity (3.29), we rewrite $T_{2,1} = T_{2,a} + T_{2,b}$ where

$$T_{2,a} = \frac{1}{2} \sum_{e \in E} [v]_e^2 \zeta_e \beta_e \kappa_e - \frac{1}{2} s_{\zeta\beta}^\nu(\mathbf{v}, \mathbf{v}) \\ T_{2,b} = s_\beta^\nu(\mathbf{v}, \zeta\mathbf{v}) - \frac{1}{2} \sum_{e \in E} [v]_e^2 \zeta_e \beta_e \kappa_e + \langle \mathbf{v}, \widetilde{\text{DIV}}(J_\beta^\nu(\zeta\mathbf{v})) \rangle_\nu - \langle \mathbf{v}, \widetilde{\text{DIV}}(J_{\zeta\beta}^\nu(\mathbf{v})) \rangle_\nu.$$

where ζ_e denotes the mean value of ζ along the edge e . Then, using the identity (3.31), the fact that $\zeta \geq 1$ and the definition of s_β^ν , we have

$$T_{2,a} \geq \frac{1}{2} s_\beta^\nu(\mathbf{v}, \mathbf{v}) - \frac{1}{2} s_{\zeta\beta}^\nu(\mathbf{v}, \mathbf{v}), \\ T_{2,b} = \frac{1}{2} \sum_{\mathbf{v} \in V} \sum_{e \in E_{\mathbf{v}}} v_{\mathbf{v}} [v]_e (\zeta(\mathbf{x}_{\mathbf{v}}) - \zeta_e) \beta_e \kappa_e + \frac{1}{2} \sum_{e \in E} [v]_e \sum_{\mathbf{v} \in V_e} v_{\mathbf{v}} (1 + \kappa_{ve}) \int_{\tilde{\mathbf{f}}(e)} (\zeta(\mathbf{x}_{\mathbf{v}}) - \zeta) \beta \cdot \mathbf{n}_{\tilde{\mathbf{f}}(e)}.$$

Next, using the Lipschitz regularity of ζ and assumptions (ii_κ) and (3.35), it follows that

$$|T_{2,b}| \leq 2s_\beta^\nu(\mathbf{v}, \mathbf{v})^{\frac{1}{2}} \left(C_{\nu,a} \tau^{-1} \|\mathbf{v}\|_{V,2}^2 \right)^{\frac{1}{2}}.$$

Turning now to $T_{2,2}$ and $T_{2,3}$ and still using regularity of ζ , we easily infer that

$$|T_{2,2}| \leq \frac{1}{2} L_\zeta h \sum_{\mathbf{v} \in V} |\tilde{c}(\mathbf{v})| |\mathbf{v}_{\mathbf{v}}|^2 \|\nabla \cdot \beta\|_{L^\infty(\tilde{c}(\mathbf{v}))} \leq \frac{1}{2} L_\zeta h \|\nabla \cdot \beta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{V,2}^2,$$

and

$$|T_{2,3}| \leq \langle H_{|\beta\cdot\mathbf{n}|}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu^{\frac{1}{2}} \left(C_{\nu,a} \tau^{-1} \|\mathbf{v}\|_{V,2}^2 \right)^{\frac{1}{2}}.$$

Hence, there is $C_{\zeta,2} > 0$ such that $|T_{2,2} + T_{2,3}| \leq C_{\zeta,2} \left(s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \langle H_{|\beta\cdot\mathbf{n}|}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu \right)^{\frac{1}{2}} \left(C_{\nu,a} \tau^{-1} \|\mathbf{v}\|_{V,2}^2 \right)^{\frac{1}{2}}$. In addition, observing that $A_{\beta,\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v}) = \frac{1}{2} s_\beta^\nu(\mathbf{v}, \mathbf{v}) + \frac{1}{2} \langle H_{|\beta\cdot\mathbf{n}|}^{\nu,\partial}(\mathbf{v}), \mathbf{v} \rangle_\nu$ owing to the identity (3.37), we deduce

$$|T_{2,2} + T_{2,3}| \leq C_{\zeta,2} A_{\beta,\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v})^{\frac{1}{2}} \left(C_{\nu,a} \tau^{-1} \|\mathbf{v}\|_{V,2}^2 \right)^{\frac{1}{2}},$$

As a result, collecting T_1 and T_2 and using Young's inequality along with assumption $(\mathcal{H}')$, we obtain

$$T_1 + T_2 \geq \frac{1}{2} \|\mathbf{v}\|_{V,a}^2 - C_{\zeta,2}^2 A_{\beta,\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v}) - \frac{1}{2} L_\zeta h \|\nabla \cdot \beta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{V,2}^2.$$

Now, recalling that $\mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \zeta\mathbf{v}) = T_1 + T_2 + T_3$, it follows that

$$\mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \zeta\mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V},a}^2 - \mathbb{C}_{\zeta,2}^2 \mathbb{A}_{\beta,\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v}) + \langle \mathbb{H}_{\sigma_{\beta,\mu}}^\nu(\mathbf{v}), \zeta\mathbf{v} \rangle_{\mathcal{V}} - \frac{1}{2} L_\zeta h \|\nabla\cdot\beta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{\mathcal{V},2}^2.$$

Now, since $\mathbb{A}_{\beta,\frac{1}{2}\nabla\cdot\beta}^\nu(\mathbf{v}, \mathbf{v}) = \mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \mathbf{v}) - \langle \mathbb{H}_{\sigma_{\beta,\mu}}^\nu(\mathbf{v}), \mathbf{v} \rangle_{\mathcal{V}}$ by definition and observing that we have $\langle \mathbb{H}_{\sigma_{\beta,\mu}}^\nu(\mathbf{v}), \zeta\mathbf{v} \rangle_{\mathcal{V}} \geq 0$ and $\langle \mathbb{H}_{\sigma_{\beta,\mu}}^\nu(\mathbf{v}), \mathbf{v} \rangle_{\mathcal{V}} \geq 0$ since $\text{ess inf}_\Omega \sigma_{\beta,\mu} = 0$ by assumption, we arrive at

$$\mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \zeta\mathbf{v} + \theta_a \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V},a}^2 + (\theta_a - \mathbb{C}_{\zeta,2}^2) \mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \mathbf{v}) - \frac{1}{2} L_\zeta h \|\nabla\cdot\beta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{\mathcal{V},2}^2,$$

so that choosing $\theta_a = \mathbb{C}_{\zeta,2}^2$ yields $\mathbb{A}_{\beta,\mu}^\nu(\mathbf{v}, \zeta\mathbf{v} + \theta_a \mathbf{v}) \geq \frac{1}{2} \left(1 - \frac{h}{h_0}\right) \|\mathbf{v}\|_{\mathcal{V},a}^2$. The expected result holds as soon as $h < h_0$ with h_0 defined by (3.36). $\square$

Error bound and *a priori* error estimate The study of the consistency error of the scheme (3.21) relies on commutators in the spirit of Bossavit (2000), Hiptmair (2001), and Bonelle & Ern (2014a).

Definition 3.22 (Commutators). *For all $v \in W^{s,p}(\Omega)$ with $sp > 1$, we define*

$$[\mathbb{J}_\beta^\nu, \mathbb{R}](v) := \mathbb{J}_\beta^\nu(\widehat{\mathbb{R}}_\mathcal{V}(v)) - \widetilde{\mathbb{R}}_\mathcal{E}(\beta v), \quad (3.39a)$$

with $\widehat{\mathbb{R}}_\mathcal{V}$ and $\widetilde{\mathbb{R}}_\mathcal{E}$ defined by (3.7) and (3.8b) respectively,

$$[\mathbb{H}_{\mu-\nabla\cdot\beta}^\nu, \mathbb{R}](v) := \mathbb{H}_{\mu-\nabla\cdot\beta}^\nu(\widehat{\mathbb{R}}_\mathcal{V}(v)) - \widetilde{\mathbb{R}}_\mathcal{V}((\mu - \nabla\cdot\beta)v), \quad (3.39b)$$

with $\widetilde{\mathbb{R}}_\mathcal{V}$ defined by (3.8a) and

$$[\mathbb{H}_{(\beta\cdot\mathbf{n})^\oplus}^{\nu,\partial}, \mathbb{R}](v) := \mathbb{H}_{(\beta\cdot\mathbf{n})^\oplus}^{\nu,\partial}(\widehat{\mathbb{R}}_\mathcal{V}(v)) - \widetilde{\mathbb{R}}_\mathcal{V}^\partial((\beta\cdot\mathbf{n})^\oplus v), \quad (3.39c)$$

with $\widetilde{\mathbb{R}}_\mathcal{V}^\partial$ defined by (3.8c). Note that (3.39b) holds for all $v \in L^1(\Omega)$.

Lemma 3.23 (Consistency error). *Let u be the unique solution of (3.12). Assume that $u \in W^{s,p}(\Omega)$ with $sp > 1$. Then, for all $\mathbf{w} \in \mathcal{V}$, the consistency error $\mathbb{E}_\mathcal{V}(u, \mathbf{w}) = \mathbb{A}_{\beta,\mu}^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u), \mathbf{w}) - \mathbb{S}(s, u_D; \mathbf{w})$ is given by*

$$\mathbb{E}_\mathcal{V}(u, \mathbf{w}) = \langle [\mathbb{H}_{\mu-\nabla\cdot\beta}^\nu, \mathbb{R}](u), \mathbf{w} \rangle_{\mathcal{V}} - \langle [\mathbb{J}_\beta^\nu, \mathbb{R}](u), \text{GRAD}(\mathbf{w}) \rangle_{\mathcal{V}} - \langle [\mathbb{H}_{(\beta\cdot\mathbf{n})^\oplus}^{\nu,\partial}, \mathbb{R}](u), \mathbf{w} \rangle_{\mathcal{V}}.$$

Proof. In the context of Friedrichs' systems, the derivation of the consistency error bound hinges on integration by parts. In the CDO framework, we use the continuous and the discrete Leibniz formulae, as well as the properties of the discrete differential operators. Owing to the definition (3.22) of the source term, we observe that

$$\begin{aligned} \mathbb{S}(s, u_D; \mathbf{w}) &= \langle \widetilde{\mathbb{R}}_\mathcal{V}(\nabla\cdot(\beta u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}((\mu - \nabla\cdot\beta)u), \mathbf{w} \rangle_{\mathcal{V}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}^\partial((\beta\cdot\mathbf{n})^\ominus u_D), \mathbf{w} \rangle_{\mathcal{V}} \\ &= \langle \widetilde{\text{DIV}} \widetilde{\mathbb{R}}_\mathcal{E}(\beta u), \mathbf{w} \rangle_{\mathcal{V}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}((\mu - \nabla\cdot\beta)u), \mathbf{w} \rangle_{\mathcal{V}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}^\partial((\beta\cdot\mathbf{n})^\oplus u_D), \mathbf{w} \rangle_{\mathcal{V}} \\ &= -\langle \widetilde{\mathbb{R}}_\mathcal{E}(\beta u), \text{GRAD}(\mathbf{w}) \rangle_{\mathcal{E}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}((\mu - \nabla\cdot\beta)u), \mathbf{w} \rangle_{\mathcal{V}} + \langle \widetilde{\mathbb{R}}_\mathcal{V}^\partial((\beta\cdot\mathbf{n})^\oplus u_D), \mathbf{w} \rangle_{\mathcal{V}}, \end{aligned}$$

where we have used the continuous Leibniz formula, the commutation property of Proposition 3.5, the fact that $(\beta\cdot\mathbf{n}) = (\beta\cdot\mathbf{n})^\oplus - (\beta\cdot\mathbf{n})^\ominus$, and the discrete anti-adjunction identity inferred in Proposition 3.3. Moreover, we observe that

$$\begin{aligned} \mathbb{A}_{\beta,\mu}^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u), \mathbf{w}) &= \langle \mathbb{J}_\beta^\nu \text{GRAD}(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \mathbb{H}_\mu^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \mathbb{H}_{(\beta\cdot\mathbf{n})^\ominus}^{\nu,\partial}(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} \\ &= \langle \widetilde{\text{DIV}} \mathbb{J}_\beta^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \mathbb{H}_{\mu-\nabla\cdot\beta}^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \mathbb{H}_{(\beta\cdot\mathbf{n})^\oplus}^{\nu,\partial}(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} \\ &= -\langle \mathbb{J}_\beta^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u)), \text{GRAD}(\mathbf{w}) \rangle_{\mathcal{V}} + \langle \mathbb{H}_{\mu-\nabla\cdot\beta}^\nu(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}} + \langle \mathbb{H}_{(\beta\cdot\mathbf{n})^\oplus}^{\nu,\partial}(\widehat{\mathbb{R}}_\mathcal{V}(u)), \mathbf{w} \rangle_{\mathcal{V}}, \end{aligned}$$

where we have used the discrete Leibniz formula from Proposition 3.13, the fact that $\mathbb{H}_\alpha^\nu$ and $\mathbb{H}_\alpha^{\nu,\partial}$ depends linearly on α together with $(\beta\cdot\mathbf{n}) = (\beta\cdot\mathbf{n})^\oplus - (\beta\cdot\mathbf{n})^\ominus$, and the discrete anti-adjunction identity from Proposition 3.3. The conclusion is straightforward. $\square$

Theorem 3.24 (*A priori estimate*). *Let u be the unique solution of (3.12) and let $\mathbf{u}$ be the unique solution of (3.21). Assume that one of the stability assumption from Table 3.1 holds. Assume that $u \in H^1(\Omega)$. Then, the following estimate holds:*

$$\|\mathbf{u} - \widehat{\mathbf{R}}_{\mathcal{V}}(u)\|_{\mathcal{V},a} \lesssim \left(\tau^{\frac{1}{2}} \|\mu - \nabla \beta\|_{L^\infty(\Omega)} h^{\frac{1}{2}} + \|\beta\|_{L^\infty(\Omega)}^{\frac{1}{2}} \right) h^{\frac{1}{2}} |u|_{H^1(\Omega)}. \quad (3.40)$$

Remark 3.25 (Regularity). *Note that the regularity assumption $u \in H^1(\Omega)$, which is made to achieve the quasi-optimal $\frac{1}{2}$ -th order convergence rate, is stronger than the intermediate regularity required to compute the consistency error, where we only assume $u \in W^{s,p}(\Omega)$ with $sp > 1$. Recalling Remark 3.4, the situation would have been different if we had considered, instead of $\widehat{\mathbf{R}}_{\mathcal{V}}$, the classical de Rham map $\mathbf{R}_{\mathcal{V}}$ requiring $u \in W^{s,p}(\Omega)$ with $sp > 3$.*

The estimate (3.40) is obtained by bounding the three right-hand side terms which composes the consistency error $\mathbb{E}_{\mathcal{V}}(u, \mathbf{w})$ from Lemma 3.23 and using the following interpolation error estimates proved in Chapter 7.

Proposition 3.26 (Interpolation error estimates). *Let $\widehat{\mathbf{l}}_{\mathcal{V}} : L^1(\Omega) \rightarrow \mathbb{P}_0(\widetilde{\mathcal{C}})$ be defined as $\widehat{\mathbf{l}}_{\mathcal{V}} = \mathbf{L}_{\mathcal{V}} \circ \widehat{\mathbf{R}}_{\mathcal{V}}$ with the reconstruction map $\mathbf{L}_{\mathcal{V}}$ defined by (3.18) and the reduction map $\widehat{\mathbf{R}}_{\mathcal{V}}$ defined by (3.7). Then, for all $v \in H^1(\Omega)$,*

$$\sum_{v \in \mathcal{V}} \|v - \widehat{\mathbf{l}}_{\mathcal{V}}(v)\|_{L^2(\widetilde{\mathcal{C}}(v))} \lesssim h |v|_{H^1(\Omega)} \quad \text{and} \quad \sum_{e \in \mathcal{E}} \|v - \widehat{\mathbf{l}}_{\mathcal{V}}(v)\|_{L^2(\widetilde{\mathbf{f}}(e))} \lesssim h^{\frac{1}{2}} |v|_{H^1(\Omega)}. \quad (3.41)$$

To prove the estimate (3.40), we also need the following result obtained using the Lipschitz regularity of the vector field β .

Proposition 3.27 (Approximation of β_e). *Let $v \in \mathcal{V}$. Then for all $e \in \mathcal{E}_v$,*

$$\left| \int_{\widetilde{\mathbf{f}}(e)} |\beta \cdot \mathbf{n}_{\widetilde{\mathbf{f}}(e)}| - |\beta_e| \right| \lesssim |\beta_e| + L_{\beta} |\widetilde{\mathcal{C}}(v)|.$$

Proof. Let $e \in \mathcal{E}$ and consider the local dual faces $\widetilde{\mathbf{f}}_c(e) = \widetilde{\mathbf{f}}(e) \cap c$ for all $c \in \mathcal{C}_e$. Observe that each dual face $\widetilde{\mathbf{f}}(e)$ can be decomposed into two triangles $\mathbf{f}_{ef,c}$ with $\mathbf{f} \in \mathcal{F}_c \cap \mathcal{F}_e$, defined by the three points $\{\mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$; see Figure 3.4.

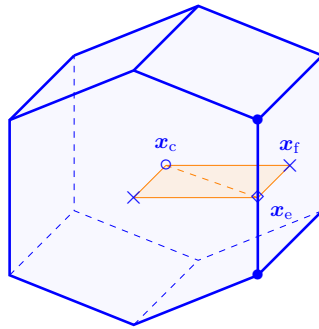


Figure 3.4 – The dual face $\widetilde{\mathbf{f}}_c(e)$ composed of two triangles.

Then, denoting

$$\delta_{\mathbf{f},c}(e) := \int_{\mathbf{f}_{ef,c}} |\beta \cdot \mathbf{n}_{ef,c}| - \int_{\mathbf{f}_{ef,c}} \beta \cdot \mathbf{n}_{ef,c},$$

with $\mathbf{n}_{ef,c}$ the unit normal attached to the elementary sub-face $\mathbf{f}_{ef,c}$, we have $0 \leq \delta_{\mathbf{f},c}(e) = 2 \int_{\mathbf{f}_{ef,c}} (\beta \cdot \mathbf{n}_{ef,c})^\ominus$. If $(\beta \cdot \mathbf{n}_{ef,c})^\ominus$ takes positive values on $\mathbf{f}_{ef,c}$, then $0 \leq \delta_{\mathbf{f},c}(e) \leq 2|\beta_e|$; otherwise, $\beta \cdot \mathbf{n}_{ef,c}$ vanishes at some point in $\mathbf{f}_{ef,c}$ so that, using the fact the Lipschitz regularity of $\beta \cdot \mathbf{n}_{ef,c}$

in $\mathbf{f}_{\text{ef},c}$, we have $0 \leq \delta_{f,c}(e) \lesssim L_\beta h_e |\mathbf{f}_{\text{ef},c}|$. As a result, since $|\mathbf{f}_{\tilde{f}(e)}(\cdot)| \geq \mathbf{f}_{\tilde{f}(e)}|\cdot|$, we sum these bounds over $c \in C_e$ and the faces $f \in F_c \cap F_e$, to obtain

$$0 \leq \int_{\tilde{f}(e)} |\boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}| - \left| \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)} \right| \leq \sum_{c \in C_e} \sum_{f \in F_c \cap F_e} \delta_{f,c}(e) \lesssim |\beta_e| + L_\beta h_e |\tilde{f}(e)|,$$

whence the result follows using mesh regularity. $\square$

Proof of Theorem 3.24. The estimate (3.40) is obtained by bounding the three right-hand side terms of Lemma 3.23 for all $\mathbf{w} \in \mathcal{V}$ such that $\|\mathbf{w}\|_{\mathcal{V};a} = 1$. A direct calculation shows that

$$\langle\langle [\mathbf{H}_{\mu-\nabla \cdot \boldsymbol{\beta}}^\nu, \mathbf{R}](u), \mathbf{w} \rangle\rangle_{\mathcal{V}} = \sum_{v \in V} \mathbf{w}_v \int_{\tilde{c}(v)} (\mu - \nabla \cdot \boldsymbol{\beta})(u - \hat{\mathbf{l}}_v(u)).$$

Hence, using the Cauchy–Schwarz inequality together with the first interpolation error estimate from Proposition 3.26, we infer that

$$|\langle\langle [\mathbf{H}_{\mu-\nabla \cdot \boldsymbol{\beta}}^\nu, \mathbf{R}](u), \mathbf{w} \rangle\rangle_{\mathcal{V}}| \leq \mathbf{C}_{\text{INT}} \left(\tau^{-\frac{1}{2}} \|\mathbf{w}\|_{\mathcal{V},2} \right) \left(h^{\frac{1}{2}} \tau^{\frac{1}{2}} \|\mu - \nabla \cdot \boldsymbol{\beta}\|_{L^\infty(\Omega)} \right) h^{\frac{1}{2}} |u|_{H^1(\Omega)}.$$

Turning to the second term, owing to assumption (i_κ) and the fact that u is single-valued on $\tilde{f}(e)$ (since $u \in H^1(\Omega)$), a direct calculation shows that

$$-\langle\langle [\mathbf{J}_\beta^\nu, \mathbf{R}](u), \text{GRAD}(\mathbf{w}) \rangle\rangle_{\mathcal{V}} = \frac{1}{2} \sum_{e \in E} [\mathbf{w}]_e \sum_{v \in V_e} (1 + \kappa_{ve}) \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)} (u - \hat{\mathbf{l}}_v(u)) \quad (3.42a)$$

$$= \sum_{e \in E} [\mathbf{w}]_e \int_{\tilde{f}(e)} \boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)} (u - \hat{\mathbf{l}}_v(u)). \quad (3.42b)$$

Hence, since $|\kappa_{ve}| \leq 1$, and using the regularity of $\boldsymbol{\beta}$ and the second interpolation error estimate from Proposition 3.26, we infer that

$$|\langle\langle [\mathbf{J}_\beta^\nu, \mathbf{R}](u), \text{GRAD}(\mathbf{w}) \rangle\rangle_{\mathcal{V}}| \lesssim \left(\sum_{e \in E} [\mathbf{w}]_e^2 \int_{\tilde{f}(e)} |\boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}| \right)^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}^{\frac{1}{2}} h^{\frac{1}{2}} |u|_{H^1(\Omega)}.$$

Next, using the triangle inequality along with Lemma 3.14 and assumption (ii_κ) , we end up with

$$\left(\sum_{e \in E} [\mathbf{w}]_e^2 \int_{\tilde{f}(e)} |\boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}| \right)^{\frac{1}{2}} \leq \mathbf{C}_\kappa^{-\frac{1}{2}} \mathbf{s}_\beta^\nu(\mathbf{w}, \mathbf{w})^{\frac{1}{2}} + \left(\sum_{e \in E} [\mathbf{w}]_e^2 \left(\int_{\tilde{f}(e)} |\boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}| - |\beta_e| \right) \right)^{\frac{1}{2}}.$$

To bound the second term on the right-hand side, we use Proposition 3.27 and the definition of the discrete norm $\|\cdot\|_{\mathcal{V},2}$ to infer that

$$\sum_{e \in E} [\mathbf{w}]_e^2 \left(\int_{\tilde{f}(e)} |\boldsymbol{\beta} \cdot \mathbf{n}_{\tilde{f}(e)}| - |\beta_e| \right) \lesssim \mathbf{s}_\beta^\nu(\mathbf{w}, \mathbf{w}) + L_\beta \|\mathbf{w}\|_{\mathcal{V},2}.$$

Then, collecting these bounds yields

$$|\langle\langle [\mathbf{J}_\beta^\nu, \mathbf{R}](u), \text{GRAD}(\mathbf{w}) \rangle\rangle_{\mathcal{V}}| \lesssim \left(\mathbf{s}_\beta^\nu(\mathbf{w}, \mathbf{w}) + \mathbf{C}_{\mathcal{V},a} \tau^{-1} \|\mathbf{w}\|_{\mathcal{V},2} \right)^{\frac{1}{2}} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}^{\frac{1}{2}} h^{\frac{1}{2}} |u|_{H^1(\Omega)}.$$

Finally, observing that

$$\langle\langle [\mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n})^\oplus}^{\nu,\partial}, \mathbf{R}](u), \mathbf{w} \rangle\rangle_{\mathcal{V}} = \sum_{v \in V^\partial} \mathbf{w}_v \int_{\tilde{f}^\partial(v)} (\boldsymbol{\beta} \cdot \mathbf{n})^\oplus (u - \hat{\mathbf{l}}_v(u)),$$

we infer that $|\langle\langle [\mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n})^\oplus}^{\nu,\partial}, \mathbf{R}](u), \mathbf{w} \rangle\rangle_{\mathcal{V}}| \lesssim \langle\langle \mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n})^\oplus}^{\nu,\partial}(\mathbf{w}), \mathbf{w} \rangle\rangle_{\mathcal{V}} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\Omega)}^{\frac{1}{2}} h^{\frac{1}{2}} |u|_{H^1(\Omega)}$. We obtain the expected result by combining these estimates. $\square$

3.3 Numerical results

In this section, we numerically investigate the performance and the reliability of the discrete problem (3.21). The contraction operator J_β^ε , defined by (3.25), is built using the *upwind* parameters corresponding to $\kappa_{ve} = \text{sign}(\ell_{\tilde{f}(e), \tilde{c}(v)} \beta_e)$. We consider two test cases. The first one aims to approximate a solution expressed as combination of sine functions under a rotating advective field and constant reaction coefficients. The second test case considers the case of a solution presenting a sharp internal layer under a constant advective field and a null reaction coefficient.

3.3.1 Computational setting

This section presents the computational setting, namely the mesh sequences, the computed quantities and the linear algebra.

Three-dimensional mesh sequences. Computations are run on three-dimensional mesh sequences of the unit cube $\Omega = [0, 1]^3$. Each mesh sequence is obtained as an uniform refinement of an initial mesh. We consider the four mesh sequences for which one element of the sequence is represented in Figure 3.1. These mesh sequences have been initially proposed by Eymard

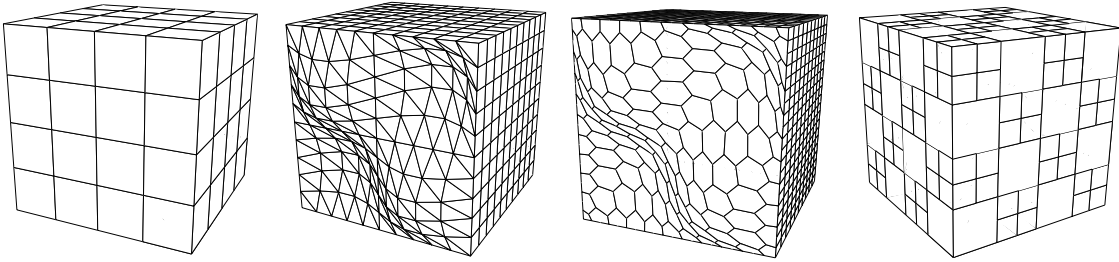


Figure 3.5 – The four mesh sequences H, PrT, PrG and CB, respectively.

et al. (2011) in the FVCA benchmark for diffusion problems. The first sequence, denoted by H, is composed of hexahedral cells, the second, denoted by PrT, is composed of prismatic cells with a triangular basis, the third, denoted by PrG, is composed of prismatic cells with an polygonal basis and the fourth, denoted by CB, is composed of checkerboard cells with non-conforming interfaces. These mesh sequences are also considered in (Bonelle, 2014, Appendix A) where mesh regularity criteria, such as cardinals, aspect ratios or non-orthogonality measurements are reported.

Remark 3.28 (Kershaw mesh sequences). *Another mesh sequence is also considered in Appendix A, corresponding to the Kershaw mesh sequence. These meshes are composed of skewed hexahedral cells. Since this sequence is not suited to study properly the convergence of our scheme, numerical results are not presented here.*

Computed quantities and linear algebra. Following Theorem 3.24, the convergence study is performed using the relative discrete L^2 -error given by:

$$\widehat{\mathbf{Er}}_{\mathcal{V}}(u) := \left(\frac{\sum_{v \in \mathcal{V}} |\tilde{c}(v)| \left(u_v - \widehat{R}_{\mathcal{V}}(u)|_v \right)^2}{\sum_{v \in \mathcal{V}} |\tilde{c}(v)| \widehat{R}_{\mathcal{V}}(u)|_v^2} \right)^{\frac{1}{2}}, \quad (3.43)$$

where u is the solution of the continuous problem (3.1) and $\mathbf{u}$ is the solution of the discrete problem (3.21). Recall that $\widehat{R}_{\mathcal{V}}(u)$ evaluates the mean value of the exact solution over dual

mesh cells (see (3.7)). To examine the cost of solving the discrete problem (3.21), we also evaluate the quantity Co defined by

$$\text{Co} = \text{nnz} \times n_{\text{ite}}, \quad (3.44)$$

where nnz denotes the total number of non-zeroes in the system matrix and n_{ite} denotes the number of iterations needed to bring the algebraic residual below a tolerance of 10^{-12} . The linear solver used in our simulation is the bi-Conjugate Gradient method preconditioned with an incomplete LU factorization from the PETSc library; see Balay *et al.* (2016). Considering two successive mesh refinements M_n and M_{n+1} , we evaluate the convergence rate of a quantity Q with respect to a quantity S by computing the quantity $-3 \log(Q_{n+1}/Q_n) / \log(S_{n+1}/S_n)$.

CDO schemes are interesting regarding their implementation since each discrete operator is computed separately from the others, and the final scheme is obtained by multiplying these operators together. For instance, to implement the discrete problem (3.21), the discrete gradient operator GRAD is computed once and for all since it only depends on the connectivity of the mesh. The contraction operator J_β^ε is also computed once and for all for steady problems (while the parameters κ_{ve} and β_e may need to be updated in unsteady problems).

3.3.2 Test case 1. Rotating advective field

The solution $u : \Omega \rightarrow \mathbb{R}$ of this first test case is defined by

$$u(x, y, z) = \sin(\pi x) \sin(2\pi y) \sin(\pi z), \quad (3.45)$$

vanishing at the boundary $\partial\Omega$ of the unit cube Ω . The advection field $\beta : \Omega \rightarrow \mathbb{R}^3$ is given by

$$\beta(x, y, z) = \begin{pmatrix} y - 1/2 \\ 1/2 - x \\ z + 1 \end{pmatrix}, \quad (3.46)$$

and the reaction coefficient μ is such that $\mu \in \{\frac{1}{2}, 5\}$. Figure 3.6 presents some flow lines of the rotating field β as well as the inflow boundary $\partial\Omega^-$. The scalar Friedrichs tensor is $\sigma_{\beta, \mu} = \mu - \frac{1}{2}$, so that the continuous problem is well-posed owing to Theorem 2.7 if $\mu > \frac{1}{2}$ and owing to Theorem 2.11 if $\mu = \frac{1}{2}$ e.g., with the potential $\zeta = (z-2)^2$. Consequently, the discrete bilinear form (3.16) is coercive if $\mu > \frac{1}{2}$ and inf-sup stable if $\mu = \frac{1}{2}$ if the mesh size is such that $h < h_0$ with h_0 defined by (3.36). However, this mesh-size restriction is always satisfied here in practice since a short calculation shows that $L_\zeta = \frac{1}{2}$ and $-\frac{1}{2}\beta \cdot \nabla \zeta = -(z-2)(z+1)$ so that $(\mathcal{H}')$ holds with $\tau = \frac{1}{2}$, yielding $h_0 = \frac{1}{4}$.

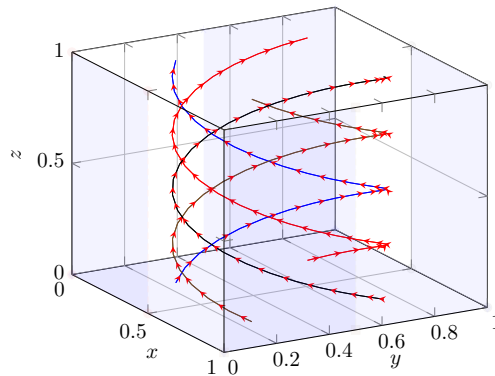


Figure 3.6 – Test case 1. Some flow lines of the advective field β and the corresponding inflow boundary $\partial\Omega^-$ in blue.

The discrete error $\widehat{\text{Er}}_V(u)$ with respect to $\#V$ is depicted in 3.7 and the corresponding legend is collected in Table 3.2.

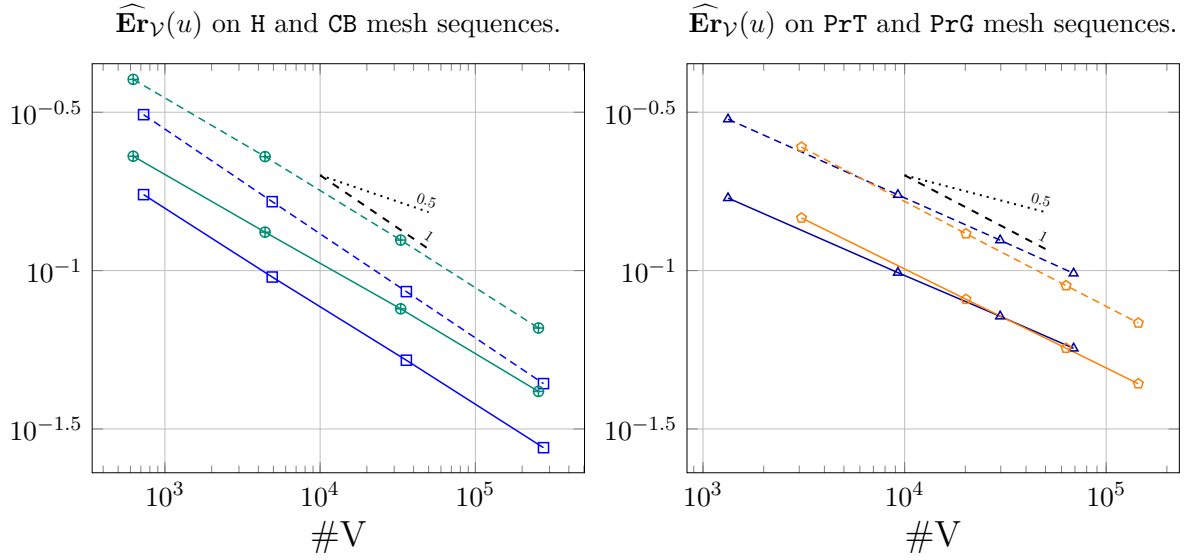


Figure 3.7 – Test case 1. Discrete relative error $\widehat{\mathbf{Er}}_{\mathcal{V}}(u)$ with $\mu = 5$ (plain lines) and $\mu = 0.5$ (dashed lines).

$\mu = 5$				$\mu = 0.5$			
H	PrT	PrG	CB	H	PrT	PrG	CB

Table 3.2 – Test case 1. Legend for Figures 3.7 and 3.8.

Accuracy. These numerical results are in agreement with the theoretical results derived in Section 3.2.2. First, we observe that the scheme converge on all mesh sequences, even when the Friedrichs tensors is uniformly equal to 0. This reflects the robustness of the proposed approach with respect to the mesh quality criteria. We also observe that the scheme is more accurate when the assumption $(\mathcal{H})$ is satisfied. This observation comes from the fact that the inf-sup stability constant ϱ from Lemma 3.19 is smaller than the coercivity constant (which is equal to $\frac{1}{2}$ from Lemma 3.18). In addition, we observe a super-convergence behavior in the discrete L^2 -norm on all mesh sequences since the convergence rates are closer to 1 than to $\frac{1}{2}$, the latter corresponding to the theoretical convergence order inferred by Theorem 3.24.

In the right panel of Figure 3.8, we compare the discrete relative error $\widehat{\mathbf{Er}}_{\mathcal{V}}(u)$ with the discrete relative error $\mathbf{Er}_{\mathcal{V}}(u)$ defined by (3.43) replacing $\widehat{\mathbf{R}}_{\mathcal{V}}$ with the de Rham map $\mathbf{R}_{\mathcal{V}}$ evaluating point-wise the exact solution at each mesh vertex. We observe that these two errors converge approximately at the same order and that $\widehat{\mathbf{Er}}_{\mathcal{V}}(u) < \mathbf{Er}_{\mathcal{V}}(u)$ on all mesh sequences. In terms of accuracy, the hierarchy is $\text{H} > \text{PrT} > \text{PrG} > \text{CB}$.

Efficiency and discret min./max principle. On the left panel of Figure 3.8, we present the computational efficiency. It is described by representing the discrete error $\widehat{\mathbf{Er}}_{\mathcal{V}}(u)$ with respect to the computational cost Co : the closer of the origin, the more efficient the computation.

We observe that the computation is more expensive for the CB mesh sequence containing non-conforming interfaces, and not surprisingly, the cheapest computation is observed for the H mesh sequence. For this test case, the efficiency hierarchy is then $\text{H} > \text{PrT} > \text{PrG} > \text{CB}$. Finally, we mention that the discrete minimum/maximum principle is satisfied on all these mesh sequences.

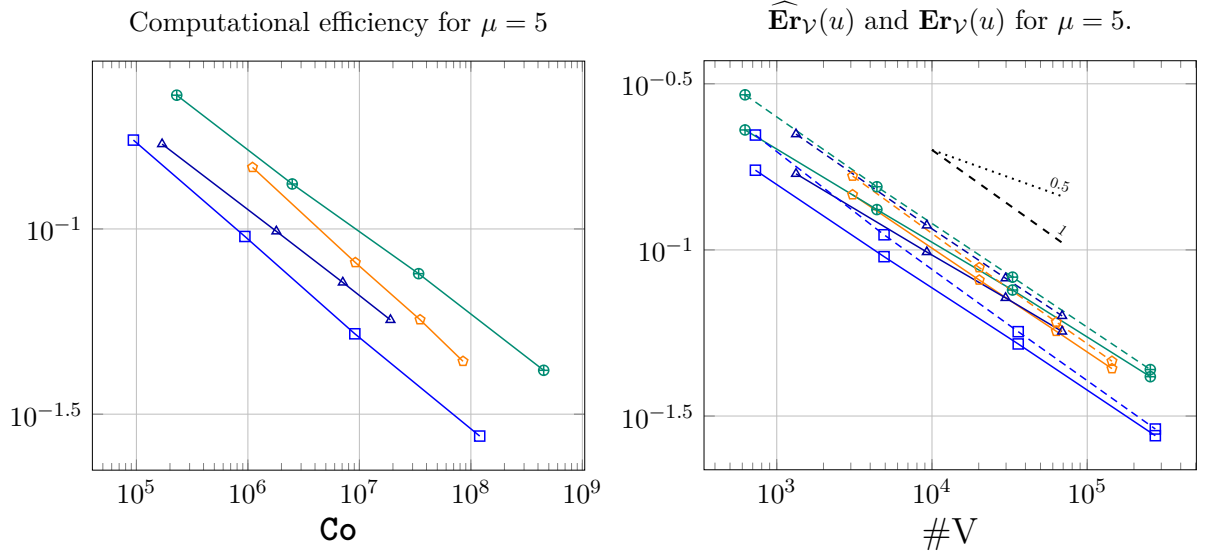


Figure 3.8 – Test case 1. On the left, the discrete relative error $\widehat{\mathbf{Er}}_V(u)$ with respect to computational cost $\mathbf{Co}$ for $\mu = 5$. On the right, comparison between the discrete relative error $\widehat{\mathbf{Er}}_V(u)$ (plain lines) with $\mathbf{Er}_V(u)$ (dashed lines).

3.3.3 Test case 2. Sharp internal layer

Adapting the test case proposed by Burman (2005), we consider the following smooth solution presenting an internal layer in the vicinity of the hyper plane $\frac{x}{2} + y + z = \sqrt{2}$:

$$u(x, y, z) = xy \tanh \left(\frac{x}{2a} + \frac{y + z - \sqrt{2}}{a} \right), \quad (3.47)$$

with $a = 0.05$. This solution is represented in Figure 3.9 for different value of z . We do not

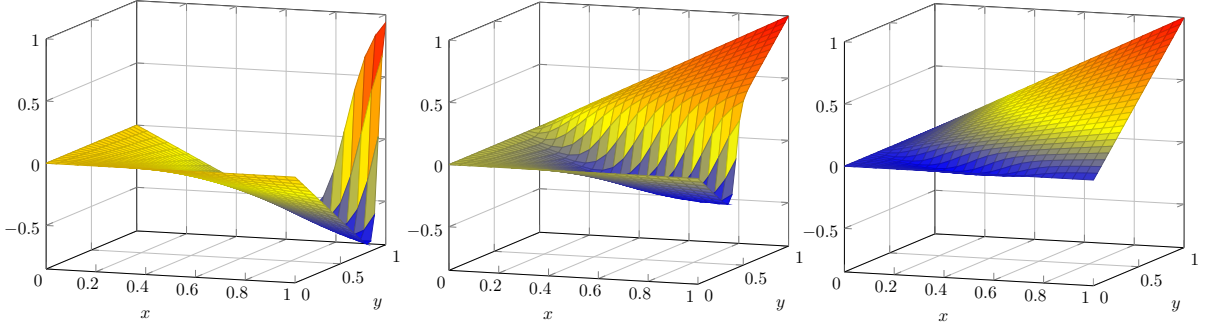


Figure 3.9 – Test case 2. From the left to the right, exact solution $(x, y) \in [0, 1]^2 \mapsto u(x, y, z_0)$ defined by (3.47) for $z_0 = 0$, $z_0 = 0.5$ and $z_0 = 1$, respectively.

consider reaction coefficient, so that $\mu = 0$, and we consider a constant advection field $\beta = (1, 1, 0)^T$. Since $\nabla \cdot \beta = 0$ a.e. in Ω , the scheme is unconditionally stable owing to Lemma 3.19.

In Figure 3.10, we present the discrete relative error $\widehat{\mathbf{Er}}_V(u)$ with respect to $\#V$ on the left panel and with respect to the computational cost $\mathbf{Co}$ on the right panel.

Accuracy, efficiency and discrete min./max. principle. Although these mesh sequences are not adapted to capture the internal layer, we observe that our scheme converges with a rate closer to 1 than to $\frac{1}{2}$. The hierarchy in terms of accuracy for this test case is $\mathbf{H} > \mathbf{CB} \geq \mathbf{PrG} > \mathbf{PrT}$, which is different from that obtained in Section 3.3.2. In terms of efficiency, the hierarchy is the same as for the first test case: $\mathbf{H} > \mathbf{PrT} > \mathbf{PrG} > \mathbf{CB}$

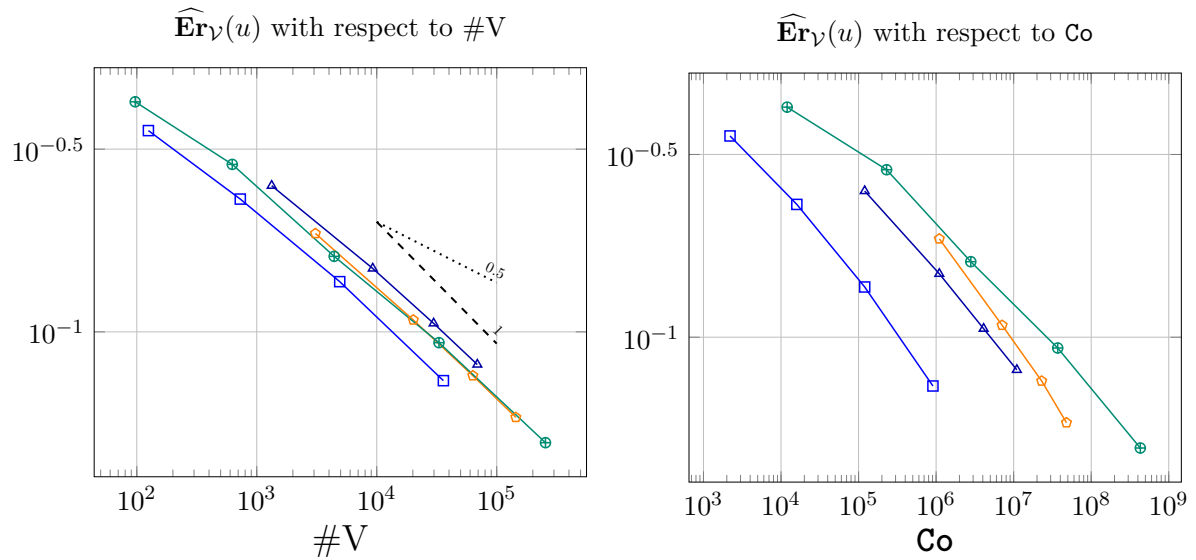


Figure 3.10 – Test case 2. Discrete relative error $\widehat{\mathbf{Er}}_V(u)$ with respect to $\#V$ (left) and with respect to the computational cost $\mathbf{Co}$ (right).

Finally, it is worthwhile to mention that the discrete minimum/maximum principle is respected for the H, PrT and PrG mesh sequences, whereas it is violated around the internal layer on the most refined mesh of the sequence CB by up to 4.3%.

Chapter 4

Péclet-robust vertex-based scheme for diffusion-advection-reaction

Contents

4.1	Discrete setting	48
4.1.1	Mesh and degrees of freedom	48
4.1.2	Edge-based partition	49
4.1.3	Reduction maps	49
4.2	Pure Diffusion problem	50
4.2.1	Discrete problem and weak boundary conditions	50
4.2.2	Analysis	52
4.3	Diffusion-advection-reaction problem	56
4.3.1	Péclet-based upwind scheme	57
4.3.2	Analysis	57
4.4	Numerical results	62
4.4.1	Test case 3. Anisotropic diffusion tensor and rotating advective field	63
4.4.2	Test case 4. Boundary layer on graded meshes	64

In this chapter, we extend the scheme presented in Chapter 3 so as to devise a vertex-based scheme approximating the scalar-valued solution of the following diffusion-advection-reaction problem:

$$-\nabla \cdot (\lambda \nabla u) + \beta \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (4.1a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega. \quad (4.1b)$$

This chapter contains three main contributions concerning CDO schemes. The first one is to devise a CDO scheme for the pure diffusion problem with weakly enforced boundary conditions. Indeed, as for stabilized finite elements, weak enforcement of Dirichlet conditions yields better results for under-resolved outflow layers (see e.g., Bazilevs & Hughes (2007); Schieweck (2008) or Burman (2012)). To this purpose, we extend Nitsche's boundary penalty method to the CDO framework. The second contribution is to propose a vertex-based scheme for (4.1) which is robust with respect to the Péclet number by using a Péclet-dependent stabilization technique so as to obtain a final *a priori* estimate on the discrete error depending on the local Péclet number. The third contribution of this chapter is to establish the well-posedness of the discrete problem when the Friedrichs' tensor $\sigma_{\beta,\mu}$ takes null values in the presence of diffusion.

Our scheme defines an extension on polyhedral meshes of the classical finite element/finite volume method combining a finite element treatment of the diffusive term and a finite volume treatment of the advective term. This approach was initially proposed by Baba & Tabata (1981) on triangular meshes using dual cells around vertices as control volumes and by Ohmori

& Ushijima (1984) using diamond cells around mesh edges. More recently, Angot *et al.* (1998) and Eymard & Hilhorst (2010) extended this approach to treat non-linear problems and Hilhorst & Vohralík (2011) performed an *a posteriori* error analysis.

The present scheme seems to be the first polyhedral discretization that is only vertex-based and that is robust for dominant advection up to the limit of zero diffusion. Recently, Beirão da Veiga *et al.* (2013) applied the Virtual Element Method (VEM) to handle vertex-based schemes on polyhedral meshes for the pure diffusion problems. These results were extended in Beirão da Veiga *et al.* (2016a,b) to the full problem (4.1) for diffusion dominant regime. Here the approach is different since we use explicit reconstruction maps and our treatment includes dominant advection regimes. However, our approach is only of lowest-order. An alternative to vertex-based schemes is the face-based scheme, proposed by Di Pietro *et al.* (2015), which is of arbitrary order and Péclet-robust. This latter work can be seen as an extension of the lowest-order HMM scheme proposed by Beirão da Veiga *et al.* (2011) for the diffusion-advection problem in the diffusive regime.

This chapter is organized as follows. First, we recall from Chapter 3 the main concepts and notation used to devise CDO schemes. Next, we extend and analyze the diffusion scheme proposed by Bonelle & Ern (2014a) by weakly enforcing the boundary conditions using the boundary penalty method of Nitsche (1971). Then, in Section 4.3, we derive and analyze a Péclet-robust discrete problem approximating the solution of (4.1). Finally, Section 4.4 collects the numerical results of two three-dimensional test cases on polyhedral meshes.

The content of this chapter is an extended version of some parts of the paper "*Vertex-based compatible Discrete Operator Schemes on Polyhedral meshes for Advection-Diffusion Equations*" by P. Cantin & A. Ern, published in Computational Methods in Applied Mathematics, 2016.

4.1 Discrete setting

This section recalls and completes the discrete setting presented in Section 3.1 in the context of CDO schemes for the scalar advection-reaction problem. A broader presentation can be found in Chapter 7.

4.1.1 Mesh and degrees of freedom

Recalling that Ω denotes an open, bounded, connected polyhedral subset of $\mathbb{R}^3$, we consider $M := \{V, E, F, C\}$ a polyhedral mesh, collecting vertices $v \in V$, edges $e \in E$, faces $f \in F$ and cells $c \in C$. This mesh is assumed to be regular in the sense that it satisfies assumptions **(C)**, **(St)** and **(Sh)**, i.e., it defines a cellular complex, it is star-shaped and it is shape-regular.

In addition to the primal mesh M , the dual mesh $\tilde{M}$ is also considered. The dual mesh collects the dual faces $\tilde{f}(e) \in \tilde{F}$ for all $e \in E$, that are in general non planar, and the dual cells $\tilde{c}(v) \in \tilde{C}$ for all $v \in V$. Since boundary conditions are weakly enforced in our scheme, we also introduce boundary dual faces, covering the dual cells on the boundary $\partial\Omega$. These boundary dual faces are defined as $\tilde{f}^\partial(v) = \partial\tilde{c}(v) \cap \partial\Omega$ for all boundary vertices $v \in V^\partial$, and are possibly broken. Primal, dual and boundary meshes are depicted in Figure 4.1.

Although our scheme is vertex-based, we also consider dofs attached to primal mesh edges. The finite-dimensional dof spaces $\mathcal{V}$ and $\mathcal{E}$ collect dofs attached to V and E , respectively. Since our scheme is built cell-wise, we also consider local definitions. For all $c \in C$, $\mathcal{V}_c$ and $\mathcal{E}_c$ denote the sub-spaces of $\mathcal{V}$ and $\mathcal{E}$ collecting dofs attached to V_c and E_c , respectively. In the context of CDO schemes, we also recall the inner products on these dof spaces:

$$\langle\langle v, w \rangle\rangle_{\mathcal{V}} = \sum_{v \in V} v_v w_v, \quad \forall v, w \in \mathcal{V} \text{ and } \forall c \in C, \quad \langle\langle v, w \rangle\rangle_{\mathcal{V}_c} = \sum_{v \in V_c} v_v w_v, \quad \forall v, w \in \mathcal{V}_c. \quad (4.2)$$

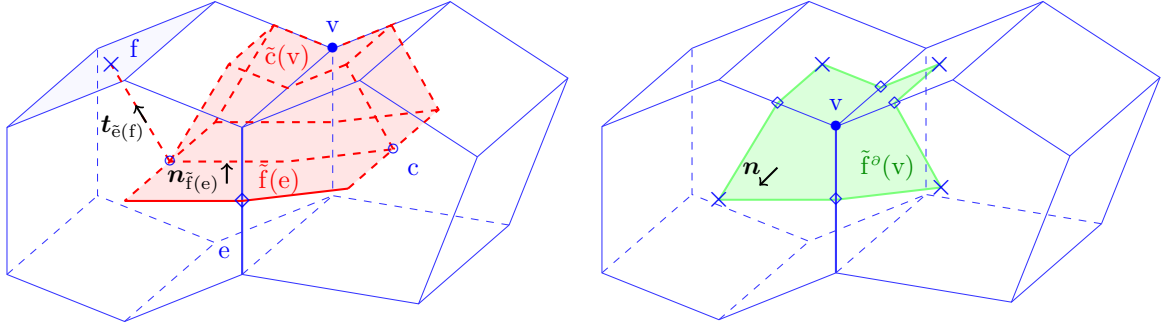


Figure 4.1 – Left panel: the primal mesh M and the dual mesh $\tilde{M}$. Right panel: the boundary dual mesh.

$$\langle\langle v, w \rangle\rangle_{\mathcal{E}} = \sum_{e \in E} v_e w_e, \quad \forall v, w \in \mathcal{E} \quad \text{and} \quad \forall c \in C, \quad \langle\langle v, w \rangle\rangle_{\mathcal{E}_c} = \sum_{e \in E_c} v_e w_e, \quad \forall v, w \in \mathcal{E}_c. \quad (4.3)$$

Local dual faces in c are denoted by $\tilde{f}_c(e)$ and are defined as $\tilde{f}_c(e) = \tilde{f}(e) \cap c$ for all $e \in E_c$, with unit normal vector $\mathbf{n}_{\tilde{f}_c(e)}$ oriented by $\mathbf{t}_e$. For any boundary face $f \in F^\partial$, $c(f)$ denotes the unique cell containing f .

4.1.2 Edge-based partition

In addition to the primal and the dual mesh, our scheme considers the partition composed of local diamonds around mesh edges. For all $c \in C$, $\mathfrak{C}_{e,c}$ denotes the partition of c composed of the sub-cells (also called diamonds) $\{\mathfrak{c}_{e,c}\}_{e \in E_c}$ such that

$$\mathfrak{c}_{e,c} = \text{int} \left(\bigcup_{f \in F_e \cap F_c} \bigcup_{v \in V_e} \overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\} \right), \quad \forall e \in E_c, \quad (4.4)$$

where we have denoted by $\text{int}(\omega)$ the interior of any set $\omega \subset \mathbb{R}^3$, $\mathbf{x}_x$ the barycenter of any mesh entity $x \in M$ and $\overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$ the closed convex hull of the 4-uplet $\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$. As observed in Figure 4.2, diamonds $\{\mathfrak{c}_{e,c}\}_{e \in E_c}$ are obtained collecting the 4 simplices attached to each edge $e \in E_c$. Owing to the star-shaped assumption **(Sh)**, we observe that $\bar{c} = \cup \{\bar{c}_{e,c} \mid e \in E_c\}$.

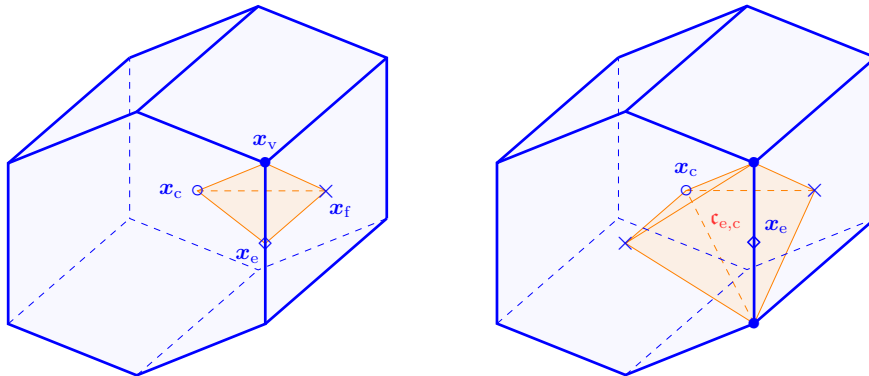


Figure 4.2 – The sub-cell $\mathfrak{c}_{e,c}$ attached to $e \in E_c$ (right panel) containing the simplex depicted in left panel.

4.1.3 Reduction maps

Reduction maps are used to reduce on mesh entities continuous fields, such as the data or the exact solution. In addition to the reduction maps $\widehat{R}_V : L^1(\Omega) \rightarrow \mathcal{V}$, $\widetilde{R}_V : L^1(\Omega) \rightarrow \mathcal{V}$, $\widetilde{R}_E : \mathbf{W}^{s,p}(\Omega) \rightarrow \mathcal{E}$ with $sp > 1$ and $\widetilde{R}_V^0 : L^1(\partial\Omega) \rightarrow \mathcal{V}$ defined by (3.7), (3.8a), (3.8b) and (3.8c), respectively, we consider in this Chapter the classical de Rham maps on primal mesh entities. The de Rham reduction map on mesh vertices $R_V : W^{s,p}(\Omega) \rightarrow \mathcal{V}$ with $sp > 3$ is defined by

$$\forall v \in V, \quad R_V(w)|_v = w(\mathbf{x}_v), \quad \forall w \in W^{s,p}(\Omega), \quad (4.5a)$$

and the de Rham reduction map on mesh edges $R_E : \mathbf{W}^{s,p}(\Omega) \rightarrow \mathcal{E}$ with $sp > 2$ is defined by

$$\forall e \in E, \quad R_E(\mathbf{w})|_e = \int_e \mathbf{w} \cdot \mathbf{t}_e, \quad \forall \mathbf{w} \in \mathbf{W}^{s,p}(\Omega). \quad (4.5b)$$

Considering these two reduction maps, Proposition 4.1 states that the continuous gradient ∇ exactly commutes with the discrete gradient operator $\text{GRAD} : \mathcal{V} \rightarrow \mathcal{E}$ defined by (3.5a) as follows:

$$\forall e \in E, \quad \text{GRAD}(\mathbf{w})|_e := \sum_{v \in V_e} \iota_{v,e} \mathbf{w}_v, \quad \forall \mathbf{w} \in \mathcal{V}. \quad (4.6)$$

Proposition 4.1 (Primal commutation). *For all $w \in W^{s,p}(\Omega)$ with $sp > 3$ such that $\nabla w \in \mathbf{W}^{s-\frac{1}{p},p}(\Omega)$, we have*

$$R_E(\nabla w) = \text{GRAD } R_V(w). \quad (4.7)$$

Proof. This identity follows from the fundamental theorem of calculus. $\square$

For all $c \in C$, the local de Rham reduction maps are denoted by $R_{V_c} : W^{s,p}(c) \rightarrow \mathcal{V}_c$ and $R_{E_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathcal{E}_c$ with $sp > 3$ and $sp > 2$ respectively, and act as their global counterpart, namely as (4.5a) and (4.5b), respectively.

4.2 Pure Diffusion problem

This section extends the CDO scheme proposed by Bonelle & Ern (2014a) to approximate the solution of the pure diffusion problem

$$-\nabla \cdot (\boldsymbol{\lambda} \nabla u) = s \quad \text{a.e. in } \Omega, \quad (4.8a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega, \quad (4.8b)$$

so as to weakly enforce the Dirichlet boundary condition (4.8b). Weak enforcement of boundary conditions is essential to couple properly the treatment of the diffusion term with the treatment of the advection-reaction terms whenever the exact solution exhibits outflow layers. Recall also that boundary conditions are weakly enforced (on the inflow boundary $\partial\Omega^-$) in CDO schemes for advection-reaction problems.

Assume that the data satisfy $s \in L^2(\Omega)$ and $u_D \in H^{1+\epsilon}(\partial\Omega)$ with $\epsilon > 0$, and consider a diffusion tensor $\boldsymbol{\lambda} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ taking symmetric and uniformly positive definite values. Then, the continuous problem (4.8) is well-posed in $H^1(\Omega)$. For simplicity, we assume that $\boldsymbol{\lambda}$ is piecewise constant in each mesh cell $c \in C$, i.e., $\boldsymbol{\lambda} \in \mathbb{P}_0(C; \mathbb{R}^{3 \times 3})$, and we denote $\lambda_{b,c}$ and $\lambda_{\sharp,c}$ its minimal and maximal eigenvalue in c , respectively. The local anisotropy ratio is denoted by $\rho_c = \lambda_{\sharp,c}/\lambda_{b,c}$. Note that the present analysis can also be extended to locally Lipschitz diffusion tensors.

4.2.1 Discrete problem and weak boundary conditions

Vertex-based CDO schemes approximating the elliptic problem (4.8) are formulated in terms of a discrete Hodge operator $\mathbf{H}_\lambda^\varepsilon : \mathcal{E} \rightarrow \mathcal{E}$, which is the discrete counterpart of the map $\mathbf{w} \mapsto \lambda \cdot \mathbf{w}$. Since boundary conditions are weakly enforced, our CDO scheme is also expressed using two discrete boundary operators. The first one is denoted by $\mathbf{N}_\lambda^{\varepsilon, \partial} : \mathcal{E} \rightarrow \mathcal{V}$ and is the discrete counterpart of the normal flux map $\mathbf{w} \mapsto \mathbf{n} \cdot \lambda \cdot \mathbf{w}$ at the boundary $\partial\Omega$. This map is essential to preserve the consistency of the scheme when the boundary conditions are weakly enforced. The second operator enforces the boundary conditions à la Nitsche on $\partial\Omega$ and is denoted by $\mathbf{H}_\nu^{\nu, \partial} : \mathcal{V} \rightarrow \mathcal{V}$ where $\nu \in \mathbb{P}_0(\mathbf{F}^\partial; \mathbb{R})$ denotes the Nitsche penalty parameter such that $\nu|_f = \lambda_{\sharp; c(f)} / h_{c(f)}$ for all $f \in \mathbf{F}^\partial$. This operator is the discrete counterpart of the map $w \mapsto \nu w$ at the boundary. As for the discrete problem (3.21), boundary operator maps to $\mathcal{V}$ to simplify the notation and the entries of $\mathbf{N}_\lambda^{\varepsilon, \partial}$ and $\mathbf{H}_\nu^{\nu, \partial}$ attached to internal vertices $v \in \mathbf{V}^\circ$ are always 0.

Bilinear forms and reconstruction maps. The CDO scheme approximating the continuous problem (4.8) is formulated with the bilinear map $\mathbf{A}_\lambda^\nu : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ such that

$$\mathbf{A}_\lambda^\nu(v, w) := \langle \mathbf{H}_\lambda^\varepsilon \text{GRAD}(v), \text{GRAD}(w) \rangle_\mathcal{E} - \langle \mathbf{N}_\lambda^{\varepsilon, \partial} \text{GRAD}(v), w \rangle_\mathcal{V} + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial}(v), w \rangle_\mathcal{V}, \quad (4.9)$$

for all $v, w \in \mathcal{V}$. The first term approximates (4.8a) and the two last terms weakly enforce the boundary condition (4.8b) with the user-dependent real number $\nu_0 > 0$ to be chosen large enough (see Lemmata 4.7, 4.17 and 4.18). The map $\mathbf{A}_\lambda^\nu$ extends that of Bonelle & Ern (2014a) where the Dirichlet boundary condition was strongly enforced. The operator $\mathbf{H}_\lambda^\varepsilon : \mathcal{E} \rightarrow \mathcal{E}$ is built cell-wise as follows:

$$\langle \mathbf{H}_\lambda^\varepsilon(v), w \rangle_\mathcal{E} = \sum_{c \in \mathbf{C}} \langle \mathbf{H}_{\lambda; c}^\varepsilon(v), w \rangle_{\mathcal{E}_c}, \quad \forall v, w \in \mathcal{E}, \quad (4.10)$$

with $\mathbf{H}_{\lambda; c}^\varepsilon : \mathcal{E}_c \rightarrow \mathcal{E}_c$. Strictly speaking, we should consider the full-rank transfer map $\mathbf{T}_{\mathcal{E}; c} : \mathcal{E} \rightarrow \mathcal{E}_c$ from global to local dofs, so that the global operator $\mathbf{H}_\lambda^\varepsilon$ results from the cell-wise assembly $\mathbf{H}_\lambda^\varepsilon = \sum_{c \in \mathbf{C}} \mathbf{T}_{\mathcal{E}; c}^\top \cdot \mathbf{H}_{\lambda; c}^\varepsilon \cdot \mathbf{T}_{\mathcal{E}; c}$. However, to alleviate the notation, these maps are omitted. The local operator $\mathbf{H}_{\lambda; c}^\varepsilon$ is defined as

$$\langle \mathbf{H}_{\lambda; c}^\varepsilon(v), w \rangle_{\mathcal{E}_c} = \int_c \mathbf{L}_{\mathcal{E}_c}(v) \cdot \lambda \cdot \mathbf{L}_{\mathcal{E}_c}(w), \quad \forall v, w \in \mathcal{E}_c. \quad (4.11)$$

where we have introduced the reconstruction map $\mathbf{L}_{\mathcal{E}_c} : \mathcal{E}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{\mathcal{E}; c}; \mathbb{R}^d)$ on the diamond partition $\mathfrak{C}_{\mathcal{E}; c} := \{\mathfrak{c}_{e, c}\}_{e \in \mathbf{E}_c}$ introduced in Section 4.1.2, and acting as follows:

$$\mathbf{L}_{\mathcal{E}_c}(w)(x) := \sum_{e \in \mathbf{E}_c} w_e \ell_{e, c}(x), \quad \forall w \in \mathcal{E}_c. \quad (4.12)$$

The local shape functions $\{\ell_{e, c}\}_{e \in \mathbf{E}_c}$, spanning $\mathbb{P}_0(\mathfrak{C}_{\mathcal{E}; c}; \mathbb{R}^d)$, are such that

$$\forall e', e \in \mathbf{E}_c, \quad \ell_{e, c}|_{\mathfrak{c}_{e', c}} = \left(\text{Id} - \frac{\tilde{\mathbf{f}}_c(e') \otimes \mathbf{e}'}{d|\mathfrak{c}_{e', c}|} \right) \frac{\tilde{\mathbf{f}}_c(e)}{|c|} + \frac{\tilde{\mathbf{f}}_c(e)}{d|\mathfrak{c}_{e, c}|} \delta_{e, e'}, \quad (4.13)$$

where we have denoted $d = 3$, $\mathbf{e} := \int_e \mathbf{t}_e$ and $\tilde{\mathbf{f}}_c(e) := \int_{\tilde{\mathfrak{f}}_c(e)} \mathbf{n}_{\tilde{\mathfrak{f}}_c(e)}$. This reconstruction map was first considered by Codecasa & Trevisan (2007) in the context of the Discrete Geometric Approach (DGA).

Remark 4.2 (Regularity along the edges). *As mentioned in Bonelle (2014), the tangential component of the reconstructed function $\mathbf{L}_{\mathcal{E}_c}(w)$ along the edges contained in the interior of the cell c and induced by the partition $\mathfrak{C}_{\mathcal{E}; c}$, is not continuous, so that $\mathbf{L}_{\mathcal{E}_c}(w) \notin \mathbf{H}(\text{curl}; c)$ in general.*

The boundary condition (4.8b) is weakly enforced by means of the two last terms in (4.9). The maps $\mathbf{N}_{\lambda}^{\varepsilon, \partial} : \mathcal{E} \rightarrow \mathcal{V}$ and $\mathbf{H}_{\nu}^{\nu, \partial} : \mathcal{V} \rightarrow \mathcal{V}$ are built face-wise at the boundary as follows:

$$\langle\langle \mathbf{N}_{\lambda}^{\varepsilon, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V}} := \sum_{f \in F^{\partial}} \langle\langle \mathbf{N}_{\lambda;f}^{\varepsilon, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V};c(f)}, \quad \forall \mathbf{v} \in \mathcal{E}, \quad \forall \mathbf{w} \in \mathcal{V}, \quad (4.14)$$

and

$$\langle\langle \mathbf{H}_{\nu}^{\nu, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V}} := \sum_{f \in F^{\partial}} \langle\langle \mathbf{H}_{\nu;f}^{\nu, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V};c(f)}, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}, \quad (4.15)$$

where $c \equiv c(f)$ is the unique cell containing the face $f \in F^{\partial}$. The local maps $\mathbf{N}_{\lambda;f}^{\varepsilon, \partial} : \mathcal{E}_c \rightarrow \mathcal{V}_c$ and $\mathbf{H}_{\nu;f}^{\nu, \partial} : \mathcal{V}_c \rightarrow \mathcal{V}_c$ are defined as

$$\langle\langle \mathbf{N}_{\lambda;f}^{\varepsilon, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V}_c} = \int_f \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \boldsymbol{\lambda} \cdot \mathbf{n} \mathbf{L}_{\mathcal{V}_c}(\mathbf{w}), \quad \forall \mathbf{v} \in \mathcal{E}_c, \quad \forall \mathbf{w} \in \mathcal{V}_c, \quad (4.16)$$

and

$$\langle\langle \mathbf{H}_{\nu;f}^{\nu, \partial}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{V}_c} = \int_f \nu \mathbf{L}_{\mathcal{V}_c}(\mathbf{v}) \mathbf{L}_{\mathcal{V}_c}(\mathbf{w}) \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{V}_c, \quad (4.17)$$

where $\mathbf{L}_{\mathcal{V}_c} : \mathcal{V}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{\mathcal{V}_c}; \mathbb{R})$ is the restriction to c of the reconstruction map $\mathbf{L}_{\mathcal{V}}$ defined by (3.18) and such that

$$\forall \mathbf{v} \in \mathcal{V}_c, \quad \mathbf{L}_{\mathcal{V}_c}(\mathbf{w})|_{\tilde{f}^{\partial}(\mathbf{v}) \cap f} = \mathbf{w}_{\mathbf{v}}, \quad \forall \mathbf{w} \in \mathcal{V}_c. \quad (4.18)$$

Observe that the definition (4.18) leads to $\mathbf{N}_{\lambda;f}^{\varepsilon, \partial}(\cdot)|_{\mathbf{v}} = 0$ and $\mathbf{H}_{\nu;f}^{\nu, \partial}(\cdot)|_{\mathbf{v}} = 0$ for all $\mathbf{v} \notin V_f$.

Discrete problem. The approximation of the continuous problem (4.8) with weakly enforced Dirichlet boundary condition reads:

$$\text{Find } \mathbf{u} \in \mathcal{V} \text{ s.t. } \mathbf{A}_{\lambda}^{\nu}(\mathbf{u}, \mathbf{v}) = \mathbb{S}(s, u_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{V}, \quad (4.19)$$

with the linear source form $\mathbb{S}(s, u_D; \cdot) : \mathcal{V} \rightarrow \mathbb{R}$ defined, for all $\mathbf{v} \in \mathcal{V}$, by

$$\mathbb{S}(s, u_D; \mathbf{v}) = \langle\langle \tilde{\mathbf{R}}_{\mathcal{V}}(s), \mathbf{v} \rangle\rangle_{\mathcal{V}} + \nu_0 \langle\langle \mathbf{H}_{\nu}^{\nu, \partial} \mathbf{R}_{\mathcal{V}}^{\partial}(u_D), \mathbf{v} \rangle\rangle_{\mathcal{V}}, \quad (4.20)$$

where $\mathbf{R}_{\mathcal{V}}^{\partial}(\cdot)|_{\mathbf{v}} = \mathbf{R}_{\mathcal{V}}(\cdot)|_{\mathbf{v}}$ for all $\mathbf{v} \in V^{\partial}$ and $\mathbf{R}_{\mathcal{V}}^{\partial}(\cdot)|_{\mathbf{v}} = 0$ for all $\mathbf{v} \in V \setminus V^{\partial}$.

Remark 4.3 (Symmetric boundary penalty term). *It is possible to consider a symmetric bilinear form $\mathbf{A}_{\lambda}^{\nu}$ by adding the term $-\langle\langle \mathbf{N}_{\lambda}^{\varepsilon, \partial} \text{GRAD}(\mathbf{w}), \mathbf{v} \rangle\rangle_{\mathcal{V}}$. The consistency of the scheme is then preserved by adding the term $-\langle\langle \mathbf{N}_{\lambda}^{\varepsilon, \partial} \text{GRAD}(\mathbf{w}), \mathbf{R}_{\mathcal{V}}(u_D) \rangle\rangle_{\mathcal{V}}$ to the source term. Symmetry is an important property when invoking duality arguments for pure diffusion problem, e.g., to apply the Aubin-Nitsche Lemma (see the Theorem 4.1 of Bonelle & Ern (2014a)). It is also a relevant property when inverting the system matrix. However, it is less important in the presence of advection.*

Remark 4.4 (Regularity of the boundary condition). *For simplicity, we have assumed that the Dirichlet condition satisfies $u_D \in H^{1+\epsilon}(\partial\Omega)$ with $\epsilon > 0$ so as to be able to discretize it by the reduction map $\mathbf{R}_{\mathcal{V}}$. Discontinuous boundary conditions in $H^{\frac{1}{2}}(\partial\Omega)$ could be considered as well by using the reduction map $\hat{\mathbf{R}}_{\mathcal{V}}^{\partial} : L^1(\partial\Omega) \rightarrow \mathcal{V}$ defined by*

$$\forall \mathbf{v} \in V^{\partial}, \quad \hat{\mathbf{R}}_{\mathcal{V}}^{\partial}(w)|_{\mathbf{v}} := \frac{1}{|\tilde{f}^{\partial}(\mathbf{v})|} \int_{\tilde{f}^{\partial}(\mathbf{v})} w, \quad \forall w \in L^1(\partial\Omega),$$

and with $\hat{\mathbf{R}}_{\mathcal{V}}^{\partial}(\cdot)|_{\mathbf{v}} = 0$ for all $\mathbf{v} \in V \setminus V^{\partial}$.

4.2.2 Analysis

This section is devoted to the analysis of the discrete problem (4.19). First, we introduce the discrete 2-norm on the dof space $\mathcal{E}$ as

$$\|\mathbf{w}\|_{\mathcal{E},2}^2 := \sum_{c \in \mathcal{C}} \|\mathbf{w}\|_{\mathcal{E}_c,2}^2 \quad \text{with} \quad \|\mathbf{w}\|_{\mathcal{E}_c,2}^2 = \sum_{e \in \mathcal{E}_c} |\mathbf{c}_{e,c}| \left(\frac{|\mathbf{w}_e|}{|e|} \right)^2, \quad \forall \mathbf{w} \in \mathcal{E}. \quad (4.21)$$

Stability. The stability of the approximation (4.19) is analyzed using the following norms on defined on $\mathcal{E}$ and $\mathcal{V}$:

$$\|\mathbf{w}\|_{\mathcal{E},\lambda}^2 := \sum_{c \in \mathcal{C}} \|\mathbf{w}\|_{\mathcal{E}_c,\lambda}^2 \quad \text{with} \quad \|\mathbf{w}\|_{\mathcal{E}_c,\lambda}^2 := \langle \mathbf{H}_{\lambda;c}^\varepsilon(\mathbf{w}), \mathbf{w} \rangle_{\mathcal{E}_c}, \quad \forall \mathbf{w} \in \mathcal{E}, \quad (4.22)$$

$$\|\mathbf{w}\|_{\mathcal{V},\nu}^2 := \sum_{f \in \mathcal{F}^\partial} \|\mathbf{w}\|_{\mathcal{V}_f,\nu}^2 \quad \text{with} \quad \|\mathbf{w}\|_{\mathcal{V}_f,\nu}^2 := \langle \mathbf{H}_{\nu;f}^{\nu,\partial}(\mathbf{w}), \mathbf{w} \rangle_{\mathcal{E}_{c(f)}}, \quad \forall \mathbf{w} \in \mathcal{V}. \quad (4.23)$$

The first discrete norm (4.22) corresponds to the classical *energy* norm while the second norm (4.23) is related to the treatment à la Nitsche of the boundary condition. The diffusion related stability norm is then

$$\|\mathbf{w}\|_{\mathcal{V},d}^2 := \|\text{GRAD}(\mathbf{w})\|_{\mathcal{E},\lambda}^2 + \|\mathbf{w}\|_{\mathcal{V},\nu}^2, \quad \forall \mathbf{w} \in \mathcal{V}. \quad (4.24)$$

The two important properties in the analysis is the stability of the reconstruction map $\mathbf{L}_{\mathcal{E}_c}$ and the continuity of the normal flux operator $\mathbf{N}_{\lambda}^{\varepsilon,\partial} : \mathcal{E} \rightarrow \mathcal{V}$.

Proposition 4.5 (Stability of $\mathbf{L}_{\mathcal{E}_c}$). *There exists $\mathcal{C}_\# > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|\mathbf{w}\|_{\mathcal{E}_c,2} \leq \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{w})\|_{L^2(c)} \leq \mathcal{C}_\# \|\mathbf{w}\|_{\mathcal{E}_c,2}, \quad \forall \mathbf{w} \in \mathcal{E}_c.$$

Proof. See the proof of Proposition 7.39. □

Proposition 4.6 (Continuity of $\mathbf{N}_{\lambda}^{\varepsilon,\partial}$). *There is $\mathcal{C}_N > 0$ such that*

$$\langle \mathbf{N}_{\lambda}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{V}} \leq \mathcal{C}_N \|\mathbf{v}\|_{\mathcal{E},\lambda} \|\mathbf{w}\|_{\mathcal{V},\nu}, \quad \forall \mathbf{v} \in \mathcal{E}, \quad \forall \mathbf{w} \in \mathcal{V}.$$

Proof. Let $\mathbf{v} \in \mathcal{E}$ and let $f \in \mathcal{F}^\partial$ with $c = c(f)$. Applying the Cauchy–Schwarz inequality, we observe that

$$\begin{aligned} \sum_{v \in \mathcal{V}_f} |\tilde{f}^\partial(v) \cap f|^{-1} \left(\mathbf{N}_{\lambda;f}^{\varepsilon,\partial}(\mathbf{v})|_v \right)^2 &= \sum_{v \in \mathcal{V}_f} |\tilde{f}^\partial(v) \cap f|^{-1} \left(\int_{\tilde{f}^\partial(v) \cap f} \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \boldsymbol{\lambda} \cdot \mathbf{n} \right)^2 \\ &\leq \sum_{v \in \mathcal{V}_f} \lambda_{\#;c} \|\boldsymbol{\lambda}^{1/2} \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(\tilde{f}^\partial(v) \cap f)}^2 = \lambda_{\#;c} \|\boldsymbol{\lambda}^{1/2} \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(f)}^2. \end{aligned}$$

Then, using a multiplicative trace inequality (see Lemma 7.25) (since $\boldsymbol{\lambda}$ is constant and $\mathbf{L}_{\mathcal{E}_c}(\mathbf{w})$ is piece-wise constant) and definition (4.11) of the operator $\mathbf{H}_{\lambda;c}^\varepsilon$, we obtain

$$\sum_{v \in \mathcal{V}_f} |\tilde{f}^\partial(v) \cap f|^{-1} \left(\mathbf{N}_{\lambda;f}^{\varepsilon,\partial}(\mathbf{w})|_v \right)^2 \leq \mathcal{C}_T^2 h_c^{-1} \lambda_{\#;c} \|\boldsymbol{\lambda}^{1/2} \mathbf{L}_{\mathcal{E}_c}(\mathbf{w})\|_{L^2(c(f))}^2 = \mathcal{C}_T^2 h_c^{-1} \lambda_{\#;c} \|\mathbf{w}\|_{\mathcal{E}_c,\lambda}^2.$$

Assembling face-wise these estimates and recalling definition (4.23) of the norm $\|\cdot\|_{\mathcal{V}_f,\nu}$ leads to

$$\langle \mathbf{N}_{\lambda}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{V}} = \sum_{f \in \mathcal{F}^\partial} \langle \mathbf{N}_{\lambda;f}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{V}_{c(f)}} = \sum_{f \in \mathcal{F}^\partial} \sum_{v \in \mathcal{V}_f^\partial} \mathbf{N}_{\lambda;f}^{\varepsilon,\partial}(\mathbf{v})|_v \mathbf{w}|_v \leq \mathcal{C}_T \sum_{f \in \mathcal{F}^\partial} \|\mathbf{v}\|_{\mathcal{E}_{c(f)},\lambda} \|\mathbf{w}\|_{\mathcal{V}_f,\nu},$$

whence the result with $\mathcal{C}_N = \mathcal{C}_T \max_{c \in \mathcal{C}} \#(\mathcal{F}^\partial \cap \mathcal{F}_c)$, that is uniformly bounded owing to mesh regularity. □

Lemma 4.7 (Coercivity). *Assume that $\nu_0 \geq 1 + \frac{1}{2}\mathcal{C}_N^2$. Then,*

$$\mathbb{A}_\lambda^\nu(\mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V},d}^2, \quad \forall \mathbf{v} \in \mathcal{V}.$$

Consequently, the discrete problem (4.19) is well-posed.

Proof. Applying Proposition 4.6, we infer that

$$\mathbb{A}_\lambda^\nu(\mathbf{v}, \mathbf{v}) \geq \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda}^2 - \mathcal{C}_N \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda} \|\mathbf{v}\|_{\mathcal{V},\nu} + \nu_0 \|\mathbf{v}\|_{\mathcal{V},\nu}^2. \quad (4.25)$$

Then, owing to the quadratic inequality $x^2 - 2axy + by^2 \geq \frac{b-a^2}{1+b}(x^2 + y^2)$ (valid for any real numbers x, y, a, b with $b \geq 0$) with $a = \frac{1}{2}\mathcal{C}_N$ and $b = \nu_0$, we observe that the choice $\nu_0 \geq 1 + \frac{1}{2}\mathcal{C}_N^2$ implies $b \geq 1 + 2a^2$ so that $\frac{b-a^2}{1+b} \geq \frac{1}{2}$, yielding the expected result. $\square$

Consistency and *a priori* error estimate. We now address the consistency of the scheme (4.19). This analysis uses on the one hand the $\mathbb{P}_0$ -consistency of the reconstruction map $\mathbf{L}_{\mathcal{E}_c}$ and on the other hand, the $\mathbb{P}_0$ -consistency of the normal flux operator $\mathbf{N}_\lambda^{\varepsilon,\partial}$.

Proposition 4.8 (Consistency of $\mathbf{L}_{\mathcal{E}_c}$). *For all $c \in C$, $\mathbf{l}_{\mathcal{E}_c}(\mathbf{V}) = \mathbf{V}$ for all $\mathbf{V} \in \mathbb{P}_0(c; \mathbb{R}^3)$, where the interpolation map is $\mathbf{l}_{\mathcal{E}_c} = \mathbf{L}_{\mathcal{E}_c} \circ \mathbf{R}_{\mathcal{E}_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathbb{P}_0(c; \mathbb{R}^3)$ with $sp > 2$.*

Proof. See the proof of Proposition 7.40. $\square$

Proposition 4.9 (Consistency of the normal flux).

$$\forall \mathbf{f} \in \mathbf{F}^\partial, \quad \tilde{\mathbf{R}}_{\mathcal{V}_f}^\partial(\mathbf{V} \cdot \boldsymbol{\lambda} \cdot \mathbf{n}) = \mathbf{N}_{\lambda;\mathbf{f}}^{\varepsilon,\partial}(\mathbf{R}_{\mathcal{E}_c}(\mathbf{V})), \quad \forall \mathbf{V} \in \mathbb{P}_0(c(\mathbf{f}); \mathbb{R}^3),$$

where the reduction map $\tilde{\mathbf{R}}_{\mathcal{V}_f}^\partial : L^1(\mathbf{f}) \rightarrow \mathcal{V}_{c(\mathbf{f})}$ acts as follows:

$$\forall \mathbf{v} \in \mathcal{V}_{c(\mathbf{f})}, \quad \tilde{\mathbf{R}}_{\mathcal{V}_f}^\partial(w)|_{\mathbf{v}} = \int_{\tilde{\mathbf{f}}^\partial(\mathbf{v}) \cap \mathbf{f}} w, \quad \forall w \in L^1(\mathbf{f}).$$

Proof. This result is a consequence of Proposition 4.8. Let $\mathbf{v} \in \mathcal{V}_f$. Then, Proposition 4.8 yields

$$\mathbf{N}_{\lambda;\mathbf{f}}^{\varepsilon,\partial}(\mathbf{R}_{\mathcal{E}_c}(\mathbf{U}))|_{\mathbf{v}} = \int_{\tilde{\mathbf{f}}^\partial(\mathbf{v}) \cap \mathbf{f}} \mathbf{L}_{\mathcal{E}_c}(\mathbf{R}_{\mathcal{E}_c}(\mathbf{U})) \cdot \boldsymbol{\lambda} \cdot \mathbf{n} = \int_{\tilde{\mathbf{f}}^\partial(\mathbf{v}) \cap \mathbf{f}} \mathbf{U} \cdot \boldsymbol{\lambda} \cdot \mathbf{n} = \tilde{\mathbf{R}}_{\mathcal{V}_f}^\partial(\mathbf{U} \cdot \boldsymbol{\lambda} \cdot \mathbf{n})|_{\mathbf{v}}.$$

$\square$

Let us now introduce the following commutators:

Definition 4.10 (Commutators). *For all $\mathbf{v} \in \mathbf{W}^{1,p}(\Omega)$ with $p \in (\frac{3}{2}, 2]$ and such that $\nabla \times \mathbf{v} \in \mathbf{L}^{\frac{2p}{3-p}}(\Omega)$, we define the two commutators*

$$[\mathbf{H}_\lambda^\varepsilon, \mathbf{R}](\mathbf{v}) := \tilde{\mathbf{R}}_\mathcal{E}(\boldsymbol{\lambda} \cdot \mathbf{v}) - \mathbf{H}_\lambda^\varepsilon(\mathbf{R}_\mathcal{E}(\mathbf{v})), \quad (4.26a)$$

$$[\mathbf{N}_\lambda^{\varepsilon,\partial}, \mathbf{R}](\mathbf{v}) := \tilde{\mathbf{R}}_\mathcal{V}^\partial(\mathbf{n} \cdot \boldsymbol{\lambda} \cdot \mathbf{v}) - \mathbf{N}_\lambda^{\varepsilon,\partial}(\mathbf{R}_\mathcal{E}(\mathbf{v})), \quad (4.26b)$$

where $\mathbf{R}_\mathcal{E}$ is defined by (4.5b), $\tilde{\mathbf{R}}_{\mathcal{V}^\partial} : L^1(\partial\Omega) \rightarrow \mathcal{V}$ is defined (see (3.8c)) as

$$\forall \mathbf{v} \in \mathcal{V}, \quad \tilde{\mathbf{R}}_\mathcal{V}^\partial(w)|_{\mathbf{v}} := \int_{\tilde{\mathbf{f}}^\partial(\mathbf{v})} w, \quad \forall w \in L^1(\partial\Omega),$$

and where $\tilde{\mathbf{R}}_\mathcal{E} : \mathbf{W}^{s,p}(\Omega) \rightarrow \mathcal{E}$ with $sp > 1$ is defined (see (3.8b)) as

$$\forall \mathbf{e} \in \mathbf{E}, \quad \tilde{\mathbf{R}}_\mathcal{E}(\mathbf{w})|_{\mathbf{e}} = \int_{\tilde{\mathbf{f}}(\mathbf{e})} \mathbf{w} \cdot \mathbf{n}_{\tilde{\mathbf{f}}(\mathbf{e})}, \quad \forall \mathbf{w} \in \mathbf{W}^{s,p}(\Omega).$$

Lemma 4.11 (Consistency error). *Let u be the unique solution of (4.8) and assume that $u \in W^{2,p}(\Omega)$ with $p \in (\frac{3}{2}, 2]$. Then,*

$$\mathbb{A}_\lambda^\nu(\mathbf{R}_\nu(u), \mathbf{w}) - \mathbb{S}(s, u_D; \mathbf{w}) = \langle \llbracket \mathbf{H}_\lambda^\varepsilon, \mathbf{R} \rrbracket(\nabla u), \text{GRAD}(\mathbf{w}) \rangle_\varepsilon - \langle \llbracket \mathbf{N}_\lambda^{\varepsilon, \partial}, \mathbf{R} \rrbracket(\nabla u), \mathbf{w} \rangle_\nu, \quad \forall \mathbf{w} \in \mathcal{V}.$$

Proof. By definition, we have $\nabla u \in \mathbf{W}^{1,p}(\Omega)$ so that the two commutators (4.26) are well-defined. Let $\mathbf{w} \in \mathcal{V}$. Owing to definition (4.20) of the source term and using successively the commutation of $\widetilde{\text{DIV}}$ with $\nabla \cdot$ from Proposition 3.5 and the anti-adjunction between GRAD and $\widetilde{\text{DIV}}$ from Proposition 3.3, it follows that

$$\begin{aligned} \mathbb{S}(s, u_D; \mathbf{w}) &= \langle \widetilde{\mathbf{R}}_\nu(s), \mathbf{w} \rangle_\nu + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial} \mathbf{R}_\nu^\partial(u_D), \mathbf{w} \rangle_\nu \\ &= -\langle \widetilde{\mathbf{R}}_\nu(\nabla \cdot (\lambda \cdot \nabla u)), \mathbf{w} \rangle_\nu + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial} \mathbf{R}_\nu^\partial(u_D), \mathbf{w} \rangle_\nu \\ &= -\langle \widetilde{\text{DIV}} \widetilde{\mathbf{R}}_\varepsilon(\lambda \cdot \nabla u), \mathbf{w} \rangle_\nu - \langle \widetilde{\mathbf{R}}_\nu^\partial(\nabla u \cdot \lambda \cdot \mathbf{n}), \mathbf{w} \rangle_\nu + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial} \mathbf{R}_\nu^\partial(u_D), \mathbf{w} \rangle_\nu \\ &= \langle \widetilde{\mathbf{R}}_\varepsilon(\lambda \cdot \nabla u), \text{GRAD}(\mathbf{w}) \rangle_\nu - \langle \widetilde{\mathbf{R}}_\nu^\partial(\nabla u \cdot \lambda \cdot \mathbf{n}), \mathbf{w} \rangle_\nu + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial} \mathbf{R}_\nu^\partial(u_D), \mathbf{w} \rangle_\nu. \end{aligned}$$

Moreover, recalling that

$$\mathbb{A}_\lambda^\nu(\mathbf{R}_\nu(u), \mathbf{v}) = \langle \mathbf{H}_\lambda^\varepsilon \text{GRAD} \mathbf{R}_\nu(u), \text{GRAD}(\mathbf{w}) \rangle_\varepsilon - \langle \mathbf{N}_\lambda^{\varepsilon, \partial} \text{GRAD}(\mathbf{R}_\nu(u)), \mathbf{w} \rangle_\nu + \nu_0 \langle \mathbf{H}_\nu^{\nu, \partial} \mathbf{R}_\nu^\partial(u), \mathbf{w} \rangle_\nu,$$

the expected result follows from the exact commutation of GRAD with ∇ from Proposition 4.1 and observing that $u|_{\partial\Omega} = u_D$ holds point-wise. $\square$

Theorem 4.12 (Diffusive convergence rate). *Let u be the unique solution of (4.8) and let $\mathbf{u}$ be the unique solution of (4.19). Assume that $\nu_0 \geq 1 + \frac{1}{2} \mathcal{C}_\mathbf{N}^2$ (see Lemma 4.7) and that $u \in H^2(\Omega)$. Then,*

$$\|\mathbf{u} - \mathbf{R}_\nu(u)\|_{\mathcal{V}, \mathbf{d}} \lesssim \left(\sum_{c \in \mathcal{C}} \rho_c \lambda_{\sharp, c} h_c^2 |u|_{H^2(c)}^2 \right)^{\frac{1}{2}}. \quad (4.27)$$

Remark 4.13 (Regularity). *Observe that the regularity $H^2(\Omega)$ on the exact solution is required to achieve the first order convergence rate in the discrete coercivity norm $\|\cdot\|_{\mathcal{V}, \mathbf{d}}$. This assumption is stronger than the regularity needed to compute the consistency error $\mathbb{A}_\lambda^\nu(\mathbf{R}_\nu(u), \mathbf{w}) - \mathbb{S}(s, u_D; \mathbf{w})$ in Lemma 4.11. Note also that the estimate (4.27) holds as well if $u \in H^2(\mathcal{C})$, i.e., if u is only piece-wise H^2 on mesh cells.*

The estimate (4.27) is a consequence of the following estimates:

Proposition 4.14 (Approximation on mesh edges and dual faces). *Let $\mathbf{w} \in \mathbf{H}^1(\Omega)$. Then,*

$$\forall c \in \mathcal{C}, \quad \sum_{e \in \mathcal{E}_c} \left\| \mathbf{w} - \frac{1}{|c|} \int_c \mathbf{w} \right\|_{L^1(e)} \lesssim h_c^{\frac{1}{2}} |\mathbf{w}|_{\mathbf{H}^1(c)} \quad \text{and} \quad \sum_{e \in \mathcal{E}_c} \left\| \mathbf{w} - \frac{1}{|c|} \int_c \mathbf{w} \right\|_{L^1(\tilde{f}_c(e))} \lesssim h_c^{\frac{3}{2}} |\mathbf{w}|_{\mathbf{H}^1(c)}.$$

Proof. These estimates are obtained by proceeding as in the proof of (Bonelle & Ern, 2014a, Theorem 3.3) and as in the proof of Lemma 7.41, using successive extensions from mesh edges and mesh dual faces to mesh cells. $\square$

Proof of Theorem 4.12. Let T_1 and T_2 be the two terms in the right-hand side of the consistency error identified in Lemma 4.11. The term T_1 has already been bounded in Bonelle & Ern (2014a). We present here a somewhat simpler proof avoiding the algebraic identity on the inverse of the discrete Hodge operator. Let $\mathbf{g}_c$ denote the mean-value of ∇u in c . Owing to the local assembly (4.10) and the $\mathbb{P}_0$ -consistency of the reconstruction map $\mathbf{L}_{\mathcal{E}_c}$ from Proposition 4.8, we infer that

$$\begin{aligned} T_1 &= \sum_{c \in \mathcal{C}} \langle \widetilde{\mathbf{R}}_{\mathcal{E}_c}(\lambda \cdot \nabla u) - \mathbf{H}_{\lambda; c}^\varepsilon(\mathbf{R}_{\mathcal{E}_c}(\nabla u)), \text{GRAD}(\mathbf{v}) \rangle_{\mathcal{E}_c} \\ &= \sum_{c \in \mathcal{C}} \langle \widetilde{\mathbf{R}}_{\mathcal{E}_c}(\lambda \cdot (\nabla u - \mathbf{g}_c)), \text{GRAD}(\mathbf{v}) \rangle_{\mathcal{E}_c} - \sum_{c \in \mathcal{C}} \langle \mathbf{H}_{\lambda; c}^\varepsilon(\mathbf{R}_{\mathcal{E}_c}(\nabla u - \mathbf{g}_c)), \text{GRAD}(\mathbf{v}) \rangle_{\mathcal{E}_c} = T_{1,1} + T_{1,2}. \end{aligned}$$

Owing to the discrete Cauchy–Schwarz inequality, shape-regularity and the fact that the diffusion tensor is positive, we have

$$|T_{1,1}|^2 \lesssim \left(\sum_{c \in C} \lambda_{b;c} \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E}_c,2}^2 \right) \left(\sum_{c \in C} h_c^{-1} \lambda_{b;c}^{-1} \|\boldsymbol{\lambda} \cdot (\nabla u - \mathbf{g}_c)\|_{\mathbf{L}^1(\tilde{f}_c(e))}^2 \right),$$

so that the stability of $\mathbf{L}_{\mathcal{E}_c}$ from Proposition 4.5 together with the definition of the discrete energy norm $\|\cdot\|_{\mathcal{E},\lambda}$ and with the anisotropy ratio $\rho_c = \lambda_{\sharp;c}/\lambda_{b;c}$ yield

$$|T_{1,1}|^2 \leq \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda}^2 \left(\sum_{c \in C} \sum_{e \in \mathbf{E}_c} \rho_c \lambda_{\sharp;c} h_c^{-1} \|\nabla u - \mathbf{g}_c\|_{\mathbf{L}^1(\tilde{f}_c(e))}^2 \right).$$

Hence, owing to Proposition 4.14, we obtain

$$|T_{1,1}|^2 \lesssim \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda}^2 \left(\sum_{c \in C} \rho_c \lambda_{\sharp;c} h_c^2 |u|_{H^2(c)}^2 \right).$$

Proceeding similarly, the discrete Cauchy–Schwarz inequality $\langle \mathbf{H}_{\lambda;c}^\varepsilon(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{E}_c} \leq \|\mathbf{v}\|_{\lambda;c} \|\mathbf{w}\|_{\lambda;c}$ for all $\mathbf{v}, \mathbf{w} \in \mathcal{E}_c$, the upper bound in Proposition 4.5 and Proposition 4.14 yield

$$|T_{1,2}|^2 \leq \mathbf{c}_{\sharp} \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda}^2 \left(\sum_{c \in C} \sum_{e \in \mathbf{E}_c} \lambda_{\sharp;c} h_c \|\nabla u - \mathbf{g}_c\|_{\mathbf{L}^1(e)}^2 \right) \lesssim \|\text{GRAD}(\mathbf{v})\|_{\mathcal{E},\lambda}^2 \left(\sum_{c \in C} \rho_c \lambda_{\sharp;c} h_c^2 |u|_{H^2(c)}^2 \right).$$

Turning to T_2 , we use the local assembly (4.14) and Proposition 4.9 to infer that

$$\begin{aligned} T_2 &= - \sum_{f \in \mathbf{F}^\partial} \langle \tilde{\mathbf{R}}_{\mathcal{V}_f^\partial}(\mathbf{n} \cdot \boldsymbol{\lambda} \cdot \nabla u) - \mathbf{N}_{\lambda;f}^{\varepsilon,\partial} \mathbf{R}_{\mathcal{E}_c}(\nabla u), \mathbf{v} \rangle_{\mathcal{V}_c} \\ &= - \sum_{f \in \mathbf{F}^\partial} \langle \tilde{\mathbf{R}}_{\mathcal{V}_f^\partial}(\mathbf{n} \cdot \boldsymbol{\lambda} \cdot (\nabla u - \mathbf{g}_c), \mathbf{v} \rangle_{\mathcal{V}_c} + \sum_{f \in \mathbf{F}^\partial} \langle \mathbf{N}_{\lambda;f}^{\varepsilon,\partial} \mathbf{R}_{\mathcal{E}_c}(\nabla u - \mathbf{g}_c), \mathbf{v} \rangle_{\mathcal{V}_c} = T_{2,1} + T_{2,2}, \end{aligned}$$

with $c = c(f)$. Proceeding as above, we end up with

$$|T_{2,1}|^2 \leq \|\mathbf{v}\|_{\mathcal{V},\nu}^2 \left(\sum_{f \in \mathbf{F}^\partial} \sum_{v \in \mathbf{V}_f^\partial} \lambda_{\sharp;c} h_c^{-1} \|\nabla u - \mathbf{g}_c\|_{\mathbf{L}^1(\tilde{f}(v) \cap f)}^2 \right) \lesssim \|\mathbf{v}\|_{\mathcal{V},\nu}^2 \left(\sum_{f \in \mathbf{F}^\partial} \lambda_{\sharp;c} h_c^2 |u|_{H^2(c)}^2 \right),$$

while using Lemma 4.9 we infer a similar bound on $T_{2,2}$. The proof is completed since $\rho_c \geq 1$ by definition. $\square$

Remark 4.15 (On the proof of Theorem 4.12). *Observe that this proof does not use explicitly any estimate on the interpolation error $\|\mathbf{w} - \mathbf{l}_{\mathcal{E}_c}(\mathbf{w})\|_{\mathbf{L}^2(c)}$ for all $c \in C$. We only use the fact that $\mathbf{L}_{\mathcal{E}_c}$ preserves exactly piece-wise constant polynomials in $\mathbb{P}_0(C; \mathbb{R}^3)$ from Proposition 4.8 and the approximation results stated by Proposition 4.14.*

4.3 Diffusion-advection-reaction problem

In this section, we devise a robust approximation with respect to the Péclet number of the solution of the scalar diffusion-advection-reaction problem

$$-\nabla \cdot (\boldsymbol{\lambda} \nabla u) + \boldsymbol{\beta} \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (4.28a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega. \quad (4.28b)$$

The data are such that $s \in L^2(\Omega)$ and $u_D \in H^{1+\epsilon}(\Omega)$ with $\epsilon > 0$. We recall that $\boldsymbol{\lambda} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ takes symmetric, positive, and piece-wise constant values on the mesh cells and we denote

$\lambda_{b;c}$ and $\lambda_{\sharp;c}$ its minimal and maximal eigenvalue on $c \in \mathcal{C}$, respectively. The continuous problem (4.28) is well-posed if

$$\mathbf{C}_P \lambda_b + \operatorname{ess\,inf}_{x \in \Omega} \sigma_{\beta,\mu}(x) > 0, \quad (4.29)$$

with $\lambda_b = \min_{c \in \mathcal{C}} \lambda_{b;c}$, $\sigma_{\beta,\mu} = \mu - \frac{1}{2} \nabla \cdot \beta$ and where $\mathbf{C}_P > 0$ denotes the Poincaré constant in $H_0^1(\Omega)$. Assuming that assumption $(\mathcal{H})$ holds, i.e., $\operatorname{ess\,inf}_{\Omega} \sigma_{\beta,\mu} > 0$, the condition (4.29) is always satisfied since λ is positive. Then, it is possible to consider the limite case when the diffusive tensor uniformly vanishes, i.e., when $\lambda_b \rightarrow 0$. However, if the model parameters β and μ only satisfy the assumption $(\mathcal{H}')$ considered in Section (3.2), stating that $\operatorname{ess\,inf}_{\Omega} \sigma_{\beta,\mu} = 0$ and that there exists a Lipschitz function $\zeta : \Omega \rightarrow \mathbb{R}$ such that $\operatorname{ess\,inf}_{\Omega} -\frac{1}{2} \beta \cdot \nabla \zeta > 0$, the condition (4.29) does not allow us to consider vanishing diffusive tensors. The condition (4.29) then must be replaced by

$$\mathbf{C}_P \lambda_b + \operatorname{ess\,inf}_{x \in \Omega} \left(-\frac{1}{2} \beta \cdot \nabla \zeta(x) \right) > 0. \quad (4.30)$$

4.3.1 Péclet-based upwind scheme

Vertex-based CDO schemes for the continuous problem (4.28) readily follow, on the one hand, from the scheme (4.19) for the diffusion with weak boundary condition enforcement and on the other hand, from the scheme (3.21) for the advection-reaction. We introduce the global bilinear form $\mathbb{A}_{\lambda,\beta,\mu}^v : \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$, defined by

$$\mathbb{A}_{\lambda,\beta,\mu}^v := \mathbb{A}_{\lambda}^v + \mathbb{A}_{\beta,\mu}^v, \quad (4.31)$$

where $\mathbb{A}_{\lambda}^v$ is defined by (4.9) and where $\mathbb{A}_{\beta,\mu}^v$ by (3.16) is modified by means of a local Péclet number (see below).

Péclet-based upwinding. In the presence of diffusive and advective phenomena, we introduce an algebraic local Péclet number on all mesh edges:

$$\forall e \in \mathcal{E}, \quad \Pi_e = \beta_e h_e |\tilde{f}(e)|^{-1} \lambda_e^{-1}, \quad (4.32)$$

where $\lambda_e = \max_{c \in \mathcal{C}_e} \lambda_{b;c}$ and where $\beta_e = \tilde{\mathbf{R}}_{\mathcal{E}}(\beta)|_e$ denotes the reduction on dual mesh faces using the reduction map $\tilde{\mathbf{R}}_{\mathcal{E}}$ defined by (3.23). Then, recalling (see definition (3.25)) that

$$\forall v \in \mathcal{V}, \quad \mathbf{J}_{\beta}^{\mathcal{E}}(w)|_v = \frac{1}{2} \sum_{e \in \mathcal{E}_v} w_e \beta_e (1 - \kappa_{ve}), \quad \forall w \in \mathcal{E}, \quad (4.33)$$

we consider the Péclet-dependent upwinding parameter $\kappa_{ve} = \pi(\iota_{\tilde{f}(e), \tilde{c}(v)} \Pi_e)$ where the function $\pi : \mathbb{R} \rightarrow \mathbb{R}$ satisfies

(i_π) $\pi(x) + \pi(-x) = 0$ and $\pi(x) \geq 0$ for all $x \in \mathbb{R}_{\geq 0}$.

(ii_π) There exists $\mathbf{C}_{\pi} > 0$ such that $\pi(x) \geq \mathbf{C}_{\pi}$ for all $x \geq 1$.

Observe that (i_π) implies (i_κ) since $\beta_e \kappa_e = \frac{1}{2} \lambda_e |\tilde{f}(e)| h_e^{-1} \sum_{v \in \mathcal{V}_e} \iota_{\tilde{f}(e), \tilde{c}(v)} \Pi_e \pi(\iota_{\tilde{f}(e), \tilde{c}(v)} \Pi_e) \geq 0$. As a consequence, since (i_κ) holds, Propositions 3.13 and 3.14 hold. An example for the function π is the Sharfetter-Gummel map defined as $\pi(x) = \coth\left(\frac{x}{2}\right) - \frac{2}{x}$; see Roos *et al.* (2008) for further insights and examples. The lower bound on x in the second assumption (ii_π) is arbitrary; changing its value only changes the constants in the error bounds.

Remark 4.16 (Link with the HHO treatment of advection terms). *It is also possible to consider positive Péclet numbers with the symmetric function $|A|(x) = x\pi(x)$ introduced by Di Pietro et al. (2015) in the context of high order face-based (HHO) discretizations, where the function π satisfies assumptions (i_π)–(ii_π).*

Discrete problem. The scheme approximating (4.28) reads:

$$\text{Find } \mathbf{u} \in \mathcal{V} \text{ s.t. } \mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}(\mathbf{u}, \mathbf{v}) = \mathbb{S}(s, u_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{V}, \quad (4.34)$$

where the source term $\mathbb{S}(s, u_D; \cdot) : \mathcal{V} \rightarrow \mathbb{R}$ is defined as

$$\mathbb{S}(s, u_D; \mathbf{v}) = \langle \tilde{\mathbf{R}}_{\mathcal{V}}(s), \mathbf{v} \rangle_{\mathcal{V}} + \nu_0 \langle \mathbf{H}_{\mathcal{V}}^{\mathcal{V}, \partial} \mathbf{R}_{\mathcal{V}}^{\partial}(u_D), \mathbf{v} \rangle_{\mathcal{V}} + \langle \tilde{\mathbf{R}}_{\mathcal{V}}((\boldsymbol{\beta} \cdot \mathbf{n})^{\ominus} u_D), \mathbf{v} \rangle_{\mathcal{V}}, \quad \forall \mathbf{v} \in \mathcal{V}. \quad (4.35)$$

4.3.2 Analysis

Following the definition (4.31) of $\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}$ and the analysis of Sections 3.2.2 and 4.2.2, related to the diffusion and the advection-reaction problems, respectively, the stability norm used to analyze the discrete problem (4.34) is defined on $\mathcal{V}$ as

$$\|\mathbf{w}\|_{\mathcal{V}, \text{da}}^2 := \|\mathbf{w}\|_{\mathcal{V}, \text{a}}^2 + \|\mathbf{w}\|_{\mathcal{V}, \text{d}}^2, \quad \forall \mathbf{w} \in \mathcal{V}, \quad (4.36)$$

where the advection-related stability norm $\|\cdot\|_{\mathcal{V}, \text{a}}$ is defined by (3.34) and where the diffusion-related stability norm $\|\cdot\|_{\mathcal{V}, \text{d}}$ is defined by (4.24).

Stability. The following Lemma 4.17 analyzes the stability of the bilinear form $\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}$ under the assumption $(\mathcal{H})$, so that the well-posedness condition (4.29) holds for all Péclet number.

Lemma 4.17 (Coercivity). *Assume that assumption $(\mathcal{H})$ holds and that $\nu_0 \geq 1 + \frac{1}{2} \mathcal{C}_{\text{N}}^2$. Then,*

$$\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}(\mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V}, \text{da}}^2, \quad \forall \mathbf{v} \in \mathcal{V}.$$

Consequently, the discrete problem (4.34) is well-posed.

Proof. This estimate is a direct consequence of Lemmata 3.18 and 4.7. $\square$

The stability of the bilinear form $\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}$ is now analyzed under the assumption $(\mathcal{H}')$. In order to facilitate the tracking of the constant, we assume that there exists a constant $\mathbf{C}_{\mathcal{V}, \text{da}} > 0$, independent of the mesh-size and the model parameter, such that

$$\tau \max \left(L_{\beta}, L_{\zeta} \|\mu\|_{L^{\infty}(\Omega)} h, L_{\zeta}^2 \|\beta\|_{L^{\infty}(\Omega)} h, L_{\zeta}^2 \lambda_{\sharp} \right) \leq \mathbf{C}_{\mathcal{V}, \text{da}}, \quad (4.37)$$

with $\lambda_{\sharp} = \max_{\mathbf{c} \in \mathbf{C}} \lambda_{\sharp; \mathbf{c}}$. We also recall the definition of the reference length h_0 defined as

$$h_0 = \left(L_{\zeta} \tau \|\nabla \cdot \beta\|_{L^{\infty}(\Omega)} \right)^{-1}, \quad (4.38)$$

with the convention $h_0 = +\infty$ if β is divergence-free.

Lemma 4.18 (Inf-sup stability). *Assume that assumption $(\mathcal{H}')$ holds. Assume that (4.37) holds and that the mesh size satisfies $0 < 2h < h_0$. Then, provided ν_0 large enough, there exists $\varrho > 0$ such that*

$$\sup_{\mathbf{w} \in \mathcal{V}; \|\mathbf{w}\|_{\mathcal{V}, \text{da}}=1} \mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{V}}(\mathbf{v}, \mathbf{w}) \geq \varrho \|\mathbf{v}\|_{\mathcal{V}, \text{da}}, \quad \forall \mathbf{v} \in \mathcal{V}.$$

Consequently, the discrete problem (3.21) is well-posed.

Remark 4.19 (Comparison with Cantin & Ern (2016b)). *Observe that this Lemma extends Lemma 5.3 in Cantin & Ern (2016b) where the simpler case $\nabla \cdot \beta = 0$ and $\mu = 0$ a.e. in Ω is considered with no mesh-size restriction.*

Case	$\text{ess inf}_\Omega \sigma_{\beta,\mu} > 0$ and $\lambda_b \geq 0$	$\text{ess inf}_\Omega \sigma_{\beta,\mu} = 0$ and $\lambda_b \geq 0$	
Assumption	$(\mathcal{H})$	$(\mathcal{H}')$	
Advective field		$\nabla \cdot \beta = 0$	$\nabla \cdot \beta \neq 0$
Nitsche penalty term	$\nu_0 \geq 1 + \frac{1}{2} \mathcal{C}_N^2$	$\nu_0 \geq \nu_{0,\sharp}$	
Mesh-size	$0 < h$	$0 < h$	$0 < 2h < h_0$

Table 4.1 – Stability of the discrete problem (4.34) with h_0 defined by (4.38) and $\nu_{0,\sharp}$ defined by (4.39).

Remark 4.20 (Lower bound on ν_0). *More precisely, Lemma 4.18 holds if $\nu_0 \geq \nu_{0,\sharp}$ with*

$$\nu_{0,\sharp} = \frac{1 + \mathcal{C}_N^2 (2\mathcal{C}_\sharp^2 \mathcal{C}_{\nu,da} n_{E,V} + \|\zeta\|_{L^\infty(\Omega)})^2}{3(1 + 2\mathcal{C}_\sharp^2 \mathcal{C}_{\nu,da} n_{E,V})} \quad (4.39)$$

with $n_{E,V} = \max_{V \in \mathcal{V}} \#E_V$. This lower bound slightly differs, up to a numerical factor, from that obtained in Lemma (4.7) for zero advection since both proofs have not been optimized regarding the lower bound in the quadratic inequality.

Table 4.1 recapitulates the different conditions from Lemmata 4.17 and 4.18 for which the discrete problem (4.34) is well-posed.

Proof of Lemma 4.18. Following the proof of Proposition 3.19, there exist two real numbers $\theta_a, \mathcal{C}_{\zeta,a} > 0$ such that

$$\mathbb{A}_{\beta,\mu}^\nu(v, \zeta v + \theta_a v) \geq \frac{1}{2} \|v\|_{V,a}^2 - \frac{1}{2} L_\zeta h \|\nabla \cdot \beta\|_{L^\infty(\Omega)} \|v\|_{V,2}^2, \quad (4.40)$$

and such that $\|\zeta v\|_{V,a} \leq \mathcal{C}_{\zeta,a} \|v\|_{V,a} \leq \mathcal{C}_{\zeta,a} \|v\|_{V,da}$ since (i_κ) and assumption (4.37) hold. Now, let us prove that there exists $\mathcal{C}_{\zeta,d} > 0$ such that $\|\zeta v\|_{V,d} \leq \mathcal{C}_{\zeta,d} \|v\|_{V,da}$. First, recalling the definition (4.15) of $H_{\nu,\theta}^\nu$, it follows that $\|\zeta v\|_{V,\nu} \leq \|\zeta\|_{L^\infty(\Omega)} \|v\|_{V,\nu}$. Furthermore, owing to the cell-wise assembly (4.10) of H_λ^ε , we infer that

$$\begin{aligned} \|\text{GRAD}(\zeta v)\|_\lambda^2 &= \sum_{c \in \mathcal{C}} \langle H_{\lambda;c}^\varepsilon \text{GRAD}(\zeta v), \text{GRAD}(\zeta v) \rangle_{\mathcal{E}_c} \\ &= \sum_{c \in \mathcal{C}} \zeta_c^2 \langle H_{\lambda;c}^\varepsilon \text{GRAD}(v), \text{GRAD}(v) \rangle_{\mathcal{E}_c} + \sum_{c \in \mathcal{C}} \|\text{GRAD}((\zeta - \zeta_c)v)\|_{\mathcal{E}_c,\lambda}^2 \\ &\leq \|\zeta\|_{L^\infty(\Omega)}^2 \|\text{GRAD}(v)\|_{\mathcal{E},\lambda}^2 + \sum_{c \in \mathcal{C}} \|\text{GRAD}((\zeta - \zeta_c)v)\|_{\mathcal{E}_c,\lambda}^2, \end{aligned}$$

where ζ_c denotes the mean-value of ζ in c . Then, recalling the definition (4.22) of the local norm $\|\cdot\|_{\mathcal{E}_c,\lambda}$, it follows that

$$\|\text{GRAD}((\zeta - \zeta_c)v)\|_{\mathcal{E}_c,\lambda}^2 = \int_c \mathbf{L}_{\mathcal{E}_c}(\text{GRAD}((\zeta - \zeta_c)v)) \cdot \lambda \cdot \mathbf{L}_{\mathcal{E}_c}(\text{GRAD}((\zeta - \zeta_c)v)),$$

so that owing to the upper bound of Proposition 4.5, the Lipschitz regularity of ζ and mesh-regularity, we infer that

$$\|\text{GRAD}((\zeta - \zeta_c)v)\|_{\mathcal{E}_c,\lambda}^2 \leq \mathcal{C}_\sharp^2 \lambda_{\sharp;c} h_c \sum_{e \in E_c} \left(\sum_{v \in V_e} \iota_{v,e} (\zeta(\mathbf{x}_v) - \zeta_c) v_v \right)^2 \leq 2\mathcal{C}_\sharp^2 \lambda_{\sharp;c} L_\zeta^2 h_c^3 \sum_{e \in E_c} \sum_{v \in V_e} v_v^2.$$

Exchanging summations, introducing $n_{E,V} = \max_{V \in \mathcal{V}} \#E_V$ and recalling the assumption (4.37) imply that

$$\|\text{GRAD}((\zeta - \zeta_c)v)\|_{\mathcal{E}_c,\lambda}^2 \leq 2\mathcal{C}_\sharp^2 \lambda_{\sharp;c} L_\zeta^2 n_{E,V} \|v\|_{V_c,2}^2 \leq 2\mathcal{C}_\sharp^2 n_{E,V} \mathcal{C}_{\nu,da} \tau^{-1} \|v\|_{V_c,2}^2,$$

whence the bound $\|\zeta \mathbf{v}\|_{\mathcal{V},d} \leq \mathbf{C}_{\zeta,d} \|\mathbf{v}\|_{\mathcal{V},da}$. Next, let us bound from below $\mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v})$. Using Proposition 4.6 and $\zeta \geq 1$, we first have

$$\begin{aligned} \mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v}) &= \langle \mathbf{H}_{\lambda}^{\varepsilon} \text{GRAD}(\mathbf{v}), \text{GRAD}(\zeta \mathbf{v}) \rangle_{\varepsilon} - \langle \mathbf{N}_{\lambda}^{\varepsilon, \theta} \text{GRAD}(\mathbf{v}), \zeta \mathbf{v} \rangle_{\mathcal{V}} + \nu_0 \langle \mathbf{H}_{\nu}^{\mathcal{V}, \theta}(\mathbf{v}), \zeta \mathbf{v} \rangle_{\mathcal{V}} \\ &\geq \langle \mathbf{H}_{\lambda}^{\varepsilon} \text{GRAD}(\mathbf{v}), \text{GRAD}(\zeta \mathbf{v}) \rangle_{\varepsilon} - \mathbf{C}_N \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda} \|\zeta \mathbf{v}\|_{\mathcal{V}, \nu} + \nu_0 \|\mathbf{v}\|_{\mathcal{V}, \nu}^2 \\ &\geq \langle \mathbf{H}_{\lambda}^{\varepsilon} \text{GRAD}(\mathbf{v}), \text{GRAD}(\zeta \mathbf{v}) \rangle_{\varepsilon} - \mathbf{C}_N \|\zeta\|_{L^{\infty}(\Omega)} \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda} \|\mathbf{v}\|_{\mathcal{V}, \nu} + \nu_0 \|\mathbf{v}\|_{\mathcal{V}, \nu}^2. \end{aligned}$$

Using now the cell-wise assembly of $\mathbf{H}_{\lambda}^{\varepsilon}$ and proceeding as above, we infer that

$$\begin{aligned} \langle \mathbf{H}_{\lambda}^{\varepsilon} \text{GRAD}(\mathbf{v}), \text{GRAD}(\zeta \mathbf{v}) \rangle_{\varepsilon} &\geq \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda}^2 + \sum_{c \in \mathcal{C}} \langle \mathbf{H}_{\lambda; c}^{\varepsilon} \text{GRAD}(\mathbf{v}), \text{GRAD}((\zeta - \zeta_c) \mathbf{v}) \rangle_{\varepsilon} \\ &\geq \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda}^2 - \left(\sum_{c \in \mathcal{C}} \|\text{GRAD}((\zeta - \zeta_c) \mathbf{v})\|_{\varepsilon, \lambda}^2 \right)^{\frac{1}{2}} \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda} \\ &\geq \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda}^2 - (2\mathbf{C}_{\#}^2 \mathbf{C}_{\mathcal{V}, da} n_{E, \mathcal{V}})^{\frac{1}{2}} \left(\tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2 \right)^{\frac{1}{2}} \|\text{GRAD}(\mathbf{v})\|_{\varepsilon, \lambda}. \end{aligned}$$

As a result, considering $\mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v} + \theta_d \mathbf{v})$ with $\theta_d > 0$ (to be chosen below) along with the estimate (4.25), we arrive at

$$\begin{aligned} \mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v} + \theta_d \mathbf{v}) &\geq (1 + \theta_d) \|\mathbf{w}\|_{\varepsilon, \lambda}^2 - (\|\zeta\|_{L^{\infty}(\Omega)} + \theta_d) \mathbf{C}_N \|\mathbf{w}\|_{\varepsilon, \lambda} \|\mathbf{v}\|_{\mathcal{V}, \nu} \\ &\quad - (2\mathbf{C}_{\#}^2 \mathbf{C}_{\mathcal{V}, da} n_{E, \mathcal{V}})^{\frac{1}{2}} \left(\tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2 \right)^{\frac{1}{2}} \|\mathbf{w}\|_{\varepsilon, \lambda} + (1 + \theta_d) \nu_0 \|\mathbf{v}\|_{\mathcal{V}, \nu}^2, \end{aligned}$$

where we have denoted $\mathbf{w} = \text{GRAD}(\mathbf{v})$. Young's inequality for the third term on the right-hand side yields

$$\mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v} + \theta_d \mathbf{v}) \geq \|\mathbf{w}\|_{\varepsilon, \lambda}^2 - (\|\zeta\|_{L^{\infty}(\Omega)} + \theta_d) \mathbf{C}_N \|\mathbf{w}\|_{\varepsilon, \lambda} \|\mathbf{v}\|_{\mathcal{V}, \nu} - \frac{\mathbf{C}_{\#}^2 \mathbf{C}_{\mathcal{V}, da} n_{E, \mathcal{V}}}{2\theta_d} \tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2 + (1 + \theta_d) \nu_0 \|\mathbf{v}\|_{\mathcal{V}, \nu}^2.$$

Hence, choosing $\theta_d = 2\mathbf{C}_{\#}^2 \mathbf{C}_{\mathcal{V}, da} n_{E, \mathcal{V}}$, so that $\frac{\mathbf{C}_{\#}^2 \mathbf{C}_{\mathcal{V}, da} n_{E, \mathcal{V}}}{2\theta_d} = \frac{1}{4}$, along with quadratic identity $x^2 - 2axy + by^2 \geq \frac{b-a^2}{1+b}(x^2 + y^2)$ with $a = \frac{1}{2}(\mathbf{C}_N + \|\zeta\|_{L^{\infty}(\Omega)})$ and $b = (1 + \theta_d)\nu_0$, and observing that the choice $\nu_0 \geq \frac{1 + \mathbf{C}_N^2(\theta_d + \|\zeta\|_{L^{\infty}(\Omega)})^2}{3(1 + \theta_d)}$ implies that $b \geq \frac{1}{3} + \frac{4}{3}a^2$ so that $\frac{b-a^2}{1+b} \geq \frac{1}{4}$, we infer that

$$\mathbb{A}_{\lambda}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v} + \theta_d \mathbf{v}) \geq \frac{1}{4} \|\mathbf{v}\|_{\mathcal{V}, d}^2 - \frac{1}{4} \tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2. \quad (4.41)$$

As a result, collecting (4.40) and (4.41) and denoting $\theta_{da} = \max\{\theta_d, \theta_a\}$, we obtain

$$\begin{aligned} \mathbb{A}_{\lambda, \beta, \mu}^{\mathcal{V}}(\mathbf{v}, \zeta \mathbf{v} + \theta_{da} \mathbf{v}) &\geq \frac{1}{4} \|\mathbf{v}\|_{\mathcal{V}, d}^2 - \frac{1}{4} \tau^{-1} \|\mathbf{v}\|_{\mathcal{V}, 2}^2 + \frac{1}{2} \|\mathbf{v}\|_{\mathcal{V}, a}^2 - \frac{1}{2} L_{\zeta} h \|\nabla \cdot \beta\|_{L^{\infty}(\Omega)} \|\mathbf{v}\|_{\mathcal{V}, 2}^2 \\ &\geq \frac{1}{4} \left(1 - \frac{2h}{h_0} \right) \|\mathbf{v}\|_{\mathcal{V}, da}^2. \end{aligned}$$

Hence, since $\|\zeta \mathbf{v} + \theta_{da} \mathbf{v}\|_{\mathcal{V}, da} \leq (\mathbf{C}_{\zeta, a} + \mathbf{C}_{\zeta, d} + \theta_{da}) \|\mathbf{v}\|_{\mathcal{V}, da}$, the proof is completed. $\square$

A priori error estimate We now establish an *a priori* error estimate on the discrete error with respect to the local Péclet numbers $\{\Pi_e\}_{e \in \mathcal{E}}$ defined by (4.32).

Theorem 4.21 (Péclet dependent *a priori* error estimate). *Let u be the exact solution of (4.28) and let $\mathbf{u}$ be the discrete solution of (4.34). Assume that one of the stability assumption from Table 4.1 holds. Assume that $u \in H^2(\Omega)$. Then,*

$$\|\mathbf{u} - \mathbf{R}_{\mathcal{V}}(u)\|_{\mathcal{V}, da} \lesssim \left(\sum_{c \in \mathcal{C}} \rho_c \lambda_{\#; c} h_c^2 |u|_{H^2(c)}^2 \right)^{\frac{1}{2}} + \left(\sum_{c \in \mathcal{C}} \varpi_c h_c \left(|u|_{H^1(c)}^2 + h_c^2 |u|_{H^2(c)}^2 \right) \right)^{\frac{1}{2}}, \quad (4.42a)$$

where

$$\varpi_c = \tau \|\mu - \nabla \cdot \beta\|_{L^\infty(c)}^2 h_c + \sum_{e \in E_c} \|\beta\|_{L^\infty(\tilde{f}_c(e))} \min \left(1, |\Pi_e| + \frac{L_\beta h_e^2}{\lambda_e} \right). \quad (4.42b)$$

Remark 4.22 (Convergence rate with respect to the Péclet Number). *In the advection-dominant regime defined by $\Pi_e > 1$ for all $e \in E$, the error estimate (4.42a) behaves as*

$$\|u - R_V(u)\|_{V,da} = \mathcal{O} \left(\|\beta\|_{L^\infty(\Omega)}^{\frac{1}{2}} h^{\frac{1}{2}} \left(|u|_{H^1(\Omega)} + h^{\frac{1}{2}} |u|_{H^2(\Omega)} \right) \right) \quad (4.43a)$$

since $\varpi_c = \mathcal{O}(\|\beta\|_{L^\infty(c)})$. *In the diffusion-dominant regime defined by $\Pi_e \leq h_e L_\Omega^{-1}$ for all $e \in E$, where L_Ω is a characteristic length of Ω , the estimate (4.42a) behaves as*

$$\|u - R_V(u)\|_{V,da} = \mathcal{O} \left(\lambda_\#^{\frac{1}{2}} h |u|_{H^2(\Omega)} \right) \quad (4.43b)$$

since $\varpi_c = \mathcal{O}(\|\beta\|_{L^\infty(c)} h_c L_\Omega^{-1})$. *The intermediary case $h_e L_\Omega^{-1} < \Pi_e \leq 1$ corresponds to transition regimes and intermediate orders of convergence.*

Proposition 4.23 (Interpolation error estimates). *Let $c \in C$ and define the interpolation map $\mathbf{l}_{V_c} = \mathbf{L}_{V_c} \circ R_{V_c} : W^{s,2}(c) \rightarrow \mathbb{P}_0(\mathfrak{C}_{V,c}; \mathbb{R})$ with $s > \frac{3}{2}$ where R_{V_c} is defined by (4.5a) and where $\mathbf{L}_{V_c}(\mathbf{w})|_{\tilde{c}(V) \cap c} = \mathbf{w}_V$ for all $\mathbf{w} \in V_c$. Then, for all $w \in H^2(c)$,*

$$\|w - \mathbf{l}_{V_c}(w)\|_{L^2(c)} \lesssim h_c \left(|w|_{H^1(c)} + h_c |w|_{H^2(c)} \right), \quad (4.44a)$$

and

$$\sum_{e \in E_c} \|w - \mathbf{l}_{V_c}(w)\|_{L^2(\tilde{f}_c(e))} \lesssim h_c^{\frac{1}{2}} \left(|w|_{H^1(c)} + h_c |w|_{H^2(c)} \right). \quad (4.44b)$$

Proof of Theorem 4.21. The bound on the diffusion-related terms derived in Theorem 4.12 still holds. For the advection-related terms, there are two adaptations from the proof of Theorem 3.24. The first one is that we consider the reduction map R_V defined by (7.30) in lieu of $\hat{R}_V$, since we are now bounding the error $u - R_V(u)$. The second adaptation is related to the change in the semi-norm induced by the bilinear form s_β^V owing to the use of Péclet-based upwinding. We bound again the three terms on the right-hand side of Lemma 3.23 for all $\mathbf{w} \in V$ with $\|\mathbf{w}\|_{V,da} = 1$. For the first term, we readily infer using the Cauchy-Schwarz inequality and the interpolation error estimate (4.44a) that

$$|\langle [\mathbf{H}_{\mu-\nabla \cdot \beta}^V, R_V](u), \mathbf{w} \rangle_V | \lesssim \left(\tau^{-\frac{1}{2}} \|\mathbf{w}\|_{V,2} \right) \left(\sum_{c \in C} \tau \|\mu - \nabla \cdot \beta\|_{L^\infty(c)}^2 h_c^2 \left(|u|_{H^1(c)}^2 + h_c^2 |u|_{H^2(c)}^2 \right) \right)^{\frac{1}{2}}.$$

Turning to the second term, we introduce disjoint sets $E_{>1} := \{e \in E \mid \Pi_e + L_\beta \lambda_e^{-1} h_e^2 > 1\}$ and $E_{\leq 1} := \{e \in E \mid \Pi_e + L_\beta \lambda_e^{-1} h_e^2 \leq 1\}$ and we split the summation in the right-hand side of (3.42b) as $\sum_{e \in E_{>1}} (\cdot) + \sum_{e \in E_{\leq 1}} (\cdot)$. Proceeding as in the proof of Theorem (3.24), we infer that

$$\sum_{e \in E_{>1}} (\cdot) \lesssim \left(\sum_{e \in E_{>1}} [\mathbf{w}]_e^2 \int_{\tilde{f}(e)} |\beta \cdot \mathbf{n}_{\tilde{f}(e)}| \right)^{\frac{1}{2}} \left(\sum_{e \in E_{>1}} \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}(e))} \|u - \mathbf{l}_V(u)\|_{L^2(\tilde{f}(e))}^2 \right)^{\frac{1}{2}},$$

so that, owing to the property (ii_π) (which implies that $\kappa_e \beta_e \geq \mathbf{C}_\pi |\beta_e|$ for all $e \in E_{>1}$) and the Proposition 3.27, we infer that

$$\sum_{e \in E_{>1}} (\cdot) \lesssim \left(s_\beta^V(\mathbf{w}, \mathbf{w}) + \mathbf{C}_{V,a} \tau^{-1} \|\mathbf{w}\|_{V,2}^2 \right)^{\frac{1}{2}} \left(\sum_{e \in E_{>1}} \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}(e))} h_e |u|_{H^1+(C_e)}^2 \right)^{\frac{1}{2}},$$

where $|u|_{H^{1+}(C_e)}^2 = \sum_{c \in C_e} |u|_{H^1(c)}^2 + h_c^2 |u|_{H^2(c)}^2$. Furthermore, we observe that

$$\sum_{e \in E_{\leq 1}} (\cdot) \lesssim \left(\sum_{e \in E_{\leq 1}} [w]_e^2 h_e \lambda_e \right)^{\frac{1}{2}} \left(\sum_{e \in E_{\leq 1}} h_e^{-1} \lambda_e^{-1} \sum_{c \in C_e} \|\beta \cdot \mathbf{n}_{\tilde{f}_c(e)}(u - \mathbf{l}_{V_c}(u))\|_{L^1(\tilde{f}_c(e))}^2 \right)^{\frac{1}{2}}, \quad (4.45)$$

and observe that the first factor in the right-hand side of this inequality (4.45) is bounded by $\|\text{GRAD}(\mathbf{w})\|_{\mathcal{E}, \lambda}$ owing to mesh regularity, the definition of λ_e and the Proposition (4.5). Next, still using Proposition 3.27 and mesh regularity, we infer that

$$\|\beta \cdot \mathbf{n}_{\tilde{f}_c(e)}(u - \mathbf{l}_{V_c}(u))\|_{L^1(\tilde{f}_c(e))}^2 \leq \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}_c(e))} \left(|\beta_e| + L_\beta h_e^3 \right) \|u - \mathbf{l}_{V_c}(u)\|_{L^2(\tilde{f}_c(e))}^2, \quad (4.46)$$

so that, combining the interpolation error (4.44b) from Proposition 4.23 with the estimates (4.45) and (4.46) yields

$$\sum_{e \in E_{\leq 1}} (\cdot) \lesssim \|\text{GRAD}(\mathbf{w})\|_{\mathcal{E}, \lambda} \left(\sum_{e \in E_{\leq 1}} h_e^{-1} \lambda_e^{-1} \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}(e))} (|\beta_e| + L_\beta h_e^3) h_e |u|_{H^{1+}(C_e)}^2 \right)^{\frac{1}{2}}.$$

As a result, recalling the definition (4.32) of the Péclet number and owing to mesh regularity, we end up with

$$\sum_{e \in E_{\leq 1}} (\cdot) \lesssim \|\text{GRAD}(\mathbf{w})\|_{\mathcal{E}, \lambda} \left(\sum_{e \in E_{\leq 1}} \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}(e))} \left(|\Pi_e| + L_\beta \lambda_e^{-1} h_e^2 \right) h_e |u|_{H^{1+}(C_e)}^2 \right)^{\frac{1}{2}}.$$

Collecting the bounds on $\sum_{e \in E_{< 1}}$ and on $\sum_{e \in E_{\leq 1}}$ yields

$$|\langle \llbracket J_\beta^\gamma, R \rrbracket(u), \text{GRAD}(\mathbf{w}) \rangle_{\mathcal{V}}| \lesssim \|\mathbf{w}\|_{\mathcal{V}, \text{da}} \left(\sum_{e \in E} \|\beta \cdot \mathbf{n}\|_{L^\infty(\tilde{f}(e))} h_e \min \left(1, |\Pi_e| + \frac{L_\beta h_e^2}{\lambda_e} \right) |u|_{H^{1+}(C_e)}^2 \right)^{\frac{1}{2}}.$$

Since the boundary term is bounded as before, we infer the expected result by exchanging the sum symbols and using mesh regularity. $\square$

4.4 Numerical results

In this section, we investigate the reliability and the efficiency of the discrete problem (4.34) for the diffusion-advection-reaction problem. In particular, we illustrate the advantage of considering a Péclet-based upwinding. Numerical results are presented on two test cases. The first one considers an anisotropic diffusion tensor with a rotating advective field when the solution is expressed as a combination of sine functions. The second one evaluates the numerical behavior of our scheme when the solution exhibits a sharp boundary layer. The diffusion tensor and the advective field are constant. We specifically study the influence of the Péclet number by varying the value of the diffusion. Another specificity is that the mesh sequence is well-adapted to capture the boundary layer.

Remark 4.24 (Pure diffusion problem and weak boundary conditions). *We do not present test cases evaluating the performance of the scheme (4.19) for the pure diffusion problem since we numerically observed, when ν_0 is large enough, that the solution coincides (up to an error of order 10^{-16} corresponding to the machine precision) with the solution of the discrete problem proposed by Bonelle & Ern (2014a), when the boundary conditions are strongly enforced. In addition, the computation of the constant $\mathcal{C}_N$ in Proposition 4.6 gives*

$$\mathcal{C}_N = \max_{c \in C} \max_{f \in F_c} \max_{e \in E_f} \left(h_c \frac{|\mathbf{c}_{e,c} \cap f|}{|\mathbf{c}_{e,c}|} \right). \quad (4.47)$$

On hexahedral mesh sequences, this quantity is equals to $3\sqrt{3}/2$, so that choosing $\nu_0 = 10$ owing to Lemma 4.7 is sufficient. However in practice, we choose $\nu_0 = 500$, so that the scheme is stable on all mesh sequences.

Following the analysis proposed in Sections 4.2.2 and 4.3.2, the convergence study is performed using the following relative discrete L^2 -error:

$$\mathbf{Er}_V(u) := \left(\frac{\sum_{v \in V} |\tilde{c}(v)| \left(u_v - R_V(u)|_v \right)^2}{\sum_{v \in V} |\tilde{c}(v)| R_V(u)|_v^2} \right)^{\frac{1}{2}}, \quad (4.48)$$

with u the exact solution of the continuous problem (3.12) and u the solution of the discrete problem (4.34). Compared to the discrete error $\widehat{\mathbf{Er}}_V(u)$ defined by (3.43), the discrete relative error now relies on the point-wise evaluation of the exact solution at all mesh vertices, using the de Rham reduction map R_V .

4.4.1 Test case 3. Anisotropic diffusion tensor and rotating advective field

In this first test case, the computations are run on the unit cube $\Omega = [0, 1]^3$ using the three-dimensional mesh sequences presented in Section 3.3, denoted by H, PrT, PrG and CB and depicted in Figure 4.3.

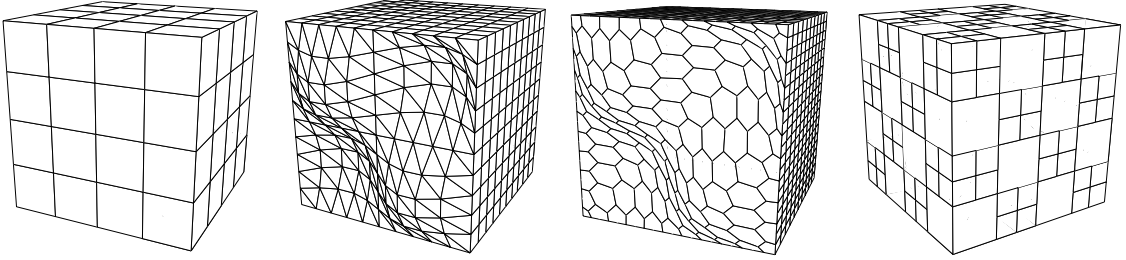


Figure 4.3 – The four mesh sequences H, PrT, PrG and CB, respectively.

The diffusion tensor λ and the advective field are given by

$$\lambda = \begin{pmatrix} 1 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} y - 1/2 \\ 1/2 - x \\ -z \end{pmatrix}, \quad (4.49)$$

where the former corresponds to the diffusion tensor of anisotropic ratio 3 (see the FVCA benchmark Eymard *et al.* (2011)), and the latter is adapted from the first test case presented in Section 3.3.2, replacing $z + 1$ with $-z$. The reaction coefficient μ is set to 0 so that the scalar-valued Friedrichs tensor is $\sigma_{\beta, \mu} = \frac{1}{2}$. As a result, the discrete problem (4.34) is well-posed owing to Lemma 4.17. The exact solution is

$$u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi\left(y + \frac{1}{2}\right)\right) \sin\left(\pi\left(z + \frac{1}{3}\right)\right). \quad (4.50)$$

Hereafter, we compare the two schemes when the discrete contraction operators J_β^ε is defined using either the full upwinding function by choosing $\pi(x) = \text{sign}(x)$ or the Sharfetter-Gummel upwinding function by choosing $\pi(x) = \coth(\frac{x}{2}) - \frac{2}{x}$.

Remark 4.25 (Samariskii map). *Other Péclet-based upwinding strategies have also been implemented, for example using the Samariskii map defined by $\pi(x) = e^x - 1$ if $x < 0$ and $\pi(x) = 1 - e^{-x}$ otherwise. However, these results are not presented here since they are very close to those obtained with the Sharfetter-Gummel map.*

The discrete relative error $\mathbf{Er}_V(u)$ is represented in Figure 4.4 with respect to $\#V$ on the left panel and with respect to the computational cost $\text{Co} = \text{nnz} \times n_{\text{ite}}$ on the right panel. The corresponding legend is collected in Table 4.2.

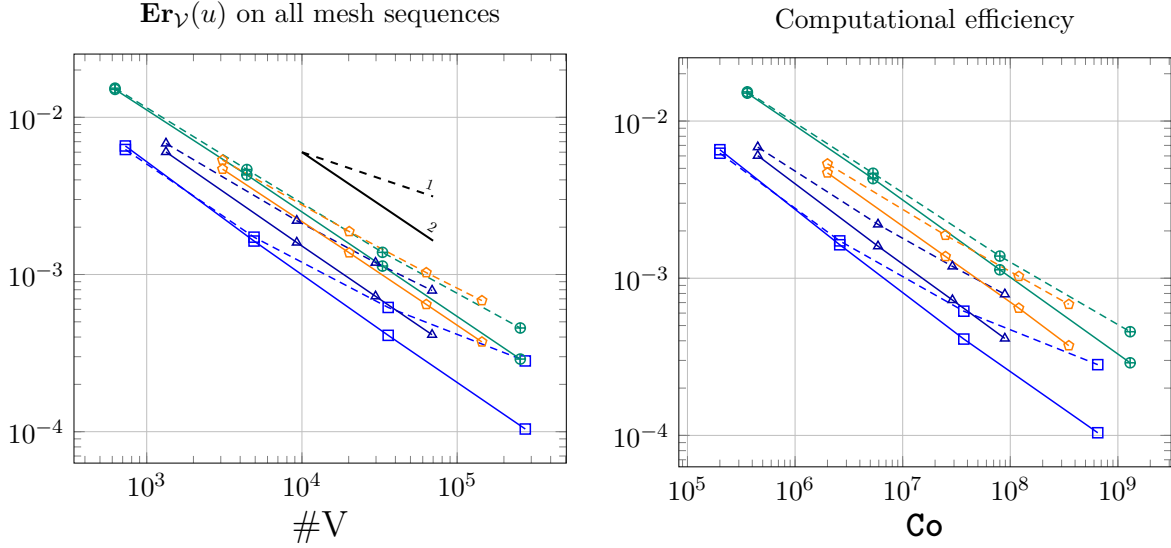


Figure 4.4 – Test case 3. Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost Co .

Upwind				Sharfetter-Gummel			
H	PrT	PrG	CB	H	PrT	PrG	CB
$\text{---}\square\text{---}$	$\text{---}\triangle\text{---}$	$\text{---}\circ\text{---}$	$\text{---}\oplus\text{---}$	$\text{---}\square\text{---}$	$\text{---}\triangle\text{---}$	$\text{---}\circ\text{---}$	$\text{---}\oplus\text{---}$

Table 4.2 – Test case 3: Legend of Figure 4.4.

Accuracy and efficiency. These numerical results are in agreement with the theoretical results derived in Section 4.3.2. In essence, the convergence rates are of order 1 on the finer mesh when we use the full upwinding strategy and approximately of order 2 when we use a Péclet-dependent upwinding using the Sharfetter-Gummel map. In particular, this figure illustrates that adapting the quantity of upwinding with respect to the Péclet-number significantly reduces the discrete relative error $\mathbf{Er}_V(u)$ than using the classical full upwinding strategy. Indeed, observing that the ratio of advective effects to diffusive effects is approximately of order 1, so that the local Péclet number is of order h , the use of Péclet-dependent upwinding reduces the amount of artificial stabilization introduced to stabilize the scheme as the mesh size goes to 0. These observations are particularly verified for the H and on the PrG mesh sequences. Finally, let us mention that the discrete minimum/maximum principle is satisfied on all meshes.

4.4.2 Test case 4. Boundary layer on graded meshes

This test case considers the numerical behavior of the scheme (4.34) on the domain $\Omega = [0, 2L] \times [-1, 1] \times [0, L]$ with $L = 3.2$ when the numerical solution is given by

$$u(2Lx, y, Lz) = x^2 z \left(y^2 - e^{\frac{y-1}{\lambda}} \right), \quad (4.51)$$

where the diffusion tensor is constant and isotropic of magnitude $\lambda > 0$: $\boldsymbol{\lambda} = \lambda \text{Id}$. This solution presents a boundary layer of length $\mathcal{O}(\lambda)$ in the plan $\{\mathbf{x} \in \partial\Omega \mid y = 1\}$ when the magnitude of diffusion goes to 0. Then, we examine the approximation of the solution u when $\lambda \in \{1, 10^{-1}, 10^{-2}, 10^{-4}\}$. Figure 4.5 illustrates the function $(x, y) \in [0, 1] \times [-1, 1] \mapsto u(2Lx, y, L)$ with respect to λ .

The advection field and the reaction coefficient are constant and are given by $\boldsymbol{\beta} = (0, 1, 0)^T$ and $\mu = 0$. Owing to Lemma 4.18, the discrete problem is well-posed for all mesh-size $h > 0$. For this test case, we consider a mesh sequence that is refined in the y direction when $y = \pm 1$,

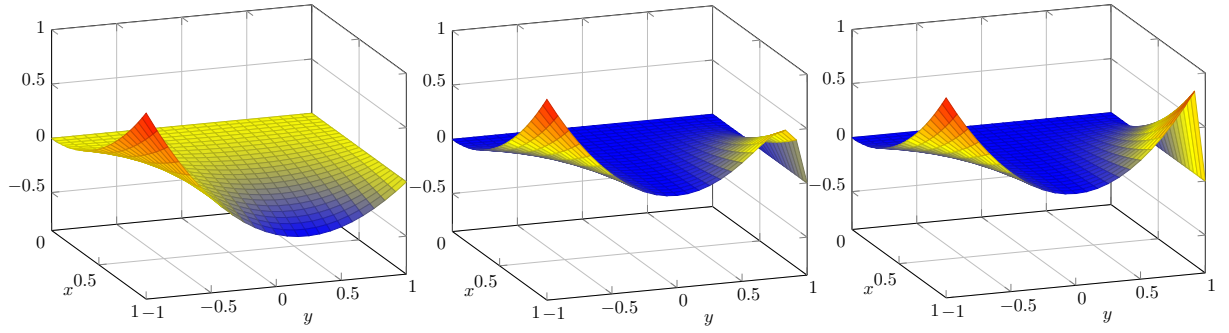


Figure 4.5 – Test case 4. From the left to the right, exact solution $(x, y) \in [0, 1] \times [-1, 1] \mapsto u(2Lx, y, L)$ for $\lambda = 1$, $\lambda = 10^{-1}$ and $\lambda = 10^{-2}$.

accordingly to the boundary layer (see Figure 4.6). The discrete relative error $\mathbf{Er}_V(u)$ is

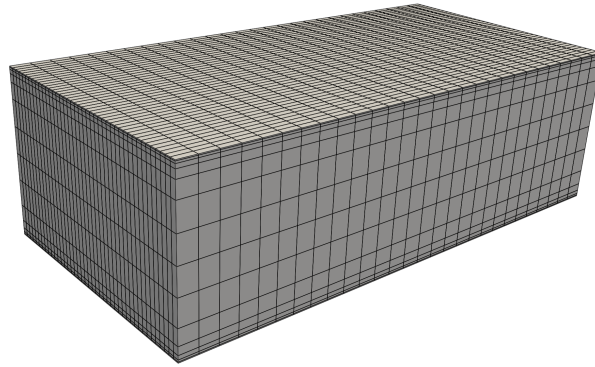


Figure 4.6 – Test case 4. Graded mesh of the domain $\Omega = [0, 2L] \times [-1, 1] \times [0, L]$ refined for $y = \pm 1$.

represented on this mesh sequence with respect to $\lambda \in \{1, 10^{-1}, 10^{-2}, 10^{-4}\}$ using the full upwinding and the Sharfetter-Gummel upwinding approaches.

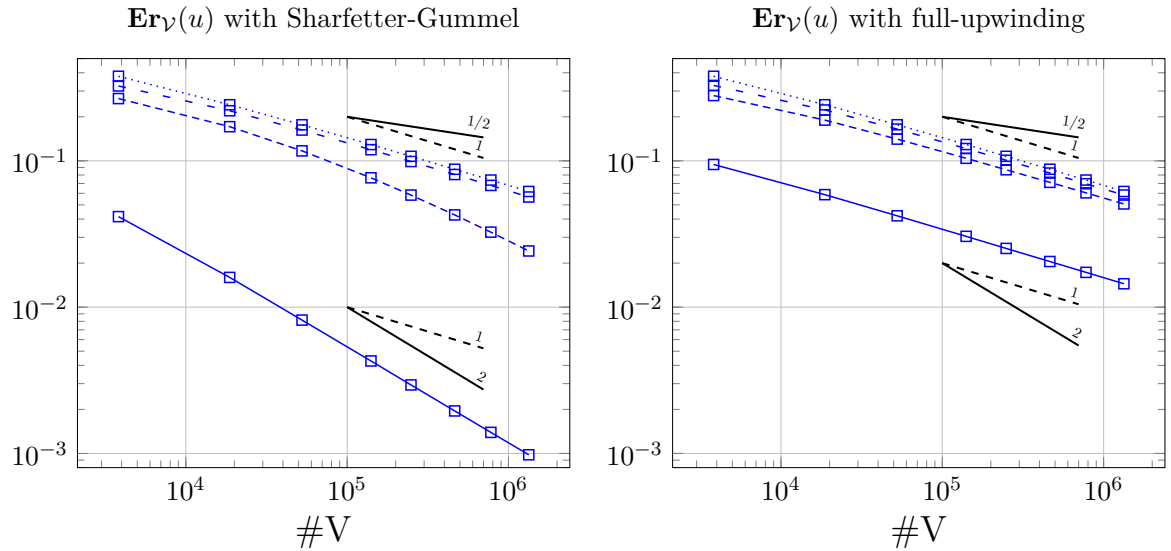


Figure 4.7 – Test case 4. $\mathbf{Er}_V(u)$ using the Sharfetter-Gummel upwinding (on the left) and the full-upwinding (on the right), for $\lambda = 1$ (—□—), $\lambda = 10^{-1}$ (- -□- -), $\lambda = 10^{-2}$ (··□··) and $\lambda = 10^{-4}$ (- ·□· -).

Accuracy. When $\lambda = 1$, the diffusive and the advective effects are of the same order and the numerical results depicted in the left panel of Figure 4.7 using the Sharfetter-Gummel upwinding show that the discrete error $\mathbf{Er}_V(u)$ decreases at order 2 as the mesh size goes to 0. On the right panel, we observe that the use of full-upwinding leads to an upper discrete error (almost of a factor 10) and reduces the convergence rate to 1. When the magnitude λ reduces, the numerical results presented in the left panel are in agreement with our theoretical results since we observe that the scheme converges at the order 1 instead of 0.5, even when $\lambda = 10^{-4}$. Obviously, in the advection dominant regime, the choice between the Sharfetter-Gummel upwinding or the full upwinding strategy is less important. We also observe that the presence of the boundary layer directly impacts the accuracy of our scheme; reducing of one order the diffusion tensor almost increases the discrete relative error of two orders on the finer mesh.

Discrete min./max. principle. Finally, Table 5.4 collects the results on the discrete minimum/maximum principle for this test case. We write Y if this principle is satisfied on all meshes of the sequence. If not, we indicate how much this principle is not satisfied on the coarsest mesh with respect to the exact solution.

λ	Full upwind		Sharfetter-Gummel	
	Min	Max	Min	Max
1	Y	Y	Y	Y
10^{-1}	Y	Y	Y	Y
10^{-2}	$0.003\% <$	Y	$0.05\% <$	Y
10^{-4}	$0.4\% <$	Y	$0.4\% <$	Y

Table 4.3 – Test case 4. Discrete minimum/maximum principle with respect to λ .

When the diffusion and the advection effects are approximately of the same order (i.e., $\lambda = 1$ and $\lambda = 10^{-1}$), the minimum/maximum principle is satisfied for both full-upwind and Sharfetter-Gummel upwind strategies. However, when the diffusion effects diminish, we observe that the minimal bounds are not exactly satisfied, due to the presence of the boundary layer (see Figure 4.5). Numerically, we observe that refining the mesh significantly diminishes the violation of the minimal bound.

Chapter 5

Improved vertex-based scheme for scalar transport

Contents

5.1	Discrete setting	68
5.1.1	Primal mesh and degrees of freedom	68
5.1.2	Edge-face-based partition	69
5.2	Advection-reaction problem	69
5.2.1	Discrete problem and piece-wise affine reconstruction	70
5.2.2	Analysis	72
5.3	Diffusion-advection-reaction problem	77
5.3.1	Discrete problem	78
5.3.2	Analysis	78
5.4	Numerical results	79
5.4.1	Implementation aspects	79
5.4.2	Computational settings	81
5.4.3	Test case 1. Rotating advective field	81
5.4.4	Test case 2. Sharp internal layer	84
5.4.5	Test case 3. Anisotropic diffusion tensor and rotating advective field	85

This chapter extends the vertex-based scheme presented in Chapter 3 to approximate the continuous problem

$$\beta \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (5.1a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega^-. \quad (5.1b)$$

The main idea to devise our scheme consists of introducing, in addition to the vertex-based degrees of freedom (dofs), one dof per mesh cell. These additional cell-based dofs are then eliminated locally using a Schur complement technique (also called static condensation). This step entails modest marginal costs since no matrix inversion is required. Indeed, we will see that dofs attached to mesh cells are only coupled with vertex-based dofs while remaining uncoupled from each other. At the end, the size of the final system matrix to invert is just the number of mesh vertices, as the linear system resulting from the scheme presented in Chapter 3.

The first main ingredient in this chapter is the use of a polyhedral reconstruction map that defines piece-wise linear polynomials from dofs attached to mesh vertices and mesh cells. This map is defined from a simple simplicial partition of the mesh cells. The second important ingredient is the stabilization technique. The crucial point is to devise a local stabilization term so that it does not hamper the possibility to eliminate locally the cell-based dofs. We achieve this point by adapting to our scheme the Continuous Interior Penalty (CIP) stabilization technique,

first introduced by Burman & Hansbo (2004) and by Burman (2005). Compared to these references, the major modification is that we only penalize the advective derivative across intra-cell sub-faces resulting from the simplicial partition of the mesh cells. As observed in the different context related to composite finite elements by Burman & Schieweck (2016), the CIP technique with this type of modification still provides enough stabilization. Compared to the stability analysis proposed in Chapter 3, the stability of the scheme is now inferred from an inf-sup argument instead of a coercivity argument when the Friedrichs tensor takes positive values. Hence, similarly to dG methods (see Ern & Guermond (2006a)), we additionally control the weighted advective derivative of the discrete solution.

To the best of our knowledge, there does not exist in the literature any vertex-based scheme of order $O(h^{\frac{3}{2}})$ on polyhedral meshes for the problem (5.1). In addition to the previously mentioned CIP method, many examples of stabilized $\mathbb{P}_1$ Lagrange finite elements can be found on matching simplicial meshes. Among various choices for the stabilization, we mention the Streamline Diffusion method by Johnson *et al.* (1984), the Subgrid Viscosity method by Guermond (1999, 2001), the Local Projection Stabilization method by Becker & Braack (2001) (see also Braack *et al.* (2007); Matthies *et al.* (2007, 2008)).

Since we consider dofs attached to mesh vertices and mesh cells, an advantage of the present scheme for the advection-reaction problem is that it can be combined with the recent Vertex Approximate Gradient (VAG) schemes, introduced by Eymard *et al.* (2012a) for diffusion problems in porous media and extended by Eymard *et al.* (2012b, 2014) for multiphase flows. The following approach would then provide a higher-order alternative to the more usual finite-volume treatment of advective terms based on standard upwinding.

This chapter is organized as follows. First, we present in Section 5.2 the approximation of the pure advection-reaction problem (5.1). Then, we extend the scheme so as to approximate the solution of the diffusion-advection-reaction problem in Section 5.3. Finally, we address some implementation aspects and present numerical results on the three-dimensional test cases already considered in Chapters 3 and 4.

The content of this chapter is an extended version of the paper "*A vertex-based scheme on polyhedral meshes for advection-reaction equations with sub-mesh stabilization*", by P. Cantin, J. Bonelle, E. Burman & A. Ern, to appear in *Computers and Mathematics with Applications*, 2016.

5.1 Discrete setting

This first section introduces and recalls some notation underlying the discrete setting. The reader can refer to Chapter 7 for further insights.

5.1.1 Primal mesh and degrees of freedom

Let Ω denote an open, bounded, connected, polyhedral subset of $\mathbb{R}^3$ and let $M = \{V, E, F, C\}$ be a polyhedral mesh of Ω , composed of polyhedral cells $c \in C$, polygonal faces $f \in F$, straight edges $e \in E$, and vertices $v \in V$. Boundary faces are collected in the set $F^\partial = \{f \in F \mid f \subset \partial\Omega\}$. In this chapter, we consider families of polyhedral meshes satisfying the mesh assumptions **(C)**, **(St)** and **(Sh)** (see Chapter 3), namely, the mesh M defines a cellular complex, all mesh entities are star-shaped with respect to their barycenter and there exists a simplicial sub-mesh that is shape-regular in the usual sense of Ciarlet (1978). We denote $\mathbf{x}_v$, $\mathbf{x}_e$, $\mathbf{x}_f$, and $\mathbf{x}_c$ the barycenters of $v \in V$, $e \in E$, $f \in F$, and $c \in C$, respectively.

Our scheme is formulated using dofs attached to mesh vertices and mesh cells. These dofs are collected in the finite-dimensional space $\mathcal{P} = \mathcal{V} \times \mathcal{C}$, where $\mathcal{V}$ collects scalar dofs attached to mesh vertices and where $\mathcal{C}$ collects scalar dofs attached to mesh cells. For all $w \in \mathcal{P}$, w_v denotes the entry attached to $v \in V$ and w_c denotes the entry attached to $c \in C$. For all $c \in C$, $\mathcal{P}_c = \mathcal{V}_c \times \mathcal{C}_c$ denotes the local dofs sub-space attached to the set $V_c \times \{c\}$.

In this chapter, we only consider the reduction map $R_{\mathcal{P}} : W^{s,p}(\Omega) \rightarrow \mathcal{P}$ with $sp > 3$ acting as follows:

$$\forall v \in V, \quad R_{\mathcal{P}}(w)|_v := w(\mathbf{x}_v) \text{ and } \forall c \in C, \quad R_{\mathcal{P}}(w)|_c := w(\mathbf{x}_c), \quad \forall w \in W^{s,p}(\Omega). \quad (5.2)$$

The local reduction map is denoted by $R_{\mathcal{P}_c} : W^{s,p}(c) \rightarrow \mathcal{P}_c$ and acts as the global reduction map. The reduction map $R_{\mathcal{P}}$ generalizes the classical de Rham map $R_{\mathcal{V}}$ defined by (4.5a) to the dof space $\mathcal{P}$.

5.1.2 Edge-face-based partition

In addition to the primal mesh M , our scheme also considers a simplicial sub-mesh. For all $c \in C$, $\mathfrak{C}_{\text{EF},c}$ denotes the simplicial partition of c composed of the simplices $\{\mathbf{c}_{\text{ef},c}\}_{(e,f) \in \text{EF}_c}$ with $\text{EF}_c = \{(e,f) \in E \times F \mid e \subset \partial f\}$, such that

$$\mathbf{c}_{\text{ef},c} := \text{int} \left(\bigcup_{v \in V_e} \overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\} \right), \quad \forall (e,f) \in \text{EF}_c, \quad (5.3)$$

where we recall that $\text{int}(\omega)$ denotes the interior of any set $\omega \subset \mathbb{R}^3$ and that $\overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$ denotes the closed convex hull of the 4-uplet $\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$, which defines also a simplex (see the left panel of Figure 5.1). Observe that $\#\mathfrak{C}_{\text{EF},c} = 2\#E_c$ since each mesh edge is shared by two mesh faces. An important notion to devise our scheme is the use of *intra-cell sub-faces*, defined as the sub-faces of the simplicial sub-mesh $\mathfrak{C}_{\text{EF},c}$ that are contained in the cell c . This sub-faces are collected in the set $\mathfrak{F}_{\text{EF},c}$ which is defined as

$$\begin{aligned} \mathfrak{F}_{\text{EF},c} = \{ & \mathbf{f}_{\text{vf},c} = \text{int}(\partial \mathbf{c}_{\text{ef},c} \cap \partial \mathbf{c}_{\text{e'f},c}) \mid f \in F_c, v \in V_f \text{ and } e, e' \in E_f \cap E_v\} \\ & \cup \{ \mathbf{f}_{\text{e},c} = \text{int}(\partial \mathbf{c}_{\text{ef},c} \cap \partial \mathbf{c}_{\text{ef'},c}) \mid e \in E_c, f, f' \in F_e \cap F_c \}. \end{aligned} \quad (5.4)$$

As depicted in the right panel of Figure 5.1, we observe that this set collects two kind of sub-faces: those that are shared by two simplices attached to one edge $e \in E_c$, denoted by $\mathbf{f}_{\text{e},c}$, and those that are shared by two simplices connected by one vertex $v \in V_f$ for all $f \in F_c$, denoted by $\mathbf{f}_{\text{vf},c}$.

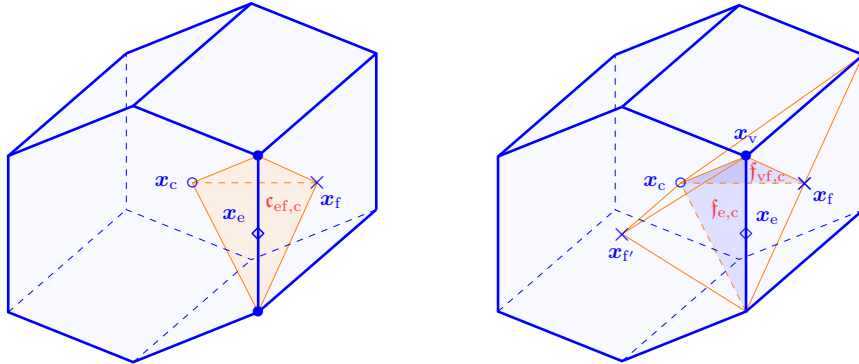


Figure 5.1 – Left panel: the sub-cell $\mathbf{c}_{\text{ef},c} \in \mathfrak{C}_{\text{EF},c}$ attached to $f \in F_c$ and $e \in E_f$. Right panel: two intra-cell sub-faces $\mathbf{f}_{\text{e},c} \in \mathfrak{F}_{\text{EF},c}$ with $e \in E_c$ and $\mathbf{f}_{\text{vf},c} \in \mathfrak{F}_{\text{EF},c}$ with $f \in F_c$ and $v \in V_f$.

5.2 Advection-reaction problem

This section focuses on the approximation of the solution of the scalar advection-reaction problem

$$\beta \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (5.5a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega^-, \quad (5.5b)$$

where we assume that $s \in L^2(\Omega)$ and $u_D \in H^{1+\epsilon}(\partial\Omega)$ with $\epsilon > 0$. We assume as before that the advective field $\beta : \Omega \rightarrow \mathbb{R}^3$ is Lipschitz-continuous and that the reaction coefficient $\mu : \Omega \rightarrow \mathbb{R}$ is bounded in Ω . Denoting $\sigma_{\beta,\mu} : \Omega \rightarrow \mathbb{R}$ the Friedrichs tensor defined by

$$\sigma_{\beta,\mu} = \mu - \frac{1}{2} \nabla \cdot \beta, \quad (5.6)$$

we assume that the following positivity assumption holds:

$$(\mathcal{H}) \quad \text{ess inf}_{\Omega} \sigma_{\beta,\mu} > 0. \quad \text{We define the reference time } \tau = (\text{ess inf}_{\Omega} \sigma_{\beta,\mu})^{-1}.$$

Owing to Theorem (2.7), the continuous problem (5.5) is well-posed.

5.2.1 Discrete problem and piece-wise affine reconstruction

The discrete problem approximating (5.5) hinges on the bilinear map $A_{\beta,\mu}^{\mathcal{P}} : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ defined as

$$A_{\beta,\mu}^{\mathcal{P}}(v, w) = \sum_{c \in \mathcal{C}} A_{\beta,\mu;c}^{\mathcal{P}}(v, w) + \sum_{f \in F^{\partial}} A_{(\beta \cdot n)^{\ominus};f}^{\mathcal{P},\partial}(v, w), \quad \forall v, w \in \mathcal{P}. \quad (5.7)$$

The first term on the right hand-side approximates (5.5a) on every mesh cell and the second term weakly enforces the boundary condition on every boundary face $f \in F^{\partial}$ lying in the sub-set $\partial\Omega^- \subset \partial\Omega$.

Bilinear forms and reconstruction map. Let $c \in \mathcal{C}$. The local bilinear form $A_{\beta,\mu;c}^{\mathcal{P}} : \mathcal{P}_c \times \mathcal{P}_c \rightarrow \mathbb{R}$ is composed of two bilinear forms, also defined on $\mathcal{P}_c \times \mathcal{P}_c$:

$$A_{\beta,\mu;c}^{\mathcal{P}}(v, w) = g_{\beta,\mu;c}(v, w) + \gamma s_{\beta;c}(v, w), \quad \forall v, w \in \mathcal{P}_c, \quad (5.8)$$

where $\gamma \in (\gamma_b, 1]$ is stabilization parameter with $\gamma_b > 0$. The first map $g_{\beta,\mu;c}$ is devised as the Galerkin approximation of (5.5a) using the reconstruction map $L_{\mathcal{P}_c} : \mathcal{P}_c \rightarrow \mathbb{P}_1(\mathfrak{C}_{\text{EF},c}; \mathbb{R}) \cap \mathcal{C}^0(c)$ as follows:

$$g_{\beta,\mu;c}(v, w) := \int_c \beta \cdot \nabla L_{\mathcal{P}_c}(v) L_{\mathcal{P}_c}(w) + \int_c \mu L_{\mathcal{P}_c}(v) L_{\mathcal{P}_c}(w), \quad \forall v, w \in \mathcal{P}_c. \quad (5.9)$$

The reconstruction map $L_{\mathcal{P}_c}$ is defined as

$$\forall x \in c, \quad L_{\mathcal{P}_c}(w)(x) := \sum_{v \in V_c} w_v \ell_{v,c}(x) + w_c \ell_c(x), \quad \forall w \in \mathcal{P}_c, \quad (5.10)$$

where the local shape functions $((\ell_{v,c})_{v \in V_c}, \ell_c)$ are obtained by combining the usual Courant (or $\mathbb{P}_1$ -Lagrange) basis functions $((\theta_v)_{v \in V_c}, (\theta_f)_{f \in F_c}, \theta_c)$ associated with the simplicial partition $\mathfrak{C}_{\text{EF},c} = \{\mathfrak{c}_{\text{ef},c}\}_{f \in F_c, e \in E_f}$ defined by (5.3). We define

$$v \in V_c, \quad \ell_{v,c} := \theta_v + \sum_{f \in F_v} \frac{|f \cap \tilde{c}(v)|}{|f|} \theta_f, \quad \text{and } \ell_c := \theta_c, \quad (5.11)$$

where we recall that $\tilde{c}(v)$ denotes the dual mesh cell attached to $v \in V$.

The local stabilization bilinear form $s_{\beta;c}$ penalizes jumps of the advective derivative of the reconstruction map $L_{\mathcal{P}_c}$ across the internal sub-faces of c induced by the sub-mesh $\mathfrak{C}_{\text{EF},c}$. We define

$$s_{\beta;c}(v, w) = h_c^2 |\beta_c|^{-1} \sum_{f \in \mathfrak{F}_{\text{EF};c}} \int_f (\beta_c \cdot \nabla L_{\mathcal{P}_c}(v)) (\beta_c \cdot \nabla L_{\mathcal{P}_c}(w)), \quad \forall v, w \in \mathcal{P}_c, \quad (5.12)$$

with $\beta_c = \beta(x_c)$.

Finally, the boundary condition (5.5b) is weakly enforced by means of the local bilinear maps $A_{(\beta \cdot n)^{\ominus};f}^{\mathcal{P},\partial} : \mathcal{P}_c \rightarrow \mathcal{P}_c$ on every boundary face $f \in F^{\partial}$, where $c \equiv c(f)$ denotes the unique cell containing the boundary face $f \in F^{\partial}$. The map $A_{(\beta \cdot n)^{\ominus};f}^{\mathcal{P},\partial}$ is defined by

$$A_{(\beta \cdot n)^{\ominus};f}^{\mathcal{P},\partial}(v, w) = \int_f (\beta \cdot n)^{\ominus} L_{\mathcal{P}_c}(v) L_{\mathcal{P}_c}(w), \quad \forall v, w \in \mathcal{P}_c. \quad (5.13)$$

Remark 5.1 (Hodge operator). *Extending the CDO framework proposed in Section 3.1.3, one may equip the dof space $\mathcal{P}_c$ with the discrete inner product*

$$\langle\langle \mathbf{v}, \mathbf{w} \rangle\rangle_{\mathcal{P}_c} = \mathbf{v}_c \mathbf{w}_c + \sum_{\mathbf{v} \in V_c} \mathbf{v}_v \mathbf{w}_v, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{P}_c, \quad (5.14)$$

so that the rightmost term of (5.9) can be represented as $\langle\langle \mathbf{H}_{\mu;c}^{\mathcal{P}}(\mathbf{v}), \mathbf{w} \rangle\rangle_{\mathcal{P}_c}$, where the entries of $\mathbf{H}_{\mu;c}^{\mathcal{P}} : \mathcal{P}_c \rightarrow \mathcal{P}_c$ are such that

$$\forall \mathbf{v}, \mathbf{v}' \in V_c, \quad (\mathbf{H}_{\mu;c}^{\mathcal{P}})_{\mathbf{v}, \mathbf{v}'} = \int_c \mu \ell_{\mathbf{v},c} \ell_{\mathbf{v}',c}, \quad (\mathbf{H}_{\mu;c}^{\mathcal{P}})_{\mathbf{v},c} = (\mathbf{H}_{\mu;c}^{\mathcal{P}})_{c,\mathbf{v}} = \int_c \mu \ell_{\mathbf{v},c} \ell_c, \quad (5.15)$$

and $(\mathbf{H}_{\mu;c}^{\mathcal{P}})_{c,c} = \int_c \mu \ell_c \ell_c$. Moreover, using the stability of $\mathcal{L}_{\mathcal{P}_c}$ from Proposition 5.4, this operator defines a Hodge operator if $\mu > 0$ a.e. in c . Similarly, the map (5.13) can be represented as a boundary Hodge operator $\mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus; \mathbf{f}}^{\mathcal{P}, \partial} : \mathcal{P}_c \rightarrow \mathcal{P}_c$ defined as $\langle\langle \mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus; \mathbf{f}}^{\mathcal{P}, \partial} \mathbf{v}, \mathbf{w} \rangle\rangle_{\mathcal{P}_c} = \int_{\mathbf{f}} (\beta \cdot \mathbf{n})^\ominus \mathcal{L}_{\mathcal{P}_c}(\mathbf{v}) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w})$ for all $\mathbf{v}, \mathbf{w} \in \mathcal{P}_c$ with $c = c(\mathbf{f})$.

Discrete problem. The scheme approximating (5.5) reads:

$$\text{Find } \mathbf{u} \in \mathcal{P} \text{ s.t. } \mathbf{A}_{\beta, \mu}^{\mathcal{P}}(\mathbf{u}, \mathbf{v}) = \mathbb{S}(s, u_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{P}, \quad (5.16)$$

with the right-hand side linear form $\mathbb{S}(s, u_D; \cdot) : \mathcal{P} \rightarrow \mathbb{R}$ defined as

$$\mathbb{S}(s, u_D; \mathbf{v}) = \sum_{c \in \mathcal{C}} \int_c s \mathcal{L}_{\mathcal{P}_c}(\mathbf{v}) + \sum_{\mathbf{f} \in \mathcal{F}^\partial} \int_{\mathbf{f}} (\beta \cdot \mathbf{n})^\ominus u_D \mathcal{L}_{\mathcal{P}_{c(\mathbf{f})}}(\mathbf{v}). \quad (5.17)$$

Alternative definitions of the source term resulting from the use of quadratures are discussed in Section 5.4.1.

Example 5.2 (Simplicial mesh). *Let us briefly discuss our scheme in the case of a simplicial primal mesh where each mesh cell c is a tetrahedron, see Figure 5.2. In our scheme, the tetrahedron c is then subdivided into 12 sub-tetrahedra composing the set $\mathfrak{C}_{\text{EF},c}$, and there are 6 internal sub-faces in the set $\mathfrak{F}_{\text{EF},c}$ where the jump of the advective derivative is penalized; instead, this jump is not penalized on the four faces of c . Obviously, standard stabilized finite element methods can be used on simplicial meshes as well. However, the present scheme leads to a smaller stencil and add less stabilization.*

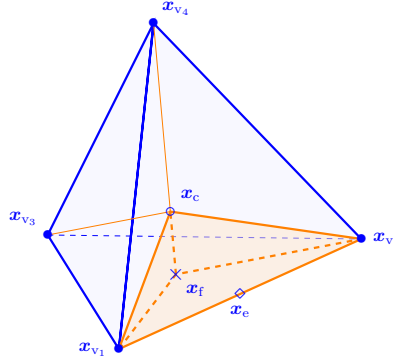


Figure 5.2 – Simplicial partition $\mathfrak{C}_{\text{EF},c}$ on a tetrahedron c .

Remark 5.3 (Vertex-based scheme of order $\mathcal{O}(h^{\frac{3}{2}})$ in the spirit of dG methods). *In order to devise a vertex-based scheme of order $\frac{3}{2}$, one may consider other reconstruction maps considering directly degrees of freedom in $\mathcal{V}_c$ so as to avoid the static condensation of the additional cell-based unknowns. For example, it is possible to consider the reconstruction map $\mathcal{L}_{\mathcal{V}_c}^1 : \mathcal{V}_c \rightarrow \mathbb{P}_1(c)$ defined as*

$$\forall \mathbf{x} \in c, \quad \mathcal{L}_{\mathcal{V}_c}^1(\mathbf{w})(\mathbf{x}) = \sum_{\mathbf{v} \in V_c} \frac{|\tilde{c}(\mathbf{v}) \cap c|}{|c|} \mathbf{w}_v + (\mathbf{x} - \mathbf{x}_c) \cdot \left(\frac{1}{|c|} \int_c \mathcal{L}_{\mathcal{E}_c}(\text{GRAD}(\mathbf{w})) \right), \quad \forall \mathbf{w} \in \mathcal{V}_c, \quad (5.18)$$

where $\mathbf{L}_{\mathcal{E}_c} : \mathcal{E}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{E,c})$ is the DGA reconstruction map considered in Section 4.2 and GRAD is the discrete gradient operator defined by (3.5a). One can prove that the interpolation map $\mathbf{L}_{\mathcal{V}_c}^1 \circ \mathbf{R}_{\mathcal{V}_c}$ exactly preserves piece-wise affine polynomials in $\mathbb{P}_1(\mathcal{C})$. However, since this reconstruction map jumps across mesh faces $\mathbf{f} \in \mathbf{F}$, it is necessary to add a consistency and a penalty term attached to these faces (in the spirit of dG schemes) so as to preserve the stability of the bilinear form. Compared to our scheme (5.16), the final system matrix will have more entries since one vertex belonging to a given cell will be connected to all the vertices belonging to the cells touching this cell. In other words, the matrix line associated with the vertex $\mathbf{v} \in \mathbf{V}$ in the final system will contain at the most

$$\#\{\mathbf{v}' \in \mathbf{V} \mid \exists \mathbf{c} \in \mathcal{C}_{\mathbf{v}} \text{ and } \exists \mathbf{c}' \in \mathcal{C}_{\mathbf{v}'} \text{ s.t. } \partial \mathbf{c} \cap \partial \mathbf{c}' \in \mathbf{F}\}, \quad (5.19)$$

i.e., 125 entries for a hexahedral mesh. Here, owing to the additional cell-based unknowns, we significantly contain the stencil (see Section 5.4.1), e.g., 27 entries for a hexahedral mesh.

5.2.2 Analysis

This section analyzes the discrete problem (5.16). For all $\mathbf{c} \in \mathcal{C}$, the discrete 2-norm on the local dof space $\mathcal{P}_c$ is defined as

$$\|\mathbf{w}\|_{\mathcal{P}_c,2}^2 = h_c^3 |\mathbf{w}_c|^2 + h_c^3 \sum_{\mathbf{v} \in \mathcal{V}_c} |\mathbf{w}_{\mathbf{v}}|^2, \quad \forall \mathbf{w} \in \mathcal{P}_c, \quad (5.20)$$

and we denote $\|\mathbf{w}\|_{\mathcal{P},2}^2 = \sum_{\mathbf{c} \in \mathcal{C}} \|\mathbf{w}\|_{\mathcal{P}_c,2}^2$ for all $\mathbf{w} \in \mathcal{P}$ the discrete 2-norm on $\mathcal{P}$.

Stability. The coercivity of the bilinear form (5.7) is established using the following discrete norm defined on $\mathcal{P}$:

$$\|\mathbf{w}\|_{\mathcal{P},a}^2 := \sum_{\mathbf{c} \in \mathcal{C}} \|\mathbf{w}\|_{\mathcal{P}_c,a}^2 \text{ with } \|\mathbf{w}\|_{\mathcal{P}_c,a}^2 := \tau^{-1} \|\mathbf{w}\|_{\mathcal{P}_c,2}^2 + \gamma_b \mathbf{s}_{\beta;c}(\mathbf{w}, \mathbf{w}) + \sum_{\mathbf{f} \in \mathbf{F}^\partial \cap \mathbf{F}_c} \mathbf{A}_{|\beta \cdot \mathbf{n}|;\mathbf{f}}^{\mathcal{P},\partial}(\mathbf{w}, \mathbf{w}), \quad (5.21)$$

for all $\mathbf{w} \in \mathcal{P}$, where the reference time τ is defined by assumption $(\mathcal{H})$, $\mathbf{s}_{\beta;c}$ by (5.12) and where $\mathbf{A}_{|\beta \cdot \mathbf{n}|;\mathbf{f}}^{\mathcal{P},\partial}$ is defined by (5.13), replacing $(\beta \cdot \mathbf{n})^\ominus$ by $|\beta \cdot \mathbf{n}|$. An important property to analyze the stability of the bilinear map $\mathbf{A}_{\beta,\mu}^{\mathcal{P}}$ is the stability of the reconstruction maps $\mathbf{L}_{\mathcal{P}_c}$ for all $\mathbf{c} \in \mathcal{C}$.

Proposition 5.4 (Stability). *There exist $\mathcal{C}_b, \mathcal{C}_\sharp > 0$ such that*

$$\forall \mathbf{c} \in \mathcal{C}, \quad \mathcal{C}_b \|\mathbf{v}\|_{\mathcal{P}_c,2} \leq \|\mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(\mathbf{c})} \leq \mathcal{C}_\sharp \|\mathbf{v}\|_{\mathcal{P}_c,2}, \quad \forall \mathbf{v} \in \mathcal{P}_c. \quad (5.22)$$

Proof. See the proof of Proposition 7.35. $\square$

Proposition 5.5 (Coercivity). *Assume that $(\mathcal{H})$ holds. Then,*

$$\mathbf{A}_{\beta,\mu}^{\mathcal{P}}(\mathbf{v}, \mathbf{v}) \geq \frac{\min(1, \mathcal{C}_b^2)}{2} \|\mathbf{v}\|_{\mathcal{P},a}^2, \quad \forall \mathbf{v} \in \mathcal{P}.$$

As a consequence, the problem (5.16) is well-posed.

Proof. Let $\mathbf{c} \in \mathcal{C}$ and let $\mathbf{v} \in \mathcal{P}_c$. Applying the Leibniz rule to integrate by parts the advective derivative in (5.9) and recalling that $\mathbf{L}_{\mathcal{P}_c}(\mathbf{v})$ is a continuous function in $\mathbf{c}$, we infer that

$$\mathbf{A}_{\beta,\mu;c}^{\mathcal{P}}(\mathbf{v}, \mathbf{v}) = \int_{\mathbf{c}} \sigma_{\beta,\mu} \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})^2 + \gamma \mathbf{s}_{\beta;c}(\mathbf{v}, \mathbf{v}) + \frac{1}{2} \int_{\mathbf{c}} \nabla \cdot (\beta \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})^2),$$

with $\sigma_{\beta,\mu} = \mu - \frac{1}{2} \nabla \cdot \beta$. Then, summing this relation for all $\mathbf{c} \in \mathcal{C}$ yields

$$\sum_{\mathbf{c} \in \mathcal{C}} \mathbf{A}_{\beta,\mu;c}^{\mathcal{P}}(\mathbf{v}, \mathbf{v}) = \sum_{\mathbf{c} \in \mathcal{C}} \left(\int_{\mathbf{c}} \sigma_{\beta,\mu} \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})^2 + \gamma \mathbf{s}_{\beta;c}(\mathbf{v}, \mathbf{v}) \right) + \frac{1}{2} \sum_{\mathbf{c} \in \mathcal{C}} \int_{\mathbf{c}} \nabla \cdot (\beta \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})^2).$$

Let $f \in F^\circ$ and let $c, c' \in C_f$ the two cells sharing f . Observing that $L_{\mathcal{P}_c}(v)|_f = L_{\mathcal{P}_{c'}}(v)|_f$ since these two functions are uniquely determined by the dofs of v attached to the mesh vertices $v \in V_f$, we infer that

$$\int_f \beta \cdot \mathbf{n}_{f,c} L_{\mathcal{P}_c}(v)^2 = - \int_f \beta \cdot \mathbf{n}_{f,c'} L_{\mathcal{P}_{c'}}(v)^2, \quad (5.23)$$

with $\mathbf{n}_{f,c}$ the normal vector attached to f pointing outward c . Hence, applying the divergence theorem yields

$$\sum_{c \in C} A_{\beta, \mu; c}^{\mathcal{P}}(v, v) = \sum_{c \in C} \left(\int_c \sigma_{\beta, \mu} L_{\mathcal{P}_c}(v)^2 + \gamma s_{\beta; c}(v, v) \right) + \frac{1}{2} \sum_{f \in F^\partial} \int_f (\beta \cdot \mathbf{n}) L_{\mathcal{P}_{c(f)}}(v)^2,$$

so that the expected result follows by adding the term $\sum_{f \in F^\partial} A_{(\beta \cdot \mathbf{n})^\ominus; f}^{\mathcal{P}, \partial}(v, v)$ with $(\beta \cdot \mathbf{n})^\ominus = \frac{1}{2}(|\beta \cdot \mathbf{n}| - \beta \cdot \mathbf{n})$ and using the stability of $L_{\mathcal{P}_c}$ from Lemma 5.4 along with assumption $(\mathcal{H})$ and $\gamma \geq \gamma_b > 0$. $\square$

The coercivity norm (5.21) is not strong enough to establish an *a priori* error estimate of order $\frac{3}{2}$. To this purpose, the following stronger norm, also called *inf-sup stability* norm, is introduced:

$$\|w\|_{\mathcal{P}, \#a}^2 := \|w\|_{\mathcal{P}, a}^2 + \sum_{c \in C} h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla L_{\mathcal{P}_c}(w)\|_{L^2(c)}^2, \quad \forall w \in \mathcal{P}. \quad (5.24)$$

To avoid the proliferation of constants in the analysis, we assume that there exists $\mathcal{C}_{\mathcal{P}, a} > 0$ independent of the mesh size and the physical parameters, such that

$$\max \left(\tau L_\beta, \|\mu\|_{L^\infty(\Omega)} \tau, \tau^{-1} \max_{c \in C} \left(h_c |\beta_c|^{-1} \right) \right) \leq \mathcal{C}_{\mathcal{P}, a}, \quad (5.25)$$

where we denote as before L_β the Lipschitz constant of β , satisfying $\|\beta - \beta_c\|_{L^\infty(c)} \leq L_\beta h_c$ and $\|\nabla \cdot \beta\|_{L^\infty(c)} \leq L_\beta$ for all $c \in C$.

Lemma 5.6 (Inf-sup stability). *Assume that $(\mathcal{H})$ and (5.25) hold. Then, there exists $\varrho > 0$ such that*

$$\sup_{w \in \mathcal{P}; \|w\|_{\mathcal{P}, \#a} = 1} A_{\beta, \mu}^{\mathcal{P}}(v, w) \geq \varrho \|v\|_{\mathcal{P}, \#a}, \quad \forall v \in \mathcal{P}. \quad (5.26)$$

The proof of Lemma 5.6 hinges on the following inequalities.

Proposition 5.7 (Face inequalities). *Let $c \in C$ and let $w \in \mathcal{P}_c$. Then, there exist $\mathcal{C}_{\text{INV}} > 0$, $\mathcal{C}_T > 0$ and $\mathcal{C}_{\text{AVG}} > 0$ such that*

$$\|\nabla L_{\mathcal{P}_c}(w)\|_{L^2(c)} \leq \mathcal{C}_{\text{INV}} h_c^{-1} \|L_{\mathcal{P}_c}(w)\|_{L^2(c)}, \quad (5.27a)$$

$$\forall f \in \mathfrak{F}_{\text{EF}, c}, \quad \|\beta_c \cdot \nabla L_{\mathcal{P}_c}(w)\|_{L^2(f)} \leq \mathcal{C}_T h_c^{-\frac{1}{2}} \|\beta_c \cdot \nabla L_{\mathcal{P}_c}(w)\|_{L^2(c)}, \quad (5.27b)$$

$$\left\| \beta_c \cdot \nabla L_{\mathcal{P}_c}(w) - \frac{1}{\#\mathfrak{C}_{\text{EF}, c}} \sum_{c \in \mathfrak{C}_{\text{EF}, c}} \beta_c \cdot \nabla L_{\mathcal{P}_c}(w)|_c \right\|_{L^2(c)} \leq \mathcal{C}_{\text{AVG}} h_c^{\frac{1}{2}} \sum_{f \in \mathfrak{F}_{\text{EF}, c}} \|[\beta_c \cdot \nabla L_{\mathcal{P}_c}(w)]\|_{L^2(f)}. \quad (5.27c)$$

Proof. These inequalities are proved in Chapter 7. In essence, the first inequality (5.27a) is the classical inverse inequality, the second one (5.27b) is a discrete trace inequality and the last one (5.27c) is the error estimate of the Oswald interpolation map. $\square$

Proof of Lemma 5.6. Let $v \in \mathcal{P}$ and let $S := \sup_{w \in \mathcal{P}; \|w\|_{\mathcal{P}, \#a} = 1} A_{\beta, \mu}^{\mathcal{P}}(v, w)$. Owing to the Proposition 5.5, we first have

$$\|v\|_{\mathcal{P}, a}^2 \lesssim A_{\beta, \mu}^{\mathcal{P}}(v, v) \leq S \|v\|_{\mathcal{P}, \#a}. \quad (5.28)$$

Hence, it only remains to control the advective derivative of $\|\cdot\|_{\mathcal{P},\sharp\mathbf{a}}$. To do so, recall that $\mathbf{L}_{\mathcal{P}_c}(\mathbf{v})$ is a continuous piece-wise affine function on the partition $\mathfrak{C}_{\text{EF},c}$ of c and define the discrete test function $\mathbf{w} \in \mathcal{P}$ defined by $\mathbf{w}_v = 0$ for all $v \in V$ and by

$$\mathbf{w}_c := h_c |\beta_c|^{-1} \frac{1}{\#\mathfrak{C}_{\text{EF},c}} \sum_{\mathfrak{c} \in \mathfrak{C}_{\text{EF},c}} \beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})|_{\mathfrak{c}}, \quad \forall c \in C, \quad (5.29)$$

First, let us prove that $\|\mathbf{w}\|_{\mathcal{P},\sharp\mathbf{a}} \lesssim \|\mathbf{v}\|_{\mathcal{P},\sharp\mathbf{a}}$. Owing to the definition (5.20) of $\|\cdot\|_{\mathcal{P},2}$, we observe that $\|\mathbf{w}\|_{\mathcal{P},2}^2 = h_c^3 \mathbf{w}_c^2$. Then, it follows that $h_c^{-1} |\beta_c| \|\mathbf{w}\|_{\mathcal{P},2}^2 = h_c^2 |\beta_c| \mathbf{w}_c^2$, leading to

$$h_c^{-1} |\beta_c| \|\mathbf{w}\|_{\mathcal{P},2}^2 \leq h_c^4 |\beta_c|^{-1} \frac{1}{\#\mathfrak{C}_{\text{EF},c}} \sum_{\mathfrak{c} \in \mathfrak{C}_{\text{EF},c}} |\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})|_{\mathfrak{c}}^2,$$

owing to the Jensen inequality. Then, since $\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})$ is piece-wise constant on all $\mathfrak{c} \in \mathfrak{C}_{\text{EF},c}$ and owing to mesh regularity, we finally infer that

$$h_c^{-1} |\beta_c| \|\mathbf{w}\|_{\mathcal{P},2}^2 \lesssim h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2. \quad (5.30)$$

In parallel, observe that for all $\mathbf{f} \in \mathbf{F}^\partial$, the boundary contribution of the norm $\|\mathbf{w}\|_{\mathcal{P},\sharp\mathbf{a}}$ vanishes, i.e., $A_{\alpha;\mathbf{f}}^{\mathcal{P},\partial}(\mathbf{w}, \mathbf{w}) = 0$, since $\mathbf{L}_{\mathcal{P}_c}(\mathbf{w})$ vanishes at the boundary of all mesh cells $c \in C$ since $\mathbf{w}_v = 0$ for all $v \in V$. Hence, owing to the definition (5.24) of the inf-sup norm, we have

$$\|\mathbf{w}\|_{\mathcal{P},\sharp\mathbf{a}}^2 = \sum_{c \in C} \left(\tau^{-1} \|\mathbf{w}\|_{\mathcal{P},2}^2 + h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2 + \gamma_b \mathbf{s}_{\beta;c}(\mathbf{w}, \mathbf{w}) \right),$$

and denote by T_1 , T_2 and T_3 the above three local contributions. First, using the inequality (5.30) and observing that $\tau^{-1} h_c |\beta_c|^{-1} \leq \mathcal{C}_{\mathcal{P},a}$ owing to (5.25), we observe that

$$T_1 = \tau^{-1} \|\mathbf{w}\|_{\mathcal{P},2}^2 \leq \mathcal{C}_{\mathcal{P},a} h_c^{-1} |\beta_c| \|\mathbf{w}\|_{\mathcal{P},2}^2 \lesssim \mathcal{C}_{\mathcal{P},a} h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2. \quad (5.31)$$

Turning to T_2 , we use the inverse inequality (5.27a) and the upper bound from Proposition 7.35, to infer that

$$T_2 = h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2 \leq \mathcal{C}_{\text{INV}}^2 \mathcal{C}_{\sharp}^2 h_c^{-1} |\beta_c| \|\mathbf{w}\|_{\mathcal{P},2}^2,$$

so that combined with the bound (5.31) on $\|\mathbf{w}\|_{\mathcal{P},2}$, this leads to

$$h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2 \lesssim \mathcal{C}_{\mathcal{P},a} \mathcal{C}_{\text{INV}}^2 \mathcal{C}_{\sharp}^2 h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2. \quad (5.32)$$

Consider the last term T_3 . The definition (5.12) of $\mathbf{s}_{\beta;c}$ along with the trace inequality (5.27b) yields

$$\mathbf{s}_{\beta;c}(\mathbf{w}, \mathbf{w}) \lesssim \mathcal{C}_T^2 h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2.$$

Hence, using the bound (5.32) on T_2 and recalling that $\gamma_b \leq 1$ yield

$$\gamma_b \mathbf{s}_{\beta;c}(\mathbf{w}, \mathbf{w}) \lesssim \mathcal{C}_{\mathcal{P},a} \mathcal{C}_{\text{INV}}^2 \mathcal{C}_{\sharp}^2 h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2. \quad (5.33)$$

As a first result, collecting (5.31), (5.32) and (5.33) yields

$$\|\mathbf{w}\|_{\mathcal{P},\sharp\mathbf{a}}^2 \lesssim \sum_{c \in C} h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \lesssim \|\mathbf{v}\|_{\mathcal{P},\sharp\mathbf{a}}^2.$$

Now, let us prove that $A_{\beta,\mu}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) \gtrsim \|\mathbf{v}\|_{\mathcal{P},\sharp\mathbf{a}}^2$. First, observing that there is $C_\theta > 0$ such that $C_\theta \|\phi\|_{L^2(c)}^2 \leq \int_c \theta_c \phi^2$ for any piece-wise affine function ϕ in c with θ_c the Courant basis function attached to c , we observe that

$$C_\theta h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \leq h_c |\beta_c|^{-1} \int_c (\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})) (\theta_c \beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})).$$

Next, we rewrite the right-hand side as $\int_c (\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \theta_c \mathbf{w}_c + \Delta_c$ with $\mathbf{w}_c$ defined by (5.29) and with the perturbation term

$$\Delta_c = h_c |\beta_c|^{-1} \int_c (\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \theta_c \left(\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v}) - h_c^{-1} |\beta_c| \mathbf{w}_c \right).$$

Observing now that $\mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) = \mathbf{w}_c \theta_c$ (recalling that $\mathbf{w}_v = 0$ for all $v \in V$), we finally obtain

$$C_\theta h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \leq \int_c (\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) + \Delta_c.$$

Next, the perturbation term is bounded using the Cauchy–Schwarz inequality with $\theta_c \leq 1$ as

$$|\Delta_c|^2 \leq \left(h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \right) \left(h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v}) - h_c^{-1} |\beta_c| \mathbf{w}_c\|_{L^2(c)}^2 \right).$$

Recalling now the definition (5.29) of the discrete test function $\mathbf{w} \in \mathcal{P}$ corresponding to the point-wise reduction of the discrete Oswald interpolation of $\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})$, we use the interpolation error estimate (5.27c) to infer that

$$|\Delta_c| \leq \left(h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \right)^{\frac{1}{2}} (C_{\text{avg}} s_{\beta;c}(\mathbf{v}, \mathbf{v}))^{\frac{1}{2}},$$

so that applying Young’s inequality leads to

$$\frac{1}{2} C_\theta h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \leq \int_c (\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) + \frac{1}{2} C_\theta^{-1} C_{\text{avg}} s_{\beta;c}(\mathbf{v}, \mathbf{v}).$$

Next, recalling the definition (5.8) of $A_{\beta, \mu; c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w})$, we rewrite the first right-hand side term of this above bound as

$$\int_c (\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) = A_{\beta, \mu; c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) - \Delta'_c,$$

where

$$\Delta'_c = \int_c \mu \mathcal{L}_{\mathcal{P}_c}(\mathbf{v}) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) + \int_c ((\beta - \beta_c) \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})) \mathcal{L}_{\mathcal{P}_c}(\mathbf{w}) + \gamma s_{\beta; c}(\mathbf{v}, \mathbf{w}).$$

Then, owing to the inverse inequality from Proposition (7.26) and the upper bound of the stability of $\mathcal{L}_{\mathcal{P}_c}$ from Proposition 7.35 together with the Cauchy–Schwarz inequality, we bound the perturbation term Δ'_c as

$$|\Delta'_c| \leq C_{\sharp}^2 \|\mu\|_{L^\infty(c)} \|\mathbf{v}\|_{\mathcal{P}_c, 2} \|\mathbf{w}\|_{\mathcal{P}_c, 2} + C_{\text{inv}} C_{\sharp}^2 h_c^{-1} \|\beta - \beta_c\|_{L^\infty(c)} \|\mathbf{v}\|_{\mathcal{P}_c, 2} \|\mathbf{w}\|_{\mathcal{P}_c, 2} + \gamma s_{\beta; c}(\mathbf{v}, \mathbf{v})^{\frac{1}{2}} s_{\beta; c}(\mathbf{w}, \mathbf{w})^{\frac{1}{2}},$$

so that owing to the Lipschitz regularity of β , and assumption (5.25), it follows that

$$|\Delta'_c| \lesssim C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}_c, a} \|\mathbf{w}\|_{\mathcal{P}_c, a}.$$

As a result, $\sum_{c \in C} |\Delta'_c| \lesssim C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, a} \|\mathbf{w}\|_{\mathcal{P}, a} \lesssim C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a}$ since $\|\mathbf{w}\|_{\mathcal{P}, a} \lesssim C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a}$ owing to the first step of the proof. Finally, collecting the above bounds yields

$$\sum_{c \in C} h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathcal{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^2(c)}^2 \lesssim \sum_{c \in C} A_{\beta, \mu; c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) + C_{\mathcal{P}, a}^2 \|\mathbf{v}\|_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a}.$$

Recalling from the definition of $\mathbf{w}$ that

$$\sum_{c \in C} A_{\beta, \mu; c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) = A_{\beta, \mu}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) \leq S \|\mathbf{w}\|_{\mathcal{P}, \sharp a} \lesssim S C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a},$$

we end up with $\|\mathbf{v}\|_{\mathcal{P}, \sharp a}^2 \lesssim S C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a} + C_{\mathcal{P}, a}^2 \|\mathbf{v}\|_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a}$, so that $\|\mathbf{v}\|_{\mathcal{P}, \sharp a}^2 \lesssim S C_{\mathcal{P}, a} \|\mathbf{v}\|_{\mathcal{P}, \sharp a} + C_{\mathcal{P}, a}^2 \|\mathbf{v}\|_{\mathcal{P}, a}^2$ owing to Young’s inequality and then

$$\|\mathbf{v}\|_{\mathcal{P}, \sharp a} \lesssim (C_{\mathcal{P}, a} + C_{\mathcal{P}, a}^2) S,$$

using the first estimate (5.28). The proof is completed. $\square$

Remark 5.8 (Analysis under assumption $(\mathcal{H}')$). *Proving the inf-sup condition as stated in Lemma 5.6 but under assumption $(\mathcal{H}')$ (that is, if $\text{ess inf}_\Omega \sigma_{\beta,\mu} = 0$) is still open for our scheme. This assumption is barely addressed in the literature in the context of CIP methods (see e.g., Burman (2014)), while in the context of dG methods, this assumption can be easily handled using the classical orthogonal subscales argument (see (Ern & Guermond, 2006a, Lemma 4.3) and Ayuso & Marini (2009)).*

Consistency and *a priori* error estimate. We measure the consistency error of the discrete problem (5.16) with the reduction map $R_{\mathcal{P}_c} : W^{s,p}(c) \rightarrow \mathcal{P}_c$ with $sp > 3$ defined by (5.2). Instead of introducing commutators as in Sections 3.2.2 and 4.2.2, the following Lemma 5.9 directly bounds the global consistency error

$$\mathbb{E}_{\mathcal{P}}(u) = \sup_{\mathbf{v} \in \mathcal{P}; \|\mathbf{v}\|_{\mathcal{P},\#a}=1} \left| \mathbb{A}_{\beta,\mu}^{\mathcal{P}}(R_{\mathcal{P}}(u), \mathbf{v}) - \mathbb{S}(s, u_D; \mathbf{v}) \right|. \quad (5.34)$$

Lemma 5.9 (Bound on the consistency error). *Let u the exact solution of (5.5) and assume that it satisfies $u \in H^2(C)$. Assume that (5.25) holds. Then,*

$$\mathbb{E}_{\mathcal{P}}(u) \lesssim \left(\sum_{c \in C} (|\beta_c| + L_\beta) h_c^{-1} \|u - \mathbf{l}_{\mathcal{P}_c}(u)\|_{L^2(c)}^2 + |\beta_c| h_c \left(\|u - \mathbf{l}_{\mathcal{P}_c}(u)\|_{H^1(c)}^2 + h_c^2 \|u\|_{H^2(c)}^2 \right) \right)^{\frac{1}{2}},$$

with the interpolation map $\mathbf{l}_{\mathcal{P}_c} = \mathbf{L}_{\mathcal{P}_c} \circ R_{\mathcal{P}_c} : H^s(c) \rightarrow \mathbb{P}_1(\mathfrak{C}_{\text{EF},c}) \cap \mathcal{C}^0(c)$ with $s > \frac{3}{2}$.

Proof. Let us set $y|_c = u|_c - \mathbf{l}_{\mathcal{P}_c}(u)$ for all $c \in C$. Recalling the properties of the exact solution, the definition (5.17) of the linear form $\mathbb{S}(s, u_D; \cdot)$ and owing to the definition (5.7) of $\mathbb{A}_{\beta,\mu}^{\mathcal{P}}$, we infer that

$$\mathbb{S}(s, u_D; \mathbf{v}) - \mathbb{A}_{\beta,\mu}^{\mathcal{P}}(R_{\mathcal{P}}(u), \mathbf{v}) = T_1 + T_2 + T_3,$$

where we have defined

$$\begin{aligned} T_1 &:= \sum_{c \in C} - \int_c y (\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) + (\beta - \beta_c) \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) - (\mu - \nabla \cdot \beta) \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})), \\ T_2 &:= \sum_{c \in C} \gamma h_c^2 |\beta_c|^{-1} \sum_{\mathbf{f} \in \mathfrak{F}_{\text{EF};c}} \int_{\mathbf{f}} (\beta_c \cdot \llbracket \nabla y \rrbracket) (\beta_c \cdot \llbracket \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) \rrbracket), \\ T_3 &:= \sum_{\mathbf{f} \in F^\partial} \int_{\mathbf{f}} (\beta \cdot \mathbf{n})^\oplus y \mathbf{L}_{\mathcal{P}_{c(\mathbf{f})}}(\mathbf{v}). \end{aligned}$$

Indeed, using the identity

$$\sum_{c \in C} \int_c (\beta \cdot \nabla y) \mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) = \sum_{c \in C} - \int_c y (\beta \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) + (\nabla \cdot \beta) \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})) + \sum_{\mathbf{f} \in F^\partial} \int_{\mathbf{f}} (\beta \cdot \mathbf{n}) y \mathbf{L}_{\mathcal{P}_{c(\mathbf{f})}}(\mathbf{v}),$$

we obtain T_1 , while $\llbracket \nabla u \rrbracket_{\mathbf{f}} = 0$ for all $\mathbf{f} \in \mathfrak{F}_{\text{EF};c}$ owing to the regularity of the exact solution, yields the second term T_2 . Finally, since $u_D = u|_{\partial\Omega}$, we obtain the third term. Next, applying the Cauchy–Schwarz inequality along with the inverse inequality (5.27a) and the Lipschitz regularity of β lead to

$$|T_1|^2 \lesssim \left(\sum_{c \in C} (|\beta_c| h_c^{-1} + L_\beta + \|\mu - \nabla \cdot \beta\|_{L^\infty(c)} \tau) \|y\|_{L^2(c)}^2 \right) \|\mathbf{v}\|_{\mathcal{P},\#a}^2,$$

so that using (5.25), we obtain

$$|T_1|^2 \lesssim \left(\sum_{c \in C} (|\beta_c| + \mathbf{c}_\beta h_c L_\beta) h_c^{-1} \|y\|_{L^2(c)}^2 \right) \|\mathbf{v}\|_{\mathcal{P},\#a}^2.$$

To bound the second term, we use the multiplicative trace inequality from Lemma 7.25 (with $\mathbf{f} \in \mathfrak{F}_{\text{EF},c}$) leading to

$$\begin{aligned} |T_2|^2 &\lesssim \left(\sum_{c \in C} \sum_{\mathbf{f} \in \mathfrak{F}_{\text{EF},c}} h_c^2 |\beta_c|^{-1} \|\beta_c \cdot \llbracket \nabla y \rrbracket\|_{L^2(\mathbf{f})}^2 \right) \|\mathbf{v}\|_{\mathcal{P},\#a}^2 \\ &\lesssim \left(\sum_{c \in C} h_c |\beta_c| |y|_{H^1(c)} \left(|y|_{H^1(c)} + h_c |u|_{H^2(c)} \right) \right) \|\mathbf{v}\|_{\mathcal{P},\#a}^2. \end{aligned}$$

Finally, still owing to this trace inequality (this time with $\mathbf{f} = \mathbf{f} \in \mathbf{F}^\partial$), we infer that

$$|T_3|^2 \lesssim \left(\sum_{\mathbf{f} \in \mathbf{F}^\partial} h_c |\beta_c| |y|_{H^1(c)} \left(|y|_{H^1(c)} + h_c |u|_{H^2(c)} \right) \right) \|\mathbf{v}\|_{\mathcal{P},\#a}^2,$$

with the corresponding mesh cell $c = c(\mathbf{f})$. The expected result is then obtained collecting these bounds. $\square$

Theorem 5.10 (*A priori error estimate*). *Let u be the unique solution of (5.5) and let $\mathbf{u} \in \mathcal{P}$ be the unique solution of (5.16). Assume that assumptions $(\mathcal{H})$ and (5.25) hold, and that $u \in H^2(C)$. Then,*

$$\|\mathbf{u} - \mathbf{R}_{\mathcal{P}}(u)\|_{\mathcal{P},\#a} \lesssim \left(\sum_{c \in C} (|\beta_c| + h_c L_\beta) h_c^3 |u|_{H^2(c)}^2 \right)^{\frac{1}{2}}.$$

This theorem extends to polyhedral meshes the convergence rate of order $\frac{3}{2}$ observed on simplicial meshes using the stabilized $\mathbb{P}_1$ -Lagrange finite element method. One also retrieves the classical *a priori* estimates obtained with the dG method of order $k = 1$.

Proof. The proof of Theorem 5.10 is a direct consequence of the proposition below. $\square$

Proposition 5.11 (Interpolation error). *There exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in C, \quad \|v - \mathbf{l}_{\mathcal{P}_c}(v)\|_{L^2(c)} + h_c |v - \mathbf{l}_{\mathcal{P}_c}(v)|_{H^1(c)} \leq \mathcal{C}_{\text{INT}} h_c^2 |v|_{H^2(c)}, \quad \forall v \in H^2(c). \quad (5.35)$$

Proof. See the proof of Proposition 7.37. $\square$

Proceeding as the proof of Lemma 6.9, this result can be generalized to achieve an *a priori* error estimate for solutions that do not belong to $H^2(C)$.

Theorem 5.12 (*A priori error estimate for rough solutions*). *Let u be the unique solution of (5.5) and let $\mathbf{u} \in \mathcal{P}$ be the unique solution of (5.16). Assume that assumptions $(\mathcal{H})$ and (5.25) hold, and that $u \in W^{2,p}(C)$ with $p \in \left(\frac{3}{2}, 2\right]$. Then,*

$$\|\mathbf{u} - \mathbf{R}_{\mathcal{P}}(u)\|_{\mathcal{P},\#a} \lesssim \left(\sum_{c \in C} \left(|\beta_c|^{\frac{p}{2}} + h_c^{\frac{p}{2}} L_\beta^{\frac{p}{2}} \right) h_c^{3(p-1)} |u|_{W^{2,p}(c)}^p \right)^{\frac{1}{p}}.$$

Remark 5.13 (Regularity). *Note that the restriction $p > \frac{3}{2}$ comes from the regularity of the domain of the reduction map $\mathbf{R}_{\mathcal{P}_c}$ evaluating point-wise the exact solution. Considering instead a reduction map evaluating the mean value of the exact solution (similarly to $\hat{\mathbf{R}}_{\mathcal{V}_c}$), it is reasonable to think that Theorem 5.10 can be extended so as to infer an *a priori* error estimate if the exact solution satisfies $u \in W^{2,p}(C)$ for all $p \in (1, 2]$.*

5.3 Diffusion-advection-reaction problem

This section extends the scheme (5.16) devised for the advection-reaction problem so as to approximate the solution of the following diffusion-advection-reaction problem:

$$-\nabla \cdot (\boldsymbol{\lambda} \nabla u) + \boldsymbol{\beta} \cdot \nabla u + \mu u = s \quad \text{a.e. in } \Omega, \quad (5.36a)$$

$$u = u_D \quad \text{a.e. on } \partial\Omega, \quad (5.36b)$$

with $s \in L^2(\Omega)$, $u_D \in H^{1+\epsilon}(\partial\Omega)$ with $\epsilon > 0$ and where the diffusion tensor $\boldsymbol{\lambda} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ is assumed to take symmetric and positive values. We recall that $\lambda_{\sharp;c}$ and $\lambda_{\flat;c}$ denote the minimal and the maximal eigenvalue of $\boldsymbol{\lambda}$ in c , respectively. The coefficient μ and $\boldsymbol{\beta}$ satisfy the same assumption as in Section 5.2. In particular, the Friedrichs tensor $\sigma_{\beta,\mu}$ satisfies the assumption $(\mathcal{H})$, i.e., $\text{ess inf}_{\Omega} \sigma_{\beta,\mu} > 0$. In what follows, proofs are omitted since they straightforwardly follow from those of Chapter 4 and of Section 5.2.

5.3.1 Discrete problem

Proceeding similarly to Section 4.2, we first introduce the diffusion related bilinear form $\mathbb{A}_{\boldsymbol{\lambda}}^{\mathcal{P}} : \mathcal{P} \times \mathcal{P} \rightarrow \mathbb{R}$ which is composed of two terms:

$$\mathbb{A}_{\boldsymbol{\lambda}}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) = \sum_{c \in \mathcal{C}} \mathbb{A}_{\boldsymbol{\lambda};c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) + \sum_{f \in \mathcal{F}^{\partial}} \mathbb{A}_{\boldsymbol{\lambda};f}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{P}. \quad (5.37)$$

The first right-hand side term results from the Galerkin approximation of the diffusive-related term in (5.36a) using the reconstruction maps $\mathbb{L}_{\mathcal{P}_c}$:

$$\forall c \in \mathcal{C}, \quad \mathbb{A}_{\boldsymbol{\lambda};c}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) = \int_c \nabla \mathbb{L}_{\mathcal{P}_c}(\mathbf{v}) \cdot \boldsymbol{\lambda} \cdot \nabla \mathbb{L}_{\mathcal{P}_c}(\mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{P}_c. \quad (5.38)$$

The boundary condition (5.36b) is weakly enforced using the local bilinear form $\mathbb{A}_{\boldsymbol{\lambda};f}^{\mathcal{P},\partial} : \mathcal{P}_c \rightarrow \mathcal{P}_c$ (with $c \equiv c(f)$) defined as

$$\forall f \in \mathcal{F}^{\partial}, \quad \mathbb{A}_{\boldsymbol{\lambda};f}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) = - \int_f \nabla \mathbb{L}_{\mathcal{P}_c}(\mathbf{v}) \cdot \boldsymbol{\lambda} \cdot \mathbf{n}_{\mathcal{P}_c}(\mathbf{w}) + \nu_0 \frac{\lambda_{\sharp;c}}{h_c} \int_f \mathbb{L}_{\mathcal{P}_c}(\mathbf{v}) \mathbb{L}_{\mathcal{P}_c}(\mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{P}_c, \quad (5.39)$$

with $\nu_0 > 0$ the Nitsche penalty term to be chosen large enough to guarantee the well-posedness of the discrete problem (see Proposition 5.14).

Approximation of (5.36). The approximation of the continuous problem (5.36) hinges on the bilinear form

$$\mathbb{A}_{\boldsymbol{\lambda},\boldsymbol{\beta},\mu}^{\mathcal{P}} = \mathbb{A}_{\boldsymbol{\lambda}}^{\mathcal{P}} + \mathbb{A}_{\boldsymbol{\beta},\mu}^{\mathcal{P}}, \quad (5.40)$$

where $\mathbb{A}_{\boldsymbol{\beta},\mu}^{\mathcal{P}}$ is the advection-reaction related bilinear form defined by (5.7). The discrete problem then reads:

$$\text{Find } \mathbf{u} \in \mathcal{P} \text{ s.t. } \mathbb{A}_{\boldsymbol{\lambda},\boldsymbol{\beta},\mu}^{\mathcal{P}}(\mathbf{u}, \mathbf{v}) = \mathbb{S}(s, u_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{P}, \quad (5.41)$$

where the source term $\mathbb{S}(s, u_D; \cdot) : \mathcal{P} \rightarrow \mathbb{R}$ is defined, for all $\mathbf{v} \in \mathcal{P}$, as

$$\mathbb{S}(s, u_D; \mathbf{v}) = \sum_{c \in \mathcal{C}} \int_c s \mathbb{L}_{\mathcal{P}_c}(\mathbf{v}) + \sum_{f \in \mathcal{F}^{\partial}} \left(\nu_0 \frac{\lambda_{\sharp;c(f)}}{h_{c(f)}} \int_f u_D \mathbb{L}_{\mathcal{P}_{c(f)}}(\mathbf{v}) + \int_f (\boldsymbol{\beta} \cdot \mathbf{n})^{\ominus} u_D \mathbb{L}_{\mathcal{P}_{c(f)}}(\mathbf{v}) \right). \quad (5.42)$$

5.3.2 Analysis

The dofs space $\mathcal{P}$ is equipped with the diffusive-related discrete norm defined as

$$\|\mathbf{w}\|_{\mathcal{P},d}^2 := \sum_{c \in \mathcal{C}} \lambda_{b;c} \|\nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2 + \sum_{f \in \mathcal{F}^\partial} \frac{\lambda_{\sharp;c(f)}}{h_{c(f)}} \|\mathbf{L}_{\mathcal{P}_{c(f)}}(\mathbf{w})\|_{L^2(f)}^2, \quad \forall \mathbf{w} \in \mathcal{P}, \quad (5.43)$$

and, for all $\mathbf{w} \in \mathcal{P}$, we denote $\|\mathbf{w}\|_{\mathcal{P},da}^2 := \|\mathbf{w}\|_{\mathcal{P},d}^2 + \|\mathbf{w}\|_{\mathcal{P},a}^2$. The following lemma establishes the coercivity of the bilinear form (5.37).

Lemma 5.14 (Coercivity of $\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{P}}$). *Assume that $(\mathcal{H})$ holds and that ν_0 is large enough. Then,*

$$\mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{P}}(\mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \min(1, \mathcal{C}_b^2) \|\mathbf{v}\|_{\mathcal{P},da}^2, \quad \forall \mathbf{v} \in \mathcal{P}.$$

This stability result is extended in order to additionally control the advective derivative of the discrete solution. We define the inf-sup stability norm as follows:

$$\|\mathbf{w}\|_{\mathcal{P},\sharp da}^2 = \|\mathbf{w}\|_{\mathcal{P},da}^2 + \sum_{c \in \mathcal{C}} h_c |\beta_c|^{-1} \|\beta_c \cdot \nabla \mathbf{L}_{\mathcal{P}_c}(\mathbf{w})\|_{L^2(c)}^2, \quad \forall \mathbf{w} \in \mathcal{P}. \quad (5.44)$$

Lemma 5.15 (Inf-sup stability). *Assume that $(\mathcal{H})$ and (5.25) hold, and that ν_0 is large enough. Then, there exists $\varrho > 0$ such that*

$$\sup_{\mathbf{w} \in \mathcal{P}; \|\mathbf{w}\|_{\mathcal{P},\sharp da}=1} \mathbb{A}_{\lambda,\beta,\mu}^{\mathcal{P}}(\mathbf{v}, \mathbf{w}) \geq \varrho \|\mathbf{v}\|_{\mathcal{P},\sharp da}, \quad \forall \mathbf{v} \in \mathcal{P}. \quad (5.45)$$

Theorem 5.16. *Let u be the unique solution of (5.36) and let $\mathbf{u} \in \mathcal{P}$ be the unique solution of (5.41). Assume that $(\mathcal{H})$ and (5.25) hold, and that $u \in H^2(\Omega)$. Then,*

$$\|\mathbf{u} - \mathbf{R}_{\mathcal{P}}(u)\|_{\mathcal{P},\sharp da} \lesssim \left(\sum_{c \in \mathcal{C}} \left(h_c |\beta_c| + h_c^2 L_\beta + \rho_c \lambda_{\sharp;c} \right) h_c^2 |u|_{H^2(c)}^2 \right)^{\frac{1}{2}},$$

with ρ_c the anisotropy ratio of λ defined as $\rho_c = \lambda_{\sharp;c} / \lambda_{b;c}$.

5.4 Numerical results

In this section, we investigate some implementation aspects for the discrete problems (5.16) and (5.41). Numerical results for the test cases 1, 2 and 3 described in Chapters 3 and 4 are considered so as to be compared with the previous results.

5.4.1 Implementation aspects

Elimination of cell-based unknowns. The algebraic realization of the discrete problems (5.16) and (5.41) is the linear system $\mathbb{A}^{\mathcal{P}} \mathbf{u} = \mathbb{S}$ with $\mathbf{u} = (\mathbf{u}_v, \mathbf{u}_c) \in \mathcal{P}$,

$$\mathbb{A}^{\mathcal{P}} = \begin{pmatrix} \mathbb{A}_{vv}^{\mathcal{P}} & \mathbb{A}_{vc}^{\mathcal{P}} \\ \mathbb{A}_{cv}^{\mathcal{P}} & \mathbb{A}_{cc}^{\mathcal{P}} \end{pmatrix} \quad \text{and} \quad \mathbb{S} = \begin{pmatrix} \mathbb{S}_v \\ \mathbb{S}_c \end{pmatrix}, \quad (5.46)$$

where $\mathbb{S}_{\mathcal{X}} \in \mathcal{X}$ is the restriction of $\mathbb{S}(s, u_D; \cdot)$ to $\mathcal{X}$ and $\mathbb{A}_{\mathcal{X}\mathcal{Y}}^{\mathcal{P}} : \mathcal{Y} \rightarrow \mathcal{X}$ the restriction of $\mathbb{A}^{\mathcal{P}}$ to $\mathcal{X}, \mathcal{Y}$ for $\mathcal{X}, \mathcal{Y} \in \{\mathcal{V}, \mathcal{C}\}$. Observing that $\mathbb{A}_{cc}^{\mathcal{P}}$ is diagonal since cell-based dofs are not coupled to each other, one can easily compute its *Schur complement* so as to express $\mathbf{u}_c$ in terms of $\mathbf{u}_v$ and obtain a linear system in terms of $\mathbf{u}_v$ only. This operation, which is often called *static condensation* in the finite element context, leads to the equivalent formulation $\underline{\mathbb{A}}^{\mathcal{P}} \mathbf{u} = \underline{\mathbb{S}}$ where

$$\underline{\mathbb{A}}^{\mathcal{P}} = \begin{pmatrix} \mathbb{A}_{vv}^{\mathcal{P}} - \mathbb{A}_{vc}^{\mathcal{P}} (\mathbb{A}_{cc}^{\mathcal{P}})^{-1} \mathbb{A}_{cv}^{\mathcal{P}} & \mathbf{0}_{cv} \\ (\mathbb{A}_{cc}^{\mathcal{P}})^{-1} \mathbb{A}_{cv}^{\mathcal{P}} & \text{Id}_{cc} \end{pmatrix} \quad \text{and} \quad \underline{\mathbb{S}} = \begin{pmatrix} \mathbb{S}_v - \mathbb{A}_{vc}^{\mathcal{P}} (\mathbb{A}_{cc}^{\mathcal{P}})^{-1} \mathbb{S}_c \\ (\mathbb{A}_{cc}^{\mathcal{P}})^{-1} \mathbb{S}_c \end{pmatrix}. \quad (5.47)$$

In order to compare the stencils associated with the two blocks $\mathbb{A}_{\mathcal{V}\mathcal{V}}^{\mathcal{P}}$ and $\underline{\mathbb{A}}_{\mathcal{V}\mathcal{V}}^{\mathcal{P}} = \mathbb{A}_{\mathcal{V}\mathcal{V}}^{\mathcal{P}} - \mathbb{A}_{\mathcal{V}\mathcal{C}}^{\mathcal{P}}(\mathbb{A}_{\mathcal{C}\mathcal{C}}^{\mathcal{P}})^{-1}\mathbb{A}_{\mathcal{C}\mathcal{V}}^{\mathcal{P}}$, we introduce the following sub-sets of $\mathcal{V}$:

$$\forall v \in \mathcal{V}, \quad \mathbf{St}_{\mathcal{V}}(v) = \{v' \in \mathcal{V} \mid \mathbb{A}_{\mathcal{V}\mathcal{V}'}^{\mathcal{P}} \neq 0\} \quad \text{and} \quad \underline{\mathbf{St}}_{\mathcal{V}}(v) = \{v' \in \mathcal{V} \mid \underline{\mathbb{A}}_{\mathcal{V}\mathcal{V}'}^{\mathcal{P}} \neq 0\}.$$

One can verify that $\underline{\mathbf{St}}_{\mathcal{V}}(v) = \{v' \in \mathcal{V} \mid C_v \cap C_{v'} \neq \emptyset\}$, yielding

$$\forall v \in \mathcal{V}, \quad \#\underline{\mathbf{St}}_{\mathcal{V}}(v) = \sum_{c \in C_v} \sum_{v' \in V_c} \frac{1}{\#(C_v \cap C_{v'})}. \quad (5.48)$$

In addition, we observe that

$$\mathbf{St}_{\mathcal{V}}(v) = \underline{\mathbf{St}}_{\mathcal{V}}(v) \cap \{v' \in \mathcal{V} \mid \exists v'' \in \mathcal{V}, F_v \cap F_{v''} \neq \emptyset \text{ and } F_{v'} \cap F_{v''} \neq \emptyset\},$$

so that $\mathbf{St}_{\mathcal{V}}(v) \subset \underline{\mathbf{St}}_{\mathcal{V}}(v)$. The converse inclusion holds in the following situation.

Proposition 5.17 (Vertex stencil). *Let $v \in \mathcal{V}$ and assume that for all $c \in C_v$, all the vertices in V_c are connected to v by a maximum of two faces of c , i.e., it is possible to find two faces $f, f' \in F_c$ having in common at least one vertex such that $v \subset \partial f$ and $v' \subset \partial f'$. Then, $\mathbf{St}_{\mathcal{V}}(v) = \underline{\mathbf{St}}_{\mathcal{V}}(v)$.*

Proof. Let $v' \in \underline{\mathbf{St}}_{\mathcal{V}}(v)$. Reformulating the definition of $\underline{\mathbf{St}}_{\mathcal{V}}(v)$ as $\{v' \in \mathcal{V} \mid \exists c \in C_v, v' \in V_c\}$, it follows, using the assumption of the proposition, that there exist $f, f' \in F_c$ having in common at least one vertex and such that $v \subset \partial f$ and $v' \subset \partial f'$. Denoting $v'' \in \partial f \cap \partial f'$ this common vertex, it follows that $\{v, v''\} \subset \partial f$ and $\{v', v''\} \subset \partial f'$, so that $F_v \cap F_{v''} \neq \emptyset$ and $F_{v'} \cap F_{v''} \neq \emptyset$. Hence, we infer that $v' \in \mathbf{St}_{\mathcal{V}}(v)$. $\square$

The assumption of Proposition 5.17 is often met in practice, as long as the mesh cells do not have too many vertices. An example of a cell that does not satisfy this assumption is shown in Figure 5.3. In this case, we stress that the static condensation is still possible, but it slightly increases the original stencil of the block matrix $\mathbb{A}_{\mathcal{V}\mathcal{V}}^{\mathcal{P}}$.

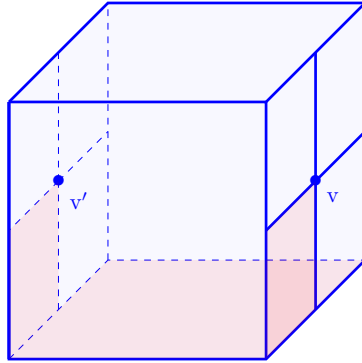


Figure 5.3 – Example of cell (cf. CB sequences) which does not satisfy assumption of Proposition 5.17: v and v' are connected by at least 3 faces.

Computation of the source term. Quadrature rules may be used in the computation of the source terms (5.17) and (5.42) and the system matrix (5.7). Practically, the source term $\mathbb{S}$ related to the advection-reaction problem (5.16) (and similarly for the problem (5.41)) is approximated as follows:

$$\forall v \in \mathcal{P}, \quad \mathbb{S}^h(s, u_D; v) := \sum_{c \in \mathcal{C}} \int_c l_{\mathcal{P}_c}(s) L_{\mathcal{P}_c}(v) + \sum_{f \in \mathcal{F}^\partial} \int_f (\beta \cdot \mathbf{n})^\ominus l_{\mathcal{P}_c(f)}(\widetilde{u_D}) L_{\mathcal{P}_c}(v), \quad (5.49)$$

where $\mathsf{I}_{\mathcal{P}_c} = \mathsf{L}_{\mathcal{P}} \circ \mathsf{R}_{\mathcal{P}}$ is the interpolation map defined in Lemma 5.9 and where $\widetilde{u_D}$ is a lifting of u_D inside Ω such that $\widetilde{u_D}|_f = u_D|_f$ for all $f \in \mathsf{F}^\partial$. Note that $\mathsf{I}_{\mathcal{P}_c(f)}(\widetilde{u_D})|_f$ is independent of the choice of the lifting $\widetilde{u_D}$ inside Ω . The source term $\mathbb{S}^h$ has the advantage of being exactly computed using a second-order quadrature as soon as the boundary mesh is compatible with the normal component of β in the following sense:

Definition 5.18 (Compatible boundary mesh). *The boundary mesh M is compatible with the normal component of β if, for all $f \in \mathsf{F}^\partial$, the sign of $(\beta \cdot \mathbf{n})$ is constant on f .*

However, the approximate source term $\mathbb{S}^h$ introduces an additional consistency error, leading to the following additional term in the error bound inferred in Lemma 5.9:

$$\sup_{\mathbf{v} \in \mathcal{P}; \|\mathbf{v}\|_{\mathcal{P}, \#a} = 1} |(\mathbb{S} - \mathbb{S}^h)(s, u_D; \mathbf{v})|.$$

Then, the error estimate from Theorem 5.10 still holds with the following proposition.

Proposition 5.19. *Assume that the boundary mesh M is compatible in the sense of Definition 5.18. Assume that $s \in H^2(\mathsf{C})$ and that $u_D \in H^{3/2}(\mathsf{F}^\partial)$. Then,*

$$\sup_{\mathbf{v} \in \mathcal{P}; \|\mathbf{v}\|_{\mathcal{P}, \#a} = 1} |(\mathbb{S} - \mathbb{S}^h)(s, u_D; \mathbf{v})| \lesssim \left(\sum_{c \in \mathsf{C}} \tau h_c^4 |s|_{H^2(c)}^2 + \sum_{f \in \mathsf{F}^\partial \cap \partial\Omega^-} \|\beta\|_{L^\infty(f)} h_{c(f)}^3 |u_D|_{H^{3/2}(f)}^2 \right)^{\frac{1}{2}}.$$

Remark 5.20 (Approximation of the source term). *Note that it is possible to treat independently the two right-most terms in (5.49), with corresponding modifications in the estimate of Proposition 5.19.*

Implementation of the stabilizing term. Numerically, we observe that the scheme (5.16) is sensitive to the choice of the user-dependent parameter γ . To reduce this dependency and to avoid numerical divisions by zero if the absolute value of the advective field reduces to 0 at some mesh cell coordinates, we numerically consider instead of definition (5.12) the following expression:

$$\forall c \in \mathsf{C}, \quad \mathsf{s}_{\beta; c}(\mathbf{v}, \mathbf{w}) = h^2 \frac{h_c r_c}{h_{\mathsf{F}; c} r_{\mathsf{F}; c}} |\beta_c| \int_{\mathcal{J}_{\mathsf{EF}, c}} \left(\frac{\beta_c}{|\beta_c|} \cdot \llbracket \mathsf{L}_{\mathcal{P}_c}(\mathbf{v}) \rrbracket \right) \left(\frac{\beta_c}{|\beta_c|} \cdot \llbracket \mathsf{L}_{\mathcal{P}_c}(\mathbf{w}) \rrbracket \right), \quad (5.50)$$

for all $\mathbf{v}, \mathbf{w} \in \mathcal{P}_c$, where h_c and r_c denotes the diameter of the circumscribed and inscribed ball in c , respectively, and where we have defined $h_{\mathsf{F}; c} = \max_{f \in \mathsf{F}_c} h_f$ and $r_{\mathsf{F}; c} = \min_{f \in \mathsf{F}_c} r_f$. Owing to mesh regularity, we have $\frac{h_c r_c}{h_{\mathsf{F}; c} r_{\mathsf{F}; c}} = \mathcal{O}(1)$ as numerically checked for all mesh sequences H , PrT , PrG and CB (whose definition is recalled below), so that the stabilization is less dependent on the shape of the mesh cells.

5.4.2 Computational settings

Hereafter, we compare the numerical results obtained for the test case 1, 2 and 3 presented in Chapters 3 and 4 with the numerical solution of the two discrete problems (5.16) and (5.41). We recall that the computational domain corresponds to $\Omega = [0, 1]^3$ and that we consider the four mesh sequences denoted by H , PrT , PrG and CB (see Figure 5.4). Following the analysis performed in Section 5.2.2 and 5.3.2, we study the convergence of our schemes by computing the following relative discrete L^2 -norm on $\mathcal{V}$

$$\mathbf{Er}_{\mathcal{V}}(u) := \left(\frac{\sum_{\mathbf{v} \in \mathcal{V}} |\tilde{c}(\mathbf{v})| \left(\mathbf{u}_{\mathbf{v}} - \mathsf{R}_{\mathcal{V}}(u)|_{\mathbf{v}} \right)^2}{\sum_{\mathbf{v} \in \mathcal{V}} |\tilde{c}(\mathbf{v})| \mathsf{R}_{\mathcal{V}}(u)|_{\mathbf{v}}^2} \right)^{\frac{1}{2}}, \quad (5.51)$$

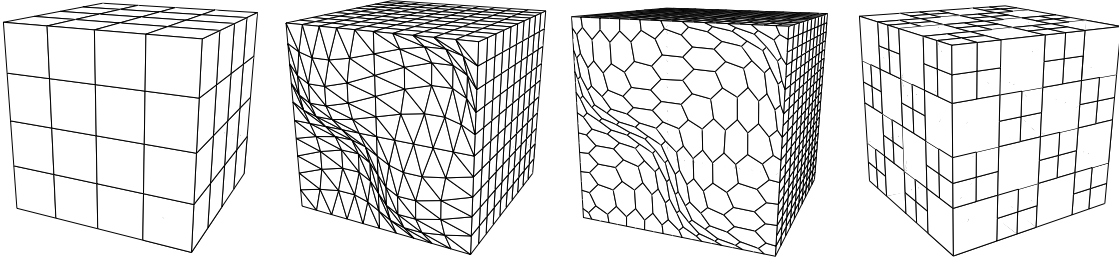


Figure 5.4 – The four mesh sequences H, PrT, PrG and CB, respectively.

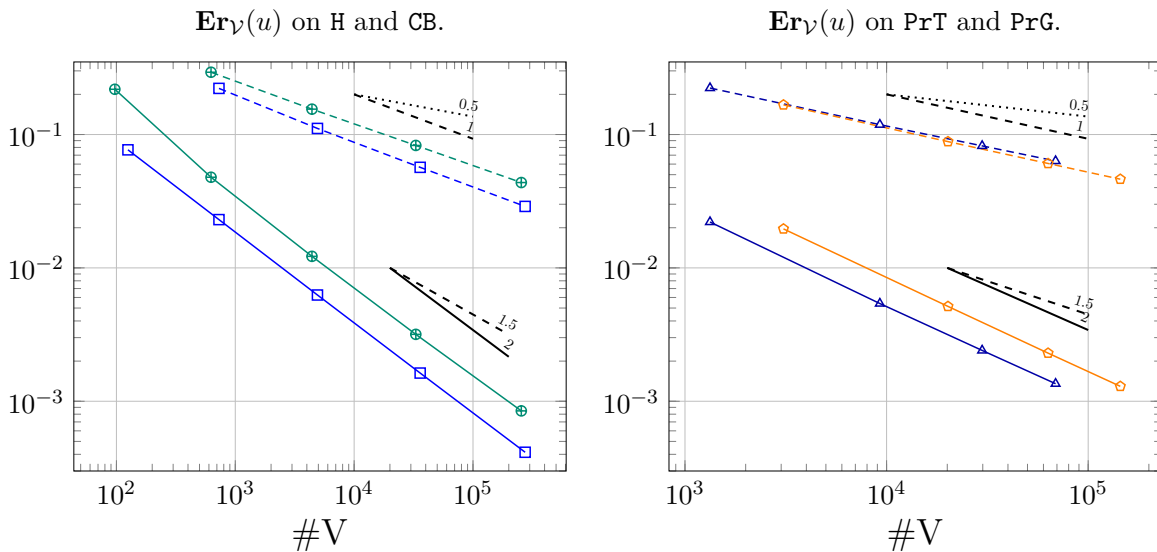
where u is the exact solution and $\mathbf{u}$ the discrete solution. We always statically condensate cell-based unknowns, even when the Proposition 5.17 is not satisfied (for the CB mesh sequence), since it significantly reduces the size of the final system. We also recall the definition of the computational cost $\mathbf{Co} = \mathbf{nnz} \times n_{\text{ite}}$ used to measure the computational efficiency of the scheme, where $\mathbf{nnz}$ is the total number of non-zeroes in the final system matrix and where n_{ite} is the number of iterations needed to bring the relative residual below 10^{-12} , using a bi-Conjugate gradient method preconditioned with an incomplete LU factorization.

5.4.3 Test case 1. Rotating advective field

This section presents the numerical results obtained with the advection-reaction scheme (5.16) for the three-dimensional test case considered in Section 3.3.2, where the exact solution u and the advective field β are defined as

$$u(x, y, z) = \sin(\pi x) \sin(2\pi y) \sin(\pi z) \quad \text{and} \quad \beta = (y - 1/2, 1/2 - x, z + 1)^T. \quad (5.52)$$

Even if the analysis has only been performed under assumption $(\mathcal{H})$ where we assume that the Friedrichs tensor satisfies $\text{ess inf}_{\Omega} \sigma_{\beta, \mu} > 0$ (see Remark 5.8), numerical results are presented when the reaction coefficient is given by $\mu \in \{5, 0.5\}$, yielding $\sigma_{\beta, \mu} = \frac{9}{2}$ and $\sigma_{\beta, \mu} = 0$, respectively. The discrete relative error $\mathbf{Er}_V(u)$ with respect to $\#V$ is represented on Figures 5.5 and 5.6 for the two schemes (3.21) and (5.16), when $\mu = 5$ and $\mu = 0.5$. The legend is collected in Table 5.1.


 Figure 5.5 – Test case 1. Numerical results for the discrete problems (3.21) and (5.16) with $\mu = 5$.

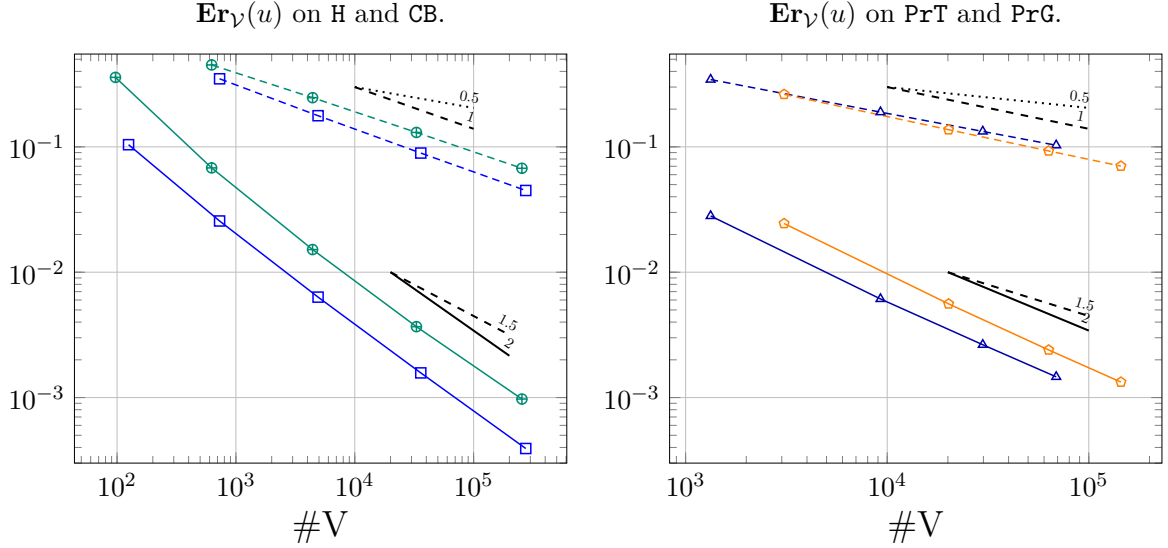


Figure 5.6 – Test case 1. Numerical results for the discrete problems (3.21) and (5.16) with $\mu = 0.5$.

Scheme (3.21)				Scheme (5.16)			
H	PrT	PrG	CB	H	PrT	PrG	CB

Table 5.1 – Test case 1. Legend of Figures 5.5 and 5.6.

Accuracy. The numerical results represented in Figures 5.5 and 5.6 show that the discrete error $\mathbf{Er}_V(u)$ obtained for the two schemes (3.21) and (5.16) converges for all mesh sequences. For the discrete problem (3.21), one retrieves the conclusions of Section 3.3.2, stating that the convergence rates are closer to 1 than to $\frac{1}{2}$. Regarding now the error obtained with the discrete problem (5.16), a super-convergence behavior is also observed since the convergence rates are closer to 2 than to $\frac{3}{2}$. This behavior is not surprising since they are also observed on simplicial meshes using the $\mathbb{P}_1$ Lagrange finite element method stabilized with the CIP approach; see Burman (2005). Comparing now the accuracy of the two schemes (5.16) and (3.21), we observe that for all mesh sequences, the gain is almost of two orders on the finest meshes when using the scheme (5.16). In terms of accuracy, the hierarchy is then $H > \text{PrT} > \text{PrG} > \text{CB}$. Since all these observations hold in both cases $\mu = 5$ and $\mu = 0.5$, these results are encouraging for the validity of the inf-sup Lemma 5.6 under assumption $(\mathcal{H}')$.

Efficiency. In the left panel of Figure 5.7, we compare the computational efficiency of the two schemes (5.16) and (3.21) by plotting $\mathbf{Er}_V(u)$ with respect to the computational cost Co . Not surprisingly, the scheme (5.16) is more efficient than the scheme (3.21). For a fixed computational cost, the scheme (5.16) reduces the discrete relative error by more than one order on all the finest meshes. We also observe that the total number of non-zeroes nnz for these two methods is identical, so that this figure indicates that (5.16) is slightly more expensive, since we need more iterations. This observation says in particular that the matrix realization of (3.21) is better conditioned than that of (5.16). The hierarchy in terms of efficiency is the same for the two schemes: $H > \text{PrT} > \text{PrG} > \text{CB}$. In the right panel of Figure 5.7, we compare the size of the final system matrix for dG methods of order $k = 1$, where 4 dofs per mesh cells are considered, with the scheme (5.16), that considers only 1 dof per mesh vertices. Clearly, for all mesh sequences, the ratio $4\#C/\#V$ is greater than 1 indicating that the dofs positioning in the scheme (5.16) is more competitive regarding the size of the final system. The advantage is

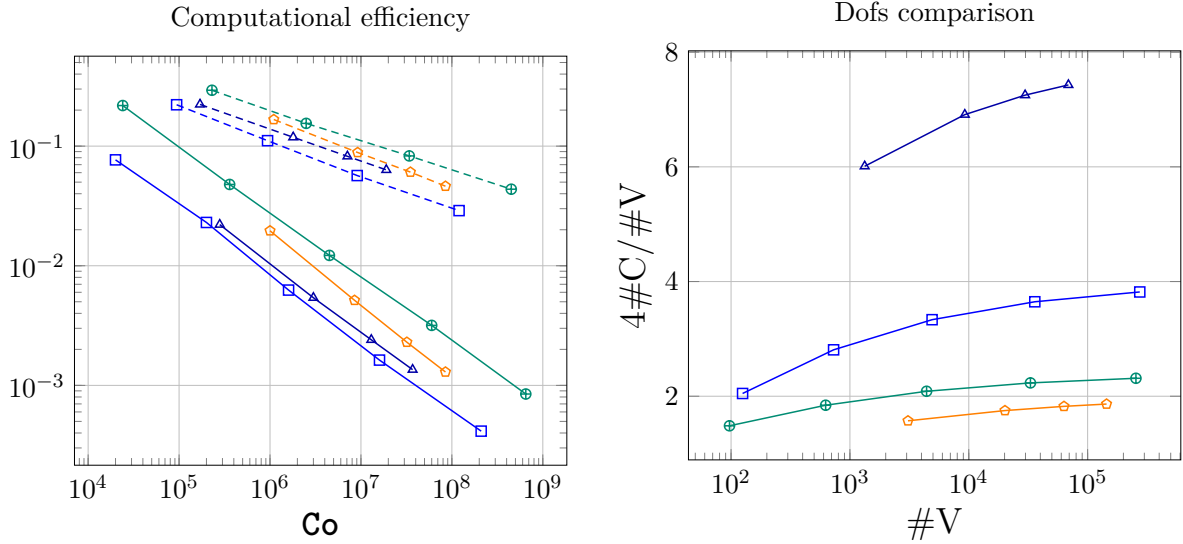


Figure 5.7 – Test case 1. Left panel: Computational efficiency of the two schemes (5.16) and (3.21). Right Panel: Comparison with dG methods of order $k = 1$ regarding the total number of dofs.

in particular substantial for the PrT mesh sequence, containing the lowest number of vertices per mesh cell.

Discrete min./max. principle. In Tables 5.2 and 5.3, we collect results on the discrete minimum (denoted **Min**) and the discrete maximum (denoted **Max**) principle for the scheme (5.16) when $\mu = 5$ and $\mu = 0.5$, respectively. We write **Y** if this principle is satisfied. If not, we indicate with a precision of order 10^{-1} , how much the bound is not respected with respect to the magnitude of the exact solution.

H		PrT		PrG		CB	
Min	Max	Min	Max	Min	Max	Min	Max
2.3%	2.3%	1.6%	2.0%	1.5%	1.4%	1.1%	1.1%
0.6%	0.6%	0.4%	0.4%	0.4%	0.4%	0.6%	0.6%
0.2%	0.2%	0.2%	0.3%	0.2%	0.2%	0.2%	0.2%
Y	Y	0.1%	0.2%	0.1%	0.1%	0.1%	0.1%

Table 5.2 – Discrete minimum and maximum principle for the scheme (5.16) with $\mu = 5$.

H		PrT		PrG		CB	
Min	Max	Min	Max	Min	Max	Min	Max
2.5%	2.5%	1.7%	1.9%	1.7%	1.1%	1.8%	1.8%
0.6%	0.6%	0.4%	0.2%	0.4%	0.3%	0.7%	0.7%
0.2%	0.2%	0.2%	0.3%	0.2%	0.2%	0.2%	0.2%
Y	Y	0.1%	0.2%	0.1%	0.1%	0.1%	0.1%

Table 5.3 – Discrete minimum and maximum principle for the scheme (5.16) with $\mu = 0.5$.

We recall that the discrete solutions obtained with the scheme (3.21) respect this principle on all mesh sequences. Here, the conclusion slightly differs since this principle is never exactly satisfied. However, we observe that refining the mesh reduces the violation of this principle.

5.4.4 Test case 2. Sharp internal layer

This section considers the test case 2 presented in Section 3.3.3. We recall that the exact solution is defined as

$$u(x, y, z) = xy \tanh\left(\frac{x}{2a} + \frac{y + z - \sqrt{2}}{a}\right) \quad (5.53)$$

with $a = 0.05$, that the advective field is $\beta = (1, 1, 0)^\top$ and that the reaction coefficient is set to 0. We also recall that the solution u presents an internal layer in the vicinity of the hyper plane $x + 2y + 2z = 2\sqrt{2}$. The discrete relative error $\mathbf{Er}_V(u)$ is represented in Figure 5.8 with respect to $\#V$.

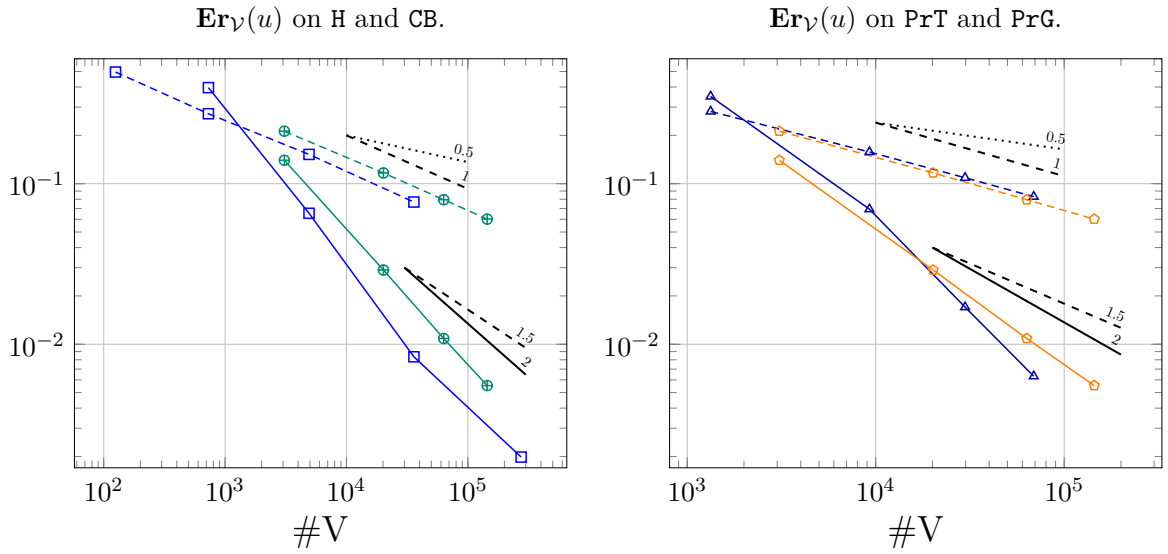


Figure 5.8 – Test case 2. Numerical errors with the schemes (5.16) and (3.21).

Accuracy and discrete min./max. principle. Regarding the accuracy and the convergence rates of the two schemes (5.16) and (3.21), the same observations as for the previous test case hold. We observe that the scheme (5.16) leads to lower errors than the scheme (3.21) and that the convergence rates are closer to 2 than to $\frac{3}{2}$. However, Table 5.4 shows that the discrete minimum/maximum principle are not satisfied on the coarser meshes. This can be explained by the sharp variation of the exact solution that are not correctly captured on these meshes. When we consider finer meshes, this principle tends to be satisfied. By comparison, we recall that this principle is satisfied on all mesh sequences using the scheme (3.21), excepted on the finer CB meshes, where the maximum principle is violated by up to 4.3%.

H		PrT		PrG		CB	
Min	Max	Min	Max	Min	Max	Min	Max
50.5%	29.7%	28.1%	11.0%	24.3%	11.8%	25.9%	19.2%
12.2%	2.1%	14.7%	0.7%	17.3%	1.3%	26.8%	10.4%
0.3%	Y	5.7%	0.4%	11.8%	0.6%	11.6%	0.6%
0.1%	Y	0.6%	0.1%	5.1%	0.3%	1.4%	Y

Table 5.4 – Test case 2. Discrete minimum and maximum principle with the scheme (5.16).

5.4.5 Test case 3. Anisotropic diffusion tensor and rotating advective field

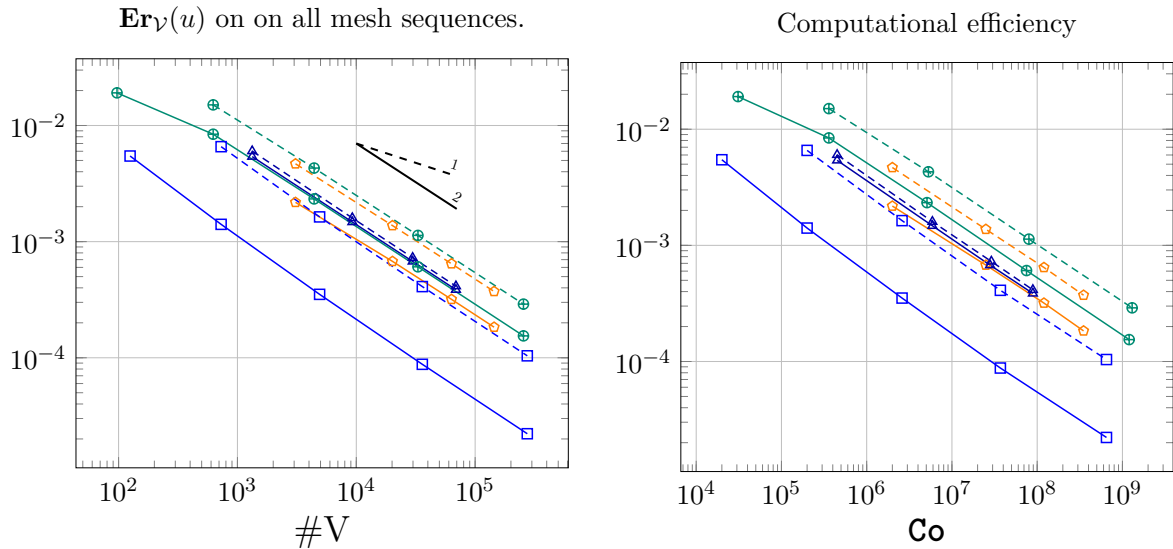
This third test case evaluates the numerical behavior of the scheme (5.41) approximating the solution of the diffusion-advection-reaction test case presented in Section 4.4.1. We recall that the exact solution is a combination of sine functions

$$u(x, y, z) = 1 + \sin(\pi x) \sin\left(\pi\left(y + \frac{1}{2}\right)\right) \sin\left(\pi\left(z + \frac{1}{3}\right)\right), \quad (5.54)$$

and that we consider the following anisotropic diffusion tensor λ and rotating advective field β :

$$\lambda = \begin{pmatrix} 1 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{pmatrix} \quad \text{and} \quad \beta = \begin{pmatrix} y - 1/2 \\ 1/2 - x \\ -z \end{pmatrix}. \quad (5.55)$$

The discrete relative error $\mathbf{Er}_V(u)$ with respect to $\#V$ and to the computational cost $\mathbf{Co}$ are represented in the left and in the right panel of Figure 5.9, respectively. The legend is collected in Table 5.5.


 Figure 5.9 – Test case 3. Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost $\mathbf{Co}$ for the discrete problems (5.41) and (4.34).

Accuracy. The left panel of Figure 5.9 show that the discrete error $\mathbf{Er}_V(u)$ obtained for the two schemes (5.41) and (4.34) converges for all mesh sequences at the order 2. Moreover, the scheme (5.41) leads to lower discrete relative error on H, PrG and CB mesh sequences. Even if the treatment of the diffusion term is not exactly the same in the two schemes (4.34) and (5.41),

Scheme (5.41)				Scheme (4.34)			
H	PrT	PrG	CB	H	PrT	PrG	CB
							

Table 5.5 – Legend of Figure 5.9.

this figure illustrates that the scheme (5.41) approximates more accurately the advection term. For the PrT mesh sequence, this advantage is less pronounced. Observe also that the mesh sequence CB yields more accurate results with the scheme (5.41) than the PrG mesh sequence with the scheme (5.41), which is the contrary of the scheme (4.34). This can be a consequence of the regularity of the supports of the reconstruction maps L_V and L_P on these two mesh sequences.

Discrete min./max. principle and stencil. We observe that the minimum/maximum principle is preserved at the precision 10^{-4} for all meshes. Finally, Table 5.6 collect the stencils of our scheme. We denote $\overline{St}$ the means stencil, defined as $nnz/\#V$ and $St.max$ the maximum stencil. In particular, we observe that the two schemes (4.34) and (5.41) have the same stencil.

H			PrT			PrG			CB		
$\#V$	$\overline{St}$	$St.max$	$\#V$	$\overline{St}$	$St.max$	$\#V$	$\overline{St}$	$St.max$	$\#V$	$\overline{St}$	$St.max$
7.3e+02	21	27	1.3e+03	18	21	3.1e+03	32	39	6.2e+02	36	93
4.9e+03	24	27	9.3e+03	19	21	2.0e+04	35	39	4.4e+03	43	93
3.6e+04	25	27	3.0e+04	20	21	6.3e+04	36	39	3.3e+04	46	93
2.7e+05	26	27	6.9e+04	20	21	1.4e+05	37	39	2.5e+05	48	93

Table 5.6 – Test case 5. Size of the system, mean stencil and maximum stencil.

Chapter 6

Edge-based scheme for vector advection-reaction

Contents

6.1	Discrete setting	88
6.1.1	Primal mesh and degrees of freedom	88
6.1.2	Edge-based partition	89
6.2	Advection-reaction problem	89
6.2.1	Discrete problem	90
6.2.2	Analysis	92
6.3	Numerical results	101
6.3.1	Test case 5. Vortex solution	102

This chapter devises and analyzes an edge-based scheme approximating the $\mathbb{R}^3$ -valued solution of the vector advection-reaction problem

$$\nabla(\boldsymbol{\beta} \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \boldsymbol{\beta} + \boldsymbol{\mu} \mathbf{u} = \mathbf{s} \quad \text{a.e. in } \Omega, \quad (6.1a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{a.e. on } \partial\Omega^-. \quad (6.1b)$$

Edge-based schemes are rarely addressed in the literature, although they are the natural way to discretize differential 1-forms in manifolds of $\mathbb{R}^3$ since they consider dofs attached to 1-chains. In the context of electromagnetism, we mention the work of Zaglmayr (2006) considering edge and face finite elements to solve the Maxwell problem and the work of Campos Pinto *et al.* (2014), proposing an edge finite element scheme for simulating plasma which preserves the charge balance at the discrete level. Edge elements are also relevant to solve the Stokes and the Navier-Stokes problems with *non-standard* boundary conditions; see Girault (1990). Starting from the differential geometry, Heumann *et al.* (2015) proposed for the continuous problem (6.1), an $\mathbf{H}(\text{curl}; \Omega)$ -conforming discretization using the edge finite elements on simplicial meshes with a stabilization penalizing jumps of the normal component of the discrete solution in the spirit of the interior penalty method, see Lesaint & Raviart (1974) or Johnson & Pitkäranta (1986). In the same paper, they also devise a scheme based on cell-based fully discontinuous polynomials. In the context of mimetic spectral element method, first introduced by Gerritsma (2012), Palha (2013) treats a similar problem on square meshes motivated by the discretization of the Lie derivative. Similarly, Mullen *et al.* (2011) devised an approximation of (6.1) using the edge-based extrusion process, along the flow lines of the vector field $\boldsymbol{\beta}$, proposed by Bossavit (2003).

To our knowledge, there does not exist in the literature any edge-based scheme approximating the problem (6.1) on polyhedral meshes. One reason to consider dofs attached to mesh edges is the possibility to combine the present scheme with the CDO scheme proposed

by Bonelle & Ern (2014b) for the Stokes problem, where the velocity is attached to mesh edges and the pressure to mesh vertices. This would then lead to a discretization of the Oseen problem using edge-based and vertex-based degrees of freedom on polyhedral meshes.

This chapter contains two main contributions. The first one is to provide an edge-based scheme, with scalar dofs attached to mesh edges, with an $\mathcal{O}(h^p)$ convergence rate as soon as the solution belongs to $\mathbf{W}^{1,p}(\Omega)$ with $p \in (\frac{3}{2}, 2]$. To reach this goal, we consider an edge-based reconstruction map defining piece-wise constant polynomials on the diamond cells surrounding the mesh edges. Then, the scheme stems from a discrete Galerkin formulation stabilized by penalizing jumps on sub-faces of these edge diamond sub-cells. Note that there is a substantial difference with the finite volume approach since we do not consider one dof per vector component on all mesh edge diamonds but only one dof per mesh edges.

Following the continuous analysis proposed in Section 2.3, the second main contribution of this chapter is to extend at the discrete level the stability analysis when the $\mathbb{R}^{3 \times 3}$ -valued Friedrichs tensor takes null or slightly negative values. Under this assumption, the analysis is more elaborated since the stability now hinges on an inf-sup condition which is satisfied if the mesh size is smaller than a reference length that linearly depends on $\|\nabla \beta^\top + \mu\|_{L^\infty(\Omega)}^{-1}$. In particular, if $\mu = -\nabla \beta^\top$, the scheme is unconditionally stable.

The content of this chapter is an extended version of the paper "*Edge-based low-order scheme on polyhedral meshes for vector-reaction equations*", by P. Cantin & A. Ern, submitted to *ESAIM: Mathematical Modeling and Numerical Analysis*, 2016.

6.1 Discrete setting

This section recalls and introduces new concepts and notation to devise and analyze our edge-based scheme approximating the solution of (6.1).

6.1.1 Primal mesh and degrees of freedom

We assume as before that Ω denotes an open, bounded, connected, polyhedral subset of $\mathbb{R}^3$ and we define $\mathcal{M} = \{\mathcal{V}, \mathcal{E}, \mathcal{F}, \mathcal{C}\}$ a primal mesh of Ω collecting vertices $v \in \mathcal{V}$, edges $e \in \mathcal{E}$, faces $f \in \mathcal{F}$ and cells $c \in \mathcal{C}$. Boundary faces are collected in the set $\mathcal{F}^\partial = \{f \in \mathcal{F} \mid f \subset \partial\Omega\}$. The mesh $\mathcal{M}$ is regular in the sense that it defines a planar cellular complex, its entities are star-shaped with respect to their barycenter, and there exists a simplicial sub-mesh that is shape-regular in the usual sense of Ciarlet (1978). In essence, the mesh satisfies **(C)**, **(St)** and **(Sh)** (see Chapter 3). For any mesh entity $x \in \mathcal{M}$, $\mathbf{x}_x$ denotes its barycenter.

In this chapter, we consider scalar degrees of freedom (dofs) attached to mesh edges collected in the finite-dimensional space $\mathcal{E} \equiv \mathbb{R}^{\#\mathcal{E}}$. For all $v \in \mathcal{E}$, v_e denotes the entry attached to $e \in \mathcal{E}$. For any mesh cell $c \in \mathcal{C}$, $\mathcal{E}_c$ denotes the restriction of $\mathcal{E}$ collecting the dofs attached to the edges $e \in \mathcal{E}_c$.

Two different reduction maps attached to mesh edges are considered in this chapter. The first one is the classical de Rham map $R_{\mathcal{E}} : \mathbf{W}^{s,p}(\Omega) \rightarrow \mathcal{E}$ with $sp > 2$ acting as follows:

$$\forall e \in \mathcal{E}, \quad R_{\mathcal{E}}(\mathbf{w})|_e := \int_e \mathbf{w} \cdot \mathbf{t}_e, \quad \forall \mathbf{w} \in \mathbf{W}^{s,p}(\Omega), \quad (6.2)$$

where we recall that $\mathbf{t}_e$ denotes a fixed unit tangential vector orienting the edge e . The local reduction map in c is denoted by $R_{\mathcal{E}_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathcal{E}_c$ with $sp > 2$ and acts as the global reduction map $R_{\mathcal{E}}$. The second reduction map considered in this chapter is defined by $\hat{R}_{\mathcal{E}} : \mathbf{L}^1(\Omega) \rightarrow \mathcal{E}$ and acts as follows:

$$\forall e \in \mathcal{E}, \quad \hat{R}_{\mathcal{E}}(\mathbf{w})|_e := \frac{|e|}{|\mathbf{c}_e|} \int_{\mathbf{c}_e} \mathbf{w} \cdot \mathbf{t}_e, \quad \forall \mathbf{w} \in \mathbf{L}^1(\Omega), \quad (6.3)$$

where we have introduced the macro-diamond $\mathbf{c}_e := \cup\{\mathbf{c}_{e,c} \mid c \in \mathcal{C}_e\}$ collecting the local diamonds $\mathbf{c}_{e,c}$ defined below. The local version of this map in the cell $c \in \mathcal{C}$ is denoted by

$\widehat{R}_{\mathcal{E}_c} : \mathbf{L}^1(\widehat{c}) \rightarrow \mathcal{E}_c$ with the patch cell $\widehat{c} = \cup\{\mathbf{c}_e \mid e \in E_c\}$ and acts similarly to the global operator (6.3).

Remark 6.1 (Regularity). *The operator (6.3) is important to establish a quasi-optimal convergence rate without requiring too much regularity on the exact solution. Typically, low-order methods for first-order problems require that the exact solution is H^1 to achieve a convergence rate of order $\frac{1}{2}$ in L^2 -norm, and the reduction map (6.2) is not well-defined in $\mathbf{H}^1(\Omega)$.*

6.1.2 Edge-based partition

In addition to the primal mesh, our scheme also considers the edge-based partition already presented in Chapter 4 for the treatment of diffusion terms in the context of CDO schemes. For all $c \in \mathcal{C}$, $\mathfrak{C}_{e,c}$ denotes the partition of c composed of the diamonds $\{\mathbf{c}_{e,c}\}_{e \in E_c}$ such that

$$\mathbf{c}_{e,c} = \text{int} \left(\bigcup_{f \in F_e \cap F_c} \bigcup_{v \in V_e} \overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\} \right), \quad \forall e \in E_c, \quad (6.4)$$

and shown in the left panel of Figure 6.1. Observe that $\#\mathfrak{C}_{e,c} = \#E_c$ and that $\bar{c} = \cup\{\mathbf{c}_{e,c} \mid e \in E_c\}$, owing to assumption (St). For all $e \in E_c$, the boundary of the diamonds $\mathbf{c}_{e,c}$ is composed of four *intra-cell sub-faces* and one *inter-cell sub-face*. Intra-cell sub-faces are collected in the set $\mathfrak{F}_{e,c}$, defined as

$$\mathfrak{F}_{e,c} := \{f_{vf,c} = \text{int}(\partial\mathbf{c}_{e,c} \cap \partial\mathbf{c}_{e',c}) \mid f \in F_c, v \in V_f \text{ and } e, e' \in E_f \cap E_v\}, \quad (6.5)$$

where we recall that $\text{int}(\omega)$ denotes the interior of any set $\omega \subset \mathbb{R}^3$. This set collects the sub-faces shared by two diamonds connected by one vertex; see the middle panel of Figure 6.1. For all $f \in F_c$, inter-cell sub-faces are collected in the set $\mathfrak{F}_{e,f}$ defined as

$$\mathfrak{F}_{e,f} := \{f_{e,f} = \text{int}(\partial\mathbf{c}_{e,c} \cap \partial\mathbf{c}_{e',c'}) \mid e \in E_f \text{ and } c, c' \in \mathcal{C}_f\}, \quad (6.6)$$

collecting the sub-faces shared by two diamonds connected by the mesh face f ; see the right panel of Figure 6.1.

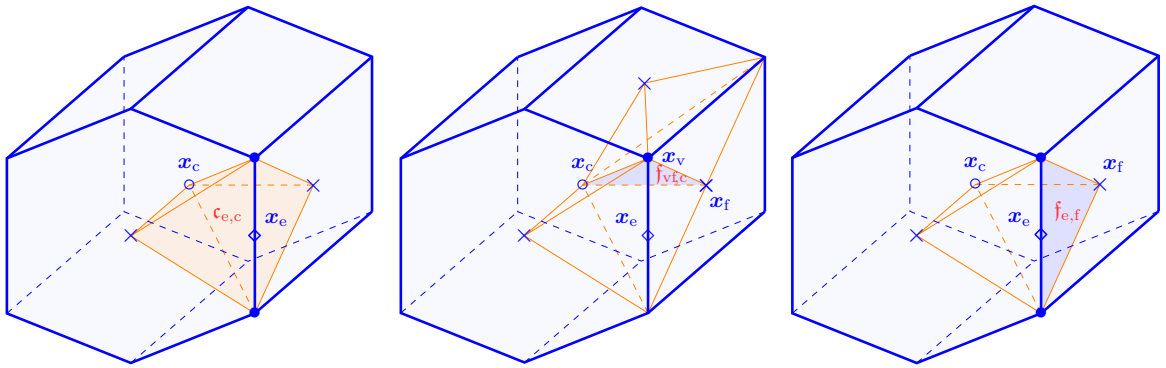


Figure 6.1 – From left to right. One edge diamond sub-cell $\mathbf{c}_{e,c} \in \mathfrak{C}_{e,c}$. One intra-cell sub-face $f_{vf,c} \in \mathfrak{F}_{e,c}$. One inter-cell sub-face $f_{e,f} \in \mathfrak{F}_{e,f}$.

6.2 Advection-reaction problem

This section is concerned with the derivation and the analysis of our edge-based scheme approximating the solution $\mathbf{u} : \Omega \rightarrow \mathbb{R}^3$ of the vector advection-reaction problem

$$\nabla(\beta \cdot \mathbf{u}) + (\nabla \times \mathbf{u}) \times \beta + \mu \mathbf{u} = \mathbf{s} \quad \text{a.e. in } \Omega, \quad (6.7a)$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{a.e. on } \partial\Omega^-, \quad (6.7b)$$

with the data $\mathbf{s} \in \mathbf{L}^2(\Omega)$ and $\mathbf{u}_D \in \mathbf{L}^2(\partial\Omega^-; |\boldsymbol{\beta} \cdot \mathbf{n}|)$. The advective field $\boldsymbol{\beta} : \Omega \rightarrow \mathbb{R}^3$ is assumed to be Lipschitz-continuous and the reaction tensor $\boldsymbol{\mu} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ is assumed to be bounded and not necessarily symmetric. Following the analysis in Section 2.3, the Friedrichs tensor $\boldsymbol{\sigma}_{\boldsymbol{\beta}, \boldsymbol{\mu}} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ is defined as

$$\boldsymbol{\sigma}_{\boldsymbol{\beta}, \boldsymbol{\mu}} = \frac{\nabla \boldsymbol{\beta} + \nabla \boldsymbol{\beta}^\top}{2} - \frac{1}{2}(\nabla \cdot \boldsymbol{\beta}) \text{Id} + \frac{\boldsymbol{\mu} + \boldsymbol{\mu}^\top}{2}, \quad (6.8)$$

whose lowest-eigenvalue is denoted by

$$\aleph(\mathbf{x}) = \min_{\mathbf{y} \in \mathbb{R}^d \setminus \{\mathbf{0}\}} \frac{(\boldsymbol{\sigma}_{\boldsymbol{\beta}, \boldsymbol{\mu}}(\mathbf{x}) \mathbf{y}, \mathbf{y})}{|\mathbf{y}|^2}, \quad \forall \mathbf{x} \in \Omega, \quad (6.9)$$

where $(\cdot, \cdot)$ denotes the classical Euclidean inner product in $\mathbb{R}^3$. In this chapter, the analysis is performed under one of the two following assumptions:

- ($\mathcal{H}$) $\text{ess inf}_\Omega \aleph > 0$. We define the reference time $\tau = (\text{ess inf}_\Omega \aleph)^{-1}$.
- ($\mathcal{H}'$) $-\mathbf{C}_\aleph < \text{ess inf}_\Omega \aleph \leq 0$, where $\mathbf{C}_\aleph$ is a constant defined by (6.32) below, and there exists a non-dimensional function $\zeta \in \text{Lip}(\Omega)$ satisfying $\zeta \geq 1$ and

$$\text{ess inf}_\Omega \left(-\frac{1}{2} \boldsymbol{\beta} \cdot \nabla \zeta \right) > -\text{ess inf}_\Omega (\zeta \aleph) \geq 0. \quad (6.10)$$

We define the reference time $\tau = \left(\text{ess inf}_\Omega \left(-\frac{1}{2} \boldsymbol{\beta} \cdot \nabla \zeta \right) \right)^{-1}$.

Owing to Theorem 2.17, the continuous problem (6.7) is well-posed since ($\mathcal{H}'$) implies ($\mathcal{H}'_2$). Indeed, observe that

$$\text{ess inf}_\Omega \left(\zeta \aleph - \frac{1}{2} \boldsymbol{\beta} \cdot \nabla \zeta \right) \geq \text{ess inf}_\Omega (\zeta \aleph) + \text{ess inf}_\Omega \left(-\frac{1}{2} \boldsymbol{\beta} \cdot \nabla \zeta \right), \quad (6.11)$$

which is strictly positive owing to (6.10), so that assumption ($\mathcal{H}'_2$) holds.

Remark 6.2 (On assumption ($\mathcal{H}'$)). *Compared to the assumption ($\mathcal{H}'$) considered in Chapters 3 and 4, assumption ($\mathcal{H}'$) is more general since it allows the Friedrichs tensor to take null or negative values. In addition, observe that we have assumed that $\zeta \geq 1$. This assumption simplify the proofs in the discrete analysis and changing the lower bound only change the numerical constant.*

6.2.1 Discrete problem

Our edge-based scheme hinges on the bilinear form $\mathbb{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}} : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$ such that

$$\mathbb{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^\mathcal{E}(\mathbf{v}, \mathbf{w}) = \mathbf{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^\mathcal{E}(\mathbf{v}, \mathbf{w}) + \sum_{\mathbf{f} \in \mathbf{F}^\partial} \langle \mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n})^\ominus; \mathbf{f}}^{\mathcal{E}, \partial}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{E}_{c(\mathbf{f})}}, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.12)$$

where the first term $\mathbf{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^\mathcal{E}$ approximates (6.7a) and the second term weakly enforces the boundary condition (6.7b).

Bilinear forms and reconstruction maps. The bilinear form $\mathbf{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^\mathcal{E} : \mathcal{E} \times \mathcal{E} \rightarrow \mathbb{R}$ is composed of three bilinear forms, also defined on $\mathcal{E} \times \mathcal{E}$, as follows:

$$\mathbf{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^\mathcal{E}(\mathbf{v}, \mathbf{w}) := \mathbf{g}_{\boldsymbol{\beta}, \boldsymbol{\mu}}(\mathbf{v}, \mathbf{w}) + \mathbf{n}_\beta(\mathbf{v}, \mathbf{w}) + \gamma \mathbf{s}_\beta(\mathbf{v}, \mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.13)$$

where $\gamma \in [\gamma_b, 1]$ is a user-dependent penalty parameter with $\gamma_b > 0$. The bilinear form $\mathbf{g}_{\boldsymbol{\beta}, \boldsymbol{\mu}}$ is assembled cell-wise as

$$\mathbf{g}_{\boldsymbol{\beta}, \boldsymbol{\mu}}(\mathbf{v}, \mathbf{w}) = \sum_{\mathbf{c} \in \mathbf{C}} \mathbf{g}_{\boldsymbol{\beta}, \boldsymbol{\mu}; \mathbf{c}}(\mathbf{v}, \mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.14)$$

where each local bilinear form $\mathbf{g}_{\beta,\mu;c} : \mathcal{E}_c \times \mathcal{E}_c \rightarrow \mathbb{R}$ corresponds to the standard Galerkin approximation of (6.1a) in the cell c :

$$\mathbf{g}_{\beta,\mu;c}(\mathbf{v}, \mathbf{w}) = \sum_{c \in \mathfrak{C}_{E,c}} \int_c \left(\nabla(\beta \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})) + (\nabla \times \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})) \times \beta \right) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{w}) + \langle \mathbf{H}_{\mu;c}^\varepsilon \mathbf{v}, \mathbf{w} \rangle_{\mathcal{E}_c}, \quad (6.15)$$

for all $\mathbf{v}, \mathbf{w} \in \mathcal{E}_c$, using the DGA reconstruction map $\mathbf{L}_{\mathcal{E}_c} : \mathcal{E}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{E,c}; \mathbb{R}^3)$ (see Section 4.2) such that

$$\forall \mathbf{x} \in c, \quad \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})(\mathbf{x}) := \sum_{e \in E_c} \mathbf{v}_e \ell_{e,c}(\mathbf{x}), \quad \forall \mathbf{v} \in \mathcal{E}_c. \quad (6.16)$$

The local shape functions $\{\ell_{e,c}\}_{e \in E_c}$ span $\mathbb{P}_0(\mathfrak{C}_{E,c}; \mathbb{R}^3)$ and are defined as

$$\forall e, e' \in E_c, \quad \ell_{e,c}|_{\mathfrak{c}_{e',c}} = \left(\text{Id} - \frac{\tilde{\mathbf{f}}_c(e') \otimes \mathbf{e}'}{d|\mathfrak{c}_{e',c}|} \right) \frac{\tilde{\mathbf{f}}_c(e)}{|c|} + \frac{\tilde{\mathbf{f}}_c(e)}{d|\mathfrak{c}_{e,c}|} \delta_{e,e'}, \quad (6.17)$$

with $d = 3$, $\mathbf{e} = \int_e \mathbf{t}_e$ and $\tilde{\mathbf{f}}_c(e) = \int_{\tilde{\mathfrak{f}}_c(e)} \mathbf{n}_{\tilde{\mathfrak{f}}_c(e)}$ where $\tilde{\mathfrak{f}}_c(e) = \tilde{\mathfrak{f}}(e) \cap c$ denotes the local dual face on c attached to $e \in E_c$. Using this reconstruction map and recalling from Chapter 3 the definition of the inner product $\langle \mathbf{v}, \mathbf{w} \rangle_{\mathcal{E}_c} = \sum_{e \in E_c} \mathbf{v}_e \mathbf{w}_e$ for all $\mathbf{v}, \mathbf{w} \in \mathcal{E}_c$, the reaction-related term in (6.15) corresponds to

$$\langle \mathbf{H}_{\mu;c}^\varepsilon(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{E}_c} = \int_c \mu \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}_c. \quad (6.18)$$

Using classical identities from vector calculus, the bilinear form (6.15) can be reformulated as

$$\mathbf{g}_{\beta,\mu;c}(\mathbf{v}, \mathbf{w}) = \sum_{c \in \mathfrak{C}_{E,c}} \int_c \beta \cdot \nabla \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{w}) + \langle \mathbf{H}_{\nabla \beta}^\varepsilon \mathbf{v}, \mathbf{w} \rangle_{\mathcal{E}_c} = \langle \mathbf{H}_{\nabla \beta}^\varepsilon \mathbf{v}, \mathbf{w} \rangle_{\mathcal{E}_c}, \quad (6.19)$$

for all $\mathbf{v}, \mathbf{w} \in \mathcal{E}_c$, where the last equality follows since $\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})$ is piece-wise constant on the sub-mesh $\mathfrak{C}_{E,c}$. Because $\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})$ jumps across inter-cell and intra-cell sub-faces induced by the partition $\mathfrak{C}_{E,c}$, we consider in expression (6.13) the bilinear form $\mathbf{n}_\beta$ such that

$$\mathbf{n}_\beta(\mathbf{v}, \mathbf{w}) = \sum_{c \in C} \mathbf{n}_{\beta;c}(\mathbf{v}, \mathbf{w}) + \sum_{f \in F^\circ} \mathbf{n}_{\beta;f}(\mathbf{v}, \mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.20)$$

where the local bilinear forms $\mathbf{n}_{\beta;x}$, with $x = f$ or $x = c$, are defined as

$$\mathbf{n}_{\beta;x}(\mathbf{v}, \mathbf{w}) = - \sum_{f \in \mathfrak{F}_{E;x}} \int_f (\beta \cdot \mathbf{n}_f) [\![\mathbf{L}_{\mathcal{E}}(\mathbf{v})]\!] \cdot \langle \mathbf{L}_{\mathcal{E}}(\mathbf{w}) \rangle, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.21)$$

with $[\![\cdot]\!]$ and $\langle \cdot \rangle$ denoting the jump and average operator, respectively, and where the sets $\mathfrak{F}_{E;x}$ are defined by (6.5) and (6.6), for $x = c$ and $x = f$, respectively. The stabilization bilinear form $\mathbf{s}_\beta$ is built similarly as follows:

$$\mathbf{s}_\beta(\mathbf{v}, \mathbf{w}) = \sum_{c \in C} \mathbf{s}_{\beta;c}(\mathbf{v}, \mathbf{w}) + \sum_{f \in F^\circ} \mathbf{s}_{\beta;f}(\mathbf{v}, \mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}, \quad (6.22)$$

where each local bilinear form $\mathbf{s}_{\beta;x}$, with $x = f$ or $x = c$, penalizes the jump across the faces in the sets $\mathfrak{F}_{E;x}$ as follows:

$$\mathbf{s}_{\beta;x}(\mathbf{v}, \mathbf{w}) = \sum_{f \in \mathfrak{F}_{E;x}} \int_f |\beta \cdot \mathbf{n}_f| [\![\mathbf{L}_{\mathcal{E}}(\mathbf{v})]\!] \cdot [\![\mathbf{L}_{\mathcal{E}}(\mathbf{w})]\!], \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}. \quad (6.23)$$

Remark 6.3 (Centered and upwind fluxes). *The bilinear forms $\mathbf{n}_\beta$ and $\mathbf{s}_\beta$ are devised similarly to the discontinuous Galerkin method. The bilinear form $\mathbf{n}_\beta$ corresponds to the use of centered fluxes while the bilinear form $\mathbf{n}_\beta + \mathbf{s}_\beta$ corresponds to the use of upwind fluxes.*

Finally, the Dirichlet boundary condition (6.7b) is weakly enforced for all $f \in F^\partial$ using the boundary operator $\mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus;f}^{\varepsilon,\partial} : \mathcal{E}_c \rightarrow \mathcal{E}_c$, where we recall that $c \equiv c(f)$ denotes the unique cell containing the boundary face f :

$$\forall f \in F^\partial, \quad \langle \mathbf{H}_{(\beta \cdot \mathbf{n})^\ominus;f}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{E}_c} := \int_f (\beta \cdot \mathbf{n})^\ominus \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{w}), \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}_c. \quad (6.24)$$

Discrete problem. The discrete problem approximating the solution of (6.7) using edge-based scalar degrees of freedom reads:

$$\text{Find } \mathbf{u} \in \mathcal{E} \quad \text{s.t.} \quad \mathbf{A}_{\beta, \mu}^{\mathcal{E}}(\mathbf{u}, \mathbf{v}) = \mathbb{S}(\mathbf{s}, \mathbf{u}_D; \mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{E}, \quad (6.25)$$

with the source term linear form $\mathbb{S}(\mathbf{s}, \mathbf{u}_D; \cdot) : \mathcal{E} \rightarrow \mathbb{R}$ defined as

$$\mathbb{S}(\mathbf{s}, \mathbf{u}_D; \mathbf{v}) := \sum_{c \in \mathcal{C}} \int_c \mathbf{s} \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) + \sum_{f \in \mathcal{F}^\partial} \int_f (\boldsymbol{\beta} \cdot \mathbf{n})^\ominus \mathbf{u}_D \cdot \mathbf{L}_{\mathcal{E}_{c(f)}}(\mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{E}. \quad (6.26)$$

6.2.2 Analysis

This section is devoted to the analysis of the discrete problem (6.25). The dof space $\mathcal{E}$ is equipped with the following discrete 2-norm:

$$\|\mathbf{w}\|_{\mathcal{E}, 2}^2 = \sum_{c \in \mathcal{C}} \|\mathbf{w}\|_{\mathcal{E}_c, 2}^2 \quad \text{with} \quad \|\mathbf{w}\|_{\mathcal{E}_c, 2}^2 := \sum_{e \in \mathcal{E}_c} |\mathbf{c}_{e,c}| \left(\frac{|\mathbf{w}_e|}{|e|} \right)^2, \quad \forall \mathbf{w} \in \mathcal{E}. \quad (6.27)$$

Stability. The analysis of (6.25) relies on the Friedrichs assumptions $(\mathcal{H})$ or $(\mathcal{H}')$ defining the reference time $\tau > 0$. The coercivity norm on the dofs space $\mathcal{E}$ is defined as

$$\|\mathbf{w}\|_{\mathcal{E}, a}^2 := \tau^{-1} \|\mathbf{w}\|_{\mathcal{E}, 2}^2 + \gamma_b \mathbf{s}_\beta(\mathbf{w}, \mathbf{w}) + \langle \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|}^{\mathcal{E}, \partial}(\mathbf{w}), \mathbf{w} \rangle_{\mathcal{E}}, \quad \forall \mathbf{w} \in \mathcal{E}, \quad (6.28)$$

where $\mathbf{s}_\beta$ is defined by (6.22) and where $\langle \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|}^{\mathcal{E}, \partial}(\mathbf{w}), \mathbf{w} \rangle_{\mathcal{E}} = \sum_{f \in \mathcal{F}^\partial} \langle \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|; f}^{\mathcal{E}, \partial}(\mathbf{w}), \mathbf{w} \rangle_{\mathcal{E}_c}$, for all $\mathbf{w} \in \mathcal{E}$ with $c \equiv c(f)$, following the definition (6.24) with $(\boldsymbol{\beta} \cdot \mathbf{n})^\ominus$ in lieu of $|\boldsymbol{\beta} \cdot \mathbf{n}|$. An important result to prove the stability of the bilinear form (6.12) is the stability of the reconstruction maps $\mathbf{L}_{\mathcal{E}_c}$ for all $c \in \mathcal{C}$.

Proposition 6.4 (Stability of $\mathbf{L}_{\mathcal{E}_c}$). *There exists $\mathcal{C}_\sharp > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|\mathbf{w}\|_{\mathcal{E}_c, 2} \leq \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{w})\|_{L^2(c)} \leq \mathcal{C}_\sharp \|\mathbf{w}\|_{\mathcal{E}_c, 2}, \quad \forall \mathbf{w} \in \mathcal{E}_c.$$

Proof. See the proof of Proposition 7.39. □

Lemma 6.5 (Coercivity). *Assume that $(\mathcal{H})$ holds. Then,*

$$\mathbf{A}_{\beta, \mu}^{\mathcal{E}}(\mathbf{v}, \mathbf{v}) \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{E}, a}^2, \quad \forall \mathbf{v} \in \mathcal{E}.$$

Consequently, the discrete problem (6.25) is well-posed.

The stability of the bilinear form $\mathbf{A}_{\beta, \mu}^{\mathcal{E}}$ is now studied under assumption $(\mathcal{H}')$. First, we introduce the reference length

$$h_0 = \left(\mathcal{C}_\sharp^2 L_\zeta \tau \|\boldsymbol{\mu} + \nabla \boldsymbol{\beta}^\top\|_{L^\infty(\Omega)} \right)^{-1}, \quad (6.29)$$

with $\mathcal{C}_\sharp$ the constant appearing in the upper bound of the stability of $\mathbf{L}_{\mathcal{E}_c}$ in Proposition 6.4 and with $L_\zeta = |\zeta|_{W^{1, \infty}(\Omega)}$ the Lipschitz constant of ζ . If $\boldsymbol{\mu} = -\nabla \boldsymbol{\beta}^\top$, we set conventionally $h_0 = +\infty$. To avoid the proliferation of constants and to simplify the proofs in the analysis, we assume that there exists a constant $\mathcal{C}_{\mathcal{E}, a} > 0$, independent of the mesh size and the physical parameters, such that

$$L_\zeta \max \left(h, \|\boldsymbol{\beta}\|_{L^\infty(\Omega)} \tau \right) \leq \mathcal{C}_{\mathcal{E}, a}. \quad (6.30)$$

In addition, we assume that the minimal eigenvalue $\aleph$ of the Friedrichs tensor $\boldsymbol{\sigma}_{\beta, \mu}$ satisfies the following assumptions:

$$\frac{1}{4} + \vartheta \tau (\text{ess inf}_\Omega \aleph) > 0 \quad \text{and} \quad h < h_0 \left(\frac{1}{4} + \vartheta \tau (\text{ess inf}_\Omega \aleph) \right), \quad (6.31)$$

where the constant $\vartheta > 0$ linearly depends on the quantity $\|\zeta\|_{L^\infty(\Omega)} + \mathbf{C}_{\mathcal{E},a}^2$. From the first condition in (6.31), we observe that the constant in assumption $(\mathcal{H}')$ is defined as

$$\mathbf{C}_{\mathcal{N}} = (4\tau\vartheta)^{-1}. \quad (6.32)$$

By convention, the second condition in (6.31) is void if $\boldsymbol{\mu} = -\nabla\boldsymbol{\beta}^\top$.

Lemma 6.6 (Inf-sup stability). *Assume that $(\mathcal{H}')$, (6.30) and (6.31) hold. Then, there exists $\varrho > 0$ such that*

$$\sup_{\mathbf{w} \in \mathcal{E}; \|\mathbf{w}\|_{\mathcal{E},a}=1} \mathbb{A}_{\boldsymbol{\beta},\boldsymbol{\mu}}^\varepsilon(\mathbf{v}, \mathbf{w}) \geq \varrho \|\mathbf{v}\|_{\mathcal{E},a}, \quad \forall \mathbf{v} \in \mathcal{E}.$$

Consequently, the discrete problem (6.25) is well-posed.

Before presenting the proof of Lemmata 6.5 and 6.6, Table 6.1 recapitulates the different conditions for which the discrete problem (6.25) is well-posed.

Case	$\text{ess inf}_\Omega \mathcal{N} > 0$	$-\mathbf{C}_{\mathcal{N}} < \text{ess inf}_\Omega \mathcal{N} \leq 0$	
Assumption	$(\mathcal{H})$	$(\mathcal{H}')$	
Advective field		$\boldsymbol{\mu} = -\nabla\boldsymbol{\beta}^\top$	$\boldsymbol{\mu} \neq -\nabla\boldsymbol{\beta}^\top$
Mesh-size	$0 < h$	$0 < h$	$0 < h < h_0 (1 + \mathbf{C}_{\mathcal{N}}^{-1} \text{ess inf}_\Omega \mathcal{N}) / 4$

Table 6.1 – Stability of the discrete problem (6.25) with h_0 defined by (6.29).

Proof of Lemma 6.5. Let $\mathbf{c} \in \mathbf{C}$ and consider $\mathbf{v}, \mathbf{w} \in \mathcal{E}_{\mathbf{c}}$. The definition (6.19) of the bilinear form $\mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}$ together with the definition (6.8) of the tensor $\boldsymbol{\sigma}_{\boldsymbol{\beta},\boldsymbol{\mu}}$ yield

$$\begin{aligned} \mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}(\mathbf{v}, \mathbf{w}) + \mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}(\mathbf{w}, \mathbf{v}) &= 2 \int_{\mathbf{c}} \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \cdot \boldsymbol{\sigma}_{\boldsymbol{\beta},\boldsymbol{\mu}} \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w}) + \int_{\mathbf{c}} (\nabla \cdot \boldsymbol{\beta}) \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w}) \\ &\quad + \sum_{\mathbf{c} \in \mathcal{C}_{\mathbf{E},\mathbf{c}}} \int_{\mathbf{c}} \left(\boldsymbol{\beta} \cdot \nabla \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w}) + \boldsymbol{\beta} \cdot \nabla \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w}) \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \right). \end{aligned} \quad (6.33)$$

Hence, combining the last two terms on the right-hand side yields

$$\mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}(\mathbf{v}, \mathbf{w}) + \mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}(\mathbf{w}, \mathbf{v}) = 2 \int_{\mathbf{c}} \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \cdot \boldsymbol{\sigma}_{\boldsymbol{\beta},\boldsymbol{\mu}} \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w}) + \sum_{\mathbf{c} \in \mathcal{C}_{\mathbf{E},\mathbf{c}}} \int_{\mathbf{c}} \nabla \cdot (\boldsymbol{\beta} \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{w})). \quad (6.34)$$

Considering now the last term of (6.34), we choose $\mathbf{w} = \mathbf{v}$, leading to

$$\frac{1}{2} \sum_{\mathbf{c} \in \mathcal{C}_{\mathbf{E},\mathbf{c}}} \int_{\mathbf{c}} \nabla \cdot (\boldsymbol{\beta} |\mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v})|^2) = \frac{1}{2} \sum_{\mathbf{f} \in \mathbf{F}_{\mathbf{c}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{c}}) |\mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v})|^2 + \sum_{\mathbf{f} \in \mathcal{F}_{\mathbf{E},\mathbf{c}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \{\mathbf{L}_{\mathcal{E}}(\mathbf{v})\},$$

since $[\mathbf{L}_{\mathcal{E}}(\mathbf{v})]^2 = 2[\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \{\mathbf{L}_{\mathcal{E}}(\mathbf{v})\}$ and with $\mathbf{n}_{\mathbf{c}}$ the unit outward normal vector to $\mathbf{c}$. Observing that the rightmost term of this identity is equal to $-2\mathbf{n}_{\boldsymbol{\beta};\mathbf{c}}(\mathbf{v}, \mathbf{v})$ owing to definition (6.21), we combine this identity with (6.34) to obtain

$$\sum_{\mathbf{c} \in \mathbf{C}} (\mathbf{g}_{\boldsymbol{\beta},\boldsymbol{\mu};\mathbf{c}}(\mathbf{v}, \mathbf{v}) + \mathbf{n}_{\boldsymbol{\beta};\mathbf{c}}(\mathbf{v}, \mathbf{v})) = \sum_{\mathbf{c} \in \mathbf{C}} \int_{\mathbf{c}} \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) \cdot \boldsymbol{\sigma}_{\boldsymbol{\beta},\boldsymbol{\mu}} \cdot \mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v}) + \frac{1}{2} \sum_{\mathbf{c} \in \mathbf{C}} \sum_{\mathbf{f} \in \mathbf{F}_{\mathbf{c}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{c}}) |\mathbf{L}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{v})|^2.$$

The above rightmost term is now reformulated as

$$\frac{1}{2} \sum_{\mathbf{c} \in \mathbf{C}} \sum_{\mathbf{f} \in \mathbf{F}_{\mathbf{c}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{c}}) |\mathbf{L}_{\mathcal{E}}(\mathbf{v})|^2 = \frac{1}{2} \sum_{\mathbf{f} \in \mathbf{F}^\partial} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}) |\mathbf{L}_{\mathcal{E}_{\mathbf{c}(\mathbf{f})}}(\mathbf{v})|^2 - \sum_{\mathbf{f} \in \mathbf{F}^\circ} \sum_{\mathbf{f} \in \mathcal{F}_{\mathbf{E},\mathbf{f}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \{\mathbf{L}_{\mathcal{E}}(\mathbf{v})\},$$

so that we obtain, using the definition (6.21) of $\mathbf{n}_{\beta;\mathbf{f}}$,

$$\frac{1}{2} \sum_{\mathbf{c} \in \mathbf{C}} \sum_{\mathbf{f} \in \mathbf{F}_{\mathbf{c}}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{c}}) |\mathbf{L}_{\mathcal{E}}(\mathbf{v})|^2 = \frac{1}{2} \sum_{\mathbf{f} \in \mathbf{F}^{\partial}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}) |\mathbf{L}_{\mathcal{E}(\mathbf{f})}(\mathbf{v})|^2 + \sum_{\mathbf{f} \in \mathbf{F}^{\circ}} \mathbf{n}_{\beta;\mathbf{f}}(\mathbf{v}, \mathbf{v}).$$

Finally, recalling the definitions (6.20) and (6.24) of $\mathbf{n}_{\beta}$ and of $\langle\langle \mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n});\mathbf{f}}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{E}(\mathbf{f})}$ for all $\mathbf{f} \in \mathbf{F}^{\partial}$, respectively, we end up with

$$\mathbf{g}_{\beta,\mu}(\mathbf{v}, \mathbf{v}) + \mathbf{n}_{\beta}(\mathbf{v}, \mathbf{v}) = \sum_{\mathbf{c} \in \mathbf{C}} \int_{\mathbf{c}} \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \cdot \boldsymbol{\sigma}_{\beta,\mu} \cdot \mathbf{L}_{\mathcal{E}}(\mathbf{v}) + \frac{1}{2} \sum_{\mathbf{f} \in \mathbf{F}^{\partial}} \langle\langle \mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n});\mathbf{f}}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\mathcal{E}(\mathbf{f})},$$

whence

$$\mathbf{A}_{\beta,\mu}^{\varepsilon}(\mathbf{v}, \mathbf{v}) = \sum_{\mathbf{c} \in \mathbf{C}} \int_{\mathbf{c}} \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \cdot \boldsymbol{\sigma}_{\beta,\mu} \cdot \mathbf{L}_{\mathcal{E}}(\mathbf{v}) + \frac{1}{2} \langle\langle \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle\rangle_{\varepsilon} + \frac{1}{2} \gamma \mathbf{s}_{\beta}(\mathbf{v}, \mathbf{v}), \quad (6.35)$$

owing to the definition (6.12) of $\mathbf{A}_{\beta,\mu}^{\varepsilon}$ and observing that $\mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n})^{\ominus};\mathbf{f}}^{\varepsilon,\partial} + \frac{1}{2} \mathbf{H}_{(\boldsymbol{\beta} \cdot \mathbf{n});\mathbf{f}}^{\varepsilon,\partial} \equiv \frac{1}{2} \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|;\mathbf{f}}^{\varepsilon,\partial}$ for all $\mathbf{f} \in \mathbf{F}^{\partial}$. The expected result then follows from assumption $(\mathcal{H})$ along with Proposition 6.4 and $\gamma \geq \gamma_b$. $\square$

In the proof of Lemma 6.6, we consider the function $\boldsymbol{\delta}(\mathbf{v})$ defined as

$$\forall \mathbf{c} \in \mathbf{C}, \quad \boldsymbol{\delta}(\mathbf{v})|_{\mathbf{c}} = \mathbf{L}_{\mathcal{E}}(\zeta \mathbf{v}) - \zeta \mathbf{L}_{\mathcal{E}}(\mathbf{v}), \quad \forall \mathbf{v} \in \mathcal{E}_{\mathbf{c}}, \quad (6.36)$$

with $\zeta \in \text{Lip}(\Omega)$ the potential defined in $(\mathcal{H}')$ and where $\zeta \mathbf{v} \in \mathcal{E}$ is defined as $(\zeta \mathbf{v})_{\mathbf{e}} = \zeta(\mathbf{x}_{\mathbf{e}}) \mathbf{v}_{\mathbf{e}}$, for all $\mathbf{e} \in \mathbf{E}$. This function satisfies the two following propositions.

Proposition 6.7 (Bounds on $\boldsymbol{\delta}$). *For all $\mathbf{c} \in \mathbf{C}$, we have*

$$\|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})} \leq 2\mathcal{C}_{\sharp} L_{\zeta} h_{\mathbf{c}} \|\mathbf{v}\|_{\mathcal{E}_{\mathbf{c}},2}, \quad \forall \mathbf{v} \in \mathcal{E}_{\mathbf{c}}. \quad (6.37a)$$

and for all $\mathbf{f} \in \mathbf{F}_{\mathbf{c}}$,

$$\|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{f})} \leq 2\mathcal{C}_{\text{T}} \mathcal{C}_{\sharp} L_{\zeta} h_{\mathbf{c}}^{\frac{1}{2}} \|\mathbf{v}\|_{\mathcal{E}_{\mathbf{c}},2}, \quad \forall \mathbf{v} \in \mathcal{E}_{\mathbf{c}}. \quad (6.37b)$$

Proof. Let us consider $\mathbf{v} \in \mathcal{E}$.

i) Proof of (6.37a). Let $\mathbf{c} \in \mathbf{C}$. Let $\zeta_{\mathbf{c}}$ denote the mean-value of ζ over $\mathbf{c}$ defined by $\zeta_{\mathbf{c}} = |\mathbf{c}|^{-1} \int_{\mathbf{c}} \zeta$. Since $\mathbf{L}_{\mathcal{E}}(\zeta_{\mathbf{c}} \mathbf{v}) = \zeta_{\mathbf{c}} \mathbf{L}_{\mathcal{E}}(\mathbf{v})$ because $\zeta_{\mathbf{c}}$ is constant, we infer that $\boldsymbol{\delta}(\mathbf{v})|_{\mathbf{c}} = (\zeta - \zeta_{\mathbf{c}}) \mathbf{L}_{\mathcal{E}}(\mathbf{v}) - \mathbf{L}_{\mathcal{E}}((\zeta - \zeta_{\mathbf{c}}) \mathbf{v})$. Using the triangle inequality, the Hölder inequality and the upper bound in Proposition 6.4 yield

$$\begin{aligned} \|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})} &\leq \|\zeta - \zeta_{\mathbf{c}}\|_{L^{\infty}(\mathbf{c})} \|\mathbf{L}_{\mathcal{E}}(\mathbf{v})\|_{L^2(\mathbf{c})} + \|\mathbf{L}_{\mathcal{E}}((\zeta - \zeta_{\mathbf{c}}) \mathbf{v})\|_{L^2(\mathbf{c})} \\ &\leq \mathcal{C}_{\sharp} \|\zeta - \zeta_{\mathbf{c}}\|_{L^{\infty}(\mathbf{c})} \|\mathbf{v}\|_{2,\mathbf{c}} + \mathcal{C}_{\sharp} \|(\zeta - \zeta_{\mathbf{c}}) \mathbf{v}\|_{2,\mathbf{c}} \leq 2\mathcal{C}_{\sharp} \|\zeta - \zeta_{\mathbf{c}}\|_{L^{\infty}(\mathbf{c})} \|\mathbf{v}\|_{2,\mathbf{c}}. \end{aligned}$$

Observing that $\|\zeta - \zeta_{\mathbf{c}}\|_{L^{\infty}(\mathbf{c})} \leq L_{\zeta} h_{\mathbf{c}}$, the expected result follows.

ii) Proof of (6.37b). Let $\mathbf{f} \in \mathbf{F}$ and let $\mathbf{c} \in \mathbf{C}_{\mathbf{f}}$. Let $\mathbf{f} \in \mathfrak{F}_{\mathbf{E},\mathbf{f}}$ and $\mathbf{c} \in \mathfrak{C}_{\mathbf{E},\mathbf{c}}$ such that $\mathbf{f} \in \partial \mathbf{c}$. Owing to the multiplicative trace inequality from Lemma (7.25), we have

$$\|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{f})} \leq \mathcal{C}_{\text{T}} \|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})}^{\frac{1}{2}} \left(h_{\mathbf{c}}^{-\frac{1}{2}} \|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})}^{\frac{1}{2}} + |\boldsymbol{\delta}(\mathbf{v})|_{\mathbf{H}^1(\mathbf{c})}^{\frac{1}{2}} \right).$$

Recall that $|\boldsymbol{\delta}(\mathbf{v})|_{\mathbf{H}^1(\mathbf{c})} = \|\nabla \boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})}$ with $|\nabla \boldsymbol{\delta}(\mathbf{v})|^2 = \sum_{i,j}^3 |\partial_j \boldsymbol{\delta}(\mathbf{v})_i|^2$. Since $\mathbf{L}_{\mathcal{E}}(\mathbf{v})$ is piecewise constant on $\mathfrak{C}_{\mathbf{E},\mathbf{c}}$, it then follows that $|\nabla \boldsymbol{\delta}(\mathbf{v})|^2 = \sum_{i,j}^3 |\mathbf{L}_{\mathcal{E}}(\mathbf{v})_i \partial_j \zeta|^2 = |\mathbf{L}_{\mathcal{E}}(\mathbf{v})|^2 |\nabla \zeta|^2$. As a result, $|\boldsymbol{\delta}(\mathbf{v})|_{\mathbf{H}^1(\mathbf{c})} \leq L_{\zeta} \|\mathbf{L}_{\mathcal{E}}(\mathbf{v})\|_{L^2(\mathbf{c})}$. Moreover, proceeding as in *i)*, we infer that $\|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{c})} \leq 2L_{\zeta} h_{\mathbf{c}} \|\mathbf{L}_{\mathcal{E}}(\mathbf{v})\|_{L^2(\mathbf{c})}$. Collecting these bounds yields

$$\|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\mathbf{f})} \leq 2\mathcal{C}_{\text{T}} L_{\zeta} h_{\mathbf{c}}^{\frac{1}{2}} \|\mathbf{L}_{\mathcal{E}}(\mathbf{v})\|_{L^2(\mathbf{c})}.$$

Then, the summation over $\mathbf{f} \in \mathfrak{F}_{\mathbf{E},\mathbf{f}}$ since $\bar{\mathbf{f}} = \cup \{\bar{\mathbf{f}} \mid \mathbf{f} \in \mathfrak{F}_{\mathbf{E},\mathbf{f}}\}$ owing to assumption $(\mathbf{St})$, and the upper bound of Proposition 6.4 yield the expected result. $\square$

Proposition 6.8 (Multiplicative stability). *Assume that (6.30) holds. Then, there exists $\mathcal{C}_\zeta > 0$ independent of the mesh size and the model parameters, such that*

$$\|\zeta \mathbf{v}\|_{\mathcal{E},a} \leq \mathcal{C}_\zeta \left(\|\zeta\|_{L^\infty(\Omega)} + \mathcal{C}_{\mathcal{E},a} \right) \|\mathbf{v}\|_{\mathcal{E},a}, \quad \forall \mathbf{v} \in \mathcal{E}.$$

Proof. Let $\mathbf{v} \in \mathcal{E}$ and let us rewrite $\|\zeta \mathbf{v}\|_{\mathcal{E},a}^2$ as

$$\begin{aligned} \|\zeta \mathbf{v}\|_{\mathcal{E},a}^2 &= \sum_{c \in \mathcal{C}} \tau^{-1} \|\zeta \mathbf{v}\|_{2,c}^2 + \sum_{c \in \mathcal{C}} \gamma_b \mathbf{s}_{\beta;c}(\zeta \mathbf{v}, \zeta \mathbf{v}) + \sum_{f \in \mathcal{F}^\partial} \langle \mathbf{H}_{|\beta \cdot \mathbf{n}|;f}^{\mathcal{E},\partial}(\zeta \mathbf{v}), \zeta \mathbf{v} \rangle_{\mathcal{E}_c(f)} + \sum_{f \in \mathcal{F}^\circ} \gamma_b \mathbf{s}_{\beta;f}(\zeta \mathbf{v}, \zeta \mathbf{v}) \\ &= T_1 + T_2 + T_3 + T_4. \end{aligned}$$

The idea of the proof is to use the Lipschitz regularity of ζ to bound separately these terms by $\|\mathbf{v}\|_{\mathcal{E},a}^2$. We recall the notation $\zeta_c = |\mathbf{c}|^{-1} \int_c \zeta$ from the proof of Proposition 6.7.

i) *Bound on T_1 .* First, we start using the triangle inequality to obtain

$$\frac{1}{2} T_1 \leq \sum_{c \in \mathcal{C}} \tau^{-1} \|\zeta_c \mathbf{v}\|_{\mathcal{E}_c,2}^2 + \sum_{c \in \mathcal{C}} \tau^{-1} \|(\zeta - \zeta_c) \mathbf{v}\|_{\mathcal{E}_c,2}^2 = T_{1,1} + T_{1,2}.$$

Since $|\zeta_c| \leq \|\zeta\|_{L^\infty(c)}$, we infer that $T_{1,1} \leq \sum_{c \in \mathcal{C}} \tau^{-1} \|\zeta\|_{L^\infty(c)}^2 \|\mathbf{v}\|_{\mathcal{E}_c,2}^2 \leq \|\zeta\|_{L^\infty(\Omega)}^2 \|\mathbf{v}\|_{\mathcal{E},a}^2$ and the bound on $T_{1,2}$ easily follows from the Lipschitz regularity of ζ : $T_{1,2} \leq \sum_{c \in \mathcal{C}} \tau^{-1} L_\zeta^2 h_c^2 \|\mathbf{v}\|_{\mathcal{E}_c,2}^2 \leq L_\zeta^2 h_c^2 \|\mathbf{v}\|_{\mathcal{E},a}^2$. Then, recalling assumption (6.30) yields

$$T_1 \leq 2 \left(\mathcal{C}_{\mathcal{E},a}^2 + \|\zeta\|_{L^\infty(\Omega)}^2 \right) \|\mathbf{v}\|_{\mathcal{E},a}^2.$$

ii) *Bound on T_2 .* Since the bilinear form $\mathbf{s}_{\beta;c}$ is symmetric and positive, we claim that

$$\frac{1}{2} T_2 \leq \sum_{c \in \mathcal{C}} \gamma_b \mathbf{s}_{\beta;c}(\zeta_c \mathbf{v}, \zeta_c \mathbf{v}) + \sum_{c \in \mathcal{C}} \gamma_b \mathbf{s}_{\beta;c}((\zeta - \zeta_c) \mathbf{v}, (\zeta - \zeta_c) \mathbf{v}) = T_{2,1} + T_{2,2}.$$

Obviously, we have $T_{2,1} \leq \sum_{c \in \mathcal{C}} \gamma_b \|\zeta\|_{L^\infty(c)}^2 \mathbf{s}_{\beta;c}(\mathbf{v}, \mathbf{v}) \leq \|\zeta\|_{L^\infty(\Omega)}^2 \|\mathbf{v}\|_{\mathcal{E},a}^2$. By definition, we have

$$\mathbf{s}_{\beta;c}((\zeta - \zeta_c) \mathbf{v}, (\zeta - \zeta_c) \mathbf{v}) = \sum_{f \in \mathcal{F}_{E,c}} \int_f |\beta \cdot \mathbf{n}_f| \|\mathbf{L}_{\mathcal{E}_c}((\zeta - \zeta_c) \mathbf{v})\|^2$$

whence, using the multiplicative trace inequality from Lemma (7.25) and that $\mathbf{L}_{\mathcal{E}_c}$ is piece-wise constant,

$$\mathbf{s}_{\beta;c}((\zeta - \zeta_c) \mathbf{v}, (\zeta - \zeta_c) \mathbf{v}) \leq 2\mathcal{C}_T^2 \|\beta\|_{L^\infty(c)} \sum_{f \in \mathcal{F}_{E,c}} \sum_{\mathbf{c} \in \mathcal{C}_{E,f}} h_c^{-1} \|\mathbf{L}_{\mathcal{E}_c}((\zeta - \zeta_c) \mathbf{v})\|_{L^2(c)}^2,$$

with $\mathcal{C}_{E,f} = \{\mathbf{c} \in \mathcal{C}_{E,c} \mid f \subset \partial \mathbf{c}\}$. Recalling now that the boundary of each diamond $\mathbf{c} \in \mathcal{C}_{E,c}$ is composed of 4 sub-faces in $\mathcal{F}_{E,c}$, we exchange summation order to obtain

$$\begin{aligned} \mathbf{s}_{\beta;c}((\zeta - \zeta_c) \mathbf{v}, (\zeta - \zeta_c) \mathbf{v}) &\leq 8\mathcal{C}_T^2 \|\beta\|_{L^\infty(c)} \sum_{\mathbf{c} \in \mathcal{C}_{E,c}} h_c^{-1} \|\mathbf{L}_{\mathcal{E}_c}((\zeta - \zeta_c) \mathbf{v})\|_{L^2(c)}^2 \\ &= 8\mathcal{C}_T^2 \|\beta\|_{L^\infty(c)} h_c^{-1} \|\mathbf{L}_{\mathcal{E}_c}((\zeta - \zeta_c) \mathbf{v})\|_{L^2(c)}^2. \end{aligned}$$

Hence, owing to upper bound from Proposition 6.4, the Lipschitz regularity of ζ , and assumption (6.30), it follows that

$$\mathbf{s}_{\beta;c}((\zeta - \zeta_c) \mathbf{v}, (\zeta - \zeta_c) \mathbf{v}) \leq 8\mathcal{C}_T^2 \mathcal{C}_\#^2 \|\beta\|_{L^\infty(c)} h_c^{-1} \|(\zeta - \zeta_c) \mathbf{v}\|_{\mathcal{E}_c,2}^2 \leq 8\mathcal{C}_T^2 \mathcal{C}_\#^2 \mathcal{C}_{\mathcal{E},a}^2 \tau^{-1} \|\mathbf{v}\|_{\mathcal{E}_c,2}^2.$$

As a result, collecting these two bounds leads to

$$T_2 \leq 2 \left(\|\zeta\|_{L^\infty(\Omega)}^2 + 8\mathcal{C}_T^2 \mathcal{C}_\#^2 \mathcal{C}_{\mathcal{E},a}^2 \right) \|\mathbf{v}\|_{\mathcal{E},a}^2.$$

iii) *Bound on T_3 .* Proceeding as in the previous step ii), we infer that

$$T_3 \leq 2 \left(\|\zeta\|_{L^\infty(\Omega)}^2 + n_{F,\partial} \mathbf{C}_T^2 \mathbf{C}_\#^2 \mathbf{C}_{\mathcal{E},a}^2 \right) \|\mathbf{v}\|_{\mathcal{E},a}^2,$$

where $n_{F,\partial} = (\max_{c \in C} \#(F_c \cap F^\partial))$ is the maximal number of boundary faces per mesh cells.

iv) *Bound on T_4 .* We use a different decomposition to bound this last term. Namely,

$$T_4 = \sum_{f \in F^\circ} \gamma_b \mathbf{s}_{\zeta^2 \beta; f}(\mathbf{v}, \mathbf{v}) + \sum_{f \in F^\circ} \gamma_b \Delta_f(\mathbf{v}) = T_{4,1} + T_{4,2},$$

with $\Delta_f(\mathbf{v}) = \mathbf{s}_{\beta; f}(\zeta \mathbf{v}, \zeta \mathbf{v}) - \mathbf{s}_{\zeta^2 \beta; f}(\mathbf{v}, \mathbf{v})$. Observing that $\mathbf{s}_{\zeta^2 \beta; f}(\mathbf{v}, \mathbf{v}) \leq \|\zeta\|_{L^\infty(f)}^2 \mathbf{s}_{\beta; f}(\mathbf{v}, \mathbf{v})$ for all $f \in F_c$, it follows that $T_{4,1} \leq \|\zeta\|_{L^\infty(\Omega)}^2 \|\mathbf{v}\|_{\mathcal{E},a}^2$. Now, introducing $\delta(\mathbf{v})$ defined by (6.36), we observe that

$$\Delta_f(\mathbf{v}) = \int_f |\boldsymbol{\beta} \cdot \mathbf{n}_f| \left(\left(\llbracket \delta(\mathbf{v}) \rrbracket + \zeta \llbracket \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \rrbracket \right)^2 - \zeta^2 \llbracket \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \rrbracket^2 \right).$$

Then, applying the Young's inequality along with the trace inequality (6.37b), it follows that

$$\begin{aligned} |\Delta_f(\mathbf{v})| &\leq 2 \int_f |\boldsymbol{\beta} \cdot \mathbf{n}_f| \llbracket \delta(\mathbf{v}) \rrbracket^2 + \int_f |\boldsymbol{\beta} \cdot \mathbf{n}_f| \zeta^2 \llbracket \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \rrbracket^2 \\ &\leq 4(2\mathbf{C}_T \mathbf{C}_\# L_\zeta)^2 \|\boldsymbol{\beta}\|_{L^\infty(f)} \sum_{c \in C_f} h_c \|\mathbf{v}\|_{e_c,2}^2 + \|\zeta\|_{L^\infty(f)}^2 \mathbf{s}_{\beta; f}(\mathbf{v}, \mathbf{v}). \end{aligned}$$

Hence, since $\#C_f = 2$ for all $f \in F^\circ$ and recalling (6.30), we infer that

$$T_{4,2} \leq 32\mathbf{C}_T^2 \mathbf{C}_\#^2 \mathbf{C}_{\mathcal{E},a}^2 \tau^{-1} \|\mathbf{v}\|_2^2 + \|\zeta\|_{L^\infty(\Omega)}^2 \sum_{f \in F^\circ} \gamma_b \mathbf{s}_{\beta; f}(\mathbf{v}, \mathbf{v}) \leq \left(32\mathbf{C}_T^2 \mathbf{C}_\#^2 \mathbf{C}_{\mathcal{E},a}^2 + \|\zeta\|_{L^\infty(\Omega)}^2 \right) \|\mathbf{v}\|_{\mathcal{E},a}^2,$$

whence

$$T_4 \leq 2 \left(16\mathbf{C}_T^2 \mathbf{C}_\#^2 \mathbf{C}_{\mathcal{E},a}^2 + \|\zeta\|_{L^\infty(\Omega)}^2 \right) \|\mathbf{v}\|_{\mathcal{E},a}^2.$$

v) The expected result then follows by collecting the bounds on T_1 , T_2 , T_3 and T_4 . $\square$

We now turn to the proof of the inf-sup stability Lemma 6.6.

Proof of Lemma 6.6. Let $\mathbf{v} \in \mathcal{E}$ and define $S = \sup_{\mathbf{w} \in \mathcal{E}; \|\mathbf{w}\|_{\mathcal{E},a}=1} \mathbb{A}_{\beta,\mu}^\mathcal{E}(\mathbf{v}, \mathbf{w})$. Let us take $\mathbf{w} = \zeta \mathbf{v} + \theta \mathbf{v}$ with $\theta > 0$ to be chosen below. We infer from Proposition 6.8 that

$$\mathbb{A}_{\beta,\mu}^\mathcal{E}(\mathbf{v}, \mathbf{w}) \leq S \|\mathbf{w}\|_{\mathcal{E},a} \leq S \left(\theta + \mathbf{C}_\zeta \left(\|\zeta\|_{L^\infty(\Omega)} + \mathbf{C}_{\mathcal{E},a} \right) \right) \|\mathbf{v}\|_{\mathcal{E},a},$$

so that it remains to prove that $\mathbb{A}_{\beta,\mu}^\mathcal{E}(\mathbf{v}, \mathbf{w}) \gtrsim \|\mathbf{v}\|_{\mathcal{E},a}^2$. First, split $\mathbb{A}_{\beta,\mu}^\mathcal{E}$ as follows:

$$\mathbb{A}_{\beta,\mu}^\mathcal{E}(\mathbf{v}, \zeta \mathbf{v}) = \mathbb{A}_{\beta, -\nabla \beta^T + \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}, \mathbf{w}) + \langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}), \mathbf{w} \rangle_\mathcal{E} = T_1 + T_2,$$

using the definition (6.18) of $\mathbb{H}_{\mu;c}^\mathcal{E}$ (replacing μ by $\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}$) and defining

$$\langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}), \mathbf{w} \rangle_\mathcal{E} = \sum_{c \in C} \langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}; c}^\mathcal{E}(\mathbf{v}), \mathbf{w} \rangle_{\mathcal{E}_c}, \quad \forall \mathbf{v}, \mathbf{w} \in \mathcal{E}. \quad (6.38)$$

i) *Bound on T_1 .* We bound from below T_1 by considering the decomposition

$$\begin{aligned} \mathbb{A}_{\beta, -\nabla \beta^T + \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}, \zeta \mathbf{v}) &= \mathbb{A}_{\zeta \beta, -\nabla(\zeta \beta)^T + \frac{1}{2}\zeta(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}, \mathbf{v}) \\ &\quad + \mathbb{A}_{\beta, -\nabla \beta^T}^\mathcal{E}(\mathbf{v}, \zeta \mathbf{v}) - \mathbb{A}_{\zeta \beta, -\nabla(\zeta \beta)^T}^\mathcal{E}(\mathbf{v}, \mathbf{v}) \\ &\quad + \langle \mathbb{H}_{\frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}), \zeta \mathbf{v} \rangle_\mathcal{E} - \langle \mathbb{H}_{\frac{1}{2}\zeta(\nabla \cdot \beta) \text{Id}}^\mathcal{E}(\mathbf{v}), \mathbf{v} \rangle_\mathcal{E} = T_{1,1} + T_{1,2} + T_{1,3}. \end{aligned}$$

Regarding $T_{1,1}$, observe that $\sigma_{\zeta\beta, -\nabla(\zeta\beta)^T + \frac{1}{2}\zeta(\nabla\cdot\beta)\text{Id}} = -\frac{1}{2}\beta\cdot\nabla\zeta\text{Id}$ and use the relation (6.35) to infer that

$$\begin{aligned} T_{1,1} &= \sum_{c \in \mathcal{C}} \int_c \left(-\frac{1}{2}\beta\cdot\nabla\zeta \right) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) + \frac{1}{2} \langle \mathbf{H}_{|\zeta\beta\cdot\mathbf{n}|}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon} + \frac{1}{2} \gamma_b s_{\zeta\beta}(\mathbf{v}, \mathbf{v}) \\ &\geq \sum_{c \in \mathcal{C}} \int_c \left(-\frac{1}{2}\beta\cdot\nabla\zeta \right) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) + \frac{1}{2} \langle \mathbf{H}_{|\beta\cdot\mathbf{n}|}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon} + \frac{1}{2} \gamma_b s_{\beta}(\mathbf{v}, \mathbf{v}), \end{aligned}$$

since $\zeta \geq 1$. Then, owing to assumption $(\mathcal{H}')$ together with the lower bound of Proposition 6.4, we infer that $T_{1,1} \geq \frac{1}{2} \|\mathbf{v}\|_{\mathcal{E},a}^2$. The next step is to bound the perturbation term $T_{1,2}$. First, we recall the identity (6.19) for $\mathbf{g}_{\beta,\mu;c}$, and we observe that $\mathbf{g}_{\beta, -\nabla\beta^T;c} \equiv 0$ and $\mathbf{g}_{\zeta\beta, -\nabla(\zeta\beta)^T;c} \equiv 0$, so that $T_{1,2}$ solely consists of surfacic terms:

$$T_{1,2} = \left(n_{\beta}(\mathbf{v}, \zeta\mathbf{v}) - n_{\zeta\beta}(\mathbf{v}, \mathbf{v}) \right) + \gamma_b \left(s_{\beta}(\mathbf{v}, \zeta\mathbf{v}) - s_{\zeta\beta}(\mathbf{v}, \mathbf{v}) \right) + \left(\langle \mathbf{H}_{(\beta\cdot\mathbf{n})^{\ominus}}^{\varepsilon,\partial}(\mathbf{v}), \zeta\mathbf{v} \rangle_{\varepsilon} - \langle \mathbf{H}_{(\zeta\beta\cdot\mathbf{n})^{\ominus}}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon} \right).$$

Now, recalling that $\beta \in \mathbf{Lip}(\Omega)$, $\zeta \geq 1$, and $\zeta \in \text{Lip}(\Omega)$, so that $\zeta \{\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\} = \{\zeta \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\}$, we observe that

$$\begin{aligned} n_{\beta;x}(\mathbf{v}, \zeta\mathbf{v}) - n_{\zeta\beta;x}(\mathbf{v}, \mathbf{v}) &= \sum_{\mathbf{f} \in \mathfrak{F}_x} \int_{\mathbf{f}} (\beta\cdot\mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \{\boldsymbol{\delta}(\mathbf{v})\}, \\ s_{\beta;x}(\mathbf{v}, \zeta\mathbf{v}) - s_{\zeta\beta;x}(\mathbf{v}, \mathbf{v}) &= \sum_{\mathbf{f} \in \mathfrak{F}_x} \int_{\mathbf{f}} |\beta\cdot\mathbf{n}_{\mathbf{f}}| [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \{\boldsymbol{\delta}(\mathbf{v})\}, \end{aligned}$$

for all $x \in F_c$ or $x \in C$, where the function $\boldsymbol{\delta}(\mathbf{v})$ is defined by (6.36). Similarly, we have

$$\langle \mathbf{H}_{(\beta\cdot\mathbf{n})^{\ominus};\mathbf{f}}^{\varepsilon,\partial}(\mathbf{v}), \zeta\mathbf{v} \rangle_{\varepsilon_{c(\mathbf{f})}} - \langle \mathbf{H}_{(\zeta\beta\cdot\mathbf{n})^{\ominus};\mathbf{f}}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon_{c(\mathbf{f})}} = \int_{\mathbf{f}} (\beta\cdot\mathbf{n})^{\ominus} \mathbf{L}_{\mathcal{E}}(\mathbf{v}) \cdot \boldsymbol{\delta}(\mathbf{v}),$$

for all $\mathbf{f} \in F^{\partial}$. Then, applying the Cauchy–Schwarz inequality to these three terms yields

$$T_{1,2} \leq 6 \left(\langle \mathbf{H}_{|\beta\cdot\mathbf{n}|}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon} + \gamma_b s_{\beta}(\mathbf{v}, \mathbf{v}) \right)^{\frac{1}{2}} \left(2 \sum_{c \in \mathcal{C}} \|\beta\|_{L^{\infty}(c)} \sum_{\mathbf{c} \in \mathfrak{C}_{E,c}} \|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\partial c)}^2 \right)^{\frac{1}{2}}. \quad (6.39)$$

Observing now that $\sigma_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}} \equiv \mathbf{0}$ by definition (see (6.8)), and using the identity (6.35), we observe that

$$\mathbb{A}_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}}^{\varepsilon}(\mathbf{v}, \mathbf{v}) = \frac{1}{2} \left(\langle \mathbf{H}_{|\beta\cdot\mathbf{n}|}^{\varepsilon,\partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon} + \gamma_b s_{\beta}(\mathbf{v}, \mathbf{v}) \right),$$

so that combining this expression with the estimate (6.39) yields

$$T_{1,2} \leq 12 \left(\mathbb{A}_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}}^{\varepsilon}(\mathbf{v}, \mathbf{v}) \right)^{\frac{1}{2}} \left(\sum_{c \in \mathcal{C}} \|\beta\|_{L^{\infty}(c)} \sum_{\mathbf{c} \in \mathfrak{C}_{E,c}} \|\boldsymbol{\delta}(\mathbf{v})\|_{L^2(\partial c)}^2 \right)^{\frac{1}{2}}.$$

Finally, we use the inequalities (6.37a)-(6.37b) together with assumption (6.30), to infer that

$$T_{1,2} \leq \mathbf{C}_{\delta} \mathbf{C}_{\varepsilon,a} \left(\mathbb{A}_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}}^{\varepsilon}(\mathbf{v}, \mathbf{v}) \right)^{\frac{1}{2}} \left(\tau^{-1} \|\mathbf{v}\|_{\mathcal{E}}^2 \right)^{\frac{1}{2}}$$

where $\mathbf{C}_{\delta} > 0$ depends exclusively on the numerical constants $\mathbf{C}_T$ and $\mathbf{C}_{\sharp}$. Now, we collect the bounds on $T_{1,1}$ and $T_{1,2}$ and we apply the Young's inequality to obtain

$$\mathbb{A}_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}}^{\varepsilon}(\mathbf{v}, \mathbf{w}) \geq \frac{1}{4} \|\mathbf{v}\|_{\mathcal{E},a}^2 + (\theta - \mathbf{C}_{\delta}^2 \mathbf{C}_{\varepsilon,a}^2) \mathbb{A}_{\beta, -\nabla\beta^T + \frac{1}{2}(\nabla\cdot\beta)\text{Id}}^{\varepsilon}(\mathbf{v}, \mathbf{v}) + T_{1,3},$$

where we recall that $\mathbf{w} = \zeta \mathbf{v} + \theta \mathbf{v}$. As a result, choosing $\theta = \mathbf{C}_\delta^2 \mathbf{C}_{\varepsilon,a}^2$ yields

$$T_1 = \mathbb{A}_{\beta, -\nabla \beta^T + \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\varepsilon(\mathbf{v}, \mathbf{w}) \geq \frac{1}{4} \|\mathbf{v}\|_{\varepsilon,a}^2 + T_{1,3}. \quad (6.40)$$

ii) *Bound on T_2 .* First, we rewrite this term as:

$$\begin{aligned} T_2 &= \langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\varepsilon(\mathbf{v}), \mathbf{w} \rangle_\varepsilon = \theta \langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\varepsilon(\mathbf{v}), \mathbf{v} \rangle_\varepsilon \\ &\quad + \langle \mathbb{H}_{\zeta(\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id})}^\varepsilon(\mathbf{v}), \mathbf{v} \rangle_\varepsilon \\ &\quad + \langle \mathbb{H}_{\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id}}^\varepsilon(\mathbf{v}), \zeta \mathbf{v} \rangle_\varepsilon - \langle \mathbb{H}_{\zeta(\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id})}^\varepsilon(\mathbf{v}), \mathbf{v} \rangle_\varepsilon \\ &= T_{2,1} + T_{2,2} + T_{2,3}, \end{aligned}$$

and we denote $\aleph_b = \text{ess inf}_\Omega \aleph$, so that $-\mathbf{C}_\aleph < \aleph_b \leq 0$ by assumption. Considering $T_{2,1}$, we observe that

$$2\sigma_{\beta,\mu} = \left(\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id} \right) + \left(\mu + \nabla \beta^T - \frac{1}{2}(\nabla \cdot \beta) \text{Id} \right)^T$$

so that we have

$$T_{2,1} = \theta \sum_{c \in \mathbf{C}} \int_c \sigma_{\beta,\mu} \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \geq \theta \aleph_b \sum_{c \in \mathbf{C}} \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(c)}^2 \geq \mathbf{C}_\sharp^2 \theta \aleph_b \|\mathbf{v}\|_{\varepsilon,2}^2,$$

where we have used the upper bound from Proposition 6.4 (since $\aleph \leq 0$). The second term $T_{2,2}$ is treated similarly as follows:

$$T_{2,2} = \sum_{c \in \mathbf{C}} \int_c \zeta \sigma_{\beta,\mu} \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \geq \mathbf{C}_\sharp^2 \aleph_b \|\zeta\|_{L^\infty(\Omega)} \|\mathbf{v}\|_{\varepsilon,2}^2.$$

Hence, we infer that

$$T_2 \geq \aleph_b \vartheta \|\mathbf{v}\|_2^2 + T_{2,3}, \quad (6.41)$$

with $\vartheta = \mathbf{C}_\sharp^2(\theta + \|\zeta\|_{L^\infty(\Omega)})$.

iii) *Bound on $T_1 + T_2$.* Collecting estimates (6.40) and (6.41), we obtain

$$\mathbb{A}_{\beta,\mu}^\varepsilon(\mathbf{v}, \mathbf{w}) = T_1 + T_2 \geq \frac{1}{4} \|\mathbf{v}\|_{\varepsilon,a}^2 + \aleph_b \vartheta \|\mathbf{v}\|_2^2 + T_{1,3} + T_{2,3}.$$

In addition, from the definition of $T_{1,3}$ and $T_{2,3}$, we have

$$T_{1,3} + T_{2,3} = \langle \mathbb{H}_{\mu + \nabla \beta^T}^\varepsilon(\mathbf{v}), \zeta \mathbf{v} \rangle_\varepsilon - \langle \mathbb{H}_{\zeta(\mu + \nabla \beta^T)}^\varepsilon(\mathbf{v}), \mathbf{v} \rangle_\varepsilon = \sum_{c \in \mathbf{C}} \int_c (\mu + \nabla \beta^T) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \delta(\mathbf{v}).$$

Applying the Hölder inequality, the inequality (6.37a) and the upper bound from Proposition 6.4, it then follows that

$$|T_{1,3} + T_{2,3}| \leq \sum_{c \in \mathbf{C}} \|\mu + \nabla \beta^T\|_{L^\infty(c)} \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(c)} \|\delta(\mathbf{v})\|_{L^2(c)} \leq \|\mu + \nabla \beta^T\|_{L^\infty(\Omega)} \mathbf{C}_\sharp^2 L_\zeta h \|\mathbf{v}\|_2^2.$$

By inference,

$$\mathbb{A}_{\beta,\mu}^\varepsilon(\mathbf{v}, \mathbf{w}) \geq \frac{1}{4} \|\mathbf{v}\|_{\varepsilon,a}^2 + \left(\tau \aleph_b \vartheta - \frac{h}{h_0} \right) \tau^{-1} \|\mathbf{v}\|_{\varepsilon,2}^2,$$

with the reference length $h_0 = \left(\mathbf{C}_\sharp^2 \|\mu + \nabla \beta^T\|_{L^\infty(\Omega)} \tau L_\zeta \right)^{-1}$. Hence, there exists $\varrho' > 0$ such that $\mathbb{A}_{\beta,\mu}^\varepsilon(\mathbf{v}, \mathbf{w}) \geq \varrho' \|\mathbf{v}\|_{\varepsilon,a}^2$, as soon as $\aleph_b$ and h satisfy (6.31), yielding the expected result. $\square$

Consistency error and *a priori* error estimate. We now study the consistency of the discrete problem (6.25) so as to achieve an *a priori* error estimate. First, we define the consistency error $\widehat{\mathbb{E}}_{\mathcal{E}}$ using the reduction map $\widehat{\mathbf{R}}_{\mathcal{E}} : \mathbf{L}^1(\Omega) \rightarrow \mathcal{E}$ defined by (6.3) as follows:

$$\widehat{\mathbb{E}}_{\mathcal{E}}(\mathbf{v}) = \sup_{\mathbf{v} \in \mathcal{E}; \|\mathbf{v}\|_{\mathcal{E},a}=1} \left| \mathbb{A}_{\beta,\mu}^{\varepsilon}(\widehat{\mathbf{R}}_{\mathcal{E}}(\mathbf{v}), \mathbf{v}) - \mathbb{S}(\mathbf{s}, \mathbf{u}_D; \mathbf{v}) \right|, \quad \forall \mathbf{v} \in \mathbf{L}^1(\Omega). \quad (6.42)$$

Lemma 6.9 (Consistency error). *Let $\mathbf{u}$ be the exact solution of (6.25) and assume that $\mathbf{u} \in \mathbf{W}^{1,p}$ with $p > 1$. Then, for all $p \in (1, 2]$,*

$$\widehat{\mathbb{E}}_{\mathcal{E}}(\mathbf{u}) \lesssim \left(\sum_{c \in C} N_{\infty;c}^p \tau^{\frac{p}{2}} h_c^{\frac{d}{2}(p-2)} \|\mathbf{u} - \widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{u})\|_{L^p(c)}^p + \sum_{c \in C} \sum_{\mathbf{c} \in \mathfrak{C}_{E,c}} \|\beta\|_{L^\infty(c)}^{\frac{p}{2}} h_c^{\frac{(d-1)}{2}(p-2)} \|\mathbf{u} - \widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{u})\|_{L^p(\partial c)}^p \right)^{\frac{1}{p}},$$

with $d = 3$, $N_{\infty;c} = \|\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}\|_{L^\infty(c)}$ and with the interpolation map $\widehat{\mathbf{l}}_{\mathcal{E}_c} = \mathbf{L}_{\mathcal{E}_c} \circ \widehat{\mathbf{R}}_{\mathcal{E}_c} : \mathbf{L}^1(\hat{c}) \rightarrow \mathbb{P}_0(\mathfrak{C}_{E,c}; \mathbb{R}^3)$ for all $c \in C$.

Proof of Lemma 6.9. Let $\mathbf{y}|_c = (\mathbf{u} - \widehat{\mathbf{l}}_{\mathcal{E}}(\mathbf{u}))|_c$ for all $c \in C$. Note that $\mathbf{y}|_{\partial c} \in \mathbf{L}^p(\partial c)$ for all $c \in \mathfrak{C}_{E,c}$. Let $\mathbf{v} \in \mathcal{E}$. Owing to the definitions of $\mathbb{A}_{\beta,\mu}^{\varepsilon}$ and $\mathbb{S}$, it follows that $\mathbb{S}(\mathbf{s}, \mathbf{u}_D; \mathbf{v}) - \mathbb{A}_{\beta,\mu}^{\varepsilon}(\widehat{\mathbf{R}}_{\mathcal{E}}(\mathbf{u}), \mathbf{v}) = T_1 + T_2 + T_3 + T_4$, with

$$\begin{aligned} T_1 &:= \sum_{c \in C} \int_c (\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y}, \quad T_2 := \sum_{X \in \{F^\circ, C\}} \sum_{x \in X} \sum_{\mathbf{f} \in \mathfrak{F}_{E;x}} \int_{\mathbf{f}} \beta \cdot \mathbf{n} [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot [\mathbf{y}], \\ T_3 &:= \sum_{X \in \{F^\circ, C\}} \sum_{x \in X} \sum_{\mathbf{f} \in \mathfrak{F}_{E;x}} \gamma_{\mathbf{f}} \int_{\mathbf{f}} |\beta \cdot \mathbf{n}| [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot [\mathbf{y}] \quad \text{and} \quad T_4 := \sum_{\mathbf{f} \in F^\partial} \sum_{\mathbf{f} \in \mathfrak{F}_{E,\mathbf{f}}} \int_{\mathbf{f}} (\beta \cdot \mathbf{n})^\ominus \mathbf{L}_{\mathcal{E}_f}(\mathbf{v}) \cdot \mathbf{y}. \end{aligned}$$

Indeed, the first term T_1 is obtained using the identity (6.37b) of $\mathbf{g}_{\beta,\mu;c}$ since $\mathbf{L}_{\mathcal{E}_c}$ are piece-wise constant, together with the integration by part formulae (6.33) and the identity

$$\sum_{c \in \mathfrak{C}_{E,c}} \int_c ((\beta \cdot \nabla) \mathbf{y}) \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) = - \sum_{c \in \mathfrak{C}_{E,c}} \int_c ((\beta \cdot \nabla) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})) \cdot \mathbf{y} - \int_c (\nabla \cdot \beta) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y} + \sum_{c \in \mathfrak{C}_{E,c}} \int_c \nabla \cdot (\beta \mathbf{y} \cdot \mathbf{L}_{\mathcal{E}_c}(\mathbf{v})),$$

holding for all $c \in C$ and all $\mathbf{v} \in \mathcal{E}_c$. The terms T_2 and T_3 result from the rightmost term of the relation (6.33) and the fact that $(\beta \cdot \mathbf{n})[\mathbf{u}]|_{\mathbf{f}} \equiv 0$ for all $\mathbf{f} \in \mathfrak{F}_{E;x}$, owing to Lemma 2.13. Finally, the term T_4 is inferred observing that $\mathbf{u}_D = \mathbf{u}|_{\partial\Omega}$. Let $p \in (1, 2]$ and use the Hölder inequality to bound T_1 as follows:

$$\left| \int_c (\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y} \right| \leq N_{\infty;c} \|\mathbf{y}\|_{L^p(c)} \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^{p'}(c)},$$

with $N_{\infty;c} = \|\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}\|_{L^\infty(c)}$. Using a local inverse inequality (see (Ern & Guermond, 2004, Lemma 1.138)), we infer that

$$\left| \int_c (\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y} \right| \leq N_{\infty;c} h_c^\theta \|\mathbf{y}\|_{L^p(c)} \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(c)},$$

with $\theta = d\left(\frac{1}{2} - \frac{1}{p}\right)$, so that applying once more the Hölder inequality, we get

$$\left| \sum_{c \in C} \int_c (\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y} \right| \leq \left(\sum_{c \in C} N_{\infty;c}^p h_c^{\theta p} \|\mathbf{y}\|_{L^p(c)}^p \right)^{\frac{1}{p}} \left(\sum_{c \in C} \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^2(c)}^{p'} \right)^{\frac{1}{p'}}.$$

Moreover, since $p' \geq 2$ and using the upper bound in Proposition 6.4, we infer that

$$|T_1| = \left| \sum_{c \in C} \int_c (\nabla \beta + \mu^T - (\nabla \cdot \beta) \text{Id}) \mathbf{L}_{\mathcal{E}_c}(\mathbf{v}) \cdot \mathbf{y} \right| \lesssim \left(\sum_{c \in C} N_{\infty;c}^p h_c^{\theta p} \|\mathbf{y}\|_{L^p(c)}^p \right)^{\frac{1}{p}} \|\mathbf{v}\|_{\mathcal{E},2}.$$

To bound the two terms T_2 and T_3 , we consider a sub-face $\mathbf{f} \in \mathfrak{F}_{\mathbf{E};\mathbf{X}}$ for all $\mathbf{x} \in \mathbf{X}$ with $\mathbf{X} \in \{\mathbf{F}^\circ, \mathbf{C}\}$. As above, the Hölder inequality yields

$$\left| \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \langle \mathbf{y} \rangle \right| \leq \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{f})}^{\frac{1}{2}} \|\langle \mathbf{y} \rangle\|_{\mathbf{L}^p(\mathbf{f})} \| |\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}|^{\frac{1}{2}} [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \|_{\mathbf{L}^{p'}(\mathbf{f})},$$

so that using a local inverse inequality, we obtain

$$\left| \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \langle \mathbf{y} \rangle \right| \leq h_{\mathbf{f}}^{\theta'} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{f})}^{\frac{1}{2}} \|\langle \mathbf{y} \rangle\|_{\mathbf{L}^p(\mathbf{f})} \| |\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}|^{\frac{1}{2}} [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \|_{\mathbf{L}^2(\mathbf{f})},$$

with $\theta' = (d-1) \left(\frac{1}{p} - \frac{1}{2} \right)$ and $h_{\mathbf{f}}$ the size of the sub-face $\mathbf{f}$. Hence, denoting $\sum_{\mathbf{f}} = \sum_{\mathbf{X} \in \{\mathbf{F}^\circ, \mathbf{C}\}} \sum_{\mathbf{x} \in \mathbf{X}} \sum_{\mathbf{f} \in \mathfrak{F}_{\mathbf{E};\mathbf{X}}}$, it follows from the triangle inequality, the Hölder inequality and $p' \geq 2$ that

$$\left| \sum_{\mathbf{f}} \int_{\mathbf{f}} (\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}) [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \cdot \langle \mathbf{y} \rangle \right| \leq \left(\sum_{\mathbf{f}} h_{\mathbf{f}}^{\theta' p} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{f})}^{\frac{p}{2}} \|\langle \mathbf{y} \rangle\|_{\mathbf{L}^p(\mathbf{f})}^p \right)^{\frac{1}{p}} \left(\sum_{\mathbf{f}} \| |\boldsymbol{\beta} \cdot \mathbf{n}_{\mathbf{f}}|^{\frac{1}{2}} [\mathbf{L}_{\mathcal{E}}(\mathbf{v})] \|_{\mathbf{L}^2(\mathbf{f})}^2 \right)^{\frac{1}{2}}.$$

Next, owing to the definitions (6.20) and (6.22) of $\mathbf{n}_{\boldsymbol{\beta}}$ and $\mathbf{s}_{\boldsymbol{\beta}}$ respectively, mesh regularity and the inequality $|a \pm b|^p \leq 2^{p-1}(|a|^p + |b|^p)$, we infer that

$$|T_2 + T_3| \lesssim \left(\sum_{\mathbf{c} \in \mathbf{C}} \sum_{\mathbf{c} \in \mathfrak{C}_{\mathbf{E},\mathbf{c}}} h_{\mathbf{c}}^{\theta' p} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{c})}^{\frac{p}{2}} \|\mathbf{y}\|_{\mathbf{L}^p(\partial \mathbf{c})}^p \right)^{\frac{1}{p}} (\gamma_{\mathbf{b}} \mathbf{s}_{\boldsymbol{\beta}}(\mathbf{v}, \mathbf{v}))^{\frac{1}{2}}.$$

Proceeding similarly, we end up with

$$|T_4| \lesssim \left(\sum_{\mathbf{f} \in \mathbf{F}^\partial} h_{\mathbf{c}(\mathbf{f})}^{\theta' p} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{f})}^{\frac{p}{2}} \|\mathbf{y}\|_{\mathbf{L}^p(\mathbf{f})}^p \right)^{\frac{1}{p}} \langle \mathbf{H}_{|\boldsymbol{\beta} \cdot \mathbf{n}|}^{\varepsilon, \partial}(\mathbf{v}), \mathbf{v} \rangle_{\varepsilon}^{\frac{1}{2}},$$

yielding the final result. $\square$

When the solution $\mathbf{u}$ is smoother, so as to be in the domain of the de Rham reduction map $\mathbf{R}_{\mathcal{E}}$ defined by (6.2), we can bound instead the consistency error

$$\mathbb{E}_{\mathcal{E}}(\mathbf{u}) = \sup_{\mathbf{v} \in \mathcal{E}; \|\mathbf{v}\|_{\varepsilon, \mathbf{a}}=1} \left| \mathbf{A}_{\boldsymbol{\beta}, \boldsymbol{\mu}}^{\varepsilon}(\mathbf{R}_{\mathcal{E}}(\mathbf{u}), \mathbf{v}) - \mathbb{S}(\mathbf{s}, \mathbf{u}_D; \mathbf{v}) \right|. \quad (6.43)$$

We omit the proof since it follows the same ideas as the proof of Lemma 6.9.

Lemma 6.10 (Consistency error). *Let $\mathbf{u}$ be the exact solution of (6.25) and assume that $\mathbf{u}$ satisfies $\mathbf{u} \in \mathbf{W}^{1,p}(\mathbf{C})$ and $\nabla \times \mathbf{u} \in \mathbf{L}^{\frac{2p}{3-p}}(\mathbf{C})$ with $p \in \left(\frac{3}{2}, 2 \right]$. Then,*

$$\mathbb{E}_{\mathcal{E}}(\mathbf{u}) \lesssim \left(\sum_{\mathbf{c} \in \mathbf{C}} N_{\infty; \mathbf{c}}^p \tau^{\frac{p}{2}} h_{\mathbf{c}}^{\frac{d}{2}(p-2)} \|\mathbf{u} - \mathbf{l}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{u})\|_{\mathbf{L}^p(\mathbf{c})}^p + \sum_{\mathbf{c} \in \mathbf{C}} \sum_{\mathbf{c} \in \mathfrak{C}_{\mathbf{E}, \mathbf{c}}} \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{c})}^{\frac{p}{2}} h_{\mathbf{c}}^{\frac{(d-1)}{2}(p-2)} \|\mathbf{u} - \mathbf{l}_{\mathcal{E}_{\mathbf{c}}}(\mathbf{u})\|_{\mathbf{L}^p(\partial \mathbf{c})}^p \right)^{\frac{1}{p}},$$

with $d = 3$, $N_{\infty; \mathbf{c}} = \|\nabla \boldsymbol{\beta} + \boldsymbol{\mu}^T - (\nabla \cdot \boldsymbol{\beta}) \text{Id}\|_{\mathbf{L}^\infty(\mathbf{c})}$ and with the interpolation map $\mathbf{l}_{\mathcal{E}_{\mathbf{c}}} = \mathbf{L}_{\mathcal{E}_{\mathbf{c}}} \circ \mathbf{R}_{\mathcal{E}_{\mathbf{c}}}$.

Theorem 6.11 (A priori error estimate). *Let $\mathbf{u}$ be the unique solution of (6.7) and let $\mathbf{u}$ be the discrete solution of (6.25). Assume that one of the stability assumption from Table 6.1 holds. Assume that the exact solution satisfies $\mathbf{u} \in \mathbf{W}^{1,p}(\Omega)$ with $p \in \left(\frac{3}{2}, 2 \right]$. Then,*

$$\|\mathbf{u} - \widehat{\mathbf{R}}_{\mathcal{E}}(\mathbf{u})\|_{\varepsilon, \mathbf{a}} \lesssim \left(\sum_{\mathbf{c} \in \mathbf{C}} \left(N_{\infty; \mathbf{c}}^p \tau^{\frac{p}{2}} h_{\mathbf{c}}^{\frac{p}{2}} + \|\boldsymbol{\beta}\|_{\mathbf{L}^\infty(\mathbf{c})}^{\frac{p}{2}} \right) h_{\mathbf{c}}^{2p-3} |\mathbf{u}|_{\mathbf{W}^{1,p}(\hat{\mathbf{c}})}^p \right)^{\frac{1}{p}},$$

with $N_{\infty; \mathbf{c}} = \|\nabla \boldsymbol{\beta} + \boldsymbol{\mu}^T - (\nabla \cdot \boldsymbol{\beta}) \text{Id}\|_{\mathbf{L}^\infty(\mathbf{c})}$.

Proof. This result follows from Lemma 6.9 with Proposition 6.12 below, along with the multiplicative trace inequality from Lemma 7.25. $\square$

Proposition 6.12 (Interpolation error). *Let $p \in [1, \infty)$. Then, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|\mathbf{v} - \hat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v})\|_{L^p(c)} \leq \mathcal{C}_{\text{INT}} h_c |\mathbf{v}|_{\mathbf{W}^{1,p}(\hat{c})}, \quad \forall \mathbf{v} \in \mathbf{W}^{1,p}(\hat{c}). \quad (6.44)$$

Proof. See the proof of Proposition 7.44. $\square$

Theorem 6.13 (*A priori error estimate*). *Let $\mathbf{u}$ be the unique solution of (6.7) and let $\mathbf{u}$ be the discrete solution of (6.25). Assume that one of the stability assumption from Table 6.1 holds. Assume that the exact solution $\mathbf{u}$ satisfies $\mathbf{u} \in \mathbf{W}^{1,p}(\mathcal{C})$ and $\nabla \times \mathbf{u} \in \mathbf{L}^{\frac{2p}{3-p}}(\mathcal{C})$ with $p \in (\frac{3}{2}, 2]$. Then,*

$$\|\mathbf{u} - \mathbf{R}_{\mathcal{E}}(\mathbf{u})\|_{\mathcal{E},a} \lesssim \left(\sum_{c \in \mathcal{C}} \left(N_{\infty;c}^p \tau^{\frac{p}{2}} h_c^{\frac{p}{2}} + \|\beta\|_{L^\infty(c)}^{\frac{p}{2}} \right) h_c^{2p-3} \left(|\mathbf{u}|_{\mathbf{W}^{1,p}(c)}^p + h_c^{\frac{3(p-1)}{2}} \|\nabla \times \mathbf{u}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)}^p \right) \right)^{\frac{1}{p}},$$

with $N_{\infty;c} = \|\nabla \beta + \boldsymbol{\mu}^T - (\nabla \cdot \beta) \text{Id}\|_{L^\infty(c)}$.

Proof. This result follows from Lemma 6.9 with Proposition 6.14 below, along with the multiplicative trace inequality from Lemma 7.25. $\square$

Proposition 6.14 (Interpolation error). *Let $p \in (\frac{3}{2}, 2]$ and let $c \in \mathcal{C}$. Then, for all $\mathbf{v} \in \mathbf{W}^{1,p}(c)$ be such that $\nabla \times \mathbf{v} \in \mathbf{L}^{\frac{2p}{3-p}}(c)$, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\|\mathbf{v} - \mathbf{l}_{\mathcal{E}_c}(\mathbf{v})\|_{L^p(c)} \leq \mathcal{C}_{\text{INT}} h_c \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right). \quad (6.45)$$

Proof. See the proof of Proposition 7.42. $\square$

Remark 6.15 (Regularity of the exact solution). *Observe that the regularity assumption in Theorem 6.11 is stronger than the regularity needed in Lemma 6.9 to compute the consistency error.*

6.3 Numerical results

In this section, we investigate numerically the edge-based scheme (6.25) for the vector advection-reaction problem (6.1). To our knowledge, there does not exist any three-dimensional test case for this problem in the literature. Hence, we propose to study the performance of our scheme when the exact solution is adapted from the Green-Taylor vortex velocity field. The computational domain is $\Omega = [0, 1]^3$ and we consider the four mesh sequences H, PrT, PrG and CB, previously considered in Chapters 3-5, and depicted in Figure 6.2.

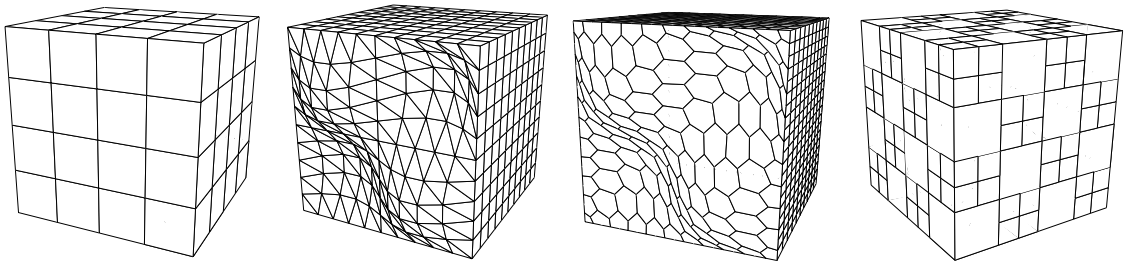


Figure 6.2 – The four mesh sequences H, PrT, PrG and CB, respectively.

Following the analysis of Section 6.2.2, the convergence analysis is performed by computing the following relative discrete L^2 -norms:

$$\widehat{\mathbf{Er}}_{\mathcal{E}}(\mathbf{u}) = \left(\frac{\sum_{e \in \mathcal{E}} |\mathbf{c}_e| |e|^{-2} (\mathbf{u}_e - \widehat{\mathbf{R}}_{\mathcal{E}}(\mathbf{u})|_e)^2}{\sum_{e \in \mathcal{E}} |\mathbf{c}_e| |e|^{-2} \widehat{\mathbf{R}}_{\mathcal{E}}(\mathbf{u})|_e^2} \right)^{\frac{1}{2}}, \quad (6.46a)$$

$$\mathbf{Er}_{\mathcal{E}}(\mathbf{u}) = \left(\frac{\sum_{e \in \mathcal{E}} |\mathbf{c}_e| |e|^{-2} (\mathbf{u}_e - \mathbf{R}_{\mathcal{E}}(\mathbf{u})|_e)^2}{\sum_{e \in \mathcal{E}} |\mathbf{c}_e| |e|^{-2} \mathbf{R}_{\mathcal{E}}(\mathbf{u})|_e^2} \right)^{\frac{1}{2}}, \quad (6.46b)$$

where $\mathbf{u}$ denotes the exact solution and $\mathbf{u}$ the discrete solution. We also recall the definition of the computational cost $\mathbf{Co} = \mathbf{nnz} \times n_{\text{ite}}$ used to measure the computational efficiency of the scheme, where $\mathbf{nnz}$ is the total number of non-zeroes in the final system matrix and where n_{ite} is the number of iterations needed to bring the relative residual below 10^{-12} , using a bi-Conjugate gradient method preconditioned with an incomplete LU factorization (see Balay *et al.* (2016)).

6.3.1 Test case 5. Vortex solution

The exact solution $\mathbf{u} : \Omega \mapsto \mathbb{R}^3$ of this test case is defined as a combination of sine and cosine functions on the three components:

$$\mathbf{u}(x, y, z) = \begin{pmatrix} \sin(\pi x) \cos(\pi y/2) \cos(\pi z/2) \\ \cos(\pi x/2) \sin(\pi y) \cos(\pi z/2) \\ \cos(\pi x/2) \cos(\pi y/2) \sin(\pi z) \end{pmatrix}. \quad (6.47)$$

This solution vanishes at the boundary $\partial\Omega$ of the unit cube. The advective field $\boldsymbol{\beta} : \Omega \rightarrow \mathbb{R}^3$ and the reaction tensor $\boldsymbol{\mu} : \Omega \rightarrow \mathbb{R}^{3 \times 3}$ are defined as follows:

$$\boldsymbol{\beta}(x, y, z) = \frac{1}{4} \begin{pmatrix} x - 2y \\ y - 2x \\ -2(z + 1) \end{pmatrix} \quad \text{and} \quad \boldsymbol{\mu} = \mu \mathbf{Id}. \quad (6.48)$$

with $\mu : \Omega \rightarrow \mathbb{R}$. In Figure 6.3, we plot some flow lines of $\boldsymbol{\beta}$ and we highlight the inflow boundary $\partial\Omega^-$ in blue.

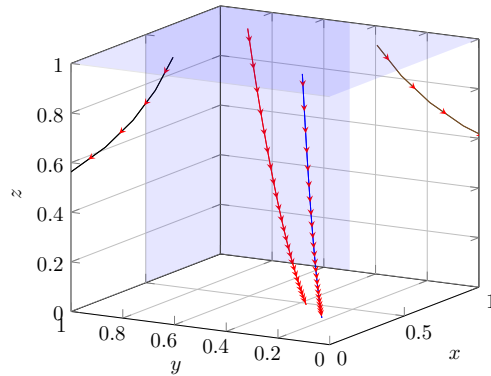


Figure 6.3 – Test case 5. Four flow lines of the advective field $\boldsymbol{\beta}$ and the corresponding inflow boundary $\partial\Omega^-$ in blue.

that we have

$$\nabla \boldsymbol{\beta}^T = \begin{pmatrix} 1/4 & -1/2 & 0 \\ -1/2 & 1/4 & 0 \\ 0 & 0 & -1/2 \end{pmatrix}.$$

Hence, the eigenvalues of the Friedrichs tensor $\sigma_{\beta,\mu}$ defined by (6.8), are $\{\mu - \frac{1}{2}, \mu, 2 + \mu\}$. The discrete problem (6.25) is then well-posed if $\mu > \frac{1}{2}$ owing to Lemma 6.5. If $\mu = \frac{1}{2}$, the discrete problem is still well-posed owing to Lemma 6.6 if the mesh-size satisfies $h < 4h_0$, for example with the potential $\zeta(\mathbf{x}) = (z + 1)^2$.

In Figure 6.4, we plot the discrete relative errors $\widehat{\mathbf{Er}}_{\mathcal{E}}(\mathbf{u})$ and $\mathbf{Er}_{\mathcal{E}}(\mathbf{u})$ with respect to $\#E$ when the reaction tensor is of magnitude $\mu \in \{5, 0.5\}$. The legend is collected in Table 6.2.

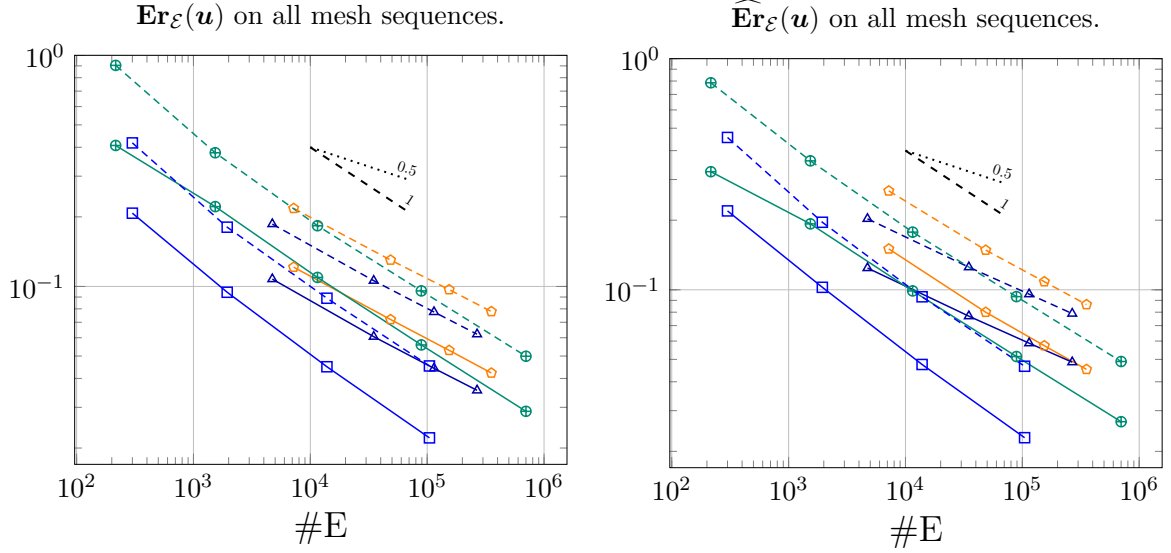


Figure 6.4 – Test case 5. The discrete relative error $\mathbf{Er}_{\mathcal{E}}(\mathbf{u})$ on the left and $\widehat{\mathbf{Er}}_{\mathcal{E}}(\mathbf{u})$ on the right, for the discrete problem (6.25) with $\mu = 5$ and $\mu = 0.5$.

H	PrT	PrG	CB	H	PrT	PrG	CB
$\mu = 5$				$\mu = 0.5$			

Table 6.2 – Test case 5. Legend of Figure 6.4.

Accuracy. The numerical results presented in Figure (6.4) show that the two discrete relative errors take similar values and converge for all mesh sequences. Similarly to our schemes devised in Chapters 3 and 5 for the scalar advection-reaction problem, we observe that if the lowest eigenvalue of the Friedrichs tensor $\sigma_{\beta,\mu}$ is positive, the scheme is more accurate than when the minimal eigenvalue is null (i.e., when $\mu = 0.5$). However, the hierarchy in terms of accuracy depends on the magnitude of μ . In the case $\mu = 5$, we have $H > \text{PrT} > \text{CB} > \text{PrG}$, whereas if $\mu = 0.5$, we have $H > \text{CB} > \text{PrT} > \text{PrG}$. Regarding the convergence rates, we expect from Theorems 6.11 and 6.13, a convergence rate of order $\frac{1}{2}$ for the two discrete relative error, since the solution is smooth. Figure 6.4 shows actually that the convergence rates are closer to 1 than to $\frac{1}{2}$.

Efficiency. In the left panel of Figure 6.5, we represent the error $\mathbf{Er}_{\mathcal{E}}(\mathbf{u})$ with respect to the computational cost Co . For a given mesh sequence, changing the value of the reaction coefficient does not significantly increases the cost to invert the system matrix, corresponding to the number of iterations n_{ite} since the two system have the same nnz . In essence, this indicates that the conditioning of the final matrix is not significantly influenced by the value of μ . In the right panel of Figure 6.5, we compare the total number of dofs when considering one dof per mesh edges, as in our scheme, with the total number of dofs when considering three

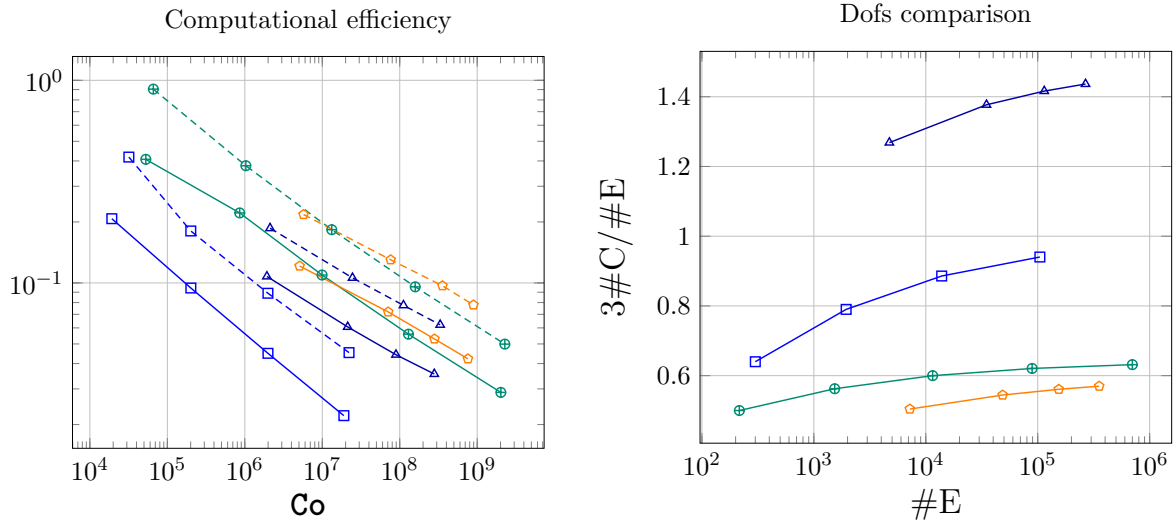


Figure 6.5 – Test case 5. Left panel: computational efficiency with the discrete relative error $\mathbf{Er}_{\mathcal{E}}(\mathbf{u})$. Right panel: comparison with dG methods of order $k = 0$ regarding the total number of degrees of freedom.

dofs per mesh cell for the three vector components, as in the standard finite volume schemes. This comparison shows that edge-based schemes are well-suited to meshes with cells having few edges (e.g., the **PrT** sequence). However, this is the converse for the **CB** and the **PrG** mesh sequences. Finally, edge-based and cell-based schemes leads asymptotically to a similar size of the final system for **H** mesh sequences.

Stencil. In Table 6.3, we collect the stencils of our scheme. We denote $\overline{\text{St}}$ the mean stencil, defined as $\text{nnz}/\#E$, and St.max the maximum stencil. This table illustrates in particular that the stencil of our scheme can take important values, especially if the mesh contains cells with a lot of edges (e.g., in the **PrG** and in the **CB** sequence). Investigations to reduce this stencil are left for future work.

H			PrT			PrG			CB		
$\#E$	$\overline{\text{St}}$	St.max	$\#E$	$\overline{\text{St}}$	St.max	$\#E$	$\overline{\text{St}}$	St.max	$\#E$	$\overline{\text{St}}$	St.max
3.0e+2	21	39	4.7e+3	38	70	7.2e+3	83	149	1.5e+3	112	272
1.9e+3	25	39	3.5e+4	46	70	4.9e+4	110	174	1.2e+4	144	272
1.4e+4	28	39	1.1e+5	48	70	1.5e+5	120	176	8.9e+4	162	272
1.0e+5	30	39	2.7e+5	49	70	3.5e+5	125	177	7.0e+5	180	276

Table 6.3 – Test case 5. Size of the system, mean stencil and maximum stencil.

Chapter 7

Analysis on polyhedral meshes

Contents

7.1 Polyhedral meshes	105
7.1.1 Primal mesh	105
7.1.2 Star-shaped and shape-regular mesh	107
7.1.3 Mesh regularity	109
7.2 Mesh partitions	109
7.2.1 Vertex-based partition	109
7.2.2 Edge-face-based partition	110
7.2.3 Edge-based partition	110
7.2.4 Abstract partition	111
7.2.5 The dual mesh	112
7.3 Functional inequalities in polyhedral meshes	113
7.3.1 Poincaré-Steklov inequality and polynomial approximation	114
7.3.2 Face and inverse inequalities on polyhedral cells	115
7.4 Reconstruction and approximation	116
7.4.1 Piece-wise constant vertex reconstruction	117
7.4.2 Piece-wise affine vertex-cell reconstruction	119
7.4.3 Piece-wise constant edge reconstruction	122

This chapter summarizes and generalizes the notions used in Chapters 3, 4, 5 and 6 to approximate on polyhedral meshes the solution of the scalar and the vector transport problems.

Section 7.1 introduces the main concepts attached to a polyhedral mesh M and introduces the notion of regular mesh. Section 7.2 presents the different partitions used in this thesis, with a special focus on the dual mesh. These partitions are employed in Section 7.3 to extend on general meshes the classical analytic inequalities (e.g., the Poincaré or the multiplicative trace inequalities) and in Section 7.4 to devise reconstruction and reduction maps.

7.1 Polyhedral meshes

Let Ω denote an open, connected, bounded, polyhedral subset of $\mathbb{R}^d$, with d the dimension. Following the work of Christiansen (2008) and of Bonelle (2014), this section introduces the essential concepts used to describe a primal mesh of Ω . These concepts are illustrated in particular for $d = 3$.

7.1.1 Primal mesh

We consider a mesh sequence $M_{\mathbb{N}} := (M_n)_{n \in \mathbb{N}}$, such that for all $n \in \mathbb{N}$, $M_n := \{M_{k,n} \mid k \in \llbracket 0, d \rrbracket\}$, where $M_{k,n}$ is a finite collection of non-empty open k -cells, defined as being homeomorphic by a bi-Lipschitz map to the open unit ball of $\mathbb{R}^k$ for the Euclidean metric.

Remark 7.1 (Generics names). *Generally speaking, for all dimension $d \geq 3$ and for all $n \in \mathbb{N}$, an element $m_0 \in M_{0,n}$ is called a vertex, an element $m_1 \in M_{1,n}$ is called an edge, an element $m_{d-1} \in M_{d-1,n}$ is called a facet and an element $m_d \in M_{d,n}$ is called a cell.*

In this thesis, mesh sequences are always assumed to define a refinement, i.e., satisfying $\#M_{d,n} \rightarrow +\infty$ as $n \rightarrow \infty$, where $\#$ denotes the cardinal. Henceforth, indices $n \in \mathbb{N}$ are omitted to alleviate the reading, so that $M = \{M_k \mid k \in \llbracket 0, d \rrbracket\}$.

Definition 7.2 (Boundary and closure of a k -cell). *Let $k \in \llbracket 1, d \rrbracket$. The boundary of the k -cell $m_k \in M_k$ is denoted by ∂m_k and is defined as*

$$\partial m_k = \cup \{m_s \in M_s \mid m_s \subset \overline{m_k}, s < k\},$$

where $\overline{m_k}$ denotes the closure of m_k . Conventionally, we set $\partial m_0 = \emptyset$ for all $m_0 \in M_0$.

Definition 7.3 (Cellular complex). *The mesh M defines a cellular complex of Ω if:*

- (Ca) *The union of all elements of M generates $\overline{\Omega}$: $\overline{\Omega} = \cup \{m \mid m \in M\}$.*
- (Cb) *For all $k \in \llbracket 1, d \rrbracket$, the boundary of a k -cell is composed of a uniformly bounded number of $(k-1)$ -cells.*
- (Cc) *Distinct elements have empty intersection.*

The first assumption (Ca) is equivalent to saying that the closure of the elements of M_d generates $\overline{\Omega}$, i.e., $\overline{\Omega} = \cup \{\overline{m} \mid m \in M_d\}$. Recursively, Assumption (Cb) means that for all $m \in M$, ∂m is composed of a uniformly bounded number of elements of M .

Definition 7.4 (Planar cellular complex). *The mesh M defines a planar cellular complex of Ω if it defines a cellular complex of Ω in the sense of Definition 7.3 and if:*

- (Cd) *For all $k \in \llbracket 1, d-1 \rrbracket$, each element of M_k is contained in some affine sub-space of dimension k .*

We now distinguish boundary and internal k -cells.

Proposition 7.5 (Boundary mesh). *For all $k \in \llbracket 0, d-1 \rrbracket$, the subset M_k^∂ of M_k collects all k -cells lying at the boundary $\partial\Omega$, i.e., $M_k^\partial := \{m_k \in M_k \mid m_k \subset \partial\Omega\}$. By convention, $M_d^\partial = \emptyset$. In addition, if M defines a cellular complex of Ω , the boundary mesh $M^\partial := \{M_k^\partial \mid k \in \llbracket 0, d-1 \rrbracket\}$ defines a cellular complex of $\partial\Omega$.*

Definition 7.6 (Internal mesh). *For all $k \in \llbracket 0, d-1 \rrbracket$, the subset M_k° of M_k collects all internal k -cells, i.e., $M_k^\circ = M_k \setminus M_k^\partial$. The internal mesh is then $M^\circ := M \setminus M^\partial$ and $M_d^\circ = M_d$.*

The following last definition introduces geometric subsets of M_k for all $k \in \llbracket 0, d \rrbracket$.

Definition 7.7 (Geometric mesh subsets). *Let $k, s \in \llbracket 0, d \rrbracket$ such that $k \neq s$ and let $m_k \in M_k$. The subset $M_{s;m_k}$ of M_s is defined by*

$$M_{s;m_k} := \{m_s \in M_s \mid m_s \subset \partial m_k\} \quad \text{if } k > s, \quad (7.1)$$

or by

$$M_{s;m_k} := \{m_s \in M_s \mid m_k \subset \partial m_s\} \quad \text{if } s > k. \quad (7.2)$$

Let us illustrate these concepts for $d = 3$. Recalling Remark 7.1, we use the following classical terminology in $\mathbb{R}^3$: an element $m_0 \in M_0$ is called a vertex and is denoted by $v \in V$, an element $m_1 \in M_1$ is called an edge and is denoted by $e \in E$, an element $m_2 \in M_2$ is called a face and is denoted by $f \in F$ and a element $m_3 \in M_3$ is called a cell and is denoted by $c \in C$. A primal mesh of the subset Ω of $\mathbb{R}^3$ is then denoted by $M := \{V, E, F, C\}$. A simple three-dimensional polyhedral mesh composed of two polygonal prisms is depicted in Figure 7.1. Owing to Definition 7.7, we obtain e.g., for all $v \in V$ and for all $e \in E$, the subsets

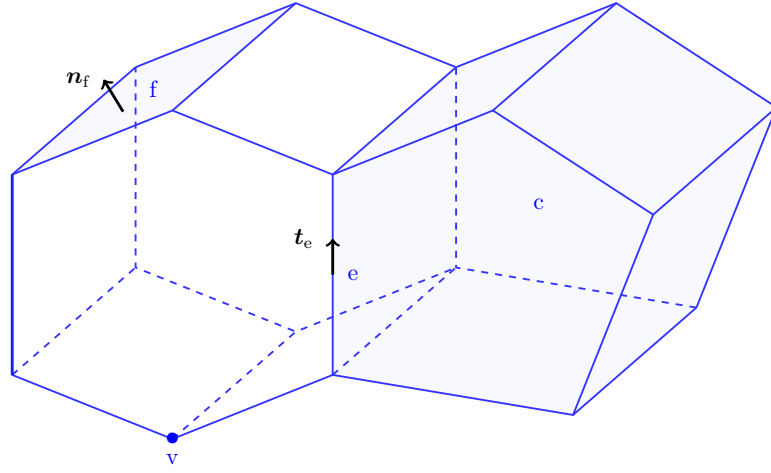


Figure 7.1 – An oriented polyhedral mesh containing the cell $c \in \mathcal{C}$, the face $f \in \mathcal{F}$, the edge $e \in \mathcal{E}$ and the vertex $v \in \mathcal{V}$.

$V_e := \{v \in \mathcal{V} \mid v \subset \partial e\}$ and $E_v := \{e \in \mathcal{E} \mid v \subset \partial e\}$.

We additionally introduce the notion of orientation: for all $e \in \mathcal{E}$, we assign an arbitrary tangential unit vector $\mathbf{t}_e$ and for all $f \in \mathcal{F}$ we assign an arbitrary normal unit vector $\mathbf{n}_f$. In particular, we denote

$$\mathbf{e} = \int_e \mathbf{t}_e \quad \text{and} \quad \mathbf{f} = \int_f \mathbf{n}_f. \quad (7.3)$$

Remark 7.8 (Orientation in the general case). *The notion of orientation can be extended to all dimension $d \geq 1$ and to all mesh elements using the notion of oriented atlas. Further insight can be found e.g., in Gerritsma (2012).*

7.1.2 Star-shaped and shape-regular mesh

Definition 7.9 (Barycenters). *For all $m \in \mathcal{M}$, $\mathbf{x}_m$ denotes the barycenter of m in the Cartesian basis of $\mathbb{R}^d$, i.e.,*

$$\mathbf{x}_m := \frac{1}{|m|} \int_m \mathbf{x},$$

where $|m|$ denotes the $\dim(m)$ -dimensional Lebesgue measure of m .

Definition 7.10 (Star-shaped mesh). *The mesh $\mathcal{M}$ is star-shaped if:*

(**St**) *For all $k \in \llbracket 1, d \rrbracket$, each element of $\mathcal{M}_k$ is star-shaped with respect to its barycenter.*

Such an assumption (**St**) is sufficient to assert optimal polynomial approximation (see Brenner & Scott (1994)). Here, we have another route to prove polynomial approximation and the motivation is different since assumption (**St**) allows us to consider sub-meshes (see Section 7.2). In Figure 7.2, we present two non-convex polyhedral cells with two sustentation faces; the left cell satisfies the star-shaped assumption whereas the right cell does not, since this cell is not star-shaped with respect to its barycenter.

Definition 7.11 (Elements size). *Let $\mathcal{M}$ be a planar cellular complex. Let $k \in \llbracket 1, d \rrbracket$ and let $m_k \in \mathcal{M}_k$. The circumscribed diameter (also called the size) of m_k is denoted by h_{m_k} and is defined as the diameter of the smallest k -ball containing m_k . The inscribed diameter of m_k is denoted by r_{m_k} and is defined as the diameter of the largest k -ball that can be inscribed in m_k .*

Definition 7.12 (Shape regularity). *The mesh $\mathcal{M}$ is shape regular if:*

(**Sha**) *There is a uniform constant $\mathcal{C}_M > 0$ with respect to the mesh refinement such that*

$$\forall k \in \llbracket 1, d \rrbracket, \forall m_k \in \mathcal{M}_k, \quad \mathcal{C}_M h_{m_k} \leq r_{m_k}.$$

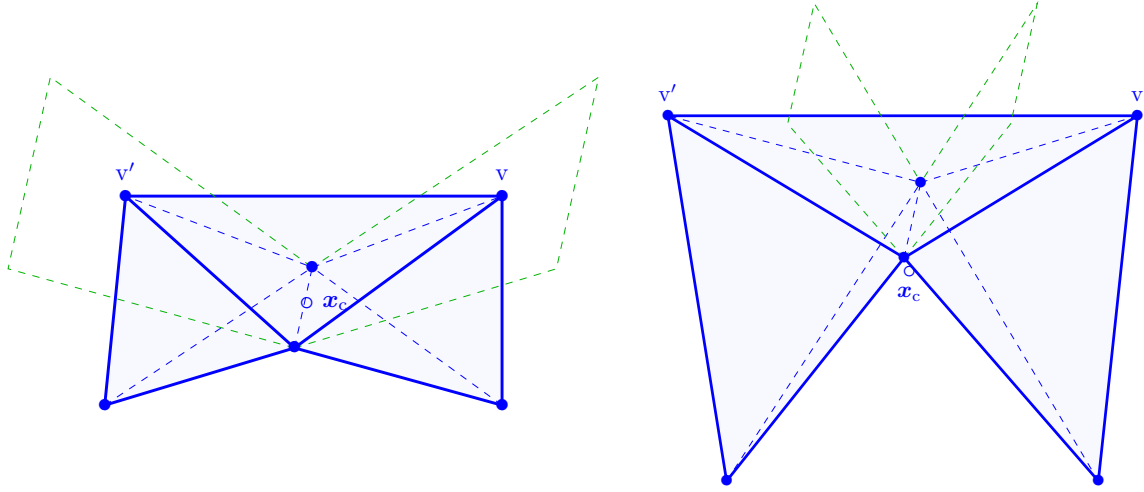


Figure 7.2 – Two polyhedral cells in blue with two sustentation faces. The two vertices v and v' lie above and below these planes for the left cell and the right cell, respectively. Hence, the left cell does satisfy **(St)** whereas the right cell does not.

(Shb) For all $(m_0, \dots, m_d) \in \times_{k \in \llbracket 0, d \rrbracket} M_k$ such that $m_k \in M_{k; m_{k+1}}$ for all $k \in \llbracket 0, d-1 \rrbracket$, there are uniform constants $\mathcal{C}_d, \dots, \mathcal{C}_2 > 0$ such that

$$\begin{aligned} \mathcal{C}_d h_{m_d}^d &\leq |\text{CO}\{\mathbf{x}_{m_0}, \dots, \mathbf{x}_{m_d}\}|, \\ \mathcal{C}_{d-1} h_{m_{d-1}}^{d-1} &\leq |\text{CO}\{\mathbf{x}_{m_0}, \dots, \mathbf{x}_{m_{d-1}}\}|, \\ &\dots \\ \mathcal{C}_2 h_{m_2}^2 &\leq |\text{CO}\{\mathbf{x}_{m_0}, \mathbf{x}_{m_1}, \mathbf{x}_{m_2}\}|, \end{aligned}$$

where for any set $\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$, $\text{CO}\{\mathbf{x}_1, \dots, \mathbf{x}_k\}$ denotes the interior of its convex hull (note that for all $k \in \llbracket 2, d \rrbracket$, $\text{CO}\{\mathbf{x}_{m_0}, \dots, \mathbf{x}_{m_k}\}$ turns out to be the smallest k -simplex containing the points $\mathbf{x}_{m_0}, \dots, \mathbf{x}_{m_k}$).

In dimension $d = 3$, Assumption **(Shb)** means that for all $c \in \mathcal{C}$, for all $f \in \mathcal{F}_c$, for all $e \in \mathcal{E}_f$ and for all $v \in \mathcal{V}_e$, there are $\mathcal{C}_3, \mathcal{C}_2 > 0$ such that

$$\mathcal{C}_3 h_c^3 \leq |\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}|, \text{ and } \mathcal{C}_2 h_f^2 \leq |\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f\}|,$$

where the 3-simplex $\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$ and the 2-simplex $\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f\}$ are depicted in Figure 7.3.

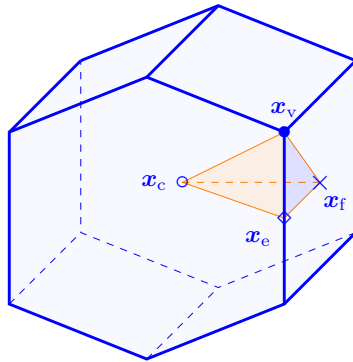


Figure 7.3 – For a cell c , a face $f \in \mathcal{F}_c$, an edge $e \in \mathcal{E}_f$ and a vertex $v \in \mathcal{V}_e$, we highlight the 3-simplex $\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\}$ in orange and the 2-simplex $\text{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f\}$ in blue.

7.1.3 Mesh regularity

Definition 7.13 (Mesh regularity ($\mathbf{M}$)). *We say that the mesh M is regular if it defines a planar cellular complex in the sense of Definition 7.4, if it is star-shaped regular in the sense of Definition 7.10 and if it is shape regular in the sense of Definition 7.12.*

We keep this assumption for the rest of this Chapter.

7.2 Mesh partitions

Mesh partitions play a important role in this thesis, in particular to define reconstruction maps mapping dofs spaces to functional spaces (see Section 7.4). These partitions are also very useful in other contexts, as in co-volume methods (see e.g., Hu & Nicolaides (1992)) or to extend the discrete Poincaré inequality for certain non-conforming functions in a general domain (see e.g., Vohralík (2005)). A particular mesh partition is the dual mesh, explicitly considered e.g., in the *Discrete Duality Finite Volume* (DDFV) method proposed by Hermeline (2000) and by Andreianov *et al.* (2012)) or in the *Mimetic Spectral Element* (MSE) method introduced by Kreeft *et al.* (2011) along with the recent *compatible isogeometric* schemes of Back & Sonnendrücker (2012, 2014).

This section introduces the concept of X -partition of Ω based on a given subset X of M . To simplify the reading, we only consider the case $d = 3$; however, we still write d instead of 3.

7.2.1 Vertex-based partition

For all $c \in C$, we define the set $\mathfrak{C}_{v,c} = \{\mathfrak{c}_{v,c}\}_{v \in V_c}$ which is composed of the sub-cells

$$\mathfrak{c}_{v,c} = \text{int} \left(\bigcup_{e \in E_v \cap E_c} \bigcup_{f \in F_e \cap F_c} \overline{CO}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\} \right), \quad \forall v \in V_c, \quad (7.4)$$

as shown in the left panel of Figure 7.4, and where $\text{int}(\omega)$ denotes the interior of any subset $\omega \in \mathbb{R}^d$. Note that $\mathfrak{C}_{v,c}$ is a V_c -partition of c owing to assumption ($\mathbf{St}$), i.e., $\bar{c} = \bigcup \{\overline{\mathfrak{c}_{v,c}} \mid v \in V_c\}$. The set $\mathfrak{F}_{v,c}$ collects intra-cell sub-faces induced by this partition and is defined as

$$\mathfrak{F}_{v,c} = \{\mathfrak{f}_{e,c} = \text{int}(\partial \mathfrak{c}_{v,c} \cap \partial \mathfrak{c}_{v',c}) \mid e \in E_c \text{ with } v, v' \in V_e\}, \quad (7.5)$$

and one element is depicted in the middle panel of Figure 7.4. Finally, for all $f \in F_c$, the set $\mathfrak{F}_{v,f}$ collects inter-cell sub-faces and is defined as

$$\mathfrak{F}_{v,f} = \{\mathfrak{f}_{v,f} = \text{int}(\partial \mathfrak{c}_{v,c} \cap f) \mid v \in V_f\}, \quad (7.6)$$

and one element is depicted in the right panel of Figure 7.4. The partition $\mathfrak{C}_{v,c}$ is particularly useful to express the coordinates of the barycenter of c and of its faces $f \in F_c$ as a convex combination of the coordinates of the cell vertices $v \in V_c$.

Proposition 7.14 (Cell and face barycenters). *For all $c \in C$ and for all $f \in F_c$, the following identities hold:*

$$\sum_{v \in V_c} |\mathfrak{c}_{v,c}|(\mathbf{x}_v - \mathbf{x}_c) = 0 \quad \text{and} \quad \sum_{v \in V_f} |\mathfrak{f}_{v,f}|(\mathbf{x}_v - \mathbf{x}_f) = 0. \quad (7.7)$$

Proof. See (Bonelle, 2014, Proposition 5.23). □

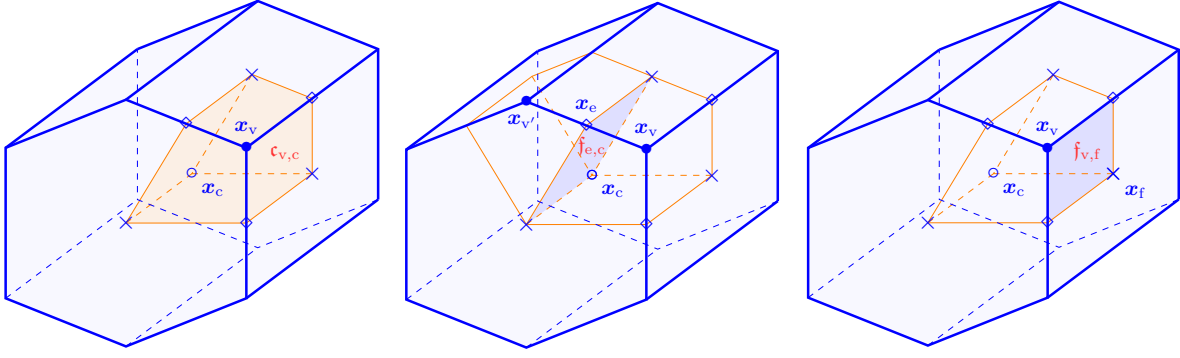


Figure 7.4 – From left to right: the sub-cell $c_{v,c}$ attached to the vertex v within the cell c , one intra-cell sub-face $f_{e,c} \in \mathfrak{F}_{v,c}$ and one inter-cell sub-face $f_{v,f} \in \mathfrak{F}_{v,f}$.

7.2.2 Edge-face-based partition

For all $c \in \mathcal{C}$, we define the set $EF_c = \{(e, f) \in E_c \times F_c \mid e \subset \partial f\}$. The edge-face-based partition $\mathfrak{C}_{EF,c} := \{c_{ef,c}\}_{(e,f) \in EF_c}$ is composed of the sub-cells $c_{ef,c}$ such that

$$c_{ef,c} := \text{int} \left(\bigcup_{v \in V_e} \overline{\text{CO}}\{x_v, x_e, x_f, x_c\} \right), \quad \forall (e, f) \in EF_c, \quad (7.8)$$

as shown in the left panel of Figure 7.5. Owing to assumption **(St)**, $\mathfrak{C}_{EF,c}$ is a EF_c -partition, i.e., $\bar{c} = \bigcup \{\bar{c}_{ef,c} \mid (e, f) \in EF_c\}$. This partition is of particular interest, since it is composed of simplices. Observe that each cell edge is connected to two tetrahedra $c_{ef,c}$ since $\#(F_e \cap F_c) = 2$, so that we have $\#\mathfrak{C}_{EF,c} = 2\#E_c$. The set collecting intra-cell sub-faces induced by this partition is defined as

$$\begin{aligned} \mathfrak{F}_{EF,c} = \{ & f_{ff',c} = \text{int}(\partial c_{ef,c} \cap \partial c_{ef',c}) \mid f, f' \in F_e \cap F_c \text{ for all } e \in E_c\} \\ & \cup \{f_{ee',c} = \text{int}(\partial c_{ef,c} \cap \partial c_{e'f,c}) \mid e, e' \in E_f \cap E_v \text{ for all } v \in V_f\}. \end{aligned} \quad (7.9)$$

with two elements illustrated in the middle panel of Figure 7.5. For all $f \in F_c$, the set collecting inter-cell sub-faces set is defined as

$$\mathfrak{F}_{EF,f} = \{f_{e,f} = \text{int}(\partial c_{ef,c} \cap f) \mid e \in E_f\}, \quad (7.10)$$

where one element is depicted in the right panel of Figure 7.5.

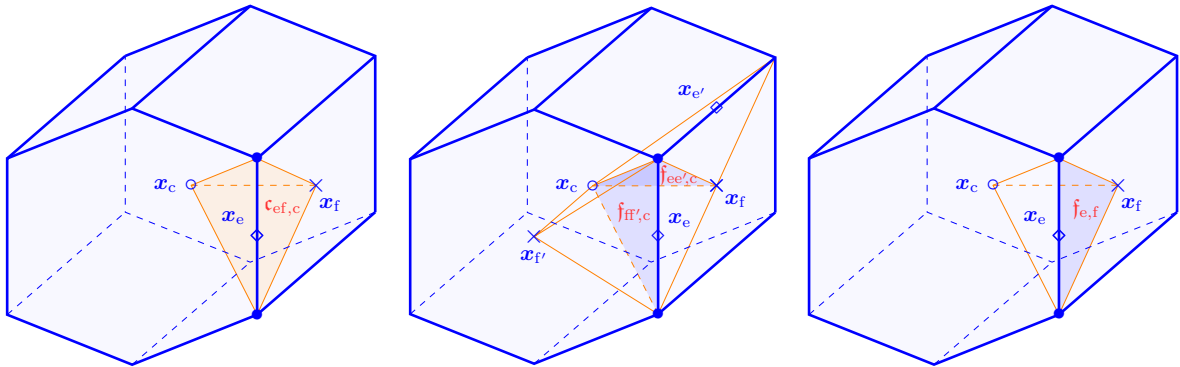


Figure 7.5 – From left to right: the sub-cell $c_{ef,c}$ attached to $(e, f) \in EF_c$, two intra-cell sub-faces $f_{ff',c} \in \mathfrak{F}_{EF,c}$ and $f_{ee',c} \in \mathfrak{F}_{EF,c}$ and one inter-cell sub-face $f_{e,f} \in \mathfrak{F}_{EF,f}$ with $f \in F_c$.

7.2.3 Edge-based partition

For all $c \in C$, the edge-based partition $\mathfrak{C}_{E,c} = \{\mathfrak{c}_{e,c}\}_{e \in E_c}$ collects all the sub-cells defined by

$$\mathfrak{c}_{e,c} = \text{int} \left(\bigcup_{f \in F_e \cap F_c} \bigcup_{v \in V_e} \overline{\text{CO}}\{\mathbf{x}_v, \mathbf{x}_e, \mathbf{x}_f, \mathbf{x}_c\} \right), \quad \forall e \in E_c, \quad (7.11)$$

as depicted in the left panel of Figure 7.6. Owing to assumption **(St)**, $\mathfrak{C}_{E,c}$ defines a E_c -partition, i.e., $\bar{c} = \bigcup \{\bar{\mathfrak{c}}_{e,c} \mid e \in E_c\}$. Observe that $\mathfrak{c}_{e,c} = \bigcup \{\mathfrak{c}_{ef,c} \mid f \in F_e \cap F_c\}$ and that $\#\mathfrak{C}_{E,c} = \#E_c$, so that $\mathfrak{C}_{EF,c}$ can be interpreted as a sub-partition of $\mathfrak{C}_{E,c}$. The set collecting intra-cell sub-faces corresponds to

$$\mathfrak{F}_{E,c} = \{\mathfrak{f}_{ee',c} = \text{int}(\partial\mathfrak{c}_{e,c} \cap \partial\mathfrak{c}_{e',c}) \mid e, e' \in E_f \cap E_v \text{ for all } f \in F_c \text{ and } v \in V_f\}, \quad (7.12)$$

with one element illustrated in the middle panel of Figure 7.6. Note that $\mathfrak{F}_{E,c} \subset \mathfrak{F}_{EF,c}$. For all $f \in F_c$, the set collecting inter-cell sub-faces of f is finally

$$\mathfrak{F}_{E,f} = \{\mathfrak{f}_{e,f} = \text{int}(\partial\mathfrak{c}_{e,c} \cap f) \mid e \in E_f\}, \quad (7.13)$$

with one element depicted in the right panel of Figure 7.6.

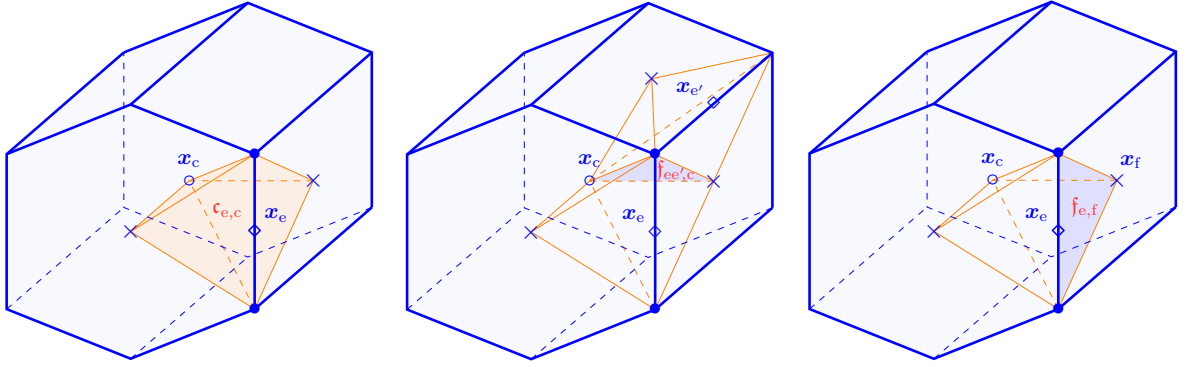


Figure 7.6 – From left to right: a sub-cell $\mathfrak{c}_{e,c}$ attached to $e \in E_c$, one intra-cell sub-face $\mathfrak{f}_{ee',c} \in \mathfrak{F}_{E,c}$ and one inter-cell sub-face $\mathfrak{f}_{e,f} \in \mathfrak{F}_{E,f}$.

Proposition 7.15 (Volume of a sub-cell). *Let $c \in C$ and let $e \in E_c$ with $\mathfrak{f}_{e,c} \in \mathfrak{F}_{v,c}$ defined by (7.5). Then, $d|\mathfrak{c}_{e,c}| = |\mathbf{e} \cdot \mathfrak{f}_{e,c}|$ with $\mathbf{e} = \int_e \mathbf{t}_e$ and $\mathfrak{f}_{e,c} = \int_{\mathfrak{f}_{e,c}} \mathbf{n}_{\mathfrak{f}_{e,c}}$.*

Proof. Observing that $\mathfrak{c}_{e,c}$ is composed of two pyramids of basis $\mathfrak{f}_{e,c}$ and of apex one of the two vertex of e , the identity follows from the volume of each of them. $\square$

7.2.4 Abstract partition

Let us consider $c \in C$ a cell of the mesh M and let X be a subset of $\{M_k \mid k \in \llbracket 0, d-1 \rrbracket\}$. The set $\mathfrak{C}_{X,c}$ is composed of the open d -cells $\mathfrak{c}_{x,c}$ with $x \in X_c$, also called *sub-cells*, where X_c is the corresponding subset of $\{M_{k;c} \mid k \in \llbracket 0, d-1 \rrbracket\}$. Then, we have $\mathfrak{C}_{X,c} = \{\mathfrak{c}_{x,c}\}_{x \in X_c}$.

Definition 7.16 (X_c -partition). *$\mathfrak{C}_{X,c}$ defines an X_c -partition of c if $\bar{c} = \bigcup \{\bar{\mathfrak{c}} \mid \mathfrak{c} \in \mathfrak{C}_{X,c}\}$.*

Owing to this definition, an X_c -partition defines a sub-mesh of c and it is possible to define the associated k -sub-cells for all $k \in \llbracket 0, d-1 \rrbracket$. For the sake of brevity, we only detail the $(d-1)$ -sub-cells, also called sub-faces.

Definition 7.17 ($(d-1)$ -connected sub-cells). *Two sub-cells are said to be $(d-1)$ -connected if their intersection is a $(d-1)$ -sub-cell, i.e., a sub-face.*

Two kind of sub-faces induced by the X_c -partition are considered: intra-cell sub-faces lying inside a cell c and inter-cell sub-faces lying on the boundary ∂c .

Definition 7.18 (Intra-cell sub-faces). *For all $c \in C$, Intra-cell sub-faces of c are collected in the set*

$$\mathfrak{F}_{X,c} = \{f = \text{int}(\partial c \cap \partial c') \mid c, c' \in \mathfrak{C}_{X,c} \text{ such that } c \text{ and } c' \text{ are } (d-1)\text{-connected}\}. \quad (7.14)$$

Definition 7.19 (Inter-cell sub-faces). *For all $c \in C$ and for all $f \in F_c$, inter-cell sub-faces of c are collected in the set*

$$\mathfrak{F}_{X,f} = \{f = \text{int}(\partial c \cap f) \mid c \in \mathfrak{C}_{X,c} \text{ such that } c \text{ and } f \text{ are } (d-1)\text{-connected}\}. \quad (7.15)$$

Collecting these local X_c -partitions, the global X -partition is defined by $\mathfrak{C}_X := \{\mathfrak{C}_{X,c}\}_{c \in C}$. Moreover, the set of internal sub-faces is defined by $\mathfrak{F}_X^\circ := \{\{\mathfrak{F}_{X,c}\}_{c \in C}, \{\mathfrak{F}_{X,f}\}_{f \in F^\circ}\}$ and that of boundary sub-face by $\mathfrak{F}_X^\partial := \{\mathfrak{F}_{X,f}\}_{f \in F^\partial}$.

As for the primal faces, sub-faces $f \in \{\mathfrak{F}_X^\circ, \mathfrak{F}_X^\partial\}$ are oriented along a unit normal vector $\mathbf{n}_f$. Since mesh faces are planar, the normal $\mathbf{n}_f$ is chosen such that $\mathbf{n}_f \cdot \mathbf{n}_f \geq 0$ in f for all $f \in \mathfrak{F}_{X,f}$ with $f \in F^\circ$ and such that $\mathbf{n} \cdot \mathbf{n}_f \geq 0$ in f for all $f \in \mathfrak{F}_X^\partial$, with $\mathbf{n}$ the outward normal to Ω .

Definition 7.20 (Jumps and averages on $\mathfrak{F}_X^\circ$). *For all $f \in \mathfrak{F}_X^\circ$ such that $f = \text{int}(\partial c \cap \partial c')$ with $c, c' \in \mathfrak{C}_X$ and such that $\mathbf{n}_f$ points from c to c' , we define, respectively, the jump and the average on f of a function v as*

$$[[v]]_f = v|_c - v|_{c'} \quad \text{and} \quad \{v\}_f = \frac{1}{2} (v|_c + v|_{c'}).$$

7.2.5 The dual mesh

The dual mesh is built from the vertex based partitions $\{\mathfrak{C}_{V,c}\}_{c \in C}$. The material of this section slightly differs from that of Bonelle (2014).

Internal dual mesh. As for the primal mesh, the internal dual mesh is denoted by $\tilde{M}$ and is composed of dual vertices $\tilde{v} \in \tilde{V}$, dual edges $\tilde{e} \in \tilde{E}$, dual faces $\tilde{f} \in \tilde{F}$, and dual cells $\tilde{c} \in \tilde{C}$, so that $\tilde{M} = \{\tilde{V}, \tilde{E}, \tilde{F}, \tilde{C}\}$. The key concept of dual meshes is the one-to-one pairing between primal and dual entities: each primal vertex $v \in V$ is associated with a unique dual cell $\tilde{c}(v) \in \tilde{C}$, each primal edge $e \in E$ is associated with a unique dual face $\tilde{f}(e) \in \tilde{F}$, and so on. Following this principle, the definition of a dual mesh is however not unique. Hereafter, we only present the so-called *fully barycentric* dual mesh based on the barycentric coordinates of the elements of M . Recalling the construction of the vertex-based partitions $\mathfrak{C}_{V,c}$ for all $c \in C$ in Section 7.2.1, we define dual cells as

$$\tilde{c}(v) = \text{int} \left(\bigcup \{ \overline{c_{v,c}} \mid c \in C_v \} \right), \quad \forall v \in V, \quad (7.16)$$

with $c_{v,c}$ the sub-cell defined by (7.4). Dual faces correspond to the $(d-1)$ -dimensional boundary of these dual cells and are defined as

$$\tilde{f}(e) = \text{int} \left(\bigcup \{ \overline{f_{e,c}} \mid c \in C_e \} \right), \quad \forall e \in E, \quad (7.17)$$

with the intra-cell sub-face $f_{e,c}$ defined by (7.5). Owing to (7.16) and (7.17), we have $c_{v,c} = \tilde{c}(v) \cap c$ and $f_{e,c} = \tilde{f}(e) \cap c = \tilde{f}_c(e)$. Hence, $c_{v,c}$ can be interpreted as the local dual cell attached to the vertex v , whereas $f_{e,c}$ corresponds to the local dual face attached to the edge e in the cell $c \in C$. For the sake of completeness, we also introduce dual edges and dual vertices, although they are not used in this thesis. For all $f \in F$, dual edges are defined by $\tilde{e}(f) = \text{int}(\bigcup \{ \overline{C_O\{x_f, x_c\}} \mid c \in C_f \})$ and for all $c \in C$, dual vertices are defined by $\tilde{v}(c) \equiv x_c$. Figure 7.7 highlights a portion of the dual mesh of the primal mesh depicted in Figure 7.1. In particular, we observe that dual faces and dual edges are oriented along the unit vector attached to their respective primal entity. The proposition below expresses the volume of each primal mesh cell in terms of its primal edges and its local dual faces.

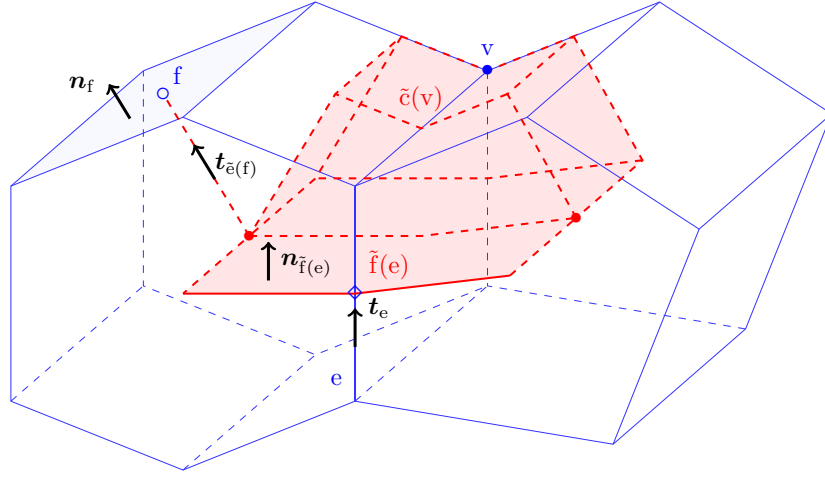


Figure 7.7 – The oriented primal mesh and a portion of the oriented dual mesh.

Proposition 7.21. *Let M be a primal mesh and let $\widetilde{M}$ be its fully barycentric dual mesh. For all $c \in C$, we have*

$$\sum_{e \in E_c} \mathbf{e} \otimes \tilde{\mathbf{f}}_c(e) = \sum_{e \in E_c} \tilde{\mathbf{f}}_c(e) \otimes \mathbf{e} = |c| \text{Id}. \quad (7.18)$$

with $\mathbf{e} = \int_e \mathbf{t}_e$ and $\tilde{\mathbf{f}}_c(e) = \int_{\tilde{f}_c(e)} \mathbf{n}_{\tilde{f}_c(e)}$.

Proof. See e.g., Codecasa & Trevisan (2007) or Droniou & Eymard (2006). $\square$

A consequence of the one-to-one pairing principle is that the dual mesh $\widetilde{M}$ is composed of a finite collection of k -cells, $k \in \llbracket 0, d \rrbracket$, since $\#\widetilde{C} = \#V$, $\#\widetilde{F} = \#E$, and so on. However, it does not define a cellular complex in the sense of Definition 7.3 since **(Ca)** is not satisfied. Moreover, we observe that, even if the primal mesh satisfies the planar condition **(Cd)**, the dual mesh does not.

Boundary dual mesh. In order to satisfy the condition **(Ca)**, we additionally introduce the dual mesh attached to the boundary $\partial\Omega$ and denoted by $\widetilde{M}^\partial$. It is composed of boundary dual vertices $\tilde{v}^\partial \in \widetilde{V}^\partial$, boundary dual edges $\tilde{e}^\partial \in \widetilde{E}^\partial$ and boundary dual faces $\tilde{f}^\partial \in \widetilde{F}^\partial$, so that $\widetilde{M}^\partial = \{\widetilde{V}^\partial, \widetilde{E}^\partial, \widetilde{F}^\partial\}$. These entities are built following the same construction principle as for the internal dual mesh $\widetilde{M}$, namely the one-to-one principle, but within the $(d-1)$ -manifold $\partial\Omega$. In essence, each primal boundary vertex $v \in V^\partial$ is attached to a unique boundary dual face $\tilde{f}^\partial(v) \in \widetilde{F}^\partial$, each primal boundary edge $e \in E^\partial$ is attached to a unique boundary dual edge $\tilde{e}^\partial(e) \in \widetilde{E}^\partial$, and so on. Dual boundary faces are then defined by

$$\tilde{f}^\partial(v) := \text{int} \left(\bigcup \{ \overline{f_{v,f}} \mid f \in F^\partial \cap F_v \} \right), \quad \forall v \in V^\partial, \quad (7.19)$$

where the inter-cell sub-faces $f_{v,f}$ are defined by (7.6). Similarly, for all $e \in E^\partial$, boundary dual edges are defined by $\tilde{e}^\partial(e) := \text{int} \left(\bigcup \{ \overline{\text{CO}\{\mathbf{x}_e, \mathbf{x}_f\}} \mid f \in F_e \cap F^\partial \} \right)$ and for all $f \in F^\partial$, boundary dual vertices are defined by $\tilde{v}^\partial(f) \equiv \mathbf{x}_f$. Figure 7.8 highlights a portion of the boundary dual mesh $\widetilde{M}^\partial$ associated with the primal mesh presented in Figure 7.1.

Proposition 7.22 (Dual cellular complex). *Owing to Definition 7.3, the boundary dual mesh $\widetilde{M}^\partial$ defines a cellular complex of the boundary $\partial\Omega$ and the dual mesh $\{\widetilde{M}, \widetilde{M}^\partial\}$ defines a cellular complex of Ω .*

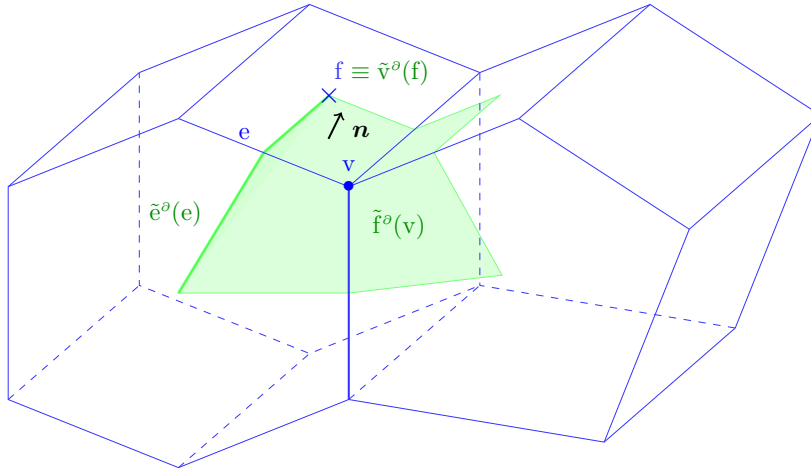


Figure 7.8 – Portion of the dual boundary mesh.

7.3 Functional inequalities in polyhedral meshes

This section briefly recalls functional inequalities in polyhedral meshes used to analyze our schemes, namely, the Poincaré(-Steklov) inequality, the multiplicative trace inequality, the inverse inequality, and the interpolation error of the Oswald operator (see Oswald (1993)). Following Ern & Guermond (2016), the following results can be extended to fractional Sobolev spaces.

7.3.1 Poincaré-Steklov inequality and polynomial approximation

Lemma 7.23 (Poincaré-Steklov inequality). *Let $p \in [1, \infty)$, let $q \in \llbracket 1, d \rrbracket$, and let $c \in \mathcal{C}$. Then, there is $\mathcal{C}_p > 0$, such that*

$$\|v - \bar{v}^c\|_{L^p(c; \mathbb{R}^q)} \leq \mathcal{C}_p h_c |v|_{W^{1,p}(c; \mathbb{R}^q)}, \quad v \in W^{1,p}(c; \mathbb{R}^q), \quad (7.20)$$

with $\bar{v}^c := |c|^{-1} \int_c v$.

The proof which relies on the compact embedding $W^{1,p}(c; \mathbb{R}^q) \hookrightarrow L^p(c; \mathbb{R}^q)$ does not define explicitly the constant $\mathcal{C}_p$. Whenever the mesh cell c is a convex set, one can take $\mathcal{C}_p = 1$ if $p = 1$ (see Acosta & Duran (2004)) and $\mathcal{C}_p = \pi^{-1}$ if $p = 2$ (see Bebendorf (2003)). We also recall that the estimate $\mathcal{C}_p \leq 2^{1-\frac{1}{p}} p^{\frac{1}{p}}$ holds in the general case. If the cell c is not convex, the constant $\mathcal{C}_p$ can take larger values; however, following Ern & Guermond (2016), it is possible to prove that it remains bounded if the mesh is regular in the sense of Definition 7.13.

A direct consequence of the Poincaré inequality is the following Corollary.

Corollary 7.24 (Polynomial approximation). *Let $p \in [1, \infty)$, let $q \in \llbracket 1, d \rrbracket$, let $k \in \{0, 1\}$, and let $c \in \mathcal{C}$. Then, there is $\mathcal{C}_{\text{POL}} > 0$ such that,*

$$\inf_{V \in \mathbb{P}_k(c; \mathbb{R}^q)} \left(\sum_{r=0}^k h_c^r |v - V|_{W^{r,p}(c; \mathbb{R}^q)} \right) \leq \mathcal{C}_{\text{POL}} h_c^{k+1} |v|_{W^{k+1,p}(c; \mathbb{R}^q)}, \quad \forall v \in W^{k+1,p}(c; \mathbb{R}^q). \quad (7.21)$$

Proof. Let $c \in \mathcal{C}$ and let us consider the case $q = 1$; the proof is identical for other choices of q . Owing to the Poincaré inequality (7.20), there exists $\mathcal{C}_p > 0$ such that

$$\|v - \bar{v}^c\|_{L^p(c)} \leq \mathcal{C}_p h_c |v|_{W^{1,p}(c)}, \quad \forall v \in W^{1,p}(c),$$

with $\bar{v}^c := \frac{1}{|c|} \int_c v \in \mathbb{P}_0(c; \mathbb{R})$, whence (7.21) for $k = 0$.

Let now $v \in W^{2,p}(\mathbf{c})$ and let us consider the affine polynomial $V(\mathbf{x}) := \bar{v}^{\mathbf{c}} + \bar{\nabla} v^{\mathbf{c}} \cdot (\mathbf{x} - \mathbf{x}_{\mathbf{c}})$, for all $\mathbf{x} \in \mathbf{c}$, where $\mathbf{x}_{\mathbf{c}}$ is the barycenter of $\mathbf{c}$. Then, applying the Poincaré inequality (7.20) with $q = d$, we infer that

$$|v - V|_{W^{1,p}(\mathbf{c})} = \|\nabla v - \bar{\nabla} v^{\mathbf{c}}\|_{L^p(\mathbf{c})} \leq \mathbf{C}_P h_{\mathbf{c}} |\nabla v|_{W^{1,p}(\mathbf{c})} \leq 2\mathbf{C}_P h_{\mathbf{c}} |v|_{W^{2,p}(\mathbf{c})},$$

since cross-derivatives are counted only once in the $W^{2,p}$ semi-norm. Moreover, observing that the function $v - V$ has zero mean-value in $\mathbf{c}$ by construction and applying once more the Poincaré inequality, we infer that

$$\|v - V\|_{L^p(\mathbf{c})} \leq \mathbf{C}_P h_{\mathbf{c}} |v - V|_{W^{1,p}(\mathbf{c})} \leq \mathbf{C}_P^2 h_{\mathbf{c}}^2 |v|_{W^{2,p}(\mathbf{c})},$$

owing to the above bound on $|v - V|_{W^{1,p}(\mathbf{c})}$, so that (7.21) follows for $k = 1$. $\square$

7.3.2 Face and inverse inequalities on polyhedral cells

Lemma 7.25 (Multiplicative trace inequality). *Let $p \in [1, \infty)$, let $q \in \llbracket 1, d \rrbracket$, and let $\mathbf{c} \in \mathcal{C}$. Let $\mathbf{f}$ be any planar sub-face contained in $\bar{\mathbf{c}}$. Then, there is $\mathbf{C}_T > 0$ such that*

$$\|v\|_{L^p(\mathbf{f}; \mathbb{R}^q)} \leq \mathbf{C}_T \|v\|_{L^p(\mathbf{c}; \mathbb{R}^q)}^{1-\frac{1}{p}} \left(|v|_{W^{1,p}(\mathbf{c}; \mathbb{R}^q)} + dh_{\mathbf{c}}^{-1} \|v\|_{L^p(\mathbf{c}; \mathbb{R}^q)} \right)^{\frac{1}{p}}, \quad \forall v \in W^{1,p}(\mathbf{c}; \mathbb{R}^q). \quad (7.22)$$

We present here the proof inspired from Carstensen & Funken (2000) using the Raviart-Thomas-Nédélec shape functions.

Proof of Lemma 7.25. Let $\mathbf{c} \in \mathcal{C}$ and let $q = 1$ with $\mathbf{f} \subset \mathbf{f}$ for all $\mathbf{f} \in \mathcal{F}_{\mathbf{c}}$. The case $q \in \llbracket 2, d \rrbracket$ follows similarly.

First, let $\mathbf{c}$ be the (non-degenerate) pyramid of base $\mathbf{f}$ and of apex $\mathbf{x}_{\mathbf{c}}$. Let us denote $\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}$ the Raviart-Thomas-Nédélec shape function in this pyramid attached to $\mathbf{f}$:

$$\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}(\mathbf{x}) = \frac{(\mathbf{x} - \mathbf{x}_{\mathbf{c}})}{(\mathbf{x}_{\mathbf{f}} - \mathbf{x}_{\mathbf{c}}) \cdot \mathbf{n}_{\mathbf{f}}}, \quad \forall \mathbf{x} \in \mathbf{c},$$

where $\mathbf{x}_{\mathbf{f}}$ denotes any point belonging to $\mathbf{f}$. Let $\mathbf{n}_{\partial\mathbf{c}}$ be the outward unit normal on the boundary $\partial\mathbf{c}$ and observe that $\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \cdot \mathbf{n}_{\partial\mathbf{c}} \equiv 1$ on $\mathbf{f}$ (since this face is planar) and that $\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \cdot \mathbf{n}_{\partial\mathbf{c}} \equiv 0$ a.e. on $\partial\mathbf{c} \setminus \mathbf{f}$. Hence, for all $v \in W^{1,p}(\mathbf{c})$, we have

$$\|v\|_{L^p(\mathbf{f})}^p = \int_{\partial\mathbf{c}} |v|^p \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \cdot \mathbf{n}_{\partial\mathbf{c}} = \int_{\mathbf{c}} \nabla \cdot (|v|^p \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}),$$

owing to the divergence theorem. Then, applying twice the chain rule and Hölder's inequality, we obtain

$$\begin{aligned} \|v\|_{L^p(\mathbf{f})}^p &= \int_{\mathbf{c}} \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \cdot \nabla |v|^p + \int_{\mathbf{c}} |v|^p \nabla \cdot \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} = p \int_{\mathbf{c}} v |v|^{p-2} \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \cdot \nabla v + \int_{\mathbf{c}} |v|^p \nabla \cdot \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} \\ &\leq p \|\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}\|_{L^\infty(\mathbf{c})} \|v\|_{L^p(\mathbf{c})}^{\frac{p}{p'}} \|\nabla v\|_{L^p(\mathbf{c})} + \|\nabla \cdot \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}\|_{L^\infty(\mathbf{c})} \|v\|_{L^p(\mathbf{c})}^p. \end{aligned}$$

Observing that $\nabla \cdot \boldsymbol{\theta}_{\mathbf{f},\mathbf{c}} = \frac{dh_{\mathbf{c}}}{(\mathbf{x}_{\mathbf{f}} - \mathbf{x}_{\mathbf{c}}) \cdot \mathbf{n}_{\mathbf{f}}} h_{\mathbf{c}}^{-1}$, $\|\boldsymbol{\theta}_{\mathbf{f},\mathbf{c}}\|_{L^\infty(\mathbf{c})} \leq \frac{h_{\mathbf{c}}}{(\mathbf{x}_{\mathbf{f}} - \mathbf{x}_{\mathbf{c}}) \cdot \mathbf{n}_{\mathbf{f}}}$ and invoking mesh regularity, we infer that there exists $\mathbf{C}'_T > 0$ such that

$$\|v\|_{L^p(\mathbf{f})}^p \leq \mathbf{C}'_T \left(p \|v\|_{L^p(\mathbf{c})}^{\frac{p}{p'}} \|\nabla v\|_{L^p(\mathbf{c})} + dh_{\mathbf{c}}^{-1} \|v\|_{L^p(\mathbf{c})}^p \right).$$

The result then follows since $\mathbf{c} \subset \mathbf{c}$. $\square$

Proposition 7.26 (Inverse inequality). *Let $\mathbf{c} \in \mathcal{C}$, let $q \in \llbracket 1, d \rrbracket$, and let $\mathfrak{C}_{X,\mathbf{c}}$ be a partition of $\mathbf{c}$ in the sense of Definition 7.16. Let $\mathbf{c} \in \mathfrak{C}_{X,\mathbf{c}}$. Then, there exists $\mathbf{C}_{\text{INV}} > 0$ such that*

$$|v|_{W^{1,p}(\mathbf{c}; \mathbb{R}^q)} \leq \mathbf{C}_{\text{INV}} h_{\mathbf{c}}^{-1} \|v\|_{L^p(\mathbf{c}; \mathbb{R}^q)}, \quad \forall v \in \mathbb{P}_1(\mathfrak{C}_{X,\mathbf{c}}; \mathbb{R}^q). \quad (7.23)$$

Proof. Split the polyhedral cell $\mathfrak{c}$ into simplices and use the classical inverse inequality, e.g., from (Ern & Guermond, 2004, Sect. 1.7). $\square$

The last important inequality needed in the analysis (in particular in Chapter 5), is the estimate of the interpolation error of the Oswald of the average operator (see Oswald (1993)), usually considered in the analysis of finite volume method (see e.g., Achdou *et al.* (2003)) or to construct quasi-interpolation operator (see Ern & Guermond (2016)).

Proposition 7.27 (Oswald inequality). *Let $c \in \mathcal{C}$, let $p \in [1, \infty)$ and let $\mathfrak{C}_{X,c}$ be a partition of c in the sense of Definition 7.16. Then, there is $\mathcal{C}_{\text{avg}} > 0$ such that*

$$\left\| v - \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c} \in \mathfrak{C}_{X,c}} v|_{\mathfrak{c}} \right\|_{L^p(c)} \leq \mathcal{C}_{\text{avg}} h_c^{\frac{1}{p}} \left(\sum_{\mathfrak{f} \in \mathfrak{F}_{X,c}} \|\llbracket v \rrbracket\|_{L^p(\mathfrak{f})}^p \right)^{\frac{1}{p}}, \quad \forall v \in \mathbb{P}_0(\mathfrak{C}_{X,c}; \mathbb{R}). \quad (7.24)$$

Proof. Owing to Definition 7.16, we have

$$\left\| v - \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c}' \in \mathfrak{C}_{X,c}} v|_{\mathfrak{c}'} \right\|_{L^p(c)}^p = \sum_{\mathfrak{c} \in \mathfrak{C}_{X,c}} \left\| v - \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c}' \in \mathfrak{C}_{X,c}} v|_{\mathfrak{c}'} \right\|_{L^p(c)}^p = \sum_{\mathfrak{c} \in \mathfrak{C}_{X,c}} \left\| \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c}' \in \mathfrak{C}_{X,c}} (v|_{\mathfrak{c}} - v|_{\mathfrak{c}'}) \right\|_{L^p(c)}^p.$$

Hence, owing to Jensen formula, we infer that

$$\left\| v - \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c}' \in \mathfrak{C}_{X,c}} v|_{\mathfrak{c}'} \right\|_{L^p(c)}^p \leq \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c} \in \mathfrak{C}_{X,c}} \sum_{\mathfrak{c}' \in \mathfrak{C}_{X,c}} \|v|_{\mathfrak{c}} - v|_{\mathfrak{c}'}\|_{L^p(c)}^p. \quad (7.25)$$

Let us now consider a cyclic path $\psi : \mathfrak{C}_{X,c} \rightarrow \mathfrak{C}_{X,c}$, visiting at least once each sub-cell of $\mathfrak{C}_{X,c}$ and satisfying, for all $\mathfrak{c} \in \mathfrak{C}_{X,c}$, $\text{int}(\overline{\mathfrak{c}} \cap \psi(\mathfrak{c})) \subset \mathfrak{F}_{X,c}$, meaning that two consecutive sub-cells are $(d-1)$ -connected in $\mathfrak{F}_{X,c}$. Denoting $\eta \geq \#\mathfrak{C}_{X,c}$ the cycle index integer defined by $\psi^\eta \equiv \text{Id}$ which is uniformly bounded owing to mesh regularity, we observe that for all $\mathfrak{c} \in \mathfrak{C}_{X,c}$, $\overline{\mathfrak{c}} = \cup \{\overline{\psi^{(i)}(\mathfrak{c})} \mid i \in \llbracket 1, \eta \rrbracket\}$. Introducing now the length $l_{\mathfrak{c},\mathfrak{c}'} = \min\{j \in \llbracket 1, \eta \rrbracket \mid \psi^j(\mathfrak{c}) = \mathfrak{c}'\}$ separating $\mathfrak{c}$ and $\mathfrak{c}'$ along this path (in short l), we infer that

$$|v|_{\mathfrak{c}} - v|_{\mathfrak{c}'}|^p = \left| \sum_{k=1}^l v|_{\psi^{k-1}(\mathfrak{c})} - v|_{\psi^k(\mathfrak{c})} \right|^p \leq l^p \sum_{k=1}^l |v|_{\psi^{k-1}(\mathfrak{c})} - v|_{\psi^k(\mathfrak{c})}|^p.$$

Hence, observing that $l \leq \frac{\eta+1}{2}$ and that $|v|_{\psi^{k-1}(\mathfrak{c})} - v|_{\psi^k(\mathfrak{c})}|^p = |\mathfrak{f}_{k-1,k}|^{-1} \|\llbracket v \rrbracket\|_{L^p(\mathfrak{f}_{k-1,k})}^p$ with $\mathfrak{f}_{k-1,k} = \text{int}(\overline{\psi^{k-1}(\mathfrak{c})} \cap \overline{\psi^k(\mathfrak{c})})$, we infer, owing to mesh regularity, that

$$\|v|_{\mathfrak{c}} - v|_{\mathfrak{c}'}\|_{L^p(c)}^p \leq \left(\frac{\eta+1}{2} \right)^p |\mathfrak{c}| \sum_{\mathfrak{f} \in \mathfrak{F}_{X,c}} |\mathfrak{f}|^{-1} \|\llbracket v \rrbracket\|_{L^p(\mathfrak{f})}^p \lesssim \left(\frac{\eta+1}{2} \right)^p h_c \sum_{\mathfrak{f} \in \mathfrak{F}_{X,c}} \|\llbracket v \rrbracket\|_{L^p(\mathfrak{f})}^p.$$

Finally, collecting these bounds along with estimate (7.25) yields

$$\left\| v - \frac{1}{\#\mathfrak{C}_{X,c}} \sum_{\mathfrak{c} \in \mathfrak{C}_{X,c}} v|_{\mathfrak{c}} \right\|_{L^p(c)} \lesssim \frac{\eta+1}{2} (\#\mathfrak{C}_{X,c})^{\frac{1}{p}} h_c^{\frac{1}{p}} \left(\sum_{\mathfrak{f} \in \mathfrak{F}_{X,c}} \|\llbracket v \rrbracket\|_{L^p(\mathfrak{f})}^p \right)^{\frac{1}{p}}.$$

The expected result then follows since $\#\mathfrak{C}_{X,c}$ is bounded owing to mesh regularity and so is η as well. Note that the optimal situation corresponds to $\eta = \#\mathfrak{C}_{X,c}$, meaning that the path visit each sub-cell exactly once. In other words, it means that ψ defines a Hamiltonian path on the graph defined by the $(d-1)$ connectivity between the sub-cells of $\mathfrak{C}_{X,c}$. $\square$

7.4 Reconstruction and approximation

This section analyzes the reduction and the reconstruction maps used to derive and analyze our schemes. These notions were first introduced by the mimetic community, e.g., in the work of Hyman & Scovel (1990), Bossavit (1998, 2000), or Arnold *et al.* (2006). In this context, these maps aim to discretize differential forms as *co-chains*. Hereafter, reduction and reconstruction maps somehow extend to polyhedral meshes the classical notion of a finite element based on the triplet $\{K, P, \Sigma\}$ introduced by Ciarlet (1978).

Reduction maps $R_{\mathcal{X}}$ reduce a continuous field onto mesh entities of a mesh set X (corresponding hereafter to V , E or $V \times C$) in order to define the finite-dimensional space $\mathcal{X} \equiv \mathbb{R}^{\#X}$ collecting degrees of freedom attached to $x \in X$. Reconstruction maps $L_{\mathcal{X}}$ generate continuous fields from a dof vector in $\mathcal{X}$. Combining these two operators, we obtain the interpolation map $I_{\mathcal{X}} = L_{\mathcal{X}} \circ R_{\mathcal{X}}$. The interpolation error is then of order $m \in \mathbb{N}_{>0}$ if the reconstruction map $L_{\mathcal{X}}$ is stable and if the reduction map $R_{\mathcal{X}}$ is the right inverse of $L_{\mathcal{X}}$ on the space of polynomials of degree at most m . We can then consider the extended finite element formalized by the triplet $\{\mathcal{X}, L_{\mathcal{X}}, R_{\mathcal{X}}\}$.

The material of this section extends that presented by Bonelle (2014) since the analysis is now performed in Lebesgue spaces of exponent $p \in [1, \infty)$ and not just in the Hilbert setting. We also adopt a different point of view to emphasize the role of each of the two operators $R_{\mathcal{X}}$ and $L_{\mathcal{X}}$, especially regarding the regularity needed to obtain the convergence of the interpolation error as the mesh size goes to 0. Hereafter, we only consider low-order elements attached to mesh vertices, mesh edges and mesh vertices-cells; other elements are available in the literature, such as the face elements presented by Lemaire (2013) and by Di Pietro & Ern (2015), extending to polyhedral meshes the Crouzeix-Raviart elements.

7.4.1 Piece-wise constant vertex reconstruction

Let $c \in C$ and let $\mathcal{V}_c$ denote the finite-dimensional space collecting degrees of freedom attached to cell vertices $v \in V_c$. The reconstruction map on $\mathcal{V}_c$ is denoted by $L_{\mathcal{V}_c} : \mathcal{V}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{V,c}; \mathbb{R})$ and generates piece-wise constant polynomials on the partition $\mathfrak{C}_{V,c}$ introduced in Section 7.2.1. It is defined by

$$L_{\mathcal{V}_c}(v)(x) := \sum_{v \in V_c} v_v \ell_{v,c}(x), \quad \forall v \in \mathcal{V}_c, \quad \forall x \in c, \quad (7.26)$$

where for all $v \in V_c$, the piece-wise constant shape function $\ell_{v,c} \in \mathbb{P}_0(\mathfrak{C}_{V,c}; \mathbb{R})$ is given by

$$\ell_{v,c}|_{\mathfrak{c}_{v',c}} = \delta_{v,v'}, \quad \forall v' \in V_c, \quad (7.27)$$

with $\delta_{v,v'}$ the Kronecker symbol equal to 1 if $v \equiv v'$ and 0 otherwise.

Remark 7.28 (General definition). *From Bonelle et al. (2015), these shape functions are obtained choosing $\alpha = 1$ in the more general definition $\ell_{v,c}|_{\mathfrak{c}_{v',c}} = \alpha \delta_{v,v'} + \frac{|\mathfrak{c}_{v,c}|}{|c|}(1 - \alpha)$ for all $v, v' \in V_c$ where $\alpha > 0$ is a stabilization parameter.*

Stability. For all $p \in [1, \infty)$, the local discrete p -norm on $\mathcal{V}_c$ is defined by

$$\|v\|_{\mathcal{V}_c,p} := \left(\sum_{v \in V_c} |\mathfrak{c}_{v,c}| |v_v|^p \right)^{\frac{1}{p}}, \quad \forall v \in \mathcal{V}_c, \quad (7.28)$$

with $\mathfrak{c}_{v,c}$ defined by (7.4).

Proposition 7.29 (p -Stability). *Let $p \in [1, \infty)$. Then,*

$$\forall c \in C, \quad \|v\|_{\mathcal{V}_c,p} = \|L_{\mathcal{V}_c}(v)\|_{L^p(c)}, \quad \forall v \in \mathcal{V}_c. \quad (7.29)$$

Proof. Owing to definitions (7.27) and (7.28), it follows that

$$\|\mathbf{L}_{\mathcal{V}_c}(\mathbf{v})\|_{L^p(c)}^p = \sum_{\mathbf{v} \in \mathbf{V}_c} \int_{\mathfrak{c}_{\mathbf{v},c}} |\mathbf{L}_{\mathcal{V}_c}(\mathbf{v})|^p = \sum_{\mathbf{v} \in \mathbf{V}_c} \int_{\mathfrak{c}_{\mathbf{v},c}} |\mathbf{v}_{\mathbf{v}}|^p = \|\mathbf{v}\|_{\mathcal{V}_c,p}^p.$$

□

Consistency and interpolation error. Let us introduce the local reduction map $\mathbf{R}_{\mathcal{V}_c} : W^{s,p}(c) \rightarrow \mathcal{V}_c$ with $sp > d$ acting as follows:

$$\forall \mathbf{v} \in \mathbf{V}_c, \quad \mathbf{R}_{\mathcal{V}_c}(v)|_{\mathbf{v}} := v(\mathbf{x}_{\mathbf{v}}), \quad \forall v \in W^{s,p}(c). \quad (7.30)$$

This reduction map is referred as the de Rham map on 0-chains. The interpolation operator resulting from the combination of $\mathbf{L}_{\mathcal{V}_c}$ with $\mathbf{R}_{\mathcal{V}_c}$ is defined by $\mathbf{l}_{\mathcal{V}_c} = \mathbf{L}_{\mathcal{V}_c} \circ \mathbf{R}_{\mathcal{V}_c} : W^{s,p}(c) \rightarrow \mathbb{P}_0(\mathfrak{C}_{\mathcal{V}_c}; \mathbb{R})$ with $sp > d$.

Proposition 7.30 ($\mathbb{P}_0$ -consistency). *For all $c \in \mathcal{C}$, $\mathbf{l}_{\mathcal{V}_c}(U) = U$ for all $U \in \mathbb{P}_0(c; \mathbb{R})$.*

Proof. This property readily follows from definitions (7.30) and (7.26). □

Proposition 7.31 (Interpolation error). *Let $p \in (\frac{d}{2}, \infty)$. Then, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|v - \mathbf{l}_{\mathcal{V}_c}(v)\|_{L^p(c)} \leq \mathcal{C}_{\text{INT}} \left(h_c |v|_{W^{1,p}(c)} + h_c^2 |v|_{W^{2,p}(c)} \right), \quad \forall v \in W^{2,p}(c). \quad (7.31)$$

Proof. Let $c \in \mathcal{C}$ and let $v \in W^{2,p}(c)$. First, let us prove that there is $\mathcal{C}_R > 0$ such that

$$\|\mathbf{R}_{\mathcal{V}_c}(v)\|_{\mathcal{V}_c,p} \leq \mathcal{C}_R \left(\|v\|_{L^p(c)} + h_c |v|_{W^{1,p}(c)} + h_c^2 |v|_{W^{2,p}(c)} \right). \quad (7.32)$$

To do so, let us consider the local edge-face based partition $\mathfrak{C}_{\text{EF},c}$ defined in Section 7.2.2 which is composed of tetrahedra $\mathfrak{c} \in \mathfrak{C}_{\text{EF},c}$. Proceeding as for finite element proofs, we use the reference tetrahedron and the continuous embedding $W^{2,p}(\mathfrak{c}) \hookrightarrow L^\infty(\mathfrak{c})$ for all $p > \frac{d}{2}$ (see, e.g., (Ern & Guermond, 2004, §1.5)) to infer that

$$\|v\|_{L^\infty(\mathfrak{c})}^p \lesssim h_c^{-d} \left(\|v\|_{L^p(\mathfrak{c})}^p + h_c^p |v|_{W^{1,p}(\mathfrak{c})}^p + h_c^{2p} |v|_{W^{2,p}(\mathfrak{c})}^p \right). \quad (7.33)$$

The estimate (7.32) then follows since

$$\|\mathbf{R}_{\mathcal{V}_c}(v)\|_{\mathcal{V}_c,p}^p = \sum_{\mathbf{v} \in \mathbf{V}_c} |\mathfrak{c}_{\mathbf{v},c}| |v(\mathbf{x}_{\mathbf{v}})|^p \lesssim h_c^d \sum_{\mathfrak{c} \in \mathfrak{C}_{\text{EF},c}} 2 \|v\|_{L^\infty(\mathfrak{c})}^p,$$

owing to mesh regularity. Now, let $V \in \mathbb{P}_0(c; \mathbb{R})$. Observe that

$$\|v - \mathbf{l}_{\mathcal{V}_c}(v)\|_{L^p(c)} \leq \|v - V\|_{L^p(c)} + \|\mathbf{l}_{\mathcal{V}_c}(v - V)\|_{L^p(c)} \leq \|v - V\|_{L^p(c)} + \|\mathbf{R}_{\mathcal{V}_c}(v - V)\|_{\mathcal{V}_c,p},$$

using the triangle equality along with Propositions 7.29 and 7.30. Hence, combining this estimate with (7.32) yields

$$\|v - \mathbf{l}_{\mathcal{V}_c}(v)\|_{L^p(c)} \leq (1 + \mathcal{C}_R) \|v - V\|_{L^p(c)} + \mathcal{C}_R h_c |v - V|_{W^{1,p}(c)} + \mathcal{C}_R h_c^2 |v - V|_{W^{2,p}(c)}.$$

Then, observing that $V \in \mathbb{P}_0(c; \mathbb{R})$ yields $|v - V|_{W^{1,p}(c)} = |v|_{W^{1,p}(c)}$ and $|v - V|_{W^{2,p}(c)} = |v|_{W^{2,p}(c)}$. We infer that

$$\|v - \mathbf{l}_{\mathcal{V}_c}(v)\|_{L^p(c)} \lesssim \inf_{V \in \mathbb{P}_0(c; \mathbb{R})} \|v - V\|_{L^p(c)} + h_c |v|_{W^{1,p}(c)} + h_c^2 |v|_{W^{2,p}(c)}.$$

The conclusion then follows using the polynomial approximation bound (7.21) with $k = 0$. □

The interpolation error estimate (7.31) is not optimal since the regularity of v is over-constrained by the regularity of the domain of the de Rham reduction map (7.30). To circumvent this difficulty, we introduce the patch cell $\hat{c} = \cup\{\tilde{c}(v) \mid v \in V_c\}$ with the dual cell $\tilde{c}(v)$ defined by (7.16) for all $v \in V_c$ and the reduction map $\hat{R}_{V_c} : L^1(\hat{c}) \rightarrow \mathcal{V}_c$ acting as follows:

$$\forall v \in V_c, \quad \hat{R}_{V_c}(v)|_v := \frac{1}{|\tilde{c}(v)|} \left(\int_{\tilde{c}(v)} v \right), \quad \forall v \in L^1(\hat{c}). \quad (7.34)$$

The interpolation map obtained with this reduction map is now denoted by $\hat{I}_{V_c} = L_{V_c} \circ \hat{R}_{V_c} : L^1(\hat{c}) \rightarrow \mathbb{P}_0(\mathfrak{C}_{V,c}; \mathbb{R})$.

Proposition 7.32 ($\mathbb{P}_0$ -consistency). *For all $c \in C$, $\hat{I}_{V_c}(U) = U|_c$ for all $U \in \mathbb{P}_0(\hat{c}; \mathbb{R})$.*

Proof. This property readily follows from definitions (7.30) and (7.26). $\square$

Proposition 7.33 (Interpolation error). *Let $p \in [1, \infty)$. Then, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in C, \quad \|v - \hat{I}_{V_c}(v)\|_{L^p(c)} \leq \mathcal{C}_{\text{INT}} h_c |v|_{W^{1,p}(\hat{c})}, \quad \forall v \in W^{1,p}(\hat{c}). \quad (7.35)$$

Proof. Let $V \in \mathbb{P}_0(\hat{c}; \mathbb{R})$. Owing to the triangle inequality and Proposition 7.32, we infer that

$$\|v - \hat{I}_{V_c}(v)\|_{L^p(c)} \leq \|v - V\|_{L^p(c)} + \|\hat{I}_{V_c}(v - V)\|_{L^p(c)}.$$

Using the definition (7.30) of $\hat{R}_{V_c}$, we observe that for all $w \in L^p(\hat{c})$,

$$\|\hat{R}_{V_c}(w)\|_{\mathcal{V}_{c,p}}^p = \sum_{v \in V_c} |\mathfrak{c}_{v,c}| \left| \frac{1}{|\tilde{c}(v)|} \int_{\tilde{c}(v)} w \right|^p \leq \sum_{v \in V_c} \frac{|\mathfrak{c}_{v,c}|}{|\tilde{c}(v)|^p} \|w\|_{L^1(\tilde{c}(v))}^p \leq \sum_{v \in V_c} \frac{1}{|\tilde{c}(v)|^{p-1}} \|w\|_{L^1(\tilde{c}(v))}^p,$$

where we have used that $|\mathfrak{c}_{v,c}| \leq |\tilde{c}(v)|$ to infer the last inequality. Owing to Hölder's inequality, we also have $\|w\|_{L^1(\tilde{c}(v))}^p \leq \|w\|_{L^p(\tilde{c}(v))}^p \|1\|_{L^{p'}(\tilde{c}(v))}^p = \|w\|_{L^p(\tilde{c}(v))}^p |\tilde{c}(v)|^{p-1}$. Hence, recalling the definition of $\hat{c}$, we infer that

$$\|\hat{R}_{V_c}(w)\|_{\mathcal{V}_{c,p}} \leq \|w\|_{L^p(\hat{c})},$$

so that, using Proposition 7.29, it follows that

$$\|\hat{I}_{V_c}(v - V)\|_{L^p(c)} = \|\hat{R}_{V_c}(v - V)\|_{\mathcal{V}_{c,p}} \leq \|v - V\|_{L^p(\hat{c})}.$$

As a result, $\|v - \hat{I}_{V_c}(v)\|_{L^p(c)} \leq 2\|v - V\|_{L^p(\hat{c})}$ and the expected result is obtained with the estimate (7.21) with $r = 1$ by replacing c by $\hat{c}$ (observe that this estimate still holds up to a different numerical constant). $\square$

Remark 7.34 (Regularity). *The benefit of using $\hat{R}_{V_c}$ instead of R_{V_c} is that the regularity of v is now optimal. The price to pay is a slight loss of locality since the upper bound now involves $\hat{c}$ instead of c .*

7.4.2 Piece-wise affine vertex-cell reconstruction

Hereafter, an extension of the classical $\mathbb{P}_1$ -Lagrange finite element is proposed on polyhedral meshes. Let $c \in C$ and let $\mathcal{P}_c$ denote the finite-dimensional space collecting the scalar degrees of freedom attached $v \in V_c$ and to c . The reconstruction map $L_{\mathcal{P}_c} : \mathcal{P}_c \rightarrow \mathbb{P}_1(\mathfrak{C}_{\text{EF},c}, \mathbb{R}) \cap \mathcal{C}^0(c)$ generates a piece-wise affine continuous polynomial on the sub-mesh $\mathfrak{C}_{\text{EF},c}$ defined in Section 7.2.2. This operator is defined as

$$L_{\mathcal{P}_c}(v)(x) = v_c \ell_c(x) + \sum_{v \in V_c} v_v \ell_{v,c}(x), \quad \forall v \in \mathcal{P}_c, \quad \forall x \in c, \quad (7.36)$$

where the shape functions $((\ell_{v,c})_{v \in V_c}, \ell_c)$ span $\mathbb{P}_1(\mathfrak{C}_{\text{EF},c}, \mathbb{R}) \cap \mathcal{C}^0(c)$ and are defined by

$$\forall v \in V_c, \quad \ell_{v,c}(\mathbf{x}) := \theta_v(\mathbf{x}) + \sum_{f \in F_v} \frac{|\mathbf{f}_{v,f}|}{|f|} \theta_f(\mathbf{x}), \quad \text{and} \quad \ell_c(\mathbf{x}) := \theta_c(\mathbf{x}), \quad \forall \mathbf{x} \in c, \quad (7.37)$$

with $\mathbf{f}_{v,f}$ the inter-cell sub-face attached to v on f defined by (7.6), and with the classical Courant (or $\mathbb{P}_1$ Lagrange, or nodal, or hat) basis functions $((\theta_v)_{v \in V_c}, (\theta_f)_{f \in F_c}, \theta_c)$ on the simplicial sub-mesh $\mathfrak{C}_{\text{EF},c}$.

Stability. For all $p \in [1, \infty)$, the discrete p -norm on $\mathcal{P}_c$ is defined by

$$\|\mathbf{v}\|_{\mathcal{P}_c,p} := h_c^{\frac{d}{p}} \left(|\mathbf{v}_c|^p + \sum_{v \in V_c} |\mathbf{v}_v|^p \right)^{\frac{1}{p}}. \quad (7.38)$$

Similarly to the definition (7.28) of $\|\cdot\|_{\mathcal{V}_c,p}$ and owing to mesh regularity, an equivalent definition is $\|\mathbf{v}\|_{\mathcal{P}_c,p} := (|c| |\mathbf{v}_c|^p + \sum_{v \in V_c} |\mathbf{c}_{v,c}| |\mathbf{v}_v|^p)^{\frac{1}{p}}$.

Proposition 7.35 (p -Stability). *Let $p \in [1, \infty)$. Then, there exist $\mathcal{C}_b, \mathcal{C}_\sharp > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \mathcal{C}_b \|\mathbf{v}\|_{\mathcal{P}_c,p} \leq \|\mathbf{L}_{\mathcal{P}_c}(\mathbf{v})\|_{L^p(c)} \leq \mathcal{C}_\sharp \|\mathbf{v}\|_{\mathcal{P}_c,p}, \quad \forall \mathbf{v} \in \mathcal{P}_c. \quad (7.39)$$

Proof. Let $c \in \mathcal{C}$. Let us consider real numbers $(\alpha_v)_{v \in V_c}$, $(\alpha_f)_{f \in F_c}$ and α_c . Proceeding as in the proof of (Ern & Guermond, 2004, Lemma 9.17), we infer from the spectral properties of the mass matrix of $\mathbb{P}_1$ -Lagrange finite elements and from mesh regularity that there are uniform constants $0 < \mathcal{C}_1 \leq \mathcal{C}_2$, such that the function $f_\alpha = \sum_{v \in V_c} \alpha_v \theta_v + \sum_{f \in F_c} \alpha_f \theta_f + \alpha_c \theta_c$ satisfies

$$\mathcal{C}_1 |\alpha|_{\ell^p} \leq h_c^{-\frac{d}{p}} \|f_\alpha\|_{L^p(c)} \leq \mathcal{C}_2 |\alpha|_{\ell^p}, \quad (7.40)$$

where $|\alpha|_{\ell^p} = (\sum_{v \in V_c} |\alpha_v|^p + \sum_{f \in F_c} |\alpha_f|^p + |\alpha_c|^p)^{\frac{1}{p}}$. Considering now $\mathbf{v} \in \mathcal{P}_c$, we observe that

$$\mathbf{L}_{\mathcal{P}_c}(\mathbf{v}) = \sum_{v \in V_c} \mathbf{v}_v \theta_v + \sum_{f \in F_c} \left(\sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|f|} \mathbf{v}_v \right) \theta_f + \mathbf{v}_c \theta_c,$$

owing to the definition (7.27) of $\ell_{v,c}$. Applying now the lower bound of (7.40) to $f_\alpha = \mathbf{L}_{\mathcal{P}_c}(\mathbf{v})$, we obtain the expected lower bound with $\mathcal{C}_b = \mathcal{C}_1$. Now, turning to the upper bound, we observe that $\frac{|\mathbf{f}_{v,f}|}{|f|} < 1$ since $\mathbf{f}_{v,f} \subset f$, so that the discrete Hölder inequality yields

$$\sum_{f \in F_c} \left| \sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|f|} \mathbf{v}_v \right|^p \leq \sum_{f \in F_c} \left(\sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|^{p'}}{|f|^{p'}} \right)^{\frac{p}{p'}} \left(\sum_{v \in V_f} |\mathbf{v}_v|^p \right) \leq \sum_{f \in F_c} \left((\#V_f)^{\frac{p}{p'}} \sum_{v \in V_f} |\mathbf{v}_v|^p \right).$$

Next, exchanging the summations leads to

$$\sum_{f \in F_c} \left| \sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|f|} \mathbf{v}_v \right|^p \leq \sum_{v \in V_c} |\mathbf{v}_v|^p \left(\sum_{f \in F_v \cap F_c} (\#V_f)^{\frac{p}{p'}} \right),$$

so that the expected upper bound holds with

$$\mathcal{C}_\sharp = \mathcal{C}_2 \left(1 + \max_{v \in V_c} \left(\sum_{f \in F_v \cap F_c} (\#V_f)^{\frac{p}{p'}} \right)^{\frac{1}{p}} \right),$$

which is uniformly bounded owing to mesh regularity. $\square$

Consistency and interpolation error. Let us introduce the reduction map $R_{\mathcal{P}_c} : W^{s,p}(c) \rightarrow \mathcal{P}_c$ with $sp > d$ acting as follows:

$$\forall v \in V_c, \quad R_{\mathcal{P}_c}(v)|_c = v(\mathbf{x}_c) \quad \text{and} \quad R_{\mathcal{P}_c}(v)|_v = v(\mathbf{x}_v), \quad \forall v \in W^{s,p}(c) \quad (7.41)$$

and let us define the interpolation operator $l_{\mathcal{P}_c} = L_{\mathcal{P}_c} \circ R_{\mathcal{P}_c} : W^{s,p}(c) \rightarrow \mathbb{P}_1(\mathfrak{C}_{\text{EF},c}; \mathbb{R}) \cap \mathcal{C}^0(c)$ with $sp > d$.

Proposition 7.36 ($\mathbb{P}_1$ -consistency). *For all $c \in \mathcal{C}$, $l_{\mathcal{P}_c}(U) = U$ for all $U \in \mathbb{P}_1(c; \mathbb{R})$.*

Proof. Let $c \in \mathcal{C}$ and let $U \in \mathbb{P}_1(c; \mathbb{R})$ such that $U(\mathbf{x}) = \mathbf{a} \cdot \mathbf{x} + b$ with $(\mathbf{a}, b) \in \mathbb{R}^d \times \mathbb{R}$ for all $\mathbf{x} \in c$. By definition, we have

$$\begin{aligned} l_{\mathcal{P}_c}(U)(\mathbf{x}) &= \sum_{v \in V_c} (\mathbf{a} \cdot \mathbf{x}_v + b) \ell_{v,c}(\mathbf{x}) + (\mathbf{a} \cdot \mathbf{x}_c + b) \ell_c(\mathbf{x}) \\ &= \mathbf{a} \cdot \left(\sum_{v \in V_c} \mathbf{x}_v \ell_{v,c}(\mathbf{x}) + \mathbf{x}_c \ell_c(\mathbf{x}) \right) + \left(\sum_{v \in V_c} \ell_{v,c}(\mathbf{x}) + \ell_c(\mathbf{x}) \right) b. \end{aligned}$$

Using definition (7.37) of the local shape functions $\ell_{v,c}$ and ℓ_c , we observe that

$$\sum_{v \in V_c} \ell_{v,c}(\mathbf{x}) + \ell_c(\mathbf{x}) = \sum_{v \in V_c} \theta_v(\mathbf{x}) + \sum_{f \in F_c} \left(\sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|\mathbf{f}|} \right) \theta_f(\mathbf{x}) + \theta_c(\mathbf{x}).$$

Since $\sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|\mathbf{f}|} = 1$ owing to the star-shaped property, the partition of the unity satisfied by the Courant basis functions yields

$$\sum_{v \in V_c} \ell_{v,c}(\mathbf{x}) + \ell_c(\mathbf{x}) = \sum_{v \in V_c} \theta_v(\mathbf{x}) + \sum_{f \in F_c} \theta_f(\mathbf{x}) + \theta_c(\mathbf{x}) \equiv 1.$$

Similarly, recalling the identity $\sum_{v \in V_f} |\mathbf{f}_{v,f}| \mathbf{x}_v = |\mathbf{f}| \mathbf{x}_f$ from Proposition 7.14 and the linear exactness of the Courant basis functions, it follows that

$$\begin{aligned} \sum_{v \in V_c} \mathbf{x}_v \ell_{v,c}(\mathbf{x}) + \mathbf{x}_c \ell_c(\mathbf{x}) &= \sum_{v \in V_c} \mathbf{x}_v \theta_v(\mathbf{x}) + \sum_{f \in F_c} \left(\sum_{v \in V_f} \frac{|\mathbf{f}_{v,f}|}{|\mathbf{f}|} \mathbf{x}_v \right) \theta_f(\mathbf{x}) + \theta_c(\mathbf{x}) \\ &= \sum_{v \in V_c} \mathbf{x}_v \theta_v(\mathbf{x}) + \sum_{f \in F_c} \mathbf{x}_f \theta_f(\mathbf{x}) + \mathbf{x}_c \theta_c(\mathbf{x}) \equiv \mathbf{x}, \end{aligned}$$

whence the result. $\square$

Proposition 7.37 (Interpolation error). *Let $p \in (\frac{d}{2}, \infty)$. Then, there is $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|v - l_{\mathcal{P}_c}(v)\|_{L^p(c)} + h_c |v - l_{\mathcal{P}_c}(v)|_{W^{1,p}(c)} \leq \mathcal{C}_{\text{INT}} h_c^2 |v|_{W^{2,p}(c)}, \quad \forall v \in W^{2,p}(c). \quad (7.42)$$

Proof. We only sketch the proof since it follows the same ideas as those used to prove Proposition 7.31. Let $v \in W^{2,p}(c)$ and let $c \in \mathcal{C}$. There is $\mathcal{C}_R > 0$ such that

$$\|R_{\mathcal{P}_c}(v)\|_{\mathcal{P}_c,p} \leq \mathcal{C}_R \left(\|v\|_{L^p(c)} + h_c |v|_{W^{1,p}(c)} + h_c^2 |v|_{W^{2,p}(c)} \right). \quad (7.43)$$

Then, considering $V \in \mathbb{P}_1(c; \mathbb{R})$ we infer that

$$\|v - l_{\mathcal{P}_c}(v)\|_{L^p(c)} \leq \|v - V\|_{L^p(c)} + \|l_{\mathcal{P}_c}(v - V)\|_{L^p(c)} \leq \|v - V\|_{L^p(c)} + \mathcal{C}_{\#} \|R_{\mathcal{P}_c}(v - V)\|_{\mathcal{P}_c,p}$$

owing to Propositions 7.36 and 7.35. Then using (7.43), we obtain

$$\|v - l_{\mathcal{P}_c}(v)\|_{L^p(c)} \leq (1 + \mathcal{C}_{\#} \mathcal{C}_R) \|v - V\|_{L^p(c)} + \mathcal{C}_{\#} \mathcal{C}_R h_c |v - V|_{W^{1,p}(c)} + \mathcal{C}_{\#} \mathcal{C}_R h_c^2 |v - V|_{W^{2,p}(c)}.$$

Moreover, still using Proposition 7.36 along with triangle inequality, we have $|v - \mathbf{l}_{\mathcal{P}_c}(v)|_{W^{1,p}(c)} \leq |v - V|_{W^{1,p}(c)} + |\mathbf{l}_{\mathcal{P}_c}(v - V)|_{W^{1,p}(c)}$, so that owing to the inverse inequality (7.23) and the identity (7.43), we infer that

$$\begin{aligned} |v - \mathbf{l}_{\mathcal{P}_c}(v)|_{W^{1,p}(c)} &\leq |v - V|_{W^{1,p}(c)} + \mathbf{C}_{\text{INV}} h_c^{-1} \|\mathbf{l}_{\mathcal{P}_c}(v - V)\|_{L^p(c)} \\ &\leq |v - V|_{W^{1,p}(c)} + \mathbf{C}_{\text{INV}} h_c^{-1} \|v - V\|_{L^p(c)} + \mathbf{C}_{\text{INV}} h_c^{-1} \|v - \mathbf{l}_{\mathcal{P}_c}(v)\|_{L^p(c)} \end{aligned}$$

Finally, combining these two bounds, using that $|v - V|_{W^{2,p}(c)} = |v|_{W^{2,p}(c)}$ and that V is arbitrary in $\mathbb{P}_1(c; \mathbb{R})$ yields

$$\|v - \mathbf{l}_{\mathcal{P}_c}(v)\|_{L^p(c)} + h_c |v - \mathbf{l}_{\mathcal{P}_c}(v)|_{W^{1,p}(c)} \lesssim \inf_{P \in \mathbb{P}_1(c; \mathbb{R})} (\|v - P\|_{L^p(c)} + h_c |v - P|_{W^{1,p}(c)}) + h_c^2 |v|_{W^{2,p}(c)},$$

whence the result using the polynomial approximation bound (7.21) with $k = 1$. $\square$

7.4.3 Piece-wise constant edge reconstruction

Let $c \in \mathcal{C}$ and let $\mathcal{E}_c$ denote the finite-dimensional space collecting the scalar degrees of freedom attached $e \in E_c$. $\mathbf{L}_{\mathcal{E}_c}$ denotes the edge-based reconstruction map from $\mathcal{E}_c$, generating piece-wise constant polynomials on the sub-mesh $\mathfrak{C}_{E,c}$, defined in Section 7.2.3. $\mathbf{L}_{\mathcal{E}_c} : \mathcal{E}_c \rightarrow \mathbb{P}_0(\mathfrak{C}_{E,c}; \mathbb{R}^d)$ is such that

$$\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})(\mathbf{x}) := \sum_{e \in E_c} \mathbf{v}_e \ell_{e,c}(\mathbf{x}), \quad \forall \mathbf{v} \in \mathcal{E}_c, \quad \forall \mathbf{x} \in c, \quad (7.44)$$

where for all $e \in E_c$, the shape function $\ell_{e,c} \in \mathbb{P}_0(\mathfrak{C}_{E,c}; \mathbb{R}^d)$ is defined as

$$\ell_{e,c}|_{\mathfrak{c}_{e',c}} = \left(\text{Id} - \frac{\tilde{\mathbf{f}}_c(e') \otimes \mathbf{e}'}{d|\mathfrak{c}_{e',c}|} \right) \frac{\tilde{\mathbf{f}}_c(e)}{|c|} + \frac{\tilde{\mathbf{f}}_c(e)}{d|\mathfrak{c}_{e,c}|} \delta_{e,e'}, \quad \forall e' \in E_c, \quad (7.45)$$

where $\delta_{e,e'}$ denotes the Kronecker symbol equals to 1 if $e = e'$ and 0 otherwise, $\tilde{\mathbf{f}}_c(e) = \int_{\tilde{\mathfrak{f}}_c(e)} \mathbf{n}_{\tilde{\mathfrak{f}}_c(e)}$ and $\mathbf{e} = \int_e \mathbf{t}_e$. This reconstruction map corresponds to that proposed by Codecasa *et al.* (2010) in the context of the Discrete Geometric Approach (DGA).

Remark 7.38 (General definition). *From Bonelle et al. (2015), these basis functions are obtained choosing $\alpha = 1$ in*

$$\ell_{e,c}|_{\mathfrak{c}_{e',c}} = \left(\text{Id} - \alpha \frac{\tilde{\mathbf{f}}_c(e') \otimes \mathbf{e}'}{d|\mathfrak{c}_{e',c}|} \right) \frac{\tilde{\mathbf{f}}_c(e)}{|c|} + \alpha \frac{\tilde{\mathbf{f}}_c(e)}{d|\mathfrak{c}_{e,c}|} \delta_{e,e'}, \quad \forall e, e' \in E_c, \quad (7.46)$$

where $\alpha > 0$ is a free parameter related to the stabilization. Let us also mention that choosing $\alpha = \sqrt{d}$ yields the so-called *SUSHI-like* edge-based reconstruction map devised by Eymard et al. (2010).

Stability. For all $p \in [1, \infty)$, we define the discrete p -norm on the space $\mathcal{E}_c$ as

$$\|\mathbf{v}\|_{\mathcal{E}_c,p} := \left(\sum_{e \in E_c} |\mathfrak{c}_{e,c}| \left(\frac{|\mathbf{v}_e|}{|e|} \right)^p \right)^{\frac{1}{p}}, \quad \forall \mathbf{v} \in \mathcal{E}_c. \quad (7.47)$$

Proposition 7.39 (p -Stability). *Let $p \in [1, \infty)$. Then, there exists $\mathbf{C}_\sharp > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|\mathbf{v}\|_{\mathcal{E}_c,p} \leq \|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{L^p(c)} \leq \mathbf{C}_\sharp \|\mathbf{v}\|_{\mathcal{E}_c,p}, \quad \forall \mathbf{v} \in \mathcal{E}_c. \quad (7.48)$$

Proof. Owing to the definition (7.44) of $\mathbf{L}_{\mathcal{E}_c}$, we infer that

$$\|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)}^p = \sum_{\mathbf{e} \in \mathbf{E}_c} \|\mathbf{v}_{\mathbf{e}} \mathbf{a}_{\mathbf{e}} + \mathbf{b}_{\mathbf{e}}\|_{\mathbf{L}^p(\mathbf{c}_{\mathbf{e},c})}^p,$$

where

$$\mathbf{a}_{\mathbf{e}} = \frac{\mathbf{e}}{|\mathbf{e}|^2} \text{ and } \mathbf{b}_{\mathbf{e}} = \left(\ell_{\mathbf{e},c} - \frac{\mathbf{e}}{|\mathbf{e}|^2} \right) \mathbf{v}_{\mathbf{e}} + \sum_{\mathbf{e}' \in \mathbf{E}_c \setminus \{\mathbf{e}\}} \mathbf{v}_{\mathbf{e}'} \ell_{\mathbf{e}',c}.$$

Recalling the identity $d|\mathbf{c}_{\mathbf{e},c}| = |\mathbf{e} \cdot \tilde{\mathbf{f}}_c(\mathbf{e})|$ from Proposition 7.15, we observe that for all $\mathbf{e}, \mathbf{e}' \in \mathbf{E}_c$, we have $\ell_{\mathbf{e},c}(\mathbf{x}) \cdot \mathbf{e}' = \delta_{\mathbf{e},\mathbf{e}'}$ for all $\mathbf{x} \in \mathbf{c}_{\mathbf{e}',c}$. Hence, we have $\mathbf{a}_{\mathbf{e}} \cdot \mathbf{b}_{\mathbf{e}} \equiv 0$ on $\mathbf{c}_{\mathbf{e},c}$, so that $|\mathbf{v}_{\mathbf{e}} \mathbf{a}_{\mathbf{e}} + \mathbf{b}_{\mathbf{e}}| \geq |\mathbf{v}_{\mathbf{e}} \mathbf{a}_{\mathbf{e}}|$ and then

$$\|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)}^p \geq \sum_{\mathbf{e} \in \mathbf{E}_c} \|\mathbf{v}_{\mathbf{e}} \mathbf{a}_{\mathbf{e}}\|_{\mathbf{L}^p(\mathbf{c}_{\mathbf{e},c})}^p = \sum_{\mathbf{e} \in \mathbf{E}_c} |\mathbf{v}_{\mathbf{e}}|^p \|\mathbf{a}_{\mathbf{e}}\|_{\mathbf{L}^p(\mathbf{c}_{\mathbf{e},c})}^p.$$

Hence, the expected lower bound follows observing that $\|\mathbf{a}_{\mathbf{e}}\|_{\mathbf{L}^p(\mathbf{c}_{\mathbf{e},c})}^p = \frac{|\mathbf{c}_{\mathbf{e},c}|}{|\mathbf{e}|^p}$. Now, turning to the upper bound, the discrete Hölder inequality yields

$$\|\mathbf{L}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)}^p \leq (\#\mathbf{E}_c)^{p-1} \sum_{\mathbf{e} \in \mathbf{E}_c} |\mathbf{v}_{\mathbf{e}}|^p \|\ell_{\mathbf{e},c}\|_{\mathbf{L}^p(c)}^p.$$

Since $\|\ell_{\mathbf{e},c}\|_{\mathbf{L}^p(c)}^p \leq |c| \|\ell_{\mathbf{e},c}\|_{\mathbf{L}^\infty(c)}^p$, we infer that $\|\ell_{\mathbf{e},c}\|_{\mathbf{L}^p(c)}^p \leq \mathbf{C}_{\sharp}^p (\#\mathbf{E}_c)^{1-p} \frac{|\mathbf{c}_{\mathbf{e},c}|}{|\mathbf{e}|^p}$ where the constant

$$\mathbf{C}_{\sharp} = (\#\mathbf{E}_c)^{1-\frac{1}{p}} \max_{\mathbf{e} \in \mathbf{E}_c} \left(\left(\frac{|c|}{|\mathbf{c}_{\mathbf{e},c}|} \right)^{\frac{1}{p}} |\mathbf{e}| \|\ell_{\mathbf{e},c}\|_{\mathbf{L}^\infty(c)} \right),$$

is uniformly bounded owing to mesh regularity. The expected upper bound thus holds. Specifically, a calculation shows that

$$|\ell_{\mathbf{e},c}|_{\mathbf{c}_{\mathbf{e},c}}| \leq \frac{|\tilde{\mathbf{f}}_c(\mathbf{e})|}{|c|} \left(\frac{|c|}{d|\mathbf{c}_{\mathbf{e},c}|} \right) \quad \text{and} \quad |\ell_{\mathbf{e},c}|_{\mathbf{c}_{\mathbf{e}',c}}| \leq \frac{|\tilde{\mathbf{f}}_c(\mathbf{e})|}{|c|} \left(1 + \frac{1}{\cos^2(\mathbf{t}_{\mathbf{e}'}, \mathbf{n}_{\tilde{\mathbf{f}}_c(\mathbf{e}')})} \right)^{\frac{1}{2}},$$

leading to

$$|\mathbf{e}| \|\ell_{\mathbf{e},c}\|_{\mathbf{L}^\infty(c)} \leq \left(\frac{|\mathbf{e}| |\tilde{\mathbf{f}}_c(\mathbf{e})|}{|c|} \right) \max \left\{ \left(\frac{|c|}{d|\mathbf{c}_{\mathbf{e},c}|} \right), \max_{\mathbf{e}' \in \mathbf{E}_c, \mathbf{e}' \neq \mathbf{e}} \left(1 + \frac{1}{\cos^2(\mathbf{t}_{\mathbf{e}'}, \mathbf{n}_{\tilde{\mathbf{f}}_c(\mathbf{e}')})} \right)^{\frac{1}{2}} \right\},$$

where $(\mathbf{t}_{\mathbf{e}'}, \mathbf{n}_{\tilde{\mathbf{f}}_c(\mathbf{e}')})$ denotes the angle defined by these two vector. $\square$

Consistency and interpolation error. Let us now introduce the de Rham reduction map $\mathbf{R}_{\mathcal{E}_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathcal{E}_c$ with $sp > d - 1$ on mesh edges acting as

$$\forall \mathbf{e} \in \mathbf{E}_c, \quad \mathbf{R}_{\mathcal{E}_c}(\mathbf{v})|_{\mathbf{e}} := \frac{1}{|\mathbf{e}|} \int_{\mathbf{e}} \mathbf{v} \cdot \mathbf{e}, \quad \forall \mathbf{v} \in \mathbf{W}^{s,p}(c). \quad (7.49)$$

Define the interpolation map $\mathbf{l}_{\mathcal{E}_c} = \mathbf{L}_{\mathcal{E}_c} \circ \mathbf{R}_{\mathcal{E}_c} : \mathbf{W}^{s,p}(c) \rightarrow \mathbb{P}_0(\mathbf{c}_{\mathbf{E},c}; \mathbb{R}^d)$ with $sp > d - 1$.

Proposition 7.40 ($\mathbb{P}_0$ -consistency). *For all $c \in \mathbf{C}$, $\mathbf{l}_{\mathcal{E}_c}(\mathbf{U}) = \mathbf{U}$ for all $\mathbf{U} \in \mathbb{P}_0(c; \mathbb{R}^d)$.*

Proof. Let $\mathbf{U} \in \mathbb{P}_0(c; \mathbb{R}^d)$ and let $\mathbf{x} \in \mathbf{c}_{\mathbf{e}',c}$ with $\mathbf{e}' \in \mathbf{E}_c$. Owing to the definition (7.49) of $\mathbf{R}_{\mathcal{E}_c}$, it follows that

$$\mathbf{l}_{\mathcal{E}_c}(\mathbf{U})(\mathbf{x}) = \sum_{\mathbf{e} \in \mathbf{E}_c} \mathbf{R}_{\mathcal{E}_c}(\mathbf{U})|_{\mathbf{e}} \ell_{\mathbf{e},c}(\mathbf{x}) = \sum_{\mathbf{e} \in \mathbf{E}_c} (\mathbf{U} \cdot \mathbf{e}) \ell_{\mathbf{e},c}(\mathbf{x}) = \left(\sum_{\mathbf{e} \in \mathbf{E}_c} \ell_{\mathbf{e},c}(\mathbf{x}) \otimes \mathbf{e} \right) \mathbf{U}.$$

Then, owing to the identity $\sum_{\mathbf{e} \in \mathbf{E}_c} \mathbf{e} \otimes \tilde{\mathbf{f}}_c(\mathbf{e}) = |c| \mathbf{Id}$ from Proposition 7.21 and proceeding as in Bonelle (2014), we infer that $\sum_{\mathbf{e} \in \mathbf{E}_c} \ell_{\mathbf{e},c}(\mathbf{x}) \otimes \mathbf{e} = \mathbf{Id}$, whence the result. $\square$

In order to estimate the interpolation error of $\mathbf{l}_{\mathcal{E}_c}$, let us prove that the domain of the reduction map (7.49) can also be defined for all $p \in (\frac{3}{2}, 2]$ as $\{\mathbf{v} \in \mathbf{W}^{1,p}(c) \mid \nabla \times \mathbf{v} \in \mathbf{L}^{\frac{2p}{3-p}}(c)\}$ (note that $\frac{2p}{3-p} \in (2, 4]$ if $p \in (\frac{3}{2}, 2]$).

Proposition 7.41. *Let $p \in (\frac{3}{2}, 2]$ and let $c \in \mathcal{C}$. Then, for all $\mathbf{v} \in \mathbf{W}^{1,p}(c)$ such that $\nabla \times \mathbf{v} \in \mathbf{L}^{\frac{2p}{3-p}}(c)$, there exists $\mathcal{C}_R > 0$ such that*

$$\|\mathbf{R}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathcal{E}_c,p} \leq \mathcal{C}_R h_c \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right). \quad (7.50)$$

Proof. Let $c \in \mathcal{C}$ and let $e \in \mathcal{E}_c$. Let $f \in F_c \cap F_e$ and denote $\mathbf{c} \in \mathcal{C}_{\text{EF},c}$ the simplex associated with (e, f) (see definition (7.8)) and $\mathbf{f} = \partial \mathbf{c} \cap f$.

Let $q > 2$ and denote q' its conjugate number. We denote $\varphi_{e,\partial \mathbf{f}}$ the characteristic function of e on $\partial \mathbf{f}$. Since $qq' = (q-1)^{-1} < 1$ by assumption, the zero extension of $\varphi_{e,\partial \mathbf{f}}$ to $\partial \mathbf{f}$ satisfies (with the same notation) $\varphi_{e,\partial \mathbf{f}} \in W^{\frac{1}{q},q'}(\partial \mathbf{f})$. Let $\varphi_{e,\mathbf{f}}$ denote the lifting from $W^{\frac{1}{q},q'}(\partial \mathbf{f})$ to $W^{1,q'}(\mathbf{f})$. Next, extending $\varphi_{e,\mathbf{f}}$ to ∂c , we obtain $\varphi_{e,\partial c} \in W^{\frac{1}{q},q'}(\partial c)$ (by the same arguments) and we finally denote $\varphi_{e,\mathbf{c}}$ the lifting of $\varphi_{e,\partial c}$ from $W^{\frac{1}{q},q'}(\partial c)$ to $W^{1,q'}(\mathbf{c})$. Owing to mesh regularity (using the reference tetrahedron) and the surjectivity of trace maps, we infer that

$$|\varphi_{e,\mathbf{f}}|_{W^{1,q'}(\mathbf{f})} \lesssim h_c^{\frac{2-q'}{q'}} \quad \text{and} \quad |\varphi_{e,\mathbf{c}}|_{W^{1,q'}(\mathbf{c})} \lesssim h_c^{\frac{1}{q'}} h_c^{\frac{2-q'}{q'}}. \quad (7.51)$$

Denoting $T_e = |e|^{-1} \int_e \mathbf{v} \cdot \mathbf{e}$ and following Amrouche *et al.* (1998), we obtain

$$T_e = \int_{\mathbf{c}} (\nabla \times \mathbf{v}) \cdot \nabla \varphi_{e,\mathbf{c}} + \int_{\mathbf{f}} (\mathbf{v} \times \mathbf{n}_{\mathbf{f}}) \cdot \nabla_{\mathbf{f}} \varphi_{e,\mathbf{f}},$$

where $\nabla_{\mathbf{f}}$ denotes the tangential gradient on $\mathbf{f}$. Owing to Hölder's inequality, it follows that

$$|T_e| \leq \|\nabla \times \mathbf{v}\|_{\mathbf{L}^q(\mathbf{c})} |\varphi_{e,\mathbf{c}}|_{W^{1,q'}(\mathbf{c})} + \|\mathbf{v} \times \mathbf{n}_{\mathbf{f}}\|_{\mathbf{L}^q(mff)} |\varphi_{e,\mathbf{f}}|_{W^{1,q'}(\mathbf{f})},$$

so that, using the estimates (7.51), we infer that

$$|T_e| \lesssim h_c^{\frac{2-q'}{q'}} \left(\|\nabla \times \mathbf{v}\|_{\mathbf{L}^q(\mathbf{c})} h_c^{\frac{1}{q'}} + \|\mathbf{v} \times \mathbf{n}_{\mathbf{f}}\|_{\mathbf{L}^q(\mathbf{f})} \right).$$

Now, recalling the continuous trace embedding $\mathbf{W}^{1,p}(\mathbf{c}) \hookrightarrow \mathbf{L}^q(\mathbf{f})$ with $q = \frac{2p}{3-p}$ (see Adams (1975)) satisfying $q > 2$ for all $p \in (\frac{3}{2}, 2]$, we obtain

$$|T_e| \lesssim h_c^{\frac{2p-3}{p}} \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right),$$

where we have used the identities $\frac{1}{q'} = \frac{3(p-1)}{2p}$, $\frac{2-q'}{q'} = \frac{2p-3}{p}$, and the inclusion $\mathbf{c} \subset c$ owing to the star-shaped property. As a result, owing to the definition of $\mathbf{R}_{\mathcal{E}_c}$ and using mesh regularity, the expected result follows from the estimate

$$\|\mathbf{R}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathcal{E}_c,p} \lesssim \left(h_c^{3-p} \sum_{e \in \mathcal{E}_c} |T_e|^p \right)^{\frac{1}{p}} \lesssim h_c^{(\#\mathcal{E}_c)^{\frac{1}{p}}} \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right).$$

□

Combining Proposition 7.41 with the stability of $\mathbf{L}_{\mathcal{E}_c}$ from Proposition 7.39, we now estimate the interpolation error $\|\mathbf{v} - \mathbf{l}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)}$.

Proposition 7.42 (Interpolation error). *Let $p \in \left(\frac{3}{2}, 2\right]$ and let $c \in \mathcal{C}$. Then, for all $\mathbf{v} \in \mathbf{W}^{1,p}(c)$ such that $\nabla \times \mathbf{v} \in \mathbf{L}^{\frac{2p}{3-p}}(c)$, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\|\mathbf{v} - \mathbf{l}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)} \leq \mathcal{C}_{\text{INT}} h_c \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right). \quad (7.52)$$

Proof. Let $c \in \mathcal{C}$ and let $\mathbf{v}$ satisfy the assumptions of the proposition. Owing to Proposition 7.40 and the triangle inequality, we infer that $\|\mathbf{v} - \mathbf{l}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)} \leq \|\mathbf{v} - \mathbf{V}\|_{\mathbf{L}^p(c)} + \|\mathbf{l}_{\mathcal{E}_c}(\mathbf{V} - \mathbf{v})\|_{\mathbf{L}^p(c)}$ for all $\mathbf{V} \in \mathbb{P}_0(c; \mathbb{R}^d)$. Then, recalling the definition of $\mathbf{l}_{\mathcal{E}_c}$ and using successively Propositions 7.41 and 7.39, the second term on the right-hand side is bounded as follows:

$$\|\mathbf{l}_{\mathcal{E}_c}(\mathbf{V} - \mathbf{v})\|_{\mathbf{L}^p(c)} \leq \mathcal{C}_{\#} \|\mathbf{R}_{\mathcal{E}_c}(\mathbf{V} - \mathbf{v})\|_{\mathcal{E}_{c,p}} \leq \mathcal{C}_{\#} \mathcal{C}_{\mathbf{R}} h_c \left(h_c^{\frac{3(p-1)}{2p}} \|\nabla \times \mathbf{v}\|_{\mathbf{L}^{\frac{2p}{3-p}}(c)} + |\mathbf{v}|_{\mathbf{W}^{1,p}(c)} \right),$$

since $\mathbf{V}$ is piece-wise constant on c . Finally, the expected result follows from (7.21) with $r = 0$ and $q = 3$. $\square$

The above estimate holds only for smooth functions, and is false e.g., for functions in $\mathbf{H}^1(\Omega)$. Similarly to Section 7.4.1, solutions that are not smooth enough to be reduced by the de Rham map are handled using the reduction map $\widehat{\mathbf{R}}_{\mathcal{E}_c} : \mathbf{L}^1(\hat{c}) \rightarrow \mathcal{E}_c$ with the patch cell $\hat{c} = \cup \{\mathbf{c}_e; e \in E_c\}$ with $\mathbf{c}_e = \cup \{\mathbf{c}_{e,c}; c \in C_e\}$, acting as follows:

$$\forall e \in E_c, \quad \widehat{\mathbf{R}}_{\mathcal{E}_c}(\mathbf{v})|_e := \frac{1}{|\mathbf{c}_e|} \int_{\mathbf{c}_e} \mathbf{v} \cdot \mathbf{e}, \quad \forall \mathbf{v} \in \mathbf{L}^1(\hat{c}). \quad (7.53)$$

The interpolation map is now denoted by $\widehat{\mathbf{l}}_{\mathcal{E}_c} = \mathbf{l}_{\mathcal{E}_c} \circ \widehat{\mathbf{R}}_{\mathcal{E}_c} : \mathbf{L}^1(\hat{c}) \rightarrow \mathbb{P}_0(\mathcal{C}_{E,c}; \mathbb{R}^d)$.

Proposition 7.43 ($\mathbb{P}_0$ -consistency). *For all $c \in \mathcal{C}$, $\widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{U}) = \mathbf{U}|_c$ for all $\mathbf{U} \in \mathbb{P}_0(\hat{c}; \mathbb{R}^d)$.*

Proof. Similar to the proof of Proposition 7.40. $\square$

Proposition 7.44 (Interpolation errors). *Let $p \in [1, \infty)$. Then, there exists $\mathcal{C}_{\text{INT}} > 0$ such that*

$$\forall c \in \mathcal{C}, \quad \|\mathbf{v} - \widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)} \leq \mathcal{C}_{\text{INT}} h_c |\mathbf{v}|_{\mathbf{W}^{1,p}(\hat{c})}, \quad \forall \mathbf{v} \in \mathbf{W}^{1,p}(\hat{c}). \quad (7.54)$$

Proof. Let $c \in \mathcal{C}$ and let $\mathbf{v} \in \mathbf{W}^{1,p}(\hat{c})$ with $p \in [1, \infty)$. Owing to the triangle inequality and Proposition 7.43, we infer that

$$\|\mathbf{v} - \widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)} \leq \|\mathbf{v} - \mathbf{V}\|_{\mathbf{L}^p(c)} + \|\widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v} - \mathbf{V})\|_{\mathbf{L}^p(c)}$$

for $\mathbf{V} \in \mathbb{P}_0(\hat{c}; \mathbb{R}^d)$. In addition, we observe that, for all $\mathbf{w} \in \mathbf{L}^p(\hat{c})$,

$$\|\widehat{\mathbf{R}}_{\mathcal{E}_c}(\mathbf{w})\|_{\mathcal{E}_{c,p}}^p = \sum_{e \in E_c} \frac{|\mathbf{c}_{e,c}|}{|e|^p} \left| \frac{1}{|\mathbf{c}_e|} \int_{\mathbf{c}_e} \mathbf{w} \cdot \mathbf{e} \right|^p \leq \sum_{e \in E_c} \frac{|\mathbf{c}_{e,c}|}{|\mathbf{c}_e|^p} \|\mathbf{w}\|_{\mathbf{L}^1(\mathbf{c}_e)}^p \leq \sum_{e \in E_c} \frac{1}{|\mathbf{c}_e|^{p-1}} \|\mathbf{w}\|_{\mathbf{L}^1(\mathbf{c}_e)}^p,$$

where we have used that $|\mathbf{c}_{e,c}| \leq |\mathbf{c}_e|$ to infer the last inequality. Owing to Hölder inequality, it follows that $\|\mathbf{w}\|_{\mathbf{L}^1(\mathbf{c}_e)}^p \leq \|\mathbf{w}\|_{\mathbf{L}^p(\mathbf{c}_e)}^p \|1\|_{\mathbf{L}^{p'}(\mathbf{c}_e)}^p$ with $\|1\|_{\mathbf{L}^{p'}(\mathbf{c}_e)}^p = |\mathbf{c}_e|^{p-1}$, so that we obtain $\|\widehat{\mathbf{R}}_{\mathcal{E}_c}(\mathbf{w})\|_{\mathcal{E}_{c,p}}^p \leq \|\mathbf{w}\|_{\mathbf{L}^p(\hat{c})}^p$. Combining this estimate with the upper bound from Proposition 7.39, we obtain

$$\|\widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v} - \mathbf{V})\|_{\mathbf{L}^p(c)} \leq \mathcal{C}_{\#} \|\widehat{\mathbf{R}}_{\mathcal{E}_c}(\mathbf{v} - \mathbf{V})\|_{\mathcal{E}_{c,p}} \leq \mathcal{C}_{\#} \|\mathbf{v} - \mathbf{V}\|_{\mathbf{L}^p(c)}.$$

Hence,

$$\|\mathbf{v} - \widehat{\mathbf{l}}_{\mathcal{E}_c}(\mathbf{v})\|_{\mathbf{L}^p(c)} \leq (1 + \mathcal{C}_{\#}) \inf_{\mathbf{V} \in \mathbb{P}_0(\hat{c}; \mathbb{R}^d)} \|\mathbf{v} - \mathbf{V}\|_{\mathbf{L}^p(c)},$$

and the expected result follows from Corollary 7.24 with $k = 1$, $q = d$ and replacing c with $\hat{c}$ (still valid up to a different numerical constant). $\square$

Chapter 8

Conclusions et perspectives

Les travaux menés dans cette thèse ont permis d'étendre l'analyse au niveau continu des problèmes d'advection-réaction scalaire et vectoriel, de proposer de nouveaux schémas sur maillages polyédriques approchant la solution de ces problèmes et d'étendre leur analyse en présence de diffusion et/ou pour des tenseurs de Friedrichs à valeurs nulles ou négatives.

Dans un premier temps, nous avons étudié au niveau continu ces problèmes dans le cadre des systèmes de Friedrichs. Cette analyse a permis d'établir la bonne position des problèmes d'advection-réaction scalaire et vectoriel dans les espaces du graphe associés aux espaces de Banach L^p pour des exposants $p \in (1, 2]$ et pour des tenseurs de Friedrichs prenant des valeurs positives, nulles ou raisonnablement négatives. Cette analyse a également permis de prouver l'unicité de la solution faible dans le cas $p \in (1, \infty)$, là encore pour des tenseurs de Friedrichs positifs, nuls ou raisonnablement négatifs.

En parallèle de ces travaux, nous avons proposé dans cette thèse plusieurs nouveaux schémas approchant la solution des problèmes de transport scalaire et vectoriel en prenant inspiration des méthodes CDO, mais également des méthodes plus "classiques", comme la méthode des éléments finis stabilisés, la méthode de Galerkin discontinue ou la méthode des volumes finis.

Tout d'abord, nous avons proposé deux schémas permettant d'approcher la solution du problème d'advection-réaction scalaire utilisant des degrés de liberté associés aux sommets d'un maillage, avec une précision en $\mathcal{O}(h^{\frac{1}{2}})$ et $\mathcal{O}(h^{\frac{3}{2}})$, respectivement. Le premier schéma, assimilable à un schéma volumes finis sur les cellules duales attachées aux sommets, a permis d'identifier d'une part une nouvelle structure géométrique en introduisant la notion d'opérateurs de contraction discrets et d'autre part d'étendre l'analyse discrète dans le cas où le tenseur de Friedrichs prend des valeurs nulles. Le second schéma est nouveau et ouvre de notre point de vue de nouvelles perspectives quant à l'amélioration de la précision du schéma upwind sur maillages polyédriques.

Concernant le problème de diffusion-advection-réaction scalaire, nous avons proposé deux nouveaux schémas utilisant des degrés de liberté associés aux sommets du maillage. Le premier schéma traite de manière robuste le régime de l'écoulement en fonction du nombre de Péclet. Pour des tenseurs de Friedrichs à valeurs positives ou nulles, ce schéma est précis à l'ordre $\mathcal{O}(h^{\frac{1}{2}+s})$ avec $s = 0$ si les effets advectifs sont dominants et $s = \frac{1}{2}$ si les effets diffusifs sont dominants. Le second schéma permet quant à lui d'améliorer la précision du schéma à l'ordre $\mathcal{O}(h^{\frac{3}{2}})$ lorsque les effets advectifs sont dominants.

Enfin, nous avons également proposé un nouveau schéma pour le problème d'advection-réaction vectoriel. Ce schéma considère un seul degré de liberté scalaire associé à chaque arête du maillage et permet d'approcher la solution de ce problème à l'ordre $\mathcal{O}(h^{\frac{1}{2}})$ lorsque le tenseur de Friedrichs prend des valeurs positives, nulles ou raisonnablement négatives.

Rappelons que les briques élémentaires à partir desquelles ces schémas sont construits et analysés correspondent aux opérateurs de réduction et de reconstruction, dont nous avons étendu l'analyse dans le cadre des espaces de Banach. Ces opérateurs étendent aux maillages polyédriques la vision offerte par la méthode des éléments finis.

Pour finir, soulignons que tous les schémas mentionnés ci-dessus ont été implémentés dans un prototype de *Code_Saturne* et que les résultats numériques obtenus sur les différents maillages polyédriques tridimensionnels ont permis d’illustrer les résultats théoriques et de mesurer l’impact de la qualité du maillage sur différents cas tests.

Plusieurs développements complémentaires de nature mathématique et numérique ont été identifiés au cours de cette thèse. Tout d’abord, en ce qui concerne l’analyse proposée dans le Chapitre 2, il reste à établir la surjectivité de la trace des espaces du graphe $V_{\beta;p}(\Omega)$ dans les espaces $L^p(|\beta \cdot \mathbf{n}|; \partial\Omega)$ dans le cas général des espaces de Banach. Ce résultat est important puisqu’il permet d’étendre notre analyse pour des données au bord de Dirichlet non-homogènes.

Dans le Chapitre 3, nous avons identifié dans le cas scalaire deux opérateurs de contraction discrets. Dans le cas du problème vectoriel, ces opérateurs n’ont pu être identifiés de telle manière à préserver au niveau discret les relations de nilpotence que l’opérateur de contraction vérifie au niveau continu. Le parallèle avec les relations de nilpotence vérifiées par les opérateurs différentiels CDO (voir Proposition 3.11 de Bonelle (2014)) semble indiquer qu’il existe une granularité (ne correspondant par forcément à la dérivée de Lie) permettant d’exhiber une structure topologique des opérateurs de contraction discrets.

Dans le Chapitre 5, une extension du schéma (5.41) pourra être envisagée afin de traiter de manière robuste la transition entre la diffusion et l’advection. À partir du schéma proposé dans la Section 4.3, cette extension consiste selon nous à déterminer une pondération de la forme linéaire de stabilisation s_β en fonction du nombre de Péclet. Une seconde amélioration de ce schéma consiste à traiter symétriquement les conditions à la limite dans le but d’étendre le Théorème 4.1 de Bonelle & Ern (2014a) afin d’obtenir une estimation *a priori* de l’erreur en norme L^2 à l’ordre 2, correspondant à ce qui est observé numériquement. Enfin, une dernière amélioration que nous avons identifiée concerne l’analyse dans le cas où le tenseur de Friedrichs prend des valeurs nulles ou négatives. Observant que cette hypothèse peut être traitée dans le cadre des méthodes de Galerkin discontinues (voir Ayuso & Marini (2009)), cette situation nous semble envisageable dans le cadre des méthodes CIP s’il est possible de fournir une estimation *a priori* en norme L^2 de la solution à l’ordre $\frac{3}{2}$ sans contrôle de la dérivée advective.

Dans le Chapitre 6, le stencil du schéma que nous proposons pour le problème d’advection-réaction vectoriel est assez étendu, du fait de la pénalisation du saut à travers toutes les faces des sous-diamants. Plusieurs pistes sont envisageables afin de réduire ce stencil, notamment en considérant d’autres fonctions de reconstruction considérant moins de degrés de liberté scalaires associés aux arêtes du maillage. Il est par exemple possible de proposer une reconstruction n’utilisant que l’information portée par trois arêtes non-coplanaires afin de reconstruire localement des fonctions constantes par morceaux à valeurs dans $\mathbb{R}^3$. Cependant, dans le cas des maillages non-conformes, cette approche n’est à l’heure actuelle pas très claire du fait de la présence d’hyperplans contenant plusieurs faces. Une autre alternative permettant de réduire le stencil, inspirée du schéma proposé dans le Chapitre 5, consiste à introduire des degrés de liberté arêtes supplémentaires, qui soient condensables statiquement, dans le but de ne pas pénaliser le saut de la solution entre les cellules du maillage. La difficulté réside alors dans l’analyse du problème discret, dont la stabilité repose dorénavant sur une condition inf-sup utilisant les bulles associées aux arêtes portant ces degrés de liberté supplémentaires.

À partir de l’analyse proposée dans le Chapitre 2, il semble également intéressant d’étendre l’analyse du problème de Oseen proposée par Amrouche & Razafison (2007) et Amrouche *et al.* (2011) permettant de prouver l’existence de solutions faibles et très faibles pour des nombres de Reynolds élevés. En ce qui concerne l’approximation du problème de Oseen utilisant des degrés de liberté associés aux sommets (pour la pression) et aux arêtes du maillage (pour la vitesse), la réduction du stencil du schéma proposé dans le Chapitre 6 est à notre avis importante avant de le coupler avec le schéma CDO proposé par Bonelle & Ern (2014b) pour le problème de Stokes. Notons qu’une approche pragmatique permettant d’approcher la solution du problème de Oseen consiste à utiliser composante par composante le schéma proposé dans

le Chapitre 5 en adaptant le traitement faible des conditions limites, notamment à partir des travaux de Burman *et al.* (2006).

Finalement, signalons que des travaux sont actuellement en cours concernant une approximation du problème d'advection-réaction scalaire instationnaire préservant le principe du maximum au niveau discret sur maillages polyédriques. À partir des travaux Burman & Ern (2005), l'approche que nous avons retenu consiste à introduire un terme de diffusion non-linéaire artificielle dans les zones où la solution varie fortement.

Appendix A

Additional numerical results on Kershaw mesh sequences

In this appendix, we collect the numerical results for the test cases 1 (see Section 3.3.2), 2 (see Section 3.3.3) and 3 (see Section 4.4.1) on the **Kershaw** mesh sequence. This sequence is composed of distorted hexahedral meshes, as observed in Figure A.1.

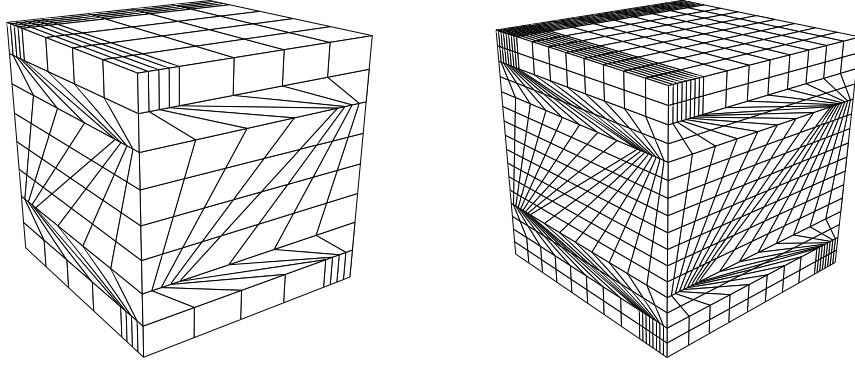


Figure A.1 – Two elements of the **Kershaw** mesh sequence.

This sequence is treated separately from the others since it only illustrates the robustness of our schemes with respect to mesh quality criteria. Indeed, this mesh sequence is not really suited to study the convergence of a scheme since it does not exactly satisfy the mesh regularity assumptions M (see Section 7.1.3) and it does not have constant mesh regularity criteria. In particular, we observe that the aspect ratio of the barycentric subdivision (see definition (A.2) in Bonelle (2014)) decreases as the mesh size goes to 0.

Hereafter, we collect the numerical results obtained on this mesh sequence, along with the numerical results obtained on **H** and **CB** mesh sequences for comparison purposes. We use the same labels as those defined in the previous chapters for these latter sequences, and we use the labels $\textcolor{red}{\diamond}$ and $\textcolor{red}{-\diamond-}$ for the **Kershaw** mesh sequence.

Test case 1. Rotating advective field

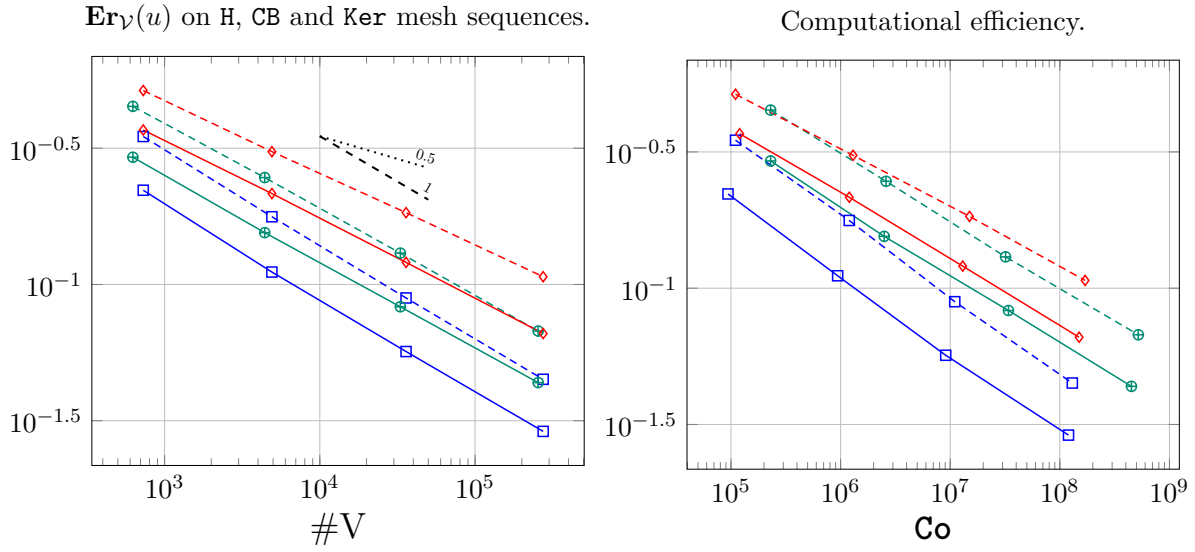


Figure A.2 – Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost $\mathbf{Co}$ when $\mu = 5$ (plain lines) and $\mu = 0.5$ (dashed lines) for the scheme (3.21).

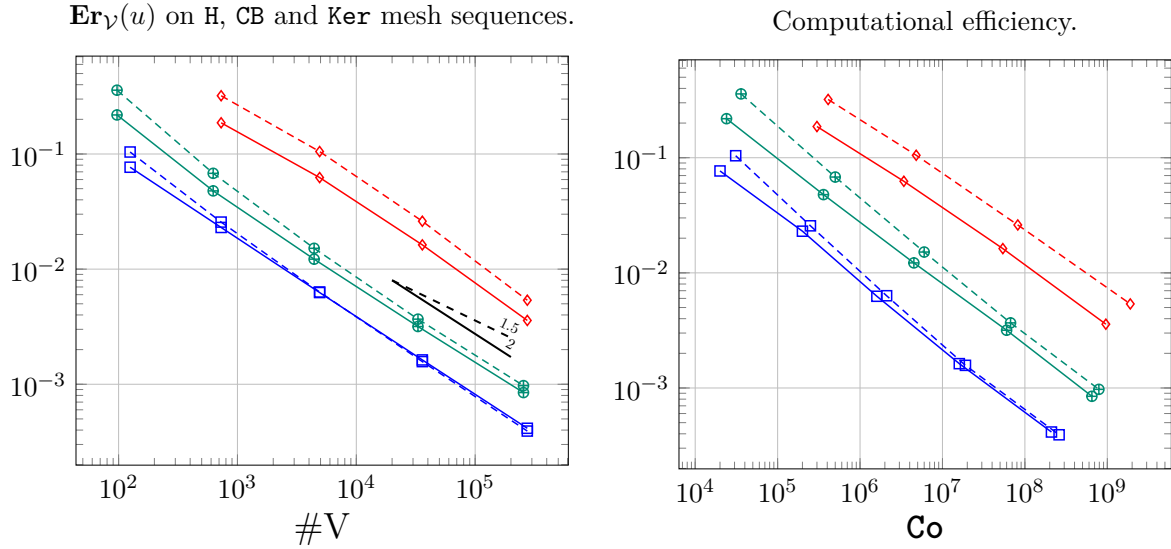


Figure A.3 – Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost $\mathbf{Co}$ when $\mu = 5$ (plain lines) and $\mu = 0.5$ (dashed lines) for the scheme (5.16).

(3.21), $\mu = 5$		(3.21), $\mu = 0.5$		(5.16), $\mu = 5$		(5.16), $\mu = 0.5$	
Min	Max	Min	Max	Min	Max	Min	Max
Y	Y	Y	Y	2.1%	6.9%	Y	20.8%
Y	Y	Y	Y	7.9%	7.6%	18.4%	13.3%
Y	Y	Y	Y	1.8%	1.3%	4.9%	2.1%
Y	Y	Y	Y	0.2%	0.2%	0.3%	0.3%

Table A.1 – Discrete minimum and maximum principle on the *Kershaw* mesh sequence for the schemes (3.21) and (5.16) with $\mu = 0.5$ and $\mu = 5$.

Test case 2. Sharp internal layer

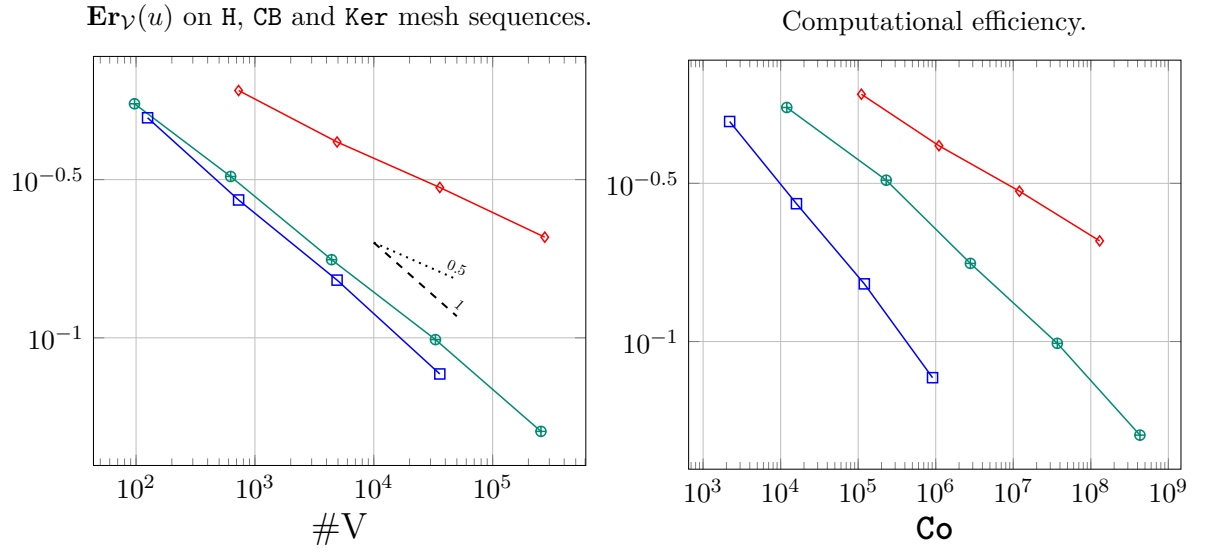


Figure A.4 – Numerical error $\text{Er}_V(u)$ with respect to $\#V$ and the computational cost Co for the scheme (3.21).

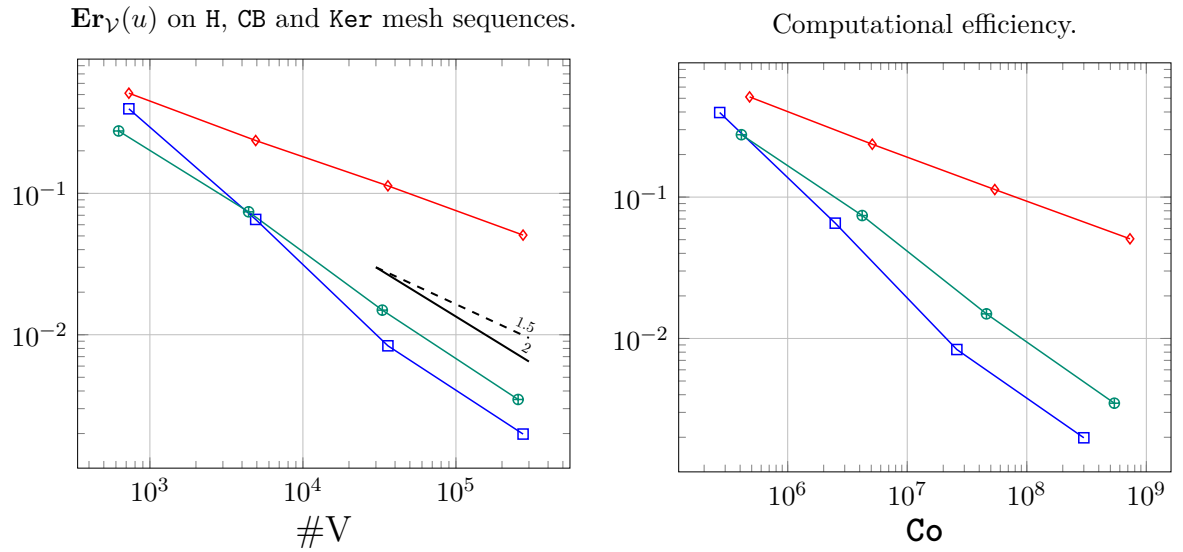


Figure A.5 – Numerical error $\text{Er}_V(u)$ with respect to $\#V$ and the computational cost Co for the scheme (5.16).

(3.21)		(5.16)	
Min	Max	Min	Max
Y	21.8%	35.3%	120.5%
Y	26.0%	15.3%	47.6%
Y	32.8%	0.8%	26.6%
Y	30.6%	Y	10.5%

Table A.2 – Discrete minimum and maximum principle on the *Kershaw* mesh sequence for the schemes (3.21) and (5.16).

Test case 3. Anisotropic diffusion tensor and rotating advective field

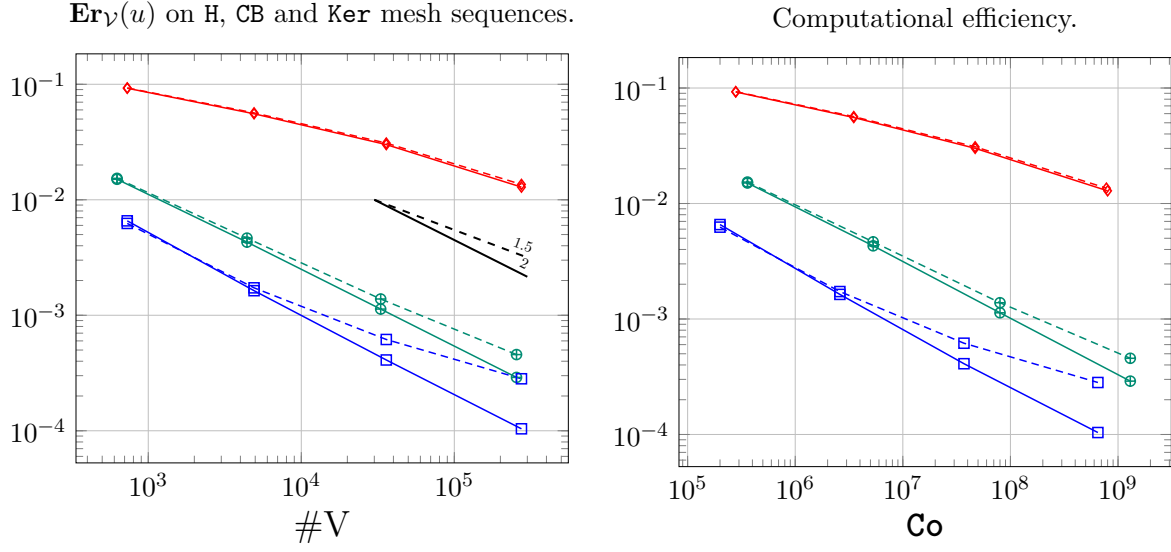


Figure A.6 – Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost $\mathbf{Co}$ for the scheme (4.34) using full-upwinding (dashed lines) and Sharfetter-Gummel upwinding (plain lines).

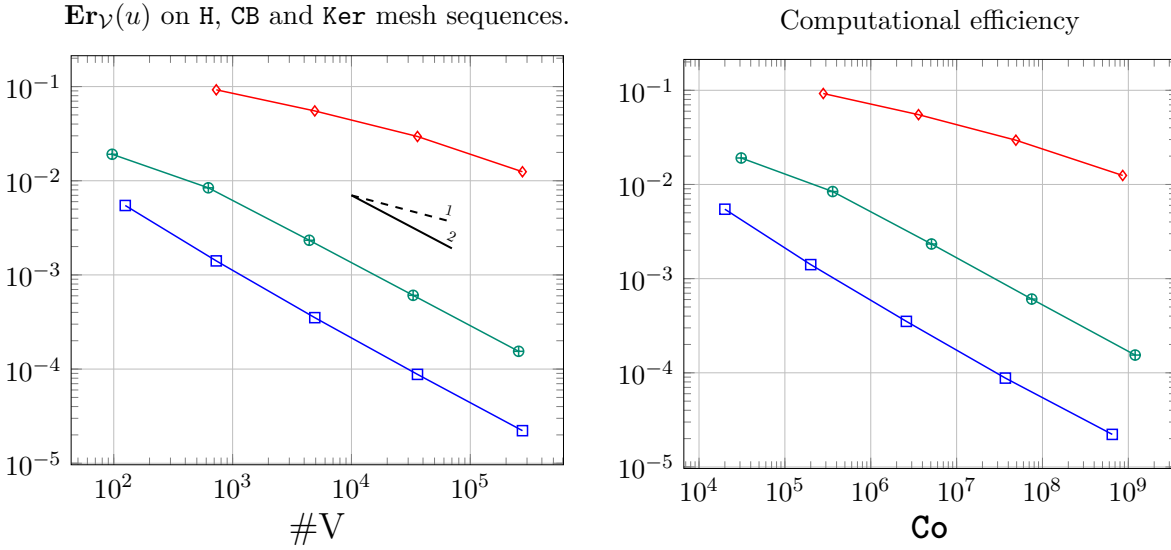


Figure A.7 – Test case 3. Numerical error $\mathbf{Er}_V(u)$ with respect to $\#V$ and the computational cost $\mathbf{Co}$ for the discrete problems (5.41).

Bibliography

- ABRAHAM, R., MARSDEN, J. E. & RATIU, T. (1988) *Manifolds, tensor analysis, and applications*. Applied Mathematical Sciences, vol. 75, second edn. Springer-Verlag, New York, pp. x+654.
- ACHDOU, Y., BERNARDI, C. & COQUEL, F. (2003) A priori and a posteriori analysis of finite volume discretizations of Darcy's equations. *Numer. Math.*, **96**, 17–42.
- ACOSTA, G. & DURAN, R. G. (2004) An optimal Poincaré inequality in L^1 for convex domains. *Proceedings of the American Mathematical Society*, **132**, 195–202.
- ADAMS, R. A. (1975) *Sobolev Spaces*. Pure and Applied Mathematics Press, vol. 65. Academic Press.
- AMROUCHE, C., BERNARDI, C., DAUGE, M. & GIRAULT, V. (1998) Vector potentials in three-dimensional non-smooth domains. *Math. Meth. Appl. Sci.*, **21**, 823–864.
- AMROUCHE, C., & RODRÍGUEZ-BELLIDO, M. A. (2011) Stationary Stokes, Oseen and Navier–Stokes equations with singular data. *Archive for Rational Mechanics and Analysis*, **199**, 597–651.
- AMROUCHE, C. & RAZAFISON, U. (2007) The stationary Oseen equations in $\mathbb{R}^3$. An approach in weighted Sobolev spaces. *Journal of Mathematical Fluid Mechanics*, **9**, 211–225.
- ANDREIANOV, B., BENDAHMANE, M., HUBERT, F. & KRELL, S. (2012) On 3D DDFV discretization of gradient and divergence operators. I. Meshing, operators and discrete duality. *IMA Journal of Numerical Analysis*, **32**, 1574–1603.
- ANGOT, P., DOLEJŠÍ, V., FEISTAUER, M. & FELCMAN, J. (1998) Analysis of a combined barycentric finite volume-nonconforming finite element method for nonlinear convection-diffusion problems. *Appl. Math.*, **43**, 263–310.
- APEL, T. & LUBE, G. (1998) Anisotropic mesh refinement for a singularly perturbed reaction diffusion model problem. *Applied Numerical Mathematics*, **26**, 415 – 433.
- ARCHAMBEAU, F., MECHITOUA, N. & SAKIZ, M. (2004) A finite volume code for the computation of turbulent incompressible flows - industrial applications. *International Journal on Finite Volumes*.
- ARNOLD, D. N., FALK, R. S. & WINTHER, R. (2006) Finite element exterior calculus, homological techniques, and applications. *Acta Numer.*, **15**, 1–155.
- AYUSO, B. & MARINI, L. D. (2009) Discontinuous Galerkin methods for advection-diffusion-reaction problems. *SIAM J. Numer. Anal.*, **47**, 1391–1420.
- AZÉRAD, P. & POUSIN, J. (1996) Inégalité de Poincaré courbe pour le traitement variationnel de l'équation de transport. *C. R. Acad. Sci. Paris Sér. I Math.*, **322**, 721–727.
- BABA, K. & TABATA, M. (1981) On a conservative upwind finite element scheme for convective diffusion equations. *RAIRO Anal. Numér.*, **15**, 3–25.
- BACK, A. & SONNENDRÜCKER, E. (2012) Spline discrete differential forms. *ESAIM: Proc.*, **35**, 197–202.
- BACK, A. & SONNENDRÜCKER, E. (2014) Finite element Hodge for spline discrete differential forms. application to the Vlasov–Poisson system. *Applied Numerical Mathematics*, **79**, 124 – 136. Workshop on Numerical Electromagnetics and Industrial Applications (NELIA 2011).

- BALAY, S., ABHYANKAR, S., ADAMS, M.-F., BROWN, J., BRUNE, P., BUSCHELMAN, K., DALCIN, L., EIJKHOUT, V., GROPP, W.-D., KAUSHIK, D., KNEPLEY, M.-G., CURFMAN MCINNES, L., RUPP, K., SMITH, B.-F., ZAMPINI, S., ZHANG, H. & ZHANG, H. (2016) PETSc users manual. *Technical Report* ANL-95/11 - Revision 3.7. Argonne National Laboratory.
- BARDOS, C. (1970) Problèmes aux limites pour les équations aux dérivées partielles du premier ordre à coefficients réels; théorèmes d'approximation; application à l'équation de transport. *Annales scientifiques de l'École Normale Supérieure*, **3**, 185–233.
- BAZILEVS, Y. & HUGHES, T. (2007) Weak imposition of dirichlet boundary conditions in fluid mechanics. *Computers & Fluids*, **36**, 12 – 26. Challenges and Advances in Flow Simulation and Modeling.
- BEBENDORF, M. (2003) A note on the Poincaré inequality for convex domains. *Z. Anal. Anwendungen*, **22**, 751–756.
- BECKER, R. & BRAACK, M. (2001) A finite element pressure gradient stabilization for the Stokes equations based on local projections. *CALCOLO*, **38**, 173–199.
- BEIRÃO DA VEIGA, H. (1988) Boundary-value problems for a class of first order partial differential equations in Sobolev spaces and applications to the Euler flow. *Rendiconti del Seminario Matematico della Università di Padova*, **79**, 247–273.
- BEIRÃO DA VEIGA, L., DRONIOU, J. & MANZINI, G. (2011) A unified approach for handling convection terms in finite volumes and mimetic discretization methods for elliptic problems. *IMA Journal of Numerical Analysis*, **31**, 1357–1401.
- BEIRÃO DA VEIGA, L., BREZZI, F., CANGIANI, A., MANZINI, G., MARINI, L. D. & RUSSO, A. (2013) Basic principles of virtual element methods. *Math. Models Methods Appl. Sci.*, **23**, 199–214.
- BEIRÃO DA VEIGA, L., BREZZI, F., MARINI, L. D. & RUSSO, A. (2016a) Mixed virtual element methods for general second order elliptic problems on polygonal meshes. *ESAIM: M2AN*, **50**, 727–747.
- BEIRÃO DA VEIGA, L., BREZZI, F., MARINI, L. D. & RUSSO, A. (2016b) Virtual element method for general second-order elliptic problems on polygonal meshes. *Mathematical Models and Methods in Applied Sciences*, **26**, 729–750.
- BONELLE, J. (2014) Compatible discrete operator schemes on polyhedral meshes for elliptic and stokes equations. *Ph.D. thesis*, Université Paris Est.
- BONELLE, J., DI PIETRO, D. & ERN, A. (2015) Low-order reconstruction operators on polyhedral meshes: application to compatible discrete operator schemes. *Comput. Aided Geom. Design*, **35/36**, 27–41.
- BONELLE, J. & ERN, A. (2014a) Analysis of compatible discrete operator schemes for elliptic problems on polyhedral meshes. *ESAIM Math. Model. Numer. Anal.*, **48**, 553–581.
- BONELLE, J. & ERN, A. (2014b) Analysis of compatible discrete operator schemes for Stokes problems on polyhedral meshes. *IMA J. Numer. Anal.*, **34**, 553–581.
- BOSSAVIT, A. (1998) *Computational electromagnetism*. Electromagnetism. Academic Press, Inc., San Diego, CA, pp. xx+352. Variational formulations, complementarity, edge elements.
- BOSSAVIT, A. (1999-2000) Computational electromagnetism and geometry. *J. Japan Soc. Appl. Electromagn. & Mech.*, **7-8**, 150–9 (no 1), 294–301 (no 2), 401–8 (no 3), 102–9 (no 4), 203–9 (no 5), 372–7 (no 6).
- BOSSAVIT, A. (2003) Extrusion, contraction: their discretization via Whitney forms. *COMPEL*, **22**, 470–480. Selected papers from the 10th International IGTE Symposium on Numerical Field Computation (Graz, 2002).
- BRAACK, M., BURMAN, E., JOHN, V. & LUBE, G. (2007) Stabilized finite element methods for the generalized Oseen problem. *Comput. Methods Appl. Mech. Engrg.*, **196**, 853–866.

- BRENNER, S. C. & SCOTT, L. R. (1994) *The mathematical theory of finite element methods*. Texts in Applied Mathematics, vol. 15. Springer-Verlag, New York, pp. xii+294.
- BREZIS, H. (2010) *Functional analysis, Sobolev spaces and partial differential equations*. Springer.
- BURMAN, E. (2005) A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty. *SIAM J. Numer. Anal.*, **43**, 2012–2033.
- BURMAN, E., FERNÁNDEZ, M. A. & HANSBO, P. (2006) Continuous interior penalty finite element method for Oseen’s equations. *SIAM Journal on Numerical Analysis*, **44**, 1248–1274.
- BURMAN, E. (2012) A penalty-free nonsymmetric nitsche-type method for the weak imposition of boundary conditions. *SIAM Journal on Numerical Analysis*, **50**, 1959–1981.
- BURMAN, E. (2014) Stabilized finite element methods for nonsymmetric, noncoercive, and ill-posed problems. part II: Hyperbolic equations. *SIAM Journal on Scientific Computing*, **36**, A1911–A1936.
- BURMAN, E. & ERN, A. (2005) Stabilized Galerkin approximation of convection-diffusion-reaction equations: Discrete maximum principle and convergence. *Math. Comput.*, **74**, 1637–1652.
- BURMAN, E. & ERN, A. (2007) A continuous finite element method with face penalty to approximate Friedrichs’ systems. *ESAIM: Mathematical Modelling and Numerical Analysis*, **41**, 55–76.
- BURMAN, E. & HANSBO, P. (2004) Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, **193**, 1437–1453.
- BURMAN, E. & SCHIEWECK, F. (2016) Local CIP stabilization for composite finite elements. *SIAM Journal on Numerical Analysis*, **54**, 1967–1992.
- CAMPOS PINTO, M., JUND, S., SALMON, S. & SONNENDRÜCKER, S. (2014) Charge-conserving FEM–PIC schemes on general grids . *Comptes Rendus Mécanique*, **342**, 570 – 582. Theoretical and numerical approaches for Vlasov-maxwell equations.
- CANTIN, P., BONELLE, J., BURMAN, E. & ERN, A. (2016) A vertex-based scheme on polyhedral meshes for advection–reaction equations with sub-mesh stabilization. *Computers & Mathematics with Applications*, **72**, 2057 – 2071.
- CANTIN, P. & ERN, A. (2016a) Edge-based low-order schemes on polyhedral meshes for vector advection-reaction equations.
- CANTIN, P. & ERN, A. (2016b) Vertex-based compatible discrete operator schemes on polyhedral meshes for advection-diffusion equations. *Comput. Meth. Appl. Math.*, **16**, 187–212.
- CARSTENSEN, C. & FUNKEN, S. A. (2000) Constants in Clément-interpolation error and residual based a posteriori estimates in finite element methods. *East-West J. Numer. Math.*, **8**, 153–175.
- CHRISTIANSEN, S. H. (2008) A construction of spaces of compatible differential forms on cellular complexes. *Math. Models Methods Appl. Sci.*, **18**, 739–757.
- CIARLET, P. G. (1978) *The Finite Element Method for Elliptic Problems*. SIAM.
- COCKBURN, B., GOPALAKRISHNAN, J. & LAZAROV, R. (2009) Unified Hybridization of Discontinuous Galerkin, Mixed, and Continuous Galerkin Methods for Second Order Elliptic Problems. *SIAM Journal on Numerical Analysis*, **47**, 1319–1365.
- CODECASA, L., SPECOGNA, R. & TREVISAN, F. (2010) A new set of basis functions for the discrete geometric approach. *J. Comput. Phys.*, **229**, 7401–7410.
- CODECASA, L. & TREVISAN, F. (2007) Constitutive equations for discrete electromagnetic problems over polyhedral grids. *J. Comput. Phys.*, **225**, 1894–1918.

- DARROZES, J. & FRANÇOIS, C. (1982) *Mécanique des fluides incompressibles*. Springer.
- DEURING, P., EYMARD, R. & MILDNER, M. (2015) L^2 -stability independent of diffusion for a finite element-finite volume discretization of a linear convection-diffusion equation. *SIAM J. Numer. Anal.*, **53**, 508–526.
- DEVINATZ, A., ELLIS, R. & FRIEDMAN, A. (1973–1974) The asymptotic behavior of the first real eigenvalue of second order elliptic operators with a small parameter in the highest derivatives. II. *Indiana Univ. Math. J.*, **23**, 991–1011.
- DI PIETRO, D., DRONIOU, J. & ERN, A. (2015) A discontinuous-skeletal method for advection-diffusion-reaction on general meshes. *SIAM J. Num. Anal.*, **53**, 2135–2157.
- DI PIETRO, D. A. & ERN, A. (2015) A hybrid high-order locking-free method for linear elasticity on general meshes. *Computer Methods in Applied Mechanics and Engineering*, **283**, 1–21.
- DI PIETRO, D. & ERN, A. (2012) *Mathematical aspects of discontinuous Galerkin methods*. Mathématiques & Applications (Berlin) [Mathematics & Applications], vol. 69. Springer, Heidelberg, pp. xviii+384.
- DIPERNA, R. & LIONS, P. (1989) Ordinary differential equations, transport theory and Sobolev spaces. *Inventiones mathematicae*, **98**, 511–548.
- DRONIOU, J. (2010) Remarks on discretizations of convection terms in Hybrid Mimetic Mixed methods. *Networks and Heterogeneous Media*, **5**, 545–563.
- DRONIOU, J., EYMARD, R., GALLOUËT, T. & HERBIN, R. (2010) A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *Mathematical Models and Methods in Applied Sciences*, **20**, 265–295.
- DRONIOU, J., EYMARD, R. & HERBIN, R. (2016) Gradient schemes: Generic tools for the numerical analysis of diffusion equations. *ESAIM: M2AN*, **50**, 749–781.
- DRONIOU, J. & EYMARD, R. (2006) A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numerische Mathematik*, **105**, 35–71.
- ERN, A., GUERMOND, J.-L. & CAPLAIN, G. (2007) An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs’ systems. *Comm. Partial Differential Equations*, **32**, 317–341.
- ERN, A. & GUERMOND, J.-L. (2004) *Theory and practice of finite elements*. Applied Mathematical Sciences, vol. 159. Springer-Verlag, New York, pp. xiv+524.
- ERN, A. & GUERMOND, J.-L. (2006a) Discontinuous Galerkin methods for Friedrichs’ systems. I. General theory. *SIAM J. Numer. Anal.*, **44**, 753–778.
- ERN, A. & GUERMOND, J.-L. (2006b) Discontinuous Galerkin methods for Friedrichs’ systems. II. Second-order elliptic PDEs. *SIAM J. Numer. Anal.*, **44**, 2363–2388.
- ERN, A. & GUERMOND, J.-L. (2008) Discontinuous Galerkin methods for Friedrichs’ systems. III. Multifield theories with partial coercivity. *SIAM J. Numer. Anal.*, **46**, 776–804.
- ERN, A. & GUERMOND, J.-L. (2016) Finite element quasi-interpolation and best approximation. *ESAIM: M2AN*.
- EYMARD, R., GALLOUËT, T. & HERBIN, R. (2000) Finite volume methods. *Handbook of numerical analysis, Vol. VII*. Handb. Numer. Anal., VII. North-Holland, Amsterdam, pp. 713–1020.
- EYMARD, R., GALLOUËT, T. & HERBIN, R. (2010) Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, **30**, 1009–1043.
- EYMARD, R., HENRY, G., HERBIN, R., HUBERT, F., KLOFKORN, R. & MANZINI, G. (2011) 3D benchmark on discretization schemes for anisotropic diffusion problems on general grids. *Proceedings of Finite Volumes for Complex Applications VI*. Springer, pp. 895–930.

- EYMARD, R., GUICHARD, C. & HERBIN, R. (2012a) Small-stencil 3D schemes for diffusive flows in porous media. *ESAIM Math. Model. Numer. Anal.*, **46**, 265–290.
- EYMARD, R., GUICHARD, C., HERBIN, R. & MASSON, R. (2012b) Vertex-centred discretization of multiphase compositional Darcy flows on general meshes. *Comput. Geosci.*, **16**, 987–1005.
- EYMARD, R., GUICHARD, C., HERBIN, R. & MASSON, R. (2014) Gradient schemes for two-phase flow in heterogeneous porous media and Richards equation. *ZAMM Z. Angew. Math. Mech.*, **94**, 560–585.
- EYMARD, R. & HILHORST, D. AND VOHRALÍK, M. (2010) A combined finite volume–finite element scheme for the discretization of strongly nonlinear convection–diffusion–reaction problems on nonmatching grids. *Numerical Methods for Partial Differential Equations*, **26**, 612–646.
- FRIEDRICHS, K. (1958) Symmetric positive linear differential equations. *Comm. Pure and Appl. Math.*, **11**, 333–418.
- GALDI, G. (1994) *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*. Springer New York.
- GERRITSMAN, M. (2012) An introduction to a compatible spectral discretization method. *Mechanics of Advanced Materials and Structures*, **19**, 48–67.
- GIRAULT, V. (1990) *The Navier-Stokes Equations theory and numerical methods: proceedings of a conference held at Oberwolfach, Sept. 18–24, 1988*. Springer Berlin Heidelberg, pp. 201–218.
- GIRAULT, V. & TARTAR, L. (2010) L^p and $W^{1,p}$ regularity of the solution of a steady transport equation. *Comptes Rendus Mathématique*, **348**, 885 – 890.
- GRISVARD, P. (2011) *Elliptic Problems in Nonsmooth Domains*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics.
- GUERMOND, J.-L. (1999) Stabilization of Galerkin approximations of transport equations by subgrid modeling. *ESAIM: M2AN*, **33**, 1293–1316.
- GUERMOND, J.-L. (2001) Subgrid stabilization of Galerkin approximations of monotone operators. *ZAMM - Journal of Applied Mathematics and Mechanics / Zeitschrift für Angewandte Mathematik und Mechanik*, **79**, 29–32.
- GUERMOND, J. L. (2005) A finite element technique for solving first-order pdes in L^p . *SIAM Journal on Numerical Analysis*, **42**, 714–737.
- GUERMOND, J. L. & MINEV, P. D. (2003) Mixed finite element approximation of an MHD problem involving conducting and insulating regions: the 3D case. *Numer. Methods Partial Differential Equations*, **19**, 709–731.
- GUICHARD, C. (2011) Schémas Volumes Finis sur maillages généraux en milieux hétérogènes anisotropes pour les écoulements polyphasiques en milieux poreux. *Ph.D. thesis*, Université Paris-Est.
- HERMELINE, F. (2000) A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comput. Phys.*, **160**, 481–499.
- HEUER, N. & KARKULIK, M. (2015) A robust DPG method for singularly perturbed reaction-diffusion problems. arXiv:1509.07560.
- HEUMANN, H. (2011) Eulerian en semi-Lagrangian methods for advection-diffusion of differential forms. *Ph.D. thesis*, ETH Zürich.
- HEUMANN, H., HIPTMAIR, R. & PAGLIANTINI, C. (2015) Stabilized Galerkin for transient advection of differential forms. *Research Report*. SAM, ETH Zürich.
- HEUMANN, H. & HIPTMAIR, R. (2008) Extrusion contraction upwind schemes for convection-diffusion problems. *Technical Report* 2008–30. ETH Zürich.

- HILHORST, D. & VOHRALÍK, M. (2011) A posteriori error estimates for combined finite volume–finite element discretizations of reactive transport equations on nonmatching grids. *Comput. Methods Appl. Mech. Engrg.*, **200**, 597–613.
- HIPTMAIR, R. (2001) Discrete Hodge operators. *Numer. Math.*, **90**, 265–289.
- HIPTMAIR, R. (2002) Finite elements in computational electromagnetism. *Acta Numerica*, **11**, 237–339.
- HU, X. & NICOLAIDES, R. A. (1992) Covolume techniques for anisotropic media. *Numerische Mathematik*, **61**, 215–234.
- HYMAN, J., MOREL, J., SHASHKOV, M. & STEINBERG, S. (2002) Mimetic finite difference methods for diffusion equations. *Computational Geosciences*, **6**, 333–352.
- HYMAN, M. & SCOVEL, J.-C. (1990) Deriving mimetic difference approximations to differential operators using algebraic topology. *Technical Report*. Los Alamos National Laboratory.
- JENSEN, M. (2004) Discontinuous Galerkin methods for Friedrichs Systems with irregular solutions. *Ph.D. thesis*, University of Oxford.
- JIANG, B. (1993) Non-oscillatory and non-diffusive solution of convection problems by the iteratively reweighted least-squares finite element method. *Journal of Computational Physics*, **105**, 108 – 121.
- JOHNSON, C., NÄVERT, U. & PITKÄRANTA, J. (1984) Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, **45**, 285–312.
- JOHNSON, C. & PITKÄRANTA, J. (1986) An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation. *Math. Comp.*, **46**, 1–26.
- KREEFT, J., PALHA, A. & GERRITSMAN, M. (2011) Mimetic framework on curvilinear quadrilaterals of arbitrary order. arXiv:1111.4304.
- LEMAIRE, S. (2013) Discrétisation non-conformes d’un modèle poromécanique sur maillages généraux. *Ph.D. thesis*, Université Paris Est.
- LESAINT, P. & RAVIART, P.-A. (1974) On a finite element method for solving the neutron transport equation. *Mathematical Aspects of Finite Elements in Partial Differential Equations*. Academic Press, pp. 89 – 123.
- MATTHIES, G., SKRZYPACZ, P. & TOBISKA, L. (2007) A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.*, **41**, 713–742.
- MATTHIES, G., SKRZYPACZ, P. & TOBISKA, L. (2008) Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.*, **32**, 90–105.
- MATTIUSI, C. (2002) A reference discretization strategy for the numerical solution of physical field problems. *Electron Microscopy and Holography* (P. W. Hawkes ed.). Advances in Imaging and Electron Physics, vol. 121. Elsevier, pp. 143 – 279.
- MULLEN, P., MCKENZIE, A., PAVLOV, D., DURANT, L., TONG, Y., KANSO, E., MARSDEN, J. E. & DESBRUN, M. (2011) Discrete Lie advection of differential forms. *Found. Comput. Math.*, **11**, 131–149.
- NÉDÉLEC, J. C. (1982) Incompressible mixed finite elements for Stokes equations. *Numerische Mathematik*, **39**, 97–112.
- NITSCHKE, J. (1971) Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind. *Abh. Math. Sem. Univ. Hamburg*, **36**, 9–15. Collection of articles dedicated to Lothar Collatz on his sixtieth birthday.
- OHMORI, K. & USHIJIMA, T. (1984) A technique of upstream type applied to a linear nonconforming finite element approximation of convective diffusion equations. *RAIRO Anal. Numér.*, **18**, 309–332.

- OSWALD, P. (1993) On a BPX-preconditioner for P1 elements. *Computing*, **51**, 125–133.
- PALHA, A., KREEFT, K. & GERRITSMA, M. (2010) Numerical solution of advection equations with the discretization of the Lie derivative. Proceedings of the V European Conference on Computational Fluid Dynamics ECCOMAS CFD 2010, pp. 1–23.
- PALHA, A. (2013) High order mimetic discretization. *Ph.D. thesis*, Technische Universiteit Delft.
- REED, W. & HILL, T. (1973) Triangular mesh methods for the neutron transport equation. *Technical Report*. Los Alamos Scientific Laboratory.
- RISCH, U. (1990) An upwind finite element method for singularly perturbed elliptic problems and local estimates in the L^∞ -norm. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, **24**, 235–264.
- ROOS, H.-G., STYNES, M. & TOBISKA, L. (2008) *Robust numerical methods for singularly perturbed differential equations*. Springer Series in Computational Mathematics, vol. 24, second edn. Springer-Verlag, Berlin, pp. xiv+604. Convection-diffusion-reaction and flow problems.
- SCHIEWECK, F. (2008) On the role of boundary conditions for CIP stabilization of higher order finite elements. *ETNA. Electronic Transactions on Numerical Analysis*, **32**, 1–16.
- TARHASAARI, T., KETTUNEN, L. & BOSSAVIT, A. (1999) Some realizations of a discrete Hodge operator: A reinterpretation of finite element techniques. *IEEE Transactions on magnetics*, **35**, 1494–1497.
- TEMAM, R. (1977) *Navier-Stokes Equations: Theory and Numerical Analysis*. North-Holland, Amsterdam.
- TONTI, E. (1975) On the formal structure of physical theories. *Technical Report*. Monograph of the Italian National Research Council.
- TONTI, E. (2013) *The mathematical structure of classical and relativistic physics*. Modeling and Simulation in Science, Engineering and Technology. Birkhäuser/Springer, New York, pp. xxxvi+514. A general classification diagram.
- VOHRALÌK, M. (2005) On the discrete Poincaré–Friedrichs inequalities for nonconforming approximations of the Sobolev Space H^1 . *Numerical Functional Analysis and Optimization*, **26**, 925–952.
- XENOPHONTOS, C. & FULTON, S. R. (2003) Uniform approximation of singularly perturbed reaction-diffusion problems by the finite element method on a Shishkin mesh. *Numerical Methods for Partial Differential Equations*, **19**, 89–111.
- ZAGLMAYR, S. (2006) High Order Finite Element Methods for Electromagnetic Field Computation. *Ph.D. thesis*, Johannes Kepler University.

Résumé

Cette thèse étudie, au niveau continu et au niveau discret sur des maillages polyédriques, les équations de transport tridimensionnelles scalaire et vectorielle. Ces équations sont constituées d'un terme diffusif, d'un terme advectif et d'un terme réactif. Dans le cadre des systèmes de Friedrichs, l'analyse mathématique est effectuée dans les espaces du graphe associés aux espaces de Lebesgue. Les conditions de positivité usuelles sur le tenseur de Friedrichs sont étendues au niveau continu et au niveau discret afin de prendre en compte les cas d'intérêt pratique où ce tenseur prend des valeurs nulles ou raisonnablement négatives. Un nouveau schéma convergeant à l'ordre $\frac{3}{2}$ est proposé pour le problème d'advection-réaction scalaire en considérant des degrés de liberté scalaires associés aux sommets du maillage. Deux nouveaux schémas considérant également des degrés de libertés aux sommets sont proposés pour le problème de transport scalaire en traitant de manière robuste les différents régimes dominants. Le premier schéma converge à l'ordre $\frac{1}{2}$ si les effets advectifs sont dominants et à l'ordre 1 si les effets diffusifs sont dominants. Le second schéma améliore la précision de ce schéma en convergeant à l'ordre $\frac{3}{2}$ lorsque les effets advectifs sont dominants. Enfin, un nouveau schéma convergeant à l'ordre $\frac{1}{2}$ est obtenu pour le problème d'advection-réaction vectoriel en considérant un seul et unique degré de liberté scalaire sur chaque arête du maillage. La précision et les performances de tous ces schémas sont examinées sur plusieurs cas tests utilisant des maillages polyédriques tridimensionnels.

Mot-clés. *maillages polyédriques, systèmes de Friedrichs, advection, diffusion, transport scalaire, transport vectoriel.*

Abstract

This thesis analyzes, at the continuous and at the discrete level on polyhedral meshes, the scalar and the vector transport problems in three-dimensional domains. These problems are composed of a diffusive term, an advective term, and a reactive term. In the context of Friedrichs systems, the continuous problems are analyzed in Lebesgue graph spaces. The classical positivity assumption on the Friedrichs tensor is generalized so as to consider the case of practical interest where this tensor takes null or slightly negative values. A new scheme converging at the order $\frac{3}{2}$ is devised for the scalar advection-reaction problem using scalar degrees of freedom attached to mesh vertices. Two new schemes considering as well scalar degrees of freedom attached to mesh vertices are devised for the scalar transport problem and are robust with respect to the dominant regime. The first scheme converges at the order $\frac{1}{2}$ when advection effects are dominant and at the order 1 when diffusion effects are dominant. The second scheme improves the accuracy by converging at the order $\frac{3}{2}$ when advection effects are dominant. Finally, a new scheme converging at the order $\frac{1}{2}$ is devised for the vector advection-reaction problem considering only one scalar degree of freedom per mesh edge. The accuracy and the efficiency of all these schemes are assessed on various test cases using three-dimensional polyhedral meshes.

Keywords. *polyhedral meshes, Friedrichs systems, advection, diffusion, scalar transport, vector transport.*



