Numerical methods for the study of fluctuations in multi-scale materials and related problems

Pierre-Loïk Rothé

UNIVERSITÉ PARIS-EST

École doctorale Science de l'Ingénierie et Environnement (SIE)

Thèse de doctorat

Discipline : Sciences de l'ingénieur

présentée par

Pierre-Loïk ROTHÉ

Méthodes numériques pour l'étude des fluctuations dans les matériaux multi-échelles et problèmes reliés

---

Thèse dirigée par Frédéric LEGOLL
préparée au CERMICS et au Laboratoire Navier, ENPC

Soutenue le 12 décembre 2019 devant un Jury composé de :

Rapporteur	Anthony Nouy	Centrale Nantes
Rapporteur	Alexei Lozinski	Université de Franche-Comté
Examinateur	Sébastien Brisard	École des Ponts ParisTech
Examinateur	Julian Fischer	IST Austria
Examinatrice	Sonia Fliss	ENSTA Paris
Directeur de Thèse	Frédéric LEGOLL	École des Ponts ParisTech

« Tout jugement oscille sur la pointe de l'erreur, dit Leto. Prétendre à l'absolue connaissance, c'est devenir un monstre. La connaissance est une perpétuelle aventure à la lisière de l'incertitude. »

Frank Herbert - Dune (1965)

REMERCIEMENTS

Je tiens tout d'abord à remercier mon directeur de thèse Frédéric Legoll, sans qui je n'aurai jamais pu arriver au bout de ce travail. Sa rigueur scientifique, ses connaissances et son exigence ont largement participé au bon déroulement de cette thèse. Encore merci de m'avoir donné cette opportunité qui m'a permis de voyager, d'enseigner et d'évoluer tant sur le plan humain que scientifique.

Je remercie Alexei Lozinski et Anthony Nouy d'avoir bien voulu être rapporteurs de ce travail, ainsi que Sébastien Brisard, Julian Fischer et Sonia Fliss d'avoir accepté de faire partie du jury.

Je suis également très reconnaissant à Claude Le Bris, à l'ED SIE pour m'avoir permis de partir en mobilité à l'Université de Washington pour deux mois. Je remercie également Ulrich Hetmaniuk pour m'avoir chaleureusement accueilli lors de cette aventure américaine.

Le CERMICS est un environnement de travail très stimulant et enrichissant non seulement sur le plan scientifique mais également sur le plan humain. C'est pourquoi, je remercie l'ensemble du laboratoire et toutes les personnes que j'ai pu côtoyer pendant ces trois années. Je remercie tout particulièrement Isabelle Simunic pour son soutien et son efficacité. Je tiens à remercier l'ensemble des doctorants qui ne sont, pour la plupart, plus simplement des collègues mais de vrais amis : Laura, Ling-Ling, Adel, Grégoire, Frédéric, Adrien, Sami, Sofiane, Robert, Inass...

Je tiens particulièrement à remercier Thomas, en tant qu'ami et colocataire, pour m'avoir supporté pendant presque quatre ans, et pour les intermèdes musicaux qui ont rendu le quotidien plus agréable. Un grand merci au cercle des amis des Ponts - Maud, Pierre, Laurent, Laure, Auréliane et Pierre-Adrien - pour leur soutien sans faille et ce même dans les moments difficiles.

Je dédie également cette thèse à ma famille qui m'a soutenu et aidé depuis le début. Je remercie en particulier mes parents et ma sœur Anne-Gaëlle. Une pensée particulière pour mon grand-père qui a toujours cru en moi, et ce jusqu'à la fin.

Sujet Méthodes numériques pour l'étude des fluctuations dans les matériaux multi-échelles et problèmes reliés

Résumé Le travail de cette thèse a porté sur la simulation numérique des matériaux multi-échelles. On considère des matériaux hétérogènes dont les propriétés physiques ou mécaniques (conductivité thermique, tenseur d'élasticité, ...) varient à une échelle petite par rapport à la taille du matériau. La thèse s'articule en deux parties qui correspondent à deux aspects différents des problèmes multi-échelles.
Dans la première partie, on se place dans le cadre de l'homogénéisation aléatoire et on s'intéresse à une question plus fine que la caractérisation d'un comportement moyen : on cherche à étudier les fluctuations de la réponse. Plus généralement, nous visons à comprendre : (i) quels paramètres de la distribution des coefficients du matériau à l'échelle fine affectent la distribution de la réponse à l'échelle macroscopique, et (ii) s'il est possible d'estimer cette distribution sans utiliser une méthode type Monte-Carlo, très couteuse. Sur le plan théorique, nous avons considéré un matériau faiblement aléatoire (micro-structure périodique avec ajout d'une perturbation aléatoire petite). Nous avons montré qu'en utilisant le correcteur standard issu de la théorie de l'homogénéisation aléatoire, nous sommes capables de calculer un tenseur $\mathcal{Q}$ qui gouverne complètement les fluctuations de la réponse. Ce tenseur, défini par une formule explicite, permet d'estimer la fluctuation de la réponse sans résoudre le problème fin pour de nombreuses réalisations. Une stratégie d'approximation numérique de ce tenseur a ensuite été développée et testée numériquement dans des cas plus généraux.
Dans la deuxième partie de la thèse, on considère un matériau hétérogène déterministe fixé où les hypothèses classiques d'homogénéisation (périodicité, ...) ne sont pas vérifiées. Les méthodes de résolution standard type Éléments Finis donnent de mauvaises approximations. Pour pallier cette difficulté, la Méthode des Éléments Finis Multi-échelles (MsFEM) a été introduite il y une vingtaine d'année. La méthode MsFEM se décompose en deux étapes : (i) créer un espace d'approximation grossier engendré par les solutions de problèmes locaux bien choisis ; (ii) approximer la solution avec une approche de Galerkine peu couteuse sur l'espace construit dans (i). Dans cette deuxième partie, plusieurs taches ont été réalisées. Tout d'abord, une implémentation de plusieurs variantes MsFEM a été effectuée sous forme de template dans le logiciel de calcul Éléments Finis FreeFem++. Par ailleurs, plusieurs variantes des MsFEM pâtissent d'une erreur dite de résonance : lorsque la taille des hétérogénéités est proche de la taille du maillage grossier, la méthode devient très imprécise. Pour pallier ce problème, une méthode MsFEM enrichie a été développée : à la base MsFEM classique on rajoute des solutions de problèmes locaux ayant pour conditions aux limites des polynômes de haut degré. L'utilisation de polynômes nous permet d'obtenir une convergence de l'approche à des coûts de calcul raisonnables.

Mots-clefs Multi-échelles, Homogénéisation aléatoire, Éléments finis multi-échelles, Équations elliptiques

Title Numerical methods for the study of fluctuations in multi-scale materials and related problems

Summary This thesis is about the numerical approximation of multi-scale materials. We consider heterogeneous materials whose physical or mechanical (thermal conductivity, elasticity tensor, ...) vary on a small scale compared to the material length. This thesis is composed of two parts describing two different aspects of multi-scale problems.
In the first part, we consider the stochastic homogenization framework. The aim here is to go beyond the identification of an effective behavior, by attempting to characterize the fluctuations of the response. Generally speaking we strive to understand: (i) what parameters of the distribution of the material coefficient affect the distribution of the response and (ii) if it is possible to approximate this distribution without resorting to a costly Monte-Carlo method. On the theoretical standpoint, we consider a weakly random material (the micro-structure is periodic and presents some small random defects). We show that we are able to compute a tensor $\mathcal{Q}$ that governs completely the fluctuations of the response, thanks to the use of standard corrector functions from the stochastic homogenization theory. This tensor is defined by an explicit formula and allows us to estimate the fluctuation of the response without solving the fine problem for many realizations. A numerical approximation of this tensor has been proposed and numerical experiments have been performed in broader random frameworks to assess the effectiveness of the approach.
In the second part, we consider a heterogeneous deterministic material where classical homogenization (periodicity, ...) assumptions are not satisfied. Standard methods such as Finite Elements give bad approximations. In order to solve this issue the Multi-scale Finite Element Method (MsFEM) can be used. This approach proceeds in two steps: (i) design a coarse approximation space spanned by solutions to well-chosen local problems; (ii) approximate the solution by an inexpensive Galerkin approach on the space designed in (i). On this topic, we first implemented the main variants of the MsFEM methods in the Finite Element software FreeFem++ on template form. Second, many MsFEM approaches suffer from resonance error: when the size of the heterogeneities is close to the coarse mesh size the accuracy decreases. In order to circumvent this issue, we designed an enriched MsFEM method: to the classical MsFEM basis, we add solutions to local problems with high degree polynomial boundary conditions. The use of polynomials allows us to obtain a converging approach for a limited computational cost.

Keywords Multi-scale, Random homogenization, Multi-scale Finite Elements, Elliptic equations.

LIST OF TALKS

- International Congress on Industrial and Applied Mathematics ICIAM 2019, Valencia, Spain, 07/15/2019 - 07/19/2019. (two oral presentations)
- International Conference on Adaptive Modeling and Simulation ADMOS 2019, EL Campello, Spain, 05/25/2019 - 05/29/2019. (Oral presentation)
- Congrès SMAI 2019, Société de Mathématiques Appliquées et Industrielles, Guidel Plages, France, 05/13/2019 - 05/17/2019. (Oral presentation)
- Groupe de Travail des Thésards du Laboratoire Jacques Louis Lions (LJLL), Sorbonne université, Paris, France, 05/07/2019. (Oral presentation)
- Arbeitsgemeinschaft Applied Analysis, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, 04/11/2019 (Oral presentation)
- FreeFem++ days 2018 10th edition - Paris Sorbonne Université, Laboratoire Jacques-Louis Lions, Paris, France, 12/12/2018 - 12/14/2018. (Oral presentation)
- Inria's Junior Seminar at INRIA Paris, Paris, France, 06/19/2018. (Oral presentation)
- EDP normandie 2017, Caen, France, 10/25/2017 -10/26/2017. (poster)
- SciCADE 2017, International Conference on Scientific Computation and Differential Equations, Bath, UK, 09/11/2017 -09/15/2017. (Oral presentation)
- USNCCM14, 14th U.S. National Congress on Computational Mechanics, Montreal, Canada, 07/17/2017 – 07/20/2017. (Oral presentation)
- Congrès SMAI 2017, Société de Mathématiques Appliquées et Industrielles, La Tremblade, France, 06/05/2017 – 06/09/2017. (Oral presentation)

CONTENTS

List of Figures xvi

Résumé de la thèse xvii

1 Introduction 1
- 1.1 General introduction . . . 1
- 1.2 Context and motivation . . . 2
- 1.3 Homogenization . . . 5
  - 1.3.1 Periodic homogenization . . . 7
  - 1.3.2 Stochastic homogenization . . . 9
  - 1.3.3 Contribution: Estimation of the fluctuations in a weakly stochastic regime . . . 11
- 1.4 Numerical approaches . . . 16
  - 1.4.1 General principle and main approaches . . . 17
  - 1.4.2 MsFEM . . . 20
  - 1.4.3 Contribution: MsFEM enriched with polynomials . . . 28
- 1.5 Perspectives . . . 33
  - 1.5.1 Stochastic homogenization . . . 33
  - 1.5.2 MsFEM enriched method . . . 34
  - 1.5.3 MsFEM implementation . . . 35

2 Numerical approximation of fluctuations 37
- 2.1 Introduction . . . 37
- 2.2 Weakly random case and main results . . . 41
  - 2.2.1 First main result . . . 42
  - 2.2.2 Second main result . . . 45
- 2.3 Limit of (the leading order term of) $I_\varepsilon(f,g)$ . . . 47
  - 2.3.1 Technical lemmas . . . 47
  - 2.3.2 Proof of Proposition 2.12 . . . 49
- 2.4 Limit of (the leading order term of) $\mathcal{Q}^L$ . . . 55
  - 2.4.1 Proof of Lemma 2.18 . . . 60
  - 2.4.2 Proof of Lemma 2.19 . . . 63
- 2.5 Numerical approximation of $\mathcal{Q}^L$ . . . 66
  - 2.5.1 Green function with Q_N-periodic boundary conditions . . . 68
  - 2.5.2 Proof of Lemma 2.25 . . . 69
  - 2.5.3 Proof of Lemma 2.26 . . . 74

2.6 Numerical results . . . 79
2.6.1 Approximation of $\mathcal{Q}$ . . . 80
2.6.2 Estimation of the asymptotic law of $I_\varepsilon(f,g)$ . . . 84
2.A Estimates for the Green function of the laplacian operator with periodic boundary conditions . . . 90
2.A.1 Analytical expression . . . 90
2.A.2 Estimates on G_N . . . 92

3 A MsFEM approach using high order polynomials 97
3.1 Introduction . . . 97
3.2 Discretization approach . . . 99
3.3 *A priori* and *a posteriori* estimates . . . 103
3.4 Numerical experiments . . . 106
3.4.1 Comparison with other MsFEM approaches . . . 106
3.4.2 *A posteriori* estimator . . . 111
3.5 Proofs . . . 115
3.5.1 Proof of Lemma 3.9 . . . 115
3.5.2 Proof of Lemma 3.10 . . . 117
3.5.3 Proof of Proposition 3.12 . . . 120

4 MsFEM Implementation in FreeFem++ 129
4.1 Linear MsFEM and MsFEM oversampling . . . 130
4.2 MsFEM à la Crouzeix Raviart . . . 132
4.3 Coupling MsFEM . . . 134
4.4 Using MsFEM approach as a second level preconditioner . . . 137

A Codes 147
A.1 Offline step: Basis creation . . . 147
A.1.1 Linear MsFEM . . . 147
A.1.2 Oversampling MsFEM . . . 151
A.1.3 MsFEM Crouzeix-Raviart . . . 156
A.2 Online step: Computing approximation . . . 162
A.2.1 Linear MsFEM and oversampling MsFEM . . . 162
A.2.2 MsFEM Crouzeix-Raviart . . . 165
A.3 Coupling linear MsFEM with P1 . . . 169
A.4 MsFEM as a second level preconditioner . . . 174

B Trace results and Sobolev interpolation results 177
B.1 Sobolev spaces on boundaries and Traces operator . . . 177
B.1.1 Characterization of the regularity of a domain Ω . . . 177
B.1.2 Review of fractional Sobolev spaces . . . 178
B.1.3 Definition of Sobolev spaces on the boundary . . . 179
B.1.4 Sobolev spaces on $\Gamma_0 \subset \Gamma$ . . . 179
B.1.5 Trace theorems . . . 180
B.2 Regularity of elliptic equations on convex domains . . . 182
B.3 Sobolev interpolation of linear operators . . . 183
B.4 Polynomial interpolation results and properties . . . 183
B.4.1 Interpolation of smooth functions . . . 184
B.4.2 Interpolation of non-smooth function . . . 185

LIST OF FIGURES

Chapter 1: Introduction **1**

1.1 Example of a two phase heterogeneous material 3

1.2 Plot of solution for different H (left), H^1 error for the 1D heterogeneous problem and Poisson problem as a function of $1/H$ (right) 5

1.3 Solutions to problem (1.2) in 1D with periodic coefficient and multiple values of ε 6

1.4 Example of a random stationary material: random checkerboard for different values of ε ($\varepsilon = 1/10$, $\varepsilon = 1/50$, $\varepsilon \ll 1$, from left to right). . . 10

1.5 Sketch of MsFEM basis function design in 2D (left), Example of MsFEM basis function for an oscillating coefficient (middle) and P1 piecewise function (right) 21

1.6 Relative energy error of linear MsFEM approximation function of $1/H$ for the 2D problem (1.2) where A_ε is a ε-periodic function ($\varepsilon = 1/32$) . 22

1.7 Decomposition of u solution to (1.2) posed in $(0.1)^2$, for a periodic A_ε with $\varepsilon = 1/32$ and $H = 1/4$. Solution u (on the left), u^B the bubble part (in the middle) and u^Γ the interface part (on the right) 24

1.8 Sketching of oversampling MsFEM basis function design 25

1.9 Left - Material with perforations, Right - Mesh used 27

1.10 Edge-based basis function (Left), Bubble basis function (Right) 27

1.11 Sketch of the design of an enrichment 31

Chapter 2: Numerical approximation of fluctuations **37**

2.1 Schematic representation (when $L < N < 2L$) of a case of some $k \in \mathbb{Z}^d$ with $|k|_\infty \leq N$ such that $k+Q_L \not\subset Q_N$. The domain $(k+Q_L)\setminus Q_N$ is represented in blue. By Q_N-periodicity, this blue domain is mapped back in a open set in Q_N, denoted $R_k^{N,L}$, and which is represented in green. 70

2.2 Two realizations of the checkerboard: $\varepsilon = 1/10$ (left) and $\varepsilon = 1/50$ (right). 80

2.3 Procedure to approximate the tensor $\mathcal{Q}$ by $\mathcal{Q}^{L,N,M}$: $N \geq L$ denotes the size of the domain on which we compute the corrector, L denotes the size of the domain on which we consider the corrected energy density, while M denotes the number of realizations we consider to approximate the covariance. 81
2.4 $\mathcal{Q}^{L,N,M}_{1111}$ as a function of N ($L = 5$ and $M = 10^4$). We also plot confidence intervals (CI) computed from the M realizations. 81
2.5 $\mathcal{Q}^{L,N,M}_{1111}$ (left) and $\mathcal{Q}^{L,N,M}_{1122}$ (right) as a function of N ($L = 10$ and $M = 10^4$). 82
2.6 $\mathcal{Q}^{N-10,N,M}_{1111}$ (left) and $\mathcal{Q}^{N-10,N,M}_{1122}$ (right) as a function of N ($L = N-10$ and $M = 10^4$). 82
2.7 Evolution of $\mathcal{Q}^{L,L,M}_{1111}$ and $\mathcal{Q}^{L,L+10,M}_{1111}$ as a function of L ($M = 10^4$). . . . 83
2.8 Variance $(\sigma^{L,N,M})^2$ computed from the tensor $\mathcal{Q}^{L,N,M}$ as a function of N ($L = N - 10$ and $M = 10^4$). 83
2.9 [Test case 1] Right-hand side $f(x,y) = 10e^{-80(x-0.5)^2}$ (left) and test function $g(x,y) = 10e^{-80(y-0.5)^2}$ (right). 85
2.10 [Test case 1] Empirical distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations. 85
2.11 [Test case 1] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations. 85
2.12 [Test case 1] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation. 86
2.13 [Test case 2] Right-hand side $f(x,y) = 10e^{-40(x^2+y^2)}$ (left) and test function $g(x,y) = 10e^{-40(x^2+(y-0.5)^2)}$ (right). 86
2.14 [Test case 2] Empirical distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/100$) computed from $\mathcal{M} = 10^3$ realizations. 87
2.15 [Test case 2] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$: right: $\varepsilon = 1/100$) computed from $\mathcal{M} = 10^3$ realizations. 87
2.16 [Test case 2] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation. 88
2.17 [Neumann test case] Right-hand side $f(x,y) = \cos(2\pi x)\sin(2\pi y)$ (left) and test function $g(x,y) = \cos(2\pi y)\sin(2\pi x)$ (right). 89
2.18 [Neumann test case] Empirical distributions of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations. 89
2.19 [Neumann test case] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations. 89
2.20 [Neumann test case] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation. 90

Chapter 3: A Multi-scale Finite Element approach using high order polynomials **97**

3.1 Local problem defining an edge enrichment . . . 101
3.2 Error on the approximation of the gradient for MsFEM-lin (left) and for our approach with 4 polynomials (right): the accuracy is poor in the yellow regions, much better in the light blue regions, and excellent in the dark blue regions. . . . 108
3.3 Compared performances in the regime $H \approx \varepsilon$. . . . 109
3.4 Comparison of our approach with classical MsFEM approaches . . . 110
3.5 Comparison of our approach with the Special Element Method, at equal number of enrichments per edge (for instance, "Eigen 1" and "Legendre N=2" both correspond to adding one enrichment per edge vs the MsFEM-lin approach). . . . 110
3.6 Our approach for triangles and quadrangles, in terms of $1/H$ (left) or of the number of degrees of freedom (right). The approaches "Triangle N=1" and "Quadrangle N=1" both correspond to the MsFEM-lin approach, on triangular (resp. quadrangular) meshes. The approaches "Triangle N=2" and "Quadrangle N=2" both correspond to adding one enrichment per edge vs the MsFEM-lin approach. . . . 111
3.7 Left: *a posteriori* error (3.20) and relative interface error (3.19) as a function of N for $H = 1/4, 1/8$. Right: *a posteriori* error (3.20) and relative interface error (3.19) as a function of N for $H = 1/16, 1/32, 1/64$. 112
3.8 *A posteriori* error (3.20) and relative interface error (3.19) as a function of $1/H$ for different polynomial degrees $N = \{1, 2, 4, 6, 8\}$. . . . 113
3.9 Error maps edge by edge for $N = 1$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale). . . . 114
3.10 Error maps edge by edge for $N = 5$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale). . . . 114
3.11 Error maps edge by edge for $N = 10$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale). . . . 114

Chapter 4: MsFEM Implementation in FreeFem++ **129**

4.1 Left: Mesh of the domain ($H = 1/16$): we use both MsFEM functions (in orange) and P1 FE functions (in red). Right: coefficient A_ε considered 135
4.2 Basis functions ϕ_i associated with interface degrees of freedom: Left - elements where ϕ_i is supported (in the red element, ϕ_i is a P1 basis function; in the yellow element, ϕ_i is a MsFEM basis function), Right - plot of basis function associated with degrees of freedom i . . . 135
4.3 Graph of error for P1 coupled with MsFEM lin (Left, relative H^1 error 28.3%), P1 coupled with MsFEM oversampling (Middle, relative H^1 error 23.5%) and P1 (Right relative H^1 error 34%) . . . 136
4.4 Separation of the domain into the P1 subdomain (red) and a large (left), medium (center) or small (right) MsFEM subdomain . . . 136

4.5 Error corresponding to a large MsFEM subdomain (Left relative H^1 error 23.5%), a medium MsFEM subdomain (Middle relative H^1 error 23.8%) and small MsFEM subdomain (Right relative H^1 error 31.1%) . 137
4.6 Left - Coarse MsFEM approximation ($H = 1/8$), Middle - fine approximation obtained with GMRES preconditioned by our approach where convergence is achieved (150 iterations, $h = 1/256$), Right - fine approximation obtained with GMRES preconditioned by Jacobi after 150 iterations and $h = 1/256$. 140

Chapter B: Trace results and Sobolev interpolation results **177**

B.1 Left: example of Lipschitz domain, Right: a non-Lipschitz domain . . . 178

RÉSUMÉ DE LA THÈSE

De nombreux problèmes dans l'industrie concernent des matériaux multi-échelles. Par exemple dans le cadre de l'ingénierie de l'aviation ou du bâtiment des matériaux de plus en plus complexes sont utilisés. En général, il y a une séparation d'échelles : on possède de l'information sur la composition du matériau à une échelle fine et on veut en déduire son comportement effectif en termes de propriétés physiques (conductivité électrique ou thermique, comportement mécanique) à une échelle plus large. Les techniques classiques d'approximation numérique comme les Eléments finis donnent de mauvais résultats au sens que le problème doit être résolu à l'échelle fine pour avoir une précision acceptable. Ceci conduit à des calculs trop lourds. Il y a donc un besoin de créer des approches multi-échelles : des techniques qui utilisent la connaissance de la micro-structure pour construire une approximation qui donne des résultats précis avec un coût de calcul raisonnable.

Ce travail de thèse a consisté à concevoir des méthodes numériques pour calculer des approximations de solutions de problèmes multi-échelles à des coûts de calcul raisonnables. En particulier, on s'est intéressé à des problèmes elliptiques avec des coefficients hautement oscillants.

Une première façon d'approximer la solution de problèmes hétérogènes est de considérer un régime asymptotique, c'est-à-dire étudier ce qui se passe quand la séparation d'échelle tend vers l'infini. Sous certaines hypothèses sur les coefficients, le matériau se comporte comme un matériau homogène dit effectif ne comportant plus de petites échelles, et donc plus facile à discrétiser numériquement. Cette façon de procéder est au coeur des techniques dites d'homogénéisation : à partir de la micro-structure, on en déduit un comportement effectif. Dans le cas déterministe, en particulier quand les coefficients sont supposés périodiques, on peut construire des approximations très efficaces qui convergent en fonction de la séparation d'échelle. Dans le cadre de l'homogénéisation stochastique, quand les coefficients sont caractérisés par une loi supposée invariante par translation on trouve des résultats similaires au cas déterministe. Asymptotiquement, le matériau se comporte comme un matériau effectif déterministe. En revanche, quand la séparation d'échelle n'est pas infinie, la solution est une fonction aléatoire et une autre question d'intérêt est de savoir comment cette solution fluctue autour de son comportement moyen. Dans la première partie de cette thèse, en s'inspirant de M. Duerinckx, A. Gloria et F. Otto [31], on étudie les fluctuations de la solution dans un cas dit faiblement aléatoire introduit dans [26]. Le but est de caractériser la loi de quantités d'intérêt dépendant de la solution.

Le point de vue de l'homogénéisation repose sur des hypothèses assez restrictives telles que la stationarité ou la périodicité des coefficients. En dehors de ces hypothèses, la théorie donne encore des résultats de compacité, mais ceux-ci sont difficiles à exploiter numériquement. C'est pourquoi, il est intéressant d'élaborer de nouvelles méthodes numériques qui ne sont pas aussi efficaces mais qui peuvent être utilisées dans une gamme de problèmes plus large. Pour répondre à cet objectif, de multiples approches ont été développées. Celles-ci sont fondées sur un principe dit ascendant : des problèmes sont résolus localement à une échelle fine et sont utilisés pour améliorer une approximation construite sur une échelle plus grossière. La méthode des Éléments Finis Multi-échelle (MsFEM) développée par T. Y. Hou et X.-H. Wu dans [55] est dans cette ligne de pensée. En résumé, L'idée est de créer des fonctions de bases adaptées au problème pour construire un espace d'approximation grossier (similaire aux fonctions affines par morceaux dans le cas élément finis P1). Ces fonctions sont solutions de problèmes locaux. A partir de cet espace engendré par les fonctions de base précalculées, et de petite dimension, on résout le problème de Galerkin associé dont la solution sera notre approximation. Les problèmes locaux à résoudre peuvent être choisis différemment (notamment conditions aux limites) et chaque choix de problème local à résoudre conduit à une unique variante MsFEM. Ainsi, l'approche MsFEM ne correspond pas à une méthode en particulier mais à une classe de méthodes. Pour la plupart de ces variantes l'approximation pâtit d'une erreur dite de résonance, tout particulièrement quand la taille des problèmes locaux est proche de la taille de l'échelle fine. Pour palier ce problème, U. Hetmaniuk et R. Lehoucq dans [53] ont proposé d'enrichir la base MsFEM classique (MsFEM linéaire) par des solutions de problèmes aux valeurs propres généralisés. U. Hetmaniuk et A. Klawonn dans [52] ont prouvé par la suite que l'erreur de résonance peut être fortement réduite avec un nombre suffisant d'enrichissements. Cette méthode dite des Éléments finis spéciaux est efficace bien que la résolution de problèmes aux valeurs propres puisse conduire à des coûts de calcul importants. La deuxième partie de cette thèse a consisté à élaborer une méthode d'enrichissement inspirée de la méthode des Éléments finis spéciaux, et fondée sur des polynômes de haut degré. Ce travail a été réalisé en collaboration avec en particulier U. Hetmaniuk de l'université de Washington dans le cadre d'une mobilité financée par l'Université Paris Est, l'Inria et University of Washington. Sur un thème proche, un travail sur l'implémentation d'approches MsFEM dans le logiciel Freefem++ (voir [50]) sous forme de *template* a aussi été réalisé en collaboration avec F. Hecht.

Les contributions originales de cette thèse sont

- La caractérisation des fluctuations de la solution de problèmes hétérogènes et aléatoires, et particulièrement dans le cas faiblement aléatoire (voir le Chapitre 2)
- Le développement et l'analyse d'une méthode d'éléments finis multi-échelles enrichie par des polynômes de haut degré (voir Chapitre 3)
- L'implémentation de MsFEM dans le logiciel FreeFem++ (voir Chapitre 4)

CHAPTER 1

INTRODUCTION

1.1 General introduction

Many problems in the industry involve multi-scale materials. For instance the aircraft industry makes use of composites materials, whereas concrete (obviously very much used in civil engineering) is actually a very complex and multi-scale material. Usually there is a *separation of scales*: we have information on the composition of the material at the micro-scale and want to infer its effective physical properties on a larger scale. Classical numerical techniques such as Finite Elements perform poorly in the sense that the problem has to be solved at the micro-scale in order to get accurate results, leading to a large computational load. In order to address this issue, multi-scale approaches have been introduced fifteen years ago. Such techniques use the knowledge at the micro-scale to build an approximation space yielding accurate results at an affordable cost.

This thesis is about the design of numerical methods to compute affordable approximation of solutions to multi-scale problems. In particular, we will be interested in solving elliptic problems with highly oscillating coefficients.

One way to approximate the solution of multi-scale problems is to study the asymptotic regime, that is the limit when the separation of scales is going to infinity. Under some assumptions on the coefficients, the material behaves like a simple effective one. Such property is at the heart of *homogenization techniques*: from the micro-structure, we infer the effective behavior. In the deterministic case, in particular when the coefficients are assumed periodic, we can build very effective approximations that converge when the separation of scales increases. In the *stochastic homogenization* framework, when the coefficients are characterized by a probability law which is assumed to be ergodic and stationary (and hence invariant by translation), then we can derive results similar to those of the deterministic case: the material behaves asymptotically as an effective material homogeneous and deterministic. However, when the separation of scales is not infinite, the solution is random and another question of interest is how it fluctuates around its mean behavior. In the first part of this thesis, inspired by the work of M. Duerinckx, A. Gloria and F. Otto [31], we study the fluctuations of the solution in a weakly stochastic case introduced in [26]. The aim is then to characterize the probability law of some quantities of interest which depend linearly upon the solution.

The homogenization point of view relies on some geometric assumptions (such as periodicity or stationarity) on the coefficients. Beyond this case, the theory till provides compactness results, but they are not easily amenable to practical implementation. It is therefore interesting to design numerical techniques that can be applied to a broader range of problems up to possibly a slight loss of efficiency. To address this issue, multiple approaches have been developed. Those usually are bottom-up approaches: problems are solved locally at a finer scale and are used to improve the computations at a coarser scale. The Multi-scale Finite Element method (MsFEM for short) developed by T.Y. Hou and X.-H. Wu in [55] falls within this line of thinking. The idea is to define adapted basis functions solution to local problems in order to build a coarse approximation space (similar to piecewise affine functions in the Finite Element case) and finally compute an approximation by solving the associated Galerkin problem. Each precise formulation of the local problems to solve leads to a different MsFEM approach. In most of these methods, the approximation suffers from a *resonance error* when the size of the local problems is close to the micro-scale. To circumvent this issue U. Hetmaniuk and R. Lehoucq in [53] proposed to enrich the standard basis with solutions to eigenvalue problems. U. Hetmaniuk and A. Klawonn in [52] proved that the *resonance error* can be canceled provided sufficiently many enrichments are considered. This method is effective, though solving eigenvalue problems can be computationally challenging. Inspired by this approach, the second part of the thesis considers the design of an enrichment method based on high order polynomials.
This work has been accomplished in collaboration with in particular U. Hetmaniuk from the University of Washington where I was invited for two months thanks to the partial funding of of Université Paris Est, Inria and University of Washington. Also, in collaboration with F. Hecht we studied how to introduce the MsFEM methods in the software Freefem++ (see [50]) in a template format.

The main contributions of this thesis are:

- The characterization of the fluctuations in the weakly stochastic case (see CHAPTER 2)
- The design and analysis of an enriched Multi-scale Finite Element Method with high order polynomials (see CHAPTER 3)
- Implementation of MsFEM into Freefem++ (see CHAPTER 4)

1.2 Context and motivation

Many materials operate on a multi-scale basis. One can think of composite material that possess an underlying structure leading to interesting properties. Also, there are a lot of physical processes whose behavior are well understood on a microscopic basis but for which it is difficult to infer what happens in a macroscopic framework. The interest of studying multi-scale materials is two-fold: first a scientific interest, that is to understand better processes that occur on multiples scales; second an engineering interest to design new materials that have a broader range of physical properties, and that are hence more likely to meet some engineering requirements at an affordable cost.
Usually, physical phenomena are modeled by Partial Differential equations. The materials properties can be encoded in such equations as coefficients functions, boundary

conditions or external forces.

We will consider here a simple steady heterogeneous diffusion problem with Dirichlet boundary conditions, a bounded domain D of $\mathbb{R}^d$, with d the dimension of the ambient space.
This problem reads as

$$\begin{cases} -\mathrm{div}(A\nabla u) = f \text{ in } D, \\ u = 0 \text{ on } \partial D. \end{cases} \tag{1.1}$$

In (1.1), the diffusion coefficient A is a matrix of size $d \times d$. Solutions to such problems can for instance represent the equilibrium temperature in a material, whose thermal properties are encoded in the coefficient A, and subject to the heat source f. For instance for $d = 2$, if the material is composed of two materials m_1 and m_2 in the domain D of thermal conductance a_1 and a_2 respectively (see FIGURE 1.1), then $A = a_1 \mathbb{1}_{m_1} I_2 + a_2 \mathbb{1}_{m_2} I_2$, with $\mathbb{1}_{m_1}$ the indicator function of the material m_1 location. The solution to (1.1) can also represent the electrical potential in electrical conductance problems or the displacement of the material in a linear elasticity context.

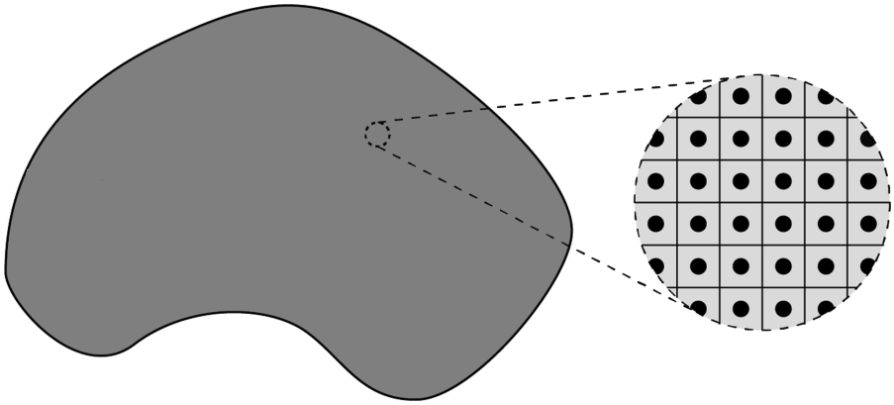


Figure 1.1: Example of a two phase heterogeneous material

We aim at solving (1.1) in a multi-scale context when the material is heterogeneous. We assume that the physical properties of the material undergo changes at a scale ε that is small compared to the characteristic length of the domain D. We denote by ε the characteristic length of the micro-scale, the smallest scale of our problem, and we denote by $|D|$ the volume of the domain that will be the characteristic length of the macro-scale, the largest scale of our problem. Mathematically this means that the coefficient A in (1.1) varies at the scale ε. Making this dependency explicit, we aim at solving the following problem:

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f \text{ in } D, \\ u_\varepsilon = 0 \text{ on } \partial D. \end{cases} \tag{1.2}$$

We can also write the corresponding variational formulation:

Find $u_\varepsilon \in H_0^1(D)$ such that

$$\forall v \in H_0^1(D), \quad a_\varepsilon(u, v) = \int_D (A_\varepsilon \nabla u) \cdot \nabla v = b(v) = \int_D fv \tag{1.3}$$

In order for the problem (1.2) to be well posed, the coefficient A_ε should satisfy some assumptions. In order to apply the Lax-Milgram theorem, we usually assume that A_ε is elliptic and bounded almost everywhere and uniformly in ε:

$$\exists\, C,\ c \in \mathbb{R}^+\ \forall \xi \in \mathbb{R}^d, \quad c\|\xi\|^2 \leq (A_\varepsilon(x)\xi)\cdot\xi \leq C\|\xi\|^2 \quad a.e. \tag{1.4}$$

When $\varepsilon \ll D$, standard approximation techniques are not efficient. For instance, the Finite Element method usually performs poorly. Let us recall the principle of conformal Finite Element methods: it is a Galerkin approach on a finite dimensional space V_H that is the span of piecewise polynomial basis functions of degree N (piecewise affine functions in a P1 formulation) built on a mesh of size H. For instance for P1 Finite Elements, denoting by ϕ_i the piecewise affine function associated with the interior vertex i. Denoting $V_H = \mathrm{Span}(\phi_i)$, the P1 FE approximation $u_H \in V_H$ satisfies

$$\forall v \in V_H, \quad a_\varepsilon(u_H, v) = \int_D (A_\varepsilon \nabla u_H)\cdot\nabla v = b(v) = \int_D fv \tag{1.5}$$

This is equivalent to solve the following linear system:

$$\begin{cases} KU = B \\ K_{i,j} = (A_\varepsilon\nabla\phi_i)\cdot\nabla\phi_j, B_i = b(\phi_i) \\ u_H = \sum_{i=1}^{Nbpt} U_i\phi_i \end{cases} \tag{1.6}$$

For a thorough monograph of Finite Element theory one can refer to the book [36].

Theorem 1.1 (see e.g. [36] THEOREM 3.16)**.** *To approximate the solution u to (1.1), we consider a uniform mesh of size H of the polygonal domain $D \subset \mathbb{R}^d$, and introduce the set V_H of piecewise affine functions. Denoting by u_H the solution to (1.5), we have*

$$\|u - u_H\|_{H^1(D)} \leq CH|u|_{H^2(D)}$$

In the multi-scale context such an estimate is usually not satisfactory since $|u_\varepsilon|_{H^2(D)}$ may not be bounded independently of the size of the heterogeneities ε. For instance if we consider $d = 1$ and take a coefficient of the form $a_\varepsilon = a_{per}(\frac{x}{\varepsilon})$ with a_{per} a $\mathbb{Z}$-periodic function, usually $|u_\varepsilon|_{H^2(D)}$ scales as $1/\varepsilon$.

Numerical results illustrate this behaviour, showing that the estimate is indeed sharp. Let us consider a 1D example on $(0,1)$:

$$\begin{cases} \left(\left(\sin(\frac{x}{\varepsilon}) + 1.1\right)u_\varepsilon'\right)' = 1 \\ u_\varepsilon(0) = u_\varepsilon(1) = 0 \end{cases} \tag{1.7}$$

We set $\varepsilon = 1/128$ (corresponds to an oscillation period $T \simeq 1/20$) and compare the P1 FE approximation for multiple H to a reference solution.

We can see on FIGURE 1.2 that the error decreases linearly for the Poisson problem. However, for the heterogeneous case, we see that the error stagnates until $H \simeq T/10$ and then decreases linearly. Hence, a FE method gives accurate results only if the mesh size discretizes well the heterogeneities. But in most multi-scale problems the computational load associated with solving the whole problem at the smallest scale ε is much too large.

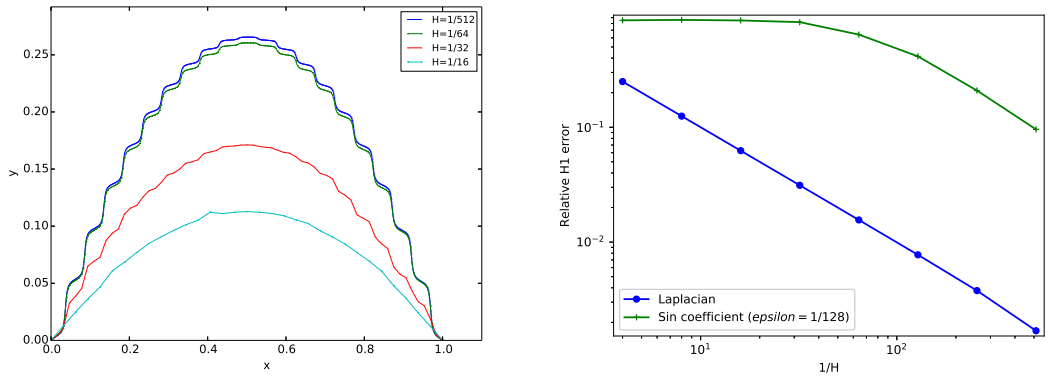


Figure 1.2: Plot of solution for different H (left), H^1 error for the 1D heterogeneous problem and Poisson problem as a function of $1/H$ (right)

The need to get accurate solutions at a reasonable computational cost has driven and led to the development of multi-scale techniques. Usually, Multi-scale approaches use the information at the micro-scale (for instance the knowledge of the coefficient A_ε) to improve accuracy on the coarser scale. Such methods are also called bottom-up approaches. Roughly stated, we can separate such techniques into two categories: methods that rely on the particular geometric assumption of the coefficient A_ε (periodicity, stationarity, ...), and more generic methods that integrate fine features into a coarse approximation space through various means.

The approaches relying on particular assumptions on the coefficient A_ε are mainly based on the mathematical homogenization process. The homogenization process stands for the asymptotic study of (1.2) when the separation of scales becomes infinite (that is when ε goes to 0). Then, usually the solution to (1.2) converges to an asymptotic behavior in some sense, and quantitative estimates may sometimes be established. The homogenization methods encompass both the theoretical framework and the numerical approaches derived from this framework. Such approaches will be discussed in SECTION 1.3. The periodic and stochastic homogenization frameworks will be presented as well as the main contributions of this thesis regarding the fluctuations in the stochastic case that will be more detailed in CHAPTER 2.

However, sometimes A_ε does not satisfy any structure assumptions though heterogeneities are present at a small scale ε. In this case other approaches should be considered. Such approaches usually enrich a coarse formulation by inserting local fine features. These methods will be further explored in SECTION 1.4. Several popular approaches will be presented though the focus will be put on the Multi-scale Finite Element methods (MsFEM) and on the enrichment method we designed during this thesis.

1.3 Homogenization

In multi-scale problems where there is some structure, the goal is to infer the macroscopic behavior of a material knowing its micro-structure. One way to achieve that is to consider a part of the material at a meso-scale δ called a representative volume element (RVE) where $\varepsilon \ll \delta \ll |D|$, and compute the behavior of the material on

δ. Assuming that δ is representative of the whole material behavior, one uses this information to perform the macroscopic computation. One can refer to the works [49], [24], [54],[74], for examples of such an homogenization approach in the physics and mechanics communities.

It gives intuition that there exists, in some cases, an effective behavior of the material that can be obtained through an averaging process.

As an illustration, we can consider the problem (1.2) with $A_\varepsilon = 1.1 + \sin(\frac{2\pi x}{\varepsilon})$ for $D = (0, 1)$. Here the micro-scale length, that is the frequency of the oscillations, is characterized by the parameter ε. FIGURE 1.3 shows solutions to (1.2) for different values of ε. We can see that the solution oscillates more and more when ε decreases. However, the amplitude of the oscillation is also decreasing. Indeed, for $\varepsilon = 1/512$ it seems that the solution does not oscillate and is the solution to a PDE with a constant coefficient. There seems to be an asymptotic regime when the separation of scale goes to infinity.

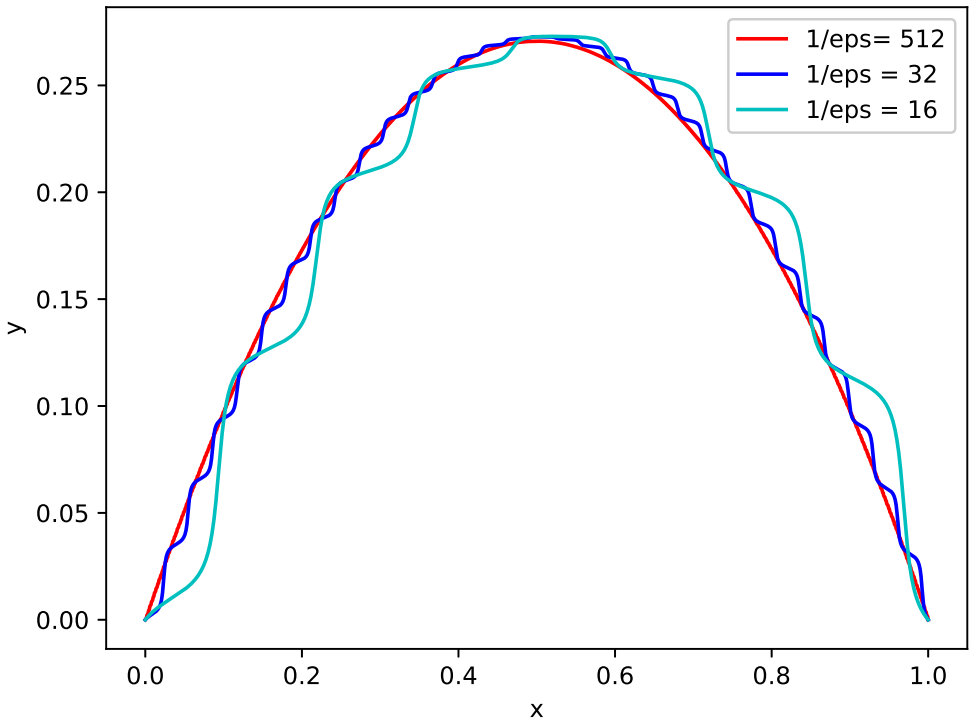


Figure 1.3: Solutions to problem (1.2) in 1D when $A_\varepsilon = 1.1 + \sin(\frac{2\pi x}{\varepsilon})$ and $\varepsilon = 1/16, 1/32, 1/512$

This empirical observation gives some intuition of the meaning of the homogenization process. However it is not satisfactory since it does not allow us to understand fully from a mathematical standpoint what is this averaging process and what are the conditions for it to happen. Especially, is computing the behavior on an RVE enough to infer the macroscopic properties? How to choose the RVE? How the result is dependent of other parameters such as the external input (right-hand side f in the

problem (1.2))? Thus, we will take a step back in order to find a mathematical framework that will allow us to fully understand this phenomenon.

Usually, multi-scale problems with separation of scales involve a standard partial differential equation where some parameter oscillates at a micro-scale ε that is small compared to the domain D. Homogenization is the mathematical study of the behavior of the solutions to the multi-scale problem when the separation of scales become infinite (that is ε goes to 0). Hence, homogenization can be seen as the mathematical asymptotic study of the behavior of a sequence of problems parameterized by the separation of scale that tends to infinity. For simple general linear elliptic PDEs, the most general result comes from the work of Spagnolo and then Murat and Tartar (see e.g. [73]), it is called H-convergence and G-convergence. They managed to prove that for a bounded elliptic coefficient the solutions to this sequence of problems converge to an asymptotic problem up to a subsequence extraction. Such a result is not constructive but is the foundation of the periodic homogenization theory where one can obtain quantitative results. Also, another framework, the so called Γ-convergence, was developed by De Giorgi in [28]. Roughly put, if the solution to the PDE is the minimizer of some energy then the minimizers converge also to the minimizer of some asymptotic energy, ensuring that an effective regime exists (one can see the monograph [27] for more details). Such theory can be applied in a broad range of physical problem. There is also the case where the coefficients are taken as random variables, that is the case of stochastic homogenization. This framework has been developed by Jikov, Papanicolaou and Varadhan (one can refer to [58] for a thorough review).

Although homogenization theory can be applied to a broad range of PDEs, we will restrict ourselves to the homogenization of multi-scale problems similar to problem (1.2), that is elliptic equations in divergence form with highly oscillating coefficients. Such problems are of interest because they cover a lot of physical phenomena (mechanics, thermal conductivity, ...). Moreover, for such problems the effective problem in the regime ε goes to 0 takes a similar form. Finally, studying elliptic problems is often a good start regarding the study of dynamical problems such as parabolic problems.

Under some geometric assumptions on A_ε, Homogenization theory applied to elliptic equations of the form (1.2) gives quantitative convergence results and an homogenized behaviour. Such estimates are the cornerstone of many efficient approximation techniques in the multi-scale context. We will mainly review here two such cases, the periodic and stochastic cases.

1.3.1 Periodic homogenization

Denoting $Q = (0,1)^d$ the unit square, the main assumption in periodic homogenization is that $A_\varepsilon(x) = A(\frac{x}{\varepsilon})$, with A a matrix valued function that is Q-periodic.

Under this assumption, it holds that u_ε the solution to (1.2) converges weakly in $H^1(D)$ and strongly in $L^2(D)$ toward $u^\star$ solution to

$$\begin{cases} -\mathrm{div}(A^\star \nabla u^\star) = f \text{ in } D, \\ u^\star = 0 \text{ on } \partial D. \end{cases} \tag{1.8}$$

We put the emphasis, that in the periodic case, the whole sequence u_ε (and not only a subsequence) converges to $u^\star$.

We still need to characterize the effective coefficient $A^\star$. In the periodic homogenization case, though there are usually no analytical formulas for $A^\star$, it can be computed by some auxiliary functions called correctors. We denote by w_i the corrector function in the direction e_i solution to

$$\begin{cases} -\text{div}(A(\nabla w_i + e_i)) = 0 \text{ in } Q, \\ w_i \text{ is } Q\text{-periodic.} \end{cases} \tag{1.9}$$

The function w_i is unique up to the addition of a constant.

Then the homogenized coefficient is given by:

$$A^\star e_i = \int_Q A\left(\nabla w_i + e_i\right) \tag{1.10}$$

When ε is small one can approximate u_ε by $u^\star$. However, we read that, in $H^1(D)$, u_ε only converges weakly (and not strongly) to $u^\star$. Indeed, ∇u_ε oscillates with period $1/\varepsilon$ with an amplitude independent of ε, hence the approximation of ∇u_ε by $\nabla u^\star$ cannot be accurate in general, $\nabla u_\star$ does not oscillate at all.

Formally we can expand u_ε in powers of ε, in order to better understand the asymptotic regime and motivated by the result in the 1D case, we write

$$u_\varepsilon(x) = u_0(x) + \varepsilon u_\varepsilon^1(x, x/\varepsilon) + \varepsilon^2 u_\varepsilon^2(x, x/\varepsilon) + \dots \tag{1.11}$$

where u_i $(1 \leq i)$ is periodic with respect to the second variable. This two-scale expansion is only a formal way to guess the homogenization result. Rigorous proofs use either the two-scale convergence framework introduced by G. Allaire (see the monograph [1]), the compensated compactness or the oscillating test function approach introduced by Murat and Tartar.

Denoting $u_\varepsilon^1(x) = \sum_{i=1}^{d} w_i(\frac{x}{\varepsilon})\frac{\partial u^\star}{\partial x_i}(x)$, then it holds that:

$$\|u_\varepsilon - u^\star - \varepsilon u_\varepsilon^1\|_{H^1(D)} \leq C\sqrt{\varepsilon} \tag{1.12}$$

Then we have a good approximation of u_ε and its gradient (recall that u_ε converges to $u^\star$ in H^1 only weakly). We note that the correctors functions are used to get back the H^1 convergence, justifying the corrector appellation. Some results can also be proven in stronger norms (e.g. $W^{1,\infty}(D)$) following the work of Avellaneda and Lin in [4].

The behavior of the solution in the periodic homogenization framework is well understood and offers an interesting set of test-cases for checking the effectiveness of numerical multi-scale methods or to study more general cases in homogenization theory. Indeed, the framework of periodic homogenization can be extended by considering perturbative cases such as a periodic coefficient with some defects. This setting has been studied by X. Blanc, C. Le Bris, P.-L. Lions and M. Josien in the following works [17], [15] and [16].

1.3.2 Stochastic homogenization

We consider here a similar problem to (1.2), though in this case the matrix A is a random function. Hence, we want to approximate the random function u_ε solution to

$$\begin{cases} -\text{div}\left(A(\frac{x}{\varepsilon},\omega)\right)\nabla u_\varepsilon(x,\omega) = f(x) \text{ in } D, \\ u_\varepsilon(x) = 0 \text{ almost surely on } \partial D. \end{cases} \tag{1.13}$$

where x embodies the space variable and ω the random realization.

Without further assumption, we do not have a separation between the micro-scale and macro-scale when ε goes to 0. Analogously with periodic case, in order to have constructive convergence results, some order in the randomness of A should be assumed.

To this end the stationarity framework can be introduced. Roughly put, this assumption ensures that the law of the coefficient A stays the same up to translations in $\mathbb{Z}^d$ (discrete stationarity) or to any translation (continuous stationarity).

More precisely, we define a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the action $(\tau_k)_{k\in\mathbb{Z}^d}$ from the group $(\mathbb{Z}^d, +)$ acts on Ω. This action is assumed measure preserving and ergodic that is for all $k \in \mathbb{Z}^d$ (for all $x \in \mathbb{R}^d$ for continuous stationarity) and $B \in \mathcal{F}$, then $\mathbb{P}(\tau_k B) = \mathbb{P}(B)$ and if $B \in \mathcal{F}$ is preserved by any τ_k then $\mathcal{P}(B) = 0$ or 1.

Definition 1.2. *A function $F \in L^1_{loc}(\mathbb{R}^d, L^1(\Omega))$ is said to be discrete stationary if*

$$\forall k \in \mathbb{Z}^d, \quad F(x+k,\omega) = F(x,\tau_k\omega) \text{ almost everywhere and almost surely.} \tag{1.14}$$

Remark 1.3. *Discrete stationary function are easy to design. For instance, if one considers X_k a sequence of i.i.d. random variables then the function*

$$F(x,\omega) = \sum_{k\in Z^d} \mathbb{1}_{Q+k}(x) X_k(\omega)$$

where $Q = (0,1)^d$ is a discrete stationary random function.

Remark 1.4. *A function which is discrete stationary and independent of ω is actually Q-periodic. The discrete stationary setting, thus naturally includes the periodic setting.*

The discrete stationarity framework allows us to use similar results as in the periodic homogenization. Indeed we can get average on large volumes.

Theorem 1.5. *Let $F \in L^\infty(\mathbb{R}^d, L^1(\Omega))$ be a stationary function. For $k \in \mathbb{Z}^d$, define $|k|_\infty = \sup_{1\le i\le d} |k_i|$, then*

$$\frac{1}{(2N+1)^d} \sum_{|k|_\infty \le N} F(x,\tau_k\omega) \underset{N\to\infty}{\longrightarrow} \mathbb{E}[F(x,\cdot)] \text{ in } L^\infty(\mathbb{R}^d), \text{ almost surely}$$

Hence, denoting $Q = (0,1)^d$ it holds that

$$F(\frac{x}{\varepsilon},\omega) \underset{\varepsilon\to 0}{\overset{\star}{\rightharpoonup}} \mathbb{E}[\int_Q F(x,\cdot)\, dx] \quad a.s. \tag{1.15}$$

Remark 1.6. *This result is analog to the Riemann Lebesgue lemma when F is Q-periodic:*

$$\forall \phi \in L^1(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} \phi(x) F(\frac{x}{\varepsilon})\, dx \underset{\varepsilon \to 0}{\longrightarrow} \int_Q F(y)\, dy \int_{\mathbb{R}^d} \phi(x)\, dx$$

We have the same behavior in the asymptotic regime compared to the periodic case. When ε goes to 0, u_ε the random function solution to (1.13) converges in some sense to $u^\star$ solution to

$$\begin{cases} -\text{div}(A^\star \nabla u^\star) = f \text{ in } D, \\ u^\star = 0 \text{ in } \partial D, \end{cases} \tag{1.16}$$

where $A^\star$ is again a constant deterministic homogenized matrix.

Figure 1.4: Example of a random stationary material: random checkerboard for different values of ε ($\varepsilon = 1/10$, $\varepsilon = 1/50$, $\varepsilon \ll 1$, from left to right).

Remark 1.7. *Compared to periodic homogenization, here we start from a random problem and get a deterministic homogenized problem in the asymptotic regime $\varepsilon \simeq 0$. This is a law of large number result.*

As in the periodic case, $A^\star$ can also be expressed in function of correctors functions. However, the corrector problem in the stochastic case cannot be reduced to a periodic cell, it has to be expressed on $\mathbb{R}^d$ as A is not periodic anymore.

$$\begin{cases} -\text{div}(A(\cdot,\omega)\,(\nabla w_i(\cdot,\omega) + e_i) = 0 \text{ in } \mathbb{R}^d, \\ \nabla w_i \text{ is stationary}, \\ \mathbb{E}[\displaystyle\int_Q \nabla w_i] = 0. \end{cases} \tag{1.17}$$

Then it holds that

$$A^\star e_i = \mathbb{E}\left[\int_Q A(w_i + e_i)\, dx\right] \tag{1.18}$$

Usually $A^\star$ does not have analytical formula, hence the usual way to approximate it is to compute w_i and use the formula (1.18). Contrary to the periodic case, estimating $A^\star$ is difficult in the random case: on one hand w_i is defined on $\mathbb{R}^d$ so it is not computable in practice; on the other hand, one must compute an average to get $A^\star$, that is use a costly Monte-Carlo approach.

In practice, w_i is approximated by w_i^N solution to

$$\begin{cases} -\text{div}\left(A(\cdot,\omega)\left(\nabla w_i^N(\cdot,\omega) + e_i\right)\right) = 0 \text{ in } Q_N = (-N,N)^d, \\ \nabla w_i \text{ is } Q_N\text{-periodic}. \end{cases} \tag{1.19}$$

Then $A^\star$ is approximated by the random matrix $A_N^\star$ defined by

$$A_N^\star e_i(\omega) = \frac{1}{|Q_N|}\int_{Q_N} A(\cdot,\omega)\left(\nabla w_i^N(\omega) + e_i\right). \tag{1.20}$$

Hence, we can decompose the error in two parts: the systematic error and the statistical error:

$$A^\star - A_N^\star(\omega) = \underbrace{A^\star - \mathbb{E}[A_N^\star]}_{\text{Systematic error}} + \underbrace{\mathbb{E}[A_N^\star] - A_N^\star(\omega)}_{\text{Statistical error}} \tag{1.21}$$

The systematic error was studied in the works from A. Gloria, F. Otto and collaborators to establish convergence rates and improve the systematic error, one can see for instance [46].

Usually $\mathbb{E}[A_N^\star]$ is estimated by a Monte-Carlo approach, that is a mean over M realizations: $\frac{1}{M}\sum_{m=1}^{M} A_N^{\star,m}(\omega)$. Such approaches are usually costly as the error usually decreases with a rate given by a Central limit theorem: $|A_N^{\star,M} - \mathbb{E}[A_N^\star]| \leq \frac{C}{\sqrt{M}}$. An efficient way to reduce this error is to apply variance reduction techniques, see for instance the works [65] by C. Le Bris, F. Legoll and W. Minvielle and [39] by J. Fischer. In practice, the statistical error is usually higher than the systematic error. Thus, variance reduction technique are critical to reduce the number of realizations required.

1.3.3 Contribution: Estimation of the fluctuations in a weakly stochastic regime

Homogenization theory describes what happens in the asymptotic regime ε goes to 0. The effective behavior is deterministic and the solution can be approximated by $u^\star$. However, if we take a step back and consider the regime where ε is small but not small enough to be in the asymptotic regime, then u_ε is still random. Hence, one can wonder if it is possible to characterize the law of u_ε from the knowledge of the distribution of A_ε or at least how to determine the mean behavior and how does u_ε fluctuates around its mean behavior.

As for the estimation of $A^\star$, we can separate the error between u_ε and $u^\star$ into two parts:

$$u^\star - u_\varepsilon(\cdot,\omega) = \underbrace{u^\star - \mathbb{E}[u_\varepsilon(x,\cdot)]}_{\text{Systematic error}} + \underbrace{\mathbb{E}[u_\varepsilon(x,\cdot)] - u_\varepsilon(x,\omega)}_{\text{Statistical error}} \tag{1.22}$$

As u_ε converges to $u^\star$ almost surely, both errors converge to 0 when $\varepsilon \to 0$. We consider later a scaling of these errors in $\varepsilon^{-d/2}$. Such scaling correspond to a similar one present in the Central limit theorem allowing to study fluctuations that are not vanishing in the asymptotic regime $\varepsilon \approx 0$. For instance, if we have a random checkerboard, A_ε depends on $N = \varepsilon^{-d}$ random variables associated with the scaling $\sqrt{N} = \varepsilon^{-d/2}$. At this scale the quantity $u_\varepsilon - u^\star$ diverges (explodes to ∞) when $d \geq 2$ and the systematic error problem that is how to approximate $\mathbb{E}(u_\varepsilon)$ thanks to $u^\star$ is still an open problem. In this thesis, we will only consider the statistical error.

For the problem of interest, in a one dimensional setting, there is a complete understanding of the fluctuations of u_ε. we refer to the works of Bal, Bourgeat and Piatninski in [19] and [8].

For the following problem

$$\begin{cases} -\Delta u_\varepsilon + V_\varepsilon(x,\omega)u_\varepsilon = f \text{ in } D, \\ u_\varepsilon(\cdot,\omega) = 0 \text{ on } \partial D, \end{cases} \tag{1.23}$$

there is also complete understanding in any dimension, one can refer to works [7] and [59].

Physicists and mechanicians are interested in the behavior of quantities of interest depending on the solution to problem (1.13). For instance, the randomness in the coefficient can embody a material with defects stemming from some industrial process. In this case, there is a need to quantify the fluctuations of the quantity of interest considered in order to ensure security or constraints requirements (take the civil engineering where a material is supposed to resist some external force).

To that end we consider the following quantity of interest:

$$I^\varepsilon(\omega, g) = \varepsilon^{-\frac{d}{2}} \int_D (u_\varepsilon(x,\omega) - \mathbb{E}[u_\varepsilon(x,\cdot)])\, g\, dx \tag{1.24}$$

Remark 1.8. *g can be taken as a localization function (for instance $\mathbb{1}_B$ with $B \subset D$) in order to get the local variations of the solution u_ε. If $g \in L^2(D)$ is the divergence of some vector field G then I^ε can be rewritten as*

$$I^\varepsilon(\omega, g) = \varepsilon^{-\frac{d}{2}} \int_D (\nabla u_\varepsilon - \mathbb{E}[\nabla u_\varepsilon]) \cdot G \tag{1.25}$$

This formulation can be useful as it can express the fluctuations of a flux or stress.

In the work [31], F. Otto, M. Duerinckx and A. Gloria, managed to characterize of the fluctuations of (1.24) for a variant of the problem (1.16) posed on $\mathbb{R}^d$ with $f \in C_c^\infty(\mathbb{R}^d)$ and with discrete operators (finite differences on the cartesian grid $\varepsilon\mathbb{Z}^d$ rather than true derivatives). They showed that

$$I^\varepsilon \xrightarrow[\varepsilon\to 0]{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

where the variance σ^2 is given by

$$\sigma^2 = \int_{\mathbb{R}^d} (\nabla u^\star \otimes \nabla v^\star) : \mathcal{Q} : (\nabla u^\star \otimes \nabla v^\star)\,.$$

$\mathcal{Q}$ is a constant fourth-order tensor depending only on A, $A^\star$ and the corrector function w_p, while $u^\star$ and $v^\star$ are given by

$$-\mathrm{div}\,(A^\star \nabla u^\star) = f, \qquad -\mathrm{div}\,(A^\star \nabla u^\star) = g.$$

One aim of this thesis is to explore and extend this result for the case of continuous PDEs with derivatives (instead of finite differences). Another objective is to design a numerical approach to approximate the fourth-order tensor $\mathcal{Q}$. These results are detailed in CHAPTER 2.

To perform this study, we will consider a restricted framework: *the weakly stochastic case*. Such framework allows us to get quantitative estimates.

Weakly stochastic case

We recall the weakly stochastic case introduced in [26], [66] and [14]. We assume that the coefficient A_ε is defined by

$$A_{\eta,\varepsilon}(x,\omega) = A_{per}(\frac{x}{\varepsilon}) + \eta\, \chi(\frac{x}{\varepsilon},\omega), \tag{1.26}$$

with A_{per} a deterministic matrix-valued function that is Q-periodic, χ an almost surely elliptic bounded matrix-valued function that is stationary, and η a small parameter. We also assume that A_{per} is symmetric.

We define χ by

$$\begin{cases} \chi(x,\omega) = \displaystyle\sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x) X_k(\omega)\, I_d, \\ X_k \text{ are i.i.d., almost surely bounded variables,} \\ \mathbb{E}[X_0] = 0, \end{cases} \tag{1.27}$$

where I_d is the identity matrix in dimension d.

Remark 1.9. *In material science, in materials (for instance when, manufacturing) defects can appear, so the resulting material will be the perturbation of some ideal material by some defects. Usually this translates as a matrix-valued coefficient that will have a deterministic part and a small random perturbation. There are however several ways to formalize the notion of small random perturbation.*

First, one can consider that the defect often occurs but is very small. Alternatively, one can consider a significant defect but with little chance of happening.

The weakly stochastic case (1.26) that we consider here corresponds to the first case. For instance, for industry material design, there can be intrinsic uncertainties in the process related (material components properties are known up to a measurement). In this case there will be many defects, though they each introduce little changes. The other case has to do with rare events, faulty design due to for instance some machine failure.

These two approaches seem similar at first glance. However, mathematically and with regard to the homogenization process they do not give the same results.

When $\eta \ll 1$, we can expand the problem in a series in powers of η. We then have:

$$\begin{cases} u_\varepsilon = u^0_\varepsilon + \eta\, u^1_\varepsilon + O(\eta^2) \\ \nabla w_i = \nabla w^0_i + \eta \nabla\, w^1_i + O(\eta^2) \\ A^\star = A^\star_{per} + O(\eta^2) \\ u_\star = u^0_\star + O(\eta^2) \end{cases} \tag{1.28}$$

where u_ε is solution to (1.2), u^0_ε is also solution to (1.2) for the coefficient A_{per} and u^1_ε satisfies $-\mathrm{div}(A_{per}(\frac{\cdot}{\varepsilon})\nabla u^1_\varepsilon) = \mathrm{div}(A_1(\frac{\cdot}{\varepsilon},\omega)\nabla u^0_\varepsilon)$ with homogeneous Dirichlet boundary conditions. The function w^0_i is the corrector in the direction e_i solution to 1.17 with the deterministic coefficient A_{per} and w^1_i is solution to

$$\begin{cases} -\mathrm{div}(A\nabla w^1_i) = \mathrm{div}(\chi(\nabla w^0_i + e_i)) \text{ in } \mathbb{R}^d, \\ \nabla w^1_i \text{ is stationary,} \\ \mathbb{E}[\int_Q \nabla w^1_i] = 0. \end{cases} \tag{1.29}$$

Finally, $A^\star_{per}$ and $u^0_\star$ are defined as the effective coefficient (1.18) and the corresponding homogenized limit (1.16) for the deterministic coefficient periodic A_{per} respectively.

Remark 1.10. *Here one can notice that $A^\star_1$ denoting the first order term in the expansion of $A^\star$ in power of η does not appear. Indeed, we have $\mathbb{A} = 0$ since $\mathbb{E} = 0$. By simple computations $A^\star_1 = \int_Q (\nabla w^0_i + e_i) \cdot \mathbb{E}(A_1)(\nabla w^0_i + e_i) = 0$*

Determine the asymptotic law

The main result is the characterization of the law of I_ε at the first order in η.

We formally expand our quantity of interest in a power series of η:

$$I_\varepsilon = I^0_\varepsilon + \eta I^1_\varepsilon + h.o.t$$

where I^0_ε and $I^1_\varepsilon = \varepsilon^{-\frac{d}{2}} \int_D \left(u^1_\varepsilon - \mathbb{E}[u^1_\varepsilon]\right) g$. From here we investigate the behavior of the random variable I^1_ε when $\varepsilon \to 0$ and $\eta \ll 1$.

Theorem 1.11. *Assume that (1.26) and (1.27) hold and that and A_{per} is an Hölder continuous function.*

Defining

$$I^1_\varepsilon = \varepsilon^{-\frac{d}{2}} \int_D \left(u^1_\varepsilon - \mathbb{E}[u^1_\varepsilon]\right) g. \tag{1.30}$$

Then it holds that

$$I^1_\varepsilon \xrightarrow[\varepsilon \to 0]{\mathcal{L}} \mathcal{N}(0, \sigma^2), \tag{1.31}$$

where $\mathcal{N}(0, \sigma^2)$ is a centered Gaussian random variable of variance

$$\sigma^2 = \int_D (\nabla u^\star_0 \otimes \nabla v^\star_0) : \mathcal{Q}^1 : (\nabla u^\star_0 \otimes \nabla v^\star_0) \tag{1.32}$$

where $u^\star_0$ and $v^\star_0$ the solutions to

$$-\operatorname{div}(A^\star_{per} \nabla u^\star_0) = f \text{ in } D\,,\ u^\star_0 = 0 \text{ on } \partial D, \tag{1.33}$$

and

$$-\operatorname{div}(A^\star_{per} \nabla v^\star_0) = g \text{ in } D\,,\ v^\star_0 = 0 \text{ on } \partial D, \tag{1.34}$$

and $\mathcal{Q}^1$ a fourth order tensor given by

$$\mathcal{Q}^1_{i,j,k,l} = \operatorname{Var}(X_0) \left(\int_Q (e_i + \nabla w^0_i) \cdot (e_j + \nabla w^0_j) \right) \left(\int_Q (e_k + \nabla w^0_k) \cdot (e_l + \nabla w^0_l) \right). \tag{1.35}$$

This is a similar result as in the work [31] at the first order in η for continuous PDE on bounded domains. This result shows that the law of our quantity of interest becomes Gaussian (and we have a formula for the variance) as ε goes to 0 and $\eta \ll 1$. So we have a characterization of the law of our quantity of interest in the regime $\varepsilon \ll 1$. In this case all laws and values are explicit, hence one can use it as a test-case to design numerical approximations for broader frameworks.

Estimating the variance: Computation of $\mathcal{Q}$

The tensor $\mathcal{Q}^1$ obtained in THEOREM 1.11 is directly derived from the study of I_1^ε. In order to build a numerical approach estimating $\mathcal{Q}$ that is valid in the general context of discrete stationarity (and not only in the weakly stochastic case), we need to introduce a general formula for $\mathcal{Q}$ and show that its leading order term (when $\eta \ll 1$) is given by $\mathcal{Q}^1$ defined by (1.35).

In the work [31], $\mathcal{Q}$ is defined as the limit of an object $\mathcal{Q}_L$ when L goes to infinity:

$$Q^L = \frac{1}{|Q_L|}\mathrm{Cov}\left(\int_{Q_L} \rho_{i,j}, \int_{Q_L} \rho_{k,l}\right) \tag{1.36}$$

with $\rho_{i,j}$ a random function defined as

$$\rho_{i,j} = (\nabla w_i + e_i)\cdot A(\nabla w_j + e_j) - \nabla w_i A^\star e_j - \nabla w_j A^\star e_i \tag{1.37}$$

We can expand $\rho_{i,j}$ in power of η and likewise we expand $\mathcal{Q}^L$ in power of η:

$$\rho_{i,j} = \eta\rho_{i,j}^1 + O(\eta^2) \tag{1.38}$$

$$\mathcal{Q}^L = \mathcal{Q}^{L,1} + O(\eta^3) \tag{1.39}$$

Theorem 1.12. *Assume that, we are in the weakly stochastic framework (1.27) and (1.26). Introduce*

$$Q^{L,1} = \frac{1}{|Q_L|}\mathrm{Cov}\left(\int_{Q_L} \rho_{i,j}^1, \int_{Q_L} \rho_{k,l}^1\right) \tag{1.40}$$

where $\rho_{i,j}^1$, is the first order term in η of the function $\rho_{i,j}$. Then $\mathcal{Q}^{L,1}$ is the leading order term (in the expansion in η) of $\mathcal{Q}$ and we have

$$\mathcal{Q}^1 = \lim_{L\to\infty} \mathcal{Q}^{L,1} \tag{1.41}$$

where $\mathcal{Q}^\infty$ is defined by 1.35.

So we have shown that in the weakly stochastic case the asymptotic variance is governed by a fourth order tensor $\mathcal{Q}^{L,1}$ whose definition is consistent with that given in [31].

Usually, the correctors in stochastic homogenization cannot be computed exactly as the associated equation is posed on $\mathbb{R}^d$. Hence, w_i is often approximated by w_i^N solution to (1.19). We define $\mathcal{Q}^{L,N,1}$ the approximation of $\mathcal{Q}^{L,N,1}$ where w_i^N is used instead of w_i and shows the following result.

Theorem 1.13. *Assume we are in the weakly stochastic framework, that A_{per} is an Hölder continuous function that N and L are chosen such that $N > L$. Then*

$$\lim_{L\to\infty} \mathcal{Q}^{L,N,1} = \mathcal{Q}^1 \tag{1.42}$$

Moreover, whenever $N > L$, it holds that

$$|\mathcal{Q}^{L,N,1} - \mathcal{Q}^1| \leq C\frac{\ln(L)^2}{L} \tag{1.43}$$

This shows that in the weakly random case, the approximation strategy for estimating $\mathcal{Q}^1$ is converging. We can explore this new strategy for broader frameworks.

In the definition of $\mathcal{Q}^{L,N,1}$ the covariance is used. In practice, we do not have access to this value and have to use the empirical covariance computed from M realizations. Hence for the general case we approach $\mathcal{Q}$ by $\mathcal{Q}^{L,N,M}$, with N the length of the definition domain in the corrector problem (1.19), L the integration domain used in the definition of $\mathcal{Q}^L$ and M the number of realizations to compute the empirical covariance.

Approximate the variance of I^ε: Random Checkerboard

During this thesis, we performed extensive numerical tests to show that the approach designed for the weakly stochastic case can be applied in a general discrete stationary framework and give accurate results.

We considered the checkerboard case

$$A_(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x) X_k(\omega), \tag{1.44}$$

with X_k i.i.d, such that $\mathbb{P}(X_k = \alpha_{min}) = \mathbb{P}(X_k = \alpha_{max}) = 0.5$.

For such choice, we considered for multiple f and g the quantity of interest I_ε defined by (1.24). We computed the empirical distribution of I_ε by using a brute force Monte-Carlo approach for multiple values of ε.

Remark 1.14. *One can notice that I_ε depends on the right-hand side f in the problem (1.13) and on the test function g in (1.24). Hence, if f or g changes the Monte-Carlo approximation must be computed again.*

We computed $\mathcal{Q}^{L,N,M}$ for large values of L, N, M in order to reach an asymptotic regime. First, we check that the distribution of I_ε seems to be a Gaussian when ε is small by looking at histograms, QQ-plots, and by performing the Shapiro-Wink test. Second, we check that the empirical variance is close to the variance computed from $\mathcal{Q}$ that is

$$\sigma^2 = \int_D (\nabla_u^\star \otimes \nabla v^\star) : \mathcal{Q} : (\nabla_u^\star \otimes \nabla v^\star) \tag{1.45}$$

It turns out that for different f and g the asymptotic regime seems to be reached quite quickly for reasonably small values of ε ($\varepsilon < 1/50$) and that (1.45) is an accurate approximation of the variance of I^ε. Indeed, the variance estimated with $\mathcal{Q}$ and the empirical variance shows overlapping confidence intervals and relative errors under 10%. Although the influence of N and L is consistent with the weakly stochastic case, we observe that the number of realizations M must be very large in order to give small enough confidence intervals.

So in this work we designed a numerical approach that allows to quantify the fluctuations of the quantity of interest I^ε without resorting to costly Monte-Carlo approaches, just by computing an approximation of $\mathcal{Q}$ that governs the fluctuations.

1.4 Numerical approaches

The homogenization approach presented in the last section gives very effective approximations and quantitative results. However, this mathematical theory requires some structure assumptions on A_ε mostly a separation of scales between the micro-scale and the macro-scale. In practice such assumptions are not always met. For instance, one can have a high definition (with resolution h) image of a composite material and notices heterogeneities at the scale ε and want to solve the problem (1.2). In such case, the homogenization theory will not apply and generic multi-scale techniques have to be developed.

One can distinguish two cases regarding what the practitioner wants to achieve. The first goal can be to compute an accurate approximation of u_ε at a meso-scale H

that is larger than h the size of the coefficient data. In this case, one will mostly use bottom-up approaches, some local computations will be used to enrich a coarse approximation. Another possibility can be to get a very accurate approximation at the scale h. Usually, it involves iterative approaches from the domain decomposition field using preconditioners linking local and coarse formulations of the problem.

We will consider mostly applications associated with the first case: we have a multi-scale problem, and we want to get an approximation with accuracy at a meso-scale H.

1.4.1 General principle and main approaches

There exists numerous numerical approaches to tackle multi-scale problems. We will only present here three types of methods. We will present shortly the Heterogeneous multi-scale method (HMM) introduced by W. E and B. Engquist in [32]. Then we will present the Local Orthogonal Decomposition (LOD) method developed by D. Peterseim and A. Målqvist in the work [70]. Finally, we will present the Multi-scale Finite Element method (MsFEM) introduced by T. Y. Hou in [55]. The MsFEM review will be more thorough as one goal to this thesis is to improve this method (see CHAPTER 3) and implement several of its variants (see CHAPTER 4).

HMM

This method was introduced by W. E and B. Engquist in the work [32] and analyzed in [33]. It shares similar features with the FE^2 approach developed by F. Feyel and J.-L.Chaboche in [37]. The principle of the approach is somehow inspired by homogenization results though it can be applied in broader frameworks. It is a bottom-up approach: computations on a fine grid of size h are used to increase the accuracy of a coarser approximation based on a coarse grid of size H. In a classical P1 FE formulation, the associated Galerkin problem gives the following stiffness matrix

$$\begin{aligned} A_{i,j} &= \int_D \nabla\phi_i \cdot (A_\varepsilon \nabla\phi_j) \\ &\simeq \sum_{K\in\mathcal{T}_H} \int_K \nabla\phi_i \cdot (A_\varepsilon \nabla\phi_j) \\ &\simeq \sum_{K\in\mathcal{T}_H} \sum_{x_k\in Quad(K)} w_k \left(\nabla\phi_i \cdot (A_\varepsilon \nabla\phi_j)\right)(x_k) \end{aligned}$$

Indeed, the integral over the whole domain is computed by summing components element by element. Each of these integrals cannot be computed analytically in practice, and their value is approximated by a quadrature formula with weights w_k and quadrature points x_k.

The main idea of HMM is to replace the evaluation $A_\varepsilon(x_k)$ by an effective coefficient $A^\star_{HMM}(x_k)$ computed at a fine scale h in a small patch ω_k around the quadrature point x_k.

More precisely, $A^\star_{HMM}(x_k)$ is defined as

$$A^\star_{HMM}(x_k) = \frac{1}{|\omega_k|} \int_{\omega_k} A_\varepsilon \nabla w_i \tag{1.46}$$

where w_i is a solution to

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla w_i) = 0 \text{ in } \omega_k, \\ \dfrac{1}{|\omega_k|}\displaystyle\int_{\omega_k} \nabla w_i = e_i. \end{cases} \tag{1.47}$$

We denote by u_{HMM} the HMM approximation of $u^\star$.

If one applies such idea in the periodic homogenization case where $A_\varepsilon = A(x, \frac{x}{\varepsilon})$ with A periodic with respect to its second variable. Then $A^\star$ is still defined though it is a function of x now. Hence, if we take $\omega_k = \varepsilon Q + x_k$ where $Q = (-1/2, 1/2)^d$, then $A^\star_{HMM}(x_k)$ corresponds exactly to the homogenized coefficient. Then classical periodic homogenization results allows to control the error between $u^\star$ and u_ε. This reasoning leads to the following result

Theorem 1.15 (see [33] THEOREM 1.2). *Assume $A_\varepsilon = A(x/\varepsilon)$ where A is $\mathbb{Z}^d$-periodic. Denoting $u^\star$ the solution to the homogenized problem (1.8), assuming $u^\star \in H^2(D)$, we have*

$$\|u^\star - u_{HMM}\|_{H^1(D)} \leq C(\varepsilon + H) \tag{1.48}$$

with C independent of ε and H.

Remark 1.16. *Similar results can be obtained if the coefficient is locally periodic i.e. $A_\varepsilon = A(x, \frac{x}{\varepsilon})$ where $y \mapsto A(x, y)$ is $\mathbb{Z}^d-$periodic.*

This approach allows to get an accurate approximation at the scale H. We note here that the computational load is manageable since fine scale computations are performed only on small patches around the quadratures points. Moreover, the fine scale computations associated with quadrature points are independent of each other and can be performed in parallel.

This method can adapt to changes of scale within A_ε by changing locally the size of the patches. Hence, it applies to a broad range of functions A_ε. The specificity of their approach is that this is not a Galerkin approximation of the problem (1.3): the variational formulation is changed to enable local homogenization on quadrature points. This is in sharp contrast with thd LOD and MsFEM methods that aim at designing relevant approximation spaces and solve the original heterogeneous problem.

LOD

This method has been developed by D. Peterseim and A. Målqvist, see [70]. It is inspired by the Variational Multi-scale approach introduced by Hughes in [57]. The principle of the approach is to exploit the symmetry of the coefficient A_ε in order to design a finite dimensional approximation space with good properties.

We consider a quasi-interpolant $\mathcal{I}_H$ on a shape regular triangulation $\mathcal{T}_H$:

$$\|u - \mathcal{I}_H(u)\|_{H^1(K)} + H_K^{-1}\|u - \mathcal{I}_H(u)\|_{L^2(K)} \leq C_{\mathcal{I}_H}\|\nabla u\|_{L^2(\omega_K)}, \tag{1.49}$$

for all $K \in \mathcal{T}_H$, and $\omega_K = T \in \mathcal{T}_H \,|\, K \cap T \neq \emptyset$ and for all $u \in H^1_0(D)$, where C_{I_H} is bounded with respect to H.

We can consider $\mathcal{I}_H$ the nodal weighted Clement interpolant defined by

$$\mathcal{I}_H(u)(x_k) = \frac{\int_D (u\phi_k)}{\int_D \phi_k}, \tag{1.50}$$

with ϕ_k the P1 FE basis function associated with the vertex k.

We consider $V^f = \{f \in H_0^1(D) : \mathcal{I}_H(f) = 0\}$, the kernel of the interpolant, that is the fine scales that are not captured by the coarse interpolant.

A_ε is symmetric, hence the associated bilinear form a_ε defines a scalar product on $H_0^1(D)$. Thus, we can define V_{LOD}^H the orthogonal of V^f with respect to a_ε in $H_0^1(D)$:

$$H_0^1(D) = V_{LOD}^H \oplus V^f \tag{1.51}$$

One can define $\mathcal{P}$ the a_ε-projection from the P1 FE space V_H to V^f. By definition, $V_{LOD}^H = V_H - \mathcal{P}V_H$. Hence, V_{LOD}^H is of the same dimension as V_H and is good candidate for a coarse space approximation.

One can design correction functions $\phi_i^{LOD} \in V_f$ defined by

$$a_\varepsilon(\phi_i^{LOD}, v) = a_\varepsilon(\phi_i, v) \text{ for all } v \in V^f. \tag{1.52}$$

Then we have a basis of V_{LOD}^H: $V_{LOD}^H = \text{Span}(\phi_i^{LOD} - \phi^i, i = 1..Nb_{vertex})$.

The correction functions ϕ_i^{LOD} do not have a compact support in contrast to P1 FE functions ϕ^i.

We consider new corrections $\phi_i^{LOD,k}$, which are localized versions of ϕ_i^{LOD}. We denote by $w_{i,k}$ the patch around the vertex i which is enlarged by k layers of coarse elements and define $\phi_i^{LOD,k} \in V^f(w_{i,k})$ solution to

$$a_\varepsilon(\phi_i^{LOD}, v) = a_\varepsilon(\phi_i, v) \text{ for all } v \in V^f(\omega_{i,k}) \tag{1.53}$$

with $V^f(\omega_{i,k}) = \{v \in V^f : v|_{D\setminus\omega_{i,k}} = 0\}$

In the case where the global corrections are available then it holds that

Theorem 1.17 (see LEMMA 3.1 in [70]). *Denote by u_{LOD}^H the approximation computed by solving the Galerkin problem on the approximation space V_{LOD}^H. Then*

$$\|u_\varepsilon - u_{LOD}^H\|_{H^1(D)} \leq CH\|f\|_{L^2(D)} \tag{1.54}$$

The above result holds as soon as $A_\varepsilon \in L^\infty(D)$, no additional regularity is required.

In the practical case, for localization correction functions it holds that

Theorem 1.18 (THEOREM 3.6 in [70]). *Denote by $u_{LOD,k}^H$ the approximation computed by solving the Galerkin problem on the approximation space $V_{LOD,k}^H$ for $k \simeq \ln(1/H)$. Then*

$$\|u - u_{LOD,k}^H\|_{H^1(D)} \leq CH \tag{1.55}$$

with C again independent of the characteristic period of oscillations of A_ε.

Such result is interesting as the LOD approximation error does not depend at all on ε and on the regularity of A_ε. The specificity of this approach is that the approximation space contains nodal basis functions that are solutions to problems with enlarged support $O(\ln(1/H))$. Though ε does not appear in the analysis, the correction functions must be solved at the fine scale on the patches $w_{i,k}$. The computational cost can be reduced as the correction functions are solution to independent problems. The proof relies critically on the symmetry of A_ε.

1.4.2 MsFEM

Principle

The Multi-scale Finite Element method has been introduced by Hou [55] and analyzed in [56]. We also refer to the monograph [34].

The Multi-scale Finite element method is a two-step approach: first, one designs an adapted basis that encodes the material heterogeneities; second, one solve the Galerkin problem on the approximation space spanned by the basis built in the first step.

We consider meshes of two sizes: a coarse mesh of size H, and a fine mesh of size h. Denoting ε by the smallest characteristic length of the heterogeneities, we assume that $h < \varepsilon < H$. h is supposed small enough to completely capture the fluctuations of the coefficient A_ε, so that the resulting approximation u_h of the solution u_ε to (1.2) would be accurate.

The goal here is not to approach u_h, but to design an approximation at the coarse scale H that would behave like the Finite Element in the laplacian case (see 1.2).

One way to do that is to design basis functions that satisfy similar properties as the Finite Element basis functions on the coarse mesh but oscillate like the coefficient A_ε. Denoting $i = 1..Nb_{vertex}$ the index of the collection of interior vertices in the coarse mesh, then in the P1 FE case basis functions are just piecewise affine functions such that $\phi_i(x_j) = \delta_{i,j}$. One way to design such basis functions would be to mimic the P1 FE on the edges of the coarse mesh and introduce oscillations similar to A_ε inside of each coarse element. In order to introduce oscillations, the new basis functions would satisfy some PDE inside each element (which would be solved in practice on a fine mesh of size $h \ll \varepsilon$).

The choice of the boundary conditions on the edges and the choice of the equation to be solved give birth to numerous variants of the MsFEM approach. We will only consider the following variants: the linear-MsFEM, the oversampling MsFEM (see [55]) and the "MsFEM à la Crouzeix-Raviart" (see [63] and [64]).

The first approach is conformal, that is the span of the basis functions V_{MsFEM} is included in $H_0^1(D)$, so that a classical Galerkin method can be used to study the error. The two other approaches are not conformal as the basis functions are not continuous across the edges.

We will give a quick review of these three methods in the subsequent sections. The implementation of the three methods will be further detailed in CHAPTER 4.

Linear version

It is the simplest MsFEM variant. The design of the MsFEM basis functions consists in mimicking the P1 FE boundary conditions on each element of the coarse mesh.

Indexing by $i = 1..Nb_{vertex}$ the set of interior vertices of the coarse mesh, we define ϕ_i^{MsFEM} the MsFEM basis function associated with vertex i. On each element

K containing the vertex i, ϕ_i^{MsFEM} satisfies:

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla \phi_i^{MsFEM}) = 0 \quad \text{in } K \\ \phi_i^{MsFEM} \text{ is affine on the edges} \\ \phi_i^{MsFEM}(x_j) = \delta_{i,j} \end{cases} \tag{1.56}$$

In practice, we do not have access to ϕ_i^{MsFEM}. We build $\phi_i^{MsFEM,h}$, an approximation of ϕ_i^{MsFEM} on a finer embedded grid of mesh size h, with P1 FE.

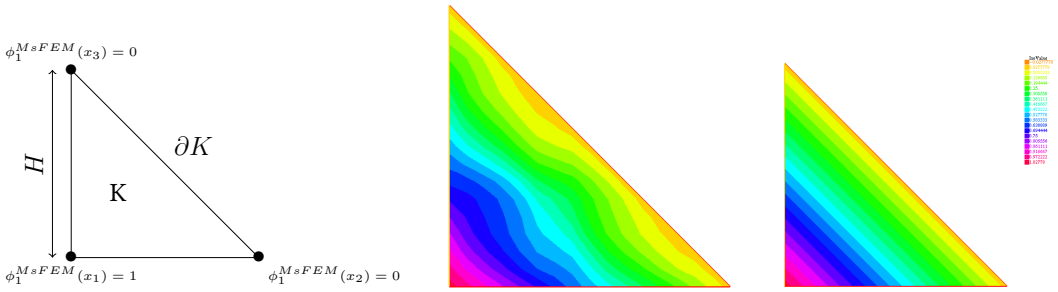


Figure 1.5: Sketch of MsFEM basis function design in 2D (left), Example of MsFEM basis function for an oscillating coefficient (middle) and P1 piecewise function (right)

Remark 1.19. *When A_ε is proportional to the identity matrix and constant then the basis functions are exactly the P1 functions of the standard FE approach. Hence in the regime $H \ll \varepsilon$, the linear MsFEM approximation will behave like the P1 Finite Element approximation (yielding linear convergence in H).*

Then we introduce $V_{MsFEM} = \operatorname{Span}\{\phi_i^{MsFEM}, i = 1..Nb_{vertex}\}$ and solve the associated Galerkin problem: Find $u_{MsFEM} \in V_{MsFEM}$ such that

$$a_\varepsilon(u_{MsFEM}, v) = \int_D (A_\varepsilon \nabla u_{MsFEM}) \cdot \nabla v = b(v) = \int_D fv \quad \text{for all } v \in V_{MsFEM} \tag{1.57}$$

Denoting K_{MsFEM}, B_{MsFEM} and U_{MsFEM} by

$$\begin{cases} K_{MsFEM,i,j} = a_\varepsilon(\phi_i^{MsFEM}, \phi_j^{MsFEM}), \quad B_{MsFEM,i} = b(\phi_i^{MsFEM}) \\ U_{MsFEM} \text{ solution to } K_{MsFEM} U_{MsFEM} = B_{MsFEM}, \end{cases}$$

We have $u_{MsFEM}(x) = \sum_{i=1}^{Nb_{vertex}} U_{MsFEM} \phi_i^{MsFEM}(x)$.

Theorem 1.20 (see [56], THEOREM 5.1). *Assuming $A_\varepsilon = A_{per}(\frac{x}{\varepsilon})$ with A_{per} a $\mathbb{Z}^d$ periodic function, it holds that*

$$\|u_\varepsilon - u_{MsFEM}\|_{H^1(D)} \leq C \left(\sqrt{\varepsilon} + H + \sqrt{\frac{\varepsilon}{H}} \right) \tag{1.58}$$

FIGURE 1.6 illustrates THEOREM 1.20. Indeed, we consider here the problem (1.2) posed on $D = (0,1)^2$ with A_ε an ϵ-periodic matrix function. In this case we can see that the numerical results are consistent with the bound of THEOREM 1.20, we can identify the three regimes: $H \gg \varepsilon$, $H \ll \varepsilon$ and $H \simeq \varepsilon$.

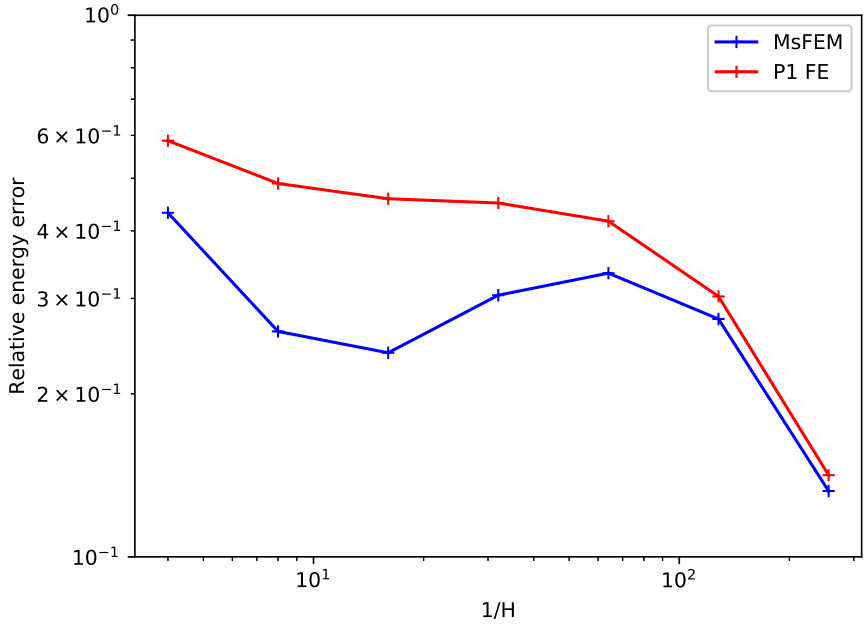


Figure 1.6: Relative energy error of linear MsFEM approximation function of $1/H$ for the 2D problem (1.2) where A_ε is a ε-periodic function ($\varepsilon = 1/32$)

When $H \gg \varepsilon$, the MsFEM approximation error decreases linearly and gives lower errors than the P1 FE error which displays a plateau. However, when $H \simeq \varepsilon$, the error is increasing again showing that the term $\sqrt{\frac{\varepsilon}{H}}$ controls the behavior of the error. Finally, when $H \ll \varepsilon$ the MsFEM approximation error is decreases linearly in H and is similar to the P1 FE. Such behavior can be explained since on each element K, A_ε is close to a constant when $H \ll \varepsilon$ and the MsFEM basis functions correspond to the P1 basis functions.

In the regime $H \gg \varepsilon$, the linear MsFEM approximation displays an interesting behavior as it decreases linearly in H. However, when $R = \frac{H}{\varepsilon}$ is very large, the computation of the basis functions solution to (1.56) can be very expensive because one has to solve a linear system with a number of DOF that is higher than R^d (and thus very large). Usually this method is used when R is one order of magnitude (close to 10). In that case the resonance effect can take place and dampen the MsFEM gain. That is why, though the linear MsFEM is useful, there is a need for better variants.

The conformal MsFEM methods such as the linear MsFEM have a particular interpretation in the symmetric case (that is when the matrix function A_ε is symmetric). Indeed, in this case the bilinear form associated with the variational formulation (1.3) defines a scalar product in $H_0^1(D)$ and the problem (1.2) is equivalent to a minimization problem.

Interpretation in the symmetric case

When the matrix function A_ε is symmetric, then solving (1.2) is equivalent to solve the following energy minimizing problem

$$u_\varepsilon = \underset{v \in H_0^1(D)}{\operatorname{argmin}} \left(\frac{1}{2} a_\varepsilon(v, v) - b(v) \right), \tag{1.59}$$

where a_ε and b are the bilinear form and linear form associated with the variational formulation (1.3).

The bilinear form a_ε is a scalar product with respect to the $H_0^1(D)$. Considering a 2D regular coarse mesh $\mathcal{T}_H$ with mesh size H, we denote its interior edges by $\Gamma = \bigcup_{K\in\mathcal{T}_H} \partial K \setminus \partial D$. Then $H_0^1(D)$ can be decomposed into two a_ε-orthogonal spaces.

$$H_0^1(D) = \{\oplus V_K\}_{K\in\mathcal{T}_H} \oplus V_\Gamma = V_B \oplus V_\Gamma, \tag{1.60}$$

where

- V_B is the space of functions that are in $H_0^1(D)$ such that for each element $K \in \mathcal{T}_H$ their restriction to K belong to $H_0^1(K)$, and that we call bubble functions
- V_Γ is the space of functions which are a_ε-harmonic in each H, and that we call interface functions.

We define the a_ε-lifting operator

$$E_D : \overset{H^{1/2}(\Gamma)}{\tau} \mapsto \overset{H_0^1(D)}{E_D(\tau)}$$

such that for any $\tau \in H^{1/2}(\Gamma)$n the function $E_D(\tau)$ satisfies

$$\begin{cases} -\mathrm{div}\left(A_\varepsilon \nabla E_D(\tau)\right) = 0 \text{ in } K \text{ for all } K \in \mathcal{T}_H, \\ E_D(\tau) = \tau \text{ on } \Gamma, \\ E_D(\tau) = 0 \text{ on } \partial D. \end{cases} \tag{1.61}$$

Remark 1.21. *Denote $V_\gamma = \{\gamma_\Gamma(u) : u \in H_0^1(D)\}$ with γ_Γ the trace operator from $H_0^1(D)$ to $H^{1/2}(\Gamma)$. Then V_Γ can be seen as the a_ε-lifting of V_γ, that is $V_\Gamma = \{E_D(\tau) : \tau \in V_\gamma\}$. Moreover, the a_ε-lifting can be seen as the solution to the following minimization problem*

$$E_D(\tau) = \underset{v\in H_0^1(D)}{\mathrm{argmin}}\, a_\varepsilon(v,v) \text{ subject to } v|_\Gamma = \tau \tag{1.62}$$

Remark 1.22. *The MsFEM approximation of u_ε (solution to problem (1.57)) belongs to V_Γ as the basis functions can be seen as the a_ε-liftings of affine and piecewise affine continuous functions on Γ.*

In addition to the orthogonal decomposition (1.60), the scalar product a_ε provides us with a norm on $H_0^1(D)$: one define the Energy norm $\|\cdot\|_E$ as

$$\|u\|_E^2 = E(u) = a_\varepsilon(u,u) \tag{1.63}$$

Using the boundedness and ellipticity assumptions on A_ε and thanks to the Poincaré inequality, it holds that the energy norm is equivalent to the $H^1(D)$ norm.

Hence, we can use the energy norm to study the error in the particular context of the linear MsFEM approach.

Thanks to the orthogonal decomposition (1.60), we have

$$u_\varepsilon = u_\varepsilon^B + u_\varepsilon^\Gamma \tag{1.64}$$

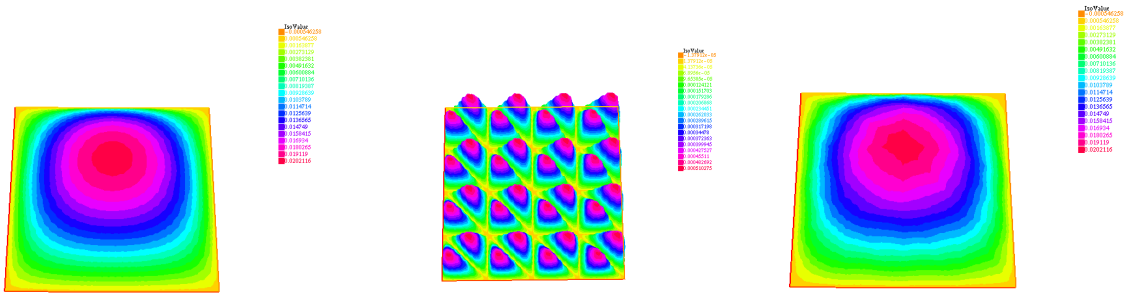


Figure 1.7: Decomposition of u solution to (1.2) posed in $(0.1)^2$, for a periodic A_ε with $\varepsilon = 1/32$ and $H = 1/4$. Solution u (on the left), u^B the bubble part (in the middle) and u^Γ the interface part (on the right)

with $u_\varepsilon^\Gamma = E_D(u_\varepsilon|_\Gamma)$ and $u_\varepsilon^B \in V_B$ satisfying

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla u_\varepsilon^B) = f \text{ in } K, \text{ for all } K \in \mathcal{T}_H, \\ u_\varepsilon^B = 0 \text{ on } \partial K, \text{ for all } K \in \mathcal{T}_H, \end{cases} \tag{1.65}$$

See FIGURE 1.7 for a representation of this decomposition.

The energy norm can also be decomposed into an interface and a bubble part. Indeed, it holds that

$$\|u_\varepsilon\|_E^2 = a_\varepsilon(u_\varepsilon, u_\varepsilon) = a_\varepsilon(u_\varepsilon^B, u_\varepsilon^B) + a_\varepsilon(u_\varepsilon^\Gamma, u_\varepsilon^\Gamma) = \|u_\varepsilon^B\|_E^2 + \|u_\varepsilon^\Gamma\|_E^2 \tag{1.66}$$

Recalling that u_{MsFEM} the MsFEM linear approximation of u_ε is in V_Γ the error in energy norm can be written as

$$\|u_\varepsilon - u_{MsFEM}\|_E^2 = \|u_\varepsilon^B\|_E^2 + \|u_\varepsilon^\Gamma - u_{MsFEM}\|_E^2 \tag{1.67}$$

Knowing that u_ε^B is solution to (1.65), thanks to the Poincaré inequality and the Lax-Milgram theorem it holds that

Theorem 1.23. *Consider u_ε^B solution to (1.65) then*

$$\|u_\varepsilon^B\|_{H^1(D)} \leq C\ H\ \|f\|_{L^2(D)},$$

with C independent of H and A_ε.

Thus, the bubble error decreases linearly in H. This result shows that approximating u_ε^B is not useful needed to obtain a convergence similar to the laplacian case. Hence, one way to get an efficient approximation is to use a Galerkin method with a space that represents V_Γ well enough.

Remark 1.24. *The linear MsFEM basis functions span a subspace of V_Γ explaining the better accuracy of linear MsFEM compared to standard FE. However, recalling* THEOREM *1.20 in the regime $H \simeq \varepsilon$ the approximation becomes inaccurate. In this case the approximation subspace V_{MsFEM} is not large enough to represent V_Γ, otherwise we would get a linear decrease of the error with respect to H.*

The main flaw of the linear MsFEM method is that on the edges u_ε is approximated by affine functions whereas u_ε usually oscillates at scale ε. Some alternatives have been designed to circumvent this issue. For instance, one can use oscillatory boundary conditions that are consistent with the oscillations of the coefficient. Such approach gives little improvement and still suffers from the resonance effect. One can also turn to non-conformal approaches in order for the artificial boundary conditions to have a smaller impact.

Oversampling approach

The oversampling approach have been introduced in the work [55] and further analyzed [35] and [45]. One can also refer to the following review [51] for further analysis.

It is a two step approach like the linear MsFEM method: constructing of the basis and resolution of a Galerkin problem. The basis is computed as follows, we agian consider a coarse mesh of size H and denote by ϕ_i^{OS} the unique MsFEM basis function associated with the vertex i. We consider for each element $K \in \mathcal{T}_H$, $\widetilde{K}$ an enlargement such that $K \subset \widetilde{K}$. We solve an equation on $\widetilde{K}$ and ϕ_i^{OS} is taken as the restriction of this solution over K. This way the basis function ϕ_i^{OS} oscillates on the edges forming ∂K.

More precisely, we define $\psi_i \in H^1(\widetilde{K})$ solution to

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla \psi_i) = 0 \quad \text{in } \widetilde{K} \supset K \\ \psi_i \text{ is affine on the edges of } \widetilde{K} \supset K \\ \psi_i(\widetilde{x_j}) = \delta_{i,j} \end{cases} \tag{1.68}$$

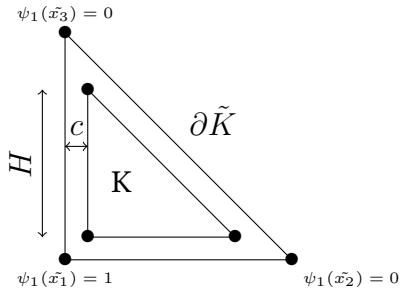


Figure 1.8: Sketching of oversampling MsFEM basis function design

Remark 1.25. *The basis functions ϕ_i^{OS} are usually discontinuous across the edges. Indeed, basis functions are taken as restriction of solution to problems on a smaller element. There is no guarantee, that for a given edge, the enlarged solutions will match for the elements sharing this edge. Hence, ϕ_i^{OS} does not belong to $H_0^1(D)$ and thus the approach is not conformal.*

Then we perform the online step that is solve the coarse Galerkine problem for any source term f, using the approximation space

$$V_{MsFEM-OS} = \mathrm{Span}(\{\phi_i^{OS}\}, i = 1..Nb_{vertex}) \tag{1.69}$$

and then find $U_{MsFEM-OS} \in V_{MsFEM-OS}$ such that for any $v \in V_{MsFEM-OS}$

$$A_\varepsilon(U_{MsFEM-OS}, v) = b(v)$$

In the work [45] a convergence result has been presented.

Theorem 1.26. *Let $f \in L^2(D)$ and $A_\varepsilon \in L^\infty(D, \mathbb{R}^{d\times d})$ be a sequence of elliptic and bounded matrices. We hence assume that there exists C and c such that*

$$\forall x \in D,\ \forall \xi \in \mathbb{R}^d,\ \forall \varepsilon,\ c|\xi|^2 \leq \xi^T A_\varepsilon \xi(x) \leq C|\xi|^2$$

We assume that A_ε is H-convergent (i.e. convergent in the sense of homogenization). We denote by $u_{MsFEM-OS}$ the oversampling MsFEM approximation. Denoting by $\tilde{K}$ the enlarged element and $|\tilde{K}|$ its associated volume, we let

$$\frac{|\tilde{K}| - |K|}{|K|} \underset{H\to 0}{\longrightarrow} 0.$$

Then it holds that

$$\lim_{H\to 0}\lim_{\varepsilon\to 0} \|u_\varepsilon - u_{MsFEM-OS}\|_{L^2(D)} = 0.$$

Remark 1.27. *This result is interesting as it ensures the convergence of the approach in the broadest sense of homogenization that is H-convergence. However, it does not provide insight as how to choose the enlargement.*

A more practical convergence result has been proven in [35] though in the more restrictive framework of periodic homogenization.

Theorem 1.28. *Let $d = 2$, $f \in L^2(D)$ and $A_\varepsilon = A_{per}(\frac{x}{\varepsilon})$ where A_{per} is $\mathbb{Z}^d$-periodic, bounded, elliptic, symmetric and $C^3(\bar{D})$. We denote by $u_{MsFEM-OS}$ the oversampling MsFEM approximation and by $\tilde{H}$ the characteristic size of the elements enlargements that is $\widetilde{H} = \min_{K\in\mathcal{T}_H} d(\partial\tilde{K}, K)$. Then it holds that*

$$\|u_\varepsilon - u_{MsFEM-OS}\|_{L^2(D)} \leq C\left(\frac{\varepsilon}{\widetilde{H}} + H + \varepsilon(\log H)^{\frac{1}{2}}\right),$$

$$\left(\sum_{K\in\mathcal{T}_H} \|\nabla u_\varepsilon - \nabla u_{MsFEM-OS}\|^2_{L^2(D)}\right)^{\frac{1}{2}} \leq C\left(\frac{\varepsilon}{\widetilde{H}} + H + \sqrt{\varepsilon}\right)$$

Remark 1.29. *This result shows an explicit rate in function of the enlargement rate, though in a more restrictive framework. However, in the resonance regime, $\widetilde{H}$ must be taken very large in order to have a good approximation. Numerically, one can observe that taking $\widetilde{H} = H + k\varepsilon$, with $k < 10$ is sufficient to significantly dampen the resonance error (see [55]).*

Although the convergence results seems to suffer from the resonance error like linear MsFEM unless a prohibitive enlargement is used ($\widetilde{H} \simeq 1/\varepsilon$), the method is very effective in practice even with small enlargement (only a few ε). The resonance error does occur (in the sense that the approach is not converging if H and ε go to 0 and $\frac{H}{\varepsilon}$ is fixed) but the overall error is significantly reduced in comparison to the linear MsFEM.

Crouzeix Raviart approach

Consider a material with small perforations and a coarse mesh as shown in FIGURE 1.9. In such case the perforations can intersect the coarse mesh. Hence, the linear or oversampling MsFEM cannot be applied since linear boundary conditions are enforced on subset of the edges where the solution is supposed to vanish.

In order to tackle such problems, an edge-based MsFEM called "MsFEM à la Crouzeix-Raviart" has been designed in the works [63] and [64]. We denote the set perforations on the domain D by B_ε. In the variational formulation of the problem (1.3), instead of computing the integrals on the whole domain D, the integrals are computed on $D\backslash B_\varepsilon$. During the offline stage two types of basis functions are computed: edge and bubble

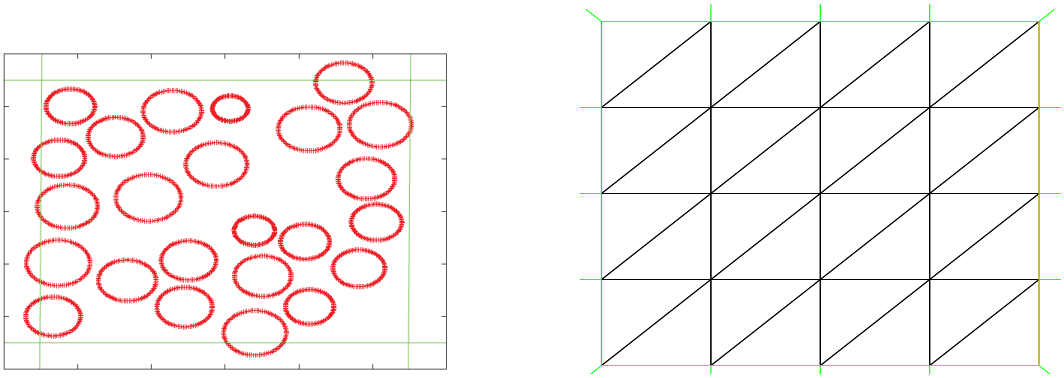


Figure 1.9: Left - Material with perforations, Right - Mesh used

basis functions. Similarly, to MsFEM oversampling, the method is not conformal, only a weak continuity property is enforced along the edges.

In each element K of the coarse mesh, the edge-based basis function ϕ_i^{CR} associated with the edge e_i is solution to the problem

$$\begin{cases} -\text{div}(A_\varepsilon \nabla \phi_i^{CR}) = 0 \quad \text{in } K, \\ \int_{e_i} \phi_i^{CR} = 1, \\ \int_{e_j} \phi_i^{CR} = 0, \ if \ i \neq j, \\ A_\varepsilon \nabla \phi_i^{CR} \cdot n_j = \lambda_{i,j} \text{ on } e_j, \text{ for any } j \end{cases} \tag{1.70}$$

where $\lambda_{i,j}$ is a constant (which can take different values on each side of the edge).

The bubble basis function ψ_K is supported by K and solves

$$\begin{cases} -\text{div}(A_\varepsilon \nabla \psi_K) = 1 \quad \text{in } K, \\ \psi_K = 0 \text{ on } \partial K. \end{cases} \tag{1.71}$$

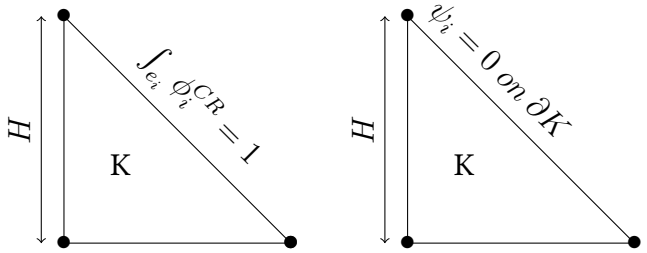


Figure 1.10: Edge-based basis function (Left), Bubble basis function (Right)

Edge-based basis functions are not continuous across the edges, as only the integral over the edges is prescribed. However, denoting by $[[\cdot]]$ the jump of the function across the edge, such functions satisfy a weak continuity property, that is for each edge e of the mesh, and for any basis function ϕ_i^{CR} we have $\int_e [[\phi_i^{CR}]] = 0$, likewise

$$\int_e [[u_{MsFEM}]] = 0 \tag{1.72}$$

as a direct consequence. Then a coarse Galerkin problem is solved in the space

$$V_{MsFEM-CR} = \text{Span}(\{\phi_i^{CR}\}, \{\psi_j\}, i = 1..Nb_{edges}, j = 1..Nb_{elements})$$

Regarding the convergence of the approach the following result has been established in [63].

Theorem 1.30. *Let u_ε solution to $-\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f$ with homogeneous Dirichlet boundary conditions on the domain $D \setminus B_\varepsilon$, for $d = 2$, with periodic perforations and $f \in H^2(D)$. We assume that the equation of any internal edge e can be written as $x_2 = \frac{p_e}{q_e} x_1 + c_e$ for $p_e \in \mathbb{Z}$, $q_e \in \mathbb{N}^\star$ that are coprime numbers such that $|q_e| \leq C$ with C independent of e and the mesh size H. Then it holds that*

$$\|u_\varepsilon - u_{MsFEM-CR}\|_{H^1_H(D\setminus B_\varepsilon)} \leq C\varepsilon \left(\sqrt{\varepsilon} + H + \sqrt{\frac{\varepsilon}{H}} \right) \|f\|_{H^2(D)}$$

where $\|u\|^2_{H^1_H(D\setminus B_\varepsilon)} = \sum_{K \in \mathcal{T}_H} \int_{K \cap D \setminus B_\varepsilon} |\nabla u|^2$ and C is independent of H, ε and f.

Remark 1.31. *The assumption on the rationality of the slopes in the mesh is necessary see [63, REMARK 2.5] to treat traces of periodic functions on the edges of the mesh. In full generality, such traces are almost periodic. In the case of rational slopes these traces are periodic simplifying the proof. This assumption is not very restrictive in practice, in particular because computers only manipulate rational numbers (and therefore, in practice, the mesh slopes are always rational).*

Remark 1.32. *This result is interesting as usually the MsFEM convergence results depend on the contrast that is the ratio between the maximum value of the coefficient and its minimum value. One can regard the problem posed in the perforated domain as the limit of a problem with an increasing contrast.*

Remark 1.33. *One could consider enriching the basis in order to improve the accuracy of the method. For instance, one can enrich the basis adding functions ϕ_i^{CR} satisfying $\int_e \phi_i^{CR} x = 0$ or 1 in addition to satisfying $\int_e \phi_i^{CR} = 0$ or 1. In such case, the weak continuity property would be ensured and the other constraints would make the solution closer and closer to a true continuous property. The resulting approximation would improve with respect to N the degree of the constraints and the regularity of the exact solution. In the case of perforations, the solution is not so regular. Indeed, perforations induces a loss in convexity and in regularity of the domain reducing the overall regularity of the solution. Thus, such an enrichment approach would give little improvement.*

1.4.3 Contribution: MsFEM enriched with polynomials

The main MsFEM methods (such as Linear, Oversampling and Crouzeix-Raviart) suffer from a resonance effect: the error does not decrease if H and ε go to 0 with $\varepsilon \simeq H$. Such behavior dampen the effectiveness of MsFEM methods. We aim at designing an MsFEM type method that would cancel this error and yield a linear convergence with respect to the coarse mesh size H.

We place ourselves in the symmetric framework that is the coefficient matrix A_ε is symmetric. Such framework allows us to use orthogonality and energy properties. We recall that in this particular the orthogonal decomposition (1.60) of $H^1_0(D)$ into the interface space V_Γ and the bubble space V_B. The MsFEM basis functions in the linear case belong to the interface space. Hence, when considering u_ε the solution of (1.2) and THEOREM 1.23, it holds that the resonance error is due to a bad approximation of the space V_Γ by the MsFEM basis functions.

The idea of enrich the space V_Γ has already been explored in the work [53] and leads to a new numerical approach: the special finite element method based on component mode synthesis. This approach was thoroughly analyzed in the work [52].

Special finite element method based on component mode synthesis

This approach developed in the work [53] is similar to a linear MsFEM method. It is a two-step approach: design of a conformal basis and then resolution of a coarse Galerkin problem. The difference with the linear MsFEM lies in the design of the basis: the linear MsFEM basis functions will be used but will be complemented by spectral enrichments on the edges. We emphasize on the fact that the method is conformal though the enrichments are edge-based.

The complementing basis functions are solution to local eigenvalues problems defined on each edge. Recalling the definition of a_ε-liftings in (1.61) and that a_ε is a scalar product on $H_0^1(D)$, for each interior edge e, we can define a generalized eigenvalue problem for traces that can be extended by 0. We define $H_{00}^{1/2}(e) = \{v \in L^2(e) : \widetilde{v} \in H^{\frac{1}{2}}(\Gamma)\}$ with $\widetilde{v}$ the extension by 0 of v in Γ. In that case, for each edge e, the following generalized eigenvalue problem is well-defined: find $(\tau_{e,i}, \lambda_{e,i}) \in V_\Gamma \times \mathbb{R}$ such that $\forall \eta \in H_{00}^{1/2}(e)$

$$a_\varepsilon(E_D(\tau_{e,i}), E_D(\eta)) = \int_D \nabla E_D(\tau_{e,i}) \cdot (A_\varepsilon \nabla E_D(\eta)) = \lambda_{e,i} \int_e \tau_{e,i}\eta \tag{1.73}$$

We next sort the eigenvalues $(\lambda_{e,i})$ in decreasing order. Then we define the approximation space as

$$V_{ACMS} = \text{Span}\left(\{\phi_i^{MsFEM}\}, \{\tau_{e,j}\}, e \text{ in } \Gamma, j = 1..I_e, i = 1..Nb_{vertex}\right)$$

Error estimates have been proved in the work [52]. Denoting the Special finite element approximation by u_{ACMS}, we have the following result:

Theorem 1.34. *Denoting by σ, the shape regularity of the coarse mesh, and by A_ε the coefficient in (1.2) it holds that if the solution to (1.2) is such that $u_\varepsilon \in H^{s_0}(D) \cap H_0^1(D)$ with $s_0 > \frac{3}{2}$ then*

$$|u_\varepsilon - u_{ACMS}|^2_{H^1(D)} \leq CH^2 + C_{s_0,\sigma,A_\varepsilon} H^{2s_0-3} \sum_{K \in \mathcal{T}_H} \frac{\|u_\varepsilon\|^2_{H^{s_0}(D)}}{\min\limits_{e \subset \partial K \cap \Gamma} \lambda_{I_e,e}}, \tag{1.74}$$

where Γ is the set of interior edges, C depends only on f and the shape of the elements and $C_{s_0,\sigma,A_\varepsilon}$ depend on s_0, σ, A_ε.

For particular geometries (such as regular shaped elements of size H), the work [20] suggests that there exists α_{min} independent of H and i such that $\lambda_{e,i} \geq C\alpha_{min}\frac{i}{H}$.

In this particular context we can write the following corollary

Corollary 1.35. *If we assume that $u \in H^2(D) \cap H_0^1(D)$ and that there exists $\alpha_{min} > 0$ such that*

$$\forall e \subset \Gamma,\ \lambda_{e,i} \geq \alpha_{min}\frac{i}{H}, \tag{1.75}$$

with α_{min} independent of $e \in \Gamma$, i and H then it holds that

$$|u_\varepsilon - u_{ACMS}|^2_{H^1(D)} \leq CH^2 \left(1 + \frac{\|u_\varepsilon\|^2_{H^2(D)}}{I}\right), \tag{1.76}$$

where I is the minimal number of enrichments taken in all edges belonging to Γ, and C depends only on the shape of the elements σ and on the coefficient A_ε.

Remark 1.36. *One can see that the result given by* THEOREM *1.34 does not rely on a particular structure assumption (periodicity, ...) for the coefficient A_ε. However, it requires that solution u_ε is of regularity at least $H^{s_0}(D)$ with $s_0 > \frac{3}{2}$. This implies some implicit assumptions regarding the regularity of the coefficient A_ε, the geometry of D and the right-hand side f.*

Remark 1.37. *Such result is an improvement compared to the linear MsFEM, indeed in the periodic case usually $\|u\|_{H^2(D)} \simeq \frac{1}{\varepsilon}$. Then, the estimate in* COROLLARY *1.35 becomes in the regime $\varepsilon \simeq H$:*

$$|u_\varepsilon - u_{ACMS}|^2_{H^1(D)} \leq CH^2 + \frac{C_{\sigma,A_\varepsilon}}{I}.$$

Then the resonance error associated with the linear MsFEM of order $O(1)$ can be decreased by adjusting I, the number of enrichments.

Remark 1.38. *Numerical experiments show a significant decrease of the error in the resonance regime compared to the linear MsFEM even for a few enrichments. The behavior of the error is consistent with the* ASSUMPTION *1.75.*

The special element method is effective and decreases the resonance error greatly compared to conformal versions of MSFEM (for instance linear MsFEM). Moreover, the enrichments have support in the elements across the edge simplifying the implementation of the method. We also put the emphasis on the fact that the convergence of the method is proven for generic A_ε without any structure assumptions. However, there are some drawbacks. First, the decrease of the eigenvalues plays a significant role in the effectiveness of the approach and is not yet understood (ASSUMPTION 1.75 has yet to be proven). Second, the eigenvectors have to be approximated by an FE approach on finer grid h. The stability of eigenvalues and eigenvectors regarding the value of h has yet to be studied. Finally, solving eigenproblems can be a computational challenge even as an offline step.

This approach allows us to understand better how to enrich V_Γ efficiently, and is the cornerstone to our new method with enriched polynomials.

The MsFEM enriched method with polynomials

This section presents the main results of CHAPTER 3 regarding the design of a new MsFEM enriched method based on polynomials. It is a two-step approach and is similar to the special finite element presented in the previous section. The main difference lies in the design of the enrichments. Indeed, our enrichments will be a_ε-liftings of polynomials instead of solutions to eigenproblems.

Let e be an interior edge of the coarse mesh ($e \subset \Gamma$) that is supposed regular in the sense (3.2), we define $\phi^{\Gamma}_{e,k}$ the edge enrichment of degree k with $1 < k \leq N$ such that on each element K containing the edge e

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla \phi^{\Gamma}_{e,k}) = 0 \text{ in } K \\ \phi^{\Gamma}_{e,k} = P_k \text{ on } e \\ \phi^{\Gamma}_{e,k} = 0 \text{ on } \partial K \setminus e \end{cases} \tag{1.77}$$

with P_k a polynomial of order k that vanishes at the vertices of the edge. The support of $\phi^{\Gamma}_{e,k}$ is thusly the two triangles sharing the edge e.

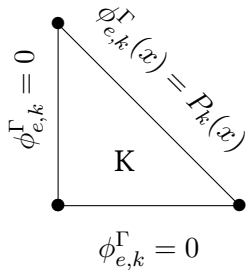


Figure 1.11: Sketch of the design of an enrichment

We define the approximation space

$$V_{MsFEM,N} = \operatorname{Span}\left(\{\phi_j^{MsFEM}\}_{j=1..Nb_{vertex}}, \{\phi^{\Gamma}_{e,k}\}_{e\subset\Gamma,\ k=2..N}\right) \tag{1.78}$$

with ϕ_j^{MsFEM} the nodal basis function associated with the vertex i in the linear MsFEM approach.

Remark 1.39. *By definition $\phi^{\Gamma}_{e,k}$ is an a_ε-lifting and thus belongs to V_Γ. Hence, adding $\phi^{\Gamma}_{e,k}$ brings our approximation space closer to V_Γ provided the enrichments are different enough. Moreover, the enrichments can be computed independently for both elements sharing the edge, hence a significant speed up of the offline phase. Note that for the special element method, the enrichments are solution to eigenvalues problems that cannot be solved independently for both elements sharing the edges. In our case, the offline cost is thus reduced for two reasons: we solve right-hand side problems rather than eigenvalue problems, and the problems are posed on a single element rather than two of them.*

We proved a similar convergence result as THEOREM 1.34

Theorem 1.40. *Assume that u_ε solution to (1.2) belongs to $H^1_0(D) \cap H^s(D)$ for some $s > \frac{3}{2}$ and that $f \in L^2(D)$. We consider $\mathcal{T}_H$ a regular mesh of D in the sense (3.2) with quadrangular (or triangular) elements and characteristic length H. Denoting by $u_{H,N}$ the solution of the Galerkin problem associated with (1.2) on the space $V_{MsFEM,N}$, it holds that*

$$\|u_\varepsilon - u_{H,N}\|^2_{H^1(D)} \leq C\left(H^2 + \|u_\varepsilon\|^2_{H^s(D)} \frac{H^{2(\min(s,N+1)-1)}}{N^{2(s-1)}}\right), \tag{1.79}$$

where C depends on the contrast of A_ε and on s the regularity of u. It is to be noted that C is independent of H, N and u.

Remark 1.41. *The result is similar to THEOREM 1.34. However, the convergence rate is explicit with respect to N which is the minimum degree of enrichments and H the size of the coarse mesh. Although there are no particular assumptions on the regularity of the*

coefficient A_ε or any geometry assumption (such as periodicity, ...)in order to get this result, the solution has to be continuous on the edges that is $H^s(D)$ with $1 < s$. The assumption $3/2 < s$ is made to encompass the case of triangles and quadrangles though it can be relaxed to $s > 1$ in the case of quadrangles. Such regularity is usually achieved when $A_\varepsilon \in L^\infty(D)$ and D is smooth enough, usually a convex domain or a polygonal domain without inward cusps.

Corollary 1.42. *Under the same assumptions as* THEOREM 1.40 *and asssuming that A_ε is $\mathbb{Z}^d\varepsilon$-periodic, $s = 2$, $N > 2$ and $H \simeq \varepsilon$ then we have*

$$\|u_\varepsilon - u_{H,N}\|^2_{H^1(D)} \leq C(H^2 + \frac{1}{N^2}), \tag{1.80}$$

with C that depends on the contrast of A_ε and of the shape of the elements.

Remark 1.43. *As for the special element method, one can adjust the number of enrichments in order to reduce the resonance error. Compared to the result of* COROLLARY 1.35*, the rate with respect to the number of enrichments is better. Numerical experiments on periodic cases show that our method performs better than the special element methods.*

Numerical tests show that this method is efficient and effectively cancels the resonance error provided that the degree N of enrichments is high enough (i.e. of the order of $1/\varepsilon$). An a posteriori estimator has been designed though it concerns only the global degree and cannot be used to refine the degree (number of enrichments) locally edge by edge.

The method performs well on classical periodic examples as well as cases when A_ε is not periodic.

This method is local and conformal. All enrichments and basis functions are computed locally element by element in contrast to MsFEM oversampling method. However, the coarse system to solve is larger as we have added enrichments. For instance, we consider a degree N and a coarse mesh of $Q = (0,1)^2$ with quadrangular elements of size H. The TABLE 1.1 describes the differences in term of degree of freedom and sparsity of the coarse system where MsFEM-lin corresponds to the linear MsFEM approach and MsFEM-N corresponds to our approach with enrichments of degree N.

	MsFEM-lin	MsFEM-N
Dof	$1/H^2$	$(1+2N)/H^2$
Matrix coeff	$1/H^4$	$4N^2/H^4$
Non-zeros	$9/H^2$	$7N^2/H^2$
Ratio	$9H^2$	$7/4H^2$

Table 1.1: Number of DOF, non-zeros coefficients in the coarse system and its ratio compared to the size of the system for MsFEM-lin and MsFEM-N methods

When we compare oversampling MsFEM or linear MsFEM to our method MsFEM-N, the number of degrees of freedom is multiplied by roughly $1 + 2N$. Such increase slows down the online phase as a bigger linear system has to be solved. However, we recall that H is supposed to be coarse so even with such a ratio the online phase remains fast.

We consider now the sparsity of the coarse linear system associated with the online phase. In term of memory cost the ratio of non-null coefficient over the size of

the matrix is better for the online coarse linear system associated with our enriched method than for the one with standard linear MsFEM approach. Also, during numerical experiment we did not notice big disparities between the time of resolution for our coarse systems and its equivalent in terms of FE element coarse system.

Remark 1.44. *All these results on sparsity and number of degrees of freedom is also true for the Special finite element method. Indeed, the size of the coarse linear system to solve during the online phase is the same as our method when the same number of enrichment is used (a degree N corresponds to $N-1$ eigenvector enrichment) since the support of the enrichment functions associated with one edge are the same.*

1.5 Perspectives

In this thesis, the work was divided into three parts: the study of the fluctuations in the context of stochastic homogenization (see CHAPTER 2), the design of an enriched multi-scale numerical approach (see CHAPTER 3) and the implementation of multi-scale methods in a Finite Element software (see CHAPTER 4).

1.5.1 Stochastic homogenization

Following CHAPTER 2, a theoretical study of the fluctuations in the weakly stochastic context allowed us to show that in this case the law of a family of quantities of interest can be inferred by knowing the structure of the random coefficient A_ε. Indeed, for such quantities of interest the law can be characterized simply in function of a fourth-order tensor $\mathcal{Q}$. This abstract object has no analytical formula in general. That is why, we designed and studied a numerical approach to approximate $\mathcal{Q}$. Numerical experiments showed that our approach yields an accurate approximation of $\mathcal{Q}$, even in non-weakly stochastic cases.

One could consider studying the fluctuations in another weakly stochastic framework. We considered a periodic coefficient with random defects with high probability of occurrence but very small effect. We could also study a periodic coefficient with defects which have a significant impact but a small probability to occur:

$$A_\varepsilon(x,\omega) = A_{per}(\frac{x}{\varepsilon}) + \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(\frac{x}{\varepsilon}) X_k(\omega), \tag{1.81}$$

with X_k i.i.d. variables such that $P(X_k = 0) = 1-\eta$ and $P(X_k = M) = \eta$ with M of order 1 and $\eta \ll 1$.

The defect would be considered as a rare event. Industry-wise, this case would be interesting as it would allow us to improve risk assessment in the case of rare failures in the material design.

Another question of interest would be to determine how to compute the mean of u_ε, Note indeed that, in (1.24), we substract $\mathbb{E}[u_\varepsilon]$, an object that is difficult to compute. It would be useful to design a quantity of interest depending on u_ε and $u_\star$ rather than u_ε and $\mathbb{E}[u_\varepsilon]$. We refer to REMARK 2.1 in CHAPTER 2 for a more detailed discussion in that direction.

Also, the heart of our numerical approach is to approximate $\mathcal{Q}$ by solving PDE on large domains with periodic conditions for a number of realizations in parallel and then to compute an empirical covariance. We studied the behavior of our approximation with respect to the size of the domain with quantitative convergence estimates. However, numerical experiments seem to indicate that a very large number of realizations is needed in order to get accurate results. Hence, there is a need to find better ways to compute the covariance. To that end, one could think of variance reduction methods such a control variate or importance sampling. The ideas developed in [14] and [65] to reduce the variance when approximating $A^\star$ could perhaps be used here.

1.5.2 MsFEM enriched method

In CHAPTER 3, we designed an enriched MsFEM method that cancels the resonance error when enough enrichments are added. The implementation of the method is straightforward, the enrichments can be computed independently element by element in parallel. It is an improvement compared to spectral type methods as the enrichments are neither solution to coupled problems on two elements nor solution to eigenproblems but solutions to elliptic problems with Dirichlet boundary conditions reducing significantly the computation time. The method is also conformal and does not suffer from reconstruction in order to get fluxes or stresses. The error estimates neither depend on the regularity of A_ε nor on its structure (no periodicity or stationarity required) provided the solution u_ε is continuous on the domain D.

The proofs of convergence rely heavily on the symmetry of the problem as energy minimizing arguments are used to get error estimates. In addition, the orthogonal decomposition (1.60) is pivotal in this approach. As linear MsFEM gets error estimates even in the non-symmetric case and in the regime $\varepsilon \ll H$, though for A_ε that are periodic, one could try to adapt the proofs to our approach. The advantage would be two-fold: show that the lack of symmetry does not impair the effectiveness of the method; and provide error estimates when $H \gg \varepsilon$, the current behavior is not known when $\varepsilon \to 0$ and H is fixed. It only describes the behavior best in the resonance regime. It is obvious to transpose the proof from the linear MsFEM error estimate to our case. The main idea is to compare the approximation to the homogenized solution. However, at some point a crude triangle inequality is used and one term does not depend on our approach but rather on a periodic homogenization result. Finally, in the regime $\varepsilon \to 0$ and when H is fixed we get the same estimate as the linear MsFEM in the periodic case (see THEOREM 1.20).

Another limitation is that the method can only be used for 2D cases as we define our enrichments on edges. However, it could be possible to design a similar method in 3D if we consider multiple type of enrichments: edges and faces interface functions. If the mesh is regular (for instance cube shaped), the trace and polynomials results would still apply. However, computationally speaking the method would be less interesting as the number of enrichments would increase (6 enrichments per degree per element instead of 4 in the 2D case) and the coarse linear system could reach quickly a critical size.

Finally, an a posteriori estimator has been designed, it allows us to refine the degree edge by edge with respect to the polynomial degree. This estimator, though useful, lacks accuracy in some regimes (for instance when high polynomial degrees are considered). This a posteriori estimator is based on a simple approach with residuals and finding a more relevant estimator would be possible using more sophisticated ap-

proaches. Getting an efficient a posteriori estimator is crucial as one drawback of our method is the associated increase of the size of the coarse linear system with respect to the polynomial degree used. If one chooses uniform refinement, the number of DOF of the coarse problem is increased by the number of edges times the polynomial degree considered. Hence, the online step can lead to solve large linear systems when high degree are used, especially compared to other methods such as MsFEM oversampling that gives similar errors with much smaller systems to solve. An effective a posteriori estimator would limit the increase in size of the coarse system while ensuring good accuracy.

1.5.3 MsFEM implementation

Following CHAPTER 4, multiple MsFEM variants have been implemented as templates in the Finite Element software FreeFem++. The MsFEM linear, oversampling and Crouzeix-Raviart methods are available. MsFEM approaches are intrusive, making them difficult to insert in a code. Also, with the numerous variants existing it would not be wise to implement it as a hard part of a code. However, some interesting implementation and theoretical projects regarding MsFEM can be explored.

Firstly, it would be interesting to combine MsFEM to Domain Decomposition Methods. If one aim at solving a multi-scale problem at the fine scale, usually direct methods to solve the system fail because the number of degrees of freedom is too large. Hence, using an iterative method is necessary. A critical step to get affordable time is to precondition the system in order to reduce the number of iterations. Domain decomposition methods such as the Schwarz method could be used as preconditioners. Sometimes such preconditioners are not sufficient and second-level preconditioners have to be used. Usually, the main preconditioner results from local independent computations to maximize parallelization and reducing execution time. However, it often lacks global information to link together the local computations that impairs the preconditioning effectiveness. Hence, introducing a coarse model (inexpensive to solve) as a second-level preconditioner could significantly improve the approach. In that sense MsFEM approximation are really suited to that role as it falls in a multi-query context: we can afford to have an offline phase as we need a quick online phase that will be repeated a lot of times. Numerical experiments (see CHAPTER 4) show that a second level preconditioning with Jacobi method as a fine preconditioner and linear MsFEM as coarse space gives good results: the number of steps needed by GMRES decreases sharply (factor 1 to 3). Also, the implementation of such a preconditioner is adaptable easily within the FreeFem++ framework (see CHAPTER 4). Some works have been initiated in that direction. One can see the works of Gander (see [40]) and Kornhuber (see [62]).

Finally, it would be also interesting to explore coupled formulations where multiple multi-scale approaches are used with respect to their specificity and the local change of the material that is in the function A_ε. For instance, when ε is very small, we are close to an homogenized regime so HMM method would be more appropriate. The MsFEM method is efficient when ε is small but not too small. Last, in the subdomains where A_ε is not oscillating too much using P1 FE would be sufficient. An attempt to couple P1 FE with MsFEM is presented in Chapter 4, though the numerical analysis has yet to be performed. Hence, locally choosing the method according to the oscillations of the coefficient would seem a good idea. However, with regard to proofs of convergence, since different arguments are used from one approach to an-

other, finding a general formulation that would encompass all these arguments could be a challenge.

CHAPTER 2

NUMERICAL APPROXIMATION OF FLUCTUATIONS IN STOCHASTIC HOMOGENIZATION

This chapter corresponds to a manuscript in preparation, co-authored with F. Legoll.

We study a method to approximate the fluctuations of the solution to an elliptic partial differential equation with highly oscillatory and random coefficients. Considering a weakly random setting (i.e. the case of periodic coefficients perturbed by a small random contribution), we show that the fluctuations of the solution are fully characterized by a fourth-order tensor, which is deterministic and independent of the right-hand side of the highly oscillatory problem. We also discuss how to practically approximate this tensor. We provide an extensive set of numerical experiments that illustrate our theoretical results, and also explore numerically the case of fully random (i.e. non weakly random) problems.

2.1 Introduction

We consider the problem

$$\begin{cases} -\operatorname{div}\left[A\left(\frac{\cdot}{\varepsilon},\omega\right)\nabla u_\varepsilon(\cdot,\omega)\right] = f \ \text{ in } D, \\ u_\varepsilon(\cdot,\omega) = 0 \ \text{ on } \partial D, \end{cases} \tag{2.1}$$

where $D \subset \mathbb{R}^d$ is a bounded domain and $f \in L^2(D)$. In this equation, the matrix-valued coefficient A is assumed to be bounded and bounded away from 0, random and stationary. For the sake of simplicity, we furthermore assume that A is symmetric. The limit behaviour, as ε goes to zero, of the solution u_ε to (2.1) is of major practical interest. It is described by homogenization theory (see e.g. the classical monographs [1, 58, 75] for some general exposition), that we now briefly recall.

Let

$$Q = \left(-\frac{1}{2},\frac{1}{2}\right)^d$$

and let w_p be the corrector function in the direction $p \in \mathbb{R}^d$, that is the solution (unique up to the addition of a random constant) to

$$\begin{cases} -\operatorname{div}[A(p+\nabla w_p)] = 0 \ \text{ in } \mathbb{R}^d, \\ \nabla w_p \text{ is stationary}, \qquad \mathbb{E}\left[\displaystyle\int_Q \nabla w_p\right] = 0, \end{cases} \tag{2.2}$$

where the notion of stationarity is defined by (2.5) below. The homogenization theory states that, as ε vanishes, u_ε approaches $u_\star$, the solution to

$$\begin{cases} -\operatorname{div}[A^\star \nabla u_\star] = f \ \text{ in } D, \\ u_\star = 0 \ \text{ on } \partial D, \end{cases} \tag{2.3}$$

where the homogenized matrix $A^\star$ is deterministic, constant and given by

$$\forall p \in \mathbb{R}^d, \quad A^\star p = \mathbb{E}\left[\int_Q A(p + \nabla w_p)\right]. \tag{2.4}$$

Since A is symmetric, $A^\star$ is also symmetric.

The notion of (discrete) stationarity employed in (2.2) is defined as follows. We consider the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We assume that the group $(\mathbb{Z}^d, +)$ acts on Ω, denote this action by $(\tau_k)_{k\in\mathbb{Z}^d}$, and assume that this action preserves the measure $\mathbb{P}$, in the sense that

$$\forall k \in \mathbb{Z}^d, \quad \forall \mathcal{A} \in \mathcal{F}, \qquad \mathbb{P}(\tau_k \mathcal{A}) = \mathbb{P}(\mathcal{A}),$$

and that it is ergodic: for any $\mathcal{A} \in \mathcal{F}$, we have

$$\Big[\forall k \in \mathbb{Z}^d, \ \ \tau_k \mathcal{A} = \mathcal{A}\Big] \Longrightarrow \mathbb{P}(\mathcal{A}) = 0 \text{ or } 1.$$

A function $F \in L^1_{\mathrm{loc}}(\mathbb{R}^d, L^1(\Omega))$ is said to be (discrete) stationary if

$$\forall k \in \mathbb{Z}^d, \quad F(x + k, \omega) = F(x, \tau_k \omega) \ \text{ a.e. in } x, \text{ almost surely.} \tag{2.5}$$

In practice, when $d \geq 2$, the solution w_p to (2.2) (and hence $A^\star$) cannot be computed. The homogenized matrix $A^\star$ is often approximated by $A^\star_N(\omega)$, defined by

$$\forall p \in \mathbb{R}^d, \quad A^\star_N(\omega) p = \frac{1}{|Q_N|} \int_{Q_N} A(\cdot, \omega)\big(p + \nabla w^N_p(\cdot, \omega)\big), \tag{2.6}$$

where w^N_p is the solution (unique up to the addition of a random constant) to the following random equation, posed on a *finite* domain:

$$\begin{cases} -\operatorname{div}\Big[A(\cdot, \omega)\big(p + \nabla w^N_p(\cdot, \omega)\big)\Big] = 0 \ \text{ in } Q_N, \\ w^N_p(\cdot, \omega) \text{ is } Q_N\text{-periodic}, \end{cases} \tag{2.7}$$

with, for instance,

$$Q_N = \bigcup_{|k|_\infty \leq N} Q + k = \left(-N - \frac{1}{2}, N + \frac{1}{2}\right)^d \tag{2.8}$$

where we have set $|k|_\infty := \max\limits_{1\leq i\leq d} |k_i|$ for any $k \in \mathbb{Z}^d$.

Besides the averaged behavior of u_ε on large space scales (which is given by $u_\star$ solution to (2.2)–(2.3)–(2.4)), a question of interest is to understand how much u_ε fluctuates around its expectation $\mathbb{E}[u_\varepsilon]$. This question has been comprehensively studied (see [7, 59]) for the problem

$$-\Delta u_\varepsilon + q_\varepsilon(x, \omega) u_\varepsilon = f \text{ in } \Omega, \quad u_\varepsilon = 0 \text{ on } \partial\Omega.$$

For the equation of interest here, namely (2.1), the question has been studied in the one-dimensional case (see [8, 19]). We also wish to cite the recent, theoretically oriented contribution [31], addressing (2.1) in the case when the differential operators are discrete (the differential operators in (2.1) are replaced by finite differences) and the equation is posed on the whole space. It is shown in [31] that the following properties hold true (for notational simplicity, we again use here the symbol ∇ for finite difference). For any $1 \leq i, j \leq d$, denote by

$$\rho_{i,j}(x,\omega) = (e_i + \nabla w_i) \cdot A(e_j + \nabla w_j) - e_j \cdot A^\star \nabla w_i - e_i \cdot A^\star \nabla w_j \tag{2.9}$$

the corrected energy function (we follow here the terminology of [31]), where w_i denotes the corrector, solution to (2.2) in the direction e_i, and $A^\star$ is the homogenized matrix (2.4). Define the fourth order tensor

$$\mathcal{Q} = \lim_{L \to +\infty} \mathcal{Q}^L \tag{2.10}$$

with, for any $1 \leq i, j, k, \ell \leq d$,

$$\mathcal{Q}^L_{i,j,k,\ell} = \mathrm{Cov}\left(\frac{1}{|Q_L|} \int_{Q_L} \rho_{i,j}, \int_{Q_L} \rho_{k,\ell} \right). \tag{2.11}$$

Given a right-hand side function f in (2.1) and a test function g, and assuming that they both are regular, compactly supported functions and that they both are the divergence of some vector field, the authors of [31] consider the quantity of interest

$$I_\varepsilon(f,g) = \varepsilon^{-d/2} \int_{\mathbb{R}^d} (u_\varepsilon(\cdot,\omega) - \mathbb{E}[u_\varepsilon])\, g, \tag{2.12}$$

where the integral over $\mathbb{R}^d$ means the sum over all lattice points (remember that, in [31], $u_\varepsilon(\cdot,\omega)$ is defined on a lattice). The quantity I_ε allows to understand the local fluctuations of u_ε. For instance, g can be the indicator function 1_K of a domain of interest K, and then $I_\varepsilon(f,g)$ measures the fluctuations of the average of u_ε in K around its mean. Note that, when ε is small, the fluctuations of u_ε are small, since u_ε is close to its deterministic limit $u_\star$. This motivates the rescaling factor $\varepsilon^{-d/2}$ in (2.12), in order for $I_\varepsilon(f,g)$ to converge to a non-trivial limit.

Remark 2.1. *Note that, in the definition of $I_\varepsilon(f,g)$, the integrand is $\varepsilon^{-d/2}(u_\varepsilon - \mathbb{E}[u_\varepsilon])$, which is different from $\varepsilon^{-d/2}(u_\varepsilon - u_\star)$, a quantity which is easier to compute. For $d > 1$, it turns out that $\varepsilon^{-d/2}(\mathbb{E}[u_\varepsilon] - u_\star)$ does not converge to 0. Considering for instance the periodic case (see e.g. [2]), it holds that $u_\varepsilon - u_\star$ is of order ε for any dimension d. In the random case, it is thus expected that $\varepsilon^{-d/2}(\mathbb{E}[u_\varepsilon] - u_\star)$ converges to a non-trivial limit for $d = 2$ and diverges for $d > 2$. It is thus not interesting to consider the quantity of interest $\varepsilon^{-d/2} \displaystyle\int_{\mathbb{R}^d} (u_\varepsilon(\cdot,\omega) - u_\star)\, g$. However, it might be possible to consider the quantity $J_\varepsilon(f,g) := \varepsilon^{-d/2} \displaystyle\int_{\mathbb{R}^d} (u_\varepsilon(\cdot,\omega) - \mathbb{E}[u_{\varepsilon,1}])\, g$ for a well-chosen random function $u_{\varepsilon,1}$, such that $\mathbb{E}[u_{\varepsilon,1}]$ is easier to compute than $\mathbb{E}[u_\varepsilon]$ and such that the laws of $I_\varepsilon(f,g)$ and $J_\varepsilon(f,g)$ converge to the same limit when $\varepsilon \to 0$. We refer e.g. to [9, 38] and do not pursue in that direction.*

It is shown in [31] that the random variable I_ε converges in law to a Gaussian random variable:

$$I_\varepsilon(f,g) \xrightarrow[\varepsilon \to 0]{\mathcal{L}} \mathcal{N}(0, \sigma^2), \tag{2.13}$$

where $\mathcal{N}(0,\sigma^2)$ is a Gaussian random variable with zero mean and variance given by

$$\sigma^2 = \int_{\mathbb{R}^d} (\nabla u_\star \otimes \nabla v_\star) : \mathcal{Q} : (\nabla u_\star \otimes \nabla v_\star). \tag{2.14}$$

In the above formula, $\mathcal{Q}$ is the fourth order tensor defined by (2.10) and we have denoted by $u_\star$ and $v_\star$ the solutions to the homogenized equation with right-hand sides f and g, respectively.

The practical interest of (2.13) is that the tensor $\mathcal{Q}$ is independent from f and g. Once evaluated, the computation of σ defined by (2.14) is inexpensive, since it only involves solving homogenized problems.

Somewhat schematically stated, the conclusion of the above mentioned contributions (namely [7, 59, 8, 19, 31]) is therefore that the fluctuations of the solution can essentially be determined independently from the right-hand side f, in some appropriate regime at least. Our aim is to elaborate on all these theoretical contributions to build an efficient numerical strategy.

The purpose of this article is threefold.

First, we prove (this is the content of Sections 2.3 and 2.4) that essentially (and in a sense to be made precise below) the theoretical results obtained in [31] for *discrete differential operators* carry over to the case of *continuous differential operators*, at least in the setting of weakly random problems.

The latter notion has been introduced in [14, 26, 66] and is recalled in Section 2.2. In short, the weakly random case consists in assuming that the random matrix A is the sum of a periodic coefficient with a small random perturbation:

$$A(x,\omega) = A_{\rm per}(x) + \eta A_1(x,\omega) \tag{2.15}$$

where $A_{\rm per}$ is a $\mathbb{Z}^d$-periodic matrix, bounded from above and bounded away from 0, and where A_1 a stationary bounded matrix. The parameter η is assumed to be small: $\eta \ll 1$. We also assume $A_{\rm per}$ and A_1 to be symmetric.

In this particular setting made precise in Section 2.2, we prove our first main result, stated in Theorem 2.7 below. It extends to our setting the results of [31], at least when we truncate our quantities of interest to the first order in the formal expansion in η. The proof of Theorem 2.7 is performed in Sections 2.3 and 2.4.

The second purpose of this article stems from the fact, already pointed out above, that one cannot access in practice to the exact correctors w_p solutions to (2.2). It is thus not possible to compute the tensor $\mathcal{Q}^L$ defined by (2.9) and (2.11). Using the truncated correctors w_p^N solutions to (2.7), it is natural to introduce an approximation $\mathcal{Q}^{L,N}$ of $\mathcal{Q}^L$ (see (2.42) below). In the weakly random case, we then show that this approximation is consistent, in the following sense. At the leading order in η, $\mathcal{Q}^{L,N} \approx \eta^2 \mathcal{Q}^{L,N,1}$ and $\mathcal{Q}^L \approx \eta^2 \mathcal{Q}^{L,1}$. We then show (see Theorem 2.10 below) that $\mathcal{Q}^{L,N,1}$ converges to $\mathcal{Q}^1$ when $L \to \infty$ and N satisfies $N > L$. The proof of Theorem 2.10 is given in Section 2.5.

The third purpose of this article (which is achieved in Section 2.6) is to provide an extensive set of numerical experiments that suggest that, for *fully random continuous differential operators*, the conclusions of Theorems 2.7 and 2.10 are again true. Although, to the best of our knowledge, there is no theoretical proof of this, and although we have been unable so far to extend our own proof of the weakly random case to the fully random case, we believe that the numerical tests presented in Section 2.6 are a strong indication toward the fact the formula holds with a large degree of generality. We hope that our observations will motivate further research in this direction.

2.2 Weakly random case and main results

We consider problem (2.1) with a coefficient A of the form (2.15). For simplicity, we choose

$$\begin{cases} A_1(x,\omega) = \displaystyle\sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)\, X_k(\omega)\ \mathrm{Id}_d, \\ X_k \text{ are i.i.d., almost surely bounded random variables,} \\ \mathbb{E}[X_0] = 0, \end{cases} \tag{2.16}$$

where Id_d is the d-dimensional identity matrix and where we recall that

$$Q = \left(-\frac{1}{2}, \frac{1}{2}\right)^d.$$

Remark 2.2. *Note that the assumption $\mathbb{E}[X_0] = 0$ is made without any loss of generality. Indeed, should $\mathbb{E}[X_0]$ be different from 0, it is always possible to write (2.15)–(2.16) in the form*

$$A(x,\omega) = A_{\mathrm{per}}(x) + \eta\,\mathbb{E}[X_0] + \eta \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)\,\big(X_k(\omega) - \mathbb{E}[X_0]\big)\ \mathrm{Id}_d,$$

where the random variables $X_k(\omega) - \mathbb{E}[X_0]$ have a vanishing expectation and where $A_{\mathrm{per}} + \eta\,\mathbb{E}[X_0]$ is a $\mathbb{Z}^d$-periodic matrix, which is bounded from above and bounded away from 0 when η is sufficiently small.

Remark 2.3. *Other cases alternative to (2.16) could be considered, such as for example*

$$A_1(x,\omega) = \sum_{k\in\mathbb{Z}^d} \varphi_{\mathrm{per}}(x)\, \mathbb{1}_{Q+k}(x)\, X_k(\omega)\ \mathrm{Id}_d$$

for some $\mathbb{Z}^d$-periodic function φ_{per}. For the sake of simplicity, we do not pursue in that direction.

As shown in [14, 66], an expansion of the oscillatory solution, the corrector, the homogenized matrix and the homogenized solution in power of η can be obtained, respectively. More precisely, we have

$$\begin{aligned} u_\varepsilon &= u_\varepsilon^0 + \eta\, u_\varepsilon^1 + O(\eta^2), \\ \nabla w_p &= \nabla w_p^0 + \eta\, \nabla w_p^1 + O(\eta^2), \\ A^\star &= A^\star_{\mathrm{per}} + O(\eta^2), \\ u_\star &= u_\star^0 + O(\eta^2). \end{aligned} \tag{2.17}$$

In the above expansion, u_ε^0 corresponds to the solution to (2.1) for the deterministic periodic coefficient $A_{\mathrm{per}}(\cdot/\varepsilon)$, that is

$$-\operatorname{div}\left[A_{\mathrm{per}}\left(\frac{\cdot}{\varepsilon}\right)\nabla u_\varepsilon^0\right] = f \ \text{ in } D, \qquad u_\varepsilon^0 = 0 \ \text{ on } \partial D, \tag{2.18}$$

while u_ε^1 is the term of first order in the expansion (in powers of η) of u_ε, namely the solution to

$$\begin{cases} -\operatorname{div}\left[A_{\mathrm{per}}\left(\frac{\cdot}{\varepsilon}\right)\nabla u_\varepsilon^1(\cdot,\omega)\right] = \operatorname{div}\left[A_1\left(\frac{\cdot}{\varepsilon},\omega\right)\nabla u_\varepsilon^0\right] \ \text{ in } D, \\ u_\varepsilon^1(\cdot,\omega) = 0 \ \text{ on } \partial D. \end{cases} \tag{2.19}$$

The function w_p^0 is the corrector in the direction p solution to (2.2) for the deterministic periodic coefficient $A_{\rm per}$, namely a solution to

$$\begin{cases} -\operatorname{div}[A_{\rm per}(p+\nabla w_p^0)] = 0 \ \text{ in } \mathbb{R}^d, \\ w_p^0 \text{ is } \mathbb{Z}^d\text{-periodic}, \end{cases} \tag{2.20}$$

while w_p^1 is the solution to

$$\begin{cases} -\operatorname{div}[A_{\rm per}\nabla w_p^1] = \operatorname{div}[A_1(p+\nabla w_p^0)] \ \text{ in } \mathbb{R}^d, \\ \nabla w_p^1 \text{ is stationary}, \qquad \mathbb{E}\left[\displaystyle\int_Q \nabla w_p^1\right] = 0. \end{cases} \tag{2.21}$$

Lastly, $A^\star_{\rm per}$ and $u^0_\star$ are the effective coefficient (2.4) and the corresponding homogenized limit (2.3) for the deterministic periodic coefficient $A_{\rm per}$, respectively.

Remark 2.4. *The topology in which the expansions (2.17) hold can be made more precise: we have*

$$\left\| u_\varepsilon(\cdot,\omega) - \left[u_\varepsilon^0 + \eta\, u_\varepsilon^1(\cdot,\omega)\right]\right\|_{H^1(D)} \le C\eta^2 \quad \text{a.s.},$$
$$\sqrt{\mathbb{E}\int_Q \left|\nabla w_p - \nabla(w_p^0 + \eta\, w_p^1)\right|^2} \le C\eta^2,$$
$$\left\|u_\star - u_\star^0\right\|_{H^1(D)} \le C\eta^2,$$

for some C independent of η, ε and ω.

Remark 2.5. *Note that, in (2.17), the first-order term in the expansion of $A^\star$ vanishes. In full generality, we have, for any $p \in \mathbb{R}^d$,*

$$A^\star p = A^\star_{\rm per} p + \eta\mathbb{E}\left[\int_Q A_1(p+\nabla w_p^0)\right] + \eta\mathbb{E}\left[\int_Q A_{\rm per}\nabla w_p^1\right] + O(\eta^2).$$

In our setting, we now recall that the expectation of A_1 vanishes, and hence that of w_p^1 also, in view of (2.21). The last two terms in the above equation thus vanish.

2.2.1 First main result

In the spirit of the quantity (2.12) considered in [31], we consider here the quantity of interest

$$I_\varepsilon(f,g) = \varepsilon^{-d/2}\int_D (u_\varepsilon(\cdot,\omega) - \mathbb{E}[u_\varepsilon])\, g \tag{2.22}$$

where $g \in L^2(D)$ and u_ε is the solution to problem (2.1) with right-hand side $f \in L^2(D)$ (we recall that the differential operators in (2.1) are *continuous* differential operators, in contrast to those in [31]; note also that the integral in (2.22) is a true integral and not a discrete sum as in (2.12)).

Using (2.17), we can expand (2.22) in powers of η, and find that

$$I_\varepsilon(f,g) = \eta\, I_\varepsilon^1(f,g) + \varepsilon^{-d/2}\,\eta^2\, C_\varepsilon^\eta(\omega) \tag{2.23}$$

with $|C_\varepsilon^\eta(\omega)| \le C\eta^2$ almost surely (for some C independent of η, ε and ω), and where

$$I_\varepsilon^1(f,g) = \varepsilon^{-d/2}\int_D u_\varepsilon^1\, g. \tag{2.24}$$

Indeed, we compute that

$$I_\varepsilon(f,g) = \varepsilon^{-d/2}\left(\eta \int_D (u_\varepsilon^1 - \mathbb{E}[u_\varepsilon^1])\, g + O(\eta^2)\right)$$

and we observe that $\mathbb{E}[u_\varepsilon^1] = 0$ in view of (2.19) and (2.16).

As in [31], we introduce the corrected energy function $\rho_{i,j}$ defined by (2.9) (where now all the differential operators are *continuous* differential operators) and the fourth order tensor $\mathcal{Q}^L_{i,j,k,\ell}$ defined by (2.11):

$$\mathcal{Q}^L_{i,j,k,\ell} = \mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L}\rho_{i,j}, \int_{Q_L}\rho_{k,\ell}\right). \tag{2.25}$$

This quantity can be expanded in a series in powers of η, as shown below.

Lemma 2.6. *Assume that (2.15) and (2.16) hold. Then the fourth order tensor (2.25) satisfies*

$$\left|\mathcal{Q}^L - \eta^2\,\mathcal{Q}^{L,1}\right| \le C|Q_L|\eta^3, \tag{2.26}$$

for some fourth order tensor $\mathcal{Q}^{L,1}$ (the expression of which is given by (2.29) below) and where C is independent of η and L.

Proof of Lemma 2.6. Using the definition (2.9) of $\rho_{i,j}$ and the expansions (2.15) and (2.17), we expand the corrected energy function as

$$\rho_{i,j} = \rho^{\mathrm{per}}_{i,j} + \eta\,\rho^1_{i,j} + O(\eta^2)$$

with

$$\rho^{\mathrm{per}}_{i,j} = (e_i + \nabla w_i^0)\cdot A_{\mathrm{per}}(e_j + \nabla w_j^0) - e_j\cdot A^\star_{\mathrm{per}}\nabla w_i^0 - e_i\cdot A^\star_{\mathrm{per}}\nabla w_j^0$$

and

$$\begin{aligned}\rho^1_{i,j} = {}& \nabla w_i^1\cdot A_{\mathrm{per}}(e_j+\nabla w_j^0) + (e_i+\nabla w_i^0)\cdot A_{\mathrm{per}}\nabla w_j^1\\ &+ (e_i+\nabla w_i^0)\cdot A_1(e_j+\nabla w_j^0) - e_j\cdot A^\star_{\mathrm{per}}\nabla w_i^1 - e_i\cdot A^\star_{\mathrm{per}}\nabla w_j^1.\end{aligned} \tag{2.27}$$

More precisely, we have

$$\sqrt{\mathbb{E}\int_Q \left|\rho_{i,j} - (\rho^{\mathrm{per}}_{i,j} + \eta\,\rho^1_{i,j})\right|^2} \le C\eta^2, \tag{2.28}$$

for some C independent of η.

We note that $\rho^{\mathrm{per}}_{i,j}$ is deterministic and that the expectation of $\rho^1_{i,j}$ vanishes, because $\mathbb{E}[A_1] = 0$. We define

$$\mathcal{Q}^{L,1}_{i,j,k,\ell} = \mathbb{E}\left(\frac{1}{|Q_L|}\int_{Q_L}\rho^1_{i,j}\int_{Q_L}\rho^1_{k,\ell}\right). \tag{2.29}$$

We introduce $r_{i,j} = \rho_{i,j} - (\rho^{per}_{i,j}) - \eta\rho^1_{i,j}$ and using (2.25) it holds that

$$\begin{aligned}|\mathcal{Q}^L_{i,j,k,\ell} - \eta^2\mathcal{Q}^{L,1}_{i,j,k,\ell}| = {}& \mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L}\eta\rho^1_{i,j}\,,\int_{Q_L} r_{k,l}\right)\\ &+ \mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L} r_{i,j}\,,\,\eta\int_{Q_L}\rho^1_{k,l}\right) + \mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L} r_{i,j}\,,\,\int_{Q_L} r_{k,l}\right)\end{aligned} \tag{2.30}$$

We have two crossed terms with r and ρ and one term where only r appears.

By using definition of the covariance, the Cauchy-Scwharz inequality and (2.28), it holds that

$$\begin{aligned}\mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L} r_{i,j}\,,\,\eta\int_{Q_L}\rho^1_{k,l}\right) &= \frac{1}{|Q_L|}\mathbb{E}\left[\left(\int_{Q_L} r_{i,j}-\mathbb{E}[\int_{Q_L} r_{i,j}]\right)\left(\int_{Q_L}\eta\rho^1_{k,l}\right)\right]\\ &\leq \frac{C_L\eta}{|Q_L|}\mathbb{E}\left[\int_Q r^2_{i,j}\right]\left[\int_Q(\rho^1_{k,l})^2\right]\\ &\leq C_L\eta^3, \end{aligned}\tag{2.31}$$

where C_L depends on L but not on η. The same estimate holds for the other crossed term.

We also have by similar arguments

$$\begin{aligned}\mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L} r_{i,j}\,,\,\int_{Q_L} r_{k,l}\right) &= \frac{1}{|Q_L|}\mathbb{E}\left[\left(\int_{Q_L} r_{i,j}-\mathbb{E}[\int_{Q_L} r_{i,j}]\right)\left(\int_{Q_L} r_{k,l}-\mathbb{E}[\int_{Q_L} r_{k,l}]\right)\right]\\ &\leq C_L\mathbb{E}\left[\int_Q r^2_{i,j}\right]\mathbb{E}\left[\int_Q r^2_{k,l}\right]\\ &\leq C_L\eta^4, \end{aligned}\tag{2.32}$$

where C_L depends on L but not on η. Then collecting (2.31) and (2.32) inserting it in (2.30) we get (2.26). This concludes the proof of Lemma 2.6. □

Inpired by [31], we expect that a result similar to (2.13) holds in our case. More precisely, we expect that, when ε tends to 0, the quantity $I_\varepsilon(f,g)$ defined by (2.22) converges to a Gaussian random variable,

$$I_\varepsilon(f,g)\xrightarrow[\varepsilon\to 0]{\mathcal{L}}\mathcal{N}(0,\sigma^2),\tag{2.33}$$

of mean zero and of variance

$$\sigma^2=\lim_{L\to\infty}(\sigma^L)^2\tag{2.34}$$

with

$$(\sigma^L)^2:=\int_D(\nabla u_\star\otimes\nabla v_\star)\colon\mathcal{Q}^L\colon(\nabla u_\star\otimes\nabla v_\star),\tag{2.35}$$

where $\mathcal{Q}^L$ is given by (2.25) and where $u_\star$ and $v_\star$ are the solutions to (2.3) with the right-hand sides f and g, respectively.

We have not been able to show (2.33) in a general random setting. However, we have shown this result in the case of *weakly random* problems, as stated in the following theorem.

Theorem 2.7. *Assume that (2.15) and (2.16) hold, that A_{per} is an Hölder continuous function, and that f and g are Hölder continuous. Let $I^1_\varepsilon(f,g)$, defined by (2.23), be the truncation at the first order (in the expansion in powers of η) of the quantity of interest $I_\varepsilon(f,g)$ defined by (2.22). When ε goes to 0, we have*

$$I^1_\varepsilon(f,g)\xrightarrow[\varepsilon\to 0]{\mathcal{L}}\mathcal{N}(0,\sigma_1^2)$$

i.e. $I^1_\varepsilon(f,g)$ converges in law to a centered Gaussian random variable of variance

$$\sigma_1^2=\int_D(\nabla u^0_\star\otimes\nabla v^0_\star)\colon\mathcal{Q}^1\colon(\nabla u^0_\star\otimes\nabla v^0_\star),\tag{2.36}$$

where $u_\star^0$ (resp. $v_\star^0$) is the homogenized solution (see (2.3)) associated to the homogenized matrix $A_{\rm per}^\star$ of the periodic coefficient $A_{\rm per}$, with the right-hand side f (resp. g):

$$-\operatorname{div}[A^\star \nabla u_\star^0] = f \ \text{ in } D, \quad u_\star^0 = 0 \ \text{ on } \partial D, \tag{2.37}$$

and

$$-\operatorname{div}[A^\star \nabla v_\star^0] = g \ \text{ in } D, \quad v_\star^0 = 0 \ \text{ on } \partial D. \tag{2.38}$$

In (2.36), the fourth order tensor $\mathcal{Q}^1$ is given by

$$\mathcal{Q}^1_{i,j,k,\ell} = \mathrm{Var}(X_0) \left(\int_Q (e_i + \nabla w_i^0) \cdot (e_j + \nabla w_j^0) \right) \left(\int_Q (e_k + \nabla w_k^0) \cdot (e_\ell + \nabla w_\ell^0) \right). \tag{2.39}$$

In addition, denoting $\mathcal{Q}^{L,1}$ defined by (2.29) the truncation at second order (in the expansion in powers of η) of $\mathcal{Q}^L$ (see (2.26)), we have that

$$\mathcal{Q}^1 = \lim_{L\to\infty} \mathcal{Q}^{L,1}. \tag{2.40}$$

This result shows that, at least when one truncates the formal expansions in η at the first order, the fluctuations of $I_\varepsilon(f,g)$ are governed by the fourth order tensor $\mathcal{Q}$. This thus generalizes the results of [31] for *continuous* differential operators, at the leading order in η.

The proof of Theorem 2.7 falls in two parts:

- in the first part (see Section 2.3), we consider the random variable $I_\varepsilon(f,g)$ (and more precisely its leading order term $I_\varepsilon^1(f,g)$) and establish that it converges in law to some centered Gaussian random variable (see Proposition 2.12). We also give an explicit expression for the variance of that Gaussian variable.

- in the second part (see Section 2.4), we consider the fourth order tensor $\mathcal{Q}$ (and more precisely its leading order $\mathcal{Q}^1$), and show that the variance identified in the first step can actually be expressed in terms of $\mathcal{Q}$, $u_\star^0$ and $v_\star^0$, thereby proving (2.36). This is the purpose of Proposition 2.20.

The proof of Theorem 2.7 is a direct consequence of Propositions 2.12 and 2.20.

Remark 2.8. *The Hölder regularity assumption on $A_{\rm per}$ is useful for two purposes. First, it implies some regularity on the corrector solution to (2.20), namely that $w_p^0 \in C^{1,\alpha}(Q)$ (see Lemma 2.13 below). Second, with $A_{\rm per}$ periodic and Hölder continuous, it is shown in [18] that the Green function of the operator $\mathcal{L} := -\operatorname{div}[A_{\rm per}\nabla\cdot]$ defined on the whole space $\mathbb{R}^d$ satisfies bounds similar to the Green function of the Laplace equation. These estimates are next useful in the proof of Proposition 2.20.*

2.2.2 Second main result

As pointed out in the introduction, we do not have access to the corrector function, solution to (2.2) on $\mathbb{R}^d$. In practice, we only consider the truncated corrector problem (2.7), posed on the bounded domain Q_N. We thus do not have access to $\rho_{i,j}$ and $\mathcal{Q}^L$, respectively defined by (2.9) and (2.11).

As for the approximation $A_N^\star(\omega)$ of $A^\star$ (see (2.6)), it is natural to introduce (similarly to (2.9)) the random function

$$\rho_{i,j}^N(x,\omega) = (e_i + \nabla w_i^N) \cdot A (e_j + \nabla w_j^N) - e_j \cdot A_N^\star(\omega) \nabla w_i^N - e_i \cdot A_N^\star(\omega) \nabla w_j^N \tag{2.41}$$

and the fourth order tensor $\mathcal{Q}^{L,N}$ defined by

$$\mathcal{Q}^{L,N}_{i,j,k,\ell} = \mathrm{Cov}\left(\frac{1}{|Q_L|}\int_{Q_L}\rho^N_{i,j}, \int_{Q_L}\rho^N_{k,\ell}\right). \tag{2.42}$$

We hope that $\mathcal{Q}^{L,N}$ converges to $\mathcal{Q}^L$ when $N \to \infty$. We are going to show such a result in a weakly stochastic setting, namely under assumptions (2.15) and (2.16).

We first expand in η the tensor $\mathcal{Q}^{L,N}$. We have the following result, similar to Lemma 2.6.

Lemma 2.9. *Assume that (2.15) and (2.16) hold. Then the fourth order tensor (2.42) satisfies*

$$\left|\mathcal{Q}^{L,N} - \eta^2\,\mathcal{Q}^{L,N,1}\right| \le C|Q_L|\eta^3, \tag{2.43}$$

where C is independent of η, L and N and where the fourth order tensor $\mathcal{Q}^{L,N,1}$ is given by

$$\mathcal{Q}^{L,N,1}_{i,j,k,\ell} = \mathbb{E}\left(\frac{1}{|Q_L|}\int_{Q_L}\rho^{N,1}_{i,j}\int_{Q_L}\rho^{N,1}_{k,\ell}\right) \tag{2.44}$$

with

$$\begin{aligned}\rho^{N,1}_{i,j} &= \nabla w^{N,1}_i \cdot A_{\mathrm{per}}(e_j+\nabla w^0_j) + (e_i+\nabla w^0_i)\cdot A_{\mathrm{per}}\nabla w^{N,1}_j \\ &\quad + (e_i+\nabla w^0_i)\cdot A_1(e_j+\nabla w^0_j) - e_j\cdot A^\star_{\mathrm{per}}\nabla w^{N,1}_i - e_i\cdot A^\star_{\mathrm{per}}\nabla w^{N,1}_j\end{aligned} \tag{2.45}$$

and

$$\begin{cases} -\operatorname{div}[A_{\mathrm{per}}\nabla w^{N,1}_p] = \operatorname{div}[A_1(p+\nabla w^0_p)] \ \text{ in } Q_N, \\ w^{N,1}_p \ \text{ is } Q_N\text{-periodic.}\end{cases} \tag{2.46}$$

In Section 2.5, we prove the following theorem.

Theorem 2.10. *Assume that (2.15) and (2.16) hold, and that A_{per} is an Hölder continuous function. Then, whenever $N > L$, we have*

$$\lim_{L\to\infty}\mathcal{Q}^{L,N,1} = \mathcal{Q}^1 \tag{2.47}$$

where $\mathcal{Q}^1$ is defined by (2.39). More precisely, we show that $\left|\mathcal{Q}^{L,N,1} - \mathcal{Q}^1\right| \le C\dfrac{(\ln L)^2}{L}$ for some C independent of N and L.

This theorem hence means that, at the leading order in η, the computable tensor $\mathcal{Q}^{L,N}$ indeed converges to $\mathcal{Q}$ when N and L tend to ∞ with $N > L$.

Remark 2.11. *In Theorem 2.10, we only consider the case $N > L$. Indeed, since $w^{N,1}_p$ is Q_N-periodic, we observe that*

$$\int_{Q_N}\rho^{N,1}_{i,j} = \int_{Q_N}(e_i+\nabla w^0_i)\cdot A_1(e_j+\nabla w^0_j) = \overline{\Lambda}_{ij}\sum_{|k|_\infty\le N} X_k(\omega)$$

with

$$\overline{\Lambda}_{ij} = \int_Q (e_i+\nabla w^0_i)\cdot(e_j+\nabla w^0_j).$$

Recalling that X_k are i.i.d. and centered random variables, we observe that, when $N = L$,

$$\mathcal{Q}^{N,N,1}_{i,j,k,\ell} = \frac{1}{|Q_N|}\overline{\Lambda}_{ij}\overline{\Lambda}_{k\ell}\mathbb{E}\left(\sum_{|k|_\infty\le N}X_k(\omega)\sum_{|p|_\infty\le N}X_p(\omega)\right) = \overline{\Lambda}_{ij}\overline{\Lambda}_{k\ell}\,\mathbb{E}(X_0^2).$$

The tensor $\mathcal{Q}^{N,N,1}$ hence happens to be independent of N, and equal (see Eq. (2.114), Lemmas 2.25 and 2.26 below) to the limit of $\mathcal{Q}^{L,N,1}$ when N and L go to ∞ with $N > L$. This equality of course strongly relies on the fact that we consider a weakly random case, and that $w_p^{N,1}$ is Q_N-periodic. This equality can hence not be expected in the full (namely, not weakly) random case: in general, the tensor $\mathcal{Q}^{N,N}$ depends on N.

2.3 Limit of (the leading order term of) $I_\varepsilon(f,g)$

In this section, we compute the asymptotic law of $I_\varepsilon^1(f,g)$ when $\varepsilon \to 0$, where we recall that $I_\varepsilon^1(f,g)$ is the leading order term in the expansion (2.23) in η of the quantity of interest $I_\varepsilon(f,g)$.

Proposition 2.12. *Assume that (2.15) and (2.16) hold, that $A_{\rm per}$ is an Hölder continuous function, and that f and g are Hölder continuous. Let $I_\varepsilon^1(f,g)$ be defined by (2.23)–(2.24). Then*

$$I_\varepsilon^1(f,g) \xrightarrow[\varepsilon\to 0]{\mathcal{L}} \mathcal{N}(0,\sigma_1^2) \tag{2.48}$$

where σ_1 is defined by (2.36), where $u_\star^0$, $v_\star^0$ and $\mathcal{Q}^1$ are respectively given by (2.37), (2.38) and (2.39).

The remainder of this section is devoted to proving Proposition 2.12.

2.3.1 Technical lemmas

We first collect here some useful technical results. The two following lemmas can e.g. be found in [43].

Lemma 2.13. *Assume that $A_{\rm per}$ is an Hölder continuous function. Then the periodic corrector w_p^0 solution to (2.20) belongs to $C^{1,\alpha}(Q)$ for some $\alpha > 0$.*

Lemma 2.14. *Consider $u_\star^0$ defined by (2.37), where we assume that the right-hand side f belongs to $C^{0,\alpha}(D)$. Then $u_\star^0$ belongs to $C^{2,\beta}(D)$ for some $\beta > 0$.*

Lemma 2.15 (Theorem 4 of [4])**.** *Assume that $A_{\rm per}$ is an Hölder continuous function and that there exists $q > d$ such that $g \in L^q(D)$. Let v_ε^0 be the solution to (2.1) with the periodic coefficient $A_{\rm per}(\cdot/\varepsilon)$ and the right-hand side g, namely the solution to*

$$-\operatorname{div}\left[A_{\rm per}\left(\frac{\cdot}{\varepsilon}\right)\nabla v_\varepsilon^0\right] = g \ \text{ in } D, \qquad v_\varepsilon^0 = 0 \ \text{ on } \partial D. \tag{2.49}$$

Then, there exists $C < \infty$ such that, for any ε, we have $\|\nabla v_\varepsilon^0\|_{L^\infty(D)} \le C$.

This result is also shown in [66, Eq. (32)].

Lemma 2.16. *Consider the solution u_ε^0 to (2.18). Let $u_\star^0$ be its homogenized limit, solution to (2.37). Introduce the two-scale expansion*

$$\widetilde{u}_\varepsilon^0 = u_\star^0 + \varepsilon \sum_{i=1}^d w_i^0\left(\frac{\cdot}{\varepsilon}\right)\frac{\partial u_\star^0}{\partial x_i}$$

of u_ε^0, and consider also the remainder

$$r_\varepsilon^u = \nabla u_\varepsilon^0 - \nabla u_\star^0 - \sum_{i=1}^d \nabla w_i^0\left(\frac{\cdot}{\varepsilon}\right)\frac{\partial u_\star^0}{\partial x_i}. \tag{2.50}$$

Assume that $A_{\rm per}$ and f are Hölder continuous functions.

Then, we have that $\|u^0_\varepsilon - \widetilde{u}^0_\varepsilon\|_{H^1(D)} \leq C\sqrt{\varepsilon}$ and

$$\|r^u_\varepsilon\|_{L^2(D)} \leq C\sqrt{\varepsilon} \tag{2.51}$$

for some C independent of ε.

Proof of Lemma 2.16. By a classical result of periodic homogenization (see for instance [11, Theorem 5.13 pp. 41-42] or [58, p. 28]), we have $\|u^0_\varepsilon - \widetilde{u}^0_\varepsilon\|_{H^1(D)} \leq C\sqrt{\varepsilon}$. We next note that $r^u_\varepsilon = \nabla(u^0_\varepsilon - \widetilde{u}^0_\varepsilon) + \varepsilon \displaystyle\sum_{i=1}^d w_i\left(\frac{\cdot}{\varepsilon}\right)\nabla\frac{\partial u^0_\star}{\partial x_i}$. We hence obtain

$$\|r^u_\varepsilon\|_{L^2(D)} \leq \|u^0_\varepsilon - \widetilde{u}^0_\varepsilon\|_{H^1(D)} + \varepsilon\|\nabla^2 u^0_\star\|_{L^2(D)} \sum_{i=1}^d \|w_i\|_{L^\infty(Q)}.$$

Using Lemmas 2.13 and 2.14, we deduce (2.51). □

Lemma 2.17. *Consider $u^0_\star$ and $v^0_\star$ defined by (2.37) and (2.38), where we assume that the right-hand sides f and g belong to $C^{0,\alpha}(D)$. We consider $\Lambda : y \mapsto \Lambda(y) \in \mathbb{R}^{d\times d}$ a $\mathbb{Z}^d$-periodic matrix, with $\Lambda \in L^\infty(Q)^{d\times d}$. Denote by*

$$S_\varepsilon = \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \left(\int_Q \nabla u^0_\star(\varepsilon(y+k))\cdot\Lambda(y+k)\,\nabla v^0_\star(\varepsilon(y+k))\,dy\right)^2. \tag{2.52}$$

We have

$$\lim_{\varepsilon\to 0} S_\varepsilon = \int_D \left(\nabla u^0_\star(x)\cdot\overline{\Lambda}\,\nabla v^0_\star(x)\right)^2 dx \tag{2.53}$$

with $\overline{\Lambda} = \displaystyle\int_Q \Lambda(y)\,dy$.

Proof. We have pointed out in Lemma 2.14 that $u^0_\star$ belongs to $C^{2,\beta}(D)$ for some $\beta > 0$, and likewise for $v^0_\star$. We can then write the following Taylor expansion, for any $y \in Q$ and any k such that $Q + k \subset D/\varepsilon$:

$$\nabla u^0_\star(\varepsilon(y+k)) = \nabla u^0_\star(\varepsilon k) + \varepsilon a^u_\varepsilon(y,k)$$

where

$$a^u_\varepsilon(y,k) = \int_0^1 \nabla^2(u^0_\star)(\varepsilon(ty+k)))\cdot y\,dt$$

and where $\nabla^2(u^0_\star)$ is the Hessian matrix of $u^0_\star$. Using the regularity of $u^0_\star$, we have

$$|a^u_\varepsilon(y,k)| \leq \|\nabla^2(u^0_\star)\|_{C^0(D)}\,|y| \leq C_d\|u^0_\star\|_{C^2(D)} \tag{2.54}$$

where C_d only depends on d. Let

$$J(k,\varepsilon) = \int_Q \nabla u^0_\star(\varepsilon(y+k))\cdot\Lambda(y+k)\,\nabla v^0_\star(\varepsilon(y+k))\,dy$$

and

$$J^0(k,\varepsilon) = \int_Q \nabla u^0_\star(\varepsilon k)\cdot\Lambda(y+k)\,\nabla v^0_\star(\varepsilon k)\,dy = \nabla u^0_\star(\varepsilon k)\cdot\overline{\Lambda}\,\nabla v^0_\star(\varepsilon k).$$

We first see that

$$|J(k,\varepsilon)| \le \|\nabla u_\star^0\|_{C^0(D)} \|\Lambda\|_{L^\infty(Q)} \|\nabla v_\star^0\|_{C^0(D)} \tag{2.55}$$

and likewise for $J^0(k,\varepsilon)$. By definition, we have $S_\varepsilon = \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \big(J(k,\varepsilon)\big)^2$. Let

$$T_\varepsilon = \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \big(J^0(k,\varepsilon)\big)^2.$$

We see that

$$S_\varepsilon - T_\varepsilon = \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \Big(J(k,\varepsilon) - J^0(k,\varepsilon)\Big)\Big(J(k,\varepsilon) + J^0(k,\varepsilon)\Big),$$

hence, using the bounds (2.55) on $J(k,\varepsilon)$ and $J^0(k,\varepsilon)$,

$$\begin{aligned}
&|S_\varepsilon - T_\varepsilon| \\
&\le 2\,\varepsilon^d \,\|\nabla u_\star^0\|_{C^0(D)}\, \|\Lambda\|_{L^\infty(Q)}\, \|\nabla v_\star^0\|_{C^0(D)} \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \Big|J(k,\varepsilon) - J^0(k,\varepsilon)\Big| \\
&\le C\varepsilon^{d+1} \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \int_Q \big|a_\varepsilon^u(y,k)\cdot\Lambda(y+k)\,\nabla v_\star^0(\varepsilon k)\big|\,dy \\
&\quad + C\varepsilon^{d+1} \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \int_Q \big|\nabla u_\star^0(\varepsilon k)\cdot\Lambda(y+k)\,a_\varepsilon^v(y,k)\big|\,dy \\
&\quad + C\varepsilon^{d+2} \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \int_Q |a_\varepsilon^u(y,k)\cdot\Lambda(y+k)\,a_\varepsilon^v(y,k)|\,dy.
\end{aligned}$$

Using the bound (2.54), we deduce that

$$|S_\varepsilon - T_\varepsilon| \le C\varepsilon^{d+1} \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \|u_\star^0\|_{C^2(D)} \|\Lambda\|_{L^\infty(Q)} \|v_\star^0\|_{C^2(D)} \le C\varepsilon. \tag{2.56}$$

We next observe that

$$\begin{aligned}
T_\varepsilon &= \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \big(J^0(k,\varepsilon)\big)^2 \\
&= \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \Big(\nabla u_\star^0(\varepsilon k)\cdot\overline{\Lambda}\,\nabla v_\star^0(\varepsilon k)\Big)^2.
\end{aligned}$$

We therefore note that T_ε is a Riemann sum. When ε goes to 0, it thus converges to $\int_D \big(\nabla u_\star^0 \cdot \overline{\Lambda}\,\nabla v_\star^0\big)^2$. Collecting this limit with (2.56), we obtain (2.53). This concludes the proof of Lemma 2.17. □

2.3.2 Proof of Proposition 2.12

We are now in position to prove Proposition 2.12. For conciseness, the dependence of $I_\varepsilon^1(f,g)$ upon f and g is not be made explicit here.

The leading order term I_ε^1, defined by (2.24), of the quantity of interest (2.22), satisfies (in view of the variational form associated to (2.49))

$$I_\varepsilon^1 = \varepsilon^{-d/2} \int_D u_\varepsilon^1\, g = \varepsilon^{-d/2} \int_D (\nabla u_\varepsilon^1)^T A_{\rm per}\left(\frac{\cdot}{\varepsilon}\right) \nabla v_\varepsilon^0.$$

Using the symmetry of $A_{\rm per}$, we obtain

$$I_\varepsilon^1 = \varepsilon^{-d/2} \int_D (\nabla v_\varepsilon^0)^T A_{\rm per}\left(\frac{\cdot}{\varepsilon}\right) \nabla u_\varepsilon^1.$$

Using the equation (2.19) satisfied by u_ε^1 and next (2.16), we obtain

$$I_\varepsilon^1 = -\varepsilon^{-d/2} \sum_{k\in\mathbb{Z}^d} X_k(\omega) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \nabla v_\varepsilon^0 \cdot \nabla u_\varepsilon^0,$$

where we note that only a finite number of indices k contribute to the above sum.

Using Lemma 2.16 and denoting by ∇W^0 the matrix given by $(\nabla W^0)_{i,j} = \dfrac{\partial w_j^0}{\partial x_i}$, we have

$$\nabla u_\varepsilon^0 = \left[\mathrm{Id}_d + \nabla W^0\left(\frac{\cdot}{\varepsilon}\right)\right] \nabla u_\star^0 + r_\varepsilon^u$$

and likewise for v_ε^0. It follows that I_ε^1 reads as

$$I_\varepsilon^1 = -(C_1^\varepsilon + C_u^\varepsilon + C_v^\varepsilon - C_r^\varepsilon), \tag{2.57}$$

with C_1^ε, C_u^ε, C_v^ε and C_r^ε respectively defined by

$$C_1^\varepsilon = \varepsilon^{-d/2} \sum_{k\in\mathbb{Z}^d} X_k(\omega) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \times \left[\left(\mathrm{Id}_d + \nabla W^0\left(\frac{\cdot}{\varepsilon}\right)\right) \nabla v_\star^0\right] \cdot \left[\left(\mathrm{Id}_d + \nabla W^0\left(\frac{\cdot}{\varepsilon}\right)\right) \nabla u_\star^0\right]$$

and

$$C_u^\varepsilon = \varepsilon^{-d/2} \sum_{k\in\mathbb{Z}^d} X_k(\omega) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) r_\varepsilon^u \cdot \nabla v_\varepsilon^0,$$

$$C_v^\varepsilon = \varepsilon^{-d/2} \sum_{k\in\mathbb{Z}^d} X_k(\omega) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) r_\varepsilon^v \cdot \nabla u_\varepsilon^0,$$

$$C_r^\varepsilon = \varepsilon^{-d/2} \sum_{k\in\mathbb{Z}^d} X_k(\omega) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) r_\varepsilon^v \cdot r_\varepsilon^u.$$

We successively study the limit when $\varepsilon \to 0$ of C_1^ε, C_u^ε, C_v^ε and C_r^ε.

Step 1: limit of C_u^ε and C_v^ε

We show that $\lim_{\varepsilon\to 0} \mathbb{E}\big[(C_u^\varepsilon)^2\big] = 0$. Since X_k are i.i.d. centered random variables, we compute that

$$\begin{aligned}\mathbb{E}\big[(C_u^\varepsilon)^2\big] &= \varepsilon^{-d} \sum_{k\in\mathbb{Z}^d} \mathrm{Var}(X_0) \left[\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) r_\varepsilon^u \cdot \nabla v_\varepsilon^0\right]^2 \\ &\le \varepsilon^{-d}\, \mathrm{Var}(X_0) \sum_{k\in\mathbb{Z}^d} \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \left|r_\varepsilon^u\right|^2 \left|\nabla v_\varepsilon^0\right|^2.\end{aligned}$$

Using Lemma 2.15 and the fact that $\displaystyle\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \le \varepsilon^d$, we get

$$\begin{aligned}\mathbb{E}\big[(C_u^\varepsilon)^2\big] &\le \mathrm{Var}(X_0)\, \|\nabla v_\varepsilon^0\|_{L^\infty(D)}^2 \sum_{k\in\mathbb{Z}^d} \int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) \left|r_\varepsilon^u\right|^2 \\ &= \mathrm{Var}(X_0)\, \|\nabla v_\varepsilon^0\|_{L^\infty(D)}^2 \|r_\varepsilon^u\|_{L^2(D)}^2.\end{aligned}$$

Using Lemma 2.16, we deduce that

$$\mathbb{E}\big[(C_u^\varepsilon)^2\big] \leq C\varepsilon. \tag{2.58}$$

Of course the same result holds for C_v^ε.

Step 2: limit of C_r^ε

By definition (see (2.50)), we have $r_\varepsilon^u = \nabla u_\varepsilon^0 - \nabla u_\star^0 - \sum_{i=1}^d \nabla w_i^0\left(\frac{\cdot}{\varepsilon}\right)\frac{\partial u_\star^0}{\partial x_i}$. In view of Lemmas 2.13 and 2.15, we know that $\|\nabla w_i^0\|_{L^\infty(\mathbb{R}^d)} \leq C$ and that $\|\nabla u_\varepsilon^0\|_{L^\infty(D)} \leq C$. Furthermore, since f is a Hölder continuous function, we know by elliptic regularity (see Lemma 2.14) that $u_\star^0$ belongs to $C^{2,\beta}(D)$ for some $\beta > 0$. This implies that $u_\star^0$ belongs to $W^{1,\infty}(D)$. We thus deduce that

$$\|r_\varepsilon^u\|_{L^\infty(D)} \leq C \tag{2.59}$$

for some C independent of ε.

We now compute $\mathbb{E}\big[(C_r^\varepsilon)^2\big]$. We have that

$$\begin{aligned}
\mathbb{E}\big[(C_r^\varepsilon)^2\big] &= \varepsilon^{-d}\sum_{k\in\mathbb{Z}^d}\operatorname{Var}(X_0)\left[\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right) r_\varepsilon^u \cdot r_\varepsilon^v\right]^2 \\
&\leq \varepsilon^{-d}\operatorname{Var}(X_0)\sum_{k\in\mathbb{Z}^d}\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right)\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right)\left|r_\varepsilon^u\right|^2\left|r_\varepsilon^v\right|^2 \\
&\leq \operatorname{Var}(X_0)\,\|r_\varepsilon^u\|_{L^\infty(D)}^2\sum_{k\in\mathbb{Z}^d}\int_D \mathbb{1}_{Q+k}\left(\frac{\cdot}{\varepsilon}\right)\left|r_\varepsilon^v\right|^2 \\
&= \operatorname{Var}(X_0)\,\|r_\varepsilon^u\|_{L^\infty(D)}^2\,\|r_\varepsilon^v\|_{L^2(D)}^2.
\end{aligned}$$

Using the bound (2.59) for $\|r_\varepsilon^u\|_{L^\infty(D)}$ and Lemma 2.16, from which we infer that $\|r_\varepsilon^v\|_{L^2(D)} \leq C\sqrt{\varepsilon}$, we get that

$$\mathbb{E}\big[(C_r^\varepsilon)^2\big] \leq C\varepsilon. \tag{2.60}$$

Step 3: limit of C_1^ε

We prove that C_1^ε converges in law to a Gaussian random variable. To that aim, we introduce

$$\Lambda(y) = \Big(\mathrm{Id}_d + \nabla W^0(y)\Big)^T\Big(\mathrm{Id}_d + \nabla W^0(y)\Big)$$

and

$$\psi(x,y) = (\nabla v_\star^0(x))^T\Lambda(y)\,\nabla u_\star^0(x), \tag{2.61}$$

so that C_1^ε can be recast as

$$C_1^\varepsilon = \varepsilon^{-d/2}\sum_{k\in\mathbb{Z}^d} X_k(\omega)\int_D \mathbb{1}_{Q+k}\left(\frac{x}{\varepsilon}\right)\psi\left(x,\frac{x}{\varepsilon}\right)dx.$$

We note that ψ is $\mathbb{Z}^d$-periodic with respect to its second variable. In view of Lemma 2.13, we have $\Lambda \in L^\infty(\mathbb{R}^d)$. As pointed out in Lemma 2.14, we have that $\nabla u_\star^0 \in L^\infty(D)$, and similarly for $\nabla v_\star^0$. The function ψ is thus uniformly bounded with respect to its two variables.

We introduce

$$Z_k^\varepsilon = \varepsilon^{-d/2} \int_D \mathbb{1}_{Q+k}\left(\frac{x}{\varepsilon}\right) \psi\left(x, \frac{x}{\varepsilon}\right) dx,$$

which also reads

$$Z_k^\varepsilon = \varepsilon^{d/2} \int_{(Q+k)\cap(D/\varepsilon)} \psi(\varepsilon y, y)\, dy.$$

Using the above bound on ψ, we have that, for any ε and k,

$$|Z_k^\varepsilon| \leq \varepsilon^{d/2} \|\psi\|_{L^\infty(D\times\mathbb{R}^d)}. \tag{2.62}$$

We also note that Z_k^ε does not vanish for a number of indices k which is bounded by $C\varepsilon^{-d}$.

The random variable C_1^ε then reads

$$C_1^\varepsilon = \sum_{k\in\mathbb{Z}^d} Z_k^\varepsilon\, X_k(\omega).$$

We show that it converges in law toward a Gaussian by showing the convergence of its characteristic function $\theta_{C_1^\varepsilon}$, defined for $\xi \in \mathbb{R}$ by

$$\theta_{C_1^\varepsilon}(\xi) = \mathbb{E}\Big[\exp(i\xi C_1^\varepsilon)\Big] = \mathbb{E}\left[\exp\left(i\xi\left(\sum_{k\in\mathbb{Z}^d} Z_k^\varepsilon\, X_k(\omega)\right)\right)\right].$$

Since the random variables X_k are i.i.d., we have

$$\theta_{C_1^\varepsilon}(\xi) = \prod_{k\in\mathbb{Z}^d} \mathbb{E}\Big[\exp(i\xi\, Z_k^\varepsilon\, X_k(\omega))\Big] = \prod_{k\in\mathbb{Z}^d} \theta_{X_0}(\xi\, Z_k^\varepsilon) \tag{2.63}$$

where $\theta_{X_0}(\zeta) = \mathbb{E}\Big[\exp(i\zeta X_0)\Big]$ is the characteristic function of $X_0(\omega)$. Since X_0 is a.s. bounded, θ_{X_0} is a smooth function. We observe that, for any fixed $\xi \in \mathbb{R}$, we have, in view of (2.62), that $|\xi\, Z_k^\varepsilon| \leq \varepsilon^{d/2}\, |\xi|\, \|\psi\|_{L^\infty(D\times\mathbb{R}^d)} \leq C_\xi \varepsilon^{d/2}$ for any $k \in \mathbb{Z}^d$ and any ε, where C_ξ is independent of k and ε. The quantities $\{\xi\, Z_k^\varepsilon\}_{k\in\mathbb{Z}^d}$ thus remain in a neighbourhood $\mathcal{V}_\xi$ of the origin, and the quantities $\{\theta_{X_0}(\xi\, Z_k^\varepsilon)\}_{k\in\mathbb{Z}^d}$ thus remain in a neighbourhood of $\theta_{X_0}(0) = 1$. We now introduce the real-valued functions a and b such that, for any $\zeta \in \mathbb{R}$, we have $\theta_{X_0}(\zeta) = \exp(a(\zeta) + ib(\zeta))$, with $b(\zeta) \in [-\pi, \pi)$. We next set $\phi(\zeta) = a(\zeta) + ib(\zeta)$. When restricting ourselves to $\zeta \in \mathcal{V}_\xi$, it is possible to define the function b (and thus the function ϕ) in a manner such that b (and thus ϕ) is smooth, similarly to θ_{X_0}.

By definition,

$$\theta_{X_0}(0) = 1, \quad \theta'_{X_0}(0) = \mathbb{E}[iX_0] = 0, \quad \theta''_{X_0}(0) = \mathbb{E}[-X_0^2] = -\mathrm{Var}(X_0),$$

and thus

$$\phi(0) = 0, \quad \phi'(0) = 0, \quad \phi''(0) = -\mathrm{Var}(X_0).$$

The Taylor expansion of ϕ thus reads

$$\phi(\zeta) = -\frac{\zeta^2}{2}\mathrm{Var}(X_0) + \int_0^\zeta \frac{(\zeta - t)^2}{2}\, \phi^{(3)}(t)\, dt.$$

We insert this expansion in (2.63):

$$\theta_{C_1^\varepsilon}(\xi) = \prod_{k\in\mathbb{Z}^d} \theta_{X_0}(\xi Z_k^\varepsilon) = \prod_{k\in\mathbb{Z}^d} \exp\Big[\phi(\xi Z_k^\varepsilon)\Big] = \phi_G^\varepsilon(\xi)\, \phi_r^\varepsilon(\xi) \tag{2.64}$$

with

$$\phi_G^\varepsilon(\xi) = \exp\left(-\frac{\mathrm{Var}(X_0)}{2}\,\xi^2 \sum_{k\in\mathbb{Z}^d} (Z_k^\varepsilon)^2\right), \tag{2.65}$$

$$\phi_r^\varepsilon(\xi) = \exp\left(\sum_{k\in\mathbb{Z}^d} \int_0^{\xi Z_k^\varepsilon} \frac{(\xi Z_k^\varepsilon - t)^2}{2}\,\phi^{(3)}(t)\,dt\right).$$

We successively study the two above quantities.

We first claim that, for any $\xi \in \mathbb{R}$,

$$\lim_{\varepsilon\to 0} \phi_r^\varepsilon(\xi) = 1. \tag{2.66}$$

We fix some $\xi \in \mathbb{R}$. In view of (2.62), for any $k \in \mathbb{Z}^d$ and any ε, we have $|\xi\, Z_k^\varepsilon| \le \varepsilon^{d/2}\,|\xi|\,\|\psi\|_{L^\infty(D\times\mathbb{R}^d)} \le C\varepsilon^{d/2}$. Since $\phi^{(3)}$ is bounded on the interval $[-C,C]$ which contains the interval $[0,\xi\, Z_k^\varepsilon]$, we obtain that there exists some C_ξ independent of ε and k such that

$$\left|\int_0^{\xi Z_k^\varepsilon} \frac{(\xi Z_k^\varepsilon - t)^2}{2}\,\phi^{(3)}(t)\,dt\right| \le C_\xi\,\varepsilon^{3d/2}.$$

In addition, as pointed out above, Z_k^ε does not vanish for a number of indices k which is bounded by $C\varepsilon^{-d}$. We thus get that

$$\left|\sum_{k\in\mathbb{Z}^d} \int_0^{\xi Z_k^\varepsilon} \frac{(\xi Z_k^\varepsilon - t)^2}{2}\,\phi^{(3)}(t)\,dt\right| \le \overline{C}_\xi\,\varepsilon^{d/2},$$

which implies (2.66).

Second, we turn to $\phi_G^\varepsilon(\xi)$. We note that

$$\sum_{k\in\mathbb{Z}^d} (Z_k^\varepsilon)^2 = \varepsilon^d \sum_{k\in\mathbb{Z}^d} \left(\int_{Q\cap(D/\varepsilon-k)} \psi\big(\varepsilon(y+k),y+k\big)\,dy\right)^2 = S_\varepsilon + R_\varepsilon \tag{2.67}$$

with

$$R_\varepsilon = \varepsilon^d \sum_{k \text{ s.t. } Q+k\not\subset D/\varepsilon} \left(\int_{Q\cap(D/\varepsilon-k)} \psi\big(\varepsilon(y+k),y+k\big)\,dy\right)^2$$

and

$$\begin{aligned} S_\varepsilon &= \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \left(\int_{Q\cap(D/\varepsilon-k)} \psi\big(\varepsilon(y+k),y+k\big)\,dy\right)^2 \\ &= \varepsilon^d \sum_{k \text{ s.t. } Q+k\subset D/\varepsilon} \left(\int_{Q} \psi\big(\varepsilon(y+k),y+k\big)\,dy\right)^2. \end{aligned}$$

In view of Lemma 2.17 and of the definition (2.61) of ψ, we have that

$$\lim_{\varepsilon\to 0} S_\varepsilon = \sigma_{\mathrm{per}}^2 \tag{2.68}$$

with

$$\sigma_{\mathrm{per}}^2 = \int_D \Big(\nabla u_\star^0(x)\cdot\overline{\Lambda}\,\nabla v_\star^0(x)\Big)^2\,dx \tag{2.69}$$

and where the $d \times d$ symmetric matrix $\overline{\Lambda}$ is given by

$$\overline{\Lambda} = \int_Q \Lambda(y)\, dy = \int_Q \Big(\mathrm{Id}_d + \nabla W^0(y) \Big)^T \Big(\mathrm{Id}_d + \nabla W^0(y) \Big)\, dy. \tag{2.70}$$

In addition, the number of indices $k \in \mathbb{Z}^d$ such that $Q \cap (D/\varepsilon - k) \neq \emptyset$ and $Q + k \not\subset D/\varepsilon$ is bounded by $C\varepsilon^{1-d}$ (this corresponds to cells $Q + k$ close to the boundary of D/ε). In the sum R_ε, there are hence at most $C\varepsilon^{1-d}$ non-vanishing terms, and each term can be bounded by $\|\psi\|^2_{L^\infty(D\times\mathbb{R}^d)}$. We thus have $R_\varepsilon \leq C\,\varepsilon$. Collecting this bound with (2.67) and (2.68), we deduce that

$$\lim_{\varepsilon\to 0} \sum_{k\in\mathbb{Z}^d} (Z_k^\varepsilon)^2 = \sigma^2_{\mathrm{per}}.$$

Inserting this result in (2.65) and collecting (2.64) and (2.66), we get that, for any $\xi \in \mathbb{R}$,

$$\lim_{\varepsilon\to 0} \theta_{C_1^\varepsilon}(\xi) = \exp\left(-\frac{\mathrm{Var}(X_0)}{2}\, \xi^2 \sigma^2_{\mathrm{per}} \right),$$

which means that the random variable C_1^ε converges in law to a centered Gaussian random variable of variance $\mathrm{Var}(X_0)\,\sigma^2_{\mathrm{per}}$.

Conclusion: limit of I_ε^1

We recall (see (2.57)) that

$$I_1^\varepsilon = -C_1^\varepsilon - C_u^\varepsilon - C_v^\varepsilon + C_r^\varepsilon,$$

and we have shown (see (2.58) and (2.60)) that

$$\mathbb{E}\big[(C_u^\varepsilon + C_v^\varepsilon - C_r^\varepsilon)^2\big] \leq C\varepsilon.$$

This implies that $C_u^\varepsilon + C_v^\varepsilon - C_r^\varepsilon$ converges in probability to 0. Indeed, for any $\kappa > 0$, we have

$$\mathbb{P}\Big[\Big|C_u^\varepsilon + C_v^\varepsilon - C_r^\varepsilon\Big| \geq \kappa\Big] \leq \frac{\mathbb{E}\big[(C_u^\varepsilon + C_v^\varepsilon - C_r^\varepsilon)^2\big]}{\kappa^2} \underset{\varepsilon\to 0}{\longrightarrow} 0.$$

Using the Slutsky theorem and the fact that C_1^ε converges in law, we get that I_1^ε converges in law to a centered Gaussian variable of variance $\sigma_1^2 = \mathrm{Var}(X_0)\,\sigma^2_{\mathrm{per}}$.

We now provide a more explicit expression for σ_1^2. By definition of σ^2_{per} (see (2.69)), we have

$$\begin{aligned}
\sigma_1^2 &= \mathrm{Var}(X_0) \int_D \Big(\nabla u_\star^0(x) \cdot \overline{\Lambda}\, \nabla v_\star^0(x) \Big)^2\, dx \\
&= \mathrm{Var}(X_0) \int_D \left(\sum_{i,j=1}^d \frac{\partial u_\star^0}{\partial x_i}\, \overline{\Lambda}_{ij}\, \frac{\partial v_\star^0}{\partial x_j} \right)^2 \\
&= \mathrm{Var}(X_0) \sum_{i,j,k,\ell=1}^d \int_D \frac{\partial u_\star^0}{\partial x_i}\, \frac{\partial v_\star^0}{\partial x_j}\, \overline{\Lambda}_{ij}\, \overline{\Lambda}_{k\ell}\, \frac{\partial u_\star^0}{\partial x_k}\, \frac{\partial v_\star^0}{\partial x_\ell} \\
&= \int_D \big(\nabla u_\star^0 \otimes \nabla v_\star^0 \big) : \mathcal{Q}^1 : \big(\nabla u_\star^0 \otimes \nabla v_\star^0 \big)
\end{aligned}$$

with the fourth order tensor $\mathcal{Q}^1$ defined by

$$\mathcal{Q}^1_{i,j,k,\ell} = \mathrm{Var}(X_0)\, \overline{\Lambda}_{ij}\, \overline{\Lambda}_{k\ell}.$$

In view of (2.70), we obtain the convergence (2.48) with σ_1^2 indeed defined by (2.36) and (2.39). This concludes the proof of Proposition 2.12.

2.4 Limit of (the leading order term of) $\mathcal{Q}^L$

In this section, we consider the fourth order tensor $\mathcal{Q}^{L,1}$, which is (see (2.26)) the first order term (in the expansion in powers of η) of $\mathcal{Q}^L$, and study its limit when $L \to \infty$. The main result of this section is Proposition 2.20.

We have shown (see (2.29)) that

$$\mathcal{Q}^{L,1}_{i,j,m,n} = \frac{1}{|Q_L|}\,\mathbb{E}\left(\int_{Q_L} \rho^1_{i,j} \int_{Q_L} \rho^1_{m,n}\right) \tag{2.71}$$

where $\rho^1_{i,j}$ is given (using (2.27) and the symmetry of $A_{\rm per}$ and $A^\star_{\rm per}$) by

$$\begin{aligned}\rho^1_{i,j} = (e_i + \nabla w^0_i)\cdot A_1(e_j + \nabla w^0_j) + \nabla w^1_i \cdot \left(A_{\rm per}(e_j + \nabla w^0_j) - A^\star_{\rm per} e_j\right) \\ + \nabla w^1_j \cdot \left(A_{\rm per}(e_i + \nabla w^0_i) - A^\star_{\rm per} e_i\right).\end{aligned} \tag{2.72}$$

The function w^1_i is the solution (unique up to the addition of a random constant) to (2.21). In view of (2.16), we are in position to use [14, Lemma 3.2]. Recalling that the expectation of X_k (and thus of A_1) vanishes, we have that

$$\nabla w^1_i(\cdot,\omega) = \sum_{\ell\in\mathbb{Z}^d} \nabla\phi_i(\cdot - \ell)\, X_\ell(\omega), \tag{2.73}$$

where the sum is a convergent series in $L^2(Q\times\Omega)$, and where ϕ_i is a deterministic function, which is the (unique up to the addition of a constant) solution to

$$\begin{cases} -\operatorname{div}[A_{\rm per}\nabla\phi_i] = \operatorname{div}[\mathbb{1}_Q(e_i + \nabla w^0_i)] \ \text{ in } \mathbb{R}^d, \\ \phi_i \in L^2_{\rm loc}(\mathbb{R}^d), \quad \nabla\phi_i \in (L^2(\mathbb{R}^d))^d. \end{cases} \tag{2.74}$$

We recall the fact (see [14, Lemma 3.1]) that there exists a solution ϕ_i to (2.74) which satisfies

$$\forall y\in\mathbb{R}^d \text{ with } |y|\ge 1, \qquad |\nabla\phi_i(y)| \le \frac{C}{|y|^d}, \tag{2.75}$$

$$\forall y\in\mathbb{R}^d, \qquad |\phi_i(y)| \le \frac{C}{1+|y|^{d-1}}, \tag{2.76}$$

for some finite C. In the sequel, we always consider for ϕ_i a solution to (2.74) satisfying the above two bounds.

We first compute $\displaystyle\int_{Q+k} \rho^1_{i,j}(\cdot,\omega)$ for any $k\in\mathbb{Z}^d$. In view of (2.72) and (2.16), we have

$$\begin{aligned}\int_{Q+k} \rho^1_{i,j}(\cdot,\omega) &= \sum_{\ell\in\mathbb{Z}^d} X_\ell(\omega) \int_{Q+k} \nabla\phi_i(\cdot-\ell)\cdot\left(A_{\rm per}(e_j+\nabla w^0_j) - A^\star_{\rm per}e_j\right) \\ &+ \sum_{\ell\in\mathbb{Z}^d} X_\ell(\omega) \int_{Q+k} \nabla\phi_j(\cdot-\ell)\cdot\left(A_{\rm per}(e_i+\nabla w^0_i) - A^\star_{\rm per}e_i\right) \\ &+ X_k(\omega)\int_Q (e_i+\nabla w^0_i)\cdot(e_j+\nabla w^0_j).\end{aligned} \tag{2.77}$$

Introduce

$$\begin{aligned}\beta^{k,\ell}_{i,j} &= \int_{Q+k} \nabla\phi_i(\cdot-\ell)\cdot\left(A_{\rm per}(e_j+\nabla w^0_j) - A^\star_{\rm per}e_j\right) \\ &= \int_{Q+k-\ell} \nabla\phi_i\cdot\left(A_{\rm per}(e_j+\nabla w^0_j) - A^\star_{\rm per}e_j\right).\end{aligned} \tag{2.78}$$

We then recast (2.77) as

$$\int_{Q+k} \rho^1_{i,j}(\cdot,\omega) = X_k(\omega)\,\overline{\Lambda}_{ij} + \sum_{\ell\in\mathbb{Z}^d} X_\ell(\omega)\,\beta^{k,\ell}_{i,j} + \sum_{\ell\in\mathbb{Z}^d} X_\ell(\omega)\,\beta^{k,\ell}_{j,i}$$

where the matrix $\overline{\Lambda}$ is defined by (2.70).

In view of the definition (2.8) of Q_L, we thus deduce that

$$\int_{Q_L} \rho^1_{i,j}(\cdot,\omega) = \sum_{|k|_\infty\le L} \left(X_k(\omega)\,\overline{\Lambda}_{ij} + \sum_{\ell\in\mathbb{Z}^d} X_\ell(\omega) \left(\beta^{k,\ell}_{i,j} + \beta^{k,\ell}_{j,i}\right)\right),$$

where we recall that $|k|_\infty = \max_{1\le i\le d} |k_i|$.

In view of (2.71), we thus get

$$\begin{aligned}
&\mathcal{Q}^{L,1}_{i,j,m,n} \\
&= \frac{1}{|Q_L|}\,\mathbb{E}\left[\int_{Q_L}\rho^1_{i,j}\int_{Q_L}\rho^1_{m,n}\right] \\
&= \frac{1}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\mathbb{E}\left[X_k(\omega)\,X_q(\omega)\,\overline{\Lambda}_{ij}\,\overline{\Lambda}_{mn}\right] \\
&+ \frac{1}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\sum_{\ell\in\mathbb{Z}^d}\sum_{\ell'\in\mathbb{Z}^d}\mathbb{E}\left[X_\ell(\omega)\,X_{\ell'}(\omega)\,\left(\beta^{k,\ell}_{i,j}+\beta^{k,\ell}_{j,i}\right)\left(\beta^{q,\ell'}_{m,n}+\beta^{q,\ell'}_{n,m}\right)\right] \\
&+ \frac{1}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\sum_{\ell'\in\mathbb{Z}^d}\mathbb{E}\left[X_k(\omega)\,X_{\ell'}(\omega)\,\overline{\Lambda}_{ij}\,\left(\beta^{q,\ell'}_{m,n}+\beta^{q,\ell'}_{n,m}\right)\right] \\
&+ \frac{1}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\sum_{\ell\in\mathbb{Z}^d}\mathbb{E}\left[X_q(\omega)\,X_\ell(\omega)\,\overline{\Lambda}_{mn}\,\left(\beta^{k,\ell}_{i,j}+\beta^{k,\ell}_{j,i}\right)\right].
\end{aligned}$$

Since X_k are i.i.d. and centered, we get

$$\begin{aligned}
\mathcal{Q}^{L,1}_{i,j,m,n} &= \mathrm{Var}(\mathrm{X}_0)\,\overline{\Lambda}_{ij}\,\overline{\Lambda}_{mn} \\
&+ \frac{\mathrm{Var}(\mathrm{X}_0)}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\sum_{\ell\in\mathbb{Z}^d}\left(\beta^{k,\ell}_{i,j}+\beta^{k,\ell}_{j,i}\right)\left(\beta^{q,\ell}_{m,n}+\beta^{q,\ell}_{n,m}\right) \\
&+ \frac{\mathrm{Var}(\mathrm{X}_0)}{|Q_L|}\,\overline{\Lambda}_{ij}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\left(\beta^{q,k}_{m,n}+\beta^{q,k}_{n,m}\right) \\
&+ \frac{\mathrm{Var}(\mathrm{X}_0)}{|Q_L|}\,\overline{\Lambda}_{mn}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\left(\beta^{k,q}_{i,j}+\beta^{k,q}_{j,i}\right).
\end{aligned} \tag{2.79}$$

Introducing

$$B^L_{1,i,j} = \frac{1}{|Q_L|}\sum_{|k|_\infty\le L}\sum_{|q|_\infty\le L}\beta^{k,q}_{i,j}, \tag{2.80}$$

$$B^L_{2,i,j,m,n} = \frac{1}{|Q_L|}\sum_{\ell\in\mathbb{Z}^d}\left(\sum_{|k|_\infty\le L}\left(\beta^{k,\ell}_{i,j}+\beta^{k,\ell}_{j,i}\right)\sum_{|q|_\infty\le L}\left(\beta^{q,\ell}_{m,n}+\beta^{q,\ell}_{n,m}\right)\right), \tag{2.81}$$

we can recast (2.79) as

$$\begin{aligned}
\mathcal{Q}^{L,1}_{i,j,m,n} = \mathrm{Var}(\mathrm{X}_0)\Big(&\overline{\Lambda}_{ij}\,\overline{\Lambda}_{mn} + B^L_{2,i,j,m,n} \\
&+ \overline{\Lambda}_{ij}\left(B^L_{1,m,n}+B^L_{1,n,m}\right) + \overline{\Lambda}_{mn}\left(B^L_{1,i,j}+B^L_{1,j,i}\right)\Big).
\end{aligned} \tag{2.82}$$

We have the following results.

Lemma 2.18. *Assume that (2.15) and (2.16) hold, and that $A_{\rm per}$ is an Hölder continuous function. Then, for any $1 \le i, j \le d$, we have*

$$\lim_{L\to\infty} B^L_{1,i,j} = 0.$$

The proof of Lemma 2.18 is postponed until Section 2.4.1. It actually shows that, for any $1 \le i, j \le d$, we have

$$\left| B^L_{1,i,j} \right| \le \frac{C \ln L}{L} \tag{2.83}$$

for some C independent of L.

Lemma 2.19. *Assume that (2.15) and (2.16) hold, and that $A_{\rm per}$ is an Hölder continuous function. Then, for any $1 \le i, j, m, n \le d$, we have*

$$\lim_{L\to\infty} B^L_{2,i,j,m,n} = 0.$$

The proof of Lemma 2.19 is postponed until Section 2.4.2. It actually shows that, for any $1 \le i, j, m, n \le d$, we have

$$\left| B^L_{2,i,j,m,n} \right| \le \frac{C(\ln L)^2}{L} \tag{2.84}$$

for some C independent of L.

The main result of this section is the following.

Proposition 2.20. *Assume that (2.15) and (2.16) hold, and that $A_{\rm per}$ is an Hölder continuous function. Then, for any $1 \le i, j, m, n \le d$, we have*

$$\lim_{L\to\infty} \mathcal{Q}^{L,1}_{i,j,m,n} = \mathcal{Q}^1_{i,j,m,n}$$

with $\mathcal{Q}^1_{i,j,m,n}$ defined by (2.39).

Proof of Proposition 2.20. Using Lemmas 2.18 and 2.19, we infer from (2.82) that

$$\lim_{L\to\infty} \mathcal{Q}^{L,1}_{i,j,m,n} = \mathrm{Var}(\mathrm{X}_0)\, \overline{\Lambda}_{ij}\, \overline{\Lambda}_{mn}.$$

In view of the definition (2.70) of the matrix $\overline{\Lambda}$, we get (2.40) with (2.39). Note also that (2.83) and (2.84) provide a convergence rate: $|\mathcal{Q}^{L,1}_{i,j,m,n} - \mathcal{Q}^1_{i,j,m,n}| \le C(\ln L)^2/L$. □

We are now left with proving Lemmas 2.18 and 2.19. To that aim, we need the following result (see [61, Proposition 4.1]) to represent divergence-free vector fields.

Lemma 2.21. *Consider $G : \mathbb{R}^d \mapsto \mathbb{R}^d$ a $\mathbb{Z}^d$-periodic vector field such that $\operatorname{div} G = 0$ in $\mathbb{R}^d$, $\displaystyle\int_Q G = 0$ and $G \in C^{0,\alpha}(Q)$.*

Then there exists a matrix field $\tau : Q \mapsto \mathbb{R}^{d\times d}$ with $\tau \in C^{1,\alpha}(Q)$, which is skew-symmetric, $\mathbb{Z}^d$-periodic, such that each component $\tau_{m,\ell}$ belongs to $H^1_{\rm loc}(\mathbb{R}^d)$ (for any $1 \le m, \ell \le d$) and such that

$$\forall 1 \le \ell \le d, \quad G \cdot e_\ell = \sum_{m=1}^{d} \frac{\partial \tau_{m,\ell}}{\partial x_m}.$$

This result is also given in [58, p. 27] for fields G in $L^2(Q)$ (and therefore $\tau \in H^1(Q)$).

In the following, we apply Lemma 2.21 for the vector field

$$G^j = A_{\rm per}(e_j + \nabla w_j^0) - A^\star_{\rm per} e_j, \tag{2.85}$$

which indeed satisfies the assumptions of Lemma 2.21 (the regularity of G^j stems from the assumption that $A_{\rm per}$ is Hölder continuous and from Lemma 2.13). For any $1 \le j \le d$, there hence exists a matrix field τ^j satisfying the above properties and such that

$$\forall 1 \le \ell \le d, \quad G^j \cdot e_\ell = \sum_{m=1}^d \frac{\partial \tau^j_{m,\ell}}{\partial x_m}. \tag{2.86}$$

Remark 2.22. *Our proofs of Lemmas 2.18 and 2.19, given in the subsequent sections, essentially follow the same arguments as those used in [14, Proof of Proposition 3.1]. It is proven there that the quantity*

$$\overline{B}^L_{2,i,j,m,n} = \frac{1}{|Q_L|}\left[\sum_{\ell\in\mathbb{Z}^d}\left(\sum_{|k|_\infty\le L}\left(\overline{\beta}^{k,\ell}_{i,j} + \overline{\beta}^{k,\ell}_{j,i}\right)\sum_{|q|_\infty\le L}\left(\overline{\beta}^{q,\ell}_{m,n} + \overline{\beta}^{q,\ell}_{n,m}\right)\right)\right],$$

where

$$\overline{\beta}^{k,\ell}_{i,j} = \int_{Q+k} \nabla\phi_i(\cdot - \ell)\cdot A_{\rm per}(e_j + \nabla w_j^0),$$

is uniformly bounded with respect to L. The proof in [14] uses the fact that $A_{\rm per}(e_j + \nabla w_j^0)$ is a divergence-free field, a direct consequence of the corrector equation.

In the expression of $\beta^{k,\ell}_{i,j}$ defined by (2.78) that we manipulate here, the key quantity is $G^j = A_{\rm per}(e_j + \nabla w_j^0) - A^\star_{\rm per} e_j$. This vector-field is not only divergence-free, but it has also a vanishing mean, by definition of the homogenized tensor $A^\star_{\rm per}$. This is the reason why we are able to show a stronger result on $B^L_{2,i,j,m,n}$ (namely the fact that it converges to 0 when $L \to \infty$) than on $\overline{B}^L_{2,i,j,m,n}$, which is only shown in [14] to be bounded.

We also need the following technical results.

Lemma 2.23. *For any $1 \le i, j \le d$ and any $k \in \mathbb{Z}^d$, let*

$$P_{i,j}(L,k) = \int_{k+Q_L} \nabla\phi_i \cdot G^j, \tag{2.87}$$

where G^j is defined by (2.85), ϕ_i is a solution to (2.74) satisfying (2.75)–(2.76), and $Q_L = (-L-1/2, L+1/2)^d$ (see (2.8)).

Assume that $|k|_\infty < L$. Then we have

$$\left|P_{i,j}(L,k)\right| \le C\int_{\partial(k+Q_L)} \frac{dy}{|y|^d} \tag{2.88}$$

for some C independent of L and k.

Proof of Lemma 2.23. We know that $\phi_i \in H^1_{\rm loc}(\mathbb{R}^d)$. For any $|k|_\infty < L$, we have $Q \subset\subset k + Q_L$ (this property may not be true if $|k|_\infty = L$). There thus exists a smooth neighboorhood $D_{L,k}$ of $\partial(k+Q_L)$ for which $\overline{Q} \cap \overline{D_{L,k}} = \emptyset$. In view of (2.74), we thus have $\operatorname{div}[A_{\rm per}\nabla\phi_i] = 0$ on $D_{L,k}$. In view of [43, Corollary 8.36], we get that $\phi_i \in C^{1,\alpha}(D_{L,k})$.

Since $C^\infty\big((k+Q_L)\cup D_{L,k}\big)$ is dense in $H^1\big((k+Q_L)\cup D_{L,k}\big)$, there exists $\phi_i^\eta \in C^2\big((k+Q_L)\cup D_{L,k}\big)$ such that $\phi_i^\eta \underset{\eta\to 0}{\longrightarrow} \phi_i$ for the norm $H^1\big((k+Q_L)\cup D_{L,k}\big)$. Moreover, since ϕ_i is $C^{1,\alpha}$ on $D_{L,k}$, we can choose ϕ_i^η such that ϕ_i^η converges to ϕ_i in the $C^{1,\alpha}(D_{L,k})$ norm, and hence such that $\lim_{\eta\to 0} \|\phi_i^\eta - \phi_i\|_{C^1(D_{L,k})} = 0$.

As pointed out underneath Lemma 2.21, for any $1 \le j \le d$, there exists a matrix field τ^j, satisfying the properties of Lemma 2.21, and such that the divergence-free vector field G^j can be written (see (2.86)) as $G^j \cdot e_\ell = \sum_{m=1}^d \dfrac{\partial \tau^j_{m,\ell}}{\partial x_m}$.

We then have

$$\begin{aligned} P^\eta_{i,j}(L,k) &:= \int_{k+Q_L} \nabla\phi_i^\eta \cdot G^j \\ &= \int_{k+Q_L} \sum_{m=1}^d \sum_{\ell=1}^d \frac{\partial \phi_i^\eta}{\partial x_\ell}\frac{\partial \tau^j_{m,\ell}}{\partial x_m}. \\ &= -\int_{k+Q_L} \sum_{m=1}^d \sum_{\ell=1}^d \frac{\partial^2 \phi_i^\eta}{\partial x_\ell \partial x_m}\tau^j_{m,\ell} + \int_{\partial(k+Q_L)} \sum_{m=1}^d \sum_{\ell=1}^d \frac{\partial \phi_i^\eta}{\partial x_\ell}\tau^j_{m,\ell}\, n_m. \end{aligned}$$

Since τ^j is skew-symmetric, the first term above vanishes, and we thus get

$$P^\eta_{i,j}(L,k) = \int_{\partial(k+Q_L)} \sum_{m=1}^d \sum_{\ell=1}^d \frac{\partial \phi_i^\eta}{\partial x_\ell}\tau^j_{m,\ell}\, n_m. \tag{2.89}$$

We now pass to the limit $\eta \to 0$ in (2.89). We first have

$$\Big|P^\eta_{i,j}(L,k) - P_{i,j}(L,k)\Big| = \left|\int_{k+Q_L} \nabla(\phi_i^\eta - \phi_i)\cdot G^j\right| \le C\|\phi_i^\eta - \phi_i\|_{H^1(k+Q_L)},$$

and hence, by definition of ϕ_i^η, we obtain that $P^\eta_{i,j}(L,k) \underset{\eta\to 0}{\longrightarrow} P_{i,j}(L,k)$. For the right-hand side of (2.89), we write, for any $1 \le m,\ell \le d$, that

$$\left|\int_{\partial(k+Q_L)} \left(\frac{\partial \phi_i^\eta}{\partial x_\ell} - \frac{\partial \phi_i}{\partial x_\ell}\right)\tau^j_{m,\ell}\, n_m\right| \le \|\phi_i^\eta - \phi_i\|_{C^1(D_{L,k})}\, \|\tau^j_{m,\ell}\|_{C^0(D_{L,k})}.$$

By definition of ϕ_i^η, we obtain that the above right-hand side converges to 0 when $\eta \to 0$. Passing to the limit in (2.89), we hence get that

$$P_{i,j}(L,k) = \int_{\partial(k+Q_L)} \sum_{m=1}^d \sum_{\ell=1}^d \frac{\partial \phi_i}{\partial x_\ell}\tau^j_{m,\ell}\, n_m.$$

Since τ^j is $\mathbb{Z}^d$ periodic and $C^{1,\alpha}$, this implies that

$$\Big|P_{i,j}(L,k)\Big| \le C_d\, \|\tau^j\|_{C^0(Q)} \int_{\partial(k+Q_L)} |\nabla\phi_i|. \tag{2.90}$$

Using again that $|k|_\infty < L$, we see that any $y \in \partial(k+Q_L)$ satisfies $|y| > 1$. We are thus in position to bound $\nabla\phi_i$ using (2.75). This yields (2.88) and concludes the proof of Lemma 2.23. □

Lemma 2.24. *Under the same assumptions as in Lemma 2.23, except that we now assume that $|k|_\infty \ge L+2$ instead of $|k|_\infty < L$, we again have (2.88).*

Proof of Lemma 2.24. The proof follows exactly the same arguments as that of Lemma 2.23. Recalling that $Q_L = (-L - 1/2, L + 1/2)^d$, we note that Q is at a positive distance from $k + Q_L$ whenever $|k|_\infty \geq L + 2$. There thus exists a smooth neighboorhood $D_{L,k}$ of $k + Q_L$ for which $\overline{Q} \cap \overline{D_{L,k}} = \emptyset$. In view of (2.74), we thus have $\operatorname{div}[A_{\rm per} \nabla \phi_i] = 0$ on $D_{L,k}$. In view of [43, Corollary 8.36], we get that $\phi_i \in C^{1,\alpha}(D_{L,k})$.

We can thus perform the same computations as in the proof of Lemma 2.23, and we obtain (2.90), namely

$$\Big| P_{i,j}(L,k) \Big| \leq C_d \, \|\tau^j\|_{C^0(Q)} \int_{\partial(k+Q_L)} |\nabla \phi_i|.$$

Using again that $|k|_\infty \geq L + 2$, we see that any $y \in \partial(k + Q_L)$ satisfies $|y| > 1$. We are thus in position to bound $\nabla \phi_i$ using (2.75). This yields (2.88) and concludes the proof of Lemma 2.24. □

2.4.1 Proof of Lemma 2.18

In view of (2.80), (2.78) and (2.85), we have

$$B_1^L = \frac{1}{|Q_L|} \sum_{|k|_\infty \leq L} \sum_{|q|_\infty \leq L} \beta_{i,j}^{k,q} = \frac{1}{|Q_L|} \sum_{|k|_\infty \leq L} \int_{k+Q_L} \nabla \phi_i \cdot G^j, \tag{2.91}$$

where we recall (see (2.8)) that $Q_L = \cup_{|q|_\infty \leq L} \, Q + q$. For the sake of simplicity, we omit in this proof the dependence of B_1^L with respect to (i,j). We split the sum (2.91) in two parts:

$$B_1^L = B_{1,\rm bndry}^L + B_{1,\rm bulk}^L, \tag{2.92}$$

with

$$\begin{aligned} B_{1,\rm bndry}^L &= \frac{1}{|Q_L|} \sum_{|k|_\infty = L} \int_{k+Q_L} \nabla \phi_i \cdot G^j, \\ B_{1,\rm bulk}^L &= \frac{1}{|Q_L|} \sum_{|k|_\infty < L} \int_{k+Q_L} \nabla \phi_i \cdot G^j. \end{aligned} \tag{2.93}$$

Bound on $B_{1,\rm bndry}^L$

Using that G^j is divergence-free, we have

$$B_{1,\rm bndry}^L = \frac{1}{|Q_L|} \sum_{|k|_\infty = L} \int_{\partial(k+Q_L)} \phi_i \, G^j \cdot n.$$

We have observed underneath Lemma 2.21 that G^j is periodic and Hölder continuous. Using (2.76), we get that

$$\Big| B_{1,\rm bndry}^L \Big| \leq \frac{C}{|Q_L|} \sum_{|k|_\infty = L} \int_{\partial(k+Q_L)} \frac{dy}{1 + |y|^{d-1}}. \tag{2.94}$$

Let $\mathbb{Z}_{\rm bndry}^L = \{k \in \mathbb{Z}^d, \ |k|_\infty = L\}$. This set is finite and its cardinal is smaller than $C_d \, (2L+1)^{d-1}$. For any $k \in \mathbb{Z}_{\rm bndry}^L$, we note that the origin $O_{\mathbb{R}^d}$ of $\mathbb{R}^d$ belongs to the interior of the cube $k + Q_L = k + (-L - 1/2, L + 1/2)^d$ and we have $d\big(O_{\mathbb{R}^d}, \partial(k + Q_L)\big) = 1/2$.

For any facet F_n (with $1 \le n \le 2^d$) of the cube $k + Q_L$, let O_n be the orthogonal projection of the origin $O_{\mathbb{R}^d}$ on that facet. Each facet is a cube in dimension $d-1$ of side length $2L+1$: it is thus included in the ball (of $\mathbb{R}^{d-1}$) of center O_n and of radius $R_L = (2L+1)\sqrt{d-1}$, that we denote $B_{d-1}(O_n, R_L)$. For any $y \in B_{d-1}(O_n, R_L)$, we have that

$$|y|^2 = |y - O_{R^d}|^2 = |y - O_n|^2 + |O_n - O_{R^d}|^2 \ge |y - O_n|^2,$$

and thus

$$\begin{aligned}
\int_{F_n} \frac{dy}{1+|y|^{d-1}} &\le \int_{B_{d-1}(O_n,R_L)} \frac{dy}{1+|y|^{d-1}} \\
&\le \int_{B_{d-1}(O_n,R_L)} \frac{dy}{1+|y-O_n|^{d-1}} \\
&= C_d \int_0^{R_L} \frac{r^{d-2}}{1+r^{d-1}}\, dr \\
&= C_d \ln\left(1 + R_L^{d-1}\right).
\end{aligned}$$

We hence get that, for any $|k|_\infty = L$,

$$\begin{aligned}
\int_{\partial(k+Q_L)} \frac{dy}{1+|y|^{d-1}} &= \sum_{\text{facet of } \partial(k+Q_L)} \int_{F_n} \frac{dy}{1+|y|^{d-1}} \\
&\le C_d \ln\left(1 + R_L^{d-1}\right). \qquad (2.95)
\end{aligned}$$

Inserting this bound in (2.94), we get that

$$\begin{aligned}
\left|B^L_{1,\text{bndry}}\right| &\le \frac{C}{|Q_L|} \sum_{k \in \mathbb{Z}^L_{\text{bndry}}} \ln\left(1 + R_L^{d-1}\right) \\
&\le \frac{C}{L} \ln\left(1 + R_L^{d-1}\right) \\
&\le \frac{C \ln L}{L}. \qquad (2.96)
\end{aligned}$$

Bound on $B^L_{1,\text{bulk}}$

In view of (2.93) and (2.87), we have

$$B^L_{1,\text{bulk}} = \frac{1}{|Q_L|} \sum_{|k|_\infty < L} P_{i,j}(L,k).$$

Using Lemma 2.23, we get

$$\left|B^L_{1,\text{bulk}}\right| \le \frac{C}{|Q_L|} \sum_{|k|_\infty < L} \int_{\partial(k+Q_L)} \frac{dy}{|y|^d} = \frac{C}{|Q_L|} \sum_{|k|_\infty < L} \int_{\partial Q_L} \frac{dy}{|y-k|^d}. \qquad (2.97)$$

For any k such that $|k|_\infty < L$ and any $y \in \partial Q_L$, we recall that $|y - k| > 1$. We wish to find some $\widetilde{y}_k$ defined as a convex linear combination of k and y such that

$$B(\widetilde{y}_k, 1/8) \subset B(y, |y-k|), \qquad (2.98)$$

$$B(\widetilde{y}_k, 1/8) \subset k + Q. \qquad (2.99)$$

To that aim, we set $\widetilde{y}_k = k + \dfrac{y-k}{8|y-k|}$.

The inclusion (2.98) holds true. Indeed, we note that $\widetilde{y}_k - y = k - y + \dfrac{y-k}{8|y-k|}$ and thus, recalling that $|y-k| > 1$, we obtain $|\widetilde{y}_k - y| = |y-k| - 1/8$. Considering now any $z \in B(\widetilde{y}_k, 1/8)$, we write

$$|z-y| \le |z-\widetilde{y}_k| + |\widetilde{y}_k - y| = |z-\widetilde{y}_k| + |y-k| - \frac{1}{8} < |y-k|,$$

and thus $z \in B(y, |y-k|)$. This proves (2.98).

The inclusion (2.99) also holds true. Indeed, for any $z \in B(\widetilde{y}_k, 1/8)$, we have

$$|z-k| \le |z-\widetilde{y}_k| + |\widetilde{y}_k - k| = |z-\widetilde{y}_k| + 1/8 < 1/4$$

and thus $z \in k + Q$. This proves (2.99).

Denote by V_d the volume of the ball $B(\widetilde{y}_k, 1/8)$. For any $z \in B(\widetilde{y}_k, 1/8)$, we have, using (2.98), that $|z-y| \le |y-k|$, hence

$$\frac{1}{|y-k|^d} \le \frac{1}{V_d} \int_{B(\widetilde{y}_k, 1/8)} \frac{dz}{|y-z|^d} \le \frac{1}{V_d} \int_{k+Q} \frac{dz}{|y-z|^d}, \tag{2.100}$$

where the second inequality stems from (2.99). We thus deduce from (2.97) that

$$\begin{aligned}
\left|B^L_{1,\mathrm{bulk}}\right| &\le \frac{C}{|Q_L|} \sum_{|k|_\infty < L} \int_{\partial Q_L} \frac{dy}{|y-k|^d} \\
&\le \frac{C}{|Q_L|\,V_d} \int_{y \in \partial Q_L} \sum_{|k|_\infty < L} \int_{z \in k+Q} \frac{dz\,dy}{|y-z|^d} \\
&= \frac{C}{|Q_L|\,V_d} \int_{y \in \partial Q_L} \int_{z \in Q_{L-1}} \frac{dz\,dy}{|y-z|^d} \\
&= \frac{C\,(2L+1)^{2d-1}}{(2L+1)^d\,|Q_L|\,V_d} \int_{y \in \partial Q} \int_{z \in (Q_{L-1})/(2L+1)} \frac{dz\,dy}{|y-z|^d} \\
&\le \frac{C}{L} \int_{y \in \partial Q} \int_{B(y,\sqrt{d}) \setminus B(y, \frac{1}{2L+1})} \frac{dz\,dy}{|y-z|^d},
\end{aligned}$$

where we have used that

$$(Q_{L-1})/(2L+1) = \left(-\frac{2L-1}{2(2L+1)}, \frac{2L-1}{2(2L+1)}\right)^d \subset B\big(y, \sqrt{d}\big) \setminus B\big(y, 1/(2L+1)\big)$$

whenever $y \in \partial Q$. We hence deduce that

$$\left|B^L_{1,\mathrm{bulk}}\right| \le \frac{C}{L} \int_{y \in \partial Q} \int_{\frac{1}{2L+1}}^{\sqrt{d}} \frac{dr}{r} \le \frac{C \ln L}{L}. \tag{2.101}$$

Conclusion

Collecting (2.92), (2.96) and (2.101), we deduce that $\lim_{L\to\infty} B^L_1 = 0$ (and more precisely that (2.83) holds), which concludes the proof of Lemma 2.18.

2.4.2 Proof of Lemma 2.19

For the sake of simplicity, we omit in this proof the dependency of B_2^L with respect to (i,j,m,n). We split the sum (2.81) in two parts:

$$B_2^L = B_{2,\mathrm{long}}^L + B_{2,\mathrm{short}}^L, \tag{2.102}$$

with

$$\begin{aligned} B_{2,\mathrm{long}}^L &= \frac{1}{|Q_L|} \sum_{|\ell|_\infty > 2L} \left(\sum_{|k|_\infty \le L} \left(\beta_{i,j}^{k,\ell} + \beta_{j,i}^{k,\ell} \right) \sum_{|q|_\infty \le L} \left(\beta_{m,n}^{q,\ell} + \beta_{n,m}^{q,\ell} \right) \right) \\ B_{2,\mathrm{short}}^L &= \frac{1}{|Q_L|} \sum_{|\ell|_\infty \le 2L} \left(\sum_{|k|_\infty \le L} \left(\beta_{i,j}^{k,\ell} + \beta_{j,i}^{k,\ell} \right) \sum_{|q|_\infty \le L} \left(\beta_{m,n}^{q,\ell} + \beta_{n,m}^{q,\ell} \right) \right). \end{aligned} \tag{2.103}$$

Bound on $B_{2,\mathrm{long}}^L$

In view of (2.78), (2.85) and (2.87), we have

$$\sum_{|k|_\infty \le L} \beta_{i,j}^{k,\ell} = \int_{Q_L - \ell} \nabla \phi_i \cdot G^j = P_{i,j}(L, -\ell),$$

hence

$$B_{2,\mathrm{long}}^L = \frac{1}{|Q_L|} \sum_{|\ell|_\infty > 2L} \Big(P_{i,j}(L,\ell) + P_{j,i}(L,\ell) \Big) \Big(P_{m,n}(L,\ell) + P_{n,m}(L,\ell) \Big).$$

Since $|\ell|_\infty > 2L \ge L+2$, we are in position to use Lemma 2.24, which yields

$$\left| B_{2,\mathrm{long}}^L \right| \le \frac{C}{|Q_L|} \sum_{|\ell|_\infty > 2L} \left[\int_{\partial(\ell + Q_L)} \frac{dy}{|y|^d} \right]^2 \le \frac{C}{|Q_L|} \sum_{|\ell|_\infty > 2L} \left[\frac{L^{d-1}}{(|\ell| - L - 1/2)^d} \right]^2,$$

where we have used that any $y \in \partial(\ell + Q_L)$ with $|\ell|_\infty > 2L$ satisfies $|y| \ge |\ell| - L - 1/2$. We thus deduce that

$$\left| B_{2,\mathrm{long}}^L \right| \le \frac{C}{L^2} \tag{2.104}$$

for some C independent of L.

Bound on $B_{2,\mathrm{short}}^L$

For the sum (2.103), we essentially argue as for the sum B_1^L defined by (2.91) and we split it in two parts:

$$B_{2,\mathrm{short}}^L = B_{2,\mathrm{short,bndry}}^L + B_{2,\mathrm{short,bulk}}^L \tag{2.105}$$

with

$$\begin{aligned} B_{2,\mathrm{short,bndry}}^L &= \frac{1}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left(\sum_{|k|_\infty \le L} \left(\beta_{i,j}^{k,\ell} + \beta_{j,i}^{k,\ell} \right) \sum_{|q|_\infty \le L} \left(\beta_{m,n}^{q,\ell} + \beta_{n,m}^{q,\ell} \right) \right), \\ B_{2,\mathrm{short,bulk}}^L &= \frac{1}{|Q_L|} \sum_{\substack{|\ell|_\infty \le 2L, \\ |\ell|_\infty \ne L,\, |\ell|_\infty \ne L+1}} \left(\sum_{|k|_\infty \le L} \left(\beta_{i,j}^{k,\ell} + \beta_{j,i}^{k,\ell} \right) \sum_{|q|_\infty \le L} \left(\beta_{m,n}^{q,\ell} + \beta_{n,m}^{q,\ell} \right) \right). \end{aligned}$$

We successively bound these two terms.

Step 1: bound on $B^L_{2,\mathrm{short,bndry}}$. Using an integration by parts as in Section 2.4.1 and the bound (2.76), we get

$$\begin{aligned}
&\left|B^L_{2,\mathrm{short,bndry}}\right| \\
&\le \frac{1}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left| \int_{\partial(\ell+Q_L)} \left(\phi_i\, G^j \cdot n + \phi_j\, G^i \cdot n\right) \right| \\
&\qquad\qquad \times \left| \int_{\partial(\ell+Q_L)} \left(\phi_m\, G^n \cdot n + \phi_n\, G^m \cdot n\right) \right| \\
&\le \frac{C}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left(\int_{\partial(\ell+Q_L)} \frac{dy}{1+|y|^{d-1}} \right)^2 . \qquad (2.106)
\end{aligned}$$

In the case when $|\ell|_\infty = L$, we are in position to use (2.95), which yields that

$$\int_{\partial(k+Q_L)} \frac{dy}{1+|y|^{d-1}} \le C_d \ln L. \qquad (2.107)$$

In the case $|\ell|_\infty = L+1$ (there are at most $C_d\, L^{d-1}$ such indices), we note that the origin of $\mathbb{R}^d$ is outside the cube $\ell + Q_L = \ell + (-L-1/2, L+1/2)^d$. For any facet F_n (with $1 \le n \le 2^d$) of the cube $\ell + Q_L$, let O_n be the orthogonal projection of the origin $O_{\mathbb{R}^d}$ on the hyperplane containing this facet. Each facet is a cube in dimension $d-1$ of side length $2L+1$: it is thus included in the ball (of $\mathbb{R}^{d-1}$) of center O_n and of radius $R_L = (2L+1)\sqrt{d-1}$, that we denote $B_{d-1}(O_n, R_L)$. For any $y \in B_{d-1}(O_n, R_L)$, we have that

$$|y|^2 = |y - O_{R^d}|^2 = |y - O_n|^2 + |O_n - O_{R^d}|^2 \ge |y - O_n|^2,$$

and thus, following the same computations as in Section 2.4.1, we get that, for any $|\ell|_\infty = L+1$,

$$\int_{\partial(\ell+Q_L)} \frac{dy}{1+|y|^{d-1}} \le C_d \ln L. \qquad (2.108)$$

Collecting (2.106), (2.107) and (2.108), we obtain

$$\left|B^L_{2,\mathrm{short,bndry}}\right| \le C\, \frac{(\ln L)^2}{L}. \qquad (2.109)$$

Step 2: bound on $B^L_{2,\mathrm{short,bulk}}$. We have

$$B^L_{2,\mathrm{short,bulk}} = \frac{1}{|Q_L|} \sum_{\substack{|\ell|_\infty \le 2L,\\ |\ell|_\infty \ne L,\, |\ell|_\infty \ne L+1}} \left(P_{i,j}(L,\ell) + P_{j,i}(L,\ell)\right)\left(P_{m,n}(L,\ell) + P_{n,m}(L,\ell)\right).$$

Using Lemmas 2.23 and 2.24, we deduce that

$$\begin{aligned}
\left|B^L_{2,\mathrm{short,bulk}}\right| &\le \frac{C}{|Q_L|} \sum_{\substack{|\ell|_\infty \le 2L,\\ |\ell|_\infty \ne L,\, |\ell|_\infty \ne L+1}} \left(\int_{\partial(\ell+Q_L)} \frac{dy}{|y|^d} \right)^2 \\
&\le \frac{C}{|Q_L|\,(2L+1)^2} \sum_{\substack{|\ell|_\infty \le 2L,\\ |\ell|_\infty \ne L,\, |\ell|_\infty \ne L+1}} \left(\int_{\partial Q} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} \right)^2 . \qquad (2.110)
\end{aligned}$$

For any facet F_n (with $1 \leq n \leq 2^d$) of the cube Q, let O_n be the orthogonal projection of $\ell/(2L+1)$ on that facet. For any ℓ such that $|\ell|_\infty \neq L$ and $|\ell|_\infty \neq L+1$, the distance between $\ell/(2L+1)$ and O_n is larger than $C_d/(2L+1)$ for some constant C_d only depending on the dimension. Each facet is a cube in dimension $d-1$ of unit side length: it is thus included in the ball (of $\mathbb{R}^{d-1}$) of center O_n and of radius $R_d = \sqrt{d-1}$, that we denote $B_{d-1}(O_n, R_d)$. For any $y \in B_{d-1}(O_n, R_d)$, we have that

$$\left| y - \frac{\ell}{2L+1} \right|^2 = |y - O_n|^2 + \left| O_n - \frac{\ell}{2L+1} \right|^2 \geq |y - O_n|^2 + \frac{C_d}{(2L+1)^2}$$

and thus

$$\begin{aligned}
\int_{F_n} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} &\leq \int_{B_{d-1}(O_n,R_d)} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} \\
&\leq \int_{B_{d-1}(O_n,R_d)} \frac{dy}{\left||y-O_n|^2 + \frac{C_d}{(2L+1)^2}\right|^{d/2}} \\
&= C_d \int_0^{R_d} \frac{r^{d-2}\, dr}{\left|r^2 + \frac{C_d}{(2L+1)^2}\right|^{d/2}} \\
&= C_d\,(2L+1) \int_0^{(2L+1)R_d/C_d} \frac{r^{d-2}\, dr}{\left|r^2+1\right|^{d/2}} \\
&\leq C_d\,(2L+1) \left(1 + \int_1^{(2L+1)R_d/C_d} \frac{dr}{r^2} \right) \\
&\leq C_d\,(2L+1).
\end{aligned}$$

We hence get that, for any $|\ell|_\infty \neq L$ and $|\ell|_\infty \neq L+1$,

$$\int_{\partial Q} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} = \sum_{\text{facet of } \partial Q} \int_{F_n} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} \leq C_d\,(2L+1). \tag{2.111}$$

Inserting this bound in (2.110), we get

$$\begin{aligned}
\left| B^L_{2,\text{short,bulk}} \right| &\leq \frac{C}{|Q_L|\,(2L+1)} \sum_{\substack{|\ell|_\infty \leq 2L, \\ |\ell|_\infty \neq L,\, |\ell|_\infty \neq L+1}} \int_{\partial Q} \frac{dy}{\left|y - \frac{\ell}{2L+1}\right|^d} \\
&= \frac{C}{|Q_L|} \sum_{\substack{|\ell|_\infty \leq 2L, \\ |\ell|_\infty \neq L,\, |\ell|_\infty \neq L+1}} \int_{\partial Q_L} \frac{dy}{|y-\ell|^d}.
\end{aligned}$$

Using the estimate (2.100) (which we proved in Section 2.4.1 for $|k|_\infty < L$ but which

is also valid for $|k|_\infty \le 2L$ as soon as $|k|_\infty \neq L$ and $|k|_\infty \neq L+1$), we obtain

$$\begin{aligned}
\left|B^L_{2,\mathrm{short,bulk}}\right| &\le \frac{C}{|Q_L|\,V_d}\int_{y\in\partial Q_L}\sum_{|\ell|_\infty<L}\int_{z\in\ell+Q}\frac{dz\,dy}{|y-z|^d} \\
&\quad + \frac{C}{|Q_L|\,V_d}\int_{y\in\partial Q_L}\sum_{L+1<|\ell|_\infty\le 2L}\int_{z\in\ell+Q}\frac{dz\,dy}{|y-z|^d} \\
&\le \frac{C}{|Q_L|\,V_d}\int_{y\in\partial Q_L}\int_{z\in Q_{L-1}}\frac{dz\,dy}{|y-z|^d} \\
&\quad + \frac{C}{|Q_L|\,V_d}\int_{y\in\partial Q_L}\int_{z\in Q_{2L}\setminus Q_{L+1}}\frac{dz\,dy}{|y-z|^d} \\
&\le \frac{C\,(2L+1)^{2d-1}}{|Q_L|\,V_d\,(2L+1)^d}\int_{y\in\partial Q}\int_{z\in(Q_{L-1}))/(2L+1)}\frac{dz\,dy}{|y-z|^d} \\
&\quad + \frac{C\,(2L+1)^{2d-1}}{|Q_L|\,V_d\,(2L+1)^d}\int_{y\in\partial Q}\int_{z\in(Q_{2L}\setminus Q_{L+1})/(2L+1)}\frac{dz\,dy}{|y-z|^d}.
\end{aligned}$$

We now use that, whenever $y\in\partial Q$, we have

$$\begin{aligned}
(Q_{L-1}))/(2L+1) &\subset B\big(y,\sqrt{d}\big)\setminus B\big(y,1/(2L+1)\big), \\
(Q_{2L}\setminus Q_{L+1})/(2L+1) &\subset B\big(y,\sqrt{d}\big)\setminus B\big(y,1/(2L+1)\big).
\end{aligned}$$

We thus deduce that

$$\begin{aligned}
\left|B^L_{2,\mathrm{short,bulk}}\right| &\le \frac{C}{L}\int_{y\in\partial Q}\int_{z\in B(y,\sqrt{d})\setminus B(y,1/(2L+1))}\frac{dz\,dy}{|y-z|^d} \\
&= \frac{C}{L}\int_{y\in\partial Q}\int_{1/(2L+1)}^{\sqrt{d}}\frac{dr}{r} \\
&\le C\frac{\ln L}{L}.
\end{aligned} \tag{2.112}$$

Conclusion

Collecting (2.102), (2.104), (2.105), (2.109) and (2.112), we deduce that $\lim_{L\to\infty} B^L_2 = 0$ (and more precisely that (2.84) holds), which concludes the proof of Lemma 2.19.

2.5 Numerical approximation of $\mathcal{Q}^L$

We consider now the practical case when we manipulate the approximate corrector w_p^N, solution to the problem (2.7) posed on the bounded domain Q_N, instead of the exact corrector w_p, solution to the corrector problem (2.2) posed on the whole space. This section is devoted to the proof of Theorem 2.10.

Similarly to expansion (2.73) of the first-order (in η) term w_p^1 in the corrector w_p, we have the expansion

$$\forall x\in Q_N,\qquad \nabla w_i^{N,1}(x) = \sum_{|\ell|_\infty\le N}\nabla\phi_i^N(x-\ell)\,X_\ell(\omega)$$

where we have used that $\big\{\ell\in\mathbb{Z}^d,\ \ Q+\ell\subset Q_N\big\} = \big\{\ell\in\mathbb{Z}^d,\ \ |\ell|_\infty\le N\big\}$ and where ϕ_i^N is a solution (unique up to the addition of a constant) to

$$\begin{cases} -\operatorname{div}[A_{\mathrm{per}}\nabla\phi_i^N] = \operatorname{div}[\mathbb{1}_Q(e_i+\nabla w_i^0)] \ \text{ in } Q_N, \\ \phi_i^N \text{ is } Q_N\text{-periodic.} \end{cases} \tag{2.113}$$

In view of (2.45) and (2.16), we have, for any $k \in \mathbb{Z}^d$ such that $Q + k \subset Q_N$, that

$$\int_{Q+k} \rho_{i,j}^{N,1}(\cdot,\omega) = X_k(\omega)\,\overline{\Lambda}_{ij} + \sum_{|\ell|_\infty \leq N} X_\ell(\omega)\,\beta_{i,j}^{N,k,\ell} + \sum_{|\ell|_\infty \leq N} X_\ell(\omega)\,\beta_{j,i}^{N,k,\ell}$$

where the matrix $\overline{\Lambda}$ is defined by (2.70) and where

$$\beta_{i,j}^{N,k,\ell} = \int_{Q+k} \nabla\phi_i^N(\cdot - \ell) \cdot \left(A_{\rm per}(e_j + \nabla w_j^0) - A_{\rm per}^\star e_j\right).$$

For any $L \leq N$, we hence get that

$$\int_{Q_L} \rho_{i,j}^{N,1}(\cdot,\omega) = \sum_{|k|_\infty \leq L} \left(X_k(\omega)\,\overline{\Lambda}_{ij} + \sum_{|\ell|_\infty \leq N} X_\ell(\omega) \left(\beta_{i,j}^{N,k,\ell} + \beta_{j,i}^{N,k,\ell}\right)\right).$$

In view of (2.44), we deduce (similarly to (2.82)) that

$$\begin{aligned}\mathcal{Q}_{i,j,m,n}^{L,N,1} = \mathrm{Var}(\mathrm{X}_0)\Big(&\overline{\Lambda}_{ij}\,\overline{\Lambda}_{mn} + B_{2,i,j,m,n}^{L,N} \\ &+ \overline{\Lambda}_{ij}\left(B_{1,m,n}^{L,N} + B_{1,n,m}^{L,N}\right) + \overline{\Lambda}_{mn}\left(B_{1,i,j}^{L,N} + B_{1,j,i}^{L,N}\right)\Big)\end{aligned} \tag{2.114}$$

with

$$B_{1,i,j}^{L,N} = \frac{1}{|Q_L|} \sum_{|k|_\infty \leq L} \sum_{|q|_\infty \leq L} \beta_{i,j}^{N,k,q}, \tag{2.115}$$

$$B_{2,i,j,m,n}^{L,N} = \frac{1}{|Q_L|} \sum_{|\ell|_\infty \leq N} \left(\sum_{|k|_\infty \leq L} \left(\beta_{i,j}^{N,k,\ell} + \beta_{j,i}^{N,k,\ell}\right) \sum_{|q|_\infty \leq L} \left(\beta_{m,n}^{N,q,\ell} + \beta_{n,m}^{N,q,\ell}\right)\right). \tag{2.116}$$

Theorem 2.10 is a direct consequence of the following two lemmas.

Lemma 2.25. *Assume that (2.15) and (2.16) hold, and that $A_{\rm per}$ is an Hölder continuous function. Then, for any $1 \leq i, j \leq d$ and whenever $N > L$, we have*

$$\left|B_{1,i,j}^{L,N}\right| \leq \frac{C \ln L}{L} \tag{2.117}$$

for some C independent of L and N.

Lemma 2.26. *Assume that (2.15) and (2.16) hold, and that $A_{\rm per}$ is an Hölder continuous function. Then, for any $1 \leq i, j, m, n \leq d$ and whenever $N > L$, we have*

$$\left|B_{2,i,j,m,n}^{L,N}\right| \leq \frac{C(\ln L)^2}{L} \tag{2.118}$$

for some C independent of L and N.

The proof of Lemma 2.25 (resp. Lemma 2.26) is postponed until Section 2.5.2 (resp. Section 2.5.3). Before turning to them, we recall now some useful results on Green functions.

2.5.1 Green function with Q_N-periodic boundary conditions

Let G_N be the Green function of the operator $\mathcal{L} := -\operatorname{div}[A_{\rm per}\nabla\cdot]$ supplied with periodic boundary conditions on Q_N. We recall that $G_N(\cdot, y)$ is solution to

$$\begin{cases} -\operatorname{div}[A_{\rm per}\nabla G_N(\cdot,y)] = -\dfrac{1}{|Q_N|} + \displaystyle\sum_{k\in\mathbb{Z}^d} \delta(\cdot - y - Nk) \ \text{ in } \mathbb{R}^d, \\ G_N(\cdot,y) \ \text{ is } Q_N\text{-periodic.} \end{cases} \tag{2.119}$$

We recall (see [60, Proposition 1.2]) that, under the assumption that $A_{\rm per}$ is Hölder continuous, the Green function G_N satisfies the following estimates:

$$\forall x, y \in \mathbb{R}^d, \quad |\nabla_x G_N(x,y)| + |\nabla_y G_N(x,y)| \le \frac{C}{|d_N(x,y)|^{d-1}}, \tag{2.120}$$

$$\forall x, y \in \mathbb{R}^d, \quad |\nabla_x \nabla_y G_N(x,y)| \le \frac{C}{|d_N(x,y)|^{d}}, \tag{2.121}$$

for some C independent of N, where $d_N(x,y) = \inf_{k\in\mathbb{Z}^3} |x - y - Nk|$.

We note that estimates similar to (2.120) and (2.121) are well-known for the case of the operator $\mathcal{L} := -\operatorname{div}[A_{\rm per}\nabla\cdot]$ supplied with homogeneous Dirichlet boundary conditions on the boundary of some bounded domain Ω, and for the case of the operator $\mathcal{L}$ in the whole space $\mathbb{R}^d$. We refer to [18] for a recent review. The case of periodic boundary conditions on Q_N has been studied in [60], where (2.120)–(2.121) are established. For the sake of completeness, we provide a proof of (2.120)–(2.121) in the case when $d = 3$ and $A_{\rm per} = \mathrm{Id}$ in Appendix 2.A below.

In the sequel, we use the following result, which is a direct consequence of (2.120)–(2.121). This result provides estimates on ϕ_i^N similar to the estimates (2.75) and (2.76) on ϕ_i.

Lemma 2.27. *Assume that (2.120)–(2.121) hold. Then there exists a solution ϕ_i^N to (2.113) which satisfies*

$$\forall x \in Q_N, \qquad |\phi_i^N(x)| \le \frac{C}{1 + |x|^{d-1}}, \tag{2.122}$$

$$\forall x \in Q_N \text{ with } |x| \ge 1, \qquad |\nabla\phi_i^N(x)| \le \frac{C}{|x|^d} \tag{2.123}$$

for some C independent of N.

Proof of Lemma 2.27. The proof follows the same argument as that of [14, Lemma 3.1]. Let G_N be a solution to (2.119). Then

$$\phi_i^N(x) = \int_{Q_N} G_N(x,\cdot)\ \operatorname{div}[\mathbb{1}_Q(e_i + \nabla w_i^0)]$$

is a solution to (2.113). By an integration by part, we obtain that

$$\phi_i^N(x) = -\int_Q (e_i + \nabla w_i^0)\cdot\nabla_y G_N(x,\cdot). \tag{2.124}$$

Let $x \in Q_N$ and $y \in Q$. We know from (2.120) that $|\nabla_y G_N(x,y| \le \dfrac{C}{|x-y|^{d-1}}$, and thus $\nabla_y G_N(x,\cdot) \in L^1(Q)$. Since $\nabla w_i^0 \in L^\infty(Q)$ (see Lemma 2.13), we deduce that

$\phi_i^N(x)$ is (uniformly in N and $x \in Q_N$) bounded. Furthermore, we deduce from (2.120) and (2.124) that, when $|x| \geq 1$, we have $|\phi_i^N(x)| \leq C/(1+|x|^{d-1})$ for some C independent of N. This proves (2.122).

We next infer from (2.124) that

$$\nabla \phi_i^N(x) = -\int_Q (e_i + \nabla w_i^0) \cdot \nabla_x \nabla_y G_N(x, \cdot),$$

hence, using (2.121), we get

$$|\nabla \phi_i^N(x)| \leq C \int_Q \frac{dy}{|x-y|^d}.$$

For any $x \in Q_N$ with $|x| \geq 1$, we write

$$|\nabla \phi_i^N(x)| \leq \frac{C}{|x|^d} \int_Q \frac{dy}{|\frac{x}{|x|} - \frac{y}{|x|}|^d} \leq \frac{C}{|x|^d},$$

where the last inequality stems from the fact that the integral is bounded. This proves (2.123) and concludes the proof of Lemma 2.27. □

2.5.2 Proof of Lemma 2.25

We have

$$\sum_{|q|_\infty \leq L} \beta_{i,j}^{N,k,q} = \int_{k+Q_L} \nabla \phi_i^N \cdot G^j, \tag{2.125}$$

and thus, similarly to (2.91), we have

$$B_1^{L,N} = \frac{1}{|Q_L|} \sum_{|k|_\infty \leq L} \int_{k+Q_L} \nabla \phi_i^N \cdot G^j,$$

where we recall that G^j is defined by (2.85) (we again omit the dependency of $B_1^{L,N}$ with respect to (i,j)).

If $N \geq 2L$, then the set $k+Q_L$ (for any $|k|_\infty \leq L$) is a subset of Q_N, on which we have the estimates (2.122) and (2.123) on ϕ_i^N. These estimates are uniform in N, and identical to the estimates (2.75) and (2.76) on ϕ_i. We can hence proceed as in the proof of Lemma 2.18 (see Section 2.4.1) and deduce (similarly to (2.83)) that

$$\text{if } N \geq 2L, \text{ then } \qquad \left| B_1^{L,N} \right| \leq \frac{C \ln L}{L} \tag{2.126}$$

for some C independent of N and L.

In the sequel of the proof, we address the case $L < N < 2L$. In that situation, some sets $k+Q_L$ are a subset of Q_N (this corresponds to the case $|k|_\infty \leq N-L$), while some others are not (in particular for the case $|k|_\infty = L$). In that latter case, we can translate the cells of $(k+Q_L) \setminus Q_N$ (shown in blue on Figure 2.1) by Q_N periodicity to some cells in Q_N, that we denote $R_k^{N,L}$ (shown in green on Figure 2.1).

We split the sum $B_1^{L,N}$ as

$$B_1^{L,N} = B_{1,\text{bulk,int}}^{L,N} + B_{1,\text{bulk,ext}}^{L,N} + B_{1,\text{bndry}}^{L,N} \tag{2.127}$$

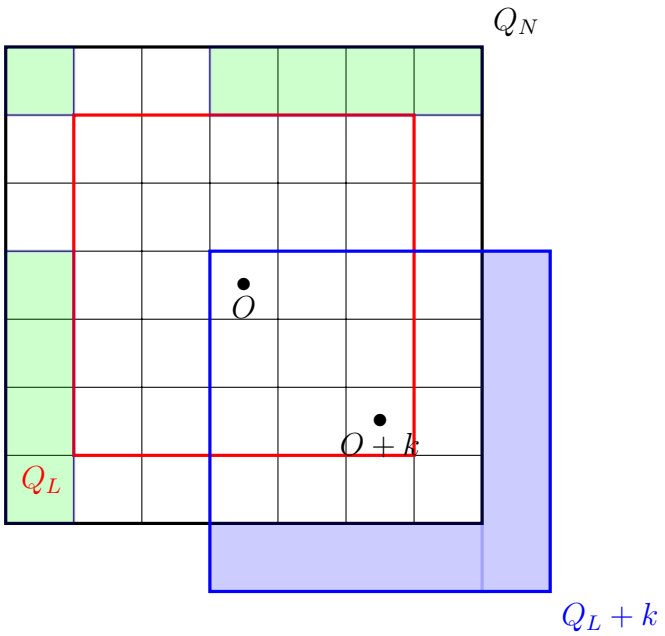


Figure 2.1: Schematic representation (when $L < N < 2L$) of a case of some $k \in \mathbb{Z}^d$ with $|k|_\infty \leq N$ such that $k + Q_L \not\subset Q_N$. The domain $(k + Q_L) \setminus Q_N$ is represented in blue. By Q_N-periodicity, this blue domain is mapped back in a open set in Q_N, denoted $R_k^{N,L}$, and which is represented in green.

with

$$B^{L,N}_{1,\text{bulk,int}} = \frac{1}{|Q_L|} \sum_{|k|_\infty \leq N-L} \int_{k+Q_L} \nabla \phi_i^N \cdot G^j, \tag{2.128}$$

$$B^{L,N}_{1,\text{bulk,ext}} = \frac{1}{|Q_L|} \sum_{N-L<|k|_\infty<L} \int_{k+Q_L} \nabla \phi_i^N \cdot G^j,$$

$$B^{L,N}_{1,\text{bndry}} = \frac{1}{|Q_L|} \sum_{|k|_\infty=L} \int_{k+Q_L} \nabla \phi_i^N \cdot G^j.$$

Using the Q_N periodicity of $\nabla \phi_i^N \cdot G^j$, we have

$$B^{L,N}_{1,\text{bulk,ext}} = \frac{1}{|Q_L|} \sum_{N-L<|k|_\infty<L} \left(\int_{(k+Q_L)\cap Q_N} \nabla \phi_i^N \cdot G^j + \int_{R_k^{L,N}} \nabla \phi_i^N \cdot G^j \right) \tag{2.129}$$

and

$$B^{L,N}_{1,\text{bndry}} = \frac{1}{|Q_L|} \sum_{|k|_\infty=L} \left(\int_{(k+Q_L)\cap Q_N} \nabla \phi_i^N \cdot G^j + \int_{R_k^{L,N}} \nabla \phi_i^N \cdot G^j \right). \tag{2.130}$$

Bound on $B^{L,N}_{1,\text{bulk,int}}$

For this first term (2.128), we have $k + Q_L \subset Q_N$ and $|k|_\infty \leq N - L < L$. We can hence use Lemma 2.23 (written for ϕ_i^N instead of ϕ_i) to bound the integral of $\nabla \phi_i^N \cdot G^j$

on $k + Q_L$ (since ϕ_i^N satisfies the same estimates as ϕ_i). We hence get

$$\begin{aligned}\left|B^{L,N}_{1,\text{bulk,int}}\right| &\leq \frac{C}{|Q_L|}\sum_{|k|_\infty \leq N-L}\int_{\partial(k+Q_L)}\frac{dy}{|y|^d}\\ &= \frac{C}{|Q_L|}\sum_{|k|_\infty \leq N-L}\int_{\partial Q_L}\frac{dy}{|y-k|^d}\\ &\leq \frac{C}{|Q_L|}\sum_{|k|_\infty < L}\int_{\partial Q_L}\frac{dy}{|y-k|^d}.\end{aligned}$$

We can next proceed as underneath (2.100) in Section 2.4.1, and obtain (similarly to (2.101)) that

$$\left|B^{L,N}_{1,\text{bulk,int}}\right| \leq \frac{C\,\ln L}{L}. \tag{2.131}$$

Limit of $B^{L,N}_{1,\text{bulk,ext}}$

For this second term, written in the form (2.129), we consider $k \in \mathbb{Z}^d$ such that $N - L < |k|_\infty < L$. We hence have that $Q \subset\subset (k + Q_L) \cap Q_N$. We can thus again use a result similar to that shown in Lemma 2.23 to bound the integral of $\nabla\phi_i^N \cdot G^j$ on $(k + Q_L) \cap Q_N$. We hence get

$$\left|\int_{(k+Q_L)\cap Q_N}\nabla\phi_i^N \cdot G^j\right| \leq C\int_{\partial\left((k+Q_L)\cap Q_N\right)}\frac{dy}{|y|^d}.$$

The boundary of $(k + Q_L) \cap Q_N$ is made of a part of the boundary of $k + Q_L$ and a part of the boundary of Q_N:

$$\partial\big((k + Q_L) \cap Q_N\big) \subset \partial(k + Q_L) \cup \partial_{k,L}Q_N, \tag{2.132}$$

where

$$\begin{aligned}&\partial_{k,L}Q_N \subset \partial Q_N \text{ is a part of the boundary of } Q_N\\ &\text{of measure bounded by } CL^{d-1}.\end{aligned} \tag{2.133}$$

The property (2.133) will be useful in Section 2.5.3 below. Here, we simply observe that the above inclusion of course implies that $\partial\big((k + Q_L) \cap Q_N\big) \subset \partial(k + Q_L) \cup \partial Q_N$, and we hence have

$$\left|\int_{(k+Q_L)\cap Q_N}\nabla\phi_i^N \cdot G^j\right| \leq C\int_{\partial(k+Q_L)}\frac{dy}{|y|^d} + C\int_{\partial Q_N}\frac{dy}{|y|^d}. \tag{2.134}$$

In the same fashion, the distance between Q and $R_k^{L,N}$ is bounded away from 0. We can hence again use a result similar to that shown in Lemma 2.24 to bound the integral of $\nabla\phi_i^N \cdot G^j$ on $R_k^{L,N}$. We hence get

$$\left|\int_{R_k^{L,N}}\nabla\phi_i^N \cdot G^j\right| \leq C\int_{\partial R_k^{L,N}}\frac{dy}{|y|^d}.$$

In general, the set $R_k^{L,N}$ is disconnected (see the green set on Figure 2.1), and composed of at most C_d connected components ($C_d = 3$ when $d = 2$). The boundary of each of these connected components is contained in $\partial Q_N \cup (\cup_{1\leq i\leq d}\partial Q_{\mathcal{L}_i})$, with $\mathcal{L}_i = 2N - L - |k_i|$. We hence have

$$\partial R_k^{L,N} \subset \partial_{k,L}Q_N \cup \left(\cup_{1\leq i\leq d}\partial_{k,L,N}Q_{\mathcal{L}_i}\right), \tag{2.135}$$

where $\partial_{k,L} Q_N$ satisfies (2.133) and where

$$\begin{array}{c} \partial_{k,L,N} Q_{\mathcal{L}_i} \subset \partial Q_{\mathcal{L}_i} \text{ is a part of the boundary of } Q_{\mathcal{L}_i} \\ \text{of measure bounded by } C L^{d-1}. \end{array} \tag{2.136}$$

The property (2.136) will be useful in Section 2.5.3 below. Here, we again simply observe that the above inclusion of course implies that $\partial R_k^{L,N} \subset \partial Q_N \cup (\cup_{1 \le i \le d} \partial Q_{\mathcal{L}_i})$, and thus

$$\left| \int_{R_k^{L,N}} \nabla \phi_i^N \cdot G^j \right| \le C \int_{\partial Q_N} \frac{dy}{|y|^d} + C \sum_{i=1}^d \int_{\partial Q_{\mathcal{L}_i}} \frac{dy}{|y|^d}.$$

Since $\mathcal{L}_i \ge 2N - L - |k|_\infty$, we deduce that

$$\left| \int_{R_k^{L,N}} \nabla \phi_i^N \cdot G^j \right| \le C \int_{\partial Q_N} \frac{dy}{|y|^d} + C \int_{\partial Q_{2N-L-|k|_\infty}} \frac{dy}{|y|^d}. \tag{2.137}$$

Indeed, there exists a constant C_d (which only depends on the dimension d) such that

$$\int_{\partial Q_{\mathcal{L}_i}} \frac{dy}{|y|^d} = \frac{C_d}{\mathcal{L}_i} \le \frac{C_d}{2N - L - |k|_\infty} = \int_{\partial Q_{2N-L-|k|_\infty}} \frac{dy}{|y|^d}.$$

Collecting (2.129), (2.134) and (2.137), we thus get that

$$\begin{aligned} & \left| B^{L,N}_{1,\mathrm{bulk,ext}} \right| \\ & \le \frac{C}{|Q_L|} \sum_{N-L<|k|_\infty<L} \left(\int_{\partial(k+Q_L)} \frac{dy}{|y|^d} + \int_{\partial Q_N} \frac{dy}{|y|^d} + \int_{\partial Q_{2N-L-|k|_\infty}} \frac{dy}{|y|^d} \right) \\ & \le \frac{C}{|Q_L|} \sum_{N-L<|k|_\infty<L} \left(\int_{\partial Q_L} \frac{dy}{|y-k|^d} + \frac{1}{N} + \frac{1}{2N-L-|k|_\infty} \right) \\ & \le \left(\frac{C}{|Q_L|} \sum_{|k|_\infty<L} \int_{\partial Q_L} \frac{dy}{|y-k|^d} \right) + \frac{C}{|Q_L|} \frac{L^d - (N-L)^d}{N} \\ & \qquad + \frac{C}{|Q_L|} \sum_{j=N-L+1}^{L-1} \frac{j^{d-1}}{2N-L-j}. \end{aligned} \tag{2.138}$$

The first term of (2.138) is estimated as underneath (2.100) in Section 2.4.1, and hence bounded by $(C \ln L)/L$ (see (2.101)). The second term is obviously bounded by C/N. Turning to the third term, we write

$$\begin{aligned} & \frac{C}{|Q_L|} \sum_{j=N-L+1}^{L-1} \frac{j^{d-1}}{2N-L-j} \\ & = \frac{C}{|Q_L|} \sum_{j=2N-2L+1}^{N-1} \frac{(2N-L-j)^{d-1}}{j} \\ & = \frac{C}{|Q_L|} \sum_{p=0}^{d-1} \binom{d-1}{p} (2N-L)^{d-1-p} (-1)^p \sum_{j=2N-2L+1}^{N-1} j^{p-1}. \end{aligned}$$

Bounding from above each sum in j, we get

$$\frac{C}{|Q_L|}\sum_{j=N-L+1}^{L-1}\frac{j^{d-1}}{2N-L-j} \leq \frac{C}{L^d}(2N-L)^{d-1}\ln N + \frac{C}{L^d}\sum_{p=1}^{d-1}\begin{pmatrix} d-1 \\ p \end{pmatrix}(2N-L)^{d-1-p}N^p.$$

Since $1 \leq N/L \leq 2$, we deduce that

$$\begin{aligned}&\frac{C}{|Q_L|}\sum_{j=N-L+1}^{L-1}\frac{j^{d-1}}{2N-L-j}\\ &\leq \frac{C}{L}\left(2\frac{N}{L}-1\right)^{d-1}\ln(2L) + \frac{C}{L}\sum_{p=1}^{d-1}\begin{pmatrix} d-1 \\ p \end{pmatrix}\left(2\frac{N}{L}-1\right)^{d-1-p}\left(\frac{N}{L}\right)^p\\ &\leq \frac{C\ln L}{L}+\frac{C}{L}.\end{aligned}$$

We therefore infer from (2.138) that

$$\left|B^{L,N}_{1,\mathrm{bulk,ext}}\right| \leq \frac{C\ln L}{L}+\frac{C}{N}+\frac{C}{L} \leq \frac{C\ln L}{L}. \tag{2.139}$$

Bound on $B^{L,N}_{1,\mathrm{bndry}}$

For that term (2.130), we proceed as in Section 2.4.1. Using that G^j is divergence-free, we recast (2.130) as

$$B^{L,N}_{1,\mathrm{bndry}} = \frac{1}{|Q_L|}\sum_{|k|_\infty=L}\left(\int_{\partial\left((k+Q_L)\cap Q_N\right)}\phi_i^N\, G^j\cdot n + \int_{\partial R_k^{L,N}}\phi_i^N\, G^j\cdot n\right).$$

Using that G^j is periodic and Hölder continuous and the estimate (2.122), we get that

$$\begin{aligned}&\left|B^{L,N}_{1,\mathrm{bndry}}\right|\\ &\leq \frac{C}{|Q_L|}\sum_{|k|_\infty=L}\left(\int_{\partial\left((k+Q_L)\cap Q_N\right)}\frac{dy}{1+|y|^{d-1}}+\int_{\partial R_k^{L,N}}\frac{dy}{1+|y|^{d-1}}\right)\\ &\leq \frac{C}{|Q_L|}\sum_{|k|_\infty=L}\left(\int_{\partial(k+Q_L)}\frac{dy}{1+|y|^{d-1}}+\int_{\partial Q_N}\frac{dy}{1+|y|^{d-1}}+\sum_{i=1}^{d}\int_{\partial Q_{2N-L-|k_i|}}\frac{dy}{1+|y|^{d-1}}\right),\end{aligned}$$

where we have used the same remarks about the boundaries of $(k+Q_L)\cap Q_N$ and $R_k^{L,N}$ as in Section 2.5.2 (see (2.132) and (2.135)). The last two integrals can be bounded by a constant independent of N and L. We thus obtain

$$\left|B^{L,N}_{1,\mathrm{bndry}}\right| \leq \frac{C}{L}+\frac{C}{|Q_L|}\sum_{|k|_\infty=L}\int_{\partial(k+Q_L)}\frac{dy}{1+|y|^{d-1}}.$$

We can then proceed as underneath (2.94) and deduce (similarly to (2.96)) that

$$\left|B^{L,N}_{1,\mathrm{bndry}}\right| \leq \frac{C}{L}+\frac{C\ln L}{L} \leq \frac{C\ln L}{L}. \tag{2.140}$$

Conclusion

Collecting (2.126) (in the case $N \geq 2L$) and (2.127), (2.131), (2.139) and (2.140) (in the case $L < N < 2L$), we deduce (2.117).

2.5.3 Proof of Lemma 2.26

In view of (2.125), we recast the sum (2.116) as

$$B^{L,N}_{2,i,j,m,n} = \frac{1}{|Q_L|} \sum_{|\ell|_\infty \leq N} \left(\int_{\ell+Q_L} \nabla \phi^N_i \cdot G^j + \int_{\ell+Q_L} \nabla \phi^N_j \cdot G^i \right) \times \left(\int_{\ell+Q_L} \nabla \phi^N_m \cdot G^n + \int_{\ell+Q_L} \nabla \phi^N_n \cdot G^m \right)$$

with G^j defined by (2.85) and ϕ^N_i defined by (2.113).

Similarly to (2.102) and (2.105) (and again omitting the dependency of $B^{L,N}_2$ with respect to (i,j,m,n)), we split the sum as

$$B^{L,N}_2 = B^{L,N}_{2,\mathrm{long}} + B^{L,N}_{2,\mathrm{short,bndry}} + B^{L,N}_{2,\mathrm{short,bulk}} \tag{2.141}$$

with

$$B^{L,N}_{2,\mathrm{long}} = \frac{1}{|Q_L|} \sum_{2L < |\ell|_\infty \leq N} \left(\int_{\ell+Q_L} \nabla \phi^N_i \cdot G^j + \int_{\ell+Q_L} \nabla \phi^N_j \cdot G^i \right) \times \left(\int_{\ell+Q_L} \nabla \phi^N_m \cdot G^n + \int_{\ell+Q_L} \nabla \phi^N_n \cdot G^m \right), \tag{2.142}$$

$$B^{L,N}_{2,\mathrm{short,bndry}} = \frac{1}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left(\int_{\ell+Q_L} \nabla \phi^N_i \cdot G^j + \int_{\ell+Q_L} \nabla \phi^N_j \cdot G^i \right) \times \left(\int_{\ell+Q_L} \nabla \phi^N_m \cdot G^n + \int_{\ell+Q_L} \nabla \phi^N_n \cdot G^m \right) \tag{2.143}$$

and

$$B^{L,N}_{2,\mathrm{short,bulk}} = \frac{1}{|Q_L|} \sum_{\substack{|\ell|_\infty \leq \min(2L,N), \\ |\ell|_\infty \neq L,\, |\ell|_\infty \neq L+1}} \left(\int_{\ell+Q_L} \nabla \phi^N_i \cdot G^j + \int_{\ell+Q_L} \nabla \phi^N_j \cdot G^i \right) \times \left(\int_{\ell+Q_L} \nabla \phi^N_m \cdot G^n + \int_{\ell+Q_L} \nabla \phi^N_n \cdot G^m \right). \tag{2.144}$$

If $N \leq 2L$, then the sum in $B^{L,N}_{2,\mathrm{long}}$ is simply void.

Limit of $B^{L,N}_{2,\mathrm{short,bndry}}$

Assume first that $N \geq 2L+1$. Then, in (2.143), all the cubes $\ell + Q_L$ (for any $\ell \in \mathbb{Z}^d$ with $|\ell|_\infty = L$ or $L+1$) are included in Q_N. In Q_N, we have the same estimates for ϕ^N_i as we have for ϕ_i. Following the same arguments as in Section 2.4.2 (Step 1), we hence get (similarly to (2.109)) that

$$\text{if } N \geq 2L+1, \text{ then } \qquad \left| B^{L,N}_{2,\mathrm{short,bndry}} \right| \leq C \, \frac{(\ln L)^2}{L}. \tag{2.145}$$

We now consider the case $N < 2L+1$. For any $\ell \in \mathbb{Z}^d$ with $|\ell|_\infty = L$ or $L+1$, we write

$$\int_{\ell+Q_L} \nabla\phi_i^N \cdot G^j = \int_{(\ell+Q_L)\cap Q_N} \nabla\phi_i^N \cdot G^j + \int_{R_\ell^{L,N}} \nabla\phi_i^N \cdot G^j, \tag{2.146}$$

where $R_\ell^{L,N} \subset Q_N$ is obtained by Q_N periodicity from $(\ell+Q_L)\setminus Q_N$. Using an integration by parts as in Section 2.4.2 and the bound (2.122), we get

$$\left|\int_{\ell+Q_L} \nabla\phi_i^N \cdot G^j\right| \le \int_{\partial\left((\ell+Q_L)\cap Q_N\right)} \frac{C\,dy}{1+|y|^{d-1}} + \int_{\partial R_\ell^{L,N}} \frac{C\,dy}{1+|y|^{d-1}}.$$

Using next the inclusions (2.132) and (2.135), we infer that

$$\begin{aligned}
&\left|\int_{\ell+Q_L} \nabla\phi_i^N \cdot G^j\right| \\
&\le \int_{\partial(\ell+Q_L)} \frac{C\,dy}{1+|y|^{d-1}} + \int_{\partial Q_N} \frac{C\,dy}{1+|y|^{d-1}} + \sum_{i=1}^d \int_{\partial Q_{2N-L-|\ell_i|}} \frac{C\,dy}{1+|y|^{d-1}} \\
&\le C\left(1 + \int_{\partial(\ell+Q_L)} \frac{dy}{1+|y|^{d-1}}\right).
\end{aligned}$$

We thus obtain that

$$\begin{aligned}
\left|B^{L,N}_{2,\mathrm{short,bndry}}\right| &\le \frac{C}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left(1 + \int_{\partial(\ell+Q_L)} \frac{dy}{1+|y|^{d-1}}\right)^2 \\
&\le \frac{C}{L} + \frac{C}{|Q_L|} \sum_{|\ell|_\infty = L \text{ or } L+1} \left(\int_{\partial(\ell+Q_L)} \frac{dy}{1+|y|^{d-1}}\right)^2.
\end{aligned}$$

We next proceed as underneath (2.106) and deduce (similarly to (2.109)) that

$$\text{if } N < 2L+1, \text{ then} \qquad \left|B^{L,N}_{2,\mathrm{short,bndry}}\right| \le C\,\frac{(\ln L)^2}{L}. \tag{2.147}$$

Limit of $B^{L,N}_{2,\mathrm{short,bulk}}$

Assume first that $N \ge 3L$. Then, in (2.144), all the cubes $\ell+Q_L$ (for any $\ell \in \mathbb{Z}^d$ with $|\ell|_\infty \le \min(2L, N) \le 2L$) are included in Q_N. In Q_N, we have the same estimates for ϕ_i^N as we have for ϕ_i. Following the same arguments as in Section 2.4.2 (Step 2), we hence get (similarly to (2.112)) that

$$\text{if } N \ge 3L, \text{ then} \qquad \left|B^{L,N}_{2,\mathrm{short,bulk}}\right| \le C\,\frac{\ln L}{L}. \tag{2.148}$$

We now consider the case $L < N < 3L$ and split the sum (2.144) into

$$B^{L,N}_{2,\mathrm{short,bulk}} = B^{L,N}_{2,\mathrm{short,bulk,int}} + B^{L,N}_{2,\mathrm{short,bulk,ext}} \tag{2.149}$$

with

$$\begin{aligned}
B^{L,N}_{2,\mathrm{short,bulk,int}} = \frac{1}{|Q_L|} \sum_{\substack{|\ell|_\infty \le N-L,\\ |\ell|_\infty \ne L,\, |\ell|_\infty \ne L+1}} &\left(\int_{\ell+Q_L} \nabla\phi_i^N \cdot G^j + \int_{\ell+Q_L} \nabla\phi_j^N \cdot G^i\right) \\
&\times \left(\int_{\ell+Q_L} \nabla\phi_m^N \cdot G^n + \int_{\ell+Q_L} \nabla\phi_n^N \cdot G^m\right)
\end{aligned} \tag{2.150}$$

and

$$B^{L,N}_{2,\mathrm{short},\mathrm{bulk},\mathrm{ext}} = \frac{1}{|Q_L|} \sum_{\substack{N-L<|\ell|_\infty\le \min(2L,N),\\ |\ell|_\infty\neq L,\,|\ell|_\infty\neq L+1}} \left(\int_{\ell+Q_L} \nabla\phi_i^N\cdot G^j + \int_{\ell+Q_L}\nabla\phi_j^N\cdot G^i\right) \times \left(\int_{\ell+Q_L}\nabla\phi_m^N\cdot G^n + \int_{\ell+Q_L}\nabla\phi_n^N\cdot G^m\right). \quad (2.151)$$

The cubes $\ell+Q_L$ appearing in (2.150) are also contained in Q_N. We can thus proceed as in Section 2.4.2 (Step 2) and get (similarly to (2.112)) that

$$\left|B^{L,N}_{2,\mathrm{short},\mathrm{bulk},\mathrm{int}}\right| \le C\,\frac{\ln L}{L}. \quad (2.152)$$

We next turn to (2.151). For the cubes $\ell+Q_L$ which are a subset of Q_N, we proceed as in Section 2.4.2. For the others, we again write (2.146), and note that ϕ_i^N is smooth on a neighborhood of the boundary of $(\ell+Q_L)\cap Q_N$ and on a neighborhood of the boundary of $R_\ell^{L,N}$. We can hence proceed as in Lemma 2.23 and deduce from (2.146) that

$$\left|\int_{\ell+Q_L}\nabla\phi_i^N\cdot G^j\right| \le \int_{\partial\left((\ell+Q_L)\cap Q_N\right)}\frac{C\,dy}{|y|^d} + \int_{\partial R_\ell^{L,N}}\frac{C\,dy}{|y|^d}.$$

Using the inclusions (2.132) and (2.135), we obtain that

$$\begin{aligned}
&\left|\int_{\ell+Q_L}\nabla\phi_i^N\cdot G^j\right| \\
&\le \int_{\partial(\ell+Q_L)}\frac{C\,dy}{|y|^d} + \int_{\partial Q_N}\frac{C\,dy}{|y|^d} + \sum_{i=1}^d \int_{\partial Q_{2N-L-|\ell_i|}}\frac{C\,dy}{|y|^d} \\
&\le \int_{\partial(\ell+Q_L)}\frac{C\,dy}{|y|^d} + \frac{C}{N} + \sum_{i=1}^d \frac{C}{2N-L-|\ell_i|} \\
&\le \int_{\partial(\ell+Q_L)}\frac{C\,dy}{|y|^d} + \frac{C}{N} + \frac{C}{2N-L-|\ell|_\infty}.
\end{aligned}$$

Inserting this estimate in (2.151), we thus obtain that

$$\begin{aligned}
&\left|B^{L,N}_{2,\mathrm{short},\mathrm{bulk},\mathrm{ext}}\right| \\
&\le \frac{C}{|Q_L|}\sum_{\substack{N-L<|\ell|_\infty\le \min(2L,N),\\ |\ell|_\infty\neq L,\,|\ell|_\infty\neq L+1}}\left[\left(\int_{\partial(\ell+Q_L)}\frac{dy}{|y|^d}\right)^2 + \frac{1}{N^2} + \frac{1}{(2N-L-|\ell|_\infty)^2}\right] \\
&\le \frac{C}{|Q_L|}\sum_{\substack{N-L<|\ell|_\infty\le 2L,\\ |\ell|_\infty\neq L,\,|\ell|_\infty\neq L+1}}\left(\int_{\partial(\ell+Q_L)}\frac{dy}{|y|^d}\right)^2 + \frac{C}{L^d\,N^2}\left(\big(\min(2L,N)\big)^d - (N-L)^d\right) \\
&\qquad + \frac{C}{L^d}\sum_{j=N-L+1}^{\min(2L,N)}\frac{j^{d-1}}{(2N-L-j)^2}.
\end{aligned} \quad (2.153)$$

The first term in the right-hand side of (2.153) is estimated as underneath (2.110). We hence obtain, similarly to (2.112), that it is bounded by $C\,(\ln L)/L$. The second term is

estimated by C/L^2, since we work in the regime $L < N < 3L$. To estimate the third term of the right-hand side of (2.153), we write that

$$\begin{aligned}
&\frac{C}{L^d}\sum_{j=N-L+1}^{\min(2L,N)}\frac{j^{d-1}}{(2N-L-j)^2}\\
&=\frac{C}{L^d}\sum_{j=2N-L-\min(2L,N)}^{N-1}\frac{(2N-L-j)^{d-1}}{j^2}\\
&=\frac{C}{L^d}\sum_{p=0}^{d-1}\begin{pmatrix} d-1\\ p\end{pmatrix}(2N-L)^{d-1-p}\,(-1)^p\sum_{j=2N-L-\min(2L,N)}^{N-1} j^{p-2}.
\end{aligned}$$

Bounding from above each sum in j, we get

$$\begin{aligned}
&\frac{C}{L^d}\sum_{j=N-L+1}^{\min(2L,N)}\frac{j^{d-1}}{(2N-L-j)^2}\\
&\le\frac{C}{L^d}\sum_{p=0}^{d-1}\begin{pmatrix} d-1\\ p\end{pmatrix}(2N-L)^{d-1-p}\sum_{j=2N-L-\min(2L,N)}^{N-1} j^{p-2}\\
&\le\frac{C}{L^d}(2N-L)^{d-1}\sum_{j=2N-L-\min(2L,N)}^{N-1} j^{-2}+\frac{C}{L^d}(d-1)(2N-L)^{d-2}\sum_{j=2N-L-\min(2L,N)}^{N-1} j^{-1}\\
&\qquad+\frac{C}{L^d}\sum_{p=2}^{d-1}\begin{pmatrix} d-1\\ p\end{pmatrix}(2N-L)^{d-1-p}\,N^{p-1}\\
&\le\frac{C}{L^d}(2N-L)^{d-1}+\frac{C}{L^d}(d-1)(2N-L)^{d-2}\,\ln N\\
&\qquad+\frac{C}{L^d\,N}\sum_{p=0}^{d-1}\begin{pmatrix} d-1\\ p\end{pmatrix}(2N-L)^{d-1-p}\,N^{p}\\
&\le\frac{C\,N^{d-1}}{L^d}+\frac{C\,N^{d-2}\,\ln N}{L^d}+\frac{C}{L^d\,N}(3N-L)^{d-1}\\
&\le\frac{C\,N^{d-1}}{L^d}+\frac{C\,N^{d-2}\,\ln N}{L^d}+\frac{C\,N^{d-2}}{L^d}.
\end{aligned}$$

Using that $L < N < 3L$, we bound from above this third term of the right-hand side of (2.153) by $C/L + C\,\ln L/L^2 + C/L^2 \le C/L$. Inserting the previous results in (2.153), we deduce that, when $L < N < 3L$,

$$\left|B^{L,N}_{2,\text{short,bulk,ext}}\right| \le \frac{C\,\ln L}{L}+\frac{C}{L^2}+\frac{C}{L}\le\frac{C\,\ln L}{L}. \tag{2.154}$$

Collecting (2.149), (2.152) and (2.154), we deduce that

$$\text{if } L<N<3L, \text{ then}\qquad \left|B^{L,N}_{2,\text{short,bulk}}\right|\le C\,\frac{\ln L}{L}. \tag{2.155}$$

Limit of $B^{L,N}_{2,\text{long}}$

For this term (2.142) not to be void, we assume that $N \ge 2L+1$. For the cubes $\ell + Q_L$ which are a subset of Q_N, we proceed as in Section 2.4.2. For the others, we again write (2.146), and note that ϕ_i^N is smooth on a neighborhood of the boundary of $(\ell +$

$Q_L) \cap Q_N$ and on a neighborhood of the boundary of $R_\ell^{L,N}$, because Q is at a positive distance away from these boundaries. We can hence proceed as in Lemma 2.24 and deduce from (2.146) that

$$\left| \int_{\ell+Q_L} \nabla \phi_i^N \cdot G^j \right| \leq \int_{\partial\left((\ell+Q_L)\cap Q_N\right)} \frac{C\,dy}{|y|^d} + \int_{\partial R_\ell^{L,N}} \frac{C\,dy}{|y|^d}.$$

Using the inclusions (2.132) and (2.135) and the properties (2.133) and (2.136), we obtain that

$$\begin{aligned}
&\left| \int_{\ell+Q_L} \nabla \phi_i^N \cdot G^j \right| \\
&\leq \int_{\partial(\ell+Q_L)} \frac{C\,dy}{|y|^d} + \int_{\partial_{\ell,L} Q_N} \frac{C\,dy}{|y|^d} + \sum_{i=1}^d \int_{\partial_{\ell,L,N} Q_{2N-L-|\ell_i|}} \frac{C\,dy}{|y|^d} \\
&\leq \int_{\partial(\ell+Q_L)} \frac{C\,dy}{|y|^d} + \frac{C}{N^d} \left|\partial_{\ell,L} Q_N\right| + \sum_{i=1}^d \frac{C}{(2N-L-|\ell_i|)^d} \left|\partial_{\ell,L,N} Q_{2N-L-|\ell_i|}\right| \\
&\leq \int_{\partial(\ell+Q_L)} \frac{C\,dy}{|y|^d} + \frac{C}{N^d} \left|\partial_{\ell,L} Q_N\right| + \sum_{i=1}^d \frac{C\,L^{d-1}}{(2N-L-|\ell_i|)^d} \\
&\leq \int_{\partial(\ell+Q_L)} \frac{C\,dy}{|y|^d} + \frac{C\,L^{d-1}}{N^d} + \frac{C\,L^{d-1}}{(2N-L-|\ell|_\infty)^d}.
\end{aligned}$$

We thus obtain that

$$\begin{aligned}
&\left| B_{2,\mathrm{long}}^{L,N} \right| \\
&\leq \frac{C}{|Q_L|} \sum_{2L<|\ell|_\infty \leq N} \left[\frac{L^{2d-2}}{N^{2d}} + \frac{L^{2d-2}}{(2N-L-|\ell|_\infty)^{2d}} + \left(\int_{\partial(\ell+Q_L)} \frac{dy}{|y|^d} \right)^2 \right] \\
&\leq \frac{C}{L^d} \frac{L^{2d-2}}{N^{2d}} \left(N^d - (2L)^d\right) + \frac{C\,L^{2d-2}}{L^d} \sum_{j=2L+1}^N \frac{j^{d-1}}{(2N-L-j)^{2d}} \\
&\qquad + \frac{C}{|Q_L|} \sum_{2L<|\ell|_\infty \leq N} \left(\int_{\partial(\ell+Q_L)} \frac{dy}{|y|^d} \right)^2. \qquad (2.156)
\end{aligned}$$

The first term of the above right-hand side is bounded by C/L^2, simply using that $L \leq N$. The third term is estimated by C/L^2, similarly to (2.104) in Section 2.4.2. Turning to the second term, we write that

$$\begin{aligned}
&\frac{C\,L^{2d-2}}{L^d} \sum_{j=2L+1}^N \frac{j^{d-1}}{(2N-L-j)^{2d}} \\
&= \frac{C\,L^{2d-2}}{L^d} \sum_{j=N-L}^{2N-3L-1} \frac{(2N-L-j)^{d-1}}{j^{2d}} \\
&= \frac{C\,L^{2d-2}}{L^d} \sum_{p=0}^{d-1} \binom{d-1}{p} (2N-L)^{d-1-p} (-1)^p \sum_{j=N-L}^{2N-3L-1} \frac{1}{j^{2d-p}}.
\end{aligned}$$

Bounding from above each sum in j, we get

$$\begin{aligned}
&\frac{C\,L^{2d-2}}{L^d}\sum_{j=2L+1}^{N}\frac{j^{d-1}}{(2N-L-j)^{2d}}\\
&\leq \frac{C\,L^{2d-2}}{L^d}\sum_{p=0}^{d-1}\left(\begin{array}{c} d-1 \\ p \end{array}\right)(2N-L)^{d-1-p}\,\frac{1}{(N-L)^{2d-p-1}}\\
&= \frac{C\,L^{2d-2}}{L^d}\sum_{p=0}^{d-1}\left(\begin{array}{c} d-1 \\ p \end{array}\right)\left(\frac{2N-L}{N-L}\right)^{d-1-p}\,\frac{1}{(N-L)^{d}}.
\end{aligned}$$

Using that using that $\dfrac{2N-L}{N-L}\leq 3$ and that $N-L\geq L$ (both bounds being consequences of the fact that $N\geq 2L+1$), we deduce that

$$\frac{C\,L^{2d-2}}{L^d}\sum_{j=2L+1}^{N}\frac{j^{d-1}}{(2N-L-j)^{2d}}\leq\frac{C\,L^{2d-2}}{L^d}\sum_{p=0}^{d-1}\left(\begin{array}{c} d-1 \\ p \end{array}\right)3^{d-1-p}\,\frac{1}{L^d}\leq\frac{C}{L^2}.$$

We thus infer from (2.156) that

$$\left|B_{2,\mathrm{long}}^{L,N}\right|\leq\frac{C}{L^2}. \tag{2.157}$$

Conclusion

Collecting (2.141), (2.157), (2.145), (2.147), (2.148) and (2.155), we obtain that

$$\left|B_{2}^{L,N}\right|\leq\frac{C}{L^2}+C\,\frac{(\ln L)^2}{L}+C\,\frac{\ln L}{L}\leq C\,\frac{(\ln L)^2}{L}.$$

This proves (2.118).

2.6 Numerical results

We now turn to numerical experiments. Our aim is twofold:

- investigate whether (2.33) holds in a fully random setting, i.e. beyond the weakly random setting considered in Theorem 2.7.
- investigate whether $\mathcal{Q}^{L,N}$ defined by (2.42) converges to $\mathcal{Q}$ in a fully random setting (recall that Theorem 2.10 states this convergence result in a weakly random setting).

The conclusions of the tests discussed below is that we indeed numerically observe, even for strongly random problems, that $\mathcal{Q}^{L,N}$ provides an accurate approximation (for large values of N and L) of the limit of the variance of $I_\varepsilon(f,g)$ when ε vanishes.

We now proceed in details. As explained above, we assume here that the randomness is *not* small. In line with the expected convergence (2.33), we compare the variance of $I_\varepsilon(f,g)$ (computed by a reference, computationally expensive method) with σ^2 defined by (2.34).

We have considered the classical two-dimensional random checkerboard test-case: the random matrix A in (2.1) is diagonal, $A(x,\omega)=a(x,\omega)\,\mathrm{Id}_d$, and the function a is

piecewise constant and takes the values $a = 0.2$ or $a = 1.8$ with equal probability $1/2$ (see Figure 2.2):

$$a(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)\, X_k(\omega)$$

where the random variables X_k are i.i.d. and satisfy $\mathbb{P}(X_k = 1.8) = \mathbb{P}(X_k = 0.2) = 1/2$. We consider (2.1) on $D = (0,1)^2$ for some given right-hand side f, and we fix some test function g in (2.22). The functions f and g will be made precise below.

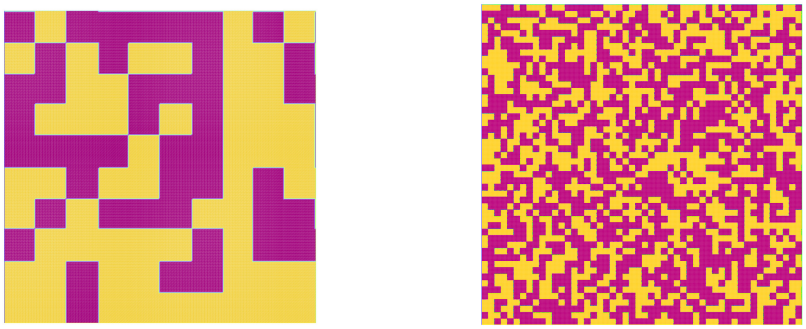


Figure 2.2: Two realizations of the checkerboard: $\varepsilon = 1/10$ (left) and $\varepsilon = 1/50$ (right).

The tensor $\mathcal{Q}^{L,N}$ is defined (see (2.42)) as a covariance. In practice, this covariance is approximated by an empirical mean: we hence define

$$\mathcal{Q}^{L,N,M}_{i,j,k,\ell} := \mathrm{Cov}_M\left(\frac{1}{|Q_L|}\int_{Q_L}\rho^N_{i,j}, \int_{Q_L}\rho^N_{k,\ell}\right), \tag{2.158}$$

where, for any random variables X and Y,

$$\mathrm{Cov}_M(X,Y) = \frac{1}{M}\sum_{m=1}^{M}(X_m - \overline{X})\,(Y_m - \overline{Y})$$

with $\overline{X} = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} X_m$, where $\{X_m\}_{1\le m\le M}$ are M i.i.d. realizations of the random variable $X(\omega)$ (and likewise for Y). The quantity $\mathcal{Q}^{L,N,M}$ can be computed in practice (see Figure 2.3 for a schematic representation of the approximation procedure).

In Section 2.6.1, we investigate the convergence of $\mathcal{Q}^{L,N,M}$ to some $\mathcal{Q}$ when L, N and M increase. In Section 2.6.2, we consider several choices of functions f and g. For each choice, we show that the law of $I_\varepsilon(f,g)$ indeed converges to a Gaussian law when $\varepsilon \to 0$, and we show that its variance can indeed be computed from $\mathcal{Q}$.

2.6.1 Approximation of $\mathcal{Q}$

We first investigate the convergence of $\mathcal{Q}^{L,N,M}$ when the size N of the truncated domain increases. To that aim, we fix $L = 5$ and $M = 10^4$ and we show one component of $\mathcal{Q}^{L,N,M}$ as a function of N on Figure 2.4 (the conclusions for the other components of the tensor is identical). We observe that the approximation $\mathcal{Q}^{L,N,M}$ quickly converges when we increase N, and that setting $N = 10$ is sufficient to reach the large N limit for the example we have considered.

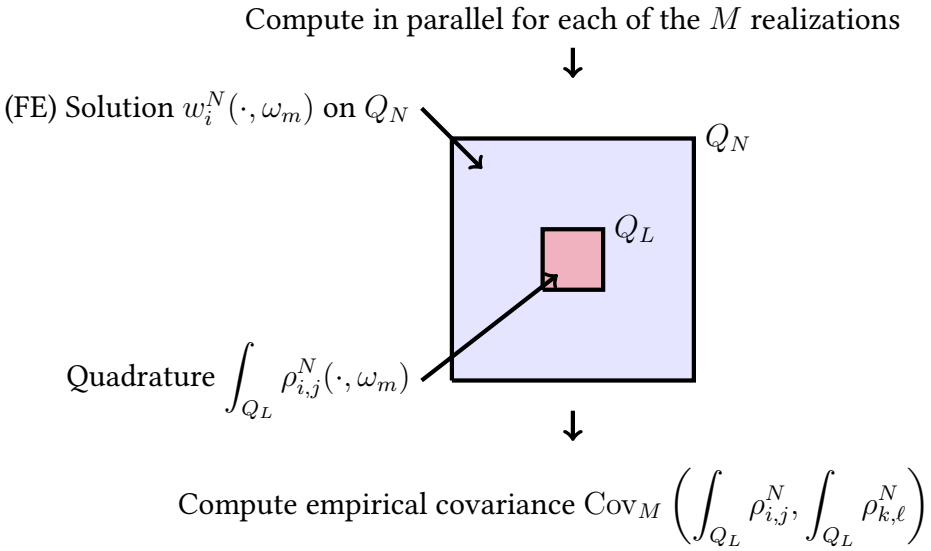


Figure 2.3: Procedure to approximate the tensor $\mathcal{Q}$ by $\mathcal{Q}^{L,N,M}$: $N \geq L$ denotes the size of the domain on which we compute the corrector, L denotes the size of the domain on which we consider the corrected energy density, while M denotes the number of realizations we consider to approximate the covariance.

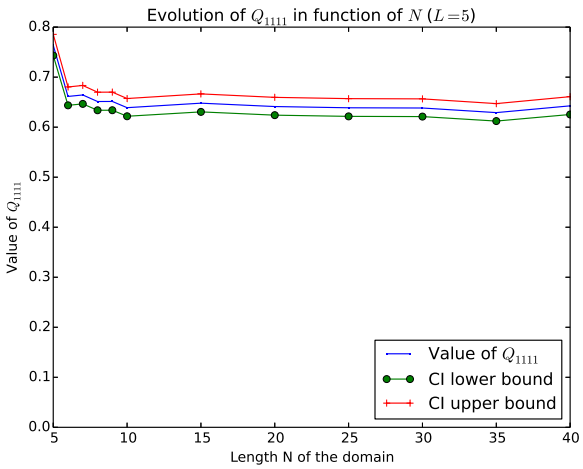


Figure 2.4: $\mathcal{Q}_{1111}^{L,N,M}$ as a function of N ($L = 5$ and $M = 10^4$). We also plot confidence intervals (CI) computed from the M realizations.

Remark 2.28. *The above computations have been performed with $M = 10^4$ realizations, which is a huge number. As a consequence, the confidence interval on Figure 2.4 is small. It is currently unclear to us how to reduce this number M of realizations (and thus the cost of the procedure) while keeping the same accuracy on the evaluation of the tensor $\mathcal{Q}$.*

We next perform the same study for $L = 10$ (with again $M = 10^4$ realizations). The results shown on Figure 2.5 again show that setting N to a small value (here $N = 15$) is sufficient to reach the large N limit.

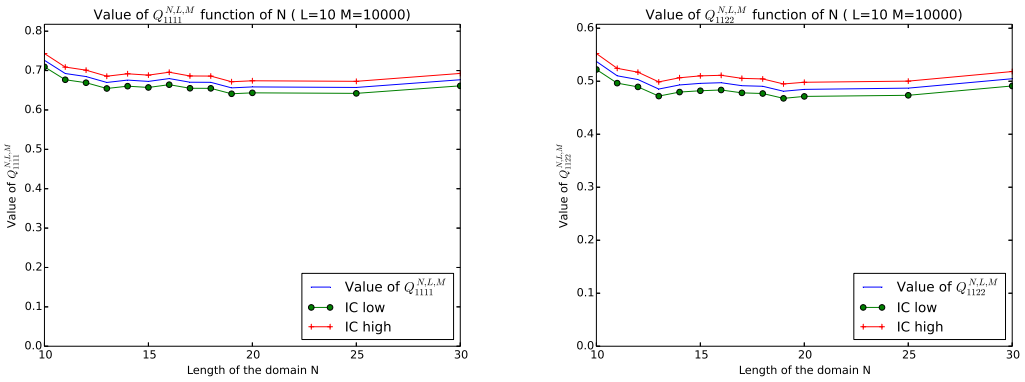


Figure 2.5: $\mathcal{Q}_{1111}^{L,N,M}$ (left) and $\mathcal{Q}_{1122}^{L,N,M}$ (right) as a function of N ($L = 10$ and $M = 10^4$).

The comparaison of the results of Figures 2.4 and 2.5 suggest to choose N slightly larger than L (since convergence is reached for $N \approx 10$, resp. $N \approx 15$, when $L = 5$, resp. $L = 10$). We thus consider the choice $N = L + 10$, again with $M = 10^4$ realizations and for increasing values of L, on Figure 2.6.

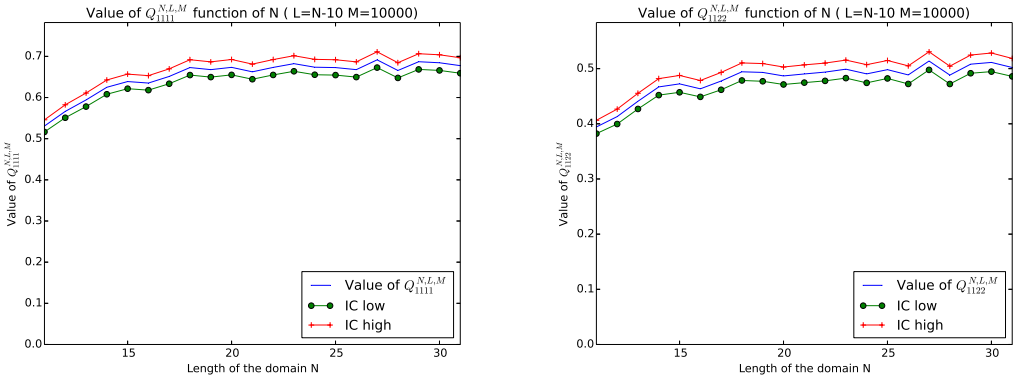


Figure 2.6: $\mathcal{Q}_{1111}^{N-10,N,M}$ (left) and $\mathcal{Q}_{1122}^{N-10,N,M}$ (right) as a function of N ($L = N - 10$ and $M = 10^4$).

Theorem 2.10 states a convergence in the regime $N > L$. As pointed out in Remark 2.11, the choice $N = L$ leads to a very particular situation in our weakly stochastic case. Our theoretical analysis thus does not provide insights on the behavior of $\mathcal{Q}^{L,N,M}$ when $N = L$. This motivates the numerical results shown on Figure 2.7, where we compare the evolution of $\mathcal{Q}^{L,N,M}$ as a function of N, in the case when $N = L$ and when $N = L + 10$. We can see that both approximations seem to converge to the same limit. Of course, for a fixed value of L and M, computing $\mathcal{Q}^{L,N,M}$ is cheaper in the case $N = L$ than in the case $N = L + 10$, since the corrector

problem (2.7) needs to be solved in a smaller domain. Making the choice $N = L$ hence seems to lead to the most efficient computations.

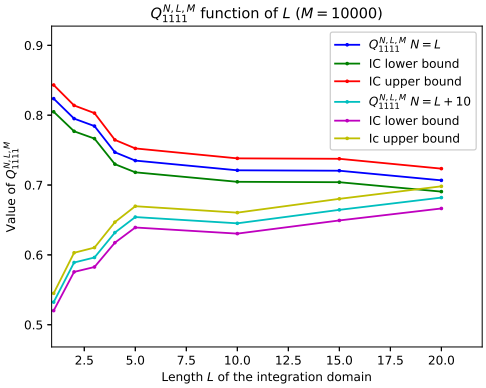


Figure 2.7: Evolution of $\mathcal{Q}^{L,L,M}_{1111}$ and $\mathcal{Q}^{L,L+10,M}_{1111}$ as a function of L ($M = 10^4$).

The fourth order tensor $\mathcal{Q}^{L,N,M}$ (and more precisely its limit when L, N and M go to ∞) is eventually used in (2.34) and (2.35) to compute the variance σ^2. We show on Figure 2.8 the approximation

$$(\sigma^{L,N,M})^2 = \int_D (\nabla u_\star \otimes \nabla v_\star) : \mathcal{Q}^{L,N,M} : (\nabla u_\star \otimes \nabla v_\star), \tag{2.159}$$

of σ^2. We have chosen $f(x, y) = 10e^{-80(x-0.5)^2}$ and $g(x, y) = 10e^{-80(y-0.5)^2}$. In practice, we approximate $u_\star$ in (2.159) by the solution to

$$-\text{div}\left(A^\star_N(\omega)\,\nabla u^N_\star(\omega)\right) = f \text{ in } D, \quad u^N_\star(\omega) = 0 \text{ on } \partial D,$$

where $A^\star_N(\omega)$ is defined by (2.6), and likewise for $v_\star$.

The quantity $(\sigma^{L,N,M})^2$ (which depends on all the components of the tensor $\mathcal{Q}^{L,N,M}$, and not only on one of them as shown on the above figures) again reaches its asymptotic limit for limited values of N and L. The confidence interval computed with $M = 10^4$ realizations is again small.

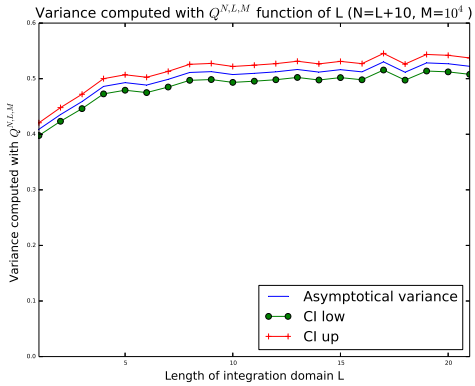


Figure 2.8: Variance $(\sigma^{L,N,M})^2$ computed from the tensor $\mathcal{Q}^{L,N,M}$ as a function of N ($L = N - 10$ and $M = 10^4$).

2.6.2 Estimation of the asymptotic law of $I_\varepsilon(f,g)$

In this section, we present numerical experiments that show that the theoretical results established in the weakly stochastic case (namely that the fourth order tensor $\mathcal{Q}^1$ characterizes the fluctuations of $I^1_\varepsilon(f,g)$) also hold (as least for the numerical examples considered here) in the full, non weakly random setting.

We proceed as follows:

1. We approximate $\mathcal{Q}$ by $\mathcal{Q}^{L,N,M}$ with $L=20$, $N=30$ and $M=10^4$.
2. We consider multiple choices of functions f and g.
3. For each choice of (f,g), we compute $\mathcal{M}$ realizations of $I_\varepsilon(f,g)$ for several values of ε. To do so, for each of these $\mathcal{M}$ realizations of A, we solve (2.1) (in practice, we use P1 finite elements and the software FreeFEM++ [50]).
4. We check whether $I_\varepsilon(f,g)$ is distributed according to a Gaussian law. To do so, we plot the empirical distribution of $I_\varepsilon(f,g)$ (computed from its $\mathcal{M}$ realizations), the associated QQ-plot and we perform a Shapiro-Wilk test [79].
5. If the Gaussian approximation is sufficiently accurate, we compute the empirical variance of $I_\varepsilon(f,g)$ and compare it with the asymptotic value $(\sigma^{L,N,M})^2$ defined by (2.159).

In what follows, we consider $\mathcal{M}=10^4$ realizations of $I_\varepsilon(f,g)$ for intermediate values of ε, and $\mathcal{M}=10^3$ realizations when ε becomes small. Recall indeed that the smaller ε is, the more expensive the computation of $I_\varepsilon(f,g)$ is (the meshsize we use to solve (2.1) is $h\simeq\varepsilon/10$).

In the following Sections 2.6.2 and 2.6.2, we consider two cases with various f and g, where the heterogeneous problem (2.1) is complemented with homogeneous Dirichlet boundary conditions. In Section 2.6.2, we explore the case when the heterogeneous problem (2.1) is complemented with homogeneous Neumann boundary conditions.

Case of a quantity of interest localized in $D_{f,g}\subset\subset D$

We start by considering some functions f and g (see Figure 2.9) so that the integrand of $I_\varepsilon(f,g)$ is (essentially) supported in a domain $D_{f,g}$ strictly included in D (the quantity $(u_\varepsilon(\cdot,\omega)-\mathbb{E}[u_\varepsilon])g$ thus essentially vanishes in a neighbourhood of ∂D). Such a choice is motivated by the fact that, in general, the solution u_ε to (2.1) (as always for heterogeneous problems) has a specific behaviour close to the boundary of D: presence of boundary layers, ... Our choice of (f,g) is aimed at making the impact of such particular features, occurring in a neighbourhood of ∂D, as small as possible.

First, we investigate whether the law of I_ε becomes closer to a Gaussian law when ε decreases. To do so, we plot the empirical distribution of I_ε on Figure 2.10 and the QQ-plot on Figure 2.11. We have considered several values of ε and only provide the results for the largest ($\varepsilon=1/10$) and the smallest ($\varepsilon=1/70$) values. We also perform a Shapiro-Wilk test with a 5% p-value.

When ε is not sufficiently small (e.g. $\varepsilon=1/10$), we can see that I_ε does not follow a Gaussian distribution, as expected. The QQ-plot shows that the extreme quantiles do not match with those of a normal distribution. This is confirmed by the Shapiro-Wilk test that rejects the hypothesis that the distribution of I_ε is Gaussian with a probability of 95%. On the other hand, when ε is small enough (here smaller than $1/70$), the distribution of I_ε is close to that of a Gaussian: the QQ-plot shows that the

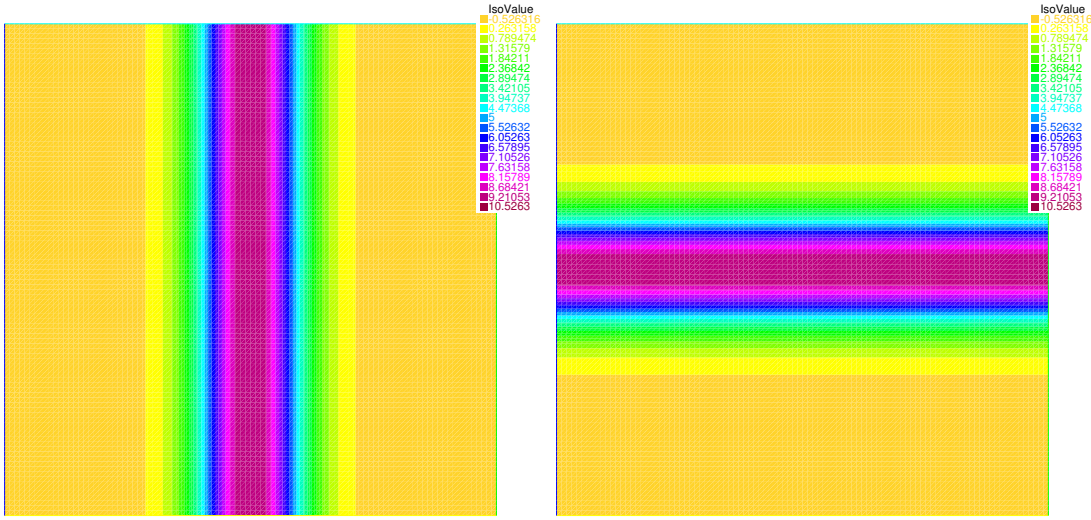


Figure 2.9: [Test case 1] Right-hand side $f(x,y) = 10e^{-80(x-0.5)^2}$ (left) and test function $g(x,y) = 10e^{-80(y-0.5)^2}$ (right).

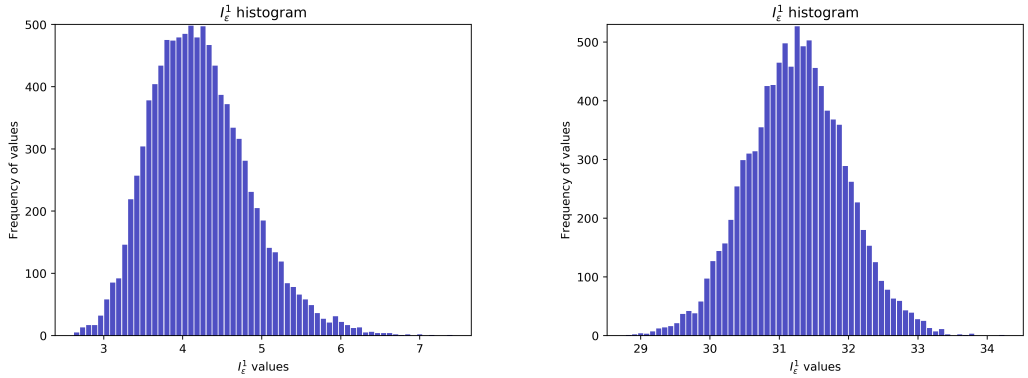


Figure 2.10: [Test case 1] Empirical distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations.

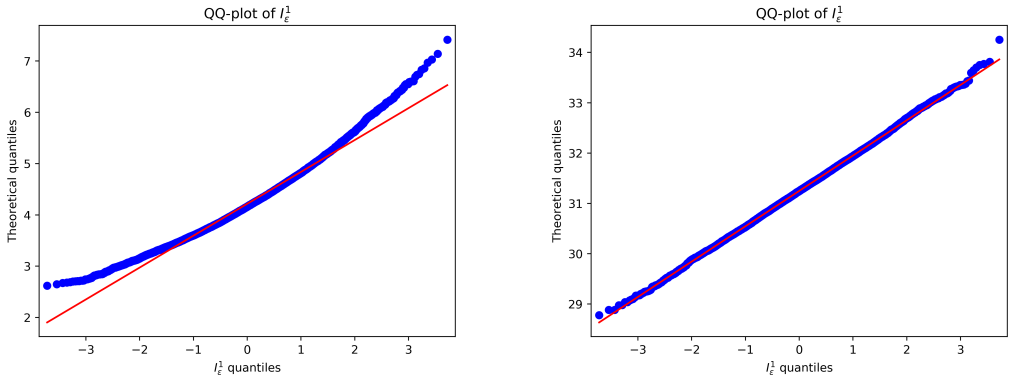


Figure 2.11: [Test case 1] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations.

quantiles of the distribution accurately match those of a Gaussian and the Shapiro-Wilk test cannot reject the Gaussian hypothesis with enough confidence.

We next compare the variance σ^2_{emp} of $I_\varepsilon(f,g)$ (computed empirically from $\mathcal{M}$ realizations of u_ε) with the asymptotic variance σ^2_{theo} (computed from $\mathcal{Q}^{L,N,M}$ with

$L = 20$, $N = 30$ and $M = 10^4$). Figure 2.12 shows that, when ε is not small enough (namely $1/40 \leq \varepsilon \leq 1/10$), the asymptotic variance σ^2_{theo} is an inaccurate approximation of the variance σ^2_{emp} of $I_\varepsilon(f,g)$: they differ by at least 10%. This is not unexpected, since our result only holds in the limit $\varepsilon \to 0$. On the other hand, when ε is small enough (say $\varepsilon < 1/40$), then the confidence intervals associated to both estimations are close to each other. For instance, when $\varepsilon = 1/70$, the two variances only differ by 5%. The relative error decreases when ε decreases. These numerical results hence seem to confirm that the empirical variance σ^2_{emp} converges, when $\varepsilon \to 0$, to the asymptotic variance computed from the tensor $\mathcal{Q}$.

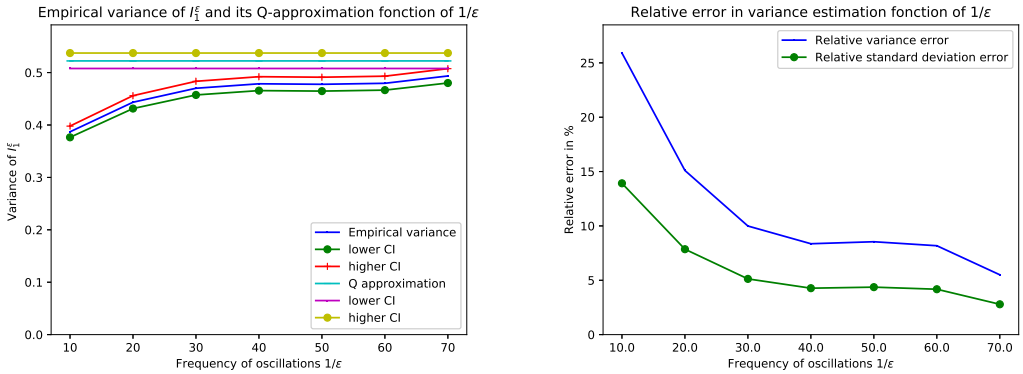


Figure 2.12: [Test case 1] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation.

Case when f or g are localized on the boundary of D

We now consider another choice for f and g (see Figure 2.13), so that the integrand in $I_\varepsilon(f,g)$ is not localized in a subdomain of D.

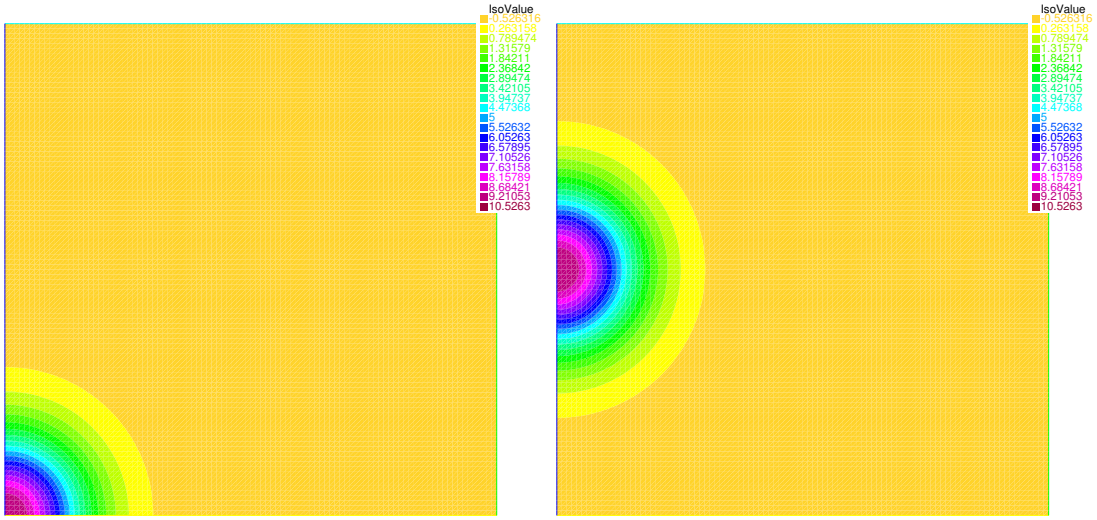


Figure 2.13: [Test case 2] Right-hand side $f(x,y) = 10e^{-40(x^2+y^2)}$ (left) and test function $g(x,y) = 10e^{-40(x^2+(y-0.5)^2)}$ (right).

As in Section 2.6.2, we first check whether the distribution of $I_\varepsilon(f,g)$ is close to a Gaussian distribution, for severak values of ε. To that aim, we plot the (empirically

computed) distribution of $I_\varepsilon(f,g)$ (see Figure 2.14) and the QQ-plot (see Figure 2.15). We also perform a Shapiro-Wilk test with a 5% p-value.

Again, for values of ε not small enough (e.g. $\varepsilon = 1/10$), the distribution of $I_\varepsilon(f,g)$ is not a Gaussian. This is obvious from the left-hand side of Figure 2.14, and quantitatively confirmed by the left-hand side of Figure 2.15, where we see on the QQ-plot that the extreme quantiles do not match with those of a normal distribution. The Shapiro-Wilk test rejects the hypothesis that the distribution of I_ε is Gaussian with a probability of 95%. On the other hand, when ε is small enough (here smaller than $1/70$), the distribution of I_ε is very close to a Gaussian distribution, as can be seen from the right-hand side of Figure 2.14. On the right-hand side of Figure 2.15, the QQ-plot shows that the quantiles of the distribution of I_ε are close to those of a Gaussian distribution (though the extreme quantiles are still slightly different). Somewhat unexpectedly, the Shapiro-Wilk test again rejects the Gaussian hypothesis, even for this small value of ε (the hypothesis is actually rejected for any ε between $1/10$ and $1/100$), although the test scores decreases when ε decreases (meaning that the rejection is performed with a smaller probability as ε decreases). The convergence towards a Gaussian distribution hence seems to be reached for smaller values of ε than in Section 2.6.2 (compare e.g. the right-hand sides of Figures 2.11 and 2.15).

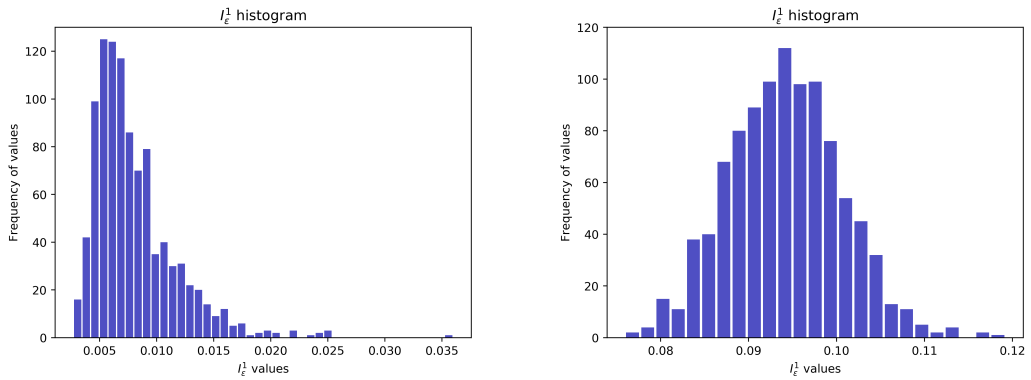


Figure 2.14: [Test case 2] Empirical distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/100$) computed from $\mathcal{M} = 10^3$ realizations.

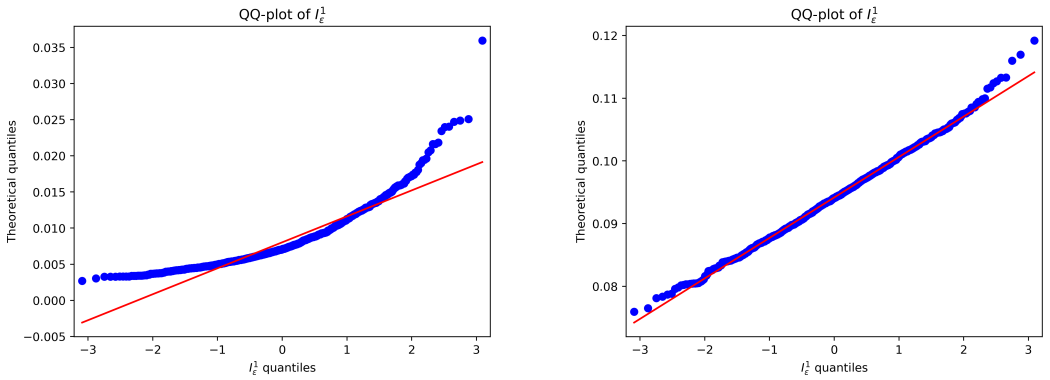


Figure 2.15: [Test case 2] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$: right: $\varepsilon = 1/100$) computed from $\mathcal{M} = 10^3$ realizations.

We next compare the variance σ^2_{emp} of $I_\varepsilon(f,g)$ (computed empirically from $\mathcal{M}$ realizations of u_ε) with the asymptotic variance σ^2_{theo} (computed from $\mathcal{Q}^{L,N,M}$ with

$L = 20$, $N = 30$ and $M = 10^4$). Figure 2.16 shows that, when ε is not small enough (here $\varepsilon \geq 1/70$), the asymptotic variance σ^2_{theo} is very different (say by 70%) from the variance σ^2_{emp} of $I_\varepsilon(f,g)$. On the other hand, when ε is sufficiently small (here $\varepsilon < 1/80$), then the confidence intervals associated to both estimations are close to each other. When $\varepsilon \leq 1/90$, the confidence intervals (in the estimation of σ^2_{emp} and σ^2_{theo}) have a non-empty overlap, yielding an accurate estimate of the variance (i.e. an estimate with a relative error smaller than 5%). The relative error between σ^2_{emp} and σ^2_{theo} decreases when ε decreases. These results seem to show that, when $\varepsilon \to 0$, the law of I_ε indeed converges to a Gaussian law and that the empirical variance σ^2_{emp} converges to the theoretically predicted asymptotic variance σ^2_{theo}, all convergences being reached for smaller values of ε than in the case of Section 2.6.2.

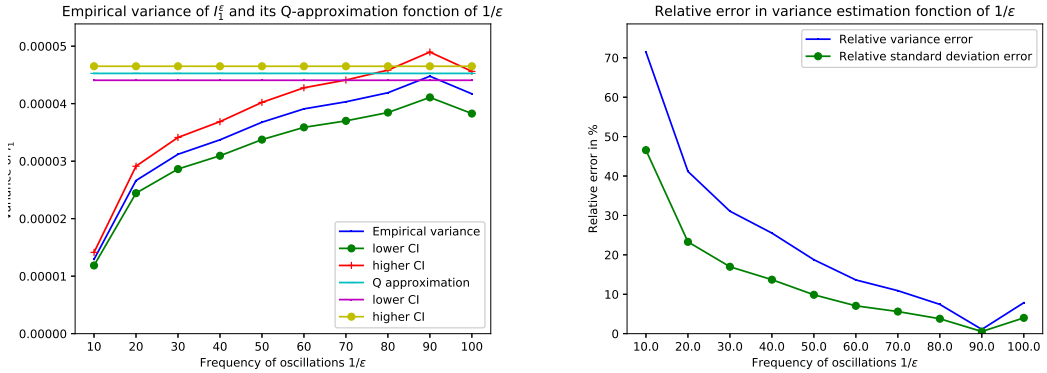


Figure 2.16: [Test case 2] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation.

The case of Neumann boundary conditions

We consider now the oscillatory problem (2.1) where the homogeneous Dirichlet boundary condition is replaced by a homogeneous Neumann boundary condition. We then of course need to consider functions f and g with vanishing mean, and we make the choice shown on Figure 2.17. The solution u_ε is only defined up to an additive (possibly random) constant, which is irrelevant in our quantity of interest $I_\varepsilon(f,g)$ since the mean of g vanishes.

We proceed as in Sections 2.6.2 and 2.6.2, first investigating whether the distribution of I_ε becomes closer to a Gaussian distribution when ε decreases (see Figures 2.18 and 2.19). We observe that the convergence seems to be reached very quickly (the threshold in terms of ε seems to be larger here than in the Dirichlet case). This is confirmed by the Shapiro-Wilk test (again with a 5% p-value), which does not reject the Gaussian hypothesis whenever $\varepsilon < 1/30$.

We next compare the variance σ^2_{emp} of $I_\varepsilon(f,g)$ (computed empirically from $\mathcal{M}$ realizations of u_ε) with the asymptotic variance σ^2_{theo} (computed from $\mathcal{Q}^{L,N,M}$ with $L = 20$, $N = 30$ and $M = 10^4$). Results are shown on Figure 2.20. They are very similar to the ones obtained in the Dirichlet case, and confirm the fact that the threshold ε_0 below which convergence is reached can be estimated at $\varepsilon_0 = 1/40$, a larger value than in the Dirichlet case.

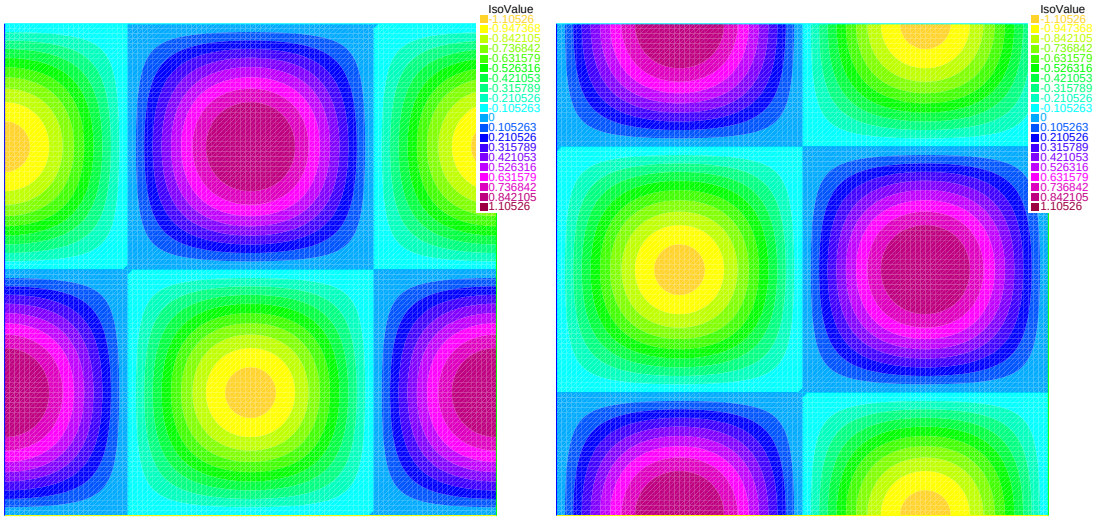


Figure 2.17: [Neumann test case] Right-hand side $f(x, y) = \cos(2\pi x)\sin(2\pi y)$ (left) and test function $g(x, y) = \cos(2\pi y)\sin(2\pi x)$ (right).

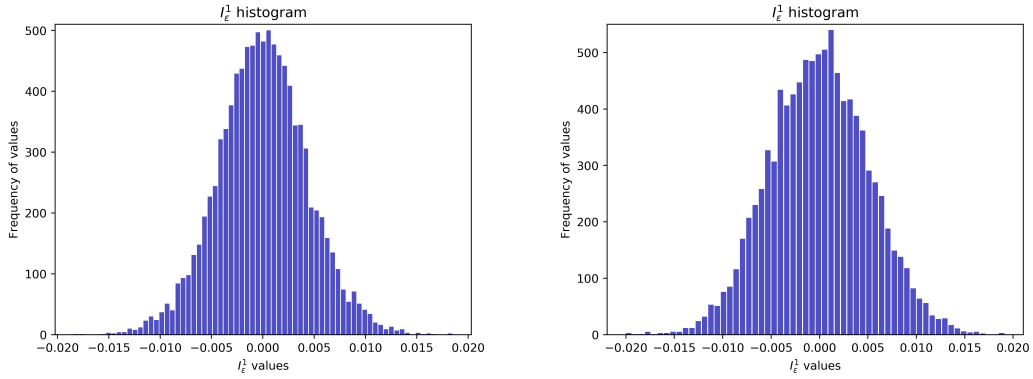


Figure 2.18: [Neumann test case] Empirical distributions of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations.

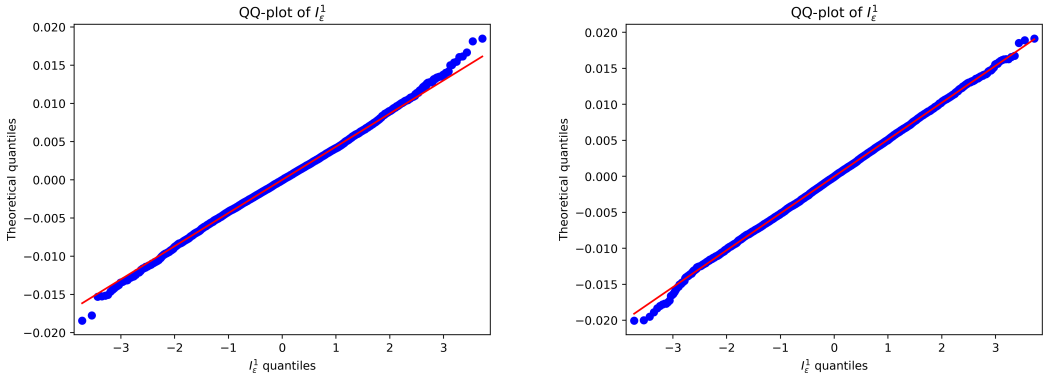


Figure 2.19: [Neumann test case] QQ-plot for the distribution of I_ε (left: $\varepsilon = 1/10$; right: $\varepsilon = 1/70$) computed from $\mathcal{M} = 10^4$ realizations.

Acknowledgements

We thank Claude Le Bris and Julian Fischer for several enlightening discussions on that work. We also had the opportunity to present an earlier version of this work

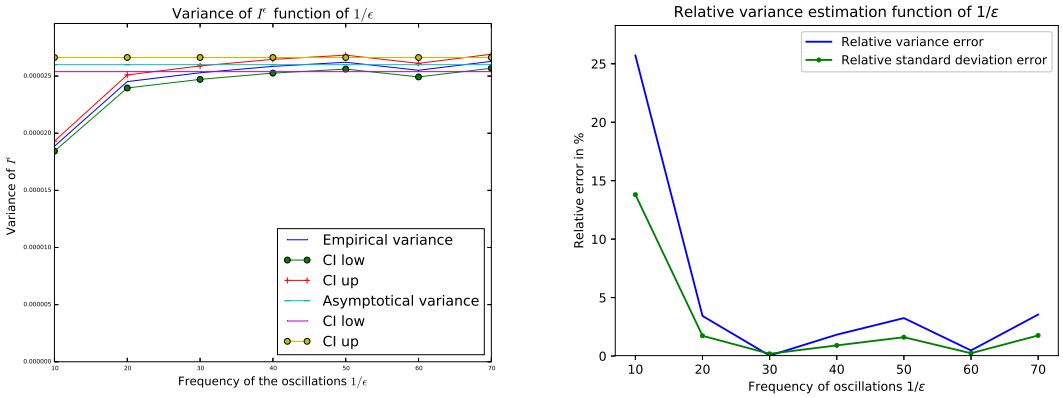


Figure 2.20: [Neumann test case] Left: comparison between the empirical variance σ^2_{emp} of I_ε and the variance σ^2_{theo} obtained with our approximation of $\mathcal{Q}$ in function of ε. Right, blue curve: relative error $|\sigma^2_{\text{emp}} - \sigma^2_{\text{theo}}|/\sigma^2_{\text{emp}}$ on the variance. Right, green curve: relative error $|\sigma_{\text{emp}} - \sigma_{\text{theo}}|/\sigma_{\text{emp}}$ on the standard deviation.

while visiting the MPI in Leipzig in April 2019, and we are grateful to Felix Otto for his kind invitation and the informative discussions we had there. The work of the authors is partially supported by ONR under grant N00014-15-1-2777 and by EOARD under grant FA9550-17-1-0294.

2.A Estimates for the Green function of the laplacian operator with periodic boundary conditions

Our aim is to show (2.120) and (2.121), in the case $d = 3$ and when $A_{\text{per}} = \text{Id}$. The Green function G_N is thus here the solution to

$$\begin{cases} -\Delta G_N(\cdot, y) = -\dfrac{1}{|Q_N|} + \displaystyle\sum_{k\in\mathbb{Z}^3} \delta(\cdot - y - Nk) \ \text{ in } \mathbb{R}^3, \\ G_N(\cdot, y) \text{ is } Q_N\text{-periodic.} \end{cases} \tag{2.160}$$

2.A.1 Analytical expression

The existence and uniqueness (up to the addition of a constant) of G_N solution to (2.160) is shown in [67] (see also [23, 29]). We recall the following result, which provides an analytic expression of G_N.

Lemma 2.29. *Consider Z^P_N defined by*

$$Z^P_N(x) = \sum_{k\in\mathbb{Z}^3,\ |k|\le P} f_N(x - Nk),$$

where

$$f_N(x) = \frac{1}{4\pi|x|} - \frac{1}{|Q_N|}\int_{Q_N} \frac{dy}{4\pi|x-y|}.$$

We have that $Z^P_N \in L^2_{\text{loc}}(\mathbb{R}^3)$ and that Z^P_N converges in $L^2_{\text{loc}}(\mathbb{R}^3)$, when $P \to \infty$, to some $\overline{G}_N$, which is a solution to (2.160) with $y = 0$.

Proof of Lemma 2.29. We first show that Z_N^P indeed converges in $L^2_{\mathrm{loc}}(\mathbb{R}^3)$, and we next prove that the limit is a solution to (2.160).

Step 1: Bound on f_N. We claim that there exists $C_f > 0$ independent of N such that

$$|f_N(x)| \leq \frac{C_f N^3}{|x|^4} \qquad \text{whenever } |x| \geq N. \tag{2.161}$$

We set $\rho(x) = \dfrac{1}{4\pi|x|}$. In dimension $d = 3$, we note that $Q = (-1/2, 1/2)^3 \subset B(0, r_\star)$ for some $r_\star < 1$. Thus, for any $z \in Q$ and any $r \in \mathbb{R}^3$ with $|r| = 1$, we write

$$\rho(r - z) = \rho(r) - z \cdot \nabla\rho(r) + \frac{1}{2} z \cdot \nabla^2\rho(r) z + R(r, z),$$

with

$$|R(r, z)| \leq C_{Taylor}|z|^3,$$

where C_{Taylor} is independent of r and z. For any $x \in \mathbb{R}^3$ with $|x| \geq N$, we then have

$$\begin{aligned}
&\int_{Q_N} \frac{dy}{4\pi|x - y|} \\
&= \frac{1}{|x|} \int_{Q_N} \rho\left(\frac{x}{|x|} - \frac{y}{|x|}\right) dy \\
&= \frac{1}{|x|}\left[\int_{Q_N} \rho\left(\frac{x}{|x|}\right) dy - \int_{Q_N} \frac{y}{|x|} \cdot \nabla\rho\left(\frac{x}{|x|}\right) dy \right. \\
&\qquad \left. + \int_{Q_N} \frac{1}{2}\frac{y}{|x|} \cdot \nabla^2\rho\left(\frac{x}{|x|}\right)\frac{y}{|x|}\, dy + \int_{Q_N} R\left(\frac{x}{|x|}, \frac{y}{|x|}\right) dy\right]
\end{aligned}$$

where we have used that $y/|x| \in Q$. We successively consider the different terms in the above right-hand side. The second term vanishes by symmetry. The third term reads

$$\begin{aligned}
\int_{Q_N} \frac{y}{|x|} \cdot \nabla^2\rho\left(\frac{x}{|x|}\right)\frac{y}{|x|}\, dy &= \sum_{i,j=1}^3 \int_{Q_N} \frac{y_i\, y_j}{|x|^2} \frac{\partial^2\rho}{\partial x_i \partial x_j}\left(\frac{x}{|x|}\right) dy \\
&= \sum_{i=1}^3 \frac{1}{|x|^2} \frac{\partial^2\rho}{\partial x_i^2}\left(\frac{x}{|x|}\right) \int_{Q_N} y_i^2\, dy \\
&= \frac{C_N}{|x|^2} \Delta\rho\left(\frac{x}{|x|}\right) \\
&= 0,
\end{aligned}$$

where we have noted $C_N = \displaystyle\int_{Q_N} y_1^2\, dy$. We thus deduce that

$$\frac{1}{|Q_N|}\int_{Q_N} \frac{dy}{4\pi|x - y|} = \frac{1}{4\pi|x|} + \frac{1}{|Q_N|}\frac{1}{|x|}\int_{Q_N} R\left(\frac{x}{|x|}, \frac{y}{|x|}\right) dy,$$

where the last term is estimated as

$$\left|\int_{Q_N} R\left(\frac{x}{|x|}, \frac{y}{|x|}\right) dy\right| \leq C_{Taylor} \int_{Q_N} \left|\frac{y}{|x|}\right|^3 dy \leq \frac{C}{|x|^3} N^6,$$

for some C independent of N. We thus deduce that $|f_N(x)| \leq \dfrac{C\,N^3}{|x|^4}$, which is the claim (2.161).

Step 2: Convergence of Z_N^P in $L^2_{\rm loc}(\mathbb{R}^3)$. The function $x \in \mathbb{R}^3 \mapsto \displaystyle\int_{Q_N} \frac{dy}{4\pi|x-y|}$ is continuous, hence f_N is continuous on $\mathbb{R}^3$ except at the origin, where it behaves like $1/(4\pi|x|)$. Hence $f_N \in L^2_{\rm loc}(\mathbb{R}^3)$ and thus $Z_N^P \in L^2_{\rm loc}(\mathbb{R}^3)$.

Choose $x \in \mathbb{R}^3$, $x \notin N\mathbb{Z}^3$. The terms $f_N(x-Nk)$ are all defined, and, in view of (2.161), we see that the series $\displaystyle\sum_{k\in\mathbb{Z}^3} f_N(x-Nk)$ is absolutely convergent. We can thus introduce $\overline{G}_N(x) = \displaystyle\sum_{k\in\mathbb{Z}^3} f_N(x-Nk)$, which is well-defined almost everywhere.

We now prove that Z_N^P converges in $L^2_{\rm loc}(\mathbb{R}^3)$ towards $\overline{G}_N$. To that aim, choose a compact $S \subset B(0,s)$ and consider the difference, for any $x \in S$,

$$r_P(x) = \left(Z_N^P(x) - \overline{G}_N(x)\right)^2 = \left(\sum_{|k|>P} f_N(x-Nk)\right)^2.$$

Using (2.161), we see that there exists $P_0 = s+1$ such that, when $P \geq P_0$ and for any $x \in S$, we have

$$r_P(x) \leq \left(\sum_{|k|>P} \frac{CN^3}{|x-Nk|^4}\right)^2 \leq \left(\sum_{|k|>P} \frac{CN^3}{(|k|N-s)^4}\right)^2 \leq \mathcal{C}_N < \infty,$$

where $\mathcal{C}_N$ is independent of x and P. We thus have that $\overline{G}_N \in L^2(S)$ and that Z_N^P converges in $L^2(S)$ towards $\overline{G}_N$.

Step 3: $\overline{G}_N$ is a solution to (2.160) with $y = 0$. Since Z_N^P converges in $L^2_{\rm loc}(\mathbb{R}^3)$ towards $\overline{G}_N$, we have that, in the sense of distributions on $\mathbb{R}^3$,

$$-\Delta\overline{G}_N = \lim_{P\to\infty} -\Delta Z_N^P = \lim_{P\to\infty} \sum_{k\in\mathbb{Z}^3,\ |k|\leq P} (-\Delta f_N)(\cdot - Nk).$$

Writing $f_N(x) = \dfrac{1}{4\pi|x|} - \dfrac{1}{|Q_N|}\displaystyle\int_{\mathbb{R}^3} 1_{Q_N}(y)\frac{dy}{4\pi|x-y|}$, we see that $-\Delta f_N = \delta_0 - \dfrac{1_{Q_N}}{|Q_N|}$. We therefore get

$$-\Delta\overline{G}_N = \lim_{P\to\infty} \sum_{k\in\mathbb{Z}^3,\ |k|\leq P} \left[-\frac{1_{Nk+Q_N}}{|Q_N|} + \delta_{Nk}\right] = -\frac{1}{|Q_N|} + \sum_{k\in\mathbb{Z}^3} \delta_{Nk}.$$

By construction, $\overline{G}_N$ is $N\mathbb{Z}^3$-periodic. We have thus shown that $\overline{G}_N$ is a solution to (2.160) with $y = 0$. This concludes the proof of Lemma 2.29. □

2.A.2 Estimates on G_N

Lemma 2.30. *Consider the function $\overline{G}_N$ built in Lemma 2.29. There exists C independent of N such that*

$$\textit{for any } x \in Q_N, \textit{ we have } \left|\overline{G}_N(x)\right| \leq \frac{C}{|x|}.$$

Proof of Lemma (2.30). We write that

$$\left|\overline{G}_N(x)\right| \leq \sum_{k\in\mathbb{Z}^3,\ |k|<2} |f_N(x-Nk)| + \sum_{k\in\mathbb{Z}^3,\ |k|\geq 2} |f_N(x-Nk)|. \tag{2.162}$$

We note that, when $|k|\geq 2$ and $x\in Q_N$, we have $|x-Nk|\geq N|k|-|x|\geq 2N-|x|\geq N$ where we have used that $Q_N\subset B(0,N)$. We are thus in position to use (2.161) in the second term of (2.162) and write

$$\sum_{k\in\mathbb{Z}^3,\ |k|\geq 2} |f_N(x-Nk)| \leq \sum_{k\in\mathbb{Z}^3,\ |k|\geq 2} \frac{C_f N^3}{|x-Nk|^4} \leq \sum_{k\in\mathbb{Z}^3,\ |k|\geq 2} \frac{C_f N^3}{N^4(|k|-1)^4} \leq \frac{C}{N},$$

where C is independent of N. For the first term of (2.162), we write

$$\sum_{k\in\mathbb{Z}^3,\ |k|<2} |f_N(x-Nk)| \leq C \sum_{k\in\mathbb{Z}^3,\ |k|<2} \left(\frac{1}{|x-Nk|} + \frac{1}{|Q_N|}\int_{Q_N} \frac{dy}{|x-Nk-y|}\right). \tag{2.163}$$

For the second term of (2.163), we use that there exists $\overline{r}$ independent of N such that, for any $x\in Q_N$ and any $|k|<2$, we have $Q_N\subset B(x-Nk,\overline{r}N)$. Hence

$$\begin{aligned}\sum_{k\in\mathbb{Z}^3,\ |k|<2} \frac{1}{|Q_N|}\int_{Q_N} \frac{dy}{|x-Nk-y|} &\leq \sum_{k\in\mathbb{Z}^3,\ |k|<2} \frac{1}{|Q_N|}\int_{B(x-Nk,\overline{r}N)} \frac{dy}{|x-Nk-y|}\\ &= \sum_{k\in\mathbb{Z}^3,\ |k|<2} \frac{1}{|Q_N|}\int_{B(0,\overline{r}N)} \frac{dy}{|y|}\\ &\leq \frac{C}{N}\end{aligned}$$

for some C independent of N. For the first term of (2.163), we write

$$\sum_{k\in\mathbb{Z}^3,\ |k|<2} \frac{1}{|x-Nk|} = \frac{1}{|x|} + \sum_{k\in\mathbb{Z}^3,\ 1\leq|k|<2} \frac{1}{|x-Nk|},$$

and we recall that $Q\subset B(0,r^\star)$ for some $r^\star<1$. Hence any $x\in Q_N$ and any $1\leq|k|<2$, we have $|x-Nk|\geq N|k|-|x|\geq N-r^\star N$, hence

$$\sum_{k\in\mathbb{Z}^3,\ |k|<2} \frac{1}{|x-Nk|} \leq \frac{1}{|x|} + \frac{C}{N}.$$

Collecting the above estimates, we deduce from (2.163) that

$$\sum_{k\in\mathbb{Z}^3,\ |k|<2} |f_N(x-Nk)| \leq \frac{1}{|x|} + \frac{C}{N}.$$

We then deduce from (2.162) that, for any $x\in Q_N$,

$$\left|\overline{G}_N(x)\right| \leq \frac{1}{|x|} + \frac{C}{N},$$

which yields the claimed result. This concludes the proof of Lemma 2.30. □

We recall (see Section 2.5.1) that, for any x and y in $\mathbb{R}^3$, we denote $d_N(x,y) = \inf_{k\in\mathbb{Z}^3} |x-y-Nk|$. It is easy to check that, for any x, y and z in $\mathbb{R}^3$, we have the triangular inequality

$$d_N(x,y) \leq d_N(x,z) + d_N(z,y). \tag{2.164}$$

Lemma 2.31. *Let G_N be a solution to (2.160). Then, for any x and y in $\mathbb{R}^3$, we have*

$$|\nabla_x G_N(x,y)| \leq \frac{C}{|d_N(x,y)|^2}, \tag{2.165}$$

$$|\nabla_y G_N(x,y)| \leq \frac{C}{|d_N(x,y)|^2}, \tag{2.166}$$

$$|\nabla_y \nabla_x G_N(x,y)| \leq \frac{C}{|d_N(x,y)|^3}, \tag{2.167}$$

where C is independent of N, x and y.

For the proof of Lemma 2.31, we need the following result from Avellaneda and Lin (see [4, Lemma 16]).

Lemma 2.32 (Lemma 16 of [4]). *Consider a $\mathbb{Z}^d$ periodic matrix field A, with $\lambda \leq A(x) \leq M$ a.e. on Y, for some constants $0 < \lambda < M$. We also assume that $A \in C^{0,\gamma}(Y)$ for some $0 < \gamma \leq 1$ with $\|A\|_{C^{0,\gamma}(Y)} \leq M$. Let $\delta > 0$, $r > 0$ and take $f \in L^{d+\delta}(B(0,r))$. Consider u_ε solution to*

$$-\operatorname{div}\left[A\left(\frac{\cdot}{\varepsilon}\right)\nabla u_\varepsilon\right] = f \quad \text{in } B(0,r),$$

and assume that $\|u_\varepsilon\|_{L^\infty(B(0,r))} < \infty$.

Then there exists a constant C depending only on λ, M, γ, d and δ such that, for any $\varepsilon > 0$ and any $r > 0$, we have

$$\|\nabla u_\varepsilon\|_{L^\infty(B(0,r/2))} \leq C\left(r^{-1}\|u_\varepsilon\|_{L^\infty(B(0,r))} + r^\mu \|f\|_{L^{d+\delta}(B(0,r))}\right),$$

where $\mu = 1 - d/(d+\delta)$.

Proof of Lemma 2.31. The function $\overline{G}_N$ built in Lemma (2.29) satisfies (2.160) with $y = 0$. Let $x \in Q_N$ and $r = |x|/2$. The ball $B(x,r)$ does not intersect $N\mathbb{Z}^3$. Thus, on $B(x,r)$, we have $-\Delta \overline{G}_N = -\dfrac{1}{|Q_N|}$. Using Lemma 2.32 and next the periodicity of $\overline{G}_N$ and Lemma 2.30, we write, for a constant C independent of N and x (and with $d = 3$), that

$$\begin{aligned}
\left\|\nabla \overline{G}_N\right\|_{L^\infty(B(x,r/2))} &\leq C\left(r^{-1}\left\|\overline{G}_N\right\|_{L^\infty(B(x,r))} + r^\mu \left\|\frac{1}{|Q_N|}\right\|_{L^{d+\delta}(B(x,r))}\right), \\
&\leq C\left(r^{-1}\left(\sup_{z\in B(x,r)} \frac{1}{d_N(z,0)}\right) + \frac{r^\mu}{|Q_N|}\, r^{d/(d+\delta)}\right), \\
&\leq C\left(r^{-1}\left(\sup_{z\in B(x,r)} \frac{1}{d_N(z,0)}\right) + \frac{r}{|Q_N|}\right).
\end{aligned}$$

We next observe that, for any $x \in Q_N$ and $z \in B(x,r)$, we have

$$\begin{aligned}
|x| = d_N(x,0) \leq d_N(x,z) + d_N(z,0) &\leq |x-z| + d_N(z,0) \\
&\leq r + d_N(z,0) = \frac{|x|}{2} + d_N(z,0),
\end{aligned}$$

hence $d_N(z,0) \geq r$ and thus

$$\begin{aligned}
\left\|\nabla \overline{G}_N\right\|_{L^\infty(B(x,r/2))} \leq C\left(\frac{1}{r^2} + \frac{r}{|Q_N|}\right) &\leq C\left(\frac{1}{|x|^2} + \frac{|x|}{|Q_N|}\right) \\
&\leq C\left(\frac{1}{|x|^2} + \frac{1}{N^2}\right) \leq \frac{C}{|x|^2}.
\end{aligned}$$

In view of the equation satisfied by $\overline{G}_N$, we know that $\overline{G}_N \in C^\infty(B(x, r/2))$. We thus deduce from the above that

$$\forall x \in Q_N, \qquad \left|\nabla \overline{G}_N(x)\right| \leq \frac{C}{|x|^2}. \tag{2.168}$$

Let G_N be a solution to (2.160). We know that the solution to (2.160) is unique, up to the addition of a constant. We thus have $G_N(x, y) = \overline{G}_N(x - y) + C_N(y)$ where $C_N(y)$ only depends on y and N, and thus $\nabla_x G_N(x, y) = \nabla \overline{G}_N(x - y)$. Consider x and y such that $x - y \in Q_N$. We have $d_N(x, y) = |x - y|$ and the bound (2.168) implies (2.165) for such x and y. The case $x - y \notin Q_N$ is obtained using the Q_N-periodicity of $G_N(\cdot, y)$.

The function $G_N^\star(x, y) = G_N(y, x)$ is the Green function associated to the adjoint operator. We thus deduce (2.166) from (2.165).

We now prove (2.167). Let $y \in \mathbb{R}^3$ and $x \in \mathbb{R}^3$ with $x \notin y + N\mathbb{Z}^3$. For any $r_0 < 3d_N(x, y)/4$, we have $-\Delta G_N(\cdot, y) = -\dfrac{1}{|Q_N|}$ on $B(x, r_0)$. We can differentiate with respect to y, which shows that $-\Delta \nabla_y G_N(\cdot, y) = 0$ on $B(x, r)$ with $r = d_N(x, y)/2$. Using again Lemma 2.32 and next (2.166), we write, for a constant C independent of N, x and y, that

$$\begin{aligned} \left\|\nabla_x \nabla_y G_N(\cdot, y)\right\|_{L^\infty(B(x, r/2))} &\leq \frac{C}{r} \left\|\nabla_y G_N(\cdot, y)\right\|_{L^\infty(B(x, r))}, \\ &\leq \frac{C}{r} \sup_{z \in B(x, r)} \frac{1}{|d_N(z, y)|^2}. \end{aligned}$$

Using the triangular inequality (2.164), we now write, for any $z \in B(x, r)$, that

$$2r = d_N(x, y) \leq d_N(x, z) + d_N(z, y) \leq |x - z| + d_N(z, y) \leq r + d_N(z, y),$$

thus

$$\left\|\nabla_x \nabla_y G_N(\cdot, y)\right\|_{L^\infty(B(x, r/2))} \leq \frac{C}{r^3} \leq \frac{C}{|d_N(x, y)|^3}.$$

This concludes the proof of (2.167) and thus that of Lemma 2.31. □

CHAPTER 3

A MULTI-SCALE FINITE ELEMENT APPROACH USING HIGH ORDER POLYNOMIALS

This chapter corresponds to a manuscript in preparation, co-authored with U. Hetmaniuk, C. Le Bris and F. Legoll.

We consider a variant of the classical MsFEM approach with enrichments based on Legendre polynomials, both in the bulk of mesh elements and on their interfaces. A convergence analysis of the approach is presented. Numerical tests show a significant reduction in the error in comparison to classical MsFEM approaches, at a limited additional off-line cost.

3.1 Introduction

We consider the problem

$$-\operatorname{div}(A\nabla u) = f \ \text{ in } D, \qquad u = 0 \ \text{ on } \partial D, \tag{3.1}$$

where D is a bounded polygonal domain in $\mathbb{R}^2$, f is a given right-hand side and the symmetric elliptic coefficient matrix A presents heterogeneities at very small scales compared with the characteristic size of D. Classical approximation techniques such as finite elements are known to poorly perform in such cases, unless the mesh size is taken (possibly prohibitively) small. Multiple alternative dedicated approaches have therefore been introduced. Among those, the multi-scale finite element method (henceforth abbreviated as MsFEM), introduced in [34, 55], uses a Galerkin approach of (3.1) on a pre-computed basis. The basis functions are obtained by solving *local* problems mimicking (3.1) at the scale of mesh elements, with carefully chosen right-hand sides and boundary conditions. The vanilla version of the approach, called *linear MsFEM*, uses as basis functions the solutions to these local problems, posed on each mesh element, with null right-hand sides and with the coarse P1 elements as Dirichlet boundary conditions. Various improvements of that version are possible. In particular, the so-called *oversampling* variant, which solves local problems on larger domains and restricts their solutions to the considered element, is very effective. The flip side is that the approach is not conformal and the size of the oversampling area must be carefully calibrated, which can be a delicate practical issue.

Our purpose here is to introduce and study a MsFEM method improved differently. It essentially elaborates upon the *Special Finite Element Method* introduced in [53] and fully analyzed in [52]. In that approach, the linear MsFEM basis is enriched with local

eigenvectors related to the scalar product associated with the variational formulation of (3.1). The approach is very effective but the resolution of eigenproblems for each element of the coarse discretization can prove computationally challenging, even for an off-line step. This is the reason why the approach we present here complements the linear MsFEM basis with enrichments that are not eigenvectors, but solutions of edge and bulk problems using polynomials either as boundary condition or right-hand side (see SECTION 3.2). As for the other MsFEM variants, all basis functions for such enrichments can be computed in parallel. One cannot indeed too much emphasize that, if the dogma of multi-scale approaches is to drastically reduce the on-line cost at the expense of an increase of the off-line cost, it might be the case for a large class of complex enough problems that the approach is doomed because of a prohibitively computationally expensive off-line stage. Another advantage of the approach presented here is that the classical Legendre interpolation results apply, allowing one to get rigorous *a priori* and *a posteriori* error estimates for the approach more easily. A similar approach has been introduced independently in [41], for the specific case of quadrangles, Legendre polynomials and Gauss-Lobatto quadratures. The approach shows promising results in time-domain acoustic-wave modeling: it compares well with reference solutions computed with the spectral finite element method. Our aim is to push further the approach by expanding it to triangular meshes and to provide a detailed convergence analysis, along with some theoretical tools for adaptivity. We emphasize that our method is both *local* and *conformal*: the support of the enrichment function is either the two elements associated with the edge when an edge is considered, or the one element itself when a bulk element is considered. Also, as said above, it is *fully parallel* in the off-line stage. In contrast, we mention that another, very interesting and efficient, line of thought is exemplified by the approach called *Localized Orthogonal Decomposition method* (LOD) introduced in [70]. There, the classical finite elements are enriched with corrector functions that are solutions to specific PDEs. These functions have global supports, however they turn out to rapidly decay away from the element considered. This property allows one to design an approximation space with functions solution to PDEs with smaller, truncated supports (typically of size of order $O(H \ln H)$). The associated error estimates in the energy norm are then independent of the scales of the heterogeneities.

We prove (see SECTION 3.3) that, with enough enrichments, we can get a convergence rate that does not depend on the oscillations of A. Moreover, numerical experiments (see SECTION 3.4) show that already a small number of enrichment functions significantly reduces the error. Our analysis applies to both quadrangular and triangular meshes, the latter being more flexible and allowing one to discretize more complex geometries than those accessible to quadrangular meshes. Furthermore, we propose an *a posteriori* estimator that can be used to locally adapt the level of enrichment.

The numerical experiments we present in SECTION 3.4 show that the proposed approach outperforms the linear MsFEM especially in the regime where $H \simeq \varepsilon$, allowing for results of comparable quality to those obtained using the Special Element Method, and is on par with non-conformal approaches such as the variant of MsFEM using oversampling at a reasonable additional computational cost. Numerical results also seemingly indicate that the *a posteriori* estimator we propose reproduces truly the *qualitative* trend of the error in energy norm.

3.2 Discretization approach

We define $(\mathcal{T}_H)_H$ a family of conforming partitions of D into a finite number of convex quadrilaterals (or triangles) with straight edges. The mesh is assumed conformal (there is no hanging nodes and each internal edge is shared by exactly two elements of the mesh) and regular in the following sense:

$$\begin{gathered}\text{for any element } K, \text{ there exists an affine transformation } F : \overline{K} \mapsto K, \\ \text{where } \overline{K} \text{ is the reference element (here the reference square or triangle),} \\ \text{such that} \qquad \|\nabla F\|_{L^\infty} \leq \gamma\, H \qquad \text{and} \qquad \|\nabla F^{-1}\|_{L^\infty} \leq \gamma\, H^{-1}, \\ \text{where } \gamma > 1 \text{ is a constant independent of } K \text{ and } H.\end{gathered} \tag{3.2}$$

In practice, the latter property is ensured using a mesh with quadrilateral (or triangular) elements with a minimum angle condition. We denote by Γ the interior skeleton, that is $\Gamma = (\cup_{K\in\mathcal{T}_H} \partial K) \setminus \partial D$.

The variational formulation of (3.1) is expressed using, for $u, v \in H^1_0(D)$, the bilinear form $a(u,v) = \displaystyle\int_D (\nabla v)^T A \nabla u$. The associated energy norm is denoted by $\|v\|_E = \sqrt{a(v,v)}$. Since A is symmetric, the unique solution u to (3.1) satisfies also

$$u = \underset{v\in H^1_0(D)}{\operatorname{argmin}} \left(\frac{1}{2} a(v,v) - \langle f, v\rangle_{L^2(D)} \right).$$

We denote by

$$V_B = \left\{ v \in H^1_0(D),\ \ v|_K \in H^1_0(K) \text{ for any } K \in \mathcal{T}_H \right\},$$

where the subscript B stands, understandably, for *bubbles*. We also define

$$V_\Gamma = \left\{ E_D\tau \in H^1_0(D),\ \ \tau \in H^{1/2}_{00}(\Gamma) \right\},$$

which is the subspace of energy minimizing extensions of trace functions on Γ, where the extension $E_D(\tau)$ solves the minimization problem $\inf\limits_{v\in H^1_0(D)} a(v,v)$ subject to $v|_\Gamma = \tau$, that is, in the weak sense,

$$\begin{cases} -\operatorname{div}\left(A\nabla(E_D\tau)\right) = 0 & \text{in } K, \text{ for any } K \in \mathcal{T}_H, \\ E_D\tau = \tau & \text{on } \Gamma, \\ E_D\tau = 0 & \text{on } \partial D. \end{cases} \tag{3.3}$$

The following orthogonal decomposition with respect to the scalar product $a(\cdot,\cdot)$ holds:

$$H^1_0(D) = V_B \oplus V_\Gamma. \tag{3.4}$$

The decomposition is orthogonal because of the definition of the energy-minimizing extension. Indeed, it holds that

$$\forall v_B \in V_B,\ \ \forall v_\Gamma \in V_\Gamma, \quad a(v_B, v_\Gamma) = 0,$$

by using the variational formulation of problem (3.3) with a test function in $H^1_0(K)$. Although not often stated in this form, property (3.4) is at the heart of the analysis and development of domain decomposition methods for elliptic partial differential equations [42, 77, 80] and modern component mode synthesis methods [10, 21].

Following the decomposition (3.4), the solution u can be uniquely expressed as $u = u_B + u_\Gamma$ with the bubble part $u_B = \underset{w \in V_B}{\operatorname{argmin}} \left(\frac{1}{2} a(w,w) - \langle f, w \rangle_{L^2(D)} \right)$ and the interface part $u_\Gamma = \underset{v \in V_\Gamma}{\operatorname{argmin}} \left(\frac{1}{2} a(v,v) - \langle f, v \rangle_{L^2(D)} \right)$.

In our approach, instead of approximating u directly, we approximate u_B and u_Γ separately. This splitting is motivated as follows. First, the decomposition (3.4) implies a natural splitting of the error. If we indeed consider a numerical approximation $(u_H, u_{B,H}, u_{\Gamma,H})$, the error in energy norm reads as $\|u - u_H\|_E^2 = \|u_B - u_{B,H}\|_E^2 + \|u_\Gamma - u_{\Gamma,H}\|_E^2$. Second, the analysis of the classical MsFEM suggests that the interface part is more difficult to approximate than the bubble part. Approximating u_B by $u_{B,H} = 0$ gives an energy error of order $O(H)$ (see (3.11) below). Moreover, u_B is the collection of solutions to independent local problems with homogeneous Dirichlet boundary conditions. Hence, u_B can be computed effectively in parallel by using a FE solver for the Dirichlet problems. However, for u_Γ, there is no decrease of the error with respect to H when one takes $u_{\Gamma,H} = 0$: the error is of order $O(1)$. Moreover, when approximating u_Γ by $u_{\Gamma,H} = u_{MsFEM-lin}$, which is the best approximation obtained when considering extensions of continuous and piecewise affine functions on Γ (corresponding to the linear MsFEM approximation introduced in [55]), then the error is of order $O(1)$ when H is close to the small scale ε. In essence, MsFEM type methods are directed towards finding the correct bulk solutions assuming a certain, unknown shape of the solution along the interfaces. The recent history of the development of this category of methods can be revisited as the quest to determine the "right" interface conditions.

Our approach designs two independent approximation spaces: (i) on the one hand, a space to approach u_B by solving problems similar to (3.1) though localized on the elements and with high order polynomials as right-hand sides and (ii) on the other hand, a space that approximates u_Γ with an harmonic lifting defined by (3.3) of high order polynomials. We now detail these two approximation spaces, which are denoted $V_{B,H,\{M_K\}}$ (resp. $V_{\Gamma,H,\{N_e\}}$) for the space V_B (resp. V_Γ), where $\{M_K\}$ (resp. $\{N_e\}$) is a set of positive integers associating a polynomial degree M_K (resp. N_e) to each element $K \in \mathcal{T}_H$ (resp. each edge $e \subset \Gamma$).

We first consider the bubble space V_B. For any element K, we choose a positive integer M_K and consider the space of polynomial functions on K of degree smaller or equal to M_K (by degree, we mean *total degree* if K is a triangle, and *partial degree* in each variable if K is a quadrangle). We denote by $\mathcal{N}_{M_K}$ the dimension of this space of polynomials and introduce a basis of this set, that we denote $\{P_i\}_{i=1,\dots,\mathcal{N}_{M_K}}$. For any $1 \leq i \leq \mathcal{N}_{M_K}$, we introduce the function $\phi^B_{K,i} \in H^1_0(K)$, which is supported in K, and which is the solution to

$$\phi^B_{K,i} = 0 \ \text{ on } \partial K \quad \text{and} \quad \forall\, v \in H^1_0(K), \ \int_K (\nabla v)^T A \nabla \phi^B_{K,i} = \int_K P_i\, v. \tag{3.5}$$

If K is a quadrangular element, we readily note that, in practice, P_i can be chosen as the polynomial that has value 1 at the i^{th} Gauss-Lobatto point and 0 at the other Gauss-Lobatto points within K. Note that we do not consider the case $M_K = 0$.

Then, we define the finite dimensional space

$$V_{B,H,\{M_K\}} = \operatorname{Span}\left\{ \phi^B_{K,i}, \ \ 1 \leq i \leq \mathcal{N}_{M_K}, \ \ K \in \mathcal{T}_H \right\} \subset V_B \tag{3.6}$$

and the approximation $u_{B,H,\{M_K\}} \in V_{B,H,\{M_K\}}$ of $u_B \in V_B$ as the solution to

$$\forall v_{B,H,\{M_K\}} \in V_{B,H,\{M_K\}}, \quad \int_D (\nabla v_{B,H,\{M_K\}})^T A \nabla u_{B,H,\{M_K\}} = \int_D f\, v_{B,H,\{M_K\}}.$$

Besides considering the above finite dimensional space (3.6), it is also possible to choose $V_{B,H,\{M_K\}} = \{0\}$, in which case u_B is approximated by $u_{B,H,\{M_K\}} = 0$ and we have $\|u_B - u_{B,H,\{M_K\}}\|_E \leq CH$ (see (3.11) below).

Remark 3.1. *We have mentioned above that, in the case of quadrangles, we choose polynomials P_i associated with the Gauss-Lobatto points. Indeed, such a choice makes the quadrature formulas (to compute the local integrals needed to assemble the stiffness matrix and the right-hand side) particularly simple, since P_i vanishes at almost every integration point. From a theoretical viewpoint, any choice of basis is of course possible.*

We now turn to the interface space V_Γ. For any interior edge e of the coarse mesh, we choose a positive integer N_e. For any $2 \leq k \leq N_e$, we define the edge enrichment function $\phi^\Gamma_{e,k}$, which is supported on the two elements sharing the edge e, and which satisfies (see FIGURE 3.1)

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla \phi^\Gamma_{e,k}) = 0 \ \text{ in } K, \\ \phi^\Gamma_{e,k} = P_k \ \text{ on } e, \\ \phi^\Gamma_{e,k} = 0 \ \text{ on } \partial K \setminus e, \end{cases} \tag{3.7}$$

where K is any of the two elements containing the edge e, and where P_k is a polynomial function of degree k that vanishes at the vertices of the edge e. Note, for the practice, that P_k is chosen to be the *boundary-adapted* k^{th} Legendre polynomial on the edge (see e.g. [22, Fig. 2.12 p. 83]).

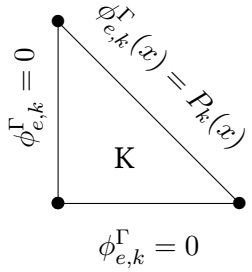


Figure 3.1: Local problem defining an edge enrichment

Formally, the cases $k = 0$ and $k = 1$ correspond to the linear MsFEM nodal basis functions associated with the two vertices of e. Denoting i_e and j_e these two vertices, we set $\phi^\Gamma_{e,0} = \phi^{MsFEM}_{i_e}$ and $\phi^\Gamma_{e,1} = \phi^{MsFEM}_{j_e}$, where ϕ^{MsFEM}_i is the solution on any element K to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla \phi^{MsFEM}_i) = 0 \ \text{ in } K, \\ \phi^{MsFEM}_i = \phi_i \ \text{ on } \partial K, \end{cases}$$

where ϕ_i is the nodal P1 Finite Element basis function associated with the vertex i. Note that the support of ϕ^{MsFEM}_i is the set of elements having i as a vertex.

We next define the finite dimensional space

$$\begin{aligned} V_{\Gamma,H,\{N_e\}} &= \operatorname{Span}\Big\{ \phi^{MsFEM}_j,\ 1 \leq j \leq Nb_{vertex}, \quad \phi^\Gamma_{e,k},\ 2 \leq k \leq N_e, e \subset \Gamma \Big\} \\ &= \operatorname{Span}\Big\{ \phi^\Gamma_{e,k},\ 0 \leq k \leq N_e, e \subset \Gamma \Big\}, \end{aligned} \tag{3.8}$$

which is a subset of V_Γ, and the approximation $u_{\Gamma,H,\{N_e\}} \in V_{\Gamma,H,\{N_e\}}$ of $u_\Gamma \in V_\Gamma$ as the solution to

$$\forall v_{\Gamma,H,\{N_e\}} \in V_{\Gamma,H,\{N_e\}}, \quad \int_D (\nabla v_{\Gamma,H,\{N_e\}})^T A \nabla u_{\Gamma,H,\{N_e\}} = \int_D f\, v_{\Gamma,H,\{N_e\}}.$$

Remark 3.2. *In order to approximate V_Γ, we decided to use liftings of polynomials defined on Γ. Such choice has been motivated by the versatility of polynomials (simplicity of implementation and good approximation properties). However, other choices could have been made. The main challenge here is to build an approximation space that accurately captures the oscillations of u_Γ on Γ. For instance, one can think of approaching such oscillating functions by a sinus basis with increasing frequencies like in Fourier approximation. It turns out that, if we enrich the MsFEM linear basis with liftings of $P_N(x) = \sin(\pi N x/H)$, then we get similar numerical results as with our polynomials. We have here chosen to work with polynomials because the derivation of approximation properties for a sinus basis proved to be more difficult than with a polynomial basis, for which we can rely on the extensive theory for polynomial approximation (see e.g. Lemma 3.18 below).*

Remark 3.3. *Note that the boundary condition which is imposed in (3.7) is continuous on ∂K, since we have considered polynomial functions P_k that vanish at the two ends of the edge e. Note that, if the boundary condition had some jumps on ∂K, then the problem (3.7) would be ill-posed in $H^1(K)$. We have chosen to work with the boundary-adapted Legendre polynomials, but other choices can be made, as long as the boundary conditions vanish at the two ends of the edge. Our specific choice is motivated by the fact that the boundary adapted Legendre polynomials are easy to compute (there is an explicit and simple recursion relation to compute their coefficients).*

Since we see our approach as an enrichment of the MsFEM linear method, the affine nodal functions $\phi_{x_i}^{MsFEM}$ must be part of the space which is spanned by our boundary conditions on ∂K. What matters for the analysis is that the space spanned by the boundary conditions on each edge e is the space of polynomial functions of degree smaller or equal to some N_e.

Denoting by $V_{H,\{M_K\},\{N_e\}} = V_{B,H,\{M_K\}} \oplus V_{\Gamma,H,\{N_e\}}$, our approximation $u_{H,\{M_K\},\{N_e\}}$ of u is defined by $u_{H,\{M_K\},\{N_e\}} = u_{B,H,\{M_K\}} + u_{\Gamma,H,\{N_e\}}$. Note that the choices $V_{B,H,\{M_K\}} = \{0\}$ and $N_e = 1$ for each edge e leads to an approximation space (and therefore a discrete solution) which is identical to the space used in the linear MsFEM approach.

The sets of positive integers $\{M_K,\ K \in \mathcal{T}_\mathcal{H}\}$ and $\{N_e,\ e \in \Gamma\}$ define the approximation space that is used in the variational problem. For the sake of clarity, in the case when we choose $M_K = M$ for any element $K \in \mathcal{T}_H$ (resp. $N_e = N$ for any edge $e \subset \Gamma$), we replace the notation $\{M_K\}$ by M (resp. $\{N_e\}$ by N).

We conclude this section by collecting several general remarks.

Remark 3.4. *Note that, although presented here in the context of the self-adjoint problem (3.1), the discretization procedure we just described can also be used in a case where the operator is* not *self-adjoint. However, the numerical analysis results that are established in the next section are, to date, restricted to the self-adjoint case.*

Remark 3.5. *In practice, one does not have access to the space $V_{H,M,N}$. Indeed, the enrichments $\phi_{e,i}^\Gamma$ or $\phi_{K,i}^B$ are solutions to local problems with no analytical expression. Usually, such functions are approximated by a finite element approach using a fine mesh of size h adapted to the characteristic length of variation of the diffusion coefficient A.*

Hence, in practice, for the numerical implementation, we use the space $V_{H,M,N,h} = V_{B,H,M,h} \oplus V_{\Gamma,H,N,h}$ *spanned by the functions* $\phi_{e,i}^{\Gamma,h}$ *and* $\phi_{K,i}^{B,h}$*, which are the approximation (on the mesh of size* h*) of* $\phi_{e,i}^{\Gamma}$ *or* $\phi_{K,i}^{B}$*. The study of the convergence of the approach with respect to the parameter* h *is outside of the scope of this article.*

Remark 3.6. *The construction of our basis, that is the offline stage, can be performed totally in parallel. Indeed, the basis functions for either the bubble or the interface approximation spaces are solutions to independent local problems. We notice that, in the special finite element approach introduced in [53] and analyzed in [52], the computation of the interface enrichments (which are eigenvectors of some operator associated with the edge* e*) requires solving a problem posed on the two elements sharing the edge* e*. In contrast, all our local problems are posed on a single element.*

In addition, the stiffness matrix and the right-hand side term for $f = 1$ *can be precomputed in parallel to further reduce the computational cost of the online stage.*

3.3 *A priori* and *a posteriori* estimates

Our global *a priori* error estimate reads as follows:

Proposition 3.7. *Assume that there exists* $0 < \alpha_{\min} \leq \alpha_{\max}$ *such that, almost everywhere in* D*, we have* $\alpha_{\min} |\xi|^2 \leq A(x)\xi \cdot \xi \leq \alpha_{\max} |\xi^2|$ *for all* $\xi \in \mathbb{R}^2$*. We also assume that the solution* u *to (3.1) belongs to* $H_0^1(D) \cap H^s(D)$ *for some* $s > 3/2$ *and that the right hand side* f *belongs to* $H^\ell(D)$ *for some* $\ell \geq 0$*. We consider our MsFEM approach in the case when* $M_K = M$ *for all elements* K *and* $N_e = N$ *for all edges* e*, for some* $M, N \in \mathbb{N}^\star$*. We then have*

$$\|u - u_{H,M,N}\|_E \leq \frac{C}{\sqrt{\alpha_{\min}}} \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} \|f\|_{H^\ell(D)} + C\sqrt{\alpha_{\max}} \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)} \quad (3.9)$$

where the constant C *is independent of* H*,* M*,* N*,* A*,* u *and* f *(but depends on* ℓ *and* s*). The above estimate holds both if we use triangular or quadrangular elements.*

In the case when no bubble enrichments are used (that is when $V_{B,H,M} = \{0\}$*), we have*

$$\|u - u_{H,M,N}\|_E \leq \frac{C}{\sqrt{\alpha_{\min}}} H \|f\|_{L^2(D)} + \sqrt{\alpha_{\max}} \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)}$$

where the constant C *is again independent of* H*,* N*,* A*,* u *and* f *(but depends on* s*).*

Some remarks are in order.

We note that, for f only in $L^2(D)$ (that is $\ell = 0$), increasing the polynomial degree M decreases the error at a rate $O(1/M)$. When f is a more regular function, the error decreases with respect to M with a better rate.

We next discuss the classical case of a rescaled periodic matrix coefficient (that is $A(x) = A_{\text{per}}(x/\varepsilon)$ for some $\mathbb{Z}^d$-periodic matrix A_{per}) and a coarse mesh of size $H \approx \varepsilon$. In that regime, it is well known that the classical linear MsFEM approach suffers from an error which does not decrease when H and ε tend to 0. In contrast, it is possible in our approach to increase N in order to still have a converging approximation. It is indeed expected that $|u|_{H^s(D)} \approx O(\varepsilon^{1-s})$. Choosing N of the order of $1/\varepsilon$ thus allows to keep a small error.

Note finally that the efficiency of our approach sensitively depends on the regularity of u *and* on the norm of its derivatives. On the optimistic side, this implies that the more regular u is, the more efficient our approach is. This unfortunately also means, on the pessimistic side, that the more oscillatory the solution is, the larger the norm of the derivatives of the solution is and thus the larger N has to be taken to obtain a given accuracy. In this respect, the LOD method [70] is way more robust, since the accuracy only depends on H, f and the contrast of A but neither on the regularity nor on the scale of the oscillations (although this is obtained at the price of computing "not so" local solutions elsewhere than in the given element).

Remark 3.8. *In the periodic case $A(x) = A_{\text{per}}(x/\varepsilon)$ for some $\mathbb{Z}^d$ periodic matrix A_{per}, we typically have that $\|u\|_{H^s(D)}$ is of the order of $1/\varepsilon^{s-1}$. In such a case, for given H, M and N, the right-hand side in the error estimate (3.9) goes to ∞ when ε goes to 0. However, the actual error does not blow up. Recall indeed that our approximation space $V_{H,M,N}$ contains the linear MsFEM approximation space, for which the estimate $\|u - u_{MsFEM-lin}\|_E \leq C\left(H + \sqrt{\varepsilon} + \sqrt{\varepsilon/H}\right)$ holds. The error in our approach being smaller that the linear MsFEM error, we hence get that our approximation does not blow up when ε goes to 0 and H, M and N are fixed.*

This observation raises the question of deriving a better estimate for our approach in the periodic case. It turns out that following the classical proof for estimating the MsFEM error (see e.g. [56]) yields the same estimate for our approach as the one obtained for the linear MsFEM approach. Indeed, in [56], an homogenization argument is used, which yields the contribution of the order of $\sqrt{\varepsilon}$ and $\sqrt{\varepsilon/H}$ in the error. This step is independent of the numerical approximation scheme used in the method and thus cannot be expected to be improved in our approach. Therefore, finding a sharper estimate in the periodic case for our approach does not seem to be an easy task, even when using periodic homogenization arguments.

The proof of PROPOSITION 3.7 is a direct consequence of the orthogonal decomposition (3.4) and the following LEMMA 3.9 and LEMMA 3.10, which respectively address the bubble approximation and the interface approximation:

Lemma 3.9. *Assume that there exists $0 < \alpha_{\min} \leq \alpha_{\max}$ such that, almost everywhere in D, we have $\alpha_{\min}|\xi|^2 \leq A(x)\xi \cdot \xi \leq \alpha_{\max}|\xi^2|$ for all $\xi \in \mathbb{R}^2$. We also assume that $f \in H^\ell(D)$. In the case when $M \geq 1$, the components u_B and $u_{B,H,M}$ satisfy*

$$\|u_B - u_{B,H,M}\|_E \leq \frac{C_\ell}{\sqrt{\alpha_{\min}}} \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} \|f\|_{H^\ell(D)} \tag{3.10}$$

for some C_ℓ independent of H, M, A and f. If $V_{B,H,M} = \{0\}$, then

$$\|u_B - u_{B,H,M}\|_E \leq \frac{C}{\sqrt{\alpha_{\min}}} H \|f\|_{L^2(D)} \tag{3.11}$$

for some universal constant C.

Lemma 3.10. *Assume that there exists $0 < \alpha_{\min} \leq \alpha_{\max}$ such that, almost everywhere in D, we have $\alpha_{\min}|\xi|^2 \leq A(x)\xi \cdot \xi \leq \alpha_{\max}|\xi^2|$ for all $\xi \in \mathbb{R}^2$. We also assume that the solution u to (3.1) belongs to $H^1_0(D) \cap H^s(D)$ for some $s > 3/2$. Then, in the case of a quadrangular mesh, the components u_Γ and $u_{\Gamma,H,N}$ satisfy*

$$\|u_\Gamma - u_{\Gamma,H,N}\|_E \leq C_s\sqrt{\alpha_{\max}} \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)}, \tag{3.12}$$

where the constant C_s is independent of H, N, A and u. The same estimate holds true for a triangular mesh.

In sharp contrast to the estimate (3.10) which does not depend on the oscillations of A, the estimate (3.12) depends on the norm of derivatives of u, hence, indirectly on the oscillations of A. As expected, the interface component is more difficult to capture than the bubble component.

Remark 3.11. *The assumption that the solution u belongs to $H^s(D)$ for some $s > 3/2$ can be relaxed in the case of quadrangular elements. Indeed, in that specific case, we perform below the proof of* LEMMA *3.10 using polynomial interpolation results at Gauss-Lobatto points, which only require the continuity of u on Γ. This continuity is satisfied as soon as $u \in H^s(D)$ for $s > d/2$, that is $s > 1$ in the case when $d = 2$. In the case of quadrangular elements, we hence only require that $u \in H^s(D)$ for some $s > 1$.*

Polynomial interpolation results are more difficult to obtain in the case of triangular elements because triangles are not cartesian product domains. Recall indeed that, in the reference square, the Gauss-Lobatto points are located at (x_k, y_k), where x_k are the Gauss-Lobatto points of the segment $[0,1] \times \{0\}$ and y_k are the Gauss-Lobatto points of the segment $\{0\} \times [0,1]$. This construction can obviously not be extended to the case of triangles. Choosing relevant interpolation points for triangles is still an open problem (a possible choice is that of Fekete points, see [76]). Hence, to recover similar interpolation properties, it can be expected that stronger constraints (namely $u \in H^s(D)$ for some $s > 3/2$ rather than $s > 1$) are required.

An estimate similar to (3.12) is obtained for the Special Finite Element Method in [52]. The estimate then depends on the k^{th} largest eigenvalue λ_e^k for the associated edge eigenproblem. The rate of decrease of λ_e^k with respect to k and H is not known, although the numerical experiments suggest it is $O(k/H)$, which would give a similar estimate as in LEMMA 3.10.

The proof of LEMMA 3.9 and LEMMA 3.10 essentially follows, and it is not unexpected, the pattern of the proof of the classical Céa's LEMMA. The best approximation is estimated using the Legendre projection (for LEMMA 3.9) or the Legendre interpolant on the bulk and the lifting of the interpolant along the edges (for LEMMA 3.10). Some technicalities arise for LEMMA 3.10 in the case of triangular meshes and an alternative proof using hp-Finite Element methods must be used. This alternative proof also extends for the quadrangular case.

We now turn to our *a posteriori* estimator. In contrast to our *a priori* estimates above, we now consider the general case when the polynomial degrees N_e (resp. M_K) associated to each edge e (resp. each element K) can be different. For some technical reasons (in particular due to the use of Scott-Zhang interpolation results, see LEMMA 3.23), we assume that the polynomial degrees of the edges are comparable on neighbouring edges, in the sense that

$$\forall e, e' \in \Gamma \text{ s.t. } \overline{e} \cap \overline{e'} \neq \emptyset, \quad \frac{N_e}{\sqrt{\gamma}} \leq N_{e'} \leq \sqrt{\gamma}\, N_e, \tag{3.13}$$

where γ is the mesh regularity constant of (3.2).

Proposition 3.12. *We assume that the diffusion coefficient matrix is of the form $A(x) = a(x)\, I$ for some scalar-valued function $a \in C^1(\overline{D})$ satisfying $\alpha_{\min} \leq a(x) \leq \alpha_{\max}$ almost everywhere in D, for some $0 < \alpha_{\min} \leq \alpha_{\max}$. We also assume that the solution u to (3.1) belongs to $H_0^1(D) \cap H^s(D)$ for some $s > 3/2$ and that, for any element K of the coarse mesh, $f \in H^{\ell_K}(K)$ for some $\ell_K \geq 0$.*

Consider the MsFEM approach on the discrete space $V_{H,\{M_K\},\{N_e\}}$, where $M_K > 0$ is the maximal degree of the polynomial functions used as right-hand sides for the bubble

basis functions in the element K, and $N_e > 0$ is the maximal degree of the polynomial functions used as boundary conditions for the interface basis functions associated to the edge e. We assume that the degrees $\{N_e\}$ satisfy (3.13).

The discrete solution $u_{H,\{M_K\},\{N_e\}}$ satisfies the a posteriori *estimate*

$$\begin{aligned}
&\|u - u_{H,\{M_K\},\{N_e\}}\|_E \\
&\leq C_{s,A} \left\{ \sum_{K\in\mathcal{T}_H} H_K^2 \, \frac{H_K^{\min(\ell_K, M_K+1)}}{M_K^{\ell_K}} \, \|f + \operatorname{div}\left(A\nabla u_{B,H,\{M_K\}}\right)\|_{L^2(K)} \, \|f\|_{H^{\ell_K}(K)} \right. \\
&\left. + \sum_{K\in\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e} \right) + \sum_{e\subset\Gamma} \frac{H_e}{p_e} \left\| J_e\left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|^2_{L^2(e)} \right\}^{1/2}
\end{aligned} \tag{3.14}$$

where $J_e(\psi)$ denotes the jump of a given function ψ across the edge e, and ν is the normal vector to the edge. In the above estimate, we have set $p_e = \min\{N_{\tilde{e}} \mid \tilde{e} \subset \partial K_e^1 \cup \partial K_e^2\}$ where K_e^1 and K_e^2 are the two elements sharing the edge e. The constant $C_{s,A}$ depends only on s (i.e. the regularity of u) and the diffusion coefficient A.

When no bubble enrichments are added, that is when $V_{B,H,\{M_K\}} = \{0\}$, we have the estimate

$$\begin{aligned}
\|u - u_{H,\{M_K\},\{N_e\}}\|_E \leq C_{s,A} &\left\{ \sum_{K\in\mathcal{T}_H} H_K^2 \, \|f\|^2_{L^2(K)} \right. \\
&\left. + \sum_{K\in\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e} \right) + \sum_{e\subset\Gamma} \frac{H_e}{p_e} \left\| J_e\left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|^2_{L^2(e)} \right\}^{1/2}.
\end{aligned} \tag{3.15}$$

Several remarks are in order.

The right-hand side of (3.14) actually defines an error indicator: the actual error is bounded from above by the product of the indicator times a constant independent of H_K, M_k and N_e.

The proof of PROPOSITION 3.12 follows the analogous proof performed for the Special Element Method in [52]. However, Scott-Zhang type polynomial interpolation approaches have to be introduced instead of classical polynomial interpolation. The restriction to scalar-valued diffusion coefficients comes from our use of LEMMA 3.22 below. The above estimate most probably also holds true in the case of a matrix-valued coefficient.

Some illustrations of the behavior of the *a posteriori* estimator are presented in SECTION 3.4.

3.4 Numerical experiments

This section is divided into two parts. We first compare our approach to standard MsFEM approximations (linear MsFEM and oversampling MSfEM approaches). Second, we investigate the performance of the *a posteriori* estimator proposed in PROPOSITION 3.12.

3.4.1 Comparison with other MsFEM approaches

In our numerical experiments, the emphasis is put on the enrichment by edge functions. As already mentioned above, the bubble error when no enrichments are used behaves like classical FE estimates for the Poisson problem: it decreases linearly with

respect to H, with a prefactor which only depends on the L^2 norm of the right-hand side and the coercivity constant of the diffusion coefficient A. In contrast, the interface error depends on the oscillations of A and has a more intricate behaviour. Moreover, in the classical MsFEM approaches (linear and oversampling), the basis functions belong to V_Γ. Hence, such approaches can also be enriched by bubble elements. In order to compare their respective effectiveness, it thus appears that it is best not to consider bubble enrichments. We therefore only act on V_Γ and no bubble enrichment is used, that is we keep $V_{B,H,\{M_K\}} = \{0\}$.

We recall that the choice $V_{H,M,N}$ corresponds to the uniform choice $\{N_e\} = N$ and $\{M_K\} = M$ for all $K \in \mathcal{T}_H$ and for all edge $e \subset \Gamma$. The linear MsFEM approach corresponds to the choice $V_{B,H,\{M_K\}} = \{0\}$ and $N = 1$.

We solve (3.1) for a classical benchmark test introduced in [55], where A is periodic and oscillates at the scale ε. More specifically, we consider

$$A_\varepsilon(x) = a\left(\frac{x}{\varepsilon}, \frac{y}{\varepsilon}\right) I_2, \qquad a(x,y) = \frac{2 + 1.8\sin(2\pi x)}{2 + 1.8\cos(2\pi y)} + \frac{2 + \sin(2\pi y)}{2 + 1.8\sin(2\pi x)} \tag{3.16}$$

on the domain $D = (0,1)^2$, and solve

$$-\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = -1 \quad \text{in } D, \qquad u_\varepsilon = 0 \quad \text{on } \partial D. \tag{3.17}$$

We consider ε ranging from $1/32$ to $1/128$.

In order to compute errors, we have computed a reference solution of (3.17) using P2 Finite Elements with a mesh of size $h = 1/2048$. Note that $h \ll \varepsilon$ for the range of values of ε that we consider. Similarly, on each element K, the interface basis functions $\phi^\Gamma_{e,i}$ have no analytical expression and are approximated using P1 Finite Elements on a mesh of a small size (of the order of h).

We present here a *selection* of the extensive volley of numerical tests we have performed to establish the performance of our approach as compared to some other existing approaches.

To start with, we wish to illustrate the somewhat intuitive statement made above regarding the fact that multiscale approaches typically do a better job at approximating the solution in the bulk than on the interfaces, thus the interest of focusing our study and our efforts on the enrichment by Legendre polynomials on the edges. To support this claim, we show on FIGURE 3.2 the typical error obtained using the linear version of MsFEM (left). Specifically, we show the relative error $\log_{10}\left[\left|\nabla\left(u_\varepsilon - u_\varepsilon^{H,M,N}\right)\right| / \left|\nabla u_\varepsilon\right|\right]$ as a function of $x \in D$. The largest errors are evidently concentrated on the interfaces. Already an enrichment of Legendre polynomials of degree $N = 4$ on the edges allow one to dramatically reduce the latter error, as shown by the right of FIGURE 3.2. We notice on FIGURE 3.2 that the relative errors seem really large in the center part of D. This is probably an artefact stemming from the fact that $|\nabla u_\varepsilon|$ is very small in that region.

Further enriching the description of the solution along the edges with a large number of Legendre polynomials, say $N = 10$, would typically render the error almost homogeneous throughout the computational domain. All in all, the above set of comments justify our tactical choice to keep $V_{B,H,M} = \{0\}$ and focus on increasing N.

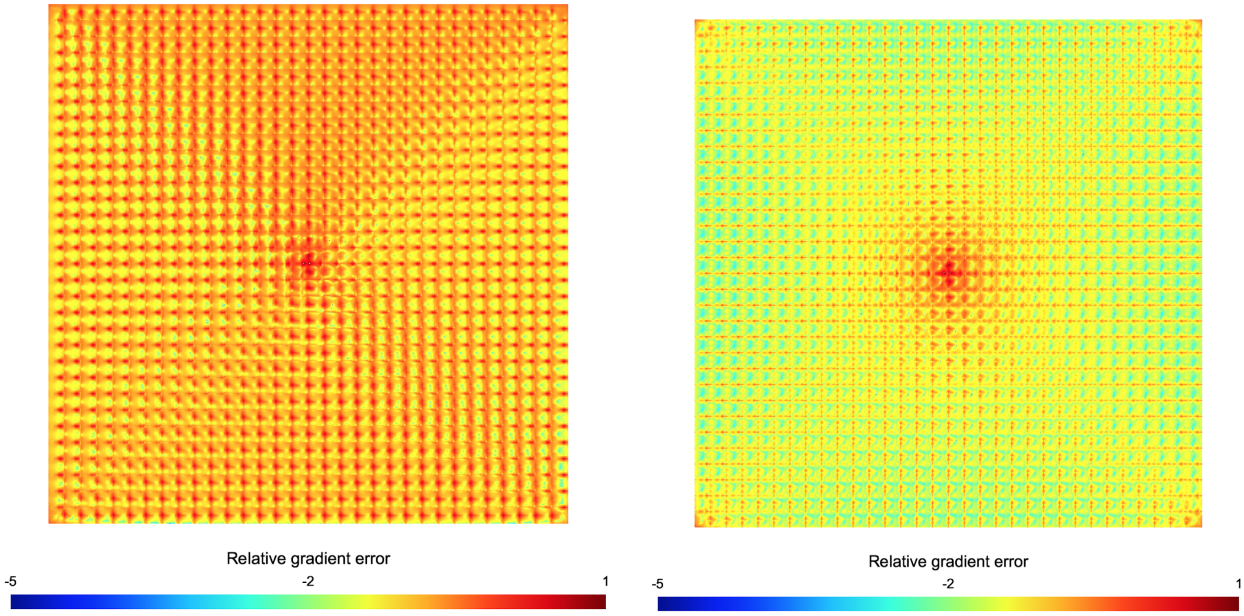


Figure 3.2: Error on the approximation of the gradient for MsFEM-lin (left) and for our approach with 4 polynomials (right): the accuracy is poor in the yellow regions, much better in the light blue regions, and excellent in the dark blue regions.

The next observation we want to make, and this is the purpose of FIGURE 3.3, is the poor performance of all previously existing multiscale approaches in the regime $H \approx \varepsilon$ (often called the *resonance regime*) where the coarse mesh size matches the typical size of oscillations. Whether one argues in terms of the meshsize H (left of FIGURE 3.3), or in terms of the number of degrees of freedom (right of that figure), traditional approaches saturate, while the Legendre enriched approach performs increasingly better. In FIGURE 3.3, and likewise in FIGURES 3.4, 3.5 and 3.6 below, all errors are relative errors in the energy norm

$$\mathcal{E}_{\text{rel}} = \sqrt{\frac{a(u_\varepsilon - u_\varepsilon^{H,M,N}, u_\varepsilon - u_\varepsilon^{H,M,N})}{a(u_\varepsilon, u_\varepsilon)}}, \tag{3.18}$$

while the approaches we test are respectively denominated as *MsFEM-lin* for the standard version of linear MsFEM, *MsFEM-OS* for its variant using oversampling (where the oversampling domain is 3 times larger in each direction than the original coarse element), *Legendre-N = ...* for the approach presented here using the corresponding degree N of Legendre polynomials on the edges, and *Eigen ...* for the Special Element approach.

The relative error (3.18) is easy to compute. Introduce indeed the energy $\mathcal{E}$, defined for any $v \in H_0^1(D)$ by

$$\mathcal{E}(v) = \frac{1}{2}\int_D (\nabla v)^T A_\varepsilon \nabla v - \int_D fv.$$

Since the matrix A_ε is symmetric, the solution to (3.1) is also the minimizer of the energy $\mathcal{E}$ in $H_0^1(D)$. Hence, denoting $\mathcal{E}^\star = \mathcal{E}(u_\varepsilon)$, it holds that $\mathcal{E}(v) - \mathcal{E}^\star = \frac{1}{2}a(u_\varepsilon - v, u_\varepsilon - v)$ for any $v \in H_0^1(D)$.

Remark 3.13. *This definition of the error is in practice very useful because computing (3.18) only requires to compare two scalars (namely $\mathcal{E}^\star$ and $\mathcal{E}(u_\varepsilon^{H,M,N})$) that can be obtained independently. We get $\mathcal{E}^\star$ by computing the energy for our reference solution and $\mathcal{E}(u_\varepsilon^{H,M,N})$ can be computed in parallel over the coarse elements K once the global problem has been solved at the online stage. In practicular, it is not needed to store the reference and numerical solutions, or compute their difference on a fine common Finite Element space, an operation which would be computationally very expensive.*

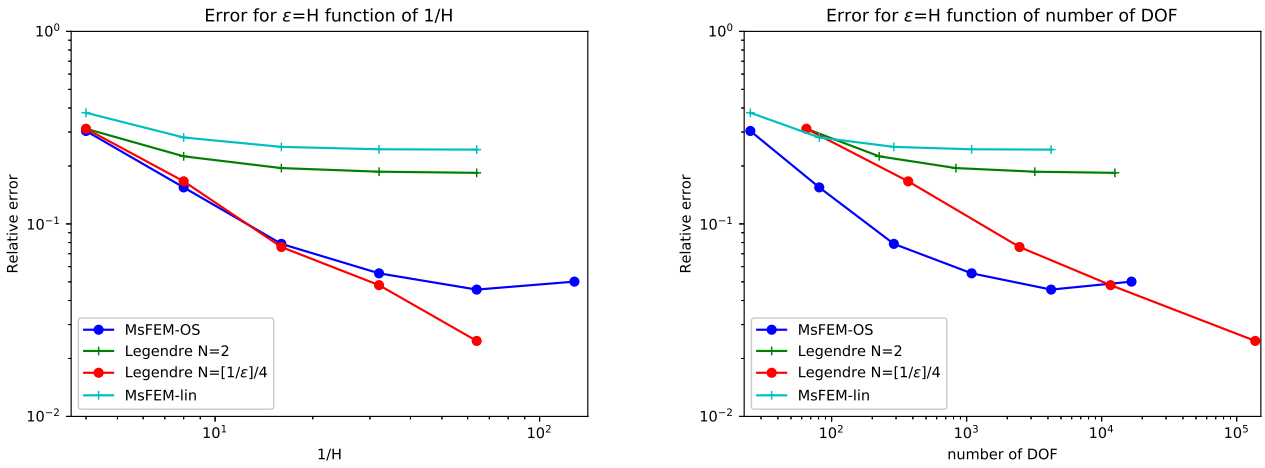


Figure 3.3: Compared performances in the regime $H \approx \varepsilon$.

We next perform, for ε fixed (namely at the value $\varepsilon = 1/32$), and H decreasing from $1/4$ to $1/64$ (or, correspondingly, the number of degrees of freedom increasing), full comparisons of the accuracy obtained for the various methods considered, including the Special Element Method. The results are shown on FIGURES 3.4 and 3.5 respectively.

For any fixed H, our approach is more accurate than the MsFEM oversampling method when N is large enough, say here $N \geq 9$ (see left side of FIGURE 3.4). For $N = 5$, our approach and the MsFEM oversampling method essentially share the same accuracy. The oversampling variant is more accurate for smaller values of N. However, for a fixed H, our approach needs more degrees of freedom than the oversampling approach. We thus compare the appoaches for a given number of degrees of freedom on the right side of FIGURE 3.4. When H is not too small (and thus the number of degrees of freedom is not too large), the MsFEM oversampling method provides better results than our approach. However, for smaller values of H (and thus larger numbers of degrees of freedom, say larger than 10^4), our approach outperforms the MsFEM oversampling method. We also notice that the oversampling approach suffers from a resonance effect (the error is essentially the same for any H between $1/128$ and $1/32$), whereas our approach provides an error which is monotonically decreasing with H.

Our tests of FIGURE 3.5 clearly show that our approach is equally accurate as (and in some cases more accurate than) the Special Element Method, for each given level of enrichment.

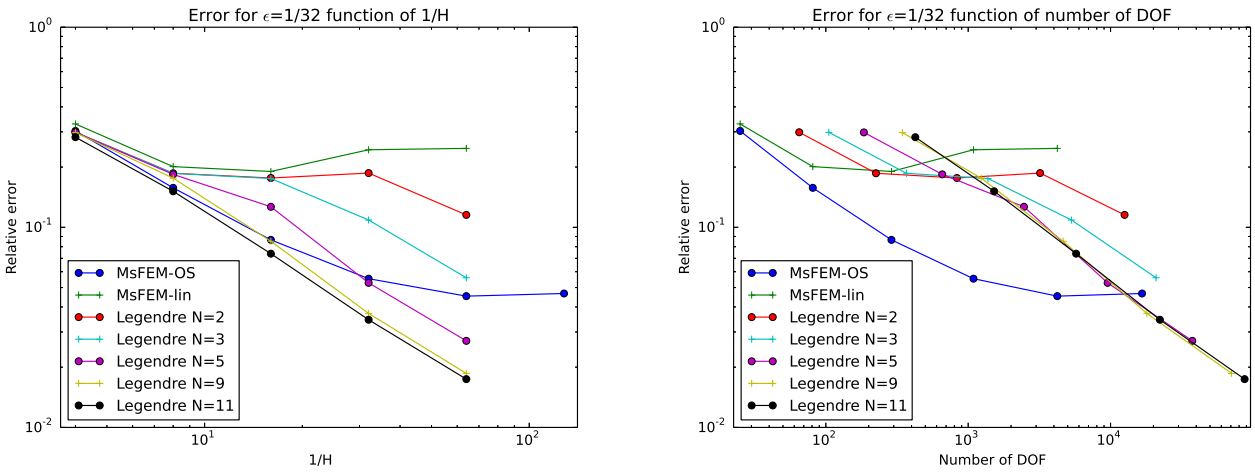


Figure 3.4: Comparison of our approach with classical MsFEM approaches

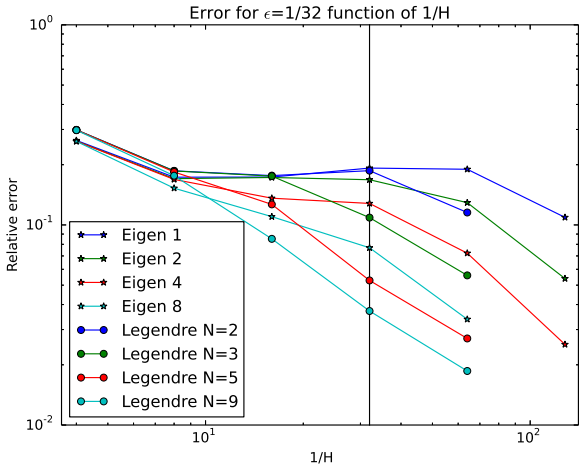


Figure 3.5: Comparison of our approach with the Special Element Method, at equal number of enrichments per edge (for instance, "Eigen 1" and "Legendre N=2" both correspond to adding one enrichment per edge vs the MsFEM-lin approach).

Our next test presented here compares the performance of our approach for triangular meshes and for quadrangular meshes. We set $\varepsilon = 1/32$ and present the relative energy error as a function of $1/H$ (left of FIGURE 3.6) and of the number of degrees of freedom (right of FIGURE 3.6). Our conclusion is that, essentially, the approach performs equally well in both cases, thereby making possible the application to a large class of computational domains, with intricate geometries for which quadrangular meshes cannot be used.

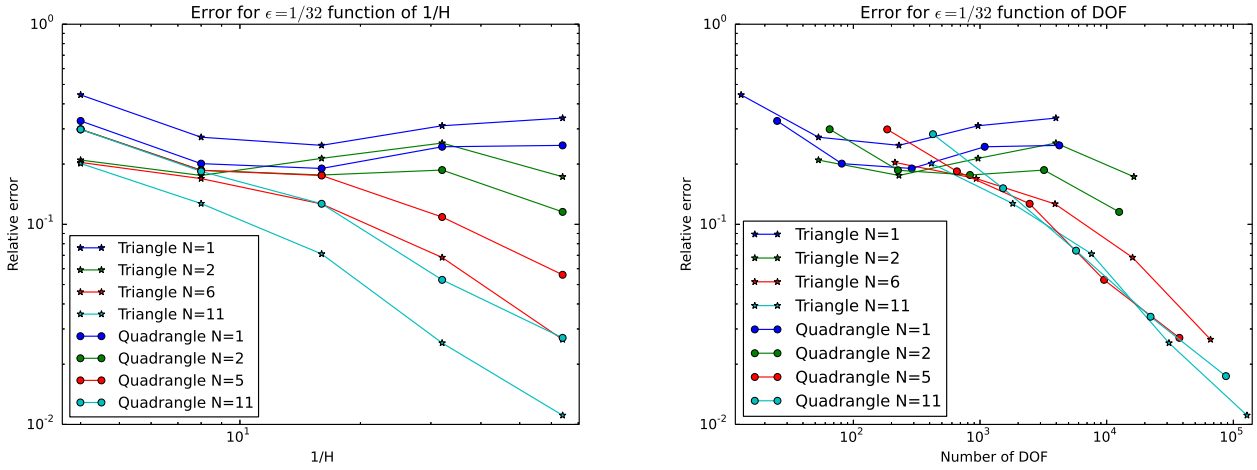


Figure 3.6: Our approach for triangles and quadrangles, in terms of $1/H$ (left) or of the number of degrees of freedom (right). The approaches "Triangle N=1" and "Quadrangle N=1" both correspond to the MsFEM-lin approach, on triangular (resp. quadrangular) meshes. The approaches "Triangle N=2" and "Quadrangle N=2" both correspond to adding one enrichment per edge vs the MsFEM-lin approach.

3.4.2 *A posteriori* estimator

We now investigate the performance of the *a posteriori* estimate given in PROPOSITION 3.12. As the previous section, we do not use bubble enrichments. This is why, instead of using as before the whole energy error, we now use the interface error defined by

$$\mathcal{E}_{\mathrm{rel},\Gamma} = \sqrt{\frac{a(u_\varepsilon^\Gamma - u_\varepsilon^{H,M,N}, u_\varepsilon^\Gamma - u_\varepsilon^{H,M,N})}{a(u_\varepsilon^\Gamma, u_\varepsilon^\Gamma)}}. \tag{3.19}$$

We compare this actual relative error with the error indicator given in (3.15), and more precisely with the indicator of the interface error, that is

$$\mathcal{E}_{\mathrm{post},\Gamma} = \left\{ \sum_{K\in\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e} \right) + \sum_{e\subset\Gamma} \frac{H_e}{p_e} \left\| J_e\big(\nu^T A \nabla(u_{\Gamma,H,\{N_e\}})\big) \right\|^2_{L^2(e)} \right\}^{1/2}. \tag{3.20}$$

Remark 3.14. *As above, we compute the relative error (3.19) thanks to the energy $\mathcal{E}$. The orthogonal decomposition (3.4) ensures that the energy of u_Γ is also a minimizer of the energy on V_Γ. Hence, the error can also be expressed as the difference between the energy of our approximation (which belongs to V_Γ as $V_{B,H,M} = \{0\}$) and the energy of u_Γ. We compute the energy of u_Γ by computing explicitly a reference solution for u_B (this simply requires to solve homogeneous Dirichlet problems in parallel in each element K) and get the associated energy. The energy of u_Γ is equal to $\mathcal{E}(u) - \mathcal{E}(u_B)$. This procedure is simpler than computing u_Γ, which would need to store the value of u on Γ.*

We consider here $f(x,y) = -10\exp(-80\,((x-0.5)^2 + (y-0.5)^2))$, keep the definition (3.16) for A_ε, and set $\varepsilon = 1/32$. On the following figures, we compare the relative interface error (3.19) with the *a posteriori* estimator (3.20) for several values of N and H. FIGURES 3.7 and 3.8 show the evolution of both errors when N increases and when $1/H$ increases, respectively.

In FIGURE 3.7, we see that, for $H = 1/4$ and $H = 1/8$, the *a posteriori* error behavior is an upper bound of the relative error interface and is a reliable indicator only for $N < 10$. When $H = 1/16, 1/32, 1/64$, the *a posteriori* error seems to represent well the relative interface error for any $N \leq 10$. For higher polynomial degrees, the relative interface error decreases sharply and the *a posteriori* indicator does not present such behavior. We are unsure to be able to trust the results for small values of H and large values of N for several reasons. First, we do not know the energy of u_Γ, but only approximate it by the energy of u_Γ^h (computed on the mesh of size $h = 1/2048$ using P2 Finite Elements). Likewise, we only manipulate numerical approximations of the basis functions. Finally, when the difference of the energies is much smaller than the energies themselves, computing a relative error may become challenging.

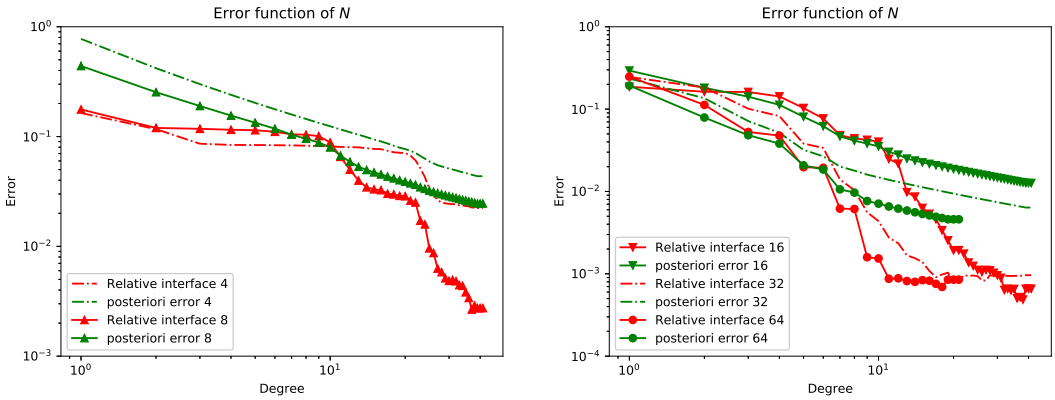


Figure 3.7: Left: *a posteriori* error (3.20) and relative interface error (3.19) as a function of N for $H = 1/4, 1/8$. Right: *a posteriori* error (3.20) and relative interface error (3.19) as a function of N for $H = 1/16, 1/32, 1/64$.

We turn now to FIGURE 3.8, which shows the behavior of the *a posteriori* estimator when H decreases for a fixed value of N. When H is large (say $H = 1/4, 1/8$), there is a significant difference between the *a posteriori* estimator and the relative error interface. For smaller values of H (say $H \leq 1/16$), the *a posteriori* error seems to behave like the relative error interface for $N = \{1, 2, 4, 6, 8\}$. We can see that the *a posteriori* estimator does not suffer from any resonance effect: it is decreasing with respect to H for all values of N tested.

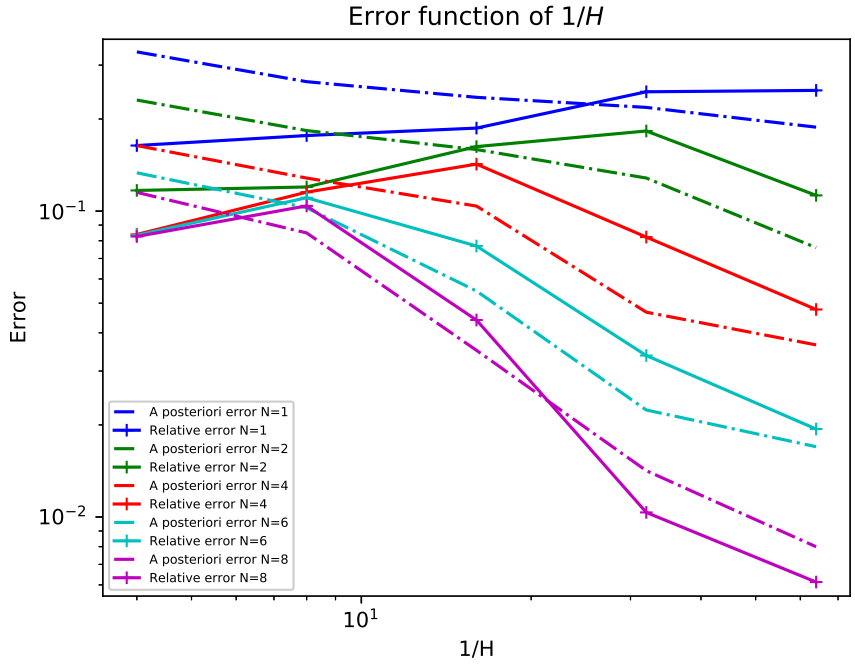


Figure 3.8: *A posteriori* error (3.20) and relative interface error (3.19) as a function of $1/H$ for different polynomial degrees $N = \{1, 2, 4, 6, 8\}$.

One of the main interest of an *a posteriori* estimator for which the error admits a local decomposition is to allow for local refinement with respect to the parameters of the method: in our case, the polynomial degree N_e of enrichments on each edge e and the size H_K of any element K. To that end, it is important to know whether the local behavior of the *a posteriori* estimator represents well the local behaviour of the actual error. This question is investigated on FIGURES 3.9, 3.10 and 3.11, where we show the error maps for $N = 1$, 5 and 10 respectively (with $H = 1/16$ fixed). We distribute the *a posteriori* estimator (3.20) onto the edges. The first term of (3.20) is element based. For each edge, we therefore add the contributions of this first term associated to the two elements sharing the edge. The second term of (3.20) is simpler to handle since it is already edge based. Using such a localization procedure, we obtain an *a posteriori* estimator which reads as a sum of contributions over the edges. Stated otherwise, we write (3.20) as

$$\mathcal{E}_{\mathrm{post},\Gamma} = \sqrt{\sum_{e\subset\Gamma} \big(\mathcal{E}_{\mathrm{post},\Gamma}(e)\big)^2}$$

with

$$\big(\mathcal{E}_{\mathrm{post},\Gamma}(e)\big)^2 = \frac{H_e}{p_e} \left\| J_e\big(\nu^T A \nabla(u_{\Gamma,H,\{N_e\}})\big)\right\|^2_{L^2(e)} + \sum_{K\in\mathcal{T}_H,\ e\subset\partial K} \|f\|^2_{L^2(K)} \left(\sum_{\tilde{e}\subset\partial K} \frac{H_{\tilde{e}}^2}{N_{\tilde{e}}\, p_{\tilde{e}}}\right).$$

We plot on the figures below the resulting values $\mathcal{E}_{\mathrm{post},\Gamma}(e)$ on a 16×16 coarse mesh. Regarding the actual error, we compute the relative energy error (3.19) elements by elements by loading a reference solution on each element and comparing it to the numerical approximation. Similarly to the first term of (3.20), we can write the numerator of (3.19), which is an element-based quantity, as a sum of contributions over the edges (the denominator of (3.19) is kept unchanged and is never localized). In some

of the figures below, we plot the error map showing the ratio between the actual local error and the local *a posteriori* estimator $\mathcal{E}_{\text{post},\Gamma}(e)$.

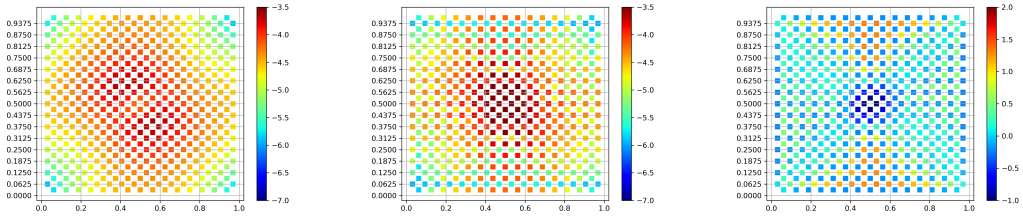


Figure 3.9: Error maps edge by edge for $N = 1$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale).

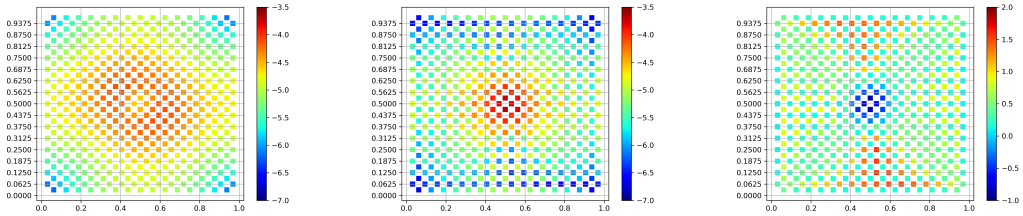


Figure 3.10: Error maps edge by edge for $N = 5$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale).

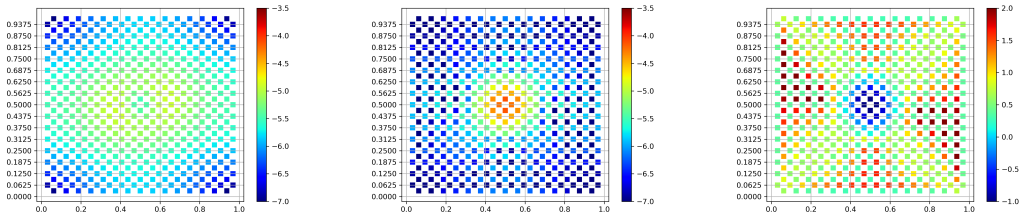


Figure 3.11: Error maps edge by edge for $N = 10$. Left: relative interface error; Center: *a posteriori* estimator; Right: ratio of the relative interface error and *a posteriori* estimator (the plots are shown in a base-10 log scale).

We see on the three error maps that the *a posteriori* estimator overestimates the actual error in the center of the domain (note also that this overestimation does not depend on N). When $N = 1$ and $N = 5$, we can see that the ratio between the local actual error and the local *a posteriori* estimator does not change much over the domain D, except near the center. It is thus possible to use the local *a posteriori* estimator to drive an adaptive discretization procedure: the edges where $\mathcal{E}_{\text{post},\Gamma}(e)$ is large are indeed the edges where the actual error is large. In contrast, when $N = 10$, the ratio between the actual and the predicted error takes very different values over the domain D. This is consistent with the above FIGURE 3.7 showing a big difference between the global actual error and the global estimated error for large values of N. For this large value of N, the quantity $\mathcal{E}_{\text{post},\Gamma}(e)$ cannnot be used to drive an adaptation procedure.

The above numerical tests hence show that the *a posteriori* estimator defined in PROPOSITION 3.12 seems to be able to represent the behaviour of the actual error in the regime H close or smaller than ε and for $N < 10$. In such regime, it can be used to locally refine the polynomial degree N_e associated with the edge e and the size H_K of the element K. However, in the regime $N > 10$ or for large values of H, the estimator fails to reproduce the actual error behavior and must be used carefully.

Acknowledgments The work of CLB, FL and PLR is partly supported by ONR under Grant N00014-15-1-2777 and by EOARD under Grant FA9550-17-1-0294. Part of this work has been completed while PLR was visiting the University of Washington in Seattle. The hospitality of that institution and the support of a "Bourse de Mobilité" of the Ecole Doctorale SIE at Université Paris-Est are gratefully acknowledged. The content of this contribution was presented, along with a larger exposition of some recent progress in multi-scale finite element methods and their relation to domain decomposition methods, in the plenary address of CLB at DD25, Saint John's, Newfoundland, July 2018. CLB wishes to thank the scientific program committee and the organization committee for their invitation.

3.5 Proofs

The proofs of the above results critically rely on results about polynomial approximation theory, fractional Sobolev spaces, traces operators and elliptic regularity. For the sake of clarity, the main results used here are also presented in ANNEX B in a more comprehensive manner.

3.5.1 Proof of LEMMA 3.9

The proof of LEMMA 3.9 needs the following approximation result (which is shown, for integer values of ℓ, in [22, EQUATION (5.8.27) p. 318] for the case of quadrangles and [22, SECTION 5.9] for the case of triangles).

Lemma 3.15. *Assume that $(\mathcal{T}_H)_H$ is a family of conforming partitions of D into a finite number of convex quadrilaterals (resp. triangles) with straight edges. We assume that the mesh is regular in the sense of (3.2). For any quadrangle K (resp. triangle K), let Π_M^K be the $L^2(K)$-orthogonal projection on the vector space of polynomials of degree in each variable (resp. total degree) at most M. Then, for any non-negative real number ℓ, there exists C_ℓ independent of H, M and of the elements K of the family of partitions such that, for any $v \in H^\ell(K)$,*

$$\left\|v - \Pi_M^K(v)\right\|_{L^2(K)} \leq C_\ell \frac{H^{\min(\ell,M+1)}}{M^\ell} |v|_{H^{\ell:M}(K)} \leq C_\ell \frac{H^{\min(\ell,M+1)}}{M^\ell} \|v\|_{H^\ell(K)}$$

where $|\cdot|^2_{H^{\ell:M}(K)} = \sum_{k=M+1}^{\lfloor \ell \rfloor} |\cdot|^2_{H^k(K)} + |\cdot|^2_{H^\ell(K)}$.

To prove LEMMA 3.15, one first considers the case when ℓ is an integer. The general case of real values of ℓ is obtained using the following Sobolev interpolation result.

Lemma 3.16. *(see [68, THEOREM 5.1]) Let $(\mathcal{X}, \mathcal{Y})$ be a couple of separable Hilbert spaces with $\mathcal{X} \subset \mathcal{Y}$, such that $\mathcal{X}$ is dense in $\mathcal{Y}$ and such that the injection from $\mathcal{X}$ to $\mathcal{Y}$ is continuous. Let (X, Y) be another couple of Hilbert spaces with analogous properties. Denote by $\mathcal{L}(X, Y)$ the set of linear continuous operators from X to Y, and likewise*

for $\mathcal{L}(\mathcal{X},\mathcal{Y})$. Let π be an operator satisfying $\pi \in \mathcal{L}(\mathcal{X},\mathcal{Y}) \cap \mathcal{L}(X,Y)$. Then, for all $0<\theta<1$, we have

$$\pi \in \mathcal{L}([\mathcal{X},X]_\theta,[\mathcal{Y},Y]_\theta)$$

where the interpolated space $[\mathcal{X},X]_\theta$ is defined in [68, DEFINITION *2.1].*

Remark 3.17. LEMMA *3.16 is useful especially when considering Sobolev spaces and allows to easily extend properties shown for Sobolev spaces with integer index to fractional Sobolev spaces. Indeed, for any $0<s<1$, one can define $H^s(D)$ as $[L^2(D),H^1(D)]_s$. As an application of* LEMMA *3.16, consider a linear operator π defined in $L^2(D)$ and which satisfies $\|\pi u\|_{L^2(D)} \leq \|u\|_{L^2(D)}$ for any $u \in L^2(D)$ and $\|\pi u\|_{L^2(D)} \leq C\|u\|_{H^1(D)}$ for any $u \in H^1(D)$. Then we know that $\pi \in \mathcal{L}(H^s(D),L^2(D))$. The proof of* LEMMA *3.16 actually provides an explicit value for the continuity constant, and yields that $\|\pi u\|_{L^2(D)} \leq C^s\|u\|_{H^s(D)}$ for any $u \in H^s(D)$.*

Proof of LEMMA 3.15. When $\ell=0$, the result comes from the fact that a projection is stable. When ℓ is a positive integer, we refer to [22, Eq. (5.8.27) p. 318] in the case of quadrangles, and [22, Sec. 5.9] in the case of triangles. For the case when ℓ is a non-integer real number, we first consider a reference element $\overline{K}$ with $\mathrm{diam}(\overline{K})=1$ and proceed by using a Sobolev interpolation argument (see LEMMA 3.16). This yields an estimate with the right power in M. We next perform a rescaling and use the regularity of the mesh, which implies the fact that the transformation between $\overline{K}$ and K is affine with a gradient bounded by H. □

Proof of LEMMA 3.9. We show first that $\|u_B\|_E \leq CH\|f\|_{L^2(D)}/\sqrt{\alpha_{\min}}$. We have that

$$\|u_B\|_E^2 = \int_D (\nabla u_B)^T A \nabla u_B = \int_D f u_B = \sum_{K\in\mathcal{T}_H} \int_K f u_B.$$

Using the Cauchy-Schwarz inequality and the Poincaré inequality (recall indeed that $u_B \in H_0^1(K)$), it holds that

$$\int_K f u_B \leq \|f\|_{L^2(K)}\|u_B\|_{L^2(K)} \leq \|f\|_{L^2(K)} CH |u_B|_{H^1(K)}$$

for some universal constant C. We hence have

$$\begin{aligned}\|u_B\|_E^2 &\leq CH \sum_{K\in\mathcal{T}_H} \|f\|_{L^2(K)}|u_B|_{H^1(K)} \\ &\leq CH\|f\|_{L^2(D)}|u_B|_{H^1(D)} \\ &\leq \frac{CH}{\sqrt{\alpha_{\min}}}\|f\|_{L^2(D)}\|u_B\|_E,\end{aligned}$$

from which we deduce that $\|u_B\|_E \leq \dfrac{CH}{\sqrt{\alpha_{\min}}}\|f\|_{L^2(D)}$ for some universal constant C. This proves (3.11).

We now consider the case when we add bubble enrichments for each element $K \in \mathcal{T}_H$ with a uniform degree $M \geq 1$. Galerkin orthogonality implies that, for any $v_{B,H,M} \in V_{B,H,M}$,

$$a(u_B - u_{B,H,M}, u_B - u_{B,H,M}) \leq a(u_B - v_{B,H,M}, u_B - v_{B,H,M}). \tag{3.21}$$

Recall that $V_{B,H,M}$ is the span of the functions $\{\phi^B_{K,i}\}_{i=1,\dots,\mathcal{N}_M}$ (see (3.6)) which solve in the element K the problem $-\mathrm{div}(A\nabla\phi^B_{K,i}) = P_i$, where $\{P_i\}_{i=1,\dots,\mathcal{N}_M}$ is a basis of

polynomial functions with (total or partial) degree M. Because of that very definition, we may uniquely define $v_{B,H,M} \in V_{B,H,M}$ as the solution to

$$\forall v \in V_B, \quad a(v_{B,H,M}, v) = \int_D \Pi_M(f)\, v,$$

where $\Pi_M(f) = \sum_{K\in\mathcal{T}_H} 1_K\, \Pi_M^K(f)$, where Π_M^K is defined in LEMMA 3.15. Using the definition of u_B, we next obtain that, for any $v \in V_B$,

$$\begin{aligned} a(u_B - v_{B,H,M}, v) &= \int_D fv - \int_D \Pi_M(f)\, v \\ &= \sum_{K\in\mathcal{T}_H} \int_K \left(f - \Pi_M^K(f)\right) v \\ &= \sum_{K\in\mathcal{T}_H} \int_K \left(f - \Pi_M^K(f)\right) \left(v - \Pi_M^K(v)\right). \end{aligned}$$

Choosing now $v = u_B - v_{B,H,M}$ in the above equality yields

$$\begin{aligned} &a(u_B - v_{B,H,M}, u_B - v_{B,H,M}) \\ &\le \sum_{K\in\mathcal{T}_H} \left\|f - \Pi_M^K(f)\right\|_{L^2(K)} \left\|u_B - v_{B,H,M} - \Pi_M^K\left(u_B - v_{B,H,M}\right)\right\|_{L^2(K)} \\ &\le C_\ell \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} \sum_{K\in\mathcal{T}_H} |f|_{H^{\ell:M}(K)}\, |u_B - v_{B,H,M}|_{H^1(K)} \end{aligned}$$

where we have used in the last line the polynomial projection properties stated in LEMMA 3.15, the fact that $f \in H^\ell(D)$ and that $M \ge 1$. We thus deduce that

$$\begin{aligned} &a(u_B - v_{B,H,M}, u_B - v_{B,H,M}) \\ &\le C_\ell \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} |f|_{H^{\ell:M}(D)} \|\nabla(u_B - v_{B,H,M})\|_{L^2(D)} \\ &\le \frac{C_\ell}{\sqrt{\alpha_{\min}}} \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} |f|_{H^{\ell:M}(D)} \sqrt{a(u_B - v_{B,H,M}, u_B - v_{B,H,M})} \end{aligned}$$

hence

$$\sqrt{a(u_B - v_{B,H,M}, u_B - v_{B,H,M})} \le \frac{C_\ell}{\sqrt{\alpha_{\min}}} \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} |f|_{H^{\ell:M}(D)}. \tag{3.22}$$

Inserting (3.22) into (3.21), we obtain

$$\sqrt{a(u_B - u_{B,H,M}, u_B - u_{B,H,M})} \le \frac{C_\ell}{\sqrt{\alpha_{\min}}} \frac{H^{\min(\ell,M+1)+1}}{M^{\ell+1}} |f|_{H^{\ell:M}(D)},$$

which is the bound (3.10). This concludes the proof of LEMMA 3.9. □

3.5.2 Proof of LEMMA 3.10

For the proof of LEMMA 3.10, we separately consider the case of quadrangles and the case of triangles. For the former case, we need the following approximation result (see the discussion in [22] between EQUATION (5.8.27) and SECTION 5.8.4).

Lemma 3.18. *Assume that $(\mathcal{T}_H)_H$ is a family of conforming partitions of D into a finite number of convex quadrilaterals with straight edges. We assume that the mesh is regular. For any quadrangle K, let i_N^K be the Legendre interpolant at the $(1+N)^2$ Gauss-Lobatto points in K (i_N^K is thus a polynomial function in $\mathbb{Q}_N$). Let $s > 1$ and $N \geq 1$. Then there exists C_s independent of H, N and of the elements K of the family of partitions such that, for any $v \in H^s(K)$,*

$$\left|v - i_N^K(v)\right|_{H^1(K)} \leq C_s \frac{H^{\min(s,N+1)-1}}{N^{s-1}} |v|_{H^s(K)}.$$

Note that a function $v \in H^s(K)$ with $s > 1$ is continuous (recall that we consider a two-dimensional setting), thus $i_N^K(v)$ is well-defined.

Proof of LEMMA 3.10 for quadrangles. As for LEMMA 3.9, Galerkin orthogonality implies that, for any $v_{\Gamma,H,N} \in V_{\Gamma,H,N}$,

$$a(u_\Gamma - u_{\Gamma,H,N}, u_\Gamma - u_{\Gamma,H,N}) \leq a(u_\Gamma - v_{\Gamma,H,N}, u_\Gamma - v_{\Gamma,H,N}). \tag{3.23}$$

Since $u \in H^s(D)$ with $s > 3/2$, u is continuous on D, hence on Γ (note here that $u \in H^s(D)$ with $s > 1$ would be sufficient). We denote by $i_N^\Gamma(u)$ the interpolant (at the Gauss-Lobatto points on Γ) of u on the set of continuous functions on Γ which are piecewise equal to polynomial functions of degree lower or equal to N.

Let $w = E_D(i_N^\Gamma(u)) \in V_{\Gamma,H,N}$ denote the harmonic lifting of $i_N^\Gamma(u)$, that is the solution to $-\operatorname{div}(A\nabla w) = 0$ on each coarse element K with the Dirichlet boundary conditions $w = i_N^\Gamma(u)$ on Γ. We then have

$$\begin{aligned}
&a\Big(u_\Gamma - E_D(i_N^\Gamma(u)), u_\Gamma - E_D(i_N^\Gamma(u))\Big) \\
&= \sum_{K\in\mathcal{T}_H} \int_K \big(\nabla u_\Gamma - \nabla E_D(i_N^\Gamma(u))\big)^T A\big(\nabla u_\Gamma - \nabla E_D(i_N^\Gamma(u))\big) \\
&\leq \sum_{K\in\mathcal{T}_H} \int_K \big(\nabla u - \nabla I_{\Gamma,H,N}(u)\big)^T A\big(\nabla u - \nabla I_{\Gamma,H,N}(u)\big) \\
&\leq \alpha_{\max} \sum_{K\in\mathcal{T}_H} |u - I_{\Gamma,H,N}(u)|^2_{H^1(K)}
\end{aligned} \tag{3.24}$$

where $I_{\Gamma,H,N}(u) \in H^1_0(D)$ is defined piecewise on each K as the Legendre interpolant of $u|_K$ at the Gauss-Lobatto points in K (it is thus a polynomial function in $\mathbb{Q}_N$). The first inequality of (3.24) holds for the following reasons:

- First, $E_D(i_N^\Gamma(u))$ and $I_{\Gamma,H,N}(u)$ agree on Γ for quadrangular mesh elements (recall indeed that the Gauss-Lobatto points of each edge of ∂K are a subset of the Gauss-Lobatto points of K; on each edge, $E_D(i_N^\Gamma(u))$ and $I_{\Gamma,H,N}(u)$ are thus two polynomial functions of degree lower than or equal to N which are equal on the $(1+N)$ Gauss-Lobatto points of the edge, and are thus equal).
- Second, u_Γ and u agree on Γ, by definition of u_Γ.
- The function $u_\Gamma - E_D(i_N^\Gamma(u))$ thus agrees with $u - I_{\Gamma,H,N}(u)$ on Γ, and the former is energy-minimizing in each element K. This hence shows the first inequality of (3.24).

Using LEMMA 3.18, we see that

$$\left(\sum_{K\in\mathcal{T}_H} |u - I_{\Gamma,H,N}(u)|^2_{H^1(K)}\right)^{1/2} \leq C_s \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)}.$$

Collecting this bound with (3.23) (that we use for $v_{\Gamma,H,N} = E_D(i_N^\Gamma(u))$) and (3.24), we deduce that

$$\sqrt{a(u_\Gamma - u_{\Gamma,H,N}, u_\Gamma - u_{\Gamma,H,N})} \leq C_s \sqrt{\alpha_{\max}} \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)},$$

which concludes the proof of LEMMA 3.10 for quadrangles. □

We now turn to the proof of LEMMA 3.10 for the case of triangles. Actually, the proof given below also holds in the case of quadrangles, and provides the same estimate as the above proof (we have kept the above proof for the case of quadrangles because the choice of $v_{\Gamma,H,N}$ is constructive, in contrast to the proof below). To address this second case, the following result is useful.

Lemma 3.19. *(see [6, THEOREM 4.1] and [5, THEOREMS 4.6, 4.8 and SECTION 4.2]) Consider a mesh $\mathcal{T}^H$ composed of quasiuniform triangular or quadrilateral elements with meshsize H. Let $u \in H^s(D) \cap H_0^1(D)$ with $s > 3/2$. Let $V_{H,0}^N = \{v \in C^0(\overline{D}) \cap H_0^1(D);\ v|_K \in P_N^K\}$ where P_N^K is the set of polynomial functions on K which are of (total or partial) degree lower or equal to N. We then have*

$$\min_{v \in V_{H,0}^N} \|u - v\|_{H^1(D)} \leq C_s \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)} \tag{3.25}$$

where C_s is independent of H, N and u.

This result will play the role in the general case of LEMMA 3.18 in the case of quadrangles.

Remark 3.20. THEOREMS *4.6 and 4.8 in [5] consider the problem of approximating the solution $u \in H_0^1(D)$ to $-\Delta u + u = f$ in D. Their proof relies of some approximation results. These approximation results have their own interest and can be stated as in* LEMMA *3.19 for any $u \in H^s(D) \cap H_0^1(D)$.*

Proof of LEMMA 3.10, alternative proof for triangles and quadrangles. Using Galerkin orthogonality, it holds, for any $v_{\Gamma,H,N} \in V_{\Gamma,H,N}$, that

$$a(u_\Gamma - u_{\Gamma,H,N}, u_\Gamma - u_{\Gamma,H,N}) \leq a(u_\Gamma - v_{\Gamma,H,N}, u_\Gamma - v_{\Gamma,H,N}). \tag{3.26}$$

Thanks to LEMMA 3.19, there exists a function $P(u) \in V_{H,0}^N$ such that, for any $s > 3/2$,

$$\|u - P(u)\|_{H^1(D)} \leq C \frac{H^{\min(s,N+1)-1}}{N^{s-1}} \|u\|_{H^s(D)}. \tag{3.27}$$

We consider the harmonic lifting $w = E_D(P(u))$ of $P(u)|_\Gamma$ on D, that is the solution to $-\operatorname{div}(A\nabla w) = 0$ on each coarse element K with the Dirichlet boundary conditions $w = P(u)$ on Γ. Note that $P(u)$ is continuous on Γ and smooth on each edge, which implies that w is well-defined and belongs to $H^1(D)$. Moreover, on each edge, $P(u)$ is a polynomial function of degree lower or equal to N. We therefore have that $w \in V_{\Gamma,H,N}$.

We now write

$$\begin{aligned}
&a\Big(u_\Gamma - E_D(P(u)), u_\Gamma - E_D(P(u))\Big) \\
&= \sum_{K\in\mathcal{T}_H} \int_K (\nabla u_\Gamma - \nabla E_D(P(u)))^T A\, (\nabla u_\Gamma - \nabla E_D(P(u))) \\
&\leq \sum_{K\in\mathcal{T}_H} \int_K (\nabla u - \nabla P(u)))^T A\, (\nabla u - \nabla P(u)) \\
&\leq \alpha_{\max} \sum_{K\in\mathcal{T}_H} |u - P(u)|^2_{H^1(K)} \\
&= \alpha_{\max} |u - P(u)|^2_{H^1(D)}, \qquad (3.28)
\end{aligned}$$

where the first inequality above again comes from the fact that $u_\Gamma - E_D(P(u))$ is energy-minimizing in each element K and agrees with $u - P(u)$ on ∂K. Collecting (3.26) (that we use for $v_{\Gamma,H,N} = E_D(P(u))$), (3.28) and (3.27), we deduce that

$$\sqrt{a(u_\Gamma - u_{\Gamma,H,N}, u_\Gamma - u_{\Gamma,H,N})} \leq C\sqrt{\alpha_{\max}}\, \frac{H^{\min(s,N+1)-1}}{N^{s-1}}\, \|u\|_{H^s(D)},$$

which concludes the general proof of LEMMA 3.10. □

3.5.3 Proof of PROPOSITION 3.12

The proof of PROPOSITION 3.12 requires the three following results, LEMMAS 3.21, 3.22 and 3.23.

Lemma 3.21. *Consider an element K of diameter H in the mesh. We assume that K is convex and that $H \leq 1$. Let $f \in L^2(K)$ and consider $z \in H^1_0(K)$ solution to*

$$-\Delta z = f \quad \text{in } K.$$

Then $z \in H^2(K)$ and there exists C, which only depends on the regularity of the mesh (in the sense of (3.2)), such that, for any edge $e \subset \partial K$, we have

$$\|\nabla z\|_{H^{1/2}(e)} \leq C\, \|f\|_{L^2(K)}. \qquad (3.29)$$

Proof of LEMMA 3.21. The fact that $z \in H^2(K)$ stems from elliptic regularity and the fact that K is convex (see [44, CHAPTER I, THEOREM 1.8] or [48, COROLLARY 2.6.8]). We proceed by scaling to show the estimate (3.29).

We first consider the case when K is obtained from the reference element by an homothetic transformation of ratio H. Let $z_{\text{ref}}(x) = z(Hx)$ and $f_{\text{ref}}(x) = f(Hx)$ be defined on the reference element K_{ref} of unit diameter. We compute that, in K_{ref},

$$-(\Delta z_{\text{ref}})(x) = -H^2(\Delta z)(Hx) = H^2 f(Hx) = H^2\, f_{\text{ref}}(x).$$

By elliptic regularity, we thus have $\|\nabla z_{\text{ref}}\|_{H^{1/2}(e_{\text{ref}})} \leq C\, H^2\, \|f_{\text{ref}}\|_{L^2(K_{\text{ref}})}$. We now proceed by scaling. We have

$$\|f_{\text{ref}}\|^2_{L^2(K_{\text{ref}})} = \int_{K_{\text{ref}}} f^2(Hx)\, dx = H^{-2} \int_K f^2(x)\, dx = H^{-2}\, \|f\|^2_{L^2(K)}.$$

Furthermore,

$$\begin{aligned}
\|\nabla z_{\text{ref}}\|^2_{L^2(e_{\text{ref}})} = \int_{e_{\text{ref}}} |(\nabla z_{\text{ref}})(x)|^2\, dx &= H^2 \int_{e_{\text{ref}}} |(\nabla z)(Hx)|^2\, dx \\
&= H \int_e |(\nabla z)(x)|^2\, dx = H\, \|\nabla z\|^2_{L^2(e)}
\end{aligned}$$

and

$$|\nabla z_{\rm ref}|^2_{H^{1/2}(e_{\rm ref})} = \int_{e_{\rm ref}}\int_{e_{\rm ref}} \frac{|\nabla z_{\rm ref}(x)-\nabla z_{\rm ref}(y)|^2}{|x-y|^2}\,dx\,dy$$
$$= H^2 \int_{e_{\rm ref}}\int_{e_{\rm ref}} \frac{|(\nabla z)(Hx)-(\nabla z)(Hy)|^2}{|x-y|^2}\,dx\,dy$$
$$= H^2 \int_e\int_e \frac{|(\nabla z)(x)-(\nabla z)(y)|^2}{|x-y|^2}\,dx\,dy = H^2\,|\nabla z|^2_{H^{1/2}(e)}\,.$$

We hence write that

$$\|\nabla z\|^2_{H^{1/2}(e)} = \|\nabla z\|^2_{L^2(e)} + |\nabla z|^2_{H^{1/2}(e)} = H^{-1}\,\|\nabla z_{\rm ref}\|^2_{L^2(e_{\rm ref})} + H^{-2}\,|\nabla z_{\rm ref}|^2_{H^{1/2}(e_{\rm ref})}$$
$$\le H^{-2}\,\|\nabla z_{\rm ref}\|^2_{H^{1/2}(e_{\rm ref})} \le C^2 H^2\,\|f_{\rm ref}\|^2_{L^2(K_{\rm ref})} = C^2\|f\|^2_{L^2(K)},$$

which is (3.29) in the simple homothetic case.

To show (3.29) in full generality, we again define $z_{\rm ref}(x) = z(F(x))$ and $f_{\rm ref}(x) = f(F(x))$ for any x in the reference element $K_{\rm ref}$ of unit diameter, where F is the affine transformation introduced at the beginning of Section 3.2. We then compute that, in $K_{\rm ref}$,

$$-\operatorname{div}(H^2\mathcal{A}\nabla z_{\rm ref}) = H^2\,f_{\rm ref},$$

where $\mathcal{A}$ is a constant matrix defined by $\mathcal{A} = (\nabla F^{-1})^T\nabla F^{-1}$. Using the bounds on ∇F and ∇F^{-1} assumed in (3.2), we observe that the matrix $H^2\mathcal{A}$ is bounded independently of H and coercive with a constant independent of H. By elliptic regularity, we thus again have $\|\nabla z_{\rm ref}\|_{H^{1/2}(e_{\rm ref})} \le C\,H^2\,\|f_{\rm ref}\|_{L^2(K_{\rm ref})}$, as in the simple homothetic case. The sequel of the proof follows the same lines as above, simply using the bounds on ∇F and ∇F^{-1}. This concludes the proof of LEMMA 3.21. □

Lemma 3.22. *Consider an element K of diameter H in the mesh. We assume that K is convex and that $H \le 1$. Let $a \in W^{1,\infty}(K)$ with $a(x) \ge \alpha_{\min} > 0$ almost everywhere in K. Let $f \in L^2(K)$ and consider $z \in H^1_0(K)$ solution to*

$$-\operatorname{div}(a\nabla z) = f \quad \text{in } K. \tag{3.30}$$

We then have the following assertions:

(i) The function z belongs to $H^2(K)$ and there exists C_a, which only depends on the regularity of the mesh, $\alpha_{\min}$ and $\|\nabla a\|_{L^\infty(K)}$, such that, for any edge $e \subset \partial K$, we have

$$\|\nabla z\|_{H^{1/2}(e)} \le C_a\,\|f\|_{L^2(K)}. \tag{3.31}$$

(ii) There exists $p_K^{\rm Sob} > 2$ depending on the largest inner angle of ∂K such that, when $f \in L^p(K)$ for some $2 \le p < p_K^{\rm Sob}$, then $z \in W^{2,p}(K)$.

In relation to the assertion (ii) above, note that, if K satisfies an exterior sphere condition and f is a Hölder function, then $z \in C^{2,\alpha}(K)$ for some $\alpha > 0$ (see [43, THEOREM 6.24]).

Proof of LEMMA 3.22. The proof of assertion (i) is performed by using LEMMA 3.21. Since a is scalar-valued, we can rewrite (3.30) as $-a\Delta z = f + \nabla a\cdot\nabla z$, that is

$$-\Delta z = F \ \text{ in } K \qquad \text{with} \qquad F = \frac{f}{a} + \frac{\nabla a}{a}\cdot\nabla z. \tag{3.32}$$

We know that $\nabla z \in L^2(K)$. Since $\nabla a \in L^\infty(K)$, we get that $F \in L^2(K)$. Using LEMMA 3.21, we hence obtain that $z \in H^2(K)$ and that

$$\|\nabla z\|_{H^{1/2}(e)} \leq C \|F\|_{L^2(K)} \leq \frac{C}{\alpha_{\min}} \left(\|f\|_{L^2(K)} + \|\nabla a\|_{L^\infty(K)} \|\nabla z\|_{L^2(K)} \right).$$

To estimate $\|\nabla z\|_{L^2(K)}$, we use the variational formulation of (3.30), which leads to

$$\alpha_{\min} \|\nabla z\|^2_{L^2(K)} \leq \int_K (\nabla z)^T a \nabla z = \int_K f\, z \leq \|f\|_{L^2(K)} \|z\|_{L^2(K)} \leq C\, H\, \|f\|_{L^2(K)} \|\nabla z\|_{L^2(K)},$$

where we have used for the final estimate a Poincaré inequality on K, which is of diameter H. We thus obtain that $\alpha_{\min} \|\nabla z\|_{L^2(K)} \leq C\, H\, \|f\|_{L^2(K)}$, and hence

$$\|\nabla z\|_{H^{1/2}(e)} \leq \frac{C}{\alpha_{\min}} \left(\|f\|_{L^2(K)} + C\, H\, \frac{\|\nabla a\|_{L^\infty(K)}}{\alpha_{\min}} \|f\|_{L^2(K)} \right).$$

Assuming that $H \leq 1$, we hence obtain (3.31).

To prove assertion (ii), we again use the formulation (3.32). Consider any $p_K^{\mathrm{Sob}} > 2$ that will be fixed later. Assume that $f \in L^p(K)$ for some $p \in [2, p_K^{\mathrm{Sob}})$. We have shown that $\nabla z \in H^1(K)$, hence $\nabla z \in L^p(K)$ using Sobolev embeddings. We hence have $F \in L^p(K)$. Using [44, CHAPTER I, THEOREM 1.8 (ii)], we deduce the existence of some $p_K^{\mathrm{Sob}} > 2$ (depending on the largest inner angle of ∂K) such that the fact that $F \in L^p(K)$ for some $2 \leq p < p_K^{\mathrm{Sob}}$ implies that $z \in W^{2,p}(K)$. This concludes the proof of LEMMA 3.22. □

Our last technical result is the following approximation result. Consider a mesh $\mathcal{T}_H$ and choose a maximal polynomial degree $p_K \in \mathbb{N}^\star$ for any element $K \in \mathcal{T}_H$. We assume that these degrees are comparable on neighbouring elements, in the sense that

$$\forall K, K' \in \mathcal{T}_H \text{ s.t. } \overline{K} \cap \overline{K'} \neq \emptyset, \quad \frac{p_K}{\gamma} \leq p_{K'} \leq \gamma\, p_K, \tag{3.33}$$

where γ is the mesh regularity constant of (3.2).

Lemma 3.23. *(Scott-Zhang type interpolation result [72, THEOREM 2.3]) Assume that $\mathcal{T}_H$ is a conformal mesh which is shape regular in the sense of (3.2). For any element $K \in \mathcal{T}_H$, we choose a maximal degree $p_K \in \mathbb{N}^\star$ and we assume that these degrees $\{p_K\}$ satisfy (3.33). Then there exists a continuous interpolation operator $\mathcal{SZ}$ from $H^1_0(D)$ to $H^1_0(D) \cap \mathcal{S}(\{p_K\})$ with*

$$\mathcal{S}(\{p_K\}) = \{u \in C^0(\overline{D});\ u|_K \text{ is a polynomial function of degree at most } p_K\}.$$

Furthermore, there exists a constant C which only depends on the mesh regularity γ of (3.2) such that, for any $u \in H^1_0(D)$ and any edge $e \subset \Gamma$, it holds that

$$\|u - \mathcal{SZ}(u)\|_{L^2(e)} \leq C \left(\frac{H_e}{p_e} \right)^{1/2} |u|_{H^1(\omega_e)} \tag{3.34}$$

where ω_e is the union of all the elements who share a vertex with the edge e, H_e is the length of the edge e and $p_e = \min\{p_K \mid e \subset \partial K\}$.

Proof of PROPOSITION 3.12. The proof falls in two steps. We first estimate the error $u_\Gamma - u_{\Gamma,H,\{N_e\}}$ and next estimate $u_B - u_{B,H,\{M_K\}}$.

Step 1: interface approximation. For the numerical solution $u_{\Gamma,H,\{N_e\}} \in V_{\Gamma,H,\{N_e\}}$, we write, using an integrating by parts over every element K and (3.7), that, for any $w_\Gamma \in V_\Gamma$,

$$\begin{aligned} a(u_{\Gamma,H,\{N_e\}}, w_\Gamma) &= \int_D \left(\nabla u_{\Gamma,H,\{N_e\}}\right)^T A \nabla w_\Gamma \\ &= \sum_{K\in\mathcal{T}_H} \sum_{e\subset\partial K} \int_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right) w_\Gamma \\ &= \sum_{e\subset\Gamma} \int_e w_\Gamma \, J_e\left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right), \end{aligned} \tag{3.35}$$

where $J_e(\psi)$ denotes the jump of a given function ψ across the edge e.

Using Galerkin orthogonality, we deduce that, for any $v_\Gamma \in V_\Gamma$ and any $v_{\Gamma,H,\{N_e\}} \in V_{\Gamma,H,\{N_e\}}$,

$$\begin{aligned} &a(u_\Gamma - u_{\Gamma,H,\{N_e\}}, v_\Gamma) \\ &= a(u_\Gamma - u_{\Gamma,H,\{N_e\}}, v_\Gamma - v_{\Gamma,H,\{N_e\}}) \\ &= a(u_\Gamma, v_\Gamma - v_{\Gamma,H,\{N_e\}}) - a(u_{\Gamma,H,\{N_e\}}, v_\Gamma - v_{\Gamma,H,\{N_e\}}) \\ &= \int_D f\,(v_\Gamma - v_{\Gamma,H,\{N_e\}}) - \sum_{e\subset\Gamma} \int_e (v_\Gamma - v_{\Gamma,H,\{N_e\}})\, J_e\left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right) \end{aligned} \tag{3.36}$$

where, in the last line, we have used the definition of the exact solution u_Γ and (3.35) for $w_\Gamma = v_\Gamma - v_{\Gamma,H,\{N_e\}}$.

We now make a specific choice for $v_{\Gamma,H,\{N_e\}}$, under the additional assumption that $v_\Gamma \in C^0(\overline{D})$. We define a function w on Γ by

$$w = \mathcal{SZ}(v_\Gamma)|_\Gamma + \sum_{e\in\Gamma} \Pi^{e,0}_{N_e}(v_\Gamma - \mathcal{SZ}(v_\Gamma)) \tag{3.37}$$

where, for each edge $e \subset \Gamma$, $\Pi^{e,0}_{N_e}$ is the L^2 projection on the polynomial functions whose values are 0 on the two vertices of the edge and of degree lower or equal than N_e on e (by construction, for any function ψ, $\Pi^{e,0}_{N_e}(\psi)$ is supported on the edge e). In (3.37), $\mathcal{SZ}$ is the Scott-Zhang type interpolant defined in LEMMA 3.23, where we choose, for each element K, the polynomial degree

$$p_K = \min\{N_e \mid e \subset \partial K\}. \tag{3.38}$$

We observe that, by construction, the degrees $\{p_K\}$ satisfy (3.33). Consider indeed two neighbouring elements K and K'. Then, denoting $\tilde{e}$ the edge shared by K and K', we have

$$\frac{p_K}{p_{K'}} = \frac{\min\{N_e \mid e \subset \partial K\}}{\min\{N_{e'} \mid e' \subset \partial K'\}} = \frac{\min\{N_e \mid e \subset \partial K\}}{N_{\tilde{e}}} \frac{N_{\tilde{e}}}{\min\{N_{e'} \mid e' \subset \partial K'\}} \leq \gamma$$

where we have used the property (3.13). We likewise have that $p_K/p_{K'} \geq 1/\gamma$. Since the degrees $\{p_K\}$ satisfy (3.33), we will be in position to use the approximation result (3.34) in the sequel.

We next observe that, on each edge e, w is a polynomial function of degree lower or equal than N_e. This is obviously the case for the second term in (3.37). This is also the case for the first term, which is indeed a polynomial function of degree $p_{K_e^1}$ (resp. $p_{K_e^2}$) on K_e^1 (resp. K_e^2), where K_e^1 and K_e^2 are the two elements sharing the edge e. By construction, $p_{K_e^1} \leq N_e$ and likewise for $p_{K_e^2}$.

Since w is continuous on Γ (because $\mathcal{SZ}(v_\Gamma)$ is continuous on $\overline{D}$ and $\Pi^{e,0}_{N_e}$ is a polynomial that vanishes at the edge boundaries) and smooth on each edge, we can consider its harmonic lifting

$$v_{\Gamma,H,\{N_e\}} = E_D(w) = E_D\Big(\mathcal{SZ}(v_\Gamma)|_\Gamma + \sum_{e\in\Gamma} \Pi^{e,0}_{N_e}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\Big). \tag{3.39}$$

Since w is a polynomial function of degree lower or equal than N_e on any edge e, we observe that w belongs to the approximation space $V_{\Gamma,H,\{N_e\}}$.

For any $v_\Gamma \in C^0(\overline{D}) \cap V_\Gamma$, we thus define $v_{\Gamma,H,\{N_e\}} \in V_{\Gamma,H,\{N_e\}}$ by (3.39). In the sequel of the proof, we bound $v_\Gamma - v_{\Gamma,H,\{N_e\}}$ in various norms, in order to bound (3.36).

Step 1a. To bound the first term of (3.36), we need to estimate $\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(K)}$ for any element $K \in \mathcal{T}_H$. To that aim, we introduce the unique solution z in $H^1_0(K)$ to

$$-\operatorname{div}(A\nabla z) = v_\Gamma - v_{\Gamma,H,\{N_e\}} \quad \text{in } K.$$

Since K is convex and $A \in C^1(\overline{D})$, we get that $z \in H^2(K)$ (see LEMMA 3.22, assertion (i)). Furthermore, we have that $v_\Gamma - v_{\Gamma,H,\{N_e\}} \in H^1(K) \subset L^q(K)$ for any finite $q \geq 2$. The assertion (ii) of Lemma 3.22 hence shows that $z \in W^{2,p}(K)$ for any $2 \leq p < p_K^{\mathrm{Sob}}$. Using Sobolev embeddings, this implies that $z \in C^1(\overline{K})$. Furthermore, since $z = 0$ on each edge, its tangential derivative vanishes on each edge. Using that $\nabla z \in C^0(\overline{K})$, we deduce that ∇z vanishes on each vertex of K.

Using the definition of z, we have

$$\begin{aligned}
&\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|^2_{L^2(K)} \\
&= \int_K (\nabla z)^T A \nabla (v_\Gamma - v_{\Gamma,H,\{N_e\}}) - \sum_{e\subset\partial K} \int_e \left(v_\Gamma - v_{\Gamma,H,\{N_e\}}\right) \nu^T A \nabla z \\
&= -\sum_{e\subset\partial K} \int_e \left(v_\Gamma - v_{\Gamma,H,\{N_e\}}\right) \nu^T A \nabla z,
\end{aligned} \tag{3.40}$$

since both v_Γ and $v_{\Gamma,H,\{N_e\}}$ are harmonic. Since $A \in C^1(\overline{D})$ and $z \in H^2(K)$, we have that $\nu^T A \nabla z \in H^{1/2}(e)$. Furthermore, since $z \in C^1(\overline{K})$ and ∇z vanishes on the vertices of K, we have $\nu^T A \nabla z \in C^0_0(\overline{e})$, where $C^0_0(\overline{e})$ is the set of continuous functions on $\overline{e}$ which vanish at both ends of e.

Introduce

$$s_e(v_\Gamma - v_{\Gamma,H,\{N_e\}}) = \sup_{w\in H^{1/2}(e)\cap C^0_0(\overline{e})} \frac{\int_e \left(v_\Gamma - v_{\Gamma,H,\{N_e\}}\right) w}{\|w\|_{H^{1/2}(e)}}.$$

We deduce from (3.40) and from LEMMA 3.22 that

$$\begin{aligned}
\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|^2_{L^2(K)} &\leq \sum_{e\subset\partial K} s_e(v_\Gamma - v_{\Gamma,H,\{N_e\}}) \left\|\nu^T A \nabla z\right\|_{H^{1/2}(e)} \\
&\leq C_A \sum_{e\subset\partial K} s_e(v_\Gamma - v_{\Gamma,H,\{N_e\}}) \|\nabla z\|_{H^{1/2}(e)} \\
&\leq C_A \sum_{e\subset\partial K} s_e(v_\Gamma - v_{\Gamma,H,\{N_e\}}) \|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(K)},
\end{aligned} \tag{3.41}$$

where C_A only depends on A and the regularity of the mesh. Using our specific choice (3.39), we get

$$\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(K)} \leq C_A \sum_{e\subset\partial K} s_e\Big(v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi^{e,0}_{N_e}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\Big). \tag{3.42}$$

We next write

$$
\begin{aligned}
&s_e\Big(v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi_{N_e}^{e,0}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\Big) \\
&= \sup_{w\in H^{1/2}(e)\cap C_0^0(\overline{e})} \frac{\int_e \left(v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi_{N_e}^{e,0}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\right) w}{\|w\|_{H^{1/2}(e)}} \\
&= \sup_{w\in H^{1/2}(e)\cap C_0^0(\overline{e})} \frac{\int_e \left(v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi_{N_e}^{e,0}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\right) \left(w - \Pi_{N_e}^{e,0}(w)\right)}{\|w\|_{H^{1/2}(e)}} \\
&\le \left\|v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi_{N_e}^{e,0}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\right\|_{L^2(e)} \sup_{w\in H^{1/2}(e)\cap C_0^0(\overline{e})} \frac{\left\|w - \Pi_{N_e}^{e,0}(w)\right\|_{L^2(e)}}{\|w\|_{H^{1/2}(e)}},
\end{aligned}
\tag{3.43}
$$

where the second equality stems from the fact that $\Pi_{N_e}^{e,0}$ is a $L^2(e)$ orthogonal projection. We now bound from above the two factors of (3.43). For the second factor, we first write, using the stability of the projection, that

$$
\forall w \in L^2(e), \quad \left\|w - \Pi_{N_e}^{e,0}(w)\right\|_{L^2(e)} \le \|w\|_{L^2(e)}\,. \tag{3.44}
$$

Second, for any $w \in C_0^0(\overline{e}) \cap H^1(e)$, we have

$$
\|w - \Pi_{N_e}^{e,0}(w)\|_{L^2(e)} \le \|w - I_{N_e}(w)\|_{L^2(e)} \le C\frac{H_e}{N_e}\,|w|_{H^1(e)} \tag{3.45}
$$

where I_{N_e} is the interpolant of degree N_e at the Gauss Lobatto points on e (the last inequality is for instance given in [22, Eq. (5.4.42)]). Note that it is critical here to assume that w vanishes at the two vertices of the edge.

By Sobolev interpolation (see LEMMA 3.16), we deduce from (3.44) and (3.45) that

$$
\forall w \in H^{1/2}(e) \cap C_0^0(\overline{e}), \quad \|w - \Pi_{N_e}^{e,0}(w)\|_{L^2(e)} \le C\sqrt{\frac{H_e}{N_e}}\,\|w\|_{H^{1/2}(e)}. \tag{3.46}
$$

Collecting (3.43), (3.44) and (3.46), we obtain that

$$
s_e\Big(v_\Gamma - \mathcal{SZ}(v_\Gamma) - \Pi_{N_e}^{e,0}(v_\Gamma - \mathcal{SZ}(v_\Gamma))\Big) \le C\sqrt{\frac{H_e}{N_e}}\,\|v_\Gamma - \mathcal{SZ}(v_\Gamma)\|_{L^2(e)}. \tag{3.47}
$$

By using Scott-Zhang type interpolation results (see LEMMA 3.23) and collecting (3.47) and (3.42), we have

$$
\begin{aligned}
\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(K)} &\le C_A \sum_{e\subset\partial K} \sqrt{\frac{H_e}{N_e}}\,\|v_\Gamma - \mathcal{SZ}(v_\Gamma)\|_{L^2(e)} \\
&\le C_A \sum_{e\subset\partial K} \frac{H_e}{\sqrt{N_e\,p_e}}\,|v_\Gamma|_{H^1(\omega_e)}
\end{aligned}
\tag{3.48}
$$

with

$$
p_e = \min\{p_{K_e^1}, p_{K_e^2}\} = \min\{N_{\tilde{e}} \mid \tilde{e} \subset \partial K_e^1 \cup \partial K_e^2\} \tag{3.49}
$$

where K_e^1 and K_e^2 are the elements sharing the edge e and $p_{K_e^1}$ and $p_{K_e^2}$ are the degrees chosen in (3.38) for the construction of the Scott-Zhang type interpolation operator $\mathcal{SZ}$.

We finally infer from (3.48) that

$$\begin{aligned}\left|\int_D f(v_\Gamma - v_{\Gamma,H,\{N_e\}})\right| &\leq \sum_{K\subset\mathcal{T}_H} \|f\|_{L^2(K)}\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(K)} \\ &\leq C_A \sum_{K\subset\mathcal{T}_H} \|f\|_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e}{\sqrt{N_e\, p_e}}\, |v_\Gamma|_{H^1(\omega_e)}\right) \\ &\leq C_A \sqrt{\sum_{K\subset\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e}{\sqrt{N_e\, p_e}}\right)^2} \sqrt{\sum_{K\subset\mathcal{T}_H} |v_\Gamma|^2_{H^1(\omega_K)}} \\ &\leq C_A\, |v_\Gamma|_{H^1(D)} \sqrt{\sum_{K\subset\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e}\right)} \end{aligned} \tag{3.50}$$

where $\omega_K = \cup_{e\subset\partial K}\omega_e$. We have thus bounded the first term of (3.36).

Step 1b. We now consider with the second term of (3.36). Using the $L^2(e)$ stability of the projection operator $\Pi^{e,0}_{N_e}$ and next (3.34), we have

$$\|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(e)} \leq \|v_\Gamma - \mathcal{SZ}(v_\Gamma)\|_{L^2(e)} \leq C\sqrt{\frac{H_e}{p_e}}\, |v_\Gamma|_{H^1(\omega_e)}, \tag{3.51}$$

with p_e again given by (3.49). We are thus in position to bound the second term of (3.36):

$$\begin{aligned}&\left|\sum_{e\subset\Gamma} \int_e (v_\Gamma - v_{\Gamma,H,\{N_e\}}) J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right| \\ &\leq \sum_{e\subset\Gamma} \left\|J_e \left(\nu^T A \nabla u_{\Gamma,H,N}\right)\right\|_{L^2(e)} \|v_\Gamma - v_{\Gamma,H,\{N_e\}}\|_{L^2(e)} \\ &\leq C \sum_{e\subset\Gamma} \sqrt{\frac{H_e}{p_e}}\, \left\|J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|_{L^2(e)} |v_\Gamma|_{H^1(\omega_e)},\end{aligned}$$

where we have used (3.51) in the last line. Using the Cauchy-Schwarz inequality, we deduce that

$$\begin{aligned}&\left|\sum_{e\subset\Gamma} \int_e (v_\Gamma - v_{\Gamma,H,\{N_e\}}) J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right| \\ &\leq C \sqrt{\sum_{e\subset\Gamma} \frac{H_e}{p_e}\, \left\|J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|^2_{L^2(e)}} \sqrt{\sum_{e\subset\Gamma} |v_\Gamma|^2_{H^1(\omega_e)}} \\ &\leq C|v_\Gamma|_{H^1(D)} \sqrt{\sum_{e\subset\Gamma} \frac{H_e}{p_e}\, \left\|J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|^2_{L^2(e)}}.\end{aligned} \tag{3.52}$$

Step 1c. Collecting (3.36), (3.50) and (3.52), we obtain, for any $v_\Gamma \in V_\Gamma \cap C^0(\overline{D})$, that

$$\begin{aligned}\frac{a(u_\Gamma - u_{\Gamma,H,\{N_e\}}, v_\Gamma)}{|v_\Gamma|_{H^1(D)}} \leq C_A \Bigg\{ &\sqrt{\sum_{K\subset\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e}\right)} \\ &+ \sqrt{\sum_{e\subset\Gamma} \frac{H_e}{p_e}\, \left\|J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}}\right)\right\|^2_{L^2(e)}} \Bigg\}.\end{aligned}$$

We use the above estimate for the choice $v_\Gamma = u_\Gamma - u_{\Gamma,H,\{N_e\}}$, which obviously belongs to V_Γ. In addition, we have assumed that $u \in H^s(D)$ for some $s > 3/2$, hence $u \in C^0(\overline{D})$. Since $u_\Gamma = u$ on Γ, this implies that $u_\Gamma \in C^0(\Gamma)$, and thus, by elliptic regularity (see [43, Theorem 8.30]), that $u_\Gamma \in C^0(\overline{K})$ for any element K. Likewise, $u_{\Gamma,H,\{N_e\}}$ is continuous on Γ and thus, using again [43, Theorem 8.30], we get that $u_{\Gamma,H,\{N_e\}} \in C^0(\overline{K})$. We thus indeed check that $v_\Gamma \in C^0(\overline{D})$. We thus deduce that

$$\|u_\Gamma - u_{\Gamma,H,\{N_e\}}\|_E \leq C_A \left\{ \sum_{K\subset\mathcal{T}_H} \|f\|^2_{L^2(K)} \left(\sum_{e\subset\partial K} \frac{H_e^2}{N_e\, p_e} \right) + \sum_{e\subset\Gamma} \frac{H_e}{p_e} \left\| J_e \left(\nu^T A \nabla u_{\Gamma,H,\{N_e\}} \right) \right\|^2_{L^2(e)} \right\}^{1/2}. \quad (3.53)$$

Step 2: bubble approximation. Regarding the bubble approximation, the *a priori* error already provides an *a posteriori* estimator since the right-hand sides in LEMMA 3.9 is independent of u_B. The following arguments yield another estimate, in the case when bubble enrichments are considered (if no enrichement is used, then we simply use the right-hand side of (3.11) as *a posteriori* estimator). We thus consider the case when $M_K \geq 1$ for any element K. For the numerical solution $u_{B,H,\{M_K\}} \in V_{B,H,\{M_K\}}$, we write, using an integrating by parts over every element K, that, for any $w_B \in V_B$,

$$\begin{aligned} a(u_{B,H,\{M_K\}}, w_B) &= \int_D \left(\nabla u_{B,H,\{M_K\}}\right)^T A \nabla w_B \\ &= -\sum_{K\in\mathcal{T}_H} \int_K w_B \ \mathrm{div}\left(A \nabla u_{B,H,\{M_K\}}\right). \end{aligned} \quad (3.54)$$

Using Galerkin orthogonality, we deduce that, for any $v_B \in V_B$ and any $v_{B,H,\{M_K\}} \in V_{B,H,\{M_K\}}$,

$$\begin{aligned} &a(u_B - u_{B,H,\{M_K\}}, v_B) \\ &= a(u_B - u_{B,H,\{M_K\}}, v_B - v_{B,H,\{M_K\}}) \\ &= a(u_B, v_B - v_{B,H,\{M_K\}}) - a(u_{B,H,\{M_K\}}, v_B - v_{B,H,\{M_K\}}) \\ &= \int_D f\,(v_B - v_{B,H,\{M_K\}}) + \sum_{K\in\mathcal{T}_H} \int_K (v_B - v_{B,H,\{M_K\}}) \ \mathrm{div}\left(A\nabla u_{B,H,\{M_K\}}\right) \\ &= \sum_{K\in\mathcal{T}_H} \int_K (v_B - v_{B,H,\{M_K\}}) \left(f + \mathrm{div}\left(A\nabla u_{B,H,\{M_K\}}\right)\right), \end{aligned} \quad (3.55)$$

where, in the fourth line, we have used the definition of the exact solution u_B and (3.54) for $w_B = v_B - v_{B,H,\{M_K\}}$.

We now make the following specific choices. For v_B, we take

$$v_B = u_B - u_{B,H,\{M_K\}}. \quad (3.56)$$

We now choose $v_{B,H,\{M_K\}}$. Let $\Pi^K_{M_K}$ be the $L^2(K)$-projection on the polynomials of degree at most M_K on the element K. We then consider $v_{B,H,\{M_K\}} \in V_{B,H,\{M_K\}}$ such that, on each K, we have

$$-\mathrm{div}(A\nabla v_{B,H,\{M_K\}}) = \Pi^K_{M_K}(z) \ \text{ in } K, \qquad v_{B,H,\{M_K\}} = 0 \ \text{ on } \partial K, \quad (3.57)$$

with

$$z = -\mathrm{div}(A\nabla v_B) \ \text{ in } K.$$

In view of (3.56) and of the definition of u_B, we see that, in K,

$$z = -\operatorname{div}(A\nabla u_B) + \operatorname{div}(A\nabla u_{B,H,\{M_K\}}) = f + \operatorname{div}(A\nabla u_{B,H,\{M_K\}}). \tag{3.58}$$

In view of (3.5), we thus see that the right-hand side $\Pi^K_{M_K}(z)$ in (3.57) satisfies

$$\Pi^K_{M_K}(z) = \Pi^K_{M_K}(f) + \operatorname{div}(A\nabla u_{B,H,\{M_K\}}). \tag{3.59}$$

In the sequel of the proof, we bound $v_B - v_{B,H,\{M_K\}}$ in order to bound (3.55).

Step 2a. We see that

$$-\operatorname{div}(A\nabla(v_{B,H,\{M_K\}} - v_B)) = \Pi^K_{M_K}(z) - z \ \text{ in } K, \qquad v_{B,H,\{M_K\}} - v_B = 0 \ \text{ on } \partial K,$$

hence

$$\alpha_{\min}\|\nabla(v_{B,H,\{M_K\}} - v_B)\|^2_{L^2(K)} \le \left\|\Pi^K_{M_K}(z) - z\right\|_{L^2(K)} \|v_{B,H,\{M_K\}} - v_B\|_{L^2(K)}.$$

Using the Poincaré inequality $\|v_{B,H,\{M_K\}} - v_B\|_{L^2(K)} \le C\, H_K \|\nabla(v_{B,H,\{M_K\}} - v_B)\|_{L^2(K)}$, we deduce that $\alpha_{\min}\|\nabla(v_{B,H,\{M_K\}} - v_B)\|_{L^2(K)} \le C\, H_K \left\|\Pi^K_{M_K}(z) - z\right\|_{L^2(K)}$ and thus

$$\|v_{B,H,\{M_K\}} - v_B\|_{L^2(K)} \le C\,\frac{H_K^2}{\alpha_{\min}} \left\|\Pi^K_{M_K}(z) - z\right\|_{L^2(K)} = C\,\frac{H_K^2}{\alpha_{\min}} \left\|\Pi^K_{M_K}(f) - f\right\|_{L^2(K)}, \tag{3.60}$$

where, for the last equality, we have used (3.58) and (3.59).

Step 2b. Collecting (3.55) and (3.60), and using that v_B is given by (3.56), we get that

$$\begin{aligned}
&a(u_B - u_{B,H,\{M_K\}}, u_B - u_{B,H,\{M_K\}}) \\
&= a(u_B - u_{B,H,\{M_K\}}, v_B) \\
&\le \sum_{K\in\mathcal{T}_H} \|v_{B,H,\{M_K\}} - v_B\|_{L^2(K)} \left\|f + \operatorname{div}\left(A\nabla u_{B,H,\{M_K\}}\right)\right\|_{L^2(K)} \\
&\le \frac{C}{\alpha_{\min}} \sum_{K\in\mathcal{T}_H} H_K^2 \left\|f + \operatorname{div}\left(A\nabla u_{B,H,M}\right)\right\|_{L^2(K)} \left\|\Pi^K_{M_K}(f) - f\right\|_{L^2(K)}.
\end{aligned} \tag{3.61}$$

Since $f \in H^{\ell_K}(K)$ for some $\ell_K \ge 0$, then we know from LEMMA 3.15 that

$$\left\|f - \Pi^K_{M_K}(f)\right\|_{L^2(K)} \le C\,\frac{H_K^{\min(\ell_K, M_K+1)}}{M_K^{\ell_K}} \|f\|_{H^{\ell_K}(K)}. \tag{3.62}$$

Collecting (3.61) and (3.62), we obtain

$$\begin{aligned}
&\|u_B - u_{B,H,\{M_K\}}\|_E \le \\
&\frac{C}{\sqrt{\alpha_{\min}}} \left\{ \sum_{K\in\mathcal{T}_H} H_K^2\, \frac{H_K^{\min(\ell_K, M_K+1)}}{M_K^{\ell_K}} \left\|f + \operatorname{div}\left(A\nabla u_{B,H,\{M_K\}}\right)\right\|_{L^2(K)} \|f\|_{H^{\ell_K}(K)} \right\}^{1/2},
\end{aligned} \tag{3.63}$$

where the right-hand side is completely computable, up to the unknown constant C.

Step 3. Collecting (3.53) and (3.63), and using the orthogonal decomposition (3.4), we obtain (3.14). This concludes the proof of PROPOSITION 3.12. □

CHAPTER 4

MSFEM IMPLEMENTATION IN FREEFEM++

This chapter discusses the implementation of some MsFEM variants in FreeFem++, a Finite Element solver developed by F. Hecht [50].

FreeFem++ is a Finite Element software that can be used for a wide range of applications. MsFEM techniques are by nature intrusive as they require one to know exactly the variational formulation in order to be applied. MsFEM approaches are two step methods: one creates coarse basis functions that are solutions to some local problems and then a Galerkin problem is solved on the space spanned by these functions.

Although the basis functions are similar to the classical finite element basis, the MsFEM approach cannot be implemented in a general fashion. Indeed, the definition of the local problems changes according to the equation to solve. Also, the local problems often cannot be solved analytically and an approximation at a fine scale h has to be made. The basis functions are not as easy to manipulate as Finite Elements functions because they depend on the local fluctuations of the coefficient. Hence, quadrature formulas depend on more parameters than in standard FE where quadrature formulas are simple and require only minimal information on the element.

All these arguments advocate for the implementation of MsFEM approaches as a template form and not as a hard-coded element in the software FreeFem++. In that sense our implementation can be seen as a simple multi-grid approach where we design a coarse approximation space with basis functions that are solutions to local problems approximated on a finer embedded grid of size h.

The implementation will be applied to the multi-scale problem (1.2):

$$\begin{cases} -\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f \text{ in } D, \\ u_\varepsilon = 0 \text{ on } \partial D, \end{cases}$$

for a coefficient A_ε that can be periodic or not.

We present here the implementation of three different MsFEM variants: Linear MsFEM, MsFEM oversampling and MsFEM à la Crouzeix-Raviart. We refer to Section 1.4.2 for the principle of these methods. Also, a template to couple these approaches with standard techniques such as Finite Elements as well as a template to

use MsFEM as a two level additive preconditioner will be presented.

This section is intended to show and explain the structure of the templates and the main guidelines of the implementation. For the sake of clarity, we do not present here the actual scripts associated with our templates but included them in ANNEX A as FREEFEM++ scripts.

4.1 Linear MsFEM and MsFEM oversampling

The Linear and oversampling MsFEM approaches are grouped together as the implementation between the two is basically the same. Indeed, the basis functions are associated with vertex degrees of freedom. The difference lies in the computation and definition of local problems. Here the techniques and definitions can be easily transposed in 3D as only degrees of freedom associated with vertices are involved.

The approach is divided into two steps: the offline phase (described in ALGORITHM 1 implemented in the FREEFEM++ script in SECTION A.1.1) and the online phase (described in ALGORITHM 2 implemented in the FREEFEM++ script in SECTION A.2.1).

Algorithm 1 Linear MsFEM: offline phase

Require: Coarse mesh $\mathcal{T}_H$, the coefficient A_ε and a small meshsize h
Ensure: $h \ll \varepsilon$ (often $h \simeq \frac{\varepsilon}{10}$ is used)
  Initialize A the stiffness matrix of the new basis and B the generic RHS
  for $K \in \mathcal{T}_H$ in parallel **do**
    Build $\mathcal{T}_h^K$ (meshing K with meshsize h)
    for $i \in K$ **do**
      Solve with P1 FE

$$-\operatorname{div}(A_\varepsilon \nabla \psi_i) = 0 \, in \, K, \; \psi_i = \phi_i \, on \, \partial K, \tag{4.1}$$

      with ϕ_i the standard P1 FE basis function and store ψ_i.
    end for
    Compute and store $A_{i,j}^{loc} = \int_K (A_\varepsilon \nabla \psi_i) \cdot \nabla \psi_j$ and $B_i = B_i + \int_K \psi_i \times 1$
    with A^{loc} the local stiffness matrix and B the generic RHS
    Assemble and store A the stiffness matrix associated with the new basis $\{\psi_i\}_{i=1..N_{vertices}}$
  end for

ALGORITHM 1 can be implemented simply in FREEFEM++. Indeed, as the degrees of freedom are based on the vertices of the coarse mesh $\mathcal{T}_H$, the numbering of the degrees of freedom and the structure of the sparse matrices associated with the right-hand side and the stiffness are exactly the same as for a P1 formulation posed on $\mathcal{T}_H$. Hence, we define a P1 dummy problem on $\mathcal{T}_H$ to initialize our matrices, then we parallelize the loop over the elements of our mesh using the Message Passing Interface (MPI) formalism. All computations are performed in parallel. Mesh and local matrices are computed on the fly in order to limit storage load. Basis functions are stored locally in files that are indexed by the element. One could store more information, if needed, such as local stiffness matrix or local mesh used to compute basis functions

for instance. Such partitioning speeds up the reading time needed to recover the basis functions in a parallel framework.

Remark 4.1. *We put the emphasis on the fact that the MsFEM solution is defined only locally. Hence, when solving 4.1 one could use other higher order finite elements instead of P1 FE in order to better adapt to A_ε changes, though in this case the information must be stored in order to recover the basis.*

Remark 4.2. *The MsFEM oversampling approach follows the same steps (see the script in SECTION (A.1.2)). However, in the loop over the element a mesh $\widehat{\mathcal{T}}_h^K$ for the enlarged element is designed instead of a considering $\mathcal{T}_h^K$ a sub-mesh of the element K. On this enlarged mesh or super-element $\widehat{K}$, we compute $\widehat{\psi}_i$ solution of equation (4.1). The boundary conditions are taken as the corresponding P1 coarse FE basis functions associated with the super element $\widehat{K}$. Then, we consider $\mathcal{T}_h^K$ a sub-mesh of our initial element K and the actual new basis functions ψ_i are defined as the restrictions of $\widehat{\psi}_i$ on this mesh.*
There are numerous ways of defining a super-element. In the current implementation, we choose an enlargement factor $C_{over} \geq 1$. Then, the super-element is the polygon defined by the vertices one gets from the initial vertices when multiplying the vector from the barycenter to the vertices by C_{over}. When the element is close to the boundary of D some truncature rule is used on the enlarged element, in order to stay in the domain. This step is crucial as it allows us to simply enforce boundary conditions by imposing values on the degrees of freedom on the boundary.

Algorithm 2 Linear MsFEM: Online phase

Require: f the right-hand side, the MsFEM basis $\{\psi_i\}$, QOF a quantity of interest to compute from the solution u, the stiffness matrix A and the generic RHS B
Ensure: QOF the quantity of interest can be computed locally
  Assemble the right-hand side $RHS_i = B_i \times f(x_i)$
  Solve the coarse linear system $AX = RHS$
  Reconstruct the solution locally:
  for $K \in \mathcal{T}_H$ in parallel **do**
    Load the fine mesh $\mathcal{T}_h^K$
    for $i \in K$ **do**
      Compute and store $U^K_{MsFEM} = U^K_{MsFEM} + X_i \times \psi_i$
    end for
    Compute $QOF_K = G(U^K_{MsFEM})$
  end for
  Return $QOF = \sum_{K \in \mathcal{T}_H} QOF_K$

The online step described by ALGORITHM 2 is more straightforward, compute the right-hand side vector RHS, solve the coarse linear formulation $AX = RHS$, recover the MsFEM approximation and compute the quantity of interest. The most expensive operation here is to compute the right-hand side $RHS_i = \int_D \psi_i f$. There are two options: either the computation can be performed in parallel inducing an extra cost or one can suppose that the RHS function f does not vary much at the coarse scale H and approximate f by $f(x_i)$ with $x_i \in \text{Supp}\phi_i$ in RHS_i. In the linear and oversampling MsFEM approach the degrees of freedom are located on the vertices $\{x_i\}$ and the support of the basis function ψ_i is centered on x_i. Thus, RHS_i can be approximated by $B_i f(x_i)$, a quantity that is less expensive to compute than the full integral RHS_i since it only requires the evaluation of f at the vertices.

Remark 4.3. *In order for the quantity of interest to be computed efficiently, it is required to be local. For instance, considering g an inline function and the associated quantity of interest $QOF = G(u_{MsFEM}) = \int_D g\ u_{MsFEM}$, the computation can be performed in parallel efficiently. However, if we consider g defined on a global FE space, then computing $QOF = G(u_{MsFEM}) = \int_D g\ u_{MsFEM}$ can be challenging as one either would have to project the MsFEM solution on the FE space of G or project G on each local sub-mesh. Both operations are very costly and will impair the MsFEM efficiency.*

Remark 4.4. *Building a global representation of the MsFEM approximation is possible, though not advisable as there will be a costly projection step from a FE space defined on a local mesh to a FE space defined on a global mesh. Such operation must be executed with special care in the oversampling case as the solution is not continuous over the coarse elements. In FreeFem++, an intermediate Discontinuous Galerkin (DG) global FE space has to be used (for linear MsFEM) since the projection element by element will give twice the value on the FE degrees of freedom associated with the edges. One has to transfer element to element the local solution to the DG FE space and then take another projection step (for instance L^2 projection) to have a result in a conformal global FE space.*

4.2 MsFEM à la Crouzeix Raviart

The implementation of MsFEM à la Crouzeix Raviart (in short MsFEM-CR) described in section 1.4.2 seems similar to the linear MsFEM and oversampling MsFEM approaches. However, the degrees of freedom are located on the edges and in the center of the elements instead of the vertices. Thus, the implementation has to be modified.

The method is divided into two parts: the offline phase (described in ALGORITHM 3 and implemented in FREEFEM++ as explained in SECTION A.1.3) and the online phase (described in ALGORITHM 4 and implemented in FREEFEM++ as explained in SECTION A.2.2).

The offline part is similar to ALGORITHM 1. The computation of the new basis will be performed locally in a parallelized loop over the element of the coarse mesh $\mathcal{T}_H$. Note however that the structure and numbering of the basis will be different as we consider degrees of freedom not on the vertices but on the edges and the center of the elements. In order to get the optimal numbering needed to simplify the resolution of the final coarse linear system on the new basis, one has to find a simple equivalent coarse problem similarly to the P1 FE coarse dummy problem in the linear MsFEM case. FREEFEM++ allows us to do that easily by using P0 FE space and P0edge FE space. The P0 FE space corresponds to the space of functions that are piecewise constant over the elements and the P0edge space is generated by an edge based basis $\{\phi_e\}_{e=1..Nb_{edge}}$ where ϕ_e equals to 1 on e, 0 on other edges and ϕ_e is piecewise constant over the elements. Then, we design a coarse dummy problem from P0 FE space and P0edge FE space to initialize our matrices and perform the computation of the new basis in parallel over the elements.

There is also a difference in the design of the basis functions associated with the edges compared to linear and oversampling MsFEM. Indeed, the local problem are not solved with classical boundary conditions (of type Dirichlet) a constraint enforced on edges by a Lagrange multiplier method. Such an operation is performed simply in FREEFEM++ as we have access to all matrices, instead of solving a $Nb_{vertices} \times Nb_{vertices}$ corresponding to a P1 discretization of the element, a $Nb_{vertices} + Nb_{edges} \times Nb_{vertices} + Nb_{edges}$ system is solved where the additional degrees of freedom corresponds to the

Algorithm 3 MsFEM-CR: offline phase

Require: Coarse mesh $\mathcal{T}_H$, the coefficient A_ε and a small meshsize h
Ensure: $h \ll \varepsilon$ (often $h \simeq \frac{\varepsilon}{10}$ is used)
for $K \in \mathcal{T}_H$ in parallel **do**
Build $\mathcal{T}_h^K$ (meshing K with meshsize h)
Solve with P1 FE

$$-\mathrm{div}(A_\varepsilon \nabla \psi_0^K) = 1 \, in \, K, \; \psi_0^K = 0 \, on \, \partial K,$$

with ψ_0^K the basis function associated with the center of K
for $e_i \subset \partial K$ **do**
Solve with P1 FE (constraint enforced with Lagrange multiplier method)

$$-\mathrm{div}(A_\varepsilon \nabla \psi_i^K) = 0 \, in \, K, \; s.t. \int_{e_j} \psi_i^K = \delta_{i,j}$$

with ψ_i the basis function associated with the edge e_i ($i = 1..N_{edge}$).
end for
Compute and store $A_{i,j}^{loc} = \int_K (A_\varepsilon \nabla \psi_i^K) \cdot \nabla \psi_j^K$ and $B_i = \int_K \psi_i^K$, for $i, j = 0..N_{edge}$),
with A^{local} the local stiffness matrix and B the generic RHS
Assemble and store A the stiffness matrix associated with the new basis $\{\psi_i^K\}_{i=0..N_{edge}}$
end for

lagrange multipliers.

We solve for the function basis associated with edge e_i a system of the form

$$\begin{bmatrix} A_K & C_K \\ C_K^T & 0 \end{bmatrix} \begin{bmatrix} X^K \\ \lambda_{ei}^K \end{bmatrix} = \begin{bmatrix} B^K \\ B_{ei}^K \end{bmatrix}$$

where $A = \int_K (A_\varepsilon \nabla \phi_i) \cdot \nabla \phi_j$, $C_{K,i,j} = \int_{e_i} \phi_j$, X_k is the vector of the basis coefficients in the P1 FE basis, $\lambda_{e_i}^K$ are the Lagrange multipliers associate to the edges of K, $B_k = \int_K 0 \times \phi_i = 0$, and $B_{e_i}^K = \delta_{e_i,e_j}$.

Remark 4.5. *When we use the Crouzeix-Raviart MsFEM method, the solution is discontinuous. There is only a weak-continuity property: the mean of the jump is equal to* 0 *on each edge. The edge based basis function associated with edge* e *satisfies this property as the mean is equal to* 1 *on the two element sharing* e *and* 0 *on the other edges.*

The online part described by Algorithm 4 follows up the offline part. In this step, one computes the right-hand side and solve the coarse linear system associated with the new basis. The most expensive tasks are the resolution of the linear system $AX = RHS$ and the assembling of the right-hand side RHS. Contrary to the linear MsFEM approach, approximating RHS is not so easy as the degrees of freedom are located on the edges and on the barycenter of the elements, not on the vertices. In the case where i corresponds to a bubble basis function associated with the element K then $RHS_i \simeq B_i \times f(x_k)$ with x_K the barycenter of K. In the case where i correspond to an edge basis function associated with the edge e_i then $RHS_i \simeq B_i \times f(x_{e_i})$ with

Algorithm 4 MsFEM-CR: Online phase

Require: f the right-hand side, the MsFEM basis $\{\psi_i^K\}$, QOF a quantity of interest to compute from the solution u, the stiffness matrix A and the generic RHS B
Ensure: QOF the quantity of interest can be computed locally
Assemble the right-hand side $RHS_i = B_i \times f(x_i)$
Solve the coarse linear system $AX = RHS$
Reconstruct the solution locally
for $K \in \mathcal{T}_H$ in parallel **do**
  Load the fine mesh $\mathcal{T}_h^K$
  for $i \in 0..N_{edge}$ **do**
    Compute and store $U_{MsFEM}^K = U_{MsFEM}^K + X_i \times \psi_i^K$
  end for
  Compute $QOF_K = G(U_{MsFEM}^K)$
end forreturn $QOF = \sum_{K \in \mathcal{T}_H} QOF_K$

x_{e_i} the middle of the edge e_i. As in the linear MsFEM case, the MsFEM approximation is built locally in parallel to compute the quantity of interest QOF.

4.3 Coupling MsFEM

MsFEM approach is really useful when the coefficient is highly heterogeneous. It is however computationally demanding if the coefficient does not vary much on the coarse mesh, especially when compared to P1 FE for instance. Hence, it would be interesting to distinguish two areas in the domain, $\mathcal{A}_{slow}$ the area where the coefficient A_ε does not vary much and $\mathcal{A}_{fast}$ the area where A_ε is highly heterogeneous. One can see such an example in FIGURE 4.1 where A_ε is oscillating in the middle square and is constant otherwise.

We will consider MsFEM approaches where the degrees of freedom are located on the vertices to simplify the coupling between standard FE functions and MsFEM basis functions. The structure of the implementation is described in ALGORITHM 5 and the corresponding FREEFEM++ script is presented in SECTION A.3.

Algorithm 5 Coupling MsFEM and P1

Require: A_ε the expression of the coefficient, rules to determine if MsFEM basis functions or P1 basis functions are used, h a small meshsize and f a right-hand side
Create $\mathcal{T}_H$ a coarse mesh of the whole domain with meshsize H
Separate the mesh into two submeshes $\mathcal{T}_H^{P1}$ and $\mathcal{T}_H^{MsFEM}$
Renumber the elements of the coarse mesh $\mathcal{T}_H$ according to the two submeshes $\mathcal{T}_H^{P1}$ and $\mathcal{T}_H^{MsFEM}$
Build the coarse stiffness matrix $A_H = \int_D (A_\varepsilon \nabla\phi_i) \cdot \nabla\phi_j$ and the right-hand side $RHS = \int_D \phi_i f$ with ϕ_i the P1 basis function associated with vertex i of $\mathcal{T}_H$
for $K \in \mathcal{T}_H^{MsFEM}$ in parallel **do**
  Build or load the $MsFEM$ basis functions associated with the vertices
  Update A_H with local MsFEM Stiffness matrix and remove P1 FE part
  Update RHS with local MsFEM right-hand side and remove P1 FE part
end for
Solve the coarse linear problem $A_H X = RHS$

The first step is to separate the mesh into two areas according to the variation of A_ε. FreeFem++ allows to do that easily, one meshes the whole domain first and then defines a submesh by using a separation function f_{sep}.
For instance, if $f_{sep}(x,y) = \mathbb{1}_{0.25<x<0.75}(x,y) \times \mathbb{1}_{0.25<y<0.75}(x,y)$ then we get in the submesh all the elements of $\mathcal{T}_H$ where $f(x,y) > 0$ see (Figure 4.1).

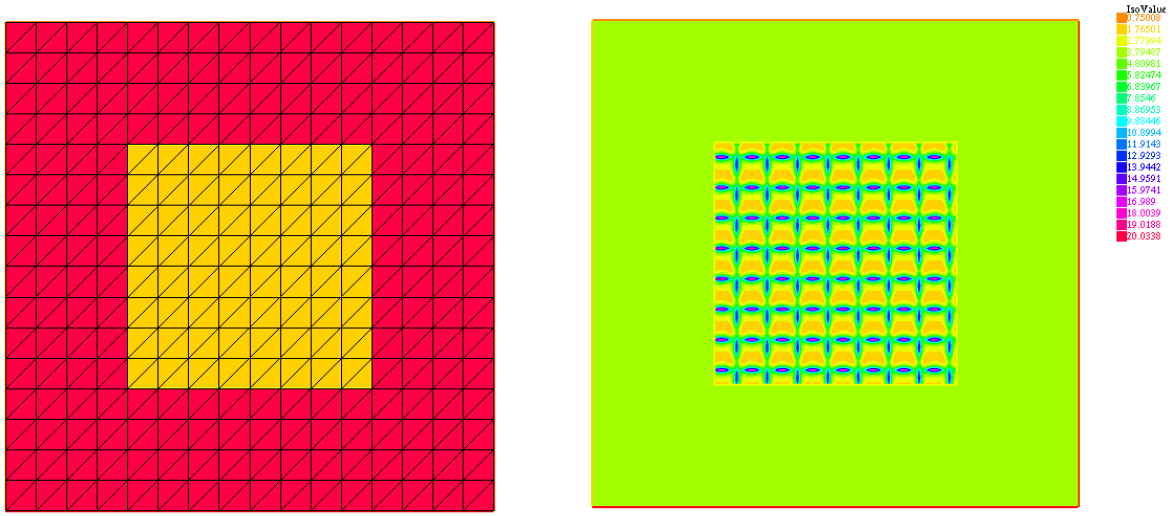


Figure 4.1: Left: Mesh of the domain ($H = 1/16$): we use both MsFEM functions (in orange) and P1 FE functions (in red). Right: coefficient A_ε considered

Then, once the areas have been defined by the two submeshes $\mathcal{T}_H^{P1}$ and $\mathcal{T}_H^{MsFEM}$, we can precompute MsFEM basis function on $\mathcal{T}_H^{MsFEM}$. Finally, the mesh of the whole domain is reconstructed by rearranging the numbering such that the elements where P1 will be used come before the elements where MsFEM basis function are used. Regarding the elements on the interface, as we have chosen to associate degrees of freedom to vertices, the corresponding basis function will simply be the P1 basis function on P1 elements and the MsFEM basis function in the MsFEM elements, see Figure 4.2.

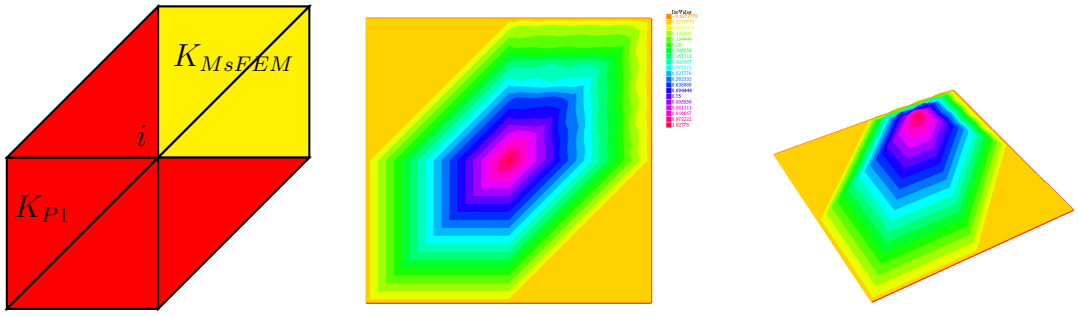


Figure 4.2: Basis functions ϕ_i associated with interface degrees of freedom: Left - elements where ϕ_i is supported (in the red element, ϕ_i is a P1 basis function; in the yellow element, ϕ_i is a MsFEM basis function), Right - plot of basis function associated with degrees of freedom i

For the assembling of the stiffness matrix, we will compute it for the whole P1 FE problem on the mesh $\mathcal{T}_H$ and modify it locally only on the elements where MsFEM will be used. The same procedure can be used for the right-hand side term. Then we

can solve the coarse linear system associated with the coupled problem and get our approximation.

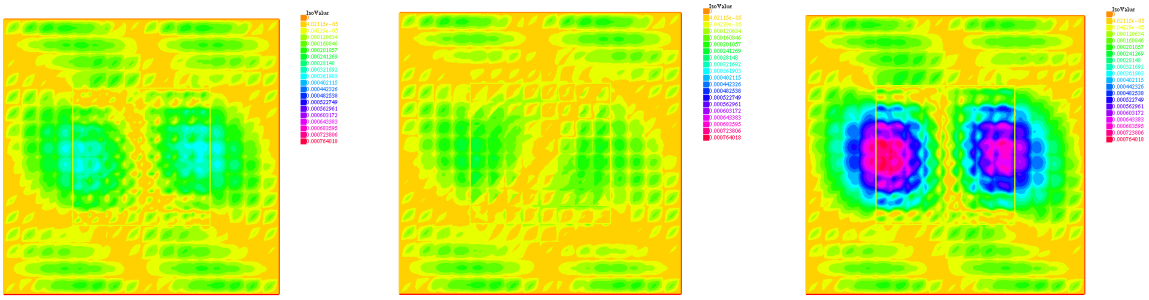


Figure 4.3: Graph of error for P1 coupled with MsFEM lin (Left, relative H^1 error 28.3%), P1 coupled with MsFEM oversampling (Middle, relative H^1 error 23.5%) and P1 (Right relative H^1 error 34%)

We consider A_ε described in FIGURE 4.1 and choose $H = 1/16$. Then, we used MsFEM linear approach coupled with P1 FE and compare it to a reference solution. The relative error in H^1 norm when the MsFEM oversampling approach is used on the whole domain is 22.8%. We can see in FIGURE 4.3 that the error of our coupled approach is close to this error and smaller than the case where only P1 FE are used.

Remark 4.6. *Choosing the separation function according to A_ε fluctuations is not an easy problem, especially when the coefficient admits rough changes (from constant to a rapidly oscillating function as on* FIGURE *4.1). Indeed, the separation must be chosen carefully since the error can be large across the separation in some cases (see* FIGURE *4.5 corresponding to the choices in* FIGURE *4.4). Recalling that the MsFEM linear basis functions correspond to P1 FE basis functions in case of constant coefficient, one could make the MsFEM part larger to account for such changes. Also, using oversampling MsFEM approach significantly mitigates this effect.*

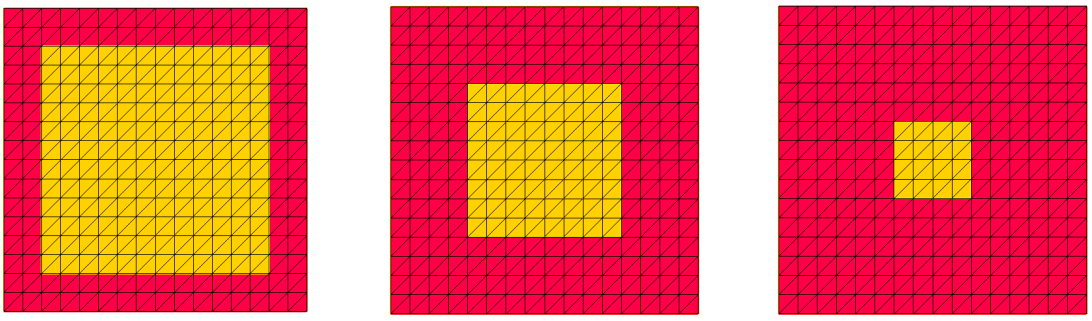


Figure 4.4: Separation of the domain into the P1 subdomain (red) and a large (left), medium (center) or small (right) MsFEM subdomain

Remark 4.7. *The coupled method using MsFEM linear approach is conformal and continuous even in the degrees of freedom located on the interface since the MsFEM basis functions share the same values as P1 FE on the edges. This approach can also be used with an oversampling MsFEM approach since the degrees of freedom are located on the vertices too. However, the approximation is no longer conformal even for the degrees of*

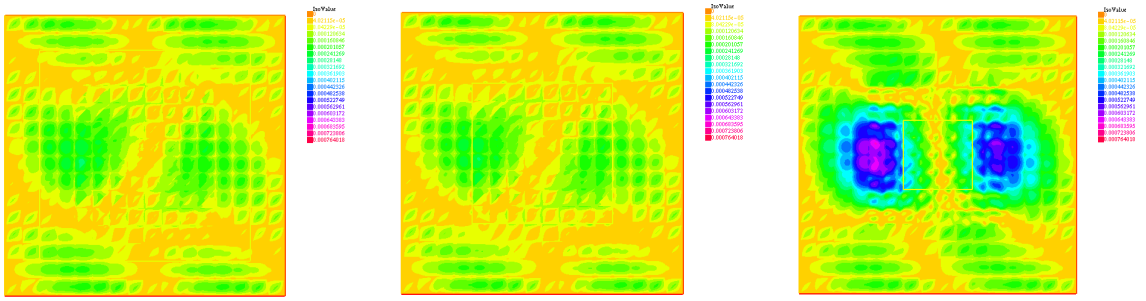


Figure 4.5: Error corresponding to a large MsFEM subdomain (Left relative H^1 error 23.5%), a medium MsFEM subdomain (Middle relative H^1 error 23.8%) and small MsFEM subdomain (Right relative H^1 error 31.1%)

freedom located on the interface between P1 and MsFEM areas since MsFEM oversampling basis functions are not linear on the edges. A functional FREEFEM++ *template has been designed for both methods.*

Remark 4.8. *We decided to consider a simple coupling where the degrees of freedom are located on the vertices. One could consider other choices, for instance Discontinuous Galerkin methods. Such approaches may be more effective, though it complicates the coupling formulation and increases the computational load needed to solve the coarse problem.*

Such work is still at a preliminary stage. Only a Freefem++ implementation for MsFEM linear and MsFEM oversampling coupled with P1 FE is available (see SECTION A.3). Currently, up to our knowledge, no error analysis has been performed and such analysis is out of the scope of this discussion.

4.4 Using MsFEM approach as a second level preconditioner

In this section, first we will do a quick review on the main tools to solve large linear systems. Then, a two level preconditioner implementation using MsFEM linear basis function as a coarse space will be presented.

The aim here is to approximate u solution to (1.2) by u_h the P1 approximation at the small scale $h \ll \varepsilon$. To that end, we define a fine grid $\mathcal{T}_h$. We then we have to solve the corresponding linear system:

$$AX = B,\ A_{i,j} = \int_D \left(A_\varepsilon \nabla \phi_i^h\right) \cdot \nabla \phi_j^h,\ B_i = \int_D \phi_i^h f, \tag{4.2}$$

with $\{\phi_i^h\}_{i=1..Nb_{vertices}}$ the P1 basis functions associated with the vertices on the mesh $\mathcal{T}_h$

.

In the multi-scale applications we have to consider $h \ll \varepsilon$, hence matrices A and B have large dimension typically $N_{rows} \simeq 1/h^2$ in 2D.

Techniques to solve linear systems fall into two categories:

- Direct solvers: solve the linear system exactly by using linear algebra. We can cite for instance the LU decomposition, Cholesky decomposition, Gaussian elimination.
- Iterative solvers: solve the system by refining an initial guess X_0 with an iterative scheme that reduces at each step the error $\|AX_i - B\|$ in a relevant space. Conjugate Gradient (CG) and Generalized minimal residual method (GMRES) fall into that category.

For a more thorough review on linear system solvers one can refer to the monographs [69] and [30].

Direct linear solvers, such as the LU decomposition, are less efficient in our case as the complexity and memory required is usually in $O(N_{rows}^2)$. Although direct solvers are very accurate, in most of multi-scale applications they cannot be applied as the memory requirement is too high.

To circumvent this issue iterative solvers such as conjugate gradient (used for symmetric definite positive matrices) or GMRES (used in more general cases) are used. Instead of solving the problem directly, the solution is updated at each step with only vector product operations. Though the methods is sure to converge after N_{rows} steps, this would induce a cost larger than when using of direct solvers. These methods are all the more effective the number of steps needed to get an accurate approximation is small. The number of steps which is needed is related to κ_A, the condition number of the matrix A, that is the ratio between its highest eigenvalue and lowest eigenvalue (if A is symmetric positive definite matrix). Usually, κ_A increases when h decreases. In order to mitigate this effect, we precondition the problem: instead of solving $AX = B$ we solve $M^{-1}AX = M^{-1}B$ and we hope that $\kappa_{M^{-1}A} \ll \kappa_A$. The best preconditioner is A^{-1}, in such case the associated condition number is 1 and the problem is solved in one step: $X = A^{-1}B$. Obviously, we do not have access to A^{-1} as it would require to have already solved the problem. So the goal is to find a preconditioner that is as close to A^{-1} as possible and that is fast to compute. One can notice that the preconditioner method is independent from the right-hand side B, hence the same preconditioner can be used in a multi-query context.

Usually in decomposition domain methods (DDM), the preconditioner M^{-1} are chosen to be solutions to local problems (inversion of small matrices) that are linked together to invert a reasonably small global problem. One of the simplest is called the Jacobi preconditioner defined by

$$M_{J,i,j} = \begin{cases} A_{i,j}, \text{ if } i = j \\ 0 \text{ if } i \neq j \end{cases}$$

Then M_J^{-1} is just the inverse of the diagonal of A. More complicated preconditioners are used in practice such as Block Jacobi preconditioning and Schwarz preconditioning. Often preconditioners from DDM are really effective in the detection of local features (called high frequency features) but are not as efficient in identifying more global patterns (low frequencies features). Hence, in practice a more efficient preconditioner is built by mixing coarse information and local information to get low and high frequencies. This is called a second level preconditioner: a local high frequency preconditioner from DDM is enriched by a coarse approximation of the solution that is fast to compute.

In the Jacobi case and considering a coarse space $V_H = \mathrm{Span}(\{\phi_i\}, i = 1..n)$ with $n \ll N_{rows}$, we define R_0 the matrix of size $n \times N_{rows}$ such that the row i is the coefficient of ϕ_i on the P1 basis functions associated with the fine mesh $\mathcal{T}_h$. We can compute the preconditioner M_C^{-1} defined by:

$$M_C^{-1} = R_0^T (R_0 A R_0^T)^{-1} R_0 + M_J^{-1} \tag{4.3}$$

The operation seems costly as one has to invert $(R_0 A R_0^T)^{-1}$. However, such matrix is of size $n \times n$ with $n \ll N_{rows}$ making the inversion affordable. Moreover, in practice one always apply the preconditioner to a vector and this operation is equivalent to solve a linear system.

In our case, the coarse MsFEM space defined as the span of our basis function seems to be a good candidate for a second level preconditioner coarse space because of its good approximation properties. The implementation of M_C^{-1} using MsFEM linear approach is described in ALGORITHM 6 with the actual implementation in SECTION A.4.

Algorithm 6 Linear MsFEM: Online phase

Require: $\mathcal{T}_h$ a fine mesh, f the right-hand side, the MsFEM basis functions $\{\psi_i\}$ expressed in the P1 FE basis associated with $\mathcal{T}_h$.
Assemble the linear system $AX = B$ on the fine mesh $\mathcal{T}_h$
Load R_0 the transformation matrix between the MsFEM basis functions and the P1 FE basis functions associated with the fine mesh $\mathcal{T}_h$
Load A_{MsFEM} the stiffness matrix associated with the MsFEM basis functions
Define the preconditioner function $P(Y) \mapsto M_C^{-1} Y$
with $M_C^{-1} = R_0^T A_{MsFEM}^{-1} R_0 + M_J^{-1}$
Apply an iterative solver preconditioned by P to $AX = B$

The implementation in FREEFEM++ is straightforward: one use ALGORITHM 1 to get the MsFEM basis functions expressed in the P1 FE basis associated with the fine mesh $\mathcal{T}_h$. Then, we get A the stiffness matrix and B the right-hand side associated with the fine discretization h. In FREEFEM++ one does not give the matrix M_C^{-1} as an argument of an iterative solver command. Indeed, the user has to define P, a preconditioning function defined by $P(Y) = M^{-1}Y$ with Y a vector having the same size as B (usually, at each step operations of type $M_C^{-1}(AX_i - B)$ have to be computed). Such an implementation is useful as since the preconditioner is allowed to change with the current number of iterations allowing adaptive schemes. Moreover, fwhen small number of iterations are considered then $M_C^{-1}Y$ can be seen as solving $MX = Y$ that is less expensive than computing M_C^{-1} (complexity of $O(n^2)$ instead of $O(n^3)$). Finally, the iterative solver is used and gives the result when the prescribed accuracy in relative residual error is reached or when the maximal number of steps is exceeded.

Considering a periodic A_ε where $\varepsilon = 1/32$ with $D = (0,1)^2$, we test this approach for $H = 1/8$, $h = 1/256$ by using linear MsFEM basis function as a coarse preconditioner and a GMRES iterative solver. It turns out that compared to a simple Jacobi preconditioning, our approach requires approximately three times less steps to converge when the tolerance is set to 10^{-6} see FIGURE 4.6.

Such a method is not intended to compete with more refined preconditioning approaches usually seen in DDM like additive Schwarz or Block Jacobi methods. Indeed,

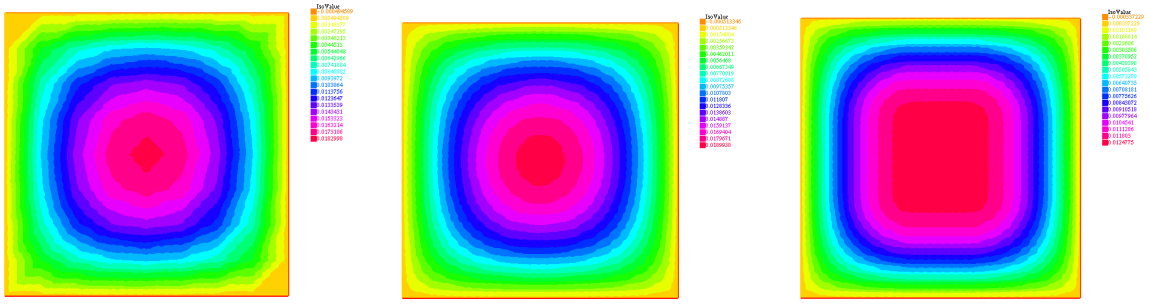


Figure 4.6: Left - Coarse MsFEM approximation ($H = 1/8$), Middle - fine approximation obtained with GMRES preconditioned by our approach where convergence is achieved (150 iterations, $h = 1/256$), Right - fine approximation obtained with GMRES preconditioned by Jacobi after 150 iterations and $h = 1/256$.

the aim here is to show how one can easily design and improve a preconditioned approach within FREEFEM++ framework by using MsFEM approaches. The analysis regarding the reduction of the condition number of the preconditioned problem and the convergence is out of the scope of this discussion.

BIBLIOGRAPHY

[1] G. Allaire. *Shape optimization by the homogenization method*, volume 146 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2002.

[2] G. Allaire and M. Amar. Boundary layer tails in periodic homogenization. *ESAIM Control Optim. Calc. Var.*, 4:209–243, 1999.

[3] D. N. Arnold, L. R. Scott, and M. Vogelius. Regular inversion of the divergence operator with dirichlet boundary conditions on a polygon. *Annali della Scuola Normale Superiore di Pisa-Classe di Scienze*, 15(2):169–192, 1988.

[4] M. Avellaneda and F.-H. Lin. Compactness methods in the theory of homogenization. *Comm. Pure and Applied Math.*, 40(6):803–847, 1987.

[5] I. Babuška and M. Suri. The hp version of the finite element method with quasiuniform meshes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 21(2):199–238, 1987.

[6] I. Babuška and M. Suri. The p and h-p versions of the finite element method, basic principles and properties. *SIAM review*, 36(4):578–632, 1994.

[7] G. Bal, J. Garnier, Y. Gu, and W. Jing. Corrector theory for elliptic equations with long-range correlated random potential. *Asymptotic Analysis*, 77(3-4):123–145, 2012.

[8] G. Bal, J. Garnier, S. Motsch, and V. Perrier. Random integrals and correctors in homogenization. *Asymptotic Analysis*, 59(1-2):1–26, 2008.

[9] P. Bella, B. Fehrman, J. Fischer, and F. Otto. Stochastic homogenization of linear elliptic equations: Higher-order error estimates in weak norms via second-order correctors. *SIAM J. Math. Anal.*, 49(6):4658–4703, 2017.

[10] J. K. Bennighof and R. B. Lehoucq. An automated multilevel substructuring method for eigenspace computation in linear elastodynamics. *SIAM J. Sci. Comput.*, 25(6):2084–2106, 2004.

[11] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 374. American Mathematical Soc., 2011.

[12] C. Bernardi and Y. Maday. *Approximations spectrales de problemes aux limites elliptiques*, volume 10. Springer, 1992.

[13] C. Bernardi and Y. Maday. Spectral, spectral element and mortar element methods. In *Theory and Numerics of Differential Equations*, pages 1–57. Springer, 2001.

[14] X. Blanc, R. Costaouec, C. Le Bris, and F. Legoll. Variance reduction in stochastic homogenization using antithetic variables. *Markov Processes and Related Fields*, 18(1):31–66, 2012.

[15] X. Blanc, M. Josien, and C. Le Bris. Precised approximations in elliptic homogenization beyond the periodic setting. *arXiv preprint arXiv:1812.07220*, 2018.

[16] X. Blanc, M. Josien, and C. Le Bris. Approximation locale précisée dans des problèmes multi-échelles avec défauts localisés. *Comptes Rendus Mathematique*, 357(2):167–174, 2019.

[17] X. Blanc, C. Le Bris, and P.-L. Lions. On correctors for linear elliptic homogenization in the presence of local defects. *Communications in Partial Differential Equations*, 43(6):965–997, 2018.

[18] X. Blanc, F. Legoll, and A. Anantharaman. Asymptotic behavior of Green functions of divergence form operators with periodic coefficients. *Applied Mathematics Research eXpress*, 2013(1):79–101, 2013.

[19] A. Bourgeat and A. Piatnitski. Estimates in probability of the residual between the random and the homogenized solutions of one-dimensional second-order operator. *Asymptotic Analysis*, 21(3-4):303–315, 1999.

[20] F. Bourquin. *Synthèse modale et analyse numérique des multistructures élastiques*. PhD thesis, Université Paris VI, 1991.

[21] F. Bourquin. Component mode synthesis and eigenvalues of second order operators: Discretization and algorithm. *Math. Model. Numer. Anal.*, 26:385–423, 1992.

[22] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods*. Springer, 2006.

[23] I. Catto, C. Le Bris, and P.-L. Lions. *The mathematical theory of thermodynamic limits: Thomas-Fermi type models*. Oxford University Press, 1998.

[24] R. M. Christensen. *Mechanics of composite materials*. Courier Corporation, 2012.

[25] P. Clément. Approximation by finite element functions using local regularization. *Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique*, 9(R2):77–84, 1975.

[26] R. Costaouec, C. Le Bris, and F. Legoll. Approximation numérique d'une classe de problèmes en homogénéisation stochastique (Numerical approximation of a class of problems in stochastic homogenization). *C.R. Acad. Sci. Paris, Série I*, 348(1-2):99–103, 2010.

[27] G. Dal Maso. *An introduction to* Γ*-convergence*, volume 8. Springer Science & Business Media, 2012.

[28] E. De Giorgi. Sulla convergenza di alcune successioni d'integrali del tipo dell'area. *Ennio De Giorgi*, 414, 1975.

[29] M. Defranceschi and C. Le Bris. *Mathematical models and methods for ab initio quantum chemistry*, volume 74. Springer Science & Business Media, 2012.

[30] V. Dolean, P. Jolivet, and F. Nataf. *An introduction to domain decomposition methods: algorithms, theory, and parallel implementation*, volume 144. SIAM, 2015.

[31] M. Duerinckx, A. Gloria, and F. Otto. The structure of fluctuations in stochastic homogenization. *arXiv preprint arXiv:1602.01717*, Feb. 2016 version.

[32] W. E and B. Engquist. The heterogeneous multiscale methods. *Commun. Math. Sci.*, 1(1):87–132, 2003.

[33] W. E, P. Ming, and P. Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *Journal of the American Mathematical Society*, 18(1):121–156, 2005.

[34] Y. Efendiev and T. Hou. *Multiscale Finite Element Methods: Theory and Applications*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer New York, first edition, 2009.

[35] Y. R. Efendiev, T. Y. Hou, and X.-H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM Journal on Numerical Analysis*, 37(3):888–910, 2000.

[36] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[37] F. Feyel and J.-L. Chaboche. FE2 multiscale approach for modelling the elastoviscoplastic behaviour of long fibre SiC/Ti composite materials. *Computer methods in applied mechanics and engineering*, 183(3-4):309–330, 2000.

[38] J. Fischer. Private communication.

[39] J. Fischer. The choice of representative volumes in the approximation of effective properties of random materials. *Archive for Rational Mechanics and Analysis*, 234(2):635–726, 2019.

[40] M. J. Gander, A. Loneland, and T. Rahman. Analysis of a new harmonically enriched multiscale coarse space for domain decomposition methods. *arXiv preprint arXiv:1512.05285*, 2015.

[41] K. Gao, S. Fu, and E. T. Chung. A high-order multiscale finite-element method for time-domain acoustic-wave modeling. *Journal of Computational Physics*, 360:120–136, 2018.

[42] P. Gervasio, E. Ovtchinnikov, and A. Quarteroni. The spectral projection decomposition method for elliptic equations in two dimensions. *SIAM J. Numer. Anal.*, 34(4):1616–1639, 1997.

[43] D. Gilbarg and N. Trudinger. *Elliptic partial differential equations of second order*. Springer, 2015.

[44] V. Girault and P.-A. Raviart. Finite Element Methods for Navier-Stokes Equations. *Springer Ser. Comput. Math*, 5, 1986.

[45] A. Gloria. An analytical framework for numerical homogenization. part II: Windowing and oversampling. *Multiscale Modeling & Simulation*, 7(1):274–293, 2008.

[46] A. Gloria, S. Neukamm, and F. Otto. Quantification of ergodicity in stochastic homogenization: optimal bounds via spectral gap on Glauber dynamics. *Inventiones mathematicae*, 199(2):455–515, 2015.

[47] P. Grisvard. *Elliptic problems in nonsmooth domains*. Pitman, 1985.

[48] P. Grisvard. *Singularities in boundary value problems*, volume 22. Springer, 1992.

[49] Z. Hashin. Analysis of composite materials—a survey. *Journal of Applied Mechanics*, 50(3):481–505, 1983.

[50] F. Hecht. New development in FreeFem++. *J. Numer. Math.*, 20(3-4):251–265, 2012.

[51] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Modeling & Simulation*, 11(4):1149–1175, 2013.

[52] U. Hetmaniuk and A. Klawonn. Error estimates for a two-dimensional special finite element method based on component mode synthesis. *Electron. Trans. Numer. Anal*, 41:109–132, 2014.

[53] U. L. Hetmaniuk and R. B. Lehoucq. A special finite element method based on component mode synthesis. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(3):401–420, 2010.

[54] U. Hornung. Models for flow and transport through porous media derived by homogenization. In *Environmental Studies*, pages 201–221. Springer, 1996.

[55] T. Hou and X. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134:169–189, 1997.

[56] T. Hou, X. 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.

[57] T. J. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy. The variational multiscale method—a paradigm for computational mechanics. *Computer methods in applied mechanics and engineering*, 166(1-2):3–24, 1998.

[58] V. Jikov, S. Kozlov, and O. Oleinik. *Homogenization of differential operators and integral functionals*. Springer-Verlag, Berlin, 1994.

[59] W. Jing. Limiting distribution of elliptic homogenization error with periodic diffusion and random potential. *Analysis & PDE*, 9(1):193–228, 2016.

[60] M. Josien. Decomposition and pointwise estimates of periodic Green functions of some elliptic equations with periodic oscillatory coefficients. *Asymptotic Analysis*, 112:227–246, 2019.

[61] T. Kato, M. Mitrea, G. Ponce, and M. Taylor. Extension and representation of divergence-free vector fields on bounded domains. *Mathematical Research Letters*, 7(5/6):643–650, 2000.

[62] R. Kornhuber, D. Peterseim, and H. Yserentant. An analysis of a class of variational multiscale methods based on subspace decomposition. *Mathematics of Computation*, 87(314):2765–2774, 2018.

[63] C. Le Bris, F. Legoll, and A. Lozinski. MsFEM à la Crouzeix-Raviart for Highly Oscillatory Elliptic Problems. *Chinese Annals of Mathematics, Series B*, 34(1):113–138, Jan 2013.

[64] C. Le Bris, F. Legoll, and A. Lozinski. An MsFEM type approach for perforated domains. *Multiscale Modeling & Simulation*, 12(3):1046–1077, 2014.

[65] C. Le Bris, F. Legoll, and W. Minvielle. Special quasirandom structures: A selection approach for stochastic homogenization. *Monte Carlo Methods and Applications*, 22(1):25–54, 2016.

[66] C. Le Bris, F. Legoll, and F. Thomines. Rate of convergence of a two-scale expansion for some "weakly" stochastic homogenization problems. *Asymptotic Analysis*, 80(3-4):237–267, 2012.

[67] E. Lieb and B. Simon. The Thomas-Fermi theory of atoms, molecules and solids. *Advances in Mathematics*, 23(1):22–116, 1977.

[68] J.-L. Lions and E. Magenes. *Problèmes aux limites non homogènes et applications*. Dunod, 1968.

[69] F. Magoulès, F.-X. Roux, and G. Houzeaux. *Parallel scientific computing*. Wiley Online Library, 2016.

[70] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Mathematics of Computation*, 83(290):2583–2603, 2014.

[71] W. McLean and W. C. H. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge university press, 2000.

[72] J. Melenk. HP–interpolation of non–smooth functions. *Newton Institute Preprint NI03050-CPD, Cambridge*, 2003.

[73] F. Murat and L. Tartar. H-convergence. In *Topics in the mathematical modelling of composite materials*, pages 21–43. Springer, 2018.

[74] S. Nemat-Nasser and M. Hori. *Micromechanics: overall properties of heterogeneous materials*, volume 37. Elsevier, 2013.

[75] G. C. Papanicolaou and S. R. S. Varadhan. Boundary value problems with rapidly oscillating random coefficients. In *Random fields, Vol. I, II (Esztergom, 1979)*, volume 27 of *Colloq. Math. Soc. János Bolyai*, pages 835–873. North-Holland, Amsterdam-New York, 1981.

[76] R. Pasquetti and F. Rapetti. Spectral element methods on unstructured meshes: which interpolation points? *Numerical Algorithms*, 55(2-3):349–366, 2010.

[77] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, first edition, 1999.

[78] S. Sauter and C. Schwab. *Boundary element methods*, volume 39 of *Springer Series in Computational Mathematics*. Springer, 2010.

[79] S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611, 1965.

[80] A. Toselli and O. Widlund. *Domain decomposition methods – algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer, 2005.

APPENDIX A

CODES

This annex presents the actual implementation in FreeFem++ of the methods presented in Chapter 4. The codes can be run on the latest FreeFem++ version (FreeFem++ 4.0) except for the transposition operator " ' " that has been replaced by " ^ T " for aesthetic purposes in the present manuscript (it shows more relevant syntax highlighting).

A.1 Offline step: Basis creation

A.1.1 Linear MsFEM

```
1  // P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria
       Matherials
2
3  // Create the basis functions for MsFEM linear approach
4
5  // run with command line: mpirun -np number_of_processor
       FreeFem++-mpi  basis_creation_parallel.edp
6
7  verbosity=0;
8
9  // MPI
10 mpiComm comm(mpiCommWorld,0,0);
11
12 int nbproc = mpiSize(comm); // number of processes in
       parallel
13 int iproc = mpiRank(comm); //current processes
14
15
16 int H=256; // size of coarse mesh
17 int h=1024; // size of fine mesh (ideally h<oscillation
       length /10)
18 int nsplit=h/H; // ratio between coarse and fine mesh
19
20
21 // Create the directory to store the basis element by element
22 if(iproc==0) // only the first process create the directory
```

```
{
    string Createrep="mkdir -p basis_repository";
    exec(Createrep);
}
mpiBarrier(comm); // wait for process one to finish


// Definition of Mesh and FE coarse space
mesh TH=square(H,H,[x,y]); // Coarse global mesh
int nbtri=TH.nt;

fespace Tri(TH,P0); // P0 on coarse mesh
fespace P1Tri(TH,P1); // P1 on coarse mesh

// Definition of oscillating coefficient
real eps = 1./32.; // size of oscillations
// coef of PDE :- div(aeps grad u)=f
func aeps =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps))
   + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps)));

// macro for the coefficients
macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
macro Grad(u) [dx(u),dy(u)] //

// Creation of the basis, storing the Stiffness matrix and
   right-hand sides

// using coarse problem formulation
varf vA(u,v)= int2d(TH)(u*v*0 )+on(1,2,3,4,u=0);// bilinear
   form of the coarse P1 problem
varf vB(unused,v)= on(1,2,3,4,unused=0);// RHS for the coarse
    P1 problem

matrix Ai=vA(P1Tri,P1Tri); // Stiffness matrix for the coarse
    P1 pb
real[int] RHSi=vB(0,P1Tri); // RHS for the coarse P1 pb
real[int] RHS=RHSi;
matrix A=Ai;

Tri ChiK; // function P0 to mark element i

// Loop to compute the linear MSFEM basis for the ith element
   ,
for (int i = 0; i < nbtri; i++)
{

     //bool to test if processor iproc deal element i
    bool elemtest= (iproc==i%nbproc); // Bool to assign
       Element i to process iproc according modulo
    if(elemtest) // If triangle assigned to current process
    {
        // Generating fine submesh of element i
        mesh THK, ThK;
```

```
69          ChiK[][i]=1; // P0 function used to mark the element
               i
70          THK=trunc(TH,ChiK>0.1,split=1);  // a mesh made of
               only one triangle
71          ThK=trunc(THK,1,split=nsplit); // each triangle
               refined by nsplit ratio
72          ChiK[][i]=0; // reinitialize the function
73
74          // Computing the basis
75          fespace VKH(THK,P1);
76          fespace VKh(ThK,P1);
77
78          VKh[int] uki(3);
79
80           // One file per element to store basis
81          string namesol="./basis_repository/basis_element"+i+"
               .txt";
82          //emptying the file
83          {
84              ofstream storebasis(namesol);
85          }
86
87          // Computing the basis functions associated to
               vertices
88          for (int j=0;j<3;j++)
89          {
90              // P1 coarse function to set the BC for the basis
                   function
91              P1Tri Test;
92              Test[][P1Tri(i,j)]=1; // set the P1 coarse
                   fonction for the the vertex j
93              varf vAK(u,v)=int2d(ThK)(Aeps(u,v))+on(1,2,3,4,u=
                   Test); // Stiffness expression for the fine
                   MSFEM pb
94              varf vBK(u,v)= on(1,2,3,4,u=Test); // RHS
                   expression for the fine MSFEM pb
95
96              real [int] bk=vBK(0,VKh); // RHS  for the fine
                   MSFEM pb
97              matrix AK=vAK(VKh,VKh); // Stiffness matrix for
                   the fine MSFEM pb
98
99              set(AK,solver=UMFPACK);
100             uki[j][]=AK^-1*bk; // solve the linear system
101
102
103             {
104             ofstream storebasis(namesol,append);
105             storebasis << uki[j][] << endl; // store basis
                   associated to vertex j of element i
106             }
107
108         }
109
```

```
110         // Assembling the stiffness matrix and RHS vector for
                the global problem
111         varf Stiffloc(u,v)= int2d(ThK)(Aeps(u,v) );
112         varf rhsloc(u,v)= int2d(ThK)(v);
113         matrix KK=Stiffloc(VKh,VKh);
114         real[int] RHSloc=rhsloc(0,VKh);
115
116         // Double loop on dof MSFEM to compute the local
               stiffness matrix and RHS
117         for (int j=0;j<3;j++)
118             {
119                 // P1 element to remove the P1 Stiffness part
                        (no splitting of DOF, DOF are vertices)
120                 real bMsFEM=uki[j][]^T*RHSloc;  // RHS MSFEM-
                       L coarse
121                 int I=P1Tri(i,j); // global numbering of
                       vertex j
122                 for (int l=0;l<3;l++)
123                     {
124                         int J=P1Tri(i,l); // global numbering
                                of vertex l
125                         real[int] dummy= KK*uki[l][];
126
127                         real KMsFEM=uki[j][]^T*dummy; //
                                local stiffness MSFEM
128                         Ai(I,J)=Ai(I,J)+KMsFEM; // new
                                stiffness
129                     }
130                 int btestj=(TH[i][j].label>0); // test if dof
                        ( vertex i) on boundary
131                 if (!btestj)
132                      {
133                         RHSi[I]+=bMsFEM; // change in global
                                RHS
134                      }
135             }
136
137
138     }
139 }
140
141 // Post process of the result concatenate results from all
       processes
142
143 mpiReduce(RHSi,RHS,processor(0,comm),mpiSUM); // transfer
       RHSi from all processes in RHS for process 0
144 mpiReduce(Ai, A,processor(0,comm),mpiSUM); // transfer Ai
       from all processes in A for process 0
145
146 // Storing the basis, Stiffness matrix and RHS performed only
        by process 0
147 if (mpiRank(comm)==0)
148 {
149
```

```
150      cout << "Post process of the result storing the basis,
            Stiffness matrix and RHS" << endl;
151      string Stiffnessmatrix="Stiffness_matrix.txt";
152      {
153          ofstream save(Stiffnessmatrix);
154          save << A << endl;
155      }
156
157     string Rhsvector="Rhs_vector.txt";
158     {
159          ofstream rhsstore(Rhsvector);
160          int nvect=RHS.n;
161          rhsstore << nvect << endl;
162          rhsstore<< RHS << endl;
163     }
164     cout << "Basis and matrices stored" << endl;
165 }
```

A.1.2 Oversampling MsFEM

```
1  // P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria
       Matherials
2
3  // Create the basis functions for MsFEM oversampling approach
4
5  // run with command line: mpirun -np number_of_processor
       FreeFem++-mpi  basis_creation_parallel.edp
6  // or ff-mpirun -np  number_of_processor
       basis_creation_parallel.edp
7
8  verbosity=0;
9
10 // MPI
11 mpiComm comm(mpiCommWorld,0,0);
12
13 int nbproc = mpiSize(comm); // number of processes in
       parallel
14 int iproc = mpiRank(comm); //current processes
15
16 int H=64; // size of coarse mesh
17 int h=1024; // size of fine mesh (ideally h<oscillation
       length /10)
18 int nsplit=h/H; // ratio between coarse and fine mesh
19 real coeffover=2; // oversampling coefficient
20
21 // Create the directory to store the basis element by element
22 if(iproc==0) // only the first process create the directory
23 {
24     string Createrep="mkdir -p basis_repository";
25     exec(Createrep);
26 }
27 mpiBarrier(comm); // wait for process one to finish
```

```
28
29
30  // Definition of Mesh and FE coarse space
31  mesh TH=square(H,H,[x,y]); // Coarse global mesh
32  int nbtri=TH.nt;
33
34
35
36  fespace Tri(TH,P0); // P0 on coarse mesh
37  fespace P1Tri(TH,P1); // P1 on coarse mesh
38
39  // Definition of oscillating coefficient
40  real eps = 1./32.; // size of oscillations
41  // coef of PDE :- div(aeps grad u)=f
42  func aeps =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps))
       + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps)));
43
44  // macro for the coeffcients
45  macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
46  macro Grad(u) [dx(u),dy(u)] //
47
48  // Creation of the basis, storing the Stiffness matrix and
       right-hand sides
49
50  // using coarse problem formulation
51  varf vA(u,v)= int2d(TH)(u*v*0 )+on(1,2,3,4,u=0);// bilinear
       form of the coarse P1 problem
52  varf vB(unused,v)= on(1,2,3,4,unused=0);// RHS for the coarse
        P1 problem
53
54  matrix Ai=vA(P1Tri,P1Tri); // Stiffness matrix for the coarse
        P1 pb
55  real[int] RHSi=vB(0,P1Tri); // RHS for the coarse P1 pb
56  real[int] RHS=RHSi;
57  matrix A=Ai;
58
59  Tri ChiK; // function P0 to mark element i
60  // Loop to compute the linear MSFEM basis for the ith element
61  for (int i = 0; i < nbtri; i++)
62  {
63
64      //bool to test if processor iproc deal element i
65      bool elemtest= (iproc==i%nbproc);
66      if(elemtest) // one element is dispatched on one
           processor according to bool value
67      {
68          // Generating fine submesh of element i
69          mesh THK, ThK;
70          ChiK[][i]=1; // P0 function used to mark the element
               i
71          THK=trunc(TH,ChiK>0.1,split=1);  // a mesh made of
               only one triangle
72          ThK=trunc(THK,1,split=nsplit); // each triangle
               divived by nsplit
```

```
        ChiK[][i]=0; // reinitialize the function


        // Computing the basis
        fespace VKH(THK,P1);
        fespace VKh(ThK,P1);

        VKh[int] uki(3);// MsFEM local basis 3 dof vertices
           of triangles

        // One file per element
        string namesol="./basis_repository/basis_element"+i+"
           .txt";
        //emptying the file
        {
            ofstream storebasis(namesol);
        }

        //generating oversampling mesh
        // triangle i vertices
        real[int] Xtriold(3);
        real[int] Ytriold(3);
        // os triangle vertices
        real[int] Xtri(3);
        real[int] Ytri(3);

        real Xb,Yb;
        for (int j=0;j<3;j++)
        {
            int I=P1Tri(i,j);
            Xtriold[j]=TH(I).x;
            Ytriold[j]=TH(I).y;
            // compute barycenter of triangle
            Xb=Xb+1./3.*Xtriold[j];
            Yb=Yb+1./3.*Ytriold[j];
        }

        for (int j=0;j<3;j++)
        {
            // expand triangle by barycenter
            Xtri[j]=Xb+(Xtriold[j]-Xb)*coeffover;
            Ytri[j]=Yb+(Ytriold[j]-Yb)*coeffover;
            // bool to test if new point out of the boundary
            int booltest=(Xtri[j]<0)+(Xtri[j]>1) + (Ytri[j
               ]<0)+ (Ytri[j]>1);
            if (booltest)
            {
            // if out then the point remains the same
            Xtri[j]=Xtriold[j];
            Ytri[j]=Ytriold[j];
            }

        }
```

```
124         // Definition of the local oversampling mesh
125         border as(t=0,1){x=t*(Xtri[1]-Xtri[0])+Xtri[0];y=t*(
               Ytri[1]-Ytri[0])+Ytri[0];label=8;};
126         border bs(t=0,1){x=t*(Xtri[2]-Xtri[1])+Xtri[1];y=t*(
               Ytri[2]-Ytri[1])+Ytri[1];label=9;};
127         border cs(t=0,1){x=t*(Xtri[0]-Xtri[2])+Xtri[2];y=t*(
               Ytri[0]-Ytri[2])+Ytri[2];label=10;};
128         real mult=1;
129         mesh ThOS=buildmesh(as(mult*ceil(coeffover)*nsplit)+
               bs(mult*ceil(coeffover)*nsplit)+cs(mult*ceil(
               coeffover)*nsplit));
130         // More refined than h to keep the same accuracy in
               the local mesh
131         mesh THOS=buildmesh(as(1)+bs(1)+cs(1));
132
133
134         fespace VHOS(THOS,P1); // coarse oversampling FE
135         fespace VhOS(ThOS,P1);// fine oversampling FE
136
137
138
139         // fine FE vector storing the local basis functions
140         VhOS[int] ukios(3);
141
142         // Loop to compute the MSFEM OS basis for the ith
               element
143         for (int j=0;j<3;j++)
144         {
145             // P1 coarse function to set the BC for the basis
                    function
146             VHOS Test;
147             Test[][j]=1; // set the P1 coarse fonction for
                   the the vertex j
148
149             varf vAKos(u,v)=int2d(ThOS)(Aeps(u,v))+on(8,9,10,
                   u=Test); // Stiffness expression for the fine
                   MSFEM OS pb
150             varf vBKos(u,v)= on(8,9,10,u=Test); // RHS
                   expression for the fine MSFEM OSpb
151             real [int] bk=vBKos(0,VhOS); // RHS  for the fine
                    MSFEM OS pb
152             matrix AK=vAKos(VhOS,VhOS); // Stiffness matrix
                   for the fine MSFEM OS pb
153             set(AK,solver=UMFPACK);
154             ukios[j][]=AK^-1*bk; // solve the linear system
155          }
156
157         // Rebuilding the solution on the old mesh
158
159         real[int,int] val(3,3);
160         for(int s=0; s<3; s++)
161         {
162             for(int t=0; t<3; t++)
163             {
```

```
                val(t,s)=ukios[s](Xtriold[t],Ytriold[t]);
        }
    }
    matrix vmat=val;
    set(vmat, solver=UMFPACK);

    // MSFEM os solution satisfying phii(xj)=deltaij on
       element i
    VhOS [int] ukiosnew(3);
    for(int s=0; s<3; s++)
       {
           real[int] e(3), c(3);
           e=0; e[s]=1;
           c=vmat^-1*e;
           ukiosnew[s]=0;
           for(int k=0; k<3; k++)
               {
                    ukiosnew[s]=ukiosnew[s]+c[k]*ukios[k
                       ];
               }
            VKh dummy=ukiosnew[s];// interpolate sol on
               local fine mesh of element i
            uki[s][]=dummy[];
               {
                    ofstream storebasis(namesol,append);
                    storebasis << uki[s][] << endl; //
                       store basis
                }

        }


    // Assembling the Stiffness matrix and RHS vector for
        the global problem
    varf Stiffloc(u,v)= int2d(ThK)(Aeps(u,v) );
    varf rhsloc(u,v)= int2d(ThK)(v);
    matrix KK=Stiffloc(VKh,VKh);
    real[int] RHSloc=rhsloc(0,VKh);

    // Double loop on dof MSFEM to compute the local
       Stiffness matrix and RHS
    for (int j=0;j<3;j++)
        {
            // P1 element to remove the P1 Stiffness part
                (no splitting of DOF, DOF are vertices)
            real bMsFEM=uki[j][]^T*RHSloc;  // RHS MSFEM-
               OS coarse
            int I=P1Tri(i,j); // global numbering of
               vertex j
            for (int l=0;l<3;l++)
                {
                    int J=P1Tri(i,l); // global numbering
                        of vertex l
                    real[int] dummy= KK*uki[l][];
```

```
208
209                         real KMsFEM=uki[j][]'T*dummy; //
                               local Stiffness MSFEM
210                         Ai(I,J)=Ai(I,J)+KMsFEM; // new
                               Stiffness
211                     }
212                 int btestj=(TH[i][j].label>0); // test si dof
                        du bord pour vertex i
213
214                 if (!btestj)
215                  {
216                     RHSi[I]+=bMsFEM; // change in global
                            RHS
217                  }
218             }
219     }
220 }
221
222 // Post process of the result concatenate results from all
    processes
223
224 mpiReduce(RHSi,RHS,processor(0,comm),mpiSUM); // transfer
    RHSi from all processes in RHS for process 0
225 mpiReduce(Ai, A,processor(0,comm),mpiSUM); // transfer Ai
    from all processes in A for process 0
226
227 // Storing the basis, Stiffness matrix and RHS performed only
     by process 0
228 if (mpiRank(comm)==0)
229 {
230
231      cout << "Post process of the result storing the basis,
            Stiffness matrix and RHS" << endl;
232      string Stiffnessmatrix="Stiffness_matrix.txt";
233      {
234         ofstream save(Stiffnessmatrix);
235         save << A << endl;
236      }
237
238     string Rhsvector="Rhs_vector.txt";
239     {
240         ofstream rhsstore(Rhsvector);
241         int nvect=RHS.n;
242         rhsstore << nvect << endl;
243         rhsstore<< RHS << endl;
244     }
245     cout << "Basis and matrices stored" << endl;
246 }
```

A.1.3 MsFEM Crouzeix-Raviart

```
1 // Pierre-Loik Rothe CERMICS/Inria Matherials
```

```
2
3  // Create the basis functions for MsFEM Crouzeix Raviart
       approach
4
5  // run with command line: mpirun -np number_of_processor
       FreeFem++-mpi  basis_creation_parallel.edp
6
7  // MPI
8  mpiComm comm(mpiCommWorld,0,0);
9
10 int nbproc = mpiSize(comm); // number of processes in
       parallel
11 int iproc = mpiRank(comm); //current processes
12
13
14 // discretization parameters
15 int H=16; // size of coarse mesh
16 int h=256; // size of fine mesh (ideally h<oscillation length
       /10)
17 int nsplit=h/H; // ratio between coarse and fine mesh
18
19
20 // Create the directory to store the basis element by element
21 if(iproc==0) // only the first process create the directory
22 {
23     string Createrep="mkdir -p basis_repository";
24     exec(Createrep);
25 }
26 mpiBarrier(comm); // wait for process one to finish
27
28
29 // Definition of Mesh and FE coarse space
30 mesh TH=square(H,H,[x,y]); // Coarse global mesh
31 int nbtri=TH.nt;
32
33 fespace P1Tri(TH,P1); // P1 on coarse mesh
34 fespace Tri(TH,P0); // P0 on coarse mesh
35 fespace P0P0edge(TH,[P0,P0edge]); // Our MsFEM basis has the
       same properties as P0-P0edge bubble and edge functions
36 P0P0edge [utest,vtest];
37
38
39 // Definition of oscillating coefficient
40 real eps = 1./64.; // size of oscillations
41 // coef of PDE :- div(aeps grad u)=f
42 func aeps =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps))
       + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps)));
43
44 // macro for the coefficients
45 macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
46 macro Grad(u) [dx(u),dy(u)] //
47
48 // P1 and P0 functions to navigate triangular element more
       easily
```

```
49 Tri ChiK=0;
50 P1Tri TestK=0;
51
52 // Creation of the basis, storing the Stiffness matrix and
      right-hand sides
53 varf VZERO(u,v)=int2d(TH)(0.*u*v);
54 matrix A=VZERO(P0P0edge,P0P0edge);  // used to define the
      skeleton of the stiffness matrix of the MsFEM solution
55 matrix Ai=A; // matrix for process i
56 real[int] RHS(P0P0edge.ndof); // structure of the coarse rhs
      vector for the global solution
57 real[int] RHSi=RHS; // RHS for process i
58
59 func bordglobal=(x==0)+(x==1.)+(y==0)+(y==1.);
60 // function 1 on dirichlet bc and 0 elsewhere
61
62 // Loop on triangles
63 for (int i = 0; i < nbtri; i++)
64 {
65
66     //bool to test if processor iproc deal with element i
67     bool elemtest= (iproc==i%nbproc); // Bool to assign
         Element i to process iproc according modulo function
68     if(elemtest) // If triangle assigned to current process
69     {
70         // Generating fine submesh of coarse element i
71         mesh THK, ThK;
72         ChiK[][i]=1; // P0 function used to mark the element
            i
73         THK=trunc(TH,ChiK>0.1,split=1);  // a mesh made of
            only one triangle
74         ThK=trunc(THK,1,split=nsplit); // each triangle
            divived by nsplit
75         ChiK[][i]=0; // reinitialize the function
76
77         // Computing the basis
78         fespace VKh(ThK,P1);
79         fespace VKH(THK,P0edge);
80
81         VKh[int] uki(4); // MsFEM CR local basis, 4 DOF : 3
            edges and one bubble
82
83         // One file per element
84         string namesol="./basis_repository/basis_element"+i+"
            .txt";
85         //emptying the file
86         {
87             ofstream storebasis(namesol);
88         }
89
90         // Construction of expressions and matrices needed
91         real tgv=1e30; // penalization to enforce Dirichlet
            BC (as on() function)
```

```
92          varf vAK(u,v)=int2d(ThK)(Aeps(u,v))+int1d(ThK)(tgv*u*
               v*bordglobal); // Stiffness expression for the
               fine MSFEM pb
93          varf vCedge(u,v)=-int1d(ThK)(u*v); // Lagrange part
               for the fine bubble and edge pb
94          varf vBK(u,v)= int2d(ThK)(1.*v); // RHS expression
               for the fine MSFEM bubble pb
95
96          matrix CK=vCedge(VKH,VKh); // Lagrange part to set
               integral value on the edges
97          matrix Asub=vAK(VKh,VKh); // Stiffness matrix for the
                fine MSFEM pb
98
99          // Computation of the bubble part
100         {
101             // right hand side for lagrange multiplier
                   integral on the edges is 0
102             real[int] bedge(3);
103             bedge=0;
104
105             real [int] bsub=vBK(0,VKh); // RHS for the fine
                   MSFEM pb Bi =int(fvi)
106             real[int] bk=[bsub, bedge]; // Global RHS for
                   solution and lagrange multipliers
107
108             matrix AK=[ [Asub,CK],[CK^T,0]]; // Global
                   Stiffness matrix for solution and lagrange
                   multipliers
109             set(AK,solver=UMFPACK);
110             real[int] xx(bk.n);
111             // Solving the linear system
112             xx=AK^-1*bk;
113
114             // Storing solution
115             [uki[0][],bedge]=xx;
116             {
117                 ofstream storebasis(namesol,append);
118                 storebasis << uki[0][] << endl; // store
                       basis
119             }
120         }
121
122         //Loop on the edges to compute the edge basis
               functions
123         for (int j=0;j<3;j++)
124         {
125             // P0 edge function on the coarse triangle to
                   locate the edge
126             VKH Test;
127             Test[][j]=1;
128
129
130             // Testing if the edge is on the global interface
                    where u=0 (homogeneous Dirichlet bc)
```

```
131             real valb=abs(int1d(THK)(bordglobal*Test)-int1d(
                    THK)(Test));
132             int boolbord=(valb<1e-15);
133             real[int] bedge(3); // edge basis is null on the
                    global interface or satisfy L phi i=0 et
                    intedgej=deltaij
134             if (boolbord)
135             {
136                 uki[j+1]=0; // if edge is part of boundary do
                        nothing and edge function 0
137             }
138             else
139             {
140                 varf vBK(u,v)=-int1d(THK)(v*Test); // RHS
                        expression for the lagrange multiplier int
                        edgej phii =deltaij
141                 bedge=vBK(0,VKH); // Rhs lagrange vector
142                 //bedge[j]=1; // could also set value to 1
                        instead of H
143                 real [int] bsub(VKh.ndof); // RHS for the
                        fine edge MSFEM pb(f=0 here)
144                 real[int] bk=[bsub, bedge]; // Global RHS
145
146                 matrix AK=[ [Asub,CK],[CK^T,0]]; // Global
                        Stiffness matrix
147                 set(AK,solver=UMFPACK);
148                 real[int] xx(bk.n);
149                 xx=AK^-1*bk; // solve the linear system
150                 [uki[j+1][],bedge]=xx;
151             }
152             {
153                 ofstream storebasis(namesol,append);
154                 storebasis << uki[j+1][] << endl; // Storing
                        basis
155             }
156         }
157
158         // Assembling the Stiffness matrix and RHS vector for
                the global problem
159         varf vAK2(u,v)= int2d(ThK)(Aeps(u,v) );
160         matrix KK=vAK2(VKh,VKh);  // Matrix to compute
                integral for local stiffness matrices
161
162         // Loop on local MsFEM basis to compute RHS and local
                Stiffness matrix
163         for (int j=0;j<4;j++)
164             {
165                 real bMsFEM=int2d(ThK)(uki[j]*(1.));  // RHS
                        with f=1 and assembling will be carried
                        out by value of f on the coarse level
166                 int I=P0P0edge(i,j); // global numbering of
                        local MsFEM function j
167                 for (int l=0;l<4;l++)
168                     {
```

```
                    int J=P0P0edge(i,l); // global numbering
                        of local MsFEM function l

                    real[int] Kinter=KK*uki[l][];
                    real KMsFEM=uki[j][]^T*Kinter; // local
                        stiffness MSFEM term
                    Ai(I,J)=Ai(I,J)+KMsFEM; // new stiffness
                    }
                // eliminating edge functions that are on the
                    global interface
                if (j>0)
                {
                    // location of the edge
                    VKH Test;
                    Test[][j-1]=1;
                    // test if edge on the global interface
                    real valb=abs(int1d(THK)(bordglobal*Test)
                        -int1d(THK)(Test));
                    int boolbord=(valb<1e-15);

                    if(boolbord)
                    {
                        Ai(I,I)=tgv; // Penalization if edge
                            on the interface
                    }

                }
                RHSi[I]+=bMsFEM; // change in global RHS
            }

            }
}

// Post process of the result concatenate results from all
    processes

mpiReduce(RHSi,RHS,processor(0,comm),mpiSUM); // transfer
    RHSi from all processes in RHS for process 0
mpiReduce(Ai, A,processor(0,comm),mpiSUM); // transfer Ai
    from all processes in A for process 0

// Storing the basis, Stiffness matrix and RHS performed only
     by process 0
if (mpiRank(comm)==0)
{

    cout << "Post process of the result storing the basis,
        Stiffness matrix and RHS" << endl;
    string Stiffnessmatrix="Stiffness_matrix.txt";
    {
        ofstream save(Stiffnessmatrix);
        save << A << endl;
    }

```

```
213     string Rhsvector="Rhs_vector.txt";
214     {
215         ofstream rhsstore(Rhsvector);
216         int nvect=RHS.n;
217         rhsstore << nvect << endl;
218         rhsstore<< RHS << endl;
219     }
220     cout << "Basis and matrices stored" << endl;
221 }
```

A.2 Online step: Computing approximation

A.2.1 Linear MsFEM and oversampling MsFEM

The degrees of freedom are the same in the formulation of both MsFEM linear and MsFEM oversampling approaches. Hence, once the basis functions are computed, they are used the same way. This is why only one script to compute the associated MsFEM approximation is presented here.

```
1  // P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria
      Matherials
2
3  // Compute the linear MsFEM solution and compute in parallel
      a QOI and points to plot the solution
4  // run with command mpirun -np 4 FreeFem++-mpi
      MsFEM_computation_parallel.edp
5  // with np number of processes
6
7  verbosity=0;
8
9  // files to load matrices needed
10 string Stiffnessmatrix="Stiffness_matrix.txt";
11 string Rhsvector="Rhs_vector.txt";
12
13 // MPI definition
14 mpiComm comm(mpiCommWorld,0,0);
15
16 int nbproc = mpiSize(comm); // number of processes
17 int iproc = mpiRank(comm); // number of the current process
18
19 // File to store the solution in x y value form element by
      element
20 if(iproc==0)
21 {
22     string Createrep="mkdir -p solution_repository";
23     exec(Createrep);
24 }
25
26 // Mesh parameters
27 // Size of mesh
28 int H=256; // size of coarse mesh
29 int h=1024; // size of fine mesh
```

```
int nsplit=h/H; // ratio between coarse and fine mesh

mesh TH=square(H,H,[x,y]); // Coarse global mesh
int nbtri=TH.nt; // number of MsFEM triangles


fespace P1Tri(TH,P1);
fespace Tri(TH,P0); // P0 on coarse mesh

// Definition of coefficient
real eps = 1./32.; // size of oscillations
func aeps =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps))
   + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps)));  // coef
   of PDE :- div(aeps grad u)=f

// macro to simplify variational formulation
macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
macro Grad(u) [dx(u),dy(u)] // EOM


// Right-hand side
func f=-1; // function that defines RHS
P1Tri frhs=f; // for coarse P1 resolution (not compulsory)

// Function to compute on the fly local meshes
Tri ChiK=0;
P1Tri TestK=0;


// Assembling the whole Stiffness matrix and the RHS for
   MsFEM basis

cout << "loading Stiffness matrix" << endl;
matrix Aglobal;
ifstream Kmat(Stiffnessmatrix);
Kmat >> Aglobal ;

int number;
ifstream RHSm(Rhsvector);
RHSm >> number;
real[int] Rhsglob(number);
RHSm >> Rhsglob;
real[int] xx=Rhsglob;

cout << "Matrices loaded" << endl;

cout << "MsFEM solution computing" << endl;
// Solving the global linear system P1- MSFEM
set(Aglobal,solver=UMFPACK);
Rhsglob=Rhsglob.*frhs[]; // new RHS with coefficients
   corresponding to values of f
xx=Aglobal'-1*Rhsglob;
cout << "MsFEM solution computed" << endl;

```

```
cout << "MsFEM fine solution reconstructing" << endl;
cout << "Reconstructing the fine solution " << endl;


real Qoi=0; // quantity of interest to compute
real Qoiglobal=0; // quantity of interest to compute
// Loop over the triangles to compute QOI in parallel and to store local values for display
  for(int k=0;k<nbtri;++k)
  {
    bool elemtest= (iproc==k%nbproc);
    if(elemtest)
    {
        mesh THK, ThK; // local fine mesh
        ChiK[][k]=1; // P0 function used to mark the element i
        THK=trunc(TH,ChiK>0.1,split=1);  // a mesh made of only one triangle
        ThK=trunc(THK,1,split=nsplit); // each triangle divived by nsplit
        ChiK[][k]=0;

        fespace VKh(ThK,P1); // fine P1 space

        VKh uloc=0; // local fine solution
        VKh[int] usol(3);
        // loading msfem basis on element i
        string basisstore="./basis_repository/basis_element"+k+".txt";
        {
            ifstream readbasis(basisstore);
            for(int i=0;i<3;i++)
            {
                readbasis>> usol[i][];
            }
        }
        // loop on MsFEM dof (3 vertices)

        for(int i=0;i<3;++i)
          {
            real ugi=xx[P1Tri(k,i)]; // coeff of the solution for the msfem basis i of element k
            real[int] dummy=usol[i][]; // loading msfem basis i of element k
            dummy=dummy*ugi;
            uloc[]=uloc[]+dummy;// build local solution
          }


        // Store the solution in x y value form
        string solstore="./solution_repository/sol_element"+k+".txt";
        {
            ofstream writesol(solstore);
```

```
126             for(int j=0;j<VKh.ndof;j++)
127             {
128                 // save coordinates for all points in the
                       fine mesh
129                 real Xs=ThK(j).x;
130                 real Ys=ThK(j).y;
131                 real valsol=uloc(Xs,Ys);
132                 writesol<< Xs << " " << Ys << " " << valsol
                       << " " << endl;
133             }
134
135         }
136         // Compute quantity of interest for instance energy
137         real dum=0.5*int2d(ThK)(Aeps(uloc,uloc))-int2d(ThK)(f
               *uloc);
138
139         Qoi=Qoi+dum;
140     }
141
142    }
143 //cout << Qoi << endl;
144 // Sum all contribution of QOI from processes to QOIglobal of
        process
145 mpiReduce(Qoi,Qoiglobal,processor(0,comm),mpiSUM);
146
147 if(iproc==0)
148 {
149 // display QOI
150 cout << "Quantity of interest = " << Qoiglobal << endl;
151 }
```

A.2.2 MsFEM Crouzeix-Raviart

```
1  // P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria
       Matherials
2
3  // Compute the CR MsFEM solution and compute in parallel a
       QOI and points to plot the solution
4  // run with command mpirun -np 4 FreeFem++-mpi
       MsFEM_computation_parallel.edp
5
6  verbosity=0;
7
8  // files to load matrices needed
9  string Stiffnessmatrix="Stiffness_matrix.txt";
10 string Rhsvector="Rhs_vector.txt";
11
12 // MPI definition
13 mpiComm comm(mpiCommWorld,0,0);
14
15 int nbproc = mpiSize(comm); // number of processes
16 int iproc = mpiRank(comm); // number of the current process
```

```
17
18  // File to store the solution in x y value form element by
        element
19  if(iproc==0)
20  {
21      string Createrep="mkdir -p solution_repository";
22      exec(Createrep);
23  }
24
25  // Mesh parameters
26  // Size of mesh
27  int H=16; // size of coarse mesh
28  int h=256; // size of fine mesh
29  int nsplit=h/H; // ratio between coarse and fine mesh
30
31  mesh TH=square(H,H,[x,y]); // Coarse global mesh
32  int nbtri=TH.nt; // number of MsFEM triangles
33
34
35  fespace P1Tri(TH,P1);
36  fespace Tri(TH,P0); // P0 on coarse mesh
37  fespace P0P0edge(TH,[P0,P0edge]); // Space for MsFEM solution
        as bubble and edge functions behave as P0-P0edge
38
39  // Definition of coefficient
40  real eps = 1./32.; // size of oscillations
41  func aeps =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps))
        + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps)));  // coef
        of PDE :- div(aeps grad u)=f
42
43  // macro to simplify variational formulation
44  macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
45  macro Grad(u) [dx(u),dy(u)] // EOM
46
47
48  // Right-hand side
49  func f=cos(x)*cos(y); // function that defines RHS
50  P1Tri frhs=f; // for coarse P1 resolution (not compulsory)
51  real[int] u1(P0P0edge.ndof); // coefficient for RHS as int(
        phimsfemi*1) already computed
52  P0P0edge [utest,u2test]; // function to locate DOF and
        compute coefficient for RHS
53
54  // loop on each dof to compute coefficient
55  for (int i=0;i<P0P0edge.ndof;i++)
56  {
57      utest[][i]=1; // Function linked mimicking dof i
58      real alpha=int2d(TH)((utest+u2test)*f)/int2d(TH)(utest+
            u2test); // compute the value associated to f
            representative of DOF considered
59      //(for bubble mean over triangle and edge mean over the
            edge)
60      u1[i]=alpha;
61      utest[][i]=0;
```

```
}


// Function to compute on the fly local meshes
Tri ChiK=0;
P1Tri TestK=0;

// Assembling the whole Stiffness matrix and the RHS for
   MsFEM basis

cout << "loading Stiffness matrix" << endl;
matrix Aglobal;
ifstream Kmat(Stiffnessmatrix);
Kmat >> Aglobal ;

int number;
ifstream RHSm(Rhsvector);
RHSm >> number;
real[int] Rhsglob(number);
RHSm >> Rhsglob;
real[int] xx=Rhsglob;

cout << "Matrices loaded" << endl;

cout << "MsFEM solution computing" << endl;
// Solving the global linear system P0-P0edge MSFEM
set(Aglobal,solver=UMFPACK);
Rhsglob=Rhsglob.*u1; // new RHS with coefficients
   corresponding to values of f
xx=Aglobal^-1*Rhsglob;
cout << "MsFEM solution computed" << endl;

cout << "MsFEM fine solution reconstructing" << endl;
cout << "Reconstructing the fine solution " << endl;



real Qoi=0; // quantity of interest to compute  ( computed
   for one process)
real Qoiglobal=0; // quantity of interest to compute (sum for
    all processes)
// Loop over the triangles to compute QOI in parallel and to
   store local values for display
  for(int k=0;k<nbtri;++k)
  {
    bool elemtest= (iproc==k%nbproc); // bool to assign one
       element to one process
    if(elemtest)
    {
        mesh THK, ThK; // local fine mesh
        ChiK[][k]=1; // P0 function used to mark the element
           i
        THK=trunc(TH,ChiK>0.1,split=1);  // a mesh made of
           only one triangle
```

```
108         ThK=trunc(THK,1,split=nsplit); // each triangle
               divived by nsplit
109         ChiK[][k]=0;
110
111         fespace VKh(ThK,P1); // fine P1 local space
112
113         VKh uloc=0; // local fine solution
114         VKh[int] usol(4);
115         // loading msfem basis on element k
116         string basisstore="./basis_repository/basis_element"+
               k+".txt";
117         {
118              ifstream readbasis(basisstore);
119              for(int i=0;i<4;i++)
120              {
121                   readbasis>> usol[i][];
122              }
123         }
124         // loop on CR MsFEM dof ( one bubble and 3 edges)
125         for(int i=0;i<4;++i)
126         {
127              real ugi=xx[P0P0edge(k,i)]; // coeff of the
                    solution for the msfem basis i of element k
128              real[int] dummy=usol[i][]; // loading msfem basis
                    i of element k
129              dummy=dummy*ugi;
130              uloc[]=uloc[]+dummy;// build local solution
131         }
132
133
134         // Store the solution in x y value form (for all
               points of the fine mesh)
135         string solstore="./solution_repository/sol_element"+k
               +".txt";
136         {
137              ofstream writesol(solstore);
138              for(int j=0;j<VKh.ndof;j++)
139              {
140                   real Xs=ThK(j).x;
141                   real Ys=ThK(j).y;
142                   real valsol=uloc(Xs,Ys);
143                   writesol<< Xs << " " << Ys << " " << valsol
                         << " " << endl;
144              }
145
146         }
147         // Compute quantity of interest (for instance here
               energy)
148         Qoi+=int2d(ThK)(Aeps(uloc,uloc))-int2d(ThK)(f*uloc);
149     }
150
151    }
152
```

```
// Sum all contribution of QOI from processes to QOIglobal of
    process 0
mpiReduce(Qoi,Qoiglobal,processor(0,comm),mpiSUM);

if(iproc==0)
{
// display QOI
cout << "Quantity of interest = " << Qoiglobal << endl;
}
```

A.3 Coupling linear MsFEM with P1

The code here is not parallelized as the aim is to show how to partition the mesh into two pieces on which we apply either an MsFEM or a P1 formulation. It also shows a way to deal with interface terms where the degrees of freedom are part of both P1 and MsFEM formulations. Only the MsFEM linear coupled approach has been presented, since the MsFEM oversampling approach would follow the same steps except for the basis functions construction (we refer to Section A.1.2 in that respect).

```
// P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria
   Matherials

// Implements a coupling method between P1-Elements and
   linear MsFEM
// In one region classical P1 are used and in the other
   region each triangle is subdivided into smaller elements
   to compute linear MsFEM.
// The solution is continuous at the vertices as the BC for
   the linear MsFEM corresponds exactly to P1 FE.
// of the coarse triangulation Thg continuous within each
   coarse triangle and discontinuous
// For the time being only the Dirichlet BC are implemented
   and it has not been yet parallelized

verbosity=0;

// geometric parameter for the separation between two areas
real hmin=0.21875; real hmin2=0.25; real valmin=hmin; real
   valmax=1-hmin; real valmin2=hmin2; real valmax2=1-hmin2;
func fint=(x>valmin)*(x<valmax)*(y>valmin)*(y<valmax); //
   function to separate P1 and MSFEM regions
func fint2=(x>valmin2)*(x<valmax2)*(y>valmin2)*(y<valmax2);
   // function to design a two area coefficient


real eps = 1./16.; // size of oscillations
func aeps =3.*(fint2<0.3)+(fint2>0.3)*((2+1.8*sin(2*pi*x/eps)
   )/ (2+1.8*cos(2*pi*y/eps)) + (2+sin(2*pi*y/eps))/(2+1.8*
   sin(2*pi*x/eps)));  // coef of PDE :- div(aeps grad u)=f

// macro for variational formulation
macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
```

```
22 macro Grad(u) [dx(u),dy(u)] // EOM
23
24 // Right-hand side
25 func f=sin(2*pi*x)*cos(2*pi*y);
26
27 // Size of mesh
28 int H=16; // size of coarse mesh
29 int h=256; // size of size mesh
30 int nsplit=h/H; // ratio between coarse and fine mesh
31
32
33 // Mesh generation
34 mesh TH=square(H,H); // Coarse global mesh
35 int interface=5; // interface number for integration
36 mesh Thg=trunc(TH,(fint<0.3),label=interface); // coarse mesh
       P1 if (fint<0.3)=1 then element i stays in mesh
37 mesh Thd=trunc(TH,(fint>0.3),label=interface); // coarse mesh
       MSFEM if (fint>0.3)=1 then element i stays in mesh
38 int nbtriP1=Thg.nt; // number of P1 triangles
39 int nbtriMsFEM=Thd.nt; // number of MsFEM triangles
40
41
42 TH=Thg+Thd; // Putting the two meshes together
43 plot(TH,wait=1); // plot of the global coarse mesh
44 fespace Tri(TH,P0); // P0 on coarse mesh
45 fespace P1Tri(TH,P1); // P1 on coarse mesh
46
47 int nbtri=TH.nt; // total number of triangles
48
49 // P1 and P0 function to navigate the mesh
50 Tri ChiK=0;
51 P1Tri TestK=0;
52
53
54 // Generating fine submeshes
55 mesh[int] THK(nbtri), ThK(nbtri);
56 for (int i=0;i<nbtri;i++)
57 {
58     ChiK[][i]=1; // P0 function used to mark the element i
59     THK[i]=trunc(TH,ChiK>0.1,split=1);  // a mesh made of
          only one triangle
60     ThK[i]=trunc(THK[i],1,split=nsplit); // each triangle
          divived by nsplit
61     ChiK[][i]=0;
62 }
63
64
65 // using coarse problem formulation
66 varf vA(u,v)= int2d(TH)(Aeps(u,v) )+on(1,2,3,4,u=0);//
     bilinear form of the coarse P1 problem
67 varf vB(unused,v)= int2d(TH)(f*v )+on(1,2,3,4,unused=0);//
     RHS for the coarse P1 problem
68
69 matrix A=vA(P1Tri,P1Tri); // Stiffness matrix for the coarse
```

```
   P1 pb -> this matrix will be used in P1 region and changed
    for MSFEM region
real[int] RHS=vB(0,P1Tri); // RHS for the coarse P1 pb
real[int] xx(RHS.n); // Coefficient for the coupled
   formulation P1-MSFEM

int nbdofKMSFEM=3; // nb of DOF in MSFEM variant for one
   element

mesh TK=THK[0];
mesh TKh=ThK[0];

fespace VKh(TKh,P1);
fespace VK(TK,P1);

// Vector of FE to store the fine solutions
VKh[int] Ui(nbtriMsFEM*nbdofKMSFEM);// FH

// number of MSFEM element currently computed
int iU=0;

// Loop on the element of the MSFEM region to buid the MsFEM
   basis (can be performed in parallel)
for (int i=nbtriP1;i<nbtri;i++)
{
    //loading meshes and P1 coarse and fine of the current
       element
    mesh TK=THK[i];
    mesh TKh=ThK[i];
    fespace VKh2(TKh,P1);
    fespace VK2(TK,P1);

    // fine FE vector storing the local basis functions
    VKh2[int] uki(nbdofKMSFEM);


    // Loop to compute the linear MSFEM basis for the ith
       element
    for (int j=0;j<nbdofKMSFEM;j++)
    {
    // P1 coarse function to set the BC for the basis
       function (P1 basis function)
    P1Tri Test;
    Test[][P1Tri(i,j)]=1;
    varf vAK(u,v)=int2d(TKh)(Aeps(u,v))+on(1,2,3,4,interface,
       u=Test); // Stiffness expression for the fine MSFEM pb
    varf vBK(u,v)= on(1,2,3,4,interface,u=Test); // RHS
       expression for the fine MSFEM pb
    real [int] bk=vBK(0,VKh2); // RHS  for the fine MSFEM pb
    matrix AK=vAK(VKh2,VKh2); // Stiffness matrix for the
       fine MSFEM pb
    set(AK,solver=UMFPACK);
    uki[j][]=AK^-1*bk; // solve the linear system
    Ui[iU][]=uki[j][]; // store it in the global solutions
```

```
113     iU++; // increasing number of MSFEM basis computed
114     }
115
116
117     // Assembling the Stiffness matrix and RHS vector for the
            global problem
118     varf vAK2(u,v)= int2d(TKh)(Aeps(u,v) );
119     matrix KK=vAK2(VKh,VKh);
120
121     // penalization term for the dirichlet BC
122     real tgv=1e30;
123
124     // Double loop on dof MSFEM to compute the local
           Stiffness matrix and RHS
125     for (int j=0;j<nbdofKMSFEM;j++)
126         {
127         // P1 element to remove the P1 Stiffness part (no
               splitting of DOF like discontinuous galerkin, DOF
               are vertices)
128         VK2 Tj;
129         P1Tri P1Tj;
130         P1Tj[][P1Tri(i,j)]=1;
131         Tj=P1Tj;
132         real bP1=int2d(TK)(Tj*f); // RHS P1 coarse
133         real bMsFEM=int2d(TK)(uki[j]*f);  // RHS MSFEM-L
               coarse
134         int I=P1Tri(i,j); // global numbering of vertex j
135         int btestj=(TH[i][j].label>0)*(TH[i][j].label!=
               interface); // test if dof du bord pour vertex i
136         RHS[I]=RHS[I]-bP1+bMsFEM; // change in global RHS
137         for (int l=0;l<nbdofKMSFEM;l++)
138             {
139             int J=P1Tri(i,l); // global numbering of vertex l
140             // P1 element to remove the P1 Stiffness part (no
                    splitting of DOF like discontinuous galerkin,
                    DOF are vertices)
141             VK2 Tl;
142             P1Tri P1Tl;
143             P1Tl[][P1Tri(i,l)]=1;
144             Tl=P1Tl;
145             func bordglobal=(x==0)+(y==0)+(x==1)+(y==1);
146             int btestjl=btestj+ (TH[i][l].label>0)*(TH[i][l].
                   label!=interface);
147             real KP1=int2d(TK)(Aeps(Tj,Tl)); // local
                   Stiffness P1 coarse
148             real KMsFEM=int2d(TKh)(Aeps(uki[j],uki[l])); //
                   local Stiffness MSFEM
149             if(btestjl==0)
150                 {
151                 A(I,J)=A(I,J)-KP1+KMsFEM; // new Stiffness=
                       ancient Stiffness+ msfem contribution -P1
                       contribution
152                 }
153             }
```

```
154         }
155 }
156
157 // Solving the global linear system P1-MSFEM
158 set(A,solver=UMFPACK);
159 xx=A^-1*RHS;
160
161 // Rebuild MSFEM solution -> the solution is evaluated
    accurately locally
162
163
164 //Global coarse solution for P1 part
165 P1Tri P1sol;
166 // Rebuild the P1 region locally
167 for(int k=0;k<nbtriP1;++k)
168 {
169     for(int i=0;i<3;++i)
170     {
171     P1sol[][P1Tri(k,i)]=xx[P1Tri(k,i)]; // fill the global p1
        part with solution
172     }
173 }
174
175 // Loop on the MSFEM region triangles
176 iU=0; // counter to load msfem basis
177 for(int k=nbtriP1;k<nbtri;++k)
178     {
179      // load the coarse and fine mesh of P1 triangle k
180     TKh=ThK[k];//FH
181     // Mesh corresponding to the global fine mesh
182     mesh TKf=trunc(TK,split=nsplit,1);
183
184     fespace VKh2(TKh,P1); // fine P1 space
185
186
187     VKh2 uloc=0; // local fine MsFEM CR solution
188
189     // loop on vertices
190     for(int i=0;i<nbdofKMSFEM;++i)
191         {
192         real ugi=xx[P1Tri(k,i)]; // coeff of the solution for
            the msfem basis i of element k
193         real[int] dummy=Ui[iU][]; // loading msfem basis i of
            element k
194         dummy=dummy*ugi;
195         uloc[]=uloc[]+dummy;// build local solution
196         iU++;
197         }
198     // perform operation with uloc, storing quantity of
        interest, ...
199     }
```

A.4 MsFEM as a second level preconditioner

```
// P.-L. Rothe - F. Legoll ENPC(CERMICS/Navier) - Inria Matherials

// Implements the second level preconditioner given by MsFEM linear basis and Jacobi fine preconditioner

verbosity=0;

// inputs : Mesh H, mesh h, diffusion coeff A known on h, Right hand side known on H, MsFEM basis functions,
// Stiffness matrix and Passage matrix from MsFEM coarse basis to global fine mesh

// Mesh def and generation
// Size of mesh
int H=8; // size of coarse mesh
int h=256; // size of size mesh
int nsplit=h/H; // ratio between coarse and fine mesh

mesh TH=square(H,H,[x,y]); // Coarse global mesh
int nbtri=TH.nt; // number of MsFEM triangles

// Global fine mesh
mesh Thf=trunc(TH,split=nsplit,1);

//P1 on global fine mesh
fespace Vhf(Thf,P1);

// Definition of coefficient
real eps = 1./32.; // size of oscillations
func defcoeff =((2+1.8*sin(2*pi*x/eps))/ (2+1.8*cos(2*pi*y/eps)) + (2+sin(2*pi*y/eps))/(2+1.8*sin(2*pi*x/eps))); // coef of PDE :- div(aeps grad u)=f
Vhf aeps=defcoeff;
macro Aeps(u,v) (Grad(u)'*Grad(v)*aeps) //
macro Grad(u) [dx(u),dy(u)] // EOM

// Right-hand side
func f=1;
P1Tri frhs=f;

Tri ChiK=0;
P1Tri TestK=0;

// Loading Precomputed MsFEM Stiffness matrix
string Stiffnessmatrix="Stiffness_matrix.txt";
 cout << "loading Stiffness matrix" << endl;
matrix Aglobal;

int number;
ifstream Kmat(Stiffnessmatrix);
```

```
Kmat >> number;
real[int] valA2(number);
int[int] IA2(number),JA2(number);
Kmat >> IA2 >> JA2 >> valA2 ;
Aglobal=[IA2,JA2,valA2];

// Loading Precomputed passage matrix between MsFEM coarse
   basis VH and fine mesh Vh
// Vh=Pglobal VH
string passagematrix="passage_matrix.txt";
cout << "loading passage matrix" << endl;
matrix Pglobal;
ifstream Pmat(passagematrix);
Pmat >> number;
real[int] valA(number);
int[int] IA(number),JA(number);
Pmat >> IA >> JA >> valA ;
Pglobal=[IA,JA,valA];

 cout << "MsFEM solution computing" << endl;

//Fine solution computed with preconditioning

cout << "defining fine problem" << endl;
//Fine problem
varf vAf(u,v)= int2d(Thf)(Aeps(u,v) )+on(1,2,3,4,u=0);
varf vBf(unused,v)= int2d(Thf)(f*v )+on(1,2,3,4,unused=0);

matrix Aref=vAf(Vhf,Vhf);
real[int] RHSref=vBf(0,Vhf);
real[int] xxref(RHSref.n);


cout << "defining Jacobi preconditioner" << endl;
//Jacobi diagonal fine preconditioner matrix
int n = Vhf.ndof;
real[int] one(n); one=1.;
matrix Id=one; //
matrix diag=one;
for(int i=0;i<n;i++)
{
diag(i,i)=1/Aref(i,i);
}

cout << "Defining whole preconditioner" << endl;

cout << "Defining  preconditioner functions" << endl;
//Jacobi diagonal preconditioner
func real[int] Preconddiag(real[int] & xx)
{
real[int] xx = diag*xx;
return xx;
}

```

```
// Function giving MsFEM solution on coarse MsFEM basis for a RHS = b
func real[int] ComputeMSFEM(real[int] b)
{
verbosity=0;
set(Aglobal,solver=UMFPACK);
real[int] xx=Aglobal^-1*b;
return xx;
}

//Second level preconditioner  M^-1 X= Pglobal^TComputeMsFEM(Pglobal X)+diag X
func real[int] Precondmsfem(real[int] & xx)
{
// MsFEM precon part
real[int] xxH=Pglobal^T*xx; //  transform RHS on RHS on MsFEM basis
real[int] solH=ComputeMSFEM(xxH); // Compute MsFEM solution on coarse MsFEM basis for RHS xxH
real[int] xxh=Pglobal*solH; // Get back to the fine mesh
// Jacobi part
real [int] MX=diag*xx; // compute the fine part of the preconditioner
xx=xxh+MX;
return xx;
}

cout << "fine solution preconditioned MSFEM computing" << endl;
//Solve the fine pb with second level preconditioner
verbosity=10; // to display the value of relative error of GMRES
real tol=1e-6; // GMRES stopped if err<tol or nb iteration > nbiter
set(Aref,solver=GMRES,nbiter=150,precon=Precondmsfem,eps=-tol);
Vhf uref;
uref[]=Aref^-1*RHSref;
plot(uref,wait=1,cmm="msfem precondition");
cout << "fine solution  preconditioned MSFEM  computed" << endl;

cout << "fine solution DIAG computing" << endl;
//Solve the fine pb with only Jacobi preconditioner
verbosity=10; // to display the value of relative error of GMRES
tol=1e-6; // GMRES stopped if err<tol or nb iteration >nbiter
set(Aref,solver=GMRES,nbiter=150,precon=Preconddiag,eps=-tol);
Vhf uref2;
uref2[]=Aref^-1*RHSref;
plot(uref2,wait=1,cmm="diag precondition");
cout << "fine solution  DIAG computed" << endl;
```

APPENDIX B

TRACE RESULTS AND SOBOLEV INTERPOLATION RESULTS

In CHAPTER 3, we have used some analytical tools to get convergence estimates. We have to consider liftings of functions defined on the boundary of local domains (solutions of elliptic PDE with prescribed Dirichlet boundary conditions) and control them in terms of Sobolev norms on the domain. This annex is intended to review and introduce the tools we used in a more comprehensive manner. First, we perform a review on the definition on fractional Sobolev spaces (on whole domains, boundary and subset of boundaries). Then, the Trace theorem and especially the scaling with regard to the length of the domain is discussed with some remarks on elliptic regularity. Finally, we present some polynomial interpolation results.

B.1 Sobolev spaces on boundaries and Traces operator

This section presents Sobolev spaces properties for specific domains (boundary or subset of a boundary) and the link between estimates on the domain and on the boundary in Sobolev norms. This review relies heavily on the following monographs: [78, SECTION 2.3 through 2.8], [71, SECTION 3], and [47, SECTION 1].

B.1.1 Characterization of the regularity of a domain Ω

Definition B.1. *(see [47, DEFINITION 1.2.1.1)] Let Ω be an open subset of $\mathbb{R}^n$. We say that its boundary Γ is continuous (respectively Lipschitz, continuously differentiable, of Class $C^{k,1}$, m times continuously differentiable) if for every $x \in \Gamma$ there exists a neighborhood V of x in $\mathbb{R}^n$ and new orthogonal coordinates $\{y_1, \ldots, y_n\}$ such that*

1. *V is an hypercube in the new coordinates:*
   $V = \{(y_1, \ldots, y_n) | -a_j < y_j < a_j,\ 1 \leq j \leq n\}$

2. *There exists a continuous (respectively Lipschitz, continuously differentiable, of Class $C^{k,1}$, m times continuously differentiable) function ϕ, defined in*
   $V' = \{(y_1, \ldots, y_{n-1}) | -a_j < y_j < a_j,\ 1 \leq j \leq n-1\}$

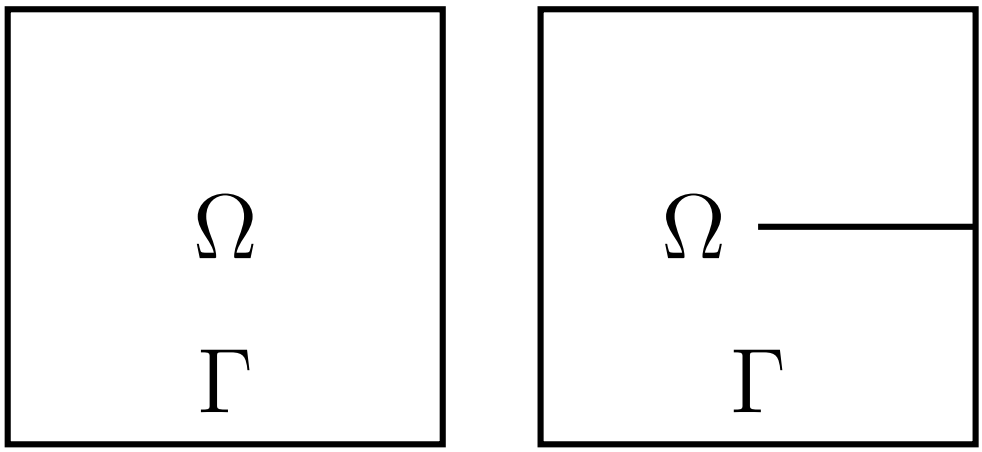


Figure B.1: Left: example of Lipschitz domain, Right: a non-Lipschitz domain

and such that

$$|\phi(y')| \leq a_n/2 \text{ for every } y' = (y_1, \ldots, y_{n-1}) \in V'$$
$$\Omega \cap V = \{y = (y', y_n) \in V | y_n < \phi(y')\}$$
$$\Gamma \cap V = \{y = (y', y_n) \in V | y_n = \phi(y')\}$$

Remark B.2. *Polygonal domains have a boundary that is Lipschitz continuous. However, the boundary is not continuously differentiable. If Ω is a bounded open convex subset of $\mathbb{R}^n$, then Ω has a Lipschitz boundary, (see [47,* COROLLARY *1.2.2.3 and* DEFINITION *1.2.1.1])*

B.1.2 Review of fractional Sobolev spaces

Considering Ω a bounded Lipschitz domain in $\mathbb{R}^d$, we define $H^\ell(\Omega)$ the fractional Sobolev spaces for $\ell \in \mathbb{R}^+$ and $\lambda = \ell - [\ell] \in]0, 1[$ by

$$H^\ell(\Omega) = \{u \in H^{[\ell]}(\Omega), |u|_\lambda < \infty\} \tag{B.1}$$

with $|u|_\lambda$ the semi-norm

$$|u|_\lambda^2 = \sum_{|\alpha|=|\ell|} \int_\Omega \int_\Omega \frac{\left(\partial^\alpha u(x) - \partial^\alpha u(y)\right)^2}{\|x - y\|^{d+2\lambda}} \, dx \, dy \tag{B.2}$$

Remark B.3. *$H^l(\Omega)$ is an Hilbert space associated with the scalar product*

$$(\phi, \psi)_\ell = \sum_{|\alpha| \leq |\ell|} \int_\Omega \partial^\alpha \phi \, \partial^\alpha \psi + \sum_{|\alpha|=|\ell|} \int_\Omega \int_\Omega \frac{\left(\partial^\alpha \phi(x) - \partial^\alpha \phi(y)\right) \left(\partial^\alpha \psi(x) - \partial^\alpha \psi(y)\right)}{\|x - y\|^{d+2\lambda}} \, dx \, dy \tag{B.3}$$

and the norm $\| \cdot \|_\ell = (\cdot, \cdot)_\ell^{1/2}$.
Moreover, it holds that the closure of $C_0^\infty(\Omega)$ with respect to the $\| \cdot \|_\ell$ norm defines $H_0^\ell(\Omega)$ as usual.

B.1.3 Definition of Sobolev spaces on the boundary

In Chapter 3, we have used interpolation results on boundaries of small domains and have expanded these results in the interior of the domains. In order to do that, we require trace theorems. Such results often involve regularities associated to fractional Sobolev spaces defined on the boundary.

All Sobolev spaces cannot be defined on any boundary of an open bounded domain. Indeed, the Sobolev regularity ℓ is bounded by a maximal order of differentiability depending on the regularity of the boundary Γ. The regularity of the boundary Γ can be seen as the maximum regularity of χ a diffeomorphism between a local part of Γ and a bounded space in $\mathbb{R}^{n-1}$ (typically B_R^0 the intersection between the ball of radius R centered at 0 and the plane $x_n = 0$). In such case, ϕ a function defined on the boundary can be seen as the sum of functions $\hat{\phi} = \chi \circ \phi$ from B_R^0 to $\mathbb{R}$ (space where Sobolev space are well defined). The regularity of such function depends obviously on the regularity of χ the diffeomorphism between the local part of Γ and B_R^0. Hence, it holds that for Lipschitz domains and C^k domains Sobolev spaces $H^\ell(\Gamma)$ are defined if $\ell \in \mathbb{R}^+$ satisfies

$$\begin{aligned} &\ell \leq 1 \text{ for Lipschitz domains } \Omega \\ &\ell \leq k \text{ for } C^k \text{ domains } \Omega \\ H^\ell(\Gamma) = \{\phi : \Gamma \mapsto \mathbb{R} |\ &\hat{\phi} \in H^\ell(B_R^0)\} \end{aligned} \tag{B.4}$$

We can also define an associated norm and scalar product similar to (B.3) except that d should be replaced by $d-1$ in (B.2) and the integral should be computed on Γ (instead of Ω).

Remark B.4. *The definition of Sobolev spaces $H^\ell(\Gamma)$ and their associated norms seems to depend on the coordinates system chosen for the diffeomorphism between Γ and B_R^0. However, it can be shown that for bounded Lipschitz domains the space contains the same set of functions and the norm are equivalent.*

Remark B.5. *For closed surfaces, then $H^\ell(\Gamma) = H_0^\ell(\Gamma)$, with $H_0^\ell(\Gamma)$ the closure of infinitely differentiable functions with support on Γ with respect to H^ℓ norm. Similarly to domains, negative Sobolev spaces on boundaries are defined as*

$$H^{-\ell}(\Gamma) = \left(H_0^\ell(\Gamma)\right)', \text{ for } \ell \geq 0 \tag{B.5}$$

B.1.4 Sobolev spaces on $\Gamma_0 \subset \Gamma$

In Chapter 3, we build basis functions on each element (convex polygons such as triangle or quadrangle) by imposing boundary conditions on edges. Hence, we need to get local approximation properties edge by edge and then get back to estimates on the whole domain by performing liftings. To that end, we will consider Sobolev spaces on $\Gamma_0 \subset \Gamma$.

For $\Gamma_0 \subset \Gamma$ measurable with $|\Gamma_0| > 0$, and $s \in [0,1]$, we define the space $H_{0,0}^s$ (also noted $\tilde{H}^s(\Gamma_0)$ in the literature) by

$$H_{0,0}^s := \{u \in H^s(\Gamma) |\ \mathrm{supp}(u) \subset \overline{\Gamma_0}\} \tag{B.6}$$

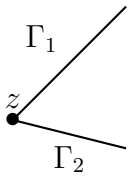
The associated norm is given by

$$\|u\|_{H_{0,0}^s} = \|\widetilde{u}\|_{H^s(\Gamma)} \tag{B.7}$$

where $\tilde{u}$ denotes the extension of u on Γ by 0.

We can also define Sobolev spaces on Γ_0, with negative exponents: $H^{-s}(\Gamma_0) = \left(H^s_{0,0}\right)'$ for $s \in [0,1]$.

Remark B.6. *One has to be careful when considering such spaces especially when one wants to compare the norms $H^s(\Gamma_0)$ and $H^s(\Gamma)$. We recall results from [3, SECTION 2] and [47, LEMMA 1.5.1.8]. For instance let us consider $\Gamma = \Gamma_1 \cup \Gamma_2$ with Γ_1 and Γ_2 two line segments intersecting at a point z.*



Then, for $0 \leq s < 1/2$

$$f \in H^s(\Gamma) \quad \text{if and only if} \quad f|_{\Gamma_1} \in H^s(\Gamma_1) \text{ and } f|_{\Gamma_2} \in H^s(\Gamma_2)$$

The norm $\|\cdot\|_{H^s(\Gamma)}$ is equivalent to $\|\cdot\|_{H^s(\Gamma_1)} + \|\cdot\|_{H^s(\Gamma_2)}$.

For $1/2 < s \leq 1$

$f \in H^s(\Gamma)$ *if and only if* $f|_{\Gamma_1} \in H^s(\Gamma_1)$, $f|_{\Gamma_2} \in H^s(\Gamma_2)$ *and f is continuous on z.*

The norm $\|\cdot\|_{H^s(\Gamma)}$ is equivalent to $\|\cdot\|_{H^s(\Gamma_1)} + \|\cdot\|_{H^s(\Gamma_2)}$.

However, for $s = 1/2$, denoting σ_1, σ_2 the unit vectors parallel to Γ_1, Γ_2 respectively pointing toward z, it holds that

$$f \in H^s(\Gamma) \quad \text{if and only if} \quad I_\Gamma(f) = \int_0^\varepsilon t^{-1} |f(z - t\sigma 1) - f(z + t\sigma_2)|^2 \, dt < \infty$$
$$\text{and} \quad f|_{\Gamma_1} \in H^s(\Gamma_1), f|_{\Gamma_2} \in H^s(\Gamma_2)$$

The norms $\|\cdot\|_{H^s(\Gamma)}$ and $\|\cdot\|_{H^s(\Gamma_1)} + \|\cdot\|_{H^s(\Gamma_2)}$ are not equivalent.

Hence, for $s = 1/2$ and Γ the boundary of a polygon in $\mathbb{R}^2$ composed by n edges $\Gamma = \cup_{j=1..n}\Gamma_j$ one cannot bound easily the $\|\cdot\|^2_{H^s(\Gamma)}$ by the sum of $\|\cdot\|^2_{H^s(\Gamma_j)}$.

B.1.5 Trace theorems

The previous section helped define regularity on the domain Ω and on the boundary Γ. However, the link between both regularity is missing, that is the answer to the following questions: If one considers a function f in $H^s(\Omega)$, what is the regularity of f on Γ ? If we have a function f that belongs to $H^s(\Gamma)$, on what conditions can one build an extension $\widetilde{f}$ in the whole domain which coincides with f on the boundary ? The Trace operator theory addresses both problems.

Theorem B.7. *(see [78, THEOREMS 2.6.8 and 2.6.9]) If Ω is a bounded Lipschitz domain with boundary Γ, then for $1/2 < \ell < 3/2$, there exists a continuous linear trace operator $\gamma : H^\ell(\Omega) \mapsto H^{\ell-1/2}(\Gamma)$ such that*

$$\gamma\phi = \phi|_\Gamma \quad \text{for all } \phi \in C^0(\bar{\Omega})$$

If Ω is a bounded C^k domain then the same result holds for $1/2 < \ell \leq k$.

We need the following corollary describing the consequence of this theorem in the case where $\ell = 1$ and the domain has a small length.

Corollary B.8. *If Ω is a bounded Lipschitz domain of $\mathbb{R}^2$ with boundary Γ, then the trace operator $\gamma : H^1(\Omega) \mapsto H^{1/2}(\Gamma)$ is continuous: for all $u \in H^1(\Omega)$, there exists C_Ω (depending on the shape and the diameter of Ω) such that*

$$\|u\|_{H^{1/2}(\Gamma)} \leq C_\Omega \|u\|_{H^1(\Omega)}$$

Moreover, if we consider $\widehat{\Omega}$ a bounded domain in $\mathbb{R}^2$ which is polygonal and convex with $\operatorname{diam}(\widehat{\Omega}) = 1$ and Ω such that $\operatorname{diam}(\Omega) = H$ and Ω can be obtained from $\widehat{\Omega}$ by an affine transformation, then it holds that

$$H^{-1}\|u\|_{L^2(\Gamma)} + |u|_{H^{1/2}(\Gamma)} \leq C\left(H^{-2}\|u\|_{L^2(\Omega)} + |u|_{H^1(\Omega)}\right)$$

with C depending only on the shape of Ω.

In the case where the domain has only Lipschitz boundary, it is impossible to define $H^s(\Gamma)$ with $s > 1$. However, in the case of polygonal domains, as the boundary is piecewise C^∞, linear continuous operator can be defined on more regular Sobolev spaces though not on the whole boundary but edge by edge.

Theorem B.9. *(see [48, Theorem 1.4.2]) If Ω is a polygonal bounded open subset of $\mathbb{R}^2$ with boundary $\Gamma = \cup_{j=1..N}\Gamma_j$ where each Γ_j is a segment. then, denoting by ν the unit outward normal vector, the mapping*

$$u \mapsto \{\gamma_j u, \gamma_j \frac{\partial u}{\partial \nu_j}, \ldots, \gamma_j \frac{\partial^k u}{\partial \nu_j^k}\}$$

which is defined for $u \in \mathcal{D}(\bar{\Omega})$ has for $k < s - 1/2$ a unique continuous extension from $H^s(\Omega)$ onto $\prod_{0 \leq p \leq k} H^{s-p-1/2}(\Gamma_j)$.

Theorem B.10. *(see [48, Theorem 1.4.6]) If Ω is a polygonal bounded open subset of $\mathbb{R}^2$ with boundary $\Gamma = \cup_{j=1..N}\Gamma_j$ where each Γ_j is a segment, then, denoting by ν the unit outward normal vector, the mapping $u \mapsto \{f_{j,\ell} = \gamma_j \frac{\partial^\ell u}{\partial \nu_j^\ell},\ 1 \leq j \leq N,\ 0 \leq l \leq m-1\}$ is linear continuous onto the subspace of $T = \prod_{1 \leq j \leq N} \prod_{1 \leq \ell \leq m-1} H^{m-l-1/2}(\Gamma_j)$ defined by the following conditions.*

Let L be any differential operator with constant coefficients and order $d \leq 1$. Denote by $P_{j,\ell}$, the differential operators tangential to Γ_j such that

$$L = \sum_\ell P_{j,\ell} \frac{\partial \ell}{\partial \nu_j^\ell}$$

then

$$\sum_\ell P_{j,\ell} f_{j,\ell}(S_j) = \sum_\ell P_{j+1,l} f_{j+1,\ell}(S_j) \text{ for } d \leq m-2$$

$$\sum_\ell P_{j,\ell} f_{j,\ell} \equiv \sum_\ell P_{j+1,l} f_{j+1,\ell} \text{ at } S_j \text{ for } d \leq m-2$$

The Trace operator is obviously not invertible. Let us consider Ω a closed space with regular boundary. Two functions in $H^1(\Omega)$ can share the same trace on the boundary (in the $H^{1/2}(\partial\Omega)$ sense) and be different on Ω (take for instance, $f(x,y) = \sqrt{x^2+y^2}$ and $g(x,y) = x^2+y^2$ on $B_1 = \{(x,y) \in \mathbb{R}\times\mathbb{R} \, |x^2+y^2 \leq 1\}$). However, it is possible to define a continuous extension operator from $H^{\ell-1/2}(\Gamma)$ to $H^\ell(\Omega)$.

Theorem B.11. *(see [78, THEOREM 2.6.11]) Let Ω be a bounded Lipschitz domain with boundary Γ. Then, for $1/2 < \ell < 3/2$, there exists a linear continuous extension operator $Z : H^{\ell-1/2}(\Gamma) \mapsto H^\ell(\Omega)$ with $(\gamma \circ Z)(\phi) = \phi$ on $H^{\ell-1/2}(\Gamma)$ for all $\phi \in H^{\ell-1/2}(\Gamma)$.*

B.2 Regularity of elliptic equations on convex domains

In CHAPTER 3, we need to assess the regularity (in terms of Sobolev spaces) of the solution to

$$\begin{cases} -\mathrm{div}(A\nabla u) = f \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{B.8}$$

Theorem B.12. *(see [47, THEOREMS 2.2.2.3 and 3.2.1.2]) If A is elliptic and Lipschitz continuous, $f \in L^2(\Omega)$ and Ω is a convex domain then the problem (B.8) admits a unique solution u in $H_0^1(\Omega)$. Moreover, u belongs to $H^2(\Omega)$.*

One can also obtain Hölder regularity results provided more regularity on f and the domain.

Theorem B.13. *(see [43, THEOREM 6.24]) Suppose that A is elliptic, in a bounded domain Ω that satisfies an exterior sphere condition, and A and f are Hölder continuous with exponent α (in $C^{0,\alpha}(\Omega)$). Suppose that f and A are bounded on $\bar{\Omega}$. Then if ϕ is continuous on $\partial\Omega$, the Dirichlet problem $-\mathrm{div}(A\nabla u) = f$ in Ω, $u = \phi$ on $\partial\Omega$ has a unique solution $u \in C^0(\bar{\Omega}) \cap C^{2,\alpha}(\Omega)$.*

Remark B.14. *The exterior domain condition is satisfied if there exists r such that for any $x \in \partial\Omega$ there exists a ball B_r of radius r satisfying $B_r \subset \mathbb{R}^n \setminus \Omega$ and $x \in \partial B_r$. For a triangle and a rectangle (most generally a convex polygonal domain) this condition is satisfied for all r.*

The Laplacian operator has unique properties allowing to control for $u \in H^2(D) \cap H_0^1(D)$ all second derivatives with only the $L^2(D)$ norm of the Laplacian. It is especially useful when $\Delta u = f$ with $u \in H_0^1(D)$ as the norms of the function can be bounded by f provided u is regular enough ($H^2(D)$).

Theorem B.15. *(see [48, THEOREM 2.2.3]) Assume that Ω is a bounded convex polygonal open subset of $\mathbb{R}^2$ such that $\mathrm{diam}(\Omega) = H$. For $u \in H^2(\Omega)$ such that the trace of u on $\partial\Omega$ vanishes it holds that*

$$H^{-2}\|u\|_{L^2(\Omega)} + H^{-1}|u|_{H^1(\Omega)} + |u|_{H^2(\Omega)} \leq C\|\Delta u\|_{L^2}, \tag{B.9}$$

with C independent of the length of Ω.

B.3 Sobolev interpolation of linear operators

Lemma B.16. *(see [68, THEOREM 5.1]) Let $(\mathcal{X}, \mathcal{Y})$ be a couple of separable Hilbert spaces with $\mathcal{X} \subset \mathcal{Y}$, such that $\mathcal{X}$ is dense in $\mathcal{Y}$ and such that the injection from $\mathcal{X}$ to $\mathcal{Y}$ is continuous. Let (X, Y) be another couple of Hilbert spaces with analogous properties. Denote by $\mathcal{L}(X, Y)$ the set of linear continuous operators from X to Y, and likewise for $\mathcal{L}(\mathcal{X}, \mathcal{Y})$. Let π be an operator satisfying $\pi \in \mathcal{L}(\mathcal{X}, \mathcal{Y}) \cap \mathcal{L}(X, Y)$. Then, for all $0 < \theta < 1$, we have*

$$\pi \in \mathcal{L}([\mathcal{X}, X]_\theta, [\mathcal{Y}, Y]_\theta)$$

where the interpolated space $[\mathcal{X}, X]_\theta$ is defined in [68, DEFINITION 2.1].

Remark B.17. LEMMA B.16 *allows to extend easily properties from integer Sobolev spaces to fractional Sobolev spaces. Indeed, with $0 < s < 1$, one can also define $H^s(\Omega)$ as $[L^2(\Omega), H^1(\Omega)]_s$. For instance, if we have a linear operator π such that for $u \in L^2(\Omega)$ $\|\pi u\|_{L^2(\Omega)} \leq \|u\|_{L^2(\Omega)}$ and for $u \in H^1(\Omega)$, $\|\pi u\|_{L^2(\Omega)} \leq C\|u\|_{H^1(\Omega)}$, then it holds that for $u \in H^s(\Omega)$, $\|\pi u\|_{L^2(\Omega)} \leq C^s\|u\|_{H^s(\Omega)}$. Such results are crucial to prove polynomial approximation results for fractional Sobolev norms spaces.*

B.4 Polynomial interpolation results and properties

For a wide range of PDE's (in our case elliptic PDE), the solution can be approximated by a Galerkin approach: solving the problem on a finite dimensional space (for instance a Finite Element space). Denoting by u the solution to the considered PDE and u_h its Galerkin approximation on space V_h, we often have results bounding the error by the best approximation error, for instance

$$\|u - u_h\|_{H^1(\Omega)} \leq C \min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)}.$$

From this point, one does not know the behavior of the error (for instance how the error decreases with respect to the meshsize h). Hence, particular functions v_h are used to get an explicit behavior of the error. Finite Element methods use polynomials functions that can be used to approach broader types of functions in a wide range of norms (norms associated to Sobolev spaces, continuous functions,...) through interpolation and projection. For example, in the case of P1 FE space provided that the solution u is regular enough, we have

$$\min_{v_h \in V_h} \|u - v_h\|_{H^1(\Omega)} \leq \|u - I_h(u)\| \leq Ch\|u\|_{H^2(\Omega)}$$

with $I_h(u)$ the P1 FE interpolant such that $I_h(u) = \sum_{i=1..N_{vertices}} u(x_i)\phi_i$ with ϕ_i the P1 basis function associated to the vertex i. Now we know that by dividing the meshsize by two, we also divide the error by two. That is why interpolation results are crucial in Galerkin approaches on FE spaces. They indeed allow to obtain an explicit behavior of the error from a best approximation error. The best approximation is often bounded by above by using explicit polynomials (projection or interpolation of the solution u), hence it is crucial to have a simple method to exhibit a relevant polynomial approaching the solution u. This section is not aimed at being exhaustive, the goal is to give an idea of what tools are available, especially regarding the polynomial approximation space. Most of the results are from the following monographs (see [36], [12], [13], [22],[72]). Although there exists approximation results on interpolation and projection on polynomial functional spaces, we present here only interpolation results.

B.4.1 Interpolation of smooth functions

There are many ways to perform interpolation. We will focus here on polynomial interpolation in 1D and 2D and consider especially the Legendre polynomials since it is the cornerstone of the analysis of the approach designed in CHAPTER 3.

Considering $\Omega \subset \mathbb{R}^N$, the interpolation of a function u is performed by using an operator going from a functional space (often $H^\ell(\Omega)$) onto the space of continuous polynomials with a total degree N. Usually, this operator use evaluation of u on special points to get a good approximation of u in the space of continuous polynomial functions. Such evaluation requires u to be continuous, that is $u \in H^s(\Omega)$ with $s > \frac{d}{2}$, with d the ambient dimension.

We will first consider interpolation results from the Lagrange Finite Element interpolant associated to a regular mesh $\mathcal{T}_H$.

Theorem B.18. *(see [36, THEOREM 1.103]) Assume we use Lagrange Finite element of degree $k > 0$ associated to a regular mesh $\mathcal{T}_h$ of a domain $\Omega \subset \mathbb{R}^n$ with reference element $\hat{K} \in \mathbb{R}^d$ (here $d = 2$) on the space of continuous functions. We consider interpolation results on element K of length h that can be obtained from $\hat{K}$ by a linear transformation.*

We define $I_K^k(u) = \sum_{i=1..Ndof} u(x_i)\phi_i$, with ϕ_i the FE basis function associated to the degree of freedom i on the Finite Element space of degree k for $\mathcal{T}_H$ ($k = 1$ piecewise affine functions on nodes).

Then, for $d/2 - 1 \leq \ell \leq k$ and for all $m \in \{0, \dots, \ell + 1\}$ it holds that

$$\forall K \in \mathcal{T}_h,\ \forall u \in H^{\ell+1}(K), \quad |u - I_K^k(u)|_{H^m(K)} \leq C h_K^{\ell+1-m} \sigma_K^m |u|_{H^{\ell+1}(K)}$$

with C independent of h and $\sigma_K = \frac{h_k}{\rho_K} \leq \sigma_0$ since $\mathcal{T}$ is regular.

Such result is interesting as we have an explicit rate of convergence with respect to the meshsize h: these estimates are the cornerstone of the h-Finite Element methods. However, if we increase the degree of Finite Element, the rate in h_K does not increase anymore as soon as the maximal regularity Sobolev index of the function u is reached (there is no need to take a degree $k > \ell$). Moreover, to get better rates with respect to the regularity ℓ one has to take element of higher degree $\ell \leq k$. There exists interpolation methods that are also taking into account the polynomial degree p used to improve estimates. Such methods are called p-Finite Element or Spectral methods (see [12] and [13]). They use specific properties of specific polynomials such as Legendre polynomials and Chebyshev polynomials. Most recent interpolation results capitalize on both aspects to gives estimates in h and p (see [22, SECTION 5]).

Proofs of such results usually follow the same pattern. First results are proved for integer Sobolev spaces by using specific properties of polynomial families (for instance, the Legendre polynomials satisfy the Legendre's differential equation and the study of the associated differential operator allows to obtain estimates) to get inequalities such as $\inf_{\phi \in P_N} |u - \phi|_{H^s(D)} \leq C|u|_{H^m(D)}$ for $s \leq m$ integers. Second, we use Sobolev interpolation arguments (see LEMMA B.16) to get the results for fractional Sobolev spaces.

Theorem B.19. *(see [22, SECTION 5.4.4]) If we consider I a 1D interval of size h, then considering the Legendre interpolant I_N^H (polynomial of degree N such that $I_N^H(u)(x_j^{GL-N}) = u(x_j^{GL-N})$ for all the $N + 1$ Gauss-Lobatto points x_j^{GL-N}), and any $u \in H^m(I)$, $1 \leq m \leq N$ and $k = 0, 1$ one has*

$$\|u - I_N^H(u)\|_{H^k(I)} \leq C h^{m-k} N^{k-m} |u|_{H^m(I)}$$

We have a similar result for Cartesian product domains, namely domains that can be obtained from $\Omega = (-1,1)^d$ with an affine transformation. It is interesting as it applies to quadrangular elements in a mesh.

Theorem B.20. *(see [22, SECTION 5.8.3]) Consider Ω a quadrangle in $\mathbb{R}^2$ of length h and the Legendre interpolant I_N^H defined as the polynomial of degree up to N in each variable (x_1, x_2) such that*

$$I_N^H(u)(x_j^{e_1,GL-N}, x_l^{e_2,GL-N}) = u(x_j^{e_1,GL-N}, x_l^{e_2,GL-N})$$

for all the $(N+1)^2$ Gauss-Lobatto points $(x_j^{e_1,GL-N}, x_l^{e_2,GL-N})$. Then for all $u \in H^m(\Omega)$, $(2+1)/2 \leq m \leq N$ and $k = 0, 1$ one has

$$\|u - I_N^H(u)\|_{H^k(\Omega)} \leq C h^{m-k} N^{k-m} |u|_{H^m(\Omega)}$$

B.4.2 Interpolation of non-smooth function

In classical polynomial interpolation, in order to define the interpolant one has to evaluate the function at specific points. However, if the function to interpolate is not continuous, the interpolant cannot be built. There are some tools to tackle such issue, the non-smooth interpolation theory.

We present the most common approach: the Clément interpolation introduced in [25].

We consider V_H the piecewise affine functions on $\mathcal{T}_H$ spanned by hat functions ϕ_i defined on the interior vertices of the mesh.

For $u \in H_0^1(D)$, the Clement interpolant $\mathcal{C}_H$ from $H_0^1(\Omega)$ is defined by

$$\mathcal{C}_H(u) = \sum_{i=1}^{Nb_{vertex}} \frac{\int_\Omega \phi_i u}{\int_\Omega \phi_i} \phi_i \tag{B.10}$$

Remark B.21. *Such interpolation does not require any evaluation of the function at a point. A weaker regularity (for instance $L^2(\Omega)$) is enough.*

Theorem B.22. *(see [36, LEMMA 1.127]) If $u \in H_0^1(\Omega)$ with Ω a polygonal domain in $\mathbb{R}^2$ and $\mathcal{T}_H$ a shape regular conformal mesh of Ω then it holds that for any edge $e \subset \Gamma$ with Γ the collection of all interior edges of the mesh, we have*

$$\|u - \mathcal{C}_H(u)\|_{L^2(e)} \leq C H_e^{1/2} \|u\|_{H^1(\omega_e)}$$

with ω_e all the elements who share a vertex with the edge e and H_e the length of the edge e.

Other non-smooth interpolation methods exist such as the Scott-Zhang interpolant that allows to take into account boundary conditions (see [36, LEMMA 1.130]).

One can also design hp-Finite Element non-smooth interpolation method where the degree of polynomial approximation is also taken into account in the estimate (see [72]).

Theorem B.23. *(Scott-Zhang type interpolation result [72, THEOREM 2.3]) Assume that $\mathcal{T}_H$ is a conformal mesh which is shape regular in the sense of (3.2). For any element $K \in \mathcal{T}_H$, we choose a maximal degree $p_K \in \mathbb{N}^\star$ and we assume that these degrees $\{p_K\}$*

satisfy (3.33). Then there exists a continuous interpolation operator $\mathcal{SZ}$ *from* $H_0^1(D)$ *to* $H_0^1(D) \cap \mathcal{S}(\{p_K\})$ *with*

$$\mathcal{S}(\{p_K\}) = \{u \in C^0(\overline{D});\ u|_K \text{ is a polynomial function of degree at most } p_K\}.$$

Furthermore, there exists a constant C *which only depends on the mesh regularity* γ *of (3.2) such that, for any* $u \in H_0^1(D)$ *and any edge* $e \subset \Gamma$*, it holds that*

$$\|u - \mathcal{SZ}(u)\|_{L^2(e)} \leq C \left(\frac{H_e}{p_e}\right)^{1/2} |u|_{H^1(\omega_e)} \tag{B.11}$$

where ω_e *is the union of all the elements who share a vertex with the edge* e*,* H_e *is the length of the edge* e *and* $p_e = \min\{p_K \mid e \subset \partial K\}$.

Résumé Le travail de cette thèse a porté sur la simulation numérique des matériaux multi-échelles. On considère des matériaux hétérogènes dont les propriétés physiques ou mécaniques (conductivité thermique, tenseur d'élasticité, ...) varient à une échelle petite par rapport à la taille du matériau. La thèse s'articule en deux parties qui correspondent à deux aspects différents des problèmes multi-échelles.
Dans la première partie, on se place dans le cadre de l'homogénéisation aléatoire et on s'intéresse à une question plus fine que la caractérisation d'un comportement moyen : on cherche à étudier les fluctuations de la réponse. Plus généralement, nous visons à comprendre : (i) quels paramètres de la distribution des coefficients du matériau à l'échelle fine affectent la distribution de la réponse à l'échelle macroscopique, et (ii) s'il est possible d'estimer cette distribution sans utiliser une méthode type Monte-Carlo, très couteuse. Sur le plan théorique, nous avons considéré un matériau faiblement aléatoire (micro-structure périodique avec ajout d'une perturbation aléatoire petite). Nous avons montré qu'en utilisant le correcteur standard issu de la théorie de l'homogénéisation aléatoire, nous sommes capables de calculer un tenseur $\mathcal{Q}$ qui gouverne complètement les fluctuations de la réponse. Ce tenseur, défini par une formule explicite, permet d'estimer la fluctuation de la réponse sans résoudre le problème fin pour de nombreuses réalisations. Une stratégie d'approximation numérique de ce tenseur a ensuite été développée et testée numériquement dans des cas plus généraux.
Dans la deuxième partie de la thèse, on considère un matériau hétérogène déterministe fixé où les hypothèses classiques d'homogénéisation (périodicité, ...) ne sont pas vérifiées. Les méthodes de résolution standard type Éléments Finis donnent de mauvaises approximations. Pour pallier cette difficulté, la Méthode des Éléments Finis Multi-échelles (MsFEM) a été introduite il y une vingtaine d'années. La méthode MsFEM se décompose en deux étapes : (i) créer un espace d'approximation grossier engendré par les solutions de problèmes locaux bien choisis; (ii) approximer la solution avec une approche de Galerkine peu couteuse sur l'espace construit dans (i). Dans cette deuxième partie, plusieurs taches ont été réalisées. Tout d'abord, une implémentation de plusieurs variantes MsFEM a été effectuée sous forme de templates dans le logiciel de calcul Éléments Finis FreeFem++. Par ailleurs, plusieurs variantes des MsFEM pâtissent d'une erreur dite de résonance : lorsque la taille des hétérogénéités est proche de la taille du maillage grossier, la méthode devient très imprécise. Pour pallier ce problème, une méthode MsFEM enrichie a été développée : à la base MsFEM classique on rajoute des solutions de problèmes locaux ayant pour conditions aux limites des polynômes de haut degré. L'utilisation de polynômes nous permet d'obtenir une convergence de l'approche à des coûts de calcul raisonnables.

Summary This thesis is about the numerical approximation of multi-scale materials. We consider heterogeneous materials whose physical or mechanical (thermal conductivity, elasticity tensor, ...) vary on a small scale compared to the material length. This thesis is composed of two parts describing two different aspects of multi-scale problems.
In the first part, we consider the stochastic homogenization framework. The aim here is to go beyond the identification of an effective behavior, by attempting to characterize the fluctuations of the response. Generally speaking we strive to understand: (i) what parameters of the distribution of the material coefficient affect the distribution of the response and (ii) if it is possible to approximate this distribution without resorting to a costly Monte-Carlo method. On the theoretical standpoint, we consider a weakly random material (the micro-structure is periodic and presents some small random defects). We show that we are able to compute a tensor $\mathcal{Q}$ that governs completely the fluctuations of the response, thanks to the use of standard corrector functions from the stochastic homogenization theory. This tensor is defined by an explicit formula and allows us to estimate the fluctuation of the response without solving the fine problem for many realizations. A numerical approximation of this tensor has been proposed and numerical experiments have been performed in broader random frameworks to assess the effectiveness of the approach.
In the second part, we consider a heterogeneous deterministic material where classical homogenization (periodicity, ...) assumptions are not satisfied. Standard methods such as Finite Elements give bad approximations. In order to solve this issue the Multi-scale Finite Element Method (MsFEM) can be used. This approach proceeds in two steps: (i) design a coarse approximation space spanned by solutions to well-chosen local problems; (ii) approximate the solution by an inexpensive Galerkin approach on the space designed in (i). On this topic, we first implemented the main variants of the MsFEM methods in the Finite Element software FreeFem++ on template form. Second, many MsFEM approaches suffer from resonance error: when the size of the heterogeneities is close to the coarse mesh size the accuracy decreases. In order to circumvent this issue, we designed an enriched MsFEM method: to the classical MsFEM basis, we add solutions to local problems with high degree polynomial boundary conditions. The use of polynomials allows us to obtain a converging approach for a limited computational cost.