Multiscale finite element methods for advection-diffusion problems

François Madiot

UNIVERSITÉ PARIS-EST

École Doctorale: Mathématiques et Sciences et Technologies de l'Information et de la Communication

THÈSE DE DOCTORAT

Spécialité : Mathématiques appliquées

Présentée par

François Madiot

Méthodes éléments finis de type MsFEM pour des problèmes d'advection diffusion

Thèse dirigée par Claude Le Bris et Frédéric Legoll au CERMICS, École des Ponts ParisTech

Soutenue le 8 décembre 2016 devant un jury composé de :

M. Jérôme DRONIOU	Monash University	Examinateur
M. Yalchin EFENDIEV	Texas A& M University	Examinateur
M. Ulrich HETMANIUK	University of Washington	Rapporteur
M. Claude LE BRIS	École des Ponts ParisTech	Directeur de thèse
M. Frédéric LEGOLL	École des Ponts ParisTech	Directeur de thèse
M. Alexei LOZINSKI	Université de Franche-Comté	Rapporteur
Mme. Anne NICOLAS	Commissariat à l'énergie atomique et aux énergies alternatives	Examinateur

Cette thèse est dédiée à mes parents.

Acknowledgements

Bien évidemment, c'est à mes deux directeurs de thèse, Claude Le Bris et Frédéric Legoll, que vont mes premiers remerciements. Je tiens à souligner leur rigueur scientifique, leurs compétences ainsi que leur exigence. Ils ont su rendre ces années enrichissantes. Leurs larges connaissances scientifiques et leur soutien ont été déterminants pour le bon déroulement de ma thèse.

Je tiens à remercier Yalchin Efendiev d'avoir accepté de présider mon jury de thèse, Ulrich Hetmaniuk et Alexei Lozinski pour l'intérêt qu'ils ont porté à mon travail en acceptant d'être les rapporteurs de ma thèse, Jérôme Droniou et Anne Nicolas d'avoir accepté de faire partie de mon jury.

Je remercie et salue l'ensemble des personnes que j'ai côtoyées pendant ces trois années, les doctorants, post-doctorants et stagiaires. Mes remerciements vont également à tous les membres du CERMICS. Un grand merci à Isabelle Simunic pour sa sympathie, son soutien et son travail efficace et remarquable.

J'aimerais remercier Alain Combrouze, Frédéric Massias et Blandine Dumoulin. Ces professeurs ont beaucoup compté pour moi durant ma formation scolaire.

Je souhaite également remercier l'ensemble des membres de ma famille pour leur confiance et leur soutien. Ils sont la raison principale de cet accomplissement et je ne peux exprimer en quelques lignes la gratitude que j'ai pour eux.

Mes derniers remerciements vont enfin à l'ensemble de mes amis de Toulouse, Paris ou d'ailleurs. Leur amitié m'a beaucoup apporté.

Titre: Méthodes éléments finis de type MsFEM pour des problèmes d'advection diffusion

Résumé: Ce travail a porté principalement sur le développement et l'étude de méthodes numériques de type éléments finis multi-échelles pour un problème d'advection diffusion multi-échelles dominé par l'advection. Deux types d'approches sont envisagées: prendre en compte l'advection dans la construction de l'espace d'approximation, ou appliquer une méthode de stabilisation. On commence par l'étude d'un problème d'advection diffusion, dominé par l'advection, dans un milieu hétérogène. On poursuit par l'étude de problèmes d'advection-diffusion, dans le régime où l'advection domine, posés dans un domaine perforé. On se focalise ici sur la condition aux bords de type Crouzeix Raviart pour la construction des éléments finis multi-échelles. On considère deux situations différentes selon la condition prescrite au bord des perforations: la condition de Dirichlet homogène ou la condition de Neumann homogène.

Dans une autre partie de ce travail, cette fois-ci sur un problème mono-échelle, on se place dans un cadre général où l'opérateur d'advection-diffusion est non coercif, possiblement dominé par l'advection. On propose une approche éléments finis basée sur une mesure invariante associée à l'opérateur adjoint. Cette approche est bien posée inconditionnellement en la taille du maillage. On la compare numériquement à une méthode standard de stabilisation.

Mots clés: Homogénéisation, Éléments finis multi-échelles, Advection dominante, Stabilisation, Mesure invariante, Non coercif, Domaine perforé

Title: Multiscale finite element methods for advection-diffusion problems

Abstract: This work essentially addresses the development and the study of multiscale finite element methods for multiscale advection-diffusion problems in the advection-dominated regime. Two types of approaches are investigated: Take into account the advection in the construction of the approximation space, or apply a stabilization method. We begin with advection-dominated advection-diffusion problems in heterogeneous media. We carry on with advection-dominated advection-diffusion problems posed in perforated domains. Here, we focus on the Crouzeix-Raviart type boundary condition for the construction of the multiscale finite elements. We consider two different choices for the condition prescribed on the boundary of the perforations: the homogeneous Dirichlet condition or the homogeneous Neumann condition.

In an other part of this work, we consider in a single-scale setting, a general framework where the advection-diffusion operator is not coercive, possibly in the advection-dominated regime. We propose a Finite Element approach based on the use of an invariant measure associated to the adjoint operator. This approach is unconditionally well-posed in the mesh size. We compare it numerically to a standard stabilization method.

Keywords: Homogenization, Multiscale Finite Elements, Advection-dominated problem, Stabilization, Invariant measure, Non-coercive problem, Perforated domain

Contents

Publications associées à ce travail

[1] C. Le Bris, F. Legoll, and F. Madiot. A numerical comparison of some Multiscale Finite Element approaches for advection-dominated problems in heterogeneous media. *Accepted in ESAIM Math. Model. Numer. Anal., arXiv preprint arXiv:1511.08453, HAL preprint hal-01235642*, 2016.

[2] C. Le Bris, F. Legoll, and F. Madiot. Stabilisation de problèmes non coercifs via une méthode numérique utilisant la mesure invariante (Stabilization of non-coercive problems using the invariant measure). *C. R. Math. Acad. Sci. Paris*, 354(8):799–803, 2016.

[3] C. Le Bris, F. Legoll, and F. Madiot. Stable approximation of the advection-diffusion equation using the invariant measure. *Submitted, arXiv preprint arXiv:1609.04777, HAL preprint hal-01367417*, 2016.

Chapter 1

Introduction

Ce travail de thèse a porté principalement sur l'étude de méthodes numériques de type éléments finis multi-échelles pour des problèmes d'advection-diffusion multi-échelles dominés par l'advection. Le caractère multi-échelles et la domination de l'advection représentent deux difficultés numériques distinctes. La littérature regorge de méthodes visant à résoudre chacune de ces deux difficultés séparément. La question est de comprendre comment adapter ces différentes méthodes pour un problème réunissant les deux difficultés.

Les principales contributions de ce travail sont les suivantes:

- La première analyse porte sur un problème d'advection-diffusion dominé par l'advection dans un milieu hétérogène de la forme

$$-\operatorname{div}\left(a_\varepsilon \nabla u^\varepsilon\right) + b \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega. \tag{1.1}$$

  On propose plusieurs approches éléments finis multi-échelles et une méthode de splitting permettant de coupler deux codes existants déjà optimisés pour chacun des sous-problèmes: un problème de diffusion multi-échelles et un problème d'advection-diffusion mono-échelle dominé par l'advection. Mathématiquement, on étudie l'erreur numérique des approches éléments finis multi-échelles dans un contexte monodimensionnel. La convergence de la méthode de splitting est analysée sans restriction sur la dimension de l'espace. Numériquement, on compare les approches de type éléments finis multi-échelles et la méthode de splitting en terme de précision et de coût de calcul.

- Dans un deuxième temps, on poursuit l'analyse dans le cas d'un problème d'advection-diffusion dominé par l'advection, posé dans un milieu perforé. Il s'agit d'un problème du type

$$-\alpha\Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \tag{1.2}$$

  où $\Omega^\varepsilon \subsetneq \Omega$ est un domaine perforé par de petits trous de taille $\varepsilon > 0$, situés à une distance d'ordre ε les uns des autres. On examine deux situations sensiblement différentes selon le type de condition prescrite aux bords des perforations. Dans la première, on impose la condition de Dirichlet homogène

$$-\alpha\Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega^\varepsilon, \tag{1.3}$$

et dans la seconde, la condition de Neumann homogène

$$\begin{cases} -\alpha\Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \\ \alpha\nabla u^\varepsilon \cdot n = 0 \quad \text{sur } \partial\Omega^\varepsilon \backslash \partial\Omega, \\ u^\varepsilon = 0 \quad \text{sur } \partial\Omega^\varepsilon \cap \partial\Omega. \end{cases} \tag{1.4}$$

Mathématiquement, on obtient des estimées d'erreur dans le cas où l'on impose la condition de Dirichlet homogène aux bords des perforations. Numériquement, on étudie les performances des différentes approches éléments finis multi-échelles vis-à-vis de la domination de l'advection et du caractère multi-échelles.

- Un troisième travail propose une approche numérique éléments finis, à notre connaissance nouvelle, basée sur la notion de mesure invariante pour les problèmes d'advection-diffusion mono-échelle non coercifs, éventuellement dominés par l'advection, de la forme

$$-\Delta u + b \cdot \nabla u = f \quad \text{dans } \Omega. \tag{1.5}$$

Mathématiquement, on montre le caractère bien posé de notre approche et on en étudie la convergence. Numériquement, on compare notre approche avec une méthode standard de stabilisation.

1.1 Contexte et motivations

Les problèmes d'advection-diffusion dans lesquels l'advection domine interviennent dans différents phénomènes physiques, tel que le transport d'un polluant dans un fluide. Des instabilités numériques apparaissent lorsqu'on utilise une méthode d'éléments finis classique. En pratique, il est nécessaire pour traiter ce régime d'utiliser des méthodes numériques adaptées. Ces méthodes, dites méthodes de stabilisation, sont bien comprises aujourd'hui.

Dans diverses applications, le problème considéré est, de plus, multi-échelle, ce qui peut rendre le problème plus difficile dans certains régimes. La pollution des eaux souterraines par l'infiltration d'un fluide à travers un milieu poreux, la propagation du vent dans une ville et l'extraction pétrolière en sont des exemples.

Les modèles étudiés se trouvent à l'intersection entre des modèles d'advection-diffusion dans un milieu homogène (dominés par l'advection, comme décrit ci-dessus) et des modèles de diffusion dans un milieu hétérogène, sans phénomène de transport. Pour ces derniers, on utilise souvent des méthodes multi-échelles de type éléments finis multi-échelles [36], où les fonctions de base des éléments finis à l'échelle macroscopique sont obtenues via un calcul à l'échelle microscopique, prenant en compte les hétérogénéités présentes dans le problème.

Dans les exemples cités précédemment, le domaine dans lequel évolue le flot contient des perforations (représentant le milieu poreux, les bâtiments, ...) qui sont à une petite distance les unes des autres. Deux petites échelles sont donc simultanément présentes dans le problème: la petite échelle de la structure, et la petite épaisseur des couches limites.

L'objectif est de comprendre comment adapter les méthodes éléments finis multi-échelles pour résoudre efficacement des problèmes multi-échelles dominés par l'advection. Le chapitre 2 concerne l'étude d'un problème d'advection-diffusion, dominé par l'advection, dans un milieu hétérogène. On présente dans l'annexe A des résultats d'homogénéisation dans le cadre périodique

pour ce problème. Le chapitre 4 est consacré à des problèmes d'advection-diffusion, sous le régime où l'advection domine, posés dans un domaine perforé.

Dans une autre partie de ce travail, on se place dans un cadre général où l'opérateur d'advection-diffusion est mono-échelle, non-coercif et le problème est dominé par l'advection. À notre connaissance, l'analyse des méthodes usuelles de stabilisation (de type *Streamline Upwind Petrov Galerkin*, SUPG) a été faite dans le cas coercif uniquement [81, Chapter III Section 3.2, p.229-255]. Plusieurs stratégies numériques ont été proposées pour ce type de problèmes (cf par exemple [32, 33]), mais aucune ne s'impose à l'heure actuelle. Dans le chapitre 3, on propose une approche éléments finis basée sur une mesure invariante associée à l'opérateur adjoint. Le contenu de ce chapitre est résumé dans l'annexe B sous la forme d'une note qui a été publiée dans Comptes Rendus Mathématique.

1.2 Introduction mathématique

Ce travail de thèse s'appuie sur des méthodes de stabilisation et la méthode des éléments finis multi-échelles. On les présente successivement dans les sections 1.2.1 et 1.2.2.

1.2.1 Méthodes de stabilisation

Soit $\Omega \subset \mathbb{R}^d$, un ouvert borné régulier. On s'intéresse pour simplifier au problème suivant

$$-\alpha\Delta u + b\cdot\nabla u = f \quad \text{dans } \Omega, \quad u = 0 \quad \text{sur } \partial\Omega, \tag{1.6}$$

avec $f \in L^2(\Omega)$, $\alpha > 0$ et $b \in L^\infty(\Omega)$ tel que

$$\operatorname{div} b = 0 \quad \text{dans } \Omega. \tag{1.7}$$

La formulation variationnelle associée à (1.6) s'écrit

$$\text{Trouver } u \in V \text{ tel que pour tout } v \in V, \quad a(u,v) = F(v),$$

avec

$$a(u,v) = \int_\Omega \alpha\nabla u\cdot\nabla v + (b\cdot\nabla u)v, \tag{1.8}$$

$$F(v) = \int_\Omega fv, \tag{1.9}$$

et $V = H_0^1(\Omega)$. Soit $\mathcal{T}_H$ un maillage régulier uniforme de taille H et V_H un espace d'approximation associé à ce maillage. On rappelle que l'approximation de Galerkin standard consiste à résoudre la formulation variationnelle

$$\text{Trouver } u_H \in V_H \text{ tel que pour tout } v_H \in V_H, \quad a(u_H, v_H) = F(v_H). \tag{1.10}$$

Soit $u \in H_0^1(\Omega)$ la solution du problème (1.6) et u_H la solution de (1.10) où V_H est la discrétisation standard en éléments finis $\mathbb{P}^1$ (par exemple). D'après [40, Theorem 8.12, p.186], on sait que

$u \in H^2(\Omega)$. On a alors l'estimée d'erreur a priori suivante:

$$|u - u_H|_{H^1(\Omega)} \leqslant CH(1 + \mathrm{Pe}\, H)|u|_{H^2(\Omega)}. \tag{1.11}$$

où $|u - u_H|_{H^1(\Omega)} = \|\nabla(u - u_H)\|_{L^2(\Omega)}$, $|u|^2_{H^2(\Omega)} = \sum_{i,j=1}^{d} \|\partial_{ij} u\|^2_{L^2(\Omega)}$, et $\mathrm{Pe} = \dfrac{\|b\|_{L^\infty(\Omega)}}{2\alpha}$. On voit que plus le produit $\mathrm{Pe}\, H$ est grand, plus l'erreur numérique peut être grande. On est dans la situation de perte de coercivité [38, Section 3.5.2] quand Pe tend vers $+\infty$. Intuitivement, le problème devient de moins en moins coercif lorsque la domination de l'advection sur la diffusion augmente. Ceci a un impact direct sur les résultats numériques: des oscillations parasites dans tout le domaine apparaissent. C'est pourquoi dans le régime où l'advection domine, c'est à dire lorsque $\mathrm{Pe}\, H > 1$, la méthode $\mathbb{P}^1$ standard aboutit à une approximation u_H de faible précision.

Le traitement numérique du régime où l'advection domine est bien connu pour les problèmes d'advection-diffusion mono-échelles. La littérature regorge de méthodes de stabilisation conçues pour ce type de problème: méthode de diffusion artificielle [50], la classe de méthodes multi-échelles variationnelles introduite dans [49]... On renvoie à [53, 78, 80, 81, 86] pour une revue des méthodes de stabilisation. Cette section est dédiée à la présentation d'une classe de méthodes de stabilisation fortement consistante.

La formulation variationnelle des méthodes dans cette classe est de la forme [78, Section 8.3.2, p.269-272]

$$\begin{aligned} &\text{Trouver } u_H \in V_H \text{ tel que pour tout } v_H \in V_H, \\ &a(u_H, v_H) + a_{\mathrm{stab}}(u_H, v_H) = F(v_H) + F_{\mathrm{stab}}(v_H), \end{aligned} \tag{1.12}$$

avec a et F définis par (1.8) et (1.9), et

$$a_{\mathrm{stab}}(u_H, v_H) = \sum_{K \in \mathcal{T}_H} (\tau_K \mathcal{L} u_H, (\mathcal{L}_{ss} + \rho \mathcal{L}_s) v_H)_K, \tag{1.13}$$

$$F_{\mathrm{stab}}(v_H) = \sum_{K \in \mathcal{T}_H} (\tau_K f, (\mathcal{L}_{ss} + \rho \mathcal{L}_s) v_H)_K, \tag{1.14}$$

où $\mathcal{L}_s u = -\alpha \Delta u$ et $\mathcal{L}_{ss} = b \cdot \nabla u$ sont respectivement la partie symétrique et antisymétrique de l'opérateur $\mathcal{L} u = -\alpha \Delta u + b \cdot \nabla u$. Le paramètre de stabilisation τ_K est à déterminer de telle sorte que $\tau_K = O\left(\dfrac{H}{\|b\|_{L^\infty(\Omega)}}\right)$ lorsque l'advection domine.

Le choix de ρ donne lieu à différentes méthodes:

- la méthode de Douglas-Wang (DW), $\rho = -1$;
- la méthode Streamline Upwind Petrov Galerkin (SUPG), $\rho = 0$;
- la méthode Galerkin Least Squares (GLS), $\rho = 1$.

Dans le cas où V_H est un espace d'approximation d'éléments finis $\mathbb{P}^1$ et div $b = 0$, les trois méthodes précédentes sont équivalentes car $(\mathcal{L}_s u_H)_{|K} = 0$ pour tout $u_H \in V_H$, pour tout $K \in \mathcal{T}_H$. Le théorème suivant donne une estimation a priori des méthodes de stabilisation dans le cas où V_H est un espace d'approximation d'éléments finis $\mathbb{P}^1$.

Théorème 1.1. *On se place dans le régime où l'advection domine*

$$Pe(x) = \frac{|b(x)|H}{2\alpha} > 1 \quad pp.\ x \in \Omega.$$

Alors, si u est la solution exacte du problème (1.6) et u_H est la solution de (1.12) avec comme paramètre

$$\tau_K(x) = \tau(x) = \frac{H}{2|b(x)|} \quad pour\ tout\ K \in \mathcal{T}_H,\ on\ a$$

$$|u - u_H|_{H^1(\Omega)} \leqslant CH\left(1 + \sqrt{\mathrm{Pe}\, H}\right)|u|_{H^2(\Omega)}. \tag{1.15}$$

En comparant les estimées (1.11) et (1.15), on peut s'attendre à une meilleure précision des méthodes stabilisées par rapport à la méthode $\mathbb{P}^1$ standard dans le régime où l'advection domine, ce qui est confirmé en pratique.

Remarque 1.2. *On a supposé ici la coercivité du problème (1.6). L'analyse numérique de ce type de méthodes repose en général sur la coercivité [19, 81], qui est souvent obtenue comme une conséquence de l'hypothèse*

$$\mathrm{div}\ b \leqslant c_0 < 0 \quad dans\ \Omega,$$

avec $b \in W^{1,\infty}(\Omega)$. À notre connaissance, l'analyse numérique des méthodes de stabilisation (1.12) n'a pas été réalisée dans le cas non coercif. Une méthode numérique stabilisée conçue pour des problèmes non symétriques, non coercifs, est proposée dans [20]. Cette méthode nécessite la résolution d'un système couplé entre le problème initial et un problème adjoint en utilisant des méthodes d'Éléments Finis stabilisés. Des estimées d'erreur en norme H^1 et L^2 sont prouvées sous l'hypothèse que le problème soit bien posé. Une méthode des moindres carrés a également été étudiée dans [14, 57].

Remarque 1.3. *Le choix du paramètre de stabilisation optimal τ_K est une question difficile et importante en pratique, car ce choix affecte la qualité de l'approximation numérique. On renvoie par exemple à [16, 39, 76, 49].*

1.2.2 Méthode des éléments finis multi-échelles

Problème diffusif On insère maintenant un caractère multi-échelles au problème (1.6) et on omet temporairement le terme d'advection. Plusieurs méthodes numériques multi-échelles ont été proposées: la méthode d'upscaling [11], la méthode Residual Free Bubbles [17], les méthodes multi-échelles variationnelles [49], la méthode des éléments finis multi-échelles (MsFEM) [48], la méthode multi-échelle hétérogène (HMM) [35], Equation-Free [56], Local generalized Finite Element [65], etc. Cette section est consacrée à la présentation de la méthode des éléments finis multi-échelles.

On s'intéresse au problème

$$-\mathrm{div}(A^\varepsilon \nabla u^\varepsilon) = f \quad \text{dans } \Omega, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega, \tag{1.16}$$

avec $f \in L^2(\Omega)$ et $A^\varepsilon \in L^\infty(\Omega)^{d\times d}$ symétrique vérifiant une condition d'ellipticité dans le sens où il existe $0 < \alpha_1 \leqslant \alpha_2$ tels que

$$0 < \alpha_1|\xi|^2 \leqslant (A^\varepsilon \xi)\cdot\xi \leqslant \alpha_2|\xi|^2, \quad \text{pour tout } \xi \in \mathbb{R}^d\backslash\{0\}. \tag{1.17}$$

La méthode des éléments finis multi-échelles (MsFEM) consiste en une approximation de Galerkin sur un espace d'approximation qui peut être précalculé et est adapté au problème considéré. Elle se décompose en trois étapes:

i) Introduire une discrétisation grossière de Ω notée $\mathcal{T}_H$; on définit l'espace d'approximation éléments finis $\mathbb{P}^1$ associé

$$V_H = \text{Vect}\left\{\phi_i^0, 1 \leqslant i \leqslant N_{V_H}\right\} \subset H_0^1(\Omega);$$

ii) Résoudre les problèmes locaux (un par élément pour chaque fonction de base ϕ_i^0 associée au maillage grossier)

$$-\text{div}\left(A^\varepsilon \nabla \psi_i^{\varepsilon,K}\right) = 0 \quad \text{dans } K, \qquad \psi_i^{\varepsilon,K} = \phi_i^0 \quad \text{sur } \partial K, \tag{1.18}$$

pour chaque élément K du maillage grossier, de manière à construire les fonctions de bases multi-échelles.

iii) Appliquer une approximation de Galerkin standard sur le problème (1.16) avec l'espace

$$V_H^\varepsilon = \text{Vect}\left\{\psi_i^\varepsilon,\ 1 \leqslant i \leqslant N_{V_H}\right\} \subset H_0^1(\Omega), \tag{1.19}$$

où ψ_i^ε est tel que $\psi_i^\varepsilon|_K = \psi_i^{\varepsilon,K}$ pour tout $K \in \mathcal{T}_H$.

L'analyse d'erreur de la méthode MsFEM dans le cas ci-dessus (1.16), pour $A^\varepsilon = A_{\text{per}}(\cdot/\varepsilon)$ où A_{per} est une matrice fixe périodique, a été faite dans [48] (voir aussi [36, Theorem 6.5] ou [63, Theorem 4.5]). Le résultat principal est énoncé dans le théorème suivant.

Théorème 1.4. *On se place dans le cadre périodique $A^\varepsilon(x) = A_{\text{per}}(x/\varepsilon)$. On suppose que A_{per} est hölderienne et que $H > \varepsilon$. On suppose également que la solution $u^\star$ du problème homogénéisé associé à (1.16) appartient à $W^{2,\infty}(\Omega)$. Soit u_H^ε l'approximation MsFEM de la solution u^ε du problème (1.16). On a alors*

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\Omega)} \leqslant C\left(H + \sqrt{\varepsilon} + \sqrt{\frac{\varepsilon}{H}}\right), \tag{1.20}$$

où C est une constante indépendante de H et ε.

Lorsque la taille du maillage grossier H est proche de l'échelle ε, un phénomène de résonance, encodé dans le terme $\sqrt{\varepsilon/H}$ dans (1.20), apparaît et dégrade la solution numérique. La méthode d'oversampling [47] est une méthode bien connue réduisant cet effet. Brièvement, cette approche, qui se trouve être non conforme, consiste à poser chaque problème local sur un domaine légèrement plus grand que l'élément de maillage grossier considéré, de manière à devenir moins sensible au choix arbitraire des conditions aux bords imposés sur le domaine plus large, puis de tronquer sur l'élément du maillage grossier les fonctions obtenues. Cette approche permet d'améliorer

considérablement les résultats numériques comparés à l'utilisation des conditions aux bords linéaires telle que dans (1.18). Dans le cas périodique, on a l'estimation suivante (voir [37]).

Théorème 1.5. *On se place dans le cadre du théorème 1.4. On suppose de plus que la distance entre un élément K et la frontière du macro élément utilisé dans l'oversampling est plus grande que H. On a alors*

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} \leqslant C\left(H + \sqrt{\varepsilon} + \frac{\varepsilon}{H}\right),$$

où $\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} = \sqrt{\sum_{K\in\mathcal{T}_H} \|u^\varepsilon - u_H^\varepsilon\|^2_{H^1(K)}}$.

Problème d'advection-diffusion On réinsère maintenant l'advection dans le problème. Plusieurs études ont été réalisées dans le régime où l'advection domine. Citons [72, 74] dans le cas de la méthode des éléments finis multi-échelles. Dans ces deux travaux, les problèmes locaux sont définis à partir de l'opérateur d'advection-diffusion. Dans cette thèse, on note cette variante l'approche Adv-MsFEM. Pour d'autres méthodes multi-échelles, on renvoie à [1, 44] pour la méthode multi-échelle hétérogène (HMM), [21] pour la méthode des éléments finis multi-échelles généralisée (GMsFEM) et [42] pour la méthode multi-échelle hybride mixte (MHM).

Dans [74], deux configurations sont analysées. Dans un premier temps, l'auteur étudie le problème dépendant du temps de la forme

$$\partial_t u^\varepsilon - \Delta u^\varepsilon + \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla u^\varepsilon = 0 \quad \text{dans } \mathbb{R}^2,$$

où b est $\mathbb{Z}^2$-périodique, de divergence nulle et de moyenne nulle. L'auteur s'intéresse seulement à obtenir des propriétés macroscopiques de la solution. Dans un second temps, toujours dans [74], le problème suivant est étudié:

$$-\Delta u^\varepsilon + \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla u^\varepsilon = f \quad \text{dans } \Omega, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega,$$

avec $f \in L^2(\Omega)$. Une estimée d'erreur est établie, uniquement dans la norme L^2 et pas dans la norme H^1 qui serait sensible à la précision de la méthode numérique pour décrire les oscillations fines de la solution. Cette estimée est obtenue sous des hypothèses vérifiées numériquement sur quelques exemples. Une étude expérimentale de convergence est réalisée et est en accord avec le résultat théorique obtenu.

Dans [72], l'auteur étudie le problème

$$\begin{cases} \rho^\varepsilon \partial_t u^\varepsilon - \operatorname{div}\left(A^\varepsilon \nabla u^\varepsilon\right) + \dfrac{1}{\varepsilon} b^\varepsilon \cdot \nabla u^\varepsilon = 0 & \text{dans } (0,T)\times(0,1)^d, \\ u^\varepsilon(0,\cdot) = u^0 \text{ dans } (0,1)^d, \qquad u^\varepsilon(t,\cdot) \text{ est } (0,1)^d\text{-périodique}, \end{cases}$$

où u^0 est régulier. Les fonctions $\rho^\varepsilon \in L^\infty((0,1)^d)$, $b^\varepsilon \in (L^\infty((0,1)^d))^d$ et $A^\varepsilon \in (L^\infty((0,1)^d))^{d\times d}$ ne dépendent pas du temps. Il est supposé qu'il existe une constante $\rho_m > 0$ telle que $\rho^\varepsilon \geqslant \rho_m$ presque partout sur $(0,1)^d$, et que b^ε est un champ de divergence nulle. Le champ de convection est plus général que dans [74] car il n'est pas supposé de moyenne nulle (mais une condition aux bords périodique est imposée sur $\partial(0,1)^d$). Dans le régime où l'advection domine, le problème est stabilisé en utilisant la méthode des caractéristiques pour intégrer l'opérateur de transport $\partial_t + \dfrac{b_H^\star}{\varepsilon}\cdot\nabla$, et une approche éléments finis multi-échelles pour la partie restante du terme

d'advection, i.e. $\frac{b^\varepsilon - \rho^\varepsilon b_H^\star}{\varepsilon} \cdot \nabla$, où $b_H^\star|_K = \frac{\int_K b^\varepsilon}{\int_K \rho^\varepsilon}$ pour tout $K \in \mathcal{T}_H$. Une estimée d'erreur est obtenue dans [72] dans le cadre périodique.

Méthode des éléments finis multi-échelles à la Crouzeix-Raviart avec fonctions bulles

On considère maintenant un problème multi-échelles posé dans un domaine perforé. On suppose que le domaine Ω^ε est obtenu en perforant le domaine Ω. Dans ce contexte, le caractère multi-échelles est amené par la géométrie du domaine. On présente une variante de la méthode des éléments finis multi-échelles particulièrement adaptée à ce type de problèmes. La méthode MsFEM à la Crouzeix-Raviart avec fonctions bulles est proposée dans [59] pour la résolution d'un problème elliptique en domaine perforé de la forme

$$-\Delta u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega^\varepsilon. \tag{1.21}$$

Dans cette version de la méthode éléments finis multi-échelles, la continuité inter-éléments est imposée de manière faible: l'intégrale du saut à l'interface de chaque arête du maillage grossier est nulle. Ceci apporte une robustesse à la méthode numérique vis-à-vis du choix arbitraire des conditions prescrites aux bords des éléments du maillage grossier dans la définition des problèmes locaux. Une estimation d'erreur est obtenue dans le cadre périodique.

Des variantes de la méthode MsFEM à la Crouzeix-Raviart avec fonctions bulles ont ensuite été proposées pour l'équation d'advection-diffusion [31] et le problème de Stokes [68]. Dans [31], les auteurs s'intéressent à un problème d'advection-diffusion de la forme

$$\begin{cases} -\Delta u^\varepsilon + b \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \\ u^\varepsilon = 0 \quad \text{sur } \partial\Omega^\varepsilon \backslash \partial\Omega, \\ u^\varepsilon = g \quad \text{sur } \partial\Omega \cap \partial\Omega^\varepsilon. \end{cases}$$

L'espace d'approximation est construit à partir de l'opérateur d'advection-diffusion. Les fonctions de base multi-échelles associées aux arêtes sont obtenues en imposant une condition de type Crouzeix Raviart. On impose la condition de Dirichlet homogène aux bords des problèmes locaux définissant les fonctions bulles. Les auteurs montrent numériquement l'importance de l'ajout de fonctions bulles dans l'espace d'approximation et explorent le cas où les domaines ne sont pas perforés périodiquement.

1.3 Résumé des travaux

1.3.1 Comparaison numérique de quelques méthodes MsFEM pour les problèmes d'advection-diffusion dans un milieu hétérogène dominés par l'advection

On s'intéresse dans le chapitre 2 de cette thèse à un problème d'advection-diffusion dans un milieu hétérogène de la forme

$$-\operatorname{div}\left(A^\varepsilon \nabla u^\varepsilon\right) + b \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega, \tag{1.22}$$

où ε représente l'échelle caractéristique de la variation de la matrice A^ε et $f \in L^2(\Omega)$. On suppose que $A^\varepsilon \in (L^\infty(\Omega))^{d\times d}$ vérifie la condition d'ellipticité (1.17). On suppose de plus que $b \in (L^\infty(\Omega))^d$ est tel que div $b = 0$ dans Ω. La limite homogénéisée dans le cadre périodique du problème (1.22) est la solution d'un problème d'advection-diffusion qui dépend, entre autres, du champ d'advection b (on renvoie à l'annexe A). On adopte une approche numérique de type éléments finis multi-échelles, telle que celle présentée dans la section 1.2.2, pour prendre en compte le caractère multi-échelles de ce problème. On se place dans le régime où l'advection domine sur la diffusion. Une méthode de stabilisation est nécessaire dans ce régime. Plusieurs approches ont été étudiées numériquement:

- encoder l'advection dans les fonctions de bases multi-échelles: Pour chaque élément du maillage $K \in \mathcal{T}_H$, on considère les problèmes locaux

$$-\text{div}\left(A^\varepsilon \nabla \phi_i^{\varepsilon,K}\right) + b\cdot\nabla\phi_i^{\varepsilon,K} = 0 \quad \text{dans } K, \quad \phi_i^{\varepsilon,K} = \phi_i^0 \quad \text{sur } \partial K.$$

  L'espace d'approximation est engendré par les ϕ_i^ε.

- appliquer une méthode de stabilisation de type Streamline-Upwind/Petrov-Galerkin (SUPG) sur un espace d'approximation standard d'éléments finis multi-échelles (voir la section 1.2.1). La méthode Stab-MsFEM est définie par la formulation variationnelle suivante:

$$\begin{aligned}&\text{Trouver } u_H^\varepsilon \in V_H^\varepsilon \text{ tel que pour tout } v_H^\varepsilon \in V_H^\varepsilon,\\ &\int_\Omega A^\varepsilon \nabla u_H^\varepsilon \cdot \nabla v_H^\varepsilon + (b\cdot\nabla u_H^\varepsilon) v_H^\varepsilon + a_{\text{stab}}(u_H^\varepsilon, v_H^\varepsilon) = \int_\Omega f v_H^\varepsilon + F_{\text{stab}}(v_H^\varepsilon),\end{aligned}$$

  où V_H^ε est défini par (1.19), a_{stab} et F_{stab} par (1.13) et (1.14) avec $\mathcal{L}u = -\text{div}\,(A^\varepsilon\nabla u) + b\cdot\nabla u$.

- une approche de type splitting entre un problème mono-échelle, dominé par l'advection et un problème de diffusion multi-échelles. La motivation principale de cette approche est son caractère non intrusif. En pratique, elle permet de coupler deux codes existants déjà optimisés pour chacun des deux sous-problèmes. Supposons qu'à l'itération n, on a calculé u_{2n} sur un maillage grossier et u_{2n+1} sur un maillage fin. On définit u_{2n+2} et u_{2n+3} par

$$\begin{cases} -\alpha_{\text{spl}}\Delta u_{2n+2} + b\cdot\nabla u_{2n+2} = f + b\cdot\nabla(u_{2n} - u_{2n+1}) & \text{dans } \Omega,\\ u_{2n+2} = 0 \quad \text{sur } \partial\Omega, \end{cases} \tag{1.23}$$

$$\begin{cases} -\text{div}\,(A^\varepsilon\nabla u_{2n+3}) = -\alpha_{\text{spl}}\Delta u_{2n+2} & \text{dans } \Omega,\\ u_{2n+3} = 0 \quad \text{sur } \partial\Omega. \end{cases} \tag{1.24}$$

  Le problème (1.23) est discrétisé avec la méthode $\mathbb{P}1$ SUPG sur un maillage grossier. Le terme $-b\cdot\nabla u_{2n+1}$, présent dans le second membre de (1.23), est intégré sur un maillage fin. Le problème (1.24) est, quant à lui, résolu numériquement avec la méthode MsFEM.

Dans le chapitre 2, on obtient une estimation d'erreur pour les méthodes de type MsFEM en dimension 1 sur le problème

$$-\frac{d}{dx}\left(A^\varepsilon\frac{du^\varepsilon}{dx}\right) + b\frac{du^\varepsilon}{dx} = f \quad \text{dans } \Omega = (0,L), \qquad u^\varepsilon(0) = u^\varepsilon(L) = 0,$$

avec un coefficient d'advection constant $b \neq 0$, $f \in L^2(0, L)$ et un coefficient de diffusion tel que $0 < \alpha_1 \leqslant A^\varepsilon(x) \leqslant \alpha_2$ presque partout sur Ω. On estime l'erreur numérique par rapport à ε, b et la taille du maillage grossier H (voir les théorèmes 2.9, 2.12 et 2.15). En supposant A^ε constant, on retrouve les estimées standards de la méthode $\mathbb{P}^1$ et sa version stabilisée à partir de celles obtenues respectivement pour la méthode MsFEM et sa version stabilisée. On étudie ensuite la convergence de l'approche de type splitting sans restriction sur la dimension de l'espace. On propose deux méthodes: la méthode (1.23)-(1.24) dont on observe qu'elle a une précision similaire à la méthode MsFEM stabilisée, et une seconde méthode, moins précise. On peut trouver des configurations où l'algorithme de la méthode (1.23)-(1.24) diverge alors que la seconde méthode dépend d'un paramètre qui garantit la convergence lorsqu'il est bien choisi.

Numériquement, les différentes approches sont comparées en terme de précision et de coût de calcul. Les méthodes les plus précises sont la méthode MsFEM stabilisée et la méthode de splitting. Leurs précisions sont relativement similaires. La différence essentielle entre les deux méthodes se trouve dans le caractère non intrusif de la méthode de splitting. Celui-ci entraîne l'augmentation du coût de la phase online qui devient prépondérant dans un contexte multi-requête (lorsque plusieurs seconds membres f doivent être considérés). Ce coût devient plus élevé lorsque seuls des solveurs itératifs peuvent être utilisés car la taille du problème ne permet plus d'utiliser des solveurs directs. Dans le cas où plusieurs champs d'advection sont étudiés, le coût de la méthode Adv-MsFEM devient considérable car les fonctions de bases multi-échelles doivent être calculées pour chaque champ d'advection. Les méthodes MsFEM stabilisée et de splitting sont encore compétitives car leurs fonctions de bases ne dépendent pas du champ d'advection. Elles sont peu précises dans la couche limite, ce qui est également le cas pour les autres approches. Lorsque la domination de l'advection est excessive, le caractère multi-échelles n'a plus d'influence sur la précision de la méthode numérique, et une méthode $\mathbb{P}^1$ stabilisée standard est aussi performante que la méthode MsFEM stabilisée. La compétitivité des méthodes MsFEM stabilisée et de splitting se trouve dans un régime intermédiaire dans lequel la précision de la méthode numérique est liée aux deux aspects, la domination de l'advection et le caractère multi-échelles.

1.3.2 Approximation stable de l'équation d'advection-diffusion utilisant la mesure invariante

L'objectif du chapitre 3 est d'étudier l'équation d'advection-diffusion mono-échelle

$$-\Delta u + b \cdot \nabla u = f \quad \text{dans } \Omega, \tag{1.25}$$

dans le régime non coercif et possiblement instable. On présente une approche numérique, à notre connaissance nouvelle, utilisant une mesure invariante associée à l'opérateur adjoint de (1.25) dont l'équation est de la forme

$$-\text{div}\,(\nabla\sigma + b\sigma) = 0 \quad \text{dans } \Omega. \tag{1.26}$$

La nature de la condition aux limites ajoutée à (1.25) a une influence sur le choix de la mesure invariante. On se place pour simplifier dans le cas de la condition aux limites de Dirichlet

homogène

$$u = 0 \quad \text{sur } \partial\Omega, \tag{1.27}$$

bien que d'autres conditions aux limites puissent être considérées.

L'approche numérique proposée s'inspire d'une preuve théorique du caractère bien posé du problème (1.25)-(1.27). Il est bien connu que le caractère bien posé de ce problème est donné par la théorie de Fredholm. D'autre part, le caractère bien posé du problème (1.25)-(1.27) est équivalent, d'après le théorème de Banach-Nečas-Babuška [38, Theorem 2.6], à un ensemble de deux conditions, la condition inf-sup

$$\exists \alpha > 0 \quad \text{tel que} \quad \inf_{u \in H_0^1(\Omega)} \sup_{v \in H_0^1(\Omega)} \frac{a(u,v)}{\|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}} \geqslant \alpha, \tag{1.28}$$

où

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v + \int_\Omega (b \cdot \nabla u) v,$$

et la seconde condition

$$\text{Pour tout } v \in H_0^1(\Omega), \quad (\text{Pour tout } w \in H_0^1(\Omega), \quad a(w,v) = 0) \Rightarrow v = 0.$$

Il s'agit alors de prouver les deux conditions énoncées ci-dessus et de rendre la constante α dans (1.28) explicite. L'étape principale de la preuve est de considérer le produit σv comme fonction test. On constate que

$$\int_\Omega (-\Delta u + b \cdot \nabla u) \sigma v = \int_\Omega \sigma \nabla u \cdot \nabla v + \int_\Omega (\nabla \sigma + \sigma b) \cdot \nabla u v - \int_{\partial\Omega} (\nabla u \cdot n) \sigma v.$$

Le terme de bord s'annule lorsque qu'on prescrit une condition de Dirichlet homogène au bord du domaine. En choisissant σ vérifiant (1.26), on obtient formellement

$$a(u, \sigma u) = \int_\Omega \sigma |\nabla u|^2 \quad \text{pour tout } u \in H_0^1(\Omega),$$

ce qui implique (1.28) si σ est positif et isolé de zéro.

L'approche numérique proposée revient à se ramener à ce problème modifié

$$-\operatorname{div}(\sigma \nabla u) + (\nabla \sigma + b\sigma) \cdot \nabla u = \sigma f \quad \text{dans } \Omega, \quad u = 0 \quad \text{sur } \partial\Omega, \tag{1.29}$$

qui est équivalent à (1.25)-(1.27) et a l'avantage d'être coercif.

Deux choix de mesures invariantes sont étudiés. La première mesure invariante σ_1 est l'unique solution du problème

$$-\operatorname{div}(\nabla \sigma_1 + b\sigma_1) = 0 \quad \text{dans } \Omega, \quad (\nabla \sigma_1 + b\sigma_1) \cdot n = 0 \quad \text{sur } \partial\Omega, \tag{1.30}$$

avec la contrainte de normalisation

$$⨍_\Omega \sigma_1 := |\Omega|^{-1} \int_\Omega \sigma_1 = 1. \tag{1.31}$$

Elle vérifie la propriété

$$\inf_{\Omega} \sigma_1 > 0. \tag{1.32}$$

Un second choix de mesure invariante est motivé ainsi. Dans le cas où div $b = 0$, choisir $\sigma = 1$ semble naturel car le problème (1.25)-(1.27) est coercif. Cependant, la mesure invariante σ_1 n'est pas nécessairement égale à 1. C'est pourquoi, on considère l'alternative

$$\begin{cases} -\operatorname{div}\,(\nabla\sigma_2 + b\sigma_2) = 0 \quad \text{dans } \Omega, \\ (\nabla\sigma_2 + b\sigma_2)\cdot n = b\cdot n - \fint_{\partial\Omega} b\cdot n \quad \text{sur } \partial\Omega, \\ \inf_{\Omega} \sigma_2 > 0. \end{cases} \tag{1.33}$$

On vérifie bien que, si div $b = 0$, alors $\sigma_2 \equiv 1$ est solution de (1.33). En pratique, la mesure invariante n'est, en général, pas connue analytiquement. On dispose seulement d'une approximation notée σ_h, où h désigne la taille du maillage associé à la discrétisation de (1.30)-(1.31) ou (1.33). On considère d'abord la mesure invariante σ_1 solution du problème (1.30)-(1.31). On utilise une méthode itérative pour imposer numériquement la contrainte de normalisation (1.31). On suppose que $\sigma_{1,h} > 0$ dans Ω ce qui est, en pratique, vérifié pour h suffisamment petit. On utilise une méthode de stabilisation (Douglas-Wang) de manière à obtenir la positivité de la solution discrète pour une taille de maillage h plus grande. La discrétisation d'une mesure invariante σ_2 solution de (1.33) est similaire à celle de σ_1. On commence par obtenir une approximation de σ_2^0, la solution du problème

$$\begin{cases} -\operatorname{div}\,(\nabla\sigma_2^0 + b\sigma_2^0) = 0 \quad \text{dans } \Omega, \\ (\nabla\sigma_2^0 + b\sigma_2^0)\cdot n = b\cdot n - \fint_{\partial\Omega} b\cdot n \quad \text{sur } \partial\Omega, \end{cases} \tag{1.34}$$

avec la contrainte de normalisation

$$\fint_{\Omega} \sigma_2^0 = 1.$$

On calcule $(\sigma_2^0)_h$ avec une méthode itérative pour imposer une contrainte de normalisation. On définit ensuite $\sigma_{2,h} = (\sigma_2^0)_h + \kappa_h\sigma_{1,h}$ où

$$\kappa_h = 1 + \inf\{\overline{\kappa} \in [0,\infty) \text{ tel que } (\sigma_2^0)_h + \overline{\kappa}\sigma_{1,h} > 0 \text{ dans } \Omega\}.$$

Revenons maintenant au problème modifié (1.29). On note U_H un espace d'approximation éléments finis $\mathbb{P}^1$ associé à un maillage régulier uniforme de taille H. On utilise une formulation antisymétrique du terme d'advection pour assurer la coercivité du problème discret. La méthode $(\mathbb{P}^1, \sigma_h\mathbb{P}^1)$ est alors définie par la formulation variationnelle suivante:

$$\begin{aligned} &\text{Trouver } u_H \in U_H \text{ tel que pour tout } v_H \in U_H, \\ &a_{\mathrm{ss}}(\sigma_h; u_H, v_H) = F(\sigma_h v_H), \end{aligned} \tag{1.35}$$

où

$$a_{\mathrm{ss}}(\sigma_h; u_H, v_H) = \int_\Omega \sigma_h \nabla u_H \cdot \nabla v_H + B_h \cdot \frac{(\nabla u_H) v_H - (\nabla v_H) u_H}{2},$$
$$B_h = \nabla \sigma_h + b\sigma_h.$$

Le problème (1.29) peut être dominé par l'advection. C'est pourquoi on étudie également la méthode $(\mathbb{P}^1, \sigma_h \mathbb{P}^1)$ GLS, une version stabilisée de la méthode $(\mathbb{P}^1, \sigma_h \mathbb{P}^1)$ par la méthode Galerkin Least Squares.

L'approche numérique proposée est bien posée inconditionnellement en la taille du maillage H. Cette propriété est utile dans les problèmes où on peut seulement obtenir une approximation grossière de u. Le cadre multi-échelle, où l'on remplace l'opérateur du Laplacien par $-\mathrm{div}(a(x/\varepsilon)\nabla\cdot)$, est un exemple particulier d'un tel contexte. On fournit également une analyse numérique détaillée de la convergence. Parmi les variantes proposées de l'approche numérique, les méthodes $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ et $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$ GLS sont les plus précises. Ces deux méthodes sont aussi stables et précises qu'une méthode de stabilisation standard. Elles sont légèrement plus robustes par rapport à la taille du maillage utilisé pour la discrétisation de u. L'approche présentée constitue une alternative intéressante, certifiée et efficace aux méthodes plus établies dans la littérature. On montre qu'elle devient compétitive dans certaines situations lorsque plusieurs seconds membres f doivent être considérés ou lorsque seul un maillage grossier peut être utilisé.

1.3.3 Méthodes MsFEM à la Crouzeix-Raviart pour des problèmes dominés par l'advection dans des domaines perforés

On examine dans le chapitre 4 des problèmes d'advection-diffusion posés dans un domaine perforé. Soit $\Omega \subset \mathbb{R}^d$, $d \geqslant 2$, un domaine ouvert régulier borné. On définit un domaine $\Omega^\varepsilon \subsetneq \Omega$ perforé par de petits trous de taille $\varepsilon > 0$, à une distance d'ordre ε les uns des autres. Sur le domaine perforé Ω^ε, on considère le problème (1.2) avec $\alpha > 0$, pour un second membre $f \in L^2(\Omega)$ et un champ d'advection $\widehat{b}^\varepsilon \in (W^{1,\infty}(\Omega^\varepsilon))^d$ tel que

$$\mathrm{div}\ \widehat{b}^\varepsilon = 0 \quad \text{dans } \Omega^\varepsilon. \tag{1.36}$$

On examine deux situations sensiblement différentes selon le type de condition prescrite aux bords des perforations: la condition de Dirichlet homogène dans le problème (1.3) et la condition de Neumann homogène dans le problème (1.4). On étudie un régime où l'advection domine sur la diffusion. La présence de perforations nécessite plus d'attention dans ce régime, particulièrement lorsque les perforations sont nombreuses, et asymptotiquement infiniment nombreuses. En effet, le choix de la condition prescrite aux bords des perforations influence fortement la nature du fluide, ce qui est attendu, et par conséquent les conclusions concernant la meilleure approche numérique à adopter. La condition de Dirichlet homogène aux bords des perforations atténue l'effet de l'advection, ce qui rend le fluide plus stable qu'en l'absence de perforations, contrairement au cas de la condition de Neumann homogène. Ce comportement peut s'illustrer en considérant la limite homogénéisée des problèmes (1.3) et (1.4) dans le cadre périodique. Si $\widehat{b}^\varepsilon = b\left(\frac{x}{\varepsilon}\right)$, avec b indépendant de ε, de divergence nulle, alors la solution u^ε du problème de Dirichlet (1.3) converge vers 0 à la vitesse ε^2. Renormalisée par le facteur ε^{-2}, la solution u^ε converge vers une limite qui ne dépend pas du champ d'advection b (on rappelle ce résultat dans le théorème 4.3). Pour

rétablir l'advection dans la limite homogénéisée du problème (1.3), ce qui numériquement revient formellement à conserver la domination de l'advection pour ε petit, on considère $\widehat{b}^\varepsilon = \frac{1}{\varepsilon} b\left(\frac{x}{\varepsilon}\right)$ (on renvoie au théorème 4.2). La situation est radicalement différente pour le problème (1.4). Si $\widehat{b}^\varepsilon = b\left(\frac{x}{\varepsilon}\right)$ (et sous des hypothèses supplémentaires), alors la solution u^ε du problème (1.4) reste d'ordre 1, au lieu de ε^2, et converge, en un certain sens, vers la solution d'un problème homogénéisé qui dépend du champ d'advection b, ce qu'on rappelle dans le théorème 4.4.

D'après [59], la méthode MsFEM à la Crouzeix-Raviart avec fonctions bulles se révèle être efficace pour le problème de diffusion dans un domaine perforé. Dans le chapitre 4, on évalue plusieurs variantes de cette méthode sur un problème d'advection-diffusion dominé par l'advection et posé dans un domaine perforé. Ce travail s'inscrit dans la suite de l'étude réalisée dans le chapitre 2, présentée dans la section 1.3.1, où l'on considère un problème d'advection-diffusion dominé par l'advection dans un milieu hétérogène.

Dans le cas du problème de Dirichlet (1.3), le fluide et les méthodes numériques sont stables, même pour des champs d'advection de forte intensité. L'approche Adv-MsFEM avec fonctions bulles également construites avec l'opérateur d'advection-diffusion est la meilleure approche parmi celles étudiées et ne requiert pas de stabilisation supplémentaire.

Dans le cas du problème (1.4) où l'on prescrit la condition de Neumann homogène aux bords des perforations, des instabilités numériques engendrées par la domination de l'advection apparaissent. On peut à nouveau utiliser l'approche Adv-MsFEM avec fonctions bulles construites avec l'opérateur d'advection-diffusion ou bien l'approche MsFEM stabilisée sans nécessairement ajouter de fonctions bulles.

1.4 Perspectives

L'exposé ci-dessus laisse des questions en suspens, et en ouvre d'autres. On en mentionne quelques unes.

On s'est intéressé à des problèmes d'advection diffusion multi-échelles dominés par l'advection. Tout d'abord, on s'est placé dans le cadre d'un problème d'advection-diffusion dans un milieu hétérogène. On a montré que le caractère multi-échelles du problème (1.1) peut être estompé lorsque la domination de l'advection est excessive. Dans un régime intermédiaire où le caractère multi-échelles influe en pratique sur la précision, les méthodes MsFEM stabilisée et de splitting sont les meilleures approches. L'analyse numérique des approches MsFEM réalisée est limitée au cas monodimensionnel. Il serait intéressant de poursuivre l'analyse numérique de ces approches sur le problème (1.1) en dimension supérieure. De plus, toutes les méthodes sont peu précises dans la région de la couche limite, à l'exception de la méthode Adv-MsFEM, lorsqu'on utilise des conditions aux limite de type Crouzeix-Raviart pour définir les fonctions de forme. Localiser la couche limite peut être un atout pour améliorer la stratégie d'approximation. On peut penser à différentes manières d'exploiter cette information comme appliquer un traitement différent (en termes de maillage ou de stabilisation par exemple) dans la couche limite ou bien introduire un couplage entre la couche limite et le reste du domaine. Ces idées sont déjà présentes dans la littérature, au moins dans l'utilisation des maillages de Shishkin [86] et la méthode hétérogène [78, Part II, Section 8.6], et représentent autant de pistes pour aborder le problème multi-échelle. Une autre piste d'amélioration possible est de construire une approximation de manière adaptative. De manière générale, l'adaptivité des méthodes multi-échelles est un sujet qu'il serait intéressant d'examiner dans le cas du problème (1.1).

Ce travail représente une étape préliminaire à l'étude des problèmes d'advection-diffusion posés dans un domaine perforé. On a abordé deux situations radicalement différentes selon le type de condition aux limites prescrites aux bords des perforations. Dans le cas du problème de Dirichlet (1.3), le fluide et les approches numériques sont stables, même pour de forts champs d'advection. L'approche Adv-MsFEM avec des fonctions bulles également construites avec l'opérateur d'advection-diffusion représente la meilleure approche possible et ne nécessite pas de stabilisation supplémentaire. Pour le problème (1.4) où l'on prescrit la condition de Neumann homogène aux bords des perforations, les difficultés numériques rencontrées sont similaires à celles du problème (1.1): des instabilités numériques apparaissent lorsque l'advection domine. On peut une nouvelle fois utiliser l'approche Adv-MsFEM avec fonctions bulles construites avec l'opérateur d'advection-diffusion ou bien l'approche MsFEM stabilisée sans nécessairement ajouter de fonctions bulles. Il serait intéressant d'obtenir une estimation d'erreur entre la solution du problème (1.4) et sa limite homogénéisée, ce qui ouvre alors la voie à une analyse numérique des approches MsFEM pour l'approximation de ce problème.
Une extension naturelle de ce travail est l'étude du problème d'Oseen en domaine perforé

$$\begin{cases} \nabla p^\varepsilon - \alpha \Delta u^\varepsilon + b \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega^\varepsilon, \\ \text{div } u^\varepsilon = 0 \quad \text{dans } \Omega^\varepsilon, \\ u^\varepsilon = 0 \quad \text{sur } \partial\Omega^\varepsilon, \end{cases}$$

avec $\alpha > 0$, $b \in (L^\infty(\Omega))^d$ et $f \in L^2(\Omega)$. On peut également envisager de remplacer la condition de Dirichlet homogène par une autre condition sur le bords des perforations.

Pour ce qui concerne le chapitre 3, on a proposé une approche éléments finis pour le problème (1.5) basée sur une mesure invariante associé à l'opérateur adjoint dans le cas où l'opérateur d'advection-diffusion n'est pas coercif et dominé par l'advection. On a montré que cette approche est bien posée inconditionnellement en la taille du maillage H et on en a analysé la convergence en détail. Il serait intéressant d'étendre au cadre multi-échelles cette approche, par exemple sur le problème (1.22) dans le cas non-coercif. Pour cela, il faut étendre le caractère isolé de zéro de la mesure invariante solution σ_1 solution du problème (1.30) au cas d'un opérateur de la forme $-\text{div}\,(A^\varepsilon \nabla \sigma_1 + b\sigma_1)$ où A^ε est régulier et vérifie (1.17). Dans le cadre du chapitre 3, on s'appuie sur le théorème [75, Theorem 1] dont la preuve repose sur une adaptation de l'inégalité de Harnack. Par ailleurs, cette extension présente une difficulté numérique majeure. Le problème que vérifie la mesure invariante est multi-échelle. Son calcul est nécessaire pour chaque milieu hétérogène, représenté dans le problème (1.22) par la matrice de diffusion A^ε. Cependant, la mesure invariante ne dépend pas du second membre du problème (1.22). Dans un contexte où l'on veut résoudre le problème (1.22) pour plusieurs seconds membres f, un seul calcul de la mesure invariante est nécessaire. On pourrait également envisager de définir la mesure invariante sur chaque élément d'un maillage grossier. Cette possibilité pourrait permettre d'éviter la discrétisation d'un problème multi-échelle posé sur le domaine entier en résolvant des problèmes locaux sur chaque élément d'un maillage grossier.

Chapter 2

A numerical comparison of some Multiscale Finite Element approaches for advection-dominated problems in heterogeneous media

Ce chapitre reprend, à quelques détails près, l'intégralité d'un article écrit en collaboration avec Claude Le Bris et Frédéric Legoll et accepté pour publication dans ESAIM: Mathematical Modelling and Numerical Analysis (M2AN) [60].

2.1 Introduction

We consider in this work an advection-diffusion equation that has both a multiscale character (encoded in an highly oscillatory diffusion coefficient) and a dominating advection. Formally, the equation reads as

$$-\operatorname{div}\left(A^{\varepsilon}\nabla u^{\varepsilon}\right)+b\cdot\nabla u^{\varepsilon}=f \quad \text{in } \Omega, \qquad u^{\varepsilon}=0 \quad \text{on } \partial\Omega. \tag{2.1}$$

Our self-explanatory notation will be made precise in the sequel, along with the mathematical setting that allows to rigorously consider this equation. Our purpose is to investigate whether numerical methods dedicated to the treatment of multiscale phenomena, such as Multiscale Finite Element Methods (henceforth abbreviated as MsFEM) and methods specifically designed to address the dominating advection, such as Streamline-Upwind/Petrov-Galerkin (SUPG) type methods, can separately adequately address the twofold problem, or, if need be, to discover how these methods may be combined to form the best possible approach in various regimes.

Equation (2.1) is practically relevant and interesting *per se*. Our study of this particular equation is nevertheless rather to be seen as a step toward the study of the following much more relevant case : an (single-scale) advection-diffusion equation, with a dominating advection term, posed on a *perforated* domain (in that vein, see [31]). In a previous, somewhat related couple of studies [58, 59], we have used with much benefit the highly oscillatory case as a test-bed for designing and studying approaches subsequently used for the more challenging perforated case.

Methods of the MsFEM type have proved efficient in a number of contexts. In essence, they are based upon choosing, as specific finite dimensional basis to expand the numerical solution upon, a set of functions that themselves are solutions to a highly oscillatory *local* problem, at scale ε, involving the differential operator present in the original equation. This problem-dependent basis set is likely to better encode the fine-scale oscillations of the solution and therefore allow to capture the solution more accurately. Numerical observation along with mathematical arguments prove that this is indeed generically the case. For the specific advection-diffusion equation (2.1) we consider here, two natural options for the construction of the basis set are (i) to pick as basis functions solutions to the (multiscale) diffusion operator only, or (ii) to also involve in the definition of the functions the advection operator. These two approaches will be among the set of approaches considered and tested below. In the former option, when the basis functions do not involve the advection operator, one may fear that, in the presence of advection, and especially in the presence of a strong advection that dominates the diffusion – a regime we focus on throughout this work –, the accuracy of the classical MsFEM dramatically deteriorates. This is for instance the case, "when $\varepsilon = 1$", for classical $\mathbb{P}^1$ finite element methods. Stabilization procedures are then in order and we will indeed adapt such a procedure to the present multiscale context. On the other hand, in the latter option, it is unclear whether the presence of the advection term *also* for the definition of the basis functions allows, or not, for the method to also perform well in the advection-dominated regime. This will be investigated below. However well such an approach performs, the fact that the advection is involved in the definition of the finite elements might create issues, and be prohibitively expensive computationally, when the advection varies and the equation needs to be solved repeatedly, either because the present steady state setting of (2.1) is in fact a time iteration within the numerical simulation of a time-dependent equation, or because equation (2.1) is part of an optimization, or inverse problem. Also, inserting the advection term in

the definition of the basis functions is a *very* invasive implementation, which might be problematic in some contexts. Both observations are sufficient motivations to also consider a splitting method, separately addressing the multiscale character with a classical MsFEM approach for the solution of the diffusion operator, and solving a single-scale advection-dominated advection-diffusion equation with a stabilized $\mathbb{P}^1$ method.

The four MsFEM-type approaches we have just mentioned (classical – that is, with basis functions constructed from diffusion only –, classical and stabilized, advection-diffusion based, splitting the advection and the multiscale character) will be studied and compared. For reference, we will also use a $\mathbb{P}^1$ finite element method, stabilized or not, in particular to investigate when the multiscale nature of the problem and the domination of the advection matter, or not.

In the context of HMM-type methods, multiscale advection-diffusion problems with dominating advection have been considered e.g. in [1].

Our article is organized as follows. Section 2.2 briefly recalls, essentially for the sake of self-consistency, some basic, classical and well-known facts on the building blocks (stabilization, multiscale approaches) we use, and describes in more details the numerical approaches we consider. We next provide, in Section 2.3, a complete numerical analysis of the approaches *in the one-dimensional setting*. We are unfortunately unable to conduct the same analysis in higher dimensions, but some of the issues we raise and discuss in the one-dimensional context are definitely useful to understand the approaches in a more general context. In particular, we point out that the direct application of an SUPG stabilization on MsFEM leads to an approach that is *not* strongly consistent (in sharp contrast to its single-scale, say $\mathbb{P}^1$ version), because the basis functions are not known analytically but only up to the numerical error present in the offline precomputation. We provide a solution to that difficulty. We show that, in spite of a lack of consistency, the method we design can be certified (and numerical observation will later show it performs efficiently). We also devote some time to the detailed study, in *any* dimension, of the convergence of the splitting approach.

Our final Section 2.4 presents a comprehensive series of numerical tests and comparisons. An executive summary of our *main* conclusions is as follows:

i) the best possible approach among all those we consider is the stabilized version of MsFEM, unless one does not want to be intrusive in which case the splitting approach performs approximately equally well, for an online computational cost that might be significantly larger, especially for problems of large size for which iterative solvers have to employed;

ii) the method using basis functions built upon the full advection-diffusion operator is not sufficiently stable to perform well in the advection-dominated regime;

iii) when advection outrageously dominates diffusion, the multiscale character of the solution (at least in the bulk of the domain) is essentially overshadowed by the advection, and a "classical" stabilized $\mathbb{P}^1$ finite element method performs as well as a MsFEM-type approach, a somewhat intuitive fact that our study allows to confirm.

Further details on the approaches considered are given in the body of the text.

2.2 Description of the numerical approaches

We describe in Section 2.2.1 the standard numerical tools we use throughout this work. We next present in Section 2.2.2 the four numerical methods we study.

2.2.1 Building blocks

In this section, we briefly recall for convenience some classical elements on the two building blocks we make use of to construct the approaches we study, namely stabilization methods (more specifically, SUPG type methods) and Multiscale Finite Element Methods (MsFEM). The reader already familiar with these notions may easily skip the present section and directly proceed to Section 2.2.2.

Stabilized methods

We temporarily consider the *single-scale* advection-diffusion problem

$$-\alpha\Delta u + b\cdot\nabla u = f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{2.2}$$

where Ω is a smooth bounded domain of $\mathbb{R}^d$, $\alpha > 0$, $b \in (L^\infty(\Omega))^d$ and $f \in L^2(\Omega)$. We suppose that

$$\operatorname{div} b = 0 \quad \text{in } \Omega, \tag{2.3}$$

so that problem (2.2) is coercive and amenable to standard numerical analysis techniques for coercive problems. We shall discuss the case of non-coercive problems in Remark 2.1 below.

Let $\mathcal{T}_H$ be a uniform regular mesh of size H discretizing Ω, and let V_H be the classical $\mathbb{P}^1$ Finite Element space associated to this mesh. The classical Galerkin approximation of (2.2) reads as the following variational formulation:

$$\text{Find } u_H \in V_H \text{ such that, for any } v_H \in V_H, \quad a(u_H, v_H) = F(v_H), \tag{2.4}$$

where

$$a(u,v) = \int_\Omega \alpha\nabla u\cdot\nabla v + (b\cdot\nabla u)v, \qquad F(v) = \int_\Omega fv. \tag{2.5}$$

Since the solution u to (2.2) is in $H^2(\Omega)$, we have the following error estimate as a direct consequence of Céa's lemma:

$$\|u - u_H\|_{H^1(\Omega)} \leqslant CH(1+\operatorname{Pe} H)\,\|u\|_{H^2(\Omega)}, \tag{2.6}$$

where C is independent of H and b. We have introduced, as is classical, the global Péclet number

$$\operatorname{Pe} = \frac{\|b\|_{L^\infty(\Omega)}}{2\alpha} \tag{2.7}$$

of problem (2.2). We thus see that the larger the product $\operatorname{Pe} H$, the larger the potential numerical error. Intuitively, the problem becomes less and less coercive as advection increasingly dominates over diffusion and, eventually, the coercivity is lost [38, Section 3.5.2] when Pe goes to $+\infty$. As is well-known, the Péclet number directly affects the quality of the numerical results. With the

standard $\mathbb{P}^1$ finite element approximation, oscillations polluting the solution are observed (see Figure 2.1 below).

Stabilization is a classical subject of numerical analysis. Many works (see e.g. [49] and the textbooks [78, 81]) have been devoted to designing stabilized methods for the advection-dominated regime. They consist in considering the following problem:

$$\begin{aligned} &\text{Find } u_H^s \in V_H \text{ such that, for any } v_H \in V_H, \\ &a(u_H^s, v_H) + a_{\text{stab}}(u_H^s, v_H) = F(v_H) + F_{\text{stab}}(v_H), \end{aligned} \tag{2.8}$$

where a and F are defined by (2.5) and a_{stab} and F_{stab} are defined by

$$a_{\text{stab}}(u_H^s, v_H) = \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} \mathcal{L} u_H^s, (\mathcal{L}_{ss} + \rho \mathcal{L}_s) v_H \right)_{\mathbf{K}}, \tag{2.9}$$

$$F_{\text{stab}}(v_H) = \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} f, (\mathcal{L}_{ss} + \rho \mathcal{L}_s) v_H \right)_{\mathbf{K}}, \tag{2.10}$$

where, for any u and v, $(u,v)_{\mathbf{K}} = \displaystyle\int_{\mathbf{K}} u\, v$, $\mathcal{L}_s u = -\alpha\Delta u$ and $\mathcal{L}_{ss} u = b \cdot \nabla u$ are the symmetric part and the skew-symmetric part of the advection-diffusion operator $\mathcal{L} v = -\alpha\Delta v + b \cdot \nabla v$, respectively (recall that b is divergence free in view of (2.3)). The stabilization parameter $\tau_{\mathbf{K}}$ is chosen, roughly, of the order of $\dfrac{H}{\|b\|_{L^\infty(\Omega)}}$. The choice of ρ leads to different stabilized methods. In the sequel, we only consider the Streamline Upwind Petrov-Galerkin method (SUPG), which corresponds to the choice $\rho = 0$.

The modification of the discrete bilinear form as in (2.8) allows to obtain the estimate

$$\|u - u_H^s\|_{H^1(\Omega)} \leqslant C H \left(1 + \sqrt{\text{Pe}\, H}\right) \|u\|_{H^2(\Omega)}, \tag{2.11}$$

where again C is independent of H and b. For large Péclet numbers (that is, $\text{Pe}\, H > 1$), this estimate is better than (2.6). More accurate numerical results are indeed obtained: see Figure 2.1 below. Note also that, in the right-hand sides of (2.6) and (2.11), $\|u\|_{H^2(\Omega)}$ depends on b, a fact that we will recall in Remark 2.11 below.

Estimate (2.11) is typically obtained under the assumptions

$$\frac{|b(x)|}{2\alpha} H \geqslant 1 \quad \text{for almost all } x \in \Omega, \tag{2.12}$$

and for the stabilization parameter

$$\tau_{\mathbf{K}}(x) = \frac{H}{2|b(x)|} \quad \text{for all } \mathbf{K} \in \mathcal{T}_H. \tag{2.13}$$

For the sake of completeness, and also because we will use similar arguments in Section 2.3 below for the multiscale setting, we provide the proof of (2.11) in Appendix 2.5.1 below.

Remark 2.1. *Notice that all the above analysis assumes that problem (2.2) is coercive (see (2.3)). This is usually the case in the literature, see [19, 81]. To the best of our knowledge, the analysis of the stabilized methods of the type (2.8) has not been performed in the non-coercive case. A stabilized numerical method designed for nonsymmetric noncoercive problems is proposed and studied in [20]. The method requires to solve the original problem coupled with an adjoint problem*

using stabilized finite element methods. Error estimates in H^1 and L^2 norms are proved under the assumption of well-posedness of the problem. Least-square methods for noncoercive elliptic problems have also been studied, see e.g. [14, 57].

Remark 2.2. *The stabilized methods (2.8) can also be understood in the framework of the Variational Multiscale Methods [49].*

Remark 2.3. *The choice of an optimal stabilization parameter $\tau_{\mathbf{K}}$ is a difficult and sensitive question, since it affects the quality of the numerical approximation. We refer e.g. to [16, 39, 76]. The Variational Multiscale Methods [49] give an interpretation of the stabilization parameter. If we assume $\tau_{\mathbf{K}}$ to be constant on each mesh element $\mathbf{K}$, the Variational Multiscale Methods yield the formula*

$$\tau_{\mathbf{K}} = \frac{1}{|\mathbf{K}|} \int_{\mathbf{K}} \int_{\mathbf{K}} g_{\mathbf{K}}, \tag{2.14}$$

where $g_{\mathbf{K}}$ is the Green's function of the operator $\mathcal{L}^\star$ (i.e. the adjoint of $\mathcal{L}$) with homogeneous Dirichlet boundary conditions on $\partial\mathbf{K}$. Simplifying assumptions are next used to infer, from (2.14), a practical expression for $\tau_{\mathbf{K}}$.

For the sake of illustration, and because it allows us to introduce notions useful for what follows, we briefly consider the one-dimensional example

$$-\alpha u'' + bu' = f \quad \text{in } (0,1), \qquad u(0) = u(1) = 0, \tag{2.15}$$

for a constant b. In that case, the expression (2.14) can be analytically computed and yields the choice

$$\tau_{\mathbf{K}} = \frac{H}{2|b|} \left(\coth(\mathrm{Pe}\, H) - \frac{1}{\mathrm{Pe}\, H} \right). \tag{2.16}$$

On Figure 2.1, we show the exact solution to (2.15) as well as two numerical approximations. We set $\alpha = 1/256$, $b = 1$, $f = 1$ and $H = 1/16$, so that we are in the advection-dominated regime. Table 2.1 shows the relative errors of the methods.

	L^2: $\mathbb{P}^1$	L^2 : $\mathbb{P}^1$ SUPG	H^1: $\mathbb{P}^1$	H^1: $\mathbb{P}^1$ SUPG
Outside the layer	0.3217	0.0624	0.8913	0.2228
Inside the layer	0.0297	0.1549	0.3722	0.7163
In the whole domain	0.3513	0.2173	1.2635	0.9391

Table 2.1 Relative errors for (2.15) with $\alpha = 1/256$, $b = 1$, $f = 1$ and $H = 1/16$. The parameter $\tau_{\mathbf{K}}$ is given by (2.16).

On Figure 2.1, we can distinguish two regions. Outside the boundary layer, the $\mathbb{P}^1$ SUPG method accurately approximates the solution. It has no spurious oscillations, in contrast to the standard $\mathbb{P}^1$ method. Inside the boundary layer, the $\mathbb{P}^1$ SUPG method only poorly performs.

MsFEM approaches

We now insert a multiscale character in our problem and temporarily erase the transport field b, which we will shortly reinstate in the next section. We consider the solution $u^\varepsilon \in H_0^1(\Omega)$ to

$$-\mathrm{div}\,(A^\varepsilon \nabla u^\varepsilon) = f \quad \text{in } \Omega, \qquad u^\varepsilon = 0 \quad \text{on } \partial\Omega. \tag{2.17}$$

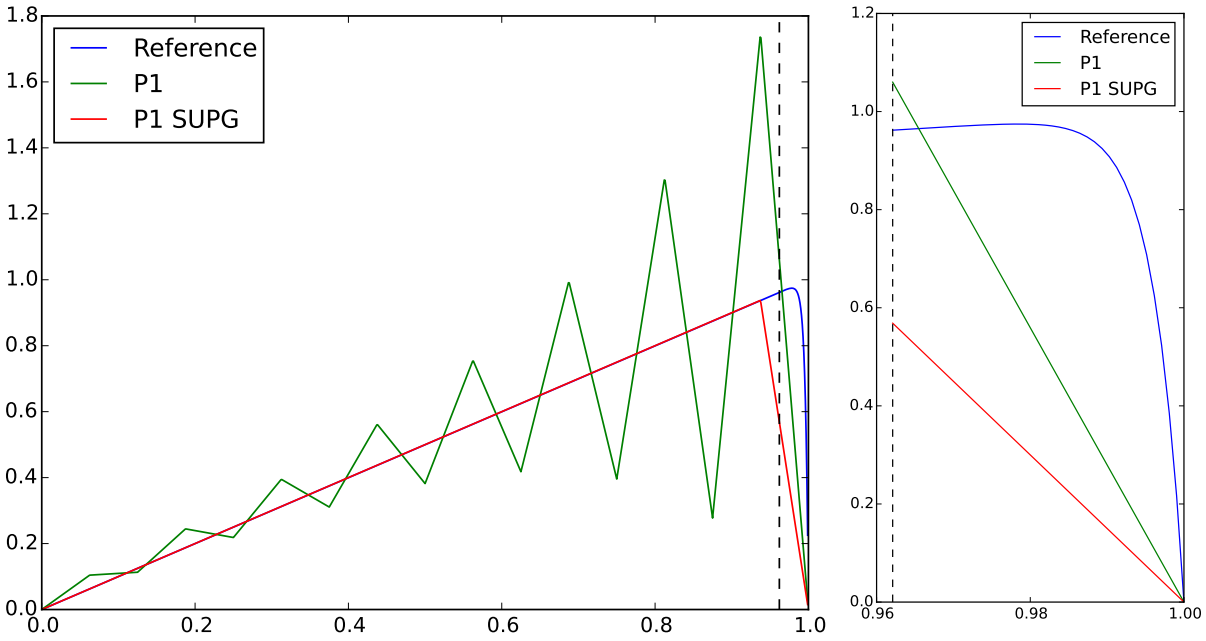

Figure 2.1 Exact and numerical solutions to (2.15) for $\mathrm{Pe}\, H = 8$. Left: Plot on the whole domain. Right: Close-up on the boundary layer. The vertical dashed line delineates the boundary layer.

We assume that the diffusion matrix A^ε, encoding the oscillations at the small scale, is elliptic in the sense that there exists $0 < \alpha_1 \leqslant \alpha_2$ such that

$$\forall \varepsilon, \quad \forall \xi \in \mathbb{R}^d, \quad \alpha_1 |\xi|^2 \leqslant (A^\varepsilon(x)\xi) \cdot \xi \leqslant \alpha_2 |\xi|^2 \quad \text{a.e. on } \Omega. \tag{2.18}$$

Throughout this article, we shall perform our theoretical analysis for general, not necessarily symmetric, matrix-valued coefficients A^ε, not necessarily either of the form $A^\varepsilon = A\,(\cdot/\varepsilon)$ for a fixed matrix A (although one may consider such a case to fix the ideas). In our numerical tests, however, we only consider a scalar coefficient A^ε.

The bottom line of the MsFEM is to perform a Galerkin approximation using specific basis functions, which are precomputed (in an offline stage) and adapted to the problem considered.

On the prototypical multiscale diffusion problem (2.17), the method, in one of its simplest variant, consists of the following three steps:

i) Introduce a discretization of Ω with a coarse mesh; throughout this article, we work with the $\mathbb{P}^1$ Finite Element space

$$V_H = \mathrm{Span}\left\{\phi_i^0, 1 \leqslant i \leqslant N_{V_H}\right\} \subset H_0^1(\Omega); \tag{2.19}$$

ii) Solve the local problems (one for each basis function for the coarse mesh)

$$-\mathrm{div}\left(A^\varepsilon \nabla \psi_i^{\varepsilon,\mathbf{K}}\right) = 0 \quad \text{in } \mathbf{K}, \qquad \psi_i^{\varepsilon,\mathbf{K}} = \phi_i^0 \quad \text{on } \partial\mathbf{K}, \tag{2.20}$$

on each element $\mathbf{K}$ of the coarse mesh, in order to build the multiscale basis functions.

iii) Apply a standard Galerkin approximation of (2.17) on the space

$$V_H^\varepsilon = \mathrm{Span}\left\{\psi_i^\varepsilon,\ 1 \leqslant i \leqslant N_{V_H}\right\} \subset H_0^1(\Omega), \tag{2.21}$$

where ψ_i^ε is such that $\psi_i^\varepsilon|_{\mathbf{K}} = \psi_i^{\varepsilon,\mathbf{K}}$ for all $\mathbf{K} \in \mathcal{T}_H$.

The error analysis of the MsFEM method in the above case (2.17), for $A^\varepsilon = A_{\text{per}}(\cdot/\varepsilon)$ with A_{per} a fixed periodic matrix, has been performed in [48] (see also [36, Theorem 6.5] or [63, Theorem 4.5]). The main result is stated in the following Theorem.

Theorem 2.4. *We consider the periodic case $A^\varepsilon(x) = A_{per}(x/\varepsilon)$. We assume that A_{per} is Hölder continuous and that $H > \varepsilon$. We also assume that the solution $u^\star$ to the homogenized problem associated to (2.17), that is the L^2-limit of u^ε solution to (2.17) when $\varepsilon \to 0$, belongs to $W^{2,\infty}(\Omega)$. Let u_H^ε be the MsFEM approximation of the solution u^ε to (2.17). Then*

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\Omega)} \leqslant C\left(H + \sqrt{\varepsilon} + \sqrt{\frac{\varepsilon}{H}}\right), \tag{2.22}$$

where C is a constant independent of H and ε.

When the coarse mesh size H is close to the scale ε, a resonance phenomenon, encoded in the term $\sqrt{\varepsilon/H}$ in (2.22), occurs and deteriorates the numerical solution. The oversampling method [47] is a popular technique to reduce this effect. In short, the approach, which is non-conforming, consists in setting each local problem on a domain slightly larger than the actual element considered, so as to become less sensitive to the arbitrary choice of boundary conditions on that larger domain, and next truncate on the element the functions obtained. That approach allows to significantly improve the results compared to using linear boundary conditions as in (2.20). In the periodic case, we have the following estimate (see [37]).

Theorem 2.5. *Assume the setting and the notation of Theorem 2.4. Assume additionally that the distance between an element $\mathbf{K}$ and the boundary of the macro element used in the oversampling is larger than H. Then*

$$\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} \leqslant C\left(H + \sqrt{\varepsilon} + \frac{\varepsilon}{H}\right),$$

where $\|u^\varepsilon - u_H^\varepsilon\|_{H^1(\mathcal{T}_H)} = \sqrt{\sum_{\mathbf{K}\in\mathcal{T}_H} \|u^\varepsilon - u_H^\varepsilon\|^2_{H^1(\mathbf{K})}}$ is the H^1 broken norm of $u^\varepsilon - u_H^\varepsilon$.

Remark 2.6. *The boundary conditions imposed in (2.20) are the so-called linear boundary conditions. Besides the linear boundary conditions, and the oversampling technique alluded to above, there are many other possible boundary conditions for the local problems. They may give rise to conforming, or non-conforming approximations. The choice sensitively affects the overall accuracy. We will explore this issue, in our specific context, in Section 2.4.2 below.*

It is important to notice that the estimates of Theorems 2.4 and 2.5 hold true assuming that the multiscale basis functions employed to compute the approximation u_H^ε are the *exact* solutions of the local problems. In practice of course, the local problems (2.20) are only approximated numerically, using a fine mesh of size h sufficiently small to capture the oscillations at scale ε.

As mentioned above, our purpose is to understand how to adapt the stabilization methods and the MsFEM methods in order to efficiently approximate

$$-\text{div}\,(A^\varepsilon \nabla u^\varepsilon) + b \cdot \nabla u^\varepsilon = f \quad \text{in } \Omega, \qquad u^\varepsilon = 0 \quad \text{on } \partial\Omega, \tag{2.23}$$

where $A^\varepsilon \in (L^\infty(\Omega))^{d\times d}$ satisfies (2.18), $b \in (L^\infty(\Omega))^d$ and $f \in L^2(\Omega)$. Notice that the transport field b is assumed to be independent of ε. We also choose it divergence-free as in (2.3). The variational formulation of (2.23) is:

$$\text{Find } u^\varepsilon \in H^1_0(\Omega) \text{ such that, for any } v \in H^1_0(\Omega), \quad a^\varepsilon(u^\varepsilon, v) = F(v), \tag{2.24}$$

where

$$a^\varepsilon(u,v) = \int_\Omega (A^\varepsilon \nabla u)\cdot\nabla v + (b\cdot\nabla u)\, v, \qquad F(v) = \int_\Omega fv. \tag{2.25}$$

We now introduce in Section 2.2.2 below the four numerical approaches we consider.

2.2.2 Our four numerical approaches

The classical MsFEM and its stabilized version

The classical MsFEM described in Section 2.2.1 is the first approach we consider. It performs a Galerkin approximation of (2.23) on the space (2.20)-(2.21). Notice that in this approximation, the transport term $b\cdot\nabla$, although present in the equation (2.23), is absent from the local problems (2.20) and thus from the definition of the basis functions. It is immediate to realize that this approach coincides with the standard $\mathbb{P}^1$ method on (2.2) when $A^\varepsilon = \alpha\,\mathrm{Id}$. Consequently, the method is expected to be unstable in the advection-dominated regime, as recalled in Section 2.2.1, and this is indeed observed in practice, as will be seen in Section 2.4.2.

This motivates the introduction of a stabilized version of this method, which is the adaptation to the multiscale context of the classical SUPG method. As we shall now see, some difficulty arises regarding the consistency of the approach, owing to the fact that the basis functions we use in practice are only approximate.

First, we consider the exact approximation space V_H^ε defined by (2.21). The SUPG stabilization, readily applied to our problem (2.24), yields the following variational formulation:

$$\begin{aligned}&\text{Find } u_H^\varepsilon \in V_H^\varepsilon \text{ such that, for any } v_H^\varepsilon \in V_H^\varepsilon,\\ &a^\varepsilon(u_H^\varepsilon, v_H^\varepsilon) + a_{\mathrm{stab}}(u_H^\varepsilon, v_H^\varepsilon) = F(v_H^\varepsilon) + F_{\mathrm{stab}}(v_H^\varepsilon),\end{aligned} \tag{2.26}$$

where we recall that the SUPG stabilization terms are (see (2.9) and (2.10))

$$\begin{aligned}a_{\mathrm{stab}}(u_H^\varepsilon, v_H^\varepsilon) &= \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}}\left(-\mathrm{div}\ (A^\varepsilon\nabla u_H^\varepsilon) + b\cdot\nabla u_H^\varepsilon\right), b\cdot\nabla v_H^\varepsilon\right)_{L^2(\mathbf{K})},\\ F_{\mathrm{stab}}(v_H^\varepsilon) &= \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} f, b\cdot\nabla v_H^\varepsilon\right)_{\mathbf{K}}.\end{aligned} \tag{2.27}$$

The method is, as is well known, strongly consistent (meaning that the exact solution u^ε solves (2.26)). Because of the definition of the approximation space V_H^ε, we have

$$a_{\mathrm{stab}}(u_H^\varepsilon, v_H^\varepsilon) = a_{\mathrm{upw}}(u_H^\varepsilon, v_H^\varepsilon) \quad \text{for any } (u_H^\varepsilon, v_H^\varepsilon) \in (V_H^\varepsilon)^2, \tag{2.28}$$

where

$$a_{\mathrm{upw}}(u_H^\varepsilon, v_H^\varepsilon) = \sum_{\mathbf{K}\in\mathcal{T}_H} (\tau_{\mathbf{K}} b\cdot\nabla u_H^\varepsilon, b\cdot\nabla v_H^\varepsilon)_{L^2(\mathbf{K})}. \tag{2.29}$$

In practice however, we only know a discrete approximation $\psi^{\varepsilon,h}$, on a fine mesh $\mathbf{K}_h$, of the solution ψ^ε to (2.20). Put differently, we manipulate $V_{H,h}^\varepsilon = \mathrm{Span}\{\psi_i^{\varepsilon,h}, 1 \leqslant i \leqslant N_{V_H}\}$ instead of V_H^ε. It follows that, for example when $A^\varepsilon \in \mathcal{C}^0(\overline{\Omega})$ and we use a $\mathbb{P}^1$ approximation on a fine mesh $\mathbf{K}_h$ for the local problem (2.20), $A^\varepsilon \nabla u_{H,h}^\varepsilon$ may be discontinuous at the edges of the mesh $\mathbf{K}_h$, and $-\mathrm{div}\,(A^\varepsilon \nabla u_{H,h}^\varepsilon) \notin L^1_{\mathrm{loc}}(\mathbf{K})$.

We may consider at least two ways to circumvent that difficulty. First, if the matrix coefficient A^ε is locally sufficiently regular, we may define the stabilization term as

$$\widetilde{a}_{\mathrm{stab}}(u_{H,h}^\varepsilon, v_{H,h}^\varepsilon) = \sum_{\mathbf{K}\in\mathcal{T}_H} \sum_{\kappa\subset K_h} \left(\tau_{\mathbf{K}} \left(-\mathrm{div}\ \left(A^\varepsilon \nabla u_{H,h}^\varepsilon\right) + b\cdot\nabla u_{H,h}^\varepsilon\right), b\cdot\nabla v_{H,h}^\varepsilon\right)_{L^2(\kappa)}.$$

When, as is the case here, we employ a $\mathbb{P}^1$ approximation on $\mathbf{K}_h$, all we need for this stabilization term to make sense is that the vector field $\mathrm{div}\,(A^\varepsilon)$ belongs to $L^1(\kappa)$ for all $\kappa \subset K_h$. This is more demanding than the simple classical assumption $A^\varepsilon \in L^\infty(\Omega)$. Under this assumption, we obtain a strongly consistent stabilized method. We will however not proceed in this direction and favor an alternate approach, to which we now turn.

Based upon the observation (2.28) for the "ideal" space V_H^ε, we may use the stabilization term (2.29) rather than (2.27). In contrast to (2.27), the quantity (2.29) is also well defined on $V_{H,h}^\varepsilon$. And this holds true without any additional regularity assumption on A^ε. The Stab-MsFEM method we employ is hence defined by the following variational formulation:

$$\begin{aligned} &\text{Find } u_{H,h}^\varepsilon \in V_{H,h}^\varepsilon \text{ such that, for any } v_{H,h}^\varepsilon \in V_{H,h}^\varepsilon, \\ &a^\varepsilon(u_{H,h}^\varepsilon, v_{H,h}^\varepsilon) + a_{\mathrm{upw}}(u_{H,h}^\varepsilon, v_{H,h}^\varepsilon) = F(v_{H,h}^\varepsilon) + F_{\mathrm{stab}}(v_{H,h}^\varepsilon). \end{aligned} \tag{2.30}$$

We emphasize that employing that stabilization comes at a price: we give up on strong consistency. We provide in Section 2.3.2, Theorem 2.14 below, an error estimate in the one-dimensional setting for this method. Despite the absence of consistency, we can still prove that the method is convergent.

The Adv-MsFEM variant

In contrast to our first two approaches, the Adv-MsFEM approach we discuss in this section accounts for the transport field in the local problems. For each mesh element $\mathbf{K} \in \mathcal{T}_H$, we indeed now consider

$$-\mathrm{div}\,\left(A^\varepsilon \nabla \phi_i^{\varepsilon,\mathbf{K}}\right) + b\cdot\nabla\phi_i^{\varepsilon,\mathbf{K}} = 0 \quad \text{in } \mathbf{K}, \qquad \phi_i^{\varepsilon,\mathbf{K}} = \phi_i^0 \quad \text{on } \partial\mathbf{K}, \tag{2.31}$$

instead of (2.20), and next the approximation space

$$V_H^{\varepsilon,Adv} = \mathrm{Span}\,\{\phi_i^\varepsilon,\ 1 \leqslant i \leqslant N_{V_H}\} \subset H_0^1(\Omega)$$

defined as in (2.21). Problem (2.31) is an advection-diffusion problem with, in principle, a high Péclet number. Nevertheless, the problem is local and is to be solved offline, so we may easily employ a mesh size sufficiently fine to avoid the issues presented in Section 2.2.1.

There is however a difficulty in considering (2.31) and b-dependent basis functions $\phi_i^{\varepsilon,\mathbf{K}}$. In the context where we want to repeatedly solve (2.23) for multiple b, for instance when b depends

on an external parameter such as time, the method becomes prohibitively expensive as we will see in Section 2.4.3.

We note in passing the following consistency. In the one-dimensional single-scale example (2.15), the stiffness matrix of the Adv-MsFEM method is

$$M_{\text{Adv-MsFEM}} = \text{Tridiag}\left(\frac{-b\exp\left(bH/\alpha\right)}{\exp\left(bH/\alpha\right)-1}, |b|\coth\left(\frac{|b|H}{2\alpha}\right), \frac{-b}{\exp\left(bH/\alpha\right)-1}\right).$$

It then coincides with the stiffness matrix $M_{\mathbb{P}^1\text{SUPG}}$ of the $\mathbb{P}^1$ SUPG method with $\tau_{\mathbf{K}}$ given by (2.16).

We also note that, in view of (2.27)–(2.31), we have that $a_{\text{stab}}(u_H^{\varepsilon,\text{Adv}}, v_H^{\varepsilon,\text{Adv}}) = 0$ for any $(u_H^{\varepsilon,\text{Adv}}, v_H^{\varepsilon,\text{Adv}}) \in (V_H^{\varepsilon,Adv})^2$. Such a stabilization is therefore void on the Adv-MsFEM method. Actually, we shall see in the numerical tests of Section 2.4.2 that the Adv-MsFEM method is only moderately sensitive to the Péclet number.

MsFEM type basis functions depending on the transport term for multiscale advection-diffusion problems have already considered in the literature. In [74], two settings are investigated. The Adv-MsFEM is first applied to the time-dependent multiscale advection-diffusion equation

$$\partial_t u^\varepsilon - \Delta u_\varepsilon + \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla u^\varepsilon = 0 \quad \text{in } \mathbb{R}^2,$$

with $b = \nabla^\perp\psi$ where $\psi(x) = \psi(x_1,x_2) = \dfrac{1}{4\pi^2}\sin(2\pi x_1)\sin(2\pi x_2)$. The field b is thus $\mathbb{Z}^2$-periodic, divergence-free and of mean zero. The purpose is then to only capture macroscopic properties of the solution u^ε. Also in [74], the Adv-MsFEM is investigated on the problem

$$-\Delta u^\varepsilon + b^\varepsilon\cdot\nabla u^\varepsilon = f,$$

with $b^\varepsilon \in (L^\infty(\Omega))^2$ and $f \in L^2(\Omega)$. Only the following L^2 error estimate

$$\frac{\|u^\varepsilon - u_H^\varepsilon\|_{L^2(\Omega)}}{\|u^\varepsilon\|_{L^2(\Omega)}} \leqslant C\frac{\varepsilon}{H} + CH^2\|f\|_{L^2(\Omega)}$$

is derived, and not an H^1 estimate which would be sensitive to how well the fine oscillations are captured by the numerical approach. It is completed in the periodic case, where $b^\varepsilon(x) = \dfrac{1}{\varepsilon}b_{\text{per}}\left(\dfrac{x}{\varepsilon}\right)$ for a fixed, periodic, divergence-free function b_{per} of mean zero, under some assumptions which have been numerically verified on some examples. An experimental study of convergence is performed and shows good agreement with the above theoretical error estimate.

A second reference we wish to cite is [72]. The author studies there the problem

$$\begin{cases} \rho^\varepsilon\partial_t u^\varepsilon - \text{div}\left(A^\varepsilon\nabla u^\varepsilon\right) + \dfrac{1}{\varepsilon}b^\varepsilon\cdot\nabla u^\varepsilon = 0 & \text{in } (0,1)^d\times(0,T), \\ u^\varepsilon(0,\cdot) = u^0 \text{ in } (0,1)^d, \qquad u^\varepsilon(t,\cdot) \text{ is } (0,1)^d\text{-periodic}, \end{cases}$$

where $u^0 \in W^{m,\infty}_{\text{per}}((0,1)^d)$ with $m \geqslant 3$. The functions $\rho^\varepsilon \in L^\infty((0,1)^d)$, $b^\varepsilon \in (L^\infty((0,1)^d))^d$ and $A^\varepsilon \in (L^\infty((0,1)^d))^{d\times d}$ do not depend on time. It is assumed that there exists a constant $\rho_m > 0$ such that $\rho^\varepsilon \geqslant \rho_m$ a.e. on $(0,1)^d$, and that b^ε is divergence-free. In contrast to [74], the mean of b^ε is not assumed to vanish (but periodic boundary conditions are imposed on $\partial(0,1)^d$). In the advection-dominated regime, the problem is stabilized using the characteristics method for

integrating the transport operator $\partial_t + \frac{b_H^\star}{\varepsilon} \cdot \nabla$, and the multiscale finite element method for the remaining part of the advection term, i.e. $\frac{b^\varepsilon - \rho^\varepsilon b_H^\star}{\varepsilon} \cdot \nabla$, where $b_H^\star|_{\mathbf{K}} = \frac{\int_{\mathbf{K}} b^\varepsilon}{\int_{\mathbf{K}} \rho^\varepsilon}$ for all $\mathbf{K} \in \mathcal{T}_H$. The MsFEM approach which is used in [72] is inspired by the variant of the Multiscale Finite Element approach introduced in [6] for purely diffusive problems. The multiscale basis functions are thus defined by $\phi_j^\varepsilon(x) = \phi_j^0\Big(w^{\varepsilon,H}(x)\Big)$ for $1 \leqslant j \leqslant N_{V_H}$, where ϕ_j^0 are the $\mathbb{P}^1$ basis functions and $w^{\varepsilon,H}|_{\mathbf{K}} = (w_1^{\varepsilon,\mathbf{K}}, \ldots, w_d^{\varepsilon,\mathbf{K}})$ for each $\mathbf{K} \in \mathcal{T}_H$, where, for any $i = 1, \ldots, d$, the function $w_i^{\varepsilon,\mathbf{K}}$ is the solution to

$$-\text{div}\,\Big(A^\varepsilon \nabla w_i^{\varepsilon,\mathbf{K}}\Big) + \frac{b^\varepsilon - \rho^\varepsilon b_H^\star}{\varepsilon} \cdot \nabla w_i^{\varepsilon,\mathbf{K}} = 0 \quad \text{in } \mathbf{K}, \qquad w_i^{\varepsilon,\mathbf{K}} = x_i \quad \text{on } \partial\mathbf{K}.$$

Note that, as in (2.31), the basis functions depend on the advection field. An error estimate is established in [72] for the periodic case.

A splitting approach

The fourth and last approach we consider is a *splitting method* that decomposes (2.23) into a single-scale, advection-dominated problem and a multiscale, purely diffusive problem. The main motivation for considering such a splitting approach is the non-intrusive character of the approach. In practice, one may couple legacy codes that are already optimized for each of the two subproblems.

Of course, splitting methods have been used in a large number of contexts. To cite only a couple of works relevant to our context, we mention [51] for a review on the splitting methods for time-dependent advection-diffusion equations, and [87] for the introduction of a viscous splitting method based on a Fourier analysis for the steady-state advection-diffusion equation.

Our splitting approach for (2.23) is the following. We define the iterations by

$$\begin{cases} -\alpha_{\text{spl}}\Delta u_{2n+2} + b \cdot \nabla u_{2n+2} = f + b \cdot \nabla(u_{2n} - u_{2n+1}) & \text{in } \Omega, \\ u_{2n+2} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.32}$$

$$\begin{cases} -\text{div}\,(A^\varepsilon \nabla u_{2n+3}) = -\alpha_{\text{spl}}\Delta u_{2n+2} & \text{in } \Omega, \\ u_{2n+3} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.33}$$

with $\alpha_{\text{spl}} > 0$. The initialization is e.g. $u_0 = u_1 = 0$.

The functions u_{2n} with even indices are approximations defined on a coarse mesh, using $\mathbb{P}^1$ finite elements, and, since our context is that of advection-dominated problems, obtained with a SUPG formulation, as explained in Section 2.2.1. Note that, in the right-hand side of (2.32), the term $-b \cdot \nabla u_{2n+1}$ is integrated on a fine mesh, as we expect this term to vary at the scale ε. The discretized variational formulation of (2.32) reads

$$\begin{aligned} &\text{Find } u_{2n+2}^H \in V_H \text{ such that, for any } v \in V_H, \\ &a^0(u_{2n+2}^H, v) + a_{\text{stab}}(u_{2n+2}^H, v) = F^1(v) + F_{\text{stab}}(v), \end{aligned} \tag{2.34}$$

where a_{stab} and F_{stab} are defined by (2.9) and (2.10), and

$$a^0(u,v) = \int_\Omega \alpha_{\text{spl}} \nabla u \cdot \nabla v + (b \cdot \nabla u) v,$$
$$F_1(v) = \int_\Omega f_1 v \quad \text{with} \quad f_1 = f + b \cdot \nabla (u^H_{2n} - u^H_{2n+1}).$$

The functions u_{2n+1} with odd indices are obtained using a MsFEM type approach. A natural choice for the discretization of the problem (2.33) is the MsFEM method presented in Section 2.2.1 above. The variational formulation is

$$\begin{aligned} &\text{Find } u^H_{2n+3} \in V^\varepsilon_H \text{ such that, for any } v \in V^\varepsilon_H, \\ &\int_\Omega (A^\varepsilon \nabla u^H_{2n+3}) \cdot \nabla v = \int_\Omega \alpha_{\text{spl}} \nabla u^H_{2n+2} \cdot \nabla v, \end{aligned} \tag{2.35}$$

where V^ε_H is defined by (2.21).

The termination criterion we use for the iterations is fixed as follows. Equation (2.34) is equivalent to the linear system $M_0[u^H_{2n+2}] = F^{0,H} + M_2[u^H_{2n}] - M_3[u^H_{2n+1}]$, where $[u^H_{2n}]$ is the vector representing the Finite Element function u^H_{2n} in V_H (i.e. $u^H_{2n}(x) = \sum_{i=1}^{N_{V_H}} [u^H_{2n}]_i \ \phi^0_i(x)$) and likewise for $[u^H_{2n+2}]$, while $[u^H_{2n+1}]$ is the vector representing the function u^H_{2n+1} in V^ε_H, that is $u^H_{2n+1}(x) = \sum_{i=1}^{N_{V_H}} [u^H_{2n+1}]_i \ \phi^\varepsilon_i(x)$. We stop the iterations when the iteration residual, defined as

$$\left\| M_0[u^H_{2n+2}] - (F^{0,H} + M_2[u^H_{2n+2}] - M_3[u^H_{2n+3}]) \right\|, \tag{2.36}$$

is smaller than a prescribed tolerance, here 10^{-9}.

We immediately note that, if we *assume* that u_{2n} and u_{2n+1} converge to some u_{even} and u_{odd}, respectively, then we have

$$-\alpha_{\text{spl}} \Delta u_{\text{even}} = f - b \cdot \nabla u_{\text{odd}} \quad \text{in } \Omega, \qquad u_{\text{even}} = 0 \quad \text{on } \partial\Omega, \tag{2.37}$$
$$-\operatorname{div}(A^\varepsilon \nabla u_{\text{odd}}) = -\alpha_{\text{spl}} \Delta u_{\text{even}} \quad \text{in } \Omega, \qquad u_{\text{odd}} = 0 \quad \text{on } \partial\Omega. \tag{2.38}$$

Adding (2.37) and (2.38), we get that u_{odd} is actually the solution to (2.23). A detailed analysis and a proof, under suitable assumptions, of the actual convergence of our splitting approach is provided in Section 2.3.4 below.

In theory however, there is no guarantee that, in all circumstances, the naive, fixed point iterations (2.32)-(2.33) above converge. In all the test cases presented in Section 2.4.2, the iterations indeed converge. With a view to address difficult cases where the iterations might not converge, we design and study in Section 2.3.4 a possible alternate iteration scheme, based on a damping, which, for a well adjusted damping parameter, unconditionally converges. As will be shown in Section 2.4.2, this unconditional convergence comes however at the price of yielding results that are generically less accurate and longer to obtain than when using the direct fixed point iteration, when the latter converges of course. We therefore only advocate this alternate approach in the difficult cases.

As will be seen in Section 2.4.2 below, the splitting method and the Stab-MsFEM method provide numerical solutions of approximately identical accuracy. The non-intrusive character of the splitting method is somehow balanced by its online cost which, owing to the iterations, is larger than that of the Stab-MsFEM method. This is especially true in a multi-query context and/or for problems of large sizes only amenable to iterative linear algebra solvers.

2.3 Elements of theoretical analysis

This section is devoted to the theoretical study of our four numerical approaches. Throughout the section, we mostly work in the one-dimensional setting (in Sections 2.3.1, 2.3.2 and 2.3.3), with the notable exception of the mathematical study in Section 2.3.4 of the iteration scheme (2.32)–(2.33) used in our splitting method and of an alternative unconditionally convergent iteration scheme, which is performed with all the possible generality. Some of our results were first established in the preliminary study [83].

The MsFEM method, the Stab-MsFEM method and the Adv-MsFEM method are studied, in Sections 2.3.1, 2.3.2 and 2.3.3 respectively, on the one-dimensional problem

$$-\frac{d}{dx}\left(A^\varepsilon \frac{du^\varepsilon}{dx}\right) + b\frac{du^\varepsilon}{dx} = f \quad \text{in } \Omega = (0,L), \qquad u^\varepsilon(0) = u^\varepsilon(L) = 0, \tag{2.39}$$

with a constant advection field $b \neq 0$, $f \in L^2(0,L)$ and a diffusion coefficient such that $0 < \alpha_1 \leqslant A^\varepsilon(x) \leqslant \alpha_2$ a.e. on Ω. We estimate the error in terms of ε, b, the macroscopic mesh size H and possibly the mesh size h used to solve the local problems.

For further use, we first establish the following two propositions, namely Propositions 2.7 and 2.8. The first one is a Céa-type result, which holds in any dimension.

Proposition 2.7. *For $d \geqslant 1$, let u_H be the numerical solution obtained by applying any conforming Galerkin method to problem (2.23) (on some finite dimensional space W_H). Then, if the matrix A^ε is symmetric, elliptic in the sense of (2.18) and b satisfies (2.3), we have*

$$|u^\varepsilon - u_H|_{H^1(\Omega)} \leqslant \inf_{v_H \in W_H} \left(\sqrt{\frac{\alpha_2}{\alpha_1}}\, |u^\varepsilon - v_H|_{H^1(\Omega)} + \frac{\left\| \sqrt{b^T (A^\varepsilon)^{-1} b} \right\|_{L^\infty(\Omega)}}{\sqrt{\alpha_1}}\, \|u^\varepsilon - v_H\|_{L^2(\Omega)} \right),$$

where $|v|_{H^1(\Omega)} = \|\nabla v\|_{L^2(\Omega)}$ for any $v \in H^1(\Omega)$.

Proof. We follow the proof of [83]. Using (2.3) and the Galerkin orthogonality, we have, for any $v_H \in W_H$,

$$\begin{aligned}
&\int_\Omega \left[\nabla(u^\varepsilon - u_H)\right]^T A^\varepsilon \nabla(u^\varepsilon - u_H) \\
&= a^\varepsilon(u^\varepsilon - u_H, u^\varepsilon - u_H) \\
&= a^\varepsilon(u^\varepsilon - u_H, u^\varepsilon - v_H) \\
&= \int_\Omega \left[\nabla(u^\varepsilon - v_H)\right]^T A^\varepsilon \nabla(u^\varepsilon - u_H) + \int_\Omega \left[b \cdot \nabla(u^\varepsilon - u_H)\right](u^\varepsilon - v_H),
\end{aligned} \tag{2.40}$$

where a^ε is defined by (2.25). Considering the square root $(A^\varepsilon)^{1/2}$ of the symmetric positive definite matrix $A^\varepsilon(x)$, and using the Cauchy-Schwarz inequality, we have, on the one hand,

$$\int_\Omega [\nabla(u^\varepsilon - v_H)]^T A^\varepsilon \nabla(u^\varepsilon - u_H) \leqslant \left(\int_\Omega [\nabla(u^\varepsilon - u_H)]^T A^\varepsilon \nabla(u^\varepsilon - u_H)\right)^{1/2} \left(\int_\Omega [\nabla(u^\varepsilon - v_H)]^T A^\varepsilon \nabla(u^\varepsilon - v_H)\right)^{1/2},$$

and, on the other hand,

$$\begin{aligned}
&\int_\Omega [b \cdot \nabla(u^\varepsilon - u_H)]\,(u^\varepsilon - v_H) \\
&= \int_\Omega \left[(A^\varepsilon)^{-1/2} b\right] \cdot \left[(A^\varepsilon)^{1/2} \nabla(u^\varepsilon - u_H)\right] (u^\varepsilon - v_H) \\
&\leqslant \left(\int_\Omega b^T (A^\varepsilon)^{-1} b\,(u^\varepsilon - v_H)^2\right)^{1/2} \left(\int_\Omega [\nabla(u^\varepsilon - u_H)]^T A^\varepsilon \nabla(u^\varepsilon - u_H)\right)^{1/2}.
\end{aligned}$$

Inserting these estimates in (2.40), we obtain

$$\left(\int_\Omega [\nabla(u^\varepsilon - u_H)]^T A^\varepsilon \nabla(u^\varepsilon - u_H)\right)^{1/2} \leqslant \left(\int_\Omega [\nabla(u^\varepsilon - v_H)]^T A^\varepsilon \nabla(u^\varepsilon - v_H)\right)^{1/2} + \left(\int_\Omega b^T (A^\varepsilon)^{-1} b\,(u^\varepsilon - v_H)^2\right)^{1/2}.$$

Using (2.18), we infer that, for any $v_H \in W_H$,

$$\sqrt{\alpha_1}|u^\varepsilon - u_H|_{H^1(\Omega)} \leqslant \sqrt{\alpha_2}|u^\varepsilon - v_H|_{H^1(\Omega)} + \left\| \sqrt{b^T (A^\varepsilon)^{-1} b} \right\|_{L^\infty(\Omega)} \|u^\varepsilon - v_H\|_{L^2(\Omega)}.$$

This concludes the proof of Proposition 2.7. □

Proposition 2.8. *Assume the ambient dimension is one. Consider $u^\varepsilon \in H^1_0(\Omega)$ the solution to (2.39). If $\dfrac{|b|L}{\alpha_2} \geqslant 1$, then*

$$|u^\varepsilon|_{H^1(\Omega)} \leqslant \frac{\sqrt{2\alpha_2 L}}{\alpha_1 \sqrt{|b|}} \|f\|_{L^2(\Omega)}.$$

Proof. Without loss of generality, we can assume that $b > 0$. We decompose the right-hand side into a zero mean part (considered in Step 1 of the proof) and a constant part (considered in Step 2).

Step 1. We first consider the case when the mean of f vanishes. Introduce $F(x) = \displaystyle\int_0^x f$ and note that $bu^\varepsilon - F \in H^1_0(\Omega)$, so that we can use it as test function in (2.39). This leads to

$$\int_\Omega A^\varepsilon (u^\varepsilon)'(bu^\varepsilon - F)' + b(u^\varepsilon)'(bu^\varepsilon - F) = \int_\Omega f(bu^\varepsilon - F),$$

which also reads as $\displaystyle\int_\Omega A^\varepsilon (u^\varepsilon)'(bu^\varepsilon - F)' + (bu^\varepsilon - F)'(bu^\varepsilon - F) = 0$, whence

$$\int_\Omega b A^\varepsilon (u^\varepsilon)'(u^\varepsilon)' = \int_\Omega A^\varepsilon (u^\varepsilon)' f.$$

Using the Cauchy-Schwarz inequality and the fact that $\alpha_2 \leqslant |b|L$, we get

$$|u^\varepsilon|_{H^1(\Omega)} \leqslant \frac{\alpha_2}{\alpha_1|b|}\|f\|_{L^2(\Omega)} \leqslant \frac{\sqrt{\alpha_2 L}}{\alpha_1\sqrt{|b|}}\,\|f\|_{L^2(\Omega)}. \tag{2.41}$$

Step 2. We now consider the case when f is constant. Without loss of generality, we can assume that $f \equiv 1$. The proof is based on the maximum principle. Introduce the function $v(x) = \dfrac{\alpha_2}{b}\displaystyle\int_0^x \frac{dy}{A^\varepsilon(y)} - u^\varepsilon(x)$. This function is such that

$$-\left(A^\varepsilon v'\right)' + bv' = \alpha_2 (A^\varepsilon)^{-1} - 1 \geqslant 0, \quad v(0) = 0 \ \text{ and } \ v(L) \geqslant 0.$$

According to the maximum principle [40, Theorem 8.1], we have that $v(x) \geqslant 0$ for all $x \in [0, L]$. We deduce that, for any $x \in [0, L]$, $u^\varepsilon(x) \leqslant \dfrac{\alpha_2}{|b|}\displaystyle\int_0^x \frac{dy}{A^\varepsilon(y)} \leqslant \frac{\alpha_2 L}{\alpha_1|b|}$. Taking u^ε as a test function in (2.39) and using that b is constant, we obtain $\displaystyle\int_\Omega A^\varepsilon (u^\varepsilon)'(u^\varepsilon)' = \int_\Omega u^\varepsilon$. Using the Cauchy-Schwarz inequality, we obtain $|u^\varepsilon|^2_{H^1(\Omega)} \leqslant \dfrac{\alpha_2 L^2}{\alpha_1^2|b|}$. Hence, for any constant f, we have

$$|u^\varepsilon|_{H^1(\Omega)} \leqslant \frac{\sqrt{\alpha_2 L}}{\alpha_1\sqrt{|b|}}\,\|f\|_{L^2(\Omega)}. \tag{2.42}$$

Step 3. For a general right-hand side f, we write $f = f_1 + f_2$ where f_1 is constant and the mean of f_2 vanishes. In view of (2.41) and (2.42), we see that

$$|u^\varepsilon|_{H^1(\Omega)} \leqslant \frac{\sqrt{\alpha_2 L}}{\alpha_1\sqrt{|b|}}\left(\|f_1\|_{L^2(\Omega)} + \|f_2\|_{L^2(\Omega)}\right).$$

We observe that $\|f_1\|^2_{L^2(\Omega)} + \|f_2\|^2_{L^2(\Omega)} = \|f\|^2_{L^2(\Omega)}$, due to the fact f_1 is constant and f_2 has zero mean. Hence, we have that $\|f_1\|_{L^2(\Omega)} + \|f_2\|_{L^2(\Omega)} \leqslant \sqrt{2}\|f\|_{L^2(\Omega)}$, and we thus deduce that

$$|u^\varepsilon|_{H^1(\Omega)} \leqslant \frac{\sqrt{2\alpha_2 L}}{\alpha_1\sqrt{|b|}}\,\|f\|_{L^2(\Omega)}.$$

This concludes the proof of Proposition 2.8. □

2.3.1 The MsFEM method

In the advection-dominated regime, the error bound of the MsFEM method, introduced in Section 2.2.2, is given by the following theorem.

Theorem 2.9. *Let u^ε be the solution to the one-dimensional problem (2.39) and $u^\varepsilon_H \in V^\varepsilon_H$ be its approximation by the MsFEM method. Assume that $\dfrac{|b|L}{\alpha_2} \geqslant 1$. Then the following estimate holds:*

$$|u^\varepsilon - u^\varepsilon_H|_{H^1(\Omega)} \leqslant H\left(\sqrt{\frac{\alpha_2}{\alpha_1}} + \frac{|b|H}{\alpha_1}\right)\left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right)\frac{\|f\|_{L^2(\Omega)}}{\alpha_1}. \tag{2.43}$$

Proof. We follow the proof of [83]. The error $u^\varepsilon - u_H^\varepsilon$ is decomposed in two parts:

$$e^I = u^\varepsilon - R_H u^\varepsilon, \qquad e_H^I = R_H u^\varepsilon - u_H^\varepsilon,$$

where $R_H u^\varepsilon$ is the interpolant of u^ε in V_H^ε. We have that $-\dfrac{d}{dx}\left(A^\varepsilon \dfrac{d(R_H u^\varepsilon)}{dx}\right) = 0$ in each mesh element $\mathbf{K} \in \mathcal{T}_H$ and $e_I \in H_0^1(\mathbf{K})$ for all $\mathbf{K} \in \mathcal{T}_H$. Using the variational formulation of (2.39), we get

$$\alpha_1 |e^I|^2_{H^1(\Omega)} \leqslant a^\varepsilon(e^I, e^I) = \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\mathcal{L}^\varepsilon e^I, e^I\right)_{\mathbf{K}} = \sum_{\mathbf{K}\in\mathcal{T}_H} \left(f - b(R_H u^\varepsilon)', e^I\right)_{\mathbf{K}}, \tag{2.44}$$

where $\mathcal{L}^\varepsilon v = -\dfrac{d}{dx}\left(A^\varepsilon \dfrac{dv}{dx}\right) + b\dfrac{dv}{dx}$. Now, since $e^I \in H_0^1(\mathbf{K})$, we have

$$0 = \left((e^I)', e^I\right)_{\mathbf{K}} = \left((u^\varepsilon)', e^I\right)_{\mathbf{K}} - \left((R_H u^\varepsilon)', e^I\right)_{\mathbf{K}}. \tag{2.45}$$

Using that b is constant, we deduce from (2.44) and (2.45) that

$$\begin{aligned}
\alpha_1 |e^I|^2_{H^1(\Omega)} &\leqslant \sum_{\mathbf{K}\in\mathcal{T}_H} \|f - b(u^\varepsilon)'\|_{L^2(\mathbf{K})} \|e^I\|_{L^2(\mathbf{K})} \\
&\leqslant \sum_{\mathbf{K}\in\mathcal{T}_H} H \|f - b(u^\varepsilon)'\|_{L^2(\mathbf{K})} |e^I|_{H^1(\mathbf{K})} \\
&\leqslant H \left(\|f\|_{L^2(\Omega)} + |b|\ |u^\varepsilon|_{H^1(\Omega)}\right) |e^I|_{H^1(\Omega)} \\
&\leqslant H \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right) \|f\|_{L^2(\Omega)}\ |e^I|_{H^1(\Omega)},
\end{aligned} \tag{2.46}$$

successively using the Poincaré inequality and Proposition 2.8.

On the other hand, using Proposition 2.7 and the Poincaré inequality, we have

$$|u^\varepsilon - u_H^\varepsilon|_{H^1(\Omega)} \leqslant \sqrt{\frac{\alpha_2}{\alpha_1}} |e^I|_{H^1(\Omega)} + \frac{|b|}{\alpha_1} \|e^I\|_{L^2(\Omega)} \leqslant \left(\sqrt{\frac{\alpha_2}{\alpha_1}} + \frac{|b|H}{\alpha_1}\right) |e^I|_{H^1(\Omega)}. \tag{2.47}$$

Collecting (2.46) and (2.47), we conclude the proof of Theorem 2.9. □

Remark 2.10. *Note that the estimate in Theorem 2.9 does not depend on the oscillation scale ε of A^ε, but only on the contrast α_2/α_1.*

Remark 2.11. *Assume that $A^\varepsilon(x) = \alpha$. Then the MsFEM method reduces to the classical $\mathbb{P}^1$ method and the estimate (2.43) then reads as*

$$|u - u_H|_{H^1(\Omega)} \leqslant H \left(1 + \frac{|b|H}{\alpha}\right) \left(1 + \sqrt{\frac{2L|b|}{\alpha}}\right) \frac{\|f\|_{L^2(\Omega)}}{\alpha}. \tag{2.48}$$

On the other hand, the classical numerical analysis result for that problem has been recalled in (2.6). It is $|u - u_H|_{H^1(\Omega)} \leqslant CH \left(1 + \dfrac{|b|H}{2\alpha}\right) |u|_{H^2(\Omega)}$. Since $-\alpha u'' + bu' = f$, $|u|_{H^2(\Omega)}$ may be bounded, using Proposition 2.8, as $\alpha|u|_{H^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)} + |b|\ |u|_{H^1(\Omega)} \leqslant \|f\|_{L^2(\Omega)} + \sqrt{2L|b|/\alpha}\ \|f\|_{L^2(\Omega)}$.

We therefore obtain

$$|u - u_H|_{H^1(\Omega)} \leqslant CH \left(1 + \frac{|b|H}{2\alpha}\right) \left(1 + \sqrt{\frac{2L|b|}{\alpha}}\right) \frac{\|f\|_{L^2(\Omega)}}{\alpha},$$

which exactly coincides, up to constants independent of b, α, H and f, with (2.48).

2.3.2 The Stab-MsFEM method

For the Stab-MsFEM method, also introduced in Section 2.2.2, we successively consider two cases. We first consider the "ideal" approach employing the *exact* multiscale basis functions, solution to (2.20). Next, we account for the discretization error when numerically solving the local problem (2.20).

When the discretization error is ignored, the error estimate is the following.

Theorem 2.12. *Let u^ε be the solution to the one-dimensional problem (2.39) and $u_H^\varepsilon \in V_H^\varepsilon$ be the solution to (2.26)-(2.27) with $\tau_{\mathbf{K}} = \dfrac{H}{2|b|}$. Assume that $\dfrac{|b|L}{\alpha_2} \geqslant 1$, and that we are in a advection-dominated regime, and hence that $\dfrac{|b|H}{2\alpha_1} \geqslant 1$. Then the following estimate holds:*

$$|u^\varepsilon - u_H^\varepsilon|_{H^1(\Omega)} \leqslant CH \left(1 + \sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}}\right) \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1}, \tag{2.49}$$

where C is a universal constant.

Remark 2.13. *In the case where A^ε is constant, the Stab-MsFEM method is simply the $\mathbb{P}^1$ SUPG method. In that case, we observe, as above, that the estimate of Theorem 2.12 is similar to the estimate (2.11) obtained for the $\mathbb{P}^1$ SUPG method.*

Note that the right-hand side of (2.49) is thought to be smaller than that of (2.43), as we think of $|b|H/\alpha_1$ as being large. Theorem 2.12 is actually established following Steps 2, 3 and 4 of the proof of Theorem 2.14 below.

Accounting now for the discretization error in the local problems and employing the method (2.30), we now have the following error estimate.

Theorem 2.14. *Let u^ε be the solution to the one-dimensional problem (2.39) and $u_{H,h}^\varepsilon \in V_{H,h}^\varepsilon$ be the solution to (2.30) with $\tau_{\mathbf{K}} = \dfrac{H}{2|b|}$. Assume that $A^\varepsilon \in W^{1,\infty}(\Omega)$ and that $\dfrac{|b|L}{\alpha_2} \geqslant 1$. Assume also that we are in a advection-dominated regime, and hence that $\dfrac{|b|H}{2\alpha_1} \geqslant 1$. Then the following estimate holds:*

$$\begin{aligned} |u^\varepsilon - u_{H,h}^\varepsilon|_{H^1(\Omega)} \leqslant C &\left(1 + \frac{H|b|}{\alpha_1} + \frac{H|b|}{\alpha_1}\sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}}\right) err(h) \\ &+ CH \left(1 + \sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}}\right) \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1}, \end{aligned} \tag{2.50}$$

where C only depends on Ω and where

$$err(h) = h\left(\sqrt{\frac{\alpha_2}{\alpha_1}} + \frac{|b|h}{\alpha_1}\right)\left(1 + \sqrt{\frac{2\alpha_2 L}{|b|}}\,\frac{\|(A^\varepsilon)' - b\|_{L^\infty(\Omega)}}{\alpha_1}\right)\frac{\|f\|_{L^2(\Omega)}}{\alpha_1}. \tag{2.51}$$

Proof. This proof is an adaptation of the analysis in [83]. We proceed as in the proof of (2.11) (see Appendix 2.5.1). We decompose the error $u^\varepsilon - u^\varepsilon_{H,h}$ in three parts:

$$e^I_h = u^\varepsilon - u^\varepsilon_h, \qquad e^I = u^\varepsilon_h - R_{H,h}u^\varepsilon_h, \qquad e^I_H = u^\varepsilon_{H,h} - R_{H,h}u^\varepsilon_h,$$

where u^ε_h is the Galerkin approximation of u^ε in V_h (the $\mathbb{P}^1$ finite element space associated to the *fine* mesh of size h) and $R_{H,h}u^\varepsilon_h$ is the Lagrange interpolant of u^ε_h in $V^\varepsilon_{H,h}$. We successively estimate e^I_h, e^I and e^I_H.

Step 1: estimation of e^I_h. Using Proposition 2.7 and the Poincaré inequality, we have

$$\begin{aligned}|e^I_h|_{H^1(\Omega)} &\leqslant \sqrt{\frac{\alpha_2}{\alpha_1}}|u^\varepsilon - I_h u^\varepsilon|_{H^1(\Omega)} + \frac{|b|}{\alpha_1}|u^\varepsilon - I_h u^\varepsilon|_{L^2(\Omega)}\\ &\leqslant \left(\sqrt{\frac{\alpha_2}{\alpha_1}} + \frac{|b|h}{\alpha_1}\right)|u^\varepsilon - I_h u^\varepsilon|_{H^1(\Omega)},\end{aligned} \tag{2.52}$$

where $I_h u^\varepsilon$ is the Lagrange interpolant of u^ε in V_h. Standard results on finite elements show that

$$|u^\varepsilon - I_h u^\varepsilon|_{H^1(\Omega)} \leqslant Ch|u^\varepsilon|_{H^2(\Omega)}. \tag{2.53}$$

Because of the equation, we have

$$\begin{aligned}|u^\varepsilon|_{H^2(\Omega)} &\leqslant \frac{\|f\|_{L^2(\Omega)} + \|(A^\varepsilon)' - b\|_{L^\infty(\Omega)}|u^\varepsilon|_{H^1(\Omega)}}{\alpha_1}\\ &\leqslant \left(1 + \sqrt{\frac{2\alpha_2 L}{|b|}}\,\frac{\|(A^\varepsilon)' - b\|_{L^\infty(\Omega)}}{\alpha_1}\right)\frac{\|f\|_{L^2(\Omega)}}{\alpha_1},\end{aligned} \tag{2.54}$$

where we have used Proposition 2.8. Collecting (2.52), (2.53) and (2.54), we obtain

$$|e^I_h|_{H^1(\Omega)} \leqslant C\,\mathrm{err}(h), \tag{2.55}$$

where $\mathrm{err}(h)$ is defined by (2.51).

Step 2: estimation of e^I. Using the coercivity of A^ε, we get

$$\alpha_1|e^I|^2_{H^1(\Omega)} \leqslant \int_\Omega A^\varepsilon \left(e^I\right)'\left(e^I\right)' = \int_\Omega A^\varepsilon (u^\varepsilon_h)'\left(e^I\right)' - \int_\Omega A^\varepsilon (R_{H,h}u^\varepsilon_h)'\left(e^I\right)'. \tag{2.56}$$

Using that e^I vanishes on the macroscopic mesh nodes and the variational formulation of the basis functions $\psi^{\varepsilon,h}_i$ of $V^\varepsilon_{H,h}$ on $\mathbf{K}$, we observe that

$$\int_\Omega A^\varepsilon (R_{H,h}u^\varepsilon_h)'\left(e^I\right)' = \sum_{\mathbf{K}\in\mathcal{T}_H}\int_{\mathbf{K}} A^\varepsilon (R_{H,h}u^\varepsilon_h)'\left(e^I\right)' = 0.$$

We thus deduce from (2.56) and the variational formulation satisfied by u_h^ε that

$$\alpha_1|e^I|^2_{H^1(\Omega)} \leqslant \int_\Omega A^\varepsilon (u_h^\varepsilon)' \left(e^I\right)' = \int_\Omega \left(f - b(u_h^\varepsilon)'\right) e^I = \int_\Omega \left(f - b(u^\varepsilon)' + b(e_h^I)'\right) e^I.$$

Using a Poincaré inequality for $e^I \in H^1_0(\mathbf{K})$ and Proposition 2.8, we deduce that

$$\begin{aligned}
\alpha_1|e^I|^2_{H^1(\Omega)} &\leqslant H\left(\|f - b(u^\varepsilon)'\|_{L^2(\Omega)} + |b|\ |e_h^I|_{H^1(\Omega)}\right)|e^I|_{H^1(\Omega)} \\
&\leqslant H\left[\left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right)\|f\|_{L^2(\Omega)} + |b|\ |e_h^I|_{H^1(\Omega)}\right]|e^I|_{H^1(\Omega)}
\end{aligned}$$

and thus

$$|e^I|_{H^1(\Omega)} \leqslant H\left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1}\right)\frac{\|f\|_{L^2(\Omega)}}{\alpha_1} + \frac{H|b|}{\alpha_1}|e_h^I|_{H^1(\Omega)}. \tag{2.57}$$

Step 3: estimation of e_H^I. We write

$$\begin{aligned}
& \alpha_1|e_H^I|^2_{H^1(\Omega)} + a_{\rm upw}(e_H^I, e_H^I) \\
\leqslant\ & a^\varepsilon(e_H^I, e_H^I) + a_{\rm upw}(e_H^I, e_H^I) \\
=\ & a^\varepsilon(u_{H,h}^\varepsilon, e_H^I) + a_{\rm upw}(u_{H,h}^\varepsilon, e_H^I) - a^\varepsilon(R_{H,h}u_h^\varepsilon, e_H^I) - a_{\rm upw}(R_{H,h}u_h^\varepsilon, e_H^I) \\
=\ & F(e_H^I) + F_{\rm stab}(e_H^I) - a^\varepsilon(R_{H,h}u_h^\varepsilon, e_H^I) - a_{\rm upw}(R_{H,h}u_h^\varepsilon, e_H^I) \\
=\ & a^\varepsilon(u_h^\varepsilon, e_H^I) + F_{\rm stab}(e_H^I) - a^\varepsilon(R_{H,h}u_h^\varepsilon, e_H^I) - a_{\rm upw}(R_{H,h}u_h^\varepsilon, e_H^I),
\end{aligned}$$

making use of the variational formulation satisfied by $u_{H,h}^\varepsilon$ and u_h^ε, respectively. Using that $e^I = u_h^\varepsilon - R_{H,h}u_h^\varepsilon$, we next obtain

$$\begin{aligned}
& \alpha_1|e_H^I|^2_{H^1(\Omega)} + a_{\rm upw}(e_H^I, e_H^I) \\
\leqslant\ & a^\varepsilon(e^I, e_H^I) + F_{\rm stab}(e_H^I) + a_{\rm upw}(e^I, e_H^I) - a_{\rm upw}(u_h^\varepsilon, e_H^I) \\
=\ & \int_\Omega \left(A^\varepsilon \left(e^I\right)' \left(e_H^I\right)' - b\left(e_H^I\right)' e^I\right) + \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} f, b(e_H^I)'\right)_{\mathbf{K}} \\
& + \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} b\left(e^I\right)', b\left(e_H^I\right)'\right)_{\mathbf{K}} - \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} b(u_h^\varepsilon)', b\left(e_H^I\right)'\right)_{\mathbf{K}} \\
=\ & \int_\Omega \left(A^\varepsilon \left(e^I\right)' \left(e_H^I\right)' - b\left(e_H^I\right)' e^I\right) + \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} \left(f - b(u_h^\varepsilon)'\right), b(e_H^I)'\right)_{\mathbf{K}} \\
& + \sum_{\mathbf{K}\in\mathcal{T}_H} \left(\tau_{\mathbf{K}} b\left(e^I\right)', b\left(e_H^I\right)'\right)_{\mathbf{K}}.
\end{aligned} \tag{2.58}$$

We now successively estimate each term of the right-hand side of (2.58). For the first part of the first term, we have

$$\left|\int_\Omega A^\varepsilon \left(e^I\right)' \left(e_H^I\right)'\right| \leqslant \int_\Omega \alpha_2 \left|\left(e^I\right)' \left(e_H^I\right)'\right| \leqslant \frac{\alpha_1}{4}|e_H^I|^2_{H^1(\Omega)} + \frac{\alpha_2^2}{\alpha_1}|e^I|^2_{H^1(\Omega)}.$$

For the second part of the first term, we obtain

$$\begin{aligned}
-\int_\Omega b\left(e_H^I\right)' e^I &\leqslant \frac{1}{4}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} + \sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{-1/2} e^I\|^2_{L^2(\mathbf{K})} \\
&\leqslant \frac{1}{4}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{2|b|}{H}H^2|e^I|^2_{H^1(\mathbf{K})} \\
&\leqslant \frac{1}{4}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} + 2|b|H|e^I|^2_{H^1(\Omega)},
\end{aligned}$$

where, in the second line, we have used the value of $\tau_{\mathbf{K}}$ and a Poincaré inequality.

We bound the second term as follows:

$$\begin{aligned}
&\sum_{\mathbf{K}\in\mathcal{T}_H}\left(\tau_{\mathbf{K}}\left(f - b(u_h^\varepsilon)'\right), b(e_H^I)'\right)_{\mathbf{K}} \\
\leqslant\ & \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{2}\|\tau_{\mathbf{K}}^{1/2}\left(f - b(u_h^\varepsilon)'\right)\|^2_{L^2(\mathbf{K})} + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{2}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} \\
=\ & \frac{H}{4|b|}\|f - b(u_h^\varepsilon)'\|^2_{L^2(\Omega)} + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{2}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} \\
\leqslant\ & \frac{H}{2|b|}\left(\|f\|^2_{L^2(\Omega)} + \|b(u_h^\varepsilon)'\|^2_{L^2(\Omega)}\right) + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{2}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} \\
\leqslant\ & \frac{H^2}{4\alpha_1}\left(\|f\|^2_{L^2(\Omega)} + \|b(u_h^\varepsilon)'\|^2_{L^2(\Omega)}\right) + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{2}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})},
\end{aligned}$$

where we have used the fact that $\dfrac{|b|H}{2\alpha_1} \geqslant 1$ in the last line, and that $\tau_{\mathbf{K}} = \dfrac{H}{2|b|}$.

For the third term, we get, using the expression of $\tau_{\mathbf{K}}$,

$$\begin{aligned}
&\sum_{\mathbf{K}\in\mathcal{T}_H}\left(\tau_{\mathbf{K}} b\left(e^I\right)', b\left(e_H^I\right)'\right)_{\mathbf{K}} \\
\leqslant\ & \sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2} b\left(e^I\right)'\|^2_{L^2(\mathbf{K})} + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{4}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})} \\
=\ & \frac{|b|H}{2}|e^I|^2_{H^1(\Omega)} + \sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{4}\|\tau_{\mathbf{K}}^{1/2} b\left(e_H^I\right)'\|^2_{L^2(\mathbf{K})}.
\end{aligned}$$

Collecting the terms, we deduce from (2.58) that

$$\begin{aligned}
\alpha_1|e_H^I|^2_{H^1(\Omega)} \leqslant \frac{\alpha_1}{4}|e_H^I|^2_{H^1(\Omega)} + \left(\frac{\alpha_2^2}{\alpha_1} + 2|b|H + \frac{|b|H}{2}\right)|e^I|^2_{H^1(\Omega)} \\
+ \frac{H^2}{4\alpha_1}\|f\|^2_{L^2(\Omega)} + \frac{H^2|b|^2}{4\alpha_1}|u_h^\varepsilon|^2_{H^1(\Omega)},
\end{aligned}$$

which yields

$$|e_H^I|_{H^1(\Omega)} \leqslant C\left[\left(\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}\right)|e^I|^2_{H^1(\Omega)} + \frac{H^2}{\alpha_1^2}\|f\|^2_{L^2(\Omega)} + \frac{H^2|b|^2}{\alpha_1^2}|u_h^\varepsilon|^2_{H^1(\Omega)}\right]^{1/2}, \tag{2.59}$$

where C is a universal constant. Using Proposition 2.8, we have

$$|u_h^\varepsilon|_{H^1(\Omega)} \leqslant |u^\varepsilon|_{H^1(\Omega)} + |e_h^I|_{H^1(\Omega)} \leqslant \frac{\sqrt{2\alpha_2 L}}{\alpha_1\sqrt{|b|}}\,\|f\|_{L^2(\Omega)} + |e_h^I|_{H^1(\Omega)},$$

and we thus deduce from (2.59) that

$$|e_H^I|_{H^1(\Omega)} \leqslant C \left[\sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}}\ |e^I|_{H^1(\Omega)} \right. \\ \left. + H \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1} + \frac{H|b|}{\alpha_1} |e_h^I|_{H^1(\Omega)} \right]. \quad (2.60)$$

Step 4: conclusion. Successively using the triangle inequality, (2.60), (2.57) and (2.55), we obtain

$$\begin{aligned}
|u^\varepsilon - u^\varepsilon_{H,h}|_{H^1(\Omega)} &\leqslant |e_H^I|_{H^1(\Omega)} + |e^I|_{H^1(\Omega)} + |e_h^I|_{H^1(\Omega)} \\
&\leqslant \left(1 + C\sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}} \right) |e^I|_{H^1(\Omega)} + \left(1 + \frac{CH|b|}{\alpha_1} \right) |e_h^I|_{H^1(\Omega)} \\
&\quad + CH \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1} \\
&\leqslant \left[1 + \frac{CH|b|}{\alpha_1} + \frac{H|b|}{\alpha_1} \left(1 + C\sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}} \right) \right] |e_h^I|_{H^1(\Omega)} \\
&\quad + \left(1 + C\sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}} \right) H \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1} \\
&\quad + CH \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1} \\
&\leqslant C \left[1 + \frac{H|b|}{\alpha_1} + \frac{H|b|}{\alpha_1} \sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}} \right] \mathrm{err}(h) \\
&\quad + CH \left(1 + \sqrt{\frac{\alpha_2^2}{\alpha_1^2} + \frac{|b|H}{\alpha_1}} \right) \left(1 + \frac{\sqrt{2\alpha_2 L|b|}}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1}.
\end{aligned}$$

This concludes the proof of Theorem 2.14. □

2.3.3 The Adv-MsFEM method

The error bound of the Adv-MsFEM method (introduced in Section 2.2.2) is given by the following theorem.

Theorem 2.15. *Let u^ε be the solution to the one-dimensional problem (2.39) and $u_H^\varepsilon \in V_H^{\varepsilon,Adv}$ be the solution to the Adv-MsFEM method. The following estimate holds:*

$$|u^\varepsilon - u_H^\varepsilon|_{H^1(\Omega)} \leqslant H \left(\sqrt{\frac{\alpha_2}{\alpha_1}} + \frac{|b|H}{\alpha_1} \right) \frac{\|f\|_{L^2(\Omega)}}{\alpha_1}.$$

The proof of this theorem follows the same pattern as the proof of Theorem 2.9. We therefore skip it.

2.3.4 Splitting approach

We now turn to the splitting method introduced in Section 2.2.2. In contrast to Sections 2.3.1, 2.3.2 and 2.3.3, we do not restrict ourselves to the one-dimensional setting. In what follows, we denote C_Ω the Poincaré constant of Ω as defined by $\|\varphi\|_{L^2(\Omega)} \leqslant C_\Omega |\varphi|_{H^1(\Omega)}$ for any $\varphi \in H_0^1(\Omega)$.

The method (2.32)–(2.33)

Lemma 2.16. *Consider the splitting method* (2.32)–(2.33). *If*

$$\frac{C_\Omega \|b\|_{L^\infty(\Omega)}}{\alpha_1} \left(\frac{\|A^\varepsilon - \alpha_{\rm spl} Id\|_{L^\infty(\Omega)}}{\alpha_{\rm spl}} \right) < 1, \tag{2.61}$$

then u_{2n+1} *converges in* $H_0^1(\Omega)$ *to* u^ε *solution to* (2.23).

Proof. Let $\widetilde{u}_n = u_{n+2} - u_n$. We reformulate the system (2.32)–(2.33) as

$$\begin{cases} -\alpha_{\rm spl}\Delta \widetilde{u}_{2n+2} + b\cdot\nabla \widetilde{u}_{2n+2} = b\cdot\nabla(\widetilde{u}_{2n} - \widetilde{u}_{2n+1}) & \text{in } \Omega, \\ \widetilde{u}_{2n+2} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.62}$$

$$\begin{cases} -\operatorname{div}\,(A^\varepsilon \nabla \widetilde{u}_{2n+3}) = -\alpha_{\rm spl}\Delta \widetilde{u}_{2n+2} & \text{in } \Omega, \\ \widetilde{u}_{2n+3} = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{2.63}$$

Using the variational formulations of (2.62) and (2.63), we have

$$\alpha_{\rm spl}|\widetilde{u}_{2n+2}|_{H^1(\Omega)} \leqslant C_\Omega \|b\|_{L^\infty(\Omega)} |\widetilde{u}_{2n} - \widetilde{u}_{2n+1}|_{H^1(\Omega)}, \tag{2.64}$$

$$\alpha_1 |\widetilde{u}_{2n+1}|_{H^1(\Omega)} \leqslant \alpha_{\rm spl} |\widetilde{u}_{2n}|_{H^1(\Omega)}, \tag{2.65}$$

where we have used (2.3) and (2.18). Letting $w_n = \widetilde{u}_{2n+1} - \widetilde{u}_{2n}$, we have

$$-\operatorname{div}\,(A^\varepsilon \nabla w_n) = -\alpha_{\rm spl}\Delta \widetilde{u}_{2n} + \operatorname{div}\,(A^\varepsilon \nabla \widetilde{u}_{2n}) \ \text{ in } \Omega, \qquad w_n = 0 \ \text{ on } \partial\Omega.$$

We deduce that

$$\alpha_1 |w_n|_{H^1(\Omega)} \leqslant \|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)} |\widetilde{u}_{2n}|_{H^1(\Omega)}. \tag{2.66}$$

Collecting (2.64) and (2.66), we get

$$|\widetilde{u}_{2n+2}|_{H^1(\Omega)} \leqslant \rho^{1+n} |\widetilde{u}_0|_{H^1(\Omega)},$$

where

$$\rho = \frac{C_\Omega \|b\|_{L^\infty(\Omega)}}{\alpha_1} \left(\frac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{\alpha_{\rm spl}} \right).$$

Because of (2.61), the sequence u_{2n} therefore converges in $H_0^1(\Omega)$ to some $u_{\rm even}$. In view of (2.65), the sequence u_{2n+1} also converges in $H_0^1(\Omega)$ to some $u_{\rm odd}$. Passing to the limit $n \to \infty$ in (2.32) and (2.33), we obtain that $u_{\rm even}$ and $u_{\rm odd}$ are the solutions to

$$-\alpha_{\rm spl}\Delta u_{\rm even} = f - b\cdot\nabla u_{\rm odd} \quad \text{in } \Omega, \quad u_{\rm even} = 0 \quad \text{on } \partial\Omega, \tag{2.67}$$

$$-\operatorname{div}\,(A^\varepsilon \nabla u_{\rm odd}) = -\alpha_{\rm spl}\Delta u_{\rm even} \quad \text{in } \Omega, \quad u_{\rm odd} = 0 \quad \text{on } \partial\Omega. \tag{2.68}$$

Adding (2.67) and (2.68), we get that $u_{\rm odd}$ is actually the solution to (2.23). □

There are unfortunately simple situations where (2.61) is not satisfied, whatever the choice of $\alpha_{\rm spl}$. Consider for instance the one-dimensional setting where A^ε is continuous. Then $\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)} = \max(|a_+ - \alpha_{\rm spl}|, |a_- - \alpha_{\rm spl}|)$ where $a_- = \inf_\Omega A^\varepsilon$ and $a_+ = \sup_\Omega A^\varepsilon$. We observe

that

$$\rho \geqslant \rho_-, \tag{2.69}$$

where $\rho_- = \dfrac{C_\Omega \|b\|_{L^\infty(\Omega)}}{\alpha_1} \dfrac{a_+ - a_-}{a_+ + a_-}$. If $\rho_- > 1$, then, for any $\alpha_{\rm spl} > 0$, condition (2.61) is not satisfied. Of course, (2.61) is only a sufficient, and not a necessary condition for the convergence of the iterations. In most cases, and even in some cases when (2.61) is not satisfied, the splitting method (2.32)–(2.33) converges, see Section 2.4.2. In some cases, it does not. Lemma 2.17 below describes such a convergence failure, for a one-dimensional example that can be easily extended to higher dimensional settings using tensor products.

Lemma 2.17. *Assume that $\Omega = (0,1)$, that $A^\varepsilon \equiv \alpha^\star$, that the initial guess for (2.32)–(2.33) is $u_0 = \cos(2\pi x) - 1$ and u_1 the solution to (2.33) with u_0 in the right-hand side. Take $\alpha^\star$ and α_{spl} such that*

$$\frac{b}{\alpha_{spl}} < \frac{b}{2\alpha^\star} - 2\pi^2 \frac{\alpha^\star}{b}. \tag{2.70}$$

Then the sequences $(u_{2n})_{n\in\mathbb{N}}$ and $(u_{2n+1})_{n\in\mathbb{N}}$ do not converge in $H^1_0(\Omega)$.

Proof. We take $f = 0$. Then equation (2.33) reads as $u_{2n+1} = (\alpha_{\rm spl}/\alpha^\star)\, u_{2n}$, so (2.32) reduces to

$$-(u_{2n+2})'' + \frac{b}{\alpha_{\rm spl}}(u_{2n+2})' = \lambda (u_{2n})' \text{ in } (0,1), \quad u_{2n+2}(0) = u_{2n+2}(1) = 0,$$

where $\lambda = \dfrac{b}{\alpha_{\rm spl}} \left(1 - \dfrac{\alpha_{\rm spl}}{\alpha^\star}\right)$. A simple calculation shows that, for any $n \in \mathbb{N}$, $(u_{2n})' = c_n \cos(2\pi x) + s_n \sin(2\pi x)$, with $[c_n, s_n]^T = (-1)^n \lambda^n A^n [0, -2\pi]^T$ and

$$A = \left[\left(\frac{b}{\alpha_{\rm spl}} \right)^2 + 4\pi^2 \right]^{-1} \begin{pmatrix} -b/\alpha_{\rm spl} & -2\pi \\ 2\pi & -b/\alpha_{\rm spl} \end{pmatrix}.$$

If $\rho(\lambda A) = \dfrac{|\lambda|}{\sqrt{\left(b/\alpha_{\rm spl}\right)^2 + 4\pi^2}} > 1$, a condition which is equivalent to (2.70), then the sequence $(u_{2n})_{n\in\mathbb{N}}$ does not converge. □

An alternate splitting method

We now present an alternate splitting method, which includes some element of damping, and which, when the damping parameter (denoted by β) is suitably adjusted, unconditionally converges. We emphasize however that we have observed in our numerical tests that the convergence of this alternate approach, although guaranteed theoretically, is much slower than that of the method (2.32)–(2.33). See Figure 2.3 below.

The iterates u_{2n+2} and u_{2n+3} are now defined by

$$\begin{cases} -(\beta + \alpha_{\rm spl})\Delta u_{2n+2} + b \cdot \nabla u_{2n+2} = f + b \cdot \nabla (u_{2n} - u_{2n+1}) - \beta \Delta u_{2n+1} & \text{in } \Omega, \\ u_{2n+2} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.71}$$

$$\begin{cases} -\operatorname{div}\left((\beta \mathrm{Id} + A^\varepsilon)\nabla u_{2n+3}\right) = -(\beta + \alpha_{\rm spl})\Delta u_{2n+2} & \text{in } \Omega, \\ u_{2n+3} = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{2.72}$$

with $\alpha_{\rm spl} > 0$ and $\beta \geqslant 0$. Of course, $\beta = 0$ yields (2.32)–(2.33).

The convergence of (2.71)–(2.72) is established in the following lemma, in the infinite dimensional setting. The discretized, finite dimensional version will be studied next.

Lemma 2.18. *Choose*

$$\beta = \underset{x\geqslant 0}{\rm argmin} \left(\frac{C_\Omega \|b\|_{L^\infty(\Omega)}}{x+\alpha_{spl}} \frac{\|A^\varepsilon - \alpha_{spl} Id\|_{L^\infty(\Omega)}}{x+\alpha_1} + \frac{x}{x+\alpha_1} \right), \tag{2.73}$$

where α_1 is such that (2.18) holds. Then u_{2n+1} converges in $H^1_0(\Omega)$ to u^ε solution to (2.23).

Proof. Following the arguments of the proof of Lemma 2.16, we have

$$|\widetilde{u}_{2n+2}|_{H^1(\Omega)} \leqslant \frac{C_\Omega \|b\|_{L^\infty(\Omega)}}{\beta+\alpha_{\rm spl}} |\widetilde{u}_{2n} - \widetilde{u}_{2n+1}|_{H^1(\Omega)} + \frac{\beta}{\beta+\alpha_{\rm spl}} |\widetilde{u}_{2n+1}|_{H^1(\Omega)}, \tag{2.74}$$

$$|\widetilde{u}_{2n+1}|_{H^1(\Omega)} \leqslant \frac{\beta+\alpha_{\rm spl}}{\beta+\alpha_1} |\widetilde{u}_{2n}|_{H^1(\Omega)}, \tag{2.75}$$

$$|w_n|_{H^1(\Omega)} \leqslant \frac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{\beta+\alpha_1} |\widetilde{u}_{2n}|_{H^1(\Omega)}, \tag{2.76}$$

where we recall that $\widetilde{u}_n = u_{n+2} - u_n$ and $w_n = \widetilde{u}_{2n+1} - \widetilde{u}_{2n}$.

Collecting (2.74), (2.75) and (2.76), we have

$$|\widetilde{u}_{2n+2}|_{H^1(\Omega)} \leqslant \rho |\widetilde{u}_{2n}|_{H^1(\Omega)},$$

where $\rho = g(\beta)$ and where the function g is defined by

$$g(x) = \frac{C_\Omega \|b\|_{L^\infty(\Omega)}}{x+\alpha_{\rm spl}} \frac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{x+\alpha_1} + \frac{x}{x+\alpha_1}.$$

We next observe that $g(x) = 1 - \dfrac{\alpha_1}{x} + O\left(\dfrac{1}{x^2}\right)$. Since $\alpha_1 > 0$, this implies that $\min\limits_{x\geqslant 0} g(x) < 1$. In view of (2.73), we have $\rho = g(\beta) = \min\limits_{x\geqslant 0} g(x) < 1$. We next conclude the proof mimicking the argument in the proof of Lemma 2.16. □

We now consider the discrete case. Given the approximations u^H_{2n} and u^H_{2n+1}, we define u^H_{2n+2} and u^H_{2n+3} as follows. First, we discretize (2.71) on a coarse mesh and use the SUPG terms to stabilize the approach. We hence define u^H_{2n+2} by the following variational formulation:

$$\begin{aligned} &\text{Find } u^H_{2n+2} \in V_H \text{ such that, for any } v \in V_H, \\ &a_1(u^H_{2n+2}, v) + a_{\rm conv}(u^H_{2n+2}, v) = \widetilde{F}^1(v) + \widetilde{F}_{\rm stab}(v) + a_{\rm conv}(P_{V^\varepsilon_H}(u^H_{2n}), v), \end{aligned} \tag{2.77}$$

where we recall that V_H is the $\mathbb{P}^1$ finite element space, and where

$$a_1(u,v) = \int_\Omega (\beta+\alpha_{\rm spl}) \nabla u \cdot \nabla v, \tag{2.78}$$

$$a_{\rm conv}(u,v) = \int_\Omega (b\cdot\nabla u)\, v + \sum_{\mathbf{K}\in\mathcal{T}_H} (\tau_{\mathbf{K}} b\cdot\nabla u, b\cdot\nabla v)_{L^2(\mathbf{K})}, \tag{2.79}$$

$$\widetilde{F}^1(v) = \int_\Omega (f - b\cdot\nabla u^H_{2n+1}) v + \int_\Omega \beta \nabla u^H_{2n+1}\cdot\nabla v,$$

$$\widetilde{F}_{\rm stab}(v) = \sum_{\mathbf{K}\in\mathcal{T}_H} (\tau_{\mathbf{K}}(f - b\cdot\nabla u^H_{2n+1}), b\cdot\nabla v)_{L^2(\mathbf{K})}.$$

In (2.77), $P_{V_H^\varepsilon}$ is the projector on the space V_H^ε defined as follows. For any $v \in H_0^1(\Omega)$, we define $P_{V_H^\varepsilon}(v) \in V_H^\varepsilon$ by

$$\forall w \in V_H^\varepsilon, \quad a_1\Big(P_{V_H^\varepsilon}(v), w\Big) = a_1(v, w). \tag{2.80}$$

Second, we discretize (2.72) using the MsFEM approach: we define u_{2n+3}^H by the following variational formulation:

$$\text{Find } u_{2n+3}^H \in V_H^\varepsilon \text{ such that, for any } w \in V_H^\varepsilon, \quad a_2(u_{2n+3}^H, w) = a_1(u_{2n+2}^H, w), \tag{2.81}$$

where

$$a_2(u, v) = \int_\Omega (\nabla v)^T (\beta \mathrm{Id} + A^\varepsilon) \nabla u. \tag{2.82}$$

Remark 2.19. *Three remarks on (2.77) are in order. First, the term $-\beta\Delta u_{2n+1}^H$ is absent from $\widetilde{F}_{stab}$ only because, as we use a $\mathbb{P}^1$ approach, that term identically vanishes in each element $\mathbf{K}$. Second, as already mentioned in Section 2.2.2, the computation of $\widetilde{F}^1(v)$ needs to be performed on a fine mesh, since u_{2n+1}^H belongs to the MsFEM space V_H^ε. Third, the introduction of the projector $P_{V_H^\varepsilon}$ in (2.77) is motivated by the need to guarantee the convergence of the iterations (2.77)–(2.81) to an accurate approximation of the solution u^ε to the reference problem (2.23). Lemma 2.20 below will clarify and establish this convergence. Note that, instead of (2.80), we could as well have defined $P_{V_H^\varepsilon}(v) \in V_H^\varepsilon$, for any $v \in H_0^1(\Omega)$, by the relation $a_2\Big(P_{V_H^\varepsilon}(v), w\Big) = a_2(v, w)$ for any $w \in V_H^\varepsilon$.*

We establish in Appendix 2.5.3 below the convergence of (2.77)–(2.81). Formally passing to the limit $n \to \infty$ in (2.77)–(2.81), we observe that, if (u_{2n}^H, u_{2n+1}^H) converges to some $(u_{\text{even}}^H, u_{\text{odd}}^H) \in V_H \times V_H^\varepsilon$, then $(u_{\text{even}}^H, u_{\text{odd}}^H)$ satisfies

$$\begin{aligned}\forall v \in V_H, \quad & a_1(u_{\text{even}}^H, v) + a_{\text{conv}}(u_{\text{even}}^H, v) \\ & = \widetilde{F}^1(v; u_{\text{odd}}^H) + \widetilde{F}_{\text{stab}}(v; u_{\text{odd}}^H) + a_{\text{conv}}(P_{V_H^\varepsilon}(u_{\text{even}}^H), v),\end{aligned} \tag{2.83}$$

and

$$\forall w \in V_H^\varepsilon, \quad a_2(u_{\text{odd}}^H, w) = a_1(u_{\text{even}}^H, w), \tag{2.84}$$

with $\widetilde{F}^1(v; u_{\text{odd}}^H) = \displaystyle\int_\Omega (f - b \cdot \nabla u_{\text{odd}}^H) v + \int_\Omega \beta \nabla u_{\text{odd}}^H \cdot \nabla v$ and $\widetilde{F}_{\text{stab}}(v; u_{\text{odd}}^H) = \displaystyle\sum_{\mathbf{K} \in \mathcal{T}_H} (\tau_{\mathbf{K}}(f - b \cdot \nabla u_{\text{odd}}^H), b \cdot \nabla v)_{L^2(\mathbf{K})}$. This convergence is rigorously stated in Lemma 2.20 below, as well as the convergence when $H \to 0$.

Lemma 2.20. *Suppose that we set the stabilization parameter to*

$$\tau_{\mathbf{K}}(x) = \frac{H}{2|b(x)|} \quad \textit{for any } \mathbf{K} \in \mathcal{T}_H.$$

Choose

$$\beta = \operatorname*{argmin}_{x \geqslant 0} \left[\left(C_\Omega + \frac{H}{2} \right) \frac{\|b\|_{L^\infty(\Omega)}}{x + \alpha_{spl}} \frac{\|A^\varepsilon - \alpha_{spl} Id\|_{L^\infty(\Omega)}}{x + \alpha_1} + \frac{x}{x + \alpha_1} \right] \tag{2.85}$$

where α_1 is such that (2.18) holds.

Then, when $n \to \infty$, (u_{2n}^H, u_{2n+1}^H) *converges in* $H_0^1(\Omega) \times H_0^1(\Omega)$ *to* $(u_{even}^H, u_{odd}^H) \in V_H \times V_H^\varepsilon$ *solutions to the variational formulation* (2.83)–(2.84).

Assume in addition that $A^\varepsilon \in W^{1,\infty}(\Omega)$ *and that*

$$\alpha_{spl} < 2\alpha_1. \tag{2.86}$$

Then, when $H \to 0$, u_{odd}^H *converges in* $H_0^1(\Omega)$ *to* u^ε *solution to* (2.23).

The proof of Lemma 2.20 is postponed until Appendix 2.5.3.

2.4 Numerical simulations

In this section, we present and discuss our numerical experiments. They have all been performed using FreeFem++ [43]. Our aim is to compare the four approaches of Section 2.2.2. Section 2.4.1 collects some preliminary material. Then we assess the accuracy and computational cost of our four numerical methods in Sections 2.4.2 and 2.4.3, respectively.

2.4.1 Test case

We work on the domain $\Omega = (0,1)^2$, discretized with a uniform coarse mesh $\mathcal{T}_H$ of size H. Let V_H be the finite dimensional vector space (2.19) associated to the classical $\mathbb{P}^1$ discretization. In (2.23), we set $b = (1,1)^T$, $f = 1$ and

$$A^\varepsilon(x_1, x_2) = \alpha \left(1 + \delta \cos\left(\frac{2\pi}{\varepsilon} x_1\right)\right) \mathrm{Id}_2, \quad \text{with } \alpha,\, \delta > 0.$$

We recall that the advection-dominated regime is defined by the condition $\mathrm{Pe}\, H > 1$, where we define here the global Péclet number Pe of problem (2.23) by (2.7). Here this regime corresponds to

$$\alpha < \frac{H}{2}. \tag{2.87}$$

In this regime, the solution exhibits the boundary layer $\Omega_{\text{layer}} = \Big((0,1) \times (1 - \delta_{\text{layer}}, 1)\Big) \cup \Big((1 - \delta_{\text{layer}}, 1) \times (0,1)\Big)$, represented on Figure 2.2, of approximate width $\delta_{\text{layer}} = \dfrac{1}{\mathrm{Pe}} \log(\mathrm{Pe})$.

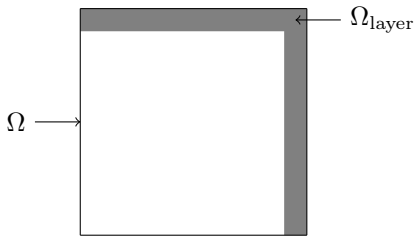

Figure 2.2 The domain Ω and the boundary layer Ω_{layer}

We choose for the splitting method the value $\alpha_{\text{spl}} = \alpha$. Motivated by the one-dimensional formula (2.16), the stabilization parameter $\tau_{\mathbf{K}}$ is chosen as

$$\tau_{\mathbf{K}}(x) = \frac{|\mathbf{K}|}{2|b(x)|} \Big(\coth\left(\mathrm{Pe}_{\mathbf{K}}(x)\right) - \left(\mathrm{Pe}_{\mathbf{K}}(x)\right)^{-1}\Big) \quad \text{for all } \mathbf{K} \in \mathcal{T}_H,$$

where $\mathrm{Pe}_{\mathbf{K}}(x) = \dfrac{|b(x)|\,H}{2\alpha}$.

Evaluation of the accuracy

Let $\mathcal{T}_h$ be a uniform fine mesh of Ω of size h such that $\mathcal{T}_h$ is a refinement of $\mathcal{T}_H$. The reference solution u_{ref} is obtained by the standard $\mathbb{P}^1$ finite element discretization on $\mathcal{T}_h$ where h is such that

$$h \leqslant \frac{1}{16}\min(\varepsilon, \delta_{\mathrm{layer}}) \quad \text{and } \mathrm{Pe}\, h \leqslant \frac{1}{4\sqrt{2}} < 1.$$

This condition ensures that the fine mesh can both resolve the oscillations throughout the domain at scale ε of the solution and the details within the boundary layer. It also ensures that, at scale h, the problem is not advection-dominated. This fine mesh is also that on which the local problems are solved, in order to determine the MsFEM basis functions.

In the sequel, the accuracies of the methods are compared using the following relative errors: $e_{H^1_{\mathrm{in}}}(u_1) = \frac{\|u_1 - u_{\mathrm{ref}}\|_{H^1(\Omega_{\mathrm{layer}})}}{\|u_{\mathrm{ref}}\|_{H^1(\Omega)}}$ inside the boundary layer, and likewise $e_{H^1_{\mathrm{out}}}(u_1) = \frac{\|u_1 - u_{\mathrm{ref}}\|_{H^1(\Omega\setminus\Omega_{\mathrm{layer}})}}{\|u_{\mathrm{ref}}\|_{H^1(\Omega)}}$ outside that layer, and, in the whole domain, $e_{L^p}(u_1) = \frac{\|u_1 - u_{\mathrm{ref}}\|_{L^p(\Omega)}}{\|u_{\mathrm{ref}}\|_{L^p(\Omega)}}$ for $p = 2$ or $p = \infty$, and $e_{H^1}(u_1) = \frac{\|u_1 - u_{\mathrm{ref}}\|_{H^1(\Omega)}}{\|u_{\mathrm{ref}}\|_{H^1(\Omega)}}$. All these relative errors are computed on the fine mesh $\mathcal{T}_h$.

Evaluation of the computational costs

The sizes of the local and global problems in the test cases we consider in Section 2.4.2 are sufficiently small to allow for the use of direct linear solvers (in our case, the UMFPACK library). This clearly favors the splitting method as opposed to the other approaches, since that method is potentially the most expensive one of all four in its online stage. The factorization of the stiffness matrices is performed once and for all in the offline stage and is repeatedly used in the iterative process during the online stage. When, for problems of larger sizes, iterative linear solvers are in order, the online cost of the splitting method correspondingly increases. To evaluate this marginal cost, we have also performed tests using iterative solvers *as if* the problem sizes were large. We have used either, for non-symmetric matrices, the GMRES solver and a value of the stopping criterion equal to 10^{-11}, or, for symmetric matrices, the conjugate gradient method with a stopping criterion at 10^{-20}. Both solvers are used with a simple diagonal preconditioner. The computations have all been performed on a Intel®Xeon®Processor E5-2667 v2. The specific function used to measure the CPU time is clock_gettime() with the clock CLOCK_PROCESS_CPUTIME_ID.

2.4.2 Accuracies

Reference test

We first consider problem (2.23), with the choices of A^ε and b described in Section 2.4.1, and the parameters $\alpha = 1/128$, $\delta = 0.5$ and $H = 1/16$. Since $\mathrm{Pe}\, H = 4$, the problem is expected to be advection-dominated and, for $\varepsilon = 1/64$, multiscale.

In order to practically check that the dominating advection is a challenge to standard approaches, we temporarily set ε to one, and compare the results obtained by the $\mathbb{P}^1$ method and the $\mathbb{P}^1$ Upwind method [50]. Table 2.2 shows that, outside the boundary layer, the relative H^1 error of the $\mathbb{P}^1$ method is approximately 20 times as large as the error of the $\mathbb{P}^1$ Upwind method. This confirms the advection-dominated regime.

	e_{L^2}	e_{L^∞}	e_{H^1}	$e_{H^1_{\mathbf{in}}}$	$e_{H^1_{\mathbf{out}}}$
$\mathbb{P}^1$	0.24	0.69	1.08	0.90	0.58
$\mathbb{P}^1$ Upwind	0.21	0.57	0.85	0.84	0.03

Table 2.2 Relative errors in the single-scale case ($\alpha = 1/128$, $\delta = 0.5$, $\varepsilon = 1$ and $H = 1/16$)

Likewise, in order to practically demonstrate the relevance of accounting for the small scale, we reinstate $\varepsilon = 1/64$ and display on Table 2.3 the relative errors for the different methods. We indeed observe that, outside the boundary layer, the relative H^1 error of the $\mathbb{P}^1$ Upwind method is about three times as large as the error of the Stab-MsFEM method.

We now compare the accuracies of the methods. The results are shown on Table 2.3. We observe that all methods have an outrageously large error within the boundary layer (close to a hundred percent). The only exception to this is discussed in Section 2.4.2 below, where we focus on the boundary layer and show that, specifically for the Adv-MsFEM method but not for the other methods, the accuracy (within the layer) is significantly improved upon changing the boundary conditions in the local problem (2.31).

As shown on Table 2.3, the Adv-MsFEM method has a relative H^1 error outside the layer about 7 times as large as the error of the Stab-MsFEM method. On this example, the methods that provide the lowest H^1 error outside the layer are the Stab-MsFEM method and the splitting method.

	e_{L^2}	e_{L^∞}	e_{H^1}	$e_{H^1_{\mathbf{in}}}$	$e_{H^1_{\mathbf{out}}}$
$\mathbb{P}^1$ Upwind	0.86	0.94	0.98	0.97	0.13
MsFEM	0.27	1.63	1.13	0.97	0.57
Stab-MsFEM	0.23	0.81	0.87	0.87	0.04
Adv-MsFEM	0.11	0.62	0.74	0.68	0.29
Splitting (2.34)–(2.35)	0.22	0.80	0.87	0.87	0.03

Table 2.3 Relative errors in the multiscale case ($\alpha = 1/128$, $\delta = 0.5$, $\varepsilon = 1/64$ and $H = 1/16$)

Comparison of the splitting methods

We specifically compare here our two variants of the splitting approach: (2.34)–(2.35) and (2.77)–(2.81).

In spite of the value of $\rho_- = 128$ in (2.69), so that assumption (2.61) of Lemma 2.16 is violated, the method (2.34)–(2.35) converges. For the approach (2.77)–(2.81), we choose β as in (2.85), that is $\beta = 1.9941$. In the numerical tests, we have not used the projection $P_{V_H^\varepsilon}$ (see Remark 2.19). The contraction factor (see (2.98) in the proof of Lemma 2.20) is $\rho = 0.99902$. Given the proof of Lemma 2.20 and that value of ρ, the convergence is expected to be slow. It is indeed *very* slow, as will now be seen, confirming that the approach (2.77)–(2.81) is only advocated in the case where the convergence of (2.34)–(2.35) fails.

Figure 2.3 shows the error (in terms of the iteration residual (2.36) for the method (2.34)–(2.35), and a similar quantity for the method (2.77)–(2.81)) in function of the number of iterations. The method (2.77)–(2.81) needs 100 times more iterations than the method (2.34)–(2.35) to reach the same tolerance with respect to that criterion.

Table 2.4 shows the accuracy of the methods with respect to the reference solution. We see that the method (2.34)–(2.35) is more accurate than the method (2.77)–(2.81) (outside the boundary layer). Both methods are equally inaccurate inside the boundary layer.

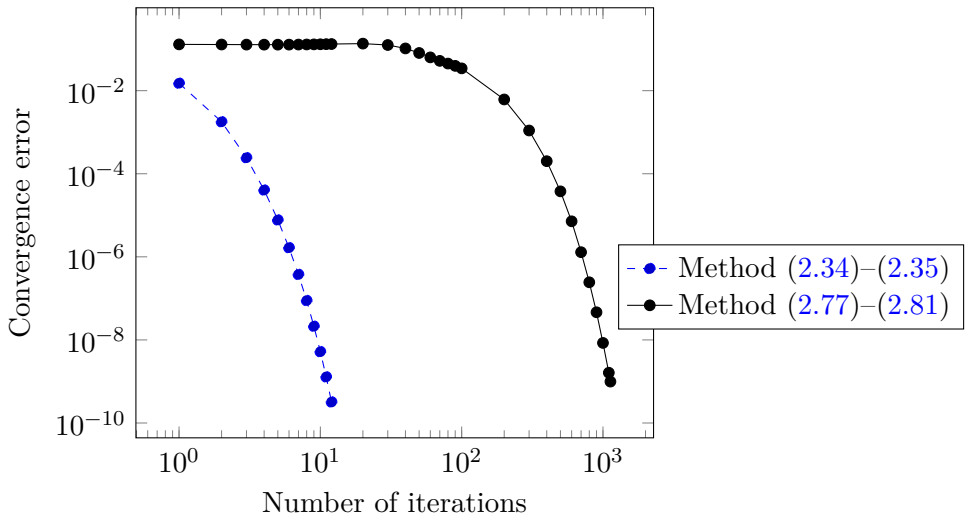

Figure 2.3 Convergence history of the two splitting methods ($\alpha = 1/128$, $\delta = 0.5$, $\varepsilon = 1/64$ and $H = 1/16$)

	e_{L^2}	e_{L^∞}	e_{H^1}	$e_{H^1_{\mathbf{in}}}$	$e_{H^1_{\mathbf{out}}}$
Method (2.34)–(2.35)	0.22	0.80	0.87	0.87	0.03
Method (2.77)–(2.81)	0.59	0.96	0.94	0.94	0.10

Table 2.4 Relative errors for the two splitting methods ($\alpha = 1/128$, $\delta = 0.5$, $\varepsilon = 1/64$ and $H = 1/16$)

In all what follows, we have only used the splitting method (2.34)–(2.35), which needs fewer iterations to converge and provides a more accurate solution.

Sensitivity with respect to the Péclet number

We set $\delta = 0.75$, $\varepsilon = 1/128$, $H = 1/16$ and $\alpha = 2^{-k}$, $k = 2, \cdots, 9$. We let α vary in order to assess the robustness of the approaches with respect to the Péclet number.

From (2.87), we suspect the advection-dominated regime corresponds to $k > 5$. To doublecheck this is indeed the case, we first set $\varepsilon = 1$ and show on Figure 2.4 the relative errors of the $\mathbb{P}^1$ method and the $\mathbb{P}^1$ Upwind method. We indeed see that, for $k > 5$, the relative H^1 error outside the layer of the $\mathbb{P}^1$ method is at least five times as large as the relative H^1 error outside the layer of the $\mathbb{P}^1$ Upwind method. In the sequel, we go back to the multiscale case with $\varepsilon = 1/128$.

Errors in the whole domain Results are shown on Figure 2.5. When α is small, all methods yield rather large errors. The error of the MsFEM method is significantly more important than the error of the Stab-MsFEM method. This indicates the presence of spurious oscillations on the solution obtained with the non stabilized MsFEM method. Hence the stabilization is important in this regime. The most robust methods are the Adv-MsFEM method, the Stab-MsFEM method and the splitting method.

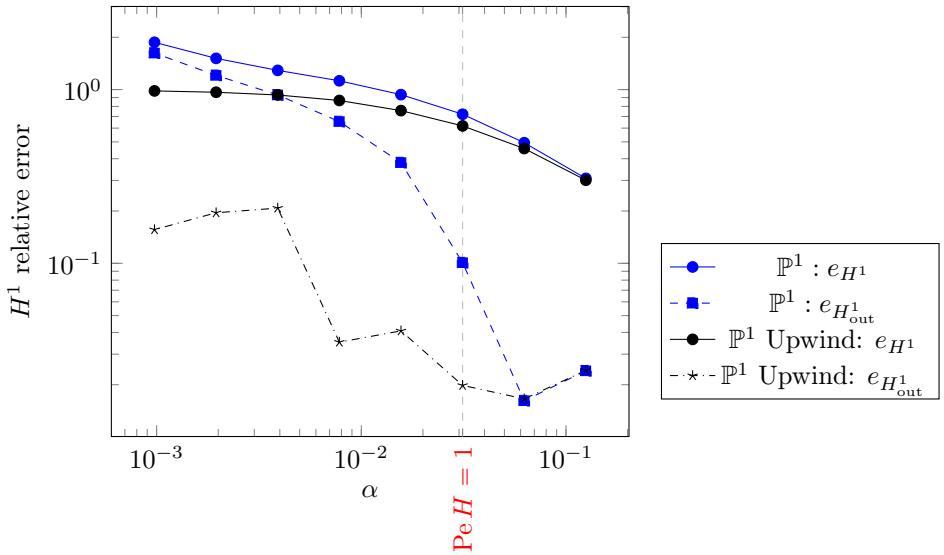

Figure 2.4 Relative errors in the single-scale case ($\delta = 0.75$, $\varepsilon = 1$ and $H = 1/16$).

When α is large, the main difficulty is to capture the oscillations at scale ε. As expected, all multiscale methods perform better than the $\mathbb{P}^1$ Upwind method. Note that the MsFEM method and the Stab-MsFEM method perform similarly. No stabilization is indeed necessary in that regime.

In both regimes, we note that the Adv-MsFEM method performs the best. We also see that the errors of the splitting method are extremely close to the errors of the Stab-MsFEM method.

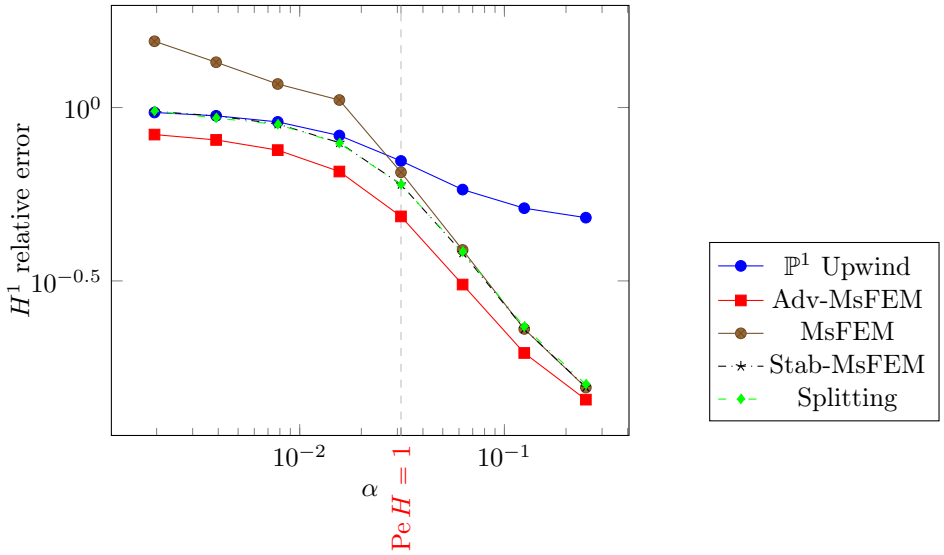

Figure 2.5 Relative error e_{H^1} ($\delta = 0.75$, $\varepsilon = 1/128$ and $H = 1/16$).

Errors outside the boundary layer It may be observed on Figure 2.6 that the Stab-MsFEM method and the splitting method are the best methods outside the boundary layer. They essentially share the same accuracy. On the other hand, the Adv-MsFEM solution is systematically less accurate than the Stab-MsFEM solution. This suggests that encoding the advection in the multiscale basis functions is not necessary to obtain a good accuracy in this subdomain, and that it may even deteriorate the quality of the numerical solution. The MsFEM

method is much less accurate than the stabilized Stab-MsFEM method when the coercivity constant α is small, and has a comparable accuracy when α is larger than 0.1.

When α is large (and hence the only difficulty is to capture the oscillation scale ε), the $\mathbb{P}^1$ Upwind method is less accurate than the Stab-MsFEM method, as expected, since the latter encodes the oscillations of A^ε in the multiscale basis functions. When α is moderately small ($10^{-2} < \alpha < 1/32$ on Figure 2.6), the problem is both advection-dominated (we indeed observe that the Stab-MsFEM method provides a better accuracy than the MsFEM method) and multiscale (the Stab-MsFEM method is more accurate than the $\mathbb{P}^1$ Upwind method). However, when α is very small (here, $\alpha < 10^{-2}$), the advection is so large that it overshadows the multiscale nature of the problem. We then observe that the $\mathbb{P}^1$ Upwind method and the Stab-MsFEM method share the same accuracy.

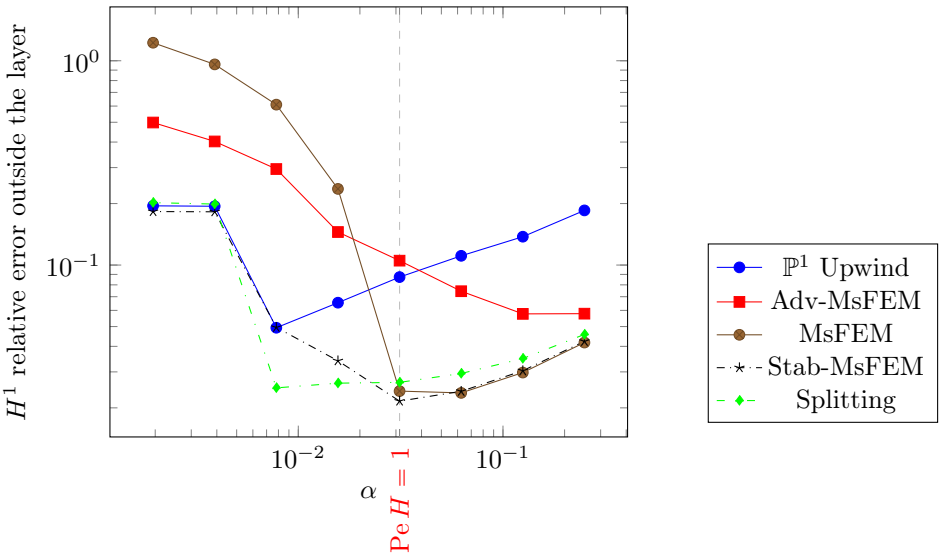

Figure 2.6 Relative error $e_{H^1_{\text{out}}}$ ($\delta = 0.75$, $\varepsilon = 1/128$ and $H = 1/16$).

Of course, the values of α that define these three regimes ((i) advection-dominated, (ii) *both* advection-dominated and multiscale, (iii) multiscale) depend on the problem considered, and in particular on the value of ε. We have checked this sensitivity by considering the following two test-cases: on Figures 2.7 and 2.8, we consider the case $\varepsilon = 1/64$ and $\varepsilon = 1/256$, respectively. The other parameters are $\delta = 0.75$ and $H = 1/32$. When $\varepsilon = 1/256$, we observe that the Stab-MsFEM method is more accurate than the $\mathbb{P}^1$ Upwind method for any $1/512 \leqslant \alpha \leqslant 1/4$. In contrast, when $\varepsilon = 1/64$, the $\mathbb{P}^1$ Upwind method and the Stab-MsFEM method perform equally well when $\alpha \leqslant 1/256$. The sensitivity with respect to ε is also investigated in Section 2.4.2 below (see e.g. Figure 2.10).

Sensitivity with respect to the oscillation scale

In this section, the sensitivity of the different numerical methods to the oscillation scale ε is assessed. We work with the parameters $\delta = 0.75$, $H = 1/32$, $\alpha = 1/128$ and $\varepsilon = 2^{-k}$, $k = 3, \cdots, 8$, so that $\text{Pe}\, H = 2 > 1$. Table 2.5 displays the relative errors of the $\mathbb{P}^1$ method and the $\mathbb{P}^1$ Upwind method for $\varepsilon = 1$. Outside the layer, the relative H^1 error of the $\mathbb{P}^1$ method is about 30 times as large as the error of the $\mathbb{P}^1$ Upwind method. The problem is advection-dominated.

Figures 2.9 and 2.10 respectively show the relative global H^1 error and the relative H^1 error outside the boundary layer. The relative global H^1 error does not seem to be sensitive to the

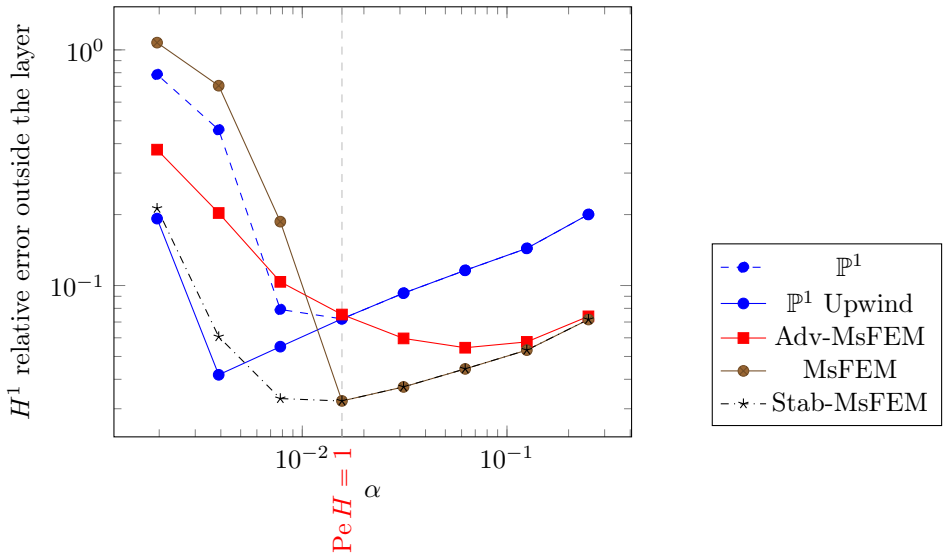

Figure 2.7 Relative error $e_{H^1_{\text{out}}}$ ($\delta = 0.75$, $\varepsilon = 1/64$ and $H = 1/32$).

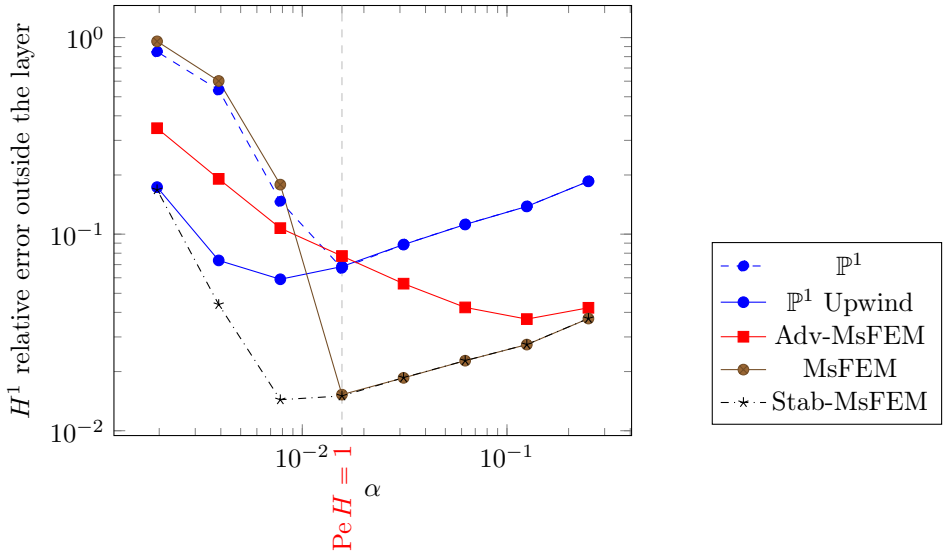

Figure 2.8 Relative error $e_{H^1_{\text{out}}}$ ($\delta = 0.75$, $\varepsilon = 1/256$ and $H = 1/32$).

	e_{L^2}	e_{L^∞}	e_{H^1}	$e_{H^1_{\mathbf{in}}}$	$e_{H^1_{\mathbf{out}}}$
$\mathbb{P}^1$	0.11	0.48	0.93	0.86	0.33
$\mathbb{P}^1$ Upwind	0.11	0.46	0.75	0.75	0.01

Table 2.5 Relative errors in the single-scale case ($\alpha = 1/128$, $\delta = 0.75$, $\varepsilon = 1$ and $H = 1/32$).

oscillation scale, as we can see on Figure 2.9. This error is dominated by the error located in the thin boundary layer due to the advection-dominated regime.

On Figure 2.10, two regions can be distinguished. In the region $\varepsilon < H$, the Stab-MsFEM method performs better than the $\mathbb{P}^1$ Upwind method. The error of the Stab-MsFEM method decreases as ε decreases (but its offline cost increases correspondingly, as the mesh to compute the highly oscillatory basis functions has to be finer), whereas the error of the $\mathbb{P}^1$ Upwind method remains constant at a large value as ε decreases. The Adv-MsFEM method yields a large error (due to the mismatch between the shape of the solution outside the boundary layer and the shape of the basis functions). The MsFEM method is also inaccurate, given the absence of any stabilization.

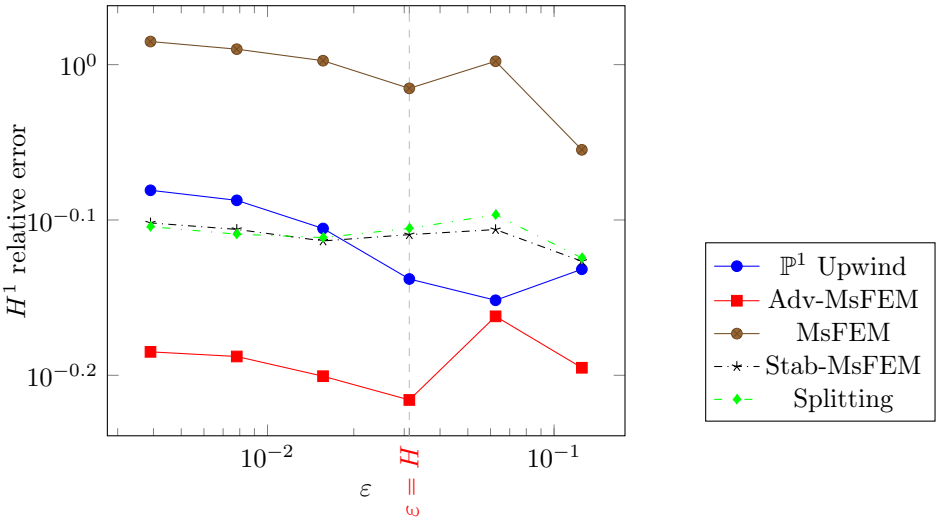

Figure 2.9 Relative error e_{H^1} ($\alpha = 1/128$, $\delta = 0.75$ and $H = 1/32$).

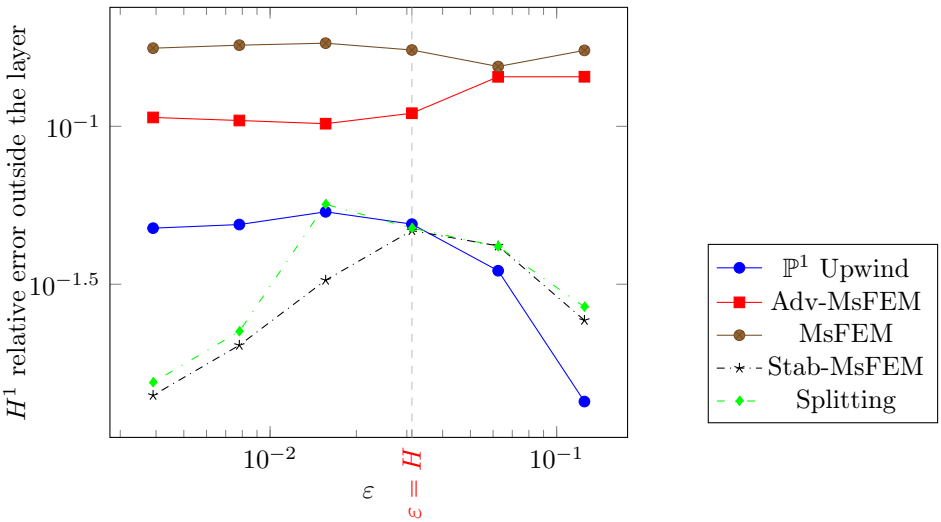

Figure 2.10 Relative error $e_{H^1_{\text{out}}}$ ($\alpha = 1/128$, $\delta = 0.75$ and $H = 1/32$).

Influence of the boundary conditions imposed on the local problems

In all the above experiments, the boundary conditions we have supplied the local problems (2.20) and (2.31) with are *linear* boundary conditions. For other choices of boundary conditions, our results remain qualitatively unchanged. We however wish to now investigate how the choice of

boundary conditions affects the accuracy of the approaches *within* the boundary layer, since this is where the approaches equally poorly perform. It is known that, in general, the oversampling method is one of the best multiscale approach available for the multiscale diffusion problem (2.17). Whether this superiority also survives in the presence of a strong advection is an interesting issue.

For clarity, the Adv-MsFEM method as presented above (i.e. based on the local problem (2.31)) is denoted here the Adv-MsFEM lin method. The other boundary conditions that we consider are

- Oversampling boundary condition, with an oversampling ratio equal to 3 (i.e. the local problems defining the basis functions are set on a quadrangle of size $3H \times 3H$). This method is denoted the Adv-MsFEM OS method;
- Crouzeix-Raviart type boundary condition. This method is denoted the Adv-MsFEM CR method.

The oversampling method is described in [47]. The MsFEM à la Crouzeix-Raviart has been introduced in [58, 59]. Both the Adv-MsFEM OS and the Adv-MsFEM CR methods are non-conforming approaches, the purpose of which is to allow for more flexible boundary conditions on the boundary of the elements. We recall that the former approach achieves this by solving the local problems on a larger domain than, and usually homothetic to, the coarse mesh element itself (see Section 2.2.1). The oversampling ratio is the homothetic factor. The latter approach imposes continuity in a weak (integral) form at the edges, see [58, 59] for more details. The relative H^1 error of those methods is computed with the broken H^1 norm

$$e_{H^1_{\text{in}}}(u_1) = \frac{\|u_1 - u_{\text{ref}}\|_{H^1(\mathcal{T}_H \cap \Omega_{\text{layer}})}}{\|u_{\text{ref}}\|_{H^1(\Omega)}},$$

with $\|u\|^2_{H^1(\mathcal{T}_H \cap \Omega_{\text{layer}})} = \sum_{\mathbf{K} \in \mathcal{T}_H} \|u\|^2_{H^1(\mathbf{K} \cap \Omega_{\text{layer}})}$.

We first study the example presented in Section 2.4.2. Table 2.6 shows the relative errors. We observe that there is at least a factor 2 between the relative H^1 error inside the layer of the Adv-MsFEM lin and the other Adv-MsFEM methods. The improvement in the accuracy outside the boundary layer is less important, although significant for the Adv-MsFEM CR method.

	e_{L^2}	e_{L^∞}	e_{H^1}	$e_{H^1_{\text{in}}}$	$e_{H^1_{\text{out}}}$
Adv-MsFEM lin	0.11	0.62	0.74	0.68	0.29
Adv-MsFEM OS	0.36	0.55	0.42	0.34	0.24
Adv-MsFEM CR	0.038	0.034	0.20	0.075	0.18

Table 2.6 Relative errors for different boundary conditions in the local problems.

Second, we consider the setting presented in Section 2.4.2. Figure 2.11 shows the relative H^1 error inside the layer for the different Adv-MsFEM methods. We observe that the boundary conditions imposed on the local problems affect the accuracy. The Adv-MsFEM lin method always has the largest error. In the advection-dominated regime, the Adv-MsFEM CR is the best method. At $\text{Pe}\, H = 16$, there is a factor 8 between the relative H^1 error of the Adv-MsFEM lin method and the relative H^1 error of the Adv-MsFEM CR method (inside the layer). This shows

that the convective profile should be encoded in some way in the boundary conditions imposed on the local problem in order for the solution to be accurate in the boundary layer region for the advection-dominated regime. For the MsFEM approaches other than the Adv-MsFEM approach, we have also performed similar experiments, which are not included here, and which do not seem to show any significant dependency of the accuracy inside the layer upon the boundary conditions of the local problems.

Figure 2.12 shows the relative H^1 error outside the layer for the different Adv-MsFEM methods. It may be observed that the Adv-MsFEM lin method and the Adv-MsFEM OS method share the same error. In the advection-dominated regime, the error of the Adv-MsFEM CR is the smallest. However, the errors outside the boundary layer of the various Adv-MsFEM methods are yet larger than the error outside the layer of the Stab-MsFEM method.

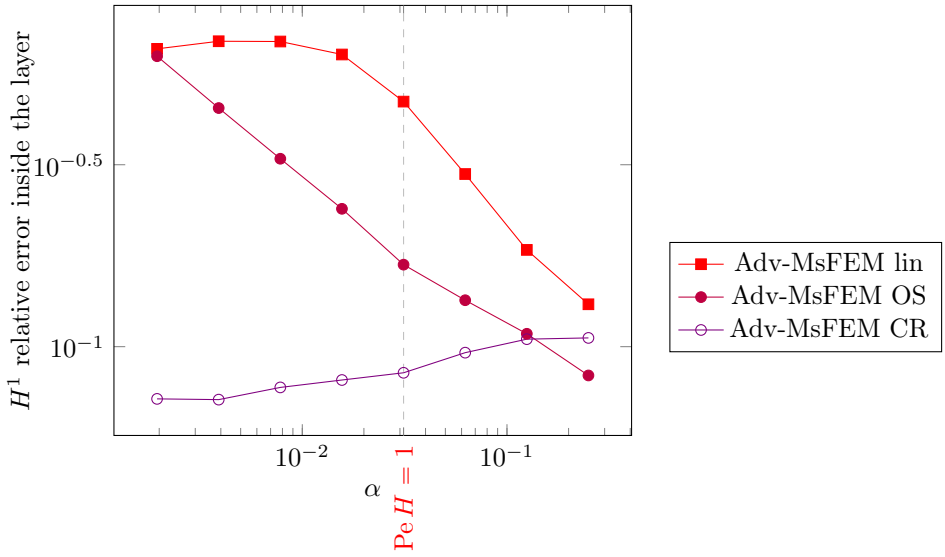

Figure 2.11 Relative error $e_{H^1_{\text{in}}}$ for the Adv-MsFEM methods.

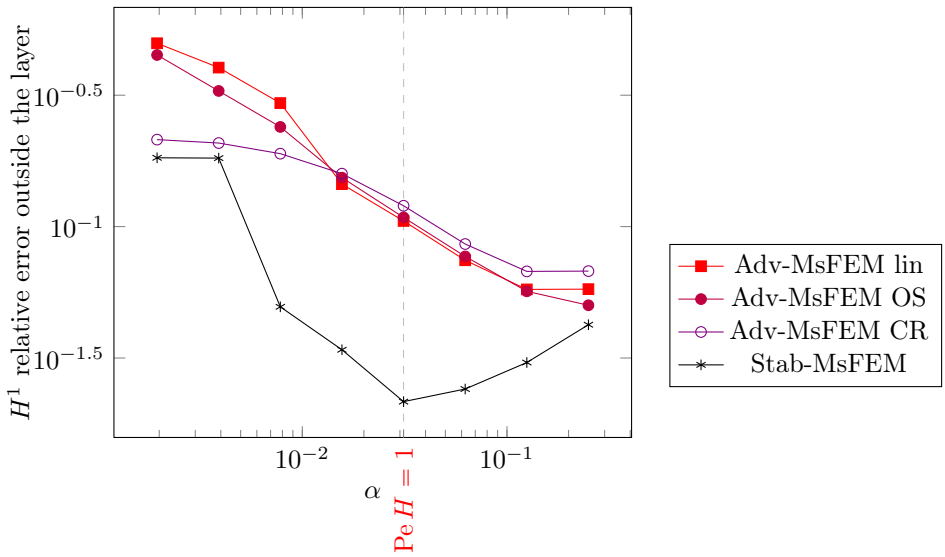

Figure 2.12 Relative error $e_{H^1_{\text{out}}}$ for the Adv-MsFEM methods.

2.4.3 Computational costs

We now turn to the computational costs of the different numerical methods. We recall that the splitting method we consider below is (2.34)–(2.35).

Reference test

We consider the reference test presented in Section 2.4.2. Table 2.7 shows the offline cost and the online cost (in seconds) of the different numerical methods.

Direct solvers	Offline (s)	Online (s)	**Iterative solvers**	Offline (s)	Online (s)
Stab-MsFEM	$1.98 \cdot 10^2$	$2.24 \cdot 10^{-4}$	Stab-MsFEM	$2.63 \cdot 10^2$	$5.78 \cdot 10^{-4}$
Splitting	$2.29 \cdot 10^2$	$3.81 \cdot 10^{-3}$	Splitting	$2.65 \cdot 10^2$	$9.03 \cdot 10^{-3}$
MsFEM	$1.80 \cdot 10^2$	$2.41 \cdot 10^{-4}$	MsFEM	$2.33 \cdot 10^2$	$1.63 \cdot 10^{-3}$
Adv-MsFEM	$1.84 \cdot 10^2$	$2.20 \cdot 10^{-4}$	Adv-MsFEM	$5.89 \cdot 10^2$	$6.99 \cdot 10^{-4}$

Table 2.7 Computational costs.

Direct solvers All the methods (but the splitting method) essentially share the same offline cost. The Stab-MsFEM method is slightly more expensive than the MsFEM variant because of the assembling of the stabilization term. The splitting method has the largest offline cost because there are more computations (two assemblings) than in the other methods.

The online cost of the splitting method is about 15 times as large as the online cost of the other methods. This corresponds to the number of iterations of the splitting method. Note that the online cost corresponds to solving the linear system from an already factorized matrix, which is negligible.

Iterative solvers The online cost of the intrusive methods (Adv-MsFEM, MsFEM, Stab-MsFEM) corresponds to calling the GMRES solver. There are some differences in these costs because the number of iterations of the GMRES solver is sensitive to the condition number of the matrix that depends on the method. The online cost of the splitting method is still the largest because of the iteration loop of the splitting method. It is again about 15 times larger than the online cost of the Stab-MsFEM method. In this particular case, the splitting method needs 12 iterations to converge. The online costs are larger now than with direct solvers, of course.

The main part of the offline cost comes from solving the local problems. The MsFEM, Stab-MsFEM, and the splitting method share the same local problems, namely (2.20). This is why they essentially share the same offline cost. In the Adv-MsFEM method, the local problem to solve is (2.31). We observe that its offline cost is about 2 times larger than for the other methods. We thus see that the computational cost of solving with the GMRES solver the non-symmetric linear system corresponding to the local problem (2.31) is higher that the cost of solving with the conjugate gradient method the symmetric linear system stemming from the local problem (2.20).

Dependency with respect to the Péclet number

We again consider the setting of Section 2.4.2 where we now vary the coefficient α and thus the Péclet number. Figures 2.13 and 2.14 respectively show the online cost (in seconds) of the

different numerical methods and the number of iterations of the splitting method as a function of α.

In Figure 2.13, we observe that the Adv-MsFEM method and the Stab-MsFEM method share the same online cost. The online costs of the two methods and the online cost of the MsFEM method (with direct solvers) do not seem to strongly depend on the Péclet number. The online cost of the MsFEM method with iterative solvers increases as α decreases, since the condition number of the stiffness matrix then increases. The splitting method is, overall, significantly more expensive than the other approaches.

Figure 2.14 shows that the number of iterations in the splitting method grows as α decreases. The number of iterations is larger when using iterative solvers than when using direct solvers, although the difference fades as the advection becomes dominant.

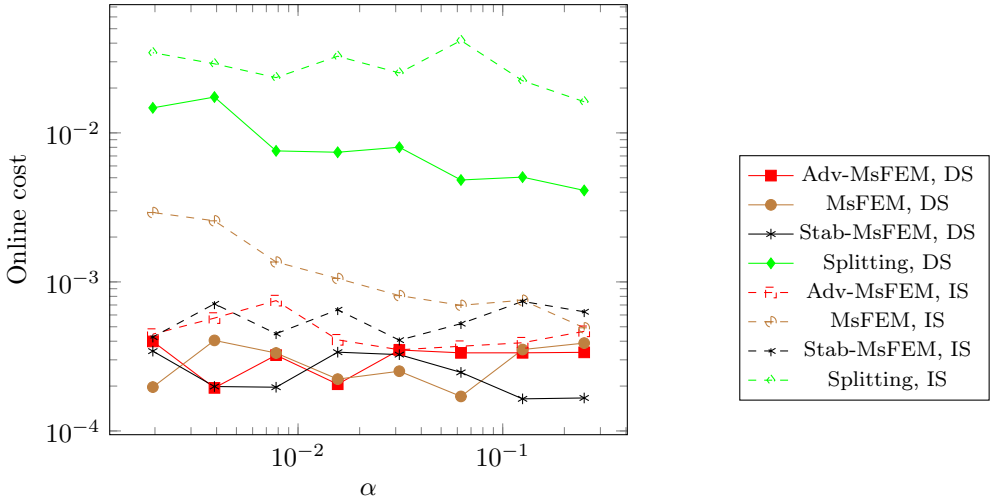

Figure 2.13 Online costs (s) for the different numerical methods, using direct (DS) or iterative (IS) solvers.

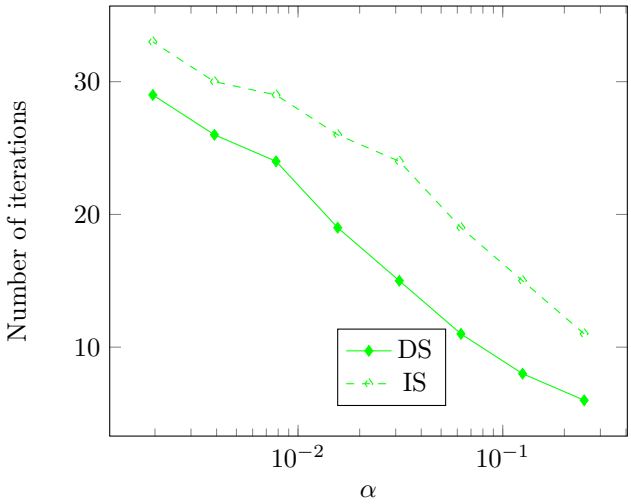

Figure 2.14 Number of iterations of the splitting method for direct (DS) and iterative (IS) solvers.

Acknowledgements

The work presented in this article elaborates on a preliminary work that explored some of the issues on a prototypical one-dimensional setting, and which was performed in the context of the internship of H. Ruffieux [83] at CERMICS, École des Ponts ParisTech. The present work benefits from this previous work. The authors wish to thank A. Quarteroni for stimulating and enlightning discussions. CLB and FL also gratefully acknowledge the long term interaction with U. Hetmaniuk (University of Washington in Seattle) and A. Lozinski (Université de Besançon) on numerical methods for multiscale problems. The authors thank the referees for their many comments on the original version of this work. The work of the authors is partially supported by ONR under Grant N00014-12-1-0383 and EOARD under Grant FA8655-13-1-3061.

2.5 Appendix: Technical proofs

2.5.1 Proof of (2.11)

By standard finite element results (see e.g. [13, end of Section IX.3] or [38, Remark 1.129 and Lemma 1.127]), we have the following best approximation property: for any H and any $v \in H^1_0(\Omega) \cap H^2(\Omega)$, there exists $I_H v \in V_H$ such that

$$\|v - I_H v\|_{L^2(\Omega)} + H\,|v - I_H v|_{H^1(\Omega)} \leqslant C_{\rm FE} H^2 |v|_{H^2(\Omega)} \tag{2.88}$$

where $C_{\rm FE}$ is independent of H and v.

Our proof of (2.11) is inspired by [83] and [77, Theorem 11.2]. We split the error $u - u^s_H$ in two parts, $e^I = u - I_H u$ and $e^I_H = u^s_H - I_H u$. Given (2.88), we have

$$|e^I|_{H^1(\Omega)} \leqslant CH|u|_{H^2(\Omega)}. \tag{2.89}$$

We now estimate the term $|e^I_H|_{H^1(\Omega)}$. The Galerkin orthogonality and (2.3) give

$$\begin{aligned}
&\alpha|e^I_H|^2_{H^1(\Omega)} + a_{\rm stab}(e^I_H, e^I_H) \\
&= a(e^I_H, e^I_H) + a_{\rm stab}(e^I_H, e^I_H) \\
&= a(e^I, e^I_H) + a_{\rm stab}(e^I, e^I_H) \\
&= \int_\Omega \left(\alpha \nabla e^I \cdot \nabla e^I_H + (b \cdot \nabla e^I) e^I_H\right) + \sum_{\mathbf{K} \in \mathcal{T}_H} \left(\tau_{\mathbf{K}} \mathcal{L} e^I, b \cdot \nabla e^I_H\right)_{\mathbf{K}} \\
&= \int_\Omega \left(\alpha \nabla e^I \cdot \nabla e^I_H - (b \cdot \nabla e^I_H) e^I\right) + \sum_{\mathbf{K} \in \mathcal{T}_H} \left(\tau_{\mathbf{K}} \mathcal{L} e^I, b \cdot \nabla e^I_H\right)_{\mathbf{K}}.
\end{aligned} \tag{2.90}$$

Estimating each term of the right-hand side of (2.90), we successively obtain:

- for the first term, using (2.89),

$$\int_\Omega \alpha \nabla e^I \cdot \nabla e^I_H \leqslant \frac{\alpha}{4} |e^I_H|^2_{H^1(\Omega)} + \alpha |e^I|^2_{H^1(\Omega)} \leqslant \frac{\alpha}{4} |e^I_H|^2_{H^1(\Omega)} + C\alpha H^2 |u|^2_{H^2(\Omega)}.$$

- for the second term,

$$\begin{aligned} -\int_\Omega (b\cdot\nabla e_H^I)e^I &\leqslant \frac{1}{4}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2+\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{-1/2}e^I\|_{L^2(\mathbf{K})}^2\\ &\leqslant \frac{1}{4}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2+\frac{2\|b\|_{L^\infty}}{H}\|e^I\|_{L^2(\Omega)}^2\\ &\leqslant \frac{1}{4}a_{\mathrm{stab}}(e_H^I,e_H^I)+C\|b\|_{L^\infty}H^3|u|_{H^2(\Omega)}^2, \end{aligned}$$

  where we have used that $\Delta e_H^I=0$ in the first term and (2.13) and (2.88) in the second term;

- for the first part of the third term, using that $\Delta e^I=\Delta u$ (because V_H is the $\mathbb{P}^1$ finite element space),

$$\begin{aligned} &\sum_{\mathbf{K}\in\mathcal{T}_H}\left(\tau_{\mathbf{K}}(-\alpha\Delta e^I),b\cdot\nabla e_H^I\right)_{\mathbf{K}}\\ \leqslant\ &\frac{1}{2}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}\alpha\Delta u\|_{L^2(\mathbf{K})}^2+\frac{1}{2}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2\\ \leqslant\ &\frac{\|b\|_{L^\infty}H^3}{16}|u|_{H^2(\Omega)}^2+\frac{1}{2}\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2\\ \leqslant\ &C\|b\|_{L^\infty}H^3|u|_{H^2(\Omega)}^2+\frac{1}{2}a_{\mathrm{stab}}(e_H^I,e_H^I), \end{aligned}$$

  where we have used (2.12) to obtain $\alpha^2\tau_{\mathbf{K}}(x)\leqslant\left(\frac{|b(x)|H}{2}\right)^2\frac{H}{2|b(x)|}\leqslant\frac{\|b\|_{L^\infty}H^3}{8}$;

- for the second part of the third term,

$$\begin{aligned} &\sum_{\mathbf{K}\in\mathcal{T}_H}\left(\tau_{\mathbf{K}}b\cdot\nabla e^I,b\cdot\nabla e_H^I\right)_{\mathbf{K}}\\ \leqslant\ &\sum_{\mathbf{K}\in\mathcal{T}_H}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e^I)\|_{L^2(\mathbf{K})}^2+\sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{4}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2\\ \leqslant\ &\frac{\|b\|_{L^\infty}H}{2}|e^I|_{H^1(\Omega)}^2+\sum_{\mathbf{K}\in\mathcal{T}_H}\frac{1}{4}\|\tau_{\mathbf{K}}^{1/2}(b\cdot\nabla e_H^I)\|_{L^2(\mathbf{K})}^2\\ \leqslant\ &C\|b\|_{L^\infty}H^3|u|_{H^2(\Omega)}^2+\frac{1}{4}a_{\mathrm{stab}}(e_H^I,e_H^I), \end{aligned}$$

  where we used that $\tau_{\mathbf{K}}(x)|b(x)\cdot\nabla e^I(x)|^2\leqslant\frac{H|b(x)|}{2}|\nabla e^I(x)|^2$.

Collecting the terms, we eventually deduce from (2.90) that

$$\alpha|e_H^I|_{H^1(\Omega)}^2+a_{\mathrm{stab}}(e_H^I,e_H^I)\leqslant\frac{\alpha}{4}|e_H^I|_{H^1(\Omega)}^2+a_{\mathrm{stab}}(e_H^I,e_H^I)+CH^2\left(\alpha+\|b\|_{L^\infty}H\right)|u|_{H^2(\Omega)}^2,$$

hence $\frac{3}{4}\alpha|e_H^I|_{H^1(\Omega)}^2\leqslant CH^2\left(\alpha+\|b\|_{L^\infty}H\right)|u|_{H^2(\Omega)}^2$, thus

$$|e_H^I|_{H^1(\Omega)}\leqslant CH\left(1+\mathrm{Pe}\,H\right)^{1/2}|u|_{H^2(\Omega)}.$$

Along with (2.89), this estimate shows (2.11).

2.5.2 Density of the MsFEM spaces V_H^ε in $H_0^1(\Omega)$

We prove here the following result:

Lemma 2.21. *Let ε be fixed and assume that $A^\varepsilon \in W^{1,\infty}(\Omega)$. Consider, for any H, the spaces V_H^ε defined by (2.21). Then, for any $g \in H_0^1(\Omega)$ and any $\eta > 0$, there exists $H_0 > 0$ such that, for any $H < H_0$, there exists some $w_H \in V_H^\varepsilon$ such that $\|g - w_H\|_{H^1(\Omega)} \leqslant \eta$.*

Note that we make no structural assumption on how A^ε depends on ε, and that we do not assume ε to be small.

Proof. We fix η and $g \in H_0^1(\Omega)$. By density of $H^2(\Omega) \cap H_0^1(\Omega)$ in $H_0^1(\Omega)$, there exists $\widetilde{g} \in H^2(\Omega) \cap H_0^1(\Omega)$ such that

$$\|g - \widetilde{g}\|_{H^1(\Omega)} \leqslant \eta. \tag{2.91}$$

By standard finite element results (see e.g. [13, end of Section IX.3] or [38, Remark 1.129 and Lemma 1.127]), for any H, there exists $v_H \in V_H$ such that

$$\|\widetilde{g} - v_H\|_{H^1(\Omega)} \leqslant C_{\rm FE} H \|\widetilde{g}\|_{H^2(\Omega)} \tag{2.92}$$

where $C_{\rm FE}$ is independent of H and $\widetilde{g}$.

Picking $H_{\rm TR1} = \eta/(C_{\rm FE}\|\widetilde{g}\|_{H^2(\Omega)})$, we therefore deduce from (2.91) and (2.92) that, for any $H < H_{\rm TR1}$, we have $v_H \in V_H$ such that

$$\|g - v_H\|_{H^1(\Omega)} \leqslant 2\eta. \tag{2.93}$$

The function v_H belongs to V_H, and hence reads $v_H = \sum_i v_i \, \phi_i^0$. We now consider $w_H = \sum_i v_i \, \psi_i^\varepsilon$, which belongs to V_H^ε. On each element $\mathbf{K}$, we have

$$\begin{cases} -\text{div}\,\big(A^\varepsilon \nabla(w_H - v_H)\big) = \text{div}\,\,(A^\varepsilon \nabla v_H) = \displaystyle\sum_{k,\ell=1}^{d} \partial_k (A^\varepsilon)_{k\ell}\, \partial_\ell v_H \quad \text{in } \mathbf{K}, \\ w_H - v_H = 0 \quad \text{on } \partial\mathbf{K}. \end{cases}$$

We hence have that

$$\int_{\mathbf{K}} \big(\nabla(w_H - v_H)\big)^T A^\varepsilon \nabla(w_H - v_H) = \int_{\mathbf{K}} \sum_{k,\ell=1}^{d} (w_H - v_H) \partial_k (A^\varepsilon)_{k\ell}\, \partial_\ell v_H$$

and therefore, using (2.18),

$$\alpha_1 \|\nabla(w_H - v_H)\|^2_{L^2(\mathbf{K})} \leqslant \|A^\varepsilon\|_{W^{1,\infty}(\Omega)} \, \|w_H - v_H\|_{L^2(\mathbf{K})} \, \|\nabla v_H\|_{L^2(\mathbf{K})}.$$

Using the Poincaré inequality on $\mathbf{K}$, we obtain that there exists a constant C (which only depends on the shape of the elements, but not on their size) such that

$$\alpha_1 \|\nabla(w_H - v_H)\|_{L^2(\mathbf{K})} \leqslant CH \|A^\varepsilon\|_{W^{1,\infty}(\Omega)} \, \|\nabla v_H\|_{L^2(\mathbf{K})}.$$

Summing over the elements, we obtain

$$\alpha_1 \|\nabla(w_H - v_H)\|_{L^2(\Omega)} \leqslant CH\|A^\varepsilon\|_{W^{1,\infty}(\Omega)} \|\nabla v_H\|_{L^2(\Omega)}$$

which implies, using the Poincaré inequality on Ω, that

$$\alpha_1 \|w_H - v_H\|_{H^1(\Omega)} \leqslant CH\|A^\varepsilon\|_{W^{1,\infty}(\Omega)} \|v_H\|_{H^1(\Omega)}.$$

Using (2.93) in the above bound, we get that, for any $H < H_{\mathrm{TR1}}$,

$$\|w_H - v_H\|_{H^1(\Omega)} \leqslant C_{\mathrm{P}} H \left[2\eta + \|g\|_{H^1(\Omega)}\right]. \tag{2.94}$$

We set $H_{\mathrm{TR2}} = \min(1/C_{\mathrm{P}}, \eta/(C_{\mathrm{P}}\|g\|_{H^1(\Omega)}))$ and $H_0 = \min(H_{\mathrm{TR1}}, H_{\mathrm{TR2}})$. Collecting (2.93) and (2.94), we deduce that, for any $H < H_0$, $w_H \in V_H^\varepsilon$ satisfies

$$\|g - w_H\|_{H^1(\Omega)} \leqslant 5\eta.$$

This concludes the proof. □

2.5.3 Proof of Lemma 2.20

We first study the convergence when $n \to \infty$, and next when $H \to 0$.

Step 1: Convergence when $n \to \infty$. Let $\widetilde{u}_n^H = u_{n+2}^H - u_n^H$. We directly infer from (2.81) that

$$|\widetilde{u}_{2n+1}^H|_{H^1(\Omega)} \leqslant \frac{\beta + \alpha_{\mathrm{spl}}}{\beta + \alpha_1} |\widetilde{u}_{2n}^H|_{H^1(\Omega)}. \tag{2.95}$$

We now estimate $|\widetilde{u}_{2n+2}^H|_{H^1(\Omega)}$ and $|\widetilde{u}_{2n+1}^H - P_{V_H^\varepsilon}(\widetilde{u}_{2n}^H)|_{H^1(\Omega)}$. Using the variational formulations (2.77) for u_{2n+2}^H and u_{2n+4}^H, we deduce a variational formulation for $\widetilde{u}_{2n+2}^H = u_{2n+4}^H - u_{2n+2}^H$. Taking $\widetilde{u}_{2n+2}^H$ as test function in that variational formulation, and setting $w_n = P_{V_H^\varepsilon}(\widetilde{u}_{2n}^H) - \widetilde{u}_{2n+1}^H$, we get

$$\begin{aligned}
&(\alpha_{\mathrm{spl}} + \beta)|\widetilde{u}_{2n+2}^H|_{H^1(\Omega)}^2 \\
&\leqslant \int_\Omega (b \cdot \nabla w_n)\widetilde{u}_{2n+2}^H + \sum_{\mathbf{K}\in\mathcal{T}_H} (\tau_{\mathbf{K}} b \cdot \nabla w_n, b \cdot \nabla \widetilde{u}_{2n+2}^H)_{L^2(\mathbf{K})} + \int_\Omega \beta \nabla \widetilde{u}_{2n+1}^H \cdot \nabla \widetilde{u}_{2n+2}^H \\
&\leqslant \|b\|_{L^\infty(\Omega)} |w_n|_{H^1(\Omega)} \|\widetilde{u}_{2n+2}^H\|_{L^2(\Omega)} \\
&\quad + \sum_{\mathbf{K}\in\mathcal{T}_H} \frac{H\|b\|_{L^\infty(\Omega)}}{2} \|\nabla w_n\|_{L^2(\mathbf{K})} \|\nabla \widetilde{u}_{2n+2}^H\|_{L^2(\mathbf{K})} + \beta |\widetilde{u}_{2n+1}^H|_{H^1(\Omega)} |\widetilde{u}_{2n+2}^H|_{H^1(\Omega)} \\
&\leqslant \|b\|_{L^\infty(\Omega)} |w_n|_{H^1(\Omega)} \|\widetilde{u}_{2n+2}^H\|_{L^2(\Omega)} + \frac{H\|b\|_{L^\infty(\Omega)}}{2} |w_n|_{H^1(\Omega)} |\widetilde{u}_{2n+2}^H|_{H^1(\Omega)} \\
&\quad + \beta |\widetilde{u}_{2n+1}^H|_{H^1(\Omega)} |\widetilde{u}_{2n+2}^H|_{H^1(\Omega)} \\
&\leqslant \left[\left(C_\Omega + \frac{H}{2}\right)\|b\|_{L^\infty(\Omega)} |w_n|_{H^1(\Omega)} + \beta |\widetilde{u}_{2n+1}^H|_{H^1(\Omega)}\right] |\widetilde{u}_{2n+2}^H|_{H^1(\Omega)}.
\end{aligned} \tag{2.96}$$

We now estimate $|w_n|_{H^1(\Omega)}$. We know that, for any $\psi \in V_H^\varepsilon$,

$$\begin{aligned} a_1(w_n,\psi) &= a_1(\widetilde{u}_{2n}^H - \widetilde{u}_{2n+1}^H,\psi) \\ &= a_2(\widetilde{u}_{2n+1}^H,\psi) - a_1(\widetilde{u}_{2n+1}^H,\psi) \\ &= \int_\Omega (\nabla\psi)^T (A^\varepsilon - \alpha_{\rm spl}{\rm Id})\nabla\widetilde{u}_{2n+1}^H, \end{aligned}$$

where we used (2.80) in the first line and (2.81) in the second line. Using that $w_n \in V_H^\varepsilon$ (this is where using the projection $P_{V_H^\varepsilon}$ is needed), we deduce that

$$|w_n|_{H^1(\Omega)} \leqslant \frac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{\beta + \alpha_{\rm spl}} |\widetilde{u}_{2n+1}^H|_{H^1(\Omega)}. \tag{2.97}$$

Collecting (2.96), (2.97) and (2.95), we obtain

$$|\widetilde{u}_{2n+2}^H|_{H^1(\Omega)} \leqslant \rho |\widetilde{u}_{2n}^H|_{H^1(\Omega)}, \tag{2.98}$$

where $\rho = \left(C_\Omega + \dfrac{H}{2}\right) \dfrac{\|b\|_{L^\infty(\Omega)}}{\beta+\alpha_{\rm spl}} \dfrac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{\beta+\alpha_1} + \dfrac{\beta}{\beta+\alpha_1}$.

As in the proof of Lemma 2.18, we introduce the function

$$g(x) = \left(C_\Omega + \frac{H}{2}\right) \frac{\|b\|_{L^\infty(\Omega)}}{x+\alpha_{\rm spl}} \frac{\|A^\varepsilon - \alpha_{\rm spl}{\rm Id}\|_{L^\infty(\Omega)}}{x+\alpha_1} + \frac{x}{x+\alpha_1},$$

observe that $g(x) = 1 - \dfrac{\alpha_1}{x} + O\left(\dfrac{1}{x^2}\right)$, which implies, since $\alpha_1 > 0$, that $\min\limits_{x\geqslant 0} g(x) < 1$. In view of (2.85), we have $\rho = g(\beta) = \min\limits_{x\geqslant 0} g(x) < 1$.

Arguing as in the proof of Lemma 2.16, we obtain that (u_{2n}^H, u_{2n+1}^H) converges in $H_0^1(\Omega)\times H_0^1(\Omega)$ to some $(u_{\rm even}^H, u_{\rm odd}^H) \in V_H \times V_H^\varepsilon$. Letting n go to $+\infty$ in (2.77) and (2.81), we obtain that $u_{\rm even}^H$ and $u_{\rm odd}^H$ satisfy the variational formulations (2.83) and (2.84).

Step 2: Convergence when $H \to 0$. We recast (2.83)–(2.84) as the following variational formulation:

$$\begin{gathered} \text{Find } (u_{\rm even}^H, u_{\rm odd}^H) \in V_H \times V_H^\varepsilon \text{ such that, for any } (v,w) \in V_H \times V_H^\varepsilon, \\ c_H\Big((u_{\rm even}^H, u_{\rm odd}^H),(v,w)\Big) = B_H(v,w), \end{gathered} \tag{2.99}$$

where the bilinear form

$$\begin{aligned} c_H\Big((u_{\rm even}^H, u_{\rm odd}^H),(v,w)\Big) \;=\; & a_1(u_{\rm even}^H, v) + a_{\rm conv}(u_{\rm even}^H, v) \\ & - \int_\Omega \beta\nabla u_{\rm odd}^H \cdot \nabla v + a_{\rm conv}(u_{\rm odd}^H, v) \\ & - a_{\rm conv}(P_{V_H^\varepsilon}(u_{\rm even}^H), v) + a_2(u_{\rm odd}^H, w) - a_1(u_{\rm even}^H, w) \end{aligned}$$

is defined on $\Big(H_0^1(\Omega)\times H_0^1(\Omega)\Big)^2$. Recall that a_1, $a_{\rm conv}$ and a_2 are defined by (2.78), (2.79) and (2.82), respectively, while the operator $P_{V_H^\varepsilon}$ is defined by (2.80). The linear form

$$B_H(v,w) = \int_\Omega fv + \sum_{\mathbf{K}\in\mathcal{T}_H} (\tau_{\mathbf{K}} f, b\cdot\nabla v)_{L^2(\mathbf{K})}$$

is defined on $H_0^1(\Omega) \times H_0^1(\Omega)$. Note that $B_H(v,w)$ does not depend on w.

The convergence proof when $H \to 0$ is based on the following arguments. First, we are going to show that, if H is sufficiently small, c_H satisfies an inf-sup condition uniformly in the mesh size H. For that purpose, we adapt the arguments of [84, Theorem 4.2.9] to our setting. We introduce the bilinear form

$$\widetilde{c}_H\Big((u,v),(\phi,\psi)\Big) = c_H\Big((u,v),(\phi,\psi)\Big) + \lambda \int_\Omega u\,\phi,$$

defined on $\Big(H_0^1(\Omega) \times H_0^1(\Omega)\Big)^2$, where $\lambda > 0$ is a parameter, and show in Step 2a below that $\widetilde{c}_H$ is coercive in the $H^1(\Omega) \times H^1(\Omega)$ norm, provided λ is large enough and H is sufficiently small. This allows us to next show, as claimed above, that c_H satisfies the inf-sup condition (see Step 2b), uniformly in H (as soon as H is sufficiently small). In contrast to the setting of [84, Theorem 4.2.9], the bilinear forms c_H and $\widetilde{c}_H$ here depend on H.

We are then in position to use classical numerical analysis arguments (see Step 2c) for estimating the discretization error (see (2.114) below). This error is bounded from above (up to some multiplicative constants) by the best approximation error and by the error introduced by the fact that c_H and B_H in (2.99) depend on H. The end of the proof amounts to showing that these two errors converge to 0 when $H \to 0$.

Step 2a: Coercivity of $\widetilde{c}_H$. We assume that

$$\lambda \geqslant \frac{4\|b\|^2_{L^\infty(\Omega)}}{\alpha_{\rm spl}} + \frac{\|b\|^2_{L^\infty(\Omega)}}{2\,(\alpha_1 - \alpha_{\rm spl}/2)} \quad \text{and} \quad H\|b\|_{L^\infty(\Omega)} < \min\Big(2\alpha_1 - \alpha_{\rm spl}, \frac{\alpha_{\rm spl}}{5}\Big) \tag{2.100}$$

where we recall that α_1 is such that (2.18) holds and that $\alpha_{\rm spl}$ is such that $2\alpha_1 > \alpha_{\rm spl}$ (see (2.86)). We claim that

$$\text{Under assumption (2.100), } \widetilde{c}_H \text{ is coercive.} \tag{2.101}$$

Note that (2.100) does not impose any restriction on β. The assumption (2.85) is only used in Step 1 above (to show the convergence when $n \to \infty$).

For any $(u,v) \in H_0^1(\Omega) \times H_0^1(\Omega)$, we have

$$\begin{aligned}\widetilde{c}_H\Big((u,v),(u,v)\Big) \;=\;& a_1(u,u) + a_2(v,v) + \lambda\|u\|^2_{L^2(\Omega)} \\ & - \int_\Omega \beta \nabla v \cdot \nabla u - a_1(u,v) \\ & + a_{\rm conv}(u - P_{V_H^\varepsilon}(u), u) + a_{\rm conv}(v,u).\end{aligned} \tag{2.102}$$

Using the coercivity of the bilinear forms a_1 and a_2, we get

$$a_1(u,u) + a_2(v,v) \geqslant (\beta + \alpha_{\rm spl})|u|^2_{H^1(\Omega)} + (\beta + \alpha_1)|v|^2_{H^1(\Omega)}. \tag{2.103}$$

We now bound the terms in the last line of (2.102). Using the fact that $\tau_{\mathbf{K}}(x) = \dfrac{H}{2|b(x)|}$ and $|P_{V_H^\varepsilon}(u)|_{H^1(\Omega)} \leqslant |u|_{H^1(\Omega)}$, we have that

$$\begin{aligned}
& |a_{\mathrm{conv}}(u - P_{V_H^\varepsilon}(u), u) + a_{\mathrm{conv}}(v, u)| \\
\leqslant\ & \|b\|_{L^\infty(\Omega)} |u - P_{V_H^\varepsilon}(u)|_{H^1(\Omega)} \|u\|_{L^2(\Omega)} + \frac{H}{2} \|b\|_{L^\infty(\Omega)} |u - P_{V_H^\varepsilon}(u)|_{H^1(\Omega)} |u|_{H^1(\Omega)} \\
& + \|b\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} \|u\|_{L^2(\Omega)} + \frac{H}{2} \|b\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} |u|_{H^1(\Omega)} \\
\leqslant\ & 2\|b\|_{L^\infty(\Omega)} |u|_{H^1(\Omega)} \|u\|_{L^2(\Omega)} + \frac{5H}{4} \|b\|_{L^\infty(\Omega)} |u|^2_{H^1(\Omega)} \\
& + \|b\|_{L^\infty(\Omega)} |v|_{H^1(\Omega)} \|u\|_{L^2(\Omega)} + \frac{H}{4} \|b\|_{L^\infty(\Omega)} |v|^2_{H^1(\Omega)} \\
\leqslant\ & \left(\frac{\alpha_{\mathrm{spl}}}{4} + \frac{5H}{4} \|b\|_{L^\infty(\Omega)} \right) |u|^2_{H^1(\Omega)} + \left(\frac{4\|b\|^2_{L^\infty(\Omega)}}{\alpha_{\mathrm{spl}}} + \frac{\|b\|^2_{L^\infty(\Omega)}}{2\left(\alpha_1 - \alpha_{\mathrm{spl}}/2\right)} \right) \|u\|^2_{L^2(\Omega)} \\
& + \left(\frac{1}{2}\left(\alpha_1 - \frac{\alpha_{\mathrm{spl}}}{2}\right) + \frac{H}{4} \|b\|_{L^\infty(\Omega)} \right) |v|^2_{H^1(\Omega)},
\end{aligned} \tag{2.104}$$

where we have used a Young inequality in the last line. We bound the terms in the second line of (2.102) by

$$\left| - \int_\Omega \beta \nabla v \cdot \nabla u - a_1(u, v) \right| \leqslant \frac{\beta}{2}(|u|^2_{H^1(\Omega)} + |v|^2_{H^1(\Omega)}) + \frac{\beta + \alpha_{\mathrm{spl}}}{2}(|u|^2_{H^1(\Omega)} + |v|^2_{H^1(\Omega)}). \tag{2.105}$$

Collecting (2.102), (2.103), (2.104) and (2.105), we get

$$\begin{aligned}
\widetilde{c}_H\Big((u, v), (u, v)\Big) \ \geqslant\ & \left(\frac{\alpha_{\mathrm{spl}}}{4} - \frac{5H}{4} \|b\|_{L^\infty(\Omega)} \right) |u|^2_{H^1(\Omega)} \\
& + \left(\frac{1}{2}\left(\alpha_1 - \frac{\alpha_{\mathrm{spl}}}{2}\right) - \frac{H}{4} \|b\|_{L^\infty(\Omega)} \right) |v|^2_{H^1(\Omega)} \\
& + \left(\lambda - \frac{4\|b\|^2_{L^\infty(\Omega)}}{\alpha_{\mathrm{spl}}} - \frac{\|b\|^2_{L^\infty(\Omega)}}{2\left(\alpha_1 - \alpha_{\mathrm{spl}}/2\right)} \right) \|u\|^2_{L^2(\Omega)}.
\end{aligned}$$

Under assumption (2.100), using a Poincaré inequality, we see that there exists $\eta > 0$ such that

$$\forall (u, v) \in H^1_0(\Omega) \times H^1_0(\Omega), \quad \widetilde{c}_H\Big((u, v), (u, v)\Big) \geqslant \eta \left(\|u\|^2_{H^1(\Omega)} + \|v\|^2_{H^1(\Omega)} \right). \tag{2.106}$$

This concludes the proof of the claim (2.101).

Step 2b: Inf-sup condition on c_H. We want to show that there exists $H_0 > 0$ and $\alpha > 0$ such that, for any $H \leqslant H_0$,

$$\inf_{U^H \in V_H \times V_H^\varepsilon} \sup_{\Phi^H \in V_H \times V_H^\varepsilon} \frac{c_H\left(U^H, \Phi^H\right)}{\|U^H\|_{H^1(\Omega)\times H^1(\Omega)} \|\Phi^H\|_{H^1(\Omega)\times H^1(\Omega)}} \geqslant \alpha. \tag{2.107}$$

We prove this statement by contradiction and therefore assume that (2.107) does not hold. Then, there exists a sequence H_n that converges to 0 and a sequence $U^{H_n} = \left(u^{H_n}_{\mathrm{even}}, u^{H_n}_{\mathrm{odd}}\right) \in V_{H_n} \times V^\varepsilon_{H_n}$ with $\|U^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)} = 1$, such that

$$\lim_{n \to +\infty} \sup_{\Phi \in V_{H_n} \times V^\varepsilon_{H_n}} \frac{c_{H_n}\left(U^{H_n}, \Phi\right)}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}} = 0. \tag{2.108}$$

As the sequence U^{H_n} is bounded in $H^1(\Omega)\times H^1(\Omega)$, it weakly converges in $H^1_0(\Omega)\times H^1_0(\Omega)$ to some $U^\star=(u^\star_{\rm even},u^\star_{\rm odd})\in H^1_0(\Omega)\times H^1_0(\Omega)$, up to the extraction of a subsequence that we still denote U^{H_n}.

Using (2.80), we also deduce from the boundedness of $u^{H_n}_{\rm even}$ that $P_{V^\varepsilon_{H_n}}(u^{H_n}_{\rm even})$ is bounded in H^1 norm. Up to an additional extraction, we hence have that $P_{V^\varepsilon_{H_n}}(u^{H_n}_{\rm even})$ weakly converges in $H^1(\Omega)$ to some $u^\Pi_{\rm even}$. We claim that $u^\Pi_{\rm even}=u^\star_{\rm even}$. Let indeed $\phi\in H^1_0(\Omega)$. By density (see Appendix 2.5.2), there exists a sequence $w_n\in V^\varepsilon_{H_n}$ converging strongly in $H^1_0(\Omega)$ to ϕ. For any n, we have $a_1(P_{V^\varepsilon_{H_n}}(u^{H_n}_{\rm even}),w_n)=a_1(u^{H_n}_{\rm even},w_n)$. Passing to the limit $n\to\infty$, we infer that $a_1(u^\Pi_{\rm even},\phi)=a_1(u^\star_{\rm even},\phi)$, which holds true for any $\phi\in H^1_0(\Omega)$. This implies that $u^\Pi_{\rm even}=u^\star_{\rm even}$. Consequently, $P_{V^\varepsilon_{H_n}}(u^{H_n}_{\rm even})-u^{H_n}_{\rm even}$ weakly converges in $H^1_0(\Omega)$ to 0.

We first show that $U^\star=0$. We fix some $\Phi=(\phi,\psi)\in H^1_0(\Omega)\times H^1_0(\Omega)$. For any $\Phi^{H_n}=(\phi^{H_n},\psi^{H_n})\in V_{H_n}\times V^\varepsilon_{H_n}$, we write

$$c_{H_n}(U^{H_n},\Phi)=c_{H_n}(U^{H_n},\Phi^{H_n})+c_{H_n}(U^{H_n},\Phi-\Phi^{H_n}). \tag{2.109}$$

We have that

$$|c_{H_n}(U^{H_n},\Phi^{H_n})|\leqslant\left(\sup_{\Psi\in V_{H_n}\times V^\varepsilon_{H_n}}\frac{c_{H_n}(U^{H_n},\Psi)}{\|\Psi\|_{H^1(\Omega)\times H^1(\Omega)}}\right)\|\Phi^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)} \tag{2.110}$$

and, since c_H is a continuous bilinear form,

$$|c_{H_n}(U^{H_n},\Phi-\Phi^{H_n})|\leqslant M\|U^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)}\,\|\Phi-\Phi^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)}. \tag{2.111}$$

By an argument of density (see Appendix 2.5.2), there exists a sequence $\psi^{H_n}\in V^\varepsilon_{H_n}$ converging strongly in $H^1_0(\Omega)$ to ψ. Likewise, by an argument of density and classical results on finite element methods, we know that there also exists a sequence $\phi^{H_n}\in V_{H_n}$ converging strongly in $H^1_0(\Omega)$ to ϕ. We thus have built a sequence $\Phi^{H_n}=\left(\phi^{H_n},\psi^{H_n}\right)\in V_{H_n}\times V^\varepsilon_{H_n}$ such that $\lim_{n\to+\infty}\|\Phi-\Phi^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)}=0$. We infer from (2.109), (2.111), (2.110) and (2.108) that

$$\lim_{n\to+\infty}c_{H_n}(U^{H_n},\Phi)=0.$$

Making use of the explicit expression of c_H and using that U^{H_n} weakly converges in $H^1_0(\Omega)\times H^1_0(\Omega)$ to $U^\star=(u^\star_{\rm even},u^\star_{\rm odd})$ and that $P_{V^\varepsilon_{H_n}}(u^{H_n}_{\rm even})-u^{H_n}_{\rm even}$ weakly converges in $H^1_0(\Omega)$ to 0, we obtain that

$$a_1(u^\star_{\rm even},\phi)-\int_\Omega\beta\nabla u^\star_{\rm odd}\cdot\nabla\phi+\int_\Omega(b\cdot\nabla u^\star_{\rm odd})\phi+a_2(u^\star_{\rm odd},\psi)-a_1(u^\star_{\rm even},\psi)=0.$$

This holds for any $(\phi,\psi)\in H^1_0(\Omega)\times H^1_0(\Omega)$. Taking $\phi=\psi$, we deduce that $u^\star_{\rm odd}=0$. This next implies that $u^\star_{\rm even}=0$, and hence that $U^\star=0$.

Second, we show the *strong* convergence in $H^1_0(\Omega)\times H^1_0(\Omega)$ of the sequence U^{H_n} to $U^\star=0$. Under assumption (2.100), we have shown in Step 2a above that $\widetilde{c}_H$ is coercive. In view of (2.106),

we thus have

$$\begin{aligned}\eta\|U^{H_n}\|^2_{H^1(\Omega)\times H^1(\Omega)} &\leqslant \widetilde{c}_{H_n}(U^{H_n},U^{H_n})\\ &= c_{H_n}(U^{H_n},U^{H_n}) + \lambda\int_\Omega \left(u^{H_n}_{\rm even}\right)^2\\ &\leqslant \left(\sup_{\Phi\in V_{H_n}\times V^\varepsilon_{H_n}} \frac{c_{H_n}(U^{H_n},\Phi)}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}}\right) + \lambda\,\|u^{H_n}_{\rm even}\|^2_{L^2(\Omega)}.\end{aligned}$$

In view of (2.108), the first term in the above right-hand side converges to 0 when $n\to\infty$. Up to the extraction of a subsequence, $u^{H_n}_{\rm even}$ (which weakly converges to 0 in $H^1(\Omega)$) strongly converges to 0 in $L^2(\Omega)$. This implies that the second term in the above right-hand side also converges to 0 when $n\to\infty$.

We then deduce that $\lim_{n\to\infty}\|U^{H_n}\|^2_{H^1(\Omega)\times H^1(\Omega)} = 0$, which is a contradiction with the fact that, by construction, $\|U^{H_n}\|_{H^1(\Omega)\times H^1(\Omega)} = 1$. This concludes the proof of (2.107).

Step 2c: Conclusion. We are now in position to use [38, Lemma 2.27], which states an upper bound on the error (see (2.114) below) under three assumptions. Assumption (i) of that lemma is that the approximation spaces are conformal. This is obviously satisfied here, as $V_H\times V^\varepsilon_H\subset H^1_0(\Omega)\times H^1_0(\Omega)$. Assumption (ii) is that c_H satisfies an inf-sup condition. It is satisfied here in view of (2.107). Assumption (iii) is that the bilinear form c_H is bounded. This is again satisfied here. The assumptions of [38, Lemma 2.27] being satisfied, we can write an error bound (see (2.114) below) between the solution to (2.99) and the solution to the corresponding infinite dimensional problem, that reads

$$\begin{aligned}&\text{Find } (u_{\rm even},u_{\rm odd})\in H^1_0(\Omega)\times H^1_0(\Omega) \text{ such that,}\\ &\text{for any } (v,w)\in H^1_0(\Omega)\times H^1_0(\Omega),\quad c\Big((u_{\rm even},u_{\rm odd}),(v,w)\Big) = B(v,w),\end{aligned}\tag{2.112}$$

where

$$\begin{aligned}c\Big((u_{\rm even},u_{\rm odd}),(v,w)\Big) = a_1(u_{\rm even},v) - \int_\Omega \beta\nabla u_{\rm odd}\cdot\nabla v + \int_\Omega (b\cdot\nabla u_{\rm odd})v\\ + a_2(u_{\rm odd},w) - a_1(u_{\rm even},w)\end{aligned}$$

and

$$B(v,w) = \int_\Omega fv.$$

It is obvious that $(u_{\rm even},u_{\rm odd})$ is a solution to (2.112) if and only if $(u_{\rm even},u_{\rm odd})$ is a solution to the system

$$\begin{cases} -(\beta+\alpha_{\rm spl})\Delta u_{\rm even} = f - b\cdot\nabla u_{\rm odd} - \beta\Delta u_{\rm odd} & \text{in } \Omega,\\ u_{\rm even} = 0 \quad \text{on } \partial\Omega,\end{cases}$$
$$\begin{cases} -\operatorname{div}\left((\beta{\rm Id} + A^\varepsilon)\nabla u_{\rm odd}\right) = -(\beta+\alpha_{\rm spl})\Delta u_{\rm even} & \text{in } \Omega,\\ u_{\rm odd} = 0 \quad \text{on } \partial\Omega.\end{cases}\tag{2.113}$$

This system is well-posed: by adding the two equations, we obtain that $u_{\rm odd}$ is a solution to (2.23), and is therefore unique. This implies the uniqueness of $u_{\rm even}$ in view of (2.113). We denote by $U = (u_{\rm even},u_{\rm odd})$ the unique solution to (2.112).

Using [38, Lemma 2.27], we obtain that

$$\|U-U^H\|_{H^1(\Omega)\times H^1(\Omega)} \leqslant \frac{1}{\alpha}\sup_{\Phi\in V_H\times V_H^\varepsilon}\frac{|B(\Phi)-B_H(\Phi)|}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}} + \inf_{G\in V_H\times V_H^\varepsilon}\left[\left(1+\frac{M}{\alpha}\right)\|U-G\|_{H^1(\Omega)\times H^1(\Omega)} + \frac{1}{\alpha}\sup_{\Phi\in V_H\times V_H^\varepsilon}\frac{|c(G,\Phi)-c_H(G,\Phi)|}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}}\right], \quad (2.114)$$

where M is the continuity constant of the bilinear form c. We successively study the two terms in the right-hand side of (2.114).

For the first term, we write, for any $\Phi=(\phi,\psi)\in V_H\times V_H^\varepsilon$, that

$$|B(\Phi)-B_H(\Phi)| \leqslant \frac{H}{2}\sum_{\mathbf{K}\in\mathcal{T}_H}\|f\|_{L^2(\mathbf{K})}\|\nabla\phi\|_{L^2(\mathbf{K})} \leqslant \frac{H}{2}\|f\|_{L^2(\Omega)}\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)},$$

which implies that

$$\lim_{H\to 0}\sup_{\Phi\in V_H\times V_H^\varepsilon}\frac{|B(\Phi)-B_H(\Phi)|}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}} = 0. \quad (2.115)$$

For the second term of the right-hand side of (2.114), we write, for any $\Phi=(\phi,\psi)\in V_H\times V_H^\varepsilon$ and any $G=(g,h)\in V_H\times V_H^\varepsilon$, that

$$c_H(G,\Phi)-c(G,\Phi) = a_{\rm conv}(g-P_{V_H^\varepsilon}(g),\phi) + \sum_{\mathbf{K}\in\mathcal{T}_H}(\tau_{\mathbf{K}}b\cdot\nabla h, b\cdot\nabla\phi)_{L^2(\mathbf{K})}.$$

We therefore deduce, using an integration by parts in the first line, that

$$\begin{aligned}
&|c_H(G,\Phi)-c(G,\Phi)|\\
\leqslant\ &\left|\int_\Omega\left[g-P_{V_H^\varepsilon}(g)\right]b\cdot\nabla\phi\right|\\
&+\frac{H\|b\|_{L^\infty(\Omega)}}{2}|g-P_{V_H^\varepsilon}(g)|_{H^1(\Omega)}|\phi|_{H^1(\Omega)} + \frac{H\|b\|_{L^\infty(\Omega)}}{2}|h|_{H^1(\Omega)}|\phi|_{H^1(\Omega)}\\
\leqslant\ &\|b\|_{L^\infty(\Omega)}\|g-P_{V_H^\varepsilon}(g)\|_{L^2(\Omega)}\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}\\
&+H\|b\|_{L^\infty(\Omega)}\left(|g|_{H^1(\Omega)}+|h|_{H^1(\Omega)}\right)\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}.
\end{aligned}$$

We hence write, for the second term of the right-hand side of (2.114), that

$$\begin{aligned}
&\left(1+\frac{M}{\alpha}\right)\|U-G\|_{H^1(\Omega)\times H^1(\Omega)} + \frac{1}{\alpha}\sup_{\Phi\in V_H\times V_H^\varepsilon}\frac{|c(G,\Phi)-c_H(G,\Phi)|}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}}\\
&\qquad\leqslant \mathcal{C}\|U-G\|_{H^1(\Omega)\times H^1(\Omega)} + \mathcal{C}\|g-P_{V_H^\varepsilon}(g)\|_{L^2(\Omega)} + \mathcal{C}H\|G\|_{H^1(\Omega)\times H^1(\Omega)}
\end{aligned}$$

where $\mathcal{C}$ is independent of H. Using the density of the families V_H and V_H^ε in $H_0^1(\Omega)$ (see Appendix 2.5.2 for the latter property), we build $G^H=(g^H,h^H)\in V_H\times V_H^\varepsilon$ such that $\lim_{H\to 0}\|U-G^H\|_{H^1(\Omega)\times H^1(\Omega)}=0$. We thus have that

$$\begin{aligned}
&\inf_{G\in V_H\times V_H^\varepsilon}\left[\left(1+\frac{M}{\alpha}\right)\|U-G\|_{H^1(\Omega)\times H^1(\Omega)} + \frac{1}{\alpha}\sup_{\Phi\in V_H\times V_H^\varepsilon}\frac{|c(G,\Phi)-c_H(G,\Phi)|}{\|\Phi\|_{H^1(\Omega)\times H^1(\Omega)}}\right]\\
&\qquad\leqslant \mathcal{C}\|U-G^H\|_{H^1(\Omega)\times H^1(\Omega)} + \mathcal{C}\|g^H-P_{V_H^\varepsilon}(g^H)\|_{L^2(\Omega)} + \mathcal{C}H\|G^H\|_{H^1(\Omega)\times H^1(\Omega)}.
\end{aligned}$$

The above three terms converge to 0 when $H \to 0$ (for the second term, this is a consequence of the fact that, for any bounded sequence $\tau_H \in H_0^1(\Omega)$, we have that $\tau^H - P_{V_H^\varepsilon}(\tau^H)$ weakly converges to 0 in $H^1(\Omega)$). Collecting this result with (2.114) and (2.115), we deduce that $\lim_{H\to 0} \|U - U^H\|_{H^1(\Omega)\times H^1(\Omega)} = 0$. This concludes the proof of Lemma 2.20.

Chapter 3

Stable approximation of the advection-diffusion equation using the invariant measure

Ce chapitre reprend l'intégralité d'un article écrit en collaboration avec Claude Le Bris et Frédéric Legoll et soumis pour publication [62]. On inclut dans l'annexe B une note résumant le contenu de ce chapitre, écrite en collaboration avec Claude Le Bris et Frédéric Legoll, et publiée dans Comptes Rendus Mathématique [61].

3.1 Introduction and motivation

Our purpose is to study the advection-diffusion equation

$$-\Delta u + b \cdot \nabla u = f \quad \text{in } \Omega, \tag{3.1}$$

more specifically in the regime where it is both non-coercive and possibly unstable. We present a new numerical strategy, based upon the utilization of the invariant measure associated to (3.1), namely the solution σ to the adjoint equation

$$-\operatorname{div}(\nabla\sigma + b\sigma) = 0 \quad \text{in } \Omega, \tag{3.2}$$

supplied with suitable boundary conditions and normalization constraints.

Equation (3.1) arises in a huge variety of contexts of the engineering sciences, either in its stationary form (3.1), or in its time-dependent form. It may also contain a reaction term $c\,u$, with $c \geqslant 0$. The second order (diffusion) operator can be chosen more general than a pure Laplacian, in the form of a divergence operator $-\operatorname{div}(A\,\nabla\cdot)$, with a suitable matrix-valued function A. All such situations proceed from straightforward applications of our discussions below, which, for brevity and clarity, we limit to the simple, stationary case (3.1).

A typical difficulty associated with equation (3.1) is the possible lack of coercivity of the bilinear form, owing to the presence of the advection term $b \cdot \nabla u$. More severely, not only coercivity, but also stability may be affected by the advection term, when the latter is "large" in a certain sense. The equation is then said advection-dominated. Studies abound in the literature, that describe the theory necessary to prove well-posedness of that problem under those difficult circumstances. Similarly, the works presenting possible numerical discretization techniques specifically targeted to this context are countless. Our purpose here is to propose yet another way of addressing the difficulties mentioned above. The computational approach we present actually originates from the theoretical proof of well-posedness of the problem.

As is well-known, a classical proof of the well-posedness of (3.1), when it is not coercive, proceeds by the Fredholm alternative. On the other hand, well-posedness is also, using the Banach-Nečas-Babuška Theorem, equivalent to a set of two conditions, namely the inf-sup condition

$$\exists \alpha > 0 \quad \text{such that} \quad \inf_{w \in W} \sup_{v \in V} \frac{a(w,v)}{\|w\|_W \|v\|_V} \geqslant \alpha > 0 \tag{3.3}$$

where

$$a(w,v) = \int_\Omega \nabla w \cdot \nabla v + \int_\Omega (b \cdot \nabla w)\, v \tag{3.4}$$

and $V = W = H^1_0(\Omega)$, and the additional condition

$$\forall v \in V, \quad (\forall w \in W,\ a(w,v) = 0) \Rightarrow v = 0. \tag{3.5}$$

We refer to, e.g., [38, p. 85] for a comprehensive exposition of the theory and general references therein for the study and approximation of (3.1). We shall recall some basic facts in Section 3.2.1 of this article.

In practice, two questions arise: to prove that the above conditions (3.3) and (3.5) hold, thereby providing a proof of well-posedness that is independent from Fredholm theory, and to make the constant α in (3.3) explicit. For this twofold purpose, the classical argument is to consider

the invariant measure associated to (3.1), namely the solution σ to (3.2), satisfying $\sigma(x) \geqslant \underline{\sigma} > 0$ almost everywhere in Ω, $|\Omega|^{-1} \int_\Omega \sigma = 1$, along with an adequate boundary condition (think of the natural Neumann boundary condition, but other boundary conditions might be considered, in particular because of the various boundary conditions (3.1) itself may be supplied with). The existence and uniqueness of a suitable σ follows from the Fredholm theory. In short, (3.3) is then obtained as follows. One multiplies (3.1) by the product $\sigma\, v$ and integrates over the domain Ω:

$$\int_\Omega (-\Delta u + b \cdot \nabla u)\, \sigma v = \int_\Omega \sigma\, \nabla u \cdot \nabla v + \int_\Omega (\nabla \sigma + \sigma\, b) \cdot \nabla u\, v - \int_{\partial\Omega} (\nabla u \cdot n)\, \sigma\, v. \tag{3.6}$$

The rightmost term cancels out when homogeneous Dirichlet boundary conditions are imposed on u (and thus on the test function v), which is the setting we adopt throughout this article. Considering (3.2), this formally yields

$$a(u, \sigma\, u) = \int_\Omega \sigma\, |\nabla u|^2 \quad \text{for any } u \in H^1_0(\Omega), \tag{3.7}$$

which readily implies (3.3), as soon as σ is positive and bounded away from zero. The classical approach is then to use a finite element discretization that also satisfies the inf-sup condition (3.3), at least for a sufficiently small mesh size h, and is therefore, "by continuity", also well-posed, using the same type of argument.

The above observation on how the consideration of the invariant measure allows to transform the original problem (3.1) into a coercive problem seems to not have been exploited computationally (except in the very specific case when b is irrotational [18], for which σ is then analytically known). This is our purpose to do so. Formally, (3.6) and (3.7) suggest a Petrov-Galerkin formulation of the problem using test functions of the form $\sigma\, v$ instead of a classical Galerkin formulation. (Formally) equivalently, one may perform a Galerkin approximation of the modified equation

$$-\operatorname{div}(\sigma\, \nabla u) + (\nabla \sigma + \sigma\, b) \cdot \nabla u = \sigma\, f. \tag{3.8}$$

The point is of course that the modified advection field

$$B = \nabla \sigma + \sigma\, b$$

is divergence-free because of (3.2). Problem (3.8), complemented by homogeneous Dirichlet boundary conditions, is consequently coercive. And the numerical analysis of its finite element approximation is amenable to standard arguments.

The definite added value of the approach is that it provides an unconditionally well-posed approximation, irrespective of the discretization parameter –the meshsize– adopted for approximating u, provided σ itself is correctly approximated (which in particular implies that some positivity of σ is preserved at the discrete level). This unconditional well-posedness may be most useful in problems where one can only afford a coarse approximation of u. Multiscale problems, where the Laplacian operator is replaced by $-\operatorname{div}(a(x/\varepsilon)\, \nabla\, \cdot)$, are prototypical examples of such a context. Problems such as inverse problems, or time-dependent problems (once semi-discretized in time using, say, an implicit Euler scheme), where the solution to the advection-diffusion equation is required repeatedly, are also problems of choice for the approach. A rather coarse approximation might be employed, while the additional computational workload to solve the

adjoint equation (3.2) is required only once. A definite improvement of the total computational time may be observed. In addition, in the advection dominated context, the approach enjoys particular stability properties that lead to numerical results qualitatively comparable to those obtained with classical, state-of-the-art stabilization approaches [39, 54, 78]. Our results show that the approach is accurate, robust and can be made effective in terms of computational cost. Applications to several other, more general contexts, may be envisioned.

On the theoretical level, one advantage of the approach is that we can establish (see Section 3.3) a complete numerical analysis. In contrast, and to the best of our knowledge, the added value of a classical *stabilized* finite element approximation is not proven theoretically when the advection-diffusion equation is *not* coercive.

Our article is articulated as follows. Section 3.2 collects some preparatory material. We need to recall a few results, first on the Banach-Nečas-Babuška theory and the inf-sup condition, and second on the invariant measures that may be associated to the problem. The former ones are very classical, and we briefly overview them in Section 3.2.1. The latter ones are slightly less standard and are the purpose of Section 3.2.2. Of course, the reader familiar with the theory may easily skip our recollection and directly proceed to Section 3.3, where we present the specific discretization we use and analyze theoretically our approximation strategy. Our main result is Theorem 3.19. An ingredient of our numerical analysis is the adaptation to the case of the advection-diffusion equation with invariant measure σ of classical arguments from [15] that allow for an error estimate in $W^{1,p}(\Omega)$, $p > 2$, of a $\mathbb{P}^1$ finite element approximation of σ (see Proposition 3.17 below). This extension is, in our opinion, nontrivial and has an interest on its own. We present its proof in Appendix 3.6. Our final Section 3.4 presents a comprehensive set of numerical tests that demonstrate the accuracy, stability and efficiency of the approach.

Our main conclusions are as follows. The approach based upon the precomputation of an approximation of σ solution to (3.2) and next the approximation of u as the solution to (3.8) provides with an approximation that is (i) well-posed unconditionally in the meshsize and amenable to a precise numerical analysis of convergence, and (ii) as stable and accurate as a direct typical stabilized solution of (3.1). It is slightly more robust with respect to the mesh size used for the approximation of u. If the precomputation time for σ is not accounted for, the approach has an equal computational cost. And if it is, then roughly ten solutions of the advection-diffusion equation are necessary to make the approach profitable. In many of the contexts we mentioned above, this is clearly the case.

In short, the approach presented here does not provide spectacular results but constitutes an interesting, certified and efficient alternative to more established approaches. It can be shown to outperform them in certain situations when only a coarse mesh can be afforded and/or when the advection-diffusion equation needs to be solved repeatedly.

The present work complements an earlier publication [61] which already summarized the approach. It provides the numerical analysis of the approach and extensive numerical tests to assess its performance.

3.2 Mathematical setting and theoretical results

We consider equation (3.1) on a domain Ω and for an advection field b that *at least* satisfy, throughout the article, the following two conditions:

$$\begin{cases} \Omega \text{ is an open bounded domain of } \mathbb{R}^d,\ d \geqslant 2; \\ b \in (L^\infty(\Omega))^d. \end{cases} \tag{3.9}$$

We additionnally assume, in this Section 3.2, that

$$\Omega \text{ is of class } \mathcal{C}^1. \tag{3.10}$$

For some of our results of Section 3.2, we will have to make stronger assumptions (see in particular (3.21) below).

The right-hand side f of equation (3.1) is assumed $H^{-1}(\Omega)$. Again, in some instances, we will need to assume a better regularity (typically $L^p(\Omega)$) on this function f.

The boundary conditions we supply (3.1) with affect the boundary conditions we need to impose in the definition of the invariant measure(s) we introduce. For simplicity, our discussion assumes homogeneous Dirichlet boundary conditions

$$u = 0 \quad \text{on } \partial\Omega, \tag{3.11}$$

although other boundary conditions may be considered. As the definition of the invariant measure σ is essentially a matter of integration by parts (in the spirit of (3.6) above), we leave to the reader the adaptation to other boundary conditions. We emphasize, however, that some of the arguments that follow might require some additional work.

For the sake of consistency, we are now about to recall a set of basic results we need, on the inf-sup theory and on the invariant measure. The results are in particular interesting to motivate the specific discretization approach we introduce. We reiterate that the reader familiar with the classical theory may easily skip the sequel and directly proceed to Section 3.3.

3.2.1 Inf-sup theory

The advection-diffusion equation (3.1) supplied with the data we have just described and the boundary condition (3.11) can be studied in the context of the Banach-Nečas-Babuška theory. Defining $U = V = H^1_0(\Omega)$,

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v + (b \cdot \nabla u)v, \tag{3.12}$$

$$F(v) = \langle f, v \rangle_{H^{-1}(\Omega), H^1_0(\Omega)}, \tag{3.13}$$

it leads to a particular case of the general variational formulation:

$$\text{Find } u \in U \text{ such that, for all } v \in V, \quad a(u, v) = F(v), \tag{3.14}$$

where U is a Banach space, V is a reflexive Banach space, $a \in \mathcal{L}(U \times V; \mathbb{R})$ and $F \in V'$. The well-posedness of (3.1)–(3.11), recast as (3.12)–(3.13)–(3.14), is known to be equivalent to the following two conditions:

(BNB1) There exists $\alpha > 0$ such that $\inf_{u\in H_0^1(\Omega)} \sup_{v\in H_0^1(\Omega)} \dfrac{a(u,v)}{\|u\|_{H^1(\Omega)}\|v\|_{H^1(\Omega)}} \geqslant \alpha$;

(BNB2) For all $v \in H_0^1(\Omega)$, $\big(\forall u \in H_0^1(\Omega),\ a(u,v) = 0\big) \Longrightarrow (v = 0)$.

Introducing a function σ satisfying

(C1) $-\mathrm{div}\,(\nabla\sigma + b\sigma) = 0$ in Ω,

(C2) $\inf_{\Omega} \sigma > 0$,

(C3) for all $u \in H_0^1(\Omega)$, ($\sigma\, u \in H_0^1(\Omega)$ and) $\|\sigma\, u\|_{H^1(\Omega)} \leqslant C_\sigma \|u\|_{H^1(\Omega)}$,

it is easy to see that the conditions (BNB1) and (BNB2) are satisfied in our case. Using (C3) and (C1), a simple calculation indeed yields, for any $u \in H_0^1(\Omega)$,

$$\begin{aligned} a(u,\sigma u) &= \int_\Omega \sigma\nabla u\cdot\nabla u - \int_\Omega \mathrm{div}\,(\nabla\sigma + b\sigma)\frac{u^2}{2} \\ &= \int_\Omega \sigma\nabla u\cdot\nabla u \\ &\geqslant \Big(\inf_\Omega \sigma\Big)\,\|\nabla u\|^2_{L^2(\Omega)} \\ &\geqslant C\,\Big(\inf_\Omega \sigma\Big)\,\|u\|^2_{H^1(\Omega)}. \end{aligned}$$

Using (C2) and (C3), we obtain

$$\frac{a(u,\sigma u)}{\|u\|_{H^1(\Omega)}\|\sigma u\|_{H^1(\Omega)}} \geqslant C\frac{(\inf_\Omega \sigma)}{C_\sigma} > 0$$

and thus the inf-sup inequality (BNB1). The second condition (BNB2) is a consequence of the maximum principle (see e.g. [40, Theorem 8.1]): a function $v \in H_0^1(\Omega)$ that satisfies $a(u,v) = 0$ for all $u \in H_0^1(\Omega)$ is a solution to $-\mathrm{div}\ (\nabla v + bv) = 0$ in Ω and therefore vanishes.

The following, very classical proposition collects the properties established.

Proposition 3.1. *We assume (3.9)–(3.10) and that there exists $\sigma \in H^1(\Omega)$ satisfying conditions (C1), (C2) and (C3). Then (3.1)–(3.11) is well-posed, that is, it has a unique solution in $H_0^1(\Omega)$, and the map $H^{-1}(\Omega) \ni f \mapsto u \in H_0^1(\Omega)$ is continuous.*

Simply observing that one may harmlessly multiply and divide by a function σ enjoying the properties (C2) and (C3), and that the following formal integration by parts holds

$$\begin{aligned} \int_\Omega (-\Delta u + b\cdot\nabla u)\,\sigma v &= -\int_{\partial\Omega} (\nabla u\cdot n)\,\sigma v + \int_\Omega \nabla u\cdot\nabla(\sigma v) + \int_\Omega (\sigma\, b\cdot\nabla u)\, v \\ &= -\int_{\partial\Omega} (\sigma\nabla u\cdot n)\, v + \int_\Omega (\sigma\nabla u)\cdot\nabla v + \int_\Omega ((\nabla\sigma + b\sigma)\cdot\nabla u)\, v \\ &= \int_\Omega (-\mathrm{div}(\sigma\nabla u))\, v + \int_\Omega ((\nabla\sigma + b\sigma)\cdot\nabla u)\, v, \end{aligned}$$

one readily obtains the following result.

Proposition 3.2. *Under the assumptions of Proposition 3.1, (3.1)–(3.11) is equivalent to the problem*

$$-\mathrm{div}(\sigma\nabla u) + (\nabla\sigma + b\sigma)\cdot\nabla u = \sigma f \quad \text{in } \Omega, \qquad u = 0 \quad \text{on } \partial\Omega, \tag{3.15}$$

which is therefore also well-posed.

Note that property (C1) is actually not required to show that (3.1)–(3.11) and (3.15) are equivalent. This equivalence thus also holds for an approximation of the invariant measure, a fact we will use in the sequel.

3.2.2 On the invariant measure

We now turn to elements of theory regarding the invariant measure σ solution to (3.2). As mentioned above, (3.2) does not completely characterize σ since a boundary condition and possibly an additional normalization need to be supplied. Because of the homogeneous Dirichlet boundary condition (3.11) we have imposed on u for (3.1), it turns out that we have some flexibility on the boundary condition we may impose on σ. In other situations, the integration by parts performed to employ the adjoint equation might require more stringent conditions on σ on the boundary. The adaptation is, as we said, left to the reader.

We are going to consider two different choices for σ, respectively studied in the next two sections.

First choice of invariant measure

The first case is (formally) defined by

$$\begin{cases} -\operatorname{div}(\nabla\sigma + b\sigma) = 0 & \text{in } \Omega, \\ (\nabla\sigma + b\sigma)\cdot n = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.16}$$

with the normalization constraint

$$\fint_\Omega \sigma := |\Omega|^{-1} \int_\Omega \sigma = 1, \tag{3.17}$$

and the property

$$\inf_\Omega \sigma \geqslant c > 0. \tag{3.18}$$

The existence (and uniqueness) of such a function σ, along with some additional regularity properties, is established below. In the subsequent sections, this function is denoted by σ_1.

The classical results concerning this case are contained in the following Lemma.

Lemma 3.3 (Theorem 1.1 of [34])**.** *In addition to* (3.9)–(3.10)*, we assume that* Ω *is connected (for the uniqueness part of our statements). Then*

i) There exists $\sigma \in H^1(\Omega)$*, unique up to a multiplicative constant, solution to* (3.16)*. Up to a change of sign, we may require* $\sigma > 0$ *a.e. on* Ω*;*

ii) Let $g \in (H^1(\Omega))'$*. The problem*

$$\begin{cases} \text{Find } v \in H^1(\Omega) \text{ such that, for any } \varphi \in H^1(\Omega), \\ \displaystyle\int_\Omega (\nabla\varphi)^T(\nabla v + bv) = \langle g, \varphi\rangle_{(H^1(\Omega))',H^1(\Omega)}, \end{cases} \tag{3.19}$$

admits at least one solution if and only if

$$\langle g, 1\rangle_{(H^1(\Omega))',H^1(\Omega)} = 0.$$

In this case, the set of all solutions is $v + \mathbb{R}\sigma$. *In addition, the application* $g \mapsto v$ *is a bounded linear map from*

$$V_{\#1} = \left\{ g \in (H^1(\Omega))', \quad \langle g, 1\rangle_{(H^1(\Omega))',H^1(\Omega)} = 0 \right\}$$

to

$$H^1_{\int=0}(\Omega) = \left\{ v \in H^1(\Omega), \quad \int_\Omega v = 0 \right\};$$

iii) Let $f \in (H^1(\Omega))'$. *The problem*

$$\begin{cases} \text{Find } u \in H^1(\Omega) \text{ such that, for any } \varphi \in H^1(\Omega), \\ \displaystyle\int_\Omega \nabla\varphi \cdot \nabla u + \varphi b \cdot \nabla u = \langle f, \varphi\rangle_{(H^1(\Omega))',H^1(\Omega)}, \end{cases} \tag{3.20}$$

admits at least one solution (which is unique up to the addition of a constant) if and only if

$$\langle f, \sigma\rangle_{(H^1(\Omega))',H^1(\Omega)} = 0.$$

In addition, the application $f \mapsto u$ *is a bounded linear map from*

$$V_{\#\sigma} = \left\{ f \in (H^1(\Omega))', \quad \langle f, \sigma\rangle_{(H^1(\Omega))',H^1(\Omega)} = 0 \right\}$$

to $H^1_{\int=0}(\Omega)$.

It is easily seen that all assertions are consequences of the Fredholm alternative. The positivity stated in (i) follows from the maximum principle. A bound from below on σ may be obtained using stronger assumptions. It is the purpose of the following lemma.

Lemma 3.4 (after Theorem 1 of [75]). *We assume* (3.9) *and that the domain* Ω *is of class* $\mathcal{C}^2$. *Then there exists a unique solution* $\sigma \in W^{1,p}(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$, *for all* $1 \leqslant p < +\infty$, *to* (3.16) *that satisfies* $\max_{\overline{\Omega}} \sigma = 1$. *In addition, there exists* $c > 0$ *so that*

$$\sigma \geqslant c \quad \text{in } \Omega.$$

This solution σ *satisfies the conditions* (C1), (C2) *and* (C3).

We immediately remark that the solution, the existence and uniqueness of which is established in Lemma 3.4, coïncides with one of the positive solutions dealt with in Lemma 3.3. We note also that a renormalization of that solution can be performed to comply with the constraint (3.17).

Before we are in position to state our main proposition regarding the measure defined in (3.16)–(3.17)–(3.18), we need to recall a technical lemma for the Neumann problem, and next to strengthen the assumptions (3.9)–(3.10). The technical lemma, which will be useful in our proof of Proposition 3.6 below, is the following.

Lemma 3.5 (Chapter I, Theorem 1.10 of [41]). *Let Ω be an open bounded domain of $\mathbb{R}^d$ with a $\mathcal{C}^{1,1}$ boundary. Let $1 < p < \infty$ and let u be a solution to*

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \nabla u \cdot n = g & \text{on } \partial\Omega, \end{cases}$$

where $f \in L^p(\Omega)$ and $g \in W^{1-1/p,p}(\partial\Omega)$ satisfy the relation $\int_\Omega f + \int_{\partial\Omega} g = 0$. Then $u \in W^{2,p}(\Omega)$ and there exists a constant C, depending on p and Ω but independent from f, g and u, such that

$$\left\| u - \fint_\Omega u \right\|_{W^{2,p}(\Omega)} \leqslant C\Big(\|f\|_{L^p(\Omega)} + \|g\|_{W^{1-1/p,p}(\partial\Omega)} \Big).$$

In the sequel of this Section 3.2, we assume that

$$\begin{cases} \text{the domain } \Omega \text{ is connected, and of class } \mathcal{C}^2; \\ b \text{ is Lipschitz-continuous on } \overline{\Omega}. \end{cases} \tag{3.21}$$

We have the following result:

Proposition 3.6. *Under assumptions (3.9)–(3.21), there exists a unique solution $\sigma \in H^1(\Omega)$ to (3.16) with the normalization (3.17). In addition, it satisfies (3.18). Furthermore, for all $1 < p < +\infty$, we have $\sigma \in W^{2,p}(\Omega) \cap \mathcal{C}^1(\overline{\Omega})$ and the estimate*

$$\|\sigma\|_{W^{2,p}(\Omega)} \leqslant C\,(1 + \|\sigma\|_{W^{1,p}(\Omega)}), \tag{3.22}$$

where C is a constant depending on p, Ω and

$$\|b\|_{Lip(\overline{\Omega})} = \sup_{x \in \overline{\Omega}} |b(x)| + \sup_{x \neq y \in \overline{\Omega}} \frac{|b(x) - b(y)|}{|x - y|}.$$

Proof. Lemma 3.4 yields the existence and uniqueness of $\sigma \in W^{1,p}(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$, $1 \leqslant p < +\infty$, solution to (3.16)–(3.17), and states that (3.18) holds. The point here is to prove that $\sigma \in W^{2,p}(\Omega)$ for all $1 < p < +\infty$, and that the estimate (3.22) holds true. Since

$$\begin{cases} -\Delta\sigma = \text{div}\,(b\sigma) \ \text{ in } \Omega, \qquad \fint_\Omega \sigma = 1, \\ \nabla\sigma \cdot n = -(b \cdot n)\sigma \ \text{ on } \partial\Omega, \end{cases}$$

we may apply Lemma 3.5. Using in particular that $b \in W^{1,\infty}(\Omega)$, we obtain that

$$\|\text{div}\,(b\sigma)\|_{L^p(\Omega)} \leqslant \|b\|_{W^{1,\infty}(\Omega)} \|\sigma\|_{L^p(\Omega)} + \|b\|_{L^\infty(\Omega)} \|\nabla\sigma\|_{L^p(\Omega)} \leqslant C\,\|\sigma\|_{W^{1,p}(\Omega)}$$

using the Leibniz formula and the Hölder inequality, while

$$\|(b \cdot n)\sigma\|_{W^{1-1/p,p}(\partial\Omega)} \leqslant C \|\sigma\|_{W^{1,p}(\Omega)}$$

is a consequence of the Sobolev trace theorems and the Lipschitz regularity of b on the closed domain $\overline{\Omega}$. Lemma 3.5 therefore implies $\sigma \in W^{2,p}(\Omega)$ and (3.22). Since p can be taken arbitrary large, we also have $\sigma \in \mathcal{C}^1(\overline{\Omega})$. □

For the numerical analysis of the approach performed in Section 3.3, we need the following extension of Proposition 3.6, making precise the continuity of the solutions to the advection-diffusion equation and its adjoint equation with respect to their respective right-hand sides.

Proposition 3.7. *Let p be such that $1 < p < +\infty$ if $d = 2$ and $2d/(d+2) \leqslant p < +\infty$ otherwise. Assuming (3.9)–(3.21) and letting σ be the unique solution to (3.16)–(3.17), we have the following results:*

i) for all $f \in L^p(\Omega)$ such that $\int_\Omega f = 0$, there exists a unique $v \in H^1(\Omega)$ solution to

$$\begin{cases} -\operatorname{div}(\nabla v + bv) = f \quad \text{in } \Omega, \qquad \fint_\Omega v = 1, \\ (\nabla v + bv)\cdot n = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{3.23}$$

In addition, $v \in W^{2,p}(\Omega)$ and we have the estimate

$$\|v - \sigma\|_{W^{2,p}(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}, \tag{3.24}$$

where C is a constant depending on p, Ω and $\|b\|_{Lip(\overline{\Omega})}$;

ii) for all $f \in L^p(\Omega)$ such that $\int_\Omega f\,\sigma = 0$, there exists a unique $u \in H^1(\Omega)$ solution to

$$\begin{cases} -\Delta u + b\cdot\nabla u = f \quad \text{in } \Omega, \qquad \fint_\Omega u = 1, \\ \nabla u \cdot n = 0 \quad \text{on } \partial\Omega. \end{cases}$$

It satisfies $u \in W^{2,p}(\Omega)$ and we have the estimate

$$\|u - 1\|_{W^{2,p}(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}, \tag{3.25}$$

where C is a constant depending on p, Ω and $\|b\|_{L^\infty(\Omega)}$.

Remark 3.8. *When $d \geqslant 3$, the range $1 < p < 2d/(d+2)$ is not covered by the above proposition, and we will not need this case in the sequel. We remark that the existence of a solution to (3.23) for such p could be shown by regularization, considering a sequence $f_n \in L^q(\Omega)$ with $q \geqslant 2d/(d+2)$ such that $\lim_{n\to\infty} \|f_n - f\|_{L^p(\Omega)} = 0$. Problem (3.23) is then well-posed for such f_n. We are then left with passing to the limit $n \to \infty$ in (3.23), which can be done using Lemma 3.5. The other assertions of the above proposition likewise hold.*

Proof. Introducing $w = v - \sigma$, the point for proving assertion (i) is to consider

$$\begin{cases} -\operatorname{div}(\nabla w + bw) = f \quad \text{in } \Omega, \qquad \fint_\Omega w = 0, \\ (\nabla w + bw)\cdot n = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{3.26}$$

The right hand side f belongs to $L^p(\Omega)$ for exponents p that have been chosen so that, by the Sobolev embeddings, $f \in (H^1(\Omega))'$. Lemma 3.3 therefore shows the existence and uniqueness of $w \in H^1(\Omega)$, and the fact that

$$\|w\|_{H^1(\Omega)} \leqslant C\|f\|_{(H^1(\Omega))'} \leqslant C\|f\|_{L^p(\Omega)}. \tag{3.27}$$

We next rewrite (3.26) as

$$\begin{cases} -\Delta w = \operatorname{div}(bw) + f \ \text{ in } \Omega, \qquad \fint_\Omega w = 0, \\ \nabla w \cdot n = -(b\cdot n) w \ \text{ on } \partial\Omega. \end{cases}$$

To apply Lemma 3.5, we distinguish two cases, whether $1 < p \leqslant 2$ or $p > 2$.

Suppose first that $1 < p \leqslant 2$. Hölder inequalities and (3.27) show that

$$\begin{aligned} \|\operatorname{div}(bw) + f\|_{L^p(\Omega)} &\leqslant \|b\|_{W^{1,\infty}(\Omega)} \|w\|_{W^{1,p}(\Omega)} + \|f\|_{L^p(\Omega)} \\ &\leqslant C\|w\|_{H^1(\Omega)} + \|f\|_{L^p(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}. \end{aligned}$$

In addition, using again (3.27) and the fact that b is Lipschitz regular on the closed domain $\overline{\Omega}$, we get

$$\|(b\cdot n)w\|_{W^{1-1/p,p}(\partial\Omega)} \leqslant C\,\|w\|_{W^{1,p}(\Omega)} \leqslant C\,\|w\|_{H^1(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}.$$

Lemma 3.5 therefore implies that $w \in W^{2,p}(\Omega)$ with

$$\|w\|_{W^{2,p}(\Omega)} \leqslant C\left(\|\operatorname{div}(bw) + f\|_{L^p(\Omega)} + \|(b\cdot n)w\|_{W^{1-1/p,p}(\partial\Omega)}\right) \leqslant C\|f\|_{L^p(\Omega)}.$$

This readily yields (3.24), in the case when $1 < p \leqslant 2$.

We now turn to the case $p > 2$. Hölder inequalities show that, for any $2 \leqslant q \leqslant p$, we have

$$\|\operatorname{div}(bw) + f\|_{L^q(\Omega)} \leqslant \|b\|_{W^{1,\infty}(\Omega)} \|w\|_{W^{1,q}(\Omega)} + \|f\|_{L^q(\Omega)}. \tag{3.28}$$

In addition, because b is Lipschitz regular on the closed domain $\overline{\Omega}$, we get

$$\|(b\cdot n)w\|_{W^{1-1/q,q}(\partial\Omega)} \leqslant C\,\|w\|_{W^{1,q}(\Omega)} \tag{3.29}$$

for any $2 \leqslant q \leqslant p$.

Setting $q = 2$ in (3.28) and (3.29) and using that $w \in H^1(\Omega)$, we are in position to use Lemma 3.5, which implies that $w \in H^2(\Omega)$ with

$$\begin{aligned} \|w\|_{H^2(\Omega)} &\leqslant C\left(\|\operatorname{div}(bw) + f\|_{L^2(\Omega)} + \|(b\cdot n)w\|_{W^{1-1/2,2}(\partial\Omega)}\right) \\ &\leqslant C(\|w\|_{H^1(\Omega)} + \|f\|_{L^2(\Omega)}) \\ &\leqslant C\|f\|_{L^p(\Omega)}. \end{aligned} \tag{3.30}$$

Using the Sobolev embeddings, we deduce that $w \in W^{1,q}(\Omega)$ for any $2 \leqslant q < \infty$ if $d = 2$, and $w \in W^{1,q}(\Omega)$ for $q^\star = 2d/(d-2)$ otherwise.

If $d = 2$, we deduce from (3.30) that

$$\|w\|_{W^{1,p}(\Omega)} \leqslant C_p\|w\|_{H^2(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}. \tag{3.31}$$

We set $q = p$ in (3.28) and (3.29) and use Lemma 3.5, which implies that $w \in W^{2,p}(\Omega)$ with

$$\begin{aligned}\|w\|_{W^{2,p}(\Omega)} &\leqslant C\left(\|\mathrm{div}\,(bw) + f\|_{L^p(\Omega)} + \|(b\cdot n)w\|_{W^{1-1/p,p}(\partial\Omega)}\right)\\ &\leqslant C(\|w\|_{W^{1,p}(\Omega)} + \|f\|_{L^p(\Omega)})\\ &\leqslant C\|f\|_{L^p(\Omega)}\end{aligned}$$

where we have used (3.31). We have thus proved (3.24), in the case when $p > 2$ and $d = 2$.

If $d > 2$, we deduce from (3.30) that

$$\|w\|_{W^{1,q^\star}(\Omega)} \leqslant C\|w\|_{H^2(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}. \tag{3.32}$$

If $q^\star \geqslant p$, we proceed as above and readily obtain (3.24). If $q^\star < p$, we set $q = q^\star$ in (3.28) and (3.29) and use Lemma 3.5, which implies that $w \in W^{2,q^\star}(\Omega)$ with

$$\begin{aligned}\|w\|_{W^{2,q^\star}(\Omega)} &\leqslant C\left(\|\mathrm{div}\,(bw) + f\|_{L^{q^\star}(\Omega)} + \|(b\cdot n)w\|_{W^{1-1/q^\star,q^\star}(\partial\Omega)}\right)\\ &\leqslant C(\|w\|_{W^{1,q^\star}(\Omega)} + \|f\|_{L^{q^\star}(\Omega)})\\ &\leqslant C\|f\|_{L^p(\Omega)}\end{aligned}$$

where we have used (3.32). Using again the Sobolev embeddings, we deduce that $w \in W^{1,q^{\star\star}}(\Omega)$ for $1/q^{\star\star} = 1/q^\star - 1/d = 1/2 - 2/d$ if $d > 4$, and for any $q^{\star\star} \geqslant 2$ otherwise. Iterating the argument a sufficient number of times, we prove (3.24) in the case $p > 2$ and $d > 2$. This concludes the proof of assertion (i).

The proof of assertion (ii) proceeds similarly. □

Second choice of invariant measure

In the case where $\mathrm{div}\, b = 0$, Problem (3.1) is coercive, and the introduction of an equivalent problem using the invariant measure seems unnecessary (our numerical results will however show that using σ_1 solution to (3.16)–(3.17) indeed shows itself useful). Put differently, one intuitive choice of invariant measure is then $\sigma = 1$. Since $\sigma = 1$ is not necessary solution to (3.16) in that case, we consider another choice of invariant measure.

The invariant measure σ that we now aim to use (and which we will denote by σ_2 later in this article) is a solution to

$$\begin{cases} -\,\mathrm{div}\,(\nabla\sigma + b\sigma) = 0 \quad \text{in } \Omega,\\ (\nabla\sigma + b\sigma)\cdot n = b\cdot n - \fint_{\partial\Omega} b\cdot n \quad \text{on } \partial\Omega,\\ \inf\limits_{\Omega} \sigma > 0. \end{cases} \tag{3.33}$$

Two remarks are in order. First, we note that σ needs not satisfy $\fint_\Omega \sigma = 1$ since, in essence, this normalization constraint does not affect (3.8) nor *a fortiori* the original problem. Second, it is evident that σ solution to (3.33) is constant if $\mathrm{div}\ b = 0$, a property that has precisely motivated the consideration of this alternate invariant measure.

Because of the constraint of positivity, we are unable to directly prove the existence of σ solution to (3.33). We therefore circumvent this theoretical difficulty by temporarily considering

the same problem, but without the sign constraint and with the specific normalization $\fint_\Omega \sigma = 1$ (see (3.34) below). In a second stage, we will modify the function (adding some term involving σ_1, see Corollary 3.11 below) in order to obtain positivity (possibly at the price of losing the normalization). We already notice that, when discretizing the problems and solving them numerically, we will proceed similarly. Since the practical implementation of (3.33), involving a sign constraint, would be delicate, we will first approximate numerically the solution to (3.34) below and next combine it with the numerical approximation of the solution to (3.16)–(3.17)–(3.18) to obtain an approximation to the solution to (3.33). This will be made precise in Section 3.3.

Let us consider the problem

$$\begin{cases} -\operatorname{div}\,(\nabla\sigma_2^0 + b\sigma_2^0) = 0 \ \text{ in } \Omega, & \fint_\Omega \sigma_2^0 = 1, \\ (\nabla\sigma_2^0 + b\sigma_2^0)\cdot n = b\cdot n - \fint_{\partial\Omega} b\cdot n \ \text{ on } \partial\Omega. \end{cases} \tag{3.34}$$

We have the following proposition, the proof of which is similar to that of Proposition 3.6 and which we therefore skip (note that the well-posedness of (3.34) in $H^1(\Omega)$ is a direct consequence of Lemma 3.3(ii)):

Proposition 3.9. *Under assumptions (3.9)–(3.21), there exists a unique $\sigma_2^0 \in H^1(\Omega)$ solution to (3.34). For any $1 < p < +\infty$, this solution satisfies $\sigma_2^0 \in W^{2,p}(\Omega) \cap \mathcal{C}^1(\overline{\Omega})$ and we have the following estimate:*

$$\|\sigma_2^0\|_{W^{2,p}(\Omega)} \leqslant C\left(1 + \|\sigma_2^0\|_{W^{1,p}(\Omega)} + \left\|b\cdot n - \fint_{\partial\Omega} b\cdot n\right\|_{W^{1-1/p,p}(\partial\Omega)}\right),$$

where C is a constant depending on p, Ω and $\|b\|_{Lip(\overline{\Omega})}$. In addition, the solution to (3.34) satisfies conditions (C1) *and* (C3).

Likewise, the following proposition holds, with a proof that mimics that of Proposition 3.7:

Proposition 3.10. *Let $1 < p < +\infty$ if $d = 2$ and $2d/(d+2) \leqslant p < +\infty$ otherwise. Assume (3.9)–(3.21). For all $f \in L^p(\Omega)$ such that $\int_\Omega f = 0$, there exists a unique $v \in H^1(\Omega)$ solution to*

$$\begin{cases} -\operatorname{div}\,(\nabla v + bv) = f \ \text{ in } \Omega, & \fint_\Omega v = 1, \\ (\nabla v + bv)\cdot n = b\cdot n - \fint_{\partial\Omega} b\cdot n \ \text{ on } \partial\Omega. \end{cases}$$

This solution belongs to $W^{2,p}(\Omega)$ and satisfies

$$\|v - \sigma_2^0\|_{W^{2,p}(\Omega)} \leqslant C\|f\|_{L^p(\Omega)}, \tag{3.35}$$

where σ_2^0 is the solution to (3.34) and C is a constant depending on p, Ω and $\|b\|_{Lip(\overline{\Omega})}$.

Of course, all what matters in the above estimation (3.35) is that $\fint_\Omega v = \fint_\Omega \sigma_2^0$ and not the actual value of that integral.

We now eventually obtain a solution to (3.33), modifying σ_2^0 in a suitable manner. This is the purpose of our next result.

Corollary 3.11. *Let σ_2^0 (resp. σ_1) be the solution to (3.34) (resp. to (3.16)–(3.17)). Under assumptions (3.9)–(3.21), the set of solutions to (3.33) reads as*

$$\left\{\sigma_2^0 + \kappa\,\sigma_1, \quad \kappa \in \mathbb{R} \text{ such that } \inf_\Omega\,(\sigma_2^0 + \kappa\,\sigma_1) > 0\right\}.$$

The proof of Corollary 3.11 is immediate. Let σ be a solution to (3.33). Then $\sigma - \sigma_2^0$ is a solution to (3.16), and we are then in position to use Lemma 3.3, noticing that the necessary value of κ is $\kappa = \fint_\Omega \sigma - 1$. The converse inclusion is straightforward.

We finally define σ_2 solution to (3.33) as

$$\sigma_2 = \sigma_2^0 + \kappa^\star\,\sigma_1, \tag{3.36}$$

where

$$\kappa^\star = 1 + \inf\left\{\kappa \in \mathbb{R} \text{ such that } \inf_\Omega\,(\sigma_2^0 + \kappa\,\sigma_1) > 0\right\}.$$

Of course this is an arbitrary choice. In practice, some suitable κ (and thus σ_2) will be used. The numerical analysis will account for this.

3.3 Discretization and numerical analysis

Practically, we implement a finite element Galerkin approximation u_H of the solution u to the coercive equivalent modified problem (3.15). Since, in most of the cases, the invariant measure σ is not known analytically, we first seek a Galerkin approximation of σ (we will describe later in this article how this approximation is obtained, for each of the two cases $\sigma \equiv \sigma_1$ and $\sigma \equiv \sigma_2$). We denote by H and h the mesh sizes for these two approximations, respectively. We have in mind that $H \gg h$, in order to be as efficient as possible. This is made possible by the uniform well-posedness of the discrete problem in u_H (see Proposition 3.12 below). In practice, we will observe that we may indeed choose H one order of magnitude larger, say, than h.

We begin with making precise the approximation u_H, for a given approximation σ_h of σ. The discrete variational formulation reads as:

$$\text{Find } u_H \in U_H \text{ such that, for all } v_H \in U_H, \quad a_{\rm ss}(\sigma_h; u_H, v_H) = F(\sigma_h v_H), \tag{3.37}$$

where F is defined by (3.13) and $a_{\rm ss}$ is defined by (3.38) below.

We assume throughout this section that the discretization space U_H in (3.37) is a subspace of $H_0^1(\Omega)$. In our actual implementation, the above formulation will be possibly slightly modified to account for a stabilization performed when computing σ_h. This will be made precise in the next section, in formulae (3.79)–(3.80). As will be mentioned there, this potential modification does not modify the numerical analysis we perform in the present section.

In the left hand side of (3.37), we have denoted

$$a_{\rm ss}(\sigma_h; u_H, v_H) = \int_\Omega \sigma_h \nabla u_H \cdot \nabla v_H + B_h \cdot \frac{(\nabla u_H) v_H - (\nabla v_H) u_H}{2}, \tag{3.38}$$

$$B_h = \nabla \sigma_h + \sigma_h \, b. \tag{3.39}$$

The classical skew-symmetric formulation of the advection part is adopted in order to ensure that $a_{\rm ss}(\sigma_h; u_H, u_H) = \int_\Omega \sigma_h \, |\nabla u_H|^2$ and thus that the problem is coercive at the discrete level whenever σ_h is positive and bounded away from zero. A simple application of standard arguments therefore shows the following well-posedness of the discretization. The point is, this well-posedness is uniform in the mesh size H, a property that is of major practical interest.

Proposition 3.12. *Assume (3.9). Consider σ_h an approximation of $\sigma \in H^1(\Omega)$ such that $\inf_\Omega \sigma_h > 0$. Then $a_{ss}(\sigma_h; \cdot, \cdot)$ is coercive in $H^1_0(\Omega)$ and (3.37) is well-posed, uniformly in H.*

We now proceed with the numerical analysis of (3.37).

3.3.1 Numerical analysis in the case when the invariant measure is analytically known

To begin with, we temporarily assume that we know σ analytically, meaning we replace B_h given by (3.39) by $B = \nabla \sigma + \sigma \, b$ in the second term of (3.38) (and we likewise replace σ_h by σ in the first term of (3.38)). Otherwise stated, we replace σ_h by σ in (3.37).

Proposition 3.13. *Assume (3.9)–(3.21) and that $\sigma_h \equiv \sigma$ in (3.37). Let u be the solution to (3.15) and u_H be the solution to (3.37). Then, for any $p > d$, we have the estimate*

$$\|u - u_H\|_{H^1(\Omega)} \leqslant C \left[\frac{\|\sigma\|_{L^\infty(\Omega)} + \|\nabla \sigma + \sigma \, b\|_{L^p(\Omega)}}{\inf_\Omega \sigma} \right] \inf_{v_H \in U_H} \|u - v_H\|_{H^1(\Omega)}, \tag{3.40}$$

with a constant C that only depends on Ω and p.

Note that, in view of Lemma 3.4, the assumptions (3.9)–(3.21) imply that σ_1 satisfies the conditions (C1), (C2) and (C3). Likewise, in view of (3.36), Lemma 3.4 and Proposition 3.9, σ_2 satisfies the conditions (C1), (C2) and (C3). In particular, for both choices $\sigma \equiv \sigma_1$ and $\sigma \equiv \sigma_2$, we have $\inf_\Omega \sigma > 0$.

Proof. As $\operatorname{div} B = 0$, we note that the problem

$$\text{Find } u \in H^1_0(\Omega) \text{ such that, for all } v \in H^1_0(\Omega), \quad a_{\rm ss}(\sigma; u, v) = F(\sigma \, v)$$

is a variational formulation of the modified problem (3.15). Since $\sigma_h \equiv \sigma$, Problem (3.37) is the Galerkin approximation of (3.15) in U_H. We note that the bilinear form $a_{\rm ss}(\sigma; \cdot, \cdot)$ is coercive, while, for all u and v in $H^1_0(\Omega)$, we have

$$\begin{aligned} a_{\rm ss}(\sigma; u, v) &\leqslant \|\sigma\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|\nabla \sigma + \sigma \, b\|_{L^p(\Omega)} \|\nabla u\|_{L^2(\Omega)} \, \|v\|_{L^q(\Omega)} \\ &\leqslant \left(\|\sigma\|_{L^\infty(\Omega)} + C_{p,\Omega} \, \|\nabla \sigma + \sigma \, b\|_{L^p(\Omega)} \right) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)}, \end{aligned} \tag{3.41}$$

where $1/p + 1/q = 1/2$ and, for the Sobolev embedding to hold, $q < 2d/(d-2)$ which amounts to $p > d$. Classical results of numerical analysis of coercive problems then allow to conclude, using the Céa lemma. □

The following corollary makes precise how the $L^p(\Omega)$ norm in the right hand side of (3.40) may be bounded from above by the $H^1(\Omega)$ norm of σ, because of the particular properties of σ. When the discretized approximation σ_h is reinstated in place of σ, this part of the argument will become substantially more difficult. We will return to this later.

Corollary 3.14. *In addition to the assumptions of Proposition 3.13, we assume that the ambient dimension is $d = 2$ or 3. Then, we have*

$$\|u - u_H\|_{H^1(\Omega)} \leqslant C \left(\frac{\|\sigma\|_{L^\infty(\Omega)} + \|\sigma\|_{H^1(\Omega)} + C_\sigma}{\inf_\Omega \sigma} \right) \inf_{v_H \in U_H} \|u - v_H\|_{H^1(\Omega)},$$

where C is a constant independent of H and

$$C_\sigma = \begin{cases} 0 & \text{if } \sigma \equiv \sigma_1, \\ \left\| b \cdot n - \fint_{\partial\Omega} b \cdot n \right\|_{H^{1/2}(\partial\Omega)} & \text{if } \sigma \equiv \sigma_2. \end{cases}$$

Proof. Using [41, Corollary 3.7], we know that $B = \nabla\sigma + \sigma\, b$ satisfies

$$\|B\|_{H^1(\Omega)} \leqslant C \left(\|B\|_{L^2(\Omega)} + \|\operatorname{div} B\|_{L^2(\Omega)} + \|\operatorname{curl} B\|_{L^2(\Omega)} + \|B \cdot n\|_{H^{1/2}(\partial\Omega)} \right).$$

We notice that, on the one hand, $\operatorname{div} B = 0$, by definition of σ, while, on the other hand, $\operatorname{curl} B = \operatorname{curl}(\sigma\, b)$, thus, given the Lipschitz regularity of b, $\|\operatorname{curl} B\|_{L^2(\Omega)} \leqslant C\, \|\sigma\|_{H^1(\Omega)}$. We therefore obtain

$$\|B\|_{H^1(\Omega)} \leqslant C \left(\|\sigma\|_{H^1(\Omega)} + \|B \cdot n\|_{H^{1/2}(\partial\Omega)} \right).$$

Finally, because $d \leqslant 3$, we may find p such that $d < p \leqslant 2d/(d-2)$, thus $\|B\|_{L^p(\Omega)} \leqslant C \|B\|_{H^1(\Omega)}$, which proves the result. □

3.3.2 Numerical analysis in the case when the invariant measure is numerically approximated

We now return to the case when the invariant measure is only approximated numerically. For simplicity, we restrict our attention to the case when the approximation space for σ is a $\mathbb{P}^1$ finite element space associated to a polyhedral mesh of Ω. In this case, Ω is thus a polygon, and it cannot be of class $\mathcal{C}^1$ or $\mathcal{C}^2$, as assumed previously in (3.10) or (3.21).

In the sequel of this Section 3.3.2, we assume, in addition to (3.9), that

$$\begin{cases} \text{the domain } \Omega \text{ is connected, convex and polyhedral;} \\ b \text{ is Lipschitz-continuous on } \overline{\Omega}; \\ \text{the ambient dimension satisfies } 2 \leqslant d \leqslant 3. \end{cases} \tag{3.42}$$

We also assume that

$$\text{The conclusions of Propositions 3.6, 3.7, 3.9 and 3.10 hold.} \tag{3.43}$$

A few remarks are in order.

First, we point out that (3.43) is not a consequence of Section 3.2, as we now do not assume (3.21) here.

Second, (3.43) obviously holds in the case $b = 0$. In that case, there exists a unique solution to (3.16)–(3.17) (resp. to (3.34)) which is $\sigma_1 = 1$ (resp. $\sigma_2^0 = 1$).

Third, when $\|b\|_{\mathrm{Lip}(\overline{\Omega})}$ is sufficiently small, then (3.43) again holds. For the sake of brevity, we only sketch the proof for Proposition 3.6. To construct the invariant measure, consider the following iterations: set $\sigma^0 = 1$, and define σ^{m+1} as the unique solution in $H^1(\Omega)$ to the problem

$$-\Delta\sigma^{m+1} = \mathrm{div}\,(b\sigma^m) \text{ in } \Omega, \qquad \nabla\sigma^{m+1}\cdot n = -b\sigma^m\cdot n \text{ on } \partial\Omega, \qquad \fint_\Omega \sigma^{m+1} = 1.$$

It is easy to see that

$$\|\nabla(\sigma^{m+1}-\sigma^m)\|_{L^2(\Omega)} \leqslant \|b\|_{L^\infty(\Omega)}\|\sigma^m-\sigma^{m-1}\|_{L^2(\Omega)}.$$

Since the mean of $\sigma^m - \sigma^{m-1}$ vanishes, we can use the Poincaré-Wirtinger (PW) inequality, from which we deduce that

$$\|\nabla(\sigma^{m+1}-\sigma^m)\|_{L^2(\Omega)} \leqslant C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)}\|\nabla(\sigma^m-\sigma^{m-1})\|_{L^2(\Omega)}.$$

Assume that b is such that $C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)} < 1$. Then σ^m converges to some $\sigma^\star$ in $H^1(\Omega)$, which is a solution to (3.16)–(3.17). The uniqueness of such a solution is easily obtained, again as a consequence of the Poincaré-Wirtinger inequality and of the fact that $C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)} < 1$. Furthermore, for any $v \in H^1(\Omega)$, we have

$$\int_\Omega \nabla(\sigma^\star-1)\cdot\nabla v = -\int_\Omega (\sigma^\star-1)\, b\cdot\nabla v - \int_\Omega b\cdot\nabla v.$$

Choosing $v = \sigma^\star - 1$, we get

$$\|\nabla(\sigma^\star-1)\|_{L^2(\Omega)} \leqslant C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)}\|\nabla(\sigma^\star-1)\|_{L^2(\Omega)} + \|b\|_{L^\infty(\Omega)},$$

and thus $\|\sigma^\star - 1\|_{H^1(\Omega)} \leqslant \sqrt{1+C_{\mathrm{PW}}^2}\,\dfrac{\|b\|_{L^\infty(\Omega)}}{1-C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)}}$.

We now prove that $\sigma^\star \in H^2(\Omega)$. Considering the Neumann problem

$$-\Delta(\sigma^\star-1) = \mathrm{div}\,(b\sigma^\star) \quad \text{in } \Omega, \qquad \nabla(\sigma^\star-1)\cdot n = -b\sigma^\star\cdot n \quad \text{on } \partial\Omega,$$

we observe, as in the proof of Proposition 3.6 and using the regularity of b, that

$$\|\mathrm{div}\,(b\sigma^\star)\|_{L^2(\Omega)} \leqslant C\,\|b\|_{W^{1,\infty}(\Omega)}\,\|\sigma^\star\|_{H^1(\Omega)}$$

while

$$\|(b\cdot n)\sigma^\star\|_{H^{1/2}(\partial\Omega)} \leqslant C\|b\|_{\mathrm{Lip}(\overline{\Omega})}\,\|\sigma^\star\|_{H^1(\Omega)},$$

where C is independent of b. Thanks to the assumption (3.42) on Ω, we are in position to use [38, Theorem 3.12], which implies that $\sigma^\star - 1 \in H^2(\Omega)$ and

$$\begin{aligned}\|\sigma^\star - 1\|_{H^2(\Omega)} &\leqslant C\left(\|\mathrm{div}\,(b\sigma^\star)\|_{L^2(\Omega)} + \|(b\cdot n)\sigma^\star\|_{H^{1/2}(\partial\Omega)}\right)\\ &\leqslant C\left(\|b\|_{W^{1,\infty}(\Omega)} + \|b\|_{\mathrm{Lip}(\overline{\Omega})}\right)\|\sigma^\star\|_{H^1(\Omega)}\\ &\leqslant C\|b\|_{\mathrm{Lip}(\overline{\Omega})}\left(1 + \frac{\|b\|_{L^\infty(\Omega)}}{1 - C_{\mathrm{PW}}\|b\|_{L^\infty(\Omega)}}\right).\end{aligned}$$

Similar estimates in $W^{2,p}(\Omega)$ can be shown using [38, Remark 3.13(ii)].

To show that Proposition 3.6 holds, we are now left with showing (3.18). Thanks to the Sobolev injections when $2 \leqslant d \leqslant 3$, we have $\|\sigma^\star - 1\|_{C^0(\Omega)} \leqslant C\|\sigma^\star - 1\|_{H^2(\Omega)}$. Thus, when b is sufficiently small, then $\|\sigma^\star - 1\|_{C^0(\Omega)}$ is small as well and (3.18) holds. We can thus conclude that Proposition 3.6 holds.

Preliminary estimate

In what follows, we proceed under the assumptions (3.9)–(3.42)–(3.43). The analogous result to that of Proposition 3.13 is stated in the following.

Proposition 3.15. *Assume (3.9)–(3.42)–(3.43). Consider σ_h an approximation of $\sigma \in H^1(\Omega)$ such that $\inf_\Omega \sigma_h > 0$. Let u be the solution to (3.1)–(3.11) (or equivalently (3.15)) and u_H be the solution to (3.37), for some $f \in L^2(\Omega)$. For any $p > d$, we have the estimate*

$$\begin{aligned}\|u - u_H\|_{H^1(\Omega)} \leqslant{}& \frac{C}{\inf_\Omega \sigma_h}\|f\|_{L^2(\Omega)}\|\sigma - \sigma_h\|_{L^p(\Omega)}\\ &+ C\inf_{v_H \in U_H}\left[\left(1 + \frac{|||\sigma|||_p}{\inf_\Omega \sigma_h}\right)\|u - v_H\|_{H^1(\Omega)} + \frac{|||\sigma - \sigma_h|||_p}{\inf_\Omega \sigma_h}\|v_H\|_{H^1(\Omega)}\right],\end{aligned} \tag{3.44}$$

where C only depends on p and Ω, and where we have used the notation

$$|||\sigma|||_p = \|\sigma\|_{L^\infty(\Omega)} + \|\nabla\sigma + b\sigma\|_{L^p(\Omega)}.$$

Proof. We note that $u \in H^1_0(\Omega)$ satisfies

$$\forall v \in H^1_0(\Omega), \quad a_{\mathrm{ss}}(\sigma; u, v) = F(\sigma v),$$

while $u_H \in U_H$ satisfies

$$\forall v_H \in U_H, \quad a_{\mathrm{ss}}(\sigma_h; u_H, v_H) = F(\sigma_h v_H).$$

Applying the first Strang Lemma (see [38, Lemma 2.27]), we have

$$\begin{aligned}\|u - u_H\|_{H^1(\Omega)} \leqslant{}& \frac{1}{C_\Omega \inf_\Omega \sigma_h}\sup_{w_H \in U_H}\frac{\left|\int_\Omega f(\sigma - \sigma_h)w_H\right|}{\|w_H\|_{H^1(\Omega)}}\\ &+ \inf_{v_H \in U_H}\left[\left(1 + \frac{\|\sigma\|_{L^\infty(\Omega)} + C_{p,\Omega}\,\|B\|_{L^p(\Omega)}}{C_\Omega \inf_\Omega \sigma_h}\right)\|u - v_H\|_{H^1(\Omega)}\right.\\ &\left.+ \frac{1}{C_\Omega \inf_\Omega \sigma_h}\sup_{w_H \in U_H}\frac{|a_{\mathrm{ss}}(\sigma - \sigma_h; v_H, w_H)|}{\|w_H\|_{H^1(\Omega)}}\right],\end{aligned} \tag{3.45}$$

where $B = \nabla\sigma + b\sigma$, C_Ω is the Poincaré constant of Ω (so that $C_\Omega \inf_\Omega \sigma_h$ is a coercivity constant of $a_{\rm ss}(\sigma_h;\cdot,\cdot)$ on U_H) and $C_{p,\Omega}$ is the constant introduced in (3.41). When we take $\sigma_h \equiv \sigma$, this estimation of course agrees with the estimation we have already established, independently, for Proposition 3.13.

For the first term of the right-hand side of (3.45), we notice that, for all $w_H \in U_H$ and $p > d$ (thus $1/q = 1/2 - 1/p < 1/2 - 1/d$), we have

$$\begin{aligned}\left|\int_\Omega f(\sigma-\sigma_h)w_H\right| &\leqslant \|f\|_{L^2(\Omega)}\|\sigma-\sigma_h\|_{L^p(\Omega)}\|w_H\|_{L^q(\Omega)}\\ &\leqslant C_{p,\Omega}\|f\|_{L^2(\Omega)}\|\sigma-\sigma_h\|_{L^p(\Omega)}\|w_H\|_{H^1(\Omega)}.\end{aligned} \tag{3.46}$$

The rightmost term of (3.45) is estimated similarly:

$$\begin{aligned}&|a_{\rm ss}(\sigma-\sigma_h;v_H,w_H)|\\ &\qquad\leqslant \left(\|\sigma-\sigma_h\|_{L^\infty(\Omega)} + C_{p,\Omega}\,\|B-B_h\|_{L^p(\Omega)}\right)\|v_H\|_{H^1(\Omega)}\,\|w_H\|_{H^1(\Omega)},\end{aligned} \tag{3.47}$$

where $B_h = \nabla\sigma_h + b\sigma_h$. Combining (3.45), (3.46) and (3.47) gives the desired estimate. □

Estimation of $\sigma - \sigma_h$

The estimation of $\nabla(\sigma - \sigma_h)$ in $L^p(\Omega)$, for some $p > d$, is the crucial ingredient we now need to proceed with the estimation of the right-hand side of (3.44). This estimation is the main purpose of Proposition 3.17 below. We emphasize that the result is not immediate and its proof instructive. Before stating this result, we first detail how σ_h is defined and provide in Proposition 3.16 below a classical error estimate on $\sigma - \sigma_h$ in $H^1(\Omega)$.

We introduce the bilinear form

$$a^\star(u,v) = \int_\Omega (\nabla u + bu)\cdot\nabla v, \tag{3.48}$$

which is formally the adjoint of the bilinear form a defined by (3.4), in the sense that $a^\star(u,v) = a(v,u)$. We note that the invariant measure σ_1 solution to (3.16)–(3.17)–(3.18) satisfies

$$\forall v \in H^1(\Omega), \qquad a^\star(\sigma_1, v) = 0$$

while the invariant measure σ_2^0 solution to (3.34) satisfies

$$\forall v \in H^1(\Omega), \qquad a^\star(\sigma_2^0, v) = \int_{\partial\Omega} g\, v$$

with $g = b\cdot n - \fint_{\partial\Omega} b\cdot n$ on $\partial\Omega$.

Proposition 3.16. *We assume that (3.9)–(3.42)–(3.43) hold. Let Σ_h be the $\mathbb{P}^1$ approximation space associated to a regular quasi-uniform polyhedral mesh of Ω and*

$$V_h = \left\{u \in \Sigma_h, \quad \fint_\Omega u = 1\right\}.$$

Let σ denote either the solution to (3.16)–(3.17)–(3.18) (in which case we set $g = 0$) or the solution to (3.34) (in which case we set $g = b \cdot n - \fint_{\partial\Omega} b \cdot n$ on $\partial\Omega$).

For h sufficiently small, there exists a unique $\sigma_h \in V_h$ (which is the Galerkin approximation of σ) solution to

$$\forall v_h \in \Sigma_h, \qquad a^\star(\sigma_h, v_h) = \int_{\partial\Omega} g\, v_h. \tag{3.49}$$

Furthermore, we have, for h sufficiently small,

$$\|\sigma - \sigma_h\|_{H^1(\Omega)} \leqslant Ch\|\sigma\|_{H^2(\Omega)} \tag{3.50}$$

where C is independent of h.

The proof of Proposition 3.16 is postponed until Appendix 3.5. We now turn to the estimation of $\nabla(\sigma - \sigma_h)$ in $L^p(\Omega)$.

Proposition 3.17. *Under the assumptions of Proposition 3.16, for all $2 < p < +\infty$, the estimate*

$$\|\sigma - \sigma_h\|_{W^{1,p}(\Omega)} \leqslant Ch\|\sigma\|_{W^{2,p}(\Omega)} \tag{3.51}$$

holds for h sufficiently small, where C is independent of h.

The proof of Proposition 3.17 is given below. We emphasize that, to the best of our knowledge, this result is not present in the literature. There exist many contributions establishing $W^{1,p}$ estimates between the solution of a linear PDE and its finite element approximation, for problems posed with homogeneous Dirichlet boundary conditions, or problems posed with Neumann boundary conditions and including a zero-order term. In [85], the author considers the Neumann problem

$$-\Delta v + v = f \ \text{ in } \Omega, \qquad \nabla v \cdot n = 0 \ \text{ on } \partial\Omega, \tag{3.52}$$

while the Dirichlet problem

$$-\Delta v = f \ \text{ in } \Omega, \qquad v = 0 \ \text{ on } \partial\Omega,$$

is studied in [70, 71, 79]. A more general PDE (including an advection term and a zero-order term, but again with homogeneous Dirichlet boundary conditions) is considered in [15, Chap. 8]. All these problems are well-posed (under appropriate assumptions) for *any* sufficiently regular right-hand side. We note that the proofs contained in the contributions we have cited consider the problem of interest (e.g. (3.52) in [85]) for several right-hand sides, and not only the right-hand side f originally considered. In contrast, Problem (3.19) is well-posed only for right-hand sides satisfying some compatibility conditions (see Lemma 3.3). This is one of the reasons why the proof of Proposition 3.17 is not immediate.

Another contribution we wish to cite is [41, Theorem A.2 p. 101]. Taking some sufficiently regular f and g such that the compatibility condition $\int_\Omega f + \int_{\partial\Omega} g = 0$ holds, the authors consider the Neumann problem

$$-\Delta v = f \ \text{ in } \Omega, \qquad \nabla v \cdot n = g \ \text{ on } \partial\Omega \tag{3.53}$$

and state a $W^{1,p}$ estimate between v and its finite element approximation v_h (chosen such that $\int_\Omega v_h = \int_\Omega v$): there exists C independent of h such that

$$\|v_h - v\|_{W^{1,p}(\Omega)} \leqslant Ch\|v\|_{W^{2,p}(\Omega)}. \tag{3.54}$$

There are (at least) two ways to prove Proposition 3.17. A first possibility is to assume that $\|b\|_{\mathrm{Lip}(\overline{\Omega})}$ is small enough. Under this assumption (which is restrictive since we precisely aim in this article at considering non-coercive problems (3.1) where b is not small) and using (3.54), the proof of (3.51) is short. For the sake of brevity, we only consider the invariant measure σ_1 solution to (3.16)–(3.17)–(3.18). We introduce the sequence $\sigma^m \in H^1(\Omega)$ defined by

$$-\Delta\sigma^{m+1} = \mathrm{div}\,(b\sigma^m) \text{ in } \Omega, \quad \nabla\sigma^{m+1}\cdot n = -b\sigma^m\cdot n \text{ on } \partial\Omega, \quad \fint_\Omega \sigma^{m+1} = 1, \tag{3.55}$$

with $\sigma^0 = 1$. Since b is small enough, it turns out that σ^m converges to σ_1, solution to (3.16)–(3.17). In addition, the above problem is of the type (3.53), so we will be in position to use (3.54).

Consider the sequence $\sigma_h^m \in \Sigma_h$ defined by

$$\forall v \in \Sigma_h, \qquad \int_\Omega \nabla\sigma_h^{m+1}\cdot\nabla v = -\int_\Omega \sigma_h^m b\cdot\nabla v, \qquad \fint_\Omega \sigma_h^{m+1} = 1, \tag{3.56}$$

with $\sigma_h^0 = 1$, which converges to $\sigma_{1,h}$, solution to (3.49) with $g = 0$. Using the result (3.54) given in [41, Theorem A.2 p. 101], one can eventually show that

$$\|\sigma_h^{m+1} - \sigma^{m+1}\|_{W^{1,p}(\Omega)} \leqslant C\|b\|_{\mathrm{Lip}(\overline{\Omega})}\left(\|\sigma_h^m - \sigma^m\|_{W^{1,p}(\Omega)} + h\|\sigma^m\|_{W^{1,p}(\Omega)}\right) \tag{3.57}$$

for some C independent of h and b. Note that the right-hand sides of (3.55) and (3.56) are different, so the intermediate problem

$$-\Delta\overline{\sigma}^{m+1} = \mathrm{div}\,(b\sigma_h^m) \text{ in } \Omega, \quad \nabla\overline{\sigma}^{m+1}\cdot n = -b\sigma_h^m\cdot n \text{ on } \partial\Omega, \quad \fint_\Omega \overline{\sigma}^{m+1} = 1,$$

has to be introduced to prove (3.57). Passing to the limit $m\to\infty$ in (3.57), and using again that $\|b\|_{\mathrm{Lip}(\overline{\Omega})}$ is sufficiently small, we obtain (3.51).

A second possibility, which is the one we follow here, is based on considering the following problem, that we write in a compact form as $L_\eta\,\sigma^{\eta,f} = f$:

$$\begin{cases} -\mathrm{div}\,(\nabla\sigma^{\eta,f} + b\sigma^{\eta,f}) + \eta\sigma^{\eta,f} = f & \text{in } \Omega, \\ (\nabla\sigma^{\eta,f} + b\sigma^{\eta,f})\cdot n = 0 & \text{on } \partial\Omega, \end{cases} \tag{3.58}$$

for any $0 < \eta \leqslant 1$. Problem (3.58) is well-posed for any sufficiently regular function f (in particular, there is no compatibility condition on f). Let $\sigma_h^{\eta,f} \in \Sigma_h$ be the P1 finite element approximation of $\sigma^{\eta,f}$. It is then possible to adapt the proof of [15, Chap. 8] to this case, and show that there exists a constant C_η independent of h and f (but a priori depending on η) such that

$$\|\sigma_h^{\eta,f} - \sigma^{\eta,f}\|_{W^{1,p}(\Omega)} \leqslant C_\eta\, h\, \|\sigma^{\eta,f}\|_{W^{2,p}(\Omega)}. \tag{3.59}$$

We now sketch the proof of (3.51), in the case of the invariant measure σ_1 solution to (3.16)–(3.17)–(3.18). The proof is based on the introduction of iterations of the type (3.55), with the operator L_η of (3.58) instead of the Laplacian operator:

$$L_\eta \, \sigma_\eta^{m+1} = \eta \sigma_\eta^m$$

for some η sufficiently small. We refer to (3.64) below for details. The proof then proceeds as in the case b small above, the pivotal estimate (3.54) being replaced by (3.59). We emphasize that we take η sufficiently small, but we do not need to take the limit $\eta \to 0$.

We now proceed in details. For any $0 < \eta \leqslant 1$, we introduce the bilinear form

$$a_\eta^\star(u, v) = \int_\Omega (\nabla u + bu) \cdot \nabla v + \eta \int_\Omega u \, v. \tag{3.60}$$

We have the following result:

Proposition 3.18. *We assume that (3.9)–(3.42)–(3.43) hold. Let $u \in H^1(\Omega)$ and $u_h \in \Sigma_h$ such that*

$$\forall v \in \Sigma_h, \quad a_\eta^\star(u - u_h, v) = 0. \tag{3.61}$$

Let $2 \leqslant p < \infty$ and assume that $u \in W^{1,p}(\Omega)$. Then, there exists C_η and $h_0(\eta)$, that both depend on η, such that, for any $0 < h < h_0(\eta)$, we have

$$\|\nabla u_h\|_{L^p(\Omega)} \leqslant C_\eta \, \|\nabla u\|_{L^p(\Omega)}. \tag{3.62}$$

Assume furthermore that $u \in W^{2,p}(\Omega)$. Then

$$\|\nabla (u - u_h)\|_{L^p(\Omega)} \leqslant C_\eta \, h \, \|u\|_{W^{2,p}(\Omega)}. \tag{3.63}$$

The proof of Proposition 3.18 is postponed until Appendix 3.6. It follows the arguments of [15, Chap. 8]. Most presumably, a similar result can be obtained when $1 < p < 2$, using duality arguments as in [15, Sec. 8.5]. We do not need such a result here, and therefore do not proceed in that direction.

We are now in position to prove Proposition 3.17.

Proof of Proposition 3.17. We define the function g on $\partial\Omega$ by $g \equiv 0$ in the case of the invariant measure σ_1 and $g = g_2 := b \cdot n - \fint_{\partial\Omega} b \cdot n$ on $\partial\Omega$ in the case of the invariant measure σ_2^0. Let $2 < p < \infty$ and let $\eta > 0$ be small enough in a sense made precise below. The proof falls in three steps.

Step 1. Consider the following iterations: set $\sigma_\eta^0 = 1$ and define σ_η^{m+1} as the unique solution to the problem

$$\left\{ \begin{array}{c} \text{Find } \sigma_\eta^{m+1} \in H^1(\Omega) \text{ such that, for any } v \in H^1(\Omega), \\ a_\eta^\star(\sigma_\eta^{m+1}, v) = \eta \displaystyle\int_\Omega \sigma_\eta^m v + \int_{\partial\Omega} g v. \end{array} \right. \tag{3.64}$$

Lemma 3.23 in Appendix 3.6 below ensures that the above problem is well-posed (the bilinear form $a_\eta^\star$ satisfying an inf-sup condition in $H^1(\Omega)$) and that, if η is sufficiently small, $\sigma_\eta^m \in W^{2,p}(\Omega)$

for any m. Taking $v \equiv 1$ in (3.64), we observe that $\fint_\Omega \sigma_\eta^m = 1$ for any m. Furthermore, we see that

$$\forall v \in H^1(\Omega), \qquad a^\star(\sigma_\eta^{m+1}, v) = \eta \int_\Omega (\sigma_\eta^m - \sigma_\eta^{m+1})\, v + \int_{\partial\Omega} gv.$$

Using Proposition 3.7 in the case $g \equiv 0$ (resp. Proposition 3.10 in the case $g = g_2$), we get

$$\|\sigma_\eta^{m+1} - \sigma\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_\eta^m - \sigma_\eta^{m+1}\|_{L^p(\Omega)}.$$

Taking η sufficiently small, this shows that

$$\|\sigma_\eta^{m+1} - \sigma\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_\eta^m - \sigma\|_{L^p(\Omega)},$$

and hence that

$$\lim_{m\to\infty} \|\sigma_\eta^m - \sigma\|_{W^{2,p}(\Omega)} = 0. \tag{3.65}$$

Step 2. We now consider the following iterations, at the discrete level: set $\sigma_h^0 = 1$ and define $\sigma_{\eta,h}^{m+1}$ as the unique solution to the problem:

$$\left\{ \begin{array}{l} \text{Find } \sigma_{\eta,h}^{m+1} \in \Sigma_h \text{ such that, for any } v \in \Sigma_h, \\ \qquad a_\eta^\star(\sigma_{\eta,h}^{m+1}, v) = \eta \displaystyle\int_\Omega \sigma_{\eta,h}^m v + \int_{\partial\Omega} gv. \end{array} \right. \tag{3.66}$$

Theorem 3.25 in Appendix 3.6 below ensures that the above problem is well-posed (the proof of Theorem 3.25 is performed in the case $g \equiv 0$, and it carries over to the case $g = g_2$). Taking $v \equiv 1$ in (3.66), we observe that $\fint_\Omega \sigma_{\eta,h}^m = 1$ for any m. Furthermore, we see that

$$\forall v \in \Sigma_h, \qquad a^\star(\sigma_{\eta,h}^{m+1} - \sigma_h, v) = \eta \int_\Omega (\sigma_{\eta,h}^m - \sigma_{\eta,h}^{m+1})\, v.$$

We show in Appendix 3.5 below (see (3.91)) that $a^\star$ satisfies an inf-sup property on functions in Σ_h of vanishing mean, with a constant γ independent of h. Since $\fint_\Omega \sigma_{\eta,h}^m = 1 = \fint_\Omega \sigma_h$ for any m, we get

$$\|\sigma_{\eta,h}^{m+1} - \sigma_h\|_{H^1(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma_{\eta,h}^{m+1}\|_{L^2(\Omega)}.$$

Taking η sufficiently small, this shows that

$$\|\sigma_{\eta,h}^{m+1} - \sigma_h\|_{H^1(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma_h\|_{L^2(\Omega)},$$

and hence that $\lim_{m\to\infty} \|\sigma_{\eta,h}^m - \sigma_h\|_{H^1(\Omega)} = 0$. By equivalence of the norms in the finite dimensional space Σ_h, this implies that

$$\lim_{m\to\infty} \|\sigma_{\eta,h}^m - \sigma_h\|_{W^{1,p}(\Omega)} = 0. \tag{3.67}$$

Step 3. We eventually introduce the following problem:

$$\left\{ \begin{array}{l} \text{Find } \overline{\sigma}_\eta^{m+1} \in H^1(\Omega) \text{ such that, for any } v \in H^1(\Omega), \\ \qquad a_\eta^\star(\overline{\sigma}_\eta^{m+1}, v) = \eta \displaystyle\int_\Omega \sigma_{\eta,h}^m v + \int_{\partial\Omega} gv. \end{array} \right. \tag{3.68}$$

We observe that (3.68) is the continuous analogue of (3.66), for the *same* right-hand side (in contrast, when going from (3.64) to (3.66), we modify both the space in which we search the solution and the right-hand side of the equation). We observe that

$$\forall v \in \Sigma_h, \qquad a_\eta^\star(\overline{\sigma}_\eta^{m+1} - \sigma_{\eta,h}^{m+1}, v) = 0.$$

We are thus in position to use Proposition 3.18, which states that, for any $h < h_0(\eta)$, we have

$$\|\overline{\sigma}_\eta^{m+1} - \sigma_{\eta,h}^{m+1}\|_{W^{1,p}(\Omega)} \leqslant C_\eta\, h\, \|\overline{\sigma}_\eta^{m+1}\|_{W^{2,p}(\Omega)}. \tag{3.69}$$

We now estimate $\|\overline{\sigma}_\eta^{m+1}\|_{W^{2,p}(\Omega)}$. We observe that

$$\forall v \in H^1(\Omega), \qquad a^\star(\overline{\sigma}_\eta^{m+1}, v) = \eta \int_\Omega (\sigma_{\eta,h}^m - \overline{\sigma}_\eta^{m+1})\, v + \int_{\partial\Omega} gv.$$

In addition, taking $v \equiv 1$ in (3.68), we see that $⨍_\Omega \overline{\sigma}_\eta^{m+1} = ⨍_\Omega \sigma_{\eta,h}^m = 1$. Using Proposition 3.7 in the case $g \equiv 0$ (resp. Proposition 3.10 in the case $g = g_2$), we get

$$\|\overline{\sigma}_\eta^{m+1} - \sigma\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \overline{\sigma}_\eta^{m+1}\|_{L^p(\Omega)}.$$

Taking η sufficiently small, this shows that

$$\|\overline{\sigma}_\eta^{m+1} - \sigma\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma\|_{L^p(\Omega)},$$

and hence

$$\|\overline{\sigma}_\eta^{m+1}\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m\|_{L^p(\Omega)} + C\|\sigma\|_{W^{2,p}(\Omega)}.$$

Inserting this estimate in (3.69), we deduce that

$$\|\overline{\sigma}_\eta^{m+1} - \sigma_{\eta,h}^{m+1}\|_{W^{1,p}(\Omega)} \leqslant C_\eta\, h\, \left(\eta \|\sigma_{\eta,h}^m\|_{L^p(\Omega)} + \|\sigma\|_{W^{2,p}(\Omega)}\right). \tag{3.70}$$

We now compare $\overline{\sigma}_\eta^{m+1}$ and σ_η^{m+1}. We observe that

$$\forall v \in H^1(\Omega), \qquad a_\eta^\star(\overline{\sigma}_\eta^{m+1} - \sigma_\eta^{m+1}, v) = \eta \int_\Omega (\sigma_{\eta,h}^m - \sigma_\eta^m)\, v,$$

hence

$$\forall v \in H^1(\Omega), \qquad a^\star(\overline{\sigma}_\eta^{m+1} - \sigma_\eta^{m+1}, v) = \eta \int_\Omega (\sigma_{\eta,h}^m - \sigma_\eta^m - \overline{\sigma}_\eta^{m+1} + \sigma_\eta^{m+1})\, v.$$

Using Proposition 3.7, we get that

$$\|\overline{\sigma}_\eta^{m+1} - \sigma_\eta^{m+1}\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma_\eta^m - \overline{\sigma}_\eta^{m+1} + \sigma_\eta^{m+1}\|_{L^p(\Omega)},$$

which implies, for η sufficiently small, that

$$\|\overline{\sigma}_\eta^{m+1} - \sigma_\eta^{m+1}\|_{W^{2,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma_\eta^m\|_{L^p(\Omega)},$$

and thus

$$\|\overline{\sigma}_\eta^{m+1} - \sigma_\eta^{m+1}\|_{W^{1,p}(\Omega)} \leqslant C\eta \|\sigma_{\eta,h}^m - \sigma_\eta^m\|_{W^{1,p}(\Omega)}. \tag{3.71}$$

Step 4. Collecting (3.70) and (3.71), we get

$$\|\sigma_\eta^{m+1} - \sigma_{\eta,h}^{m+1}\|_{W^{1,p}(\Omega)} \leqslant C_\eta\, h\, \left(\eta\|\sigma_{\eta,h}^m\|_{L^p(\Omega)} + \|\sigma\|_{W^{2,p}(\Omega)}\right) + C\eta\|\sigma_{\eta,h}^m - \sigma_\eta^m\|_{W^{1,p}(\Omega)}.$$

Using (3.65) and (3.67), we are in position to pass to the limit $m \to \infty$. We deduce that

$$\|\sigma - \sigma_h\|_{W^{1,p}(\Omega)} \leqslant C_\eta\, h\, \left(\eta\|\sigma_h\|_{L^p(\Omega)} + \|\sigma\|_{W^{2,p}(\Omega)}\right) + C\eta\|\sigma_h - \sigma\|_{W^{1,p}(\Omega)},$$

and thus, for η sufficiently small,

$$\begin{aligned}\|\sigma - \sigma_h\|_{W^{1,p}(\Omega)} &\leqslant C_\eta\, h\, \left(\eta\|\sigma_h\|_{L^p(\Omega)} + \|\sigma\|_{W^{2,p}(\Omega)}\right)\\ &\leqslant C_\eta\, h\, \left(\|\sigma_h - \sigma\|_{L^p(\Omega)} + \|\sigma\|_{W^{2,p}(\Omega)}\right).\end{aligned}$$

Taking h such that $C_\eta\, h \leqslant 1/2$, we deduce (3.51). □

Main result: estimation of $u - u_H$

Proposition 3.17 allows to deduce from Proposition 3.15 the following Theorem 3.19. To this end, we successively consider the case of our two invariant measures. For our first invariant measure σ_1, solution to (3.16)–(3.17)–(3.18), we obviously consider its $\mathbb{P}^1$ finite element approximation $\sigma_{1,h}$. For our second invariant measure, the analysis is essentially similar. There is, however, an additional subtlety in the very definition of the measure and its approximation. One, basic but crucial, remark is that the solution u to (3.1)–(3.11) does not depend on the choice of σ. More precisely, (3.1)–(3.11) is equivalent to (3.15) irrespectively of the choice of σ. We use this flexibility for our numerical analysis:

(i) we first approximate σ_1 as above by $\sigma_{1,h}$, and assume that $\sigma_{1,h} > 0$ on Ω. In practice, we have always numerically observed this property for sufficiently small h (this property actually often holds for h not asymptotically small). In addition, from a theoretical viewpoint, this bound from below is a consequence of the assumptions (3.9)–(3.42)–(3.43), as explained in the proof of Theorem 3.19 below.

(ii) we next approximate, again using $\mathbb{P}^1$ finite elements, σ_2^0 solution to (3.34) by $\sigma_{2,h}^0$. We perform both these approximations on the same regular mesh $\mathcal{T}_h$. We then define $\sigma_{2,h} = \sigma_{2,h}^0 + \kappa_h\, \sigma_{1,h}$ where

$$\kappa_h = 1 + \inf\left\{\overline{\kappa} \in [0,\infty) \text{ such that } \sigma_{2,h}^0 + \overline{\kappa}\, \sigma_{1,h} > 0 \text{ on } \Omega\right\}.$$

Precisely since, as noticed above, u does not depend on our choice of invariant measure, we correspondingly define $\sigma_2 = \sigma_2^0 + \kappa_h\, \sigma_1$. The point of our analysis is then to estimate $\sigma_2 - \sigma_{2,h}$. The detail is contained in the following proof.

Theorem 3.19. *Assume that the invariant measure (σ_1, as defined by (3.16)–(3.17)–(3.18), or σ_2 a solution to (3.33)) is approximated as we have just described in items (i) and (ii) above. Under the assumptions (3.9)–(3.42)–(3.43), we have, for h sufficiently small,*

$$\|u - u_H\|_{H^1(\Omega)} \leqslant C\, h\, \|f\|_{L^2(\Omega)} + C \inf_{v_H \in U_H} \left[\|u - v_H\|_{H^1(\Omega)} + h\, \|v_H\|_{H^1(\Omega)}\right] \tag{3.72}$$

for a constant C independent of h.

Proof. We first consider the case of our first invariant measure σ_1, solution to (3.16)–(3.17)–(3.18), approximated by its $\mathbb{P}^1$ finite element approximation $\sigma_{1,h}$. We have shown in Proposition 3.17 that $\|\sigma_1 - \sigma_{1,h}\|_{W^{1,p}(\Omega)} = O(h)$ for any $2 < p < \infty$. Using the Sobolev injections, we obtain that $\|\sigma_1 - \sigma_{1,h}\|_{C^0(\Omega)} = O(h)$. Since $\inf_\Omega \sigma_1 > 0$, we get that, for any h sufficiently small, $\inf_\Omega \sigma_{1,h} \geqslant c_{\min} > 0$ for some $c_{\min}$ independent of h. The estimate (3.44) thus holds. In order to deduce (3.72) from (3.44), we have to estimate both $\|\sigma_1 - \sigma_{1,h}\|_{L^\infty(\Omega)}$ and $\|\nabla(\sigma_1 - \sigma_{1,h})\|_{L^p(\Omega)}$ by $O(h)$ terms. Since, for p sufficiently large, the L^∞ norm is controlled by the $W^{1,p}$ norm, we will conclude our proof if we show that $\|\sigma_1 - \sigma_{1,h}\|_{W^{1,p}(\Omega)} = O(h)$. This latter bound is precisely the purpose of Proposition 3.17.

As we said, the case of the second invariant measure is essentially similar with the suitable definition of σ_2 and $\sigma_{2,h}$. Note first that, when h is sufficiently small, we have $\inf_\Omega \sigma_{1,h} \geqslant c_{\min} > 0$ for some $c_{\min}$ independent of h, as explained above. We then observe that

$$\sigma_{2,h} = \sigma_{2,h}^0 + \kappa_h \, \sigma_{1,h} = \sigma_{2,h}^0 + (1 + \overline{\kappa}_h) \, \sigma_{1,h} \geqslant c_{\min} > 0,$$

since, by definition of $\overline{\kappa}_h$, we have $\sigma_{2,h}^0 + \overline{\kappa}_h \, \sigma_{1,h} \geqslant 0$ on Ω. The estimate (3.44) thus again holds.

In our argument to establish (3.72), we use $\|\sigma_2 - \sigma_{2,h}\|_{L^\infty(\Omega)}$ and $\|\nabla(\sigma_2 - \sigma_{2,h})\|_{L^p(\Omega)}$ for those particular choices of σ_2 and $\sigma_{2,h}$. Obviously,

$$\sigma_2 - \sigma_{2,h} = (\sigma_2^0 - \sigma_{2,h}^0) + \kappa_h \, (\sigma_1 - \sigma_{1,h}).$$

The term $\sigma_1 - \sigma_{1,h}$ has just been estimated above in the suitable norms, while the term $\sigma_2^0 - \sigma_{2,h}^0$ is estimated similarly. Eventually, for h sufficiently small, because of the convergence in $C^0(\Omega)$ of the $\mathbb{P}^1$ finite elements approximations, we know that $\sup_\Omega |\sigma_{2,h}^0|$ is bounded uniformly in h while $\inf_\Omega \sigma_{1,h} \geqslant c_{\min} > 0$ for some $c_{\min}$ independent of h. Thus κ_h is bounded uniformly in h, when h is sufficiently small. The triangle inequality allows to conclude our proof. □

3.4 Implementation details and numerical results

3.4.1 Discretization of the invariant measure

The numerical approximation of σ_1, solution to (3.16)–(3.17)–(3.18) is, as we said in the previous section, obtained using a classical $\mathbb{P}^1$ finite element space associated to a uniform mesh of size h that we denote $\mathcal{T}_h$.

Problem (3.16)–(3.17)–(3.18) involves two constraints: a normalization constraint and a sign constraint. We comply with the normalization constraint by implementing an iterative algorithm. With a view to obtaining the positivity constraint, we use the adjoint equation and stabilize the problem (see (3.74) below). The stiffness matrix of the formulation employed to compute $\sigma_{1,h}$ is the adjoint matrix to the matrix of the Douglas-Wang (DW) stabilized version of the approximation of the solution to the advection-diffusion equation. Since, in most situations, it is observed that the latter approximation preserves the maximum principle, it is intuitively expected that the same applies for the adjoint formulation. Nevertheless, we have no general theoretical argument that shows our formulation guarantees positivity of the solution $\sigma_{1,h}$ (except of course if h is sufficiently small, as shown in the proof of Theorem 3.19).

For essentially all the practical computations we have performed, we observe that positivity is preserved, in the sense that $\int_K \sigma_{1,h}$ is positive for all $K \in \mathcal{T}_H$. This is all we need to proceed with a coercive bilinear form at the discrete level for the advection-diffusion equation (3.15) since we will be using $\mathbb{P}^1$ finite elements for the approximation u_H of u (see Section 3.4.2 below). We mention that Droniou [23] and Tobiska [55] have proposed a discretization that gives a positive discrete solution, but we do not proceed this way.

We implement the following iterative algorithm: $\sigma_{1,h}^{n+1}$ is defined as the solution to the problem

$$\left\{ \begin{array}{c} \text{Find } \sigma_{1,h}^{n+1} \in \Sigma_h \text{ such that, for all } \varphi \in \Sigma_h, \\ a^\star\left(\sigma_{1,h}^{n+1}, \varphi\right) + \lambda \displaystyle\int_\Omega \sigma_{1,h}^{n+1}\, \varphi + a_{\text{stab, DW}}\left(\sigma_{1,h}^{n+1}, \varphi\right) = \lambda \int_\Omega \sigma_{1,h}^{n}\, \varphi, \end{array} \right. \tag{3.73}$$

where λ is a positive parameter, Σ_h is the $\mathbb{P}^1$ finite element space associated to $\mathcal{T}_h$, $a^\star$ is defined by (3.48) and

$$a_{\text{stab,DW}}(\sigma, \varphi) = \sum_{K \in \mathcal{T}_h} \int_K \tau^\star \, [\operatorname{div}(\nabla\sigma + b\sigma)] \, (-\Delta\varphi + b \cdot \nabla\varphi), \tag{3.74}$$

with

$$\tau^\star(x) = \frac{h}{2|b(x)|}\left(\coth(\mathrm{Pe}_K^\star(x)) - \frac{1}{\mathrm{Pe}_K^\star(x)}\right), \qquad \mathrm{Pe}_K^\star(x) = \frac{|b(x)|h}{2}. \tag{3.75}$$

Note that the stabilization term (3.74) is the standard DW stabilization term for the invariant measure equation (3.16). The formulation is thus strongly consistent.

We set $\lambda = 10^{-3}$ in our numerical tests. The iterations are initialized with an approximation of the invariant measure of the potential part of b, namely $I_{\Sigma_h}(e^{-\psi_H})$, where I_{Σ_h} is the nodal interpolation operator in Σ_h and $\psi_H \in U_H$ satisfies, for any $v_H \in U_H$, $\displaystyle\int_\Omega \nabla\psi_H \cdot \nabla v_H = \int_\Omega b \cdot \nabla v_H$. The stopping criterion we use is $\left\| 1 - \dfrac{\sigma_{1,h}^{n+1}}{\sigma_{1,h}^{n}} \right\|_{L^1(\Omega)} < 10^{-3}$. Temporarily ignoring the stabilization term $a_{\text{stab,DW}}$ in (3.73), we formally see that using this stopping criterion aims at enforcing that $\dfrac{\operatorname{div}(\nabla\sigma_{1,h}^{n+1} + b\sigma_{1,h}^{n+1})}{\sigma_{1,h}^{n+1}}$ is small when we stop the iterations (3.73). This is a better criterion than enforcing that $\operatorname{div}(\nabla\sigma_{1,h}^{n+1} + b\sigma_{1,h}^{n+1})$ is small, as σ_1 may vary a lot over the domain Ω.

We similarly obtain an approximation $\sigma_{2,h}^0 \in \Sigma_h$ of σ_2^0 solution to (3.34), considering iterations where $(\sigma_2^0)_h^{n+1}$ is the solution to the problem:

$$\left\{ \begin{array}{c} \text{Find } (\sigma_2^0)_h^{n+1} \in \Sigma_h \text{ such that, for all } \varphi \in \Sigma_h, \\ a^\star\left((\sigma_2^0)_h^{n+1}, \varphi\right) + \lambda \displaystyle\int_\Omega (\sigma_2^0)_h^{n+1}\, \varphi = \lambda \int_\Omega (\sigma_2^0)_h^{n}\, \varphi + \int_{\partial\Omega} \left(b \cdot n - \frac{1}{|\partial\Omega|} \int_{\partial\Omega} b \cdot n \right) \varphi. \end{array} \right.$$

The iterations are initialized using $(\sigma_2^0)_h^0 = 1$. The same stopping criterion is adopted as for σ_1. Notice that, in that case, we need not account for the positivity, which will be obtained by the combination $\sigma_{2,h} = (\sigma_2^0)_h + \kappa_h\, \sigma_{1,h}$ described above, so no stabilization of the formulation is employed.

3.4.2 Discretization of u

A natural way to define the $(\mathbb{P}^1, \sigma_h\mathbb{P}^1)$ method would be to consider the following variational formulation (see (3.37)):

$$\text{Find } u_H \in \mathbb{P}^1(\mathcal{T}_H) \text{ s.t., for all } v_H \in \mathbb{P}^1(\mathcal{T}_H), \quad a_{\rm ss}(\sigma_h; u_H, v_H) = F(\sigma_h v_H), \tag{3.76}$$

where we recall (see (3.38)–(3.39)) that

$$\begin{aligned} a_{\rm ss}(\sigma_h; u_H, v_H) &= \int_\Omega \sigma_h \nabla u_H \cdot \nabla v_H + B_h \cdot \frac{(\nabla u_H) v_H - (\nabla v_H) u_H}{2}, \\ B_h &= \nabla \sigma_h + \sigma_h\, b. \end{aligned} \tag{3.77}$$

In the case of the invariant measure σ_1, we need to add an extra term related to the stabilized discretization used for $\sigma_{1,h}$. The reason is the following. Formally, the variational formulation (3.76)–(3.77) corresponds to the approximation

$$-\operatorname{div}(\sigma_h \nabla u) + B_h \cdot \nabla u + \frac{1}{2}(\operatorname{div} B_h)\, u = \sigma_h\, f \tag{3.78}$$

of equation (3.15). The zero order term in $\operatorname{div} B_h$ (originating from the skew-symmetric formulation that ensures coercivity at the discrete level) affects the accuracy. Now, the stabilization (3.74) introduced in the variational formulation for $\sigma_{1,h}$ amounts to modifying $B_h = (B_1)_h = \nabla\sigma_{1,h} + \sigma_{1,h}\, b$ into

$$(\overline{B}_1)_h = (B_1)_h + \left(\sum_{K \in \mathcal{T}_h} \tau^\star \operatorname{div}((B_1)_h) \mathbb{1}_K \right) b \tag{3.79}$$

with $\tau^\star$ defined by (3.75). More precisely, at convergence (i.e. when $n \to \infty$), the formulation (3.73) amounts to requesting that, for any $\varphi \in \Sigma_h$,

$$\int_\Omega (\overline{B}_1)_h \cdot \nabla\varphi = 0$$

rather than $\displaystyle\int_\Omega (B_1)_h \cdot \nabla\varphi = 0$, which is the standard discretization of (3.16). Formally, the quantity the divergence of which is zero is not $(B_1)_h$, but $(\overline{B}_1)_h$. In view of the last term of the left-hand side of (3.78), and with the aim of obtaining the best possible accuracy, we thus need to modify $(B_1)_h$ into $(\overline{B}_1)_h$ in (3.77).

In order to be consistent, we therefore define

$$a_{\rm ss}(\sigma_{1,h}; u_H, v_H) = \int_\Omega \sigma_{1,h} \nabla u_H \cdot \nabla v_H + (\overline{B}_1)_h \cdot \frac{(\nabla u_H) v_H - (\nabla v_H) u_H}{2} \tag{3.80}$$

instead of (3.77). The problem is, by construction, coercive, and may be analyzed by the standard tools of numerical analysis we have used in the previous section. We readily note that the replacement of $(B_1)_h$ by $(\overline{B}_1)_h$ does not affect this analysis. Indeed, on any $K \in \mathcal{T}_h$,

$$\operatorname{div}\left[(B_1)_h\right] = \operatorname{div}(\nabla\sigma_{1,h}) + \operatorname{div}(\sigma_{1,h}\, b),$$

where the first term vanishes for $\mathbb{P}^1$ finite elements. Thus, for any $p > d$,

$$\left\|(\overline{B}_1)_h - (B_1)_h\right\|_{L^p(\Omega)} \leqslant \left\|\tau^\star\, b\, \mathrm{div}(\sigma_{1,h}\, b)\right\|_{L^p(\Omega)} \leqslant \mathcal{C} h\, \|b\|_{W^{1,\infty}(\Omega)}\, \|\sigma_{1,h}\|_{W^{1,p}(\Omega)}$$

where $\mathcal{C}$ is a universal constant such that $|\coth(y) - y^{-1}| \leqslant \mathcal{C}$ for any $y \in \mathbb{R}$. The factor $\|\sigma_{1,h}\|_{W^{1,p}(\Omega)}$ can be bounded from above independently of h as a consequence of Proposition 3.17. Using arguments similar to those used in the proofs of Proposition 3.15 and Theorem 3.19, we thus see that (3.72) again holds when using the bilinear form (3.80) instead of (3.77).

In the case of the invariant measure σ_2, we also need to add an extra term to (3.76)–(3.77). Recall that $\sigma_{2,h} = (\sigma_2^0)_h + \kappa_h\, \sigma_{1,h}$, where no stabilization is employed to compute $(\sigma_2^0)_h$, in contrast to $\sigma_{1,h}$. Formally, and again in view of the last term of the left-hand side of (3.78), it seems advantageous to work with $(B_2^0)_h + \kappa_h\, (\overline{B}_1)_h$ rather than $(B_2^0)_h + \kappa_h\, (B_1)_h$, as we expect that the divergence of the former is smaller than that of the latter. In order to be consistent, we therefore define

$$a_{\mathrm{ss}}(\sigma_{2,h}; u_H, v_H) = \int_\Omega \sigma_{2,h} \nabla u_H \cdot \nabla v_H + (\overline{B}_2)_h \cdot \frac{(\nabla u_H) v_H - (\nabla v_H) u_H}{2} \tag{3.81}$$

with $(\overline{B}_2)_h = (B_2^0)_h + \kappa_h\, (\overline{B}_1)_h$ instead of (3.77) (recall that $(\overline{B}_1)_h$ is defined by (3.79) and that $(B_2^0)_h = \nabla(\sigma_2^0)_h + (\sigma_2^0)_h\, b$). As in the case of the invariant measure σ_1, the problem is, by construction, coercive, and may be analyzed by the standard tools of numerical analysis we have used in the previous section.

In addition to the above practical and theoretical considerations, we also need to possibly modify the formulation when the problem is advection-dominated. It turns out that we only need to use such a stabilized formulation when working with the invariant measure σ_2. In that case, we use a GLS type method and define the $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS method by the following variational formulation:

$$\left\{ \begin{array}{c} \text{Find } u_H \in \mathbb{P}^1(\mathcal{T}_H) \text{ such that, for all } v_H \in \mathbb{P}^1(\mathcal{T}_H), \\ a_{\mathrm{ss}}(\sigma_{2,h}; u_H, v_H) + a_{\mathrm{stab}}(u_H, v_H) = F(\sigma_{2,h} v_H) + F_{\mathrm{stab}}(v_H), \end{array} \right. \tag{3.82}$$

where a_{ss} is defined by (3.81), and

$$\begin{aligned} a_{\mathrm{stab}}(u_H, v_H) &= \sum_{K \in \mathcal{T}_H} \int_K \tau\, (\sigma_{2,h} b \cdot \nabla u_H)\, (\sigma_{2,h} b \cdot \nabla v_H), \\ F_{\mathrm{stab}}(v_H) &= \sum_{K \in \mathcal{T}_H} \int_K \tau\, (\sigma_{2,h} f)\, (\sigma_{2,h} b \cdot \nabla v_H), \end{aligned} \tag{3.83}$$

with

$$\tau(x) = \frac{H}{2|(B_2)_h(x)|} \left(\coth(\mathrm{Pe}_K(x)) - \frac{1}{\mathrm{Pe}_K(x)} \right), \qquad \mathrm{Pe}_K(x) = \frac{|(B_2)_h(x)| H}{2\sigma_{2,h}(x)},$$

where $(B_2)_h = (B_2^0)_h + \kappa_h\, (B_1)_h = \nabla \sigma_{2,h} + \sigma_{2,h}\, b$. Denoting $\mathcal{L}_h v = -\mathrm{div}(\sigma_h \nabla v) + B_h \cdot \nabla v$ the operator approximating that of (3.15) when only σ_h is available, we indeed see that, on any $K \in \mathcal{T}_H$, we have

$$\mathcal{L}_h u_H = \sigma_h b \cdot \nabla u_H$$

as a consequence of the fact that u_H is a $\mathbb{P}^1$ function. The term (3.83) is indeed a GLS-type stabilization term, in the sense that it reads $a_{\text{stab}}(u_H, v_H) = \sum_{K\in\mathcal{T}_H} \int_K \tau\,(\mathcal{L}_h u_H)\,(\mathcal{L}_h v_H)$.

3.4.3 Irrotational case

For all our numerical tests throughout this article, we work on the unit square $\Omega = (0,1)^2$ and choose the right-hand side $f = 1$. All computations are performed on a Intelő Xeonő Processor E5-2667 v2. We use the FreeFem++ software [43].

We assume in this Section 3.4.3 that the velocity field b is irrotational: $b = \nabla\phi$. In that case, we know that $\sigma_1 = \left(\fint_\Omega e^{-\phi}\right)^{-1} e^{-\phi}$ and that $\nabla\sigma_1 + \sigma_1\, b = 0$ in Ω. Specifically here, the velocity field b is taken of the form

$$b = \big(b_x^0, b_y^0\big)^T + \lambda_1\big(\cos(2\pi x)\sin(2\pi y), \sin(2\pi x)\cos(2\pi y)\big)^T + \lambda_2\big(\cos^2(2\pi x), 0\big)^T + \lambda_3(y,x)^T,$$

where b_x^0, b_y^0, λ_1, λ_2, $\lambda_3 > 0$. We take $b_x^0 = b_y^0 = 64$. The parameters λ_1, λ_2 and λ_3 are given in Table 3.1 for the four test cases (i) through (iv) we consider. The last column of Table 3.1 shows that the problem is not coercive in the tests (ii) to (iv).

	λ_1	λ_2	λ_3	$\inf_{v_H\in U_H} \frac{a(v_H,v_H)}{\|v_H\|^2_{L^2(\Omega)}}$
Test (i)	0	0	0	19.93
Test (ii)	0	50.34	0	−45.05
Test (iii)	0	50.34	30	−45.05
Test (iv)	20	50.34	0	−95.21

Table 3.1 Definition of the parameters for the four discrete problems (i)-(iv)

Tables 3.2 through 3.5 show the relative error

$$\text{err} = \frac{\|\nabla(u_H - u_{\text{ref}})\|_{L^2(\Omega\setminus\Omega_{\text{layer}})}}{\|\nabla u_{\text{ref}}\|_{L^2(\Omega)}} \tag{3.84}$$

for various numerical solutions u_H. In the convection-dominated regime, the solution presents a boundary layer of approximate width

$$\delta_{\text{layer}} = \frac{2}{\|b\|_{L^\infty(\Omega)}} \log \frac{\|b\|_{L^\infty(\Omega)}}{2}.$$

For the convection fields we consider, we set the boundary layer region (see Figure 3.1) as

$$\Omega_{\text{layer}} = ((0,1)\times(1-\delta_{\text{layer}},1)) \cup ((1-\delta_{\text{layer}},1)\times(0,1)) \cup ((0,1)\times(0,\delta_{\text{layer}}))$$

and we only measure the accuracy of u_H outside this layer. We have also assessed the accuracy in $L^2(\Omega)$ norm and obtained similar qualitative conclusions. The reference solution u_{ref} is computed using a $\mathbb{P}^1$ approach with a tiny mesh size.

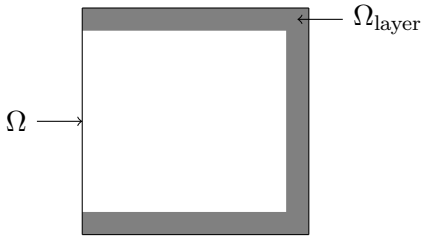

Figure 3.1 The domain Ω_{layer} coloured in grey

For all approaches, we fix the mesh size $H = 1/16$ for the approximation u_H of u. We compare six approaches: the classical $\mathbb{P}^1$ finite element approximation (which may be unstable in the advection-dominated regime), its stabilized Galerkin least-square variant $\mathbb{P}^1$-GLS, and our four approaches $(\mathbb{P}^1, \sigma_1\mathbb{P}^1)$, $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$, $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$ and $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS, respectively using the exact value of σ_1, its approximation $\sigma_{1,h}$, and the approximation $\sigma_{2,h}$ of σ_2, either classical (as in (3.76) with a_{ss} defined by (3.80) or (3.81)) or stabilized (as in (3.82)).

The comparison between $\mathbb{P}^1$ and $\mathbb{P}^1$-GLS is used as an empirical measure of the instability of the problem. Likewise, comparing $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$ and $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS allows to see the potential added value of a stabilization of the problem when using $\sigma_{2,h}$ as an approximation of the invariant measure. The approach $(\mathbb{P}^1, \sigma_1\mathbb{P}^1)$ using the exact value of the invariant measure is of course the most accurate one, and performs equally well as (and often better than) $\mathbb{P}^1$-GLS. When we forbid ourselves to use that exact value of the invariant measure, $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ is the best method to use, and it does not require stabilization (see the third column of Tables 3.2 through 3.5). Note yet that, if one has to work with $h = H$, then $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS is the best method (see the second column of Tables 3.2 through 3.5), providing results the accuracy of which is around 8%. We also note the following fact. The two rightmost columns show tests that use a mesh to approximate σ that is not a subset of the mesh used to compute u. In that case, the approach $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ deteriorates. In the present state of our understanding, we are unable to explain this phenomenon. We therefore advocate to employ $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ with meshes that are a subset of one another, or, otherwise, to switch to $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS.

Remark 3.20. *For the Test (iv) reported on in Table 3.5, and in the particular case $h = H$, it turns out that $\int_K \sigma_{1,h}$ is not positive for all $K \in \mathcal{T}_H$ (it is positive for all the other values of h considered, and for all the computations reported on in Tables 3.2 through 3.4). In that case, it is thus not possible to use the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method. However, it turns out, still for that value of h, that there exists κ_h such that $\int_K \big(\sigma^0_{2,h} + \kappa_h\, \sigma_{1,h}\big)$ is positive for all $K \in \mathcal{T}_H$. The $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS approach can thus be used.*

3.4.4 General case

We now consider the general, not necessarily irrotational case. This time, b reads as

$$b = \big(b^0_x, b^0_y\big)^T + \lambda_1\big(\cos(2\pi x)\sin(2\pi y), \sin(2\pi x)\cos(2\pi y)\big)^T + \lambda_2\big(\cos^2(2\pi x), 0\big)^T + \lambda_3(y, x)^T + \lambda_4(y, -x)^T,$$

Error (3.84)	$(h = H)$	$(h = H/5)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.191			
$\mathbb{P}^1$-GLS	0.0328			
$(\mathbb{P}^1, (\sigma_1)\mathbb{P}^1)$	0.0187			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	0.313	0.0208	0.127	0.0818
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.191	0.191	0.191	0.191
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.0328	0.0328	0.0326	0.0327

Table 3.2 Test (i)

Error (3.84)	$(h = H)$	$(h = H/7)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.479			
$\mathbb{P}^1$-GLS	0.0551			
$(\mathbb{P}^1, (\sigma_1)\mathbb{P}^1)$	0.0199			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	0.385	0.0218	0.139	0.102
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.614	0.398	0.362	0.377
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.0827	0.0532	0.0511	0.0520

Table 3.3 Test (ii)

Error (3.84)	$(h = H)$	$(h = H/9)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.536			
$\mathbb{P}^1$-GLS	0.0411			
$(\mathbb{P}^1, (\sigma_1)\mathbb{P}^1)$	0.0302			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	0.453	0.0250	0.153	0.126
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.862	0.461	0.412	0.429
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.0784	0.0420	0.0397	0.0405

Table 3.4 Test (iii)

Error (3.84)	$(h = H)$	$(h = H/7)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.468			
$\mathbb{P}^1$-GLS	0.0573			
$(\mathbb{P}^1, (\sigma_1)\mathbb{P}^1)$	0.0250			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	-	0.0266	0.153	0.111
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.612	0.401	0.379	0.388
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.0894	0.0550	0.0535	0.0544

Table 3.5 Test (iv)

where $b_x^0 = b_y^0 = 64$ and where $\lambda_1,\ \lambda_2,\ \lambda_3,\ \lambda_4 > 0$. We study the three examples defined in Table 3.6.

	λ_1	λ_2	λ_3	λ_4	$\inf_{v_H \in U_H} \frac{a(v_H, v_H)}{\|v_H\|^2_{L^2(\Omega)}}$
Test (v)	0	50.34	0	64	-45.05
Test (vi)	20	50.34	0	64	-95.21
Test (vii)	0	50.34	30	64	-45.05

Table 3.6 Definition of the parameters for the three discrete problems (v)-(vii)

Tables 3.7 through 3.9 show our results, for $H = 1/16$ as in the previous section. The approaches evaluated are identical to those of Tables 3.2 through 3.5, with the notable exception of the approach $(\mathbb{P}^1, \sigma_1\mathbb{P}^1)$ since now the exact invariant measure σ_1 is unknown. The conditions in which we perform our tests are identical. The results confirm our conclusions of the previous section.

Remark 3.21. *For the largest value of h used in Tables 3.7 through 3.9, it turns out that $\int_K \sigma_{1,h}$ is not positive for all $K \in \mathcal{T}_H$ (see also Remark 3.20). But there still exists κ_h such that $\int_K (\sigma^0_{2,h} + \kappa_h\, \sigma_{1,h})$ is positive for all $K \in \mathcal{T}_H$, which allows us to use the $(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS approach.*

Error (3.84)	$(h = 1/17)$	$(h = H/7)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.568			
$\mathbb{P}^1$-GLS	0.0704			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	-	0.0390	0.146	0.0981
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.657	0.515	0.462	0.480
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.117	0.0672	0.0658	0.0660

Table 3.7 Test (v)

Error (3.84)	$(h = H)$	$(h = H/7)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.620			
$\mathbb{P}^1$-GLS	0.0807			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	-	0.0549	0.151	0.105
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.641	0.522	0.482	0.495
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.134	0.0772	0.0764	0.0768

Table 3.8 Test (vi)

Error (3.84)	$(h = 1/17)$	$(h = H/9)$	$(h = 1/150)$	$(h = 1/230)$
$\mathbb{P}^1$	0.636			
$\mathbb{P}^1$-GLS	0.0606			
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	-	0.0285	0.159	0.116
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$	0.648	0.550	0.489	0.509
$(\mathbb{P}^1, \sigma_{2,h}\mathbb{P}^1)$-GLS	0.112	0.0588	0.0571	0.0573

Table 3.9 Test (vii)

3.4.5 Computational cost and efficiency

We now evaluate the computational cost of the most accurate of our approaches, namely the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method, given the results of the tests performed in the previous sections, as compared to the classical $\mathbb{P}^1$-GLS method.

As can be easily seen upon considering some particular situations where σ_1 is known analytically (some of these cases are considered in Section 3.4.3 above), the contrast of σ_1 over the domain, say measured by the ratio $\frac{\sup_\Omega \sigma_1}{\inf_\Omega \sigma_1}$, may be huge, especially in the advection-dominated regime. Therefore, the stiffness matrix involved in the solution procedure for the modified equation (3.15) is often ill-conditioned. We therefore use, in our tests, a direct solver (from the UMFPACK library) for the linear algebraic systems. An alternate, equally effective approach is to use an iterative inversion algorithm together with a diagonal preconditioner. We have indeed tested such an approach in other tests not reproduced here, obtaining similar conclusions. In particular, the diagonal preconditionner, although simple, turns out to be very effective in diminishing the number of iterations.

Fixed cost

We compare the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method and the $\mathbb{P}^1$-GLS method at fixed cost. Tables 3.10 through 3.12 show the accuracy of the two methods for the tests (v)-(vi)-(vii). Similar results have been obtained for our tests (i) through (iv). We observe that the $\mathbb{P}^1$-GLS is definitely more accurate than the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method. However, as already mentioned and as will be confirmed in the next tests, the latter approach is more adequate in a multiquery context, where several resolutions of the advection-diffusion equation (3.1)–(3.11) are to be performed.

	cost	Error (3.84)
$\mathbb{P}^1$-GLS ($H = 1/122$)	4.76	0.00293
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ ($H = 1/16$, $h = H/7$)	4.79	0.0390

Table 3.10 Test (v)

	cost	Error (3.84)
$\mathbb{P}^1$-GLS ($H = 1/127$)	6.42	0.00485
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ ($H = 1/16$, $h = H/7$)	6.30	0.0549

Table 3.11 Test (vi)

	cost	Error (3.84)
$\mathbb{P}^1$-GLS ($H = 1/144$)	7.43	0.00143
$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ ($H = 1/16$, $h = H/9$)	7.09	0.0285

Table 3.12 Test (vii)

Fixed meshsize h

We fix the meshsize $h = 1/2048$. In order to measure the cost of the methods in a multiquery context, we distinguish, in the computational cost of the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method, (i) the offline cost, which comprises the assembling phase of the stiffness matrix, which itself involves the

pre-computation of $\sigma_{1,h}$, and (ii) the online cost equal to the resolution time for the modified advection-diffusion equation.

Tables 3.13 through 3.15 show our results for the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method and the $\mathbb{P}^1$-GLS method for the tests (v)-(vi)-(vii). Again, similar results we do not show and which lead to similar conclusions have been obtained for our tests (i) through (iv). The two columns on the left of each table show the relative accuracy obtained for different mesh sizes H (employed, we recall, for the approximation of the advection-diffusion equation). The two columns on the right allow to compare the online cost of the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method, as defined above, with the total cost of the the $\mathbb{P}^1$-GLS method. The specific function used to measure the CPU time is clock_gettime() with the clock CLOCK_PROCESS_CPUTIME_ID.

The main two conclusions are, on the one hand, that the $(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$ method is more robust and allow for larger mesh sizes than the $\mathbb{P}^1$-GLS method, and, on the other hand, that the two approaches essentially share the same cost, if we assume that $\sigma_{1,h}$ has been precomputed. Other tests, not reported on here, show that roughly ten solutions of the advection-diffusion equation are necessary to make the approach profitable if we take into account the cost to compute $\sigma_{1,h}$.

	Error (3.84)		Online cost	
$1/H$	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS
16	0.0291	0.0704	0.00107	0.000921
24	0.0190	0.0508	0.00177	0.00171
28	0.0173	0.0235	0.00250	0.00229
32	0.0135	0.0187	0.00312	0.00294
64	0.00626	0.00608	0.0138	0.0137

Table 3.13 Test (v)

	Error (3.84)		Online cost	
$1/H$	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS
16	0.0475	0.0807	0.00102	0.000804
24	0.0308	0.0615	0.00157	0.00166
28	0.0275	0.0440	0.00238	0.00232
32	0.0226	0.0258	0.00298	0.00301
64	0.0105	0.0102	0.0143	0.0139

Table 3.14 Test (vi)

	Error (3.84)		Online cost	
$1/H$	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS	$(\mathbb{P}^1, \sigma_{1,h}\mathbb{P}^1)$	$\mathbb{P}^1$-GLS
16	0.0219	0.0606	0.000976	0.000807
24	0.0139	0.0437	0.00186	0.00171
28	0.0119	0.0189	0.00256	0.00229
32	0.00946	0.0146	0.00332	0.0296
64	0.00401	0.00388	0.0156	0.0138

Table 3.15 Test (vii)

Acknowledgments

The work of the authors is partially supported by the ONR under grant N00014-15-1-2777 and by the EOARD under grant FA8655-13-1-3061. Stimulating discussions with Y. Achdou and O. Pironneau are gratefully acknowledged. We warmly thank A. Lozinski for his remarks on a draft version of this article.

3.5 Appendix: Proof of Proposition 3.16

Proposition 3.16 is shown as a consequence of a more general result, namely Theorem 3.22 below. We introduce the space

$$V = \left\{ u \in H^1(\Omega), \quad \fint_\Omega u = 1 \right\}$$

and recall (see (3.48)) that the bilinear form $a^\star$ is defined by

$$a^\star(u,v) = \int_\Omega (\nabla u + bu) \cdot \nabla v.$$

Let $f \in L^2(\Omega)$ and g be Lipschitz-continuous on $\partial\Omega$ such that $\displaystyle\int_\Omega f + \int_{\partial\Omega} g = 0$. Consider the problem

$$\text{Find } u \in V \text{ such that, for any } v \in H^1(\Omega), \quad a^\star(u,v) = \int_\Omega fv + \int_{\partial\Omega} gv. \tag{3.85}$$

Under assumptions (3.9)–(3.42)–(3.43), we have shown above (in Proposition 3.7 for the specific case $g = 0$ and in Proposition 3.10 for the case $g = b \cdot n - \displaystyle\fint_{\partial\Omega} b \cdot n$, where we actually did not use the specific expression of g) that problem (3.85) is well-posed, and that its unique solution belongs to $H^2(\Omega)$.

We here consider the Galerkin discretization of (3.85). Consider a mesh of Ω made of elements $T \in \mathcal{T}_h$. Following Proposition 3.16, we take $\Sigma_h \subset H^1(\Omega)$ the associated finite dimensional space made of continuous piecewise affine functions and introduce

$$V_h = \left\{ u \in \Sigma_h, \quad \fint_\Omega u = 1 \right\} \subset V.$$

Theorem 3.22. *We assume that (3.9)–(3.42)–(3.43) hold. Let u denote the solution to (3.85). For h sufficiently small, there exists a unique $u_h \in V_h$ solution to*

$$\forall v_h \in \Sigma_h, \qquad a^\star(u_h, v_h) = \int_\Omega f v_h + \int_{\partial\Omega} g\, v_h. \tag{3.86}$$

Furthermore, we have, for h sufficiently small,

$$\|u - u_h\|_{H^1(\Omega)} \leqslant Ch \|u\|_{H^2(\Omega)} \tag{3.87}$$

where C is independent of h.

Theorem 3.22 obviously implies Proposition 3.16. Consider indeed the invariant measure $\sigma_1 \in H^1(\Omega)$ solution to (3.16)–(3.17). It is the solution to (3.85) with $f = g = 0$. Likewise, the invariant measure $\sigma_2^0 \in H^1(\Omega)$ solution to (3.34) is the solution to (3.85) with $f = 0$ and

$g = b \cdot n - \fint_{\partial\Omega} b \cdot n$. Theorem 3.22 then implies that (3.49) is well-posed and that the error estimate (3.50) holds.

Proof of Theorem 3.22. The proof falls in two steps.

Step 1: well-posedness of (3.86). Let $\lambda > 0$. The bilinear form $a_{\rm coer}(u,v) = \displaystyle\int_\Omega \nabla v \cdot \nabla u + \lambda \int_\Omega v\, u$ is coercive in $H^1(\Omega)$, while the bilinear form $a_{\rm comp}(u,v) = \displaystyle\int_\Omega bu \cdot \nabla v$ can be represented by a compact operator $T \in \mathcal{L}\left(H^1(\Omega), \left(H^1(\Omega)\right)'\right)$ as $a_{\rm comp}(u,v) = \langle Tu, v\rangle$. Consequently (see the proof of [84, Theorem 4.2.9]),

$$\begin{gathered}\text{when } h \text{ is sufficiently small, the bilinear form} \\ a^\star_\lambda(u,v) = a_{\rm coer}(u,v) + a_{\rm comp}(u,v) \text{ satisfies an inf-sup condition on } \Sigma_h.\end{gathered} \tag{3.88}$$

Using [38, Prop. 2.21], we thus see that the problem

$$\text{Find } u_h \in \Sigma_h \text{ such that, for all } v_h \in \Sigma_h, \quad a^\star_\lambda(u_h, v_h) = \int_\Omega \overline{f} v_h + \int_{\partial\Omega} g\, v_h,$$

is well-posed for any $\overline{f} \in L^2(\Omega)$.

We now consider the iterations

$$\left\{\begin{array}{c}\text{Find } u_h^{n+1} \in \Sigma_h \text{ such that, for all } v_h \in \Sigma_h, \\ a^\star(u_h^{n+1}, v_h) + \lambda \displaystyle\int_\Omega u_h^{n+1} v_h = a^\star_\lambda(u_h^{n+1}, v_h) = \lambda \int_\Omega u_h^n v_h + \int_\Omega f v_h + \int_{\partial\Omega} g\, v_h,\end{array}\right. \tag{3.89}$$

with the initial condition $u_h^0 = |\Omega|^{-1}$ (or any function in Σ_h and of mean equal to 1). Thanks to the above argument, these problems are well-posed and define a sequence $u_h^n \in \Sigma_h \subset H^1(\Omega)$. Furthermore, taking $v_h = 1$ as test function, we see that all the functions u_h^n share the same mean, and due to the choice of u_h^0, we get $u_h^n \in V_h$ for any n.

We next prove that the sequence $\{u_h^n\}_{n\in\mathbb{N}}$ converges to a solution to (3.86). We recall that $H^1_{\int=0}(\Omega) = \left\{v \in H^1(\Omega), \quad \displaystyle\int_\Omega v = 0\right\}$. We infer from (3.89) that, for any $v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$,

$$a^\star(u_h^{n+1} - u_h^n, v_h) + \lambda \int_\Omega (u_h^{n+1} - u_h^n)\, v_h = \lambda \int_\Omega (u_h^n - u_h^{n-1})\, v_h,$$

from which we deduce that

$$\sup_{v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a^\star(u_h^{n+1} - u_h^n, v_h)}{\|v_h\|_{H^1(\Omega)}} - \lambda \|u_h^{n+1} - u_h^n\|_{H^1(\Omega)} \leqslant \lambda \|u_h^n - u_h^{n-1}\|_{H^1(\Omega)}. \tag{3.90}$$

Using the same arguments as above (this time for $\lambda = 0$ on $\Sigma_h \cap H^1_{\int=0}(\Omega)$), we have that the bilinear form $a^\star$ satisfies an inf-sup condition on $\Sigma_h \cap H^1_{\int=0}(\Omega)$ with a constant $\gamma > 0$ independent of h:

$$\inf_{w_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \sup_{v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a^\star(w_h, v_h)}{\|w_h\|_{H^1(\Omega)}\, \|v_h\|_{H^1(\Omega)}} \geqslant \gamma. \tag{3.91}$$

We thus infer from (3.90) that

$$(\gamma - \lambda)\|u_h^{n+1} - u_h^n\|_{H^1(\Omega)} \leqslant \lambda \|u_h^n - u_h^{n-1}\|_{H^1(\Omega)}.$$

Taking λ sufficiently small (so that $0 < \lambda/(\gamma - \lambda) < 1$), we obtain that the sequence $\{u_h^n\}_{n\in\mathbb{N}}$ converges in $H^1(\Omega)$ to some $u_h^\infty \in V_h$. Passing to the limit $n \to \infty$ in (3.89), we get that u_h^∞ is a solution to (3.86).

We now prove that (3.86) has a unique solution. Consider two solutions u_h and $\overline{u}_h$ to (3.86). Then $u_h - \overline{u}_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$ and satisfies $a^\star(u_h - \overline{u}_h, v_h) = 0$ for any $v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$. We deduce from (3.91) that $u_h - \overline{u}_h = 0$.

Step 2: estimate (3.87). Introducing the interpolant $I_h u \in \Sigma_h$, we deduce from (3.91) that

$$\gamma \|I_h u - u_h - c\|_{H^1(\Omega)} \leqslant \sup_{w_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a^\star(I_h u - u_h - c, w_h)}{\|w_h\|_{H^1(\Omega)}},$$

where $c = \fint_\Omega (I_h u - u_h)$. Using (3.85) and (3.86), we deduce from the above estimate that

$$\begin{aligned} \gamma \|I_h u - u_h - c\|_{H^1(\Omega)} &\leqslant \sup_{w_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a^\star(I_h u - u - c, w_h)}{\|w_h\|_{H^1(\Omega)}} \\ &\leqslant \big(1 + \|b\|_{L^\infty(\Omega)}\big) \|I_h u - u - c\|_{H^1(\Omega)}. \end{aligned} \tag{3.92}$$

We next write that

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leqslant \|u - I_h u + c\|_{H^1(\Omega)} + \|I_h u - u_h - c\|_{H^1(\Omega)} \\ &\leqslant C \|u - I_h u + c\|_{H^1(\Omega)} \quad \text{[using (3.92)]} \\ &\leqslant C \|\nabla(u - I_h u)\|_{L^2(\Omega)}, \end{aligned}$$

where, in the last line, we have used the Poincaré-Wirtinger inequality, as a consequence of the fact that $c = \fint_\Omega (I_h u - u_h) = \fint_\Omega (I_h u - u)$. We then conclude using the approximation result $\|u - I_h u\|_{H^1(\Omega)} \leqslant Ch\|u\|_{H^2(\Omega)}$. This yields (3.87). □

3.6 Appendix: Proof of Proposition 3.18

In order to prove Proposition 3.18, we follow and adapt the arguments of [15, Chap. 8]. The proof relies on several technical results, the proof of which are given in the subsequent appendices 3.7 and 3.8.

We fix some $0 < \eta \leqslant 1$, and we recall (see (3.60)) that the bilinear form $a^\star_\eta$ is defined by

$$a^\star_\eta(u, v) = \int_\Omega (\nabla u + bu) \cdot \nabla v + \eta \int_\Omega u\, v.$$

We also define $a_\eta(u, v) = a^\star_\eta(v, u)$.

3.6.1 Elliptic regularity results

Let q be such that $1 < q < +\infty$ if $d = 2$ and $2d/(d+2) \leqslant q < +\infty$ otherwise, and let $f \in L^q(\Omega)$. We consider the problem

$$\text{Find } u \in H^1(\Omega) \text{ such that, for any } v \in H^1(\Omega), \quad a^\star_\eta(u, v) = \int_\Omega f v, \tag{3.93}$$

for which we have the following result.

Lemma 3.23. *We work under the assumptions of Proposition 3.18 and assume that* $0 < \eta \leqslant 1$ *and* $f \in L^q(\Omega)$ *with* q *chosen as above. Then Problem (3.93) has a unique solution* $u \in H^1(\Omega)$. *In addition, if* η *is sufficiently small, then* $u \in W^{2,q}(\Omega)$ *and it satisfies*

$$\left\| u - \sigma_1 \fint_\Omega u \right\|_{W^{2,q}(\Omega)} \leqslant C \|f\|_{L^q(\Omega)} \tag{3.94}$$

for some C *independent of* η *and* f, *where* σ_1 *is the invariant measure defined by (3.16)–(3.17)–(3.18).*

We only prove this result in dimension $2 \leqslant d \leqslant 3$ (see the assumptions of Proposition 3.18), but it certainly holds for larger dimensions.

Proof. The proof falls in two steps.

Step 1: existence and uniqueness of a solution. We first show that the bilinear form $a_\eta^\star$ satisfies the (BNB1) condition on $H^1(\Omega)$. Using the invariant measure σ_1 defined by (3.16)–(3.17)–(3.18), a simple computation indeed yields that, for any $v \in H^1(\Omega)$,

$$\begin{aligned} a_\eta(v, \sigma_1 v) &= \int_\Omega \sigma_1 |\nabla v|^2 + \frac{1}{2} \int_\Omega (\nabla \sigma_1 + b\sigma_1) \cdot \nabla(v^2) + \eta \int_\Omega \sigma_1 \, v^2 \\ &= \int_\Omega \sigma_1 |\nabla v|^2 + \eta \int_\Omega \sigma_1 \, v^2 \qquad (3.95) \\ &\geqslant (\inf \sigma_1) \min(1, \eta) \, \|v\|^2_{H^1(\Omega)} \\ &\geqslant c \min(1, \eta) \, \|v\|_{H^1(\Omega)} \|\sigma_1 v\|_{H^1(\Omega)}, \qquad (3.96) \end{aligned}$$

for some $c > 0$ independent of η. For any $w \in H^1(\Omega)$, we set $v = \sigma_1^{-1} w$, which belongs to $H^1(\Omega)$, and thus have

$$\begin{aligned} a_\eta^\star(w, \sigma_1^{-1} w) = a_\eta(\sigma_1^{-1} w, w) &\geqslant c \min(1, \eta) \, \|v\|_{H^1(\Omega)} \|\sigma_1 v\|_{H^1(\Omega)} \\ &= c \min(1, \eta) \, \|w\|_{H^1(\Omega)} \|\sigma_1^{-1} w\|_{H^1(\Omega)}. \qquad (3.97) \end{aligned}$$

We thus deduce the (BNB1) condition.

The bilinear form $a_\eta^\star$ also satisfies the (BNB2) condition on $H^1(\Omega)$. Indeed, if $v \in H^1(\Omega)$ is such that $a_\eta^\star(u, v) = 0$ for any $u \in H^1(\Omega)$, then we have $a_\eta^\star(\sigma_1 v, v) = 0$, and we have thus found a function $w \in H^1(\Omega)$ (namely $w = \sigma_1 v$) such that $a_\eta^\star(w, \sigma_1^{-1} w) = 0$. The estimate (3.97) shows that w (and hence v) vanishes, which implies the (BNB2) condition on $H^1(\Omega)$.

We have chosen $f \in L^q(\Omega)$ with an exponent q such that $f \in (H^1(\Omega))'$. We thus obtain that Problem (3.93) is well-posed.

Step 2: $W^{2,q}$ estimate. Introduce $\overline{u} = u + \sigma_1 - \sigma_1 \fint_\Omega u$, which satisfies

$$\begin{cases} -\operatorname{div}(\nabla \overline{u} + b\overline{u}) = F \ \text{ in } \Omega, & \fint_\Omega \overline{u} = 1, \\ (\nabla \overline{u} + b\overline{u}) \cdot n = 0 \ \text{ on } \partial\Omega, \end{cases}$$

with

$$F = f - \eta u = f - \eta \overline{u} + \eta \sigma_1 - \eta \sigma_1 \fint_\Omega u.$$

Using $v \equiv 1$ as test function in (3.93), we see that $\eta \fint_\Omega u = \fint_\Omega f$. We hence get that

$$F = f - \eta(\overline{u} - \sigma_1) - \sigma_1 \fint_\Omega f.$$

Using that $\sigma_1 \in W^{2,s}(\Omega)$ for any $1 < s < \infty$, we deduce that, for any $s > 1$,

$$\|F\|_{L^s(\Omega)} \leqslant C\|f\|_{L^s(\Omega)} + \eta\|\overline{u} - \sigma_1\|_{L^s(\Omega)}, \tag{3.98}$$

where we only know, at this stage, that $\overline{u} \in H^1(\Omega)$. To proceed and obtain a $W^{2,q}$ estimate, we distinguish two cases, whether $d = 2$ or $d = 3$.

Suppose first that $d = 2$. Using the continuous injection $H^1(\Omega) \subset L^q(\Omega)$, we deduce from (3.98) (written with $s = q$) that $F \in L^q(\Omega)$. We are thus in position to apply Proposition 3.7, which yields (see (3.24)) that there exists some C independent of η such that

$$\|\overline{u} - \sigma_1\|_{W^{2,q}(\Omega)} \leqslant C\|F\|_{L^q(\Omega)} \leqslant C\|f\|_{L^q(\Omega)} + C\eta\|\overline{u} - \sigma_1\|_{L^q(\Omega)}.$$

For η sufficiently small, this implies (3.94).

Suppose now that $d = 3$. If $q \leqslant 6$, we proceed as above, using the continuous injection $H^1(\Omega) \subset L^q(\Omega)$. We now turn to the case $q > 6$. Using (3.98) for $s = 6$, we deduce that $F \in L^6(\Omega)$. Applying Proposition 3.7, we obtain that

$$\|\overline{u} - \sigma_1\|_{W^{2,6}(\Omega)} \leqslant C\|f\|_{L^q(\Omega)} + C\eta\|\overline{u} - \sigma_1\|_{L^6(\Omega)},$$

which implies that $\|\overline{u} - \sigma_1\|_{W^{2,6}(\Omega)} \leqslant C\|f\|_{L^q(\Omega)}$. Using the continuous injection $W^{2,6}(\Omega) \subset L^\infty(\Omega)$, we deduce that $\|\overline{u} - \sigma_1\|_{L^\infty(\Omega)} \leqslant C\|f\|_{L^q(\Omega)}$. The estimate (3.98), written with $s = q$, now yields

$$\|F\|_{L^q(\Omega)} \leqslant C\|f\|_{L^q(\Omega)} + \eta\|\overline{u} - \sigma_1\|_{L^q(\Omega)} \leqslant C\|f\|_{L^q(\Omega)},$$

from which, applying again Proposition 3.7, we infer (3.94). □

Likewise, for any $f \in L^2(\Omega)$, we consider the problem

$$\text{Find } u \in H^1(\Omega) \text{ s.t., for any } v \in H^1(\Omega), \quad a_\eta^\star(v,u) = a_\eta(u,v) = \int_\Omega fv, \tag{3.99}$$

for which we have the following result.

Lemma 3.24. *We work under the assumptions of Proposition 3.18 and assume that $0 < \eta \leqslant 1$ and $f \in L^2(\Omega)$. Then Problem (3.99) has a unique solution $u \in H^1(\Omega)$. In addition, $u \in H^2(\Omega)$ and it satisfies*

$$\left\| u - \fint_\Omega u \right\|_{H^2(\Omega)} \leqslant C\|f\|_{L^2(\Omega)} \tag{3.100}$$

for some C independent of η and f.

A similar result certainly holds for $f \in L^q(\Omega)$ with q chosen such that $f \in (H^1(\Omega))'$, yielding a control on $\left\| u - \fint_\Omega u \right\|_{W^{2,q}(\Omega)}$. We will however not need such a result and therefore do not

pursue in that direction. As for Lemma 3.23, we only prove Lemma 3.24 in dimension $2 \leqslant d \leqslant 3$ (see the assumptions of Proposition 3.18), but it certainly holds for larger dimensions.

Proof. The proof falls in three steps.

Step 1: existence and uniqueness of a solution. The bilinear form a_η satisfies the (BNB1) condition on $H^1(\Omega)$, as a direct consequence of (3.96). It also satisfies the (BNB2) condition on $H^1(\Omega)$. Indeed, if $v \in H^1(\Omega)$ is such that $a_\eta(u,v) = 0$ for any $u \in H^1(\Omega)$, then we have $a_\eta(\sigma_1^{-1}v, v) = 0$, and we have thus found a function $w \in H^1(\Omega)$ (namely $w = \sigma_1^{-1}v$) such that $a_\eta(w, \sigma_1 w) = 0$. The estimate (3.96) shows that w (and hence v) vanishes, which implies the (BNB2) condition on $H^1(\Omega)$. We thus obtain that Problem (3.99) is well-posed.

Step 2: H^1 estimate. We claim that the solution u to (3.99) satisfies

$$\left\| u - \fint_\Omega u \right\|_{H^1(\Omega)} \leqslant C \|f\|_{L^2(\Omega)} \tag{3.101}$$

for some C independent of η and f. Consider indeed $\overline{u} = u - \fint_\Omega u$. Using the Poincaré-Wirtinger inequality and (3.95) for the function $\overline{u}$, we have

$$\begin{aligned}
\|\overline{u}\|^2_{H^1(\Omega)} &\leqslant C \, \|\nabla \overline{u}\|^2_{L^2(\Omega)} \\
&\leqslant C a_\eta(\overline{u}, \sigma_1 \overline{u}) \\
&= C a_\eta(u, \sigma_1 \overline{u}) - C a_\eta(1, \sigma_1 \overline{u}) \fint_\Omega u \\
&= C \int_\Omega f \, \sigma_1 \overline{u} - C\eta \int_\Omega \sigma_1 \overline{u} \fint_\Omega u \\
&\leqslant C \|\sigma_1\|_{L^\infty(\Omega)} \|f\|_{L^2(\Omega)} \|\overline{u}\|_{H^1(\Omega)} + C\eta \|\sigma_1\|_{L^\infty(\Omega)} \|\overline{u}\|_{H^1(\Omega)} \|u\|_{L^2(\Omega)}.
\end{aligned}$$

We hence deduce that

$$\|\overline{u}\|_{H^1(\Omega)} \leqslant C \left(\|f\|_{L^2(\Omega)} + \eta \|u\|_{L^2(\Omega)} \right). \tag{3.102}$$

We now write (3.95) for the function u, from which we deduce that

$$\|\nabla u\|^2_{L^2(\Omega)} + \eta \|u\|^2_{L^2(\Omega)} \leqslant C a_\eta(u, \sigma_1 u) = C \int_\Omega f \, \sigma_1 u \leqslant C \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)},$$

which implies that

$$\eta \|u\|_{L^2(\Omega)} \leqslant C \|f\|_{L^2(\Omega)}. \tag{3.103}$$

Inserting this estimate in (3.102), we obtain the claimed estimate (3.101).

Step 3: H^2 estimate. We proceed as in the proof of Proposition 3.7. Introducing again $\overline{u} = u - \fint_\Omega u$, we observe that

$$\begin{cases} -\Delta \overline{u} = F \ \text{ in } \Omega, & \fint_\Omega \overline{u} = 0, \\ \nabla \overline{u} \cdot n = 0 \ \text{ on } \partial\Omega, \end{cases}$$

with $F = f - b\nabla\overline{u} - \eta\overline{u} - \eta \fint_\Omega u$. We compute that

$$\|F\|_{L^2(\Omega)} \leqslant \|f\|_{L^2(\Omega)} + \|b\|_{L^\infty(\Omega)}\|\overline{u}\|_{H^1(\Omega)} + \eta\|\overline{u}\|_{L^2(\Omega)} + \eta\|u\|_{L^2(\Omega)}$$

and we deduce, using (3.101) and (3.103), that $\|F\|_{L^2(\Omega)} \leqslant C\|f\|_{L^2(\Omega)}$. We are then in position to use [38, Theorem 3.12], which implies that $\overline{u} \in H^2(\Omega)$ with

$$\|\overline{u}\|_{H^2(\Omega)} \leqslant C\|F\|_{L^2(\Omega)} \leqslant C\|f\|_{L^2(\Omega)}.$$

This concludes the proof of (3.100). □

3.6.2 Discretized problems

We now consider the discretization of Problems (3.93) and (3.99). Let $f \in L^2(\Omega)$ and let $\Sigma_h \subset H^1(\Omega)$ be the $\mathbb{P}^1$ approximation space associated to a regular quasi-uniform polyhedral mesh of Ω.

Theorem 3.25 (Discretization of Problem (3.93))**.** *We assume that* (3.9)–(3.42)–(3.43) *hold, and that* $0 < \eta \leqslant 1$*. Then, there exists* h_0 *independent of* η *such that, for sufficiently small* η *and any* $h \leqslant h_0$*, there exists a unique* $u_h \in \Sigma_h$ *solution to*

$$\forall v_h \in \Sigma_h, \qquad a^\star_\eta(u_h, v_h) = \int_\Omega f v_h. \tag{3.104}$$

Proof. The proof follows the lines of that of Theorem 3.22. Let $\lambda > 0$. We consider the iterations

$$\left\{ \begin{array}{c} \text{Find } u_h^{n+1} \in \Sigma_h \text{ such that, for all } v_h \in \Sigma_h, \\ a^\star(u_h^{n+1}, v_h) + \lambda \displaystyle\int_\Omega u_h^{n+1}\, v_h = a^\star_\lambda(u_h^{n+1}, v_h) = (\lambda - \eta) \int_\Omega u_h^n\, v_h + \int_\Omega f v_h, \end{array} \right. \tag{3.105}$$

with the initial condition $u_h^0 = \eta^{-1} \fint_\Omega f$ (or any function in Σ_h such that $\eta \fint_\Omega u_h^0 = \fint_\Omega f$). In view of (3.88), these problems are well-posed for any $h \leqslant h_0$, where h_0 is independent of η. They thus define a sequence $u_h^n \in \Sigma_h \subset H^1(\Omega)$. Furthermore, taking $v_h = 1$ as test function, we see that

$$\lambda \int_\Omega u_h^{n+1} = (\lambda - \eta) \int_\Omega u_h^n + \int_\Omega f.$$

Our choice of u_h^0 implies that all the functions u_h^n share the same mean.

We next prove that the sequence $\{u_h^n\}_{n\in\mathbb{N}}$ converges to a solution to (3.104). We recall that $H^1_{\int=0}(\Omega) = \left\{ v \in H^1(\Omega), \quad \displaystyle\int_\Omega v = 0 \right\}$. We infer from (3.105) that, for any $v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$,

$$a^\star(u_h^{n+1} - u_h^n, v_h) + \lambda \int_\Omega (u_h^{n+1} - u_h^n)\, v_h = (\lambda - \eta) \int_\Omega (u_h^n - u_h^{n-1})\, v_h,$$

from which we deduce that

$$\sup_{v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a^\star(u_h^{n+1} - u_h^n, v_h)}{\|v_h\|_{H^1(\Omega)}} - \lambda\|u_h^{n+1} - u_h^n\|_{H^1(\Omega)} \leqslant (\lambda - \eta)\|u_h^n - u_h^{n-1}\|_{H^1(\Omega)}. \tag{3.106}$$

Using (3.91), we infer from (3.106) that

$$(\gamma - \lambda)\|u_h^{n+1} - u_h^n\|_{H^1(\Omega)} \leqslant (\lambda - \eta)\|u_h^n - u_h^{n-1}\|_{H^1(\Omega)}.$$

Taking λ sufficiently small (so that $0 < \lambda/(\gamma - \lambda) < 1$) and $\eta < \lambda$, we obtain that the sequence $\{u_h^n\}_{n\in\mathbb{N}}$ converges in $H^1(\Omega)$ to some $u_h^\infty \in \Sigma_h$. Passing to the limit $n \to \infty$ in (3.105), we get that u_h^∞ is a solution to (3.104).

We now prove that (3.104) has a unique solution. Consider two solutions u_h and $\overline{u}_h$ to (3.104). Then $u_h - \overline{u}_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$ and satisfies $a^\star(u_h - \overline{u}_h, v_h) = -\eta \displaystyle\int_\Omega (u_h - \overline{u}_h)\, v_h$ for any $v_h \in \Sigma_h$. We deduce from (3.91) that

$$\gamma\|u_h - \overline{u}_h\|_{H^1(\Omega)} \leqslant \eta\|u_h - \overline{u}_h\|_{H^1(\Omega)},$$

which implies, whenever $\eta < \gamma$, that $u_h = \overline{u}_h$. □

Theorem 3.26 (Discretization of Problem (3.99))**.** *We assume that* (3.9)–(3.42)–(3.43) *hold, and that* $0 < \eta \leqslant 1$*. Then, there exists* h_0 *independent of* η *such that, for sufficiently small* η *and any* $h \leqslant h_0$*, there exists a unique* $u_h \in \Sigma_h$ *solution to*

$$\forall v_h \in \Sigma_h, \qquad a_\eta(u_h, v_h) = \int_\Omega f v_h. \tag{3.107}$$

Proof. The proof follows the lines of that of Theorem 3.25. We consider the iterations

$$\left\{ \begin{array}{c} \text{Find } u_h^{n+1} \in \Sigma_h \text{ such that, for all } v_h \in \Sigma_h, \\ a(u_h^{n+1}, v_h) + \lambda \displaystyle\int_\Omega u_h^{n+1}\, v_h = a_\lambda(u_h^{n+1}, v_h) = (\lambda - \eta) \int_\Omega u_h^n\, v_h + \int_\Omega f v_h, \end{array} \right. \tag{3.108}$$

with an initial condition u_h^0 such that $\eta \fint_\Omega u_h^0\, \sigma_{1,h} = \fint_\Omega f\, \sigma_{1,h}$ (where $\sigma_{1,h}$ satisfies (3.49) with $g \equiv 0$). We have shown in Proposition 3.16 that $\|\sigma_{1,h} - \sigma_1\|_{H^1(\Omega)} \leqslant Ch$, and we have shown in Lemma 3.4 that σ_1 is positive and bounded away from 0. We hence have $\fint_\Omega \sigma_{1,h} > 0$ when h is sufficiently small, and it is thus possible to pick u_h^0 as a constant function.

The problems (3.108) are well-posed for any $h \leqslant h_0$, where h_0 is independent of η. Consider indeed a basis $(\varphi_i)_{1\leqslant i\leqslant I}$ of Σ_h. Since $a_\lambda^\star$ satisfies the inf-sup condition (3.88) on Σ_h as soon as $h \leqslant h_0$, we know that the matrix K, defined by $K_{ij} = a_\lambda^\star(\varphi_j, \varphi_i)$ for any $1 \leqslant i, j \leqslant I$, is invertible. The matrix K^T is therefore invertible. This implies that (3.108) are indeed well-posed for any $h \leqslant h_0$, and thus define a sequence $u_h^n \in \Sigma_h \subset H^1(\Omega)$. Furthermore, taking $v_h = \sigma_{1,h}$ as test function, we see that

$$\lambda \int_\Omega u_h^{n+1}\, \sigma_{1,h} = (\lambda - \eta) \int_\Omega u_h^n\, \sigma_{1,h} + \int_\Omega f\, \sigma_{1,h}.$$

Our choice of u_h^0 implies that, for any n, we have

$$\eta \fint_\Omega u_h^n\, \sigma_{1,h} = \fint_\Omega f\, \sigma_{1,h}. \tag{3.109}$$

Let $\overline{u}_h^n = u_h^n - \fint_\Omega u_h^n$. We infer from (3.108) that, for any $v_h \in \Sigma_h$,

$$a(\overline{u}_h^{n+1}, v_h) + \lambda \int_\Omega u_h^{n+1} v_h = (\lambda - \eta) \int_\Omega u_h^n v_h + \int_\Omega f v_h.$$

Taking now $v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$, where we recall that $H^1_{\int=0}(\Omega) = \left\{ v \in H^1(\Omega), \quad \int_\Omega v = 0 \right\}$, we get

$$a(\overline{u}_h^{n+1}, v_h) + \lambda \int_\Omega \overline{u}_h^{n+1} v_h = (\lambda - \eta) \int_\Omega \overline{u}_h^n v_h + \int_\Omega f v_h,$$

hence

$$a(\overline{u}_h^{n+1} - \overline{u}_h^n, v_h) + \lambda \int_\Omega (\overline{u}_h^{n+1} - \overline{u}_h^n) v_h = (\lambda - \eta) \int_\Omega (\overline{u}_h^n - \overline{u}_h^{n-1}) v_h,$$

from which we deduce that

$$\sup_{v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a(\overline{u}_h^{n+1} - \overline{u}_h^n, v_h)}{\|v_h\|_{H^1(\Omega)}} - \lambda \|\overline{u}_h^{n+1} - \overline{u}_h^n\|_{H^1(\Omega)} \leqslant (\lambda - \eta) \|\overline{u}_h^n - \overline{u}_h^{n-1}\|_{H^1(\Omega)}. \tag{3.110}$$

The bilinear form $a^\star$ satisfies an infsup condition on $\Sigma_h \cap H^1_{\int=0}(\Omega)$ for h sufficiently small (see (3.91)). Considering a basis $(\varphi_i)_{1 \leqslant i \leqslant I}$ of $\Sigma_h \cap H^1_{\int=0}(\Omega)$, we get that the matrix K defined by $K_{ij} = a^\star(\varphi_j, \varphi_i)$ for any $1 \leqslant i, j \leqslant I$, is invertible. The matrix K^T is therefore invertible, which implies that the bilinear form a also satisfies an infsup condition on $\Sigma_h \cap H^1_{\int=0}(\Omega)$ (for h sufficiently small):

$$\inf_{w_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \sup_{v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)} \frac{a(w_h, v_h)}{\|w_h\|_{H^1(\Omega)} \, \|v_h\|_{H^1(\Omega)}} \geqslant \gamma_a. \tag{3.111}$$

We then infer from (3.110) that

$$(\gamma_a - \lambda) \|\overline{u}_h^{n+1} - \overline{u}_h^n\|_{H^1(\Omega)} \leqslant (\lambda - \eta) \|\overline{u}_h^n - \overline{u}_h^{n-1}\|_{H^1(\Omega)}.$$

Taking λ sufficiently small (so that $0 < \lambda/(\gamma_a - \lambda) < 1$) and $\eta < \lambda$, we obtain that the sequence $\{\overline{u}_h^n\}_{n \in \mathbb{N}}$ converges in $H^1(\Omega)$ to some $\overline{u}_h^\infty \in \Sigma_h$. We also deduce from (3.109) that

$$\eta \fint_\Omega \overline{u}_h^n \, \sigma_{1,h} + \eta \fint_\Omega u_h^n \fint_\Omega \sigma_{1,h} = \fint_\Omega f \, \sigma_{1,h}.$$

Since $\fint_\Omega \sigma_{1,h} \neq 0$, we obtain that $\fint_\Omega u_h^n$ converges to some ℓ satisfying

$$\eta \fint_\Omega \overline{u}_h^\infty \, \sigma_{1,h} + \eta \, \ell \fint_\Omega \sigma_{1,h} = \fint_\Omega f \, \sigma_{1,h}.$$

We thus get that the sequence $\{u_h^n\}_{n \in \mathbb{N}}$ converges in $H^1(\Omega)$ to $u_h^\infty := \ell + \overline{u}_h^\infty \in \Sigma_h$.

Passing to the limit $n \to \infty$ in (3.108), we get that u_h^∞ is a solution to (3.107).

We now prove that (3.107) has a unique solution. Consider two solutions $u_{1,h}$ and $u_{2,h}$ to (3.107). Introduce $\overline{u}_{1,h} = u_{1,h} - \fint_\Omega u_{1,h}$ and likewise for $u_{2,h}$. Then $\overline{u}_{1,h} - \overline{u}_{2,h} \in \Sigma_h \cap H^1_{\int=0}(\Omega)$ and satisfies $a(\overline{u}_{1,h} - \overline{u}_{2,h}, v_h) = -\eta \int_\Omega (\overline{u}_{1,h} - \overline{u}_{2,h}) v_h$ for any $v_h \in \Sigma_h \cap H^1_{\int=0}(\Omega)$. We deduce

from (3.111) that

$$\gamma_a \|\overline{u}_{1,h} - \overline{u}_{2,h}\|_{H^1(\Omega)} \leqslant \eta \|\overline{u}_{1,h} - \overline{u}_{2,h}\|_{H^1(\Omega)},$$

which implies, whenever $\eta < \gamma_a$, that $\overline{u}_{1,h} = \overline{u}_{2,h}$. The functions $u_{1,h}$ and $u_{2,h}$ are thus equal up to the addition of a constant. Taking $v_h = \sigma_{1,h}$ as test function in (3.107), we see that

$$\eta \int_\Omega u_{1,h}\, \sigma_{1,h} = \int_\Omega f\, \sigma_{1,h}$$

and likewise for $u_{2,h}$, which implies that $u_{1,h} = u_{2,h}$. □

3.6.3 Weight function

For some $\kappa \geqslant 1$, we set, for any x and z in Ω,

$$\chi_z(x) = \sqrt{|x - z|^2 + (\kappa h)^2}. \tag{3.112}$$

It is easy to show that there exists C (independent of κ and h) such that

$$\forall T \in \mathcal{T}_h, \quad \forall z \in \Omega, \quad \frac{\sup_{x \in T} \chi_z(x)}{\inf_{x \in T} \chi_z(x)} \leqslant C.$$

Likewise, for any $\beta \in \mathbb{N}^d$ and $\Lambda \in \mathbb{R}$, there exists C (independent of κ and h) such that, for any x and z in Ω,

$$\left|\partial_\beta(\chi_z^\Lambda)(x)\right| \leqslant C \chi_z^{\Lambda - |\beta|}(x). \tag{3.113}$$

The following estimate will be useful:

Lemma 3.27. *For any real number $\theta > d$, we have, for any $x \in \Omega$,*

$$\int_\Omega \chi_z^{-\theta}(x)\, dz \leqslant C_d \frac{1}{(\kappa h)^{\theta - d}} \left(\frac{1}{d} + \frac{1}{\theta - d}\right), \tag{3.114}$$

where C_d only depends on the dimension d.

Proof. Let R be the diameter of Ω, so that, for any $x \in \Omega$, we have $\Omega \subset B_R(x)$. We write

$$\int_\Omega \chi_z^{-\theta}(x) dz \leqslant \int_{B_R(x)} \chi_z^{-\theta}(x) dz = C_d \int_0^R \frac{r^{d-1}}{(r^2 + (\kappa h)^2)^{\theta/2}} dr.$$

We split the integral from $r = 0$ to $r = \kappa h$ (for which we write that $r^2 + (\kappa h)^2 \geqslant (\kappa h)^2$) and from $r = \kappa h$ to $r = R$ (for which we write that $r^2 + (\kappa h)^2 \geqslant r^2$). A straightforward computation then leads to (3.114). □

We now recall some useful properties of Σ_h, the subset of $H^1(\Omega)$ of piecewise affine functions. First, for any $\Lambda \in \mathbb{R}$, there exists C such that, for any $\psi \in H^1(\Omega)$ such that $\psi_{|T} \in H^2(T)$ for any $T \in \mathcal{T}_h$, there exists $I_h\psi \in \Sigma_h$ such that

$$\int_\Omega \chi_z^\Lambda (\psi - I_h\psi)^2 + h^2 \int_\Omega \chi_z^\Lambda |\nabla(\psi - I_h\psi)|^2 \leqslant C h^4 \sum_{T \in \mathcal{T}_h} \int_T \chi_z^\Lambda |\nabla^2 \psi|^2 \tag{3.115}$$

where C is independent of z, κ, h and ψ. Second, we have, for any $\Lambda \in \mathbb{R}$ and any $\psi_h \in \Sigma_h$, that

$$\sum_{T\in\mathcal{T}_h} \int_T \chi_z^\Lambda |\nabla \psi_h|^2 \leqslant C h^{-2} \int_\Omega \chi_z^\Lambda \, \psi_h^2. \tag{3.116}$$

3.6.4 Numerical Green functions

For any element $T \in \mathcal{T}_h$ of the mesh, we introduce a function $\overline{\delta}_T \in C_0^\infty(T)$ such that $\overline{\delta}_T \geqslant 0$ on T and $\displaystyle\int_T \overline{\delta}_T = 1$.

Let $z \in \Omega$ such that z does not lie on an edge of the mesh. We call K^z the element containing z, and set $\delta^z = \overline{\delta}_{K^z}$. For any function P which is piecewise constant on $\mathcal{T}_h$, we thus have

$$\int_\Omega \delta^z P = \int_{K^z} \delta^z P = P(z) \int_{K^z} \delta^z = P(z). \tag{3.117}$$

It is possible to build δ^z such that it satisfies the following bounds:

$$\forall k \in \mathbb{N}, \quad \|\nabla^k \delta^z\|_{L^\infty(\Omega)} \leqslant \frac{C_k}{h^{d+k}}.$$

Note that δ^z depends on h which is the diameter of K^z.

Let $\nu \in \mathbb{R}^d$ be a constant vector. Since $\delta^z \in C_0^\infty(K^z)$, we have that $\nu \cdot \nabla \delta^z \in L^2(\Omega)$. Problem (3.99) is thus well posed for the right-hand side $\nu \cdot \nabla \delta^z$ (see Lemma 3.24). We hence define $g^z \in H^1(\Omega)$ such that

$$\forall v \in H^1(\Omega), \quad a_\eta^\star(v, g^z) = -\int_\Omega (\nu \cdot \nabla \delta^z)\, v. \tag{3.118}$$

Likewise, we introduce $g_h^z \in \Sigma_h$ such that

$$\forall v \in \Sigma_h, \quad a_\eta^\star(v, g_h^z) = -\int_\Omega (\nu \cdot \nabla \delta^z)\, v. \tag{3.119}$$

In view of Theorem 3.26, we know that there exists h_0 independent of η such that, for any $h < h_0$, the above problem is well-posed.

For any $\lambda > 0$, we define

$$M_{h,\lambda} = \sup_{z\in\Omega,\ z \text{ not on edges}} \sqrt{\int_\Omega \chi_z^{d+\lambda} \Big(|g^z - g_h^z|^2 + |\nabla(g^z - g_h^z)|^2 \Big)} \tag{3.120}$$

with χ_z defined by (3.112). The following lemma will be most useful:

Lemma 3.28. *We work under the assumptions of Proposition 3.18. Then there exists $h_0 > 0$, $\lambda > 0$, $\kappa \geqslant 1$ (possibly depending on η) and $C_{\kappa,\lambda,\eta}$ (possibly depending of κ, λ and η) such that, for any h such that $0 < h \leqslant h_0$ and $\kappa h \leqslant 1$, we have*

$$M_{h,\lambda}^2 \leqslant C_{\kappa,\lambda,\eta}\, h^\lambda.$$

The proof of this lemma is postponed until Appendix 3.7. The restriction $h \leqslant h_0$ comes from the fact that the existence of g_h^z is only ensured for sufficiently small h. A careful inspection of the proof shows that one can take $\kappa = C/\eta$ and $C_{\kappa,\lambda,\eta} = C\,\kappa^{4+d+\lambda}$.

We proceed in the sequel of this Appendix 3.6 with the proof of Proposition 3.18.

3.6.5 Proof of (3.62)

Let u and u_h (resp. in $H^1(\Omega)$ and Σ_h) satisfying the assumptions of Proposition 3.18. We write, for any fixed z not lying on the mesh edges, that

$$\begin{aligned}
& \nu \cdot \nabla u_h(z) \\
= & \int_\Omega \delta^z \, (\nu \cdot \nabla u_h) \qquad \text{[eq. (3.117)]} \\
= & -\int_\Omega (\nu \cdot \nabla \delta^z) u_h \qquad \text{[int. by part and } \delta^z = 0 \text{ on } \partial\Omega] \\
= & \, a^\star_\eta(u_h, g^z_h) \qquad \text{[def. (3.119) of } g^z_h \text{ and } u_h \in \Sigma_h] \\
= & \, a^\star_\eta(u, g^z_h) \qquad \text{[Assumption (3.61) and } g^z_h \in \Sigma_h] \\
= & \, a^\star_\eta(u, g^z) + a^\star_\eta(u, g^z_h - g^z) \\
= & -\int_\Omega (\nu \cdot \nabla \delta^z) u + a^\star_\eta(u, g^z_h - g^z) \qquad \text{[def. (3.118) of } g^z \text{ and } u \in H^1(\Omega)] \\
= & \int_\Omega \delta^z \, (\nu \cdot \nabla u) + a^\star_\eta(u, g^z_h - g^z). \qquad \text{[int. by part and } \delta^z = 0 \text{ on } \partial\Omega]
\end{aligned} \tag{3.121}$$

In Sections 3.6.5 and 3.6.5 below, we successively bound the two terms of the right-hand side of (3.121). In Section 3.6.5, we conclude the proof of (3.62).

Bound on the second term of the right hand side of (3.121)

In view of the assumptions of Proposition 3.18, we know that $u \in W^{1,p}(\Omega)$ for some $p \geqslant 2$. Let $1 < q \leqslant 2$ such that

$$1 = \frac{1}{p} + \frac{1}{q}.$$

We know that $g^z_h - g^z \in H^1(\Omega) \subset W^{1,q}(\Omega)$. By Hölder inequality and since $0 < \eta \leqslant 1$, we write that

$$\begin{aligned}
& |a^\star_\eta(u, g^z_h - g^z)| \\
& \leqslant \int_\Omega |\nabla u| \, |\nabla(g^z_h - g^z)| + \|b\|_{L^\infty} \int_\Omega |u| \, |\nabla(g^z_h - g^z)| + \int_\Omega |u| \, |g^z_h - g^z| \\
& \leqslant (1 + \|b\|_{L^\infty}) \left(\|\tau_z^{-1} \nabla u\|_{L^p(\Omega)} + \|\tau_z^{-1} u\|_{L^p(\Omega)} \right) \\
& \qquad \times \left(\|\tau_z \nabla(g^z_h - g^z)\|_{L^q(\Omega)} + \|\tau_z (g^z_h - g^z)\|_{L^q(\Omega)} \right)
\end{aligned} \tag{3.122}$$

where the function τ_z is defined by (3.124) below. Since $q \leqslant 2$, there exists $s > 1$ such that $1 = 1/s + q/2$ (if $q = 2$, we take $s = \infty$). Introducing real numbers α and β such that $\alpha + \beta = 1$, we write

$$\|\tau_z \nabla(g^z_h - g^z)\|^q_{L^q(\Omega)} \leqslant \|\tau_z^{\alpha q}\|_{L^s(\Omega)} \left\| \tau_z^{\beta q} |\nabla(g^z_h - g^z)|^q \right\|_{L^{2/q}(\Omega)}.$$

We hence have

$$\|\tau_z \nabla(g^z_h - g^z)\|^2_{L^q(\Omega)} \leqslant \|\tau_z^{\alpha q}\|^{2/q}_{L^s(\Omega)} \int_\Omega \tau_z^{2\beta} |\nabla(g^z_h - g^z)|^2. \tag{3.123}$$

Inspired by [79], we take

$$\tau_z = \chi_z^{(d+\lambda)/p}, \qquad 2\beta = p, \tag{3.124}$$

where $\lambda > 0$ and the parameter $\kappa \geqslant 1$ in the definition (3.112) of χ_z are defined in Lemma 3.28. We hence have $\tau_z^{2\beta} = \chi_z^{d+\lambda}$. In view of the definition (3.120) of $M_{h,\lambda}$, we infer from (3.123) that

$$\|\tau_z \nabla(g_h^z - g^z)\|_{L^q(\Omega)}^2 \leqslant M_{h,\lambda}^2 \, \|\tau_z^{\alpha q}\|_{L^s(\Omega)}^{2/q} . \tag{3.125}$$

Our choice of β implies that $\alpha = 1-\beta = 1-p/2$ and $\alpha q s = -p$, hence $\|\tau_z^{\alpha q}\|_{L^s(\Omega)}^s = \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}$, and therefore

$$\|\tau_z^{\alpha q}\|_{L^s(\Omega)}^{2/q} = \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}^{2/(qs)} = \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}^{(p-2)/p} .$$

If $s = \infty$ (which corresponds to the case $p = q = 2$), the above estimate still holds, since $\alpha = 0$ in that case. We thus get from (3.125) that

$$\|\tau_z \nabla(g_h^z - g^z)\|_{L^q(\Omega)} \leqslant M_{h,\lambda} \, \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}^{(p-2)/(2p)} .$$

We likewise have

$$\|\tau_z (g_h^z - g^z)\|_{L^q(\Omega)} \leqslant M_{h,\lambda} \, \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}^{(p-2)/(2p)} .$$

We then deduce from (3.122) that

$$\begin{aligned} &|a_\eta^\star(u, g_h^z - g^z)| \\ &\qquad \leqslant 2\,(1 + \|b\|_{L^\infty})\, M_{h,\lambda} \, \left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)}^{(p-2)/(2p)} \, \left(\|\tau_z^{-1}\nabla u\|_{L^p(\Omega)} + \|\tau_z^{-1} u\|_{L^p(\Omega)}\right) . \end{aligned} \tag{3.126}$$

Using (3.114) with $\theta = d + \lambda$, and noting that $\chi_z(x) = \chi_x(z)$, we get

$$\left\|\chi_z^{-(d+\lambda)}\right\|_{L^1(\Omega)} \leqslant C_d \frac{1}{(\kappa h)^\lambda} \left(\frac{1}{d} + \frac{1}{\lambda}\right) .$$

Inserting this estimate in (3.126), using Lemma 3.28 and the fact that $\kappa \geqslant 1$, we obtain

$$|a_\eta^\star(u, g_h^z - g^z)| \leqslant C_{\kappa,\lambda,\eta} h^{\lambda/2} \, \left(\frac{1}{h^\lambda}\right)^{(p-2)/(2p)} \, \left(\|\tau_z^{-1}\nabla u\|_{L^p(\Omega)} + \|\tau_z^{-1} u\|_{L^p(\Omega)}\right)$$

where $C_{\kappa,\lambda,\eta}$ is independent of h, but depends on κ, λ and η. We denote it C_η in the sequel. We integrate the p-th power of the above relation with respect to z:

$$\begin{aligned} &\int_\Omega |a_\eta^\star(u, g_h^z - g^z)|^p \, dz \\ \leqslant \;& C_\eta \, h^\lambda \left(\int_\Omega \|\tau_z^{-1}\nabla u\|_{L^p(\Omega)}^p \, dz + \int_\Omega \|\tau_z^{-1} u\|_{L^p(\Omega)}^p \, dz\right) \\ = \;& C_\eta \, h^\lambda \left(\int_\Omega |\nabla u(x)|^p \left[\int_\Omega \tau_z^{-p}(x) dz\right] dx + \int_\Omega |u(x)|^p \left[\int_\Omega \tau_z^{-p}(x) dz\right] dx\right) \end{aligned} \tag{3.127}$$

Let us bound $\displaystyle\int_\Omega \tau_z^{-p}(x) dz$. We write, using again (3.114) with $\theta = d + \lambda$, that

$$\forall x \in \Omega, \quad \int_\Omega \tau_z^{-p}(x) dz = \int_\Omega \chi_z^{-(d+\lambda)} \leqslant \frac{C}{h^\lambda} .$$

We hence infer from (3.127) that

$$\int_\Omega |a_\eta^\star(u, g_h^z - g^z)|^p \, dz \leqslant C_\eta \left(\int_\Omega |\nabla u|^p + \int_\Omega |u|^p \right) \leqslant C_\eta \, \|u\|^p_{W^{1,p}(\Omega)}. \tag{3.128}$$

Bound on the first term of the right hand side of (3.121)

We denote

$$F(z) = \int_\Omega \delta^z(x) \; \nu \cdot \nabla u(x) \, dx$$

the first term of the right-hand side of (3.121). Recalling that δ^z is supported in K^z, and using the Hölder inequality, we have, for any $z \in \Omega$,

$$|F(z)| \leqslant \|\nabla u\|_{L^p(K^z)} \|\delta^z\|_{L^q_x(K^z)}.$$

We compute that

$$\|\delta^z\|^q_{L^q_x(K^z)} \leqslant \|\delta^z\|^q_{L^\infty_x(K^z)} \int_{K^z} dx \leqslant C h^{-qd} \, h^d,$$

thus $\|\delta^z\|_{L^q_x(K^z)} \leqslant C h^{-d/p}$, hence

$$|F(z)| \leqslant C \|\nabla u\|_{L^p(K^z)} \, h^{-d/p}.$$

We integrate the p-th power of the above estimate with respect to z:

$$\begin{aligned}
\int_\Omega |F(z)|^p \, dz &\leqslant C h^{-d} \int_\Omega dz \int_{K^z} |\nabla u(x)|^p \, dx \\
&= C h^{-d} \sum_{T \in \mathcal{T}_h} \int_T dz \int_{K^z} |\nabla u(x)|^p \, dx \\
&= C h^{-d} \sum_{T \in \mathcal{T}_h} \int_T dz \int_T |\nabla u(x)|^p \, dx \\
&= C h^{-d} \sum_{T \in \mathcal{T}_h} h^d \int_T |\nabla u(x)|^p \, dx \\
&= C \int_\Omega |\nabla u(x)|^p \, dx,
\end{aligned} \tag{3.129}$$

where we have used that $K^z = T$ when $z \in T$.

Proof of (3.62)

The right-hand side of (3.121) is the sum of two functions of z, the L^p norm of which is bounded from above (up to a multiplicative constant independent of h) by $\|u\|_{W^{1,p}(\Omega)}$, in view of (3.128) and (3.129). We thus get

$$\|\nabla u_h\|_{L^p(\Omega)} \leqslant C_\eta \, \|u\|_{W^{1,p}(\Omega)}$$

for any $h < h_0(\eta)$ (this restriction comes from the fact that, in Lemma 3.28, we work in the regime $\kappa h \leqslant 1$ for some κ that depends on η), where C_η and $h_0(\eta)$ a priori depend on η. This concludes the proof of (3.62).

3.6.6 Proof of (3.63)

To obtain a bound on $u - u_h$, we introduce the interpolant $I_h u \in \Sigma_h$. We then note that $u - I_h u \in H^1(\Omega)$, $u_h - I_h u \in \Sigma_h$ and that, in view of (3.61), we have, for any $v \in \Sigma_h$,

$$a_\eta^\star\Big((u - I_h u) - (u_h - I_h u), v\Big) = 0.$$

Furthermore, we observe that $u - I_h u \in W^{1,p}(\Omega)$. We are thus in position to write (3.62), that is

$$\|\nabla(u_h - I_h u)\|_{L^p(\Omega)} \leqslant C_\eta \, \|\nabla(u - I_h u)\|_{L^p(\Omega)}.$$

Thus, we get that

$$\|\nabla(u - u_h)\|_{L^p(\Omega)} \leqslant \|\nabla(u - I_h u)\|_{L^p(\Omega)} + \|\nabla(I_h u - u_h)\|_{L^p(\Omega)} \leqslant C_\eta \, \|\nabla(u - I_h u)\|_{L^p(\Omega)}$$

and we conclude using an approximation result (see e.g. [79, eq. (1.5)]), stating that, for any $u \in W^{2,p}(\Omega)$, we have $\|u - I_h u\|_{W^{1,p}(\Omega)} \leqslant Ch\|u\|_{W^{2,p}(\Omega)}$. This yields (3.63).

3.7 Appendix: Proof of Lemma 3.28

The proof of Lemma 3.28 relies on four technical results, that we state below and prove in Appendix 3.8. We next turn here to the proof of Lemma 3.28.

Proposition 3.29. *Assume that $b \in (L^\infty(\Omega))^d$ and that κ and h are such that $\kappa h \leqslant 1$. Let $w \in H^1(\Omega)$ and $w_h \in \Sigma_h$ such that*

$$\forall v \in \Sigma_h, \quad a_\eta^\star(v, w - w_h) = 0. \tag{3.130}$$

We set $e = w - w_h$. Then there exists $C > 0$, independent of η, κ and h, such that

$$\int_\Omega \chi_z^{d+\lambda} \, |\nabla e|^2 \leqslant C \int_\Omega \chi_z^{d+\lambda-2} \, |e|^2 + C \int_\Omega \chi_z^{d+\lambda} \, |\nabla(w - I_h w)|^2 + \chi_z^{d+\lambda-2} \, |w - I_h w|^2.$$

Lemma 3.30. *Assume that (3.94) holds for any q such that $2d/(d+2) < q < \infty$. Assume also that $2 \leqslant d \leqslant 3$ and that $0 < \lambda < 4 - d$. Let $\zeta > 0$ and $x_0 \in \Omega$, and set*

$$\chi(x) = \sqrt{|x - x_0|^2 + \zeta^2}. \tag{3.131}$$

Let $f \in H^1(\Omega)$. The solution $v \in H^1(\Omega)$ to the problem

$$\forall \phi \in H^1(\Omega), \qquad a_\eta^\star(v, \phi) = \int_\Omega f \, \phi \tag{3.132}$$

satisfies

$$\int_\Omega \chi^{-d-\lambda} \, \left|\nabla^2 v\right|^2 \leqslant C_\eta^2 \, \zeta^{-2} \int_\Omega \chi^{4-d-\lambda} \, \left(|f|^2 + |\nabla f|^2\right),$$

where C_η is independent of x_0 and ζ (but a priori depends on η), and where σ_1 is the invariant measure defined by (3.16)–(3.17). In addition, $C_\eta \leqslant C/\eta$ for some C independent of η.

Remark 3.31. *Using the same arguments as for the proof of Lemma 3.30, it is also possible to show that*

$$\int_\Omega \chi^{-d-\lambda} \left|\nabla^2 \left(v - \sigma_1 \fint_\Omega v\right)\right|^2 \leqslant C\zeta^{-2} \int_\Omega \chi^{4-d-\lambda} \left(|f|^2 + |\nabla f|^2\right),$$

where C is independent of η, x_0 and ζ.

Proposition 3.32. *Assume that $b \in (L^\infty(\Omega))^d$ and that (3.94) holds for any q such that $2d/(d+2) < q < \infty$. Assume also that $2 \leqslant d \leqslant 3$ and that $0 < \lambda < 4 - d$.*

Let $w \in H^1(\Omega)$ and $w_h \in \Sigma_h$ such that the Galerkin orthogonality (3.130) holds. We set $e = w - w_h$. Then, for any $\varepsilon > 0$ small enough, there exists $\kappa_1(\varepsilon, \eta) \geqslant 1$ (which a priori depends on ε and η) such that, for any $\kappa \geqslant \kappa_1(\varepsilon, \eta)$ and any h such that $\kappa h \leqslant 1$, we have

$$\int_\Omega \chi_z^{d+\lambda-2} |e|^2 \leqslant 8\varepsilon \int_\Omega \chi_z^{d+\lambda} |\nabla e|^2,$$

where χ_z is defined by (3.112).

The proof of Proposition 3.32 shows that it is sufficient to take $\kappa_1(\varepsilon, \eta) \geqslant \dfrac{C}{\varepsilon\eta}$ for some C independent of ε and η.

Lemma 3.33. *Assume that $b \in (L^\infty(\Omega))^d$, that (3.94) holds for any q such that $2d/(d+2) < q < \infty$, and that (3.100) holds. Assume also that $2 \leqslant d \leqslant 3$ and that $0 < \lambda < 4 - d$. Consider χ defined by (3.131) and assume that $\zeta \leqslant 1$.*

Let $f \in H_0^1(\Omega)$ and consider the solution $v \in H^1(\Omega)$ of the adjoint problem

$$\forall \phi \in H^1(\Omega), \qquad a_\eta^\star(\phi, v) = \int_\Omega (\nu \cdot \nabla f)\ \phi. \tag{3.133}$$

Then v satisfies

$$\int_\Omega \chi^{d+\lambda} |\nabla^2 v|^2 \leqslant C \int_\Omega \chi^{d+\lambda} |\nabla f|^2 + C_\eta^2 \zeta^{-2} \int_\Omega \chi^{d+\lambda} f^2,$$

where C is independent of η, x_0 and ζ, and C_η is independent of x_0 and ζ (but a priori depends on η). In addition, $C_\eta \leqslant C/\eta$ for some C independent of η.

Thanks to the above results, we are in position to prove Lemma 3.28.

Proof of Lemma 3.28. From the assumptions of Lemma 3.28, we know that (3.9)–(3.42)–(3.43) hold. We can thus take λ such that $0 < \lambda < 1 \leqslant 4 - d$, and all the assumptions of Proposition 3.29, Lemma 3.30, Proposition 3.32 and Lemma 3.33 are fulfilled.

Let g^z and g_h^z be the solutions to (3.118) and (3.119). They satisfy the Galerkin orthogonality (3.130), namely $a_\eta^\star(v, g^z - g_h^z) = 0$ for any $v \in \Sigma_h$. Let $\varepsilon > 0$ be small enough, as in Proposition 3.32. Combining Propositions 3.29 and 3.32, we obtain that there exists $\kappa_1(\varepsilon, \eta) \geqslant 1$

(a priori depending on ε and η) such that, for any $\kappa \geqslant \kappa_1(\varepsilon,\eta)$ and any h such that $\kappa h \leqslant 1$,

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - g_h^z)|^2 + \chi_z^{d+\lambda-2}(g^z - g_h^z)^2 \\
&\leqslant (C+1)\int_\Omega \chi_z^{d+\lambda-2}(g^z - g_h^z)^2 \\
&\qquad + C\left(\int_\Omega \chi_z^{d+\lambda-2}(g^z - I_h g^z)^2 + \int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - I_h g^z)|^2\right) \\
&\leqslant 8(C+1)\varepsilon \int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - g_h^z)|^2 \\
&\qquad + C\left(\int_\Omega \chi_z^{d+\lambda-2}(g^z - I_h g^z)^2 + \int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - I_h g^z)|^2\right)
\end{aligned}$$

where C is independent of ε, κ and η. We pick ε such that $8(C+1)\varepsilon \leqslant 1/2$ and we obtain that

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - g_h^z)|^2 + \chi_z^{d+\lambda-2}(g^z - g_h^z)^2 \\
&\qquad\qquad \leqslant C\left(\int_\Omega \chi_z^{d+\lambda-2}(g^z - I_h g^z)^2 + \int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - I_h g^z)|^2\right).
\end{aligned}$$

We have simply written that the error $g^z - g_h^z$ is bounded by the best approximation error. However, this is not a trivial estimate, as all errors are weighted.

We next proceed as follows, successively using (3.115) and the fact that $h \leqslant \kappa h \leqslant \chi_z$:

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda}|\nabla(g^z - g_h^z)|^2 + \chi_z^{d+\lambda-2}(g^z - g_h^z)^2 \\
&\qquad\qquad \leqslant C\int_\Omega \left(\chi_z^{d+\lambda-2}h^4 + \chi_z^{d+\lambda}h^2\right)|\nabla^2 g^z|^2 \leqslant Ch^2\int_\Omega \chi_z^{d+\lambda}|\nabla^2 g^z|^2. \quad (3.134)
\end{aligned}$$

We have used at the second line the regularity assumption (3.100) with $q = 2$. Applying Lemma 3.33 with $f \equiv \delta^z \in C_0^\infty(\Omega)$ and $\chi \equiv \chi_z$, we obtain

$$\begin{aligned}
\int_\Omega \chi_z^{d+\lambda}\,|\nabla^2 g^z|^2 &\leqslant C\int_\Omega \chi_z^{d+\lambda}|\nabla\delta^z|^2 + \frac{C_\eta^2}{\kappa^2 h^2}\int_\Omega \chi_z^{d+\lambda}(\delta^z)^2 \\
&\leqslant C\int_{K^z} \chi_z^{d+\lambda}|\nabla\delta^z|^2 + \frac{C}{\eta^2\kappa^2 h^2}\int_{K^z} \chi_z^{d+\lambda}(\delta^z)^2 \\
&\leqslant C\,h^{-d-2}\,\|\chi_z\|_{L^\infty(K^z)}^{d+\lambda}\left(1 + \frac{1}{\eta^2\kappa^2}\right),
\end{aligned}$$

where we have used that $\|\delta^z\|_{L^\infty(K^z)} \leqslant Ch^{-d}$ and $\|\nabla\delta^z\|_{L^\infty(K^z)} \leqslant Ch^{-d-1}$. Using that $\|\chi_z\|_{L^\infty(K^z)}^2 \leqslant Ch^2(1+\kappa^2)$, we get that

$$\int_\Omega \chi_z^{d+\lambda}\,|\nabla^2 g^z|^2 \leqslant C_{\kappa,\lambda,\eta}\,h^{\lambda-2}, \quad (3.135)$$

where $C_{\kappa,\lambda,\eta}$ depends on κ, λ and η but not on h (more precisely, since $1 \leqslant \kappa$, one can take $C_{\kappa,\lambda,\eta} = C\,\kappa^{2+d+\lambda}\eta^{-2}$). Introduce

$$\mathcal{M}_{h,\lambda}(z) = \sqrt{\int_\Omega \chi_z^{d+\lambda}\Big(|g^z - g_h^z|^2 + |\nabla(g^z - g_h^z)|^2\Big)},$$

so that $M_{h,\lambda} = \sup_z \mathcal{M}_{h,\lambda}(z)$. Using that χ_z is bounded (this is a consequence of the regime $\kappa h \leqslant 1$), we write, collecting (3.134) and (3.135), that

$$\mathcal{M}_{h,\lambda}^2(z) \leqslant C \int_\Omega \chi_z^{d+\lambda} |\nabla(g^z - g_h^z)|^2 + \chi_z^{d+\lambda-2}(g^z - g_h^z)^2 \leqslant C_{\kappa,\lambda,\eta}\, h^\lambda.$$

Taking the supremum over z yields the claimed bound on $M_{h,\lambda}$ and thus concludes the proof of Lemma 3.28. □

3.8 Appendix: Technical proofs

We collect in this Appendix the proofs of Proposition 3.29, Lemma 3.30, Proposition 3.32 and Lemma 3.33.

3.8.1 Proof of Proposition 3.29

We set $e = w - w_h$, $\widetilde{e} = I_h w - w_h$ and $\psi = \chi_z^{d+\lambda}\,\widetilde{e}$. We have

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda} |\nabla e|^2 + \eta \int_\Omega \chi_z^{d+\lambda} e^2 \\
&= \int_\Omega \nabla(\chi_z^{d+\lambda} e) \cdot \nabla e - \int_\Omega e \nabla(\chi_z^{d+\lambda}) \cdot \nabla e + \eta \int_\Omega \chi_z^{d+\lambda} e^2 \\
&= a_\eta^\star(\chi_z^{d+\lambda} e, e) - \int_\Omega (b \cdot \nabla e)\ \chi_z^{d+\lambda} e - \int_\Omega e \nabla(\chi_z^{d+\lambda}) \cdot \nabla e \\
&= a_\eta^\star\big(\chi_z^{d+\lambda}(w - I_h w + \widetilde{e}), e\big) - \int_\Omega (b \cdot \nabla e)\, \chi_z^{d+\lambda} e - \int_\Omega e \nabla(\chi_z^{d+\lambda}) \cdot \nabla e \\
&= a_\eta^\star\big(\chi_z^{d+\lambda}(w - I_h w), e\big) + a_\eta^\star(\psi, e) - \int_\Omega (b \cdot \nabla e)\, \chi_z^{d+\lambda} e - \int_\Omega e \nabla(\chi_z^{d+\lambda}) \cdot \nabla e.
\end{aligned}$$

Using the Galerkin orthogonality (3.130) and the fact that $\eta \int_\Omega \chi_z^{d+\lambda} e^2 > 0$, and next the estimate (3.113), we get

$$\begin{aligned}
\int_\Omega \chi_z^{d+\lambda} |\nabla e|^2 &\leqslant a_\eta^\star\big(\chi_z^{d+\lambda}(w - I_h w), e\big) + a_\eta^\star(\psi - I_h \psi, e) + \|b\|_{L^\infty} \int_\Omega |\nabla e|\, \chi_z^{d+\lambda}\, |e| \\
&\quad + \int_\Omega |\nabla e|\ |\nabla(\chi_z^{d+\lambda})|\ |e| \\
&\leqslant a_\eta^\star\big(\chi_z^{d+\lambda}(w - I_h w), e\big) + a_\eta^\star(\psi - I_h \psi, e) \\
&\quad + (C + \|b\|_{L^\infty} \|\chi_z\|_{L^\infty}) \int_\Omega \chi_z^{d+\lambda-1}\, |\nabla e|\, |e|.
\end{aligned}$$

Since we work in the regime $\kappa h \leqslant 1$, we have that, for any $z \in \Omega$,

$$\|\chi_z\|_{L^\infty} \leqslant C, \tag{3.136}$$

where C only depends on Ω. We deduce from the above estimate that

$$\begin{aligned}\int_\Omega \chi_z^{d+\lambda}\,|\nabla e|^2 &\leqslant a_\eta^\star\big(\chi_z^{d+\lambda}(w-I_hw),e\big)+a_\eta^\star(\psi-I_h\psi,e)+C\int_\Omega \chi_z^{d+\lambda-1}|\nabla e|\,|e|\\ &\leqslant a_\eta^\star\big(\chi_z^{d+\lambda}(w-I_hw),e\big)+a_\eta^\star(\psi-I_h\psi,e)\\ &\quad+\frac14\int_\Omega \chi_z^{d+\lambda}\,|\nabla e|^2+C\int_\Omega \chi_z^{d+\lambda-2}\,|e|^2. \end{aligned}\tag{3.137}$$

We successively estimate the first two terms of (3.137). For the first term, we write, using estimate (3.113), the Cauchy Schwarz inequality and the Young inequality,

$$\begin{aligned}&\Big|a_\eta^\star\big(\chi_z^{d+\lambda}(w-I_hw),e\big)\Big|\\ &\leqslant C\int_\Omega |\nabla e|\Big(\chi_z^{d+\lambda}\,|\nabla(w-I_hw)|+\chi_z^{d+\lambda-1}\,|w-I_hw|\Big)\\ &\quad+\|b\|_{L^\infty(\Omega)}\int_\Omega \chi_z^{d+\lambda}\,|w-I_hw|\,|\nabla e|+\eta\int_\Omega \chi_z^{d+\lambda}\,|w-I_hw|\,|e|\\ &\leqslant C\int_\Omega |\nabla e|\,\Big(\chi_z^{d+\lambda}\,|\nabla(w-I_hw)|+\chi_z^{d+\lambda-1}\,|w-I_hw|\Big)\\ &\quad+C\int_\Omega \chi_z^{d+\lambda-2}\,|w-I_hw|\,|e|\\ &\leqslant C\left(\int_\Omega \chi_z^{d+\lambda}\,|\nabla e|^2\right)^{1/2}\left(\int_\Omega \chi_z^{d+\lambda}|\nabla(w-I_hw)|^2+\chi_z^{d+\lambda-2}|w-I_hw|^2\right)^{1/2}\\ &\quad+C\left(\int_\Omega \chi_z^{d+\lambda-2}\,|w-I_hw|^2\right)^{1/2}\left(\int_\Omega \chi_z^{d+\lambda-2}\,|e|^2\right)^{1/2}\\ &\leqslant\frac14\int_\Omega \chi_z^{d+\lambda}|\nabla e|^2+C\int_\Omega \chi_z^{d+\lambda}|\nabla(w-I_hw)|^2+\chi_z^{d+\lambda-2}|w-I_hw|^2\\ &\quad+C\int_\Omega \chi_z^{d+\lambda-2}|e|^2. \end{aligned}\tag{3.138}$$

Estimating the second term of (3.137) is done in a similar fashion:

$$\begin{aligned}&|a_\eta^\star(\psi-I_h\psi,e)|\\ &\leqslant\int_\Omega |\nabla e|\,|\nabla(\psi-I_h\psi)|+\|b\|_{L^\infty(\Omega)}\int_\Omega |\psi-I_h\psi|\,|\nabla e|+\eta\int_\Omega |\psi-I_h\psi|\,|e|\\ &\leqslant C\int_\Omega |\nabla e|\Big(|\nabla(\psi-I_h\psi)|+|\psi-I_h\psi|\Big)+\int_\Omega |\psi-I_h\psi|\,|e|\\ &\leqslant C\left(\int_\Omega \chi_z^{d+\lambda}|\nabla e|^2\right)^{1/2}\left(\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big)\right)^{1/2}\\ &\quad+\left(\int_\Omega \chi_z^{d+\lambda}|e|^2\right)^{1/2}\left(\int_\Omega \chi_z^{-d-\lambda}|\psi-I_h\psi|^2\right)^{1/2}\\ &\leqslant\frac14\int_\Omega \chi_z^{d+\lambda}|\nabla e|^2+C\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big)\\ &\quad+C\int_\Omega \chi_z^{d+\lambda}|e|^2. \end{aligned}\tag{3.139}$$

Collecting (3.137), (3.138) and (3.139), we obtain

$$\begin{aligned}\frac{1}{4}\int_\Omega \chi_z^{d+\lambda}\,|\nabla e|^2 \leqslant C\int_\Omega \chi_z^{d+\lambda-2}\,|e|^2 \\ + C\int_\Omega \chi_z^{d+\lambda}|\nabla(w-I_hw)|^2+\chi_z^{d+\lambda-2}|w-I_hw|^2 \\ + C\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big).\end{aligned} \tag{3.140}$$

We now bound the last term of (3.140) by finite element estimation (note that $\psi\in H^1(\Omega)$ and $\psi_{|T}\in H^2(T)$ for any $T\in\mathcal{T}_h$, since χ_z belongs to $C^\infty(\Omega)$ and $\widetilde{e}\in\Sigma_h$; we are thus in position to use (3.115)):

$$\begin{aligned}&\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big)\\ \leqslant\ & Ch^2\sum_{T\in\mathcal{T}_h}\int_T \chi_z^{-d-\lambda}\left|\nabla^2(\chi_z^{d+\lambda}\widetilde{e})\right|^2 \quad \text{[estimate (3.115) and def. of } \psi\text{]}\\ \leqslant\ & Ch^2\int_\Omega \chi_z^{-d-\lambda}\Big(|\chi_z^{d+\lambda-2}\,\widetilde{e}|^2+|\chi_z^{d+\lambda-1}|^2\,|\nabla\widetilde{e}|^2\Big),\end{aligned}$$

where, in the last line, we have used (3.113) and the fact that $\widetilde{e}$ is piecewise affine. We next use the inverse inequality (3.116) and the fact that $\chi_z^{-2}\leqslant h^{-2}$:

$$\begin{aligned}&\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big)\\ &\qquad\leqslant Ch^2\int_\Omega \chi_z^{d+\lambda-4}\,|\widetilde{e}|^2+C\int_\Omega \chi_z^{d+\lambda-2}\,|\widetilde{e}|^2\leqslant C\int_\Omega \chi_z^{d+\lambda-2}\,|\widetilde{e}|^2.\end{aligned}$$

Since $\widetilde{e}=I_hw-w_h=I_hw-w+e$, we get

$$\begin{aligned}&\int_\Omega \chi_z^{-d-\lambda}\Big(|\nabla(\psi-I_h\psi)|^2+|\psi-I_h\psi|^2\Big)\\ &\qquad\leqslant C\int_\Omega \chi_z^{d+\lambda-2}\,|e|^2+C\int_\Omega \chi_z^{d+\lambda-2}\,(I_hw-w)^2.\end{aligned}$$

Inserting this estimate in (3.140), we obtain

$$\begin{aligned}\frac{1}{4}\int_\Omega \chi_z^{d+\lambda}\,|\nabla e|^2 \leqslant C\int_\Omega \chi_z^{d+\lambda-2}\,|e|^2 \\ + C\int_\Omega \chi_z^{d+\lambda}|\nabla(w-I_hw)|^2+\chi_z^{d+\lambda-2}|w-I_hw|^2.\end{aligned}$$

This concludes the proof of Proposition 3.29.

3.8.2 Proof of Lemma 3.30

We choose some s such that

$$1\leqslant\frac{2d}{d+2}<s<2 \tag{3.141}$$

and let $s^\star=sd/(d-s)$ (note that $s<2\leqslant d$). From the Sobolev injections, we know that there exists C_s such that

$$\forall g\in W^{1,s}(\Omega),\quad \|g\|_{L^{s^\star}(\Omega)}\leqslant C_s\|g\|_{W^{1,s}(\Omega)}. \tag{3.142}$$

The function f in the statement of Lemma 3.30 belongs to $H^1(\Omega)$. Since $s<2$, we see that $f\in W^{1,s}(\Omega)\subset L^{s^\star}(\Omega)$. Since $s>\dfrac{2d}{d+2}$, we have $s^\star>2>\dfrac{2d}{d+2}$. We can thus use the regularity assumption (3.94) for $s^\star$, from which we deduce that $v\in W^{2,s^\star}(\Omega)$. We set $q=s^\star/2>1$ and write a Hölder inequality with exponents q and q':

$$\begin{aligned}\int_\Omega \chi^{-d-\lambda}\left|\nabla^2 v\right|^2 &\leqslant \left(\int_\Omega \chi^{-(d+\lambda)q'}\right)^{1/q'} \left\|\nabla^2 v\right\|^2_{L^{2q}(\Omega)}\\ &= \left(\int_\Omega \chi^{-(d+\lambda)q'}\right)^{1/q'} \left\|\nabla^2 v\right\|^2_{L^{s^\star}(\Omega)}\\ &\leqslant C\left(\frac{1}{\zeta^{q'(d+\lambda)-d}}\right)^{1/q'} \|v\|^2_{W^{2,s^\star}(\Omega)} \quad \text{[eq. (3.114)]}\end{aligned} \tag{3.143}$$

In view of (3.94), we have

$$\|v\|_{W^{2,s^\star}(\Omega)}\leqslant \left\|v-\sigma_1\fint_\Omega v\right\|_{W^{2,s^\star}(\Omega)}+C\left|\fint_\Omega v\right|\leqslant C\|f\|_{L^{s^\star}(\Omega)}+C\left|\fint_\Omega v\right|.$$

Taking $\phi\equiv 1$ in (3.132), we see that $\eta\fint_\Omega v=\fint_\Omega f$. Since $s^\star\geqslant 1$, we hence get

$$\|v\|_{W^{2,s^\star}(\Omega)}\leqslant C_\eta\|f\|_{L^{s^\star}(\Omega)}\qquad (\text{with } C_\eta\leqslant C/\eta).$$

Inserting this estimate in (3.143), and next using (3.142) for the function $f\in W^{1,s}(\Omega)$, we deduce that

$$\int_\Omega \chi^{-d-\lambda}\left|\nabla^2 v\right|^2\leqslant \frac{C_\eta^2}{\zeta^{\lambda+d/q}}\|f\|^2_{L^{s^\star}(\Omega)}\leqslant \frac{C_\eta^2}{\zeta^{\lambda+d/q}}\|f\|^2_{W^{1,s}(\Omega)}. \tag{3.144}$$

We now define $\overline{q}=2/s$. Since $s<2$, we have $\overline{q}>1$ and we can use the Hölder inequality with exponents $\overline{q}$ and $\overline{q}'$ to bound from above $\|\nabla f\|^s_{L^s(\Omega)}$ (and likewise for $\|f\|^s_{L^s(\Omega)}$):

$$\begin{aligned}\|\nabla f\|^s_{L^s(\Omega)} &= \int_\Omega \chi^{-(4-d-\lambda)s/2}\ \chi^{(4-d-\lambda)s/2}\left|\nabla f\right|^s\\ &\leqslant \left(\int_\Omega \chi^{-(4-d-\lambda)s\overline{q}'/2}\right)^{1/\overline{q}'}\left(\int_\Omega \chi^{(4-d-\lambda)\overline{q}s/2}\left|\nabla f\right|^{\overline{q}s}\right)^{1/\overline{q}}\\ &= \left(\int_\Omega \chi^{-(4-d-\lambda)s/(2-s)}\right)^{(2-s)/2}\left(\int_\Omega \chi^{4-d-\lambda}\left|\nabla f\right|^2\right)^{1/\overline{q}}.\end{aligned} \tag{3.145}$$

We now observe that $\dfrac{2d}{4-\lambda}<2$ since $\lambda<4-d$. Consequently, we can pick a real number s satisfying (3.141) and $s>\dfrac{2d}{4-\lambda}$. This implies that $(4-d-\lambda)s/(2-s)>d$. In (3.145), we are thus in position to use (3.114) with $\theta=(4-d-\lambda)s/(2-s)$. We thus obtain

$$\|\nabla f\|^s_{L^s(\Omega)}\leqslant C\zeta^{(-4+d+\lambda)s/2+d(2-s)/2}\left(\int_\Omega \chi^{4-d-\lambda}\left|\nabla f\right|^2\right)^{1/\overline{q}}$$

and likewise for $\|f\|^s_{L^s(\Omega)}$. Inserting these estimates in (3.144), we deduce that

$$\int_\Omega \chi^{-d-\lambda}\left|\nabla^2 v\right|^2\leqslant \frac{C_\eta^2}{\zeta^{\lambda+d/q}}\zeta^{(-4+d+\lambda)+d(2-s)/s}\int_\Omega \chi^{4-d-\lambda}\left(|f|^2+|\nabla f|^2\right).$$

We have $1/q = 2/s^\star = 2/s - 2/d$, so that $\lambda + d/q = \lambda + 2d/s - 2$ while $(-4+d+\lambda) + d(2-s)/s = -4 + \lambda + 2d/s$. We then obtain

$$\int_\Omega \chi^{-d-\lambda} \left|\nabla^2 v\right|^2 \leqslant C_\eta^2 \, \zeta^{-2} \int_\Omega \chi^{4-d-\lambda} \left(|f|^2 + |\nabla f|^2\right),$$

which concludes the proof of Lemma 3.30.

3.8.3 Proof of Proposition 3.32

Consider the problem (3.93) for the right-hand side $f = \chi_z^{d+\lambda-2} e$, which is indeed in $L^2(\Omega)$. We denote $v \in H^1(\Omega)$ its solution, and thus have

$$\forall \phi \in H^1(\Omega), \quad a_\eta^\star(v,\phi) = \int_\Omega \chi_z^{d+\lambda-2} \, e \, \phi.$$

Taking e as a test function in the above problem, we get, using the Cauchy Schwarz inequality,

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda-2} |e|^2 \\
&= a_\eta^\star(v,e) \\
&= a_\eta^\star(v - I_h v, e) \quad \text{[Galerkin orthogonality (3.130)]} \\
&\leqslant C \left[\int_\Omega \chi_z^{d+\lambda} \left(e^2 + |\nabla e|^2\right)\right]^{1/2} \left[\int_\Omega \chi_z^{-d-\lambda} \Big(|\nabla(v - I_h v)|^2 + |v - I_h v|^2\Big)\right]^{1/2}.
\end{aligned}$$

Let $\varepsilon > 0$. Using the Young inequality, we deduce that

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda-2} |e|^2 \\
&\qquad \leqslant \varepsilon \int_\Omega \chi_z^{d+\lambda} \left(e^2 + |\nabla e|^2\right) + \frac{C^2}{4\varepsilon} \int_\Omega \chi_z^{-d-\lambda} \Big(|\nabla(v - I_h v)|^2 + |v - I_h v|^2\Big). \quad (3.146)
\end{aligned}$$

We now use the regularity assumption (3.94) for the problem (3.93) with the right hand side f defined above, which states that v belongs to $H^2(\Omega)$ (note indeed that $\dfrac{2d}{d+2} < 2$). We are thus in position to use the finite element estimate (3.115). Inserting (3.115) in (3.146), we get

$$\int_\Omega \chi_z^{d+\lambda-2} \, e^2 \leqslant \varepsilon \int_\Omega \chi_z^{d+\lambda} \left(e^2 + |\nabla e|^2\right) + \frac{Ch^2}{\varepsilon} \int_\Omega \chi_z^{-d-\lambda} |\nabla^2 v|^2. \quad (3.147)$$

We now use Lemma 3.30, with $\chi = \chi_z$, noting that the right-hand side $f = \chi_z^{d+\lambda-2} e$ is in $H^1(\Omega)$. We thus deduce from (3.147), successively using Lemma 3.30 and estimate (3.113), that

$$\begin{aligned}
&\int_\Omega \chi_z^{d+\lambda-2} \, e^2 \\
&\leqslant \varepsilon \int_\Omega \chi_z^{d+\lambda} \left(e^2 + |\nabla e|^2\right) + \frac{Ch^2}{\varepsilon} \frac{C_\eta^2}{\kappa^2 h^2} \int_\Omega \chi_z^{4-d-\lambda} \left(|\chi_z^{d+\lambda-2} \, e|^2 + |\nabla(\chi_z^{d+\lambda-2} \, e)|^2\right) \\
&\leqslant \varepsilon \int_\Omega \chi_z^{d+\lambda} \left(e^2 + |\nabla e|^2\right) + \frac{C_\eta^2}{\varepsilon \kappa^2} \left(\int_\Omega \chi_z^{d+\lambda} |\nabla e|^2 + \int_\Omega \chi_z^{d+\lambda-2} e^2\right),
\end{aligned}$$

where C_η only depends on η. For any fixed ε, we take $\kappa_1(\varepsilon,\eta) \geqslant 1$ such that, when $\kappa \geqslant \kappa_1(\varepsilon,\eta)$, we have $\dfrac{C_\eta^2}{\varepsilon\kappa^2} \leqslant \min(\varepsilon, 1/2)$. We thus deduce that, for any $\kappa \geqslant \kappa_1(\varepsilon,\eta)$,

$$\begin{aligned}\frac{1}{2}\int_\Omega \chi_z^{d+\lambda-2}\, e^2 &\leqslant \varepsilon \int_\Omega \chi_z^{d+\lambda} e^2 + 2\varepsilon \int_\Omega \chi_z^{d+\lambda} |\nabla e|^2 \\ &\leqslant \varepsilon \|\chi_z\|_{L^\infty}^2 \int_\Omega \chi_z^{d+\lambda-2} e^2 + 2\varepsilon \int_\Omega \chi_z^{d+\lambda} |\nabla e|^2.\end{aligned}$$

Since we work in the regime $\kappa h \leqslant 1$, we are in position to use (3.136), and thus $\varepsilon\|\chi_z\|_{L^\infty}^2 \leqslant C\varepsilon$ for a constant C that only depends on Ω. Taking ε small enough (namely such that $C\varepsilon \leqslant 1/4$), we get the claimed bound. This concludes the proof of Proposition 3.32.

3.8.4 Proof of Lemma 3.33

Since (3.100) holds and $\nu \cdot \nabla f \in L^2(\Omega)$, we have that $v \in H^2(\Omega)$. Expanding the expression $\nabla^2\left(\chi^{(d+\lambda)/2}\, v\right)$, we find (using (3.113) for χ rather than χ_z) that

$$\chi^{d+\lambda}\, |\nabla^2 v|^2 \leqslant \left|\nabla^2\left(\chi^{(d+\lambda)/2} v\right)\right|^2 + C\left(\chi^{d+\lambda-2}\, |\nabla v|^2 + \chi^{d+\lambda-4}\, v^2\right). \tag{3.148}$$

We now identify an equation satisfied by $\chi^{(d+\lambda)/2} v$, which will be useful to estimate its second derivatives. For any $\phi \in H^1(\Omega)$, we have

$$\begin{aligned}a_\eta^\star\left(\phi, \chi^{(d+\lambda)/2} v\right) &= \int_\Omega \nabla\phi \cdot \nabla\left(\chi^{(d+\lambda)/2} v\right) + \phi\, b \cdot \nabla\left(\chi^{(d+\lambda)/2}\, v\right) + \eta\, \phi\, \chi^{(d+\lambda)/2} v \\ &= a_\eta^\star\left(\chi^{(d+\lambda)/2}\phi, v\right) + \int_\Omega \nabla\phi \cdot \nabla\left(\chi^{(d+\lambda)/2} v\right) + \phi\, b \cdot \nabla\left(\chi^{(d+\lambda)/2}\, v\right) \\ &\quad - \int_\Omega \nabla\left(\chi^{(d+\lambda)/2}\phi\right) \cdot \nabla v - \int_\Omega \chi^{(d+\lambda)/2}\phi\, b \cdot \nabla v \\ &= a_\eta^\star\left(\chi^{(d+\lambda)/2}\phi, v\right) + \int_\Omega v\nabla\phi \cdot \nabla\left(\chi^{(d+\lambda)/2}\right) \\ &\quad - \int_\Omega \phi\nabla\left(\chi^{(d+\lambda)/2}\right) \cdot \nabla v + \int_\Omega \phi\, v\, b \cdot \nabla\left(\chi^{(d+\lambda)/2}\right) \\ &= \int_\Omega (\nu \cdot \nabla f)\, \chi^{(d+\lambda)/2}\phi + \int_\Omega v\nabla\phi \cdot \nabla\left(\chi^{(d+\lambda)/2}\right) \\ &\quad - \int_\Omega \phi\nabla\left(\chi^{(d+\lambda)/2}\right) \cdot \nabla v + \int_\Omega \phi\, v\, b \cdot \nabla\left(\chi^{(d+\lambda)/2}\right) \\ &= \int_\Omega F\, \phi + \int_{\partial\Omega} G\, \phi,\end{aligned}$$

where

$$F = \chi^{(d+\lambda)/2}\, (\nu \cdot \nabla f) - \operatorname{div}\left[v\nabla\left(\chi^{(d+\lambda)/2}\right)\right] - \nabla\left(\chi^{(d+\lambda)/2}\right) \cdot \nabla v + v\, b \cdot \nabla\left(\chi^{(d+\lambda)/2}\right)$$

and

$$G = v\, n \cdot \nabla\left(\chi^{(d+\lambda)/2}\right).$$

Let $\zeta = \chi^{(d+\lambda)/2} v$. We see that $\zeta \in H^1(\Omega)$ and is such that, for any $\phi \in H^1(\Omega)$,

$$a_\eta^\star(\phi, \zeta) = \int_\Omega F\, \phi + \int_{\partial\Omega} G\, \phi.$$

Due to the presence of G, we cannot directly use the regularity result (3.100). We are instead going to use Lemma 3.5, which states a regularity result for (non-homogeneous) Neumann problems. We write that ζ satisfies

$$-\Delta\zeta + b\cdot\nabla\zeta + \eta\zeta = F \quad \text{in } \Omega, \qquad \nabla\zeta\cdot n = G \quad \text{on } \partial\Omega,$$

that we recast in the form

$$-\Delta\zeta = \widetilde{F} \quad \text{in } \Omega, \qquad \nabla\zeta\cdot n = G \quad \text{on } \partial\Omega, \tag{3.149}$$

with $\widetilde{F} = F - b\cdot\nabla\zeta - \eta\zeta$, that is

$$\widetilde{F} = \chi^{(d+\lambda)/2}\,(\nu\cdot\nabla f) - \operatorname{div}\left[v\nabla\left(\chi^{(d+\lambda)/2}\right)\right] - \nabla\left(\chi^{(d+\lambda)/2}\right)\cdot\nabla v - \chi^{(d+\lambda)/2} b\cdot\nabla v - \eta\chi^{(d+\lambda)/2} v.$$

We wish to use Lemma 3.5 for the problem (3.149). Since $f\in H^1(\Omega)$, $\chi\in C^\infty$, $v\in H^1(\Omega)$ and $b\in (L^\infty(\Omega))^d$, we see that $\widetilde{F}\in L^2(\Omega)$. We have $v\in H^1(\Omega)$ thus $G\in H^{1/2}(\partial\Omega)$. We are thus in position to use Lemma 3.5 with $p=2$ on (3.149) (see also [38, Theorem 3.12 and Remark 3.13]), which implies that

$$\|\nabla^2\zeta\|_{L^2(\Omega)} = \left\|\nabla^2\left(\chi^{(d+\lambda)/2} v\right)\right\|_{L^2(\Omega)} \leqslant C\left(\left\|\widetilde{F}\right\|_{L^2(\Omega)} + \|G\|_{H^{1/2}(\partial\Omega)}\right) \tag{3.150}$$

where C is of course independent of η. We integrate (3.148) and use (3.150):

$$\int_\Omega \chi^{d+\lambda}|\nabla^2 v|^2 \leqslant C\left(\left\|\widetilde{F}\right\|^2_{L^2(\Omega)} + \|G\|^2_{H^{1/2}(\partial\Omega)} + \int_\Omega \chi^{d+\lambda-2}\,|\nabla v|^2 + \int_\Omega \chi^{d+\lambda-4}\,v^2\right).$$

Since we work in the regime $\zeta\leqslant 1$, we have that $\|\chi\|_{L^\infty}\leqslant C$ for some C that only depends on Ω. Using in addition the bounds (3.113), we deduce from the above estimate that there exists C independent of η, x_0 and ζ such that

$$\int_\Omega \chi^{d+\lambda}|\nabla^2 v|^2 \leqslant C\|G\|^2_{H^1(\Omega)} + C\int_\Omega \chi^{d+\lambda}\,(\nu\cdot\nabla f)^2 + \chi^{d+\lambda-2}\,|\nabla v|^2 + \chi^{d+\lambda-4}\,v^2. \tag{3.151}$$

For the first term above, we see that $|G|\leqslant C\,|v|\,|\chi|^{(d+\lambda)/2-1}$, thus $\|G\|^2_{L^2(\Omega)}\leqslant C\displaystyle\int_\Omega \chi^{d+\lambda-2}\,v^2 \leqslant C\int_\Omega \chi^{d+\lambda-4}\,v^2$. In addition, we have

$$\begin{aligned}|\nabla G| &\leqslant |\nabla v|\left|\nabla\left(\chi^{(d+\lambda)/2}\right)\right| + |v|\,|\nabla n|\left|\nabla\left(\chi^{(d+\lambda)/2}\right)\right| + |v|\left|\nabla^2\left(\chi^{(d+\lambda)/2}\right)\right| \\ &\leqslant C\left(|\nabla v|\;\chi^{(d+\lambda)/2-1} + |v|\;\chi^{(d+\lambda)/2-1} + |v|\;\chi^{(d+\lambda)/2-2}\right)\end{aligned}$$

and thus

$$\|\nabla G\|^2_{L^2(\Omega)} \leqslant C\int_\Omega \chi^{d+\lambda-2}\,|\nabla v|^2 + \chi^{d+\lambda-4}\,v^2.$$

We hence deduce from (3.151) that

$$\int_\Omega \chi^{d+\lambda}|\nabla^2 v|^2 \leqslant C\left(\int_\Omega \chi^{d+\lambda}\,(\nu\cdot\nabla f)^2 + \chi^{d+\lambda-4}\,v^2 + \chi^{d+\lambda-2}\,|\nabla v|^2\right). \tag{3.152}$$

We are now left with bounding the two last terms in (3.152) in terms of f. We start with the last term, and write

$$\begin{aligned}
&\int_\Omega \chi^{d+\lambda-2}\,|\nabla v|^2\\
&= \int_\Omega \nabla\left(\chi^{d+\lambda-2}v\right)\cdot\nabla v - \int_\Omega v\,\nabla\left(\chi^{d+\lambda-2}\right)\cdot\nabla v\\
&= a_\eta^\star\left(\chi^{d+\lambda-2}v, v\right) - \int_\Omega \chi^{d+\lambda-2}v\,b\cdot\nabla v - \eta\int_\Omega \chi^{d+\lambda-2}v^2 - \int_\Omega v\nabla\left(\chi^{d+\lambda-2}\right)\cdot\nabla v.
\end{aligned}$$

Using (3.133), we see that

$$a_\eta^\star\left(\chi^{d+\lambda-2}v, v\right) = \int_\Omega (\nu\cdot\nabla f)\,\chi^{d+\lambda-2}v.$$

We thus get that

$$\begin{aligned}
&\int_\Omega \chi^{d+\lambda-2}|\nabla v|^2\\
&\qquad = \int_\Omega (\nu\cdot\nabla f)\,\chi^{d+\lambda-2}v - \int_\Omega \chi^{d+\lambda-2}v\,b\cdot\nabla v - \eta\int_\Omega \chi^{d+\lambda-2}v^2 - \int_\Omega v\nabla\left(\chi^{d+\lambda-2}\right)\cdot\nabla v.
\end{aligned}$$

We next proceed using the Young inequality and (3.113):

$$\begin{aligned}
&\int_\Omega \chi^{d+\lambda-2}|\nabla v|^2\\
\leqslant\ & \frac12\int_\Omega \chi^{d+\lambda}|\nu\cdot\nabla f|^2 + \frac12\int_\Omega \chi^{d+\lambda-4}v^2 + \frac14\int_\Omega \chi^{d+\lambda-2}|\nabla v|^2 + \int_\Omega \chi^{d+\lambda-2}|b|^2v^2\\
&+\frac14\int_\Omega \chi^{d+\lambda-2}|\nabla v|^2 + C\int_\Omega \chi^{d+\lambda-4}v^2,
\end{aligned}$$

which implies that

$$\int_\Omega \chi^{d+\lambda-2}|\nabla v|^2 \leqslant C\int_\Omega \chi^{d+\lambda}|\nabla f|^2 + C\int_\Omega \chi^{d+\lambda-4}v^2. \tag{3.153}$$

We now turn to the second term of (3.152). We pick some $P' > \max\left(\frac{d}{2}, \frac{d}{4-d-\lambda}\right)$ (note that $4-d-\lambda>0$) and write the Hölder's inequality with P' and its conjugate exponent P (note that $P' > d/2 \geqslant 1$):

$$\begin{aligned}
\int_\Omega \chi^{d+\lambda-4}v^2 &\leqslant \left(\int_\Omega \chi^{(d+\lambda-4)P'}\right)^{1/P'}\left(\int_\Omega v^{2P}\right)^{1/P}\\
&\leqslant C\zeta^{(d+\lambda-4)+d/P'}\left(\int_\Omega v^{2P}\right)^{1/P},
\end{aligned} \tag{3.154}$$

where we have used (3.114) with $\theta = P'(4-d-\lambda)$, which is indeed larger than d. Note that the last factor of (3.154) is finite, as we have $v\in H^2(\Omega)\subset L^\infty(\Omega)$ (recall that $d\leqslant 3$).

We next use a duality argument to bound $\|v\|_{L^{2P}(\Omega)}$ in terms of f. Let $w \in H^1(\Omega)$ solve (3.93) with a right-hand side equal to $\mathrm{sign}(v)\,|v|^{2P-1}$. Taking v as test function, we get

$$\begin{aligned}
\|v\|_{L^{2P}}^{2P} &= \int_\Omega \left(\mathrm{sign}(v)\,|v|^{2P-1}\right) v \\
&= a_\eta^\star(w,v) \quad [\text{def. of } w] \\
&= \int_\Omega (\nu\cdot\nabla f)\, w \quad [\text{def. (3.133) of } v] \\
&= -\int_\Omega f\,(\nu\cdot\nabla w) \quad [\text{int. by parts and } f\in H_0^1(\Omega)] \\
&\leqslant \|f\|_{L^r(\Omega)}\|\nabla w\|_{L^{r'}(\Omega)}
\end{aligned}$$

with $r = \dfrac{2Pd}{2P+d}$. Taking $v \equiv 1$ in (3.93), we infer that $\eta \fint_\Omega w = \fint_\Omega \mathrm{sign}(v)|v|^{2P-1}$. Note that $r > 1$ since $P > 1 \geqslant \dfrac{d}{2(d-1)}$. We have that $r' = \left(1 - \dfrac{1}{2P} - \dfrac{1}{d}\right)^{-1}$, and we note that $W^{1,2P/(2P-1)}(\Omega) \subset L^{r'}(\Omega)$. Using that Sobolev injection, we get

$$\begin{aligned}
\|v\|_{L^{2P}}^{2P} &\leqslant C\|f\|_{L^r(\Omega)}\|\nabla w\|_{W^{1,2P/(2P-1)}(\Omega)} \\
&\leqslant C\|f\|_{L^r(\Omega)}\left\|w - \sigma_1 \fint_\Omega w\right\|_{W^{2,2P/(2P-1)}(\Omega)} + C\left|\fint_\Omega w\right| \\
&\leqslant C_\eta\|f\|_{L^r(\Omega)}\left\|\mathrm{sign}(v)|v|^{2P-1}\right\|_{L^{2P/(2P-1)}(\Omega)} \qquad (\text{with } C_\eta \leqslant C/\eta). \qquad (3.155)
\end{aligned}$$

We have used in the last line the regularity (3.94), which indeed holds since $2P/(2P-1) > 2d/(d+2)$ (this condition is equivalent to the condition $P' > d/2$, which we have enforced when choosing P'). We next deduce from (3.155) that

$$\|v\|_{L^{2P}}^{2P} \leqslant C_\eta\|f\|_{L^r(\Omega)}\|v\|_{L^{2P}(\Omega)}^{2P-1}.$$

Since $v \in L^{2P}(\Omega)$, we get $\|v\|_{L^{2P}(\Omega)} \leqslant C_\eta\|f\|_{L^r(\Omega)}$. Inserting this in (3.154), we obtain

$$\begin{aligned}
\int_\Omega \chi^{d+\lambda-4}v^2 &\leqslant C_\eta^2\zeta^{2d(1-1/r)+\lambda-2}\,\|f\|_{L^r(\Omega)}^2 \quad [\text{Writing } P' \text{ in terms of } r] \\
&\leqslant C_\eta^2\zeta^{2d(1-1/r)+\lambda-2}\left(\int_\Omega \chi^{d+\lambda}f^2\right)\left(\int_\Omega \chi^{-(d+\lambda)r/(2-r)}\right)^{(2-r)/r}
\end{aligned}$$

where we have eventually used a Hölder inequality with $\overline{q} = 2/r$. Note that $\overline{q} > 1$ (that is, $r < 2$) as a consequence of the fact that $P' > d/2$. We next see that $r > 1 > \dfrac{2d}{2d+\lambda}$, which implies that $(d+\lambda)r/(2-r) > d$, so we are in position to use (3.114), which yields

$$\begin{aligned}
\int_\Omega \chi^{d+\lambda-4}v^2 &\leqslant C_\eta^2\zeta^{2d(1-1/r)+\lambda-2}\left(\int_\Omega \chi^{d+\lambda}f^2\right)\zeta^{-(d+\lambda)+d(2-r)/r} \\
&\leqslant C_\eta^2\zeta^{-2}\int_\Omega \chi^{d+\lambda}f^2. \qquad (3.156)
\end{aligned}$$

Collecting (3.152), (3.153) and (3.156) yields the desired estimate and concludes the proof of Lemma 3.33.

Chapter 4

Multiscale Finite Element methods à la Crouzeix-Raviart for advection-dominated problems in perforated domains

4.1 Introduction

We consider a regular bounded open set $\Omega \subset \mathbb{R}^d$, in dimension $d \geqslant 2$, and its subset $\Omega^\varepsilon \subsetneq \Omega$, a domain perforated by holes of presumably small size $\varepsilon > 0$. We denote by $B^\varepsilon = \Omega \backslash \overline{\Omega^\varepsilon}$ the set of perforations (see Figure 4.1 below). Although this is by no means a limitation of our computational approaches, B^ε will often be, in the sequel, a periodic array of perforations, each of them of diameter of order ε and separated by a distance also of order ε. On the perforated domain Ω^ε, we consider the advection-diffusion equation

$$-\alpha \Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon = f \quad \text{in } \Omega^\varepsilon, \tag{4.1}$$

where $\alpha > 0$, for a right-hand side $f \in L^2(\Omega)$ and for an advection field $\widehat{b}^\varepsilon \in (W^{1,\infty}(\Omega^\varepsilon))^d$ on which we make a variety of assumptions. On the outer boundary $\partial\Omega$, we impose homogeneous Dirichlet boundary conditions. On the other hand, the equation is supplied either with homogeneous Dirichlet or homogeneous Neumann boundary conditions on the boundary of the perforations. More precisely, we concurrently consider the two problems

$$\left\{ \begin{array}{rcll} -\alpha \Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon & = & f & \text{in } \Omega^\varepsilon, \\ u^\varepsilon & = & 0 & \text{on } \partial\Omega^\varepsilon, \end{array} \right. \tag{4.2}$$

and

$$\left\{ \begin{array}{rcll} -\alpha \Delta u^\varepsilon + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon & = & f & \text{in } \Omega^\varepsilon, \\ \alpha \nabla u^\varepsilon \cdot n & = & 0 & \text{on } \partial\Omega^\varepsilon \backslash \partial\Omega, \\ u^\varepsilon & = & 0 & \text{on } \partial\Omega^\varepsilon \cap \partial\Omega. \end{array} \right. \tag{4.3}$$

The well-posedness of these problems, under various assumptions on $\widehat{b}^\varepsilon$, will be established in Section 4.3.1 below.

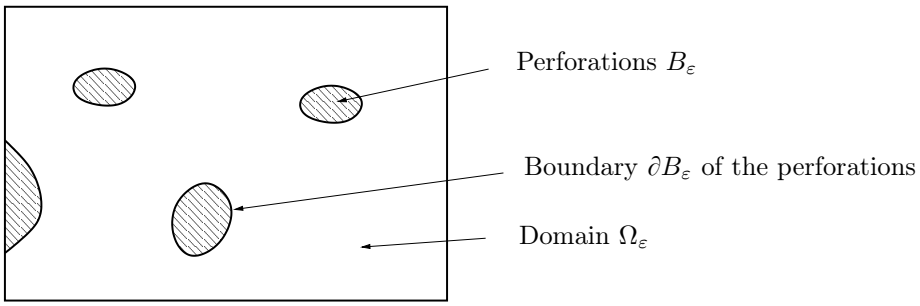

Figure 4.1 The domain Ω contains perforations B_ε, some of which may intersect $\partial\Omega$. The perforated domain is $\Omega_\varepsilon = \Omega\backslash\overline{B_\varepsilon}$. The boundary of Ω_ε is the union of $\partial B_\varepsilon \cap \overline{\Omega_\varepsilon}$ (the part of the boundary of the perforations that is included in $\overline{\Omega_\varepsilon}$) and of $\partial\Omega \cap \overline{\Omega_\varepsilon}$.

We study a regime where advection dominates diffusion. In the absence of perforations, it is well-known that numerical instabilities arise for classical Finite Element methods [86], and stabilization methods are in order. The case of perforated domains deserves more attention, all the more if the perforations are presumable many, and asymptotically infinitely many. It will be seen, and this is not unexpected, that the choice of the boundary conditions on the perforations drastically affects the nature of the flow, and therefore the conclusions regarding which numerical approach is best to adopt. In short, homogeneous Dirichlet boundary conditions on the perforations damp the effect of advection, making the flow more stable than it would be in the absence of perforations, while this is not the case for homogeneous Neumann boundary conditions. This intuitive fact, which we will investigate thoroughly at the numerical level, is particularly well exemplified, at the theoretical level, by the comparison of the respective homogenization limits of the problems (4.2) and (4.3) when ε vanishes. If $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$ is an oscillatory advection field, with b independent of ε, and if b is (say) divergence-free, then the solution u^ε to (4.2) converges to zero as ε^2. Once renormalized by a factor ε^{-2}, u^ε converges, in the limit $\varepsilon \to 0$, to the nontrivial solution of a problem where the advection field b has disappeared. We recall below the classical statement, Theorem 4.3, that formalizes this result. To reinstate advection in the (rescaled) homogenization limit, which in the numerical practice formally means keeping advection dominant in (4.2) for ε small, we therefore consider $\widehat{b}^\varepsilon = \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)$ (see the theoretical result stated in Theorem 4.2). In sharp contrast, for the Neumann problem (4.3) and an advection field of the type $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$ (enjoying specific additional properties), not only the solution u^ε stays of order 1, instead of ε^2, but it converges, in a suitable sense, to the solution of an homogenized equation that does contain advection, as shown by the classical result we recall below in Theorem 4.4.

The numerical approaches we consider are variants of the Multiscale Finite Element method (MsFEM). We recall that MsFEM (see e.g. [36]) encodes the multiscale character of the problem to be solved in the finite element basis functions by defining them as the solution of independent local problems involving a differential operator identical, or close to that of the original equation. The finite element basis functions are defined independently of the source term and therefore are precomputed. A Galerkin approximation on the resulting approximation space is then performed. The choice of the boundary conditions imposed on the local problems is a critical issue. In [58], the first two authors of the present article have introduced Crouzeix-Raviart type boundary conditions for the local problems, in the case of a prototypical diffusion problem $-\mathrm{div}\,(a_\varepsilon \nabla u^\varepsilon) = f$, for an

highly oscillatory coefficient $a_\varepsilon = a(\cdot/\varepsilon)$. The approach was then enriched with bubble functions to address the case of the same diffusion equation posed in a perforated domain, see [59]. The main advantage of this particular choice of boundary conditions has been shown there to be the robustness with respect to the location of the perforations. The approximation remains accurate, irrespective of the fact that the boundaries of the mesh elements intersect or not the perforations, a sensitive issue for other types of boundary conditions. The next step, performed in [60], has been to consider the advection diffusion equation $-\mathrm{div}\,(a_\varepsilon \nabla u^\varepsilon) + \widehat{b}^\varepsilon \cdot \nabla u^\varepsilon = f$ instead of a pure diffusion, and to assume that the advection dominates diffusion to make the case strikingly different from pure diffusion. A question of specific interest is whether or not the advection term must be introduced in the equation defining the local basis functions, and whether or not this brings more stability to the approach. Two of our main conclusions in [60] were (i) that the multiscale character of the problem may be reduced when advection outrageously dominates diffusion and (ii) that, when the multiscale character is still important, a stabilized version of the MsFEM using basis functions defined by the diffusion operator only and Crouzeix-Raviart type boundary conditions on the boundary of mesh elements is one of the most effective and accurate approaches. The present work elaborates on all those previous works for advection-dominated advection diffusion equations in perforated domains.

Our article is articulated as follows. We make precise the various numerical approaches we consider in Section 4.2. Some elements of theoretical analysis for the problems (4.2) and (4.3) (assuming the perforations are periodic) are provided in Section 4.3. In particular, we identify the homogenization limits for these two equations. We also state some error estimates in the case of (4.2). The detailed proofs of all these theoretical results are postponed until the Appendices. Our numerical tests, and our conclusions, are presented in Section 4.4. In short, these conclusions are the following:

- for problem (4.2) (i.e. with homogeneous Dirichlet boundary conditions on the perforations), the flow and the numerical solutions are both stable, even for considerably large advection fields, and the method using a basis of functions built upon the full advection-diffusion operator enriched with bubble functions built likewise is the best possible approach, without requiring any additional stabilization;
- for problem (4.3) (i.e. with homogeneous Neumann boundary conditions on the perforations), instabilities due to the dominating advection arise, and one may use either again the method using a basis of functions built upon the full advection-diffusion operator enriched with bubble functions built likewise, or a stabilized version of the approach with basis functions built with the sole diffusive part of the operator, with no necessary enrichment by bubble functions whatsoever.

These conclusions will be substantiated and commented upon in the sequel.

4.2 Presentation of our numerical approaches

We introduce in this section the different variants of the MsFEM we will consider. All the variants use Crouzeix-Raviart boundary conditions on the boundary of mesh elements for the definition of the basis functions, including bubble functions. As we recalled above, previous works of ours [58, 59] have introduced this variant.

We will consider MsFEM approaches with Crouzeix-Raviart type conditions that

- use basis functions defined with the full advection-diffusion operator (we abbreviate this into *Adv-MsFEM*), or only the diffusive part of that operator (we abbreviate this into *MsFEM*, and sometimes to avoid any ambiguity, *standard MsFEM*);
- possibly enrich the approximation space spanned by these functions by adding bubble functions, the latter being either defined with the full advection-diffusion operator, or only the diffusive part of that operator;
- possibly have stabilized variational formulations, with various options for the stabilization terms.

We have considered in our investigations all combinations of the above options, but we will only report here on the most useful ones.

Our approaches share the following setting.

First of all, at the continuous level, we note that, extending by zero inside the perforations B^ε a function in $H^1_0(\Omega^\varepsilon)$, we can see this function as a function in $H^1_0(\Omega)$. In the case of Problem (4.3), the choice of an extension in $H^1_0(\Omega)$ is more delicate. One example of such an extension procedure can be found in [28, p. 603].

For the discretization, we consider a uniform regular mesh $\mathcal{T}_H$ of Ω, with mesh size H. This mesh size is presumably much larger than what would be in order for a classical FEM applied to a problem with small scale ε. Some actual range of values will be made precise below. We denote by $\mathcal{E}_H^{\rm in}$ and $\mathcal{E}_H^{\rm ext}$, respectively, the set of inner and outer edges/faces of the mesh $\mathcal{T}_H$ ($\mathcal{E}_H^{\rm ext}$ is the set of edges lying in $\partial\Omega$). For the study of Problem (4.2), we define the following infinite-dimensional functional spaces:

$$W_H = \left\{ \begin{array}{c} u \in L^2(\Omega) \text{ such that } u|_K \in H^1(K) \text{ for all } K \in \mathcal{T}_H, \\ \displaystyle\int_E [[u]] = 0 \text{ for all } E \in \mathcal{E}_H^{\rm in}, \quad u = 0 \text{ in } B^\varepsilon \cup \partial\Omega \end{array} \right\}, \tag{4.4}$$

$$W_H^0 = \left\{ u \in W_H \text{ such that } \int_E u = 0 \text{ for all } E \in \mathcal{E}_H^{\rm in} \right\}, \tag{4.5}$$

and

$$W_{H,\rm bubble}^0 = \left\{ \begin{array}{c} u \in W_H \text{ such that } \displaystyle\int_E u = 0 \text{ for all } E \in \mathcal{E}_H^{\rm in} \\ \text{and } \displaystyle\int_K u = 0 \text{ for all } K \in \mathcal{T}_H \end{array} \right\}. \tag{4.6}$$

In the case of Neumann boundary conditions, we introduce the same functional spaces W_H, W_H^0 and $W_{H,\rm bubble}^0$ as above, except that in their definitions, K and E are respectively everywhere replaced by $K \cap \Omega^\varepsilon$ and $E \cap \Omega^\varepsilon$, while the homogeneous Dirichlet boundary condition is set only on the portion $\partial\Omega^\varepsilon \cap \partial\Omega$ of the outer boundary.

The variational formulation of the Dirichlet problem (4.2) reads as follows: find $u^\varepsilon \in H^1_0(\Omega^\varepsilon)$ such that, for any $v \in H^1_0(\Omega^\varepsilon)$,

$$a(u^\varepsilon, v) = F(v)$$

with

$$a(u,v) = \int_{\Omega^\varepsilon} \alpha \nabla u \cdot \nabla v + \left(\widehat{b}^\varepsilon \cdot \nabla u\right) v \quad \text{and} \quad F(v) = \int_{\Omega^\varepsilon} f\, v. \tag{4.7}$$

Since u^ε vanishes on $\partial\Omega^\varepsilon$, we can also consider the bilinear form

$$\begin{aligned} c(u,v) &= \int_{\Omega^\varepsilon} \alpha\nabla u\cdot\nabla v + \frac{1}{2}\left(\widehat{b}^\varepsilon\cdot\nabla u\right)v - \frac{1}{2}\mathrm{div}\left(\widehat{b}^\varepsilon v\right)u \\ &= \int_{\Omega^\varepsilon} \alpha\nabla u\cdot\nabla v + \frac{1}{2}\left(\widehat{b}^\varepsilon\cdot\nabla u\right)v - \frac{1}{2}\left(\widehat{b}^\varepsilon\cdot\nabla v\right)u - \frac{1}{2}u\,v\,\mathrm{div}\ \widehat{b}^\varepsilon, \end{aligned}$$

with a skew-symmetric formulation of the advection-term. For any u and v in $H^1_0(\Omega^\varepsilon)$, we have $a(u,v) = c(u,v)$. The variational formulation of (4.2) can thus be equivalently written: find $u^\varepsilon \in H^1_0(\Omega^\varepsilon)$ such that, for any $v \in H^1_0(\Omega^\varepsilon)$,

$$c(u^\varepsilon, v) = F(v).$$

The finite dimensional approximation spaces (that we introduce below) are not included in $H^1_0(\Omega^\varepsilon)$, since Crouzeix-Raviart boundary conditions allow for discontinuous functions. Our approximations of (4.2) are therefore not conformal approximations. For variational formulations, we therefore introduce the following three bilinear forms:

$$a_H(u,v) = \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \alpha\nabla u\cdot\nabla v + \left(\widehat{b}^\varepsilon\cdot\nabla u\right)v, \tag{4.8}$$

$$a_{\mathrm{diff},H}(u,v) = \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \alpha\nabla u\cdot\nabla v, \tag{4.9}$$

and

$$\begin{aligned} &c_H(u,v) \\ &\quad= \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \alpha\nabla u\cdot\nabla v + \frac{1}{2}\left(\widehat{b}^\varepsilon\cdot\nabla u\right)v - \frac{1}{2}\left(\widehat{b}^\varepsilon\cdot\nabla v\right)u - \frac{1}{2}u\,v\,\mathrm{div}\ \widehat{b}^\varepsilon, \end{aligned} \tag{4.10}$$

which all involve broken integrals.

We easily observe that, on the broken space

$$\left\{v\in L^2(\Omega^\varepsilon),\quad v\in H^1(K\cap\Omega^\varepsilon)\text{ for any }K\in\mathcal{T}_H,\quad v=0\text{ on }\partial\Omega^\varepsilon\right\},$$

and under the (classical) assumption $\mathrm{div}\ \widehat{b}^\varepsilon \leqslant 0$, we have $c_H(v,v) \geqslant \alpha\sum_{K\in\mathcal{T}_H}\|\nabla v\|^2_{L^2(K\cap\Omega^\varepsilon)}$. Under mild additional constraints on the broken space (such as weak continuity of functions across element edges), we will obtain that c_H is coercive. This is not the case for the bilinear form a_H. For this reason, we will favor the bilinear form c_H over a_H when considering the problem (4.2) (in that vein, see Remark 4.1 below).

We now turn to the Neumann problem (4.3), the variational formulation of which reads as follows: find $u^\varepsilon \in V^\varepsilon$ such that, for any $v \in V^\varepsilon$,

$$a(u^\varepsilon, v) = F(v)$$

with a and F defined by (4.7) and

$$V^\varepsilon = \left\{u\in H^1(\Omega^\varepsilon)\text{ such that }u=0\text{ on }\partial\Omega^\varepsilon\cap\partial\Omega\right\}. \tag{4.11}$$

For this problem, there is no reason to consider the bilinear form c, as in general $a(u,v) \neq c(u,v)$ for u and v in V^ε. Only the bilinear forms $a_H(u,v)$ and $a_{\mathrm{diff},H}(u,v)$ will be considered in that case.

The finite dimensional approximation spaces we use in practice (and which are in the sequel generically denoted by V_H with various additional subscripts or superscripts) are spanned by (numerical approximations on a finer mesh of) functions $\Phi^{\varepsilon,E}$ associated to the inner edges/faces $E \in \mathcal{E}_H^{\mathrm{in}}$ and bubble functions $\Psi^{\varepsilon,K}$ associated to each mesh element $K \in \mathcal{T}_H$. For the notation of these functions, we again use additional subscripts that depend on the specific situation considered.

4.2.1 MsFEM approaches using only the diffusion operator, and their stabilized version

We successively consider the Dirichlet problem (4.2) (with the functional spaces W_H, W_H^0 and $W_{H,\mathrm{bubble}}^0$ defined by (4.4), (4.5) and (4.6)) and the Neumann problem (4.3) (with the corresponding functional spaces W_H, W_H^0 and $W_{H,\mathrm{bubble}}^0$).

Dirichlet problem (4.2)

Variational formulations. The variational formulation of the standard MsFEM approach with Crouzeix-Raviart type boundary conditions reads as

$$\text{Find } u_H \in V_H \text{ such that, for any } v_H \in V_H, \quad c_H(u_H, v_H) = F(v_H), \tag{4.12}$$

with c_H defined in (4.10) and the finite dimensional approximation space $V_H \subset W_H$ given by

$$V_H = \left\{ u \in W_H \text{ such that } a_{\mathrm{diff},H}(u,v) = 0 \text{ for any } v \in W_H^0 \right\}.$$

The variational formulation for the variant using bubble functions reads as

$$\left\{ \begin{array}{c} \text{Find } u_H \in V_{H,\mathrm{bubble}} \text{ such that,} \\ \text{for any } v_H \in V_{H,\mathrm{bubble}}, \quad c_H(u_H, v_H) = F(v_H), \end{array} \right. \tag{4.13}$$

where the finite dimensional approximation space $V_{H,\mathrm{bubble}} \subset W_H$ is

$$V_{H,\mathrm{bubble}} = \left\{ u \in W_H \text{ such that } a_{\mathrm{diff},H}(u,v) = 0 \text{ for any } v \in W_{H,\mathrm{bubble}}^0 \right\}. \tag{4.14}$$

The stabilized version of formulation (4.13) (or, *mutatis mutandis*, of (4.12)) reads as

$$\left\{ \begin{array}{l} \text{Find } u_H \in V_{H,\mathrm{bubble}} \text{ such that, for any } v_H \in V_{H,\mathrm{bubble}}, \\ \quad c_H(u_H, v_H) + a_{\mathrm{stab}}(u_H, v_H) = F(v_H) + F_{\mathrm{stab}}(v_H), \end{array} \right. \tag{4.15}$$

where the stabilization terms are defined by

$$\begin{aligned} a_{\mathrm{stab}}(u_H, v_H) &= \sum_{K \in \mathcal{T}_H} \left(\tau_{\mathbf{K}} \left(-\alpha \Delta u_H + \widehat{b}^\varepsilon \cdot \nabla u_H \right), \widehat{b}^\varepsilon \cdot \nabla v_H \right)_{L^2(K \cap \Omega^\varepsilon)}, \\ F_{\mathrm{stab}}(v_H) &= \sum_{K \in \mathcal{T}_H} \left(\tau_{\mathbf{K}} f, \widehat{b}^\varepsilon \cdot \nabla v_H \right)_{L^2(K \cap \Omega^\varepsilon)}, \end{aligned} \tag{4.16}$$

with $\tau_{\mathbf{K}}(x) = \dfrac{H}{2\left|\widehat{b}^{\varepsilon}(x)\right|}\left[\coth\left(\dfrac{\left|\widehat{b}^{\varepsilon}(x)\right| H}{2\alpha}\right) - \dfrac{2\alpha}{\left|\widehat{b}^{\varepsilon}(x)\right| H}\right]$.

Description of the basis functions. We now make precise the definition of the basis functions. For any $E \in \mathcal{E}_H^{\text{in}}$, we introduce a function $\Phi_0^{\varepsilon,E}$ which is such that, for all mesh elements $K \in \mathcal{T}_H$,

$$\begin{cases} -\alpha\Delta\Phi_0^{\varepsilon,E} = 0 \quad \text{in } K \cap \Omega^{\varepsilon}, \\ \Phi_0^{\varepsilon,E} = 0 \quad \text{in } K \cap B^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E'} \Phi_0^{\varepsilon,E} = \delta_{E,E'} \quad \text{and} \quad \alpha\nabla\Phi_0^{\varepsilon,E}\cdot n = \lambda^{K,E'} \text{ on } E' \cap \Omega^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Phi_0^{\varepsilon,E} = 0 \quad \text{on } E' \cap \Omega^{\varepsilon}, \end{cases} \tag{4.17}$$

where $\lambda^{K,E'}$ is constant for all $E' \subset \partial K$. The function $\Phi_0^{\varepsilon,E}$ is supported in the elements K for which $E \subset \partial K$.

We also define, for $K \in \mathcal{T}_H$, the bubble function $\Psi_0^{\varepsilon,K}$, the support of which is reduced to $K \cap \Omega^{\varepsilon}$, as the solution to

$$\begin{cases} -\alpha\Delta\Psi_0^{\varepsilon,K} = 1 \quad \text{in } K \cap \Omega^{\varepsilon}, \\ \Psi_0^{\varepsilon,K} = 0 \quad \text{in } K \cap B^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E'} \Psi_0^{\varepsilon,K} = 0 \quad \text{and} \quad \alpha\nabla\Psi_0^{\varepsilon,K}\cdot n = \mu^{K,E'} \text{ on } E' \cap \Omega^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Psi_0^{\varepsilon,K} = 0 \quad \text{on } E' \cap \Omega^{\varepsilon}, \end{cases} \tag{4.18}$$

where $\mu^{K,E'}$ is constant for all $E' \subset \partial K$.

We then have

$$V_H = \text{Span}\left\{\Phi_0^{\varepsilon,E}, \quad E \in \mathcal{E}_H^{\text{in}}\right\}$$

and

$$V_{H,\text{bubble}} = \text{Span}\left\{\Phi_0^{\varepsilon,E},\ \Psi_0^{\varepsilon,K}, \quad E \in \mathcal{E}_H^{\text{in}},\ K \in \mathcal{T}_H\right\}.$$

Details on the stabilized formulations. Given the above basis functions, we can obtain a simpler expression of the term (4.16) by decomposing $u_H \in V_{H,\text{bubble}}$ as

$$u_H = \sum_{E\in\mathcal{E}_H^{\text{in}}} U_H^E\,\Phi_0^{\varepsilon,E} + \sum_{K\in\mathcal{T}_H} U_H^K\,\Psi_0^{\varepsilon,K}.$$

Following the definition of the basis functions, we have

$$a_{\text{stab}}(u_H, v_H) = \sum_{K\in\mathcal{T}_H} \left(\tau_{\mathbf{K}}\left(\widehat{b}^{\varepsilon}\cdot\nabla u_H\right), \widehat{b}^{\varepsilon}\cdot\nabla v_H\right)_{L^2(K\cap\Omega^{\varepsilon})} + \sum_{K\in\mathcal{T}_H} U_H^K \int_{K\cap\Omega^{\varepsilon}} \tau_{\mathbf{K}}\left(\widehat{b}^{\varepsilon}\cdot\nabla v_H\right). \tag{4.19}$$

In practice, we make use of a discrete approximation of the basis functions on a fine mesh K_h, and (4.16) may not be defined in general. For example, if we use a $\mathbb{P}^1$ approximation on a fine mesh K_h^{ε} for the local problems, then $\nabla u_{H,h}$ may be discontinuous at the interfaces of K_h^{ε}.

As a consequence, we have that

$$-\Delta u_{H,h} \notin L^2(K \cap \Omega^\varepsilon)$$

and the stabilization term (4.16) has no natural expression when we work with the discretized approximation space $(V_{H,\text{bubble}})_h$ rather than $V_{H,\text{bubble}}$. There are (at least) two options to circumvent the difficulty.

The first option is to use

$$\widetilde{a}_{\text{stab}}(u_{H,h}, v_{H,h}) = \sum_{K \in \mathcal{T}_H} \sum_{\kappa \subset K_h^\varepsilon} \left(\tau_K \left(-\alpha \Delta u_{H,h} + \widehat{b}^\varepsilon \cdot \nabla u_{H,h} \right), \widehat{b}^\varepsilon \cdot \nabla v_{H,h} \right)_{L^2(\kappa)} \tag{4.20}$$

rather than (4.16). This yields a strongly consistent stabilized method.

The second option, and this is the variant we adopt, is to use the stabilization term (4.19) rather than (4.16). In contrast to (4.16), the quantity (4.19) is also well defined on $(V_{H,\text{bubble}})_h$. We point out that this stabilization approach is not strongly consistent. We however point out that we have already used (for the same reasons) this type of non-strongly consistent stabilization approach in [60], where we were able to show the convergence of the approach (see [60, Section 3.2]).

Neumann problem (4.3)

Variational formulations. The variational formulations for the Neumann problem read as (4.12), (4.13) and (4.15), with c_H replaced by a_H defined in (4.8).

Basis functions and stabilized formulations. For problem (4.3), (4.17) and (4.18) are respectively replaced by the following two systems (we temporarily use the same notation for the Dirichlet and the Neumann problems):

$$\begin{cases} -\alpha \Delta \Phi_0^{\varepsilon,E} = 0 \quad \text{in } K \cap \Omega^\varepsilon, \\ \alpha \nabla \Phi_0^{\varepsilon,E} \cdot n = 0 \quad \text{in } K \cap \partial B^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E' \cap \Omega^\varepsilon} \Phi_0^{\varepsilon,E} = \delta_{E,E'} \quad \text{and} \quad \alpha \nabla \Phi_0^{\varepsilon,E} \cdot n = \lambda^{K,E'} \text{ on } E' \cap \Omega^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Phi_0^{\varepsilon,E} = 0 \quad \text{on } E' \cap \Omega^\varepsilon, \end{cases} \tag{4.21}$$

where $\lambda^{K,E'}$ is constant for all $E' \subset \partial K$, and

$$\begin{cases} -\alpha \Delta \Psi_0^{\varepsilon,K} = 1 \quad \text{in } K \cap \Omega^\varepsilon, \\ \alpha \nabla \Psi_0^{\varepsilon,K} \cdot n = 0 \quad \text{in } K \cap \partial B^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E' \cap \Omega^\varepsilon} \Psi_0^{\varepsilon,K} = 0 \quad \text{and} \quad \alpha \nabla \Psi_0^{\varepsilon,K} \cdot n = \mu^{K,E'} \text{ on } E' \cap \Omega^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Psi_0^{\varepsilon,K} = 0 \quad \text{on } E' \cap \Omega^\varepsilon, \end{cases} \tag{4.22}$$

where $\mu^{K,E'}$ is constant for all $E' \subset \partial K$.

For (4.3), we again work with the stabilization term (4.19).

4.2.2 MsFEM approaches using the full advection-diffusion operator, and their stabilized version

In this variant, we use basis functions that depend on the advection field.

Dirichlet problem (4.2)

Variational formulations. When no bubble functions are used to enrich the approximation space, the variational formulation, for the study of Problem (4.2), reads as

$$\text{Find } u_H \in V_H^{\text{adv}} \text{ such that, for any } v_H \in V_H^{\text{adv}}, \quad c_H(u_H, v_H) = F(v_H), \tag{4.23}$$

with c_H defined by (4.10) and

$$V_H^{\text{adv}} = \left\{ u \in W_H \text{ such that } c_H(u, v) = 0 \text{ for any } v \in W_H^0 \right\}. \tag{4.24}$$

When using bubble functions, we consider the variational formulation

$$\left\{ \begin{array}{c} \text{Find } u_H \in V_H^{\text{adv bubble}} \text{ such that,} \\ \text{for any } v_H \in V_H^{\text{adv bubble}}, \quad c_H(u_H, v_H) = F(v_H), \end{array} \right. \tag{4.25}$$

where

$$V_H^{\text{adv bubble}} = \left\{ u \in W_H \text{ such that } c_H(u, v) = 0 \text{ for any } v \in W_{H,\text{bubble}}^0 \right\}. \tag{4.26}$$

The stabilized version of the formulation (4.25) reads as

$$\left\{ \begin{array}{c} \text{Find } u_H \in V_H^{\text{adv bubble}} \text{ such that, for any } v_H \in V_H^{\text{adv bubble}}, \\ c_H(u_H, v_H) + a_{\text{stab}}(u_H, v_H) = F(v_H) + F_{\text{stab}}(v_H). \end{array} \right. \tag{4.27}$$

with again $V_H^{\text{adv bubble}}$ defined by (4.26). For the same reasons as those for which we favor (4.19) over (4.20), we choose the stabilization defined by

$$a_{\text{stab}}(u_H, v_H) = \sum_{K \in \mathcal{T}_H} U_H^K \int_{K \cap \Omega^\varepsilon} \tau_{\mathbf{K}} \left(\widehat{b}^\varepsilon \cdot \nabla v_H \right). \tag{4.28}$$

Note that, for the formulation (4.23), the stabilization is void as $a_{\text{stab}}(u_H, v_H) = 0$ for any $u_H \in V_H^{\text{adv}}$.

Description of the basis functions. Similarly as in Section 4.2.1, we now make explicit a basis of functions for our approximation spaces. For any $E \in \mathcal{E}_H^{\text{in}}$, the function $\Phi_{\text{D}}^{\varepsilon,E}$ is defined by

$$\left\{ \begin{array}{l} -\alpha \Delta \Phi_{\text{D}}^{\varepsilon,E} + \widehat{b}^\varepsilon \cdot \nabla \Phi_{\text{D}}^{\varepsilon,E} = 0 \quad \text{in } K \cap \Omega^\varepsilon, \\ \Phi_{\text{D}}^{\varepsilon,E} = 0 \quad \text{in } K \cap B^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E'} \Phi_{\text{D}}^{\varepsilon,E} = \delta_{E,E'} \quad \text{and} \quad \left(\alpha \nabla \Phi_{\text{D}}^{\varepsilon,E} - \frac{1}{2} \widehat{b}^\varepsilon \Phi_{\text{D}}^{\varepsilon,E} \right) \cdot n = \lambda^{K,E'} \text{ on } E' \cap \Omega^\varepsilon, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Phi_{\text{D}}^{\varepsilon,E} = 0 \quad \text{on } E' \cap \Omega^\varepsilon, \end{array} \right. \tag{4.29}$$

while the bubble function $\Psi_{\mathrm{D}}^{\varepsilon,K}$ is the solution to

$$\begin{cases} -\alpha\Delta\Psi_{\mathrm{D}}^{\varepsilon,K} + \widehat{b}^{\varepsilon}\cdot\nabla\Psi_{\mathrm{D}}^{\varepsilon,K} = 1 \quad \text{in } K\cap\Omega^{\varepsilon}, \\ \Psi_{\mathrm{D}}^{\varepsilon,K} = 0 \quad \text{in } K\cap B^{\varepsilon}, \\ \text{if } E'\in\mathcal{E}_H^{\mathrm{in}}, \quad \displaystyle\int_{E'}\Psi_{\mathrm{D}}^{\varepsilon,K} = 0 \quad \text{and} \quad \left(\alpha\nabla\Psi_{\mathrm{D}}^{\varepsilon,K} - \frac{1}{2}\widehat{b}^{\varepsilon}\Psi_{\mathrm{D}}^{\varepsilon,K}\right)\cdot n = \mu^{K,E'} \text{ on } E'\cap\Omega^{\varepsilon}, \\ \text{if } E'\in\mathcal{E}_H^{\mathrm{ext}}, \quad \Psi_{\mathrm{D}}^{\varepsilon,K} = 0 \quad \text{on } E'\cap\Omega^{\varepsilon}, \end{cases} \tag{4.30}$$

where $\lambda^{K,E'}$ and $\mu^{K,E'}$ are constant for all $E'\subset\partial K$. The well-posedness of the two problems (4.29) and (4.30) is the purpose of Section 4.7.1.

In the case of the Dirichlet problem, we then have

$$V_H^{\mathrm{adv}} = \mathrm{Span}\left\{\Phi_{\mathrm{D}}^{\varepsilon,E}, \quad E\in\mathcal{E}_H^{\mathrm{in}}\right\} \tag{4.31}$$

and

$$V_H^{\text{adv bubble}} = \mathrm{Span}\left\{\Phi_{\mathrm{D}}^{\varepsilon,E},\ \Psi_{\mathrm{D}}^{\varepsilon,K}, \quad E\in\mathcal{E}_H^{\mathrm{in}},\ K\in\mathcal{T}_H\right\}. \tag{4.32}$$

The method obtained is similar to the approach of [31]. The difference lies in the definition of the bubble functions since [31] use homogeneous Dirichlet conditions on the boundary of elements whereas we impose here Crouzeix-Raviart conditions. The work [31] shows the added value of bubble functions (which are, we emphasize it, defined using the full advection-diffusion operator, as we do here). It also explores numerically how non periodicity of the location of the perforations affects the quality of the numerical approach, an issue which we ourselves will examine later, in a specific manner, in the present article.

Remark 4.1. *In principle, from a practical viewpoint, it is possible to work with the bilinear form a_H instead of c_H. The definitions (4.24) and (4.26) of the discretization spaces should then be amended, as well as the variational formulations (4.23), (4.25) and (4.27). This variant is briefly examined in Section 4.4.1, where it is shown that it may lead to inaccurate results.*

Neumann problem (4.3)

Variational formulations. The variational formulations for the Neumann problem read as (4.23), (4.25) and (4.27), with c_H replaced by a_H both in the variational formulations and in the definition of the discretization spaces.

When no bubble functions are used to enrich the approximation space, the variational formulation thus reads as

$$\text{Find } u_H\in V_H^{\mathrm{adv}} \text{ such that, for any } v_H\in V_H^{\mathrm{adv}}, \quad a_H(u_H,v_H) = F(v_H), \tag{4.33}$$

where, instead of (4.24), the approximation space V_H^{adv} reads as

$$V_H^{\mathrm{adv}} = \left\{u\in W_H \text{ such that } a_H(u,v) = 0 \text{ for any } v\in W_H^0\right\}.$$

When using bubble functions, we use the variational formulation

$$\begin{cases} \text{Find } u_H\in V_H^{\text{adv bubble}} \text{ such that,} \\ \text{for any } v_H\in V_H^{\text{adv bubble}}, \quad a_H(u_H,v_H) = F(v_H), \end{cases} \tag{4.34}$$

where, instead of (4.26), the approximation space $V_H^{\text{adv bubble}}$ reads as

$$V_H^{\text{adv bubble}} = \left\{ u \in W_H \text{ such that } a_H(u,v) = 0 \text{ for all } v \in W_{H,\text{bubble}}^0 \right\}. \tag{4.35}$$

The stabilized version of the formulation (4.34) reads as

$$\begin{gathered} \text{Find } u_H \in V_H^{\text{adv bubble}} \text{ such that, for any } v_H \in V_H^{\text{adv bubble}}, \\ a_H(u_H, v_H) + a_{\text{stab}}(u_H, v_H) = F(v_H) + F_{\text{stab}}(v_H). \end{gathered} \tag{4.36}$$

with again $V_H^{\text{adv bubble}}$ defined by (4.35) and a_{stab} given by (4.28). As in the Dirichlet case, the stabilization is void for the formulation (4.33).

Description of the basis functions. Similarly as in Section 4.2.1, we now make explicit a basis of functions for our approximation spaces. For any $E \in \mathcal{E}_H^{\text{in}}$, the basis function $\Phi_{\text{N}}^{\varepsilon,E}$ is defined by

$$\begin{cases} -\alpha\Delta\Phi_{\text{N}}^{\varepsilon,E} + \widehat{b}^{\varepsilon} \cdot \nabla\Phi_{\text{N}}^{\varepsilon,E} = 0 \quad \text{in } K \cap \Omega^{\varepsilon}, \\ \left(\alpha\nabla\Phi_{\text{N}}^{\varepsilon,E}\right) \cdot n = 0 \quad \text{in } K \cap \partial B^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E' \cap \Omega^{\varepsilon}} \Phi_{\text{N}}^{\varepsilon,E} = \delta_{E,E'} \quad \text{and} \quad \left(\alpha\nabla\Phi_{\text{N}}^{\varepsilon,E}\right) \cdot n = \lambda^{K,E'} \text{ on } E' \cap \Omega^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Phi_{\text{N}}^{\varepsilon,E} = 0 \quad \text{on } E' \cap \Omega^{\varepsilon}, \end{cases} \tag{4.37}$$

while the bubble function $\Psi_{\text{N}}^{\varepsilon,K}$ is the solution to

$$\begin{cases} -\alpha\Delta\Psi_{\text{N}}^{\varepsilon,K} + \widehat{b}^{\varepsilon} \cdot \nabla\Psi_{\text{N}}^{\varepsilon,K} = 1 \quad \text{in } K \cap \Omega^{\varepsilon}, \\ \left(\alpha\nabla\Psi_{\text{N}}^{\varepsilon,K}\right) \cdot n = 0 \quad \text{in } K \cap \partial B^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{in}}, \quad \displaystyle\int_{E' \cap \Omega^{\varepsilon}} \Psi_{\text{N}}^{\varepsilon,K} = 0 \quad \text{and} \quad \left(\alpha\nabla\Psi_{\text{N}}^{\varepsilon,K}\right) \cdot n = \mu^{K,E'} \quad \text{on } E' \cap \Omega^{\varepsilon}, \\ \text{if } E' \in \mathcal{E}_H^{\text{ext}}, \quad \Psi_{\text{N}}^{\varepsilon,K} = 0 \quad \text{on } E' \cap \Omega^{\varepsilon}, \end{cases} \tag{4.38}$$

where $\lambda^{K,E'}$ and $\mu^{K,E'}$ are constant for all $E' \subset \partial K$. The well-posedness of the two problems (4.37) and (4.38) is the purpose of Section 4.7.2.

In the case of the Neumann problem, we then have, instead of (4.31) and (4.32),

$$V_H^{\text{adv}} = \text{Span}\left\{ \Phi_{\text{N}}^{\varepsilon,E}, \quad E \in \mathcal{E}_H^{\text{in}} \right\}$$

and

$$V_H^{\text{adv bubble}} = \text{Span}\left\{ \Phi_{\text{N}}^{\varepsilon,E}, \ \Psi_{\text{N}}^{\varepsilon,K}, \quad E \in \mathcal{E}_H^{\text{in}}, \ K \in \mathcal{T}_H \right\}.$$

4.2.3 A mixed approach and its stabilized version

In the previous two sections, we have considered basis functions that are all built using the same operator: either the full operator (in Section 4.2.2), or only the diffusion term (in Section 4.2.1), for both $\Phi^{\varepsilon,E}$ and $\Psi^{\varepsilon,K}$. The question arises to build separately functions $\Phi^{\varepsilon,E}$ associated to edges and bubble functions $\Psi^{\varepsilon,K}$ associated to elements, using two different operators for each category. In this direction, we consider here the variant where the functions associated to the

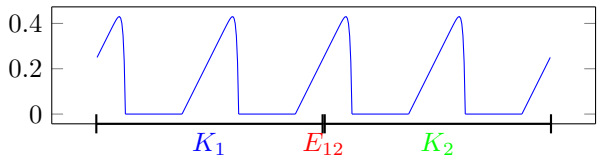

Figure 4.2 Rescaled exact solution $u^\varepsilon/\varepsilon^2$ to Problem (4.39) with $\alpha = 1/64$, $\beta = 1$, $\varepsilon = 1/32$, $f = 1$ and $|K_1| = |K_2| = 2\varepsilon$.

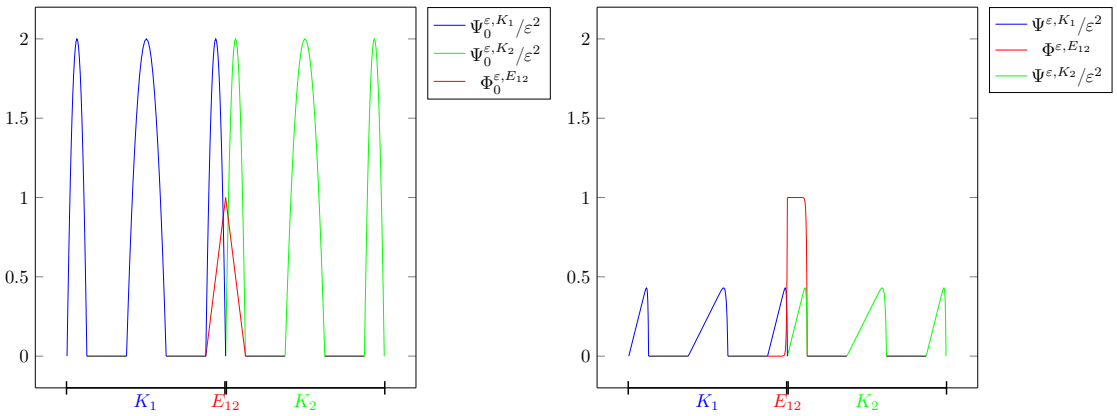

Figure 4.3 Basis functions for (left) MsFEM and (right) Adv-MsFEM as defined in Sections 4.2.1 and 4.2.2, applied to the one-dimensional problem (4.39), with $\alpha = 1/64$, $\beta = 1$, $\varepsilon = 1/32$, and $|K_1| = |K_2| = 2\varepsilon$.

edges are constructed with the diffusion operator, as in the standard MsFEM, while the bubble functions are built using the full advection-diffusion operator.

The consideration of this particular variant is based upon the following, one-dimensional observation (for the sake of conciseness, we only consider here the Dirichlet problem (4.2)).

We temporarily consider the following perforated problem in dimension 1:

$$\begin{cases} -\alpha \dfrac{d^2u^\varepsilon}{dx^2} + \dfrac{\beta}{\varepsilon}\dfrac{du^\varepsilon}{dx} = f \quad \text{in } \Omega^\varepsilon, \\ \qquad\qquad u^\varepsilon = 0 \quad \text{on } \partial\Omega^\varepsilon, \end{cases} \tag{4.39}$$

where Ω^ε is the subset of the segment $\Omega = (0,1)$ defined by periodically perforating that segment with holes of size $\varepsilon/2$. We mesh the segment using a uniform mesh consisting of mesh elements of size 2ε. The exact solution to (4.39) on two adjacent mesh elements (denoted K_1 and K_2) is displayed on Figure 4.2. We show on Figure 4.3 the basis functions corresponding respectively to the MsFEM and the Adv-MsFEM approach, as defined in Sections 4.2.1 and 4.2.2.

We see that the profile of the advective bubble functions of Adv-MsFEM seems closer to the exact solution than that of the bubble functions of MsFEM. This observation motivates the introduction of a "mixed" variant for which the approximation space collects the functions $\Phi^{\varepsilon,E}$ (associated to an edge) of MsFEM and the advective bubble functions $\Psi^{\varepsilon,K}$ of Adv-MsFEM. For the approximation of (4.2), the discrete variational formulation then writes

$$\begin{cases} \text{Find } u_H \in V_{H,\text{adv-bubble}} \text{ such that,} \\ \text{for any } v_H \in V_{H,\text{adv-bubble}}, \quad c_H(u_H, v_H) = F(v_H), \end{cases} \tag{4.40}$$

where the approximation space reads as

$$V_{H,\text{adv-bubble}} = \text{Span}\left\{\Phi_0^{\varepsilon,E},\ \Psi_{\text{D}}^{\varepsilon,K},\quad E \in \mathcal{E}_H^{\text{in}},\ K \in \mathcal{T}_H\right\},$$

where $\Phi_0^{\varepsilon,E}$ and $\Psi_{\text{D}}^{\varepsilon,K}$ are defined by (4.17) and (4.30).

The variational formulation of the stabilized version reads as:

$$\begin{cases} \text{Find } u_H \in V_{H,\text{adv-bubble}} \text{ such that, for any } v_H \in V_{H,\text{adv-bubble}}, \\ \qquad c_H(u_H, v_H) + a_{\text{stab}}(u_H, v_H) = F(v_H) + F_{\text{stab}}(v_H). \end{cases} \tag{4.41}$$

We can obtain a simplified expression of the stabilization term (4.16) by decomposing $u_H \in V_{H,\text{adv-bubble}}$ as

$$u_H = u_H^0 + \sum_{K\in\mathcal{T}_H} U_H^K\, \Psi_{\text{D}}^{\varepsilon,K} \quad \text{with} \quad u_H^0 = \sum_{E\in\mathcal{E}_H^{\text{in}}} U_H^E\, \Phi_0^{\varepsilon,E}.$$

According to the definition of the basis functions, we get

$$a_{\text{stab}}(u_H, v_H) = \sum_{K\in\mathcal{T}_H} \left(\tau_{\mathbf{K}}\left(\widehat{b}^{\varepsilon}\cdot\nabla u_H^0\right), \widehat{b}^{\varepsilon}\cdot\nabla v_H\right)_{L^2(K\cap\Omega^{\varepsilon})} + \sum_{K\in\mathcal{T}_H} U_H^K \int_{K\cap\Omega^{\varepsilon}} \tau_{\mathbf{K}}\left(\widehat{b}^{\varepsilon}\cdot\nabla v_H\right). \tag{4.42}$$

4.3 Elements of theoretical analysis

In this section 4.3, we consider periodic perforations. More precisely, let Y be the unit square and $\mathcal{O} \subset Y$ be some smooth perforation. We next scale $\mathcal{O}$ and Y by a factor ε and then periodically repeat this pattern with periods ε in all directions. The set of perforations is therefore

$$B_\varepsilon = \Omega \cap (\cup_{k\in\mathbb{Z}^d}\ \varepsilon B_k) \quad \text{with} \quad B_k = k + \mathcal{O} \tag{4.43}$$

and the perforated domain is $\Omega_\varepsilon = \Omega\backslash\overline{B_\varepsilon}$.

4.3.1 Homogenization results

For self-consistency and for the convenience of the reader, we include here some results of homogenization for the problems (4.2) and (4.3) considered. These results are useful to bear in mind the different scalings involved, and the asymptotic behavior of the solutions u^ε we approximate in the various cases. The proofs of these results are essentially contained in the literature, although some tiny details may vary. Also for convenience, we include the proof of these results in Appendix 4.5 of this article. In any event, we do not claim any originality for these results.

There is indeed a considerable body of literature for the homogenization of diffusion and advection-diffusion problems set on perforated domains. The behavior obtained in the homogenization limit drastically depend on the boundary conditions set on the boundaries of the perforations and on the density and size of these perforations. For the diffusion problem itself, we wish to cite [8, 22, 26, 27, 30, 64, 73], and more specifically [24, 27, 30, 59, 64, 73] for the case of Dirichlet boundary conditions. The advection-diffusion equation is studied in [7, 9, 10, 46, 82].

We also mention, for completeness, some of the many studies of the (Navier) Stokes equation in this setting, such as [3, 5, 67]. A general reference on such topics is the textbook [45].

In the case of homogeneous Dirichlet boundary conditions, that is problem (4.2), two different results, depending on the choice of the advection field $\widehat{b}^\varepsilon$, may be established, using standard arguments of the literature. Both results have proofs readily adapted from the already classical proofs of the same estimates for the pure diffusion operator (dating back to [64] and slightly extended in [59, Appendix A.2]).

As we briefly mentioned in the introduction, the most interesting case is when $\widehat{b}^\varepsilon = \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)$. Then, Theorem 4.2 below holds. Its proof is postponed until Appendix 4.5.1. The advection field b does affect the homogenized behavior, since the cell problem (4.46) defined there depends on b.

Theorem 4.2 (adapted from [64, 59]). *We assume (4.43), that the domain Ω^ε is connected, and that the right-hand side f belongs to $H^2(\Omega)$. Let $\widehat{b}^\varepsilon = \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)$ where b belongs to $(W^{1,\infty}(Y\backslash\overline{\mathcal{O}}))^d$, is Y-periodic and is such that $\operatorname{div} b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$. The solution u^ε to Problem (4.2) satisfies*

$$\left| u^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right) f \right|_{H^1(\Omega^\varepsilon)} \leqslant C\varepsilon^{3/2}\mathcal{N}(f), \tag{4.44}$$

with the notation $|v|_{H^1(\Omega^\varepsilon)} = \|\nabla v\|_{H^1(\Omega^\varepsilon)}$, where

$$\mathcal{N}(f) = \|f\|_{L^\infty(\Omega)} + \|\nabla f\|_{L^2(\Omega)} + \|\Delta f\|_{L^2(\Omega)} \tag{4.45}$$

and where w is the solution (actually in $C^1(\overline{Y\backslash\mathcal{O}})$) of the cell problem

$$\begin{cases} -\alpha\Delta w + b\cdot\nabla w = 1 \quad \text{in } Y\backslash\overline{\mathcal{O}}, \\ w \ \text{ is } Y\text{-periodic}, \quad w = 0 \quad \text{on } \partial\mathcal{O}. \end{cases} \tag{4.46}$$

On the other hand, when $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$, Theorem 4.3 below describes the asymptotic behavior of the solution to (4.2), which does not depend at the dominant order upon the advection field b.

Theorem 4.3 (adapted from [64, 59]). *Under the same assumptions as those of Theorem 4.2, except that here $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$, estimate (4.44) holds, where w is now defined as the solution to the cell problem*

$$\begin{cases} -\alpha\Delta w = 1 \quad \text{in } Y\backslash\overline{\mathcal{O}}, \\ w \ \text{ is } Y\text{-periodic}, \quad w = 0 \quad \text{on } \partial\mathcal{O}, \end{cases}$$

instead of (4.46).

We skip the proof of Theorem 4.3, that follows the same lines as the proof of Theorem 4.2.

For Neumann boundary conditions, the situation is different, as briefly mentioned in the introduction. No rescaling of the solution, which stays of order one, is necessary, and no enhancement of the advection field by a factor ε^{-1} is required for the advection to affect the homogenized limit. In the case of the diffusion problem, the problem was first solved in the case of *isolated* (meaning that $\overline{\mathcal{O}} \subset Y$) holes in [28]. The generalization to nonisolated holes was addressed in [2, 8, 26]. The homogenization limit for the advection-diffusion equation (4.3), in the periodic case, is the purpose of the following theorem. Its proof, which is postponed until Appendix 4.5.2, is an easy adaptation of the proof of [4, Theorem 2.9], that uses two-scale convergence and addresses the case without advection on a periodically perforated domain.

Theorem 4.4 (adapted from Theorem 2.9 of [4]). *We assume (4.43) and that Ω^ε is such that*

$$H^1(\Omega^\varepsilon) \hookrightarrow H^{1/2}(\partial\Omega^\varepsilon). \tag{4.47}$$

We also assume that the domain E, obtained by Y-periodicity from $Y\backslash\overline{\mathcal{O}}$, is a smooth connected open set of $\mathbb{R}^d$. Let $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$ where $b \in (W^{1,\infty}(Y\backslash\overline{\mathcal{O}}))^d$ is Y-periodic. We suppose either that b is curl-free or that div $b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$ and $b\cdot n \geqslant 0$ on $\partial\mathcal{O}$. We assume that $f \in L^2(\Omega)$. As ε vanishes, $u^\varepsilon - u^\star - \varepsilon\sum_{i=1}^d w_i\left(\frac{\cdot}{\varepsilon}\right)\partial_{x_i}u^\star$ vanishes in $H^1(\Omega^\varepsilon)$, where $u^\star$ is the solution to the problem

$$\begin{cases} -\text{div}\,(A^\star\nabla u^\star) + b^\star\cdot\nabla u^\star = \dfrac{|Y\backslash\overline{\mathcal{O}}|}{|Y|}f \quad \text{in } \Omega, \\ u^\star = 0 \quad \text{on } \partial\Omega, \end{cases} \tag{4.48}$$

where the matrix $A^\star$ and the vector $b^\star$ are constant and given, for $1 \leqslant i \leqslant d$, by

$$A^\star\, e_i = \frac{1}{|Y|}\int_{Y\backslash\overline{\mathcal{O}}} \alpha\,(e_i + \nabla w_i), \tag{4.49}$$

$$b^\star\cdot e_i = \frac{1}{|Y|}\int_{Y\backslash\overline{\mathcal{O}}} b\cdot(e_i + \nabla w_i), \tag{4.50}$$

and where w_i is the solution to the cell problem

$$\begin{cases} -\Delta w_i = 0 \quad \text{in } Y\backslash\overline{\mathcal{O}}, \\ w_i \ \text{ is } Y\text{-periodic}, \quad (\nabla w_i + e_i)\cdot n = 0 \quad \text{on } \partial\mathcal{O}. \end{cases} \tag{4.51}$$

As the perforations are smooth, we have a continuous injection from $H^1(Y\backslash\overline{\mathcal{O}})$ to $H^{1/2}(\partial\mathcal{O})$. The assumption (4.47) thus amounts to a geometrical assumption on the perforations that intersect the boundary of Ω.

4.3.2 Error analysis for the Dirichlet problem

The error estimate of the Adv-MsFEM approach with advective bubble functions, the variational formulation of which is (4.25), is the purpose of the following theorem. It is the extension of a similar result for the diffusion problem [59, Theorem 2.2] establishing for that case the exact analogue estimate as (4.53) below. This is not unexpected since, in both cases, the *same* differential operator is present in the original equation and in the definition of all the basis functions. The proof of Theorem 4.5 is postponed until Appendix 4.6.2.

Theorem 4.5. *We assume (4.43), that the domain Ω^ε is connected, and that $\widehat{b}^\varepsilon = \frac{1}{\varepsilon}b\left(\frac{\cdot}{\varepsilon}\right)$ where $b \in (W^{1,\infty}(Y\backslash\overline{\mathcal{O}}))^2$ is Y-periodic and is such that div $b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$. Let u^ε be the solution to (4.2) in dimension $d = 2$, with $f \in H^2(\Omega)$.*

We assume (and this is a purely technical assumption that does not matter for the numerical practice) that the slopes of the edges of the mesh elements are rational numbers, that is, more precisely, we suppose that the equation defining any internal edge E of the mesh reads as $x_2 = \frac{p_E}{q_E}x_1 + c_E$ for some $c_E \in \mathbb{R}$, $p_E \in \mathbb{Z}$ and $q_E \in \mathbb{N}^\star$ that are coprime, with

$$|q_E| \leqslant C, \tag{4.52}$$

for a constant C independent of the edge considered in the mesh and of the mesh size H. Then, the Adv-MsFEM approximation u_H^ε, solution to (4.25), satisfies

$$\|u^\varepsilon - u_H^\varepsilon\|_{H_H^1(\Omega^\varepsilon)} \leqslant C\varepsilon \left(\sqrt{\varepsilon} + H + \sqrt{\frac{\varepsilon}{H}}\right) \|f\|_{H^2(\Omega)}, \tag{4.53}$$

for a constant C independent of ε, H and f, where we have used the notation $\|v\|_{H_H^1(\Omega^\varepsilon)} = \sqrt{\sum_{K\in\mathcal{T}_H} \|v\|^2_{H^1(K\cap\Omega^\varepsilon)}}$.

As for the approach that does not account for the advection field in none of the basis functions, namely the MsFEM with bubble functions defined by the variational formulation (4.13), its error estimate is established in the following theorem, itself again an extension of the result for the diffusion problem [59, Theorem 2.2]. Again, the proof is postponed until Appendix 4.6.3. Understandably, the advection term, present in the original equation but not in the definition of the basis functions, has to be accommodated (thus the additional first term in the right-hand side of (4.54) below).

Theorem 4.6. *We make the same assumptions as those of Theorem 4.5 and let again u^ε be the solution to (4.2) in dimension $d = 2$, with $f \in H^2(\Omega)$. We additionally assume here that $b \in (W^{2,\infty}(Y\backslash\overline{\mathcal{O}}))^2$ and that $\|b\|_{L^\infty(Y\backslash\overline{\mathcal{O}})}$ is sufficiently small (in a sense made precise in the proof).*

Then, the MsFEM approximation u_H^ε, solution to (4.13), satisfies

$$\|u^\varepsilon - u_H^\varepsilon\|_{H_H^1(\Omega^\varepsilon)} \leqslant CH\frac{\|b\|_{L^\infty(Y\backslash\overline{\mathcal{O}})}}{\alpha}\left(1 + \frac{\alpha + C_P\|b\|_{L^\infty(Y\backslash\overline{\mathcal{O}})}}{\alpha - C_P\|b\|_{L^\infty(Y\backslash\overline{\mathcal{O}})}}\right)\|f\|_{L^2(\Omega)} + C\varepsilon\left(\sqrt{\varepsilon} + H + \sqrt{\frac{\varepsilon}{H}}\right)\|f\|_{H^2(\Omega)}, \tag{4.54}$$

for a constant C independent of ε, H and f.

Although we suspect that similar estimates to those of Theorems 4.5 and 4.6 above can be established for the problem with Neumann boundary conditions (4.3), we have not pursued in this direction.

4.4 Numerical results

This section presents our numerical results. They have all been performed with FreeFem++ [43], on the following test case. We consider the two-dimensional domain $\Omega = (0,1)^2$. Its subdomain Ω^ε is a periodically perforated domain defined by

$$\Omega^\varepsilon = \left\{x \in \Omega, \quad \chi\left(\frac{x}{\varepsilon}\right) = 1\right\}. \tag{4.55}$$

where χ is the extension by Y-periodicity, for the periodicity cell $Y = (0,1)^2$, of the characteristic function $\mathbb{1}_{Y\backslash\overline{\mathcal{O}}}$, where $\mathcal{O} \subset Y$ defines a perforation.

For either of the problems considered ((4.2) or (4.3)), and for either of our approaches, based on the diffusion operator only or the full advection-diffusion operator, we will investigate several

issues. The first issue is how enriching the approach with bubble functions affects the accuracy. Of course, as all our approaches are variational approaches, enriching the variational approximation space using bubble functions can only *improve* the accuracy, at the price of increasing the number of degrees of freedom. We however notice that, when a gain in accuracy is observed, that gain is much higher than that obtained by, say, reducing by a factor two the size of the coarse mesh. We therefore then safely conclude about the added value of bubble functions. Other issues that are specifically examined below are the influence of the Péclet number (measuring the relative amplitude of the advection with respect to the diffusion) and that of the small scale ε defining both the size of the perforations and their typical distance. Many of these issues are examined upon considering a range of mesh sizes H for the coarse mesh. This range is typically chosen as H varying from $\varepsilon/10$ to $10\,\varepsilon$. One must bear in mind that capturing all the details of the oscillatory solutions u^ε using a standard FEM approach would require choosing a mesh size in any event smaller, and in most cases much smaller, than $\varepsilon/10$. At the other end, choosing H larger than $10\,\varepsilon$ is hopeless. Thus the choice of our typical range of values of H.

Beside comparing the various approaches considered, and assessing their performance in function of the various parameters of the problem, we will also specifically assess their robustness with respect to the location of the perforations. To this aim, we consider two locations for the perforation within the periodicity cell $Y = (0,1)^2$:

$$\mathcal{O} = \mathcal{O}_1 = (0.25, 0.75)^2$$

and

$$\mathcal{O} = \mathcal{O}_2 = (0, 0.25) \times (0.25, 0.75) \cup (0.75, 1) \times (0.25, 0.75).$$

The shape of the perforations is the same (squares of size 0.5ε). The difference lies in the relative position of the mesh with respect to the perforations. One set of perforations is obtained from the other by shifting by $0.5\,\varepsilon$ in the x direction. When $\mathcal{O} = \mathcal{O}_1$, the perforations do not intersect the edges of the mesh elements (which are taken aligned with the periodicity cells). In contrast, when $\mathcal{O} = \mathcal{O}_2$, many edges are intersected by the perforations. In doing so, we have in mind, like in our previous works, to use these two specific periodic geometries to emphasize which approaches can easily carry over to the case of non periodic perforations, where a typical mesh may often intersect the perforations. To some extent, the two periodic geometries we consider respectively represent the best case scenario (when perforations are all interior to mesh elements) and the worst case scenario (when "half" the perforations intersect the boundaries of mesh elements).

The advection field $\widehat{b}^\varepsilon$ we consider is proportional to the constant field $b = (1,1)^T$. Depending on the situation considered, the proportionality constant is either 1 or $1/\varepsilon$, for reasons that have been made clear above. Throughout the section, the Péclet number, the non dimensionalized number that measures the relative importance of advection over diffusion, is denoted by $\mathrm{Pe} = \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega)}/(2\alpha)$.

The reference solution u_{ref}, and all the relative errors that will be defined with respect to that reference solution, are computed on the fine mesh. The reference solution itself is computed using the standard $\mathbb{P}^1$ Finite Element method on this fine mesh. We measure the accuracy using, on domains $\omega \subset \Omega^\varepsilon$, the H^1 broken norm

$$|u|_{H^1_H(\omega)} = \left(\sum_{K \in \mathcal{T}_H} \|\nabla u\|^2_{L^2(K \cap \omega)} \right)^{1/2}, \tag{4.56}$$

and the relative errors

$$e_{H^1(\omega)}(u) = \frac{|u - u_{\text{ref}}|_{H^1_H(\omega)}}{|u_{\text{ref}}|_{H^1(\omega)}}, \tag{4.57}$$

in the *whole* domain ($\omega = \Omega^\varepsilon$) and, possibly, separately inside and outside the boundary layer when there is such a boundary layer close to some portion of the boundary of the domain Ω (see Section 4.4.2).

The results for Problem (4.2), where we impose homogeneous Dirichlet boundary conditions on the perforations, are presented in Section 4.4.1, while Section 4.4.2 contains those for the Neumann problem (4.3).

In all what follows, we choose $f(x,y) = \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}y\right)$ as right-hand side for the advection-diffusion equation considered (we have checked that our results and conclusions do not sensitively depend on the choice of f).

4.4.1 Homogeneous Dirichlet boundary condition

Unless otherwise stated, the Adv-MsFEM and its variants defined in Section 4.2.2 are derived from the approximation space (4.26).

In Section 4.4.1, we study the added value of bubble functions. Sections 4.4.1 and 4.4.1 respectively explore the influence of the Péclet number and the small scale ε.

Adding bubble functions

We fix $\varepsilon = 0.03$ and $\alpha = 0.25$. The perforations are defined by the set $\mathcal{O} = \mathcal{O}_1$, but we have explicitly checked, alternately choosing $\mathcal{O} = \mathcal{O}_2$, that all our results and conclusions in this section are qualitatively insensitive to the location of the perforations.

To start with, we consider the MsFEM approach. We observe on Figure 4.4 that adding bubble functions significantly improves the accuracy, and that the best option is that with *advective* bubble functions.

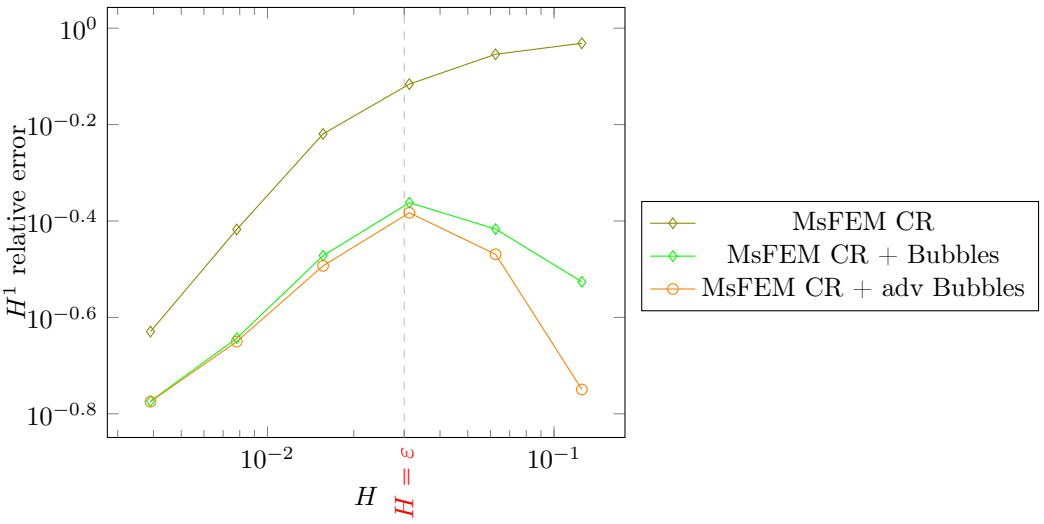

Figure 4.4 [Dirichlet Problem (4.2)] Addition of bubble functions to MsFEM CR.

We next turn to Adv-MsFEM. *Only here in the present article*, we temporarily also include in our comparison multiscale basis functions built with boundary conditions other than Crouzeix-

Raviart, namely elements with linear boundary conditions and elements using oversampling (specifically with an oversampling ratio equal to 3, see [36] for the definition). In the case of the Adv-MsFEM CR approach, the bubble functions are defined by (4.30), with Crouzeix Raviart boundary conditions. In the case of the Adv-MsFEM lin approach (with affine boundary conditions on ∂K) and of the Adv-MsFEM OS approach (with oversampling), the bubble functions are defined using Dirichlet homogeneous conditions on ∂K, that is as the solution to

$$\begin{cases} -\alpha\Delta\Psi_{\mathrm{D}}^{\varepsilon,K} + \widehat{b}^{\varepsilon}\cdot\nabla\Psi_{\mathrm{D}}^{\varepsilon,K} = 1 & \text{in } K\cap\Omega^{\varepsilon}, \\ \Psi_{\mathrm{D}}^{\varepsilon,K} = 0 \quad \text{in } K\cap B^{\varepsilon}, & \Psi_{\mathrm{D}}^{\varepsilon,K} = 0 \quad \text{on } \partial K. \end{cases}$$

Figure 4.5 displays the relative H^1 broken error of the different approaches. We again observe that adding advective bubble functions significantly improves the accuracy, and that Adv-MsFEM à la Crouzeix Raviart with advective bubble functions is the best of all the approaches considered.

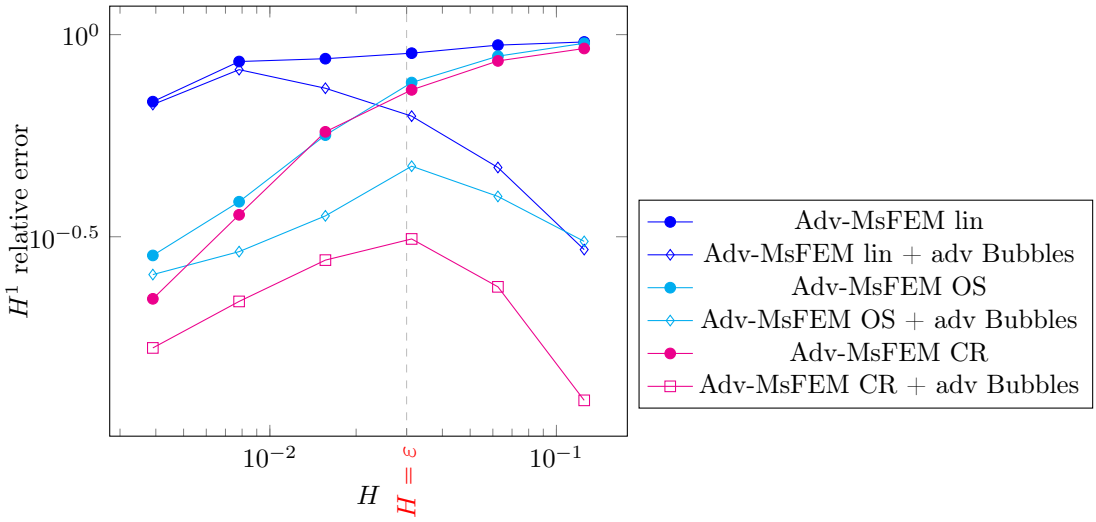

Figure 4.5 [Dirichlet Problem (4.2)] Addition of bubble functions to Adv-MsFEM and variants.

Influence of the Péclet number

We now study the influence of a large advection, quantified by the Péclet number, on the accuracy of our approaches. We fix $\varepsilon = 0.03125$ and the mesh size $H = 1/16$. We choose $\mathcal{O} = \mathcal{O}_1$, the configuration where the perforations do not intersect the coarse mesh. In order to vary the Péclet number, we let the diffusion parameter take the values $\alpha = 2^k$, for the integers $k = -5$ through 2. When α decreases, Problem (4.2) increasingly becomes advection-dominated. Given the results of our previous section, we only consider MsFEM and Adv-MsFEM with advective bubble functions. Figure 4.6 shows that the latter approach stays accurate when α decreases while the error blows up to a hundred percent for the former approach.

We have checked that our conclusions are not modified when using a stabilized formulation of the various approaches. They are not modified either when considering shifted perforations $\mathcal{O} = \mathcal{O}_2$, many of which now intersect the edges of mesh elements.

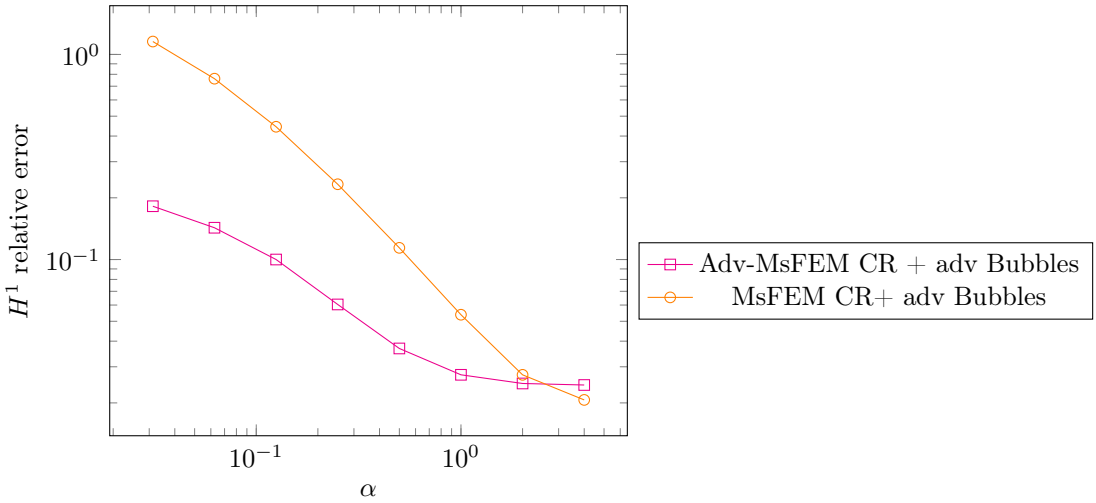

Figure 4.6 [Dirichlet Problem (4.2)] Sensitivity to the Péclet number.

Influence of the small scale ε

We fix $\alpha = 1/16$, $H = 1/16$, and, in order to evaluate the influence of the small scale ε, let ε take the values $\varepsilon = 2^{-k}$ for $k = 3, \dots, 8$. Since we know from previous observations that stabilization does not bring any added value, we only consider our approaches without stabilization. In Figure 4.7, we observe that the most accurate method is the Adv-MsFEM with advective bubble functions.

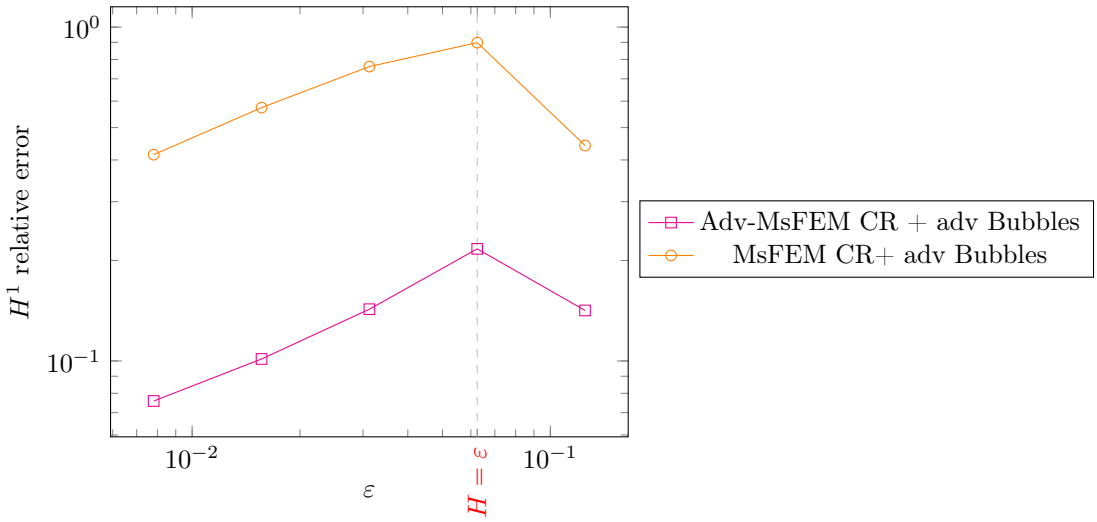

Figure 4.7 [Dirichlet Problem (4.2)] Sensitivity to the the small scale ε.

The results are shown for $\mathcal{O} = \mathcal{O}_1$, in which case the perforations do not intersect the coarse mesh, but we have also checked that the exact same conclusion holds for $\mathcal{O} = \mathcal{O}_2$.

Comparison of two variants of Adv-MsFEM

In the previous sections, we have always used the formulation (4.25) (with the bilinear form c_H defined by (4.10)) for the definition of Adv-MsFEM. As briefly mentioned in the introduction and in Remark 4.1, a formulation such as (4.34) (which we introduce and use for the Neumann

problem) can also be considered for the present Dirichlet case. The difference between the two approaches is the use of the bilinear form a_H instead of c_H in the definition of the local and global problems, which in particular implies different boundary conditions prescribed on the inner edges/faces of the mesh elements, see Section 4.2.2.

We first compare the two methods as in Section 4.4.1, that is for varying Péclet numbers (i.e. varying α). We first fix $\mathcal{O} = \mathcal{O}_1$, and next fix $\mathcal{O} = \mathcal{O}_2$. In both cases, we have observed (results not shown) that the behavior of the methods (4.25) and (4.34) when α decreases is similar.

We then consider again the setting of Section 4.4.1. We start with $\mathcal{O} = \mathcal{O}_2$. In Figure 4.8, both methods yields a reasonable accuracy for small values of ε. We now set $\mathcal{O} = \mathcal{O}_1$, all the other parameters being unchanged. We see in Figure 4.9 that method (4.34) is much less accurate than method (4.25) for small values of ε. This provides a practical motivation (in addition to the theoretical motivation outlined above) to use method (4.25).

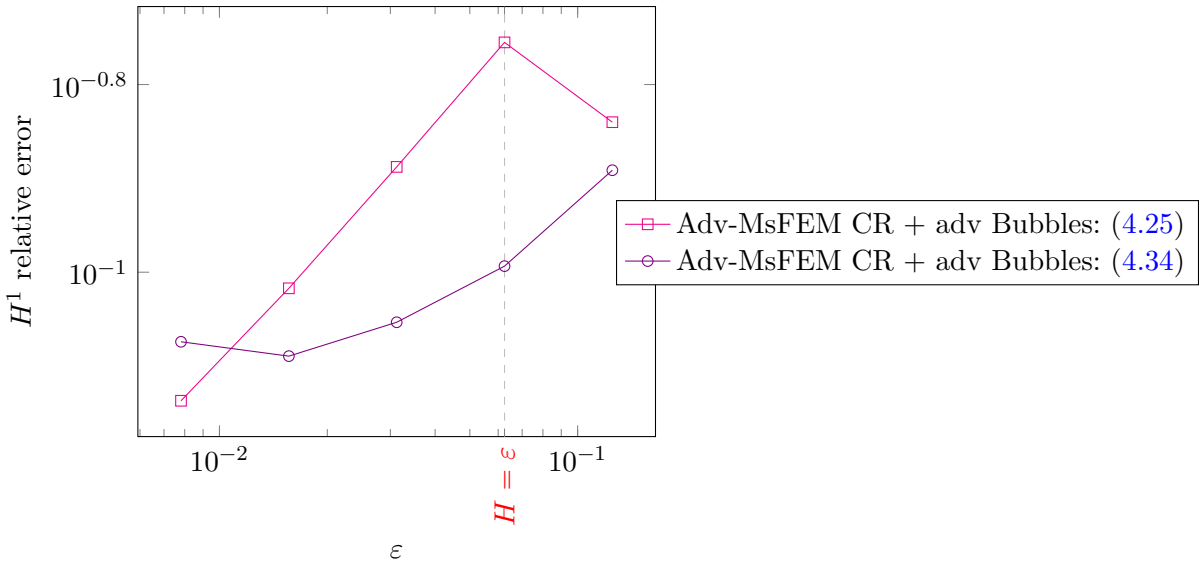

Figure 4.8 Comparison of the Adv-MsFEM CR variants: e_{H^1} ($\alpha = 1/16$, $\mathcal{O} = (0, 0.25) \times (0.25, 0.75) \cup (0.75, 1) \times (0.25, 0.75)$, $f = \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}y\right)$, $H = 1/16$).

4.4.2 Homogeneous Neumann boundary condition

In this section, we investigate the influence of the Péclet number and the small scale ε. Then, we study the effect of adding bubble functions.

We consider the Neumann problem (4.3) for a constant advection field $\widehat{b}^\varepsilon$, namely $\widehat{b}^\varepsilon = (1, 1)^T$. As expressed by Theorem 4.4, the homogenized problem is an advection-dominated problem posed in Ω. In contrast to the situation with homogeneous Dirichlet boundary conditions, the flow is not slowed down by the boundary conditions set on the boundary of perforations. It however has to comply with the Dirichlet boundary conditions on the outer boundary of the domain Ω. Given the orientation of $\widehat{b}^\varepsilon$, a boundary layer is expected close to the upper right corner of Ω. We denote by $\Omega_{\text{layer}} = \Big((0, 1) \times (1 - \delta_{\text{layer}}, 1)\Big) \cup \Big((1 - \delta_{\text{layer}}, 1) \times (0, 1)\Big)$ this expected boundary layer, of approximate width $\delta_{\text{layer}} = \dfrac{1}{\text{Pe}} \log(\text{Pe})$.

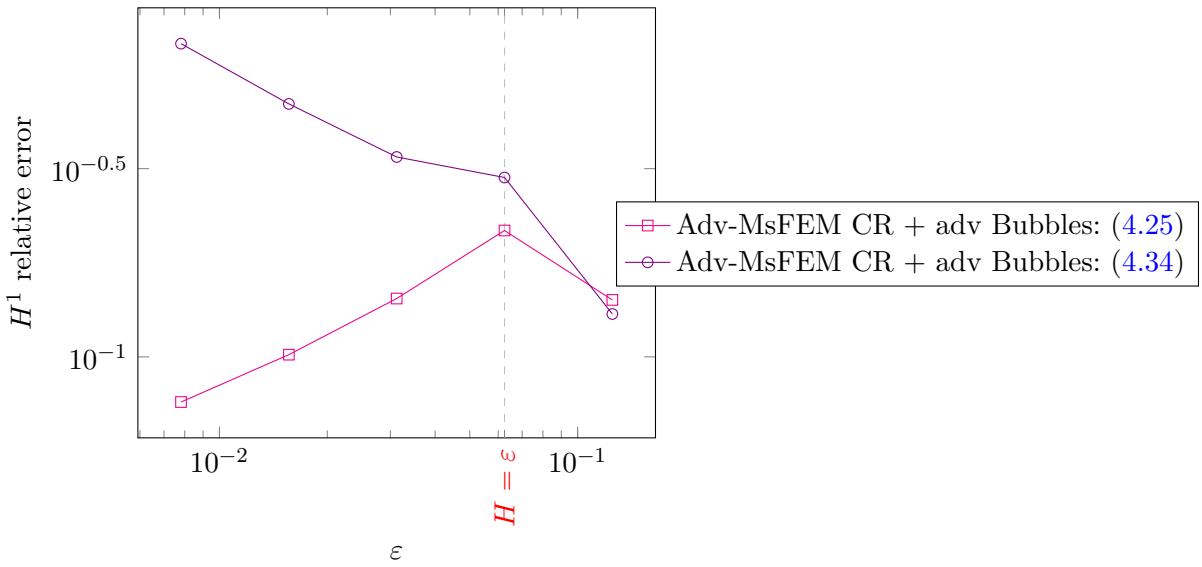

Figure 4.9 Comparison of the Adv-MsFEM CR variants: e_{H^1} ($\alpha = 1/16$, $\mathcal{O} = (0.25, 0.75)^2$, $f = \sin\left(\frac{\pi}{2}x\right)\sin\left(\frac{\pi}{2}y\right)$, $H = 1/16$).

Influence of the Péclet number

As already mentioned, one consequence of the Neumann conditions, as opposed to the homogeneous Dirichlet conditions, set on the boundary of the perforations is that the flow is not slowed down around the perforations. Thus, when advection dominates diffusion, the effect of advection is all the more acute (and a boundary layer indeed develops close to the boundary of the domain Ω). Since advection is more extreme, it is therefore important to primarily investigate how the approaches perform on Problem (4.3) when advection increasingly dominates diffusion (a study we presented in Section 4.4.1 for the Dirichlet problem). In practice, we perform our tests fixing $\varepsilon = 0.03125$, $H = 1/16$ and varying $\alpha = 2^k$, for integers $k = -9$ to -2.

It is well known that all discretization methods poorly perform within the boundary layer in the advection dominated regime. The only exceptions are methods specifically tailored to the boundary layer and we do not wish to go in that direction. We have checked that all our approaches essentially fail in the boundary layer, the error for some of them even blowing up to more than a hundred percent. Therefore, in order to discriminate between the approaches, we only consider the region outside the boundary layer (we have also adopted such a strategy in [60]). Figure 4.10 shows the relative error (4.57) (for $\omega = \Omega^\varepsilon \backslash \Omega_{\text{layer}}$ in (4.56)) calculated there, in the configuration where the perforations do not intersect the coarse mesh, i.e. when $\mathcal{O} = \mathcal{O}_1$. We observe that Adv-MsFEM performs well. As is the case for MsFEM provided it is stabilized. Figure 4.11 shows the results of the same tests for $\mathcal{O} = \mathcal{O}_2$. It confirms the same conclusions, qualitatively, and therefore the flexibility of our approaches all based upon Crouzeix-Raviart type boundary conditions.

Influence of the small scale ε

We fix $\alpha = 1/256$, $H = 1/16$ and we vary $\varepsilon = 2^{-k}$, $k = 5, \ldots, 8$. We only show here the results when the perforations do not intersect the coarse mesh, i.e. when $\mathcal{O} = \mathcal{O}_1$. The results for $\mathcal{O} = \mathcal{O}_2$ are similar.

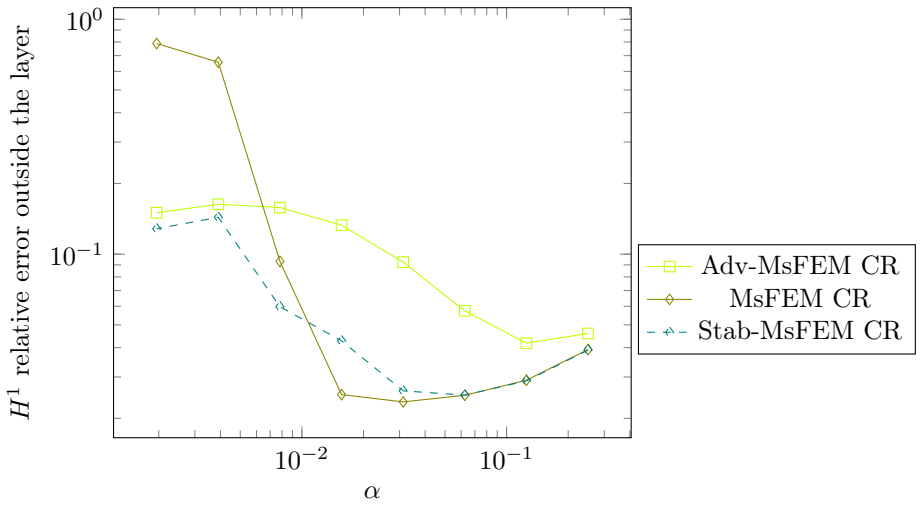

Figure 4.10 [Neumann Problem (4.3)] Sensitivity to the Péclet number: Error outside the layer when $\mathcal{O} = \mathcal{O}_1$.

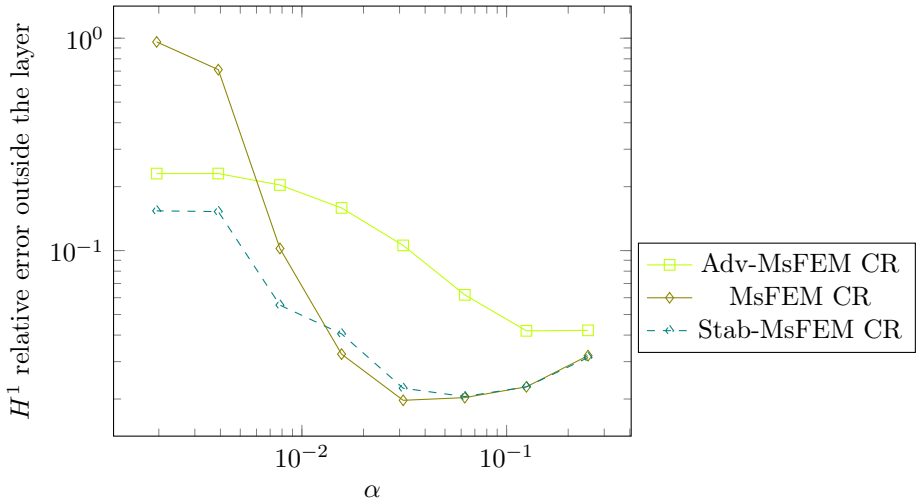

Figure 4.11 [Neumann Problem (4.3)] Sensitivity to the Péclet number: Error outside the layer when $\mathcal{O} = \mathcal{O}_2$.

Figure 4.12 and Figure 4.13 both show that the relative error, respectively on the whole domain and outside the boundary layer, is essentially insensitive to the small scale ε. The comparison of the actual size of the error in each of the two figures shows that the error within the boundary layer significantly dominates that outside the layer and is often prohibitively large, as is usually the case in the advection-dominated regime and as was mentioned in the previous section. In both figures, we observe that MsFEM is outperformed. Overall, Adv-MsFEM performs the best, but Stab-MsFEM is the most accurate method outside the layer.

Adding bubble functions

In this section, we study the added value of bubble functions for Adv-MsFEM and, given that the conclusions of the previous sections that show the inaccuracy of MsFEM itself, the stabilized variant Stab-MsFEM.

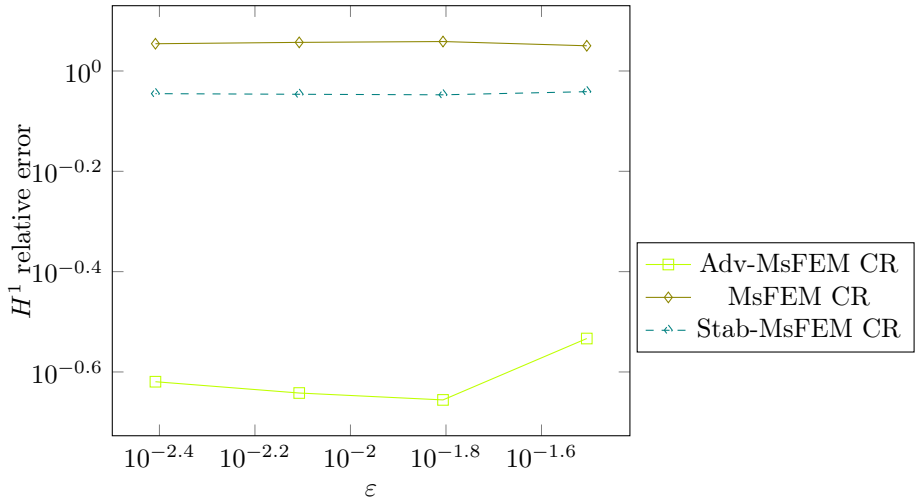

Figure 4.12 [Neumann Problem (4.3)] Sensitivity to the small scale ε: Error in the whole domain.

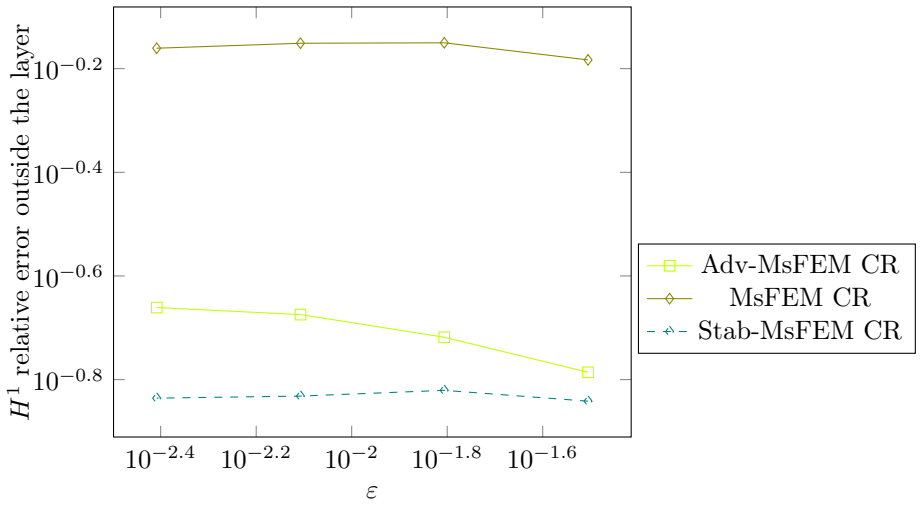

Figure 4.13 [Neumann Problem (4.3)] Sensitivity to the small scale ε: Error outside the layer.

Adv-MsFEM with advective bubbles functions We consider the test cases of Section 4.4.2. Figure 4.14 displays the relative H^1 broken error of our different approaches outside the boundary layer, when $\mathcal{O} = \mathcal{O}_1$. The case $\mathcal{O} = \mathcal{O}_2$ is shown on Figure 4.15. We observe that the Adv-MsFEM with advective bubble functions outperforms the Adv-MsFEM and the Stab-MsFEM (without bubble functions). It in fact also gives reasonable results in the whole domain (results not shown).

We next turn to the test cases of Section 4.4.2. In Figure 4.16, we observe, for $\mathcal{O} = \mathcal{O}_1$, that the Adv-MsFEM with advective bubble functions yields a reasonable accuracy. Choosing next $\mathcal{O} = \mathcal{O}_2$, we see on Figure 4.17 the relative H^1 broken error of the Adv-MsFEM with advective bubble functions inside and outside the boundary layer. Comparing Figures 4.16 and 4.17, we infer that:

- inside the boundary layer, the Adv-MsFEM with advective bubble functions is sensitive to the location of the perforations with respect to the coarse mesh;
- outside the boundary layer, the Adv-MsFEM with advective bubble functions is robust to the location of the perforations with respect to the coarse mesh.

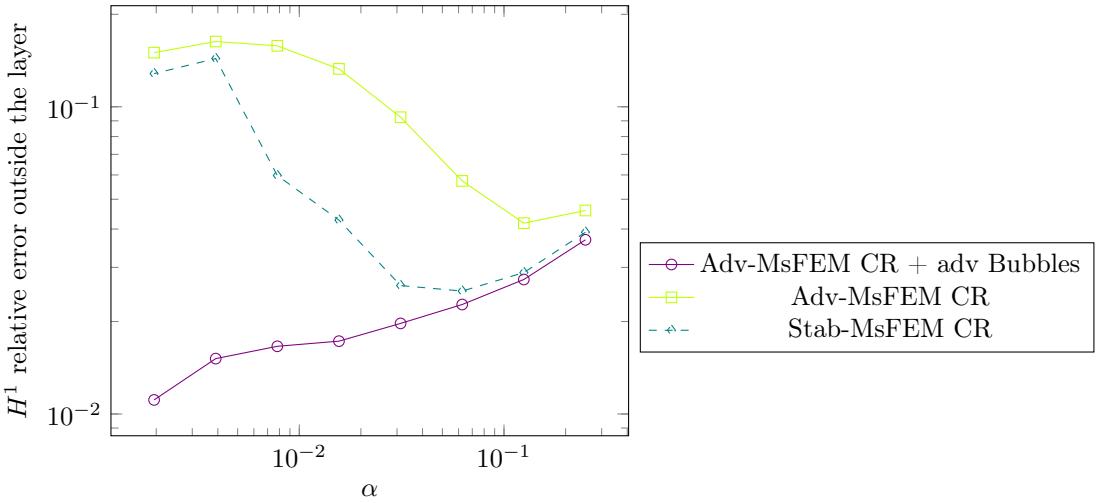

Figure 4.14 [Neumann Problem (4.3)] Adding bubble functions: Error outside the layer when $\mathcal{O} = \mathcal{O}_1$.

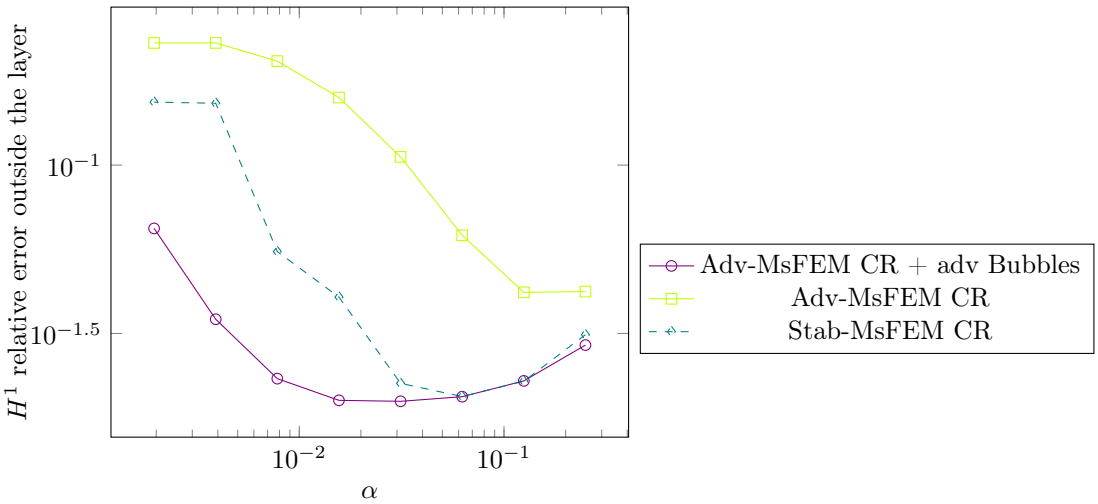

Figure 4.15 [Neumann Problem (4.3)] Adding bubble functions: Error outside the layer when $\mathcal{O} = \mathcal{O}_2$.

Stab-MsFEM with advective bubbles functions We consider the test cases of Section 4.4.2 in the case $\mathcal{O} = \mathcal{O}_1$. Figure 4.18 shows the error outside the boundary layer. We observe that adding bubbles functions does not significantly improve the accuracy. The results for $\mathcal{O} = \mathcal{O}_2$ are similar.

Nonperiodic geometry of perforations A main motivation for using MsFEM approaches is to consider nonperiodic cases. Unlike the periodic setting, homogenization theory does not yield any explicit approximation strategy. In this section, we assess the performance of our approaches on a nonperiodic geometry $\Omega^\varepsilon_{\mathrm{np}}$ depicted in Figure 4.19 and compare to an equivalent periodically perforated domain $\Omega^\varepsilon_{\mathrm{p}}$ defined by (4.55) with $\varepsilon = 0.03125$, $Y = (0,1)^2$ and $\mathcal{O} = r\mathcal{O}_1$ where $r > 0$ is such that $|\Omega^\varepsilon_{\mathrm{p}}| = |\Omega^\varepsilon_{\mathrm{np}}|$. To this aim, we choose the test case introduced in the study of the influence of the Péclet number. We recall that we fix $\varepsilon = 0.03125$, $H = 1/16$ and we vary $\alpha = 2^k$, for integers $k = -9$ to -2. Figure 4.20 displays the relative H^1 broken error outside the boundary layer of our most accurate approaches, namely the Adv-MsFEM with advective

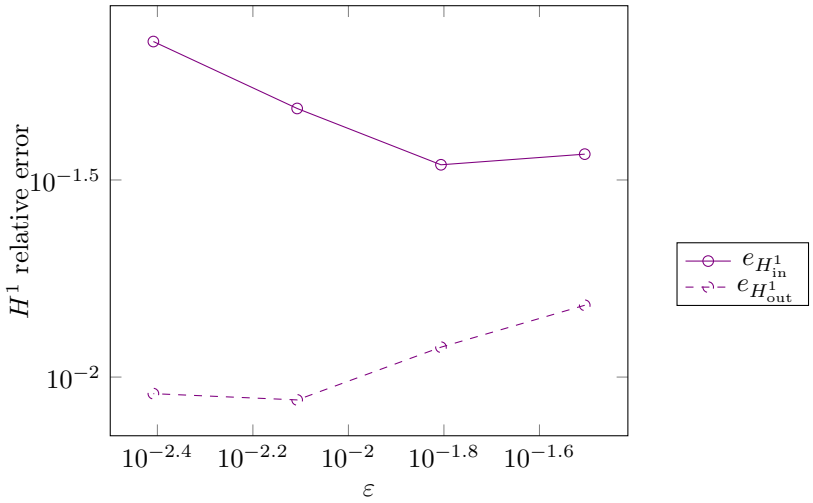

Figure 4.16 [Neumann Problem (4.3)] Sensitivity to the small scale ε: Adv-MsFEM with advective bubble functions ($\mathcal{O} = \mathcal{O}_1$).

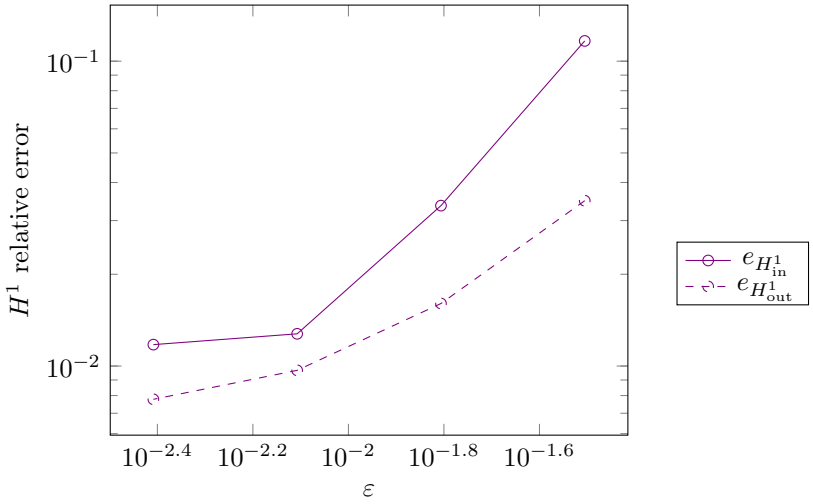

Figure 4.17 [Neumann Problem (4.3)] Sensitivity to the small scale ε: Adv-MsFEM with advective bubble functions ($\mathcal{O} = \mathcal{O}_2$).

bubble functions and the Stab-MsFEM. We observe that the Stab-MsFEM is insensitive to the nonperiodicity of the geometry. We see that the Adv-MsFEM with advective bubble functions is more sensitive to the nonperiodicity of the geometry but still outperforms the Stab-MsFEM in both cases.

Acknowledgments

The work of the authors is partially supported by the ONR under grant N00014-15-1-2777 and the EOARD under grant FA8655-13-1-3061.

4.5 Appendix: Homogenization results

We include here the proof of the homogenization limit for some of the problems we consider.

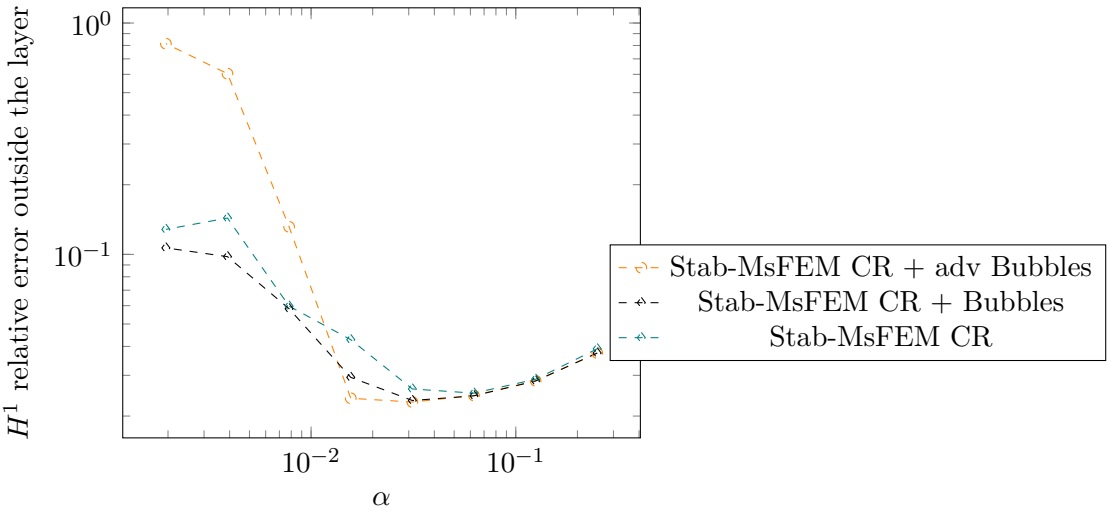

Figure 4.18 [Neumann Problem (4.3)] Adding bubbles to Stab-MsFEM: Error outside the layer.

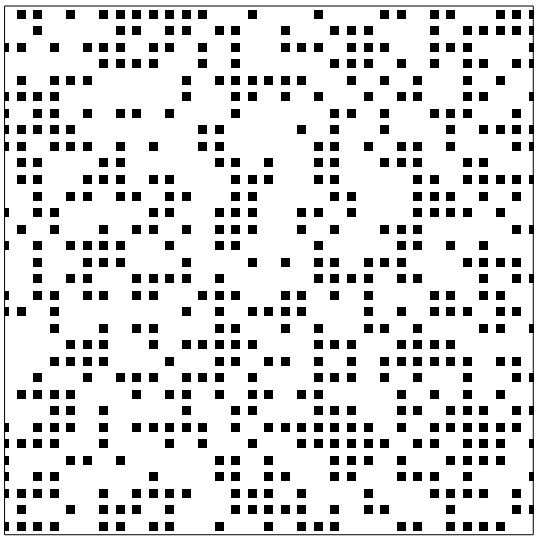

Figure 4.19 Nonperiodic geometry: Let M_ε be the set of perforations obtained by periodically perforating the domain $\Omega = (0,1)^2$ with $Y = (0,1)^2$, $\varepsilon = 0.03125$ and the motif $\mathcal{O}_2$. We include each perforation of M_ε in B_ε according to a Bernoulli distribution of parameter 1/2.

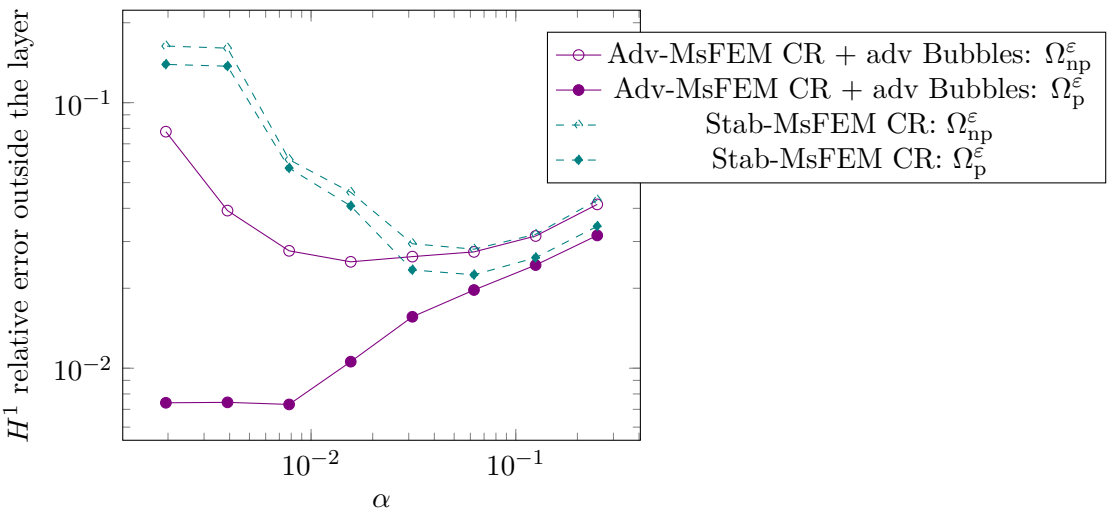

Figure 4.20 [Neumann Problem (4.3)] Sensitivity to the nonperiodicity of the geometry

4.5.1 Homogeneous Dirichlet boundary condition

We prove here Theorem 4.2. The proof of Theorem 4.3 follows the same pattern and we therefore omit it. A key ingredient in the proof below is the following Poincaré inequality (see [59, Appendix A.1]): there exists $C > 0$ independent of ε such that

$$\forall \phi \in H^1_0(\Omega^\varepsilon), \quad \|\phi\|_{L^2(\Omega^\varepsilon)} \leqslant C\varepsilon \|\nabla \phi\|_{L^2(\Omega^\varepsilon)} = C\varepsilon |\phi|_{H^1(\Omega^\varepsilon)}, \tag{4.58}$$

where we recall the notation $|v|_{H^1(\Omega^\varepsilon)} = \|\nabla v\|_{L^2(\Omega^\varepsilon)}$ for any $v \in H^1(\Omega^\varepsilon)$.

Proof of Theorem 4.2. We adapt the proof of [59, Appendix A.2], where we considered a purely diffusive problem. We first prove that Problem (4.46) is well-posed. Consider

$$V = \left\{ w \in H^1(Y \backslash \overline{\mathcal{O}}), \quad w \text{ is } Y\text{-periodic}, \quad w = 0 \text{ on } \partial\mathcal{O} \right\}.$$

The variational formulation of (4.46) reads as: find $w \in V$ such that

$$\forall v \in V, \quad a(w, v) = \int_\Omega v$$

with

$$a(w, v) = \int_{Y \backslash \overline{\mathcal{O}}} \alpha \nabla w \cdot \nabla v + \int_{Y \backslash \overline{\mathcal{O}}} (b \cdot \nabla w)\, v.$$

The bilinear form a is coercive on V. Indeed, for any $w \in V$,

$$\begin{aligned} a(w, w) &= \int_{Y \backslash \overline{\mathcal{O}}} \alpha |\nabla w|^2 + \int_{Y \backslash \overline{\mathcal{O}}} (b \cdot \nabla w)\, w \\ &= \int_{Y \backslash \overline{\mathcal{O}}} \alpha |\nabla w|^2 + \int_{Y \backslash \overline{\mathcal{O}}} b \cdot \nabla \left(\frac{w^2}{2} \right) \\ &= \int_{Y \backslash \overline{\mathcal{O}}} \alpha |\nabla w|^2 - \int_{Y \backslash \overline{\mathcal{O}}} (\operatorname{div} b) \left(\frac{w^2}{2} \right), \end{aligned}$$

where we used the periodicity of w and b and the fact that $w = 0$ on $\partial\mathcal{O}$. Using now that $\operatorname{div} b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$ and a Poincaré inequality in V, we get, for any $w \in V$,

$$a(w,w) \geqslant \alpha\|\nabla w\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} \geqslant C\|w\|^2_{H^1(Y\backslash\overline{\mathcal{O}})},$$

and thus the coercivity of a. The solution w of (4.46) is thus well defined.

Since the perforations are isolated, we can consider the cell problem on a larger bounded open domain Y^+ of class C^1 such that $\overline{Y} \subset Y^+$ and $Y^+ \cap \mathcal{O}^\star = \mathcal{O}$, where $\mathcal{O}^\star$ is the domain obtained by Y-periodicity from $\mathcal{O}$. Applying [40, Theorem 8.10] and using the regularity of b, we get that $w \in W^{k,2}(Y\backslash\overline{\mathcal{O}})$, for some $k > 1 + d/2$. This holds when $b \in W^{k-2,\infty}(Y\backslash\overline{\mathcal{O}})$. Using the Sobolev inclusions, we obtain that $w \in C^1(\overline{Y\backslash\mathcal{O}})$.

Using similar arguments, we observe that (4.2) is well-posed for any $f \in L^2(\Omega)$.

We now prove (4.44). Let η^ε be a regular function, vanishing in the neighborhood of the boundary of Ω, such that $0 \leqslant \eta^\varepsilon \leqslant 1$ on $\overline{\Omega}$, and which is equal to 1 on $\{x \in \Omega, \ \operatorname{dist}(x, \partial\Omega) > \varepsilon\}$ (see Figure 4.1). Since the domain Ω is regular, we can construct η^ε such that it satisfies

$$\begin{aligned}
&\|\eta^\varepsilon\|_{L^\infty(\Omega)} \leqslant C, \qquad \|1-\eta^\varepsilon\|_{L^2(\Omega)} \leqslant C\sqrt{\varepsilon},\\
&\|\nabla\eta^\varepsilon\|_{L^\infty(\Omega)} \leqslant \frac{C}{\varepsilon}, \qquad \|\nabla\eta^\varepsilon\|_{L^2(\Omega)} \leqslant \frac{C}{\sqrt{\varepsilon}}, \qquad \|\nabla^2\eta^\varepsilon\|_{L^2(\Omega)} \leqslant \frac{C}{\varepsilon^{3/2}},
\end{aligned}$$

where $C > 0$ is a constant independent of ε. We define $\phi^\varepsilon = u^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right) f\,\eta^\varepsilon$ and compute

$$\begin{aligned}
\nabla\phi^\varepsilon &= \nabla u^\varepsilon - \varepsilon\nabla w\left(\frac{\cdot}{\varepsilon}\right) f\eta^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right)\nabla(f\eta^\varepsilon),\\
\Delta\phi^\varepsilon &= \Delta u^\varepsilon - \Delta w\left(\frac{\cdot}{\varepsilon}\right) f\eta^\varepsilon - 2\varepsilon\nabla w\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla(f\eta^\varepsilon) - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right)\Delta(f\eta^\varepsilon).
\end{aligned}$$

Recalling that $\widehat{b}^\varepsilon = \dfrac{1}{\varepsilon} b\left(\dfrac{\cdot}{\varepsilon}\right)$, we get

$$\begin{aligned}
&-\alpha\Delta\phi^\varepsilon + \widehat{b}^\varepsilon\cdot\nabla\phi^\varepsilon\\
&= f + \alpha\Delta w\left(\frac{\cdot}{\varepsilon}\right) f\eta^\varepsilon + 2\varepsilon\alpha\nabla w\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla(f\eta^\varepsilon)\\
&\quad + \frac{1}{\varepsilon} b\left(\frac{\cdot}{\varepsilon}\right)\cdot\left(-\varepsilon\nabla w\left(\frac{\cdot}{\varepsilon}\right) f\eta^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right)\nabla(f\eta^\varepsilon)\right) + \varepsilon^2\alpha w\left(\frac{\cdot}{\varepsilon}\right)\Delta(f\eta^\varepsilon)\\
&= f + \left(\alpha\Delta w\left(\frac{\cdot}{\varepsilon}\right) - b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla w\left(\frac{\cdot}{\varepsilon}\right)\right) f\eta^\varepsilon + 2\varepsilon\alpha\nabla w\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla(f\eta^\varepsilon)\\
&\quad - \varepsilon b\left(\frac{\cdot}{\varepsilon}\right)\cdot w\left(\frac{\cdot}{\varepsilon}\right)\nabla(f\eta^\varepsilon) + \varepsilon^2\alpha w\left(\frac{\cdot}{\varepsilon}\right)\Delta(f\eta^\varepsilon)\\
&= f(1-\eta^\varepsilon) + \varepsilon\left(2\alpha\nabla w\left(\frac{\cdot}{\varepsilon}\right) - b\left(\frac{\cdot}{\varepsilon}\right) w\left(\frac{\cdot}{\varepsilon}\right)\right)\cdot\nabla(f\eta^\varepsilon) + \varepsilon^2\alpha w\left(\frac{\cdot}{\varepsilon}\right)\Delta(f\eta^\varepsilon).
\end{aligned}$$

We infer from the above that

$$\begin{aligned}
&\left\|-\alpha\Delta\phi^\varepsilon + \widehat{b}^\varepsilon\cdot\nabla\phi^\varepsilon\right\|_{L^2(\Omega^\varepsilon)}\\
&\leqslant \|f\|_{L^\infty(\Omega)}\|1-\eta^\varepsilon\|_{L^2(\Omega)}\\
&\quad + \varepsilon\|2\alpha\nabla w - bw\|_{L^\infty(Y\setminus\overline{\mathcal{O}})}\left(\|f\|_{L^\infty(\Omega)}\|\nabla\eta^\varepsilon\|_{L^2(\Omega)} + \|\nabla f\|_{L^2(\Omega)}\|\eta^\varepsilon\|_{L^\infty(\Omega)}\right)\\
&\quad + \varepsilon^2\alpha\|w\|_{L^\infty(Y\setminus\overline{\mathcal{O}})}\left(\|f\|_{L^\infty(\Omega)}\|\Delta\eta^\varepsilon\|_{L^2(\Omega)} + 2\|\nabla f\|_{L^2(\Omega)}\|\nabla\eta^\varepsilon\|_{L^\infty(\Omega)} + \|\Delta f\|_{L^2(\Omega)}\|\eta^\varepsilon\|_{L^\infty(\Omega)}\right)\\
&\leqslant C\sqrt{\varepsilon}\left(\|f\|_{L^\infty(\Omega)} + \|\nabla f\|_{L^2(\Omega)} + \|\Delta f\|_{L^2(\Omega)}\right).
\end{aligned}$$

Noticing that ϕ^ε vanishes on the boundary of Ω^ε and that $\operatorname{div} b \leqslant 0$ on $Y\setminus\overline{\mathcal{O}}$, we obtain

$$\int_{\Omega^\varepsilon}\alpha|\nabla\phi^\varepsilon|^2 \leqslant \int_{\Omega^\varepsilon}\left(-\alpha\Delta\phi^\varepsilon + \widehat{b}^\varepsilon\cdot\nabla\phi^\varepsilon\right)\phi^\varepsilon \leqslant C\sqrt{\varepsilon}\|\phi^\varepsilon\|_{L^2(\Omega^\varepsilon)}\mathcal{N}(f). \tag{4.59}$$

Combining (4.59) and (4.58), we get

$$|\phi^\varepsilon|_{H^1(\Omega^\varepsilon)} \leqslant C\varepsilon^{3/2}\mathcal{N}(f). \tag{4.60}$$

To estimate $\left|u^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right) f\right|_{H^1(\Omega^\varepsilon)}$, we use the triangle inequality and write

$$\left|u^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right) f\right|_{H^1(\Omega^\varepsilon)} \leqslant |\phi^\varepsilon|_{H^1(\Omega^\varepsilon)} + \varepsilon^2\left|w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon)\right|_{H^1(\Omega^\varepsilon)}. \tag{4.61}$$

We are thus left with bounding the quantity $\varepsilon^2\left|w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon)\right|_{H^1(\Omega^\varepsilon)}$. To this aim, we compute

$$\varepsilon^2\nabla\left[w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon)\right] = \varepsilon\nabla w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon) + \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right)\nabla f - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right)\nabla(f\eta^\varepsilon).$$

We then have

$$\begin{aligned}
&\varepsilon^2\left|w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon)\right|_{H^1(\Omega^\varepsilon)}\\
\leqslant\ &\varepsilon\left\|\nabla w\left(\frac{\cdot}{\varepsilon}\right) f(1-\eta^\varepsilon)\right\|_{L^2(\Omega)} + \varepsilon^2\left\|w\left(\frac{\cdot}{\varepsilon}\right)\nabla f\right\|_{L^2(\Omega)} + \varepsilon^2\left\|w\left(\frac{\cdot}{\varepsilon}\right)\nabla(f\eta^\varepsilon)\right\|_{L^2(\Omega)}\\
\leqslant\ &C\varepsilon^{3/2}\mathcal{N}(f).
\end{aligned} \tag{4.62}$$

We infer from (4.61), (4.60) and (4.62) that

$$\left|u^\varepsilon - \varepsilon^2 w\left(\frac{\cdot}{\varepsilon}\right) f\right|_{H^1(\Omega^\varepsilon)} \leqslant C\varepsilon^{3/2}\mathcal{N}(f).$$

This concludes the proof of Theorem 4.2. □

4.5.2 Homogeneous Neumann boundary condition

We prove Theorem 4.4 using two-scale convergence, as in [4, Theorem 2.9]. The problem considered in [4, Theorem 2.9] is a diffusion problem with a zero-order term, while the problem we consider here is an advection-diffusion problem without zero-order term. Although the problems are different, the arguments of the proof are essentially similar. We thus only detail the steps in the argument different from those in [4, Theorem 2.9].

As a preliminary step, before we are in position to prove Theorem 4.4, we establish the following Poincaré inequality. We recall (see (4.11)) that

$$V^\varepsilon = \left\{u \in H^1(\Omega^\varepsilon) \text{ such that } u = 0 \text{ on } \partial\Omega^\varepsilon \cap \partial\Omega\right\}. \tag{4.63}$$

Lemma 4.7. *We assume (4.43) and (4.47). There exists C independent of ε such that*

$$\forall u \in V^\varepsilon, \quad \|u\|_{L^2(\Omega^\varepsilon)} \leqslant C\|\nabla u\|_{L^2(\Omega^\varepsilon)}. \tag{4.64}$$

Proof of Lemma 4.7. The proof of (4.64) falls in two steps.

Step 1. For any $u \in V^\varepsilon$, we show here how to build a suitable extension v of u in the perforations B^ε. We recall that, in view of (4.43), we have $B^\varepsilon = \Omega \cap (\cup_{k\in\mathbb{Z}^d}(\varepsilon\mathcal{O} + \varepsilon k))$.

Let $\overline{u} \in H^1(Y\backslash\overline{\mathcal{O}})$. We claim that there exists $\overline{v} \in H^1(Y)$ with $\overline{v} = \overline{u}$ in $Y\backslash\overline{\mathcal{O}}$ and

$$\|\nabla\overline{v}\|_{L^2(Y)} \leqslant C\|\nabla\overline{u}\|_{L^2(Y\backslash\overline{\mathcal{O}})} \tag{4.65}$$

for some C independent of $\overline{u}$ and $\overline{v}$.

Consider indeed the function $\overline{u} - c$, where $c = \dfrac{1}{|Y\backslash\overline{\mathcal{O}}|}\displaystyle\int_{Y\backslash\overline{\mathcal{O}}} \overline{u}$. This function admits a trace on $\partial\mathcal{O}$ which belongs to $H^{1/2}(\partial\mathcal{O})$. By surjectivity of the trace, there exists $\overline{w} \in H^1(\mathcal{O})$ with $\overline{w} = \overline{u} - c$ on $\partial\mathcal{O}$ and $\|\overline{w}\|_{H^1(\mathcal{O})} \leqslant C_{\rm surj}\|\overline{u} - c\|_{H^{1/2}(\partial\mathcal{O})}$. We then define $\overline{v} \in L^2(Y)$ by $\overline{v} = \overline{u}$ in $Y\backslash\overline{\mathcal{O}}$ and $\overline{v} = \overline{w} + c$ in $\mathcal{O}$. By construction of $\overline{w}$, we have that $\overline{v} \in H^1(Y)$. Furthermore, we compute

$$\begin{aligned}
\|\nabla\overline{v}\|^2_{L^2(Y)} &= \|\nabla\overline{v}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + \|\nabla\overline{v}\|^2_{L^2(\mathcal{O})} \\
&= \|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + \|\nabla\overline{w}\|^2_{L^2(\mathcal{O})} \\
&\leqslant \|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + \|\overline{w}\|^2_{H^1(\mathcal{O})} \\
&\leqslant \|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + C^2_{\rm surj}\|\overline{u} - c\|^2_{H^{1/2}(\partial\mathcal{O})}.
\end{aligned}$$

Using the trace inequality and the Poincaré-Wirtinger inequality in $Y\backslash\overline{\mathcal{O}}$, we deduce from the above that

$$\begin{aligned}
\|\nabla\overline{v}\|^2_{L^2(Y)} &\leqslant \|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + C^2_{\rm surj}C_{\rm trace}\|\overline{u} - c\|^2_{H^1(Y\backslash\overline{\mathcal{O}})} \\
&\leqslant \|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})} + C^2_{\rm surj}C_{\rm trace}C_{\rm PW}\|\nabla\overline{u}\|^2_{L^2(Y\backslash\overline{\mathcal{O}})},
\end{aligned}$$

which concludes the proof of (4.65).

By scaling, we next deduce from (4.65) that there exists C independent of ε such that, for any $u \in H^1(\varepsilon(Y\backslash\overline{\mathcal{O}}))$, there exists $v \in H^1(\varepsilon Y)$ with $v = u$ in $\varepsilon(Y\backslash\overline{\mathcal{O}})$ and

$$\|\nabla v\|_{L^2(\varepsilon Y)} \leqslant C\|\nabla u\|_{L^2(\varepsilon(Y\backslash\overline{\mathcal{O}}))}.$$

Consider now $u \in V^\varepsilon$:

- For each perforation $\varepsilon\mathcal{O} + \varepsilon k$ included in Ω, we extend u in $\varepsilon\mathcal{O} + \varepsilon k$ as above.
- For each perforation $\varepsilon\mathcal{O} + \varepsilon k$ that intersects the boundary $\partial\Omega$, we extend u in $\Omega \cap (\varepsilon\mathcal{O} + \varepsilon k)$ by a function that is equal to u on $\partial(\varepsilon\mathcal{O} + \varepsilon k) \cap \Omega$ and that vanishes on $\partial\Omega$. To build this extension, we use the assumption (4.47) along with the same arguments as above.

Doing so, we thus construct $v \in H^1_0(\Omega)$ such that $v = u$ on Ω^ε and

$$\|\nabla v\|_{L^2(\Omega)} \leqslant C\|\nabla u\|_{L^2(\Omega^\varepsilon)} \tag{4.66}$$

where C is independent of u and v.

Step 2. Let $u \in V^\varepsilon$ and let $v \in H^1_0(\Omega)$ be the extension built in Step 1. Using the Poincaré inequality in $H^1_0(\Omega)$ and (4.66), we write that

$$\|u\|_{L^2(\Omega^\varepsilon)} \leqslant \|v\|_{L^2(\Omega)} \leqslant C\|\nabla v\|_{L^2(\Omega)} \leqslant C\|\nabla u\|_{L^2(\Omega^\varepsilon)}.$$

This concludes the proof of Lemma 4.7. □

Proof of Theorem 4.4. We recall that the variational form of (4.3) is as follows:

$$\text{Find } u^\varepsilon \in V^\varepsilon \text{ such that, for any } v \in V^\varepsilon, \quad a^\varepsilon(u^\varepsilon, v) = \int_{\Omega^\varepsilon} f\, v$$

where V^ε is defined by (4.63) and the bilinear form a^ε is defined (see (4.7); we now make explicit the dependency with respect to ε) by

$$a^\varepsilon(u, v) = \int_{\Omega^\varepsilon} \alpha \nabla u \cdot \nabla v + \int_{\Omega^\varepsilon} \left(\widehat{b}^\varepsilon \cdot \nabla u\right) v. \tag{4.67}$$

The proof falls in 3 steps.

Step 1: a priori estimates. Similarly to $\widetilde{u}^\varepsilon \in L^2(\Omega)$, the extension by zero of the solution u^ε to (4.3), we denote $\widetilde{\nabla} u^\varepsilon \in L^2(\Omega)$ the extension by zero of ∇u^ε. We want to show that the sequences $\widetilde{u}^\varepsilon$ and $\widetilde{\nabla} u^\varepsilon$ are uniformly bounded in $L^2(\Omega)$ with respect to ε. We recall that $\widehat{b}^\varepsilon = b\left(\frac{\cdot}{\varepsilon}\right)$. We distinguish two cases, depending on the assumptions on b.

We start with the case where $\operatorname{div} b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$ and $b \cdot n \geqslant 0$ on $\partial\mathcal{O}$. For any $u \in V^\varepsilon$, we have

$$\begin{aligned} a^\varepsilon(u, u) &= \int_{\Omega^\varepsilon} \alpha |\nabla u|^2 + \frac{1}{2}\int_{\partial\Omega^\varepsilon} \left(\widehat{b}^\varepsilon \cdot n\right) u^2 - \frac{1}{2}\int_{\Omega^\varepsilon} \left(\operatorname{div} \widehat{b}^\varepsilon\right) u^2 \\ &\geqslant \alpha \|\nabla u\|^2_{L^2(\Omega^\varepsilon)} \\ &\geqslant C\|u\|^2_{H^1(\Omega^\varepsilon)}, \end{aligned} \tag{4.68}$$

successively using that $\operatorname{div} b \leqslant 0$ in $Y\backslash\overline{\mathcal{O}}$, $b \cdot n \geqslant 0$ on $\partial\mathcal{O}$, $u = 0$ on $\partial\Omega^\varepsilon \cap \partial\Omega$ and the Poincaré inequality shown in Lemma 4.7. The bilinear form (4.67) is thus coercive, uniformly with respect to ε. This of course implies that the problem (4.3) is well-posed, and also that u^ε is uniformly bounded in $H^1(\Omega^\varepsilon)$.

In the case where b is curl-free, we know there exists $\psi \in W^{2,\infty}_{\text{loc}}(E)$ (where we recall that E is the domain obtained by Y-periodicity from $Y\backslash\overline{\mathcal{O}}$)) such that $b = \alpha\,\nabla\psi$ in E. Denoting by

$\widehat{\psi}^\varepsilon = \varepsilon\psi\left(\frac{x}{\varepsilon}\right)$, we have, for any u and v in $H^1(\Omega^\varepsilon)$,

$$\begin{aligned} & a^\varepsilon\left(u, e^{-\widehat{\psi}^\varepsilon} v\right) \\ &= \int_{\Omega^\varepsilon} \alpha \nabla u \cdot \nabla\left(e^{-\widehat{\psi}^\varepsilon} v\right) + \left(\widehat{b}^\varepsilon \cdot \nabla u\right) e^{-\widehat{\psi}^\varepsilon} v \\ &= \int_{\Omega^\varepsilon} \alpha e^{-\widehat{\psi}^\varepsilon} \nabla u \cdot \nabla v + \int_{\Omega^\varepsilon} \left[\left(\widehat{b}^\varepsilon - \alpha \nabla \widehat{\psi}^\varepsilon\right) \cdot \nabla u\right] e^{-\widehat{\psi}^\varepsilon} v \\ &= \int_{\Omega^\varepsilon} \alpha e^{-\widehat{\psi}^\varepsilon} \nabla u \cdot \nabla v. \end{aligned} \tag{4.69}$$

Writing b as the sum of its mean and a mean-free quantity, we see that ψ can be written as the sum of a linear function and a periodic function, which implies that

$$\widehat{\psi}^\varepsilon(x) = \frac{1}{\alpha}\langle b\rangle \cdot x + \varepsilon\, \psi_{\mathrm{per}}\left(\frac{x}{\varepsilon}\right) \tag{4.70}$$

where $\langle b\rangle$ is the mean of b over Y and ψ_{per} is Y-periodic. There thus exists C_ψ such that

$$\forall \varepsilon, \qquad \left\|\widehat{\psi}^\varepsilon\right\|_{W^{1,\infty}(\Omega^\varepsilon)} \leqslant C_\psi. \tag{4.71}$$

Furthermore, setting $\psi_0(x) = \frac{1}{\alpha}\langle b\rangle \cdot x$, we have

$$\lim_{\varepsilon \to 0} \left\|\widehat{\psi}^\varepsilon - \psi_0\right\|_{L^\infty(\Omega^\varepsilon)} = 0. \tag{4.72}$$

We deduce from (4.69) and (4.71), using again Lemma 4.7, that

$$\forall u \in V^\varepsilon, \qquad a^\varepsilon\left(u, e^{-\widehat{\psi}^\varepsilon} u\right) \geqslant \alpha\, e^{-C_\psi} \|\nabla u\|^2_{L^2(\Omega^\varepsilon)} \geqslant C \|u\|^2_{H^1(\Omega^\varepsilon)}, \tag{4.73}$$

where C is a constant independent of ε. We deduce that there exists C independent of ε such that

$$\inf_{u \in V^\varepsilon} \sup_{v \in V^\varepsilon} \frac{a^\varepsilon(u,v)}{\|u\|_{H^1(\Omega^\varepsilon)} \|v\|_{H^1(\Omega^\varepsilon)}} \geqslant C. \tag{4.74}$$

Again using (4.73), we also observe that, if $v \in V^\varepsilon$ is such that $a^\varepsilon(u,v) = 0$ for any $u \in V^\varepsilon$, then $v = 0$. Using the Banach-Nečas-Babuška Theorem, we conclude that the problem (4.3) is well-posed (we refer to, e.g., [38, p. 85] for a comprehensive exposition of the inf-sup theory). We next deduce from (4.74) that u^ε is uniformly bounded in $H^1(\Omega^\varepsilon)$.

Thus, in both cases we consider, the solution u^ε to (4.3) is well-defined and uniformly bounded in $H^1(\Omega^\varepsilon)$. Consequently, $\widetilde{u}^\varepsilon$ (resp. $\widetilde{\nabla} u^\varepsilon$) is uniformly bounded in $L^2(\Omega)$ (resp. $(L^2(\Omega))^d$).

Step 2: homogenized limit. Because of the above bounds, we know that there exists $u_0 \in L^2(\Omega \times Y)$ (resp. $\xi_0 \in (L^2(\Omega \times Y))^d$) such that the sequence $\widetilde{u}^\varepsilon$ (resp. $\widetilde{\nabla} u^\varepsilon$) two-scale converges, up to an extraction, to u_0 (resp. ξ_0). Following the proof of [4, Theorem 2.9], we obtain that there exists $u^\star \in H^1_0(\Omega)$ and $u_1 \in L^2\left(\Omega, H^1_{\mathrm{per}}(Y\backslash\overline{\mathcal{O}})/\mathbb{R}\right)$ such that

$$u_0(x,y) = u^\star(x)\, \mathbb{1}_{Y\backslash\overline{\mathcal{O}}}(y) \quad \text{and} \quad \xi_0(x,y) = \mathbb{1}_{Y\backslash\overline{\mathcal{O}}}(y)\left(\nabla u^\star(x) + \nabla_y u_1(x,y)\right). \tag{4.75}$$

To identify the equation satisfied by $u^\star$ and u_1, we consider, following [4], the following test function

$$\phi^\varepsilon(x) = \phi(x) + \varepsilon\phi_1\left(x, \frac{x}{\varepsilon}\right),$$

where $\phi \in \mathcal{C}_c^\infty(\Omega)$ and $\phi_1 \in \mathcal{C}_c^\infty(\Omega; \mathcal{C}_{\rm per}^\infty(Y))$, the gradient of which reads

$$\nabla\phi^\varepsilon(x) = \nabla\phi(x) + \nabla_y\phi_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon\nabla\phi_1\left(x, \frac{x}{\varepsilon}\right).$$

The variational formulation (4.67) of (4.3), using $v \equiv \phi^\varepsilon$, writes

$$\alpha \int_\Omega \widetilde{\nabla} u^\varepsilon \cdot \nabla\phi^\varepsilon + \int_\Omega b\left(\frac{\cdot}{\varepsilon}\right) \cdot \widetilde{\nabla} u^\varepsilon \phi^\varepsilon = \int_{\Omega^\varepsilon} f\phi^\varepsilon = \int_\Omega \mathbb{1}_E\left(\frac{\cdot}{\varepsilon}\right) f\phi^\varepsilon, \tag{4.76}$$

where we recall that E is the domain obtained by Y-periodicity from $Y\backslash\overline{\mathcal{O}}$.

We check that the functions $[\nabla\phi(x) + \nabla_y\phi_1(x,y) + \varepsilon\nabla\phi_1(x,y))]$ and $b\,(y)\,(\phi(x) + \varepsilon\phi_1(x,y))$ are admissible test functions in the sense of the two-scale convergence.

Passing to the limit $\varepsilon \to 0$ in (4.76) and using (4.75), we obtain

$$\alpha \int_{\Omega\times Y} \mathbb{1}_{Y\backslash\overline{\mathcal{O}}}(\nabla u^\star + \nabla_y u_1) \cdot (\nabla\phi + \nabla_y\phi_1) + \int_{\Omega\times Y} \mathbb{1}_{Y\backslash\overline{\mathcal{O}}}(\nabla u^\star + \nabla_y u_1) \cdot b\phi = \int_{\Omega\times Y} \mathbb{1}_{Y\backslash\overline{\mathcal{O}}} f\phi.$$

By a density argument, we have that this variational formulation holds for any $(\phi, \phi_1) \in H_0^1(\Omega) \times L^2(\Omega; H_{\rm per}^1(Y))$. We thus have

$$\forall \phi_1 \in H_{\rm per}^1(Y), \quad \int_{Y\backslash\overline{\mathcal{O}}} (\nabla u^\star + \nabla_y u_1) \cdot \nabla_y\phi_1 = 0 \tag{4.77}$$

and

$$\forall \phi \in H_0^1(\Omega), \quad \alpha \int_{\Omega\times(Y\backslash\overline{\mathcal{O}})} (\nabla u^\star + \nabla_y u_1) \cdot \nabla\phi + \int_{\Omega\times(Y\backslash\overline{\mathcal{O}})} (\nabla u^\star + \nabla_y u_1) \cdot b\phi = |Y\backslash\overline{\mathcal{O}}| \int_\Omega f\phi. \tag{4.78}$$

Let w_i, $1 \leqslant i \leqslant d$, be the corrector solution to (4.51). We deduce from (4.77) that

$$u_1(x,y) = \sum_{i=1}^d w_i(y)\, \partial_{x_i} u^\star(x).$$

Inserting this expression in (4.78), we get that $u^\star$ satisfies

$$\forall \phi \in H_0^1(\Omega), \quad \int_\Omega A^\star \nabla u^\star \cdot \nabla\phi + (b^\star \cdot \nabla u^\star)\,\phi = \frac{|Y\backslash\overline{\mathcal{O}}|}{|Y|} \int_\Omega f\phi,$$

where $A^\star$ (resp. $b^\star$) is defined by (4.49) (resp. (4.49)). This is exactly the variational formulation of (4.48).

Since the couple $(u^\star, u_1)$ is uniquely determined, we infer that the whole sequence $\widetilde{u}^\varepsilon$ (resp. $\widetilde{\nabla} u^\varepsilon$) two-scale converges to u_0 (resp. to ξ_0) in $L^2(\Omega)$ (resp. in $(L^2(\Omega))^d$).

Step 3: H^1 convergence. Let $u^{\varepsilon,1} = u^\star + \varepsilon \sum_{i=1}^d w_i\left(\frac{\cdot}{\varepsilon}\right)\partial_{x_i}u^\star$ and $\xi_1(x,y) = \sum_{i=1}^d (e_i + \nabla w_i(y))\,\partial_{x_i}u^\star(x)$, so that $\nabla u^{\varepsilon,1} = \xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right) + \varepsilon\,\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)$ with $\xi_2(x,y) = \sum_{i=1}^d w_i(y)\,\partial_{x_i}\nabla u^\star(x)$. We note that $\xi_0(x,y) = \mathbb{1}_{Y\setminus\overline{\mathcal{O}}}(y)\,\xi_1(x,y)$.

We note that $u^{\varepsilon,1}$ does not vanish on $\partial\Omega$. This is a usual difficulty in homogenization, which is standardly addressed by introducing a truncation function η^ε as in the proof of Theorem 4.2 (see Appendix 4.5.1), and considering $g^{\varepsilon,1} = u^\star + \varepsilon\eta^\varepsilon \sum_{i=1}^d w_i\left(\frac{\cdot}{\varepsilon}\right)\partial_{x_i}u^\star \in V^\varepsilon$. Under sufficient regularity assumptions on $\mathcal{O}$ and f, we have $w_i \in W^{1,\infty}(Y\setminus\overline{\mathcal{O}})$ and $u^\star \in W^{2,\infty}(\Omega)$, hence

$$\lim_{\varepsilon\to 0} \|u^{\varepsilon,1} - g^{\varepsilon,1}\|_{H^1(\Omega^\varepsilon)} = 0. \tag{4.79}$$

To estimate $\|u^\varepsilon - g^{\varepsilon,1}\|_{H^1(\Omega^\varepsilon)}$, we distinguish two cases, depending on the assumptions on b.

We start with the case where $\operatorname{div} b \leqslant 0$ in $Y\setminus\overline{\mathcal{O}}$ and $b\cdot n = 0$ on $\partial\mathcal{O}$. Using (4.68), we write

$$\begin{aligned} C\|u^\varepsilon - g^{\varepsilon,1}\|^2_{H^1(\Omega^\varepsilon)} &\leqslant a^\varepsilon(u^\varepsilon - g^{\varepsilon,1}, u^\varepsilon - g^{\varepsilon,1}) \\ &= \int_{\Omega^\varepsilon} f(u^\varepsilon - g^{\varepsilon,1}) + a^\varepsilon(g^{\varepsilon,1}, g^{\varepsilon,1}) - a^\varepsilon(g^{\varepsilon,1}, u^\varepsilon) \\ &= \int_{\Omega^\varepsilon} f(u^\varepsilon - u^{\varepsilon,1}) + a^\varepsilon(u^{\varepsilon,1}, u^{\varepsilon,1}) - a^\varepsilon(u^{\varepsilon,1}, u^\varepsilon) + R_\varepsilon, \end{aligned} \tag{4.80}$$

where

$$\lim_{\varepsilon\to 0} R_\varepsilon = 0 \tag{4.81}$$

as a consequence of (4.79).

We successively pass to the limit in the three terms of (4.80). For the first one, we have

$$\int_{\Omega^\varepsilon} f u^\varepsilon = \int_\Omega f\widetilde{u}^\varepsilon \to_{\varepsilon\to 0} \int_{\Omega\times Y} f u_0 = |Y\setminus\overline{\mathcal{O}}| \int_\Omega f u \tag{4.82}$$

and

$$\int_{\Omega^\varepsilon} f u^{\varepsilon,1} = \int_\Omega \mathbb{1}_E\left(\frac{\cdot}{\varepsilon}\right) f u^{\varepsilon,1} \to_{\varepsilon\to 0} \int_{\Omega\times Y} \mathbb{1}_{Y\setminus\overline{\mathcal{O}}} f u = |Y\setminus\overline{\mathcal{O}}| \int_\Omega f u. \tag{4.83}$$

For the second term of (4.80), we write

$$\begin{aligned} & a^\varepsilon(u^{\varepsilon,1}, u^{\varepsilon,1}) \\ = & \int_{\Omega^\varepsilon} \alpha\nabla u^{\varepsilon,1}\cdot\nabla u^{\varepsilon,1} + \int_{\Omega^\varepsilon}\left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla u^{\varepsilon,1}\right] u^{\varepsilon,1} \\ = & \int_{\Omega^\varepsilon} \alpha\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right) + 2\varepsilon\int_{\Omega^\varepsilon}\alpha\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right) + \varepsilon^2\int_{\Omega^\varepsilon}\alpha\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right) \\ & + \int_{\Omega^\varepsilon}\left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right] u^{\varepsilon,1} + \varepsilon\int_{\Omega^\varepsilon}\left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right] u^{\varepsilon,1} \\ = & \int_\Omega \alpha\xi_0\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_0\left(\cdot,\frac{\cdot}{\varepsilon}\right) + 2\varepsilon\int_{\Omega^\varepsilon}\alpha\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right) + \varepsilon^2\int_{\Omega^\varepsilon}\alpha\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right) \\ & + \int_\Omega\left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_0\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right] u^{\varepsilon,1} + \varepsilon\int_{\Omega^\varepsilon}\left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right] u^{\varepsilon,1}. \end{aligned}$$

Passing to the limit $\varepsilon \to 0$, we get that

$$\lim_{\varepsilon\to 0} a^\varepsilon(u^{\varepsilon,1}, u^{\varepsilon,1}) = \int_{\Omega\times Y} \alpha\xi_0(x,y)\cdot\xi_0(x,y) + \int_{\Omega\times Y} \left[b(y)\cdot\xi_0(x,y)\right] u(x). \quad (4.84)$$

For the third term of (4.80), we write

$$\begin{aligned} & a^\varepsilon(u^{\varepsilon,1}, u^\varepsilon) \\ = & \int_{\Omega^\varepsilon} \alpha\nabla u^{\varepsilon,1}\cdot\nabla u^\varepsilon + \int_{\Omega^\varepsilon} \left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\nabla u^{\varepsilon,1}\right] u^\varepsilon \\ = & \int_\Omega \alpha\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\widetilde{\nabla} u^\varepsilon + \varepsilon\int_\Omega \alpha\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\cdot\widetilde{\nabla} u^\varepsilon \\ & + \int_\Omega \left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_1\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right]\widetilde{u}^\varepsilon + \varepsilon\int_\Omega \left[b\left(\frac{\cdot}{\varepsilon}\right)\cdot\xi_2\left(\cdot,\frac{\cdot}{\varepsilon}\right)\right]\widetilde{u}^\varepsilon. \end{aligned}$$

Passing to the limit $\varepsilon \to 0$ and using the two-scale limits of $\widetilde{u}^\varepsilon$ and $\widetilde{\nabla} u^\varepsilon$, we get that

$$\begin{aligned} & \lim_{\varepsilon\to 0} a^\varepsilon(u^{\varepsilon,1}, u^\varepsilon) \\ = & \int_{\Omega\times Y} \alpha\xi_1(x,y)\cdot\xi_0(x,y) + \int_{\Omega\times Y} \left[b(y)\cdot\xi_1(x,y)\right] u_0(x) \\ = & \int_{\Omega\times Y} \alpha\xi_0(x,y)\cdot\xi_0(x,y) + \int_{\Omega\times Y} \left[b(y)\cdot\xi_0(x,y)\right] u(x). \end{aligned} \quad (4.85)$$

Collecting (4.80), (4.81), (4.82), (4.83), (4.84) and (4.85), we deduce that

$$\lim_{\varepsilon\to 0} C\|u^\varepsilon - g^{\varepsilon,1}\|_{H^1(\Omega^\varepsilon)} = 0.$$

Collecting this result with (4.79), we deduce the claimed H^1 convergence.

In the case where b is curl-free, using (4.73), we write

$$\begin{aligned} & C\|u^\varepsilon - g^{\varepsilon,1}\|^2_{H^1(\Omega^\varepsilon)} \\ & \leqslant a^\varepsilon\left(u^\varepsilon - g^{\varepsilon,1}, e^{-\widetilde{\psi}^\varepsilon}(u^\varepsilon - g^{\varepsilon,1})\right) \\ & = \int_{\Omega^\varepsilon} f e^{-\widetilde{\psi}^\varepsilon}(u^\varepsilon - g^{\varepsilon,1}) + a^\varepsilon\left(g^{\varepsilon,1}, e^{-\widetilde{\psi}^\varepsilon} g^{\varepsilon,1}\right) - a^\varepsilon\left(g^{\varepsilon,1}, e^{-\widetilde{\psi}^\varepsilon} u^\varepsilon\right) \\ & = \int_{\Omega^\varepsilon} f e^{-\widetilde{\psi}^\varepsilon}(u^\varepsilon - u^{\varepsilon,1}) + a^\varepsilon\left(u^{\varepsilon,1}, e^{-\widetilde{\psi}^\varepsilon} u^{\varepsilon,1}\right) - a^\varepsilon\left(u^{\varepsilon,1}, e^{-\widetilde{\psi}^\varepsilon} u^\varepsilon\right) + R_\varepsilon, \end{aligned} \quad (4.86)$$

where

$$\lim_{\varepsilon\to 0} R_\varepsilon = 0 \quad (4.87)$$

as a consequence of (4.79) and (4.71).

We successively pass to the limit in the three terms of (4.86). For the first one, we use (4.72). Arguing as in (4.82) and (4.83), we get that

$$\lim_{\varepsilon\to 0} \int_{\Omega^\varepsilon} f e^{-\widetilde{\psi}^\varepsilon}(u^\varepsilon - u^{\varepsilon,1}) = 0. \quad (4.88)$$

For the second term of (4.86), we write

$$\begin{aligned} a^\varepsilon\left(u^{\varepsilon,1}, e^{-\tilde{\psi}^\varepsilon}u^{\varepsilon,1}\right) &= \int_{\Omega^\varepsilon} \alpha e^{-\tilde{\psi}^\varepsilon}\left|\nabla u^{\varepsilon,1}\right|^2 \qquad \text{[Eq. (4.69)]} \\ &= \int_{\Omega^\varepsilon} \alpha e^{-\psi_0}\left|\nabla u^{\varepsilon,1}\right|^2 + o(1). \qquad \text{[Eq. (4.72)]} \end{aligned}$$

Arguing as in (4.84), we get

$$\lim_{\varepsilon\to 0} a^\varepsilon\left(u^{\varepsilon,1}, e^{-\tilde{\psi}^\varepsilon}u^{\varepsilon,1}\right) = \int_{\Omega\times Y} \alpha e^{-\psi_0}\xi_0(x,y)\cdot\xi_0(x,y). \tag{4.89}$$

For the third term of (4.86), we write

$$\begin{aligned} a^\varepsilon\left(u^{\varepsilon,1}, e^{-\tilde{\psi}^\varepsilon}u^{\varepsilon}\right) &= \int_{\Omega^\varepsilon} \alpha e^{-\tilde{\psi}^\varepsilon}\nabla u^{\varepsilon,1}\cdot\nabla u^{\varepsilon} \qquad \text{[Eq. (4.69)]} \\ &= \int_{\Omega^\varepsilon} \alpha e^{-\psi_0}\nabla u^{\varepsilon,1}\cdot\nabla u^{\varepsilon} + o(1). \qquad \text{[Eq. (4.72)]} \end{aligned}$$

Arguing as in (4.85), we get

$$\lim_{\varepsilon\to 0} a^\varepsilon\left(u^{\varepsilon,1}, e^{-\tilde{\psi}^\varepsilon}u^{\varepsilon,1}\right) = \int_{\Omega\times Y} \alpha e^{-\psi_0}\xi_0(x,y)\cdot\xi_0(x,y). \tag{4.90}$$

Collecting (4.86), (4.87), (4.88), (4.89) and (4.90), we deduce that

$$\lim_{\varepsilon\to 0} C\|u^\varepsilon - g^{\varepsilon,1}\|_{H^1(\Omega^\varepsilon)} = 0.$$

Collecting this result with (4.79), we deduce the claimed H^1 convergence. This concludes the proof of Theorem 4.4. □

4.6 Appendix: Proof of the error estimates

4.6.1 Some preliminary material

Before being in position to present the proofs of Theorem 4.5 and Theorem 4.6, which follow that of [59, Theorem 2.2], we need some preliminary results.

To start with, we establish a Poincaré-type inequality. As we clearly have

$$\|u\|_{L^2(\Omega^\varepsilon)} \leqslant C\varepsilon\|\nabla u\|_{L^2(\Omega^\varepsilon)} \quad \text{for all} \quad u \in H_0^1(\Omega^\varepsilon),$$

where C is a constant independent of ε, we also have

$$\|u\|_{L^2(\Omega^\varepsilon)} \leqslant C_P\,\varepsilon|u|_{H_H^1(\Omega^\varepsilon)} \quad \text{for all} \quad u \in W_{H,B^\varepsilon}, \tag{4.91}$$

where C_P is a constant independent of ε, where $|u|^2_{H_H^1(\Omega^\varepsilon)} = \sum_{K\in\mathcal{T}_H} \|\nabla u\|^2_{L^2(K\cap\Omega^\varepsilon)}$ and

$$W_{H,B^\varepsilon} = \left\{ \begin{array}{c} u \in L^2(\Omega) \text{ such that } u|_K \in H^1(K) \text{ for any } K \in \mathcal{T}_H, \\ \displaystyle\int_E [[u]] = 0 \text{ for any } E \in \mathcal{E}_H^{\text{in}}, \quad u = 0 \text{ in } B^\varepsilon \end{array} \right\}.$$

We now recall three lemmas, borrowed from [59, Section 3]. The first lemma is a trace inequality. For any domain $\omega \subset \mathbb{R}^d$, it is classical to define the space

$$H^{1/2}(\omega) = \left\{ u \in L^2(\omega), \quad \int_\omega \int_\omega \frac{|u(x)-u(y)|^2}{|x-y|^{d+1}} dxdy < +\infty \right\},$$

and the norm

$$\|u\|^2_{H^{1/2}(\omega)} = \|u\|^2_{L^2(\omega)} + |u|^2_{H^{1/2}(\omega)}$$

where

$$|u|^2_{H^{1/2}(\omega)} = \left(\int_\omega \int_\omega \frac{|u(x)-u(y)|^2}{|x-y|^{d+1}} dxdy \right)^{1/2}.$$

Lemma 4.8 (Lemma 3.2 of [59]). *There exists C (depending on the regularity of the mesh) such that, for any $K \in \mathcal{T}_H$ and any edge $E \subset \partial K$, we have*

$$\|v\|_{L^2(E)} \leqslant C \left(H^{-1} \|v\|^2_{L^2(K)} + H \|\nabla v\|^2_{L^2(K)} \right) \quad \text{for all} \quad v \in H^1(K). \tag{4.92}$$

Under the additional assumption that $\int_E v = 0$, we have

$$\|v\|^2_{L^2(E)} \leqslant CH \|\nabla v\|^2_{L^2(K)} \tag{4.93}$$

and

$$\|v\|^2_{H^{1/2}(E)} \leqslant C(1+H) \|\nabla v\|^2_{L^2(K)}. \tag{4.94}$$

Lemma 4.9 (Corollary 3.3 of [59]). *Consider an edge $E \in \mathcal{E}_H^{in}$, and let $K_E \subset \mathcal{T}_H$ denote all the triangles sharing this edge. There exists C (depending only on the regularity of the mesh) such that*

$$\| [[v]] \|_{L^2(E)} \leqslant CH \sum_{K \in K_E} \|\nabla v\|^2_{L^2(K)} \quad \text{for all} \quad v \in W_H, \tag{4.95}$$

and

$$\| [[v]] \|_{L^2(E)} \leqslant C(1+H) \sum_{K \in K_E} \|\nabla v\|^2_{L^2(K)} \quad \text{for all} \quad v \in W_H. \tag{4.96}$$

Lemma 4.10 (Lemma 3.4 of [59]). *Let $g \in L^\infty(\mathbb{R})$ be a q-periodic function with zero mean. Let $f \in W^{1,1}(0,H) \subset C^0(0,H)$ be a function defined on the interval $[0,H]$ that vanishes at least at one point of $[0,H]$. Then, for any $\varepsilon > 0$,*

$$\left| \int_0^H g\left(\frac{x}{\varepsilon}\right) f(x) dx \right| \leqslant 2\, \varepsilon\, q\, \|g\|_{L^\infty(\mathbb{R})}\, \|f'\|_{L^1(0,H)}.$$

4.6.2 Proof of Theorem 4.5

To simplify notation, we denote by u, instead of u^ε, the solution to (4.2), and u_H, instead of u_H^ε its approximation, solution to (4.25).

Upon dividing Equation (4.2) by α, we may assume that $\alpha = 1$. Let $\Pi_H f$ be the L^2-orthogonal projection of f on the space of piecewise constant functions. We recall that we have the following

interpolation estimate: there exists C independent of H and f such that

$$\|f - \Pi_H f\|_{L^2(\Omega)} \leqslant CH\|\nabla f\|_{L^2(\Omega)} \tag{4.97}$$

We define

$$v_H(x) = \sum_{K\in\mathcal{T}_H} \Pi_H f \ \Psi^{\varepsilon,K}(x) + \sum_{E\in\mathcal{E}_H^{\text{in}}} \left(\int_E u\right) \Phi^{\varepsilon,E}(x),$$

where $\Phi^{\varepsilon,E}$ and $\Psi^{\varepsilon,K}$ are respectively solutions to (4.29) and (4.30). We recall that if $K \subset B^\varepsilon$ (resp. $E \subset B^\varepsilon$), then $\Psi^{\varepsilon,K} \equiv 0$ (resp. $\Phi^{\varepsilon,E} \equiv 0$). We see from (4.26) that $v_H \in V_H^{\text{adv bubble}}$. We next decompose the exact solution into

$$u = v_H + \phi.$$

Notice that v_H satisfies

$$\begin{aligned}
&\int_E v_H = \int_E u, \quad \text{hence} \int_E \phi = 0, \quad \text{for all} \ \ E \in \mathcal{E}_H^{\text{in}} \\
&\left(\nabla v_H - \frac{1}{2}\widehat{b}^\varepsilon v_H\right)\cdot n = \text{Constant on (each side of) E}, \quad \text{for all} \ \ E \in \mathcal{E}_H^{\text{in}} \\
&-\Delta v_H + \widehat{b}^\varepsilon \cdot \nabla v_H = \Pi_H f \quad \text{on } K \cap \Omega^\varepsilon, \text{for all} \ \ K \in \mathcal{T}_H.
\end{aligned} \tag{4.98}$$

In what follows, we make use of the notation $g_{|\varepsilon}(x) = g\left(\frac{x}{\varepsilon}\right)$. The constant C denote a constant independent of ε, H and f and that may vary from one line to the next. We begin by estimating ϕ.

Step 1: Estimation of $u - v_H$: Using the approximation of u given in the homogenization result (4.44) and denoting $\widetilde{\phi} = \varepsilon^2 w^\varepsilon f - v_H$, we have

$$\begin{aligned}
c_H(\phi,\phi) &= c_H(u - \varepsilon^2 w^\varepsilon f, \phi) + c_H(\varepsilon^2 w^\varepsilon f - v_H, \phi) \qquad (4.99)\\
&= c_H(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\Delta\widetilde{\phi} + \widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi + \int_{\partial(K\cap\Omega^\varepsilon)} \left(\nabla\widetilde{\phi} - \frac{1}{2}\widehat{b}^\varepsilon\widetilde{\phi}\right)\cdot n\phi \\
&= c_H(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\Delta\widetilde{\phi} + \widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi + \int_{(\partial K)\cap\Omega^\varepsilon} \left(\nabla\widetilde{\phi} - \frac{1}{2}\widehat{b}^\varepsilon\widetilde{\phi}\right)\cdot n\phi \\
&= c_H(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\Delta\widetilde{\phi} + \widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi \\
&\quad + \varepsilon^2 \sum_{K\in\mathcal{T}_H} \int_{(\partial K)\cap\Omega^\varepsilon} \left(\nabla(w^\varepsilon f) - \frac{1}{2}\widehat{b}^\varepsilon(w^\varepsilon f)\right)\cdot n\phi, \qquad (4.100)
\end{aligned}$$

where we used the fact that $\phi = 0$ on $\partial\Omega^\varepsilon$ in the third line and (4.98) in the last line. Using that $\operatorname{div} b \leqslant 0$, we obtain the coercivity of c_H as follows

$$\begin{aligned} c_H(\phi,\phi) &= \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} |\nabla\phi|^2 + (\widehat{b}^\varepsilon\cdot\nabla\phi)\phi - \frac{1}{2}\int_{\partial(K\cap\Omega^\varepsilon)} (\widehat{b}^\varepsilon\cdot n)\phi^2 \\ &= \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} |\nabla\phi|^2 + \frac{1}{2}\widehat{b}^\varepsilon\cdot\nabla(\phi^2) - \frac{1}{2}\int_{\partial(K\cap\Omega^\varepsilon)} (\widehat{b}^\varepsilon\cdot n)\phi^2 \\ &= \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} |\nabla\phi|^2 - \frac{1}{2}(\operatorname{div}\,\widehat{b}^\varepsilon)\phi^2 \\ &\geqslant |\phi|^2_{H^1_H(\Omega^\varepsilon)}. \end{aligned} \tag{4.101}$$

Combining (4.100) and (4.101), we get

$$\begin{aligned} |\phi|^2_{H^1_H(\Omega^\varepsilon)} \leqslant c_H(u-\varepsilon^2 w^\varepsilon f,\phi) &+ \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \left(-\Delta\widetilde{\phi} + \widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi \\ &+ \varepsilon^2 \sum_{E\in\mathcal{E}_H^{\mathrm{in}}}\int_{E\cap\Omega^\varepsilon} \left(\nabla(w^\varepsilon f) - \frac{1}{2}\widehat{b}^\varepsilon(w^\varepsilon f)\right)\cdot n[[\phi]]. \end{aligned} \tag{4.102}$$

We bound successively the three terms of the right-hand side of (4.102). Roughly speaking:

- the first term is small because of the homogenization result (4.44);
- the second is small because at the leading order $-\Delta v_H + \widehat{b}^\varepsilon\cdot\nabla v_H \simeq \Pi_H f$;
- bounding the third term uses the fact that w is a periodic function.

Step 1a The first term of (4.102) is estimated as follows. Denoting $\widetilde{u} = u - \varepsilon^2 w^\varepsilon f$, we write

$$\begin{aligned} c_H(\widetilde{u},\phi) &= \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \nabla\widetilde{u}\cdot\nabla\phi + \frac{1}{2}(\widehat{b}^\varepsilon\cdot\nabla\widetilde{u})\phi - \frac{1}{2}\operatorname{div}\,(\widehat{b}^\varepsilon\phi)\widetilde{u} \\ &= \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \nabla\widetilde{u}\cdot\nabla\phi \\ &\quad + \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \frac{1}{2}(\widehat{b}^\varepsilon\cdot\nabla\widetilde{u})\phi - \frac{1}{2}(\widehat{b}^\varepsilon\cdot\nabla\phi)\widetilde{u} \\ &\quad - \sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon} \frac{1}{2}\operatorname{div}\,(\widehat{b}^\varepsilon)\phi\widetilde{u} \\ &\leqslant |\widetilde{u}|_{H^1_H(\Omega^\varepsilon)}|\phi|_{H^1_H(\Omega^\varepsilon)} \\ &\quad + \varepsilon^{-1}\|b\|_{L^\infty(Y\setminus\overline{\mathcal{O}})}(|\widetilde{u}|_{H^1_H(\Omega^\varepsilon)}\|\phi\|_{L^2(\Omega^\varepsilon)} + |\phi|_{H^1_H(\Omega^\varepsilon)}\|\widetilde{u}\|_{L^2(\Omega^\varepsilon)}) \\ &\quad + \varepsilon^{-2}\|b\|_{W^{1,\infty}(Y\setminus\overline{\mathcal{O}})}\|\widetilde{u}\|_{L^2(\Omega^\varepsilon)}\|\phi\|_{L^2(\Omega^\varepsilon)} \\ &\leqslant (1 + C\|b\|_{W^{1,\infty}(Y\setminus\overline{\mathcal{O}})})|\widetilde{u}|_{H^1_H(\Omega^\varepsilon)}|\phi|_{H^1_H(\Omega^\varepsilon)} \\ &\leqslant C\varepsilon^{3/2}\mathcal{N}(f)(1 + C\|b\|_{W^{1,\infty}(Y\setminus\overline{\mathcal{O}})})|\phi|_{H^1_H(\Omega^\varepsilon)}, \end{aligned} \tag{4.103}$$

where we used (4.91) in the fourth line and (4.44) in the last line.

Step 1b We now turn to the second term of the right-hand side of (4.102). Using the corrector equation (4.46) and (4.98), similarly as in the proof of Theorem 4.2, we have

$$\begin{aligned}
&\left|\sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon}\left(-\Delta\widetilde{\phi}+\widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi\right|\\
&\leqslant\big(\|f-\Pi_H f\|_{L^2(\Omega)}+2\varepsilon\|\nabla w\|_{L^\infty(\Omega^\varepsilon)}\|\nabla f\|_{L^2(\Omega)}\\
&\qquad+\varepsilon^2\|w\|_{L^\infty(\Omega^\varepsilon)}\|\Delta f\|_{L^2(\Omega)}+\varepsilon\|b\|_{L^\infty(Y\setminus\overline{\mathcal{O}})}\|w\|_{L^\infty(\Omega^\varepsilon)}\|\nabla f\|_{L^2(\Omega)}\big)\|\phi\|_{L^2(\Omega^\varepsilon)}\\
&\leqslant C\varepsilon(CH\|\nabla f\|_{L^2(\Omega)}+C\varepsilon\mathcal{N}(f))|\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}$$

where $\mathcal{N}(f)$ is defined by (4.45), and we used (4.91), (4.97) and the fact that

$$w\in C^1(\overline{Y\setminus\mathcal{O}}) \tag{4.104}$$

in the last line. We infer that

$$\left|\sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon}\left(-\Delta\widetilde{\phi}+\widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi\right|\leqslant C\varepsilon(H+\varepsilon)\mathcal{N}(f)|\phi|_{H^1_H(\Omega^\varepsilon)}. \tag{4.105}$$

Step 1c We then deal with the third term of the right-hand side of (4.102). In view of the assumptions on the mesh, we first observe that, for any edge $E\in\mathcal{E}_H^{\rm in}$, the function $x\in E\to n\cdot\left(\nabla w-\frac{1}{2}bw\right)\left(\frac{x}{\varepsilon}\right)$ is periodic with period $q_E\varepsilon$, for some $q_E\in\mathbb{N}^\star$ satisfying $|q_E|\leqslant C$ for some constant C independent of the mesh edges and of H. We denote $\left\langle\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E$ the average of that function over one period, and decompose the third term of the right-hand side of (4.102) as follows:

$$\begin{aligned}
&\varepsilon^2\sum_{E\in\mathcal{E}_H^{\rm in}}\int_{E\cap\Omega^\varepsilon}\left(\nabla(w^\varepsilon f)-\frac{1}{2}\widehat{b}^\varepsilon(w^\varepsilon f)\right)\cdot n[[\phi]]\\
&=\varepsilon\sum_{E\in\mathcal{E}_H^{\rm in}}\int_{E\cap\Omega^\varepsilon}\left(\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n-\left\langle\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E\right)f[[\phi]]\\
&\quad+\varepsilon\sum_{E\in\mathcal{E}_H^{\rm in}}\left\langle\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E\int_{E\cap\Omega^\varepsilon}f[[\phi]]\\
&\quad+\varepsilon^2\sum_{E\in\mathcal{E}_H^{\rm in}}\int_{E\cap\Omega^\varepsilon}(w^\varepsilon\nabla f\cdot n)\,[[\phi]].
\end{aligned}\tag{4.106}$$

We will extend the function $\phi=u-v_H$ by 0 inside the perforations B^ε, so that we can understand ϕ either as a function of $H^1_0(\Omega)$ or $H^1_0(\Omega^\varepsilon)$.

We consider the first term of the right-hand side of (4.106), which we evaluate essentially using the fact that it contains a periodic oscillatory function of zero mean. We claim that

$$\begin{aligned}
&\left|\int_{E\cap\Omega^\varepsilon}\left(\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n-\left\langle\left(\nabla w-\frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E\right)f[[\phi]]\right|\\
&\leqslant C\sqrt{\varepsilon}\|f\|_{H^1(E)}\|[[\phi]]\|_{H^{1/2}(E)},
\end{aligned}\tag{4.107}$$

where C is a constant independent of the edge E, ε and H. Indeed, we first note that ϕ vanishes on $E \cap B^\varepsilon$, hence

$$\begin{aligned}&\int_{E\cap\Omega^\varepsilon} \left(\left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n - \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \right) f[[\phi]] \\ &= \int_E \left(\left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n - \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \right) f[[\phi]]. \end{aligned} \tag{4.108}$$

Using the regularity (4.104) of w, we indeed have that

$$\begin{aligned}&\left| \int_E \left(\left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n - \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \right) f[[\phi]] \right| \\ &\leqslant C \|f\|_{L^2(E)} \|[[\phi]]\|_{L^2(E)}. \end{aligned} \tag{4.109}$$

Third, suppose momentarily that $[[\phi]] \in H^1(E) \subset C^0(E)$. We infer from the fact that $\displaystyle\int_E [[\phi]] = 0$ that $[[\phi]]$, and hence $[[\phi]]f$, vanishes at least at one point on E. In addition, the function

$$\left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n - \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E$$

is periodic on E (with a period q_E uniformly bounded with respect to $E \in \mathcal{E}_H^{\mathrm{in}}$) and of zero mean. We are then in position to apply Lemma 4.10, which yields, using (4.104),

$$\begin{aligned}&\left| \int_E \left(\left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n - \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \right) f[[\phi]] \right| \\ &\leqslant 4\varepsilon q_E C \left\| \nabla w - \frac{1}{2} bw \right\|_{L^\infty} \|\nabla_E (f[[\phi]])\|_{L^2(E)}, \\ &\leqslant C\varepsilon \|f\|_{H^1(E)} \|[[\phi]]\|_{H^1(E)} \end{aligned} \tag{4.110}$$

where, for any function g, $\nabla_E g = t_E \cdot \nabla g$ where t_E is a unit tangential vector to the edge E. By interpolation between (4.109) and (4.110), and using (4.108), we infer (4.107), with a constant C (independent of the edge) which is independent from ε and H by scaling arguments (see [58] for details).

We deduce from (4.107) that the first term of the right-hand side of (4.106) satisfies

$$\begin{aligned}
&\left|\varepsilon \sum_{E\in\mathcal{E}_H^{\rm in}} \int_{E\cap\Omega^\varepsilon} \left(\left(\nabla w - \frac{1}{2}bw\right)_{|\varepsilon}\cdot n - \left\langle \left(\nabla w - \frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E\right) f[[\phi]]\right| \\
&\leqslant C\varepsilon^{3/2} \sum_{E\in\mathcal{E}_H^{\rm in}} \|f\|_{H^1(E)} \|[[\phi]]\|_{H^{1/2}(E)} \\
&\leqslant C\varepsilon^{3/2} \left(\sum_{E\in\mathcal{E}_H^{\rm in}} \|f\|^2_{H^1(E)}\right)^{1/2} \left(\sum_{E\in\mathcal{E}_H^{\rm in}} \|[[\phi]]\|^2_{H^{1/2}(E)}\right)^{1/2} \\
&\leqslant C\varepsilon^{3/2} \left(\sum_{E\in\mathcal{E}_H^{\rm in};\text{choose one } K\in K_E} \frac{1}{H}\|f\|^2_{H^1(T)} + H\|\nabla f\|^2_{H^1(T)}\right)^{1/2} \times \\
&\quad \left(\sum_{E\in\mathcal{E}_H^{\rm in}} \sum_{K\in K_E} \|\nabla\phi\|^2_{H^1(T)}\right)^{1/2}
\end{aligned}$$

where we have used (4.92) of Lemma 4.8, and (4.96) of Lemma 4.9 (and, we recall that K_E denotes the set of triangles sharing the edge E). We therefore obtain that the first term of the right-hand side of (4.106) satisfies

$$\begin{aligned}
&\left|\varepsilon \sum_{E\in\mathcal{E}_H^{\rm in}} \int_{E\cap\Omega^\varepsilon} \left(\left(\nabla w - \frac{1}{2}bw\right)_{|\varepsilon}\cdot n - \left\langle \left(\nabla w - \frac{1}{2}bw\right)_{|\varepsilon}\cdot n\right\rangle_E\right) f[[\phi]]\right| \\
&\leqslant C\varepsilon^{3/2} \left(\frac{1}{H}\|f\|^2_{H^1(\Omega)} + H\|\nabla f\|^2_{H^1(\Omega)}\right)^{1/2} |\phi|_{H^1_H(\Omega^\varepsilon)} \\
&\leqslant C\varepsilon \left(\sqrt{\frac{\varepsilon}{H}}\|f\|^2_{H^1(\Omega)} + \sqrt{\varepsilon H}\|\nabla f\|^2_{H^1(\Omega)}\right) |\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned} \tag{4.111}$$

The second term of the right-hand side of (4.106) has no oscillatory character. This is why it is estimated using standard arguments for Crouzeix-Raviart finite elements (using that $\displaystyle\int_{E\cap\Omega^\varepsilon} [[\phi]] = 0$), and the regularity of w. Introducing, for each edge E, the constant $c_E = |E|^{-1}\displaystyle\int_E f$, we

bound the second term of the right-hand side of (4.106) by

$$
\begin{aligned}
&\left|\varepsilon \sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \int_{E\cap\Omega^\varepsilon} f[[\phi]]\right| \\
&= \left|\varepsilon \sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \left\langle \left(\nabla w - \frac{1}{2} bw\right)_{|\varepsilon} \cdot n \right\rangle_E \int_{E\cap\Omega^\varepsilon} (f-c_E)[[\phi]]\right| \\
&\leqslant C\varepsilon \sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \|[[\phi]]\|_{L^2(E)} \|f-c_E\|_{L^2(E)} \\
&\leqslant C\varepsilon \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \|[[\phi]]\|^2_{L^2(E)}\right)^{1/2} \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \|f-c_E\|^2_{L^2(E)}\right)^{1/2} \\
&\leqslant C\varepsilon \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \sum_{K\in K_E} H\|\nabla\phi\|^2_{L^2(K)}\right)^{1/2} \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}};\text{choose one } K\in K_E} \|f-c_E\|^2_{L^2(K)}\right)^{1/2} \\
&\leqslant C\varepsilon H|\phi|_{H^1_H(\Omega^\varepsilon)} \|f\|_{L^2(\Omega)},
\end{aligned}
\tag{4.112}
$$

where we have used (4.104), (4.95) of Lemma 4.9 and (4.93) of Lemma 4.8.

We are now left with the third term of the right-hand side of (4.106). This term has a prefactor ε^2 and all we have to prove is that the term itself is bounded. Using again (4.104), (4.95) of Lemma 4.9 and (4.92) of Lemma 4.8, we obtain

$$
\begin{aligned}
&\left|\varepsilon^2 \sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \int_{E\cap\Omega^\varepsilon} (w^\varepsilon \nabla f \cdot n)\,[[\phi]]\right| \\
&\leqslant C\varepsilon^2 \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \|[[\phi]]\|^2_{L^2(E)}\right)^{1/2} \left(\sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \|\nabla f\|^2_{L^2(E)}\right)^{1/2} \\
&\leqslant C\varepsilon^2 \left(\frac{1}{H}\sum_{K\in\mathcal{T}_H} \|[[\phi]]\|^2_{L^2(K)}\right)^{1/2} \left(H\sum_{K\in\mathcal{T}_H} \|\nabla f\|^2_{L^2(K)}\right)^{1/2} \\
&\leqslant C\varepsilon^2 \|\nabla f\|_{H^1(\Omega)} |\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}
\tag{4.113}
$$

Collecting (4.106), (4.111), (4.112) and (4.113) we infer that the third term of the right-hand side of (4.102) satisfies

$$
\begin{aligned}
&\left|\varepsilon^2 \sum_{E\in\mathcal{E}_H^{\mathrm{in}}} \int_{E\cap\Omega^\varepsilon} \left(\nabla(w^\varepsilon f) - \frac{1}{2}\widehat{b}^\varepsilon (w^\varepsilon f)\right) \cdot n[[\phi]]\right| \\
&\leqslant C\varepsilon \left(\sqrt{\frac{\varepsilon}{H}}\|f\|_{H^1(\Omega)} + \left(\varepsilon + \sqrt{\varepsilon H}\right)\|\nabla f\|_{H^1(\Omega)} + H\|\nabla f\|_{L^2(\Omega)}\right) |\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}
\tag{4.114}
$$

Conclusion of Step 1 Combining (4.102), (4.103), (4.105) and (4.114), we infer that

$$
|\phi|^2_{H^1_H(\Omega^\varepsilon)} \leqslant C\varepsilon \left(\sqrt{\frac{\varepsilon}{H}} + H + \sqrt{\varepsilon}\right) \left(\|f\|_{L^\infty(\Omega)} + \|\nabla f\|_{H^1(\Omega)}\right) |\phi|_{H^1_H(\Omega^\varepsilon)} \tag{4.115}
$$

This ends the first step of the proof.

Step 2: Estimation of $u_H - v_H$: Denoting $\phi_H = u_H - v_H$, we see that

$$|\phi_H|^2_{H^1_H(\Omega^\varepsilon)} \leqslant c_H(\phi_H, \phi_H) = c_H(u_H - u, \phi_H) + c_H(u - v_H, \phi_H),$$

where we recall that c_H is defined by (4.10). The first term is estimated using (4.114). We then bound the second term.

Since $\phi_H \in V_H^{\text{adv bubble}}$, we deduce from the discrete variational formulation (4.25) that

$$\begin{aligned} c_H(u_H - u, \phi_H) &= \int_{\Omega^\varepsilon} f\phi_H + c_H(\varepsilon^2 w^\varepsilon f - u, \phi_H) - c_H(\varepsilon^2 w^\varepsilon f, \phi_H) \\ &= \int_{\Omega^\varepsilon} f\phi_H + c_H(\varepsilon^2 w^\varepsilon f - u, \phi_H) \\ &\quad - \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\Delta\left(\varepsilon^2 w^\varepsilon f\right) + \widehat{b}^\varepsilon \cdot \nabla\left(\varepsilon^2 w^\varepsilon f\right)\right)\phi_H \\ &\quad + \sum_{K\in\mathcal{T}_H} \int_{\partial(K\cap\Omega^\varepsilon)} \phi_H n \cdot (\nabla(\varepsilon^2 w^\varepsilon f) - \frac{1}{2}\widehat{b}^\varepsilon(\varepsilon^2 w^\varepsilon f)) \end{aligned} \tag{4.116}$$

Since $\phi_H = 0$ on $\partial\Omega^\varepsilon$, we can take the integral in the last term of (4.116) only on $\partial K \cap \Omega^\varepsilon$. Using (4.46), we compute

$$-\Delta\left(\varepsilon^2 w^\varepsilon f\right) + \widehat{b}^\varepsilon \cdot \nabla\left(\varepsilon^2 w^\varepsilon f\right) = f + \varepsilon(-2\nabla w + bw)_{|\varepsilon} \cdot \nabla f - \varepsilon^2 w^\varepsilon \Delta f$$

We then obtain from (4.116) that

$$\begin{aligned} c_H(u_H - u, \phi_H) &= c_H(\varepsilon^2 w^\varepsilon f - u, \phi_H) \\ &\quad - \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(\varepsilon(-2\nabla w + bw)_{|\varepsilon} \cdot \nabla f - \varepsilon^2 w^\varepsilon \Delta f\right)\phi_H \\ &\quad + \sum_{K\in\mathcal{T}_H} \int_{\partial(K\cap\Omega^\varepsilon)} \phi_H n \cdot (\nabla(\varepsilon^2 w^\varepsilon f) - \frac{1}{2}\widehat{b}^\varepsilon(\varepsilon^2 w^\varepsilon f)). \end{aligned} \tag{4.117}$$

We now successively bound the three terms of the right-hand side of (4.117). Following the arguments of Step 1a, we obtain that

$$c_H(\varepsilon^2 w^\varepsilon f - u, \phi_H) \leqslant C\varepsilon^{3/2}\mathcal{N}(f)(1 + C\|b\|_{W^{1,\infty}(Y\setminus\overline{\mathcal{O}})})|\phi_H|_{H^1_H(\Omega^\varepsilon)}. \tag{4.118}$$

For the second term of the right-hand side of (4.117), we use the fact that the first factor is bounded and that the second factor satisfies a Poincaré inequality. More precisely, we use the regularity (4.104) of w and (4.91) for ϕ_H. We then obtain

$$\begin{aligned} &\left|\sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(\varepsilon(-2\nabla w + bw)_{|\varepsilon} \cdot \nabla f - \varepsilon^2 w^\varepsilon \Delta f\right)\phi_H\right| \\ &\leqslant C\varepsilon \sum_{K\in\mathcal{T}_H} (\|\nabla f\|_{L^2(K\cap\Omega^\varepsilon)} + \varepsilon\|\Delta f\|_{L^2(K\cap\Omega^\varepsilon)})\|\phi_H\|_{L^2(\Omega^\varepsilon)} \\ &\leqslant C\varepsilon\|\nabla f\|_{L^2(\Omega)}\|\phi_H\|_{L^2(\Omega^\varepsilon)} \\ &\leqslant C\varepsilon^2\|\phi_H\|_{H^1_H(\Omega^\varepsilon)}. \end{aligned} \tag{4.119}$$

Following the arguments of Step 1c, we get

$$\left|\sum_{K\in\mathcal{T}_H}\int_{\partial(K\cap\Omega^\varepsilon)}\phi_H n\cdot(\nabla(\varepsilon^2 w^\varepsilon f)-\frac{1}{2}\widehat{b}^\varepsilon(\varepsilon^2 w^\varepsilon f))\right|$$
$$\leqslant C\varepsilon\left(\sqrt{\frac{\varepsilon}{H}}\|f\|_{H^1(\Omega)}+\sqrt{\varepsilon H}\|\nabla f\|_{H^1(\Omega)}+H\|\nabla f\|_{L^2(\Omega)}+\varepsilon\|\nabla f\|_{H^1(\Omega)}\right)|\phi_H|_{H^1_H(\Omega^\varepsilon)}. \quad (4.120)$$

Combining (4.118), (4.119) and (4.120), we infer that

$$|u_H-v_H|^2_{H^1_H(\Omega^\varepsilon)}\leqslant C\varepsilon\left(\sqrt{\frac{\varepsilon}{H}}+H+\sqrt{\varepsilon}\right)\left(\|f\|_{L^\infty(\Omega)}+\|\nabla f\|_{H^1(\Omega)}\right)|\phi_H|_{H^1_H(\Omega^\varepsilon)}. \quad (4.121)$$

Conclusion We deduce from (4.115) and (4.121) that

$$|u-u_H|_{H^1_H(\Omega^\varepsilon)}\leqslant C\varepsilon\left(\sqrt{\frac{\varepsilon}{H}}+H+\sqrt{\varepsilon}\right)\left(\|f\|_{L^\infty(\Omega)}+\|\nabla f\|_{H^1(\Omega)}\right).$$

which yields the desired estimate (4.53). This concludes the proof of Theorem 4.5.

4.6.3 Proof of Theorem 4.6

We again denote by u be the solution to (4.2), and u_H its approximation, solution to (4.13).

Let $\Pi_H f$ be the L^2-orthogonal projection of f on the space of piecewise constant functions. We recall that we have the following interpolation estimate: there exists C independent of H, f such that

$$\|f-\Pi_H f\|_{L^2(\Omega)}\leqslant CH\|\nabla f\|_{L^2(\Omega)} \quad (4.122)$$

In what follows, we make use of the notation $g_{|\varepsilon}(x)=g\left(\frac{x}{\varepsilon}\right)$. The constant C denote a constant independent of ε, H, f and that may vary from one line to the next.

We will make use of the fact that

$$w\in C^2(\overline{Y\backslash\mathcal{O}}), \quad (4.123)$$

which follows from the arguments of the proof of Theorem 4.2 and the assumption that $b\in(W^{2,\infty}(Y\backslash\overline{\mathcal{O}}))^2$.

We define

$$v_H(x)=\sum_{K\in\mathcal{T}_H}\Pi_H(f(1-(b\cdot\nabla w)_{|\varepsilon}))\Psi_0^{\varepsilon,K}(x)+\sum_{E\in\mathcal{E}_H^{\rm in}}\left(\int_E u\right)\Phi_0^{\varepsilon,E}(x),$$

where $\Phi_0^{\varepsilon,E}$ and $\Psi_0^{\varepsilon,K}$ are respectively solutions to Problems (4.17) and (4.18). We recall that if $K\subset B^\varepsilon$, respectively $E\subset B^\varepsilon$ then $\Psi_0^{\varepsilon,K}\equiv 0$, $\Phi_0^{\varepsilon,E}\equiv 0$. We see from (4.14), that $v_H\in V_{H,\rm bubble}$. We next decompose the exact solution into

$$u=v_H+\phi.$$

Notice that v_H satisfies

$$\begin{aligned}
&\int_E v_H = \int_E u, \quad \text{hence} \int_E \phi = 0, \quad \text{for all} \ \ E \in \mathcal{E}_H^{\text{in}} \\
&(\alpha \nabla v_H) \cdot n = \text{Constant on (each side of) E}, \quad \text{for all} \ \ E \in \mathcal{E}_H^{\text{in}} \\
&-\alpha \Delta v_H = \Pi_H(f(1-(b\cdot\nabla w)_{|\varepsilon})) \quad \text{on } K \cap \Omega^\varepsilon, \text{for all} \ \ K \in \mathcal{T}_H.
\end{aligned} \tag{4.124}$$

For all $(u_H, v_H) \in V_{H,\text{bubble}} \times V_{H,\text{bubble}}$, we have

$$\begin{aligned}
a_H(u_H, v_H) &\leqslant \alpha |u_H|_{H^1_H(\Omega^\varepsilon)} |v_H|_{H^1_H(\Omega^\varepsilon)} + \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)} |u_H|_{H^1_H(\Omega^\varepsilon)} \|v_H\|_{L^2(\Omega^\varepsilon)} \\
&\leqslant (\alpha + C_P \varepsilon \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}) |u_H|_{H^1_H(\Omega^\varepsilon)} |v_H|_{H^1_H(\Omega^\varepsilon)},
\end{aligned}$$

where we used (4.91). Using

$$C_P \|b\|_{L^\infty(Y\setminus\overline{\mathcal{O}})} < \alpha, \tag{4.125}$$

where C_P is such that (4.91) holds, the bilinear form a_H is coercive and we have for all $u_H \in V_{H,\text{bubble}}$

$$\begin{aligned}
a(u_H, u_H) &\geqslant \alpha |u_H|^2_{H^1_H(\Omega^\varepsilon)} - \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)} |u_H|_{H^1_H(\Omega^\varepsilon)} \|u_H\|_{L^2(\Omega^\varepsilon)} \\
&\geqslant (\alpha - C_P \varepsilon \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}) |u_H|^2_{H^1_H(\Omega^\varepsilon)} \\
&\geqslant (\alpha - C_P \|b\|_{L^\infty(Y\setminus\overline{\mathcal{O}})}) |u_H|^2_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}$$

We want to apply Strang's lemma[38, Lemma 2.25] and obtain

$$\begin{aligned}
|u - u_H|^2_{H^1_H(\Omega^\varepsilon)} \leqslant &\left(1 + \frac{\alpha + C_P \varepsilon \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}}{\alpha - C_P \varepsilon \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}}\right) \inf_{w_H \in V_{H,\text{bubble}}} |u - w_H|_{H^1_H(\Omega^\varepsilon)} \\
&+ \frac{1}{\alpha - C_P \varepsilon \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}} \sup_{w_H \in V_{H,\text{bubble}}} \frac{\left|\int_{\Omega^\varepsilon} f w_H - a_H(u, w_H)\right|}{|w_H|_{H^1_H(\Omega^\varepsilon)}}.
\end{aligned} \tag{4.126}$$

Our proof is decomposed in two steps. We bound the first term of (4.126) by the norm of $\phi = u - v_H$. Then we estimate the second term of (4.126) which corresponds to the error of consistency.

Step 1: Estimation of $u - v_H$: Using the approximation of u given in the homogenization result (4.44) and denoting $\widetilde{\phi} = \varepsilon^2 w^\varepsilon f - v_H$, we have

$$\begin{aligned}
\alpha|\phi|^2_{H^1_H(\Omega^\varepsilon)} &= a_{\text{diff},H}(u - \varepsilon^2 w^\varepsilon f, \phi) + a_{\text{diff},H}(\varepsilon^2 w^\varepsilon f - v_H, \phi) \\
&= a_{\text{diff},H}(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\alpha\Delta\widetilde{\phi}\right)\phi + \int_{\partial(K\cap\Omega^\varepsilon)} (\alpha\nabla\widetilde{\phi}\cdot n)\phi \\
&= a_{\text{diff},H}(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\alpha\Delta\widetilde{\phi}\right)\phi + \int_{(\partial K)\cap\Omega^\varepsilon} (\alpha\nabla\widetilde{\phi}\cdot n)\phi \\
&= a_{\text{diff},H}(u - \varepsilon^2 w^\varepsilon f, \phi) \\
&\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} \left(-\alpha\Delta\widetilde{\phi}\right)\phi \\
&\quad + \varepsilon^2 \sum_{K\in\mathcal{T}_H} \int_{(\partial K)\cap\Omega^\varepsilon} (\alpha\nabla(w^\varepsilon f)\cdot n)\phi, \qquad (4.127)
\end{aligned}$$

where we used the fact that $\phi = 0$ on $\partial\Omega^\varepsilon$ in the third line, (4.124) in the last line and we recall that $a_{\text{diff},H}$ is defined by (4.9).

We bound successively the three terms of the right-hand side of (4.102). Roughly speaking:

- the first term is small because of the homogenization result (4.44);
- the second is of order H because at the leading order $-\alpha\Delta v_H + \widehat{b}^\varepsilon \cdot \nabla v_H \simeq \Pi_H(f(1 - (b\cdot\nabla w)_{|\varepsilon}))$;
- bounding the third term uses the fact that w is a periodic function.

Step 1a The first term of (4.127) is estimated as follows. Denoting $\widetilde{u} = u - \varepsilon^2 w^\varepsilon f$, we write

$$\begin{aligned}
a_{\text{diff},H}(\widetilde{u}, \phi) &\leqslant \alpha|\widetilde{u}|_{H^1_H(\Omega^\varepsilon)}|\phi|_{H^1_H(\Omega^\varepsilon)} \\
&\leqslant C\varepsilon^{3/2}\mathcal{N}(f)|\phi|_{H^1_H(\Omega^\varepsilon)}, \qquad (4.128)
\end{aligned}$$

where we used (4.44) in the last line.

Step 1b We now turn to the second term of the right-hand side of (4.127). Using the corrector equation (4.46) and (4.98), similarly as in the proof of Theorem 4.2, we have

$$\begin{aligned}
&\left|\sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon}\left(-\alpha\Delta\widetilde{\phi}+\widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi\right|\\
&\leqslant\left(\|f(1-(b\cdot\nabla w)_{|\varepsilon})-\Pi_H(f(1-(b\cdot\nabla w)_{|\varepsilon}))\|_{L^2(\Omega)}+2\alpha\varepsilon\|\nabla w\|_{L^\infty(\Omega^\varepsilon)}\|\nabla f\|_{L^2(\Omega)}\right.\\
&\qquad\left.+\alpha\varepsilon^2\|w\|_{L^\infty(\Omega^\varepsilon)}\|\Delta f\|_{L^2(\Omega)}\right)\|\phi\|_{L^2(\Omega^\varepsilon)}\\
&\leqslant C\varepsilon H\|\nabla(f(1-(b\cdot\nabla w)_{|\varepsilon}))\|_{L^2(\Omega^\varepsilon)}|\phi|_{H^1_H(\Omega^\varepsilon)}+C\varepsilon^2\mathcal{N}(f)|\phi|_{H^1_H(\Omega^\varepsilon)}\\
&\leqslant CH\|\nabla^2 w\|_{L^\infty(\Omega^\varepsilon)}\|b\|_{L^\infty(\Omega^\varepsilon)}\|f\|_{L^2(\Omega)}|\phi|_{H^1_H(\Omega^\varepsilon)}\\
&\quad+C\varepsilon H(\|\nabla f\|_{L^2(\Omega)}+\|f\|_{L^2(\Omega)})|\phi|_{H^1_H(\Omega^\varepsilon)}\\
&\quad+C\varepsilon^2\mathcal{N}(f)|\phi|_{H^1_H(\Omega^\varepsilon)}\\
&\leqslant CH\|b\|_{L^\infty(\Omega^\varepsilon)}\|f\|_{L^2(\Omega)}|\phi|_{H^1_H(\Omega^\varepsilon)}\\
&\quad+C\varepsilon\{CH\{\|f\|_{L^2(\Omega)}+\|\nabla f\|_{L^2(\Omega)}\}+C\varepsilon\mathcal{N}(f)\}|\phi|_{H^1_H(\Omega^\varepsilon)},
\end{aligned}$$

where $\mathcal{N}(f)$ is defined by (4.45), and we used (4.91), (4.97) in the third line and (4.123) in the fourth line. We infer that

$$\begin{aligned}
&\left|\sum_{K\in\mathcal{T}_H}\int_{K\cap\Omega^\varepsilon}\left(-\alpha\Delta\widetilde{\phi}+\widehat{b}^\varepsilon\cdot\nabla\widetilde{\phi}\right)\phi\right|\\
&\qquad\leqslant\left(CH\|b\|_{L^\infty(\Omega^\varepsilon)}\|f\|_{L^2(\Omega)}+C\varepsilon(H+\varepsilon)\mathcal{N}(f)\right)|\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}\tag{4.129}$$

Step 1c We then deal with the third term of the right-hand side of (4.127) similarly as in the previous proof. We then obtain

$$\begin{aligned}
&\left|\varepsilon^2\sum_{E\in\mathcal{E}_H^{\mathrm{in}}}\int_{E\cap\Omega^\varepsilon}(\alpha\nabla(w^\varepsilon f)\cdot n)\,[[\phi]]\right|\\
&\leqslant C\varepsilon\left(\sqrt{\frac{\varepsilon}{H}}\|f\|_{H^1(\Omega)}+\sqrt{\varepsilon H}\|\nabla f\|_{H^1(\Omega)}+H\|\nabla f\|_{L^2(\Omega)}+\varepsilon\|\nabla f\|_{H^1(\Omega)}\right)|\phi|_{H^1_H(\Omega^\varepsilon)}.
\end{aligned}\tag{4.130}$$

Conclusion of Step 1 Combining (4.128), (4.129), and (4.130), we infer that

$$\begin{aligned}
|\phi|^2_{H^1_H(\Omega^\varepsilon)}&\leqslant CH\|b\|_{L^\infty(\Omega^\varepsilon)}\|f\|_{L^2(\Omega)}\\
&\quad+C\varepsilon\left(\sqrt{\frac{\varepsilon}{H}}+H+\sqrt{\varepsilon}\right)\left(\|f\|_{L^\infty(\Omega)}+\|\nabla f\|_{H^1(\Omega)}\right)|\phi|_{H^1_H(\Omega^\varepsilon)}
\end{aligned}\tag{4.131}$$

This ends the first step of the proof.

Step 2: Estimation of the consistency error We want to estimate the error of consistency. We have for all $w_H \in V_{H,\text{bubble}}$,

$$\begin{aligned} a_H(u,w_H) - \int_{\Omega^\varepsilon} f w_H &= a_H(u-\varepsilon^2 w^\varepsilon f, w_H) + a_H(\varepsilon^2 w^\varepsilon f, w_H) - \int_{\Omega^\varepsilon} f w_H \\ &= a_H(u-\varepsilon^2 w^\varepsilon f, w_H) \\ &\quad + \sum_{K\in\mathcal{T}_H} \int_{K\cap\Omega^\varepsilon} (-\alpha\Delta(\varepsilon^2 w^\varepsilon f) + \widehat{b}^\varepsilon \cdot \nabla(\varepsilon^2 w^\varepsilon f)) w_H - \int_{\Omega^\varepsilon} f w_H \\ &\quad + \varepsilon^2 \sum_{K\in\mathcal{T}_H} \int_{\partial(K\cap\Omega^\varepsilon)} (\alpha\nabla(\varepsilon^2 w^\varepsilon f)\cdot n) w_H. \end{aligned} \tag{4.132}$$

The first term of (4.132) is bounded using (4.44). The second term of (4.132) is estimated using the corrector equation (4.46) and (4.98), similarly as in the proof of Theorem 4.2. We deal with the term of (4.132) following the arguments of [59, Section 4, Step 1c]. We then obtain for all $w_H \in V_{H,\text{bubble}}$,

$$\begin{aligned} &\left| a_H(u,w_H) - \int_{\Omega^\varepsilon} f w_H \right| \\ &\leqslant C\varepsilon \left(\sqrt{\frac{\varepsilon}{H}} + H + \sqrt{\varepsilon} \right) \left(\|f\|_{L^\infty(\Omega)} + \|\nabla f\|_{H^1(\Omega)} \right) |w_H|_{H^1_H(\Omega^\varepsilon)}. \end{aligned} \tag{4.133}$$

Conclusion Combining (4.126), (4.131) and (4.133), we get the desired estimate. This concludes the proof of Theorem 4.5.

4.7 Appendix: Definition of the Adv-MsFEM à la Crouzeix-Raviart basis functions

4.7.1 Approximation space (4.26)

In this section, we show that Problems (4.29) and (4.30) are well-posed.

They are saddle-point problems (see [38] for the theory). Let n_K be the number of inner edges associated to $K \in \mathcal{T}_H$ and $V_K = \{u \in H^1(K), u = 0 \text{ in } K\cap B^\varepsilon \text{ and } u = 0 \text{ on } \mathcal{E}_H^{\text{ext}}\}$, their variational formulation are of the form

$$\begin{aligned} &\text{Find } (u_H, [\lambda^{K,E}]_{E\in\mathcal{E}_H^{\text{in}}}) \in V_K \times \mathbb{R}^{n_K} \text{ such that} \\ &c_K(u_H, v_H) - \sum_{E\in\mathcal{E}_H^{\text{in}}} \lambda^{K,E} \int_E v_H = F(v_H) \quad \text{for all } \ v_H \in V_K, \\ &\mu^{K,E} \int_E u_H = \delta^{K,E}\mu^{K,E} \quad \text{for all } \ \mu^{K,E} \in \mathbb{R}, \text{ for all } \ E \in \mathcal{E}_H^{\text{in}}, \end{aligned}$$

where

$$c_K(u_H, v_H) = \int_{K\cap\Omega^\varepsilon} \alpha\nabla u_H \cdot \nabla v_H + \frac{1}{2}(\widehat{b}^\varepsilon \cdot \nabla u_H) v_H - \frac{1}{2}\text{div}\,(\widehat{b}^\varepsilon v_H) u_H,$$

$\delta^{K,E} = 0, 1$ and $F \in L^2(K\cap\Omega^\varepsilon)$.

We want to apply [38, Theorem 2.34, p100]. We denote

$$c_E(v_H, (\mu^{K,E})_E) = \sum_{E\in\mathcal{E}_H^{\text{in}}} \mu^{K,E} \int_E v_H.$$

The bilinear form c_K is coercive on

$$W_K = \left\{ u_H \in V_K, \int_E u_H = 0 \text{ for all } \ E \in \mathcal{E}_H^{\text{in}} \right\}$$

since we have for all $u_H \in W_K$

$$\begin{aligned} c_K(u_H, u_H) &= \int_{K\cap\Omega^\varepsilon} \alpha \nabla u_H \cdot \nabla u_H + \frac{1}{2}(\widehat{b}^\varepsilon \cdot \nabla u_H) u_H - \frac{1}{2}\text{div } (\widehat{b}^\varepsilon u_H) u_H \\ &= \int_{K\cap\Omega^\varepsilon} \alpha |\nabla u_H|^2 - \frac{1}{2}(\text{div } \widehat{b}^\varepsilon)(u_H)^2 \\ &\geqslant \alpha \int_{K\cap\Omega^\varepsilon} |\nabla u_H|^2, \end{aligned}$$

where we used that div $b \leqslant 0$ in the last line.

Let $(\mu^E)_E \in \mathbb{R}^{n_K}$, we denote $v_H = \sum_{E\in\mathcal{E}_H^{\text{in}}} \mu^E \Phi_0^{\varepsilon,E}$, where $\Phi_0^{\varepsilon,E}$ is solution to (4.17). We have

$$\frac{c_E(v_H, (\mu^E)_E)}{\|(\mu^E)_E\| \|v_H\|_{H^1(K\cap\Omega^\varepsilon)}} = \frac{\sum_{E\in\mathcal{E}_H^{\text{in}}} (\mu^E)^2}{\|(\mu^E)_E\| \|v_H\|_{H^1(K\cap\Omega^\varepsilon)}} \geqslant C,$$

where we used that $C\|v_H\|_{H^1(K\cap\Omega^\varepsilon)} \leqslant \|(\mu^E)_E\|$. We infer that there exists a constant C such that

$$\inf_{(\mu^E)_E\in\mathbb{R}^{n_K}} \sup_{v_H\in V_K} \frac{c_E(v_H, (\mu^E)_E)}{\|(\mu^E)_E\| \|v_H\|_{H^1(K\cap\Omega^\varepsilon)}} \geqslant C.$$

We then deduce that Problems (4.29) and (4.30) are well-posed.

4.7.2 Approximation space (4.35)

In this section, we show that Problems (4.37) and (4.38) are well-posed under the assumption that

$$C_P^K \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)} < \alpha, \tag{4.134}$$

where C_P^K is a constant such that $\|v\|_{L^2(K\cap\Omega^\epsilon)} \leqslant C_P^K \|\nabla v\|_{L^2(K\cap\Omega^\epsilon)}$ for all $v \in H_0^1(K \cap \Omega^\varepsilon)$. They are saddle-point problems. Let n_K be the number of inner edges associated to $K \in \mathcal{T}_H$ and $V_K = \{u \in H^1(K), u = 0 \text{ in } K \cap B^\varepsilon \text{ and } u = 0 \text{ on } \mathcal{E}_H^{\text{ext}}\}$, their variational formulation are of the form

$$\begin{aligned} &\text{Find } (u_H, [\lambda^{K,E}]_{E\in\mathcal{E}_H^{\text{in}}}) \in V_K \times \mathbb{R}^{n_K} \text{ such that} \\ &a_K(u_H, v_H) - \sum_{E\in\mathcal{E}_H^{\text{in}}} \lambda^{K,E} \int_E v_H = F(v_H) \quad \text{for all } \ v_H \in V_K, \\ &\mu^{K,E} \int_E u_H = \delta^{K,E} \mu^{K,E} \quad \text{for all } \ \mu^{K,E} \in \mathbb{R}, \text{ for all } \ E \in \mathcal{E}_H^{\text{in}}, \end{aligned}$$

where

$$a_K(u_H, v_H) = \int_{K\cap\Omega^\varepsilon} \alpha \nabla u_H \cdot \nabla v_H + (\widehat{b}^\varepsilon \cdot \nabla u_H) v_H,$$

$\delta^{K,E} = 0, 1$ and $F \in L^2(K \cap \Omega^\varepsilon)$.

We want to apply [38, Theorem 2.34, p100]. We denote $a_E(v_H, (\mu^{K,E})_E) = \sum_{E\in\mathcal{E}_H^{\text{in}}} \mu^{K,E} \int_E v_H$. The bilinear form c_K is coercive on

$$W_K = \left\{ u_H \in V_K, \int_E u_H = 0 \text{ for all } \ E \in \mathcal{E}_H^{\text{in}} \right\}$$

since we have for all $u_H \in W_K$

$$\begin{aligned} a_K(u_H, u_H) &= \int_{K\cap\Omega^\varepsilon} \alpha \nabla u_H \cdot \nabla u_H + (\widehat{b}^\varepsilon \cdot \nabla u_H) u_H \\ &= \int_{K\cap\Omega^\varepsilon} \alpha |\nabla u_H|^2 - C_P^K \|\widehat{b}^\varepsilon)\|_{L^\infty(\Omega^\varepsilon)} \|\nabla u_H\|^2_{L^2(\Omega^\varepsilon)} \\ &\geqslant (\alpha - C_P^K \|\widehat{b}^\varepsilon\|_{L^\infty(\Omega^\varepsilon)}) \int_{K\cap\Omega^\varepsilon} |\nabla u_H|^2, \end{aligned}$$

where we used that div $b \leqslant 0$ in the last line.

Following the lines of the previous section, we have that there exists a constant C such that

$$\inf_{(\mu^E)_{E\in\mathbb{R}^{n_K}}} \sup_{v_H\in V_K} \frac{c_E(v_H, (\mu^E)_E)}{\|(\mu^E)_E\| \|v_H\|_{H^1(K\cap\Omega^\varepsilon)}} \leqslant C.$$

We then deduce that Problems (4.37) and (4.38) are well-posed.

Bibliography

[1] A. Abdulle. Discontinuous Galerkin finite element heterogeneous multiscale method for elliptic problems with multiple scales. *Math. Comp.*, 81(278):687–713, 2012.

[2] E. Acerbi, V. Chiadò Piat, G. Dal Maso, and D. Percivale. An extension theorem from connected sets, and homogenization in general periodic domains. *Nonlinear Anal.*, 18(5):481–496, 1992.

[3] G. Allaire. Homogenization of the Navier-Stokes equations in open sets perforated with tiny holes II: Non-critical sizes of the holes for a volume distribution and a surface distribution of holes. *Arch. Rational Mech. Anal.*, 113(3):261–298, 1991.

[4] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23(6):1482–1518, 1992.

[5] G. Allaire. Homogenization of the unsteady Stokes equations in porous media. In *Progress in partial differential equations: calculus of variations, applications (Pont-à-Mousson, 1991)*, volume 267 of *Pitman Res. Notes Math. Ser.*, pages 109–123. Longman Sci. Tech., Harlow, 1992.

[6] G. Allaire and R. Brizzi. A multiscale finite element method for numerical homogenization. *Multiscale Model. Simul.*, 4(3):790–812, 2005.

[7] G. Allaire, A. Mikelic, and A. Piatnitski. Homogenization approach to the dispersion theory for reactive transport through porous media. *SIAM J. Math. Anal.*, 42(1):125–144, 2010.

[8] G. Allaire and F. Murat. Homogenization of the Neumann problem with nonisolated holes. *Asymptotic Anal.*, 7(2):81–95, 1993.

[9] G. Allaire and A.-L. Raphael. Homogenization of a convection-diffusion model with reaction in a porous medium. *C. R. Math. Acad. Sci. Paris*, 344(8):523–528, 2007.

[10] B. Amaziane, M. Goncharenko, and L. Pankratov. Homogenization of a convection–diffusion equation in perforated domains with a weak adsorption. *Z. Angew. Math. Phys.*, 58(4):592–611, 2007.

[11] T. Arbogast. Numerical subgrid upscaling of two-phase flow in porous media. In *Numerical treatment of multiphase flows in porous media*, pages 35–49. Springer, 2000.

[12] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York, 1978.

[13] C. Bernardi, Y. Maday, and F. Rapetti. *Discrétisations variationnelles de problèmes aux limites elliptiques*, volume 45 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2004.

[14] J. H. Bramble, R. D. Lazarov, and J. E. Pasciak. Least-squares for second-order elliptic problems. *Comput. Methods Appl. Mech. Engrg.*, 152(1-2):195–210, 1998. Symposium on Advances in Computational Mechanics, Vol. 5 (Austin, TX, 1997).

[15] S. C. Brenner and L. R. Scott. *The mathematical theory of Finite Element methods*, volume 15. Springer, 2008.

[16] F. Brezzi, M. O. Bristeau, L. P. Franca, M. Mallet, and G. Rogé. A relationship between stabilized finite element methods and the Galerkin method with bubble functions. *Comput. Methods Appl. Mech. Engrg.*, 96(1):117–129, 1992.

[17] F. Brezzi, D. Marini, and E. Süli. Residual-free bubbles for advection-diffusion problems: the general error analysis. *Numer. Math.*, 85(1):31–47, 2000.

[18] F. Brezzi, L. D. Marini, and P. Pietra. Two-dimensional exponential fitting and applications to drift-diffusion models. *SIAM J. Numer. Anal.*, 26(6):1342–1355, 1989.

[19] A. N. Brooks and T. J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982. FENOMECH '81, Part I (Stuttgart, 1981).

[20] E. Burman. Stabilized Finite Element methods for nonsymmetric, noncoercive, and ill-posed problems. Part I: Elliptic equations. *SIAM J. Sci. Comput.*, 35(6):A2752–A2780, 2013.

[21] V. M. Calo, E. T. Chung, Y. Efendiev, and W. T. Leung. Multiscale stabilization for convection-dominated diffusion in heterogeneous media. *Comput. Methods Appl. Mech. Engrg.*, 304:359 – 377, 2016.

[22] A. Capatina, H. Ene, and C. Timofte. Homogenization results for elliptic problems in periodically perforated domains with mixed-type boundary conditions. *Asymptot. Anal.*, 80(1):45–56, 2012.

[23] C. Chainais-Hillairet and J. Droniou. Finite-volume schemes for noncoercive elliptic problems with Neumann boundary conditions. *IMA J. Numer. Anal.*, 31(1):61–85, 2011.

[24] D. Cioranescu, A. Damlamian, G. Griso, and D. Onofrei. The periodic unfolding method for perforated domains and Neumann sieve models. *J. Math. Pures Appl. (9)*, 89(3):248–277, 2008.

[25] D. Cioranescu and P. Donato. *An introduction to homogenization*, volume 17 of *Oxford Lecture Series in Mathematics and its Applications*. The Clarendon Press, Oxford University Press, New York, 1999.

[26] D. Cioranescu, P. Donato, and R. Zaki. Periodic unfolding and Robin problems in perforated domains. *C. R. Math. Acad. Sci. Paris*, 342(7):469–474, 2006.

[27] D. Cioranescu and F. Murat. A strange term coming from nowhere. In *Topics in the mathematical modelling of composite materials*, pages 45–93. Springer, 1997.

[28] D. Cioranescu and J. S. J. Paulin. Homogenization in open sets with holes. *J. Math. Anal. Appl.*, 71(2):590–607, 1979.

[29] G. Dal Maso. *An introduction to* Γ*-convergence*. Progress in Nonlinear Differential Equations and their Applications, 8. Birkhäuser Boston, Inc., Boston, MA, 1993.

[30] G. Dal Maso and F. Murat. Asymptotic behaviour and correctors for linear Dirichlet problems with simultaneously varying operators and domains. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 21(4):445–486, 2004.

[31] P. Degond, A. Lozinski, B. P. Muljadi, and J. Narski. Crouzeix-Raviart MsFEM with bubble functions for diffusion and advection-diffusion in perforated media. *Commun. Comput. Phys.*, 17(4):887–907, 2015.

[32] L. Demkowicz and J. Gopalakrishnan. A class of discontinuous Petrov-Galerkin methods. II. Optimal test functions. *Numer. Methods Partial Differential Equations*, 27(1):70–105, 2011.

[33] J. Droniou and T. Gallouët. Finite volume methods for convection-diffusion equations with right-hand side in H^{-1}. *M2AN Math. Model. Numer. Anal.*, 36(4):705–724, 2002.

[34] J. Droniou and J.-L. Vázquez. Noncoercive convection–diffusion elliptic problems with Neumann boundary conditions. *Calc. Var. Partial Differential Equations*, 34(4):413–434, 2009.

[35] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[36] Y. Efendiev and T. Y. Hou. *Multiscale finite element methods*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2009. Theory and applications.

[37] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM J. Numer. Anal.*, 37(3):888–910, 2000.

[38] A. Ern. *Theory and practice of Finite Elements*, volume 159. Springer, 2004.

[39] L. P. Franca, S. L. Frey, and T. J. R. Hughes. Stabilized finite element methods. I. Application to the advective-diffusive model. *Comput. Methods Appl. Mech. Engrg.*, 95(2):253–276, 1992.

[40] D. A. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*, volume 224. Springer, 2001.

[41] 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.

[42] C. Harder, D. Paredes, and F. Valentin. On a multiscale hybrid-mixed method for advective-reactive dominated problems with heterogeneous coefficients. *Multiscale Model. Simul.*, 13(2):491–518, 2015.

[43] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[44] P. Henning and M. Ohlberger. The heterogeneous multiscale finite element method for advection-diffusion problems with rapidly oscillating coefficients and large expected drift. *Netw. Heterog. Media*, 5(4):711–744, 2010.

[45] U. Hornung. *Homogenization and porous media*, volume 6. Springer, 1997.

[46] U. Hornung and W. Jäger. Diffusion, convection, adsorption, and reaction of chemicals in porous media. *J. Differential Equations*, 92(2):199–225, 1991.

[47] T. Y. Hou and X.-H. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1):169–189, 1997.

[48] T. Y. Hou, X.-H. Wu, and Z. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comp.*, 68(227):913–943, 1999.

[49] T. J. R. Hughes. Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. *Comput. Methods Appl. Mech. Engrg.*, 127(1):387–401, 1995.

[50] T. J. R. Hughes and A. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In *Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979)*, volume 34 of *AMD*, pages 19–35. Amer. Soc. Mech. Engrs. (ASME), New York, 1979.

[51] W. Hundsdorfer and J. Verwer. *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, volume 33 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2003.

[52] V. V. Jikov, S. M. Kozlov, and O. A. Oleĭnik. *Homogenization of differential operators and integral functionals*. Springer-Verlag, Berlin, 1994. Translated from the Russian by G. A. Yosifian [G. A. Iosif′yan].

[53] V. John and P. Knobloch. On spurious oscillations at layers diminishing (sold) methods for convection–diffusion equations: Part I–A review. *Comput. Methods Appl. Mech. Engrg.*, 196(17):2197–2215, 2007.

[54] C. Johnson, U. Nävert, and J. Pitkäranta. Finite element methods for linear hyperbolic problems. *Comput. Methods Appl. Mech. Engrg.*, 45(1-3):285–312, 1984.

[55] K. Kavaliou and L. Tobiska. A Finite Element method for a noncoercive elliptic problem with Neumann boundary conditions. *Comput. Methods Appl. Math.*, 12(2):168–183, 2012.

[56] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, and C. Theodoropoulos. Equation-free, coarse-grained multiscale computation: enabling microscopic simulators to perform system-level analysis. *Commun. Math. Sci.*, 1(4):715–762, 2003.

[57] J. Ku. A least-squares method for second order noncoercive elliptic partial differential equations. *Math. Comp.*, 76(257):97–114, 2007.

[58] C. Le Bris, F. Legoll, and A. Lozinski. MsFEM à la Crouzeix-Raviart for highly oscillatory elliptic problems. *Chin. Ann. Math. Ser. B*, 34(1):113–138, 2013.

[59] C. Le Bris, F. Legoll, and A. Lozinski. An MsFEM type approach for perforated domains. *Multiscale Model. Simul.*, 12(3):1046–1077, 2014.

[60] C. Le Bris, F. Legoll, and F. Madiot. A numerical comparison of some Multiscale Finite Element approaches for advection-dominated problems in heterogeneous media. *Accepted in ESAIM Math. Model. Numer. Anal., arXiv preprint arXiv:1511.08453, HAL preprint hal-01235642*, 2016.

[61] C. Le Bris, F. Legoll, and F. Madiot. Stabilisation de problèmes non coercifs via une méthode numérique utilisant la mesure invariante (Stabilization of non-coercive problems using the invariant measure). *C. R. Math. Acad. Sci. Paris*, 354(8):799–803, 2016.

[62] C. Le Bris, F. Legoll, and F. Madiot. Stable approximation of the advection-diffusion equation using the invariant measure. *Submitted, arXiv preprint arXiv:1609.04777, HAL preprint hal-01367417*, 2016.

[63] C. Le Bris, F. Legoll, and F. Thomines. Multiscale finite element approach for "weakly" random problems and related issues. *ESAIM Math. Model. Numer. Anal.*, 48(3):815–858, 2014.

[64] J. L. Lions. Asymptotic expansions in perforated media with a periodic structure. *Rocky Mountain J. Math.*, 10:125–140, 1980.

[65] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.

[66] P. Marcellini. Convergence of second order linear elliptic operators. *Boll. Un. Mat. Ital. B (5)*, 16(1):278–290, 1979.

[67] A. Mikelić. Homogenization of nonstationary Navier-Stokes equations in a domain with a grained boundary. *Ann. Mat. Pura Appl. (4)*, 158(1):167–179, 1991.

[68] B. P. Muljadi, J. Narski, A. Lozinski, and P. Degond. Nonconforming multiscale finite element method for Stokes flows in heterogeneous media. Part I: Methodologies and numerical experiments. *Multiscale Model. Simul.*, 13(4):1146–1172, 2015.

[69] F. Murat and L. Tartar. H-convergence. In *Topics in the mathematical modelling of composite materials*, volume 31 of *Progr. Nonlinear Differential Equations Appl.*, pages 21–43. Birkhäuser Boston, Boston, MA, 1997.

[70] F. Natterer. Über die punktweise Konvergenz finiter Elemente. *Numer. Math.*, 25(1):67–77, 1975/76.

[71] J. Nitsche. L_∞-convergence of finite element approximations. In *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, pages 261–274. Lecture Notes in Math., Vol. 606. Springer, Berlin, 1977.

[72] F. Ouaki. *Etude de schémas multi-échelles pour la simulation de réservoir*. PhD thesis, Ecole Polytechnique X, http://www.theses.fr/2013EPXX0062, 2013.

[73] G. C. Papanicolaou and S. R. S. Varadhan. Diffusion in regions with many small holes. In *Stochastic Differential Systems Filtering and Control*, pages 190–206. Springer, 1980.

[74] P. J. Park and T. Y. Hou. Multiscale numerical methods for singularly perturbed convection-diffusion equations. *Int. J. Comput. Methods*, 1(01):17–65, 2004.

[75] B. Perthame. Perturbed dynamical systems with an attracting singularity and weak viscosity limits in Hamilton-Jacobi equations. *Trans. Amer. Math. Soc.*, 317(2):723–748, 1990.

[76] J. Principe and R. Codina. On the stabilization parameter in the subgrid scale approximation of scalar convection-diffusion-reaction equations on distorted meshes. *Comput. Methods Appl. Mech. Engrg.*, 199(21-22):1386–1402, 2010.

[77] A. Quarteroni. *Numerical models for differential problems*, volume 2. Springer Science & Business Media, 2010.

[78] A. Quarteroni and A. Valli. *Numerical approximation of partial differential equations*, volume 23 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1994.

[79] R. Rannacher and R. Scott. Some optimal error estimates for piecewise linear finite element approximations. *Math. Comp.*, 38(158):437–445, 1982.

[80] H.-G. Roos. Robust numerical methods for singularly perturbed differential equations: a survey covering 2008–2012. *ISRN Appl. Math.*, pages Art. ID 379547, 30, 2012.

[81] H.-G. Roos, M. Stynes, and L. Tobiska. *Numerical Methods for Singularly Perturbed Differential Equations: Convection-Diffusion and Flow Problems*, volume 159. Springer, 2004.

[82] J. Rubinstein and R. Mauri. Dispersion and convection in periodic porous media. *SIAM J. Appl. Math.*, 46(6):1018–1023, 1986.

[83] H. Ruffieux. Multiscale finite element method for highly oscillating advection-diffusion problems in convection-dominated regime. Master's thesis, Ecole Polytechnique Fédérale de Lausanne, Spring 2013.

[84] S. A. Sauter and C. Schwab. *Boundary element methods*. Springer, 2011.

[85] R. Scott. Optimal L^∞ estimates for the finite element method on irregular meshes. *Math. Comp.*, 30(136):681–697, 1976.

[86] M. Stynes. Steady-state convection-diffusion problems. *Acta Numer.*, 14(1):445–508, 2005.

[87] W. G. Szymczak. An analysis of viscous splitting and adaptivity for steady-state convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 67(3):311–354, 1988.

Appendix A

Quelques résultats d'homogénéisation périodique sur le problème d'advection diffusion stationnaire hautement oscillant

Cette annexe est consacrée à la théorie de l'homogénéisation du problème d'advection diffusion hautement oscillant en dimension $d \geqslant 2$ dans le cas périodique. On note $Y = (0,1)^d$ la cellule de périodicité. On définit Ω un domaine ouvert borné régulier de $\mathbb{R}^d$. Le problème s'écrit

$$-\operatorname{div}\left(A^\varepsilon \nabla u^\varepsilon\right) + b^\varepsilon \cdot \nabla u^\varepsilon = f \quad \text{dans } \Omega, \quad u^\varepsilon = 0 \quad \text{sur } \partial\Omega, \tag{A.1}$$

avec $b^\varepsilon \in (L^\infty(\Omega))^d$ et $f \in L^2(\Omega)$. On suppose que la matrice de diffusion A^ε est de la forme $A\left(\frac{x}{\varepsilon}\right)$ où $A \in (L^\infty(Y))^{d\times d}$ est Y-périodique, symétrique pour presque tout $y \in Y$. On suppose de plus qu'il existe $0 < \alpha_1 \leqslant \alpha_2$ tels que

$$\alpha_1|\xi|^2 \leqslant A(y)\xi \cdot \xi \leqslant \alpha_2|\xi|^2 \quad \text{pour tout } \xi \in \mathbb{R}^d, \text{ pp. } y \in Y. \tag{A.2}$$

Omettons un instant le champ d'advection b^ε dans le problème (A.1). Dans le cas du problème de diffusion pure hautement oscillant, la propriété (A.2) assure la coercivité du problème de manière uniforme par rapport à ε, ce qui permet en particulier de montrer la caractère bien posé du problème et que la suite u^ε est bornée dans $H_0^1(\Omega)$. La présence du champ d'advection b^ε représente une difficulté supplémentaire car le problème (A.1) n'est plus coercif en général. On étudie dans cette section l'homogénéisation du problème (A.1) pour différentes formes du champ d'advection b^ε. Les résultats présentés dans cette annexe sont classiques. On renvoie à [12, 25, 29, 52, 66, 69] pour une introduction à la théorie de l'homogénéisation.

i) Le cas $b^\varepsilon = b(x)$

Théorème A.1. *(adaptation de [4, Theorem 2.3, p.1494]) Soit u^ε la solution du problème (A.1) avec $b^\varepsilon = b(x)$. On suppose que $b \in (W^{1,\infty}(\Omega))^d$ et $\operatorname{div} b \leqslant 0$ dans Ω. La suite $(u^\varepsilon)_{\varepsilon>0}$ converge faiblement dans $H_0^1(\Omega)$, quand ε tend vers 0, vers $u^\star$, solution du problème*

$$-\operatorname{div}\left(A^\star \nabla u^\star\right) + b \cdot \nabla u^\star = f \quad \textit{dans } \Omega, \quad u^\star = 0 \quad \textit{sur } \partial\Omega,$$

avec

$$(A^\star)e_i = \int_Y A\,(y)\,\big(e_i + \nabla w_i\,(y)\,\big)\mathrm{d}y,$$

où w_i, $1 \leqslant i \leqslant d$, *est le correcteur solution du problème de cellule suivant*

$$\begin{cases} -\operatorname{div}\,(A(\nabla w_i + e_i)) = 0 \quad \textit{sur } Y, \\ y \to w_i(y) \quad Y\textit{-périodique.} \end{cases}$$

Démonstration. Plusieurs techniques de preuve sont possibles. On utilise ici la notion de convergence "à deux-échelles" [4]. On rappelle que la suite $(u^\varepsilon)_{\varepsilon>0}$ converge "à deux-échelles" vers $u_0 \in L^2(\Omega \times Y)$ si pour toute fonction $\psi \in C_c^\infty(\Omega, C_\#^\infty(Y))$, on a

$$\lim_{\varepsilon\to 0} \int_{\Omega^\varepsilon} u^\varepsilon(x)\psi\left(x, \frac{x}{\varepsilon}\right) = \int_{\Omega\times Y} u_0\psi.$$

On a supposé que $b \in W^{1,\infty}(\Omega)$ et $\operatorname{div} b \leqslant 0$ dans Ω, de telle sorte que le problème (A.1) est uniformément coercif. La suite u^ε est bornée dans $H_0^1(\Omega)$ et converge faiblement, à extraction près, dans $H_0^1(\Omega)$ vers u_0. On peut montrer qu'il existe $u_1 \in L^2(\Omega, H_\#^1(Y)/\mathbb{R})$ telle que la suite ∇u^ε converge, à extraction près, "à deux échelles" vers $\nabla u_0 + \nabla_y u_1$. Soient $\phi \in \mathcal{C}_c^\infty(\Omega)$ et $\phi_1 \in \mathcal{C}_c^\infty(\Omega; \mathcal{C}_\#^\infty(Y))$, on considère la fonction test suivante

$$\phi^\varepsilon(x) = \phi(x) + \varepsilon\phi_1\left(x, \frac{x}{\varepsilon}\right).$$

La formulation variationnelle du problème hautement oscillant nous donne

$$\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot \nabla\phi^\varepsilon + b \cdot \nabla u^\varepsilon \phi^\varepsilon = \int_\Omega f\phi^\varepsilon. \tag{A.3}$$

On fait apparaître le gradient de la fonction test

$$\nabla\phi^\varepsilon(x) = \nabla\phi(x) + \nabla_y\phi_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon\nabla\phi_1\left(x, \frac{x}{\varepsilon}\right).$$

On déduit alors de (A.3) que

$$\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot (\nabla\phi + \nabla_y\phi_1 + \varepsilon\nabla\phi_1) + b \cdot \nabla u^\varepsilon(\phi + \varepsilon\phi_1) = \int_\Omega f\phi^\varepsilon. \tag{A.4}$$

Les fonctions $A\,(y)\,(\nabla\phi(x) + \nabla_y\phi_1(x,y) + \varepsilon\nabla\phi_1(x,y))$ et $b\,(x)\,(\phi(x) + \varepsilon\phi_1(x,y))$ sont des fonctions test admissibles au sens de la convergence à deux échelles. On peut alors passer à la limite dans (A.4). On a alors

$$\int_{\Omega\times Y} A(\nabla u_0 + \nabla_y u_1) \cdot (\nabla\phi + \nabla_y\phi_1) + \int_{\Omega\times Y} (\nabla u_0 + \nabla_y u_1) \cdot b\phi = \int_\Omega f\phi, \tag{A.5}$$

La fonction $y \mapsto u_1(x,y)$ est Y-périodique, b et ϕ sont indépendants de y, donc $\displaystyle\int_{\Omega\times Y} (\nabla_y u_1 \cdot b)\phi = 0$. Par un argument de densité, on obtient (A.4) pour tout $(\phi, \phi_1) \in H_0^1(\Omega) \times L^2(\Omega; H_\#^1(Y)/\mathbb{R})$.

On peut découpler (A.5) de la façon suivante

$$\begin{cases} \displaystyle\int_Y A(\nabla u_0 + \nabla_y u_1)\cdot(\nabla_y \phi_1) = 0 \quad \forall \phi_1 \in H^1_{\#}(Y)/\mathbb{R}, \\ \displaystyle\int_{\Omega\times Y} A(\nabla u_0 + \nabla_y u_1)\cdot(\nabla \phi) + \int_{\Omega\times Y} (\nabla u_0 \cdot b)\phi = \int_\Omega f\phi \quad \forall \phi \in H^1_0(\Omega). \end{cases} \tag{A.6}$$

Soit w_i, $1 \leqslant i \leqslant d$, le correcteur solution du problème de cellule

$$\begin{cases} -\operatorname{div}\,(A(\nabla w_i + e_i)) = 0 \quad \text{dans } Y, \\ y \to w_i(y) \quad Y\text{-périodique.} \end{cases}$$

On peut écrire

$$u_1(x,y) = \sum_{i=1}^{d} w_i(y)\partial_{x_i} u_0(x).$$

On obtient alors à partir de (A.6) que u_0 est alors la solution du problème homogénéisé

$$\int_\Omega A^\star \nabla u_0 \cdot \nabla \phi + (b\cdot\nabla u_0)\phi = \int_\Omega f\phi \quad \forall \phi \in H^1_0(\Omega),$$

avec

$$A^\star e_i = \int_Y A\,(y)\,(e_i + \nabla w_i\,(y))\,\mathrm{d}y.$$

Le couple (u_0, u_1) est déterminé de manière unique, on peut donc en déduire que toute la suite u^ε converge faiblement vers u_0 dans $H^1_0(\Omega)$. □

ii) Le cas $b^\varepsilon = b\left(\frac{x}{\varepsilon}\right)$ avec b Y-périodique

Théorème A.2 ([12, Theorem 13.1, p.185])**.** *Soit u^ε la solution du problème (A.1) avec $b^\varepsilon(x) = b\left(\frac{x}{\varepsilon}\right)$ où b est Y-périodique. On suppose que $b \in (W^{1,\infty}(Y))^d$ et* div $b \leqslant 0$ *dans Y. La suite $(u^\varepsilon)_{\varepsilon>0}$ converge faiblement dans $H^1_0(\Omega)$, quand ε tend vers 0, vers $u^\star$ solution du problème*

$$-\operatorname{div}\,(A^\star\nabla u^\star) + b^\star\cdot\nabla u^\star = f \quad \text{dans } \Omega, \quad u^\star = 0 \quad \text{sur } \partial\Omega.$$

avec

$$(A^\star)e_i = \int_Y A\,(y)\,\big(e_i + \nabla w_i\,(y)\,\big)\mathrm{d}y,$$
$$b^\star\cdot e_i = \int_Y b(y)(e_i + \nabla w_i(y))\mathrm{d}y,$$

où w_i, $1 \leqslant i \leqslant d$, est le correcteur solution du problème de cellule suivant

$$\begin{cases} -\operatorname{div}\,(A(\nabla w_i + e_i)) = 0 \quad \text{dans } Y, \\ y \to w_i(y) \quad Y\text{-périodique.} \end{cases} \tag{A.7}$$

Démonstration. On suit le même cheminement que dans le cas précédent. On obtient que le problème est uniformément coercif par rapport à ε. La suite u^ε est alors bornée dans $H^1_0(\Omega)$ et converge faiblement, à extraction près, dans $H^1_0(\Omega)$ vers u_0. On peut montrer qu'il existe $u_1 \in L^2(\Omega, H^1_{\#}(Y)/\mathbb{R})$ telle que la suite ∇u^ε converge, à extraction près, "à deux échelles" vers

$\nabla u_0 + \nabla_y u_1$. Soient $\phi \in \mathcal{C}_c^\infty(\Omega)$ et $\phi_1 \in \mathcal{C}_c^\infty(\Omega; \mathcal{C}_\#^\infty(Y))$, on considère la fonction test suivante

$$\phi^\varepsilon(x) = \phi(x) + \varepsilon \phi_1\left(x, \frac{x}{\varepsilon}\right).$$

La formulation variationnelle du problème hautement oscillant nous donne

$$\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot \nabla \phi^\varepsilon + b^\varepsilon \cdot \nabla u^\varepsilon \phi^\varepsilon = \int_\Omega f \phi^\varepsilon. \tag{A.8}$$

On fait apparaître le gradient de la fonction test

$$\nabla \phi^\varepsilon(x) = \nabla \phi(x) + \nabla_y \phi_1\left(x, \frac{x}{\varepsilon}\right) + \varepsilon \nabla \phi_1\left(x, \frac{x}{\varepsilon}\right).$$

On déduit alors de (A.8) que

$$\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot (\nabla \phi + \nabla_y \phi_1 + \varepsilon \nabla \phi_1) + b^\varepsilon \cdot \nabla u^\varepsilon (\phi + \varepsilon \phi_1) = \int_\Omega f \phi^\varepsilon. \tag{A.9}$$

La fonction $\psi(x, y)$ est une fonction test admissible (au sens de la convergence à deux échelles) si elle est Y-périodique en y et satisfait

$$\lim_{\varepsilon \to 0} \int_\Omega \psi\left(x, \frac{x}{\varepsilon}\right)^2 = \int_{\Omega \times Y} \psi(x, y)^2.$$

Les fonctions $A(y)\left[\nabla \phi(x) + \nabla_y \phi_1(x, y) + \varepsilon \nabla \phi_1(x, y)\right]$ et $b(y)\left(\phi(x) + \varepsilon \phi_1(x, y)\right)$ sont donc des fonctions test admissibles. On peut alors passer à la limite dans (A.9), et on obtient que

$$\int_{\Omega \times Y} A(\nabla u_0 + \nabla_y u_1) \cdot (\nabla \phi + \nabla_y \phi_1) + \int_{\Omega \times Y} (\nabla u_0 + \nabla_y u_1) \cdot b \phi = \int_\Omega f \phi. \tag{A.10}$$

Par un argument de densité, on obtient (A.10) pour tout $(\phi, \phi_1) \in H_0^1(\Omega) \times L^2(\Omega; H_\#^1(Y)/\mathbb{R})$. On peut découpler la formulation de la façon suivante

$$\begin{cases} \displaystyle\int_Y A(\nabla u_0 + \nabla_y u_1) \cdot (\nabla_y \phi_1) = 0 \quad \forall \phi_1 \in H_\#^1(Y)/\mathbb{R}, \\ \displaystyle\int_{\Omega \times Y} A(\nabla u_0 + \nabla_y u_1) \cdot (\nabla \phi) + \int_{\Omega \times Y} (\nabla u_0 + \nabla_y u_1) \cdot b \phi = \int_\Omega f \phi \quad \forall \phi \in H_0^1(\Omega). \end{cases} \tag{A.11}$$

Soit w_i, $1 \leqslant i \leqslant d$, le correcteur solution du problème (A.7). En utilisant le fait que

$$u_1(x, y) = \sum_{i=1}^d w_i(y) \partial_{x_i} u_0(x),$$

on obtient à partir de (A.11) que u_0 est la solution du problème homogénéisé suivant

$$\int_\Omega A^\star \nabla u_0 \cdot \nabla \phi + (b^\star \cdot \nabla u_0) \phi = \int_\Omega f \phi \quad \forall \phi \in H_0^1(\Omega),$$

avec

$$A^\star e_i = \int_Y A\,(y)\,(e_i + \nabla w_i\,(y))\mathrm{d}y,$$
$$b^\star \cdot e_i = \int_Y b(y)(e_i + \nabla w_i(y))\mathrm{d}y.$$

Le couple (u_0, u_1) est déterminé de manière unique, on peut donc en déduire que toute la suite u^ε converge faiblement vers u_0 dans $H_0^1(\Omega)$. □

iii) Le cas $b^\varepsilon = \dfrac{b}{\varepsilon}$ avec b constant non nul

Théorème A.3. *Soit u^ε la solution du problème (A.1) avec $b^\varepsilon(x) = \dfrac{b}{\varepsilon}$ où b est constant non nul. La suite $(u^\varepsilon)_{\varepsilon>0}$ converge faiblement vers 0 dans $H_0^1(\Omega)$, quand ε tend vers 0.*

Démonstration. Le problème est coercif, ce qui implique que la suite u^ε est bornée dans $H_0^1(\Omega)$. À extraction près, la suite converge faiblement dans $H_0^1(\Omega)$ vers $u^\star$. Soit $\phi \in H_0^1(\Omega)$. La formulation variationnelle de (A.1) s'écrit

$$\varepsilon \int_\Omega A^\varepsilon \nabla u^\varepsilon \nabla \phi + \int_\Omega (b \cdot \nabla u^\varepsilon)\phi = \varepsilon \int_\Omega f\phi.$$

En passant à la limite $\varepsilon \to 0$, on obtient que

$$\forall \phi \in H_0^1(\Omega), \quad \int_\Omega (b \cdot \nabla u^\star)\phi = 0,$$

ce qui implique $b \cdot \nabla u^\star = 0$ dans Ω. On peut supposer que b est dirigé selon e_x. On a donc $u^\star = u^\star(y)$. La condition aux bord de Dirichlet homogène nous permet de déduire que $u^\star = 0$. □

iv) Le cas $b^\varepsilon = \dfrac{b\left(\frac{x}{\varepsilon}\right)}{\varepsilon}$ avec b Y-périodique, de moyenne nulle

Théorème A.4. *Soit u^ε la solution du problème (A.1) avec $b^\varepsilon(x) = \dfrac{b\left(\frac{x}{\varepsilon}\right)}{\varepsilon}$ où b est Y-périodique. On suppose que $b \in (W^{1,\infty}(Y))^d$, $\operatorname{div} b = 0$ dans Y et $\displaystyle\int_Y b = 0$. La suite $(u^\varepsilon)_{\varepsilon>0}$ converge faiblement dans $H_0^1(\Omega)$, quand ε tend vers 0, vers $u^\star$, solution du problème*

$$-\operatorname{div}\,(A^\star \nabla u^\star) = f \quad \text{dans } \Omega, \quad u^\star = 0 \quad \text{sur } \partial\Omega,$$

avec

$$A^\star e_i = \int_Y A\,(y)\,(e_i + \nabla w_i\,(y)) + b w_i\,(y)\,\mathrm{d}y,$$

où w_i, $1 \leqslant i \leqslant d$, est le correcteur solution du problème de cellule suivant

$$\begin{cases} -\operatorname{div}\,(A(\nabla w_i + e_i)) - b \cdot (\nabla w_i + e_i) = 0 \quad \text{dans } Y, \\ y \to w_i(y) \quad Y\text{-périodique.} \end{cases} \tag{A.12}$$

Démonstration. Plusieurs techniques de preuve sont possibles. On utilise la méthode de la fonction test oscillante de Tartar [25, Chapter 8]

$$\phi^\varepsilon(x) = \phi(x) + \varepsilon \sum_{i=1}^{d} w_i\left(\frac{x}{\varepsilon}\right) \partial_{x_i}\phi(x),$$

avec $\phi \in \mathcal{C}_c^\infty(\Omega)$ et où w_i, $1 \leqslant i \leqslant d$, sont les correcteur solution du problème de cellule (A.12). On peut appliquer cette méthode pour démontrer les théorèmes A.1 et A.2 sous l'hypothèse, plus restrictive, que div $b = 0$ dans Y. Avec les hypothèses div $b = 0$ et $\displaystyle\int_Y b = 0$, les correcteurs sont bien définis. La formulation variationnelle de (A.1) nous donne

$$\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot \nabla \phi^\varepsilon - b^\varepsilon u^\varepsilon \cdot \nabla \phi^\varepsilon = \int_\Omega f \phi^\varepsilon, \tag{A.13}$$

où on a intégré par parties le terme d'advection. On fait apparaître le gradient de la fonction test

$$\begin{aligned}
\nabla\phi^\varepsilon(x) &= \nabla\phi(x) + \sum_{i=1}^{d} \nabla w_i\left(\frac{x}{\varepsilon}\right) \partial_{x_i}\phi(x) + \varepsilon \sum_{i=1}^{d} w_i\left(\frac{x}{\varepsilon}\right) \nabla\partial_{x_i}\phi(x) \\
&= \sum_{i=1}^{d} \left(e_i + \nabla w_i\left(\frac{x}{\varepsilon}\right)\right) \partial_{x_i}\phi(x) + \varepsilon \sum_{i=1}^{d} w_i\left(\frac{x}{\varepsilon}\right) \nabla\partial_{x_i}\phi(x).
\end{aligned}$$

On pose

$$r^\varepsilon = -\int_\Omega (A^\varepsilon \nabla u^\varepsilon) \cdot \left(\varepsilon \sum_{i=1}^{d} w_i\left(\frac{x}{\varepsilon}\right) \nabla\partial_{x_i}\phi(x)\right) + \int_\Omega f\varepsilon \left(\sum_{i=1}^{d} w_i\left(\frac{x}{\varepsilon}\right) \partial_{x_i}\phi(x)\right).$$

On déduit alors de (A.13) que

$$\begin{aligned}
&\int_\Omega A^\varepsilon \nabla u^\varepsilon \cdot \left(\sum_{i=1}^{d} (e_i + \nabla w_i^\varepsilon)\partial_{x_i}\phi\right) - b^\varepsilon u^\varepsilon \cdot \nabla\phi^\varepsilon = \int_\Omega f\phi + r^\varepsilon \\
&\int_\Omega \nabla u^\varepsilon \cdot \left(\sum_{i=1}^{d} A^\varepsilon (e_i + \nabla w_i^\varepsilon)\partial_{x_i}\phi\right) - b^\varepsilon u^\varepsilon \cdot \nabla\phi^\varepsilon = \int_\Omega f\phi + r^\varepsilon \\
&\int_\Omega -u^\varepsilon \mathrm{div}\left(\sum_{i=1}^{d} A^\varepsilon (e_i + \nabla w_i^\varepsilon)\partial_{x_i}\phi\right) - b^\varepsilon u^\varepsilon \cdot \nabla\phi^\varepsilon = \int_\Omega f\phi + r^\varepsilon \\
&\int_\Omega -u^\varepsilon \left(\frac{1}{\varepsilon} g^\varepsilon + h^\varepsilon\right) = \int_\Omega f\phi + r^\varepsilon,
\end{aligned} \tag{A.14}$$

avec

$$\begin{aligned}
g^\varepsilon(x) &= \sum_{i=1}^{d} \left[\mathrm{div}\,(A(e_i + \nabla w_i)) + b \cdot (e_i + \nabla w_i)\right]\left(\frac{x}{\varepsilon}\right) \partial_{x_i}\phi(x), \\
h^\varepsilon(x) &= \sum_{i=1}^{d} \left\{A\left(\frac{x}{\varepsilon}\right)\left(e_i + \nabla w_i\left(\frac{x}{\varepsilon}\right)\right) + b\left(\frac{x}{\varepsilon}\right) w_i\left(\frac{x}{\varepsilon}\right)\right\} \cdot \nabla\partial_{x_i}\phi(x).
\end{aligned}$$

D'après la définition du problème du correcteur, $g^\varepsilon(x) = 0$. De plus h^ε est bornée dans $L^2(\Omega)$, $h^\varepsilon \in L^2_\#(Y; C(\Omega))$, donc la suite h^ε converge faiblement vers $\sum_{i=1}^{d} (A^\star e_i) \cdot \nabla \partial_{x_i} \phi$ dans $L^2(\Omega)$ où

$$A^\star e_i = \int_Y A(y)\left(e_i + \nabla w_i(y)\right) + b(y)\, w_i(y)\, \mathrm{d}y.$$

En utilisant la régularité de ϕ, on obtient que $\sum_{i=1}^{d} A^\star e_i \cdot \nabla \partial_{x_i} \phi = \operatorname{div}\left((A^\star)^T \nabla \phi\right)$. La forme bilinéaire du problème est coercive uniformément en ε car $\operatorname{div} b = 0$, la suite u^ε est donc bornée dans $H^1_0(\Omega)$. Par suite, on peut en déduire qu'à extraction près, la suite u^ε converge faiblement vers $u^\star$ dans $H^1_0(\Omega)$ et fortement dans $L^2(\Omega)$. On peut alors passer à la limite dans (A.14) et obtenir

$$\begin{aligned} \int_\Omega -u^\star \operatorname{div}\left((A^\star)^T \nabla \phi\right) &= \int_\Omega f\phi \\ \int_\Omega A^\star \nabla u^\star \cdot \nabla \phi &= \int_\Omega f\phi. \end{aligned} \tag{A.15}$$

Par un argument de densité on obtient (A.15) pour tout $\phi \in H^1_0(\Omega)$. Ceci implique l'unicité de $u^\star \in H^1_0(\Omega)$. On peut donc en déduire que toute la suite converge vers $u^\star \in H^1_0(\Omega)$.

□

Annexe B

Stabilisation de problèmes non coercifs via une méthode numérique utilisant la mesure invariante

Cette annexe concerne l'étude du chapitre 3 et reprend l'intégralité d'une note rédigée en collaboration avec Frédéric Legoll et Claude Le Bris et publiée dans Comptes Rendus Mathématiques [61].

Abstract

Nous nous intéressons à un problème d'advection-diffusion non coercif où l'advection domine. Nous présentons une approche numérique possible, à notre connaissance nouvelle, basée sur l'utilisation de la mesure invariante associée au problème. Nous démontrons sur l'exemple traité que l'approche permet de définir une approximation éléments finis du problème bien posée, et ce inconditionnellement en la taille du maillage. Plusieurs variantes de l'approche sont possibles, dont une, qui s'avère stable, conduit à des résultats numériques de qualité tout à fait comparable à une méthode classique de stabilisation sur l'équation considérée. Ceci suggère une piste possible, générale, pour toute une classe de problèmes non coercifs.

Abstract

We study an advection-diffusion equation that is both non-coercive and advection-dominated. We present a possible numerical approach, to our best knowledge new, and based on the invariant measure associated to the original equation. We show that the approach allows for an unconditionally well-posed finite element approximation. Two variants of the approach are studied. One of them is stable, and as accurate as a classical stabilization approach. This suggests a possible general strategy, applicable to a large class of non coercive problems.

Abridged English version

We study the advection-diffusion equation $-\Delta u + \mathbf{b} \cdot \nabla u = f$ (equation (B.1) of the French version) in the regime where the associated bilinear form is not coercive and where the advection dominates. As is well-known, a classical proof of the well-posedness of this non coercive equation proceeds by the application of the Fredholm alternative. Well-posedness is actually, using the Banach-Necas-Babuska Theorem, equivalent to the celebrated inf-sup condition and an additional condition (see (B.2) below and, e.g., [3] for a comprehensive exposition of the theory and general references therein for the study and approximation of (B.1)). Well-posedness of the Galerkin approximation of (B.1) follows, at least for a sufficiently small mesh size and for a suitable class of finite element approximations. In fact, as is also well known, the inf-sup condition may be established independently, using the notion of invariant measure associated to the original equation, namely the (positive, normalized) solution σ of the adjoint equation $-\mathrm{div}\,(\nabla\sigma + \sigma\,\mathbf{b}) = 0$ (see (B.3) below) supplied with adequate boundary conditions. The non coercive bilinear form $a(u, v)$ corresponding to the advection-diffusion equation (B.1) then reads, for a test function $v = \sigma u$, as (B.5). The bilinear form $(u, v) \mapsto a(u, \sigma v)$ is hence coercive. The inf-sup condition readily follows. Although the above is by now classical, the approach consisting in using the invariant measure as a tool for forcing coerciveness in an otherwise non coercive equation seems to not have been explored from a numerical perspective. In short, our numerical approach consists in constructing, using an auxiliary finite element computation, an approximation of the invariant

measure σ, and then performing a Petrov-Galerkin approximation of the original advection-diffusion equation (B.1) using test functions approximating the product form $\sigma\, v$. The definite added value of the approach is that it provides an unconditionally well-posed approximation, irrespective of the discretization parameter –the meshsize– adopted for approximating u, provided σ itself is correctly approximated. This property may be most useful in problems where one can only afford a coarse approximation of u. Multiscale problems are prototypical examples of such a context. In addition, in the advection dominated context, the approach enjoys particular stability properties that lead to numerical results qualitatively comparable to those obtained with classical, state-of-the-art stabilization approaches [2, 4, 5, 7]. The results of this Note, see in particular Tables B.1 and B.2, show that the approach is accurate, robust and can be made effective in terms of computational cost. Applications to several other, more general contexts, may be envisioned.

B.1 Introduction et motivation

Nous étudions dans cette Note la question très classique de l'approximation par éléments finis d'une équation d'advection-diffusion

$$-\Delta u + \mathbf{b} \cdot \nabla u = f, \tag{B.1}$$

pour des données telles que l'équation est non coercive et dominée par l'advection. Il est bien connu que l'approximation numérique classique par éléments finis est alors instable, et requiert une méthode de stabilisation. La littérature regorge de méthodes dans ce sens (cf. par exemple les célèbres contributions [2, 4, 5] et l'ouvrage [7]), au moins dans le cas coercif (les travaux dans le cas non coercif étant, semble-t-il, plus rares), et il est impossible de citer ici toutes les contributions, y compris récentes : nous dressons un état des lieux dans [6]. Nous présentons ici une approche à notre connaissance nouvelle, reproduisant au niveau numérique l'observation théorique essentielle fournissant une des approches possibles pour démontrer le caractère bien posé du problème. Nous obtenons ainsi une approche différente de celles de la littérature, systématiquement coercive et, dans une très large gamme de régimes, stable. Cette nouvelle approche se compare très favorablement aux approches déjà connues.

Supposons plus précisément (pour fixer les idées, mais la suite est largement adaptable à d'autres cas *modulo* des ajustements techniques que nous omettons, et qui sont précisés dans [6]) que l'équation (B.1) est posée sur un domaine borné régulier Ω de $\mathbb{R}^d$ et qu'on la munit de conditions au bord de Dirichlet homogènes $u = 0$ sur $\partial\Omega$. Supposons aussi que le champ $\mathbf{b}$ est donné, qu'il possède toute la régularité nécessaire pour donner un cadre rigoureux aux résultats cités ci-dessous, mais qu'il ne présente pas les propriétés classiques qui rendent le problème coercif (typiquement $\mathbf{b}$, ou sa divergence, assez petits dans une norme adéquate). Une des approches théoriques classiques pour prouver que le problème est bien posé (au sens de Hadamard) passe par l'alternative de Fredholm. Il est aussi connu que le caractère bien posé est, *via* la théorie de Banach-Nečas-Babuska, équivalent à la célèbre *condition inf-sup* (voir par exemple [3, p. 85])

$$\exists \alpha > 0 \quad \text{tel que} \quad \inf_{w\in W} \sup_{v\in V} \frac{a(w,v)}{\|w\|_W \|v\|_V} \geqslant \alpha > 0, \tag{B.2}$$

(où on a bien sûr défini ici $a(w,v) = \int_\Omega \nabla w \cdot \nabla v + \int_\Omega (\mathbf{b}\cdot\nabla w)\, v$ et $V = W = H_0^1(\Omega)$), à laquelle on adjoint une autre condition classique : $\forall v \in V,\ (\forall w \in W, a(w,v) = 0) \Rightarrow v = 0$. Soit afin de rendre *explicite* la constante α de (B.2), soit afin de fournir une preuve *autonome* du caractère bien posé, on peut aussi prouver que cette condition (B.2) est vérifiée par le problème (B.1) en considérant la mesure invariante associée au problème (B.1). Cette mesure est la solution σ de

$$-\operatorname{div}(\nabla\sigma + \sigma\,\mathbf{b}) = 0, \tag{B.3}$$

satisfaisant les propriétés $\sigma(x) \geqslant \underline{\sigma} > 0$ presque partout dans Ω, $|\Omega|^{-1}\int_\Omega \sigma = 1$ et une condition au bord bien choisie (pour simplifier, pensons à la condition de Neumann naturelle, mais d'autres choix peuvent éventuellement être faits, selon les conditions au bord imposées dans (B.1)). L'existence et l'unicité de σ convenable sont obtenues par la théorie de Fredholm. Schématiquement, la preuve de la condition inf-sup (B.2) repose alors sur la multiplication de (B.1) par la fonction produit $\sigma\, u$, et une intégration sur le domaine Ω. On écrit, à des termes de bord près qui, dans les conditions prises ci-dessus, s'annulent tous (voir le détail ci-dessous en (B.7)),

$$a(u,\sigma\,u) = \int_\Omega (-\Delta u + \mathbf{b}\cdot\nabla u)\,\sigma u = \int_\Omega \sigma\,|\nabla u|^2 - \frac{1}{2}\int_\Omega \Big(\operatorname{div}(\nabla\sigma + \sigma\,\mathbf{b})\Big)\, u^2, \tag{B.4}$$

et donc, en utilisant (B.3),

$$a(u,\sigma\,u) = \int_\Omega \sigma\,|\nabla u|^2. \tag{B.5}$$

Ceci permet d'obtenir facilement (B.2). Il reste alors, en pratique, à utiliser une discrétisation par éléments finis qui permette de vérifier aussi cette condition inf-sup au niveau discret, au moins, "par continuité", pour un pas de maillage h assez petit. Il s'ensuit un problème discret bien posé. Tout cela est désormais classique, mais, curieusement, cette technique de transformation d'un problème non coercif en un problème modifié coercif, *via* l'utilisation de la mesure invariante σ, ne semble pas avoir été exploitée au niveau numérique (sauf dans le cas où $\mathbf{b}$ est irrotationel [1], pour lequel σ est alors analytiquement connu). L'égalité (B.5) suggère pourtant une approximation de type Petrov-Galerkin sur (B.1), usant de fonctions tests produits σv, au lieu d'une approche classique Galerkin, de sorte que la coercivité (B.5) soit directement satisfaite au niveau discret. Il est facile de se rendre compte que, au moins dans le cas décrit ci-dessus, une telle approche est aussi une approche Galerkin sur l'équation *modifiée*

$$-\operatorname{div}(\sigma\,\nabla u) + (\nabla\sigma + \sigma\,\mathbf{b})\cdot\nabla u = \sigma\, f, \tag{B.6}$$

où on observe que, par construction, le champ $\nabla\sigma + \sigma\,\mathbf{b}$ est, à cause de (B.3), à divergence nulle. D'où encore la coercivité immédiate. On comprend immédiatement l'intérêt d'une telle approche, puisqu'elle donne par nature un caractère bien posé (et, qui plus est, coercif) au problème discret associé à (B.1), et ce *uniformément en la taille du maillage*. Pour des problèmes coûteux où la discrétisation ne peut être que grossière, avoir un problème discret bien posé et stable est un objectif naturel. On pense ainsi, par exemple mais pas seulement, au cas où on superposerait au problème (B.1) ci-dessus un aspect multiéchelle (en remplaçant l'opérateur de diffusion par un opérateur $\operatorname{div}(a(x/\varepsilon)\,\nabla\cdot)$ à coefficients rapidement oscillants, ou bien en le considérant sur un domaine Ω_ε à la géométrie complexe). Un autre avantage de l'approche est qu'on peut en faire

une analyse numérique en utilisant des arguments standards. Le prix à payer pour employer une telle approche est l'approximation numérique de la mesure σ, avec la précision requise (on verra qu'il suffit en pratique d'utiliser un maillage pour σ un peu plus fin que celui pour u), puisqu'il est rare que cette mesure soit connue explicitement analytiquement.

L'objectif de cette Note est d'explorer cette piste, sur l'équation (B.1) spécifiquement. Les résultats numériques présentés ci-dessous montrent que, en qualité, elle est tout à fait comparable aux approches classiques (de type SUPG, GLS ou DW, cf. [7]), voire meilleure, et que, même si le calcul additionnel de la mesure invariante est un surcoût, l'approche peut aussi être rendue globalement compétitive dans plusieurs contextes (cf. la fin de la Section 2). Rien n'interdit de penser que la même approche peut être appliquée à d'autres cas, y compris potentiellement des cas non linéaires, et des cas inaccessibles à des approches plus classiques. Les résultats résumés dans cette Note sont exposés plus en détail, et complétés, dans le chapitre 3.

B.2 Mise en oeuvre de l'approche

B.2.1 Différentes mesures possibles

Le point clé de l'approche suggérée est bien évidemment la détermination de la mesure invariante. Il convient tout d'abord de souligner une certaine liberté dans le choix de cette mesure. En effet, le passage de la forme non coercive initiale à la forme coercive suggérée par (B.5) repose, via une intégration par partie, sur trois ingrédients : (i) l'annulation, grâce à l'équation adjointe (B.3), du terme de volume causant la non coercivité, (ii) l'élimination des termes de bord, et (iii) la positivité de la mesure invariante, ou plus exactement, son caractère isolé de zéro. Typiquement, l'intégration par partie est de la forme :

$$\int_\Omega (-\Delta u + \mathbf{b}\cdot\nabla u)\,\sigma v = \int_\Omega \sigma\,\nabla u\cdot\nabla v + \int_\Omega (\nabla\sigma + \sigma\,\mathbf{b})\cdot\nabla u\, v - \int_{\partial\Omega}(\nabla u\cdot\mathbf{n})\,\sigma\, v, \qquad \text{(B.7)}$$

où le dernier terme disparaît au vu des conditions de Dirichlet homogènes posées sur u et donc v. Dans ce cas, et l'observation est en fait assez générale, on a donc un choix libre de la condition au bord à imposer sur la solution de (B.3), pourvu que la fonction σ ainsi définie existe et soit (positive et) isolée de zéro.

Deux choix au moins peuvent paraître naturels. Le premier est d'imposer la condition de Neumann associée à l'équation (B.3). On définit ainsi la mesure $\sigma_1 > 0$, normalisée par $|\Omega|^{-1}\int_\Omega \sigma_1 = 1$, solution de

$$\begin{cases} -\mathrm{div}\,(\nabla\sigma_1 + \sigma_1\,\mathbf{b}) = 0 & \text{dans } \Omega, \\ (\nabla\sigma_1 + \sigma_1\,\mathbf{b})\cdot\mathbf{n} = 0 & \text{sur } \partial\Omega, \end{cases} \qquad \text{(B.8)}$$

dont on peut montrer qu'elle existe, sous de bonnes hypothèses de régularité de Ω et $\mathbf{b}$, par l'alternative de Fredholm suivie d'une application du principe du maximum. Ce choix peut cependant paraître surprenant car, dans le cas particulier où le champ $\mathbf{b}$ est à divergence nulle, et où le problème original (B.1) est donc *ipso facto* coercif, on n'a $\sigma_1 \equiv 1$ que si $\mathbf{b}\cdot\mathbf{n} \equiv 0$ sur $\partial\Omega$. On verra pourtant que ce choix de mesure est le meilleur qu'on puisse prendre dans le cas spécifique traité ici. Motivé par la discussion ci-dessus, on peut alternativement penser à choisir σ de sorte que $\sigma \equiv 1$ dès que $\mathrm{div}\,\mathbf{b} = 0$. D'où l'idée de considérer la mesure σ_2 solution de l'équation similaire à (B.8) où on a remplacé la condition de bord $(\nabla\sigma_1 + \sigma_1\,\mathbf{b})\cdot\mathbf{n} = 0$ par $(\nabla\sigma_2 + \sigma_2\,\mathbf{b})\cdot\mathbf{n} = \mathbf{b}\cdot\mathbf{n} - |\partial\Omega|^{-1}\int_{\partial\Omega}\mathbf{b}\cdot\mathbf{n}$ sur $\partial\Omega$. Ce choix, quoique plus consistant en un

certain sens, s'avérera moins efficace dans les tests numériques menés. Retenons que, dans un cadre de travail fixé (conditions aux bord imposées pour (B.1), régularité des données), on peut utiliser la flexibilité dont on dispose pour définir la mesure invariante pour choisir la meilleure d'entre elles.

B.2.2 Discrétisation

Une fois la mesure invariante choisie, l'approche consiste à

(i) si sa solution n'est pas connue explicitement analytiquement, résoudre numériquement l'équation (B.8) (ou une équation analogue) par une méthode d'éléments finis, pour un certain maillage de taille h, obtenant ainsi une approximation σ_h de la mesure σ. Il faut veiller à ce que cette approximation soit positive, et suffisamment précise pour ne pas hypothéquer la qualité du calcul de u. Pour ce faire, nous avons utilisé une formulation variationnelle stabilisée de (B.8).

(ii) construire une approximation Petrov-Galerkin $(U_H, V_H = \sigma_h U_H)$ de l'équation (B.1) (équivalente à une approximation Galerkin de (B.6) sur U_H), à partir d'un espace d'approximation éléments finis U_H de départ, pour une certaine taille de maillage H. Dans les tests numériques présentés dans la section suivante, U_H est l'espace des éléments finis continus $\mathbb{P}1$ sur un maillage régulier de taille H. Comme le champ $\nabla\sigma_h + \sigma_h \mathbf{b}$ n'est pas à divergence nulle pour $h > 0$, on utilise l'astuce classique d'antisymétrisation du terme d'advection dans la formulation variationnelle de (B.6), afin d'assurer la coercivité au niveau discret.

Les deux tailles de maillage h et H, utilisées respectivement dans les étapes (i) et (ii), ne sont pas nécessairement identiques. Quoi qu'il en soit, l'étape (i) est un surcoût qui, bon an mal an, double le coût de (ii) pour une unique résolution de (B.1). Cependant, on doit garder à l'esprit que l'étape (i) n'est effectuée qu'une seule fois, à $\mathbf{b}$ donné, et ne dépend pas de f. Ainsi, dans un calcul répétitif comme dans un problème inverse, où on doit résoudre plusieurs fois le problème (B.1) pour des données f différentes, le surcoût de (i) devient négligeable. Il l'est de même dans le cas de la résolution de l'équation d'advection-diffusion transitoire, pour un champ $\mathbf{b}$ lui indépendant du temps, où on résout, après semi-discrétisation Euler implicite en temps, une équation du type (B.1), appelant un unique calcul de mesure invariante à l'étape (i).

Signalons aussi que l'étape (ii) requiert souvent l'utilisation d'un préconditionneur, la matrice de rigidité issue de (B.5) ayant, pour une mesure σ présentant de fortes variations, un conditionnement délicat. Dans les tests pratiqués, un simple préconditionnement diagonal a suffi à rendre la résolution très efficace.

B.3 Résultats numériques

Après une série de calculs en dimension 1, présentés dans [6], l'approche utilisant la mesure invariante a été testée en dimension 2, pour différents choix de champs $\mathbf{b}$. Nous comparons nos approches avec une méthode classique d'éléments finis $\mathbb{P}1$ et une telle méthode stabilisée par l'approche Galerkin Least-Squares (GLS, cf. [4, 5]), pour distinguer les cas stables des cas instables.

Nous commençons par considérer un champ $\mathbf{b}$ irrotationnel, c'est-à-dire, en un sens intuitif (penser à la décomposition de Helmholtz), un champ aussi éloigné que possible d'un champ à

divergence nulle, pour lequel le problème serait coercif. Dans ce cas, $\mathbf{b} = \nabla\Phi$ dérive donc d'un potentiel, et la mesure σ_1 solution de (B.8) est alors explicitement connue : $\sigma_1 = \exp(-\Phi)$, à un facteur de normalisation près. L'objet du test est alors de vérifier (i) que la connaissance exacte de la mesure permet de construire une méthode efficace, (ii) que cette efficacité reste robuste quand on remplace σ_1 par une approximation numérique –y compris relativement grossière–, (iii) que l'approche se compare favorablement à une méthode classique. Les résultats présentés confirment cela (notons que, pour le cas de tels champs $\mathbf{b}$, la question (i) a été abordée dans [1]). Parallèlement, les mêmes questions concernant σ_2 (cette fois nécessairement numériquement approchée) amènent une réponse plus nuancée, comme les résultats le montreront de manière évidente.

Un autre intérêt de ce cas irrotationnel est qu'il met en exergue un point intéressant de notre approche. Pour un tel champ $\mathbf{b} = \nabla\Phi$, non seulement σ_1 est connue explicitement, mais en plus cette mesure permet de construire une formulation qui est non seulement coercive mais inconditionnellement stable, puisque le terme de transport $\nabla\sigma_1 + \sigma_1\,\mathbf{b} = 0$ disparaît ! Cette propriété reste "quasiment" vraie pour une approximation $(\sigma_1)_h$ de σ_1, mais, sauf cas particulier, n'est plus vraie pour le choix σ_2. On comprend donc par cet argument intuitif la stabilisation automatiquement induite par l'approche *quand* la mesure est bien choisie. Dans le cas général, nous croyons que cette supériorité du choix σ_1 "stabilisant" reste globalement vraie, notamment parce que *intuitivement*, $\nabla\sigma + \sigma\,\mathbf{b}$ reste plus petit pour $\sigma = \sigma_1$ (vu la condition de nullité au bord) que pour $\sigma = \sigma_2$. Les tests numériques du paragraphe suivant montrent cette supériorité, bien que nous n'ayons pas d'argument théorique précis sur ce point.

Les données utilisées sont les suivantes : $(x, y) \in \Omega =]0,1[^2 \subset \mathbb{R}^2$, $f = 1$, $\mathbf{b} = (\delta^{-1} + \lambda\cos^2(2\pi x), \delta^{-1})^T$. En jouant sur la valeur des paramètres δ et λ, on peut rendre le problème coercif ou non, stable ou instable. Pour être bref, nous ne montrons ici les résultats que dans le cas non coercif et instable ($\delta = 1/64$, $\lambda = 50.34$) et on renvoie au chapitre 3 pour les résultats complets. La mesure invariante σ_1 est connue explicitement à partir de la donnée de $\mathbf{b}$ ci-dessus, mais on l'approche aussi. La discrétisation de l'étape (i), pour déterminer une approximation de σ_1 ou de σ_2, se fait alors en éléments finis $\mathbb{P}1$ sur un maillage régulier triangulaire de taille h. Celle de l'étape (ii) s'effectue en éléments finis $\mathbb{P}1$ sur un maillage régulier de taille $H = 1/16$. Pour simplifier, on a pris ici H/h entier, mais cela n'est pas requis par l'approche.

Les résultats sont détaillés dans la Table B.1. On mesure l'erreur relative par rapport à une solution de référence calculée sur un maillage très fin. L'erreur H^1 indiquée s'entend *hors couche limite*, comme cela est usuel pour les problèmes où l'advection domine. On observe la supériorité, l'efficacité et la stabilité de la méthode basée sur la mesure σ_1.

Méthode	$\mathbb{P}1$	$\mathbb{P}1$ GLS	$(\mathbb{P}1, \sigma_1\mathbb{P}1)$	$(\mathbb{P}1, (\sigma_1)_h\mathbb{P}1)$ $H/h = 7$	$(\mathbb{P}1, (\sigma_2)_h\mathbb{P}1)$ $H/h = 7$	$(\mathbb{P}1, (\sigma_2)_h\mathbb{P}1)$ GLS $H/h = 7$
Erreur L^2	0.208	0.214	0.200	0.207	0.195	0.226
Erreur H^1	0.479	0.0488	0.0199	0.0218	0.399	0.0536

TABLE B.1 Erreurs relatives dans le cas non coercif instable ($\mathbf{b}$ irrotationnel, $H = 1/16$).

Nous passons ensuite au cas général. Comme cas représentatif d'un champ $\mathbf{b}$ général, non irrotationnel, on choisit de prendre $\mathbf{b} = (1 + 50.34\cos^2(2\pi x) + 64y, 64(1-x))^T$, toutes autres données égales par ailleurs aux données du cas précédent. Ce champ $\mathbf{b}$ a été ajusté pour avoir

un cas non coercif et instable. Les résultats sont détaillés dans la Table B.2 et confirment les conclusions déjà indiquées.

Méthode	$\mathbb{P}1$	$\mathbb{P}1$ GLS	$(\mathbb{P}1, (\sigma_1)_h\mathbb{P}1)$ $H/h = 4$	$(\mathbb{P}1, (\sigma_2)_h\mathbb{P}1)$ $H/h = 4$	$(\mathbb{P}1, (\sigma_2)_h\mathbb{P}1)$ GLS $H/h = 4$
Erreur L^2	0.165	0.222	0.154	0.160	0.158
Erreur H^1	0.414	0.104	0.0221	0.0221	0.029

TABLE B.2 Erreurs relatives dans le cas non coercif instable (**b** "quelconque", $H = 1/16$).

Acknowledgements The authors thank Yves Achdou and Olivier Pironneau for helpful discussions. The work of the authors is partially supported by ONR under grant N00014-15-1-2777 and EOARD under grant FA8655-13-1-3061.

Références

[1] F. Brezzi, L.D. Marini, P. Pietra, Two-dimensional exponential fitting and applications to drift-diffusion models, SIAM J. Numer. Anal. 26 (6) (1989) 1342–1355.

[2] A.N. Brooks, T. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg. 32 (1-3) (1982) 199–259.

[3] A. Ern, J.-L. Guermond, **Theory and practice of finite elements**, Applied Mathematical Sciences, vol. 159, Springer-Verlag, New York, 2004.

[4] L.P. Franca, S.L. Frey, T.J.R. Hugues, Stabilized finite element methods : I. Application to the advective-diffusive model, Comput. Methods Appl. Mech. Engrg. 95 (1992) 253–276.

[5] C. Johnson, U. Nävert, J. Pitkäranta, Finite element methods for linear hyperbolic problems, Comput. Methods Appl. Mech. Engrg. 45 (1984) 285–312.

[6] F. Madiot, Multiscale finite element methods for advection diffusion problems, Université Paris-Est, thèse en préparation.

[7] A. Quarteroni, A. Valli, **Numerical approximation of partial differential equations**, Springer Series in Computational Mathematics, vol. 23, Springer-Verlag, Berlin, 1994.