Numerical resolution of partial differential equations with variable coefficients

Joubine Aghili

THÈSE

Pour obtenir le grade de

Docteur

Délivré par l'**Université de Montpellier**

Préparée au sein de l'école doctorale

Information, Structures, Systèmes

Et de l'unité de recherche

Institut Montpelliérain Alexander Grothendieck

Spécialité: **Mathématiques et Modélisation**

Présentée par **Joubine Aghili**

Numerical resolution of partial differential equations with variable coefficients

Soutenue le 2 décembre 2016 devant le jury composé de :

M. Luca Formaggia	Politecnico di Milano	Rapporteur
M. Nicolas Seguin	Université de Rennes 1	Rapporteur
M. Sébastien Boyaval	Université Paris-Est	Co-encadrant
M. Daniele Di Pietro	Université de Montpellier	Directeur
Mme Françoise Krasucki	Université de Montpellier	Examinatrice
M. Jean-Claude Latché	IRSN	Examinateur
M. Fabien Marche	Université de Montpellier	Examinateur
M. Roland Masson	Université de Nice	Examinateur

Contents

Remerciements

Le document que vous tenez entre les mains n'aurait jamais pu voir le jour sans :

— **Daniele A. Di Pietro**, toujours présent et à l'écoute. Ta perfectionnisme, ta rigueur scientifique, tes petites blagues qui tombent toujours à pic ont été plus qu'appréciés. J'ai énormément appris en très peu de temps, je tiens à te remercier pour cet encadrement exceptionnel.

— **Sébastien Boyaval**, un puits de connaissance qui ne se contente jamais de répondre simplement aux questions. Merci pour ton encadrement et surtout pour tout le temps que tu as ~~perdu~~ pris pour m'expliquer (et me réexpliquer) toutes ces choses.

Je vous remercie encore tous les deux pour votre encadrement pendant ces trois années.

Je tiens à remercier **Luca Formaggia** et **Nicolas Seguin** pour avoir accepté de rapporter cette thèse ainsi que pour vos nombreuses remarques et corrections. Merci également à **Françoise Krasucki**, **Roland Masson**, **Fabien Marche** et **Jean-Claude Latché** pour avoir accepté de constituer ce jury de thèse.

Merci également à tous **les permanents de l'IMAG**, en particulier aux membres de l'**équipe ACSIOM** avec qui j'ai pu discuter de temps à autre. Un vif remerciement à **Bernadette Lacan** pour son efficacité redoutable et avec qui tout devient plus simple. Je vous remercie tous pour m'avoir permis de travailler dans d'excellentes conditions.

Je ne peux oublier l'**équipe d'Orsay**, en particulier **Filippo Santambrogio** pour m'avoir suivi et encouragé à continuer en thèse pendant toute l'année de Master 2. Ton appui a été crucial sur la fin, je te remercie encore pour ton implication. Je remercie également la **Fondation Mathématique Jacques Hadamard** pour m'avoir soutenu pendant le stage de Master.

Une pensée va à présent à **mes profs de lycée**, en particulier les matheux : **Youri**

Beltchenko pour m'avoir transmis son goût pour les mathématiques au bon moment ainsi que **Eliane Gayout** et **Laurent Thieulin** mes profs de spé et de sup dont je garde un excellent souvenir.

Une avalanche de gratitude destinée à toute la bande de thésards, jeunes docteurs et postdocs de Montpellier : mes ex-cobureaux et actuels cobureaux : **Benjamin**, **Julien** premier interlocuteur quand il s'agit de parler de choses non-scientifiques, **Abel**[1] et **Maud** qui vient de débarquer[2]. Puis **Gautier**  , **Mickaël** pour tes fous rires permanents qui donnent la patate, **Elsa** qui est devenue maintenant une experte à Mario Kart, **Christian**, **Guillaume** pour ta capacité de déconne sans limites, **Myriam** et ton oreille toujours attentive, la team HHO : **Rita**, **Florent** et **Michele** bon courage à vous !, **Coralie**, **Wenran**, **Alexandre**, **Paul**, **Paul-Marie**, **Jocelyn**, **Julien (l'autre)** tiens tu m'passes le sel ?, **Jérémy**, **Arnaud** et son pull Naf Naf qui m'a toujours fait rêver, **Christophe** toujours prêt pour parler 3h de nos histoires sur un banc, **Stéphanie**, **Angelina**, **Ridha**, **Rodrigo**, **Louis**, **Mario**, **Etienne**, **Antoine**, **Tutu**, **Francesco** toujours d'attaque pour aller danser, **Samuel**, **Emmanuelle**, **Anis**, **Quentin** & **Théo**. Je remercie aussi [ ] que j'ai honteusement oublié.

Sortons à présent de la sphère mathématique. Une gratitude éternelle va en tout premier lieu à **ma mère**, mon premier soutien, et ça depuis toujours. Puis à toute **ma famille** là-haut, qui a toujours représenté un appui solide et permanent. Merci à vous d'exister. Je n'oublie pas **mon père**, qui n'a pu faire le déplacement pour la soutenance, je te remercie pour ses encouragements constants à distance tout au long de ce périple. Mention spéciale à **Benjamin**, mon ami et acolyte de toujours, **Anastasia**, pour ton soutien pendant nos années étudiantes, je te souhaite le meilleur pour ta thèse ! **Plume** pour tes ronrons, **Sarah & Salma** avec qui j'apprécie toujours de parler de tout, de rien mais surtout de n'importe quoi. Enfin mes anciens **camarades de fac, prépa et de terminale** qui se reconnaîtront. Merci **Gautier** pour ta relecture minutieuse ainsi qu'**Héloïse** pour ta relecture nocturne ! Merci aussi **Juline** pour ton soutien abondant pendant les derniers mois. Dédicace spéciale aux **savateurs du club Caméléon**, mon lieu de prédilection pour sortir de ma bulle et apprécier de temps à autre, l'efficacité du chassé-bas-revers-pointé-figure.

Il serait ingrat d'oublier de remercier la  **France** pour nous avoir accueilli et rendu tout ça possible.

1. Never gonna give you up. Never gonna let you down.
2. Compte sur moi pour te ralentir dans ton travail

Introduction

Contexte, motivations et structure du manuscrit

Dans cette thèse on s'intéresse à la résolution numérique d'Équations aux Derivées Partielles (EDP), en accordant une attention particulière à celles qui dépendent de coefficients physiques variables en espace. Un exemple important est l'équation de diffusion pure, consistant à trouver une fonction $u : \Omega \to \mathbb{R}$ (avec Ω domaine polyédrique de $\mathbb{R}^d$, $d \geqslant 1$) vérifiant

$$-\operatorname{div}(\boldsymbol{\kappa}\boldsymbol{\nabla} u) = f \qquad \text{dans } \Omega, \tag{1}$$

quand $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}$ et $f : \Omega \to \mathbb{R}$ sont des fonctions données suffisamment régulières et des conditions opportunes sont fixées au bord de Ω. Cette équation traduit la conservation d'un flux (ou variable vectorielle) $\boldsymbol{\sigma} = -\boldsymbol{\kappa}\boldsymbol{\nabla} u$ dépendant de manière linéaire du gradient d'un potentiel (ou variable scalaire) u à travers un coefficient physique $\boldsymbol{\kappa}$, et elle apparaît dans de nombreuses branches des sciences et de l'ingénierie : thermique, hydrodynamique, électrostatique, etc.

Le calcul *exact* des solutions de (1) pour des données générales est très souvent hors de portée, ce qui justifie le recours à la résolution numérique. Certaines méthodes se sont imposées au cours des dernières décennies comme méthodes de référence. On peut nommer, parmi d'autres, la Méthode des Éléments Finis (FEM) ou la méthode Volumes Finis (FV), qui ont permis de résoudre de nombreux problèmes des EDP. Cependant, ces méthodes se montrent de plus en plus limitées pour traiter des problèmes de nature plus complexe, comportant, p.ex., des variations importantes des coefficients physiques, des géométries complexes, ou des couplages multiphysique. Au cours des dernières années, des nouveaux paradigmes ont émergé qui ont permis de repenser les schémas numériques classiques pour mettre en place des méthodes plus modernes. Ces méthodes, supportant des maillages généraux en di-

mension d'espace quelconque ainsi que des ordres d'approximation arbitraires, ont la particularité d'être construites en embarquant ou reproduisant la physique du problème dans leurs formulations.

Le développement et l'utilisation de ces méthodes de dernière génération soulèvent de nombreuses questions à étudier. Dans ce travail, nous nous pencherons sur quelques-unes d'entre elles. Nous nous focaliserons particulièrement sur des questions liées (i) à la mise en œuvre efficace de méthodes polyédriques mixtes d'ordre arbitraire (Mixed High-Order) ; (ii) à leur application à des problèmes en mécanique des fluides incompressibles ; (iii) à leur convergence *hp* ; (iv) et à la réduction de modèle. Ces points sont traités sous la forme de quatre chapitres distincts. Par la suite, nous allons résumer brièvement le contenu de ces chapitres en mettant l'accent sur les résultats marquants.

Dans le **Chapitre 1**, tiré de [2], nous nous intéressons à la mise en œuvre efficace (par hybridation) de la méthode de discrétisation polyédrique d'ordre arbitraire *Mixed High-Order* (MHO) de [57]. La méthode MHO étend les idées des FEM mixtes classiques aux maillages constitués d'éléments polyédriques. On rencontre de tels maillages, p.ex., dans le contexte des écoulements en milieux poreux, où des éléments polyédriques et des interfaces non conformes apparaissent pour modéliser l'érosion et la formation de fractures ou failles. Dans la modélisation des réservoirs pétroliers, on retrouve également des éléments polyédriques dans l'abord des puits pour effectuer le raccord entre le maillage du puits (radial) et celui du réservoir (Corner-Point Geometry).

Le principe des méthodes d'approximation mixtes (dont MHO) consiste à approcher de manière séparée le flux $\boldsymbol{\sigma}$ et le potentiel u, ce qui donne lieu à un problème de type point-selle. D'un point de vue numérique, ce type de problème est moins agréable à résoudre par rapport à une formulation coercive en la variable u uniquement. Dans le cas des méthodes FEM, il est bien connu [6, 8, 42, 89, 107] qu'on peut se ramener à un problème coercif (donc, plus simple à résoudre) par un processus d'*hybridation*. L'hybridation consiste à imposer la continuité de la composante normale de la variable flux à l'aide de multiplicateurs de Lagrange, et à résoudre un problème où ceux-ci apparaissent comme les seules inconnues globalement couplées. Dans le Chapitre 1 nous étendons ces idées à la méthodes MHO. Les résultats les plus importants de l'analyse menée dans ce chapitre sont : (i) l'obtention d'une formulation primale équivalente pour la méthode MHO, qui en permet une mise en œuvre efficace où les seuls degrés de liberté globalement couplés sont des polynômes discontinus sur le squelette du maillage ; (ii) l'identification d'un lien avec la méthode

Hybrid High-Order (HHO) primale de [59]. On propose aussi une analyse de convergence en norme d'énergie et en norme L^2 basée directement sur la version hybridisée de la méthode MHO, qui étend les résultats de [57].

Dans le **Chapitre 2** nous étendons la version hybridée de la méthode MHO à des problèmes issus de la mécanique des fluides. Dans un premier temps, nous considérons le problème de Stokes (cette partie est tirée de [2, Section 4]). Celui-ci peut être vu comme une version vectorielle du problème de Poisson (voir (1) avec $\boldsymbol{\kappa} = \boldsymbol{I}_d$) sous la contrainte de divergence nulle pour la vitesse. Classiquement [14,31], la bonne position du problème discret repose dans ce cas sur une condition inf-sup. Dans le cas de la méthode proposée, cette condition est obtenue en exploitant l'interprétation des multiplicateurs de Lagrange introduits dans le processus d'hybridation du Chapitre 1 comme traces du potentiel. Les résultats principaux sont : (i) l'obtention d'une nouvelle méthode de discrétisation pour le problème de Stokes inf-sup stable sur maillages généraux ; (ii) des estimations d'erreur optimales en norme d'énergie et un résultat de superconvergence en norme L^2 pour la vitesse.

Dans un second temps, nous considérons le problème d'Oseen, caractérisé par l'ajout de termes advectif et réactif au problème précédent. Cette application est originale, et n'a pas été publiée ailleurs à notre connaissance. Ici, le point clé est la prise en compte du terme advectif, qui demande une attention particulière afin d'obtenir une méthode robuste par rapport au nombre de Péclet (nombre sans dimension représentant le rapport entre les effets advectifs et diffusifs). Nous nous inspirons pour cela des techniques récemment proposées dans [54] pour un problème de diffusion-advection-réaction scalaire. Les résultats principaux obtenus dans ce chapitre sont : (i) l'obtention d'une nouvelle méthode de discrétisation pour le problème d'Oseen sur maillages généraux ; (ii) des estimations d'erreurs optimales en norme d'énergie montrant la variation de l'ordre de convergence en fonction du nombre de Péclet local.

Le **Chapitre 3** concerne l'étude théorique et numérique d'une variation *hp* de la méthode HHO de [59] pour le problème de diffusion variable (1) (on rappelle que le lien entre la méthode HHO et la version hybridée de la méthode MHO a été étudié dans le Chapitre 1). La dénomination *hp* fait référence à des méthodes numériques où la finesse du maillage ainsi que l'ordre d'approximation polynomial peuvent varier, même simultanément. Les résultats marquants de ce chapitre sont les suivants : (i) l'obtention d'une nouvelle méthode *hp*-HHO permettant la variation locale du degré polynomial ; (ii) des estimations d'erreurs *hp* en norme d'énergie et L^2 robustes vis-à-vis des hétérogéneités et du coefficient de diffusion ; (iii) des résultats

d'approximation *hp* sur maillages généraux s'appliquant potentiellement à toute méthode basée sur des espaces de polynômes par morceaux.

Enfin, le **Chapitre 4** contient des perspectives sur l'approximation numérique, en de nombreuses valeurs d'un paramètre $\boldsymbol{\mu}$, des solutions paramétrées d'une EDP. Une possibilité consiste à pré-calculer un petit nombre de solutions pour des valeurs fixées du paramètre, puis à obtenir la solution pour une valeur quelconque de $\boldsymbol{\mu}$ par projection sur l'espace affine décrit par le petit nombre de solutions précalculées. C'est l'idée fondatrice de la méthode dite des Bases Réduites (BR), qui repose néanmoins sur l'hypothèse qu'une méthode numérique précise existe pour toute valeur du paramètre $\boldsymbol{\mu}$ (les valeurs pour lesquelles la méthode précise sera vraiment utile seront choisies au cours d'une phase d'apprentissage dite "hors-ligne", en référence au contexte d'application "temps réel" de la méthode).

La méthode des Bases Réduites (BR) a fait l'objet de nombreuses recherches récentes [1,16,22,28–30,34,40,77,78,81,92,93,96–98]) et a permis de construire un bon modèle réduit dans un certain nombre de cas pratiques : des problèmes avec domaines paramétrés [88,98], des problèmes paraboliques, hyperboliques ou encore non-linéaires, voir [81, 93].

Nous nous focalisons ici sur l'équation de diffusion (1) dépendant d'un paramètre $\boldsymbol{\mu}$ intervenant dans les expression du coefficient de diffusion et du terme source. On montre à l'aide d'exemples numériques que l'application usuellement faite de la méthode BR peut être potentiellement améliorée en utilisant la formulation mixte du problème plutôt que la formulation primale. Précisément, quand le terme source est dans $L^2(\Omega)$ alors on peut améliorer les performances de la réduction en appliquant BR avec le projecteur hérité de la formulation mixte plutôt que le projecteur hérité de la formulation primale (on peut potentiellement soit diminuer l'erreur liée à la projection à dimension d'espace réduit donnée, soit diminuer la dimension de l'espace réduit à tolérance fixée sur l'erreur de projection).

Le manuscrit est complété par l'**Annexe A** présentant des détails pratiques pour l'implémentation de la méthode MHO pour le problème de Poisson.

Hybridation de la méthode Mixed High-Order

Dans le Chapitre 1 de ce manuscrit nous étudions l'hybridation de la méthode Mixed High-Order (MHO) de [57]. L'hybridation est une procédure formalisée dans [8]

mettant un cadre mathématique sur une technique déjà connue auparavant des ingénieurs [50] et permettant une mise en œuvre efficace des méthodes mixtes (c.-à-d., des méthodes où on approche le flux et le potentiel de manière indépendante). L'intérêt de l'hybridation consiste à écrire un problème coercif (et, donc, plus facile à résoudre en pratique) équivalent au problème de type point selle correspondant à la méthode mixte. Dans sa forme classique, l'hybridation se décompose en deux étapes : (i) relaxer la condition de continuité de la composante normale du flux aux interfaces et l'imposer par multiplicateurs de Lagrange (ii) éliminer la variable flux localement dans le but de réduire la taille du système final. Les résultats les plus importants de l'analyse menée dans ce chapitre sont : (i) l'obtention d'une formulation primale équivalente pour la méthode MHO, qui en permet une mise en œuvre efficace où les seuls degrés de liberté globalement couplés sont des polynômes discontinus sur le squelette du maillage ; (ii) l'identification d'un lien avec la méthode Hybrid High-Order (HHO) primale de [59]. On montre, en particulier, que la formulation primale obtenue par hybridation de la méthode MHO coïncide avec la méthode HHO à la stabilisation près.

On considère ici le problème (1) dans le cas $\boldsymbol{\kappa} = \boldsymbol{I}_d$. Sous forme faible, ce problème consiste à trouver le flux $\boldsymbol{\sigma} \in \mathbf{H}(\mathrm{div};\Omega)$ et le potentiel $u \in L^2(\Omega)$ tels que

$$\begin{aligned} (\boldsymbol{\sigma},\boldsymbol{\tau}) + (\boldsymbol{\nabla}\cdot\boldsymbol{\tau},u) &= 0 && \forall \boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega),\\ -(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma},q) &= (f,q) && \forall q \in L^2(\Omega), \end{aligned} \tag{2}$$

où nous avons noté $(\cdot,\cdot)$ les produits scalaires standards de $L^2(\Omega)$ et $L^2(\Omega)^d$. L'exemple suivant sert à illustrer la notion d'hybridation pour la méthode FE classique Raviart–Thomas-$\mathbb{P}^0$. Il permettra également d'identifier plus aisément les différences par rapport à la méthode MHO considérée ici.

Exemple (Méthodes de Raviart–Thomas et de Crouzeix–Raviart)**.** Soit $\mathcal{T}_h$ un maillage simplicial conforme. En utilisant l'espace de Raviart-Thomas $\mathbb{RT}^0(\mathcal{T}_h) \subset \mathbf{H}(\mathrm{div};\Omega)$ introduit dans [94] pour le flux et l'espace des polynômes constants par morceaux $\mathbb{P}^0(\mathcal{T}_h) \subset L^2(\Omega)$ pour le potentiel, la version discrète du problème (2) s'écrit

$$\begin{aligned} (\boldsymbol{\sigma}_h,\boldsymbol{\tau}_h) + (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h,u_h) &= 0 && \forall \boldsymbol{\tau}_h \in \mathbb{RT}^0(\mathcal{T}_h),\\ -(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h,q_h) &= (f,q_h) && \forall q_h \in \mathbb{P}^0(\mathcal{T}_h). \end{aligned} \tag{3}$$

L'hybridation de (3) consiste à introduire l'espace Λ_h de fonctions constantes par morceaux sur les interfaces de $\mathcal{T}_h$ et l'espace de Raviart-Thomas *discontinu* $\mathbb{RT}^{0,d}(\mathcal{T}_h)$,

et à reformuler le problème (3) sous forme *mixte-hybride* comme suit :

$$\begin{aligned}
(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (\boldsymbol{\nabla}\cdot\boldsymbol{\tau}_h, u_h) + \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T^{\mathrm{i}}} (\boldsymbol{\tau}_h\cdot\boldsymbol{n}_{TF}, \lambda_h)_F &= 0 && \forall \boldsymbol{\tau}_h \in \mathbb{RT}^{0,d}(\mathcal{T}_h),\\
-(\boldsymbol{\nabla}\cdot\boldsymbol{\sigma}_h, q_h) &= (f, q_h) && \forall q_h \in \mathbb{P}^0(\mathcal{T}_h),\\
\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T^{\mathrm{i}}} (\boldsymbol{\sigma}_h\cdot\boldsymbol{n}_{TF}, \mu_h)_F &= 0 && \forall \mu_h \in \Lambda_h.
\end{aligned} \tag{4}$$

Ici, pour tout élément $T \in \mathcal{T}_h$ du maillage, on a noté $\mathcal{F}_T^{\mathrm{i}}$ l'ensemble de ses interfaces. La différence entre (3) et (4) est que, dans le deuxième cas, la continuité de la composante normale des flux est imposée à l'aide de multiplicateurs de Lagrange dans Λ_h. Habituellement on élimine la variable flux $\boldsymbol{\sigma}_h \in \mathbb{RT}^{0,d}(\mathcal{T}_h)$ par condensation statique (grâce à la structure diagonale par blocs) afin d'obtenir un problème en les seuls multiplicateurs de Lagrange sous la forme

$$\mathbf{A}\mathsf{\Lambda} = \mathsf{F},$$

où la matrice $\mathbf{A}$ est symétrique définie positive. Si l'on considère maintenant $\mathbb{CR}(\mathcal{T}_h)$ l'espace de Crouzeix-Raviart de [49] construit comme l'espace des polynômes affines par morceaux et continus aux points milieux des faces du maillage, l'approximation non-conforme du problème s'écrit alors

$$(\boldsymbol{\nabla}_h u_h, \boldsymbol{\nabla}_h v_h) = (f, v_h) \quad \forall v_h \in \mathbb{CR}_0(\mathcal{T}_h),$$

où $\mathbb{CR}_0(\mathcal{T}_h)$ est le sous-espace de $\mathbb{CR}(\mathcal{T}_h)$ dont les degrés de liberté aux bords s'annulent et $\boldsymbol{\nabla}_h$ est le gradient par morceaux associé à $\mathcal{T}_h$. Ce dernier problème se réécrit matriciellement sous la forme

$$\mathbf{B}\mathsf{U} = \mathsf{G},$$

avec $\mathbf{B}$ matrice symétrique définie positive. Il est bien connu [6, 8, 42, 89, 107] que les matrices $\mathbf{A}$ et $\mathbf{B}$ sont identiques ainsi que les seconds membres F et G donnant lieu à une parfaite équivalence entre ces deux approximations.

La méthode MHO considérée ici présente des analogies mais aussi des différences importantes par rapport à l'élément fini de Raviart–Thomas de l'exemple précédent. Une première différence cruciale est qu'elle s'applique à des maillages $\mathcal{T}_h$ polyédriques généraux. En outre, le paradigme des fonctions de base est remplacé par la notion de reconstruction obtenue en opérant directement sur les degrés de liberté. Ainsi,

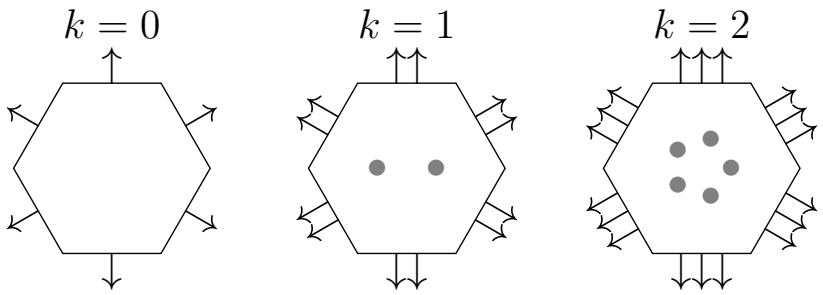


FIGURE 1 – Espace $\mathbf{\Sigma}_T^k$ des degrés de liberté pour le flux dans la méthode MHO pour $k \in \{0, 1, 2\}$. Ici, on considère l'exemple d'un élément T hexagonal.

pour tout entier $k \geqslant 0$, des espaces de degrés de liberté discrets de flux $\mathbf{\Sigma}_T^k$ (voir Figure 1) et de potentiel $U_T^k = \mathbb{P}^k(T)$ sont associés à chaque élément T du maillage $\mathcal{T}_h$ encodant les caractéristiques de l'objet continu par la donnée de polynômes. Deux opérateurs de reconstruction opérant sur l'espace $\mathbf{\Sigma}_T^k$ sont alors définis élément par élément : (i) une reconstruction D_T^k de l'opérateur de divergence utilisée dans les termes de couplage et satisfaisant une propriété de commutativité opportune et (ii) une reconstruction $\mathfrak{C}_T^k$ du flux, utilisée pour définir le pendant du produit L^2 dans l'espace des flux.

En notant $\mathbf{\Sigma}_h^k$ l'espace des degrés de liberté globaux pour le flux obtenu imposant la continuité des inconnues aux interfaces et U_h^k l'espace des polynômes de degré k discontinus sur le maillage, la méthode MHO s'écrit : Trouver $\boldsymbol{\sigma}_h \in \mathbf{\Sigma}_h^k$ et $u_h \in U_h^k$, tels que

$$\begin{aligned} H_h(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (u_h, D_h^k \boldsymbol{\tau}_h)_T &= 0 \qquad && \forall \boldsymbol{\tau}_h \in \mathbf{\Sigma}_h^k \\ -(D_h^k \boldsymbol{\sigma}_h, q_h) &= (f, q_h)_T && \forall q_h \in U_h^k. \end{aligned} \tag{5}$$

Ici, l'opérateur de divergence globale D_h^k agissant sur l'espace $\mathbf{\Sigma}_h^k$ est posé égal dans chaque élément à l'opérateur de divergence locale D_T^k décrit plus haut, tandis que le produit scalaire H_h sur $\mathbf{\Sigma}_h^k$ est obtenu par assemblage de contributions locales H_T sur $\mathbf{\Sigma}_T^k$, $T \in \mathcal{T}_h$, de la forme

$$H_T(\boldsymbol{\sigma}, \boldsymbol{\tau}) := (\mathfrak{C}_T^k \boldsymbol{\sigma}, \mathfrak{C}_T^k \boldsymbol{\tau})_T + \text{stabilisation},$$

où $(\cdot, \cdot)_T$ est le produit $L^2(T)^d$ et le terme de stabilisation sert à assurer la coercivité de la forme bilinéaire. Le caractère bien posé du problème (5) est une conséquence de cette dernière propriété combinée avec la commutativité de la divergence, qui permet de prouver la stabilité au sens de la condition *inf-sup*.

L'hybridation de la formulation MHO (5) présentée dans le Chapitre 1 de ce ma-

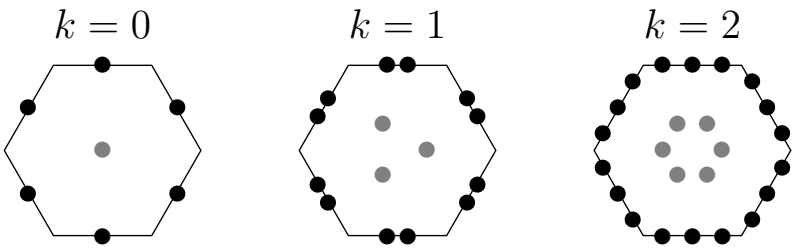


FIGURE 2 – Espace W_T^k des degrés de liberté pour le potentiel dans la version hybridée de la méthode MHO pour $k \in \{0, 1, 2\}$ et élément hexagonal comme dans la Figure 1. L'espace global W_h^k est obtenu en imposant l'unicité des multiplicateurs de Lagrange aux interfaces.

nuscrit s'effectue dans un esprit similaire aux FE classiques, mais avec quelques différences importantes. Tout d'abord, il ne s'agit plus de garantir la continuité de la composante normale du flux (qui n'est pas définie partout dans l'élément), mais l'unicité des valeurs des degrés de liberté aux interfaces grâce à des multiplicateurs de Lagrange. Ensuite, la correspondance que l'on prouve avec une méthode primale de type HHO est valable pour tout degré polynomial $k \geqslant 0$ (tandis que le résultat concernant les méthodes de Raviart–Thomas-$\mathbb{P}^0$ et de Crouzeix–Raviart n'est valable qu'à l'ordre plus bas). Le point clé consiste ici à identifier un opérateur de couplage *potentiel-vers-flux* transformant un couple de variables potentiel-trace vers une variable flux, et permettant de formaliser de manière élégante l'étape d'élimination des inconnues de flux pour le problème mixte-hybride. Notant W_h^k l'espace contenant les degrés de liberté pour le potentiel u_h et les multiplicateurs de Lagrange λ_h (cf. Figure 2), on arrive à une reformulation de (5) de la forme

$$A_h((u_h, \lambda_h), (v_h, \mu_h)) = (f_h, v_h) \qquad \forall (v_h, \mu_h) \in W_h^k, \tag{6}$$

où A est une forme bilinéaire sur $W_h^k \times W_h^k$ agissant uniquement sur les couples potentiel-traces. Un résultat important prouvé ici est que cette forme bilinéaire coïncide à la stabilisation près avec celle introduite dans [59] dans le cadre des méthodes HHO (récemment, des résultats d'équivalence plus généraux ont été obtenus dans [25]). Ce résultat est valable sur maillages généraux et pour tout degré polynomial $k \geqslant 0$. Un autre point à noter est que la taille du système linéaire correspondant à l'équation (6) peut être ultérieurement réduite en éliminant par condensation statique les inconnues de potentiel à l'intérieur de l'élément (en gris dans la Figure 2).

Le chapitre se clôture par une analyse d'erreur complète de la méthode hybridée ainsi obtenue. Cette analyse permet de retrouver des résultats qui incluent ceux

de [57] en travaillant directement sur la version hybridée de la méthode. Lorsque des polynômes de degré k sont utilisés, on prouve une convergence à l'ordre $(k+1)$ pour le flux et à l'ordre $(k+2)$ pour le potentiel. On remarquera, par ailleurs, une différence importante par rapport à la méthode Hybridizable Discontinuous Galerkin (HDG) de [45] qui, pour un choix d'inconnues similaire, obtient des ordres de convergence k et $(k+1)$, respectivement (pour plus de détails, voir [44]). Un point important issu de l'analyse d'erreur qui motive l'extension de la méthode aux problèmes de Stokes et Oseen dans le Chapitre 2 est que les multiplicateurs de Lagrange peuvent être interprétés comme des traces du potentiel.

Il est utile de terminer cette discussion par une petite section bibliographique. Les méthodes MHO et HHO sont conçues afin de traiter des maillages généraux tout en garantissant un ordre polynomial d'approximation arbitraire. Le premier schéma HHO apparaît dans [56] dans le cadre d'un problème d'élasticité linéaire quasi-incompressible. Des travaux ultérieurs en ont considéré l'extension à de nombreux problèmes linéaires et non linéaires ; voir, p.ex., [23, 41, 53, 54, 61, 63] ainsi que les références contenues dans ces articles.

La prise en compte de maillages généraux a été considérée dans un premier temps (à partir des années 2000) dans le cadre des méthodes de bas ordre. On peut commencer par citer la méthode Mimetic Finite Difference (MFD) [33]. L'approche MFD est davantage algébrique, elle repose sur la reproduction de propriétés mathématiques et physiques fondamentales comme les lois de conservation, de symétrie, de positivité des solutions et des relations clés entre opérateurs différentiels. Les inconnues sont situées aux noeuds des éléments bien qu'une version avec inconnues aux faces (et aux mailles) soit étudiée dans [85]. Il est important de citer aussi les méthodes de type FV comme la méthode Hybrid Finite Volume (HFV) [70] ou encore la méthode Mixed Finite Volume (MFV) [65]. Plus récemment, sont apparues les méthodes de type Compatible Discrete Operator (CDO) basées aussi sur des inconnues aux sommets et aux maillages et faisant intervenir le maillage dual, voir [27]. Des connexions avec la méthode MFD ainsi qu'une synthèse avec d'autres méthodes mimétiques sont présentées dans [66] et [26].

Plus récemment, on a considéré également la possibilité de monter en ordre sur maillages généraux. La possibilité d'utiliser des méthodes de type Galerkine discontinues (dG) d'ordre élevé sur maillages généraux a été mise en évidence dans [17, 55]. Bien que les méthodes dG offrent de nombreux avantages, elles présentent néanmoins des systèmes à inverser plus coûteux que les méthodes FEM classiques du fait du grand nombre de degrés de liberté. Une approche née pour essayer de contourner

cette difficulté est la méthode dite Hybridizable Discontinuous Galerkin (HDG), qui a vu le jour grâce aux travaux de Cockburn et *al.* Dans les méthodes HDG, il s'agit souvent de traiter le flux normal le long des interfaces comme une nouvelle variable devant vérifier une *relation de transmission*. Cette dernière ne dépend pas du flux ni du potentiel, ce qui permet ensuite d'éliminer localement ces variables et de réduire la taille du système, voir [46]. Il est important de noter la proximité avec la méthode HHO où les mêmes degrés de liberté sont utilisés ainsi que le processus d'hybridation. Cependant, comme on l'a fait remarquer plus haut, la méthode HHO présentée montre une vitesse de convergence d'un ordre supplémentaire ; ce sujet est traité en détail dans [44]. Une approche conforme permettant de combiner maillages généraux et ordre élevé est la méthode Virtual Element Method (VEM), inspirée de la méthode MFD nodale. Introduite initialement dans [19], elle peut être vue comme une généralisation de la méthode des EF aux maillages généraux. La différence principale par rapport aux méthodes EF classiques est que les fonctions de forme ne sont pas connues en tout point de l'élément, ce qui justifie le terme "Virtual". Par conséquent, les formes bilinéaires sont approchées par des versions discrètes obtenues par somme d'une partie consistante et d'une partie stabilisante, qui nécessitent uniquement les valeurs des degrés de libertés locaux.

Application aux problèmes de Stokes et d'Oseen

Le Chapitre 2 concerne l'application de la méthode HHO du chapitre précédent à des problèmes linéaires en mécanique des fluides.

Dans un premier temps, on considère le problème de Stokes. Celui-ci apparaît dans l'étude des écoulements de fluides visqueux où les effets inertiels sont négligeables. Ce modèle est utilisé, par exemple, pour décrire des écoulements laminaires dans des canaux étroits. Les premières traces du modèle remontent à l'étude de G.G. Stokes dans le cas d'un écoulement dans une cavité étroite autour d'un objet sphérique [103]. Étant donné un domaine Ω de $\mathbb{R}^d$, $d \in \{2,3\}$, et un champ de force $\boldsymbol{f}$ fixé, le problème de Stokes consiste à trouver le champ de vitesse $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ et un champ de pression $p : \Omega \to \mathbb{R}$ tels que

$$\begin{aligned} -\triangle \boldsymbol{u} + \boldsymbol{\nabla} p &= \boldsymbol{f} \qquad && \text{dans } \Omega, \\ \boldsymbol{\nabla}\cdot\boldsymbol{u} &= 0 && \text{dans } \Omega, \end{aligned} \tag{7}$$

avec des conditions aux bords opportunes. Par simplicité, dans ce qui suit, on considère le cas de conditions de Dirichlet homogènes pour $\boldsymbol{u}$, ce qui demande

d'ajouter la condition $\int_\Omega p = 0$ pour garantir l'unicité de la pression ; l'extension à des conditions aux bords plus générales ne pose pas de difficultés particulières. D'un point de vue mathématique, le problème (7) peut se comprendre comme la version vectorielle de (1) avec $\boldsymbol{\kappa} = \boldsymbol{I}_d$ sous contrainte de divergence nulle. La difficulté majeure dans l'approximation de ses solutions réside alors dans la structure point–selle du problème. En effet, en utilisant les méthodes classiques, il est bien connu que la bonne position du problème au niveau continu n'implique pas celle au niveau discret. Les travaux classiques de Babuška [14] et Brezzi [31] ont apporté des outils permettant de sélectionner les "bons" couples d'espaces où approcher vitesse et pression permettant ainsi de garantir l'existence et unicité des solutions. Le point clé consiste à satisfaire au niveau discret une condition *inf-sup* analogue à celle du problème continu et exprimant la surjectivité de l'opérateur de divergence discret. Ce point délicat est à l'origine d'une abondante littérature autour des méthodes d'approximation pour traiter numériquement le problème de Stokes.

La méthode MHO développée dans ce manuscrit s'inspire de la formulation variationnelle suivante de (7) : Trouver la vitesse $\boldsymbol{u} \in H^1_0(\Omega)^d$ et la pression $p \in L^2_0(\Omega)$ (avec $L^2_0(\Omega)$ espace des fonctions de carré intégrables et à moyenne nulle sur Ω) telles que

$$
\begin{aligned}
(\boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v}) - (p, \boldsymbol{\nabla}\cdot\boldsymbol{v}) &= (\boldsymbol{f}, \boldsymbol{v}) && \forall \boldsymbol{v} \in H^1_0(\Omega)^d && \text{(8a)}\\
(\boldsymbol{\nabla}\cdot\boldsymbol{u}, q) &= 0 && \forall q \in L^2_0(\Omega). && \text{(8b)}
\end{aligned}
$$

Chaque composante u_i de la vitesse $\boldsymbol{u}$ est discrétisée comme un élément $(u_{h,i}, \lambda_{h,i})$ de l'espace hybride W_h^k (cf. Figure 2), tandis que la pression est discrétisée comme un élément de l'espace $\mathcal{P}_h^k$ des polynômes par morceaux sur le maillage $\mathcal{T}_h$ à moyenne nulle sur Ω. Le terme visqueux dans (8a) est alors approché à l'aide de la forme bilinéaire A_h du Chapitre 1 appliquée terme à terme sur les composantes des vitesses. Pour le couplage vitesse-pression, on exploite l'interprétation des multiplicateurs de Lagrange comme traces des vitesses. Ainsi, on construit un opérateur de divergence discret $\mathcal{D}_h^k$ qui approche l'opérateur $\boldsymbol{\nabla}\cdot : H^1_0(\Omega)^d \to L^2_0(\Omega)$ et qui satisfait une propriété de commutativité opportune. Cet opérateur est défini dans le même esprit que D_h^k dans (5), c.-à-d. en utilisant une formule d'intégration par parties discrète. Posant $\boldsymbol{W}_h^k := (W_h^k)^d$, la formulation MHO du problème de Stokes consiste à trouver la vitesse $\boldsymbol{u}_h \in \boldsymbol{W}_h^k$ et la pression $p_h \in \mathcal{P}_h^k$ telles que

$$
\begin{aligned}
\boldsymbol{\mathcal{A}}_h((\boldsymbol{u}_h, \boldsymbol{\lambda}_h), (\boldsymbol{v}_h, \boldsymbol{\mu}_h)) - (p_h, \mathcal{D}_h^k \boldsymbol{u}_h) &= L_h(\boldsymbol{v}_h) && \forall (\boldsymbol{v}_h, \boldsymbol{\mu}_h) \in \boldsymbol{W}_h^k,\\
(\mathcal{D}_h^k \boldsymbol{u}_h, q_h) &= 0 && \forall q_h \in \mathcal{P}_h^k,
\end{aligned}
\qquad (9)
$$

où la forme bilinéaire $\boldsymbol{\mathcal{A}}_h$ est construite à partir de A_h en sommant les composantes vectorielles. On justifie alors le caractère bien posé du problème (9) grâce à une condition *inf-sup* automatiquement vérifiée par construction de $\mathcal{D}_h^k$. Une analyse de convergence est présentée en norme d'énergie où l'on retrouve la convergence à l'ordre $(k+1)$ faisant écho aux résultats du Chapitre 1. Enfin, on démontre la superconvergence d'ordre $(k+2)$ en norme L^2 pour la vitesse.

Dans la deuxième partie du Chapitre 2, on considère une extension de la méthode au modèle d'Oseen, prenant en compte les effets d'advection et de réaction. Ce développement est un travail original qui n'a pas été publié ailleurs. Ce modèle s'écrit : Trouver le champ de vitesse $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ et de pression $p : \Omega \to \mathbb{R}$ tels que

$$\begin{aligned} -\nu\triangle\boldsymbol{u} + \boldsymbol{\beta}\cdot\boldsymbol{\nabla}\boldsymbol{u} + \mu\boldsymbol{u} + \boldsymbol{\nabla}p &= \boldsymbol{f} && \text{dans } \Omega, \\ \boldsymbol{\nabla}\cdot\boldsymbol{u} &= 0 && \text{dans } \Omega, \end{aligned} \tag{10}$$

où $\boldsymbol{\beta}$ est un champ de vecteurs fixé (et suffisamment régulier) représentant les effets advectifs, $\nu \geqslant 0$ un scalaire représentant la viscosité cinématique, et $\mu > 0$ un coefficient de réaction. La présence du terme d'advection $\boldsymbol{\beta}\cdot\boldsymbol{\nabla}\boldsymbol{u}$ apporte un lot de difficultés supplémentaires lorsqu'il s'agit d'approcher une solution numériquement. Un problème particulièrement délicat consiste à garantir un bon comportement de la méthode de discrétisation dans toute la plage de valeurs pour le nombre de Péclet, représentant le rapport entre les effets advectifs et diffusifs.

La discrétisation MHO du problème d'Oseen étend les idées développées dans le cas scalaire dans [54]. D'une part, les discrétisations du terme visqueux et du couplage vitesse-pression sont identiques au cas de Stokes (mis à part, bien entendu, le fait que le terme visqueux est ici multiplié par le scalaire ν). Le terme d'advection, d'autre part, est discrétisé en introduisant, pour tout élément du maillage $T \in \mathcal{T}_h$, un opérateur de reconstruction local de dérivée advective $\boldsymbol{G}_{\boldsymbol{\beta},T}^k$ motivé par une formule d'intégration par parties discrète. La formulation HHO du problème (10) s'écrit alors comme : Trouver $\boldsymbol{u}_h \in \boldsymbol{W}_h^k$ et $p_h \in \mathcal{P}_h^k$ tels que

$$\begin{aligned} \boldsymbol{\mathcal{A}}_{\nu,\boldsymbol{\beta},\mu,h}((\boldsymbol{u}_h,\boldsymbol{\lambda}_h),(\boldsymbol{v}_h,\boldsymbol{\mu}_h)) - (p_h,\mathcal{D}_h^k\boldsymbol{u}_h) &= L_h(\boldsymbol{v}_h) && \forall(\boldsymbol{v}_h,\boldsymbol{\mu}_h) \in \boldsymbol{W}_h^k, \\ (\mathcal{D}_h^k\boldsymbol{u}_h, q_h) &= 0 && \forall q_h \in \mathcal{P}_h^k, \end{aligned} \tag{11}$$

avec la bilinéaire $\boldsymbol{\mathcal{A}}_{\nu,\boldsymbol{\beta},\mu,h}$ sur $\boldsymbol{W}_h^k \times \boldsymbol{W}_h^k$ est définie comme

$$\boldsymbol{\mathcal{A}}_{\nu,\boldsymbol{\beta},\mu,h} := \nu\boldsymbol{\mathcal{A}}_h + \boldsymbol{\mathcal{A}}_{\boldsymbol{\beta},\mu,h},$$

où la contribution visqueuse $\boldsymbol{\mathcal{A}}_h$ est la même que dans (9), tandis que la contribution advective-réactive $\boldsymbol{\mathcal{A}}_{\mu,h,\boldsymbol{\beta}}$ est une forme linéaire coercive obtenue à l'aide de l'opérateur $\boldsymbol{G}^k_{\boldsymbol{\beta},T}$ et incorporant un terme de stabilisation en amont. Ce dernier a la particularité d'être défini à l'intérieur de chaque élément en utilisant la différence entre les inconnues de maille et de face. Grâce à la présence de ce terme de stabilisation, le problème discret est stable par construction. L'analyse d'erreur proposée permet d'apprécier les variations dans l'ordre de convergence de la méthode en fonction de la valeur du nombre de Péclet local. Plus précisément, on estime l'erreur en une norme diffusive-advective-réactive entre la solution du schéma HHO et la projection sur $\boldsymbol{W}^k_h \times \mathcal{P}^k_h$ de la solution exacte. Cette étape se fait en estimant la consistance de trois termes dont deux, le terme diffusif $\boldsymbol{\mathcal{A}}_h$ et couplage vitesse-pression, sont établis plus tôt pour le problème de Stokes. La consistance du dernier terme $\boldsymbol{\mathcal{A}}_{\boldsymbol{\beta},\mu,h}$ est obtenue en découpant les intégrales en plusieurs morceaux selon si un nombre de Péclet local Pe_{TF} opportunément défini est $\leqslant 1$ ou > 1. Pour un élément du maillage $T \in \mathcal{T}_h$, on a alors une estimation de l'erreur en $\mathcal{O}(h_T^{k+1/2})$ quand l'advection est dominante et $\mathcal{O}(h_T^{k+1})$ sinon (ce qui est en accord avec le résultats du chapitre précédent).

On termine par une petite section bibliographique. La littérature concernant l'approximation numérique du problème de Stokes est très riche, et on se limitera à quelques travaux en relation avec les développements considérés ici ; pour une introduction on renvoie, p.ex., à [24, 75]. Comme déjà remarqué pour le problème de Poisson, on peut obtenir une discrétisation d'ordre élevé sur maillages généraux à l'aide de méthodes complètement discontinues avec ou sans stabilisation de la pression ; voir, p.ex., [35, 47, 52, 76, 105]. Pour le problème de Stokes, on peut citer des méthodes dG [90] ainsi que des approches HDG [39, 45, 48, 68, 82, 108]. Dans le contexte des VEM, on peut citer [71]. La prise en compte robuste de forces volumiques avec une partie irrotationnelle importante a été considéré dans le contexte des méthodes HHO dans [61]. La littérature pour le problème d'Oseen est, en revanche, moins riche. Pour le problème d'Oseen, des approches dG existent aussi [47], ainsi que HDG [39] ou bien avec des méthodes basées sur des opérateurs définis faiblement [86]. Concernant le traitement robuste des termes d'advection, on citera tout d'abord [58], où on développe une méthode dG pour un problème de diffusion-advection-réaction localement dégénéré valable pour tout nombre de Péclet dans la plage $[0, +\infty]$. Récemment, les idées de ce papier ont été reprises dans le contexte des méthodes HHO dans [54], où l'analyse montre en plus comment l'ordre de convergence des contributions d'erreurs varie en fonction d'un nombre de Péclet local opportunément défini. Ce travail a servi d'inspiration aux développements proposés ici pour le

problème d'Oseen.

Une méthode hp−HHO pour le problème de diffusion variable général

Dans le Chapitre 3 nous nous intéressons à l'analyse hp de la méthode HHO dans le cas de l'équation de diffusion scalaire (1). Il s'agit de mettre en place une formulation HHO du problème permettant d'adapter finesse du maillage et ordre polynomial dans l'optique de converger plus rapidement vers la solution. En effet, il est souvent plus efficace d'augmenter l'ordre polynomial p quand la solution recherchée est très régulière.

Le point de départ est la définition d'espaces de degrés de liberté supportant des ordres d'approximation polynomiales variables. En effet, pour plus de souplesse, on considère le cas où l'ordre n'est plus fixé globalement comme dans les méthodes HHO exposées dans les chapitres précédents, mais localement sur chaque élément. On se donne ainsi un vecteur $\underline{p}_h := (p_F)_{F\in\mathcal{F}_h} \in \mathbb{N}^{\mathcal{F}_h}$ contenant les ordres polynomiaux sur chaque faces (ici, $\mathcal{F}_h$ désigne l'ensemble des faces du maillage $\mathcal{T}_h$). Pour tout élément $T \in \mathcal{T}_h$, notant $\underline{p}_T$ la restriction à $T \in \mathcal{T}_h$ de $\underline{p}_h$, on définit alors l'espace de degrés de liberté local

$$U_T^{\underline{p}_T} := \mathbb{P}^{p_T}(T) \times \left(\bigtimes_{F\in\mathcal{F}_T} \mathbb{P}^{p_F}(F) \right), \qquad p_T := \min_{F\in\mathcal{F}_T} p_F. \tag{12}$$

En suivant la méthodologie HHO, on définit alors un opérateur de reconstruction de potentiel $r_T^{p_T+1} : U_T^{\underline{p}_T} \to \mathbb{P}^{p_T+1}(T)$ et une forme bilinéaire a_T. L'opérateur $r_T^{p_T+1}$ est construit en résolvant un problème de Neumann sur T où la diffusion $\boldsymbol{\kappa}$ est prise en compte, et dont le but est de reproduire au niveau discret une formule d'intégration par parties où le rôle de la fonction dans les intégrales de volumes et de faces est joué par les inconnues de maille et de face, respectivement. La forme bilinéaire a_T est définie sur $U_T^{\underline{p}_T} \times U_T^{\underline{p}_T}$ comme somme d'un terme consistant construit à l'aide de $r_T^{p_T+1}$ et d'un terme de stabilisation :

$$a_T(\underline{u}_T, \underline{v}_T) := (\boldsymbol{\kappa}\boldsymbol{\nabla} r_T^{p_T+1}\underline{u}_T, \boldsymbol{\nabla} r_T^{p_T+1}\underline{v}_T) + \text{stabilisation}.$$

La stabilisation est basée sur une pénalisation des résidus sur les faces au sens des moindres carrés, qui est par construction consistant jusqu'à l'ordre $p_T + 1$. L'espace

de degrés de liberté global $U_h^{\underline{p}_h}$ est construit afin de définir une forme bilinéaire a_h globale sur $U_h^{\underline{p}_h} \times U_h^{\underline{p}_h}$ par somme des contributions locales. La forme linéaire l_h définie sur $U_h^{\underline{p}_h}$ prenant en compte le second membre permet enfin de poser le problème $hp-$HHO sous la forme : Trouver $\underline{u}_h \in U_h^{\underline{p}_h}$ tel que

$$a_h(\underline{u}_h, \underline{v}_h) = l_h(\underline{v}_h) \quad \forall \underline{v}_h \in U_h^{\underline{p}_h}.$$

On montre que le problème est bien posé et une analyse de convergence hp en norme d'énergie et L^2 est présentée, dans l'esprit de Babuška et Suri [11]. De plus, on montre que l'analyse d'erreur est robuste vis-à-vis du coefficient de diffusion, avec une constante multiplicative proportionnelle à la racine carrée du rapport d'anisotropie local quand on considère la norme d'énergie. En norme d'énergie on retrouve la convergence classique des méthodes HHO lorsqu'on considère le raffinement en h, tandis qu'on a un résultat plus standard en $(p_T + 1)^{-p_T}$ lorsqu'on considère le raffinement en p. On notera que ce résultat est comparable aux meilleurs résultats obtenus pour les méthodes dG basées sur des polynômes de degré k et *sur des maillages rectangulaires* (un demi-ordre de convergence est perdu sur des maillages plus généraux). Le chapitre sera clôturé par des tests numériques, sur différents maillages, venant valider les estimations d'erreurs présentées.

Terminons par un point bibliographique. La littérature sur le sujet étant assez vaste, la liste de références donnée n'est qu'une sélection basée sur des critères de proximité avec les résultats présentés. Les premières contributions sur les méthodes éléments finis $p-$ et $hp-$ conformes sur maillages standards remontent au début des années 80 grâce aux travaux de Babuška et Suri ; voir [11–13]. Ces travaux portent essentiellement sur des méthodes conformes. Peu après, à la fin des années 90, des méthodes hp non-conformes sur maillages standards voient le jour dans le cadre des problèmes elliptiques [73,91,95,101]. Dans le cadre des problèmes liés à la mécanique des fluides, on peut citer les méthodes hp-FEM [100]. La possibilité de raffiner à la fois en h et en p sur des maillages généraux est une direction de recherche bien plus récente, voir par exemple les méthodes $hp-$composite [5, 74] ou Galerkin discontinue polyhédrique [36]. Enfin, plus récemment, une version hp de la méthode VEM est présentée dans [20]. Bien que ces méthodes présentent en général des tailles de systèmes plus faibles, elles nécessitent des fonctions de bases et formules de quadrature adaptées. L'implémentation pratique en est donc la principale difficulté. Dans la plupart des cas, on rencontre une décroissance polynomiale lors d'un raffinement en h et une décroissance exponentielle lors d'un raffinement en p. Dans le cas des

EDPs elliptiques du second ordre on observe souvent une estimation de la forme

$$\|u - u_{hp}\| \leqslant Ch^{\min(k,p+1)-1}p^{1-k}\|u\|_{H^k},$$

où u est la solution et u_{hp} son approximation par une méthode hp.

Perspectives sur la réduction de l'équation de diffusion paramétrée

Le Chapitre 4 présente quelques éléments d'un travail en cours concernant l'approximation numérique rapide d'une EDP elliptique paramétrée pour un grand nombre de valeurs du paramètre. La nécessité de calculer de nombreuses fois les solutions d'une EDP paramétrée se rencontre, par exemple, en optimisation, ou en quantification d'incertitudes (le paramètre restant de dimension faible en pratique). Les méthodes de discrétisation directes (éléments finis ou HHO/MHO) étant souvent trop coûteuses dans ce contexte, on peut avoir recours à des méthodes réduites, dont la méthode des Bases Réduites (BR) constitue un exemple important.

Sous l'hypothèse qu'une méthode numérique précise existe pour toute valeur du paramètre, la méthode BR construit un modèle réduit après une période d'apprentissage :

(i) dans un premier temps (phase *offline*) on calcule des solutions de l'EDP pour un ensemble bien choisi de valeurs du paramètre à l'aide d'une méthode numérique précise standard (cette étape préalable est en général coûteuse) ;

(ii) dans un deuxième temps (phase *online*) on peut obtenir la solution de l'EDP pour toute valeur du paramètre comme la projection de Galerkin sur l'espace des solutions pré-calculées.

Si dans la phase *offline* on a aussi pu identifier comment assembler rapidement la projection de Galerkin pour toute valeur du paramètre, alors cette dernière, utilisée *online*, constitue le modèle réduit du problème.

Le succès de la méthode BR repose sur

(i) la régularité paramétrique des solutions en fonction de l'espace de Hilbert choisi pour y plonger les quelques solutions précises calculées et construire le projecteur (hérité de la formulation variationnelle du problème considérée) ;

(ii) le calcul rapide du projecteur, qui permet par exemple d'implémenter la méthode pour des applications en temps réel ;

(iii) éventuellement, le calcul rapide de la base réduite, qui permet par exemple d'échapper au fléau de la dimension dans un certain nombre de cas pratiques.

Les deux derniers points ont fait l'objet de nombreuses publications pour des applications pratiques assez variées (il n'est pas possible de toutes les citer ici : on mentionne par exemple [1, 16, 28–30, 40, 77, 78, 81, 92, 93, 97, 98]). Mais le premier point (d'ordre plus théorique, relevant de l'analyse mathématique) a rarement été discuté dans la littérature autrement qu'implicitement, en constatant a posteriori l'efficacité d'un algorithme BR dans des cas particuliers. Quelques travaux importants [22,34,96] ont pu obtenir un lien a priori entre le succès observé de la méthode BR (en fait, d'une instance de la méthode dans un Hilbert donné) et un concept de régularité paramétrique utilisé implicitement pour construire la base réduite dans la méthode BR (les épaisseurs de Kolmogorov, dont l'algorithme glouton utilisé pour construire la base permet de calculer une approximation par borne supérieure). Les théorèmes généraux obtenus dans [22,34,96] restent toutefois non-totalement quantifiables dans des cas particuliers et ne permettent pas d'expliquer complétement le succès pratique constaté pour la méthode BR.

Nous considérons ici l'équation de diffusion (1) avec coefficient de diffusion $\boldsymbol{\kappa}$ variable, et nous nous plaçons dans le cadre où $\boldsymbol{\kappa}$ varie *lentement*, de telle sorte qu'on peut synthétiser ses variations avec peu de termes dans une somme de produits de fonctions régulières de l'espace et d'un paramètre $\boldsymbol{\mu}$ de faible dimension. Dans le cadre Hilbertien de notre exemple, les solutions pourront être calculées précisément pour tout $\boldsymbol{\mu}$ pourvu qu'elles soient toutes suffisamment proches d'un sous-espace affine (au sens de la distance donnée par la norme Hilbertienne), et pourvu qu'on sache en calculer rapidement la projection. La question que nous nous posons ici est s'il existe une différence entre deux instances de la méthode BR pour le même problème quand il est plongé dans deux espaces de Hilbert différents, en ce qui concerne la régularité paramétrique.

Le Chapitre 4 apporte une première réponse à cette question par le biais de l'investigation numérique précise d'un cas particulier. La conclusion (partielle) obtenue dans ce cadre semble indiquer que l'instance de la méthode BR utilisant le projecteur dans $H^1(\Omega)$ hérité de la formulation primale, qui est usuellement utilisée pour le problème en question, n'est pas toujours optimale par rapport à la dimension de l'espace réduit. Précisément, quand le terme source est dans $L^2(\Omega)$, alors on peut améliorer les performances de la réduction (en terme d'erreur en fonction de la di-

mension d'espace réduit) en utilisant l'instance de la méthode BR avec le projecteur hérité de la formulation mixte. Si l'explication théorique de [22,34,96] s'applique bien à notre exemple numérique, alors notre résultat indique en outre que les épaisseurs de Kolmogorov de l'espace des solutions paramétrées pour notre problème décroissent plus vite dans $\mathbf{H}(\mathrm{div};\Omega) \times L^2(\Omega)$ que dans $H^1(\Omega)$. Cette conjecture est un élément nouveau en théorie des EDPs à notre connaissance.

Chapter 1

Hybridization of the Mixed High-Order method

The material contained in this chapter is mainly taken from [2]. Let $\Omega \subset \mathbb{R}^d$, $d \geqslant 1$, be an open, bounded, connected polytopal set. For any open, connected subset $X \subset \overline{\Omega}$ with non-zero Lebesgue measure, the standard inner product and norm of the Lebesgue space $L^2(X)$ are denoted by $(\cdot,\cdot)_X$ and $\|\cdot\|_X$, respectively, with the convention that the index is omitted if $X = \Omega$. Similarly, the classical Sobolev spaces are denoted $H^m(X)$ for $m \geqslant 1$. For a given $f \in L^2(\Omega)$, we consider here the Poisson equation that consists in finding a scalar-valued field $u : \Omega \to \mathbb{R}$ such that

$$\begin{aligned} -\triangle u &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.1}$$

Other boundary conditions could be considered, but we stick to the homogeneous Dirichlet case for the sake of simplicity. Let $W := H^1_0(\Omega)$. A classical primal weak formulation of problem (1.1) consists in seeking $u \in W$ such that it holds

$$(\boldsymbol{\nabla} u, \boldsymbol{\nabla} v) = (f, v) \quad \forall v \in W. \tag{1.2}$$

Problem (1.1) can be alternatively be reformulated as a system of first-order PDEs:

$$\begin{aligned} \boldsymbol{s} &= \boldsymbol{\nabla} u \qquad \text{in } \Omega, \\ \boldsymbol{\nabla}\cdot\boldsymbol{s} &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.3}$$

Letting

$$\boldsymbol{\Sigma} := \mathbf{H}(\mathrm{div};\Omega), \qquad U := L^2(\Omega), \tag{1.4}$$

an often used variational formulation inspired by (1.3) reads: Find $(\boldsymbol{s}, u) \in \boldsymbol{\Sigma} \times U$ such that

$$\begin{aligned} (\boldsymbol{s},\boldsymbol{t}) + (u, \boldsymbol{\nabla}\cdot\boldsymbol{t}) &= 0 \qquad && \forall \boldsymbol{t} \in \boldsymbol{\Sigma}, \\ -(\boldsymbol{\nabla}\cdot\boldsymbol{s}, v) &= (f, v) \qquad && \forall v \in U. \end{aligned} \tag{1.5}$$

The unknowns $\boldsymbol{s}$ and u will be henceforth referred to as the *flux* and *potential*, respectively. The formulation (1.2) where the potential u is the sole unknown will be referred to as *primal*, whereas the formulation (1.5) where both $\boldsymbol{s}$ and u appear as unknowns will be referred to as *mixed*.

At the continuous level, the primal formulation can be recovered from the mixed formulation by eliminating the flux $\boldsymbol{s}$. A discrete counterpart of this procedure is studied here for the Mixed High-Order (MHO) method of [57]. Thanks to the peculiar structure of the MHO method, the elimination of the flux degrees of freedom (DOFs) can be carried out at the local level by solving a small coercive problem inside each element. This procedure, classically referred to as *hybridization*, consists in two steps. In the first step, we decouple interface flux DOFs and introduce Lagrange multipliers to enforce their single-valuedness. In the second step, we locally eliminate the flux DOFs to end up with an equivalent coercive problem where the DOFs are the original potential unknowns and the Lagrange multipliers (which can be alternatively interpreted as traces of the discrete potential over faces). The former can be further eliminated at the local level by *static condensation*. This coercive reformulation is clearly to be preferred in the practical implementation, as symmetric positive-definite linear systems are much easier to solve numerically than linear systems with a saddle-point structure. An important side result of the hybridization is that we can establish a link with the Hybrid High-Order (HHO) method of [59].

This chapter is organized as follows. In Section 1.1 we introduce the notion of admissible mesh sequence and recall some known basic results. In Section 1.2 we recall the MHO method. In Section 1.3 we state its mixed hybrid reformulation containing all possible flux and potential DOFs. In Section 1.4 we show how to locally eliminate flux DOFs and state the equivalent primal reformulation of the MHO method. An error analysis is carried out in Section 1.5, where we derive optimal error estimates for both the energy- and L^2-norms of the error. The extension to the Darcy problem with a variable diffusion coefficient is briefly addressed in Section 1.6.

1.1 Discrete setting

In this section we introduce the notion of admissible mesh sequences and recall a few known functional analysis results from [53, 55].

1.1.1 Admissible meshes

Denoting by $\mathcal{H} \subset \mathbb{R}_*^+$ a countable set of *meshsizes* having 0 as its unique accumulation point, we consider mesh sequences $(\mathcal{T}_h)_{h\in\mathcal{H}}$ where, for all $h \in \mathcal{H}$, $\mathcal{T}_h = \{T\}$ is a finite collection of nonempty disjoint open polyhedra T (called *elements* or *cells*) such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} \overline{T}$ and $h = \max_{T\in\mathcal{T}_h} h_T$ (h_T stands for the diameter of T).

A hyperplanar closed connected subset F of $\overline{\Omega}$ is called a *face* if it has positive $(d-1)$-dimensional measure and (i) either there exist distinct $T_1, T_2 \in \mathcal{T}_h$ such that $F \subset \partial T_1 \cap \partial T_2$ (and F is an *interface*) or (ii) there exists $T \in \mathcal{T}_h$ such that $F \subset \partial T \cap \partial\Omega$ (and F is a *boundary face*). The set of interfaces is denoted by $\mathcal{F}_h^{\mathrm{i}}$, the set of boundary faces by $\mathcal{F}_h^{\mathrm{b}}$, and we let $\mathcal{F}_h := \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\mathrm{b}}$. The diameter of a face $F \in \mathcal{F}_h$ is denoted by h_F.

For all $T \in \mathcal{T}_h$, we let $\mathcal{F}_T := \{F \in \mathcal{F}_h \mid F \subset \partial T\}$ denote the set of faces lying on the boundary of T. Symmetrically, for all $F \in \mathcal{F}_h$, $\mathcal{T}_F := \{T \in \mathcal{T}_h \mid F \subset \partial T\}$ is the set of the one (if F is a boundary face) or two (if F is an interface) elements sharing F.

For all $F \in \mathcal{F}_T$, we denote by $\boldsymbol{n}_{TF}$ the normal to F pointing out of T. For every interface $F \subset \partial T_1 \cap \partial T_2$, we adopt the following convention: an orientation is fixed once and for all by means of a unit normal vector $\boldsymbol{n}_F$, and the elements T_1 and T_2 are numbered so that $\boldsymbol{n}_F := \boldsymbol{n}_{T_1F}$.

We assume throughout the rest of this work that the mesh sequence $(\mathcal{T}_h)_{\mathcal{H}}$ is *admissible* in the sense of [55, Chapter 1].

Definition 1.1.1 (Admissible mesh sequence)**.** For all $h \in \mathcal{H}$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathfrak{T}_h$ and the following properties hold for all $h \in \mathcal{H}$ with mesh regularity parameter $\varrho > 0$ independent of h:

(i) for all simplex $S \in \mathfrak{T}_h$ of diameter h_S and inradius r_S, $\varrho h_S \leqslant r_S$;

(ii) for all $T \in \mathcal{T}_h$, and all $S \in \mathfrak{T}_T := \{S \in \mathfrak{T}_h \mid S \subset T\}$, $\varrho h_T \leqslant h_S$.

For an admissible mesh sequence, it is known from [55, Lemma 1.41] that the number

of faces of one element can be bounded uniformly in h, i.e., it holds that

$$\forall h \in \mathcal{H}, \qquad \max_{T\in\mathcal{T}_h}\big\{\mathfrak{N}_T := \mathrm{card}(\mathcal{F}_T)\big\} \leqslant \mathfrak{N}_\partial, \tag{1.6}$$

for an integer $(d+1) \leqslant \mathfrak{N}_\partial < +\infty$ depending on ϱ but independent of h. Furthermore, for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$ and all $F \in \mathcal{F}_T$, h_F is uniformly comparable to h_T in the following sense (cf. [55, Lemma 1.42]):

$$\rho^2 h_T \leqslant h_F \leqslant h_T.$$

1.1.2 Basic results

Let an integer $k \geqslant 0$ be fixed. Let X denote either a mesh element in $\mathcal{T}_h$ or a mesh face in $\mathcal{F}_h$. We denote by $\mathbb{P}^k(X)$ the space spanned by the restriction to X of d-variate polynomials of total degree k. We introduce the L^2-orthogonal projector $\pi_X^k : L^2(X) \to \mathbb{P}^k(X)$ such that, for all $v \in L^2(X)$,

$$(\pi_X^k v - v, w)_X = 0 \qquad \forall w \in \mathbb{P}^k(X).$$

The following trace and inverse inequalities hold for all $T \in \mathcal{T}_h$ and all $v \in \mathbb{P}^k(T)$:

$$\|v\|_F \leqslant C_{\mathrm{tr}} h_F^{-1/2} \|v\|_T \qquad \forall F \in \mathcal{F}_T, \tag{1.7}$$

$$\|\boldsymbol{\nabla} v\|_T \leqslant C_{\mathrm{inv}} h_T^{-1} \|v\|_T, \tag{1.8}$$

with real numbers C_{tr} and C_{inv} that are independent of $h \in \mathcal{H}$ (but depending on ϱ), cf. [55, Lemmata 1.44 and 1.46].

It follows from [53, Lemmas 3.4 and 3.6] that there exists a real number C_{app} depending on ϱ but independent of h such that, for all $T \in \mathcal{T}_h$, the L^2-orthogonal projector π_T^k on $\mathbb{P}^k(T)$ satisfies: For all $s \in \{1, \ldots, k+1\}$, and all $v \in H^s(T)$,

$$|v - \pi_T^k v|_{H^m(T)} + h_T^{1/2} |v - \pi_T^k v|_{H^m(\partial T)} \leqslant C_{\mathrm{app}} h_T^{s-m} |v|_{H^s(T)} \qquad \forall m \in \{0, \ldots, s\}. \tag{1.9}$$

At the global level, we introduce the broken polynomial space

$$\mathbb{P}^k(\mathcal{T}_h) := \big\{v \in L^2(\Omega) \mid v_{|T} \in \mathbb{P}^k(T) \quad \forall T \in \mathcal{T}_h\big\}, \tag{1.10}$$

on which we define the L^2-orthogonal operator $\pi_h^k : L^2(\Omega) \to \mathbb{P}^k(\mathcal{T}_h)$ defined such

that for $v \in L^2(\Omega)$,

$$(\pi_h^k v - v, w) = 0 \qquad \forall w \in \mathbb{P}^k(\mathcal{T}_h). \tag{1.11}$$

Clearly, for all $v \in L^2(\Omega)$, and all $T \in \mathcal{T}_h$, it holds that $\pi_T^k(v_{|T}) = (\pi_h^k v)_{|T}$, and optimal approximation properties for π_h^k follow from (1.9). The regularity assumptions on the exact solution are expressed in terms of the normed broken Sobolev spaces on $\mathcal{T}_h$,

$$H^m(\mathcal{T}_h) := \{v \in L^2(\Omega), v_{|T} \in H^m(T)\}, \quad m \geqslant 1$$

with norm $\|v\|_{H^m(\mathcal{T}_h)} := \big(\sum_{T\in\mathcal{T}_h} \|v_{|T}\|^2_{H^m(T)}\big)^{1/2}$.

Finally, we recall the following local Poincaré inequality valid for all $T \in \mathcal{T}_h$:

$$\|v - \pi_T^k v\|_T \leqslant C_{\mathrm{P}} h_T \|\boldsymbol{\nabla} v\|_T, \qquad \forall v \in H^1(T), \tag{1.12}$$

where C_{P} is independent of h but possibly depends on ϱ ($C_{\mathrm{P}} = \pi^{-1}$ for convex elements [18]).

1.2 The Mixed High-Order method

In this section, we recall the MHO method of [57] as well as a few known results that will be useful for the subsequent discussion.

1.2.1 Degrees of freedom and discrete spaces

For a given fixed integer $k \geqslant 0$, we define the following local polynomial spaces attached to mesh elements and faces, respectively:

$$\mathbb{T}_T^k := \boldsymbol{\nabla}\mathbb{P}^k(T) \quad \forall T \in \mathcal{T}_h, \qquad \mathbb{F}_F^k := \mathbb{P}^k(F) \quad \forall F \in \mathcal{F}_h. \tag{1.13}$$

Notice that, in the lowest-order case, we have $\mathbb{T}_T^0 = \{\mathbf{0}\}$, which reflects the fact that element DOFs are unnecessary. The local and global DOF spaces for the flux approximation are, respectively,

$$\boldsymbol{\Sigma}_T^k := \mathbb{T}_T^k \times \left\{ \bigtimes_{F\in\mathcal{F}_T} \mathbb{F}_F^k \right\} \quad \forall T \in \mathcal{T}_h \qquad \text{and} \qquad \check{\boldsymbol{\Sigma}}_h^k := \bigtimes_{T\in\mathcal{T}_h} \boldsymbol{\Sigma}_T^k. \tag{1.14}$$

Figure 1.1 depicts the degrees of freedom for different values of the polynomial degree

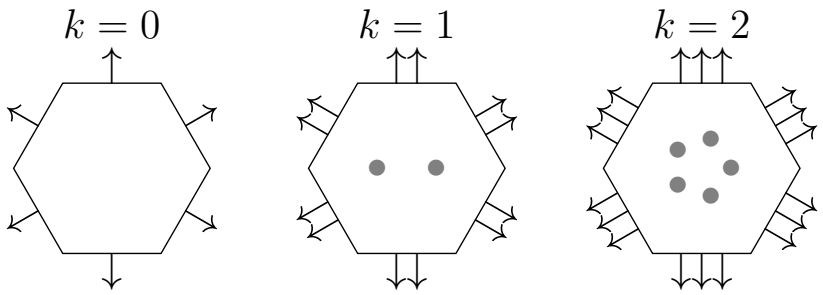


Figure 1.1 – Local DOF space for the flux $\boldsymbol{\Sigma}_T^k$ for $k = 0, 1, 2$

k. We also introduce the following patched version of $\check{\boldsymbol{\Sigma}}_h^k$:

$$\boldsymbol{\Sigma}_h^k := \left\{ \boldsymbol{\tau}_h = (\boldsymbol{\tau}_T, (\tau_{TF})_{F\in\mathcal{F}_T})_{T\in\mathcal{T}_h} \in \check{\boldsymbol{\Sigma}}_h^k \mid \textstyle\sum_{T\in\mathcal{T}_F} \tau_{TF} = 0 \quad \forall F \in \mathcal{F}_h^{\mathrm{i}} \right\}. \tag{1.15}$$

The local space $\boldsymbol{\Sigma}_T^k$ is equipped with the following local $L^2(T)$-like norm:

$$\forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k, \qquad |||\boldsymbol{\tau}|||_T^2 := \|\boldsymbol{\tau}_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F \|\tau_{TF}\|_F^2. \tag{1.16}$$

The scaling coefficient h_F in the second term ensures that the addends are dimensionally homogeneous. For all $T \in \mathcal{T}_h$, we denote by $R_{\boldsymbol{\Sigma},T}^k : \check{\boldsymbol{\Sigma}}_h^k \to \boldsymbol{\Sigma}_T^k$ the restriction operator which realizes the mapping between global and local flux DOFs, and we equip $\check{\boldsymbol{\Sigma}}_h^k$ (hence also $\boldsymbol{\Sigma}_h^k$) with the global $L^2(\Omega)$-like norm $|||\cdot|||$ such that, for all $\boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k$,

$$|||\boldsymbol{\tau}_h|||^2 := \sum_{T\in\mathcal{T}_F} |||R_{\boldsymbol{\Sigma},T}^k \boldsymbol{\tau}_h|||_T^2. \tag{1.17}$$

Let, for a fixed $s > 2$,

$$\boldsymbol{\Sigma}^+(T) := \{\boldsymbol{t} \in L^s(T)^d \mid \boldsymbol{\nabla}\cdot\boldsymbol{t} \in L^2(T)\}.$$

The regularity in $\boldsymbol{\Sigma}^+(T)$ guarantees that all terms below are well-defined. We introduce the local interpolator $I_{\boldsymbol{\Sigma},T}^k : \boldsymbol{\Sigma}^+(T) \to \boldsymbol{\Sigma}_T^k$ such that, for all $\boldsymbol{t} \in \boldsymbol{\Sigma}^+(T)$, $I_{\boldsymbol{\Sigma},T}^k \boldsymbol{t} = (\boldsymbol{\tau}_T, (\tau_{TF})_{F\in\mathcal{F}_T})$ with

$$\boldsymbol{\tau}_T = \varpi_T^k \boldsymbol{t}, \qquad \tau_{TF} = \pi_F^k(\boldsymbol{t}\cdot\boldsymbol{n}_{TF}) \quad \forall F \in \mathcal{F}_T, \tag{1.18}$$

where π_F^k denotes the L^2-orthogonal projector on $\mathbb{P}^k(F)$ while ϖ_T^k denotes the L^2-

orthogonal projector on $\mathbb{T}_T^k$ (in fact, an elliptic projector on $\mathbb{P}^k(T)$) such that

$$(\varpi_T^k \boldsymbol{t}, \boldsymbol{w})_T = (\boldsymbol{t}, \boldsymbol{w})_T \qquad \forall \boldsymbol{w} \in \mathbb{T}_T^k.$$

Note that the polynomial $w \in \mathbb{P}^{k+1}(T)$ such that $\varpi_T^k \boldsymbol{t} = \boldsymbol{\nabla} w$ is only defined up to a constant. The choice of this constant has, however, no effect on the following discussion. The global interpolator $I_{\boldsymbol{\Sigma},h}^k : \boldsymbol{\Sigma}^+ \to \check{\boldsymbol{\Sigma}}_h^k$ with

$$\boldsymbol{\Sigma}^+ := \{\boldsymbol{t} \in L^2(\Omega)^d \mid \boldsymbol{t}_{|T} \in \boldsymbol{\Sigma}^+(T),\ \forall T \in \mathcal{T}_h\}$$

is such that, for all $\boldsymbol{t} \in \boldsymbol{\Sigma}^+$,

$$R_{\boldsymbol{\Sigma},T}^k I_{\boldsymbol{\Sigma},h}^k \boldsymbol{t} = I_{\boldsymbol{\Sigma},T}^k \boldsymbol{t}_{|T} \qquad \forall T \in \mathcal{T}_h. \tag{1.19}$$

Remark 1.2.1 (Restriction of $I_{\boldsymbol{\Sigma},h}^k$ to $\boldsymbol{\Sigma}^+ \cap \mathbf{H}(\mathrm{div};\Omega)$). We observe that functions in $\boldsymbol{\Sigma}^+ \cap \mathbf{H}(\mathrm{div};\Omega)$ are mapped by $I_{\boldsymbol{\Sigma},h}^k$ to elements of the patched space $\boldsymbol{\Sigma}_h^k$, cf. (1.15).

The local and global DOF spaces for the potential are given by, respectively,

$$U_T^k := \mathbb{P}^k(T) \quad \forall T \in \mathcal{T}_h \qquad \text{and} \qquad U_h^k := \bigtimes_{T \in \mathcal{T}_h} U_T^k. \tag{1.20}$$

In what follows, we identify when needed the space U_h^k with the broken polynomial space $\mathbb{P}^k(\mathcal{T}_h)$ defined by (1.10). Both U_T^k and U_h^k are naturally endowed with the L^2-norm topology.

1.2.2 Divergence reconstruction

From this section till the end of the chapter, we use the notation $a \lesssim b$ for the inequality $a \leqslant Cb$ with real number $C > 0$ independent of the meshsize h. Let a element $T \in \mathcal{T}_h$ be fixed. We define the local discrete divergence operator $D_T^k : \boldsymbol{\Sigma}_T^k \to \mathbb{P}^k(T)$ such that, for all $\boldsymbol{\tau} = (\boldsymbol{\tau}_T, (\tau_{TF})_{F \in \mathcal{F}_T}) \in \boldsymbol{\Sigma}_T^k$,

$$(D_T^k \boldsymbol{\tau}, v)_T = -(\boldsymbol{\tau}_T, \boldsymbol{\nabla} v)_T + \sum_{F \in \mathcal{F}_T} (\tau_{TF}, v)_F, \qquad \forall v \in \mathbb{P}^k(T). \tag{1.21}$$

The right-hand side of (1.21) mimicks an integration by parts formula where the role of the function in volumetric and boundary integrals is played by element-based and face-based DOFs, respectively. This choice warrants the following commuting

property for the operator D_T^k:

$$D_T^k(I_{\boldsymbol{\Sigma},T}^k\boldsymbol{t}) = \pi_T^k(\boldsymbol{\nabla}\cdot\boldsymbol{t}) \qquad \forall \boldsymbol{t}\in\boldsymbol{\Sigma}^+(T), \tag{1.22}$$

where π_T^k denotes the L^2-orthogonal projector on $\mathbb{P}^k(T)$ such that, for all $q\in L^1(T)$,

$$(\pi_T^k q - q, r)_T = 0 \qquad \forall r\in\mathbb{P}^k(T).$$

To prove (1.22), let $\boldsymbol{t}\in\boldsymbol{\Sigma}^+(T)$ and observe that it holds for all $v\in\mathbb{P}^k(T)$,

$$\begin{aligned}(D_T^k(I_{\boldsymbol{\Sigma},T}^k\boldsymbol{t}), v)_T &= -(\varpi_T^k\boldsymbol{t}, \boldsymbol{\nabla}v)_T + \sum_{F\in\mathcal{F}_T}(\pi_F^k(\boldsymbol{t}\cdot\boldsymbol{n}_{TF}), v)_F\\ &= -(\boldsymbol{t}, \boldsymbol{\nabla}v)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{t}\cdot\boldsymbol{n}_{TF}, v)_F = (\boldsymbol{\nabla}\cdot\boldsymbol{t}, v)_T,\end{aligned}$$

where we have used the definitions of ϖ_T^k and π_F^k to pass to the second line and an integration by parts to conclude. We note the following inverse estimate, which will be needed later on in this chapter.

Proposition 1.2.2 (Inverse estimate for D_T^k)**.** *There exists a real number $C>0$ independent of h but possibly depending on the mesh regularity parameter ϱ such that, for all $T\in\mathcal{T}_h$ and all $\boldsymbol{\tau}\in\boldsymbol{\Sigma}_T^k$,*

$$\|D_T^k\boldsymbol{\tau}\|_T \leqslant Ch_T^{-1}|||\boldsymbol{\tau}|||_T. \tag{1.23}$$

Proof. Recalling (1.21) we have, for all $\boldsymbol{\tau}\in\boldsymbol{\Sigma}_T^k$,

$$\|D_T^k\boldsymbol{\tau}\|_T = \sup_{v\in\mathbb{P}^k(T),\|v\|_T=1}\left\{-(\boldsymbol{\nabla}v, \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T}(v, \tau_{TF})_F\right\}. \tag{1.24}$$

Using the Cauchy–Schwarz inequality followed by the discrete inverse inequality (1.8), it is inferred that

$$|(\boldsymbol{\nabla}v, \boldsymbol{\tau}_T)_T| \lesssim h_T^{-1}\|v\|_T\|\boldsymbol{\tau}_T\|_T.$$

Again the Cauchy–Schwarz inequality together with the discrete trace inequality (1.7) yields, for all $F\in\mathcal{F}_T$,

$$|(v, \tau_{TF})_F| \lesssim h_F^{-1}\|v\|_T h_F^{1/2}\|\tau_{TF}\|_F.$$

Inequality (1.23) follows using the discrete Cauchy–Schwarz inequality together with the above bounds to estimate the right-hand side of (1.24) and recalling the defini-

tion (1.16) of the $|||\cdot|||_T$-norm $\square$

In what follows, we will also need the global divergence operator $D_h^k : \check{\boldsymbol{\Sigma}}_h^k \to \mathbb{P}^k(\mathcal{T}_h)$ such that, for all $\boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k$ and all $T \in \mathcal{T}_h$,

$$(D_h^k \boldsymbol{\tau}_h)_{|T} = D_T^k R_{\boldsymbol{\Sigma},T}^k \boldsymbol{\tau}_h. \tag{1.25}$$

Using the definition (1.19) of the global interpolator $I_{\boldsymbol{\Sigma},h}^k$ together with the commuting property (1.22) for the local divergence operator, the following global commuting property follows

$$D_h^k(I_{\boldsymbol{\Sigma},h}^k \boldsymbol{t}) = \pi_h^k(\boldsymbol{\nabla}_h \cdot \boldsymbol{t}), \qquad \forall \boldsymbol{t} \in \boldsymbol{\Sigma}^+, \tag{1.26}$$

where $\boldsymbol{\nabla}_h \cdot$ denotes the broken divergence operator on $\mathcal{T}_h$ and π_h^k the L^2-orthogonal projector on $\mathbb{P}^k(\mathcal{T}_h)$, the space of broken (fully discontinuous) polynomials of degree k on $\mathcal{T}_h$.

1.2.3 Flux reconstruction

Let $T \in \mathcal{T}_h$. We next introduce the flux reconstruction operator $\mathfrak{C}_T^k : \boldsymbol{\Sigma}_T^k \to \boldsymbol{\nabla}\mathbb{P}^{k+1}(T)$ such that, for all $\boldsymbol{\tau} = (\boldsymbol{\tau}_T, (\tau_{TF})_{F\in\mathcal{F}_T}) \in \boldsymbol{\Sigma}_T^k$ and all $w \in \mathbb{P}^{k+1}(T)$,

$$(\mathfrak{C}_T^k \boldsymbol{\tau}, \boldsymbol{\nabla} w)_T = -(D_T^k \boldsymbol{\tau}, w)_T + \sum_{F\in\mathcal{F}_T} (\tau_{TF}, w)_F \tag{1.27a}$$

$$= (\boldsymbol{\tau}_T, \boldsymbol{\nabla}\pi_T^k w)_T + \sum_{F\in\mathcal{F}_T} (\tau_{TF}, \pi_F^k w - \pi_T^k w)_F, \tag{1.27b}$$

where we have used (1.21) to pass to the second line. Computing $y \in \mathbb{P}^{k+1}(T)$ such that $\mathfrak{C}_T^k \boldsymbol{\tau} = \boldsymbol{\nabla} y$ and (1.27) holds requires to solve a well-posed Neumann problem for which the usual compatibility condition on the right-hand side is verified. The following polynomial consistency property for $\mathfrak{C}_T^k$ is an immediate consequence of (1.27a) using the commuting property (1.22) for the first term in the right-hand side and he definition of π_F^k for the second:

$$\mathfrak{C}_T^k I_{\boldsymbol{\Sigma},T}^k \boldsymbol{\nabla} w = \boldsymbol{\nabla} w, \qquad \forall w \in \mathbb{P}^{k+1}(T). \tag{1.28}$$

Recalling [57, Lemma 3] and using (1.23), we also have continuity and partial coercivity in the following sense: For all $\boldsymbol{\tau} = (\boldsymbol{\tau}_T, (\tau_{TF})_{F\in\mathcal{F}_T}) \in \boldsymbol{\Sigma}_T^k$,

$$\|\boldsymbol{\tau}_T\|_T \leqslant \|\mathfrak{C}_T^k \boldsymbol{\tau}\|_T \leqslant |||\boldsymbol{\tau}|||_T. \tag{1.29}$$

1.2.4 The Mixed High-Order method

We let H_h denote a global bilinear form on $\check{\boldsymbol{\Sigma}}_h^k \times \check{\boldsymbol{\Sigma}}_h^k$ assembled element-wise from local contributions,

$$H_h(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) := \sum_{T\in\mathcal{T}_h} H_T(R^k_{\boldsymbol{\Sigma},T}\boldsymbol{\sigma}_h, R^k_{\boldsymbol{\Sigma},T}\boldsymbol{\tau}_h),$$

where, for all $T \in \mathcal{T}_h$, the bilinear form H_T on $\boldsymbol{\Sigma}_T^k \times \boldsymbol{\Sigma}_T^k$ is such that, for all $\boldsymbol{\sigma}, \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k$,

$$H_T(\boldsymbol{\sigma}, \boldsymbol{\tau}) := (\mathfrak{C}_T^k\boldsymbol{\sigma}, \mathfrak{C}_T^k\boldsymbol{\tau})_T + J_T(\boldsymbol{\sigma}, \boldsymbol{\tau}), \tag{1.30}$$

with local stabilization bilinear form J_T matching the following assumptions:

(H1) *Symmetry, nonnegativity and polynomial consistency.* J_T is symmetric, positive semi-definite, and it satisfies the following polynomial consistency condition:

$$\forall w \in \mathbb{P}^{k+1}(T), \qquad J_T(I^k_{\boldsymbol{\Sigma},T}\boldsymbol{\nabla} w, \boldsymbol{\tau}) = 0 \qquad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k. \tag{1.31}$$

(H2) *Stability and continuity.* There exists a real number $\eta > 0$ independent of h and of T such that H_T is coercive on $\ker(D_T^k)$ and continuous on $\boldsymbol{\Sigma}_T^k$:

$$\eta|||\boldsymbol{\tau}|||_T^2 \leqslant H_T(\boldsymbol{\tau}, \boldsymbol{\tau}) \qquad \forall \boldsymbol{\tau} \in \ker(D_T^k)\,, \tag{1.32a}$$

$$H_T(\boldsymbol{\tau}, \boldsymbol{\tau}) \leqslant \eta^{-1}|||\boldsymbol{\tau}|||_T^2 \qquad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k\,. \tag{1.32b}$$

Remark 1.2.3 (Condition (1.32b)). In view of (1.30) and of the second inequality in (1.29), and since J_T is symmetric and positive semi-definite owing to **(H1)**, condition (1.32b) holds if and only if there is a real number $C > 0$ independent of h such that, for all $T \in \mathcal{T}_h$,

$$J_T(\boldsymbol{\tau}, \boldsymbol{\tau}) \leqslant C|||\boldsymbol{\tau}|||_T^2 \qquad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k. \tag{1.33}$$

An example of stabilization bilinear form satisfying assumptions **(H1)**–**(H2)** is

$$J_T(\boldsymbol{\sigma}, \boldsymbol{\tau}) := \sum_{F\in\mathcal{F}_T} h_F(\mathfrak{C}_T^k\boldsymbol{\sigma}\cdot\boldsymbol{n}_{TF} - \sigma_{TF}, \mathfrak{C}_T^k\boldsymbol{\tau}\cdot\boldsymbol{n}_{TF} - \tau_{TF})_F. \tag{1.34}$$

In (1.34), we penalize in a least-square sense the difference between two quantities both representing the normal component of the flux variable on a face F. For further

use, we also define the global stabilization bilinear form J_h on $\check{\boldsymbol{\Sigma}}_h^k \times \check{\boldsymbol{\Sigma}}_h^k$ such that

$$J_h(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) := \sum_{T\in\mathcal{T}_h} J_T(R^k_{\boldsymbol{\Sigma},T}\boldsymbol{\sigma}_h, R^k_{\boldsymbol{\Sigma},T}\boldsymbol{\tau}_h). \tag{1.35}$$

Letting $f_h := \pi_h^k f$, the MHO method reads: Find $(\boldsymbol{\sigma}_h, u_h) \in \boldsymbol{\Sigma}_h^k \times U_h^k$ such that

$$H_h(\boldsymbol{\sigma}_h, \boldsymbol{\tau}_h) + (u_h, D_h^k\boldsymbol{\tau}_h) = 0 \qquad \forall \boldsymbol{\tau}_h \in \boldsymbol{\Sigma}_h^k, \tag{1.36a}$$

$$-(D_h^k\boldsymbol{\sigma}_h, v_h) = (f_h, v_h) \qquad \forall v_h \in U_h^k. \tag{1.36b}$$

The well-posedness of problem (1.36) is a classical consequence of the coercivity (1.32a) of H_T in the kernel of D_T^k together with the commuting property (1.26). For the details, we refer to [57].

1.3 Mixed hybrid formulation

In this section we hybridize (1.36) in the spirit of [8] by using the unpatched space $\check{\boldsymbol{\Sigma}}_h^k$ defined by (1.14) in place of the subspace $\boldsymbol{\Sigma}_h^k$ defined by (1.15), and enforcing the single-valuedness of flux DOFs located at interfaces via Lagrange multipliers. Let

$$\Lambda_h^k := \bigtimes_{F\in\mathcal{F}_h} \Lambda_F^k \qquad \text{with} \qquad \Lambda_F^k := \begin{cases} \mathbb{P}^k(F) & \text{if } F \in \mathcal{F}_h^{\mathrm{i}}, \\ \{0\} & \text{if } F \in \mathcal{F}_h^{\mathrm{b}}, \end{cases} \tag{1.37}$$

and define the following local and global hybrid DOF spaces (U_T^k and U_h^k are given by (1.20)):

$$W_T^k := U_T^k \times \left\{ \bigtimes_{F\in\mathcal{F}_T} \Lambda_F^k \right\} \quad \forall T \in \mathcal{T}_h \qquad \text{and} \qquad W_h^k := U_h^k \times \Lambda_h^k. \tag{1.38}$$

We next define a H_0^1-like discrete norm and an interpolator on the space of hybrid DOFs. For all $T \in \mathcal{T}_h$, denote by $R_{W,T}^k : W_h^k \to W_T^k$ the restriction operator that maps global to local DOFs. We equip W_h^k with the norm such that, for all $z_h \in W_h^k$,

$$\|z_h\|_{1,h}^2 := \sum_{T\in\mathcal{T}_h} \|R_{W,T}^k z_h\|_{1,T}^2, \tag{1.39}$$

with local norm such that, for all $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$,

$$\|z\|_{1,T}^2 := \|\boldsymbol{\nabla} v_T\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1}\|\mu_F - v_T\|_F^2 \qquad \forall T \in \mathcal{T}_h. \tag{1.40}$$

Proposition 1.3.1 (Norm $\|\cdot\|_{1,h}$)**.** *The map $\|\cdot\|_{1,h}$ is a norm on W_h^k.*

Proof. We only have to prove that, for all $z_h = ((v_T)_{T\in\mathcal{T}_h}, (\mu_F)_{F\in\mathcal{F}_h}) \in W_h^k$, $\|z_h\|_{1,h} = 0$ implies $z_h = 0_{W,h}$. Let us assume $\|z_h\|_{1,h} = 0$. By definition of the $\|\cdot\|_{1,h}$-norm, this implies for every $T \in \mathcal{T}_h$,

$$\boldsymbol{\nabla} v_T = 0, \quad \text{ and } \mu_F - v_T = 0 \quad \forall F \in \mathcal{F}_T. \tag{1.41}$$

By the first relation in (1.41), we infer that v_T is constant on T. Let now T be such that one of its faces F belongs to $\mathcal{F}_h^{\rm b}$, so that $\mu_F = 0$ by the definition (1.37) of Λ_F^k. Using the second relation in (1.41), we infer that $v_T = 0$ and $\mu_{F'} = 0$ for all $F' \in \mathcal{F}_T\backslash\{F\}$. Using similar arguments to proceed towards the interior of the domain, we finally conclude that $v_T = 0$ for all $T \in \mathcal{T}_h$ and $\mu_F = 0$ for all $F \in \mathcal{F}_h$. □

We introduce the bilinear form B_h on $\check{\boldsymbol{\Sigma}}_h^k \times W_h^k$ such that, for all $(\boldsymbol{\tau}_h, z_h) \in \check{\boldsymbol{\Sigma}}_h^k \times W_h^k$ with $z_h = (v_h, \mu_h)$, recalling the definitions (1.21) of D_T^k and (1.25) of D_h^k to infer the second equality,

$$B_h(\boldsymbol{\tau}_h, z_h) := (v_h, D_h^k\boldsymbol{\tau}_h) - \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\mu_F, \tau_{TF})_F \tag{1.42a}$$

$$= \sum_{T\in\mathcal{T}_h}\left\{-(\boldsymbol{\nabla} v_T, \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T}(v_T - \mu_F, \tau_{TF})_F\right\}. \tag{1.42b}$$

Problem (1.36) can be equivalently reformulated as follows: Find $\overline{\boldsymbol{\sigma}}_h \in \check{\boldsymbol{\Sigma}}_h^k$ and $\overline{w}_h := (\overline{u}_h, \lambda_h) \in W_h^k$ such that,

$$H_h(\overline{\boldsymbol{\sigma}}_h, \boldsymbol{\tau}_h) + B_h(\boldsymbol{\tau}_h, \overline{w}_h) = 0 \qquad \forall \boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k, \tag{1.43a}$$

$$-B_h(\overline{\boldsymbol{\sigma}}_h, z_h) = (f_h, v_h) \qquad \forall z_h = (v_h, \mu_h) \in W_h^k. \tag{1.43b}$$

The following result justifies the choice of the space (1.37) for the Lagrange multipliers by showing that problem (1.43) is well-posed, and establishes a link between the solutions of problems (1.36) and (1.43).

Lemma 1.3.2 (Relation between (1.36) and (1.43))**.** *The following inf-sup condition*

holds with $C > 0$ independent of h s.t. for all $z_h \in W_h^k$:

$$C\|z_h\|_{1,h} \leqslant \sup_{\boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k, |||\boldsymbol{\tau}_h||| = 1} B_h(\boldsymbol{\tau}_h, z_h), \tag{1.44}$$

with $\|\cdot\|_{1,h}$-norm on W_h^k defined by (1.39) *and $|||\cdot|||$-norm on $\check{\boldsymbol{\Sigma}}_h^k$ defined by* (1.17)*. Additionally, problem* (1.43) *has a unique solution $(\overline{\boldsymbol{\sigma}}_h, (\overline{u}_h, \lambda_h)) \in \check{\boldsymbol{\Sigma}}_h^k \times W_h^k$. Finally, denoting by $(\boldsymbol{\sigma}_h, u_h) \in \boldsymbol{\Sigma}_h^k \times U_h^k$ the unique solution to problem* (1.36)*, it holds*

$$(\overline{\boldsymbol{\sigma}}_h, \overline{u}_h) = (\boldsymbol{\sigma}_h, u_h).$$

In view of this result, we drop the bar in what follows.

Proof. Let $z_h = (v_h, \mu_h) \in W_h^k$ be given, and define $\boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k$ such that, for all $T \in \mathcal{T}_h$,

$$\tau_T = -(\boldsymbol{\nabla} v_h)_{|T} \text{ and } \tau_{TF} = h_F^{-1}(v_F - \mu_F) \quad \forall F \in \mathcal{F}_T.$$

Assume $z_h \neq 0_{W,h}$ since the other case is trivial. Then, it holds, using (1.42b), the linearity of B_h in its first argument, and denoting by \$ the supremum in (1.44),

$$\|z_h\|_{1,h}^2 = B_h(\boldsymbol{\tau}_h, z_h) = B_h\left(\frac{\boldsymbol{\tau}_h}{|||\boldsymbol{\tau}_h|||}, z_h\right) |||\boldsymbol{\tau}_h||| \leqslant \$|||\boldsymbol{\tau}_h||| = \$\|z_h\|_{1,h},$$

which proves (1.44).

The well-posedness of problem (1.43) is a classical consequence of (1.44) together with the coercivity (1.32a) of the bilinear form H_h in the kernel of D_h^k, see, e.g., [32, Section 4.2.3]. The last part of the statement is a classical result from the theory of Lagrange multipliers. □

1.4 Primal hybrid formulation

In this section, we reformulate the mixed hybrid problem (1.43) as a coercive primal hybrid problem after locally eliminating the flux DOFs, and we establish a link with the HHO method of [59].

Let

$$W(T) := \{v \in H^1(T) \mid v_{|\partial T \cap \partial\Omega} \equiv 0\}. \tag{1.45}$$

We introduce the local interpolator $I_{W,T}^k : W(T) \to W_T^k$ such that, for all $v \in W(T)$,

$$I_{W,T}^k v = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \quad \text{with} \quad v_T = \pi_T^k v \quad \text{and} \quad \mu_F = \pi_F^k v \quad \forall F \in \mathcal{F}_T. \tag{1.46}$$

The corresponding global interpolator is $I_{W,h}^k : W \to W_h^k$ (recall that $W := H_0^1(\Omega)$) such that, for all $v \in W$,

$$I_{W,h}^k v = ((v_T)_{T\in\mathcal{T}_h}, (\mu_F)_{F\in\mathcal{F}_h}) \quad \text{with} \quad v_T = \pi_T^k v \quad \forall T \in \mathcal{T}_h \quad \text{and} \quad \mu_F = \pi_F^k v \quad \forall F \in \mathcal{F}_h. \tag{1.47}$$

1.4.1 Potential-to-flux operator

The first step consists in locally eliminating the flux DOFs. We define local and global operators which allow, given a set of potential DOFs, to identify the corresponding flux DOFs. To this end, we need a stronger assumption than (1.32a), namely:

$$\eta|||\boldsymbol{\tau}|||_T^2 \leqslant H_T(\boldsymbol{\tau},\boldsymbol{\tau}) \qquad \forall \boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k, \tag{$\mathbf{H2}^+$}$$

so that H_T (resp. H_h) is actually an inner-product on $\boldsymbol{\Sigma}_T^k$ (resp. $\check{\boldsymbol{\Sigma}}_h^k$), defining a norm $\|\cdot\|_{H,T}$ (resp. $\|\cdot\|_H$) equivalent to $|||\cdot|||_T$ (resp. $|||\cdot|||$).

Proposition 1.4.1. *The stabilization bilinear form J_T defined by* (1.34) *satisfies* ($\mathbf{H2}^+$).

Proof. Recalling the first inequality in (1.29) to infer $\|\boldsymbol{\tau}_T\|_T \leqslant \|\mathfrak{C}_T^k\boldsymbol{\tau}\|_T$, and inserting $\pm\mathfrak{C}_T^k\boldsymbol{\tau}\cdot\boldsymbol{n}_{TF}$ into the second term in the right-hand side of (1.16), one has, for all $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k$,

$$\begin{aligned}|||\boldsymbol{\tau}|||_T^2 &\lesssim \|\mathfrak{C}_T^k\boldsymbol{\tau}\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F\|\mathfrak{C}_T^k\boldsymbol{\tau}\cdot\boldsymbol{n}_{TF} - \tau_{TF}\|_F^2 + \sum_{F\in\mathcal{F}_T} h_F\|\mathfrak{C}_T^k\boldsymbol{\tau}\cdot\boldsymbol{n}_{TF}\|_F^2\\ &\lesssim \|\mathfrak{C}_T^k\boldsymbol{\tau}\|_T^2 + J_T(\boldsymbol{\tau},\boldsymbol{\tau}) = H_T(\boldsymbol{\tau},\boldsymbol{\tau}),\end{aligned}$$

where we have used the definition (1.34) of J_T together with the discrete trace inequality (1.7) and the bound (1.6) on $\mathfrak{N}_T$ to pass to the second line, plus the definition (1.30) of the bilinear form H_T to conclude. □

For all $T \in \mathcal{T}_h$, a local *potential-to-flux operator* $\boldsymbol{\varsigma}_T^k : W_T^k \to \boldsymbol{\Sigma}_T^k$ can be naturally defined such that, for all $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$, it holds, for all $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k$, using

the definition (1.21) of D_T^k to pass to the second line,

$$H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) = -(v_T, D_T^k \boldsymbol{\tau})_T + \sum_{F\in\mathcal{F}_T} (\mu_F, \tau_{TF})_F \tag{1.48a}$$

$$= (\boldsymbol{\nabla} v_T, \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T} (\mu_F - v_T, \tau_{TF})_F, \tag{1.48b}$$

insofar as this yields a well-posed problem for $\boldsymbol{\varsigma}_T^k z$ in view of (**H2**$^+$). We also define the global potential-to-flux operator $\boldsymbol{\varsigma}_h^k : W_h^k \to \check{\boldsymbol{\Sigma}}_h^k$ such that, for all $z_h \in W_h^k$,

$$R_{\boldsymbol{\Sigma},T}^k(\boldsymbol{\varsigma}_h^k z_h) = \boldsymbol{\varsigma}_T^k(R_{W,T}^k z_h) \qquad \forall T \in \mathcal{T}_h.$$

An important remark is that, as a consequence of (1.48), $\boldsymbol{\varsigma}_h^k z_h$ satisfies

$$\forall z_h \in W_h^k, \boldsymbol{\tau}_h \in \check{\boldsymbol{\Sigma}}_h^k, \qquad H_h(\boldsymbol{\varsigma}_h^k z_h, \boldsymbol{\tau}_h) = -B_h(\boldsymbol{\tau}_h, z_h) \tag{1.49}$$

with bilinear form B_h defined by (1.42a).

Lemma 1.4.2 (Stability and continuity for $\boldsymbol{\varsigma}_T^k$)**.** *For all $T \in \mathcal{T}_h$ and all $z \in W_T^k$, it holds, denoting by $\|\cdot\|_{H,T}$ the norm defined by H_T on $\boldsymbol{\Sigma}_T^k$,*

$$\eta^{1/2}\|z\|_{1,T} \leqslant \|\boldsymbol{\varsigma}_T^k z\|_{H,T} \leqslant \eta^{-1/2}\|z\|_{1,T}. \tag{1.50}$$

Thus, for all $z_h \in W_h^k$, we have, with $\|\cdot\|_H$ denoting the norm defined by H on $\check{\boldsymbol{\Sigma}}_h^k$,

$$\eta^{1/2}\|z_h\|_{1,h} \leqslant \|\boldsymbol{\varsigma}_h^k z_h\|_H \leqslant \eta^{-1/2}\|z_h\|_{1,h}. \tag{1.51}$$

Proof. Let $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$. Letting $\boldsymbol{\tau}_z \in \boldsymbol{\Sigma}_T^k$ be such that

$$\boldsymbol{\tau}_z = (\boldsymbol{\nabla} v_T, (h_F^{-1}(\mu_F - v_T))_{F\in\mathcal{F}_T}),$$

so that $|||\boldsymbol{\tau}_z|||_T = \|z\|_{1,T}$, one has, using (1.48b) with $\boldsymbol{\tau} = \boldsymbol{\tau}_z$ followed by (1.32b),

$$H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}_z) = \|z\|_{1,T}^2 = |||\boldsymbol{\tau}_z|||_T \|z\|_{1,T} \geqslant \eta^{1/2}\|\boldsymbol{\tau}_z\|_{H,T}\|z\|_{1,T}.$$

Hence, the first inequality in (1.50) is proved observing that

$$\eta^{1/2}\|z\|_{1,T} \leqslant \sup_{\boldsymbol{\tau}\in\boldsymbol{\Sigma}_T^k\setminus\{\mathbf{0}_{\boldsymbol{\Sigma},T}\}} \frac{H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau})}{\|\boldsymbol{\tau}\|_{H,T}} \lesssim \sup_{\boldsymbol{\tau}\in\boldsymbol{\Sigma}_T^k\setminus\{\mathbf{0}_{\boldsymbol{\Sigma},T}\}} \frac{\|\boldsymbol{\varsigma}_T^k z\|_{H,T}\|\boldsymbol{\tau}\|_{H,T}}{\|\boldsymbol{\tau}\|_{H,T}} = \|\boldsymbol{\varsigma}_T^k z\|_{H,T},$$

where the second bound uses the fact that H_T defines an inner product on $\boldsymbol{\Sigma}_T^k$.

On the other hand, it holds for all $\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k$, bounding the right-hand side of (1.48b) with the Cauchy-Schwarz inequality, and recalling the definitions (1.40) of the $\|\cdot\|_{1,T}$-norm and (1.16) of the $|||\cdot|||_T$-norm,

$$H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) \leqslant \|z\|_{1,T} |||\boldsymbol{\tau}|||_T \leqslant \eta^{-1/2} \|z\|_{1,T} \|\boldsymbol{\tau}\|_{H,T},$$

where we have used (**H2**$^+$) to conclude. The second inequality in (1.50) then follows from the previous bound observing that

$$\|\boldsymbol{\varsigma}_T^k z\|_{H,T} = \sup_{\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k \setminus \{\mathbf{0}_{\boldsymbol{\Sigma},T}\}} \frac{H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau})}{\|\boldsymbol{\tau}\|_{H,T}} \lesssim \sup_{\boldsymbol{\tau} \in \boldsymbol{\Sigma}_T^k \setminus \{\mathbf{0}_{\boldsymbol{\Sigma},T}\}} \frac{\|z\|_{1,T} \|\boldsymbol{\tau}\|_{H,T}}{\|\boldsymbol{\tau}\|_{H,T}} = \|z\|_{1,h}.$$

Finally, (1.51) can be proved squaring (1.50) and summing over $T \in \mathcal{T}_h$. □

1.4.2 Discrete gradient and potential reconstruction operators

Let us next define the *gradient reconstruction operator*

$$\boldsymbol{G}_T^k := \mathfrak{C}_T^k \circ \boldsymbol{\varsigma}_T^k, \tag{1.52}$$

with $\mathfrak{C}_T^k$ and $\boldsymbol{\varsigma}_T^k$ defined by (1.27) and (1.48), respectively.

Proposition 1.4.3 (Characterization of $\boldsymbol{G}_T^k$)**.** *The operator $\boldsymbol{G}_T^k$ satisfies the following remarkable property: For all $z = (v_T, (\mu_F)_{F \in \mathcal{F}_T}) \in W_T^k$,*

$$(\boldsymbol{G}_T^k z, \boldsymbol{\nabla} w)_T = (\boldsymbol{\nabla} v_T, \boldsymbol{\nabla} w)_T + \sum_{F \in \mathcal{F}_T} (\mu_F - v_T, \boldsymbol{\nabla} w \cdot \boldsymbol{n}_{TF})_F \quad \forall w \in \mathbb{P}^{k+1}(T). \tag{1.53}$$

Proof. Let $w \in \mathbb{P}^{k+1}(T)$ be fixed, make $\boldsymbol{\tau} := I_{\boldsymbol{\Sigma},T}^k \boldsymbol{\nabla} w$ in (1.48b), and use the fact that $\mathfrak{C}_T^k \boldsymbol{\tau} = \boldsymbol{\nabla} w$ owing to (1.28) and that $\mathfrak{C}_T^k \boldsymbol{\varsigma}_T^k z = \boldsymbol{G}_T^k z$ and $J_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) = J_T(\boldsymbol{\varsigma}_T^k z, I_{\boldsymbol{\Sigma},T}^k \boldsymbol{\nabla} w) = 0$ owing to (1.52) and (1.31), respectively, to infer from the definition (1.30) of H_T that

$$H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) = (\mathfrak{C}_T^k \boldsymbol{\varsigma}_T^k z, \mathfrak{C}_T^k \boldsymbol{\tau}) + J_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) = (\boldsymbol{G}_T^k z, \boldsymbol{\nabla} w)_T.$$

Plugging this relation into (1.48b) yields (1.53). □

Equation (1.53) shows that the discrete gradient operator defined by (1.52) is in fact analogous to the one defined in [59, eq. (11)] in the framework of HHO methods

provided that the Lagrange multipliers are interpreted as trace unknowns. In what follows we recall some important consequences:

(i) *Euler equation.* For any function $\varphi \in H^1(T)$, the following orthogonality property holds:

$$(\boldsymbol{G}_T^k I_{W,T}^k \varphi - \boldsymbol{\nabla}\varphi, \boldsymbol{\nabla} w)_T = 0 \qquad \forall w \in \mathbb{P}^{k+1}(T). \tag{1.54}$$

Interpreting (1.54) as an Euler equation, we conclude that $\boldsymbol{G}_T^k \circ I_{W,T}^k$ is in fact the L^2-orthogonal projector on $\boldsymbol{\nabla}\mathbb{P}^{k+1}(T)$.

(ii) *Potential reconstruction.* Defining, for all $T \in \mathcal{T}_h$, the local potential reconstruction operator $r_T^k : W_T^k \to \mathbb{P}^{k+1}(T)$ such that, for all $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$,

$$\boldsymbol{\nabla} r_T^k z = \boldsymbol{G}_T^k z, \qquad (r_T^k z - v_T, 1)_T = 0 \tag{1.55}$$

there exists a real number $C > 0$ independent of h_T such that, for all $v \in W(T) \cap H^{k+2}(T)$,

$$\begin{aligned} h_T\|\boldsymbol{\nabla}(v - r_T^k I_{W,T}^k v)\|_T + h_T^{3/2}\|\boldsymbol{\nabla}(v - r_T^k I_{W,T}^k v)\|_{\partial T} \\ + \|v - r_T^k I_{W,T}^k v\|_T + h_T^{1/2}\|v - r_T^k I_{W,T}^k v\|_{\partial T} \leqslant C h_T^{k+2}\|v\|_{H^{k+2}(T)}. \end{aligned} \tag{1.56}$$

(iii) *Control of the potential-to-flux element-based DOFs.* For all $z \in W_T^k$, it holds, denoting by $(\boldsymbol{\varsigma}_T^k z)_T \in \mathbb{T}_T^k$ the element DOFs for $\boldsymbol{\varsigma}_T^k z \in \boldsymbol{\Sigma}_T^k$,

$$\|(\boldsymbol{\varsigma}_T^k z)_T\|_T \leqslant \|\boldsymbol{G}_T^k z\|_T \leqslant \|\boldsymbol{\varsigma}_T^k z\|_{H,T} \leqslant \eta^{-1/2}\|z\|_{1,T} \quad \forall z \in W_T^k. \tag{1.57}$$

We close this section by defining global gradient and potential reconstructions as follows: For all $z_h \in W_h^k$, we let

$$(\boldsymbol{G}_h^k z_h)_{|T} := \boldsymbol{G}_T^k R_{W,T}^k z_h \quad \text{and} \quad (r_h^k z_h)_{|T} = r_T^k R_{W,T}^k z_h \quad \forall T \in \mathcal{T}_h. \tag{1.58}$$

1.4.3 Primal hybrid formulation

Denoting by $(\boldsymbol{\sigma}_h, w_h) \in \check{\boldsymbol{\Sigma}}_h^k \times W_h^k$ the solution to problem (1.43) (we have removed the bar from $\boldsymbol{\sigma}_h$ as a result of Lemma 1.3.2), it is readily inferred from (1.49) and (1.43a) that

$$\boldsymbol{\sigma}_h = \boldsymbol{\varsigma}_h^k w_h. \tag{1.59}$$

Then, using (1.59), equation (1.43b) can be rewritten for all $z_h = (v_h, \mu_h) \in W_h^k$ as

$$-B_h(\boldsymbol{\varsigma}_h^k w_h, z_h) = (f_h, v_h).$$

Define the bilinear form A_h on $W_h^k \times W_h^k$ such that, for all $w_h, z_h \in W_h^k$,

$$A_h(w_h, z_h) := H_h(\boldsymbol{\varsigma}_h^k w_h, \boldsymbol{\varsigma}_h^k z_h) = (\boldsymbol{G}_h^k w_h, \boldsymbol{G}_h^k z_h) + j_h(w_h, z_h), \tag{1.60}$$

where we have introduced the bilinear form j_h on $W_h^k \times W_h^k$ such that

$$j_h(w_h, z_h) := J_h(\boldsymbol{\varsigma}_h^k w_h, \boldsymbol{\varsigma}_h^k z_h), \tag{1.61}$$

with J_h defined by (1.35). The equality in (1.60) is a straightforward consequence of (1.30) together with (1.48b). Then, recalling (1.49) and using the symmetry of the bilinear form H_h, it is inferred, for all $z_h \in W_h^k$,

$$-B_h(\boldsymbol{\varsigma}_h^k w_h, z_h) = H_h(\boldsymbol{\varsigma}_h^k w_h, \boldsymbol{\varsigma}_h^k z_h) = A_h(w_h, z_h),$$

and we conclude that:

Theorem 1.4.4 (Primal reformulation of problem (1.43))**.** *The problem* (1.43) *can be reformulated as the following coercive problem: Find* $w_h = (u_h, \lambda_h) \in W_h^k$ *such that,*

$$A_h(w_h, z_h) = (f, v_h) \qquad \forall z_h = (v_h, \mu_h) \in W_h^k, \tag{1.62}$$

and (1.59) *holds.*

It follows from (1.51) that, for all $z_h \in W_h^k$, observing that $A_h(z_h, z_h) = \|\boldsymbol{\varsigma}_h^k z_h\|_{H,T}^2$ as a consequence of (1.60),

$$\eta \|z_h\|_{1,h}^2 \leqslant A_h(z_h, z_h) := \|z_h\|_A^2 \leqslant \eta^{-1} \|z_h\|_{1,h}^2. \tag{1.63}$$

As a result, the bilinear form A_h is coercive, and the well-posedness of the primal problem (1.62) follows directly from the Lax–Milgram lemma.

Remark 1.4.5 (Implementation). From a practical viewpoint, the symmetric positive definite linear system associated to problem (1.62) can be solved more efficiently than the saddle-point system associated to problem (1.36). Moreover, when doing so, element-based DOFs can be statically condensed, leading to a global problem in the Lagrange multipliers only. The discrete flux $\boldsymbol{\sigma}_h$ can be recovered from the solution of problem (1.62) according to (1.59) by an element-by-element post-processing.

1.4.4 Link with the Hybrid High-Order method

In [59], the authors study a Hybrid High-Order method based on the following bilinear form on $W_h^k \times W_h^k$, which only differs from the one defined by (1.62) in the choice of the stabilization term:

$$A_h^{\mathrm{HHO}}(w_h, z_h) := (\boldsymbol{G}_h^k w_h, \boldsymbol{G}_h^k z_h) + j_h^{\mathrm{HHO}}(w_h, z_h), \tag{1.64}$$

where, in comparison with (1.61), no link with a mixed hybrid method is used, but

$$j_h^{\mathrm{HHO}}(w_h, z_h) := \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} (\pi_F^k(\mathfrak{r}_T^k R_{W,T}^k w_h - \lambda_F), \pi_F^k(\mathfrak{r}_T^k R_{W,T}^k z_h - \mu_F))_F,$$

and, for all $T \in \mathcal{T}_h$, the potential reconstruction operator $\mathfrak{r}_T^k : W_T^k \to \mathbb{P}^{k+1}(T)$ is such that, for all $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$,

$$\mathfrak{r}_T^k z = \big(r_T^k z - \pi_T^k(r_T^k z)\big) + v_T,$$

with r_T^k is defined by (1.55). The stabilization bilinear forms j_h defined by (1.61) and j_h^{HHO} are equivalent in that both of them (i) are polynomially consistent up to degree $k+1$ and (ii) yield stability and continuity for A_h and A_h^{HHO} in the form (1.63).

1.5 Error analysis

In this section we show how the error analysis for the MHO method (1.36) can be carried out directly based on the primal formulation (1.62). The error analysis for the mixed formulation (1.36) can be found in [57]. Additional estimates as well as a potential reconstruction of order $(k+2)$ are proposed.

1.5.1 Energy error estimate

Theorem 1.5.1 (Energy error estimate)**.** *Let $u \in W$ and $w_h = (u_h, \lambda_h) \in W_h^k$ denote the unique solutions to* (1.2) *and* (1.62)*, respectively, and set*

$$\widehat{w}_h = (\widehat{u}_h, \widehat{\lambda}_h) := I_{W,h}^k u,$$

where $I_{W,h}^k$ is the interpolation operator defined by (1.47)*. Then, provided $u \in H^{k+2}(\mathcal{T}_h)$, the following estimate holds with real number C independent of h and*

norm $\|\cdot\|_A$ *defined by* (1.63)*:*

$$\eta^{1/2}\|\widehat{w}_h - w_h\|_{1,h} \leqslant \|\widehat{w}_h - w_h\|_A \leqslant Ch^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}. \tag{1.65}$$

Proof. The first inequality in (1.65) is an immediate consequence of the coercivity of A_h, cf. (1.63). Moreover, again recalling (1.63), it is readily inferred that

$$\|\widehat{w}_h - w_h\|_A \leqslant \eta^{-1/2}\frac{A_h(\widehat{w}_h - w_h, \widehat{w}_h - w_h)}{\|\widehat{w}_h - w_h\|_{1,h}} \leqslant \eta^{-1/2}\sup_{z_h\in W_h^k,\, \|z_h\|_{1,h}=1} A_h(\widehat{w}_h - w_h, z_h).$$

Owing to (1.62), we then infer that it holds

$$\|\widehat{w}_h - w_h\|_A \leqslant \eta^{-1/2}\sup_{z_h\in W_h^k,\, \|z_h\|_{1,h}=1} \mathcal{E}_h(z_h), \tag{1.66}$$

with consistency error $\mathcal{E}_h$ such that, for all $z_h = (v_h, \mu_h) \in W_h^k$,

$$\mathcal{E}_h(z_h) := A_h(\widehat{w}_h, z_h) - (f, v_h) = \big\{(\boldsymbol{G}_h^k\widehat{w}_h, \boldsymbol{G}_h^k z_h) - (f, v_h)\big\} + j_h(\widehat{w}_h, z_h) := \mathfrak{T}_1 + \mathfrak{T}_2. \tag{1.67}$$

Let, for all $T \in \mathcal{T}_h$, $\breve{u}_T \in \mathbb{P}^{k+1}(T)$ denote the elliptic projection of u such that,

$$(\boldsymbol{\nabla}(\breve{u}_T - u), \boldsymbol{\nabla}\xi)_T = 0 \quad \text{for all } \xi \in \mathbb{P}^{k+1}(T), \quad (\breve{u}_T - u, 1)_T = 0.$$

By (1.54), it holds

$$\boldsymbol{G}_T^k R_{W,T}^k \widehat{w}_h = \boldsymbol{\nabla}\breve{u}_T, \quad \forall T \in \mathcal{T}_h.$$

Moreover, since $f = -\triangle u$ a.e. in Ω, it is readily inferred from (1.53) that

$$\begin{aligned}
\mathfrak{T}_1 &= \sum_{T\in\mathcal{T}_h}\{(\boldsymbol{\nabla}\breve{u}_T, \boldsymbol{G}_T^k R_{W,T}^k z_h)_T\} + (\triangle u, v_h)\\
&= \sum_{T\in\mathcal{T}_h}\{(\boldsymbol{\nabla}\breve{u}_T, \boldsymbol{G}_T^k R_{W,T}^k z_h)_T\} + \sum_{T\in\mathcal{T}_h}(\triangle u, v_T)_T\\
&= \sum_{T\in\mathcal{T}_h}\{(\boldsymbol{\nabla}\breve{u}_T, \boldsymbol{\nabla}v_T)_T + \sum_{F\in\mathcal{F}_T}(\mu_F - v_T, \boldsymbol{\nabla}\breve{u}_T\cdot\boldsymbol{n}_{TF})_F - (\boldsymbol{\nabla}u, \boldsymbol{\nabla}v_T)_T + \sum_{F\in\mathcal{F}_T}(v_T, \boldsymbol{\nabla}u\cdot\boldsymbol{n}_{TF})_F\}\\
&= \sum_{T\in\mathcal{T}_h}(\boldsymbol{\nabla}(\breve{u}_T - u), \boldsymbol{\nabla}v_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\nabla}(\breve{u}_T - u)\cdot\boldsymbol{n}_{TF}, \mu_F - v_T)_F\\
&= \sum_{T\in\mathcal{T}_h}\{\mathfrak{T}_{1,1}(T) + \mathfrak{T}_{1,2}(T)\},
\end{aligned}$$

where we have used the definition of $\boldsymbol{G}_T^k$ with $w = \breve{u}_T \in \mathbb{P}^{k+1}(T)$ to pass to the third line and the flux continuity across interfaces together with the fact that $\mu_F \equiv 0$ for

all $F \in \mathcal{F}_h^{\mathrm{b}}$ to insert

$$\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\boldsymbol{\nabla}u\cdot\boldsymbol{n}_{TF},\mu_F)_F = 0$$

in the fourth line. By definition of $\breve{u}_T$, and using the orthogonality relation (1.54) with $\varphi = u \in H^{k+2}(T)$ and $w = v_T \in \mathbb{P}^k(T) \subset \mathbb{P}^{k+1}(T)$, we immediately infer

$$\mathfrak{T}_{1,1}(T) = (\boldsymbol{G}_T^k I_{W,T}^k u - \boldsymbol{\nabla}u, \boldsymbol{\nabla}v_T)_T = 0.$$

The second term $\mathfrak{T}_{1,2}$ can be estimated as follows:

$$\begin{aligned}
\mathfrak{T}_{1,2}(T) &\leqslant \sum_{F\in\mathcal{F}_T}\|\boldsymbol{\nabla}(\breve{u}_T - u)\|_F\|\mu_F - v_T\|_F && (1.68)\\
&\leqslant \left(\sum_{F\in\mathcal{F}_T} h_F\|\boldsymbol{\nabla}(\breve{u}_T - u)\|_F^2\right)^{1/2}\left(\sum_{F\in\mathcal{F}_T} h_F^{-1}\|\mu_F - v_T\|_F^2\right)^{1/2} && (1.69)\\
&\lesssim h_T^{k+1}\|u\|_{H^{k+2}(T)}\|R_{W,T}^k z_h\|_{1,T}. && (1.70)
\end{aligned}$$

Finally, one obtains a bound on $\mathfrak{T}_1$ after summing over $T \in \mathcal{T}_h$ and using a Cauchy-Schwarz inequality,

$$|\mathfrak{T}_1| \lesssim h^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}\|z_h\|_{1,h}. \tag{1.71}$$

For the second term $\mathfrak{T}_2$ in (1.67), letting $\boldsymbol{\tau}_h := \boldsymbol{\varsigma}_h^k z_h$, we have, on the other hand,

$$\begin{aligned}
\mathfrak{T}_2 &= H_h(\boldsymbol{\varsigma}_h^k\widehat{w}_h, \boldsymbol{\tau}_h) - (\boldsymbol{G}_h^k\widehat{w}_h, \boldsymbol{G}_h^k z_h) && \text{Eq. (1.60)}\\
&= \sum_{T\in\mathcal{T}_h}\left\{(\boldsymbol{\nabla}\widehat{u}_T, \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T}(\widehat{\lambda}_F - \widehat{u}_T, \tau_{TF})_F - (\boldsymbol{\nabla}\breve{u}_T, \mathfrak{C}_T^k R_{\boldsymbol{\Sigma},T}^k\boldsymbol{\tau}_h)_T\right\} && \text{Eqs. (1.48b), (1.52)}\\
&= \sum_{T\in\mathcal{T}_h}\left\{(\boldsymbol{\nabla}(\widehat{u}_T - \pi_T^k\breve{u}_T), \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T}(\widehat{\lambda}_F - \pi_F^k\breve{u}_T - \widehat{u}_T + \pi_T^k\breve{u}_T, \tau_{TF})_F\right\} && \text{Eq. (1.27b)}\\
&= \sum_{T\in\mathcal{T}_h}\left\{(\boldsymbol{\nabla}\pi_T^k(u - \breve{u}_T), \boldsymbol{\tau}_T)_T + \sum_{F\in\mathcal{F}_T}(\pi_F^k(u - \breve{u}_T), \tau_{TF})_F\right.\\
&\qquad \left. + \sum_{F\in\mathcal{F}_T}(\pi_T^k(\breve{u}_T - u), \tau_{TF})_F\right\}. && \text{Eq. (1.47)}\\
&:= \sum_{T\in\mathcal{T}_h}\left\{\mathfrak{T}_{2,1}(T) + \mathfrak{T}_{2,2}(T) + \mathfrak{T}_{2,3}(T)\right\}.
\end{aligned}$$

We treat the terms within braces using the Cauchy–Schwarz, discrete inverse (1.8) and trace (1.7) inequalities, the approximation properties (1.56) of r_T^k, and the boundedness properties of the L^2-orthogonal projectors on polynomial spaces over

elements and faces. We start with the first term

$$
\begin{aligned}
|\mathfrak{T}_{2,1}(T)| &\leqslant \|\boldsymbol{\nabla}\pi_T^k(u-\breve{u}_T)\|_T\|\boldsymbol{\tau}_T\|_T && (1.72)\\
&\lesssim h_T^{-1}\|\pi_T^k(u-\breve{u}_T)\|_T\|\boldsymbol{\tau}_T\|_T && (1.73)\\
&\lesssim h_T^{-1}\|u-\breve{u}_T\|_T\|\boldsymbol{\tau}_T\|_T && (1.74)\\
&\lesssim h_T^{k+1}\|u\|_{H^{k+2}(T)}|\!|\!| R_{\boldsymbol{\Sigma},T}^k\boldsymbol{\tau}_h|\!|\!|_T. && (1.75)
\end{aligned}
$$

The second term inside the sum is treated using similar arguments, namely

$$
\begin{aligned}
|\mathfrak{T}_{2,2}(T)| &\leqslant \sum_{F\in\mathcal{F}_T}\|\pi_F^k(u-\breve{u}_T)\|_F\|\tau_{TF}\|_F\\
&\leqslant \left\{\sum_{F\in\mathcal{F}_T} h_F^{-1}\|\pi_F^k(u-\breve{u}_T)\|_F^2\right\}^{1/2}\times\left\{\sum_{F\in\mathcal{F}_T} h_F\|\tau_{TF}\|_F^2\right\}^{1/2}\\
&\lesssim \left\{h_T^{-1}\sum_{F\in\mathcal{F}_T}\|u-\breve{u}_T\|_F^2\right\}^{1/2}\times|\!|\!| R_{\boldsymbol{\Sigma},T}^k\boldsymbol{\tau}_h|\!|\!|_T\\
&\lesssim h_T^{-1/2}\|u-\breve{u}_T\|_{\partial T}|\!|\!| R_{\boldsymbol{\Sigma},T}^k\boldsymbol{\tau}_h|\!|\!|_T\\
&\lesssim h_T^{k+1}\|u\|_{H^{k+2}(T)}|\!|\!| R_{\boldsymbol{\Sigma},T}^k\boldsymbol{\tau}_h|\!|\!|_T.
\end{aligned}
$$

The term $\mathfrak{T}_{2,3}(T)$ can be estimated using similar arguments. Accounting for the previous bounds on $\mathfrak{T}_{2,1}(T)$, $\mathfrak{T}_{2,2}(T)$, and $\mathfrak{T}_{2,3}(T)$, and using the Cauchy–Schwarz inequality leads to

$$
|\mathfrak{T}_2| \lesssim h^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}|\!|\!|\boldsymbol{\tau}_h|\!|\!| = h^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}|\!|\!|\boldsymbol{\varsigma}_h^k z_h|\!|\!|, \tag{1.76}
$$

since by definition $\boldsymbol{\tau}_h = \boldsymbol{\varsigma}_h^k z_h$. Using (1.71) and (1.76) to bound the consistency error in (1.66) together with (**H2**$^+$) and the second inequality in (1.51) to infer $|\!|\!|\boldsymbol{\varsigma}_h^k z_h|\!|\!| \lesssim \|z_h\|_{1,h}$ concludes the proof. □

Corollary 1.5.2 (Convergence of the gradient reconstruction)**.** *Under the assumptions of Theorem 1.5.1 it holds with real number $C>0$ independent of h (but possibly depending on the mesh regularity parameter ϱ),*

$$
\|\boldsymbol{\nabla}u - \boldsymbol{G}_h^k w_h\| \leqslant Ch^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}.
$$

Proof. We use the triangular inequality to infer

$$
\|\boldsymbol{\nabla}u - \boldsymbol{G}_h^k w_h\| \leqslant \|\boldsymbol{\nabla}u - \boldsymbol{G}_h^k\widehat{w}_h\| + \|\boldsymbol{G}_h^k(\widehat{w}_h - w_h)\|,
$$

and estimate the first term in the right-hand side using the approximation properties (1.56) of $\boldsymbol{G}_T^k$ and the second using the a priori error estimate (1.65) after observing that $\|\boldsymbol{G}_h^k(\widehat{w}_h - w_h)\| \leqslant \|\widehat{w}_h - w_h\|_A$ (this is a consequence of (1.60) since the bilinear form j defined by (1.61) is positive semi-definite owing to **(H1)**). □

1.5.2 Error estimates with elliptic regularity

This section collects error estimates that hold under additional regularity assumptions on the problem. We assume throughout this section that elliptic regularity holds in the following form: For all $g \in L^2(\Omega)$, the unique weak solution $\zeta \in W$ to

$$(\boldsymbol{\nabla}\zeta, \boldsymbol{\nabla}v) = (g, v) \qquad \forall v \in W, \tag{1.77}$$

satisfies the a priori estimate

$$\|\zeta\|_{H^2(\Omega)} \leqslant C_{\mathrm{ell}}\|g\|$$

with C_{ell} only depending on Ω. This holds true, for instance, when Ω is convex. Then, additional error estimates can be derived which are the counterpart of the classical results for the Raviart–Thomas mixed method proved in [8, 64, 72] thanks to the well-known Aubin-Nitsche trick [9], cf. also [56, 60] for its adaptation to HHO methods.

Lemma 1.5.3 (Error estimate for the potential and the Lagrange multipliers)**.** *Under the assumptions of Theorem 1.5.1, and provided that elliptic regularity holds, the following bounds hold for $w_h = (u_h, \lambda_h) \in W_h^k$ solution to* (1.62)*, with $\widehat{u}_h \in U_h^k$ and $\widehat{\lambda}_h \in \Lambda_h^k$ defined as in Theorem 1.5.1:*

$$\|u_h - \widehat{u}_h\| \leqslant Ch^{k+2}\|u\|_{H^{k+2}(\mathcal{T}_h)}, \tag{1.78a}$$

$$|\lambda_h - \widehat{\lambda}_h|_{\mathrm{LM}} \leqslant Ch^{k+1}\|u\|_{H^{k+2}(\mathcal{T}_h)}, \tag{1.78b}$$

where $C > 0$ is a real number independent of h (but depending on the mesh regularity parameter ϱ) and we have set

$$|\mu_h|_{\mathrm{LM}}^2 := \sum_{F\in\mathcal{F}_h} h_F^{-1}\|\mu_F\|_F^2.$$

Remark 1.5.4 (Interpretation of the Lagrange multipliers). In view of (1.78b), the Lagrange multipliers can be interpreted as traces of the potential.

Proof. The bound (1.78a) can be proved for the primal hybrid formulation proceeding as in [59, Theorem 10] estimating the penalty term as in Theorem 1.5.1. To prove (1.78b), it suffices to use the estimate (1.82) below followed by (1.78a) and Theorem 1.5.1. □

The estimate (1.78a) shows that the discrete potential u_h resulting from (1.62) is superclose to the L^2-orthogonal projection of the potential on $\mathbb{P}^k(\mathcal{T}_h)$. As for classical mixed finite element methods [8, 89], we can improve this result and finally exhibit a potential reconstruction that converges as h^{k+2}.

Lemma 1.5.5 (Potential reconstruction)**.** *Under the assumptions of Lemma 1.5.3, denoting by u and $w_h = (u_h, \lambda_h)$ the unique solutions to problems* (1.2) *and* (1.62)*, respectively, it holds with real number $C > 0$ independent of h (but depending on the mesh regularity parameter ϱ)*

$$\|u - r_h^k w_h\| \leqslant Ch^{k+2}\|u\|_{H^{k+2}(T)}, \tag{1.79}$$

where the potential reconstruction operator r_h^k is defined by (1.58)*.*

Proof. Using the triangular inequality we can estimate

$$\|u - r_h^k w_h\| \leqslant \|u - r_h^k \widehat{w}_h\| + \|r_h^k(\widehat{w}_h - w_h)\| := \mathfrak{T}_1 + \mathfrak{T}_2. \tag{1.80}$$

As a result of the approximation properties (1.56), it is readily inferred

$$|\mathfrak{T}_1| \lesssim h^{k+2}\|u\|_{H^{k+2}(\mathcal{T}_h)}.$$

Additionally, using Poincaré's inequality (1.12) inside each element, one has

$$\begin{aligned}
\mathfrak{T}_2^2 &= \sum_{T\in\mathcal{T}_h} \|r_T^k I_{W,T}^k(\widehat{w}_h - w_h)\|_T^2 \\
&\lesssim \sum_{T\in\mathcal{T}_h} \left\{h_T^2\|\boldsymbol{\nabla} r_T^k I_{W,T}^k(\widehat{w}_h - w_h)\|_T^2 + \|\pi_T^0(\widehat{u}_T - u_T)\|_T^2\right\} \\
&\leqslant \sum_{T\in\mathcal{T}_h} \left\{h_T^2\|\boldsymbol{G}_T^k I_{W,T}^k(\widehat{w}_h - w_h)\|_T^2 + \|\widehat{u}_T - u_T\|_T^2\right\},
\end{aligned}$$

where, in the last line, we have used the definition (1.55) of r_T^k together with the fact that π_T^0 is a bounded operator. Hence, using the a priori bounds (1.65) and (1.78a), we infer

$$|\mathfrak{T}_2| \lesssim h^{k+2}\|u\|_{H^{k+2}(\mathcal{T}_h)}.$$

The conclusion follows plugging the bounds for $\mathfrak{T}_1$ and $\mathfrak{T}_2$ into (1.80). □

Proposition 1.5.6. *There exists a real number $C > 0$ independent of h (but depending on the mesh regularity parameter ϱ) such that, for all $T \in \mathcal{T}_h$ and all $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$, the following inequality holds for all $F \in \mathcal{F}_T$*

$$h_F^{-1}\|\mu_F\|_F^2 \leqslant C\left(h_T^{-2}\|v_T\|_T^2 + |||\boldsymbol{\varsigma}_T^k z|||_T^2\right). \tag{1.81}$$

Additionally, for all $z_h = (v_h, \mu_h) \in W_h^k$, we have

$$|\mu_h|_{\mathrm{LM}}^2 \leqslant C\sum_{T\in\mathcal{T}_h}\left(h_T^{-2}\|v_T\|_T^2 + \|\boldsymbol{\varsigma}_T^k R_{W,T}^k z_h\|_T^2\right). \tag{1.82}$$

Proof. Let an element $T \in \mathcal{T}_h$ and a face $F \in \mathcal{F}_T$ be fixed, and, for a given $z = (v_T, (\mu_F)_{F\in\mathcal{F}_T}) \in W_T^k$, let $\boldsymbol{\tau} = (\boldsymbol{\tau}_T, (\tau_{TF})_{F\in\mathcal{F}_T}) \in \boldsymbol{\Sigma}_T^k$ be such that $\tau_{TF} = h_F^{-1}\mu_F$, $\boldsymbol{\tau}_T \equiv \mathbf{0}$, and $\tau_{TF'} \equiv 0$ for all $F' \in \mathcal{F}_T\backslash\{F\}$. Using $\boldsymbol{\tau}$ as a test function in (1.48a) it is inferred

$$\begin{aligned}
h_F^{-1}\|\mu_F\|_F^2 &= (v_T, D_T^k\boldsymbol{\tau})_T + H_T(\boldsymbol{\varsigma}_T^k z, \boldsymbol{\tau}) \\
&\leqslant \|v_T\|_T\|D_T^k\boldsymbol{\tau}\|_T + \eta^{-1}|||\boldsymbol{\varsigma}_T^k z|||_T|||\boldsymbol{\tau}|||_T && \text{Cauchy–Schwarz and eq. (1.32b)} \\
&\lesssim \left(h_T^{-2}\|v_T\|_T^2 + |||\boldsymbol{\varsigma}_T^k z|||_T^2\right)^{1/2}|||\boldsymbol{\tau}|||_T, && \text{eq. (1.23)}
\end{aligned}$$

and (1.81) follows observing that, owing to (1.16), $|||\boldsymbol{\tau}|||_T = h_F^{-1/2}\|\mu_F\|_F$. Inequality (1.82) can be proved observing that $\sum_{F\in\mathcal{F}_h} h_F^{-1}\|\mu_F\|_F^2 \leqslant \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} h_F^{-1}\|\mu_F\|_F^2$ and using (1.81). □

1.6 Extension to the Darcy problem

To show how the presence of spatially varying coefficients can be taken into account, we briefly address in this section the extension to the Darcy problem. For the details we refer to [57, 60]. Let $\boldsymbol{\kappa} : \Omega \to \mathbb{R}^{d\times d}$ denote a tensor-valued, symmetric uniformly elliptic diffusion coefficient, which we assume to be piecewise constant on a fixed partition P_Ω of Ω. We further assume that, for all $h \in \mathcal{H}$, the mesh $\mathcal{T}_h$ is compliant with the partition P_Ω, so that $\boldsymbol{\kappa} \in \mathbb{P}^0(\mathcal{T}_h)^{d\times d}$, and the jumps of $\boldsymbol{\kappa}$ can only occur at interfaces. For a given $f \in L^2(\Omega)$, the model problem reads: Find $\boldsymbol{s} : \Omega \to \mathbb{R}^d$ and

$u : \Omega \to \mathbb{R}$ s.t.,

$$\begin{aligned} \boldsymbol{s} + \boldsymbol{\kappa}\boldsymbol{\nabla} u &= 0 \qquad \text{in } \Omega, \\ \boldsymbol{\nabla}\cdot\boldsymbol{s} &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega. \end{aligned} \tag{1.83}$$

For all $T \in \mathcal{T}_h$, we denote by $\underline{\kappa}_T$ and $\overline{\kappa}_T$ the (positive) smallest and largest eigenvalues of $\boldsymbol{\kappa}_T := \boldsymbol{\kappa}_{|T}$, respectively, and we define the local anisotropy ratio

$$\alpha_T := \frac{\overline{\kappa}_T}{\underline{\kappa}_T}.$$

In what follows we briefly outline the modifications required to adapt the MHO method to the Darcy problem (1.83). A first important modification is that the local space of flux DOFs is now defined as (compare with (1.13))

$$\mathbb{T}_T^k := \boldsymbol{\kappa}_T\boldsymbol{\nabla}\mathbb{P}^k(T). \tag{1.84}$$

Correspondently, the flux reconstruction operator maps on $\boldsymbol{\kappa}_T\boldsymbol{\nabla}\mathbb{P}^{k+1}(T)$ (with (1.27) remaining formally unchanged). The local interpolator $I_{\boldsymbol{\Sigma},T}^k : \boldsymbol{\Sigma}^+(T) \to \boldsymbol{\Sigma}_T^k$ is still defined by (1.18), but ϖ_T^k now denotes the L^2-orthogonal projection on the space $\mathbb{T}_T^k$ defined by (1.84). The global interpolator is still formally given by (1.19). The following energy error estimate is proved in [57, Theorem 6] (compare with Theorem 1.5.1).

Theorem 1.6.1 (Error estimate for the flux)**.** *Let $(\boldsymbol{s}, u)$ denote the weak solution to* (1.83) *and $(\boldsymbol{\sigma}_h, u_h)$ the solution of the MHO discretization applied to the Darcy problem as described above. Then, provided that $\boldsymbol{s} \in H^{k+1}(\mathcal{T}_h)^d$ and $u \in H^{k+2}(\mathcal{T}_h)$, it holds*

$$|||I_{\boldsymbol{\Sigma},h}^k\boldsymbol{s} - \boldsymbol{\sigma}_h|||_h \leqslant C\left\{\sum_{T\in\mathcal{T}_h}\overline{\kappa}_T\alpha_T h_T^{2(k+1)}\|u\|_{H^{k+2}(T)}^2\right\}^{1/2},$$

where $C > 0$ is independent of both h and $\boldsymbol{\kappa}$, but possibly depends on the mesh regularity parameter ϱ.

Remark 1.6.2 (Robustness with respect to $\boldsymbol{\kappa}$). The above estimate shows that the method is fully robust with respect to the heterogeneity of the diffusion coefficient, and it exhibits only a moderate dependence on its local anisotropy ratio α_T (with a power 1/2).

Chapter 2

Application to the Stokes and Oseen problems

In this chapter we apply the hybridized version of the MHO method of Chapter 1 to the discretization of linear problems in incompressible fluid mechanics.

Our first application, taken from [2], is to the Stokes problem. The main difficulty lies here in the enforcement of the zero-divergence constraint on the velocity. For a given polynomial degree $k \geqslant 0$, our discretization hinges on the hybrid space of degrees of freedom (DOF) defined in (1.38) for each component of the velocity, and on the space of fully discontinuos polynomials of degree k for the pressure. This choice of unknowns enables an inf-sup stable discretization on general meshes. Our error analysis shows that the error in the energy norm for the velocity and in the L^2-norm for the pressure optimally scales as h^{k+1} (with h denoting the meshsize). Additionally, under further regularity for the continuous problem, the estimate for the L^2-norm of the velocity can be improved to h^{k+2}. These theoretical estimates are confirmed by numerical experiments.

Our second application is to the development of a novel (not previously published) method for the Oseen problem. With respect to the Stokes problem, the viscous term is multiplied by a (constant) kinematic viscosity coefficient ν, and an additional convective term is added in the momentum equation. A key point is in this case to track the dependence of the constants appearing in the error estimates on the Peclet number, a dimensionless number accounting for the relative importance of advective and viscous effects. We propose here a treatment for the advective term inspired by [54], which yields robust error estimates additionally accounting for the variation

in the order of convergence in the different regimes. Specifically, we prove that the error in the energy norm for the velocity scales as h^{k+1} in the diffusion-dominated regime (a result coherent with the one found for the Stokes problem) and as $h^{k+1/2}$ in the advection-dominated regime. The error on the L^2-norm of pressure has a similar scaling, with an additional (explicit) dependence of the multiplicative constant on the global Peclet number.

Throughout this chapter, $(\mathcal{T}_h)_{h\in\mathcal{H}}$ will denote an admissible mesh sequence in the sense of Definition 1.1.1.

2.1 An inf-sup stable discretization of the Stokes problem on general meshes

The Stokes problem consists in finding the velocity field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ and the pressure field $p : \Omega \to \mathbb{R}$ such that

$$-\triangle \boldsymbol{u} + \boldsymbol{\nabla} p = \boldsymbol{f} \qquad \text{in } \Omega, \tag{2.1a}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{2.1b}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega, \tag{2.1c}$$

$$\int_\Omega p = 0. \tag{2.1d}$$

Denoting by $L^2_0(\Omega)$ the space of square-integrable functions with zero mean on Ω, and letting

$$\boldsymbol{W} := H^1_0(\Omega)^d \qquad P := L^2_0(\Omega), \tag{2.2}$$

a standard weak formulation of (2.1) reads: Find $(\boldsymbol{u}, p) \in \boldsymbol{W} \times P$ such that

$$(\boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{\nabla}\boldsymbol{v}) - (p, \boldsymbol{\nabla}\cdot\boldsymbol{v}) = (\boldsymbol{f}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{W}, \tag{2.3a}$$

$$(\boldsymbol{\nabla}\cdot\boldsymbol{u}, q) = 0 \qquad \forall q \in P. \tag{2.3b}$$

It is appearent from the weak formulation that the pressure p acts as the Lagrange multiplier for the zero-divergence constraint on the velocity $\boldsymbol{u}$. Consequently, problem (2.3) has a saddle-point structure, and its well-posedness hinges on an inf-sup condition. For classical results in this direction, we refer the reader to [75].

The key ideas are here to

(i) discretize the diffusive term in the momentum conservation equation (2.3a) using the bilinear form A_h defined by (1.60) for each component of the discrete velocity field (in view of the results in Section 1.4.4, one could alternatively use the bilinear form A_h^{HHO} defined by (1.64));

(ii) realize the velocity-pressure coupling by means of a discrete divergence operator $\mathcal{D}_h^k$ designed in the same spirit as D_h^k (cf. (1.25)) and relying on the interpretation of the Lagrange multipliers as traces of the potential; cf. Remark 1.5.4. This choice ensures discrete stability in terms of an inf-sup condition.

To alleviate the notation, throughout this section we often abridge by $a \lesssim b$ the inequality $a \leqslant Cb$ with real number $C > 0$ independent of h. Explicit names for the constant are kept in the statements for the sake of easy consultation.

2.1.1 Discrete spaces

Recalling the definition (1.38) of W_T^k, we define, for all $T \in \mathcal{T}_h$, the local DOF space for the velocity as

$$\boldsymbol{W}_T^k := (W_T^k)^d,$$

while we seek the pressure in $\mathbb{P}^k(T)$. Correspondingly, the global DOF spaces for the velocity and pressure are given by

$$\boldsymbol{W}_h^k := (W_h^k)^d, \qquad \mathcal{P}_h^k := \mathbb{P}^k(\mathcal{T}_h) \cap L_0^2(\Omega), \tag{2.4}$$

cf. again (1.38) for the definition of W_h^k. We also define the local and global velocity interpolators $I_{\boldsymbol{W},T}^k$ and $I_{\boldsymbol{W},h}^k$ obtained applying component-wise the interpolators $I_{W,T}^k$ and $I_{W,h}^k$ defined by (1.46) and (1.47), so that, for all $\boldsymbol{z} = (z_i)_{1\leqslant i\leqslant d} \in \boldsymbol{W}$,

$$I_{\boldsymbol{W},T}^k \boldsymbol{z} = (I_{W,T}^k z_{i|T})_{1\leqslant i\leqslant d} \quad \text{and} \quad I_{\boldsymbol{W},h}^k \boldsymbol{z} = (I_{W,h}^k z_i)_{1\leqslant i\leqslant d}. \tag{2.5}$$

Given a generic element $z_h \in W_h^k$ (resp., $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$) and a mesh element $T \in \mathcal{T}_h$, we denote by z_T (resp., $\boldsymbol{z}_T$) its restriction to the local space W_T^k (resp., $\boldsymbol{W}_T^k$).

2.1.2 Viscous term

The discretization of the viscous term in (2.3a) hinges on the bilinear form $\mathcal{A}_h$ on $\boldsymbol{W}_h^k \times \boldsymbol{W}_h^k$ such that, for all $\boldsymbol{w}_h = (w_{h,i})_{1\leqslant i\leqslant d}$ and all $\boldsymbol{z}_h = (z_{h,i})_{1\leqslant i\leqslant d}$ elements of

$\boldsymbol{W}_h^k$,

$$\mathcal{A}_h(\boldsymbol{w}_h, \boldsymbol{z}_h) := \sum_{i=1}^{d} A_h(w_{h,i}, z_{h,i}), \tag{2.6}$$

with bilinear form A_h defined by (1.60). The coercivity and continuity of the bilinear form $\mathcal{A}_h$ follow from the corresponding properties (1.63) of the bilinear form A_h:

$$\eta \|\boldsymbol{z}_h\|_{1,h}^2 \leqslant \mathcal{A}_h(\boldsymbol{z}_h, \boldsymbol{z}_h) := \|\boldsymbol{z}_h\|_{\mathcal{A},h}^2 \leqslant \eta^{-1} \|\boldsymbol{z}_h\|_{1,h}^2, \tag{2.7}$$

where we have introduced the $H_0^1(\Omega)^d$-like seminorm on $\boldsymbol{W}_h^k$

$$\|\boldsymbol{z}_h\|_{1,h}^2 := \sum_{i=1}^{d} \|z_{h,i}\|_{1,h}^2 \tag{2.8}$$

and we remind the reader that the scalar version of the $\|\cdot\|_{1,h}$-norm defined by (1.39) is such that, for all $z_h = (v_{h,i}, \mu_{h,i}) \in W_h^k$,

$$\|z_h\|_{1,h}^2 := \sum_{T \in \mathcal{T}_h} \|z_T\|_{1,T}^2, \qquad \|z\|_{1,T}^2 := \|\boldsymbol{\nabla} v_T\|_T^2 + \sum_{F \in \mathcal{F}_T} h_F^{-1} \|\mu_F - v_T\|_F^2 \quad \forall T \in \mathcal{T}_h.$$

The consistency properties of the bilinear form $\mathcal{A}_h$ are summarized in the following lemma.

Lemma 2.1.1 (Consistency of $\mathcal{A}_h$)**.** *There is $C > 0$ independent of h such that, for all $\boldsymbol{u} = (u_i)_{1 \leqslant i \leqslant d} \in \boldsymbol{W} \cap H^{k+2}(\Omega)^d$, it holds*

$$\sup_{\boldsymbol{z}_h = (v_{h,i}, \mu_{h,i})_{1 \leqslant i \leqslant d} \in \boldsymbol{W}_h^k,\ \|\boldsymbol{z}_h\|_{1,h} = 1} \left\{ \sum_{i=1}^{d} (\triangle u_i, v_{h,i}) + \mathcal{A}_h(I_{\boldsymbol{W},h}^k \boldsymbol{u}, \boldsymbol{z}_h) \right\} \leqslant C h^{k+1} \|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}. \tag{2.9}$$

Proof. This is a straightforward consequence of Theorem 1.5.1. The proof is not repeated here for the sake of conciseness. □

2.1.3 Velocity-pressure coupling

For all $T \in \mathcal{T}_h$, we define the local discrete divergence operator $\mathcal{D}_T^k : \boldsymbol{W}_T^k \to \mathbb{P}^k(T)$ such that, for all $\boldsymbol{z}_T = (v_{T,i}, (\mu_{F,i})_{F \in \mathcal{F}_T})_{1 \leqslant i \leqslant d} \in \boldsymbol{W}_T^k$,

$$(\mathcal{D}_T^k \boldsymbol{z}_T, q)_T = \sum_{i=1}^{d} \left\{ -(v_{T,i}, \partial_i q)_T + \sum_{F \in \mathcal{F}_T} (\mu_{F,i} n_{TF,i}, q)_F \right\} \qquad \forall q \in \mathbb{P}^k(T), \tag{2.10}$$

where ∂_i denotes the partial derivative with respect to the ith space variable. In the context of lowest-order methods for the Stokes problem, this formula for the divergence has been used, e.g., in [21, 62]. In the higher-order case, it is essentially analogous (up to the choice of the discretization space for the velocity) to the one of [68, Section 4]. We record the following equivalent expression for $\mathcal{D}_T^k$ obtained integrating by parts the first term in (2.10):

$$(\mathcal{D}_T^k \boldsymbol{z}_T, q)_T = \sum_{i=1}^d \left\{ (\partial_i v_{T,i}, q)_T + \sum_{F \in \mathcal{F}_T} ((\mu_{F,i} - v_{T,i}) n_{TF,i}, q)_F \right\} \qquad \forall q \in \mathbb{P}^k(T). \quad (2.11)$$

The velocity-pressure coupling hinges on the global discrete divergence operator $\mathcal{D}_h^k : \boldsymbol{W}_h^k \to \mathcal{P}_h^k$ such that, for all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$,

$$(\mathcal{D}_h^k \boldsymbol{z}_h)_{|T} = \mathcal{D}_T^k \boldsymbol{z}_T \qquad \forall T \in \mathcal{T}_h. \quad (2.12)$$

Remark 2.1.2 (Interpretation of $\mathcal{D}_h^k$ vs. D_h^k). The operator $\mathcal{D}_h^k$ defined by (2.12) can be regarded as the discrete counterpart of the divergence operator defined from $\boldsymbol{W} = (H_0^1(\Omega))^d$ to $P = L_0^2(\Omega)$ (cf. (2.2)), as opposed to the operator D_h^k defined by (1.25), which discretizes the divergence from $\boldsymbol{\Sigma} = \mathbf{H}(\mathrm{div}; \Omega)$ to $U = L^2(\Omega)$ (cf. (1.4)).

The following commuting property is key to the inf-sup stability of the velocity-pressure coupling.

Proposition 2.1.3 (Commuting property for $\mathcal{D}_h^k$). *Let, for all $T \in \mathcal{T}_h$,*

$$\boldsymbol{W}(T) := W(T)^d$$

where we recall that $W(T) = \{v \in H^1(T) \mid v_{|\partial T \cap \partial\Omega} = 0\}$ (cf. (1.45)). Then, we have the following commuting diagrams:

$$\begin{array}{ccc} \boldsymbol{W}(T) & \xrightarrow{\boldsymbol{\nabla}\cdot} & L^2(T) \\ \downarrow{\scriptstyle I_{\boldsymbol{W},T}^k} & & \downarrow{\scriptstyle \pi_T^k} \\ \boldsymbol{W}_T^k & \xrightarrow{\mathcal{D}_T^k} & \mathbb{P}^k(T) \end{array} \qquad \begin{array}{ccc} \boldsymbol{W} & \xrightarrow{\boldsymbol{\nabla}\cdot} & P \\ \downarrow{\scriptstyle I_{\boldsymbol{W},h}^k} & & \downarrow{\scriptstyle \pi_h^k} \\ \boldsymbol{W}_h^k & \xrightarrow{\mathcal{D}_h^k} & \mathcal{P}_h^k \end{array}$$

Proof. Let $\boldsymbol{z} = (z_1, \ldots, z_d) \in \boldsymbol{W}(T)$. Using the definition (2.10) of $\mathcal{D}_T^k$ and (2.5) of

$I^k_{\boldsymbol{W},T}$, one has for all $q \in \mathbb{P}^k(T)$,

$$\begin{aligned}(\mathcal{D}^k_T(I^k_{\boldsymbol{W},T}\boldsymbol{z}),q)_T &= \sum_{i=1}^d \left\{ -(\pi^k_T z_i, \partial_i q)_T + \sum_{F\in\mathcal{F}_T} (\pi^k_F z_i\ n_{TF,i}, q)_F \right\} \\ &= \sum_{i=1}^d \left\{ -(z_i, \partial_i q)_T + \sum_{F\in\mathcal{F}_T} (z_i\ n_{TF,i}, \partial_i q)_F \right\} \\ &= (\boldsymbol{\nabla}\cdot\boldsymbol{z}, q)_T = (\pi^k_T(\boldsymbol{\nabla}\cdot\boldsymbol{z}), q)_T,\end{aligned}$$

where we have used, for all $1 \leqslant i \leqslant d$, that $\partial_i q \in \mathbb{P}^{k-1}(T) \subset \mathbb{P}^k(T)$ and $(q\ n_{T,i})_{|F} \in \mathbb{P}^k(F)$ for all $F \in \mathcal{F}_T$ (recall that faces are (hyper)planar by assumption), together with the definitions π^k_T and π^k_F to cancel the projectors in the second like, an integration by parts to pass to the third, and the definition of π^k_T to conclude. This proves the commuting property expressed by the first diagram. Recalling the definition (2.12) of $\mathcal{D}^k_h$ and (1.11) of π^k_h, and observing that, for all $\boldsymbol{z} \in \boldsymbol{W}$, $\mathcal{D}^k_h(I^k_{\boldsymbol{W},h}\boldsymbol{z})$ has zero average on Ω since $\boldsymbol{z}$ vanishes on $\partial\Omega$ concludes the proof. □

Lemma 2.1.4 (Consistency of the pressure-velocity coupling)**.** *There is $C > 0$ independent of h such that, for all $p \in P \cap H^{k+1}(\Omega)$,*

$$\sup_{\boldsymbol{z}_h=(v_{h,i},\mu_{h,i})_{1\leqslant i\leqslant d}\in\boldsymbol{W}^k_h,\ \|\boldsymbol{z}_h\|_{1,h}=1} \left\{ \sum_{i=1}^d (\partial_i p, v_{h,i}) + (p, \mathcal{D}^k_h\boldsymbol{z}_h) \right\} \leqslant C h^{k+1} \|p\|_{H^{k+1}(\Omega)}. \tag{2.13}$$

Proof. Let $\boldsymbol{z}_h \in \boldsymbol{W}^k_h$ be such that $\|\boldsymbol{z}_h\|_{1,h} = 1$. Integrating by parts element-by element, we can reformulate the first term inside the supremum as follows:

$$\sum_{i=1}^d (\partial_i p, v_{h,i}) = -\sum_{i=1}^d \sum_{T\in\mathcal{T}_h} \left\{ (p, \partial_i v_{T,i})_T + \sum_{F\in\mathcal{F}_T} (p, (\mu_{F,i} - v_{T,i}) n_{TF,i})_F \right\},$$

where we have used the fact that, by the regularity assumption, the jumps of p vanish across interfaces while, by definition of $\boldsymbol{W}^k_h$, $\mu_{F,i} = 0$ for all $1 \leqslant i \leqslant d$ and all $F \in \mathcal{F}^{\rm b}_h$ to insert the term

$$\sum_{i=1}^d \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} (p, \mu_{F,i} n_{TF,i})_F = 0.$$

On the other hand, using the definition (2.11) of $\mathcal{D}^k_T$ with $q = \pi^k_T p$ for all $T \in \mathcal{T}_h$,

we have for the second term

$$(p, \mathcal{D}_h^k \boldsymbol{z}_h) = (\pi_h^k p, \mathcal{D}_h^k \boldsymbol{z}_h) = \sum_{i=1}^d \sum_{T\in\mathcal{T}_h} \left\{ (\pi_T^k p, \partial_i v_{T,i})_T + \sum_{F\in\mathcal{F}_T} (\pi_T^k p, (\mu_{F,i} - v_{T,i}) n_{TF,i})_F \right\}.$$

Using the above relations, we infer

$$\begin{aligned}
&\left| \sum_{i=1}^d (\partial_i p, v_{h,i}) + (p, \mathcal{D}_h^k \boldsymbol{z}_h) \right| \\
&\quad = \left| \sum_{i=1}^d \sum_{T\in\mathcal{T}_h} \left\{ \cancel{(\pi_T^k p - p, \partial_i v_{T,i})_T} + \sum_{F\in\mathcal{F}_T} (\pi_T^k p - p, (\mu_{F,i} - v_{T,i}) n_{TF,i})_F \right\} \right| \\
&\quad \leqslant \left\{ \sum_{T\in\mathcal{T}_h} h_T \|\pi_T^k p - p\|_{\partial T}^2 \right\}^{1/2} \times \|\boldsymbol{z}_h\|_{1,h} \\
&\quad \lesssim h^{k+1} \|p\|_{H^{k+1}(\Omega)},
\end{aligned}$$

where we have observed that $\partial_i v_{T,i} \in \mathbb{P}^{k-1}(T)$ and used the definition of π_T^k to cancel the first term in the second line, used the Cauchy–Schwarz inequality together with the definition (2.8) of the $\|\cdot\|_{1,h}$-norm to pass to the third line, and the approximation properties (1.9) of π_h^k to conclude. □

2.1.4 Discrete problem and well-posedness

The discretization of the Stokes problem (2.3) reads: Find $(\boldsymbol{w}_h, p_h) \in \boldsymbol{W}_h^k \times \mathcal{P}_h^k$ such that

$$\mathcal{A}_h(\boldsymbol{w}_h, \boldsymbol{z}_h) - (p_h, \mathcal{D}_h^k \boldsymbol{z}_h) = L_h(\boldsymbol{z}_h) \qquad \forall \boldsymbol{z}_h \in \boldsymbol{W}_h^k, \tag{2.14a}$$

$$(\mathcal{D}_h^k \boldsymbol{w}_h, q_h) = 0 \qquad \forall q_h \in \mathcal{P}_h^k, \tag{2.14b}$$

where the linear form L_h on $\boldsymbol{W}_h^k$ is such that, for all $\boldsymbol{z}_h = (v_{h,i}, \mu_{h,i})_{1\leqslant i\leqslant d}$,

$$L_h(\boldsymbol{z}_h) = \sum_{i=1}^d (f_i, v_{h,i}). \tag{2.15}$$

Next, we prove that problem (2.14) is well-posed. A key point is that the velocity-pressure coupling is inf–sup stable.

Lemma 2.1.5 (Well-posedness of problem (2.14))**.** *There exists a real number* $\gamma_{\rm st} >$

0 *independent of* h *such that, for all* $q_h \in \mathcal{P}_h^k$*, the following inf-sup condition holds:*

$$\gamma_{\mathrm{st}}\|q_h\| \leqslant \sup_{\boldsymbol{z}_h \in \boldsymbol{W}_h^k \setminus \{\mathbf{0}_{\boldsymbol{W},h}\}} \frac{(\mathcal{D}_h^k \boldsymbol{z}_h, q_h)}{\|\boldsymbol{z}_h\|_{1,h}}. \tag{2.16}$$

Additionally, problem (2.14) *is well-posed.*

Proof. The proof proceeds in two steps: first, we prove that $I_{\boldsymbol{W},h}^k$ is a bounded operator, then we use classical techniques based on the commuting diagram property of Proposition 2.1.3 to prove the inf-sup condition.

(i) $\|\cdot\|_{1,h}$*-boundedness of* $I_{\boldsymbol{W},h}^k$. Let $\boldsymbol{z} = (z_1, \ldots, z_d) \in \boldsymbol{W}$. Using the H^1-stability of π_T^k (cf. [53, Appendix A]), the discrete trace inequality (1.7), and the trace approximation properties of π_T^k, we have that for all $1 \leqslant i \leqslant d$,

$$\begin{aligned}
\|I_{W,h}^k z_i\|_{1,h}^2 &= \sum_{T\in\mathcal{T}_h} \Big\{ \|\boldsymbol{\nabla} \pi_T^k z_i\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|\pi_F^k (z_i - \pi_T^k z_i)\|_F^2 \Big\} \\
&\leqslant \sum_{T\in\mathcal{T}_h} \Big\{ \|\boldsymbol{\nabla} z_i\|_T^2 + \sum_{F\in\mathcal{F}_T} h_F^{-1} \|z_i - \pi_T^k z_i\|_F^2 \Big\} \\
&\lesssim \sum_{T\in\mathcal{T}_h} \Big\{ \|\boldsymbol{\nabla} z_i\|_T^2 + h_T^{-2} \|z_i - \pi_T^k z_i\|_T^2 \Big\} \\
&\lesssim \sum_{T\in\mathcal{T}_h} \|\boldsymbol{\nabla} z_i\|_T^2 = \|\boldsymbol{\nabla} z_i\|^2.
\end{aligned}$$

Thus by summing over all components of $\boldsymbol{z}$ and recalling (2.8), we finally get

$$\|I_{\boldsymbol{W},h}^k \boldsymbol{z}\|_{1,h} \lesssim \|\boldsymbol{\nabla} \boldsymbol{z}\|. \tag{2.17}$$

(ii) *Inf-sup condition* (2.16). Let now $q_h \in \mathcal{P}_h^k$. Using the surjectivity property of the divergence operator defined from $\boldsymbol{W}$ to P, we infer the existence of $\boldsymbol{v}_q \in \boldsymbol{W}$ such that $\boldsymbol{\nabla}\cdot\boldsymbol{v}_q = q_h$ with $\|\boldsymbol{\nabla}\boldsymbol{v}_q\| \lesssim \|q_h\|$. Thus, accounting for the boundedness result of the previous point, we have the following inequality:

$$\|I_{\boldsymbol{W},h}^k \boldsymbol{v}_q\|_{1,h} \lesssim \|\boldsymbol{\nabla}\boldsymbol{v}_q\| \lesssim \|q_h\|. \tag{2.18}$$

To prove (2.16), we then proceed as follows:

$$\begin{aligned}\|q_h\|^2 &= (\boldsymbol{\nabla}\cdot\boldsymbol{v}_q, q_h)\\ &= (\mathcal{D}_h^k I_{\boldsymbol{W},h}^k \boldsymbol{v}_q, q_h)\\ &\leqslant \left(\sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k\setminus\{\boldsymbol{0}_{\boldsymbol{W},h}\}} \frac{(\mathcal{D}_h^k\boldsymbol{z}_h, q_h)}{\|\boldsymbol{z}_h\|_{1,h}}\right) \times \|I_{\boldsymbol{W},h}^k\boldsymbol{v}_q\|_{1,h}\\ &\lesssim \left(\sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k\setminus\{\boldsymbol{0}_{\boldsymbol{W},h}\}} \frac{(\mathcal{D}_h^k\boldsymbol{z}_h, q_h)}{\|\boldsymbol{z}_h\|_{1,h}}\right) \times \|q_h\|\end{aligned}$$

where we have used the global commuting property for $\mathcal{D}_h^k$ to pass to the second line, a passage to the supremum in the third line, and (2.18) to conclude.

(iii) *Well-posedness of problem* (2.14). The well-posedness of problem (2.14) follows from an application of [69, Theorem 2.34] since $\mathcal{A}_h$ is coercive on $\boldsymbol{W}_h^k$ owing to (2.7) and the inf-sup condition (2.16) holds. □

Remark 2.1.6 (Static condensation for problem (2.14)). The size of the linear system corresponding to problem (2.32) can be significantly reduced by resorting to static condensation. Following the procedure hinted to in [2] and detailed in [60, Section 6.2], it can be shown that the only globally coupled variables are face DOFs for the velocity and the average value of the pressure in each element. As a result, after statically condensing all the other DOFs and eliminating the velocity unknowns on the (Dirichlet) boundary, the total unknown count yields

$$d\binom{k+d-1}{k}\operatorname{card}(\mathcal{F}_h^{\mathrm{i}}) + \operatorname{card}(\mathcal{T}_h).$$

2.1.5 Energy-norm error estimate

Lemma 2.1.7 (Basic error estimate)**.** *Let $(\boldsymbol{u}, p)\in\boldsymbol{W}\times P$ denote the unique solution to* (2.3)*, and let $(\widehat{\boldsymbol{w}}_h, \widehat{p}_h) := (I_{\boldsymbol{W},h}^k\boldsymbol{u}, \pi_h^k p)$. Then, denoting by $(\boldsymbol{w}_h, p_h)\in\boldsymbol{W}_h^k\times\mathcal{P}_h^k$ the unique solution to* (2.14)*, the following holds with $\|\cdot\|_{\mathcal{A},h}$-norm defined by* (2.7)*:*

$$\max\left(\frac{\gamma_{\mathrm{st}}\eta^{1/2}}{2}\|p_h-\widehat{p}_h\|, \|\boldsymbol{w}_h-\widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h}\right) \leqslant \sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k\setminus\{\boldsymbol{0}_{\boldsymbol{W},h}\}} \frac{\mathcal{E}_h(\boldsymbol{z}_h)}{\|\boldsymbol{z}_h\|_{\mathcal{A},h}}, \tag{2.19}$$

where the consistency error is defined as

$$\mathcal{E}_h(\boldsymbol{z}_h) := L_h(\boldsymbol{z}_h) + (\widehat{p}_h, \mathcal{D}_h^k\boldsymbol{z}_h) - \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h). \tag{2.20}$$

Proof. We denote by \$ the supremum in the right-hand side of (2.19) and proceed to estimate the error on the velocity and on the pressure.

(i) *Error on the velocity.* Observe that

$$\mathcal{D}_h^k \boldsymbol{w}_h = \mathcal{D}_h^k \widehat{\boldsymbol{w}}_h = 0$$

as a consequence of the discrete mass equation (2.14b) and the right commuting diagram in Proposition 2.1.3 together with the continuous mass equation (2.1b), respectively. As a result, making $\boldsymbol{z}_h = \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h$ in the discrete momentum equation (2.14a), and recalling the definition of the consistency error $\mathcal{E}_h$, one has

$$\begin{aligned}
\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h}^2 &= \mathcal{A}_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h, \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) \\
&= \mathcal{A}_h(\boldsymbol{w}_h, \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) - \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) \\
&= L_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) + \cancel{(p_h, \mathcal{D}_h^k(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h))} - \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) \\
&= \mathcal{E}_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) - \cancel{(\widehat{p}_h, \mathcal{D}_h^k(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h))} \\
&\leqslant \$\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h},
\end{aligned} \tag{2.21}$$

hence,

$$\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h} \leqslant \$.$$

(ii) *Error on the pressure.* Let us now estimate the error on the pressure. Using (2.14a) together with the definition of the consistency error yields, for all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$,

$$(p_h - \widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h) = (p_h, \mathcal{D}_h^k \boldsymbol{z}_h) - (\widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h) = \mathcal{A}_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h) - \mathcal{E}_h(\boldsymbol{z}_h).$$

Using the inf-sup condition (2.16) for $q_h = p_h - \widehat{p}_h$ together with the above relation followed by (2.21), the Cauchy–Schwarz inequality, and the second inequality in (2.7), it is inferred that

$$\gamma_{\mathrm{st}} \eta^{1/2} \|p_h - \widehat{p}_h\| \leqslant \sup_{\boldsymbol{z}_h \in \boldsymbol{W}_h^k \setminus \{\boldsymbol{0}_{\boldsymbol{W},h}\}} \frac{(p_h - \widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h)}{\eta^{-1/2} \|\boldsymbol{z}_h\|_{1,h}} \leqslant \|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h} + \$ \leqslant 2\$. \tag{2.22}$$

The estimate (2.19) follows from (2.21)–(2.22). □

Theorem 2.1.8 (Convergence rate for the energy-norm of the error)**.** *Under the assumptions and notations of Lemma 2.1.7, and assuming the additional regularity*

$\boldsymbol{u} \in H^{k+2}(\Omega)^d$ *and* $p \in H^{k+1}(\Omega)$*, the following holds:*

$$\max\left(\frac{\gamma_{\mathrm{st}}\eta^{1/2}}{2}\|p_h - \widehat{p}_h\|, \|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A},h}\right) \leqslant Ch^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)}\right), \quad (2.23)$$

with real number $C > 0$ *independent of* h*.*

Proof. It suffices to bound the consistency error $\mathcal{E}_h(\boldsymbol{z}_h)$ in (2.19) for a generic $\boldsymbol{z}_h = (v_{h,i}, \mu_{h,i})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_h^k$. Observing that $f_i = -\triangle u_i + \partial_i p$ for all $1 \leqslant i \leqslant d$ a.e. in Ω, we have that

$$\mathcal{E}_h(\boldsymbol{z}_h) = \underbrace{-\sum_{i=1}^d(\triangle u_i, v_{h,i}) - \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h)}_{\mathfrak{T}_1} + \underbrace{\sum_{i=1}^d(\partial_i p, v_{h,i}) + (\widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h)}_{\mathfrak{T}_2}.$$

For the first term, the consistency (2.9) of the viscous bilinear form $\mathcal{A}_h$ yields

$$|\mathfrak{T}_1| \lesssim h^{k+1}\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}\|\boldsymbol{z}_h\|_{1,h}.$$

For the second term, we use the consistency (2.13) of the discrete velocity-pressure coupling to infer

$$|\mathfrak{T}_2| \lesssim h^{k+1}\|p\|_{H^{k+1}(\Omega)}\|\boldsymbol{z}_h\|_{1,h}.$$

Using the above bounds and recalling the coercivity of $\mathcal{A}_h$ expressed by the first inequality of (2.7) to infer $\|\boldsymbol{z}_h\|_{1,h} \lesssim \|\boldsymbol{z}_h\|_{\mathcal{A},h}$, we get

$$|\mathcal{E}_h(\boldsymbol{z}_h)| \lesssim h^{k+1}\left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)}\right)\|\boldsymbol{z}_h\|_{\mathcal{A},h}.$$

Plugging the above bound into the error estimate (2.19) yields the desired result. □

2.1.6 L^2-norm error estimate for the velocity

We can obtain a sharp estimate for the L^2-norm of the error on the velocity assuming further regularity for problem (2.1). We assume in this section that Cattabriga's regularity holds (cf. [4, 38]) in the following form: There is C_{Cat} only depending on Ω such that, for all $\boldsymbol{g} \in L^2(\Omega)^d$, denoting by $(\boldsymbol{z}, r) \in \boldsymbol{W} \times P$ the unique solution to

$$(\boldsymbol{\nabla}\boldsymbol{z}, \boldsymbol{\nabla}\boldsymbol{v}) - (r, \boldsymbol{\nabla}\cdot\boldsymbol{v}) = (\boldsymbol{g}, \boldsymbol{v}) \qquad \forall \boldsymbol{v} \in \boldsymbol{W}, \quad (2.24a)$$

$$(\boldsymbol{\nabla}\cdot\boldsymbol{z}, q) = 0 \qquad \forall q \in P, \quad (2.24b)$$

it holds that

$$\|\boldsymbol{z}\|_{H^2(\Omega)^d} + \|r\|_{H^1(\Omega)} \leqslant C_{\mathrm{Cat}}\|\boldsymbol{g}\|. \tag{2.25}$$

The following result shows that supercloseness holds for the velocity element DOFs, which converge with order $(k+2)$ to the L^2-orthogonal projection of the velocity on the broken polynomial space $\mathbb{P}^k(\mathcal{T}_h)^d$.

Theorem 2.1.9 (Convergence rate for the L^2-norm of the error on the velocity)**.** *Under the assumptions and notations of Theorem 2.1.8, and assuming that Cattabriga's regularity* (2.25) *holds and that* $\boldsymbol{f} \in H^k(\Omega)^d$*, there exists a real number* $C > 0$ *independent of* h *such that, if* $k \geqslant 1$*,*

$$\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\| \leqslant Ch^{k+2}\left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)} + \|\boldsymbol{f}\|_{H^k(\Omega)^d}\right). \tag{2.26}$$

For $k = 0$*, further assuming that* $\boldsymbol{f} \in H^1(\Omega)^d$*,*

$$\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\| \leqslant Ch^2\|\boldsymbol{f}\|_{H^1(\Omega)^d}, \tag{2.27}$$

where $\boldsymbol{u}_h, \widehat{\boldsymbol{u}}_h \in \mathbb{P}^k(\mathcal{T}_h)^d$ *are obtained from element unknowns setting, for all* $T \in \mathcal{T}_h$*,*

$$\boldsymbol{u}_{h|T} = (u_{T,i})_{1\leqslant i\leqslant d}, \qquad \widehat{\boldsymbol{u}}_{h|T} = (\widehat{u}_{T,i})_{1\leqslant i\leqslant d}.$$

Proof. Let $(\boldsymbol{z}, r) \in \boldsymbol{W} \times P$ solve (2.24) with $\boldsymbol{g} = \widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h$, set $\widehat{\boldsymbol{z}}_h := I^k_{\boldsymbol{W},h}\boldsymbol{z}$, and define the error on the velocity

$$\boldsymbol{e}_h := \widehat{\boldsymbol{w}}_h - \boldsymbol{w}_h = \left((\epsilon_{T,i})_{T\in\mathcal{T}_h}, (\rho_{F,i})_{F\in\mathcal{F}_h}\right)_{1\leqslant i\leqslant d} \in \boldsymbol{W}^k_h.$$

We also introduce the following vector-valued quantities obtained from the element and face DOFs of $\boldsymbol{e}_h$, respectively:

$$\boldsymbol{\epsilon}_T = (\epsilon_{T,i})_{1\leqslant i\leqslant d} \quad \forall T \in \mathcal{T}_h \qquad \text{and} \qquad \boldsymbol{\rho}_F = (\rho_{F,i})_{1\leqslant i\leqslant d} \quad \forall F \in \mathcal{F}_h.$$

Using the fact that $-\triangle\boldsymbol{z} + \boldsymbol{\nabla} r = \widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h = \boldsymbol{\epsilon}_h$ a.e. in Ω, it holds for all $T \in \mathcal{T}_h$, integrating by parts and exploiting the flux continuity and the fact that $\boldsymbol{\rho}_F \equiv \boldsymbol{0}$ for all $F \in \mathcal{F}_h^{\mathrm{b}}$ to insert the term $0 = \sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\boldsymbol{\rho}_F, (\boldsymbol{\nabla}\boldsymbol{z} - r\boldsymbol{I}_d)\boldsymbol{n}_{TF})_F$,

$$\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\|^2 = \sum_{T\in\mathcal{T}_h}\left\{(\boldsymbol{\nabla}\boldsymbol{\epsilon}_T, \boldsymbol{\nabla}\boldsymbol{z} - r\boldsymbol{I}_d)_T + \sum_{F\in\mathcal{F}_T}((\boldsymbol{\rho}_F - \boldsymbol{\epsilon}_T), (\boldsymbol{\nabla}\boldsymbol{z} - r\boldsymbol{I}_d)\boldsymbol{n}_{TF})_F\right\}.$$

Adding to the above expression the quantity (cf. (2.14a))

$$0 = \mathcal{A}_h(\boldsymbol{w}_h, \widehat{\boldsymbol{z}}_h) - (p_h, \mathcal{D}_h^k \widehat{\boldsymbol{z}}_h) - \sum_{T\in\mathcal{T}_h} (\boldsymbol{f}, \pi_T^k \boldsymbol{z})_T = \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \widehat{\boldsymbol{z}}_h) - \mathcal{A}_h(\boldsymbol{e}_h, \widehat{\boldsymbol{z}}_h) - \sum_{T\in\mathcal{T}_h} (\boldsymbol{f}, \pi_T^k \boldsymbol{z})_T,$$

where we have used Proposition 2.1.3 together with (2.24b) to infer $\mathcal{D}_h^k \widehat{\boldsymbol{z}}_h = \pi_h^k(\boldsymbol{\nabla}\cdot\boldsymbol{z}) = 0$, we have

$$\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\|^2 = \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3, \tag{2.28}$$

with

$$\begin{aligned}
\mathfrak{T}_1 &:= \sum_{T\in\mathcal{T}_h} \left\{ (\boldsymbol{\nabla}\boldsymbol{\epsilon}_T, \boldsymbol{\nabla}\boldsymbol{z})_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{\rho}_F - \boldsymbol{\epsilon}_T, \boldsymbol{\nabla}\boldsymbol{z}\boldsymbol{n}_{TF})_F \right\} - \mathcal{A}_h(\boldsymbol{e}_h, \widehat{\boldsymbol{z}}_h),\\
\mathfrak{T}_2 &:= -\sum_{T\in\mathcal{T}_h} \left\{ (\boldsymbol{\nabla}\cdot\boldsymbol{\epsilon}_T, r)_T + \sum_{F\in\mathcal{F}_T} ((\boldsymbol{\rho}_F - \boldsymbol{\epsilon}_T)\cdot\boldsymbol{n}_{TF}, r)_F \right\},\\
\mathfrak{T}_3 &:= \mathcal{A}_h(\widehat{\boldsymbol{w}}_h, \widehat{\boldsymbol{z}}_h) - \sum_{T\in\mathcal{T}_h} (\boldsymbol{f}, \pi_T^k \boldsymbol{z})_T.
\end{aligned}$$

To bound $\mathfrak{T}_1$ we recall the definitions (2.6) of $\mathcal{A}_h$ and (1.60) of A_h, and observe that, with $\boldsymbol{\delta}_T := \left(z_{i|T} - r_T^k I_{W,T}^k z_{i|T}\right)_{1\leqslant i\leqslant d}$,

$$\mathfrak{T}_1 = \sum_{T\in\mathcal{T}_h} \left\{ (\boldsymbol{\nabla}\boldsymbol{\epsilon}_T, \boldsymbol{\nabla}\boldsymbol{\delta}_T)_T + \sum_{F\in\mathcal{F}_T} (\boldsymbol{\rho}_F - \boldsymbol{\epsilon}_T, \boldsymbol{\nabla}\boldsymbol{\delta}_T\boldsymbol{n}_{TF})_F \right\} + \mathcal{J}(\boldsymbol{e}_h, \widehat{\boldsymbol{z}}_h),$$

where, for the sake of brevity, we have introduced the bilinear form $\mathcal{J}(\boldsymbol{w}_h, \boldsymbol{v}_h) := \sum_{i=1}^d J(w_{h,i}, v_{h,i})$. Hence, we infer

$$\begin{aligned}
|\mathfrak{T}_1| &\leqslant \left\{ \|\boldsymbol{e}_h\|_{1,h}^2 + \mathcal{J}(\boldsymbol{e}_h, \boldsymbol{e}_h) \right\}^{1/2} \times \left\{ \sum_{T\in\mathcal{T}_h} \left[\|\boldsymbol{\nabla}\boldsymbol{\delta}_T\|_T^2 + h_T \|\boldsymbol{\nabla}\boldsymbol{\delta}_T\|_{\partial T}^2 \right] + \mathcal{J}(\widehat{\boldsymbol{z}}_h, \widehat{\boldsymbol{z}}_h) \right\}^{1/2}\\
&\leqslant h^{k+1} \left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)} \right) h \|\boldsymbol{z}\|_{H^2(\Omega)^d}\\
&\lesssim h^{k+2} \left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)} \right) \|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|,
\end{aligned} \tag{2.29}$$

where we have used the Cauchy–Schwarz inequality followed by the energy estimate (2.23) for the first factor, while, for the second factor, we have estimated $\boldsymbol{\delta}_T$ using (1.56), $\mathcal{J}(\widehat{\boldsymbol{z}}_h, \widehat{\boldsymbol{z}}_h)$ as the term $\mathfrak{T}_3$ in the proof of Theorem 2.1.8, and we have used Cattabriga's regularity (2.25) for $\boldsymbol{z}$ to conclude.

To estimate $\mathfrak{T}_2$, we observe that

$$\mathcal{D}_h^k \boldsymbol{e}_h = \mathcal{D}_h^k \widehat{\boldsymbol{w}}_h - \mathcal{D}_h^k \boldsymbol{w}_h = 0$$

owing to Proposition 2.1.3 together with (2.3b) and (2.14b), hence, letting $r_h := \pi_h^k r$ and using (2.11) with $\boldsymbol{z} = R^k_{\boldsymbol{W},T}\boldsymbol{e}_h$ and $q = r_T$, we infer

$$0 = (\mathcal{D}_h^k\boldsymbol{e}_h, r_h) = \sum_{T\in\mathcal{T}_h}\left\{(\boldsymbol{\nabla}\cdot\boldsymbol{\epsilon}_T, r_T)_T + \sum_{F\in\mathcal{F}_T}((\boldsymbol{\rho}_F - \boldsymbol{\epsilon}_T)\cdot\boldsymbol{n}_{TF}, r_T)_F\right\}.$$

Subtracting the above expression from $\mathfrak{T}_2$, and using the Cauchy–Schwarz inequality together with the bound (1.6) on $\mathfrak{N}_\partial$, it is inferred

$$\begin{aligned}|\mathfrak{T}_2| &\lesssim \|\boldsymbol{e}_h\|_{1,h}\left\{\sum_{T\in\mathcal{T}_h}\left[\|r - r_T\|_T^2 + h_T\|r - r_T\|_{\partial T}^2\right]\right\}^{1/2}\\ &\lesssim h^{k+2}\left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)}\right)\|r\|_{H^1(\Omega)},\end{aligned} \tag{2.30}$$

where we have used the first inequality in (2.7) together with the energy estimate (2.23) for the first factor and the approximation properties (1.9) of π_h^k for the second.

Let us now estimate $\mathfrak{T}_3$. For all $T\in\mathcal{T}_h$, we have $(\boldsymbol{f}, \pi_T^k\boldsymbol{z})_T = (\pi_T^k\boldsymbol{f}, \boldsymbol{z})_T$. Moreover, since $(\boldsymbol{f}, \boldsymbol{z}) = (\boldsymbol{\nabla}\boldsymbol{u} - p\boldsymbol{I}_d, \boldsymbol{\nabla}\boldsymbol{z})$ and, owing to (2.12), $(\pi_h^k p, \mathcal{D}_h^k\widehat{\boldsymbol{z}}_h) = (p, \pi_h^k(\boldsymbol{\nabla}\cdot\boldsymbol{z})) = (\pi_h^k p, \boldsymbol{\nabla}\cdot\boldsymbol{z})$, we infer

$$\begin{aligned}\mathfrak{T}_3 = {}&(\boldsymbol{f} - \pi_h^k\boldsymbol{f}, \boldsymbol{z})\\ &- \sum_{T\in\mathcal{T}_h}\left\{\sum_{i=1}^d\left[(\boldsymbol{\nabla}u_i, \boldsymbol{\nabla}z_i)_T - (\boldsymbol{G}_T^k I_{W,T}^k u_i, \boldsymbol{G}_T^k I_{W,T}^k z_i)\right] - (p - \pi_h^k p, \boldsymbol{\nabla}\cdot\boldsymbol{z})\right\}\\ &+ \mathcal{J}(\widehat{\boldsymbol{w}}_h, \widehat{\boldsymbol{z}}_h).\end{aligned}$$

Denote by $\mathfrak{T}_{3,1}, \mathfrak{T}_{3,2}, \mathfrak{T}_{3,3}$ the addends in the right-hand side. If $k\geqslant 1$, we can write

$$(\boldsymbol{f} - \pi_h^k\boldsymbol{f}, \boldsymbol{z}) = (\boldsymbol{f} - \pi_h^k\boldsymbol{f}, \boldsymbol{z} - \pi_h^1\boldsymbol{z}),$$

hence

$$|\mathfrak{T}_{3,1}| \lesssim h^k\|\boldsymbol{f}\|_{H^k(\Omega)^d}h^2\|\boldsymbol{z}\|_{H^2(\Omega)^d} \lesssim h^{k+2}\|\boldsymbol{f}\|_{H^k(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|_{H^2(\Omega)^d}.$$

On the other hand, for $k = 0$, we write $(\boldsymbol{f} - \pi_h^0\boldsymbol{f}, \boldsymbol{z} - \pi_h^0\boldsymbol{z})$ so that

$$|\mathfrak{T}_{3,1}| \lesssim h\|\boldsymbol{f}\|_{H^1(\Omega)^d}h\|\boldsymbol{z}\|_{H^1(\Omega)^d} \lesssim h^2\|\boldsymbol{f}\|_{H^1(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|.$$

To estimate $\mathfrak{T}_{3,2}$ we use the orthogonality property (1.54) to infer

$$\mathfrak{T}_{3,2} = \sum_{T\in\mathcal{T}_h}\sum_{i=1}^{d}\left[(\boldsymbol{\nabla} u_i - \boldsymbol{G}_T^k I_{W,T}^k u_i, \boldsymbol{\nabla} z_i - \boldsymbol{G}_T^k I_{W,T}^k z_i)\right],$$

hence, recalling (1.56) and using Cattabriga's regularity (2.25) for $\boldsymbol{z}$, it is inferred

$$|\mathfrak{T}_{3,2}| \lesssim h^{k+2}\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|.$$

Finally, using the Cauchy–Schwarz inequality, proceeding as for the estimate of $\mathfrak{T}_3$ in the proof of Theorem 2.1.8, and recalling again (2.25), it is inferred

$$\begin{aligned}|\mathfrak{T}_{3,3}| &\leqslant \mathcal{J}(\widehat{\boldsymbol{w}}_h, \widehat{\boldsymbol{w}}_h)^{1/2}\mathcal{J}(\widehat{\boldsymbol{z}}_h, \widehat{\boldsymbol{z}}_h)^{1/2}\\ &\lesssim h^{k+1}\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}h\|\boldsymbol{z}\|_{H^2(\Omega)^d}\\ &\lesssim h^{k+2}\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|.\end{aligned}$$

Gathering the above estimates, we infer for $k \geqslant 1$,

$$|\mathfrak{T}_3| \lesssim h^{k+2}\left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|\boldsymbol{f}\|_{H^k(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|,\right)$$

while, for $k = 0$, using Cattabriga's regularity for $\boldsymbol{u}$, we get

$$|\mathfrak{T}_3| \lesssim h^{k+2}\|\boldsymbol{f}\|_{H^1(\Omega)^d}\|\widehat{\boldsymbol{u}}_h - \boldsymbol{u}_h\|.$$

Using the above bounds for $\mathfrak{T}_3$ in conjunction with (2.29) and (2.30) to estimate the right-hand side of (2.28), and invoking Cattabriga's regularity for $(\boldsymbol{u}, p)$ when $k = 0$, gives the desired result. □

To close this section, we exhibit a discrete velocity reconstruction that converges with order $(k+2)$ to the exact velocity $\boldsymbol{u}$. Let, for all $T \in \mathcal{T}_h$, $\boldsymbol{r}_T^k : \boldsymbol{W}_T^k \to \mathbb{P}^{k+1}(T)^d$ denote the velocity reconstruction operator such that, for all $\boldsymbol{w} \in \boldsymbol{W}_T^k$,

$$\boldsymbol{r}_T^k\boldsymbol{w} = (r_T^k w_i)_{1\leqslant i\leqslant d}$$

with r_T^k defined by (1.55), and define its global counterpart $\boldsymbol{r}_h^k : \boldsymbol{W}_h^k \to \mathbb{P}^{k+1}(\mathcal{T}_h)^d$ such that, for all $\boldsymbol{w}_h \in \boldsymbol{W}_h^k$,

$$\boldsymbol{r}_h^k\boldsymbol{w}_{h|T} = \boldsymbol{r}_T^k(R_{\boldsymbol{W},T}^k\boldsymbol{w}_h), \qquad \forall T \in \mathcal{T}_h.$$

Corollary 2.1.10 (Convergence of $\boldsymbol{r}_h^k\boldsymbol{w}_h$)**.** *Using the notation of Theorem 2.1.8,*

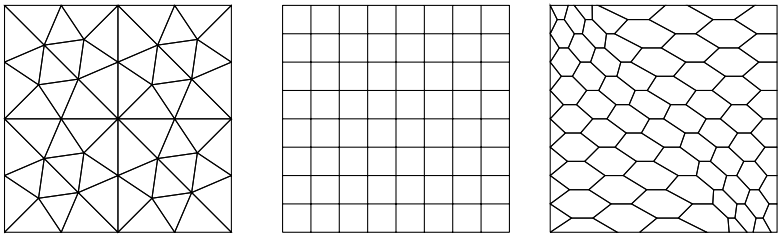


Figure 2.1 – Triangular (Tria), Cartesian (Cart) and hexagonal (Hex) mesh families for the numerical example of Section 2.1.7

and under the assumptions of Theorem 2.1.9, there is a real number C independent of h such that

$$\|\boldsymbol{u} - \boldsymbol{r}_h^k \boldsymbol{w}_h\| \leqslant C h^{k+2} \left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)} + \|\boldsymbol{f}\|_{H^k(\Omega)^d} \right).$$

Proof. Recalling that $\widehat{\boldsymbol{w}}_h = I_{\boldsymbol{W},h}^k \boldsymbol{u}$, and using the triangular inequality, one has

$$\|\boldsymbol{u} - \boldsymbol{r}_h^k \boldsymbol{w}_h\| \leqslant \|\boldsymbol{u} - \boldsymbol{r}_h^k \widehat{\boldsymbol{w}}_h\| + \|\boldsymbol{r}_h^k (\widehat{\boldsymbol{w}}_h - \boldsymbol{w}_h)\| := \mathfrak{T}_1 + \mathfrak{T}_2.$$

As a result of (1.56) it is readily inferred $|\mathfrak{T}_1| \lesssim h^{k+2} \|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d}$. Additionally, we estimate the second term $\mathfrak{T}_2$ by adding and removing $\pi_T^0(\widehat{\boldsymbol{w}}_h - \boldsymbol{w}_h)$ combined with the triangle inequality and using (1.9) such that we get

$$\begin{aligned} \mathfrak{T}_2 &= \sum_{T \in \mathcal{T}_h} \|\boldsymbol{r}_T^k R_{\boldsymbol{W},T}^k (\widehat{\boldsymbol{w}}_h - \boldsymbol{w}_h)\|_T^2 \\ &\lesssim \sum_{T \in \mathcal{T}_h} \left\{ h_T^2 \|\boldsymbol{\nabla} \boldsymbol{r}_T^k R_{\boldsymbol{W},T}^k (\widehat{\boldsymbol{w}}_h - \boldsymbol{w}_h)\|_T^2 + \|\pi_T^0 (\widehat{\boldsymbol{u}}_T - \boldsymbol{u}_T)\|_T \right\}. \end{aligned}$$

Estimating the first term between braces using (1.56), observing, for the second, that it holds

$$\|\pi_T^0 (\widehat{\boldsymbol{u}}_T - \boldsymbol{u}_T)\|_T \leqslant \|\widehat{\boldsymbol{u}}_T - \boldsymbol{u}_T\|_T$$

since π_T^0 is bounded as a projector, and recalling (2.26), we infer

$$|\mathfrak{T}_2| \lesssim h^{k+2} \left(\|\boldsymbol{u}\|_{H^{k+2}(\Omega)^d} + \|p\|_{H^{k+1}(\Omega)} + \|\boldsymbol{f}\|_{H^k(\Omega)^d} \right).$$

The desired result follows. □

2.1.7 Numerical examples

We solve the Stokes problem (2.1) on the unit square $\Omega = (0,1)^2$ with $\boldsymbol{f} \equiv \boldsymbol{0}$ and Dirichlet boundary conditions inferred from the following exact solution:

$$\boldsymbol{u}(x,y) = \big(-\exp(x)(y\cos y + \sin y), \exp(x)(y\sin y)\big), \qquad p = 2\exp(x)\sin(y) - p_0,$$

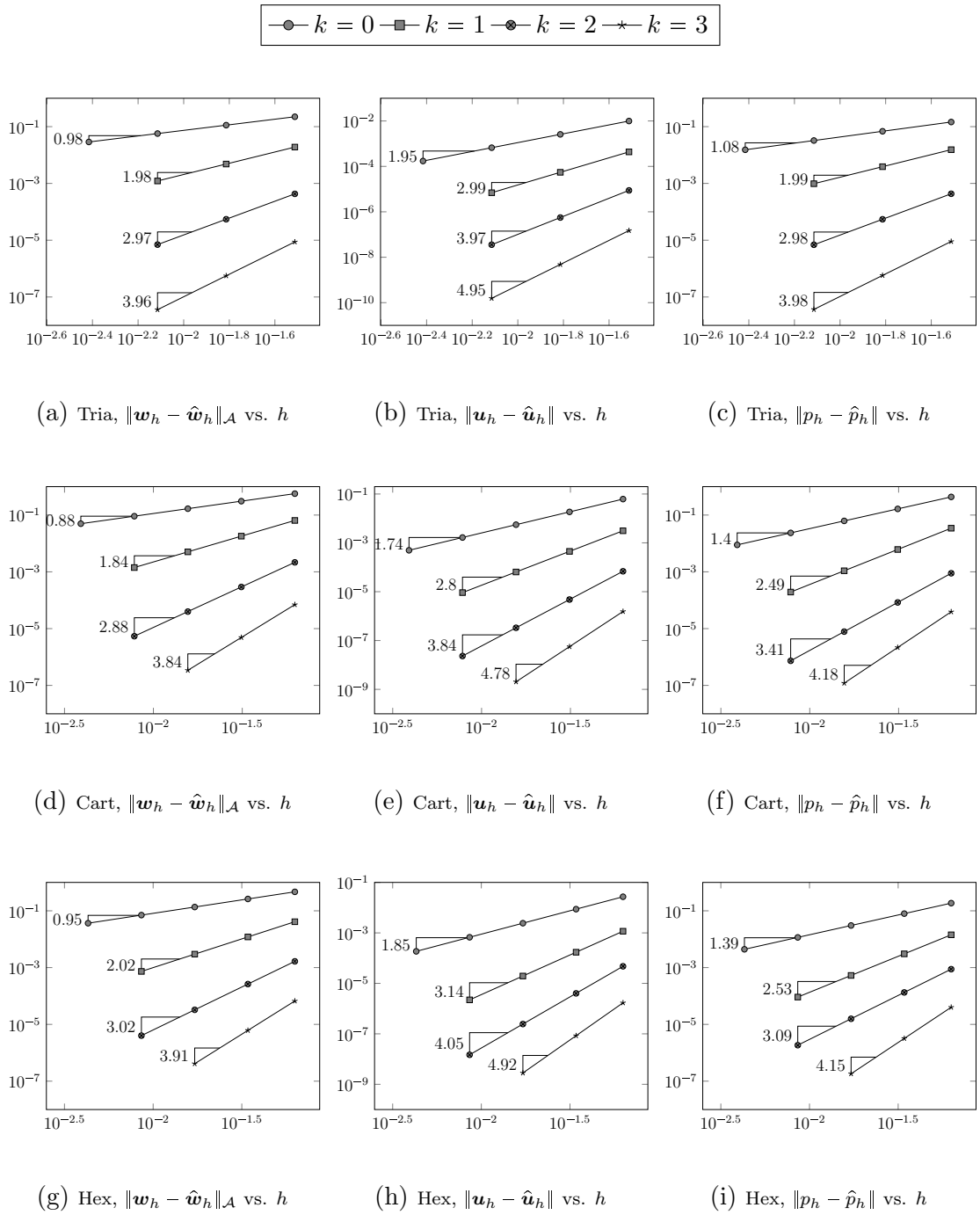
where $p_0 \in \mathbb{R}$ is chosen so as to ensure $\int_\Omega p = 0$. We consider the three mesh families depicted in Figure 2.1. The triangular and Cartesian mesh families correspond, respectively, to the mesh families 1 and 2 of the FVCA5 benchmark [80], whereas the (predominantly) hexagonal mesh family was first introduced in [62].

Figure 2.2 displays convergence results for the different meshes and polynomial degrees up to 3. Following (2.19), we display the $\|\cdot\|_{\mathcal{A},h}$-norm of the error in the velocity as well as the L^2-norm of the error both in the velocity and in the pressure. In all the cases, the numerical results match the order estimates predicted by the theory (in some cases, a slight superconvergence is observed for the pressure at the lowest orders).

Local computations are based on the linear algebra facilities provided by the boost uBLAS library [83]. The local linear systems are solved using the Cholesky factorization available in uBLAS. The global system (involving face unknowns only) is solved using SuperLU [51] through the PETSc 3.4 interface [15]. The tests have been run sequentially on a laptop computer powered by an Intel Core i7-3520 CPU clocked at 2.90 GHz and equipped with 8Gb of RAM.

2.2 A robust discretization of the Oseen problem

In this section, we extend the method (2.14) to the Oseen problem. Let $\nu \in \mathbb{R}_+^*$ denote a constant kinematic viscosity, $\boldsymbol{f} \in L^2(\Omega)^d$ a volumetric body force, $\boldsymbol{\beta} \in \mathrm{Lip}(\Omega)^d$ a given velocity field such that $\boldsymbol{\nabla}\cdot\boldsymbol{\beta} = 0$ in Ω, and $\mu \in \mathbb{R}_+$ a reaction coefficient. We consider the Oseen problem that consists in seeking the velocity



(a) Tria, $\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A}}$ vs. h (b) Tria, $\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\|$ vs. h (c) Tria, $\|p_h - \widehat{p}_h\|$ vs. h

(d) Cart, $\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A}}$ vs. h (e) Cart, $\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\|$ vs. h (f) Cart, $\|p_h - \widehat{p}_h\|$ vs. h

(g) Hex, $\|\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h\|_{\mathcal{A}}$ vs. h (h) Hex, $\|\boldsymbol{u}_h - \widehat{\boldsymbol{u}}_h\|$ vs. h (i) Hex, $\|p_h - \widehat{p}_h\|$ vs. h

Figure 2.2 – Convergence results for the numerical example of Section 2.1.7 on the mesh families of Figure 2.1. The notation is the same as in Theorems 2.1.8 and 2.1.9

field $\boldsymbol{u} : \Omega \to \mathbb{R}^d$ and the pressure field $p : \Omega \to \mathbb{R}$ such that

$$-\nu\triangle\boldsymbol{u} + (\boldsymbol{\beta}\cdot\boldsymbol{\nabla})\boldsymbol{u} + \mu\boldsymbol{u} + \boldsymbol{\nabla}p = \boldsymbol{f} \qquad \text{in } \Omega, \tag{2.31a}$$

$$\boldsymbol{\nabla}\cdot\boldsymbol{u} = 0 \qquad \text{in } \Omega, \tag{2.31b}$$

$$\boldsymbol{u} = \boldsymbol{0} \qquad \text{on } \partial\Omega, \tag{2.31c}$$

$$\int_\Omega p = 0. \tag{2.31d}$$

Notice that the reaction term is introduced here mainly to simplify the expressions of some multiplicative constants appearing in the analysis, and we do not consider the case when this term is dominant.

The main difficulty consists here in writing an appropriate discretization of the advective term, robust also when advection is dominant. Following [54], this is achieved by

(i) introducing a discrete counterpart of the directional (advective) derivative $\boldsymbol{\beta}\cdot\boldsymbol{\nabla}$ which reproduces at the discrete level a suitable integration by parts formula;

(ii) adding an upwind stabilization term which acts between element- and face-unknowns.

A key point is that static condensation in the spirit of Remark 2.1.6 remains possible for the resulting method, which makes its implementation very efficient. These developments are original, and have not been published elsewhere. Numerical tests are undergoing and will be included in a forthcoming paper.

To alleviate the notation, throughout this section we often abridge by $a \lesssim b$ the inequality $a \leqslant Cb$ with real number $C > 0$ independent of h, ν, $\boldsymbol{\beta}$, and μ. As in the previous section, named constants are used in the statements for the sake of easy consultation.

2.2.1 Discrete problem

The HHO discretization of the Oseen problem (2.31) is obtained modifying the scheme (2.14) to account for the presence of the kinematic viscosity and the advective-reactive terms. Specifically, the discrete problem now reads: Find $(\boldsymbol{w}_h, p_h) \in$

$\boldsymbol{W}_h^k \times \mathcal{P}_h^k$ such that

$$\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h) - (p_h, \mathcal{D}_h^k \boldsymbol{z}_h) = L_h(\boldsymbol{z}_h) \qquad \forall \boldsymbol{z}_h \in \boldsymbol{W}_h^k, \tag{2.32a}$$

$$(\mathcal{D}_h^k \boldsymbol{w}_h, q_h) = 0 \qquad \forall q_h \in \mathcal{P}_h^k, \tag{2.32b}$$

where the discrete global divergence operator $\mathcal{D}_h^k$ is given by (2.12), the linear form L_h on $\boldsymbol{W}_h^k$ by (2.15), while the bilinear form $\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}$ on $\boldsymbol{W}_h^k \times \boldsymbol{W}_h^k$ results from the assembly of the viscous and advective-reactive contributions:

$$\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h) := \mathcal{A}_{\nu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h) + \mathcal{A}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h). \tag{2.33}$$

For the viscous contribution, we simply set

$$\mathcal{A}_{\nu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h) := \nu \mathcal{A}_h(\boldsymbol{w}_h, \boldsymbol{z}_h), \tag{2.34}$$

where the bilinear form $\mathcal{A}_h$ on $\boldsymbol{W}_h^k \times \boldsymbol{W}_h^k$ is defined by (2.6). The bilinear form $\mathcal{A}_{\boldsymbol{\beta},\mu,h}$, on the other hand, is defined by element-by-element assembly of local contributions as

$$\mathcal{A}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h, \boldsymbol{z}_h) := \sum_{T \in \mathcal{T}_h} \mathcal{A}_{\boldsymbol{\beta},\mu,T}(\boldsymbol{w}_T, \boldsymbol{z}_T). \tag{2.35}$$

The precise definition of the local contribution $\mathcal{A}_{\boldsymbol{\beta},\mu,T}$ for a generic mesh element $T \in \mathcal{T}_h$ will be the object of the two following subsections.

For the sake of conciseness, from this point on for a given $\boldsymbol{z}_h = ((v_{T,i})_{T \in \mathcal{T}_h}, (\mu_{F,i})_{F \in \mathcal{F}_T})_{1 \leqslant i \leqslant d} \in \boldsymbol{W}_h^k$ we use the following shortcut notation for the vector-valued fields obtained from element-based and face-based DOFs, respectively:

$$\begin{aligned} \boldsymbol{v}_T &:= (v_{T,1}, \ldots, v_{T,d}) \in \mathbb{P}^k(T)^d \qquad \forall T \in \mathcal{T}_h, \\ \boldsymbol{\mu}_F &:= (\mu_{F,1}, \ldots, \mu_{F,d}) \in \mathbb{P}^k(F)^d \qquad \forall F \in \mathcal{F}_T. \end{aligned} \tag{2.36}$$

This notation carries out verbatim when considering restriction $\boldsymbol{z}_T$ of $\boldsymbol{z}_h$ to a generic mesh element $T \in \mathcal{T}_h$.

2.2.2 Discrete advective derivative

Let an element $T \in \mathcal{T}_h$ be fixed and set, for all $F \in \mathcal{F}_T$,

$$\beta_{TF} := \boldsymbol{\beta}_{|F} \cdot \boldsymbol{n}_{TF}.$$

A useful remark is that, by the regularity of $\boldsymbol{\beta}$, for all $F \in \mathcal{F}_h^{\mathrm{i}}$ such that $F \subset \partial T_1 \cap \partial T_2$,

$$\beta_{T_1F} + \beta_{T_2F} = 0. \tag{2.37}$$

We define the local advective derivative reconstruction operator $\boldsymbol{G}_{\boldsymbol{\beta},T}^k : \boldsymbol{W}_T^k \to \mathbb{P}^k(T)^d$ such that, for all $\boldsymbol{z}_T = (v_{T,i}, (\mu_{F,i})_{F\in\mathcal{F}_T})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_T^k$ and all $\boldsymbol{w} \in \mathbb{P}^k(T)^d$, using the shortcut notation introduced in (2.36),

$$\begin{aligned}(\boldsymbol{G}_{\boldsymbol{\beta},T}^k \boldsymbol{z}_T, \boldsymbol{w})_T &:= ((\boldsymbol{\beta}\cdot\boldsymbol{\nabla})\boldsymbol{v}_T, \boldsymbol{w})_T + \sum_{F\in\mathcal{F}_T} (\beta_{TF}(\boldsymbol{\mu}_F - \boldsymbol{v}_T), \boldsymbol{w})_F \\ &= -(\boldsymbol{v}_T, (\boldsymbol{\beta}\cdot\boldsymbol{\nabla})\boldsymbol{w})_T + \sum_{F\in\mathcal{F}_T} (\beta_{TF}\boldsymbol{\mu}_F, \boldsymbol{w})_F,\end{aligned} \tag{2.38}$$

where we have used integration by parts and $\boldsymbol{\nabla}\cdot\boldsymbol{\beta} = 0$ to pass to the second line. In the following proposition we prove a discrete counterpart of the following integration by parts formula: For all $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{W}$,

$$(\boldsymbol{\beta}\cdot\boldsymbol{\nabla}\boldsymbol{u}, \boldsymbol{v}) + (\boldsymbol{u}, \boldsymbol{\beta}\cdot\boldsymbol{\nabla}\boldsymbol{v}) = 0. \tag{2.39}$$

This relation will be later used in Proposition 2.2.5 to prove the stability of the advective-reactive term.

Proposition 2.2.1 (Discrete integration by parts for the advective term)**.** *For all $\boldsymbol{z}_h = ((v_{T,i})_{T\in\mathcal{T}_h}, (\mu_{F,i})_{F\in\mathcal{F}_h})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_h^k$ and all $\boldsymbol{w}_h = ((u_{T,i})_{T\in\mathcal{T}_h}, (\lambda_{F,i})_{F\in\mathcal{F}_h})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_h^k$, using the shortcut notation defined in* (2.36)*, it holds*

$$\begin{aligned}\sum_{T\in\mathcal{T}_h} \{(\boldsymbol{G}_{\boldsymbol{\beta},T}^k \boldsymbol{z}_T, \boldsymbol{u}_T)_T + (\boldsymbol{v}_T, \boldsymbol{G}_{\boldsymbol{\beta},T}^k \boldsymbol{w}_T)_T\} = \\ - \sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} (\beta_{TF}(\boldsymbol{\mu}_F - \boldsymbol{v}_T), \boldsymbol{\lambda}_F - \boldsymbol{u}_T)_F.\end{aligned} \tag{2.40}$$

Proof. We have

$$\begin{aligned}&\sum_{T\in\mathcal{T}_h} (\boldsymbol{G}_{\boldsymbol{\beta},T}^k \boldsymbol{z}_T, \boldsymbol{u}_T)_T \\ &= \sum_{T\in\mathcal{T}_h} \left\{ ((\boldsymbol{\beta}\cdot\boldsymbol{\nabla})\boldsymbol{v}_T, \boldsymbol{u}_T)_T + \sum_{F\in\mathcal{F}_T} (\beta_{TF}(\boldsymbol{\mu}_F - \boldsymbol{v}_T), \boldsymbol{u}_T)_F \right\} \\ &= \sum_{T\in\mathcal{T}_h} \left\{ -(\boldsymbol{v}_T, \boldsymbol{G}_{\boldsymbol{\beta},T}^k \boldsymbol{w}_T)_T + \sum_{F\in\mathcal{F}_T} (\beta_{TF}(\boldsymbol{\mu}_F - \boldsymbol{v}_T), \boldsymbol{u}_T)_F + \sum_{F\in\mathcal{F}_T} (\beta_{TF}\boldsymbol{\lambda}_F, \boldsymbol{v}_T)_F, \right\},\end{aligned}$$

where we have used the definition (2.38) of $\boldsymbol{G}^k_{\boldsymbol{\beta},T}\boldsymbol{z}_T$ with $\boldsymbol{u}_T$ as a test function in the first line and the definition (2.38) of $\boldsymbol{G}^k_{\boldsymbol{\beta},T}\boldsymbol{w}_T$ with $\boldsymbol{v}_T$ as a test function in the second line. The proof is concluded adding in the right hand side of the above expression

$$-\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\beta_{TF}\boldsymbol{\lambda}_F,\boldsymbol{\mu}_F)_F=0.$$

The fact that this quantity is zero can be easily proved using the fact that $\boldsymbol{\mu}_F$ vanishes for all $F\in\mathcal{F}_h^{\mathrm{b}}$ together with the continuity of the normal component of the velocity on interfaces expressed by (2.37). □

Remark 2.2.2 (Discrete integration by parts for the advective term). In the continuous integration by parts formula (2.39), the right-hand side is zero owing to (2.37) combined with the fact that $\boldsymbol{u}\in\boldsymbol{W}$ implies that the jumps of $\boldsymbol{u}$ vanish across interfaces and $\boldsymbol{u}$ is zero on $\partial\Omega$. By contrast, in the discrete counterpart (2.40) the right-hand side accounts for the difference between element-based and face-based unknowns. It is precisely because of this difference that we will need to introduce an upwind stabilization term.

2.2.3 Local advective-reactive contribution

We are now ready to define the local contribution to the advective-reactive bilinear form (cf. (2.35)): For all $\boldsymbol{w}_T=(u_{T,i},(\lambda_{F,i})_{F\in\mathcal{F}_T})_{1\leqslant i\leqslant d}\in\boldsymbol{W}_T^k$ and all $\boldsymbol{z}_T=(v_{T,i},(\mu_{F,i})_{F\in\mathcal{F}_T})_{1\leqslant i\leqslant d}\in\boldsymbol{W}_T^k$, using the shortcut notation (2.36), we let

$$\mathcal{A}_{\boldsymbol{\beta},\mu,T}(\boldsymbol{w}_T,\boldsymbol{z}_T):=-(\boldsymbol{u}_T,\boldsymbol{G}^k_{\boldsymbol{\beta},T}\boldsymbol{z}_T)_T+s_{\boldsymbol{\beta},T}(\boldsymbol{w}_T,\boldsymbol{z}_T)+\mu(\boldsymbol{u}_T,\boldsymbol{v}_T)_T,\tag{2.41}$$

where the upwind stabilization bilinear form $s_{\boldsymbol{\beta},T}$ is such that, with $\beta^{\ominus}_{TF}:=\frac{|\beta_{TF}|-\beta_{TF}}{2}$,

$$s_{\boldsymbol{\beta},T}(\boldsymbol{w}_T,\boldsymbol{z}_T):=\sum_{F\in\mathcal{F}_T}(\beta^{\ominus}_{TF}(\boldsymbol{\lambda}_F-\boldsymbol{u}_T),\boldsymbol{\mu}_F-\boldsymbol{v}_T)_F.$$

Remark 2.2.3 (Static condensation for problem (2.32)). The global advective-reactive bilinear form defined by (2.35) with local contributions given by (2.41) has the same stencil as the viscous contribution defined by (2.34). It can be proved that static condensation as in Remark 2.1.6 can be performed also for the discrete Oseen problem (2.32). A crucial point to preserve the possibility of statically condensing all element-based velocity DOFs and all but one pressure DOF per element is that the upwind stabilization acts between element-based and face-based DOFs (and not, as in finite volume or discontinuous Galerkin methods, between element-based DOFs

of adjoining elements).

2.2.4 Well-posedness

In this section we carry out the stability analysis for the HHO method (2.32) and prove that the resulting problem is well-posed.

Diffusive-advective-reactive norm

For all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$, we define the following diffusive-advective-reactive norm on $\boldsymbol{W}_h^k$:

$$|||\boldsymbol{z}_h|||_h^2 := \|\boldsymbol{z}_h\|_{\nu,h}^2 + \|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}^2, \tag{2.42}$$

where, recalling the definition (2.34) of $\mathcal{A}_{\nu,h}$,

$$\|\boldsymbol{z}_h\|_{\nu,h}^2 := \mathcal{A}_{\nu,h}(\boldsymbol{z}_h,\boldsymbol{z}_h) \quad \text{and} \quad \|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}^2 := \sum_{T\in\mathcal{T}_h} \|\boldsymbol{z}_T\|_{\boldsymbol{\beta},\mu,T}^2,$$

with, for all $T \in \mathcal{T}_h$, all $\boldsymbol{z}_T = (v_{T,i},(\mu_{F,i})_{F\in\mathcal{F}_T})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_T^k$ and $\boldsymbol{v}_T$ and $\boldsymbol{\mu}_F$ defined according to (2.36),

$$\|\boldsymbol{z}_T\|_{\boldsymbol{\beta},\mu,T}^2 := \frac{1}{2}\sum_{F\in\mathcal{F}_T} \||\beta_{TF}|^{1/2}(\boldsymbol{\mu}_F-\boldsymbol{v}_T)\|_F^2 + \tau_{\mathrm{ref},T}^{-1}\|\boldsymbol{v}_T\|_T^2. \tag{2.43}$$

In (2.43), $\tau_{\mathrm{ref},T}$ denotes the reference time such that

$$\tau_{\mathrm{ref},T} := \max(\mu, L_{\boldsymbol{\beta},T})^{-1}, \qquad L_{\boldsymbol{\beta},T} := \max_{1\leqslant i\leqslant d} \|\boldsymbol{\nabla}\beta_i\|_{L^\infty(T)^d}.$$

The norms $\|\cdot\|_{\nu,h}$ and $\|\cdot\|_{1,h}$ are uniformly equivalent on $\boldsymbol{W}_h^k$ thanks to (2.7). More precisely, as consequence of the definition (2.34) of the viscous bilinear form $\mathcal{A}_{\nu,h}$ together with the coercivity (2.7) of $\mathcal{A}_h$, it holds for all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$,

$$\nu^{1/2}\|\boldsymbol{z}_h\|_{1,h} \lesssim \|\boldsymbol{z}_h\|_{\nu,h} \lesssim \nu^{1/2}\|\boldsymbol{z}_h\|_{1,h}. \tag{2.44}$$

The fact that both the maps $\|\cdot\|_{\nu,h}$ and $|||\cdot|||_h$ define norms on $\boldsymbol{W}_h^k$ is then an immediate consequence.

We next show that the $|||\cdot|||_h$-norm can be bounded in terms of the $\|\cdot\|_{1,h}$-norm. This

bound is needed in the proof of the inf-sup condition in Lemma 2.2.7 below. We need to define the following local and global Peclet numbers:

$$\mathrm{Pe}_T := \max_{F\in\mathcal{F}_T} \frac{\|\beta_{TF}\|_{L^\infty(F)} h_T}{\nu} \quad \forall T\in\mathcal{T}_h, \qquad \mathrm{Pe}_h := \max_{T\in\mathcal{T}_h} \mathrm{Pe}_T. \tag{2.45}$$

We also introduce the global reference time such that

$$\tau_{\mathrm{ref},h}^{-1} := \max_{T\in\mathcal{T}_h} \tau_{\mathrm{ref},T}^{-1}.$$

Proposition 2.2.4 (Bound for the $|||\cdot|||_h$-norm)**.** *There is a real number $C>0$ independent of h such that, for all $\boldsymbol{z}_h\in\boldsymbol{W}_h^k$, it holds*

$$|||\boldsymbol{z}_h|||_h \le C\left[\nu(1+\mathrm{Pe}_h)+\tau_{\mathrm{ref},h}^{-1}\right]^{1/2} \|\boldsymbol{z}_h\|_{1,h}. \tag{2.46}$$

Proof. Let an element $\boldsymbol{z}_h=((v_{T,i})_{T\in\mathcal{T}_h},(\mu_{F,i})_{F\in\mathcal{F}_h})\in\boldsymbol{W}_h^k$ be fixed. The bound

$$\|\boldsymbol{z}_h\|_{\nu,h}^2 \lesssim \nu\|\boldsymbol{z}_h\|_{1,h}^2 \tag{2.47}$$

is an immediate consequence of (2.44). Let now a mesh element $T\in\mathcal{T}_h$ be fixed, denote by $\boldsymbol{z}_T$ the restriction of $\boldsymbol{z}_h$ to T, and recall the shortcut notation (2.36). By definition (2.45) of the local Peclet number Pe_T, it is readily inferred that

$$\frac12\sum_{F\in\mathcal{F}_T} \||\beta_{TF}|^{1/2}(\boldsymbol{\mu}_F-\boldsymbol{v}_T)\|_F^2 \le \frac12\nu\mathrm{Pe}_T\sum_{F\in\mathcal{F}_T} h_T^{-1}\|\boldsymbol{\mu}_F-\boldsymbol{v}_T\|_F^2 \lesssim \nu\mathrm{Pe}_T\|\boldsymbol{z}_T\|_{1,T}^2.$$

Summing over $T\in\mathcal{T}_h$, we conclude that

$$\frac12\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T} \||\beta_{TF}|^{1/2}(\boldsymbol{\mu}_F-\boldsymbol{v}_T)\|_F^2 \lesssim \nu\mathrm{Pe}_h\|\boldsymbol{z}_h\|_{1,h}^2.$$

On the other hand, the Poincaré inequality for hybrid spaces proved in [53, Proposition 5.4] yields

$$\sum_{T\in\mathcal{T}_h} \tau_{\mathrm{ref},T}^{-1}\|\boldsymbol{v}_T\|_T^2 \lesssim \tau_{\mathrm{ref},h}^{-1}\|\boldsymbol{z}_h\|_{1,h}^2.$$

From the above relations we get

$$\|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}^2 \lesssim \left(\nu\mathrm{Pe}_h+\tau_{\mathrm{ref},h}^{-1}\right)\|\boldsymbol{z}_h\|_{1,h}^2,$$

which, combined with (2.47) concludes the proof. □

Stability and well-posedness

To prove the well-posedness of the discrete problem (2.32), we use a similar argument as in the proof of Lemma 2.1.5 based on the $\|\!|\cdot|\!\|_h$-coercivity of the diffusive-advective-reactive bilinear form $\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}$ defined by (2.33) and the inf-sup stability of the pressure-velocity coupling. A preliminary result is the coercivity of the advective-reactive bilinear form defined by (2.35).

Proposition 2.2.5 (Coercivity of $\mathcal{A}_{\boldsymbol{\beta},\mu,h}$)**.** *It holds for all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$,*

$$\varsigma\|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}^2 \leqslant \mathcal{A}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h,\boldsymbol{z}_h), \tag{2.48}$$

where

$$\varsigma := \min_{T\in\mathcal{T}_h}(1,\tau_{\mathrm{ref},T}\mu). \tag{2.49}$$

Corollary 2.2.6 (Coercivity of $\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}$)**.** *There is a real number $C>0$ independent of h, ν, $\boldsymbol{\beta}$, μ such that, for all $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$,*

$$C(1+\varsigma)|\!|\!|\boldsymbol{z}_h|\!|\!|_h^2 \leqslant \mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h,\boldsymbol{z}_h), \tag{2.50}$$

with ς given by (2.49)*.*

Proof of Proposition 2.2.5. Let $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$, denote by $\boldsymbol{z}_T = (v_{T,i},(\mu_{F,i})_{F\in\mathcal{F}_T})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_T^k$ its restriction to $T\in\mathcal{T}_h$, and recall the shortcut notation introduced in (2.36). Using (2.40) with $\boldsymbol{z}_h = \boldsymbol{w}_h$, we infer

$$-\sum_{T\in\mathcal{T}_h}(\boldsymbol{v}_T,\boldsymbol{G}_{\boldsymbol{\beta},T}^k\boldsymbol{z}_T) = \frac12\sum_{T\in\mathcal{T}_h}\sum_{F\in\mathcal{F}_T}(\beta_{TF}(\boldsymbol{\mu}_F-\boldsymbol{v}_T),\boldsymbol{\mu}_F-\boldsymbol{v}_T)_F.$$

Using this relation we have

$$\begin{aligned}
&\mathcal{A}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h,\boldsymbol{z}_h)\\
&\quad=\sum_{T\in\mathcal{T}_h}\mathcal{A}_{\boldsymbol{\beta},\mu,T}(\boldsymbol{z}_T,\boldsymbol{z}_T)\\
&\quad=\sum_{T\in\mathcal{T}_h}\left\{-(\boldsymbol{v}_T,\boldsymbol{G}_{\boldsymbol{\beta},T}^k\boldsymbol{z}_T)+\sum_{F\in\mathcal{F}_T}(\beta_{TF}^{\ominus}(\boldsymbol{v}_T-\boldsymbol{\mu}_F),\boldsymbol{v}_T-\boldsymbol{\mu}_F)_F+\mu(\boldsymbol{v}_T,\boldsymbol{v}_T)_T\right\}\\
&\quad=\sum_{T\in\mathcal{T}_h}\left\{\frac12\sum_{F\in\mathcal{F}_T}(|\beta_{TF}|(\boldsymbol{v}_T-\boldsymbol{\mu}_F),(\boldsymbol{v}_T-\boldsymbol{\mu}_F))_F+\mu\|\boldsymbol{v}_T\|_T^2\right\}\\
&\quad=\sum_{T\in\mathcal{T}_h}\left\{\frac12\sum_{F\in\mathcal{F}_T}\||\beta_{TF}|^{1/2}(\boldsymbol{v}_T-\boldsymbol{\mu}_F)\|^2+\tau_{\mathrm{ref},T}^{-1}\tau_{\mathrm{ref},T}\mu\|\boldsymbol{v}_T\|^2\right\},
\end{aligned}$$

where, to pass to the third line we have observed that $\beta_{TF}^{\ominus} = \frac{1}{2}(|\beta_{TF}| - \beta_{TF})$. □

We are now ready to prove the main result of this section.

Lemma 2.2.7 (Well-posedness of problem (2.32))**.** *There is a real number $C > 0$ independent of h, ν, $\boldsymbol{\beta}$, and μ such that, for all $q_h \in \mathcal{P}_h^k$, the following inf-sup condition holds:*

$$\gamma_{\rm os}\|q_h\| \leqslant \sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k,|||\boldsymbol{z}_h|||_h\leqslant 1} (\mathcal{D}_h^k\boldsymbol{z}_h, q_h). \tag{2.51}$$

where $\gamma_{\rm os} := C\left[\nu(1+\mathrm{Pe}_h)+\tau_{{\rm ref},h}^{-1}\right]^{-1/2}$. Additionally, problem (2.32) is well-posed.

Proof. Let $q_h \in \mathcal{P}_h^k$ and $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$. Recalling (2.46) and the $\|\cdot\|_{1,h}$-boundedness of $I_{\boldsymbol{W},h}^k$ expressed by (2.17), we observe that it holds, for all $\boldsymbol{z} \in \boldsymbol{W}$,

$$|||I_{\boldsymbol{W},h}^k\boldsymbol{z}|||_h \lesssim \left[\nu(1+\mathrm{Pe}_h)+\tau_{{\rm ref},h}^{-1}\right]^{1/2}\|I_{\boldsymbol{W},h}^k\boldsymbol{z}\|_{1,h} \lesssim \left[\nu(1+\mathrm{Pe}_h)+\tau_{{\rm ref},h}^{-1}\right]^{1/2}\|\boldsymbol{\nabla}\boldsymbol{z}\|.$$

The proof of the inf-sup condition (2.51) then follows the reasoning of point (ii) in Lemma 2.1.5 replacing $\|\cdot\|_{1,h} \leftarrow |||\cdot|||_{1,h}$ and using the above relation in place of (2.17).

Finally, the well-posedness of problem (2.32) follows from (2.48) and (2.51) according to the classical theory of saddle-point problems; cf., e.g., [24]. □

2.2.5 Energy-norm error estimate

The goal of this section is to estimate the error between the solution $(\boldsymbol{w}_h, p_h) \in \boldsymbol{W}_h^k \times \mathcal{P}_h^k$ of the HHO scheme (2.32) with respect to the projection

$$(\widehat{\boldsymbol{w}}_h, \widehat{p}_h) = (I_{\boldsymbol{W},h}^k\boldsymbol{u}, \pi_h^k p) \in \boldsymbol{W}_h^k \times \mathcal{P}_h^k$$

of the weak solution $(\boldsymbol{u}, p)$ of the continuous Oseen problem (2.31).

Consistency of the advective-reactive bilinear form

In the following lemma, we study the consistency of the advective-reactive bilinear form $\mathcal{A}_{\boldsymbol{\beta},\mu,h}$ defined by (2.35) from the local contributions (2.41).

Lemma 2.2.8 (Consistency of $\mathcal{A}_{\boldsymbol{\beta},\mu,h}$)**.** *There exists $C > 0$ independent of $h, \nu, \boldsymbol{\beta}$ and μ such that, for all $\boldsymbol{u} = (u_1, \ldots, u_d) \in \boldsymbol{W} \cap H^{k+2}(\Omega)^d$, it holds*

$$\sup_{\substack{\boldsymbol{z}_h=(v_{h,i},\mu_{h,i})_{1\leqslant i\leqslant d}\in \boldsymbol{W}_h^k \\ |||\boldsymbol{z}_h|||_h=1}} \left\{ \sum_{i=1}^d [((\boldsymbol{\beta}\cdot\boldsymbol{\nabla})u_i, v_{h,i}) + (\mu u_i, v_{h,i})] - \mathcal{A}_{\boldsymbol{\beta},\mu,h}(\widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h) \right\}$$
$$\leqslant C \left\{ \sum_{T\in\mathcal{T}_h} \mathcal{N}_{1,T} h_T^{2(k+1)} + \mathcal{N}_{2,T} \min(\tfrac{1}{2}, \mathrm{Pe}_T) h_T^{2k+1} \right\}^{1/2}, \quad (2.52)$$

where $\mathcal{N}_{1,T} := \tau_{\mathrm{ref},T}^{-1} \|\boldsymbol{u}\|^2_{H^{k+1}(T)}$ and $\mathcal{N}_{2,T} := \|\boldsymbol{\beta}\|_{L^\infty(T)^d} \|\boldsymbol{u}\|^2_{H^{k+1}(T)}$.

Proof. We denote by $\mathcal{E}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h)$ the argument of the supremum. Let $\boldsymbol{z}_h = (v_{h,i}, \mu_{h,i})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_h^k$ and $\widehat{\boldsymbol{w}}_h = I^k_{\boldsymbol{W},h}\boldsymbol{u} = (\widehat{u}_{h,i}, \widehat{\lambda}_{h,i})_{1\leqslant i\leqslant d} \in \boldsymbol{W}_h^k$, where we remind the reader that $v_{h,i} = (v_{T,i})_{T\in\mathcal{T}_h}$, $\widehat{u}_{h,i} = (\widehat{u}_{T,i})_{T\in\mathcal{T}_h}$ and $\widehat{\lambda}_{h,i} = (\widehat{\lambda}_{F,i})_{F\in\mathcal{F}_h}$ for any $1 \leqslant i \leqslant d$. Integrating by parts the first term in $\mathcal{E}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h)$ and adding the quantity

$$0 = -\sum_{T\in\mathcal{T}_h} \sum_{F\in\mathcal{F}_T} (\beta_{TF}\boldsymbol{u}, \boldsymbol{v}_F)_F,$$

we have, expanding the definition (2.38) of the discrete advective derivative and of the upwind stabilization,

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h) &= \sum_{i=1}^d \left\{ ((\boldsymbol{\beta}\cdot\boldsymbol{\nabla})u_i, v_{h,i}) + (\mu u_i, v_{h,i}) \right\} - \mathcal{A}_{\boldsymbol{\beta},\mu,h}(\widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h) \\
&= \sum_{i=1}^d \left\{ \sum_{T\in\mathcal{T}_h} (\widehat{u}_{T,i} - u_i, \mu v_{T,i} + (\boldsymbol{\beta}\cdot\boldsymbol{\nabla})v_{T,i})_T + \sum_{F\in\mathcal{F}_T} (\beta_{TF}(\widehat{u}_{T,i} - u_i), \mu_{F,i} - v_{T,i})_F \right. \\
&\qquad \left. - \sum_{F\in\mathcal{F}_T} (\beta^{\ominus}_{TF}(\widehat{\lambda}_{F,i} - u_{T,i}), \mu_{F,i} - v_{T,i})_F \right\} \\
&:= \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3.
\end{aligned}$$

We use the same arguments as for the term $\mathfrak{T}_{2,1}, \mathfrak{T}_{2,2}$ and $\mathfrak{T}_{2,3}$ in the proof of [54, Theorem 10] for the scalar case. Recalling that $\widehat{u}_{T,i} = \pi_T^k u_i$ and observing that $(\pi_T^0\boldsymbol{\beta})\cdot\boldsymbol{\nabla}v_{T,i} \in \mathbb{P}^{k-1}(T) \subset \mathbb{P}^k(T)$, we have

$$\mathfrak{T}_1 = \sum_{i=1}^d \sum_{T\in\mathcal{T}_h} (\widehat{u}_{T,i} - u_i, \mu v_{T,i} + (\boldsymbol{\beta} - \pi_T^0\boldsymbol{\beta})\cdot\boldsymbol{\nabla}v_{T,i}).$$

We can now estimate the first term using repetitively the Cauchy-Schwarz inequality, inverse inequality (1.8), the definition of $\tau_{\mathrm{ref},T}$, the projection approxima-

tion estimate (1.9) and the Lipschitz continuity property of the advective velocity $\|\boldsymbol{\beta} - \pi_T^0\boldsymbol{\beta}\|_{L^\infty(T)^d} \leqslant L_{\boldsymbol{\beta},T} h_T$:

$$\begin{aligned}
|\mathfrak{T}_1| &\leqslant \sum_{i=1}^{d} \sum_{T\in\mathcal{T}_h} \|\widehat{u}_{T,i} - u_i\|_T \left\{\mu \|v_{T,i}\|_T + \|\boldsymbol{\beta} - \pi_T^0\boldsymbol{\beta}\|_{L^\infty(T)^d} \|\boldsymbol{\nabla} v_{T,i}\|_T\right\} \\
&\lesssim \sum_{i=1}^{d} \sum_{T\in\mathcal{T}_h} \tau_{\mathrm{ref},T}^{-1} h_T^{k+1} \|u_i\|_{H^{k+1}(T)} \|v_{T,i}\|_T \\
&\leqslant \left\{ \sum_{T\in\mathcal{T}_h} \tau_{\mathrm{ref},T}^{-1} h_T^{2(k+1)} \|\boldsymbol{u}\|_{H^{k+1}(T)}^2 \right\}^{1/2} \|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}.
\end{aligned} \tag{2.53}$$

The terms $\mathfrak{T}_2$ and $\mathfrak{T}_3$ are estimated using a decomposition strategy based on the local Peclet number Pe_{TF}. Precisely, we consider the following decomposition

$$\mathfrak{T}_2 + \mathfrak{T}_3 = \mathfrak{T}_2^{\mathrm{d}} + \mathfrak{T}_3^{\mathrm{d}} + \mathfrak{T}_2^{\mathrm{a}} + \mathfrak{T}_3^{\mathrm{a}},$$

where the superscript "d" stands for face integrals where $|\mathrm{Pe}_{TF}| \leqslant 1$ whereas the superscript "a" stands for face integrals where $|\mathrm{Pe}_{TF}| > 1$. This strategy allows us to bound the terms either with the diffusive or the advective part of the full norm $\|\!|\cdot|\!\|_h$ by following the exact same reasoning as in [54, step (ii) of Theorem 10]. On the one hand, for the diffusive part, we have

$$|\mathfrak{T}_2^{\mathrm{d}}| + |\mathfrak{T}_3^{\mathrm{d}}| \lesssim \left(\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\beta}\|_{L^\infty(T)^d} \min(1, \mathrm{Pe}_T) \|\boldsymbol{u}_{|T} - \widehat{\boldsymbol{u}}_T\|_F^2 \right)^{1/2} \|\boldsymbol{z}_h\|_{\nu,h}.$$

On the other hand, for the advective part, we obtain a similar estimate, but where the advective-reactive norm of $\boldsymbol{z}_h$ appears in place of its diffusive norm:

$$|\mathfrak{T}_2^{\mathrm{a}}| + |\mathfrak{T}_3^{\mathrm{a}}| \lesssim \left(\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\beta}\|_{L^\infty(T)^d} \min(1, \mathrm{Pe}_T) \|\boldsymbol{u}_{|T} - \widehat{\boldsymbol{u}}_T\|_F^2 \right)^{1/2} \|\boldsymbol{z}_h\|_{\boldsymbol{\beta},\mu,h}.$$

Using the approximation properties (1.9) of $\widehat{\boldsymbol{u}}_T = \pi_T^k \boldsymbol{u}$ we get for all $F \in \mathcal{F}_h$

$$\|\boldsymbol{u}_{|T} - \widehat{\boldsymbol{u}}_T\|_F^2 \leqslant C_{\mathrm{app}} h_T^{k+1/2} \|\boldsymbol{u}\|_{H^{k+1}(T)^d}.$$

Finally, gathering all the previous estimates, we arrive at

$$|\mathfrak{T}_2| + |\mathfrak{T}_3| \lesssim \left(\sum_{T\in\mathcal{T}_h} \|\boldsymbol{\beta}\|_{L^\infty(T)^d} \min(1, \mathrm{Pe}_T) h_T^{2k+1} \|\boldsymbol{u}\|_{H^{k+1}(T)^d}^2 \right)^{1/2} \|\!|\boldsymbol{z}_h|\!\|_h. \tag{2.54}$$

From (2.53) and (2.54), one finally obtains

$$|\mathcal{E}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h)| \lesssim \left\{ \sum_{T\in\mathcal{T}_h} \left[\tau_{\mathrm{ref},T}^{-1} h_T^{2(k+1)} + \|\boldsymbol{\beta}\|_{L^\infty(T)^d} \min(1, \mathrm{Pe}_T) h_T^{2k+1} \right] \|\boldsymbol{u}\|_{H^{k+1}(T)}^2 \right\}^{1/2} |||\boldsymbol{z}_h|||_h. \tag{2.55}$$

Taking the supremum of $\mathcal{E}_{\boldsymbol{\beta},\mu,h}(\boldsymbol{z}_h)$ over $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$ s.t. $|||\boldsymbol{z}_h|||_h = 1$ concludes the proof. □

Error estimate

Lemma 2.2.9 (Abstract error estimate)**.** *It holds*

$$\epsilon := \gamma_{\mathrm{os}} \|p_h - \widehat{p}_h\| + (1+\varsigma) |||\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h|||_h \lesssim \sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k, |||\boldsymbol{z}_h|||_h\leqslant 1} \mathcal{E}_h(\boldsymbol{z}_h), \tag{2.56}$$

with consistency error linear form

$$\mathcal{E}_h(\boldsymbol{z}_h) := L_h(\boldsymbol{z}_h) - \mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h) + (\mathcal{D}_h^k \boldsymbol{z}_h, \widehat{p}_h).$$

Proof. We denote by \$ the supremum in the right-hand side of (2.56). Using the coercivity of $\mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}$ from Corollary 2.2.6 and the same arguments as in the proof of Lemma 2.19, we infer

$$\begin{aligned} C(1+\varsigma) |||\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h|||_h^2 &\leqslant \mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h, \boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) \\ &\leqslant \mathcal{E}_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) + (p_h - \widehat{p}_h, \mathcal{D}_h^k(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h)) \\ &= \mathcal{E}_h(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h) \\ &\leqslant \$ |||\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h|||_h. \end{aligned} \tag{2.57}$$

Using the inf-sup property (2.51) with $q_h = p_h - \widehat{p}_h$, the relation $(p_h - \widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h) = \mathcal{A}_{\nu,\boldsymbol{\beta},\mu,h}(\boldsymbol{w}_h - \widehat{\boldsymbol{w}}_h, \boldsymbol{z}_h) - \mathcal{E}_h(\boldsymbol{z}_h)$ and the stability relation (2.2.5) we have,

$$\gamma_{\mathrm{os}} \|p_h - \widehat{p}_h\| \lesssim \sup_{\boldsymbol{z}_h\in\boldsymbol{W}_h^k, |||\boldsymbol{z}_h|||_h\leqslant 1} (p_h - \widehat{p}_h, \mathcal{D}_h^k \boldsymbol{z}_h) \lesssim \$, \tag{2.58}$$

which concludes the proof of the abstract estimate. □

Theorem 2.2.10 (Convergence rate)**.** *Denoting by* $(\boldsymbol{u}, p)$ *the weak solution to* (2.31)*,* $(\widehat{\boldsymbol{w}}_h, \widehat{p}_h) = (I_{\boldsymbol{W},h}^k \boldsymbol{u}, \pi_h^k p)$ *its projection, and further assuming the regularity* $(\boldsymbol{u}, p) \in$

$H^{k+2}(\Omega) \times H^{k+1}(\Omega)$, it holds for the approximation error ϵ defined by (2.56)*,*

$$\epsilon \lesssim \left\{ \sum_{T\in\mathcal{T}_h} (\|p\|^2_{H^{k+1}(T)} + \nu\|\boldsymbol{u}\|^2_{H^{k+2}(T)^d} + \mathcal{N}_{1,T})h_T^{2(k+1)} + \mathcal{N}_{2,T}\min(1,\mathrm{Pe}_T)h_T^{2k+1} \right\}^{1/2},$$

with $\mathcal{N}_{1,T}$ and $\mathcal{N}_{2,T}$, $T \in \mathcal{T}_h$, defined in Lemma 2.2.8.

Remark 2.2.11 (Convergence rate and local boundary Péclet numbers). The contribution of viscous terms to the approximation error ϵ displays the classical superconvergent behavior $\mathcal{O}(h_T^{k+1})$ typical of HHO methods, see, e.g., [59]. For advection terms, on the other hand, the order of the local contribution depends on the value of the local Peclet number Pe_T defined by (2.45): (i) elements on whose boundary viscous effects dominate ($\mathrm{Pe}_T \leqslant h_T$) contribute to the approximation error ϵ with a term which is $\mathcal{O}(h_T^{k+1})$; (ii) elements where advection dominates ($\mathrm{Pe}_T \geqslant 1$), on the other hand, contribute with a term which is $\mathcal{O}(h_T^{k+1/2})$; (iii) finally, for boundary Peclet values between h_T and 1, intermediate orders of convergence are observed.

Proof of Theorem 2.2.10. Let $\boldsymbol{z}_h \in \boldsymbol{W}_h^k$. The consistency error can be rewritten as

$$\begin{aligned}
\mathcal{E}_h(\boldsymbol{z}_h) &= \sum_{T\in\mathcal{T}_h} \left\{ (-\nu\triangle\boldsymbol{u}, \boldsymbol{v}_T)_T - \mathcal{A}_{\nu,T}(\widehat{\boldsymbol{w}}_T, \boldsymbol{z}_T) \right\} \\
&\quad + \sum_{T\in\mathcal{T}_h} \left\{ ((\boldsymbol{\beta}\cdot\boldsymbol{\nabla})\boldsymbol{u}, \boldsymbol{v}_T) + (\boldsymbol{u}, \boldsymbol{v}_T)_T - \mathcal{A}_{\boldsymbol{\beta},\mu,T}(\widehat{\boldsymbol{w}}_T, \boldsymbol{z}_T) \right\} \\
&\quad + \left\{ (\boldsymbol{\nabla} p, \boldsymbol{v}_h) + (\mathcal{D}_h^k \boldsymbol{z}_h, \widehat{p}_h) \right\}. \\
&= \mathfrak{T}_1 + \mathfrak{T}_2 + \mathfrak{T}_3.
\end{aligned}$$

The first term $\mathfrak{T}_1$ can be estimated using the consistency of $\mathcal{A}_{\boldsymbol{\kappa},h}$, a consequence of 2.9. Similarly,the second term $\mathfrak{T}_2$ is estimated using the consistency (2.52) of $\mathcal{A}_{\boldsymbol{\beta},\mu,h}$. Finally, the last term is estimated using the consistency (2.13) of the pressure-velocity coupling. This yields the desired estimation. □

Chapter 3

A *hp*-Hybrid High-Order method for variable diffusion on general meshes

3.1 Introduction

In the last few years, discretization technologies have appeared that support arbitrary approximation orders on general polytopal meshes. In this work, we focus on a particular instance of such technologies, the so-called Hybrid High-Order (HHO) methods originally introduced in [56, 59]. So far, the literature on HHO methods has focused on the *h*-version of the method with uniform polynomial degree. Our goal is to provide a first example of variable-degree *hp*-HHO method and carry out a full *hp*-convergence analysis valid for fairly general meshes and arbitrary space dimension. Let $\Omega \subset \mathbb{R}^d$, $d \geqslant 1$, denote a bounded connected polytopal domain. We consider the variable diffusion model problem

$$\begin{aligned} -\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}u) &= f \qquad \text{in } \Omega, \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{3.1}$$

where $\boldsymbol{\kappa}$ is a uniformly positive, symmetric, tensor-valued field on Ω, while $f \in L^2(\Omega)$ denotes a volumetric source. For the sake of simplicity, we assume that $\boldsymbol{\kappa}$ is piecewise constant on a partition P_Ω of Ω into polytopes. The weak formulation of

problem (3.1) reads: Find $u \in U := H_0^1(\Omega)$ such that

$$(\boldsymbol{\kappa}\boldsymbol{\nabla}u, \boldsymbol{\nabla}v) = (f, v) \qquad \forall v \in U, \tag{3.2}$$

where we have used the notation $(\cdot,\cdot)$ for the usual inner products of both $L^2(\Omega)$ and $L^2(\Omega)^d$. Here, the scalar-valued field u represents a potential, and the vector-valued field $\boldsymbol{\kappa}\boldsymbol{\nabla}u$ the corresponding flux.

For a given polytopal mesh $\mathcal{T}_h = \{T\}$ of Ω, the *hp*-HHO discretization of problem (3.2) is based on two sets of degrees of freedom (DOFs): (i) Skeletal DOFs, consisting in $(d-1)$-variate polynomials of total degree $p_F \geqslant 0$ on each mesh face F, and (ii) elemental DOFs, consisting in d-variate polynomials of degree p_T on each mesh element T, where p_T denotes the lowest degree of skeletal DOFs on the boundary of T. Skeletal DOFs are globally coupled and can be alternatively interpreted as traces of the potential on the mesh faces or as Lagrange multipliers enforcing the continuity of the normal flux at the discrete level; cf. [2, 44] for further insight. Elemental DOFs, on the other hand, are bubble-like auxiliary DOFs that can be locally eliminated by static condensation, as detailed in [44, Section 2.4] for the case where $p_F = p$ for all mesh faces F.

Two key ingredients are devised locally from skeletal and elemental DOFs attached to each mesh element T: (i) A reconstruction of the potential of degree (p_T+1) (i.e., one degree higher than elemental DOFs in T) obtained solving a small Neumann problem and (ii) a stabilisation term penalizing face-based residuals and polynomially consistent up to degree (p_T+1). The local contributions obtained from these two ingredients are then assembled following a standard, finite element-like procedure. The resulting discretization has several appealing features, the most prominent of which are summarized hereafter: (i) It is valid for fairly general polytopal meshes; (ii) the construction is dimension-independent, which can significantly ease the practical implementation; (iii) it enables the local adaptation of the approximation order, a highly desirable feature when combined with a regularity estimator (whose development will be addressed in a separate work); (iv) it exhibits only a moderate dependence on the diffusion coefficient $\boldsymbol{\kappa}$; (v) it has a moderate computational cost thanks to the possibility of eliminating elemental DOFs locally via static condensation; (vi) parallel implementations can be simplified by the fact that processes communicate via skeletal unknowns only.

The seminal works on the *p*- and *hp*-conforming finite element method on standard meshes date back the early 80s; cf. [11–13]. Starting from the late 90s, noncon-

forming methods on standard meshes supporting arbitrary-order have received a fair amount of attention; a (by far) nonexhaustive list of contributions focusing on scalar diffusive problems similar to the one considered here includes [37,73,91,95,101]. The possibility of refining both in h and in p on general meshes is, on the other hand, a much more recent research topic. We cite, in particular, *hp*-composite [5, 74] and polyhedral [36] discontinuous Galerkin methods, and the two-dimensional virtual element method proposed in [20].

The main results of this paper, summarized in Section 3.3.2, are *hp*-energy- and L^2-estimates of the error between the approximate and the exact solution. These are the first results of this kind for HHO methods, and among the first for discontinuous skeletal methods in general (a prominent example of discontinuous skeletal methods are the Hybridizable Discontinuous Galerkin methods of [46]; cf. [44] for a precise study of their relation with HHO methods). The cornerstone of the analysis is the extension of the classical Babuška-Suri *hp*-approximation results to regular mesh sequences in the sense of [55, Chapter 1] and arbitrary space dimension $d \geqslant 1$; cf. Lemma 3.2.1. Similar results had been derived in [20] for $d = 2$ and, under different assumptions on the mesh, in [36] for $d \in \{2, 3\}$. A key point is here to show that the regularity assumptions on the mesh imply uniform bounds for the Lipschitz constant of mesh elements. The resulting energy-norm estimate confirms the characteristic h-superconvergence behaviour of HHO methods, whereas we have a more standard scaling as $(p_T + 1)^{-p_T}$ with respect to the polynomial degree p_T of elemental DOFs. This scaling is analogous to the best available results for discontinuous Galerkin (dG) methods *on rectangular meshes* based on polynomials of degree p_T, cf. [73] (on more general meshes, the scaling for the symmetric interior penalty dG method is $p_T^{-(p_T-1/2)}$, half a power less than for the *hp*-HHO method studied here). Classically, when elliptic regularity holds, the h-convergence order can be increased by 1 for the L^2-norm. In our error estimates, the dependence on the diffusion coefficient is carefully tracked, showing full robustness with respect to its heterogeneity and only a moderate dependence with a power of $1/2$ on its local anisotropy when the error in the energy-norm is considered. Numerical experiments confirm the expected exponentially convergent behaviour for both isotropic and strongly anisotropic diffusion coefficients on a variety of two-dimensional meshes.

The rest of the paper is organized as follows. In Section 3.2 we introduce the main notations and prove the basic results required in the analysis including, in particular, Lemma 3.2.1 (whose proof is detailed in Appendix 3.5). In Section 3.3 we formulate the *hp*-HHO method, state our main results, and provide some numerical examples.

The proofs of the main results, preceeded by the required preparatory material, are collected in Section 3.4.

3.2 Setting

In this section we introduce the main notations and prove the basic results required in the analysis.

3.2.1 Mesh and notation

Let $\mathcal{H} \subset \mathbb{R}_*^+$ denote a countable set of meshsizes having 0 as its unique accumulation point. We consider mesh sequences $(\mathcal{T}_h)_{h\in\mathcal{H}}$ where, for all $h \in \mathcal{H}$, $\mathcal{T}_h = \{T\}$ is a finite collection of nonempty disjoint open polytopal elements such that $\overline{\Omega} = \bigcup_{T\in\mathcal{T}_h} \overline{T}$ and $h = \max_{T\in\mathcal{T}_h} h_T$ (h_T stands for the diameter of T). A hyperplanar closed connected subset F of $\overline{\Omega}$ is called a face if it has positive $(d-1)$-dimensional measure and (i) either there exist distinct $T_1, T_2 \in \mathcal{T}_h$ such that $F = \partial T_1 \cap \partial T_2$ (and F is an interface) or (ii) there exists $T \in \mathcal{T}_h$ such that $F = \partial T \cap \partial\Omega$ (and F is a boundary face). The set of interfaces is denoted by $\mathcal{F}_h^{\mathrm{i}}$, the set of boundary faces by $\mathcal{F}_h^{\mathrm{b}}$, and we let $\mathcal{F}_h := \mathcal{F}_h^{\mathrm{i}} \cup \mathcal{F}_h^{\mathrm{b}}$. For all $T \in \mathcal{T}_h$, the set $\mathcal{F}_T := \{F \in \mathcal{F}_h | F \subset \partial T\}$ collects the faces lying on the boundary of T and, for all $F \in \mathcal{F}_T$, we denote by $\boldsymbol{n}_{TF}$ the normal vector to F pointing out of T.

The following assumptions on the mesh will be kept throughout the exposition.

Assumption 1 (Admissible mesh sequence)**.** We assume that $(\mathcal{T}_h)_{h\in\mathcal{H}}$ is admissible in the sense of [55, Chapter 1], i.e., for all $h \in \mathcal{H}$, $\mathcal{T}_h$ admits a matching simplicial submesh $\mathfrak{T}_h$ and there exists a real number $\varrho > 0$ (the mesh regularity parameter) independent of h such that the following conditions hold: (i) For all $h \in \mathcal{H}$ and all simplex $S \in \mathfrak{T}_h$ of diameter h_S and inradius r_S, $\varrho h_S \leqslant r_S$; (ii) for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, and all $S \in \mathfrak{T}_h$ such that $S \subset T$, $\varrho h_T \leqslant h_S$; (iii) every mesh element $T \in \mathcal{T}_h$ is star-shaped with respect to every point of a ball of radius ϱh_T.

Assumption 2 (Compliant mesh sequence)**.** We assume that the mesh sequence is compliant with the partition P_Ω on which the diffusion tensor $\boldsymbol{\kappa}$ is piecewise constant, so that jumps only occur at interfaces and, for all $T \in \mathcal{T}_h$,

$$\boldsymbol{\kappa}_T := \boldsymbol{\kappa}_{|T} \in \mathbb{P}^0(T)^{d\times d}.$$

In what follows, for all $T \in \mathcal{T}_h$, $\overline{\kappa}_T$ and $\underline{\kappa}_T$ denote the largest and smallest eigenvalue of $\boldsymbol{\kappa}_T$, respectively, and $\lambda_{\boldsymbol{\kappa},T} := \overline{\kappa}_T/\underline{\kappa}_T$ the local anisotropy ratio.

3.2.2 Basic results

Let X be a mesh element or face. For an index q, $H^q(X)$ denotes the Hilbert space of functions which are in $L^2(X)$ together with their weak derivatives of order $\leqslant q$, equipped with the usual inner product $(\cdot,\cdot)_{q,X}$ and associated norm $\|\cdot\|_{q,X}$. When $q = 0$, we recover the Lebesgue space $L^2(X)$, and the subscript 0 is omitted from both the inner product and the norm. The subscript X is also omitted when $X = \Omega$. For a given integer $l \geqslant 0$, we denote by $\mathbb{P}^l(X)$ the space spanned by the restriction to X of d-variate polynomials of degree $\leqslant l$. For further use, we also introduce the L^2-projector $\pi_X^l : L^1(X) \to \mathbb{P}^l(X)$ such that, for all $w \in L^1(X)$,

$$(\pi_X^l w - w, v)_X = 0 \qquad \forall v \in \mathbb{P}^l(X). \tag{3.3}$$

We recall hereafter a few known results on admissible mesh sequences and refer to [55, Chapter 1] and [53] for a more comprehensive collection. By [55, Lemma 1.41], there exists an integer $\mathfrak{N}_\partial \geqslant (d+1)$ (possibly depending on d and ϱ) such that the maximum number of faces of one mesh element is bounded,

$$\max_{h\in\mathcal{H},T\in\mathcal{T}_h} \operatorname{card}(\mathcal{F}_T) \leqslant \mathfrak{N}_\partial. \tag{3.4}$$

The following multiplicative trace inequality, valid for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, and all $v \in H^1(T)$, is proved in [55, Lemma 1.49]:

$$\|v\|_{\partial T}^2 \leqslant C\left(\|v\|_T\|\boldsymbol{\nabla} v\|_T + h_T^{-1}\|v\|_T^2\right), \tag{3.5}$$

where C only depends on d and ϱ. We also note the following local Poincaré's inequality valid for all $T \in \mathcal{T}_h$ and all $v \in H^1(T)$ such that $(v,1)_T = 0$:

$$\|v\|_T \leqslant C_{\mathrm{P}} h_T \|\boldsymbol{\nabla} v\|_T, \tag{3.6}$$

where $C_{\mathrm{P}} = \pi^{-1}$ when T is convex, while it can be estimated in terms of ϱ for nonconvex elements (cf., e.g., [106]).

The following functional analysis results lie at the heart of the *hp*-analysis carried out in Section 3.4.

Lemma 3.2.1 (Approximation)**.** *There is a real number $C > 0$ (possibly depending on d and ϱ) such that, for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, all integer $l \geqslant 1$, all $s \geqslant 0$, and all $v \in H^{s+1}(T)$, there exists a polynomial $\Pi_T^l v \in \mathbb{P}^l(T)$ satisfying, for all $0 \leqslant q \leqslant s+1$,*

$$\|v - \Pi_T^l v\|_{q,T} \leqslant C \frac{h_T^{\min(l,s)-q+1}}{l^{s+1-q}} \|v\|_{s+1,T}. \tag{3.7}$$

Proof. See Section 3.5. □

Lemma 3.2.2 (Discrete trace inequality)**.** *There is a real number $C > 0$ (possibly depending on d and ϱ) such that, for all $h \in \mathcal{H}$, all $T \in \mathcal{T}_h$, all integer $l \geqslant 1$, and all $v \in \mathbb{P}^l(T)$, it holds*

$$\|v\|_{\partial T} \leqslant C \frac{l}{h_T^{1/2}} \|v\|_T. \tag{3.8}$$

Proof. When all meshes in the sequence $(\mathcal{T}_h)_{h\in\mathcal{H}}$ are simplicial and conforming, the proof of (3.8) can be found in [99, Theorem 4.76] for $d = 2$; for $d \geqslant 2$ the proof is analogous. The extension to admissible mesh sequences in the sense of Assumption 1 can be done following the reasoning in [55, Lemma 1.46]. □

3.3 Discretization

In this section, we formulate the *hp*-HHO method, state our main results, and provide some numerical examples.

3.3.1 The *hp*-HHO method

We present in this section an extension of the classical HHO method of [59] accounting for variable polynomial degrees. Let a vector $\underline{p}_h = (p_F)_{F\in\mathcal{F}_h} \in \mathbb{N}^{\mathcal{F}_h}$ of skeletal polynomial degrees be given. For all $T \in \mathcal{T}_h$, we denote by $\underline{p}_T = (p_F)_{F\in\mathcal{F}_T}$ the restriction of $\underline{p}_h$ to $\mathcal{F}_T$, and we introduce the following local space of DOFs:

$$\underline{U}_T^{\underline{p}_T} := \mathbb{P}^{p_T}(T) \times \left(\bigtimes_{F\in\mathcal{F}_T} \mathbb{P}^{p_F}(F) \right), \qquad p_T := \min_{F\in\mathcal{F}_T} p_F. \tag{3.9}$$

We use the notation $\underline{v}_T = (v_T, (v_F)_{F\in\mathcal{F}_T})$ for a generic element of $\underline{U}_T^{\underline{p}_T}$. We define the local potential reconstruction operator $r_T^{p_T+1} : \underline{U}_T^{\underline{p}_T} \to \mathbb{P}^{p_T+1}(T)$ such that, for

all $\underline{v}_T \in \underline{U}_T^{p_T}$ and $w \in \mathbb{P}^{p_T+1}(T)$

$$(\boldsymbol{\kappa}_T \boldsymbol{\nabla} r_T^{p_T+1} \underline{v}_T, \boldsymbol{\nabla} w)_T = -(v_T, \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}_T \boldsymbol{\nabla} w))_T + \sum_{F\in\mathcal{F}_T} (v_F, \boldsymbol{\kappa}_T \boldsymbol{\nabla} w \cdot \boldsymbol{n}_{TF})_F, \tag{3.10}$$

and

$$(r_T^{p_T+1} \underline{v}_T - v_T, 1)_T = 0. \tag{3.11}$$

Note that computing $r_T^{p_T+1}\underline{v}_T$ according to (3.10) requires to invert the $\boldsymbol{\kappa}_T$-weighted stiffness matrix of $\mathbb{P}^{k+1}(T)$, which can be efficiently accomplished by a Cholesky solver.

We define on $\underline{U}_T^{p_T} \times \underline{U}_T^{p_T}$ the local bilinear form a_T such that

$$\mathrm{a}_T(\underline{u}_T, \underline{v}_T) := (\boldsymbol{\kappa}_T \boldsymbol{\nabla} r_T^{p_T+1} \underline{u}_T, \boldsymbol{\nabla} r_T^{p_T+1} \underline{v}_T)_T + \mathrm{s}_T(\underline{u}_T, \underline{v}_T) \tag{3.12}$$

where

$$\mathrm{s}_T(\underline{u}_T, \underline{v}_T) := \sum_{F\in\mathcal{F}_T} \frac{\kappa_F}{h_T} (\delta_{TF}^{p_T} \underline{u}_T, \delta_{TF}^{p_T} \underline{v}_T)_F, \tag{3.13}$$

and for all $F \in \mathcal{F}_T$, we have set $\kappa_F := \boldsymbol{\kappa}_T \boldsymbol{n}_{TF} \cdot \boldsymbol{n}_{TF}$ and the face-based residual operator $\delta_{TF}^{p_T} : \underline{U}_T^{p_T} \to \mathbb{P}^{p_F}(F)$ is such that, for all $\underline{v}_T \in \underline{U}_T^{p_T}$,

$$\delta_{TF}^{p_T} \underline{v}_T := \pi_F^{p_F} \big(v_F - r_T^{p_T+1} \underline{v}_T + \pi_T^{p_T} r_T^{p_T+1} \underline{v}_T - v_T \big). \tag{3.14}$$

The first contribution in a_T is in charge of consistency, whereas the second ensures stability by a least-square penalty of the face-based residuals $\delta_{TF}^{p_T}$. This subtle form for $\delta_{TF}^{p_T}$ ensures that the residual vanishes when its argument is the interpolate of a function in $\mathbb{P}^{p_T+1}(T)$, and is required for high-order h-convergence (a detailed motivation is provided in [59, Remark 6]).

The global space of DOFs and its subspace with strongly enforced boundary conditions are defined, respectively, as

$$\underline{U}_h^{p_h} := \left(\bigtimes_{T\in\mathcal{T}_h} \mathbb{P}^{p_T}(T) \right) \times \left(\bigtimes_{F\in\mathcal{F}_h} \mathbb{P}^{p_F}(F) \right), \qquad \underline{U}_{h,0}^{p_h} := \left\{ \underline{v}_h \in \underline{U}_h^{p_h} \mid v_F \equiv 0 \quad \forall F \in \mathcal{F}_h^{\mathrm{b}} \right\}. \tag{3.15}$$

Note that interface DOFs in $\underline{U}_h^{p_h}$ are single-valued. We use the notation $\underline{v}_h = ((v_T)_{T\in\mathcal{T}_h}, (v_F)_{F\in\mathcal{F}_h})$ for a generic DOF vector in $\underline{U}_h^{p_h}$ and, for all $T \in \mathcal{T}_h$, we denote by $\underline{v}_T \in \underline{U}_T^{p_T}$ its restriction to T. For further use, we also introduce the global

interpolator $\underline{I}_h^{\underline{p}_h} : H^1(\Omega) \to \underline{U}_h^{\underline{p}_h}$ such that, for all $v \in H^1(\Omega)$,

$$\underline{I}_h^{\underline{p}_h} v = \big((\pi_T^{p_T} v)_{T\in\mathcal{T}_h}, (\pi_F^{p_F} v)_{F\in\mathcal{F}_h}\big), \tag{3.16}$$

and denote by $\underline{I}_T^{\underline{p}_T}$ its restriction to $T \in \mathcal{T}_h$.

The *hp*-HHO discretization of problem (3.2) consists in seeking $\underline{u}_h \in \underline{U}_{h,0}^{\underline{p}_h}$ such that

$$\mathrm{a}_h(\underline{u}_h, \underline{v}_h) = \mathrm{l}_h(\underline{v}_h) \qquad \forall \underline{v}_h \in \underline{U}_{h,0}^{\underline{p}_h}, \tag{3.17}$$

where the global bilinear form a_h on $\underline{U}_h^{\underline{p}_h} \times \underline{U}_h^{\underline{p}_h}$ and the linear form l_h on $\underline{U}_h^{\underline{p}_h}$ are assembled element-wise setting

$$\mathrm{a}_h(\underline{u}_h, \underline{v}_h) := \sum_{T\in\mathcal{T}_h} \mathrm{a}_T(\underline{u}_T, \underline{v}_T), \qquad \mathrm{l}_h(\underline{v}_h) := \sum_{T\in\mathcal{T}_h} (f, v_T)_T.$$

Remark 3.3.1 (Static condensation). Using a standard static condensation procedure, it is possible to eliminate element-based DOFs locally and solve (3.17) by inverting a system in the skeletal unknowns only. For the sake of conciseness, we do not repeat the details here and refer instead to [44, Section 2.4]. Accounting for the strong enforcement of boundary conditions, the size of the system after static condensation is

$$N_{\mathrm{dof}} = \sum_{F\in\mathcal{F}_h^{\mathrm{i}}} \binom{p_F + d - 1}{p_F}. \tag{3.18}$$

Remark 3.3.2 (Finite element interpretation). A finite element interpretation of the scheme (3.17) is possible following the extension proposed in [44, Remark 3] of the ideas originally developed in [10] in the context of nonconforming Virtual Element Methods. For all $F \in \mathcal{F}_h^{\mathrm{i}}$, we denote by $[\cdot]_F$ the usual jump operator (the sign is irrelevant), which we extend to boundary faces $F \in \mathcal{F}_h^{\mathrm{b}}$ setting $[\varphi]_F := \varphi$. Let

$$\mathfrak{U}_{h,0}^{\underline{p}_h} := \Big\{ \mathfrak{v}_h \in L^2(\Omega) \mid \mathfrak{v}_{h|T} \in \mathfrak{U}_T^{\underline{p}_T} \text{ for all } T \in \mathcal{T}_h \text{ and } \pi_F^{p_F}([\mathfrak{v}_T]_F) = 0 \text{ for all } F \in \mathcal{F}_h \Big\},$$

where, for all $T \in \mathcal{T}_h$, we have introduced the local space

$$\mathfrak{U}_T^{\underline{p}_T} := \big\{ \mathfrak{v}_T \in H^1(T) \mid \boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}_T \boldsymbol{\nabla}\mathfrak{v}_T) \in \mathbb{P}^{p_T}(T) \text{ and } \boldsymbol{\kappa}_T \boldsymbol{\nabla}\mathfrak{v}_{T|F}\cdot\boldsymbol{n}_{TF} \in \mathbb{P}^{p_F}(F) \text{ for all } F \in \mathcal{F}_T \big\}.$$

It can be proved that, for all $T \in \mathcal{T}_h$, $\underline{I}_T^{\underline{p}_T} : \mathfrak{U}_T^{\underline{p}_T} \to \underline{U}_T^{\underline{p}_T}$ is an isomorphism. Thus, the triplet $(T, \mathfrak{U}_T^{\underline{p}_T}, \underline{I}_T^{\underline{p}_T})$ defines a finite element in the sense of Ciarlet [43]. Additionally, problem (3.17) can be reformulated as the nonconforming finite element method:

Find $\mathfrak{u}_h \in \mathfrak{U}_{h,0}^{\underline{p}_h}$ such that

$$\mathfrak{a}_h(\mathfrak{u}_h, \mathfrak{v}_h) = \mathrm{l}_h(\mathfrak{v}_h) \qquad \forall \mathfrak{v}_h \in \mathfrak{U}_{h,0}^{\underline{p}_h},$$

where $\mathfrak{a}_h(\mathfrak{u}_h, \mathfrak{v}_h) := \mathrm{a}_h(\underline{I}_h^{\underline{p}_h}\mathfrak{u}_h, \underline{I}_h^{\underline{p}_h}\mathfrak{v}_h)$, $\mathrm{l}_h(\mathfrak{v}_h) := \mathrm{l}_h(\underline{I}_h^{\underline{p}_h}\mathfrak{v}_h)$, and it can be proved that $\mathfrak{u}_h$ is the unique element of $\mathfrak{U}_{h,0}^{\underline{p}_h}$ such that $\underline{u}_h = \underline{I}_h^{\underline{p}_h}\mathfrak{u}_h$ with $\underline{u}_h$ unique solution to (3.17).

3.3.2 Main results

We next state our main results. The proofs are postponed to Section 3.4. For all $T \in \mathcal{T}_h$, we denote by $\|\cdot\|_{\mathrm{a},T}$ and $|\cdot|_{\mathrm{s},T}$ the seminorms defined on $\underline{U}_T^{\underline{p}_T}$ by the bilinear forms a_T and s_T, respectively, and by $\|\cdot\|_{\mathrm{a},h}$ the seminorm defined by the bilinear form a_h on $\underline{U}_h^{\underline{p}_h}$. We also introduce the penalty seminorm $|\cdot|_{\mathrm{s},h}$ such that, for all $\underline{v}_h \in \underline{U}_h^{\underline{p}_h}$, $|\underline{v}_h|_{\mathrm{s},h}^2 := \sum_{T\in\mathcal{T}_h} |\underline{v}_T|_{\mathrm{s},T}^2$. Note that $\|\cdot\|_{\mathrm{a},h}$ is a norm on the subspace $\underline{U}_{h,0}^{\underline{p}_h}$ with strongly enforced boundary conditions (the arguments are essentially analogous to that of [56, Proposition 5]). We will also need the global reconstruction operator $r_h^{\underline{p}_h} : \underline{U}_h^{\underline{p}_h} \to L^2(\Omega)$ such that, for all $\underline{v}_h \in \underline{U}_h^{\underline{p}_h}$,

$$(r_h^{\underline{p}_h}\underline{v}_h)_{|T} = r_T^{p_T+1}\underline{v}_T \qquad \forall T \in \mathcal{T}_h.$$

Finally, for the sake of conciseness, throughout the rest of the paper we note $a \lesssim b$ the inequality $a \leqslant Cb$ with real number $C > 0$ independent of h, $\underline{p}_h$, and $\boldsymbol{\kappa}$.

Our first estimate concerns the error measured in energy-like norms.

Theorem 3.3.3 (Energy error estimate)**.** *Let $u \in U$ and $\underline{u}_h \in \underline{U}_{h,0}^{\underline{p}_h}$ denote the unique solutions of problems* (3.2) *and* (3.17)*, respectively, and set*

$$\widehat{\underline{u}}_h := \underline{I}_h^{\underline{p}_h} u. \tag{3.19}$$

Assuming that $u_{|T} \in H^{p_T+2}(T)$ for all $T \in \mathcal{T}_h$, it holds

$$\|\underline{u}_h - \widehat{\underline{u}}_h\|_{\mathrm{a},h} \lesssim \left(\sum_{T\in\mathcal{T}_h} \overline{\kappa}_T \lambda_{\boldsymbol{\kappa},T} \frac{h_T^{2(p_T+1)}}{(p_T+1)^{2p_T}} \|u\|_{p_T+2,T}^2\right)^{1/2}. \tag{3.20}$$

Consequently, we have, denoting by $\boldsymbol{\nabla}_h$ the broken gradient on $\mathcal{T}_h$ (whose restriction

to every element $T \in \mathcal{T}_h$ coincides with the usual gradient),

$$\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h(u - r_h^{p_h}\underline{u}_h)\|^2 + |\underline{u}_h|_{\mathrm{s},h}^2 \lesssim \sum_{T\in\mathcal{T}_h} \overline{\kappa}_T \lambda_{\boldsymbol{\kappa},T} \frac{h_T^{2(p_T+1)}}{(p_T+1)^{2p_T}} \|u\|_{p_T+2,T}^2. \tag{3.21}$$

Proof. See Section 3.4.3. □

In (3.20) and (3.21), we observe the characteristic improved h-convergence of HHO methods (cf. [59]), whereas, in terms of p-convergence, we have a more standard scaling as $(p_T+1)^{-p_T}$ (i.e., half a power more than discontinuous Galerkin methods based on polynomials of degree p_T, cf., e.g., [91]). In (3.21), we observe that the left-hand side has the same convergence rate (both in h and in p) as the interpolation error

$$\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h(u - r_h^{p_h}\widehat{\underline{u}}_h)\|^2 + |\widehat{\underline{u}}_h|_{\mathrm{s},h}^2,$$

as can be verified combining (3.26) and (3.28) below. Note that, in this case, the p-convergence is limited by the second term, which measures the discontinuity of the potential reconstruction at interfaces. An inspection of formulas (3.20) and (3.21) also shows that the method is fully robust with respect to the heterogeneity of the diffusion coefficient, while only a moderate dependence (with a power of $1/2$) is observed with respect to its local anisotropy ratio.

For the sake of completeness, we also provide an estimate of the L^2-error between the piecewise polynomial fields u_h and $\widehat{u}_h$ such that

$$u_{h|T} := u_T \quad \text{and} \quad \widehat{u}_{h|T} := \widehat{u}_T = \pi_T^{p_T} u \qquad \forall T \in \mathcal{T}_h.$$

To this end, we need elliptic regularity in the following form: For all $g \in L^2(\Omega)$, the unique element $z \in U$ such that

$$(\boldsymbol{\kappa}\boldsymbol{\nabla}z, \boldsymbol{\nabla}v) = (g, v) \qquad \forall v \in U, \tag{3.22}$$

satisfies the a priori estimate

$$\|z\|_2 \lesssim \underline{\kappa}^{-1}\|g\|_{L^2(\Omega)}, \qquad \underline{\kappa} := \min_{T\in\mathcal{T}_h} \underline{\kappa}_T. \tag{3.23}$$

The following result is proved in Section 3.4.4.

Theorem 3.3.4 (L^2-error estimate)**.** *Under the assumptions of Theorem 3.3.3, and further assuming elliptic regularity* (3.23) *and that $f \in H^{p_T+\Delta_T}(T)$ for all $T \in \mathcal{T}_h$*

with $\Delta_T = 1$ *if* $p_T = 0$ *while* $\Delta_T = 0$ *otherwise,*

$$\underline{\kappa}\|u_h - \widehat{u}_h\| \lesssim \overline{\kappa}^{1/2}\lambda_{\boldsymbol{\kappa}} h \left(\sum_{T\in\mathcal{T}_h} \lambda_{\boldsymbol{\kappa},T}\overline{\kappa}_T \frac{h_T^{2(p_T+1)}}{(p_T+1)^{2p_T}}\|u\|^2_{p_T+2,T}\right)^{1/2} + \left(\sum_{T\in\mathcal{T}_h} \frac{h_T^{2(p_T+2)}}{(p_T+\Delta_T)^{2(p_T+2)}}\|f\|^2_{p_T+\Delta_T}\right)^{1/2}. \quad (3.24)$$

with $\lambda_{\boldsymbol{\kappa}} := \max_{T\in\mathcal{T}_h}\lambda_{\boldsymbol{\kappa},T}$, $\overline{\kappa} := \max_{T\in\mathcal{T}_h}\overline{\kappa}_T$.

3.3.3 Numerical examples

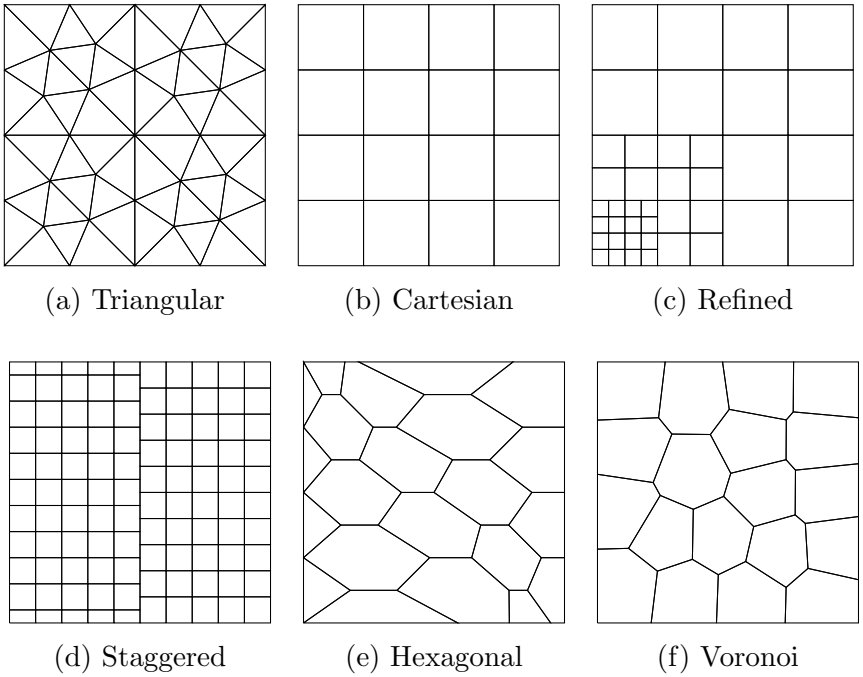
We close this section with some numerical examples. The h-convergence properties of the method (3.17) have been numerically investigated in [59, Section 4]. To illustrate its p-convergence properties, we solve on the unit square domain $\Omega = (0,1)^2$ the homogeneous Dirichlet problem with exact solution $u = \sin(\pi x_1)\sin(\pi x_2)$ and right-hand side f chosen accordingly. We consider two values for the diffusion coefficients:

$$\boldsymbol{\kappa}_1 = \boldsymbol{I}_2, \qquad \boldsymbol{\kappa}_2 = \begin{pmatrix} (x_2-\overline{x}_2)^2 + \epsilon(x_1-\overline{x}_1)^2 & -(1-\epsilon)(x_1-\overline{x}_1)(x_2-\overline{x}_2) \\ -(1-\epsilon)(x_1-\overline{x}_1)(x_2-\overline{x}_2) & (x_1-\overline{x}_1)^2 + \epsilon(x_2-\overline{x}_2)^2 \end{pmatrix},$$

where $\boldsymbol{I}_2$ denotes the identity matrix of dimension 2, $\overline{\boldsymbol{x}} := -(0.1, 0.1)$, and $\epsilon = 1\cdot 10^{-2}$. The choice $\boldsymbol{\kappa} = \boldsymbol{\kappa}_1$ ("regular" test case) yields a homogeneous isotropic problem, while the choice $\boldsymbol{\kappa} = \boldsymbol{\kappa}_2$ ("Le Potier's" test case [84]) corresponds to a highly anisotropic problem where the principal axes of the diffusion tensor vary at each point of the domain. Figures 3.2–3.3 depict the energy- and L^2-errors as a function of the number of skeletal DOFs N_{dof} (cf. (3.18)) when $p_F = p$ for all $F \in \mathcal{F}_h$ and $p \in \{0,\dots,9\}$ for the proposed choices for $\boldsymbol{\kappa}$ on the meshes of Figure 3.1. In all the cases, the expected exponentially convergent behaviour is observed. Interestingly, the best performance in terms of error vs. N_{dof} is obtained for the Cartesian and Voronoi meshes. A comparison of the results for the two values of the diffusion coefficients allows to appreciate the robustness of the method with respect to anisotropy.

3.4 Convergence analysis

In this section we prove the results stated in Section 3.3.2.



(a) Triangular (b) Cartesian (c) Refined

(d) Staggered (e) Hexagonal (f) Voronoi

Figure 3.1 – Meshes considered in the p-convergence test of Section 3.3.3. The triangular, Cartesian, refined, and staggered meshes originate from the FVCA5 benchmark [80]; the hexagonal mesh was originally introduced in [62]; the Voronoi mesh was obtained using the PolyMesher algorithm of [104].

3.4.1 Consistency of the potential reconstruction

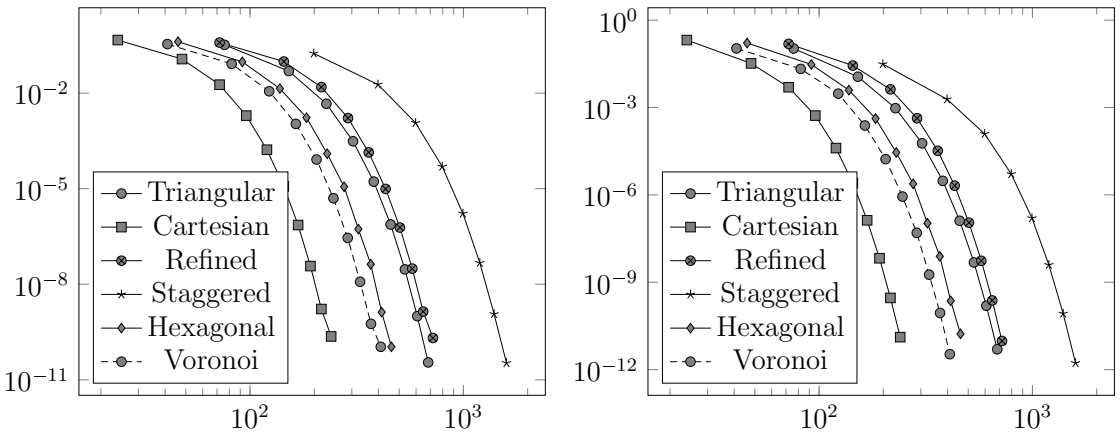
Preliminary to the convergence analysis is the study of the approximation properties of the potential reconstruction $r_T^{p_T+1}$ defined by (3.10) when its argument is the interpolate of a regular function. Let a mesh element $T \in \mathcal{T}_h$ be fixed. For any integer $l \geqslant 1$, we define the elliptic projector $\varpi_{\boldsymbol{\kappa},T}^l : H^1(T) \to \mathbb{P}^l(T)$ such that, for all $v \in H^1(T)$, $(\varpi_{\boldsymbol{\kappa},T}^l v - v, 1)_T = 0$ and it holds

$$(\boldsymbol{\kappa}_T \boldsymbol{\nabla}(\varpi_{\boldsymbol{\kappa},T}^l v - v), \boldsymbol{\nabla} w)_T = 0 \qquad \forall w \in \mathbb{P}^l(T). \tag{3.25}$$

Proposition 3.4.1 (Characterization of $(r_T^{p_T+1} \circ \underline{I}_T^{p_T})$)**.** *It holds, for all $T \in \mathcal{T}_h$,*

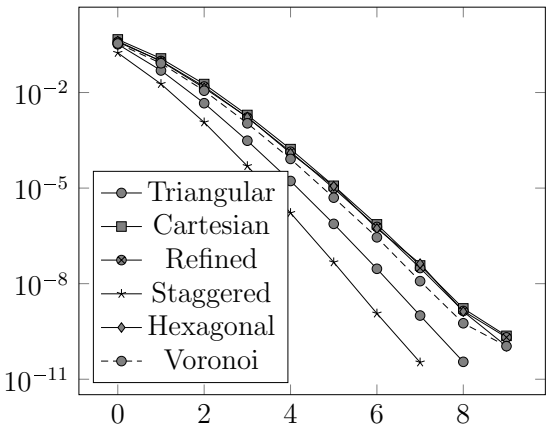
$$r_T^{p_T+1} \circ \underline{I}_T^{p_T} = \varpi_{\boldsymbol{\kappa},T}^{p_T+1}.$$

Proof. For a generic $v \in H^1(T)$, letting $\underline{v}_T = \underline{I}_T^{p_T} v$ in (3.10) we infer, for all $w \in$

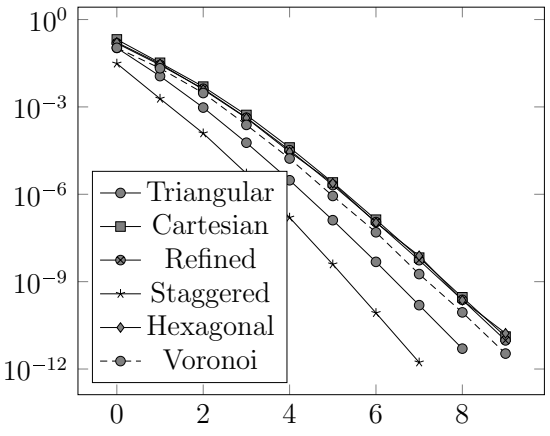


(a) $\|\underline{u}_h - \underline{\widehat{u}}_h\|_{\mathrm{a},h}$ vs. N_{dof}

(b) $\|u_h - \widehat{u}_h\|$ vs. N_{dof}

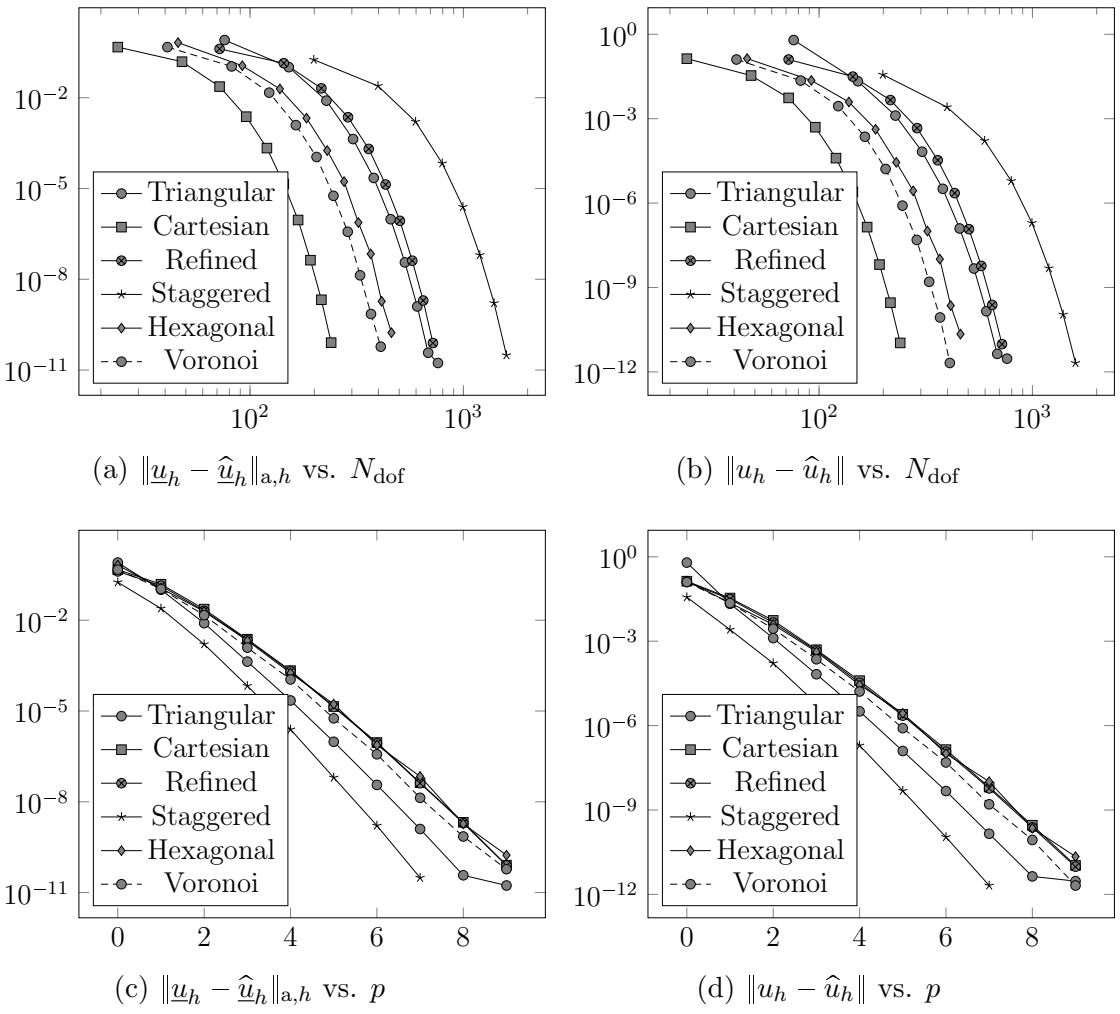


(c) $\|\underline{u}_h - \underline{\widehat{u}}_h\|_{\mathrm{a},h}$ vs. p



(d) $\|u_h - \widehat{u}_h\|$ vs. p

Figure 3.2 – Convergence with p-refinement (regular test case)



(a) $\|\underline{u}_h - \underline{\widehat{u}}_h\|_{\mathrm{a},h}$ vs. N_{dof}

(b) $\|u_h - \widehat{u}_h\|$ vs. N_{dof}

(c) $\|\underline{u}_h - \underline{\widehat{u}}_h\|_{\mathrm{a},h}$ vs. p

(d) $\|u_h - \widehat{u}_h\|$ vs. p

Figure 3.3 – Convergence with p-refinement (Le Potier's test case)

$\mathbb{P}^{p_T+1}(T)$,

$$\begin{aligned}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}(r_T^{p_T+1}\circ\underline{I}_T^{p_T})v,\boldsymbol{\nabla}w)_T &= -(\pi_T^{p_T}v,\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}_T\boldsymbol{\nabla}w))_T+\sum_{F\in\mathcal{F}_T}(\pi_F^{p_F}v,\boldsymbol{\kappa}_T\boldsymbol{\nabla}w\cdot\boldsymbol{n}_{TF})_F\\ &= -(v,\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}_T\boldsymbol{\nabla}w))_T+\sum_{F\in\mathcal{F}_T}(v,\boldsymbol{\kappa}_T\boldsymbol{\nabla}w\cdot\boldsymbol{n}_{TF})_F\\ &= (\boldsymbol{\kappa}_T\boldsymbol{\nabla}v,\boldsymbol{\nabla}w)_T,\end{aligned}$$

where we have used the fact that $\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}_T\boldsymbol{\nabla}w)\in\mathbb{P}^{p_T-1}(T)\subset\mathbb{P}^{p_T}(T)$ and $(\boldsymbol{\kappa}_T\boldsymbol{\nabla}w)_{|F}\cdot\boldsymbol{n}_{TF}\in\mathbb{P}^{p_T}(F)\subset\mathbb{P}^{p_F}(F)$ (cf. the definition (3.9) of p_T) to pass to the second line, and an integration by parts to conclude. □

We next study the approximation properties of $\varpi_{\boldsymbol{\kappa},T}^l$, from which those of $(r_T^{p_T+1}\circ\underline{I}_T^{p_T})$ follow in the light of Proposition 3.4.1.

Lemma 3.4.2 (Approximation properties of $\varpi_{\boldsymbol{\kappa},T}^l$)**.** *For all integer $l\geqslant 1$, all mesh element $T\in\mathcal{T}_h$, all $0\leqslant s\leqslant l$, and all $v\in H^{s+1}(T)$, it holds*

$$\begin{aligned}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_T+\frac{h_T^{1/2}}{l}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_{\partial T}+\frac{\kappa_T^{1/2}}{h_T}\|v-\varpi_{\boldsymbol{\kappa},T}^l v\|_T\\ +\frac{\kappa_T^{1/2}}{h_T^{1/2}}\|v-\varpi_{\boldsymbol{\kappa},T}^l v\|_{\partial T}\lesssim\overline{\kappa}_T^{1/2}\frac{h_T^{\min(l,s)}}{l^s}\|v\|_{s+1,T}.\end{aligned}\tag{3.26}$$

Proof. By definition (3.25) of $\varpi_{\boldsymbol{\kappa},T}^l$, it holds,

$$\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_T=\min_{w\in\mathbb{P}^l(T)}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-w)\|_T\leqslant\overline{\kappa}_T^{1/2}\|\boldsymbol{\nabla}(v-\Pi_T^l v)\|_T,\tag{3.27}$$

hence, using (3.7) with $q=1$, it is readily inferred

$$\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_T\lesssim\overline{\kappa}_T^{1/2}\frac{h_T^{\min(l,s)}}{l^s}\|v\|_{s+1,T}.$$

To prove the second bound in (3.26), use the triangle inequality to infer

$$\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_{\partial T}\leqslant\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v-\Pi_T^l v)\|_{\partial T}+\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(\Pi_T^l v-\varpi_{\boldsymbol{\kappa},T}^l v)\|_{\partial T}:=\mathfrak{T}_1+\mathfrak{T}_2.$$

For the first term, the multiplicative trace inequality (3.5) combined with (3.7) (with $q=1,2$) gives

$$\mathfrak{T}_1\lesssim\overline{\kappa}_T^{1/2}\frac{h_T^{\min(l,s)-1/2}}{l^{s-1/2}}\|v\|_{s+1,T}.$$

For the second term, we have,

$$\begin{aligned}\mathfrak{T}_2 &\lesssim \frac{l}{h_T^{1/2}}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(\Pi_T^l v - \varpi_{\boldsymbol{\kappa},T}^l v)\|_T\\ &\leqslant \frac{l}{h_T^{1/2}}\left(\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}(\Pi_T^l v - v)\|_T + \|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(v - \varpi_{\boldsymbol{\kappa},T}^l v)\|_T\right)\\ &\lesssim \overline{\kappa}_T^{1/2}\frac{l}{h_T^{1/2}}\|\boldsymbol{\nabla}(\Pi_T^l v - v)\|_T \lesssim \overline{\kappa}_T^{1/2}\frac{h_T^{\min(l,s)-1/2}}{l^{s-1}}\|v\|_{s+1,T},\end{aligned}$$

where we have used the discrete trace inequality (3.8) in the first line, the triangle inequality in the second line, the estimate (3.27) in the third, and the approximation result (3.7) with $q = 1$ to conclude. To obtain the third bound in (3.26), after recalling that $(v - \varpi_{\boldsymbol{\kappa},T}^l v, 1)_T = 0$, we apply the local Poincaré's inequality (3.6) to infer

$$\|v - \varpi_{\boldsymbol{\kappa},T}^l v\|_T \lesssim \frac{h_T}{\underline{\kappa}_T^{1/2}}\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}(v - \varpi_{\boldsymbol{\kappa},T}^l v)\|_T \lesssim \frac{\overline{\kappa}_T^{1/2}}{\underline{\kappa}_T^{1/2}}\frac{h_T^{\min(l,s)+1}}{l^s}\|v\|_{s+1,T},$$

where the conclusion follows from the first bound in (3.26). Finally, to obtain the last bound, we use the multiplicative trace inequality (3.5) to infer

$$\|v - \varpi_{\boldsymbol{\kappa},T}^l v\|_{\partial T}^2 \lesssim \underline{\kappa}_T^{-1/2}\|v - \varpi_{\boldsymbol{\kappa},T}^l v\|_T\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}(v - \varpi_{\boldsymbol{\kappa},T}^l v)\|_T + h_T^{-1}\|v - \varpi_{\boldsymbol{\kappa},T}^l v\|_T^2,$$

and use the first and third bound in (3.26) to estimate the various terms. □

3.4.2 Consistency of the stabilization term

The consistency properties of the stabilization bilinear form s_T defined by (3.12) are summarized in the following Lemma.

Lemma 3.4.3 (Consistency of the stabilization term)**.** *For all $T \in \mathcal{T}_h$, all $0 \leqslant q \leqslant p_T$, and all $v \in H^{q+2}(T)$, it holds*

$$|\underline{I}_T^{p_T} v|_{\mathrm{s},T} \lesssim \overline{\kappa}_T^{1/2}\lambda_{\boldsymbol{\kappa},T}^{1/2}\frac{h_T^{\min(p_T,q)+1}}{(p_T+1)^q}\|v\|_{q+2,T}. \tag{3.28}$$

Proof. Let $T \in \mathcal{T}_h$ and $v \in H^{q+2}(T)$ and set, for the sake of brevity,

$$\breve{v}_T := (r_T^{p_T+1} \circ \underline{I}_T^{p_T})v = \varpi_{\boldsymbol{\kappa},T}^{p_T+1} v$$

(cf. Proposition 3.4.1). For all $F \in \mathcal{F}_T$, recalling the definitions of the face residual $\delta_{TF}^{p_T}$ (cf. (3.14)) and of the local interpolator $\underline{I}_T^{p_T}$ (cf. (3.16)), together with the fact that $p_T \leqslant p_F$ by definition (3.9), we get

$$\delta_{TF}^{p_T}\underline{I}_T^{p_T} v = \pi_F^{p_F}(\check{v}_T - v) - \pi_T^{p_T}(\check{v}_T - v).$$

Using the triangle inequality and the $L^2(F)$-stability of $\pi_F^{p_F}$, we infer

$$\|\delta_{TF}^{p_T}\underline{I}_T^{p_T} v\|_F \leqslant \|\check{v}_T - v\|_F + \|\pi_T^{p_T}(\check{v}_T - v)\|_F := \mathfrak{T}_1 + \mathfrak{T}_2. \tag{3.29}$$

For the first term, the approximation properties (3.26) of $\varpi_{\boldsymbol{\kappa},T}^{p_T+1}$ (with $l = p_T + 1$ and $s = q + 1$) readily yield

$$\mathfrak{T}_1 \lesssim \lambda_{\boldsymbol{\kappa},T}^{1/2} \frac{h_T^{\min(p_T,q)+3/2}}{(p_T+1)^q} \|v\|_{q+2,T}. \tag{3.30}$$

For the second term, on the other hand, the discrete trace inequality (3.8) followed by the $L^2(T)$-stability of $\pi_T^{p_T}$ and (3.26) (with $l = p_T + 1$ and $s = q + 1$) gives

$$\mathfrak{T}_2 \lesssim \frac{(p_T+1)}{h_T^{1/2}} \|\pi_T^{p_T}(\check{v}_T - v)\|_T \leqslant \frac{(p_T+1)}{h_T^{1/2}} \|\check{v}_T - v\|_T \lesssim \lambda_{\boldsymbol{\kappa},T}^{1/2} \frac{h_T^{\min(p_T,q)+3/2}}{(p_T+1)^q} \|v\|_{q+2,T}. \tag{3.31}$$

The bound (3.28) follows using (3.30)–(3.31) in the right-hand side of (3.29), squaring the resulting inequality, multiplying it by κ_F/h_T, summing over $F \in \mathcal{F}_T$, and using the bound (3.4) on $\mathrm{card}(\mathcal{F}_T)$. □

3.4.3 Energy error estimate

Proof of Theorem 3.3.3. We start by noting the following abstract error estimate:

$$\|\underline{u}_h - \widehat{\underline{u}}_h\|_{\mathrm{a},h} \leqslant \sup_{\underline{v}_h \in \underline{U}_{h,0}^{p_h}, \|\underline{v}_h\|_{\mathrm{a},h}=1} \mathcal{E}_h(\underline{v}_h), \tag{3.32}$$

with consistency error

$$\mathcal{E}_h(\underline{v}_h) := \mathrm{l}_h(\underline{v}_h) - \mathrm{a}_h(\widehat{\underline{u}}_h, \underline{v}_h). \tag{3.33}$$

To prove (3.32), it suffices to observe that

$$\begin{aligned}\|\underline{u}_h - \widehat{\underline{u}}_h\|_{\mathrm{a},h}^2 &= \mathrm{a}_h(\underline{u}_h - \widehat{\underline{u}}_h, \underline{u}_h - \widehat{\underline{u}}_h)\\ &= \mathrm{a}_h(\underline{u}_h, \underline{u}_h - \widehat{\underline{u}}_h) - \mathrm{a}_h(\widehat{\underline{u}}_h, \underline{u}_h - \widehat{\underline{u}}_h)\\ &= \mathrm{l}_h(\underline{u}_h - \widehat{\underline{u}}_h) - \mathrm{a}_h(\widehat{\underline{u}}_h, \underline{u}_h - \widehat{\underline{u}}_h),\end{aligned}$$

where we have used the definition of the $\|\cdot\|_{\mathrm{a},h}$-norm in the first line, the linearity of a_h in its first argument in the second line, and the discrete problem (3.17) in the third. The conclusion follows dividing both sides by $\|\underline{u}_h - \widehat{\underline{u}}_h\|_{\mathrm{a},h}$, using linearity, and passing to the supremum.

We next bound the consistency error $\mathcal{E}_h(\underline{v}_h)$ for a generic vector of DOFs $\underline{v}_h \in \underline{U}_{h,0}^{p_h}$. A preliminary step consists in finding a more appropriate rewriting for $\mathcal{E}_h(\underline{v}_h)$. Observing that $f = -\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}\boldsymbol{\nabla}u)$ a.e. in Ω, integrating by parts element-by-element, and using the continuity of the normal component of $\boldsymbol{\kappa}\boldsymbol{\nabla}u$ across interfaces together with the strongly enforced boundary conditions in $\underline{U}_{h,0}^{p_h}$ to insert v_F into the second term in parentheses, we infer

$$\mathrm{l}_h(\underline{v}_h) = \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}u, \boldsymbol{\nabla}v_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}u\cdot\boldsymbol{n}_{TF}, v_F - v_T)_F\right). \tag{3.34}$$

Setting, for the sake of conciseness (cf. Proposition 3.4.1),

$$\breve{u}_T := r_T^{p_T+1}\widehat{\underline{u}}_T = \varpi_{\boldsymbol{\kappa},T}^{p_T+1}u, \tag{3.35}$$

and using the definition (3.10) of $r_T^{p_T+1}\underline{v}_T$ with $w = \breve{u}_T$, we have

$$\mathrm{a}_h(\widehat{\underline{u}}_h, \underline{v}_h) = \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}\breve{u}_T, \boldsymbol{\nabla}v_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}\breve{u}_T\cdot\boldsymbol{n}_{TF}, v_F - v_T)_F + \mathrm{s}_T(\widehat{\underline{u}}_T, \underline{v}_T)\right). \tag{3.36}$$

Subtracting (3.36) from (3.34), and observing that the first terms inside the summations cancel out owing to (3.25), we have

$$\mathcal{E}_h(\underline{v}_h) = \sum_{T\in\mathcal{T}_h}\left(\sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}(\breve{u}_T - u)\cdot\boldsymbol{n}_{TF}, v_F - v_T)_F + \mathrm{s}_T(\widehat{\underline{u}}_T, \underline{v}_T)\right). \tag{3.37}$$

Denote by $\mathfrak{T}_1(T)$ and $\mathfrak{T}_2(T)$ the two summands in parentheses. Using the Cauchy–Schwarz inequality followed by the approximation properties (3.26) of $\breve{u}_T$ (with

$l = s = p_T + 1$) and (3.40) below, we have for the first term

$$|\mathfrak{T}_1(T)| \leqslant h_T^{1/2}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(\breve{u}_T - u)\|_{\partial T} \times \left(\sum_{F\in\mathcal{F}_T}\frac{\kappa_F}{h_T}\|v_F - v_T\|_F^2\right)^{1/2}$$
$$\lesssim \overline{\kappa}_T^{1/2}\lambda_{\boldsymbol{\kappa},T}^{1/2}\frac{h_T^{p_T+1}}{(p_T+1)^{p_T}}\|u\|_{p_T+2,T}\|\underline{v}_T\|_{\mathrm{a},T}. \quad (3.38)$$

For the second term, the Cauchy–Schwarz inequality followed by (3.28) (with $q = p_T$) readily yields

$$|\mathfrak{T}_2(T)| \lesssim \overline{\kappa}_T^{1/2}\lambda_{\boldsymbol{\kappa},T}^{1/2}\frac{h_T^{p_T+1}}{(p_T+1)^{p_T}}\|u\|_{p_T+2,T}|\underline{v}_T|_{\mathrm{s},T} \lesssim \overline{\kappa}_T^{1/2}\lambda_{\boldsymbol{\kappa},T}^{1/2}\frac{h_T^{p_T+1}}{(p_T+1)^{p_T}}\|u\|_{p_T+2,T}\|\underline{v}_T\|_{\mathrm{a},T}. \quad (3.39)$$

Using (3.38)–(3.39) to estimate the right-hand side of (3.37), applying the Cauchy–Schwarz inequality, and passing to the supremum yields (3.20). To prove (3.21), it suffices to observe that, inserting $\widehat{\underline{u}}_h$ and using the triangle inequality,

$$\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h(u - r_h^{p_h}\underline{u}_h)\|^2 + |\underline{u}_h|_{\mathrm{s},h}^2 \lesssim \|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}_h(u - r_h^{p_h}\widehat{\underline{u}}_h)\|^2 + |\widehat{\underline{u}}_h|_{\mathrm{s},h}^2 + \|\underline{u}_h - \widehat{\underline{u}}_h\|_{\mathrm{a},h}^2,$$

and (3.21) follows using the estimates (3.26), (3.28), and (3.20) to bound the terms in the right-hand side. □

Proposition 3.4.4 (Estimate of boundary difference seminorm)**.** *It holds, for all* $\underline{v}_T \in \underline{U}_T^{p_T}$,

$$\sum_{F\in\mathcal{F}_T}\frac{\kappa_F}{h_T}\|v_F - v_T\|_F^2 \lesssim \lambda_{\boldsymbol{\kappa},T}\|\underline{v}_T\|_{\mathrm{a},T}^2. \quad (3.40)$$

Proof. Let $T \in \mathcal{T}_h$, $\underline{v}_T \in \underline{U}_T^{p_T}$, and set, for the sake of brevity $\breve{v}_T := r_T^{p_T+1}\underline{v}_T$. We have, for all $F \in \mathcal{F}_T$,

$$\begin{aligned}\|v_F - v_T\|_F &= \|\pi_F^{p_F}(v_F - v_T)\|_F\\ &= \|\pi_F^{p_F}(v_F - \breve{v}_T + \pi_T^{p_T}\breve{v}_T - v_T + \breve{v}_T - \pi_T^{p_T}\breve{v}_T)\|_F \quad (3.41)\\ &\leqslant \|\delta_{TF}^{p_T}\underline{v}_T\|_F + \|\breve{v}_T - \pi_T^{p_T}\breve{v}_T\|_F,\end{aligned}$$

where we have used the fact that $p_T \leqslant p_F$ (cf. (3.9)) to infer that $v_{T|F} \in \mathbb{P}^{p_F}(F)$ and thus insert $\pi_F^{p_F}$ in the first line, added and subtracted $(\breve{v}_T - \pi_T^{p_T}\breve{v}_T)$ in the second line, used the triangle inequality together with the definition (3.14) of the face-based residual $\delta_{TF}^{p_T}$ and the $L^2(F)$-stability of $\pi_F^{p_F}$ in the third. To conclude, we observe that, if $p_T = 0$, the discrete trace inequality (3.8) followed by Poincaré's inequality

yield $\|\check{v}_T - \pi_T^0\check{v}_T\|_F \lesssim h_T^{1/2}\underline{\kappa}_T^{-1/2}\|\boldsymbol{\kappa}\boldsymbol{\nabla}\check{v}_T\|_T$ while, if $p_T \geqslant 1$,

$$\begin{aligned}\|\check{v}_T - \pi_T^{p_T}\check{v}_T\|_F &= \|\check{v}_T - \pi_T^0\check{v}_T - \pi_T^{p_T}(\check{v}_T - \pi_T^0\check{v}_T)\|_F\\ &\lesssim \frac{p_T+1}{h_T^{1/2}}\|\check{v}_T - \pi_T^0\check{v}_T - \pi_T^{p_T}(\check{v}_T - \pi_T^0\check{v}_T)\|_T\\ &\lesssim \frac{p_T+1}{h_T^{1/2}}\frac{h_T}{p_T}\|\check{v}_T - \pi_T^0\check{v}_T\|_{1,T} \lesssim h_T^{1/2}\underline{\kappa}_T^{-1/2}\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}\check{v}_T\|_T,\end{aligned}$$

where we have inserted $\pm\pi_T^0\check{v}_T$ in the first line, used the discrete trace inequality (3.8) in the second line, the $L^2(T)$-optimality of $\pi_T^{p_T}$ together with the approximation properties (3.7) (with $l = p_T$ and $q = s = 0$) in the third line, and we have concluded observing that $\frac{p_T+1}{p_T} \leqslant 2$ and using the local Poincaré's inequality (3.6) to infer $\|\check{v}_T - \pi_T^0\check{v}_T\|_{1,T} \lesssim h_T^{1/2}\|\boldsymbol{\nabla}\check{v}_T\|_T$. Plugging the above bounds for $\|\check{v}_T - \pi_T^{p_T}\check{v}_T\|_F$ into (3.41), squaring the resulting inequality, multiplying it by κ_F/h_T, summing over $F \in \mathcal{F}_T$, and recalling the bound (3.4) on $\mathrm{card}(\mathcal{F}_T)$, (3.40) follows. □

3.4.4 L^2-error estimate

Proof of Theorem 3.3.4. We let $z \in U$ solve (3.22) with $g = \widehat{u}_h - u$ and set $\underline{\widehat{z}}_h := \underline{I}_h^{p_h} z$ and, for all $T \in \mathcal{T}_h$ (cf. Proposition 3.4.1),

$$\check{\widehat{z}}_T := r_T^{p_T+1}\underline{\widehat{z}}_T = \varpi_{\boldsymbol{\kappa},T}^{p_T+1} z. \tag{3.42}$$

For the sake of brevity, we also let $\underline{e}_h := \underline{\widehat{u}}_h - \underline{u}_h \in \underline{U}_{h,0}^{p_h}$ (recall the definition (3.19) of $\underline{\widehat{u}}_h$), so that $\widehat{u}_T - u_T = e_T$ for all $T \in \mathcal{T}_h$. We start by observing that

$$\|e_h\|^2 = -(\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}z), e_h) = \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}z, \boldsymbol{\nabla}e_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}z\cdot\boldsymbol{n}_{TF}, e_F - e_T)_F\right), \tag{3.43}$$

where we have used the fact that $-\boldsymbol{\nabla}\cdot(\boldsymbol{\kappa}z) = e_h$ a.e. in Ω followed by element-by-element partial integration together with the continuity of the normal component of $\boldsymbol{\kappa}_T\boldsymbol{\nabla}z$ across interfaces and the strongly enforced boundary conditions in $\underline{U}_{h,0}^{p_h}$ to insert e_F into the last term.

In view of adding and subtracting $\mathrm{a}_h(\underline{e}_h, \underline{\widehat{z}}_h)$ to the right-hand side of (3.43), we next

provide two useful reformulations of this quantity. First, we have

$$\begin{aligned}\mathrm{a}_h(\underline{e}_h,\widehat{\underline{z}}_h) &= \mathrm{a}_h(\widehat{\underline{u}}_h,\widehat{\underline{z}}_h) - \mathrm{a}_h(\underline{u}_h,\widehat{\underline{z}}_h) + (f,z) - (\boldsymbol{\kappa}\boldsymbol{\nabla}u,\boldsymbol{\nabla}z)\\ &= \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}\breve{u}_T,\boldsymbol{\nabla}\breve{z}_T)_T - (\boldsymbol{\kappa}_T\boldsymbol{\nabla}u,\boldsymbol{\nabla}z)_T + \mathrm{s}_T(\widehat{\underline{u}}_T,\widehat{\underline{z}}_T) + (f,z-\pi_T^{p_T}z)_T\right)\\ &= \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}(\breve{u}_T-u),\boldsymbol{\nabla}(\breve{z}_T-z))_T + \mathrm{s}_T(\widehat{\underline{u}}_T,\widehat{\underline{z}}_T) + (f-\pi_T^{p_T}f,z-\pi_T^{1-\Delta_T}z)_T\right),\end{aligned} \tag{3.44}$$

where we have added the quantity $(f,z)-(\boldsymbol{\kappa}\boldsymbol{\nabla}u,\boldsymbol{\nabla}z)=0$ (cf. (3.2)) in the first line, we have passed to the second line using the definition (3.12) of a_T (with $\underline{u}_T=\widehat{\underline{u}}_T$ and $\underline{v}_T=\widehat{\underline{z}}_T$) together with the discrete problem (3.17) to infer

$$\mathrm{a}_h(\underline{u}_h,\widehat{\underline{z}}_h) = (f,z_h) = \sum_{T\in\mathcal{T}_h}(f,\pi_T^{p_T}z)_T,$$

and we have concluded using the definitions (3.25) of $\varpi_{\boldsymbol{\kappa},T}^{p_T+1}$ (together with (3.35) and (3.42)) and (3.3) of $\pi_T^{p_T}$ and $\pi_T^{1-\Delta_T}$. Second, using the definition (3.10) of $r_T^{p_T+1}$ (with $\underline{v}_T=\underline{e}_T$), we obtain

$$\mathrm{a}_h(\underline{e}_h,\widehat{\underline{z}}_h) = \sum_{T\in\mathcal{T}_h}\left((\boldsymbol{\kappa}_T\boldsymbol{\nabla}z,\boldsymbol{\nabla}e_T)_T + \sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}\breve{z}_T,e_F-e_T)_F + \mathrm{s}_T(\widehat{\underline{z}}_T,\underline{e}_T)\right), \tag{3.45}$$

where we have additionally used the fact that $\breve{z}_T=\varpi_{\boldsymbol{\kappa},T}^{p_T+1}z$ (cf. (3.42)) together with the definition (3.25) of $\varpi_{\boldsymbol{\kappa},T}^{p_T+1}$ to replace $\breve{z}_T$ by z in the first term in parentheses.

Thus, adding (3.44) and subtracting (3.45) from (3.43), we obtain after rearranging

$$\|e_h\|^2 = \sum_{T\in\mathcal{T}_h}\left(\mathfrak{T}_1(T)+\mathfrak{T}_2(T)+\mathfrak{T}_3(T)\right), \tag{3.46}$$

with

$$\begin{aligned}\mathfrak{T}_1(T) &:= \sum_{F\in\mathcal{F}_T}(\boldsymbol{\kappa}_T\boldsymbol{\nabla}(z-\breve{z}_T)\cdot\boldsymbol{n}_{TF},e_F-e_T)_F + \mathrm{s}_T(\widehat{\underline{z}}_T,\underline{e}_T),\\ \mathfrak{T}_2(T) &:= (\boldsymbol{\kappa}_T\boldsymbol{\nabla}(\breve{u}_T-u),\boldsymbol{\nabla}(\breve{z}_T-z))_T + \mathrm{s}_T(\widehat{\underline{u}}_T,\widehat{\underline{z}}_T),\\ \mathfrak{T}_3(T) &:= (f-\pi_T^{p_T}f,z-\pi_T^{1-\Delta_T}z)_T.\end{aligned}$$

Using the Cauchy–Schwarz inequality, the approximation properties (3.26) of $\breve{z}_T$ (with $l=p_T+1$ and $s=1$) together with the consistency properties (3.28) of s_T

(with $q = 0$) for the first factor, and the bound (3.40) for the second factor, we get,

$$|\mathfrak{T}_1(T)| \leqslant \Big(h_T\|\boldsymbol{\kappa}_T^{1/2}\boldsymbol{\nabla}(z-\breve{z}_T)\|_{\partial T}^2 + |\widehat{\underline{z}}_T|_{\mathrm{s},T}^2\Big)^{1/2} \times \left(\sum_{F\in\mathcal{F}_T}\frac{\kappa_F}{h_T}\|e_F - e_T\|_F^2 + |\underline{e}_T|_{\mathrm{s},T}^2\right)^{1/2}$$
$$\lesssim \overline{\kappa}_T^{1/2}\lambda_{\boldsymbol{\kappa},T}h_T\|\underline{e}_T\|_{\mathrm{a},T}\|z\|_{2,T}. \quad (3.47)$$

For the second term, the Cauchy–Schwarz inequality followed by the approximation properties (3.26) of $\breve{u}_T$ (with $q = p_T$) and $\breve{z}_T$ (with $q = 1$), and the consistency properties (3.28) of s_T (with $q = p_T$ and $q = 0$ for the first and second factor, respectively) yield

$$\begin{aligned}|\mathfrak{T}_2(T)| &\leqslant \big(\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}(\breve{u}_T - u)\|_T^2 + |\widehat{\underline{u}}_T|_{\mathrm{s},T}^2\big)^{1/2} \times \big(\|\boldsymbol{\kappa}^{1/2}\boldsymbol{\nabla}(\breve{z}_T - z)\|_T^2 + |\widehat{\underline{z}}_T|_{\mathrm{s},T}^2\big)^{1/2}\\ &\lesssim \overline{\kappa}_T\lambda_{\boldsymbol{\kappa},T}\frac{h_T^{p_T+2}}{(p_T+1)^{p_T}}\|u\|_{p_T+2,T}\|z\|_{2,T}.\end{aligned} \quad (3.48)$$

Finally, for the third term we have, when $p_T = 0$,

$$|\mathfrak{T}_3(T)| \leqslant \|f - \pi_T^0 f\|_T\|z - \pi_T^0 z\|_T \lesssim h_T^2\|f\|_{1,T}\|z\|_{1,T} \leqslant h_T^2\|f\|_{1,T}\|z\|_{2,T}, \quad (3.49)$$

while, when $p_T \geqslant 1$,

$$\begin{aligned}|\mathfrak{T}_3(T)| &\leqslant \|f - \pi_T^{p_T} f\|_T\|z - \pi_T^1 z\|_T\\ &\leqslant \|f - \Pi_T^{p_T} f\|_T\|z - \Pi_T^1 z\|_T\\ &\lesssim \frac{h_T^{p_T+2}}{p_T^{p_T+2}}\|f\|_{p_T}\|z\|_{2,T} \leqslant \frac{h_T^{p_T+2}}{p_T^{p_T+2}}\|f\|_{p_T,T}\|z\|_{2,T},\end{aligned} \quad (3.50)$$

where we have used the optimality of $\pi_T^{p_T}$ in the $L^2(T)$-norm to pass to the second line and the approximation properties (3.7) of $\Pi_T^{p_T}$ to conclude. Using (3.47)–(3.50) to bound the right-hand side of (3.46), and recalling the energy error estimate (3.20) and elliptic regularity (3.23), the conclusion follows. □

3.5 Proof of Lemma 3.2.1

Let $\widehat{K} \subset \mathbb{R}^N$ be a L-Lipschitz set (that is, such that its boundary can be locally parametrized by means of $L-$Lipschitz functions) with $\mathrm{diam}(\widehat{K}) = 1$, and fix $r_0 > 1$ and a d-cube $R(r_0)$ containing $\hat{K}$. In the proof of [12, Lemma 4.1] it is shown the

following: Given a function $v \in H^{s+1}(\widehat{K})$, its projection $\Pi^l_{\widehat{K}} v$ on $\mathbb{P}^l(\widehat{K})$ satisfies

$$\|v - \Pi^l_{\widehat{K}} v\|_{q,\widehat{K}} \lesssim \frac{1}{l^{s+1-q}} \|v\|_{s+1,\widehat{K}}, \tag{3.51}$$

for $q \leqslant s+1$ as long as there exists an extension operator $E : H^{s+1}(\widehat{K}) \to H^{s+1}(R(2r_0))$ such that

$$\|E(v)\|_{s+1,R(2r_0)} \leqslant C\|v\|_{s+1,R(\widehat{K})}, \qquad E(v) = 0 \text{ on } R(2r_0)\backslash R\left(\frac{3}{2}r_0\right). \tag{3.52}$$

The existence of such an extension (in any dimension $d \geqslant 1$), is granted by [102, Theorem 5] provided $\widehat{K}$ satisfies some regularity conditions. Namely, by means of a careful inspection of [102, Theorems 5 & 5′], and in particular formulas (25), (30) and the end of the proof of Theorem 5 (p. 192), we get that the constant C in (3.52) depends on the Lipschitz constant L and on the (minimal) number of $L-$*Lipschitz coverings* of $\widehat{K}$, that is, the number of open sets which cover $\partial\widehat{K}$ and in each of whom $\partial\widehat{K}$ can be parametrized by means of an $L-$Lipschitz function. Thus, we get the *hp*-estimate (3.7) provided we show that replacing $\widehat{K}$ with an element T of the mesh, formula (3.51) holds with the appropriate scaling in h_T.

(i) *Proof of* (3.7) *for regular elements.* Assume, for the moment being, that the regularity of $T \in \mathcal{T}_h$ descends from Assumption 1. Let $\widehat{T} := \frac{T}{h_T}$ and suppose, without loss of generality, that the barycenter of T (and thus of $\widehat{T}$) is $\mathbf{0}$. Then, by homogeneity, we get that, for every $f \in H^r(T)$, letting $\widehat{f}(\boldsymbol{x}) := f(\boldsymbol{x}/\lambda)$,

$$\|f\|_{r,T} \lesssim \lambda^{\frac{d}{2}-r} \|\widehat{f}\|_{r,\frac{T}{\lambda}}. \tag{3.53}$$

Thus, setting $r = s+1$, $\lambda = h_T$ and $f = v - \mathfrak{q}$, where $\mathfrak{q}$ is a generic polynomial of degree l, we get by (3.51) (applied to $v - \mathfrak{q}$ and $\Pi^l_T(v-\mathfrak{q})$ in place of v and $\Pi^l_T v$, respectively),

$$\|v - \Pi^l_T v\|_{q,T} = \|(v-\mathfrak{q}) - \Pi^l_T(v-\mathfrak{q})\|_{q,T} \lesssim \frac{h_T^{\frac{d}{2}-q}}{l^{s+1-q}} \|\widehat{v} - \widehat{\mathfrak{q}}\|_{s+1,\widehat{T}}. \tag{3.54}$$

Using [67, Theorem 3.2] and again (3.53) to return to norms on T, we conclude that

$$\|v - \Pi^l_T v\|_{q,T} \lesssim \frac{h_T^{\frac{d}{2}-q}}{l^{s+1-q}} \left(\sum_{i=\min(l,s)}^{s+1} |\widehat{v}|^2_{i,\widehat{T}} \right)^{\frac{1}{2}} \lesssim \frac{h_T^{\min(l,s)-q+1}}{l^{s+1-q}} \|v\|_{s+1,T}.$$

(ii) *Proof of regularity under Assumption 1.* To conclude the proof, we are left to show that Assumption 1 entails uniform bounds only in terms of ρ for the Lipschitz constant of every element $T \in \mathcal{T}_h$. To this aim, consider $\boldsymbol{x} \in \partial T$. Then, $\boldsymbol{x} \in S$ for some (convex) element of the submesh $S \in \mathfrak{T}_h$ contained in T. Since $S \subset T$, it is clear that a bound on the Lipschitz regularity of ∂S immediately implies a bound on the Lipschitz regularity of ∂T. Thus, we focus on the regularity of S. Since S is convex, we can cover ∂S by means of $2(d+1)$ open sets U_i, such that $\partial S \cap U_i$ admits a local convex (and thus Lipschitz) parametrization ϕ_i, i.e., there exists an orthogonal coordinate system such that $\partial S \cap U_i$ is the graph of a Lipschitz function $\phi_i : I_i \subset \mathbb{R}^{d-1} \to \mathbb{R}$. This bound on the number of open sets U_i is crucial to get [102, Theorem 5] to work (clearly, thanks to (3.4), the bound on the number of Lipschitz coverings of T is bounded by a constant $2^d N_\partial = c(d,\rho)$). We claim that each ϕ_i is $1/\rho-$Lipschitz.

Suppose that $\boldsymbol{x} \in U_i =: U$ and set $\phi := \phi_i$. Up to a rotation and a rescaling, we can suppose that $\boldsymbol{x} = \mathbf{0}$ and $\phi(\boldsymbol{x}) = \phi(\mathbf{0}) = 0$. Let now r_s be the inradius of S and h_S be its diameter. By Assumption 1, we know that $\frac{h_s}{r_S} \leqslant \frac{1}{\rho}$. Let B_{r_s} be a ball contained in S of radius r_S. Up to a further rotation of center $\boldsymbol{x} = \mathbf{0}$ of the coordinate system, we can suppose that B_{r_S} is centered on the x_d axis. In place of $\phi : I \to \mathbb{R}$, it is useful to consider its Lipschitz extension $\widetilde{\phi} : \mathbb{R}^{d-1} \to \mathbb{R}$ defined by, denoting by $|\cdot|$ the usual Euclidian norm,

$$\widetilde{\phi}(\boldsymbol{x}) := \inf \left\{ \phi(\boldsymbol{y}) + Lip(\phi)|\boldsymbol{y} - \boldsymbol{x}| \mid \boldsymbol{y} \in \mathbb{R}^{d-1} \right\},$$

We know that $\widetilde{\phi}$ is Lipschitz on $\mathbb{R}^{d-1}$ and that $Lip(\widetilde{\phi}) = Lip(\phi)$ (see for instance [3, Proposition 2.12]). Moreover, it is clear that $\widetilde{\phi}$ is convex on $\mathbb{R}^{d-1}$. The fact that $B_{r_S} \subset S$ and it is centered on the x_d-axis (without loss of generality, we can suppose that its center is $\boldsymbol{\xi} = (\mathbf{0}, \xi_d)$ with $\xi_d > 0$) translates into the fact that B_{r_S} is contained in the epigraph of $\widetilde{\phi}$ and its center has distance from $\mathbf{0} \in \mathbb{R}^d$ at most h_S. Let now $\boldsymbol{p} \in \partial\widetilde{\phi}(0)$, where $\partial\widetilde{\phi}$ is the subdifferential of $\widetilde{\phi}$. Then, for every $\boldsymbol{y} \in \mathbb{R}^{d-1}$ we have

$$\widetilde{\phi}(\boldsymbol{y}) \geqslant \boldsymbol{p} \cdot \boldsymbol{y}.$$

By choosing $\boldsymbol{y} = \lambda \boldsymbol{p}$, with $\lambda \neq 0$, we get the inequality

$$\frac{\widetilde{\phi}(\lambda \boldsymbol{p})}{\lambda |\boldsymbol{p}|} \geqslant |\boldsymbol{p}|. \tag{3.55}$$

Since the epigraph of $\widetilde{\phi}$ contains B_{r_S}, which is centered at a height less than h_S on

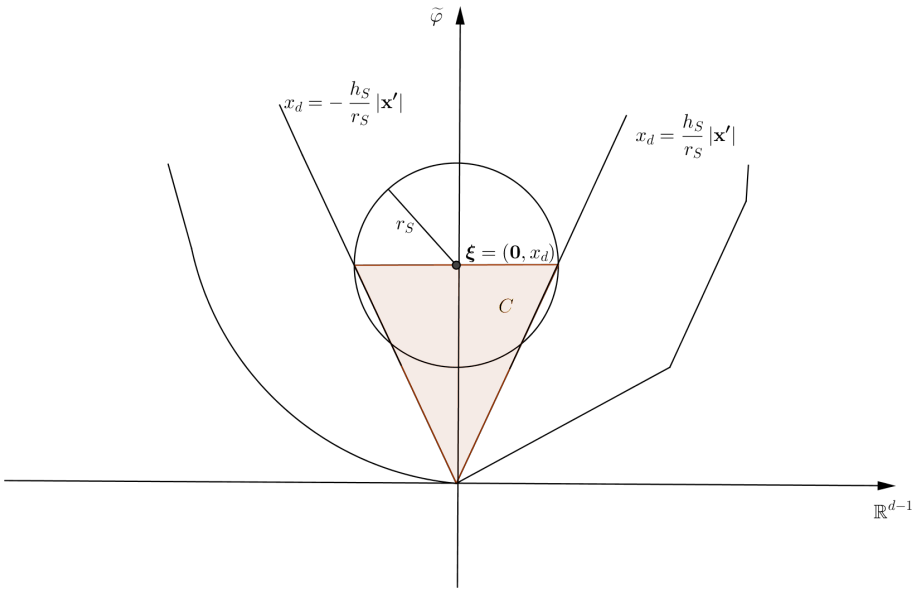


Figure 3.4 – Illustration for point (ii) in the proof of Lemma 3.2.1.

the x_d-axis, and by the convexity of $\widetilde{\phi}$, we have that the truncated cone

$$C = \left\{(\boldsymbol{x}', x_n) \in \mathbb{R}^{d-1} \times \mathbb{R} : h_S \geqslant x_n \geqslant \frac{h_S}{r_S}|\boldsymbol{x}'|\right\},$$

is contained in the epigraph of $\widetilde{\phi}$ (see Figure (3.4)). Then we get from (3.55)

$$|\boldsymbol{p}|^2 \leqslant \widetilde{\phi}(\boldsymbol{p}) \leqslant \frac{h_S}{r_S}|\boldsymbol{p}|$$

and so, by Assumption 1,

$$|\boldsymbol{p}| \leqslant \frac{h_S}{r_S} \leqslant \frac{1}{\rho}.$$

Let now $\boldsymbol{y} \in \mathbb{R}^{d-1}$. Then

$$|\widetilde{\phi}(\boldsymbol{y}) - \widetilde{\phi}(\boldsymbol{x})| \leqslant |\boldsymbol{p}\cdot(\boldsymbol{y} - \boldsymbol{x})| \leqslant |\boldsymbol{p}||\boldsymbol{y} - \boldsymbol{x}| \leqslant \rho^{-1}|\boldsymbol{y} - \boldsymbol{x}|.$$

Since $\boldsymbol{x}$ is arbitrary, this shows that $Lip(\phi) = Lip(\widetilde{\phi}) = \rho^{-1}$, and so that ∂S is $\rho^{-1}-$Lipschitz.

Chapter 4

Perspectives on the numerical reduction of the parametrized diffusion equation

This chapter contains some preliminary work on model order reduction in the context of parametric PDEs. We focus here on the Darcy equation with a parameter-dependent diffusion coefficient and source term. The need for reduced model arises, e.g., in the multi-query context, where one needs to evaluate the solution for a large number of parameters, or in real-time simulations, when the numerical solution must be obtained in a time shorter than the one associated with the system to be modeled. In both cases, resorting to (full) finite element (or HHO/MHO) approximations would lead to unduly large computational times. A possibility is then to approximate the space $\mathcal{M}$ of the parametric solutions by a subspace defined by a finite (usually small) number of "snapshot" solutions. The approximate solution for a given value of the parameter is then obtained by a Galerkin projection on the latter. These ideas are at the core of the Reduced Basis Method (RBM) [87, 92]; examples of applications can be found, e.g., in [29, 97].

Clearly, a key point in this context consists in obtaining a good approximation of the space $\mathcal{M}$. This is done following a Proper Orthogonal Decomposition (equiv. Principal Component Analysis) approach in two steps: first, a (possibly large) number of "trial" solutions is pre-computed; then, a basis is selected from the latter, typically using a greedy algorithm. It is noteworthy that RBM does not separate the two steps, but constructs the basis by successive iterations that alternate with the computation of trial solutions. In practice, one often chooses a greedy algorithm

that is fully incremental: each step of the algorithm computes one new trial solution to enrich the basis with one new dimension. We propose here two possible starting points to obtain a basis for the numerical method at hand inspired, respectively, by the primal and mixed formulation. A relevant difference between the two approaches stems from the Hilbert functional framework. The two approaches can be expected to yield different results, as they correspond to regarding $\mathcal{M}$ as a manifold of the (different) solution spaces associated with each variational formulation.

Our goal is here to investigate this point numerically on different model problems and using different measures for the error. Our numerical results suggest that, for a given number of potentials, the reduction of both the potential and the flux achieved by the (new) algorithm based on the mixed formulation can yield more precise results than more standard reduction techniques, based on the primal formulation.

The chapter is organized as follows. In Section 4.1 we formulate the model problem as well as its primal and mixed formulation. In Section 4.2 we describe the different reduction algorithms used in the computations. Finally, in Section 4.3 we provide an extensive comparison of the algorithms on different model problems.

4.1 Setting

In this section we define the model problem as well as its variational formulations.

4.1.1 Model problem

Let $\Omega \subset \mathbb{R}^d$, $d \geqslant 1$, denote a nonempty bounded connected polyhedral domain with boundary Γ and outward normal $\boldsymbol{n}$. Let $0 < k_{\min} \leqslant k_{\max}$ be two positive real numbers. Let $k(\boldsymbol{\mu}) \in L^\infty(\Omega)$, $f(\boldsymbol{\mu}) \in L^2(\Omega)$ denote two families of real-valued functions parametrized by $\boldsymbol{\mu} \in \boldsymbol{\Lambda} \subset \mathbb{R}^M$, $M \geqslant 1$. We assume that $k(\boldsymbol{\mu})$ is piecewise constant on a partition of Ω into polyhedra and that $k_{\min} \leqslant k(\boldsymbol{\mu}) \leqslant k_{\max}$ a.e. in Ω. We consider the numerical approximation of the family of functions $u(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$, where, for all $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$, $u(\boldsymbol{\mu})$ solves

$$\begin{aligned} -\operatorname{div}(k(\boldsymbol{\mu})\boldsymbol{\nabla} u(\boldsymbol{\mu})) &= f(\boldsymbol{\mu}) \qquad \text{in } \Omega, \\ u(\boldsymbol{\mu}) &= 0 \qquad \text{on } \Gamma. \end{aligned} \tag{4.1}$$

In the numerical investigation carried out in Section 4.3 we will also consider periodic boundary conditions, which are not detailed here for the sake of conciseness.

4.1.2 Primal and mixed variational formulations

For all $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$, $u(\boldsymbol{\mu})$ is the solution of the following primal problem:

Find $u(\boldsymbol{\mu}) \in H^1_0(\Omega)$ such that for all $v \in H^1_0(\Omega)$,

$$(k(\boldsymbol{\mu})\boldsymbol{\nabla} u(\boldsymbol{\mu}), \boldsymbol{\nabla} v)_\Omega = (f(\boldsymbol{\mu}), v)_\Omega, \tag{4.2}$$

whose well-posedness follows from the assumptions.

For all $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$, $u(\boldsymbol{\mu})$ is also found from the solution of the following mixed problem: Find $(\boldsymbol{\sigma}(\boldsymbol{\mu}), p(\boldsymbol{\mu})) \in \mathbf{H}(\mathrm{div};\Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} (\tfrac{1}{k}(\boldsymbol{\mu})\boldsymbol{\sigma}(\boldsymbol{\mu}), \boldsymbol{\tau})_\Omega + (p(\boldsymbol{\mu}), \mathrm{div}\, \boldsymbol{\tau})_\Omega &= 0, \\ -(\mathrm{div}\, \boldsymbol{\sigma}(\boldsymbol{\mu}), q)_\Omega &= (f(\boldsymbol{\mu}), q)_\Omega, \end{aligned} \tag{4.3}$$

for all $\boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega)$ and $q \in L^2(\Omega)$. Also in this case, well-posedness for this problems stems from classical arguments.

Throughout the rest of this chapter, we work under the assumption that the solutions of the primal and mixed formulation coincide.

4.2 The Reduced-Basis Method

In this section we briefly present the Reduced-Basis Methods (RBM) used in the numerical computations of the next section. The starting point for the RBM is a set of solutions $(u(\boldsymbol{\mu}))_{\boldsymbol{\mu} \in \boldsymbol{\Lambda}_{\mathtt{trial}}}$ of the parametric problem (4.1) corresponding to a finite family of parameter values $\boldsymbol{\Lambda}_{\mathtt{trial}} \subset \boldsymbol{\Lambda}$. In practice, these solutions are obtained numerically using, e.g., primal or mixed finite element methods; cf. [7, 32] for a comparison. A key assumption underlying the RBM method is that these approximations are sufficiently accurate: for this reason, in order to simplify the discussion, we neglect the fact that they are numerical approximations and refer to $u(\boldsymbol{\mu})$, $\boldsymbol{\mu} \in \boldsymbol{\Lambda}_{\mathtt{trial}}$, as the exact solution of (4.1).

A key step in the RBM consists in identifying a family of basis functions $(u(\boldsymbol{\mu}_n))_{1 \leqslant n \leqslant N}$

(the so-called *snapshots*) selected from the precomputed solutions $(u(\boldsymbol{\mu}))_{\boldsymbol{\mu}\in\boldsymbol{\Lambda}_{\texttt{trial}}}$ that enables a good representation of the whole family of exact solutions

$$\mathcal{M} := \{u(\boldsymbol{\mu}) \mid \boldsymbol{\mu} \in \boldsymbol{\Lambda}\}.$$

The sequence of snapshots can be defined using a greedy algorithm which constructs the basis incrementally by adding at each iteration the solution corresponding to the parameter $\boldsymbol{\mu} \in \boldsymbol{\Lambda}_{\texttt{trial}}$ that maximizes a suitably defined distance. A key element of the greedy algorithm is the projection operator of $u(\boldsymbol{\mu})$ on the linear subspace $\mathrm{Span}\{u(\boldsymbol{\mu}_n) \mid n = 1, \dots, N\}$. In particular, the projection crucially depends on the Hilbert space X in which the linear subspace $\mathrm{Span}\{u(\boldsymbol{\mu}_n) \mid n = 1, \dots, N\}$ is embedded. We consider here two possible reduction strategies inspired, respectively, by the primal (4.2) and mixed (4.3) formulations.

The approximability of the family $\mathcal{M}$ in the two cases is in general not the same, so that the choice of the approximation in the greedy algorithm can have a sizeable impact on the capability of the RBM method of approximating $\mathcal{M}$. Our goal is to investigate this point with the help of numerical computations. The rest of this section aims at providing details on the greedy algorithms that will be used in the computations.

4.2.1 A reduced-basis method based on the primal formulation

Let us first consider the case when we take inspiration from the primal formulation (4.2). Given $N \geqslant 1$ solutions $\{u(\boldsymbol{\mu}_1), \dots, u(\boldsymbol{\mu}_N)\} \subset H_0^1(\Omega)$ of (4.2) corresponding to a set of parameters $\{\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_N\} \subset \boldsymbol{\Lambda}_{\texttt{trial}}$, the RBM constructs for all $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$ an approximation $u_N(\boldsymbol{\mu})$ of the exact solution $u(\boldsymbol{\mu})$ as a linear combination of these particular solutions

$$u_N(\boldsymbol{\mu}) = \sum_{n=1}^{N} u_n^N(\boldsymbol{\mu}) u(\boldsymbol{\mu}_n),$$

with $(u_n^N(\boldsymbol{\mu}))_{1\leqslant n\leqslant N} \subset \mathbb{R}$ sequence of real numbers. For any $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$, denoting

$$\mathrm{U}^N = \mathrm{Span}\{u(\boldsymbol{\mu}_n) \mid n = 1 \dots N\}, \tag{4.4}$$

the reduced basis approximation $u_N(\boldsymbol{\mu})$ of $u(\boldsymbol{\mu})$ satisfies

$$(k(\boldsymbol{\mu})\boldsymbol{\nabla} u_N(\boldsymbol{\mu}), \boldsymbol{\nabla} v)_\Omega = (f(\boldsymbol{\mu}), v)_\Omega \quad \forall v \in \mathrm{U}^N \tag{4.5}$$

Algorithm 1 The Greedy Algorithm

Input: A set of parameter $\mathbf{\Lambda}_{\texttt{trial}}$, a tolerance error $\varepsilon_{\texttt{greedy}} > 0$.

1: Select an arbitrary $\boldsymbol{\mu}_1 \in \mathbf{\Lambda}_{\texttt{trial}}$ and compute the solution $u(\boldsymbol{\mu}_1)$
2: Define $\mathrm{U}^1 = \mathrm{Span}\{u(\boldsymbol{\mu}_1)\}$ with $u(\boldsymbol{\mu}_1)$ solution of (4.2)
3: $N \leftarrow 1$
4: **while** $\sup_{\boldsymbol{\mu}\in\mathbf{\Lambda}_{\texttt{trial}}} |||u(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})||| > \varepsilon_{\texttt{greedy}}$ **do**
5: Define $\boldsymbol{\mu}_{N+1} = \mathrm{argmax}_{\boldsymbol{\mu}\in\mathbf{\Lambda}_{\texttt{trial}}} |||u(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})|||$
6: Compute $u(\boldsymbol{\mu}_{N+1})$ from (4.2)
7: Define $\mathrm{U}^{N+1} := \mathrm{Span}\{u(\boldsymbol{\mu}_n) \mid n = 1, \ldots, N+1\}$
8: $N \leftarrow N + 1$
9: **end while**

Output: A family of basis functions $(u(\boldsymbol{\mu}_n))_{1\leqslant n\leqslant N}$ corresponding to the parameters $(\boldsymbol{\mu}_n)_{1\leqslant n\leqslant N}$, a family of nested spaces $(\mathrm{U}^n)_{1\leqslant n\leqslant N}$.

such that $u_N(\boldsymbol{\mu})$ can be seen as the Ritz-Galerkin projection of $u(\boldsymbol{\mu})$ on U^N.

As already pointed out, the key point in using RBM is the choice of the snapshots in the definition of U^N, which can be done for instance through a greedy algorithm or simply randomly. The version of the greedy algorithm used in the computations of the following section in the case where $u(\boldsymbol{\mu})$ is regarded as the solution of the primal formulation (4.2) is detailed in Algorithm 1. The choice of the triple norm measuring the distance between two elements of the solution space in lines 4 and 5 of Algorithm 1 is to some extent arbitrary. We consider here the choices $|||u(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu})||| = P_N^{\alpha,\theta}(\boldsymbol{\mu})$ where

$$P_N^{\alpha,\theta}(\boldsymbol{\mu}) := \|k(\boldsymbol{\mu})^{\theta\alpha}\boldsymbol{\nabla}^{\alpha}(u(\boldsymbol{\mu}) - u_N(\boldsymbol{\mu}))\|_{L^2(\Omega)}, \tag{4.6}$$

with $\alpha \in \{0,1\}$ and $\theta \in \{0,1\}$. Specifically, for $\alpha = 0$ we obtain the L^2-norm of the potential, for $\alpha = 1$ and $\theta = 0$ the L^2-norm of the gradient, and for $\alpha = \theta = 1$ the L^2-norm of the flux.

Remark 4.2.1 (Convergence rates for the RBM). The actual sequence of basis functions generated by the Greedy argument clearly depends on the choice of the triple norm, the trial set $\mathbf{\Lambda}_{\texttt{trial}}$, the initial (randomly selected) parameter $\boldsymbol{\mu}_1$, and on the numerical method (e.g., finite element or HHO/MHO) used to approximate $u(\boldsymbol{\mu}_n)$. Nevertheless, numerous implementations of RBM have shown approximation errors with good (typically exponential) convergence rates uniformly in $\mathbf{\Lambda}$ for small N; cf., e.g., [29, 97]. Although the fast convergence of RBM is not fully understood, the potential for exponential convergence is usually explained after interpreting the

greedy algorithm as the computation of an upper-bound for

$$d_N^{H_0^1(\Omega)}(\mathcal{M}) := \inf_{\substack{V_N \subset H_0^1(\Omega) \\ \dim(V_N)=N}} \sup_{u \in \mathcal{M}} \inf_{v^N \in V_N} |||u - v^N|||, \tag{4.7}$$

the *Kolmogorov $N-$widths* of $\mathcal{M}$ as a subset of the Hilbert space $H_0^1(\Omega)$ equipped with the $|||\cdot|||$-norm (the reduced basis of dimension N is a suboptimal solution to the infimum in (4.7)). We observe, in passing, that Kolmogorov widths which decrease fast with $N \geqslant 1$ entail a similarly fast decay of the RBM approximation error for $\mathcal{M}$, *asymptotically*, see e.g. [22].

4.2.2 Two reduced-basis methods based on the mixed formulation

We discuss here two reduction strategies based on the mixed formulation (4.3). The first reduces both potentials and fluxes. In this case, we exhibit a corresponding greedy algorithm where inf-sup stability is achieved by adding supremizers to the basis for the flux. The second addresses a potential-based reduction strategy used for comparison purposes with the first formulation.

A potential-flux reduction strategy

Given N_s flux snapshots $\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{N_s}\} \subset \mathbf{H}(\mathrm{div}; \Omega)$ and N_p potential snapshots $\{p_1, \ldots, p_{N_p}\} \subset L^2(\Omega)$, we construct in this case a flux-potential couple

$$\boldsymbol{\sigma}_{N_s,N_p}(\boldsymbol{\mu}) = \sum_{n=1}^{N_s} \sigma_n^{N_s,N_p}(\boldsymbol{\mu})\boldsymbol{\sigma}_n \quad p_{N_s,N_p}(\boldsymbol{\mu}) = \sum_{n=1}^{N_p} p_n^{N_s,N_p}(\boldsymbol{\mu})p_n,$$

where $(\sigma_n^{N_s,N_p}(\boldsymbol{\mu}))_{1 \leqslant n \leqslant N_s}$ and $(p_n^{N_s,N_p}(\boldsymbol{\mu}))_{1 \leqslant n \leqslant N_p}$ are sequences of real numbers. Denoting the spaces of reduced fluxes and potentials by

$$\boldsymbol{S}^{N_s,N_p} := \mathrm{Span}\{\boldsymbol{\sigma}_n \mid n = 1, \ldots, N_s\} \text{ and } \mathrm{Q}^{N_s,N_p} := \mathrm{Span}\{p_n \mid n = 1, \ldots, N_p\} \tag{4.8}$$

for any $\boldsymbol{\mu} \in \boldsymbol{\Lambda}$ the reduced basis approximation $(\boldsymbol{\sigma}_{N_s,N_p}(\boldsymbol{\mu}), p_{N_s,N_p}(\boldsymbol{\mu}))$ is the solution of

$$\begin{aligned} (\tfrac{1}{k}(\boldsymbol{\mu})\boldsymbol{\sigma}_{N_s,N_p}(\boldsymbol{\mu}), \boldsymbol{\tau})_\Omega + (p_{N_s,N_p}(\boldsymbol{\mu}), \mathrm{div}\, \boldsymbol{\tau})_\Omega &= 0, \\ -(\mathrm{div}\, \boldsymbol{\sigma}_{N_s,N_p}(\boldsymbol{\mu}), q)_\Omega &= (f(\boldsymbol{\mu}), q)_\Omega, \end{aligned} \tag{4.9}$$

Algorithm 2 The Greedy Algorithm combined with flux enrichment

Input: A set of parameter $\boldsymbol{\Lambda}_{\texttt{trial}}$, a tolerance error $\varepsilon_{\texttt{greedy}} > 0$.

Pick an arbitrary $\boldsymbol{\mu}_1 \in \boldsymbol{\Lambda}_{\texttt{trial}}$ and compute the solution $(p(\boldsymbol{\mu}_1), \boldsymbol{\sigma}(\boldsymbol{\mu}_1))$ of (4.3).
$N \leftarrow 1$.
Compute the supremizer $\widehat{\boldsymbol{\sigma}}_1$ associated to $p(\boldsymbol{\mu}_1)$.
Define $\boldsymbol{S}^{2N,N} := \mathrm{Span}\{\boldsymbol{\sigma}(\boldsymbol{\mu}_1), \widehat{\boldsymbol{\sigma}}_1\}$.
Define $\mathrm{Q}^{N,N} := \mathrm{Span}\{p(\boldsymbol{\mu}_1)\}$.
while $\sup_{\boldsymbol{\mu}\in\boldsymbol{\Lambda}_{\texttt{trial}}} |||\boldsymbol{\sigma}(\boldsymbol{\mu}) - \boldsymbol{\sigma}_{2N,N}(\boldsymbol{\mu})||| > \varepsilon_{\texttt{greedy}}$ **do**
    Define $\boldsymbol{\mu}_{N+1} = \mathrm{argmax}_{\boldsymbol{\mu}\in\boldsymbol{\Lambda}_{\texttt{trial}}} \sup_{\boldsymbol{\mu}\in\boldsymbol{\Lambda}_{\texttt{trial}}} |||\boldsymbol{\sigma}(\boldsymbol{\mu}) - \boldsymbol{\sigma}_{2N,N}(\boldsymbol{\mu})||| > \varepsilon_{\texttt{greedy}}$
    Compute the solution $(p(\boldsymbol{\mu}_{N+1}), \boldsymbol{\sigma}(\boldsymbol{\mu}_{N+1}))$ of (4.3).
    $N \leftarrow N + 1$.
    Compute the supremizer $\widehat{\boldsymbol{\sigma}}_N$ associated to $p(\boldsymbol{\mu}_N)$.
    Define $\boldsymbol{S}^{2N,N} := \mathrm{Span}\{\boldsymbol{\sigma}(\boldsymbol{\mu}_1), \ldots, \boldsymbol{\sigma}(\boldsymbol{\mu}_N), \widehat{\boldsymbol{\sigma}}_1, \ldots, \widehat{\boldsymbol{\sigma}}_N\}$.
    Define $\mathrm{Q}^{2N,N} := \mathrm{Span}\{p(\boldsymbol{\mu}_n) \mid 1 \leqslant n \leqslant N\}$.
end while
$N_p \leftarrow N$.

Output: A family of parameters $(\boldsymbol{\mu}_n)_{1\leqslant n\leqslant N_p}$, a family of nested spaces $(\boldsymbol{S}^{2n,n})_{1\leqslant n\leqslant N_p}$ and $(\mathrm{Q}^{2n,n})_{1\leqslant n\leqslant N_p}$.

for all $(\boldsymbol{\tau}, q) \in \boldsymbol{S}^{N_s,N_p} \times \mathrm{Q}^{N_s,N_p}$. Contrary to the primal formulation, the well-posedness of the above saddle-point problem is not automatically guaranteed and depends on the construction of the reduced spaces (4.8); see [24] for a general introduction to the analysis and approximation of mixed problems. One possible stabilization strategy is to enrich the reduced flux spaces, as suggested, e.g., in [98], with additional fluxes called *supremizers*. Precisely, let $(\boldsymbol{\sigma}(\boldsymbol{\mu}_n), p(\boldsymbol{\mu}_n))_{1\leqslant n\leqslant N_p}$ be N_p flux-potential solutions of (4.3) associated to N_p parameters $(\boldsymbol{\mu}_n)_{1\leqslant n\leqslant N_p}$. We compute N_p additional fluxes $(\widehat{\boldsymbol{\sigma}}_n)_{1\leqslant n\leqslant N_p}$ as the Riesz representants of the flux-potential coupling map with respect to the scalar product of $\mathbf{H}(\mathrm{div};\Omega)$. By construction, the problem (4.9) is well posed in the spaces $\boldsymbol{S} := \mathrm{Span}\{\boldsymbol{\sigma}_1, \ldots, \boldsymbol{\sigma}_{N_p}, \widehat{\boldsymbol{\sigma}}_1, \ldots, \widehat{\boldsymbol{\sigma}}_{N_p}\}$ and $\mathrm{Q} := \mathrm{Span}\{p(\boldsymbol{\mu}_n) \mid n = 1, \ldots, N_p\}$.

This stabilization technique can then be incorporated into a greedy algorithm, as detailed in the Algorithm 2, such that every couple of reduced spaces $\{(\boldsymbol{S}^{2N,N}, \mathrm{Q}^{2N,N})\}_{1\leqslant N\leqslant N_p}$ is inf-sup stable (thus, in our case, $N_s = 2N_p$). Given $\theta \in \{0, 1\}$, we choose the triple norm as $|||\boldsymbol{\sigma}(\boldsymbol{\mu}) - \boldsymbol{\sigma}_{2N,N}(\boldsymbol{\mu})||| = D_N^{1,\theta}(\boldsymbol{\mu})$ where

$$D_N^{1,\theta}(\boldsymbol{\mu}) := \|k(\boldsymbol{\mu})^{\theta-1}(\boldsymbol{\sigma}(\boldsymbol{\mu}) - \boldsymbol{\sigma}_{2N,N}(\boldsymbol{\mu}))\|_{L^2(\Omega)}. \tag{4.10}$$

For $\theta = 0$ this corresponds to the norm of the gradient, whereas for $\theta = 1$ we obtain the norm of the flux. Additionally, in the numerical tests, we also consider the

L^2-error on the potential given by

$$D_N^{0,0}(\boldsymbol{\mu}) = \|p(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_{L^2(\Omega)}. \tag{4.11}$$

A potential-based reduction strategy

One may also simply not reduce the flux space in the mixed formulation (4.9). using the error expressions given by (4.6) and (4.10). Given $N \geqslant 1$ potentials $(p_n)_{1\leqslant n\leqslant N}$ and denoting $\mathrm{Q}^N := \mathrm{Span}\{\, p_n \mid 1 \leqslant n \leqslant N\}$, the mixed formulation where the flux are not reduced reads: Find

$$\boldsymbol{\sigma}_N(\boldsymbol{\mu}) \in \mathbf{H}(\mathrm{div};\Omega), \quad p_N(\boldsymbol{\mu}) = \sum_{n=1}^{N} p_n^N(\boldsymbol{\mu}) p_n,$$

with coefficients $(p_n^N(\boldsymbol{\mu}))_{1\leqslant n\leqslant N}$ such that for all $\boldsymbol{\tau} \in \mathbf{H}(\mathrm{div};\Omega)$ and $q \in \mathrm{Q}^N$, it holds

$$\begin{aligned} (\tfrac{1}{k}(\boldsymbol{\mu})\boldsymbol{\sigma}_N(\boldsymbol{\mu}), \boldsymbol{\tau})_\Omega + (p_N(\boldsymbol{\mu}), \mathrm{div}\,\boldsymbol{\tau})_\Omega &= 0, \\ -(\mathrm{div}\,\boldsymbol{\sigma}_N(\boldsymbol{\mu}), q)_\Omega &= (f(\boldsymbol{\mu}), q)_\Omega. \end{aligned} \tag{4.12}$$

Even though this formulation has no practical interest in the context of real-time computations, it is interesting for comparison purposes with the mixed formulation (4.9). In this case the error measure is defined as

$$\breve{D}_N^{1,\theta}(\boldsymbol{\mu}) = \|k(\boldsymbol{\mu})^{\theta-1}(\boldsymbol{\sigma}(\boldsymbol{\mu}) - \boldsymbol{\sigma}_N(\boldsymbol{\mu}))\|_{L^2(\Omega)} \qquad \theta \in \{0,1\}. \tag{4.13}$$

For $\theta = 0$ this corresponds to the norm of the gradients, whereas for $\theta = 1$ we obtain the norm of the flux. We also consider the L^2-error on the potential given by

$$\breve{D}_N^{0,0}(\boldsymbol{\mu}) = \|p(\boldsymbol{\mu}) - p_N(\boldsymbol{\mu})\|_{L^2(\Omega)}. \tag{4.14}$$

4.3 Numerical investigation

In this section we compare the error rates observed using the three reduction strategies highlighted in the previous section for different sequences of parameter values issued from the corresponding greedy algorithm. We focus on two different parametric problems entering the framework of Section 4.1.

4.3.1 Model problems

We let $\Omega := (0,1)^2$, and we consider parameter spaces that are subsets of $\mathbb{R}^4$. Thus, the parameter $\boldsymbol{\mu}$ has components $(\mu_1, \mu_2, \mu_3, \mu_4)$ some of which may be set to zero. The diffusion coefficient $k : \Omega \times \mathbb{R}^2 \to \mathbb{R}$ depends on two real parameters μ_1 and μ_2 according to the following expression:

$$k(\boldsymbol{x}, \boldsymbol{\mu}) = \langle k \rangle + \mu_1 \Psi_1(\boldsymbol{x}) \sqrt{\lambda_1} + \mu_2 \Psi_2(\boldsymbol{x}) \sqrt{\lambda_2}, \tag{4.15}$$

where λ_1 and λ_2 are assumed to be fixed positive constants and $\langle k \rangle$ is also a positive constant, chosen in practice such that $k(\cdot, \boldsymbol{\mu}) > 0$ on Ω. The functions Ψ_1 and Ψ_2 are piecewise constants over Ω, and their values are ± 1 on particular dyadic subdivisions of Ω, precisely

$$\Psi_1(\boldsymbol{x}) := \begin{array}{|c|c|}\hline 1 & -1 \\ \hline -1 & 1 \\ \hline \end{array}, \qquad \Psi_2(\boldsymbol{x}) := \begin{array}{|c|c|c|c|c|c|c|c|}\hline -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ \hline 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ \hline -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ \hline 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ \hline -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ \hline 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ \hline -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ \hline 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ \hline \end{array}. \tag{4.16}$$

Problem 1 (Homogeneous Dirichlet problem)**.** We prescribe homogeneous Dirichlet conditions on Γ. The load function $f(\boldsymbol{\mu}) = 1$, such that the parameter dependence only appears within the coefficient $k(\boldsymbol{\mu})$. The problem reads: Find $u(\boldsymbol{\mu})$ s.t.

$$-\operatorname{div}(k(\boldsymbol{\mu}) \boldsymbol{\nabla} u(\boldsymbol{\mu})) = 1 \text{ on } \Omega, \text{ with } u(\boldsymbol{\mu}) = 0 \text{ on } \Gamma, \tag{4.17}$$

and is parametrized by the two real parameters μ_1 and μ_2 appearing in the expression (4.15) of k. The two other components of $\boldsymbol{\mu}$ are set to zero, i.e., $\mu_3 = \mu_4 = 0$.

Problem 2 (Periodic problem)**.** We consider periodic boundary conditions on Γ together with a zero-mean constraint on $u(\boldsymbol{\mu})$ and a parametrized right-hand side $f(\boldsymbol{\mu})$. The problem reads: Find $u(\boldsymbol{\mu})$ s.t.

$$-\operatorname{div}(k(\boldsymbol{\mu}) \boldsymbol{\nabla} u(\boldsymbol{\mu})) = f(\boldsymbol{\mu}) \quad \text{in } \Omega, \tag{4.18a}$$

$$\int_\Omega u(\boldsymbol{\mu}) dx, = 0 \qquad \text{with periodic B.C.}, \tag{4.18b}$$

with $f(\boldsymbol{\mu}) := \mu_3 \sin(2\pi x) \sin(2\pi y) + \mu_4 (x - y)$. In this case, we have four parameters corresponding to the components $\mu_1, \mu_2, \mu_3, \mu_4$ of $\boldsymbol{\mu}$.

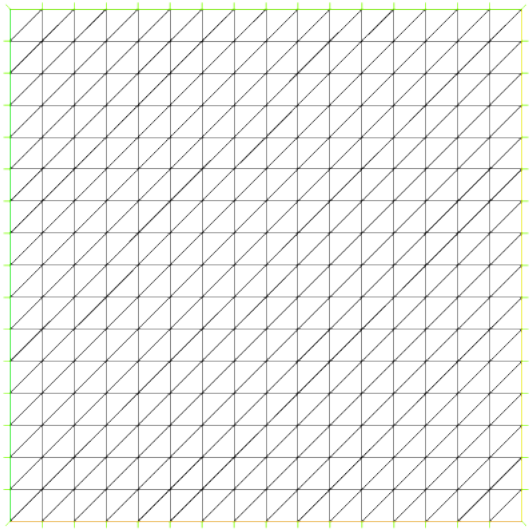


Figure 4.1 – Approximation of Ω by a triangular mesh $\mathcal{T}_h$ for numerical computations

Given a set of parameters $\mathbf{\Lambda}_{\texttt{test}}$ significantly larger than $\mathbf{\Lambda}_{\texttt{trial}}$, we investigate the decay of the following

$$\sup_{\boldsymbol{\mu}\in\mathbf{\Lambda}_{\texttt{test}}} \mathcal{E}_N, \tag{4.19}$$

when N increases and $\mathcal{E}_N$ is either $P_N^{1,\theta}, D_N^{1,\theta}, \check{D}_N^{1,\theta}, P_N^{0,0}$ or $D_N^{0,0}$ for $\theta \in \{0,1\}$. This quantity can be taken as a measure of the capability of the reduced basis to approximate the solution manifold $\mathcal{M}$. We perform these computations for the above two problems and two different ways to compute the reduced bases (4.4) and (4.8). The first is by using a fixed set of parameters, generated randomly. The second is by selecting parameters by using a Greedy Algorithm based on the primal formulation of each of the two problems, and build a flux-potential space with this specific set of parameter by taking care of adding the supremizers.

4.3.2 Numerical settings

For the numerical computations we use the `FreeFem++` software [79]. The domain Ω is approximated with a triangular mesh $\mathcal{T}_h$ consisting in 512 triangular elements (with meshsize $h \simeq \texttt{1.18E}-\texttt{2}$), see Figure 4.1. Letting $A = \sqrt[3]{3/2} \simeq 1.1447$, we assume the following bounds over the components of $\boldsymbol{\mu}$,

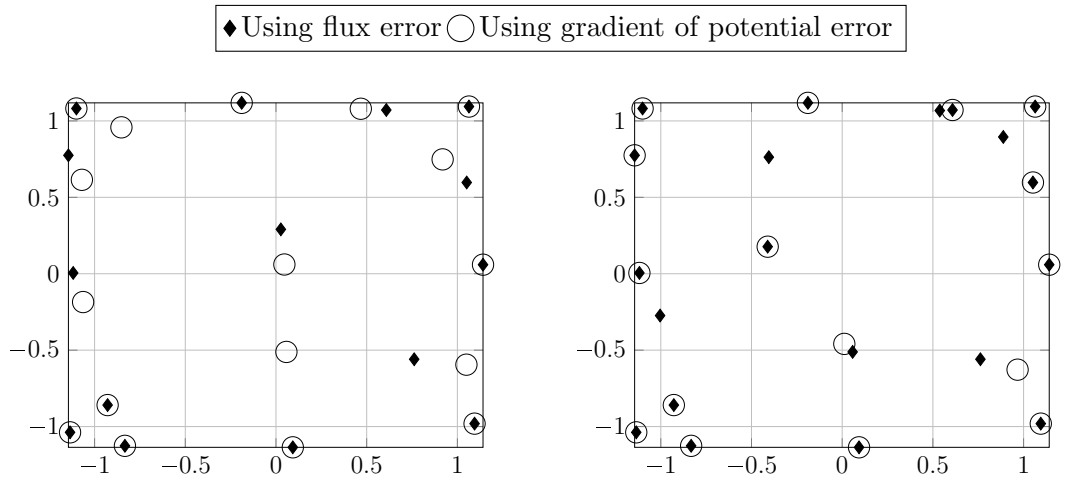
$$\boldsymbol{\mu} \in [-A, A]^4. \tag{4.20}$$

We express $k(\boldsymbol{\mu})$ in the form (4.15) with $\langle k \rangle = 10$, $\lambda_1 = 2$ and $\lambda_2 = 6$. The set $\boldsymbol{\Lambda}_{\texttt{trial}}$, on which we perform the Greedy Algorithm is made of 300 parameters satisfying the range condition (4.20). The supremum in the definition of the $N-$widths is taken over the set $\boldsymbol{\Lambda}_{\texttt{test}}$ made of 1700 parameters all satisfying the bound condition (4.20). The tolerance error $\varepsilon_{\texttt{greedy}}$ in the greedy algorithms 1 and 2 is set to $\texttt{1E}-\texttt{6}$.

4.3.3 Discussion

In what follows, we collect the results of the numerical investigation over nine Figures 4.3–4.11 based on the numerical assumptions given in Section 4.3.2. They depict the decay with respect to the number of basis functions N of the quantity defined by (4.19) for Problems 1 and 2 with different constructions of the reduced spaces U^N, $(\boldsymbol{S}^{2N,N}, \mathrm{Q}^{2N,N})$ and Q^N. The Figures 4.3, 4.4 and 4.5 correspond to the case where the reduced basis is constructed with a fixed arbitrary parameters sample. The Figures 4.6, 4.7 and 4.8 correspond to the case where the reduced basis is constructed using Algorithm 1. Finally, the Figures 4.9, 4.10 and 4.11 corresponds to the case where the reduced basis is constructed using Algorithm 2. According to the problem considered, we observe either a clear advantage using the mixed formulation, or similar decay rates between the two formulations. The primal formulation rarely gives better performances. In Figures 4.3, 4.4 and 4.5, we address the case where the reduced spaces are built from a fixed set of parameters, generated randomly (i.e., we do not use the greedy algorithm). The Figure 4.3 treats the case where the flux error is measured throught the quantities $P_N^{1,1}, D_N^{1,1}$ and $\breve{D}_N^{1,1}$. For Problem 1, Figure 4.3(a) shows better decay rates for the mixed approach when considering Problem 1. The decay rates are, on the other hand, similar for the periodic problem; cf. Figure 4.3(b). In Figure 4.4 we compare the decays using the error on the gradient given by $P_N^{1,0}, D_N^{1,0}$, and $\breve{D}_N^{1,0}$, respectively. Similar considerations hold as for the case when the flux error is considered. For the sake of completeness, we also display the decays using the L^2-error on the potential in Figure 4.5, for which again similar considerations hold between problem 1 in Figure 4.5(a) and problem 2 in Figure 4.5(b).

Figure 4.6 addresses the case where the reduced spaces are built upon a family of parameters computed with the greedy Algorithm 1 for the primal formulation (4.5). The Figure 4.2 displays the selected parameters in the range domain for different choice of the $|||\cdot|||$-norm in the context of Problem 1. Similar conclusions as for the case when no greedy algorithm is used can be drawn. Taking the triple norm $|||\cdot|||$



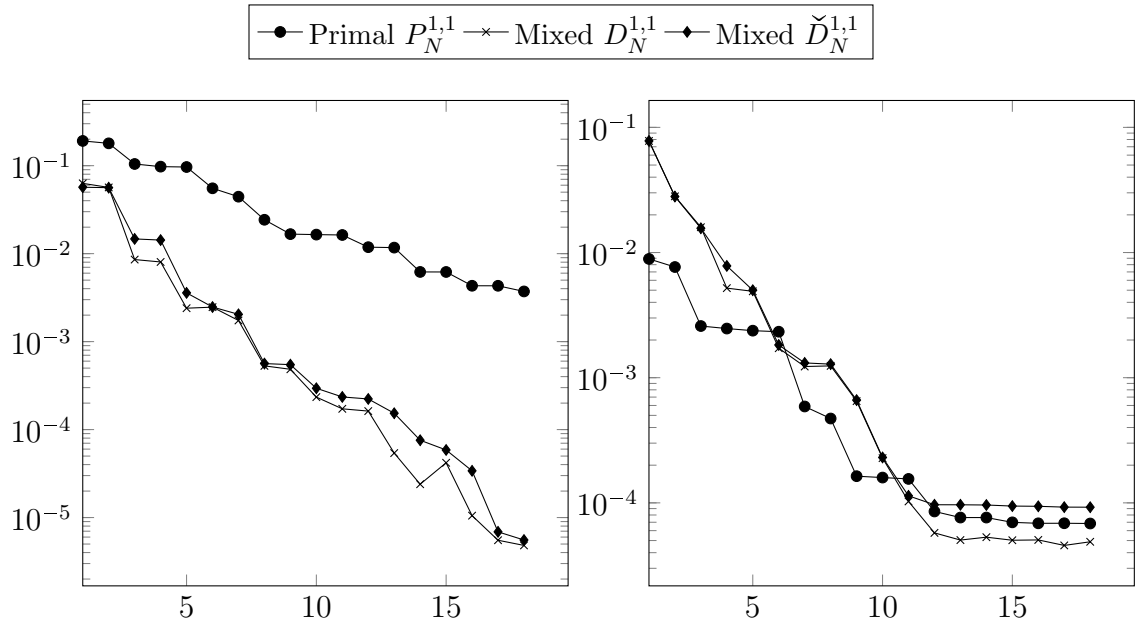
(a) Primal formulation (b) Mixed formulation

Figure 4.2 – Parameter families generated by the Greedy Algorithm 1 (left) and by the Greedy Algorithm 2 (right) in the context of the Problem 1 for the flux and gradient norms.

equal to the norm of the flux or of the gradient does not seem to have an impact on the results, as can be appreciated comparing Figures 4.6 with 4.7. However, in the case of the periodic Problem 2, a stagnation of the error is observed for both of the mixed formulation (with dot and cross marks) in Figures 4.6(b), 4.7(b) and 4.8(b), while this is not the case for the primal formulation (with dot marks).

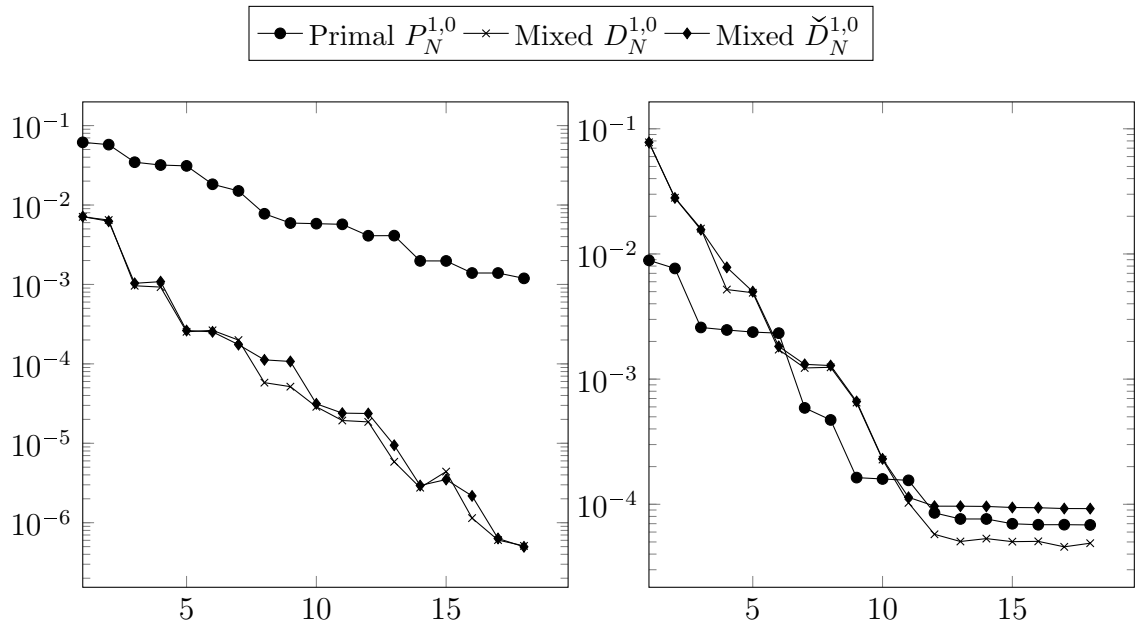
Finally, Figures 4.9, 4.10 and 4.11 treat the case where the family of parameters are built upon a Greedy Algorithm 2 adapted for the mixed formulation. Overall, the mixed formulation remains advantageous for both problems, see Figures 4.9(a), 4.10(a) and 4.11(a). This is probably due to the specific choice of parameters resulting from the Greedy Algorithm 2.

In computations not shown here, we have also considered the case of parametric Robin boundary conditions, for which results comparable to those obtained with periodic boundary conditions are observed.



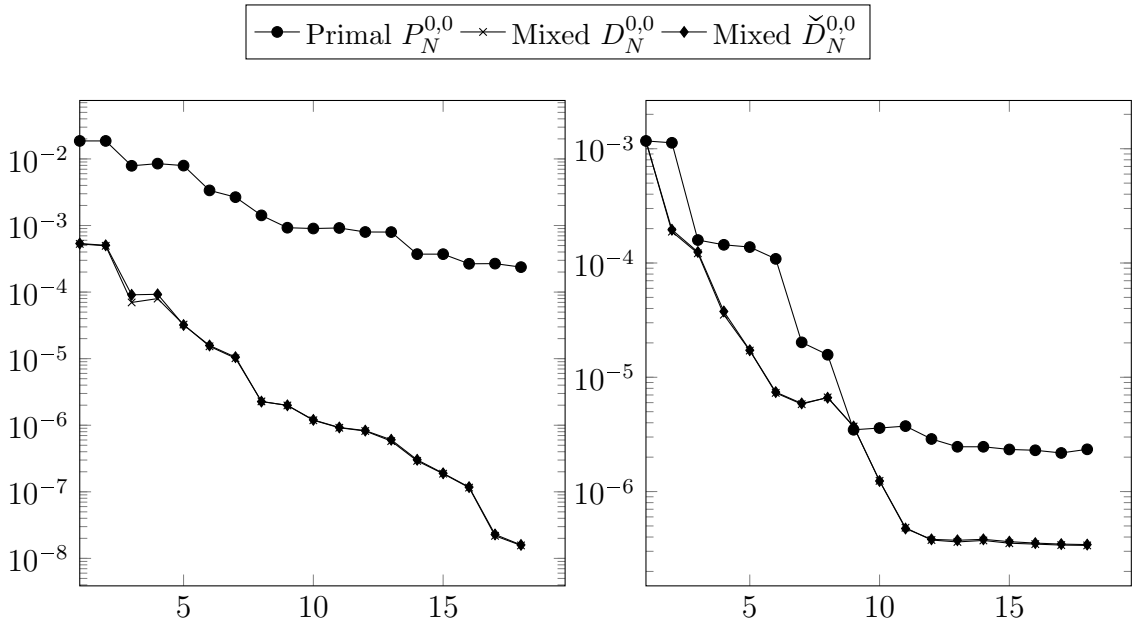
(a) Problem 1 (b) Problem 2

Figure 4.3 – Flux errors $P_N^{1,1}$, $D_N^{1,1}$ and $\breve{D}_N^{1,1}$ for the two problems when the parameter sample used to generate a reduced basis is chosen **arbitrarily**.



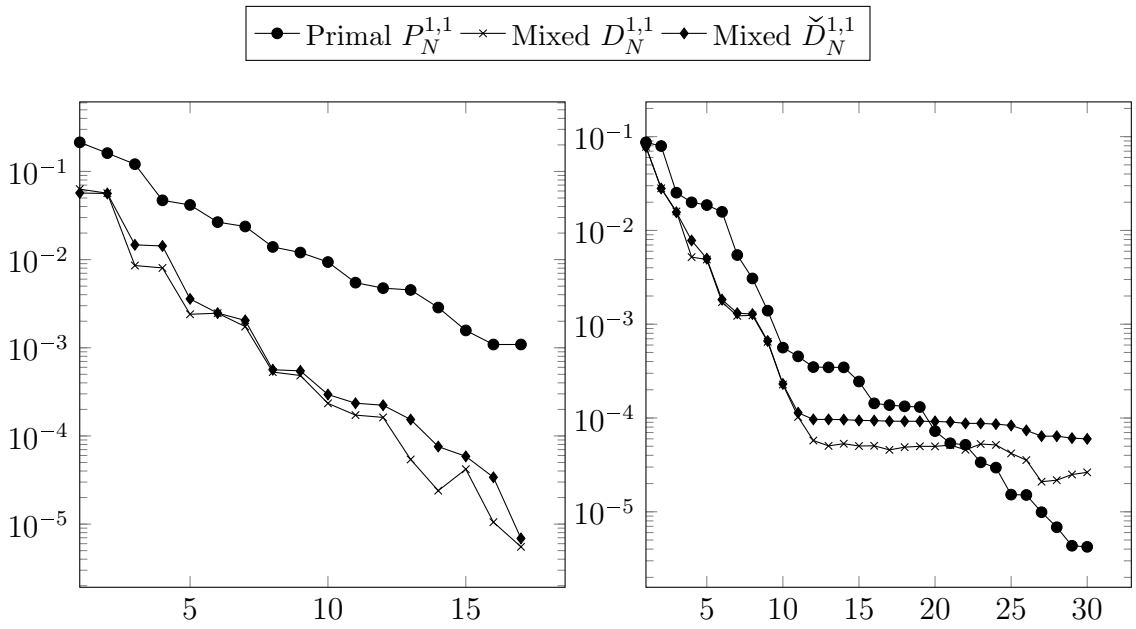
(a) Problem 1 (b) Problem 2

Figure 4.4 – Gradient of potentials errors $P_N^{1,0}$, $D_N^{1,0}$ and $\breve{D}_N^{1,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen **arbitrarily**.



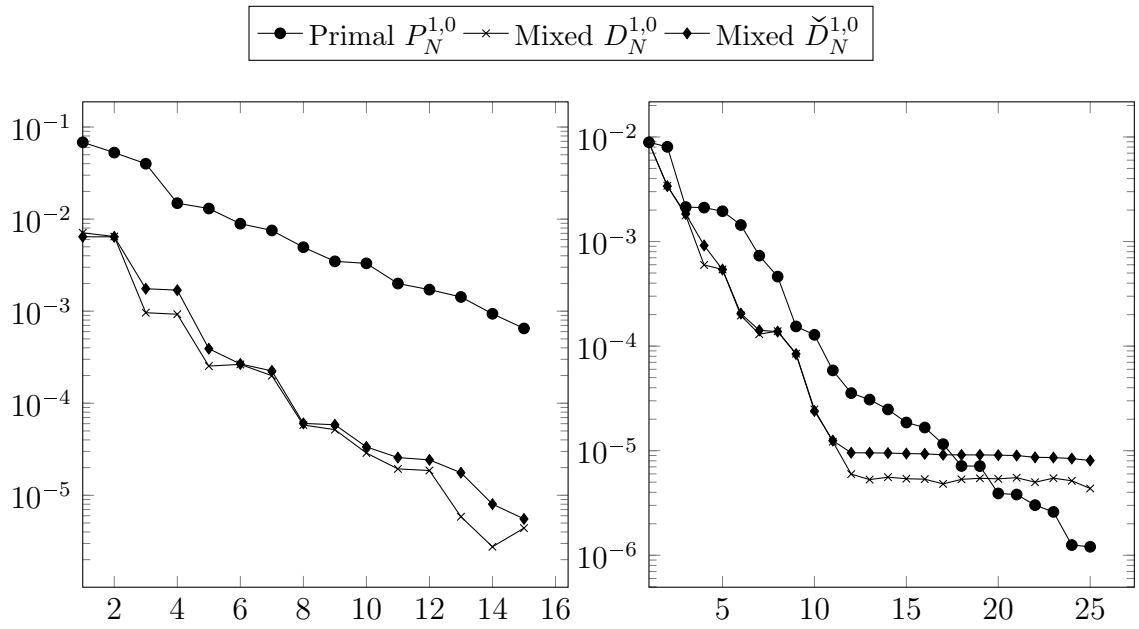
(a) Problem 1 (b) Problem 2

Figure 4.5 – Potentials errors $P_N^{0,0}$, $D_N^{0,0}$ and $\breve{D}_N^{0,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen **arbitrarily**.



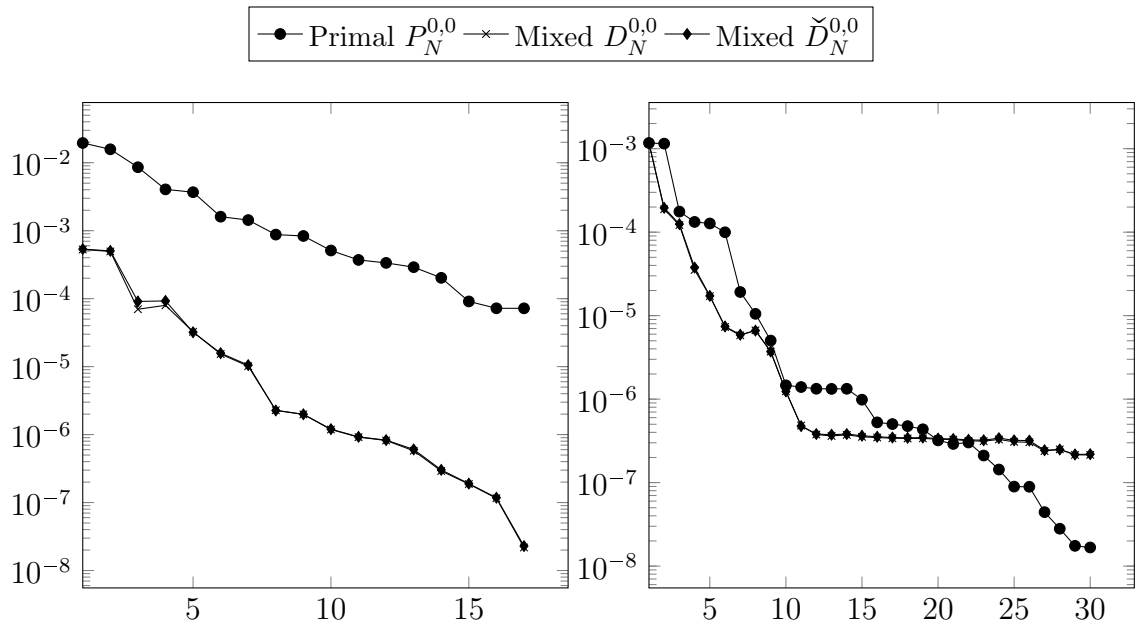
(a) Problem 1 (b) Problem 2

Figure 4.6 – Flux errors $P_N^{1,1}$, $D_N^{1,1}$ and $\breve{D}_N^{1,1}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 1.



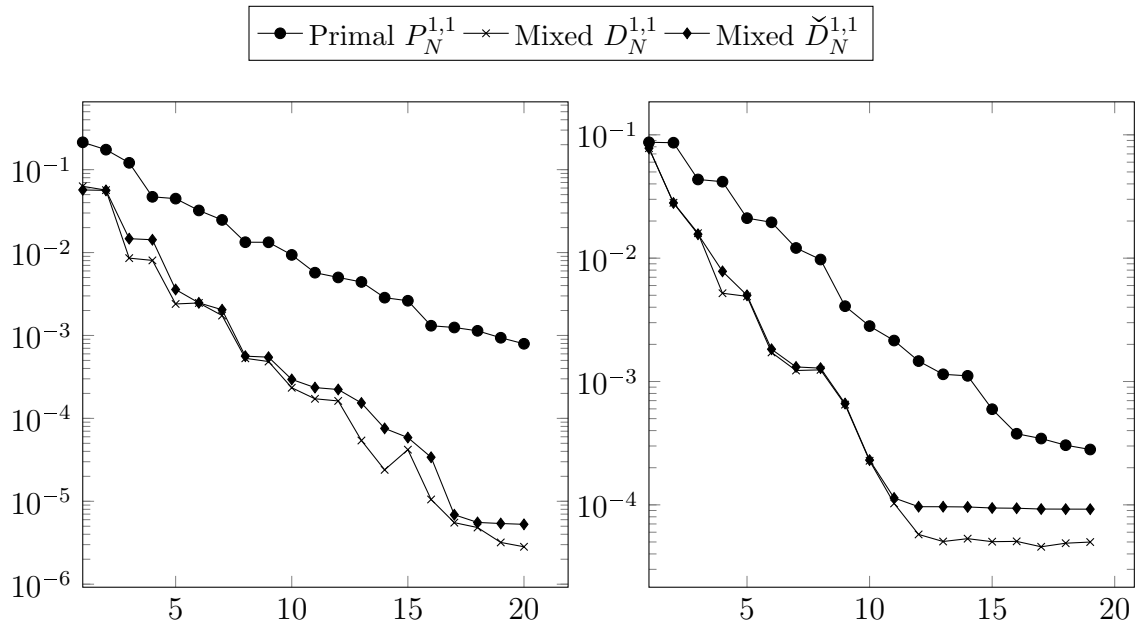
(a) Problem 1 (b) Problem 2

Figure 4.7 – Gradient of potentials errors $P_N^{1,0}$, $D_N^{1,0}$ and $\check{D}_N^{1,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 1.



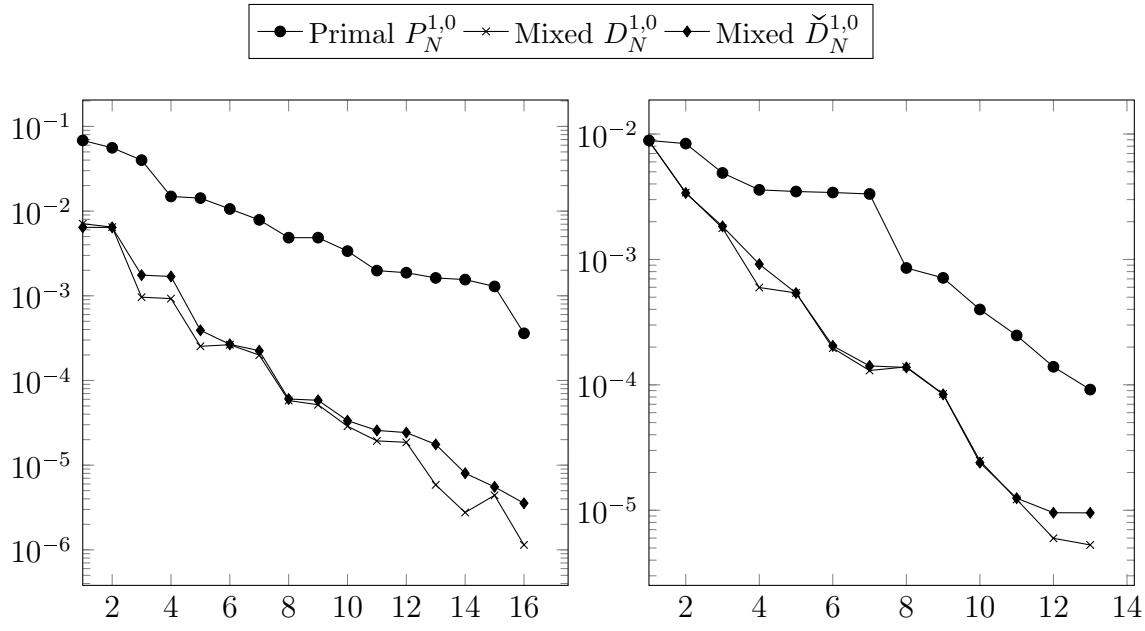
(a) Problem 1 (b) Problem 2

Figure 4.8 – Potentials errors $P_N^{0,0}$, $D_N^{0,0}$ and $\check{D}_N^{0,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 1.



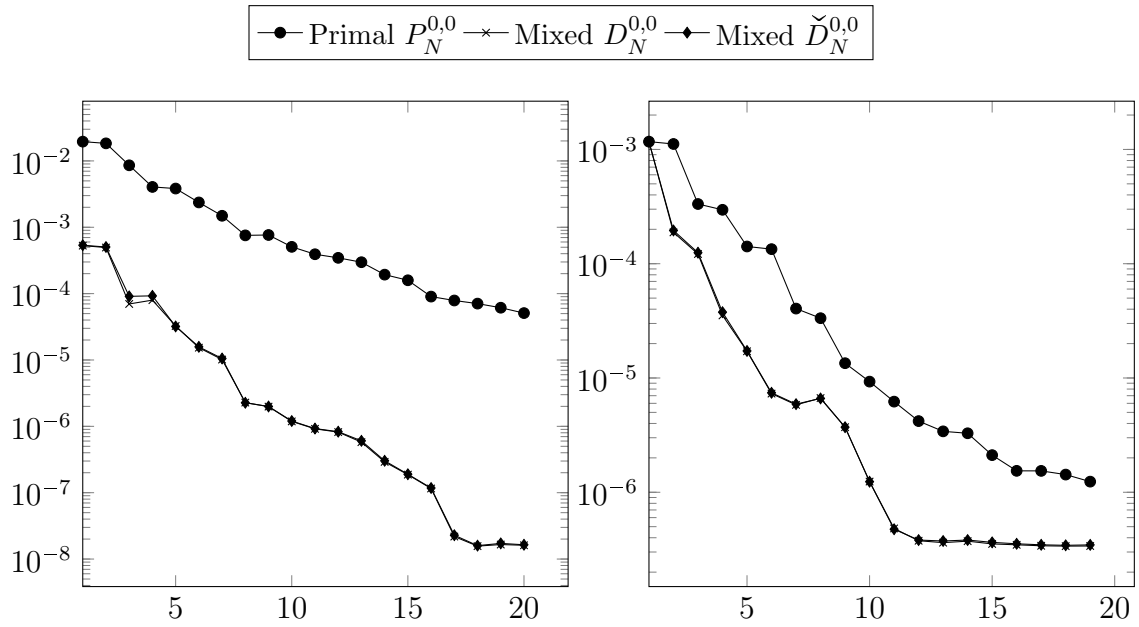
(a) Problem 1 (b) Problem 2

Figure 4.9 – Flux errors $P_N^{1,1}$, $D_N^{1,1}$ and $\breve{D}_N^{1,1}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 2.



(a) Problem 1 (b) Problem 2

Figure 4.10 – Gradient of potentials errors $P_N^{1,0}$, $D_N^{1,0}$ and $\breve{D}_N^{1,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 2.



(a) Problem 1 (b) Problem 2

Figure 4.11 – Potentials errors $P_N^{0,0}$, $D_N^{0,0}$ and $\check{D}_N^{0,0}$ for the two problems when the parameter sample used to generate a reduced basis is chosen with the Algorithm 2.

Appendix A

Implementation of the Mixed High-Order method

We discuss the practical implementation of the primal hybrid method (1.62) for the Poisson problem. The implementation of the method (2.14) for the Stokes equations follows similar principles and is not detailed here for the sake of brevity.

An essential point consists in selecting appropriate bases for the polynomial spaces on elements and faces. Particular care is required to make sure that the resulting local problems are well-conditioned, since the accuracy of the local computations may affect the overall quality of the approximation. For a given polynomial degree $l \in \{k, k+1\}$, one possibility leading to a hierarchical basis for $\mathbb{P}^l(T)$, $T \in \mathcal{T}_h$, is to choose the following family of monomial functions:

$$\Big\{\varphi_T = \prod_{i=1}^{d} \xi_{T,i}^{\alpha_i} \quad | \quad \xi_{T,i} := \tfrac{x_i - x_{T,i}}{h_T} \quad \forall 1 \leqslant i \leqslant d, \quad \underline{\alpha} \in \mathbb{N}^d, \quad \|\underline{\alpha}\|_{l^1} \leqslant l\Big\}, \tag{A.1}$$

where $\boldsymbol{x}_T$ denotes the barycenter of T. The idea is here (i) to express basis functions with respect to a reference frame local to one element, which ensures that the basis does not depend on the position of the element and (ii) to scale with respect to a local length scale. Choosing this length scale equal to h_T ensures that the basis functions take values in the interval $[-1, 1]$. For anisotropic elements, a better option would be to use the inertial frame of reference and, possibly, to perform orthonormalization, cf. [17]. Similarly, a hierarchical monomial basis can be defined for the spaces $\mathbb{P}^k(F)$, $F \in \mathcal{F}_h$, using the face barycenter $\boldsymbol{x}_F$ and the face diameter h_F.

Let, for a given polynomial degree $l \geqslant 0$ and a number of variables $n \geqslant 0$, $N_n^l := \dim(\mathbb{P}^l)$. For any element $T \in \mathcal{T}_h$, we assume for the sake of simplicity that a hierarchical basis $\mathcal{B}_T^{k+1} := \{\varphi_T^i\}_{0 \leqslant i < N_d^{k+1}}$ (not necessarily given by (A.1)) has been selected for $\mathbb{P}^{k+1}(T)$ so that φ_T^0 is the constant function on T and $(\varphi_T^i, \varphi_T^0)_T = 0$ for all $1 \leqslant i < N_d^{k+1}$. While this latter condition is not verified for general element shapes by the choice (A.1), one can obtain also in that case a well-posed local problem (1.27) for the computation of $\mathfrak{C}_T^k$ by removing φ_T^0, since the remaining functions vanish at $\boldsymbol{x}_T$. For more general choices, the zero-average condition can be enforced by a Lagrange multiplier constant over the element. Having assumed that $\mathcal{B}_T^{k+1}$ is hierarchical, a basis for $\mathbb{P}^k(T)$ is readily obtained by selecting the first N_d^k basis functions. Additionally, for any face $F \in \mathcal{F}_h$, we denote by $\mathcal{B}_F^k := \{\varphi_F^i\}_{0 \leqslant i < N_{d-1}^k}$ a basis for $\mathbb{P}^k(F)$ (not necessarily hierarchical in this case).

The definition of the discrete spaces (1.14) relies on a generalized notion of DOFs. Solving the primal hybrid problem (1.62) amounts to computing the coefficients $(u_T^i)_{0 \leqslant i < N_d^k}$ for all $T \in \mathcal{T}_h$ and $(\lambda_F^i)_{0 \leqslant i < N_{d-1}^k}$ for all $F \in \mathcal{F}_h$ of the following expansions for the local potential unknown $u_T \in \underline{U}_T^{p_T}$ and the local Lagrange multiplier $\lambda_F \in \Lambda_F^k$, respectively:

$$u_T = \sum_{0 \leqslant i < N_d^k} u_T^i \varphi_T^i, \qquad \lambda_F = \sum_{0 \leqslant i < N_{d-1}^k} \lambda_F^i \varphi_F^i. \tag{A.2}$$

For all $T \in \mathcal{T}_h$, we also introduce as intermediate unknowns the algebraic flux DOFs $(\sigma_T^i)_{1 \leqslant i < N_d^k}$ and $(\sigma_{TF}^i)_{0 \leqslant i < N_{d-1}^k}$, $F \in \mathcal{F}_T$, corresponding to the coefficients of the following expansions for the components of the local flux unknown $(\boldsymbol{\sigma}_T, (\sigma_{TF})_{F \in \mathcal{F}_T}) \in \boldsymbol{\Sigma}_T^k$:

$$\mathbb{T}_T^k \ni \boldsymbol{\sigma}_T = \sum_{1 \leqslant i < N_d^k} \sigma_T^i \boldsymbol{\nabla} \varphi_T^i \qquad \mathbb{F}_F^k \ni \sigma_{TF} = \sum_{0 \leqslant i < N_{d-1}^k} \sigma_{TF}^i \varphi_F^i \quad \forall F \in \mathcal{F}_T, \tag{A.3}$$

where we have used the fact that $(\boldsymbol{\nabla} \varphi_T^i)_{1 \leqslant i < N_d^k}$ is a basis for the DOF space $\mathbb{T}_T^k$ defined by (1.13) (the sum starts from 1 to accomodate the zero-average constraint in the definition of $\mathbb{T}_T^k$). Clearly, the total number of local flux DOFs in $\boldsymbol{\Sigma}_T^k$ (cf. (1.14)) is

$$N_{\Sigma,T}^k := (N_d^k - 1) + \mathfrak{N}_T N_{d-1}^k,$$

with $\mathfrak{N}_T$ defined in (1.6).

For a given element $T \in \mathcal{T}_h$, the discrete operators $D_T^k, \mathfrak{C}_T^k, \boldsymbol{\varsigma}_T^k$ act on and take values in finite dimensional spaces, hence they can be represented by matrices once the choice of the bases for the DOF spaces has been made. Their action on a vector of

DOFs then results from right matrix-vector multiplication. In what follows, we show how to carry out the computation of such matrices in detail and how to use them to infer the local contribution to the bilinear form A stemming from the element T.

A.1 Discrete divergence operator

The discrete divergence operator D_T^k acting on $\boldsymbol{\Sigma}_T^k$ with values in $\mathbb{P}^k(T)$ can be represented by the matrix D of size $N_d^k \times N_{\Sigma,T}^k$ with block-structure $\left[\mathsf{D}_T \,\vdots\, (\mathsf{D}_F)_{F\in\mathcal{F}_T}\right]$ induced by the geometric items to which flux DOFs in $\boldsymbol{\Sigma}_T^k$ are associated. According to the definition (1.21) of D_T^k, the matrix D can be computed as the solution of the following linear system of size N_d^k with $N_{\Sigma,T}^k$ right-hand sides:

$$\mathsf{M}_D\mathsf{D} = \mathsf{R}_D, \tag{A.4}$$

with block form

$$\overbrace{\phantom{\mathsf{M}_D}}^{N_d^k}\quad \overbrace{\phantom{\mathsf{D}_T\;\mathsf{D}_{F_1}\;\cdots\;\mathsf{D}_{F_{\mathfrak{N}_T}}}}^{N_d^k-1\;\;N_{d-1}^k}\quad \overbrace{\phantom{\mathsf{D}_{F_{\mathfrak{N}_T}}}}^{N_{d-1}^k}$$

$$N_d^k\left\{\begin{array}{|c|}\hline \mathsf{M}_D \\ \hline \end{array}\right. \quad \underbrace{\begin{array}{|c:c:c:c|}\hline \mathsf{D}_T & \mathsf{D}_{F_1} & \cdots & \mathsf{D}_{F_{\mathfrak{N}_T}} \\ \hline \end{array}}_{N_{\Sigma,T}^k} \;=\; \underbrace{\begin{array}{|c:c:c:c|}\hline \mathsf{R}_{D,T} & \mathsf{R}_{D,F_1} & \cdots & \mathsf{R}_{D,F_{\mathfrak{N}_T}} \\ \hline \end{array}}_{N_{\Sigma,T}^k}$$

where the system matrix is $\mathsf{M}_D := \left[(\varphi_T^i, \varphi_T^j)_T\right]_{0\leqslant i,j<N_d^k}$, while the right-hand side is such that

$$\mathsf{R}_{D,T} := \left[(\boldsymbol{\nabla}\varphi_T^i, \boldsymbol{\nabla}\varphi_T^j)_T\right]_{0\leqslant i<N_d^k, 1\leqslant j<N_d^k} \qquad \mathsf{R}_{D,F} := \left[(\varphi_T^i, \varphi_F^j)_F\right]_{0\leqslant i<N_d^k, 0\leqslant j<N_{d-1}^k} \quad \forall F\in\mathcal{F}_T.$$

When considering orthonormal bases such as, e.g., the ones introduced in [17], the matrix M_D is unit diagonal and numerical resolution is unnecessary.

A.2 Consistent flux reconstruction operator

The consistent flux reconstruction operator $\mathfrak{C}_T^k$ acting on $\boldsymbol{\Sigma}_T^k$ with values in $\boldsymbol{\nabla}\mathbb{P}^{k+1,0}(T)$ can be represented by the matrix C of size $(N_d^{k+1}-1)\times N_{\Sigma,T}^k$ with the block-structure $\left[\mathsf{C}_T \,\vdots\, (\mathsf{C}_F)_{F\in\mathcal{F}_T}\right]$ induced by the geometric items to which flux DOFs in $\boldsymbol{\Sigma}_T^k$ are associated. According to definition (1.27a), this requires to solve a linear system of

size $(N_d^{k+1}-1)$ with $N_{\Sigma,T}^k$ right-hand sides,

$$\mathsf{M}_C\mathsf{C} = \mathsf{Q}_C\mathsf{D} + \mathsf{R}_C := \widetilde{\mathsf{R}}_C. \tag{A.5}$$

The linear system (A.5) has the following block form:

$$\underbrace{\begin{array}{c}N_d^{k+1}-1\\ \{\\ \end{array}\overbrace{\boxed{\mathsf{M}_C}}^{N_d^{k+1}-1}}^{}\ \overbrace{\boxed{\begin{array}{c|c|c|c}\mathsf{C}_T & \mathsf{C}_{F_1} & \cdots & \mathsf{C}_{F_{\mathfrak{N}_T}}\end{array}}}^{N_{\Sigma,T}^k}\ =\ \boxed{\mathsf{Q}_C}\ \boxed{\mathsf{D}}\ +\ \overbrace{\boxed{\begin{array}{c|c|c|c}\mathsf{0} & \mathsf{R}_{C,F_1} & \cdots & \mathsf{R}_{C,F_{\mathfrak{N}_T}}\end{array}}}^{N_{\Sigma,T}^k}$$

$$\begin{array}{cccccccccccc} & N_d^k-1 & & N_{d-1}^k & & N_{d-1}^k & & N_d^k & N_{\Sigma,T}^k & N_d^k-1 & N_{d-1}^k & & N_{d-1}^k \end{array}$$

with system matrix $\mathsf{M}_C := \big[(\boldsymbol{\nabla}\varphi_T^i, \boldsymbol{\nabla}\varphi_T^j)\big]_{1\leqslant i,j<N_d^{k+1}}$ and the matrix blocks appearing in the right-hand side in addition to the matrix D obtained solving (A.4) are given by

$$\mathsf{Q}_C := \big[-(\varphi_T^i, \varphi_T^j)_T\big]_{1\leqslant i<N_d^{k+1}, 0\leqslant j<N_d^k}, \qquad \mathsf{R}_{C,F} := \big[(\varphi_T^i, \varphi_F^j)_F\big]_{1\leqslant i<N_d^{k+1}, 0\leqslant j<N_{d-1}^k} \quad \forall F\in\mathcal{F}_T.$$

A.3 Bilinear form H_T

We are now ready to compute the matrix H of size $N_{\Sigma,T}^k \times N_{\Sigma,T}^k$ representing the local bilinear form H_T defined by (1.30) as

$$\mathsf{H} = \mathsf{C}^t\widetilde{\mathsf{R}}_C + \mathsf{J}, \tag{A.6}$$

where the factors appearing in the first term are defined in (A.5), while the matrix J representing the stabilization term J_T defined by (1.34) is given by (the block partitioning is the one induced by the geometric entity to which flux DOFs are attached):

$$\mathsf{J} = \sum_{F\in\mathcal{F}_T} \mathsf{C}^t\mathsf{Q}_{J,1,F}\mathsf{C} - \Big[\,\mathsf{0}\,\Big|\,(\mathsf{C}^t\mathsf{Q}_{J,2,F})_{F\in\mathcal{F}_T}\Big] - \Big[\,\mathsf{0}\,\Big|\,(\mathsf{C}^t\mathsf{Q}_{J,2,F})_{F\in\mathcal{F}_T}\Big]^t + h_F\left[\begin{array}{c|c}\mathsf{0} & \mathsf{0}\\ \hline \mathsf{0} & \mathrm{diag}(\mathsf{M}_F)_{F\in\mathcal{F}_T}\end{array}\right],$$

where C is defined by (A.5) while, for all $F\in\mathcal{F}_T$, we have defined the auxiliary matrices

$$\mathsf{Q}_{J,1,F} := h_F\big[(\boldsymbol{\nabla}\varphi_T^i\cdot\boldsymbol{n}_{TF}, \boldsymbol{\nabla}\varphi_T^j\cdot\boldsymbol{n}_{TF})_F\big]_{1\leqslant i,j<N_d^{k+1}},$$
$$\mathsf{Q}_{J,2,F} := h_F\big[(\boldsymbol{\nabla}\varphi_T^i\cdot\boldsymbol{n}_{TF}, \varphi_F^j)_F\big]_{1\leqslant i<N_d^{k+1}, 0\leqslant j<N_{d-1}^k},$$

and face mass matrices

$$\mathsf{M}_F := \left[(\varphi_F^i, \varphi_F^j)_F\right]_{0\leqslant i,j<N_{d-1}^k}. \tag{A.7}$$

A.4 Hybridization

The first step to perform hybridization is to construct the matrix B representing the bilinear form B defined by (1.42a), which has the following block form corresponding to the geometric items to which DOFs in $\boldsymbol{\Sigma}_T^k$ (rows) and W_T^k (columns) are associated:

$$\mathsf{B} = \left[\begin{array}{c|c|c|c} \multicolumn{1}{c}{N_d^k} & \multicolumn{1}{c}{N_{d-1}^k} & & \multicolumn{1}{c}{N_{d-1}^k} \\ \hline & 0 & \cdots & 0 \\ & \mathsf{M}_{F_1} & & \\ \cline{2-4} \mathsf{R}_D^t & & \ddots & 0 \\ & 0 & & \\ \cline{4-4} & & & \mathsf{M}_{F_{\mathfrak{N}_T}} \end{array}\right] \begin{array}{l} \} N_d^k - 1 \\ \} \mathfrak{N}_T N_{d-1}^k \end{array} \} N_{\Sigma,T}^k$$

$$\underbrace{\qquad\qquad\qquad\qquad}_{N_{W,T}^k}$$

with matrix R_D as in (A.4), M_F defined by (A.7), and

$$N_{W,T}^k := N_d^k + \mathfrak{N}_T N_{d-1}^k,$$

corresponding to the number of DOFs in W_T^k.

The condition on the Lagrange multipliers in Λ_h^k on boundary faces $F \in \mathcal{F}_b$ (cf. (1.37)) is enforced via Lagrange multipliers in $\mathbb{P}^k(F)$. This choice is reflected by the fact that we include boundary faces in the definition of the matrix B.

The local contribution to the bilinear form A defined by (1.60) is finally given by

$$\mathsf{A} = \mathsf{B}^t \mathsf{H}^{-1} \mathsf{B}, \tag{A.8}$$

which requires the solution of a linear system involving the matrix H defined by (A.6). Observe that $\mathsf{H}^{-1}\mathsf{B}$ is in fact the matrix representation of the lifting operator $\boldsymbol{\varsigma}_T^k$ defined by (1.48a).

The matrix A has the following block structure induced by the geometric items to

which DOFs in W_T^k are attached:

$$\mathsf{A} = \begin{array}{c} \overbrace{\phantom{\mathsf{A}_{TT}\;\mathsf{A}_{TF}}}^{N_d^k \quad \mathfrak{N}_T N_{d-1}^k} \\ \left[\begin{array}{cc} \mathsf{A}_{TT} & \mathsf{A}_{TF} \\ \mathsf{A}_{TF}^t & \mathsf{A}_{FF} \end{array}\right] \end{array} \begin{array}{l} \} N_d^k \\ \} \mathfrak{N}_T N_{d-1}^k \end{array}$$

Observing that cell DOFs for a given element T are only linked to the face DOFs (Lagrange multipliers) attached to the faces in $\mathcal{F}_T$, one can finally obtain a problem in the sole Lagrange multipliers by computing the Schur complement of A_{TT}. This requires the numerical inversion of the symmetric positive-definite matrix A_{TT} of size $N_d^k \times N_d^k$.

Bibliography

[1] A. Abdulle and O. Budáč. A Petrov–Galerkin Reduced-Basis approximation of the Stokes equation in parameterized geometries. *Comptes Rendus Mathematique*, 353(7):641 – 645, 2015.

[2] J. Aghili, S. Boyaval, and D. A. Di Pietro. Hybridization of Mixed High-Order methods on general meshes and application to the Stokes equations. *Comput. Meth. Appl. Math.*, 15(2):111–134, 2015.

[3] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[4] C. Amrouche and V. Girault. On the existence and regularity of the solution of Stokes problem in arbitrary dimension. *Proc. Japan. Acad.*, 67:171–175, 1991.

[5] P. F. Antonietti, S. Giani, and P. Houston. *hp*-version composite discontinuous Galerkin methods for elliptic problems on complicated domains. *SIAM J. Sci. Comput.*, 35(3):A1417–A1439, 2013.

[6] T. Arbogast and Z. Chen. On the implementation of mixed methods as nonconforming methods for second-order elliptic problems. *Math. Comp.*, 64:943–972, 1995.

[7] D. N. Arnold. Mixed finite element methods for elliptic problems. *Comput. Methods Appl. Mech. Eng.*, 82(1-3):281–300, 1990.

[8] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: implementation, postprocessing and error estimates. *RAIRO Modél. Math. Anal. Num.*, 19(4):7–32, 1985.

[9] J.-P. Aubin. *Analyse fonctionnelle appliquée*. Presses Universitaires de France, Paris, 1987.

[10] B. Ayuso de Dios, K. Lipnikov, and G. Manzini. The nonconforming virtual element method. *M2AN Math. Model. Numer. Anal.*, 2015. Published online.

[11] I. Babuška and M. Suri. The *h-p* version of the finite element method with quasi-uniform meshes. *RAIRO Modél. Math. Anal. Numér.*, 21(2):199–238, 1987.

[12] I. Babuška and M. Suri. The optimal convergence rate of the *p*-version of the finite element method. *SIAM J. Numer. Anal.*, 24(4):750–776, 1987.

[13] I. Babuška, B. A. Szabo, and I. N. Katz. The *p*-version of the finite element method. *SIAM J. Numer. Anal.*, 18(3):515–545, 1981.

[14] I. Babuška. The finite element method with lagrangian multipliers. *Numer. Math.*, 20:179–192, 1973.

[15] S. Balay, J. Brown, K. Buschelman, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, B. F. Smith, and H. Zhang. PETSc Web page. `http://www.mcs.anl.gov/petsc`, 2011.

[16] M. Barrault, Y. Maday, N.C. Nguyen, and A.T. Patera. An "empirical interpolation" method: application to efficient Reduced-Basis discretization of partial differential equations. *Comptes Rendus Mathematique*, 339(9):667 – 672, 2004.

[17] F. Bassi, L. Botti, A. Colombo, D. A. Di Pietro, and P. Tesini. On the flexibility of agglomeration based physical space discontinuous Galerkin discretizations. *J. Comput. Phys.*, 231(1):45–65, 2012.

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

[19] L. Beirão da Veiga, F. Brezzi, A. Cangiani, G. Manzini, L. D. Marini, and A. Russo. Basic principles of virtual element methods. *M3AS Math. Models Methods Appl. Sci.*, 199(23):199–214, 2013.

[20] L. Beirão da Veiga, A. Chernov, L. Mascotto, and A. Russo. Basic principles of *hp* virtual elements on quasiuniform meshes. *Mathematical Models and Methods in Applied Sciences*, 26(08):1567–1598, 2016.

[21] L. Beirão da Veiga, V. Gyrya, K. Lipnikov, and G. Manzini. Mimetic finite difference method for the Stokes problem on polygonal meshes. *JCP*, 228(19):7215–7232, 2009.

[22] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for Greedy algorithms in Reduced–Basis methods. *SIAM J. Math Anal.*, June 2011.

[23] D. Boffi, M. Botti, and D. A. Di Pietro. A nonconforming high-order method for the Biot problem on general meshes. *SIAM J. Sci. Comput.*, 38(3):A1508–A1537, 2016.

[24] D. Boffi, F. Brezzi, and M. Fortin. *Mixed finite element methods and applications*, volume 44 of *Springer Series in Computational Mathematics*. Springer, Berlin Heidelberg, 2013.

[25] D. Boffi and D. A. Di Pietro. Unified formulation and analysis of mixed and primal discontinuous skeletal methods on polytopal meshes, September 2016. Submitted. Preprint arXiv:1609.04601 [math.NA].

[26] J. Bonelle. *Analysis of Compatible Discrete Operator Schemes on polyhedral meshes for elliptic and Stokes equations*. PhD thesis, 2014.

[27] J. Bonelle and A. Ern. Analysis of compatible discrete operator schemes for Stokes problems on polyhedral meshes. *IMA J. Numer. Anal.*, 34(4):553–581, 2014.

[28] S. Boyaval. Reduced-Basis approach for homogenization beyond the periodic setting. *SIAM Multiscale Modeling & Simulation*, 7(1):466–494, 2008.

[29] S. Boyaval, C. Le Bris, T. Lelièvre, Y. Maday, N.C. Nguyen, and A.T. Patera. Reduced-Basis techniques for stochastic problems. *Archives of Computational Methods in Engineering*, 17(4):435–454, 2010.

[30] S. Boyaval, C. Le Bris, Y. Maday, N.C. Nguyen, and A.T. Patera. A Reduced-Basis approach for variational problems with stochastic parameters: Application to heat conduction with variable Robin coefficient. *Computer Methods in Applied Mechanics and Engineering*, 198(41–44):3187–3206, June 2009.

[31] F. Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from lagrange multipliers. *RAIRO Ser. Rouge*, 8:129–151, 1874.

[32] F. Brezzi, M. Fortin, and D. Boffi. *Mixed and hybrid finite element methods*, volume 44 of *Springer Series in Computational Mathematics*. Springer Berlin Heidelberg, 2013.

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

[34] A. Buffa, Y. Maday, A.T. Patera, C. Prud'homme, and G. Turinici. A priori convergence of the Greedy algorithm for the parametrized Reduced-Basis method. *ESAIM: M2AN*, 46(3):595–603, 2012.

[35] E. Burman and B. Stamm. Bubble stabilized discontinuous Galerkin method for Stokes problem. *Math. Models Methods Appl. Sci.*, 20(2):297–313, 2010.

[36] A. Cangiani, E. H. Georgoulis, and P. Houston. *hp*-version discontinuous Galerkin methods on polygonal and polyhedral meshes. *Math. Models Methods Appl. Sci.*, 24(10):2009–2041, 2014.

[37] P. Castillo, B. Cockburn, D. Scötzau, and C. Schwab. Optimal a priori error estimates for the *hp*-version of the local discontinuous Galerkin method for convection-diffusion problems. *Math. Comp.*, 71(238):455–478, 2001.

[38] L. Cattabriga. Su un problema al contorno relativo al sistema di equazioni di Stokes. *Rend. Sem. Mat. Univ. Padova*, 31:308–340, 1961.

[39] A. Cesmelioglu, B. Cockburn, N. C. Nguyen, and J. Peraire. Analysis of HDG methods for Oseen equations. *Journal of Scientific Computing*, 55(2):392–431, 2013.

[40] R. Chakir and Y. Maday. Une méthode combinée d'éléments finis à deux grilles/bases réduites pour l'approximation des solutions d'une e.d.p. paramétrique. *Comptes Rendus Mathematique*, 347(7–8):435–440, 2009.

[41] F. Chave, D. A. Di Pietro, F. Marche, and F. Pigeonneau. A Hybrid High-Order method for the Cahn–Hilliard problem in mixed form. *SIAM J. Numer. Anal.*, 54(3):1873–1898, 2016.

[42] Z. Chen. Equivalence between and multigrid algorithms for nonconforming and mixed methods for second-order elliptic problems. *East-West J. Numer. Math.*, 4:1–33, 1996.

[43] P. G. Ciarlet. *The finite element method for elliptic problems*, volume 40 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2002. Reprint of the 1978 original [North-Holland, Amsterdam; MR0520174 (58 #25001)].

[44] B. Cockburn, D. A. Di Pietro, and A. Ern. Bridging the Hybrid High-Order and Hybridizable Discontinuous Galerkin methods. *ESAIM: Math. Model Numer. Anal. (M2AN)*, 50(3):635–650, 2016.

[45] B. Cockburn and J. Gopalakrishnan. The derivation of hybridizable discontinuous Galerkin methods for Stokes flow. *SIAM J. Numer. Anal.*, 47(2):1092–1125, 2009.

[46] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous Galerkin, mixed, and continuous Galerkin methods for second order elliptic problems. *SIAM J. Numer. Anal.*, 47(2):1319–1365, 2009.

[47] B. Cockburn, G. Kanschat, and D. Schötzau. The Local Discontinus Galerkin method for the Oseen equations. *Mathematics of Computation*, 73:569–593, 2003.

[48] B. Cockburn and K. Shi. Devising **HDG** methods for Stokes flow: An overview. *Comput. & Fluids*, 98:221–229, 2014.

[49] M. Crouzeix and P.-A. Raviart. Conforming and nonconforming finite element methods for solving the stationary Stokes equations. *RAIRO Modél. Math. Anal. Num.*, 7(3):33–75, 1973.

[50] B.M. De Veubeke. Displacement and equilibrium models in the finite element method. *Stress Analysis*, pages 145–197, 1977.

[51] J. W. Demmel, S. C. Eisenstat, J. R. Gilbert, X. S. Li, and J. W. H. Liu. A supernodal approach to sparse partial pivoting. *SIAM J. Matrix Analysis and Applications*, 20(3):720–755, 1999.

[52] D. A. Di Pietro. Analysis of a discontinuous Galerkin approximation of the Stokes problem based on an artificial compressibility flux. *Int. J. Num. Meth. Fluids*, 55(8):793–813, 2007.

[53] D. A. Di Pietro and J. Droniou. A Hybrid High-Order method for Leray–Lions elliptic equations on general meshes. *Math. Comp.*, 2016. Accepted for publication. Preprint arXiv 1508.01918 [math.NA].

[54] D. A. Di Pietro, J. Droniou, and A. Ern. A discontinuous-skeletal method for advection-diffusion-reaction on general meshes. *SIAM J. Numer. Anal.*, 53(5):2135–2157, 2015.

[55] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications*. Springer-Verlag, Berlin, 2012.

[56] D. A. Di Pietro and A. Ern. A Hybrid High-Order locking-free method for linear elasticity on general meshes. *Comput. Meth. Appl. Mech. Engrg.*, 283:1–21, 2015.

[57] D. A. Di Pietro and A. Ern. Arbitrary-order mixed methods for heterogeneous anisotropic diffusion on general meshes. *IMA J. Numer. Anal.*, 2016. Published online 10.1093/imanum/drw003.

[58] D. A. Di Pietro, A. Ern, and J.-L. Guermond. Discontinuous Galerkin methods for anisotropic semi-definite diffusion with advection. *SIAM J. Numer. Anal.*, 46(2):805–831, 2008.

[59] D. A. Di Pietro, A. Ern, and S. Lemaire. An arbitrary-order and compact-stencil discretization of diffusion on general meshes based on local reconstruction operators. *Comput. Meth. Appl. Math.*, 14(4):461–472, 2014.

[60] D. A. Di Pietro, A. Ern, and S. Lemaire. *A review of Hybrid High-Order methods: formulations computational aspects, comparison with other methods.*

Building bridges: Connections and challenges in modern approaches to numerical partial differential equations, G. Barrenechea, F. Brezzi, A. Cangiani, M. Georgoulis eds. Springer, 2016.

[61] D. A. Di Pietro, A. Ern, A. Linke, and F. Schieweck. A discontinuous skeletal method for the viscosity-dependent Stokes problem. *Comput. Meth. Appl. Mech. Engrg.*, 306:175–195, 2016.

[62] D. A. Di Pietro and S. Lemaire. An extension of the Crouzeix–Raviart space to general meshes with application to quasi-incompressible linear elasticity and Stokes flow. *Math. Comp.*, 84(291):1–31, 2015.

[63] D. A. Di Pietro and R. Specogna. An a posteriori-driven adaptive Mixed High-Order method with application to electrostatics. *J. Comput. Phys.*, 326(1):35–55, 2016.

[64] J. Douglas and J. E. Roberts. Mixed finite element methods for second order elliptic problems. *Math. Appl. Comp.*, 1:91–103, 1982.

[65] J. Droniou and R. Eymard. A mixed finite volume scheme for anisotropic diffusion problems on any grid. *Numer. Math.*, 105:35–71, 2006.

[66] J. Droniou, R. Eymard, T. Gallouët, and R. Herbin. A unified approach to mimetic finite difference, hybrid finite volume and mixed finite volume methods. *M3AS Mathematical Models and Methods in Applied Sciences*, 20(2):1–31, 2010.

[67] T. Dupont and R. Scott. Polynomial approximation of functions in Sobolev spaces. *Math. Comp.*, 34(150):441–463, 1980.

[68] H. Egger and C. Waluga. A Hybrid Discontinuous Galerkin method for Darcy–Stokes problems. *Domain Decomposition Methods in Science and Engineering XX*, 2009.

[69] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[70] R. Eymard, T. Gallouët, and R. Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes. SUSHI: a scheme using stabilization and hybrid interfaces. *IMA J. Numer. Anal.*, 30(4):1009–1043, 2010.

[71] P. F. Antonietti, L. Beirão da Veiga, D. Mora, and M. Verani. A stream virtual element formulation of the Stokes problem on polygonal meshes. *SIAM J. Numer. Anal.*, 52:386–404, 2014.

[72] R. S. Falk and J. E. Osborn. Error estimates for mixed methods. *RAIRO Anal. numer.*, 14:309–324, 1980.

[73] E. H. Georgoulis and E. Süli. Optimal error estimates for the *hp*-version interior penalty discontinuous Galerkin finite element method. *IMA J. Numer. Anal.*, 25:205–220, 2005.

[74] S. Giani and P. Houston. *hp*-adaptive composite discontinuous Galerkin methods for elliptic problems on complicated domains. *Numer. Methods Partial Differential Equations*, 30(4):1342–1367, 2014.

[75] V. Girault and P.-A. Raviart. *Finite element methods for Navier-Stokes equations*, volume 5 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1986. Theory and algorithms.

[76] V. Girault, B. Rivière, and M. F. Wheeler. A discontinuous Galerkin method with nonoverlapping domain decomposition for the Stokes and Navier-Stokes problems. *Math. Comp.*, 74(249):53–84 (electronic), 2004.

[77] M. A. Grepl, Y. Maday, N. C. Nguyen, and A. T. Patera. Efficient Reduced-Basis treatment of nonaffine and nonlinear partial differential equations. *M2AN (Math. Model. Numer. Anal.)*, 41(2):575–605, 2007. (doi: 10.1051/m2an:2007031).

[78] B. Haasdonk and M. Ohlberger. Reduced Basis method for finite volume approximations of parametrized linear evolution equations. *Mathematical Modelling and Numerical Analysis (M2AN)*, 42(3):277–302, 2008. (doi: 10.1051/m2an:2008001).

[79] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[80] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. In R. Eymard and J.-M. Hérard, editors, *Finite Volumes for Complex Applications V*, pages 659–692. John Wiley & Sons, 2008.

[81] J. Hesthaven, G. Rozza, and B. Stamm. *Certified Reduced-Basis Methods for Parametrized Differential Equations*. SpringerBriefs in Mathematics. Springer International Publishing, 2016.

[82] Y. Jeon, E.-J. Park, and D. Sheen. A hybridized finite element method for the Stokes problem. *Computers and Mathematics with Applications*, 68:2222–2232, 2014.

[83] W. Joerg and M. Koch. *BOOST uBLAS C++ Library*. `http://www.boost.org`.

[84] C. Le Potier. A finite volume method for the approximation of highly anisotropic diffusion operators on unstructured meshes. In *Finite volumes for complex applications IV*, pages 401–412. ISTE, London, 2005.

[85] K. Lipnikov and G. Manzini. A high-order mimetic method on unstructured polyhedral meshes for the diffusion equation. *J. Comput. Phys.*, 272:360–385, 2014.

[86] X. Liu, J. Li, and Z. Chen. A weak Galerkin finite element method for the Oseen equations. *Advances in Computational Mathematics*, pages 1–18, 2016.

[87] Y. Maday, A. T. Patera, and G. Turinici. A priori convergence theory for Reduced-Basis approximations of single-parameter elliptic partial differential equations. *Journal of Scientific Computing*, 17(1):437–446, 2002.

[88] Y. Maday and E. M. Rønquist. A Reduced-Basis element method. *Journal of Scientific Computing*, 17(1):447–459, 2002.

[89] L. D. Marini. An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method. *SIAM J. Numer. Anal.*, 22(3):493–496, 1985.

[90] A. Montlaur, S. Fernandez-Mendez, and A. Huerta. Discontinuous Galerkin methods for the Stokes equations using divergence-free approximations. *Int. J. Numer. Meth. Fluids*, 57:1071–1092, 2008.

[91] I. Perugia and D. Schötzau. A *hp*-analysis of the Local Discontinuous Galerkin method for diffusion problems. *J. Sci. Comput.*, 17(1–4):561–571, 2002.

[92] C. Prud'homme, D. Rovas, K. Veroy, Y. Maday, A.T. Patera, and G. Turinici. Reliable real-time solution of parametrized partial differential equations: Reduced-Basis output bounds methods. *Journal of Fluids Engineering*, 124(1):70–80, 2002.

[93] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced-Basis Methods for Partial Differential Equations*, volume 92 of *UNITEXT*. Springer International Publishing, 2016.

[94] P. A. Raviart and J. M. Thomas. Primal hybrid finite element methods for 2nd order elliptic equations. *Mathematics of Computation*, 31(138):pp. 391–413, 1977.

[95] B. Rivière, Wheeler M. F., and V. Girault. Improved energy estimates for interior penalty, constrained and discontinuous Galerkin methods for elliptic problems. Part I. *Comput. Geosci.*, 3:337–360, 1999.

[96] D. Ronald, P. Guergana, and W. Przemyslaw. Greedy algorithms for Reduced-Bases in Banach spaces. *Constructive Approximation*, 37(3):455–466, 2013.

[97] G. Rozza, D.B.P. Huynh, and A.T. Patera. Reduced-Basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations — application to transport and continuum mechanics. *Arch. Comput. Methods Eng.*, 15(3):229–275, 2008.

[98] G. Rozza and K. Veroy. On the stability of the reduced basis method for Stokes equations in parametrized domains. *Computer Methods in Applied Mechanics and Engineering*, 196(7):1244 – 1260, 2007.

[99] C. Schwab. *p- and hp−FEM – Theory and application to solid and fluid mechanics*. Oxford University Press. Oxford, 1998.

[100] C. Schwab. *hp-FEM for Fluid Flow Simulation*, pages 325–438. Springer Berlin Heidelberg, Berlin, Heidelberg, 1999.

[101] B. Stamm and T. P. Wihler. *hp*-optimal discontinuous Galerkin methods for linear elliptic problems. *Math. Comp.*, 79(272):2117–2133, 2010.

[102] E. M. Stein. *Singular integrals and differentiability properties of functions*. Princeton Mathematical Series, No. 30. Princeton University Press, Princeton, N.J., 1970.

[103] G.G. Stokes. On the effect of the internal friction of fluids on the motion of pendulums. *Cambridge Philos. Trans.*, 9:8–106, 1851.

[104] C. Talischi, G. H. Paulino, A. Pereira, and I. F. M. Menezes. PolyMesher: a general-purpose mesh generator for polygonal elements written in Matlab. *Structural and Multidisciplinary Optimization*, 45(3):309–328, 2012.

[105] A. Toselli. *hp*-finite element discontinuous Galerkin approximations for the Stokes problem. *M3AS*, 12(11):1565–1616, 2002.

[106] M. Vohralík. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45(4):1570–1599, 2007.

[107] M. Vohralík and B. Wohlmuth. Mixed finite element methods: implementation with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods. *M3AS Math. Models Methods Appl. Sci.*, 23(5):803–838, 2013.

[108] B. Wang and B.C. Khoo. Hybridizable discontinuous Galerkin method (HDG) for Stokes interface flow. *J. Comput. Phys.*, 247:262–278, 2013.

Résumé

Cette thèse aborde différents aspects de la résolution numérique des Équations aux Dérivées Partielles.

Le premier chapitre est consacré à l'étude de la méthode Mixed High-Order (MHO). Il s'agit d'une méthode mixte de dernière génération permettant d'obtenir des approximations d'ordre arbitraire sur maillages généraux. Le principal résultat obtenu est l'équivalence entre la méthode MHO et une méthode primale de type Hybrid High-Order (HHO).

Dans le deuxième chapitre, nous appliquons la méthode MHO/HHO à des problèmes issus de la mécanique des fluides. Nous considérons d'abord le problème de Stokes, pour lequel nous obtenons une discrétisation d'ordre arbitraire inf-sup stable sur maillages généraux. Des estimations d'erreur optimales en normes d'énergie et L^2 sont proposées. Ensuite, nous étudions l'extension au problème d'Oseen, pour lequel on propose une estimation d'erreur en norme d'énergie où on trace explicitement la dépendance du nombre de Péclet local.

Dans le troisième chapitre, nous analysons la version *hp* de la méthode HHO pour le problème de Darcy. Le schéma proposé permet de traiter des maillages généraux ainsi que de faire varier le degré polynomial d'un élément à l'autre. La dépendance de l'anisotropie locale du coefficient de diffusion est tracée explicitement dans l'analyse d'erreur en normes d'énergie et L^2.

La thèse se clôture par une ouverture sur la réduction de problèmes de diffusion à coefficients variables. L'objectif consiste à comprendre l'impact du choix de la formulation (mixte ou primale) utilisée pour la projection sur l'espace réduit sur la qualité du modèle réduit.

Mots-clés : Méthodes Mixed High-Order, méthodes Hybrid High-Order, maillages généraux, analyse *hp*, problème d'Oseen, problème de Stokes, problème de Darcy, réduction de modèle, bases réduites.

Abstract

This Ph.D. thesis deals with different aspects of the numerical resolution of Partial Differential Equations.

The first chapter focuses on the Mixed High-Order method (MHO). It is a last generation mixed scheme capable of arbitrary order approximations on general meshes. The main result of this chapter is the equivalence between the MHO method and a Hybrid High-Order (HHO) primal method.

In the second chapter, we apply the MHO/HHO method to problems in fluid mechanics. We first address the Stokes problem, for which a novel inf-sup stable, arbitrary-order discretization on general meshes is obtained. Optimal error estimates in both energy- and L^2-norms are proved. Next, an extension to the Oseen problem is considered, for which we prove an error estimate in the energy norm where the dependence on the local Peclet number is explicitly tracked.

In the third chapter, we analyse a *hp* version of the HHO method applied to the Darcy problem. The resulting scheme enables the use of general meshes, as well as varying polynomial orders on each face. The dependence with respect to the local anisotropy of the diffusion coefficient is explicitly tracked in both the energy- and L^2-norms error estimates.

In the fourth and last chapter, we address a perspective topic linked to model order reduction of diffusion problems with a parametric dependence. Our goal is in this case to understand the impact of the choice of the variational formulation (primal or mixed) used for the projection on the reduced space on the quality of the reduced model.

Keywords : Mixed High-Order methods, Hybrid High-Order methods, general meshes, *hp* analysis, Oseen problem, Stokes problem, moder reduction, reduced basis method.