HAL
open science

Some problems related to statistical error in stochastic homogenization

William Minvielle

To cite this version:

William Minvielle. Some problems related to statistical error in stochastic homogenization. Numerical Analysis [math.NA]. Université Paris-Est Marne la Vallée, 2015. English. NNT : . tel-01429837

HAL Id: tel-01429837

https://hal.science/tel-01429837v1

Submitted on 9 Jan 2017

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

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

L'École Doctorale Mathématiques et Sciences et Technologies de l'Information et de la Communication (MSTIC)

Thèse de doctorat

présentée pour obtenir le titre de

Docteur de l'Université Paris-Est
Spécialité : Mathématiques

présentée par

William Minvielle

Sujet : *Quelques problèmes liés à l'erreur statistique en homogénéisation stochastique*

(Some problems related to statistical error in stochastic homogenization)

Soutenue le 25 septembre 2015 devant le jury composé de :

Rapporteurs :	M. Yves Achdou M. James Nolen
Examinateurs :	M. Grégoire Allaire M. Antoine Lejay
Directeurs de Thèse :	Claude Le Bris Frédéric Legoll

Titre : Quelques problèmes liés à l'erreur statistique en homogénéisation stochastique.

Résumé : Le travail de cette thèse a porté sur le développement de techniques numériques pour l'homogénéisation d'équations dont les coefficients présentent des hétérogénéités aléatoires à petite échelle. Les difficultés liées à la résolution de telles équations aux dérivées partielles peuvent être résolues grâce à la théorie de l'homogénéisation stochastique. On substitue alors la résolution d'une équation dont les coefficients sont aléatoires et oscillants à l'échelle la plus fine du problème par la résolution d'une équation à coefficients constants. Cependant, une difficulté subsiste : le calcul de ces coefficients dits homogénéisés sont définis par une moyenne ergodique, que l'on ne peut atteindre en pratique. Seuls des approximations aléatoires de ces quantités déterministes sont calculables, et l'erreur commise lors de l'approximation est importante. Ces questions sont développées en détail dans le **Chapitre** 1 qui tient lieu d'introduction. L'objet du **Chapitre** 2 de cette thèse est de réduire l'erreur de cette approximation dans un cas nonlinéaire, en réduisant la variance de l'estimateur par la méthode des variables antithétiques. Dans le **Chapitre** 3, on montre comment obtenir une meilleure réduction de variance par la méthode des variables de contrôle. Cette approche repose sur un modèle approché, disponible dans le cas étudié. Elle est plus invasive et moins générique, on l'étudie dans un cas linéaire. Dans le **Chapitre** 4, à nouveau dans un cas linéaire, on introduit une méthode de sélection pour réduire l'erreur commise. Enfin, le **Chapitre** 5 porte sur l'analyse d'un problème inverse, où l'on recherche des paramètres à l'échelle la plus fine, ne connaissant que quelques quantités macroscopiques, par exemple les coefficients homogénéisés du modèle.

Mots clefs : Equations aux dérivées partielles, Homogénéisation, Matériaux aléatoires, Méthodes de Monte Carlo, Réduction de variance, Problème inverse.

Title: Some problems related to statistical error in stochastic homogenization.

Abstract: In this thesis, we design numerical techniques to address the homogenization of equations the coefficients of which exhibit small scale random heterogeneities. Solving such elliptic partial differential equations is prohibitively expensive. One may use stochastic homogenization theory to reduce the complexity of this task. We then substitute the random, fine scale oscillating coefficients of the equation with constant homogenized coefficients. These coefficients are defined through an ergodic average inaccessible to practical computation. Only random approximations thereof are available. The error committed in this approximation is significant. These issues are detailed in the introductory **Chapter** 1. In **Chapter** 2, we show how to reduce the error in this approximation, in a nonlinear case, by using an antithetic variable estimator that has a smaller variance than the standard Monte Carlo estimator. In **Chapter** 3, in a linear case, we show how to obtain an even better variance reduction with the control variate method. Such a method is based on a surrogate model. In **Chapter** 4, we use a selection method to reduce the global error. **Chapter** 5 is devoted to the analysis of an inverse problem, wherein we seek parameters at the fine scale whilst only being provided with a handful of macroscopic quantities, among which the homogenized coefficients.

Keywords: Partial differential equations, Homogenization, Random materials, Monte Carlo methods, Variance reduction, Inverse problem.

Contents

1 Introduction **11**

1.1 Introduction générale . . . 11

1.2 Introduction mathématique . . . 13

1.2.1 Modélisation . . . 13

1.2.2 Stationnarité . . . 14

1.2.3 Homogénéisation . . . 15

1.2.4 Décomposition de l'erreur . . . 16

1.2.5 Quantification de la convergence . . . 16

1.3 Résumé des travaux . . . 17

1.3.1 Réduction de variance par variables antithétiques pour un problème d'homogénéisation stochastique nonlinéaire convexe. 17

1.3.2 Une approche par variables de contrôle basée sur une théorie perturbative pour la réduction de variance en homogénéisation stochastique. 20

1.3.3 Structures quasi-aléatoires spéciales : une approche de sélection pour l'homogénéisation stochastique. . . . 22

1.3.4 Un problème d'identification de paramètres en homogénéisation stochastique. . . . 25

1.4 Perspectives . . . 28

1.4.1 Comparaison des méthodes de réduction de variance . . . 28

1.4.2 Extension à la stationnarité continue des variables de contrôle . . . 28

1.4.3 Utilisation d'autres variables de contrôle . . . 29

1.4.4 Sélection pour des problèmes nonlinéaires . . . 29

1.4.5 Convergence du problème inverse en dimension supérieure . . . 29

1.4.6 Perspectives générales dans le champ de recherche . . . 30

2 Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem **31**

2.1 Introduction . . . 32

2.1.1 Homogenization theoretical setting . . . 33

2.1.2 The questions we consider . . . 35

2.2 Description of the proposed approach and main results . . . 36

2.2.1 Statement of our main results . . . 36

2.2.2 Classical results on antithetic variables . . . 40

2.2.3 Derivatives of the corrector and of the homogenized energy density . 40

2.2.4 Monotonicity properties . . . 41

2.2.5 Proof of Propositions 2.1, 2.2 and 2.5 . . . 45

2.2.6 Examples satisfying our structure assumptions . . . 46

2.3 Numerical results . . . 48

2.3.1 Newton algorithm to solve the truncated corrector problem . . . 48

2.3.2 Overview of numerical results . . . 49
2.3.3 Test Case 1 . . . 50
2.3.4 Test Case 2 . . . 53
2.3.5 Test Case 3 . . . 55
2.4 Appendix: Proof of (2.8) . . . 55

3 A control variate approach based on a defect-type theory for variance reduction in stochastic homogenization 59
3.1 Introduction . . . 60
3.1.1 Homogenization theoretical setting . . . 62
3.1.2 Practical approximation of the homogenized matrix . . . 63
3.1.3 Control variate approach . . . 64
3.2 A weakly random setting: rare defects in a periodic structure . . . 66
3.2.1 Presentation of the model . . . 66
3.2.2 Weakly-random homogenization result . . . 67
3.3 Control variate approaches for stochastic homogenization . . . 69
3.3.1 A first-order model . . . 69
3.3.2 A second-order model . . . 72
3.4 Elements of theoretical analysis . . . 74
3.4.1 One-dimensional case . . . 74
3.4.2 Multi-dimensional case . . . 79
3.5 Numerical results . . . 81
3.5.1 Low contrast test-case . . . 82
3.5.2 High contrast test-case . . . 84
3.5.3 Using a Reduced Basis (RB) approach . . . 84

4 Special Quasirandom Structures: a selection approach for stochastic homogenization 91
4.1 Introduction . . . 92
4.1.1 Overview . . . 92
4.1.2 Theoretical setting . . . 94
4.1.3 Numerical approximation of the homogenized matrix . . . 95
4.2 Variance reduction approach . . . 96
4.2.1 Original formulation of the SQS approach . . . 96
4.2.2 Formal derivation of the SQS conditions using a perturbative setting 97
4.2.3 Practical evaluation of the SQS conditions . . . 101
4.2.4 Selection Monte Carlo sampling . . . 105
4.3 Elements of theoretical analysis . . . 107
4.3.1 Proof of convergence of the approach . . . 107
4.3.2 Complete analysis in some simple cases . . . 114
4.4 Numerical experiments . . . 117
4.4.1 Robustness of the approach . . . 117
4.4.2 Efficiency of the approach . . . 118
4.5 Proof of Lemma 4.4 . . . 125

5 A parameter identification problem in stochastic homogenization 129
5.1 Introduction . . . 131
5.2 Discrete homogenization theory . . . 133
5.2.1 Homogenization result . . . 133

5.2.2 Approximation on finite boxes . . . 135
5.2.3 Physical problem . . . 137
5.2.4 The one dimensional case . . . 141
5.3 A parameter fitting problem . . . 145
5.3.1 General case . . . 145
5.3.2 The one-dimensional case . . . 146
5.4 Numerical results . . . 150
5.4.1 Optimization algorithm . . . 151
5.4.2 Numerical results . . . 151
5.5 Appendix: Computation of the derivatives of (5.50) . . . 154

A Neumann, Dirichlet and periodic approximations of the corrector problem 155
A.1 Introduction . . . 155
A.2 Approximations of the corrector . . . 155
A.2.1 The Dirichlet approximation . . . 156
A.2.2 The periodic approximation . . . 156
A.2.3 Neumann approximation . . . 157
A.3 Flux formulation . . . 158
A.3.1 The H^{div} space . . . 158
A.3.2 The periodic $H^{\mathrm{div}}_{\mathrm{per}}$ space . . . 159
A.3.3 Flux formulation of the periodic problem (A.4) . . . 159
A.3.4 Flux formulation of the Neumann problem (A.9) . . . 160
A.4 Comparison of the approximations . . . 161
A.4.1 Comparison of the Dirichlet and periodic approximations . . . 161
A.4.2 Comparison of the Neumann and periodic approximations . . . 162

Bibliography 165

Publications

- [LM15b] F. Legoll and W. Minvielle.
  Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem.
  Discrete and Continuous Dynamical Systems - Series S, 8(1):1–27, 2015.

- [LM15a] F. Legoll and W. Minvielle.
  A control variate approach based on a defect-type theory for variance reduction in stochastic homogenization.
  Multiscale Modeling & Simulation, 13(2):519-550, 2015.

- [LMOS15] F. Legoll, W. Minvielle, A. Obliger and M. Simon.
  A parameter identification problem in stochastic homogenization.
  ESAIM: Proc., 48:190-214, 2015.

- [LBLM] C. Le Bris, F. Legoll and W. Minvielle.
  Special Quasirandom Structures: a selection approach for stochastic homogenization.
  To be submitted.

Chapter 1

Introduction

1.1 Introduction générale

Les matériaux multiéchelles sont des matériaux présentant des caractéristiques qui varient selon l'échelle à laquelle on les examine. Ainsi, un béton contient des pores inclus dans une matrice solide, la taille de ces pores pouvant être micrométrique. Le comportement (par exemple élastique) macroscopique de l'ensemble est une combinaison du comportement des différentes phases, mais ne saurait se contenter d'une simple moyenne algébrique des rigidités de chaque phase. En d'autres termes, bien que le calcul intéressant l'ingénieur (par exemple, certifier la résistance du matériau) soit fait à une échelle de l'ordre du mètre au moins, la description mécanique du matériau est effectuée à une échelle bien plus fine. Ces matériaux multiéchelles sont présents dans nombre de produits industriels (automobile, aéronautique, génie civil). Par exemple, dans le secteur de l'énergie nucléaire, des matériaux multiéchelles sont utilisés pour construire des réacteurs nucléaires (bétons, aciers) et envisagés pour le stockage des déchets radioactifs (argiles).

Comment garantir la sécurité des ouvrages et anticiper leur comportement ? Les techniques usuelles d'expérimentation physique peuvent être inadaptées si le matériau coûte trop cher à produire, ou si l'on souhaite prédire ou concevoir un matériau nouveau. L'expérimentation *numérique* prend alors toute sa place. Cependant, la résolution des équations de la mécanique sur des échelles macroscopiques pour des matériaux multiéchelles n'est pas *a priori* aisée, car reproduire fidèlement la microstructure requiert déjà de mailler à une échelle très fine, sans compter la résolution subséquente des équations. Pour autant, il existe des techniques spécialisées pour de tels problèmes, et l'homogénéisation - dont il est question dans cette thèse - en est une.

L'homogénéisation est une théorie mathématique qui permet de simplifier la description du comportement de matériaux multiéchelles. A l'échelle macroscopique, les propriétés du milieu sont décrites par des propriétés moyennes, constantes ou variant lentement selon la variable d'espace. C'est un modèle adapté pour mener des calculs à l'échelle la plus grande, car il n'est plus nécessaire d'avoir une résolution de calcul très fine. Le contrôle de l'erreur commise lors de cette substitution est exprimée par un paramètre de *séparation d'échelle* $\varepsilon \ll 1$, qui quantifie le rapport de tailles caractéristiques entre la structure microscopique et la taille macroscopique de l'échantillon sur lequel porte le calcul. Il existe des situations où un continuum d'échelles doivent être prises en compte, et la séparation d'échelle n'a pas lieu. Les techniques (pratiques ou théoriques) pour les traiter sont encore largement exploratoires, et nous ne couvrons pas ce cas dans cette thèse. Lorsqu'il y a séparation d'échelle, plus ε est petit, plus le modèle homogénéisé est précis.

Cette théorie propose ainsi de remplacer le calcul dans un milieu hétérogène par un calcul dans un milieu homogène. Le calcul des coefficients du milieu homogène est l'étape cruciale en homogénéisation, et aussi la plus coûteuse.

Une hypothèse de *structure* est nécessaire pour obtenir des expressions explicites. Dans le livre fondateur [BLP78], les auteurs supposent que la structure microscopique est la répétition *périodique* d'un même motif. Le comportement moyen est déductible de la résolution d'un problème "typique" sur la maille périodique (le problème du *correcteur*, parfois appelé problème de cellule dans ce cas).

Cependant, l'hypothèse de périodicité est parfois restrictive. Si des matériaux peuvent assez bien la satisfaire, d'autres s'en éloignent nettement. Un exemple d'hypothèse de structure peu contraignante est la stationnarité (voir les articles fondateurs [Koz79, PV81]). On suppose que le matériau observé correspond à *une réalisation* d'un matériau aléatoire, qui possède une propriété d'invariance spatiale de sa *loi*. La périodicité correspond alors à un cas particulier. Certaines situations, par exemple un réseau cristallin perturbé par un défaut local ayant une influence importante (dopage de semi-conducteurs par exemple), ne peuvent pas être modélisé comme un problème aléatoire stationnaire, toutefois cette classe est très vaste.

Cette généralité vient avec certains inconvénients. D'un point de vue pratique, il faut connaître de la loi du milieu de référence, ce qui peut être difficile s'agissant de quantités microscopiques. De plus, l'équation du correcteur permettant de calculer le comportement moyen change de niveau de complexité par rapport au cas périodique : il s'agit d'une équation à coefficients aléatoires, posée dans un domaine non borné. Intuitivement, on ne peut pas se contenter d'un domaine fini pour décrire la variété infinie de la loi de cette variable aléatoire. Ainsi, pour simplifier la description du comportement macroscopique, il faut d'abord en passer par un calcul presque aussi complexe que le problème original, avec toutefois une simplification importante : le chargement correspond à un chargement de référence et non au chargement originel pour lequel on cherche à calculer la solution. Par conséquent, d'un point de vue pratique, l'homogénéisation aléatoire ne permet pas de réduire significativement la complexité du problème de départ si l'objectif n'est d'effectuer qu'un seul calcul, pour un seul chargement et une seule occurrence des paramètres de la loi de la microstructure. Toutefois, il existe des situations où il faut effectuer des calculs *répétés*, pour de nombreux chargements différents. Ainsi, pour résoudre un problème d'évolution temporelle du matériau multiéchelle, il faudra résoudre de manière répétée, pour différents chargements, l'équation statique de ce matériau. De même, la résolution d'un problème *inverse* comprend nécessairement l'évaluation répétée, pour différentes valeurs de paramètres, du calcul *direct*. Citons enfin la décomposition de domaine, où la solution du problème est obtenue par itérations successives de la résolution du problème par sous-domaines, les itérations ayant pour but de calculer la solution aux interfaces. L'homogénéisation aléatoire réduit alors globalement la difficulté : au lieu de devoir effectuer de très nombreux calculs coûteux, un seul sera nécessaire - afin de déterminer milieu homogène associé - après quoi les suivants seront réalisés au moyen du milieu homogène, donc rapides.

D'autres méthodes multiéchelles, alternatives à l'homogénéisation, ont été développées. Citons par exemple l'approche MsFEM (méthode des éléments finis multiéchelles). L'homogénéisation constitue alors un guide pour justifier ces méthodes récentes et encore exploratoires.

L'intérêt de l'homogénéisation étant établi, il reste que le calcul du coefficient homogénéisé est coûteux. Les approches considérées dans cette thèse ont pour objectif de

rendre ce calcul plus efficace.

Puisque le problème du correcteur est posé sur un domaine non borné, il faut pour le résoudre numériquement le tronquer sur un domaine borné. Du fait de cette troncature, on obtient alors une approximation aléatoire du milieu homogène, de laquelle on approche l'espérance grâce à une méthode de Monte Carlo : on considère un grand nombre de réalisations de l'environnement pour lesquelles on résout le problème du correcteur, et l'on considère la moyenne empirique de ces coefficients homogénéisé *apparents* pour approcher le coefficient homogénéisé exact. L'erreur totale est en grande partie liée à la dispersion statistique des coefficients apparents.

Dans les **Chapitre** 2 et **Chapitre** 3, nous construisons des approximations ayant la même espérance que l'estimateur de Monte Carlo usuel, mais avec une dispersion statistique plus faible. L'approximation du coefficient homogénéisé construit selon une telle approche sera moins entaché d'erreur. Dans le **Chapitre** 2, on a recours à une méthode de réduction de variance classique et peu dépendante du problème, la technique des *variables antithétiques*. Dans le **Chapitre** 3, on utilise la technique des *variables de contrôle*, nécessitant une connaissance du milieu de référence plus précise et un travail préparatoire. Nous poursuivons vers une méthode davantage liée au problème, donc obtenant une meilleure réduction d'erreur, dans le **Chapitre** 4 où nous développons une méthode de *sélection*. L'objectif du **Chapitre** 5 est différent. On s'intéresse à un problème *inverse*, et l'on cherche les paramètres de la loi de l'environnement microscopique connaissant des quantités macroscopiques (par exemple les coefficients homogénéisés).

1.2 Introduction mathématique

1.2.1 Modélisation

Commençons par un exemple représentatif des problèmes qui seront traités dans cette thèse. Soit d la dimension ambiante. On considère un domaine borné et régulier $\mathcal{D} \subset \mathbb{R}^d$ qui représente le solide que l'on cherche à modéliser. En élasticité linéaire, la grandeur caractéristique de ce solide est le tenseur d'élasticité A_ε qui est un tenseur d'ordre 4, symétrique coercif et borné, reliant le déplacement à la contrainte selon une relation constitutive linéaire. L'indice ε souligne que cette rigidité est une caractéristique *locale* du solide, qui varie à l'échelle $\varepsilon \ll 1$. Un problème typique consiste à chercher le déplacement $u_\varepsilon \in H^1(\mathcal{D})^d$ d'un solide soumis à des forces volumiques $f \in H^{-1}(\mathcal{D})^d$ et attaché au bord $\partial\mathcal{D}$. L'équation aux dérivées partielle résultante s'écrit alors

$$\begin{cases} -\operatorname{div} A_\varepsilon : \nabla^s u_\varepsilon = f & \text{dans } \mathcal{D}, \\ u_\varepsilon = 0 & \text{sur } \partial\mathcal{D}. \end{cases} \tag{1.1}$$

On a noté $\nabla^s : H^1(\mathcal{D})^d \mapsto L^2(\mathcal{D})^{d\times d}$ la partie symétrique du gradient. Il est possible d'adapter les techniques, classiques ou nouvelles, ainsi que leurs justifications mathématiques, à cette équation. Cependant, pour plus de simplicité dans l'expérimentation numérique, nous avons choisi dans cette thèse de travailler sur une version scalaire de cette équation, qui conserve néanmoins toute la difficulté du caractère multiéchelle :

$$\begin{cases} -\operatorname{div} A_\varepsilon \nabla u_\varepsilon = f & \text{dans } \mathcal{D}, \\ u_\varepsilon = 0 & \text{sur } \partial\mathcal{D}. \end{cases} \tag{1.2}$$

Dans (1.2), le second membre $f \in H^{-1}(\mathcal{D})$ est une fonction scalaire, et non vectorielle. Le champ $A_\varepsilon \in L^\infty(\mathcal{D})^{d\times d}$ est matriciel et l'inconnue $u_\varepsilon \in H^1(\mathcal{D})$ est scalaire. De très

nombreuses situations physiques sont modélisées par (1.2). La conduction thermique en est une : l'équation (1.2) gouverne la température u_ε ; A_ε est le coefficient de diffusion thermique et f le terme source. L'électrostatique, où u_ε est le potentiel scalaire, f la distribution de charge et A_ε la perméabilité diélectrique du milieu en est une autre.

L'équation (1.2) est typique en sciences de l'ingénieur, mais d'autres modèles, parfois plus complexes, existent. En particulier, dans le **Chapitre** 2 sera explorée une version plus générale et nonlinéaire de ce problème, et dans le **Chapitre** 5 nous verrons une version discrète de l'équation aux dérivées partielles (1.2).

1.2.2 Stationnarité

Nous supposerons que A_ε s'écrit sous la forme $A_\varepsilon(x,\omega) = A\left(\frac{x}{\varepsilon},\omega\right)$ où $\omega \in \Omega$ représente la variable d'aléa, et $x \in \mathbb{R}^d$ la variable d'espace. C'est une hypothèse importante, selon laquelle le champ A_ε est identique en chaque point macroscopique. En outre, nous supposerons que A est *stationnaire*, c'est à dire possède une propriété d'invariance spatiale de sa loi : celle-ci est la même en chaque point microscopique. Le paragraphe suivant précise cette propriété.

Soit $(\Omega, \mathcal{F}, \mathbb{P})$ un espace de probabilité. On suppose que le groupe $(\mathbb{Z}^d, +)$ agit sur Ω. On note $(\tau_k)_{k\in\mathbb{Z}^d}$ cette action, et on suppose qu'elle préserve la mesure $\mathbb{P}$, c'est-à-dire que pour tout $k \in \mathbb{Z}^d$ et $B \in \mathcal{F}$, $\mathbb{P}(\tau_k B) = \mathbb{P}(B)$. Nous supposons que cette action est *ergodique* : si $B \in \mathcal{F}$ est invariant selon tous les τ_k, alors $\mathbb{P}(B) = 0$ ou 1.

Définition 1.1. *Une fonction $F \in L^1_{\text{loc}}\left(\mathbb{R}^d, L^1(\Omega)\right)$ est* stationnaire *si*

$$\forall k \in \mathbb{Z}^d, \quad F(x+k,\omega) = F(x,\tau_k\omega) \quad \textit{presque partout et presque sûrement.} \tag{1.3}$$

Les fonctions stationnaires (au sens de la stationnarité *discrète* comme ci-dessus, ou *continue* comme ci-après dans la Remarque 1.5, pour ces dernières $\mathbb{R}^d$ remplace $\mathbb{Z}^d$ dans l'équation (1.3)) fournissent le socle de la théorie de l'homogénéisation aléatoire et vérifient le théorème ergodique (voir [Kre85, Shi84, Tem72]) suivant :

Théorème 1.2. *Soit une fonction stationnaire $F \in L^\infty\left(\mathbb{R}^d, L^1(\Omega)\right)$. On note, pour $k \in \mathbb{Z}^d$, $|k|_\infty = \sup_{1\le i\le d} |k_i|$. Alors*

$$\frac{1}{(2N+1)^d} \sum_{|k|_\infty \le N} F(x,\tau_k\omega) \xrightarrow[N\to\infty]{} \mathbb{E}\left(F(x,\cdot)\right) \quad \textit{dans } L^\infty(\mathbb{R}^d), \textit{ presque sûrement.}$$

En conséquence, en appelant Q le cube unité de $\mathbb{R}^d$,

$$F\left(\frac{x}{\varepsilon},\omega\right) \xrightarrow[\varepsilon\to 0]{*} \mathbb{E}\left(\int_Q F(x,\cdot)dx\right) \quad \textit{dans } L^\infty(\mathbb{R}^d), \textit{ presque sûrement.} \tag{1.4}$$

Autrement dit, les fonctions stationnaires admettent des moyennes sur les grands volumes.

Remarque 1.3. *Dans le Théorème 1.2, si la fonction F est $(0,1)^d$-périodique, on retrouve alors une version du Lemme de Riemann-Lebesgue :*

$$\forall \varphi \in L^1(\mathbb{R}^d), \quad \int_{\mathbb{R}^d} \varphi(x) F\left(\frac{x}{\varepsilon}\right) dx \xrightarrow[\varepsilon\to 0]{} \int_{(0,1)^d} F(y)dy \int_{\mathbb{R}^d} \varphi(x)dx.$$

Remarque 1.4. *Une autre hypothèse du même type, mais différente, est la stationnarité* continue. *On suppose que $\mathbb{R}^d$ agit sur Ω selon une action $(\tau_y)_{y\in\mathbb{R}^d}$, et la stationnarité de F s'écrit alors*

$$\forall y \in \mathbb{R}^d, \quad F(x+y,\omega) = F(x,\tau_y\omega) \quad \textit{presque partout et presque sûrement.} \tag{1.5}$$

La relation (1.4) demeure en remplaçant l'opérateur de moyenne $\mathbb{E}\int_Q$ par $\mathbb{E}$.

1.2.3 Homogénéisation

Sous l'hypothèse de stationnarité de A, on peut énoncer le résultat d'homogénéisation suivant (voir [Koz79, PV81]) :

Théorème 1.5. *Soit $A \in L^\infty(\mathbb{R}^d \times \Omega)^{d\times d}$. On suppose que*

- (ellipticité) *: il existe une constante $\alpha > 0$ telle que, pour presque tout $x \in \mathbb{R}^d$, pour tout $p \in \mathbb{R}^d$ et presque sûrement en $\omega \in \Omega$,*

$$\alpha|p|^2 \leq p \cdot A(x,\omega)p.$$

- (stationnarité) *: A est stationnaire au sens de (1.3).*

Soit $\mathcal{D}$ un domaine borné de $\mathbb{R}^d$, $f \in H^{-1}(\mathcal{D})$ et $A_\varepsilon(x,\omega) := A(x/\varepsilon,\omega)$. Alors, presque sûrement, la solution du problème (1.2) converge faiblement dans $H^1_0(\mathcal{D})$ vers $u^\star \in H^1_0(\mathcal{D})$ solution de l'équation

$$\begin{cases} -\operatorname{div} A^\star \nabla u^\star = f & \textit{dans } \mathcal{D}, \\ u^\star = 0 & \textit{sur } \partial\mathcal{D}. \end{cases} \tag{1.6}$$

La matrice constante, déterministe et α-coercive $A^\star \in \mathbb{R}^{d\times d}$ est donnée par

$$\forall p \in \mathbb{R}^d, \quad A^\star p := \mathbb{E}\int_Q A(p+\nabla w_p), \tag{1.7}$$

où le correcteur $w_p \in L^2(\Omega; L^2_{\mathrm{loc}}(\mathbb{R}^d))$ est l'unique solution (à une constante additive près) dont le gradient est dans l'espace $\nabla w_p \in L^2(\Omega; L^2_{\mathrm{unif}}(\mathbb{R}^d))^d$ de l'équation

$$\begin{cases} -\operatorname{div} A(p+\nabla w_p) = 0 & \textit{dans } \mathbb{R}^d, \\ \nabla w_p \quad \textit{stationnaire et} & \mathbb{E}\int_Q \nabla w_p = 0. \end{cases} \tag{1.8}$$

La convergence donnée par le Théorème 1.5 justifie l'approximation de la solution de (1.2) par celle de (1.6) lorsque $\varepsilon \ll 1$. Au contraire de (1.2), le problème (1.6) est facile à résoudre numériquement puisque $A^\star$ est constant et que l'équation est posée sur un domaine borné.

Pour autant, il demeure une difficulté dans ce programme: calculer $A^\star$ est en général difficile. Plus précisément, pour calculer $A^\star$, il faut résoudre d problèmes du correcteur (1.8) (un dans chaque direction $p = e_i$, $1 \leq i \leq d$), lesquels sont posés dans $\mathbb{R}^d$. La solution recherchée n'est pas décroissante à l'infinie, seulement strictement sous-linéaire. Contrairement au cas périodique, dans le cas aléatoire, on ne peut pas transformer ce problème en un problème de cellule sur un domaine compact[1].

[1] Plus exactement, le domaine permettant de recouvrer assez de compacité est l'espace de probabilité. Malheureusement, cet espace est trop abstrait pour aider à la résolution numérique du problème du correcteur.

Une solution pour résoudre (1.8) consiste à tronquer le domaine de résolution. Nous résolvons l'équation sur un sous-domaine $Q_N := (0,N)^d$ de taille finie et aussi grande que possible en la munissant de conditions aux limites adéquates. Par exemple, on considère dans la suite le problème

$$\begin{cases} -\operatorname{div} A(p+\nabla w_p^N) = 0 \quad \text{dans } \mathbb{R}^d, \\ \qquad w_p^N \text{ est } Q_N \text{ périodique.} \end{cases} \tag{1.9}$$

Les conditions aux limites dans (1.9) ont été choisies périodiques. D'autres conditions peuvent être utilisées ; elles seront discutées dans la Section A.2. En conséquence de ces conditions de bords périodiques, le problème (1.9) se ramène à un problème sur le domaine compact Q_N vu comme un tore. De cette approximation du correcteur nous déduisons une approximation $A^{\star,N}(\omega)$ de $A^\star$ définie par

$$\forall p \in \mathbb{R}^d, \quad A^{\star,N}(\omega)p := \frac{1}{|Q_N|}\int_{Q_N} A(x,\omega)(p+\nabla w_p^N(x,\omega))dx. \tag{1.10}$$

Cette approximation converge presque sûrement vers $A^\star$ d'après un résultat de [BP04, Theorem 1]. Voir aussi la Section A.2.2.

1.2.4 Décomposition de l'erreur

Pour autant, cette approximation souffre de plusieurs faiblesses. Premièrement, en quantifier la convergence est un problème délicat. Nous y revenons dans la Section 1.2.5. Deuxième obstacle, il s'agit d'une approximation aléatoire d'une quantité déterministe. Pire, même l'espérance de cette approximation est distincte de $A^\star$. Commençons par décomposer l'erreur selon deux contributions orthogonales dans $L^2(\Omega)$:

$$A^\star - A^{\star,N}(\omega) \quad = \quad A^\star - \mathbb{E}[A^{\star,N}] \quad + \quad \mathbb{E}[A^{\star,N}] - A^{\star,N}(\omega). \tag{1.11}$$

La première partie de l'erreur dans (1.11), appelée erreur *systématique*, est complètement déterministe. Nous verrons dans la Section 1.2.5 qu'elle n'est pas dominante. La seconde erreur est appelée erreur *statistique*, elle dérive de la dispersion statistique de $A^{\star,N}$ autour de sa moyenne. Une première manière de réduire cette erreur est de considérer M copies indépendantes et identiquement distribuées de $A^{\star,N}$, notées $(A_j^{\star,N})_{1\le j\le M}$, en résolvant (1.9) pour différents environnements aléatoires, puis d'en prendre la moyenne empirique

$$\mathcal{I}_{MC}^M(\omega) := \frac{1}{M}\sum_{j=1}^M A_j^{\star,N}(\omega), \tag{1.12}$$

afin d'obtenir un *estimateur* (approximation aléatoire) de $\mathbb{E}[A^{\star,N}]$. Par le théorème de la limite centrale, l'erreur commise $|\mathcal{I}_{MC}^M - \mathbb{E}[A^{\star,N}]|$ décroît en loi comme $\sqrt{\frac{\mathbb{V}\mathrm{ar}A^{\star,N}}{M}}$ lorsque M tend vers l'infini. Notons que $\mathbb{V}\mathrm{ar}A^{\star,N}$ tend vers 0 lorsque N tend vers l'infini. Dans les **Chapitre** 2 et **Chapitre** 3, nous introduisons des méthodes pour réduire cette erreur statistique, en construisant une approximation de même espérance $\mathbb{E}[A^{\star,N}]$ que $\mathcal{I}_{MC}^M$ mais de variance plus faible. L'erreur statistique est donc réduite, et l'erreur systématique reste la même. Dans le **Chapitre** 4, nous réduisons globalement l'erreur totale.

1.2.5 Quantification de la convergence

La motivation principale d'une partie des travaux de cette thèse est l'amélioration de la convergence d'estimateurs particuliers vers $A^\star$, plus performants quee $\mathcal{I}_{MC}^M$ et que nous introduirons ci-dessous. Étudier théoriquement l'avantage de ces nouveaux estimateurs n'est

pas simple. En effet, la simple *quantification* théorique de l'erreur commise dans (1.11), par exemple sous la forme d'une estimation d'erreur $|A^\star - \mathbb{E}[A^{\star,N}]| \leq CN^{-\alpha}$, est complexe. Les premiers travaux en ce sens proviennent de [Yur86]. Ces résultats non optimaux n'ont été que récemment complétés, d'abord en homogénéisation discrète (voir le **Chapitre** 5 pour une présentation de cette modélisation). Nous renvoyons à l'article [GNO14] pour une revue récente de ces avancées. Dans un second temps, ils ont été pour partie étendus à des cas continus, certaines études étant toujours en cours.

L'intérêt pour notre travail, est multiple. Nous l'utilisons premièrement comme justification de nos motivations : él'analyse d'erreur confirme qu'il faut réduire l'erreur statistique, principale source de l'erreur. Deuxièmement, il s'agit d'une source d'inspiration pour établir des résultats nouveaux, et pour comparer les hypothèses requises.

1.3 Résumé des travaux

1.3.1 Réduction de variance par variables antithétiques pour un problème d'homogénéisation stochastique nonlinéaire convexe.

Dans le **Chapitre** 2, nous nous intéressons au problème d'optimisation multiéchelle et nonlinéaire suivant :

$$\mathcal{E}_\varepsilon(f,\mathcal{D}) \ := \inf\left\{\int_{\mathcal{D}} W\left(\frac{x}{\varepsilon},\omega,\nabla u(x)\right)dx - \int_{\mathcal{D}} fu, \quad u \in W_0^{1,p}(\mathcal{D})\right\}, \tag{1.13}$$

pour un certain $1 < p < \infty$, et un chargement $f \in W_0^{1,p}(\mathcal{D})'$.

Caractère bien posé de (1.13)

Considérons tout d'abord la version mono échelle et déterministe de ce problème :

$$\inf\left\{\int_{\mathcal{D}} W\left(x,\nabla u(x)\right)dx - \int_{\mathcal{D}} fu, \quad u \in W_0^{1,p}(\mathcal{D})\right\}. \tag{1.14}$$

Sous les conditions de croissance et convexité suivantes de W,

$$\exists\ 2 \leq p < \infty,\ \exists c, C > 0,\ \forall x,\xi \in \mathbb{R}^d, \quad c|\xi|^p \leq W(x,\xi) \leq C(1+|\xi|^p), \tag{1.15}$$

$$\text{L'application partielle } \xi \mapsto W(x,\xi) \text{ est convexe pour presque tout } x, \tag{1.16}$$

le problème d'optimisation (1.14) est bien posé.

Remarque 1.6. *Sous les hypothèses (1.15)–(1.16), $\xi \mapsto W(x,\xi)$ est strictement convexe.*

Ce problème est une déclinaison très simplifiée d'un problème de mécanique nonlinéaire.

Remarque 1.7. *Dans l'article [Bal76], l'auteur considère un problème dont l'inconnue $u : \mathbb{R}^d \mapsto \mathbb{R}^d$ est vectorielle plutôt que scalaire. Il s'agit du champ de déplacement d'un solide soumis à un chargement par des forces volumiques $f : \mathcal{D} \mapsto \mathbb{R}^d$, et l'énergie (1.14) représente l'énergie de déformation, qu'il convient de minimiser étant donné un chargement. L'auteur discute des hypothèses de croissance et convexité.*

Nous avons choisi l'hypothèse de croissance (1.15), alors même qu'elle n'a pas de contrepartie physique : il s'agit d'une hypothèse technique. Nous renvoyons à [JKO94, Chapter 15] pour des hypothèses plus générales.

Remarque 1.8. *Sous l'hypothèse (1.16), le problème (1.14) est équivalent à l'équation d'optimalité d'Euler-Lagrange, qui s'écrit*

$$\begin{cases} -\operatorname{div}[\partial_\xi W(x, \nabla u(x))] = f(x) & \text{dans } \mathcal{D}, \\ u(x) = 0 & \text{sur } \partial\mathcal{D}. \end{cases} \tag{1.17}$$

Cette équation est linéaire dans le cas où W s'écrit sous la forme $W(x,\xi) = \frac{1}{2}\xi \cdot A(x)\xi$ et vérifie (1.15)–(1.16) pour $p = 2$.

Homogénéisation stochastique

Afin de discuter l'homogénéisation du problème (1.13), nous supposons que

$$\text{Pour presque tout } x \in \mathbb{R}^d, \quad \xi \mapsto W(x,\omega,\xi) \quad \text{est presque sûrement continue,} \tag{1.18}$$

$$\forall \xi \in \mathbb{R}^d, \quad (x,\omega) \mapsto W(x,\omega,\xi) \quad \text{est stationnaire au sens de (1.3).} \tag{1.19}$$

L'hypothèse de stationnarité (1.19) est une hypothèse de *structure*, permettant d'obtenir des moyennes sur des grands volumes. Elle a été discutée dans la Section 1.2.2. L'hypothèse (1.18) est une condition technique classique. L'application $(x,\xi) \mapsto W(x,\omega,\xi)$, mesurable par rapport à la variable ξ et continue par rapport à la variable x, est appelée fonction de Carathéodory. Grâce à cette hypothèse, une fonction de la forme $x \mapsto W(x,\omega,g(x))$ est mesurable si g l'est.

Théorème 1.9 ([DMM86a, DMM86b])**.** *On suppose (1.15)–(1.16) et (1.18)–(1.19). Alors le problème (1.13) admet un unique minimiseur u_ε, lequel converge presque sûrement, faiblement dans $W_0^{1,p}(\mathcal{D})$, vers la solution du problème d'optimisation* homogénéisé

$$\mathcal{E}_0(f,\mathcal{D}) := \inf\left\{\int_{\mathcal{D}} W^\star(\nabla u^\star) - \int_{\mathcal{D}} fu, \quad u \in W_0^{1,p}(\mathcal{D})\right\}, \tag{1.20}$$

où la densité d'énergie homogénéisée $W^\star$ est donnée par

$$W^\star(\xi) := \lim_{N\to\infty} \inf\left\{\frac{1}{|Q_N|}\int_{Q_N} W(x,\omega,p+\nabla w(x)), \quad w \in W^{1,p}_{\text{per}}(Q_N)\right\}. \tag{1.21}$$

A l'instar du cas linéaire, on ne peut toutefois espérer atteindre cette limite, aussi nous considérons dans la suite l'approximation standard suivante

$$W_N^\star(\omega,\xi) := \inf\left\{\frac{1}{|Q_N|}\int_{Q_N} W(x,\omega,p+\nabla w(x)), \quad w \in W^{1,p}_{\text{per}}(Q_N)\right\}, \tag{1.22}$$

qui est l'analogue nonlinéaire de (1.10). La limite (1.21) a lieu presque sûrement, et par définition, $W^\star(\xi)$ est la limite presque sûre de $W_N^\star(\omega,\xi)$. Les conditions aux limites dans (1.22) peuvent être modifiées pour construire d'autres approximations convergentes. En particulier, on peut remplacer $w \in W^{1,p}_{\text{per}}(Q_N)$ par $w \in W_0^{1,p}(Q_N)$ avec le même résultat. Nous renvoyons au Lemma 2.16 pour plus de détails. Le Théorème 1.9 généralise la situation linéaire *symétrique*, réalisée pour $p = 2$ et $W(x,\omega,\xi) := \frac{1}{2}\xi \cdot A(x,\omega)\xi$, quand A est symétrique. Dans ce cas, la formulation énergétique du problème (1.2) (respectivement (1.6)) correspond exactement au problème (1.13) (respectivement (1.20)), et le Théorème 1.9 se réduit au Théorème 1.5.

Réalisations antithétiques

La technique des variables antithétiques est une méthode simple et facile à mettre en oeuvre pour réduire la variance d'une approximation. Notons d'abord qu'une variable aléatoire peut souvent s'écrire comme une certaine fonction de plusieurs variables aléatoires indépendantes et identiquement distribuées, uniformes sur $(0,1)$. Nous appellerons $X_k \sim \mathcal{U}(0,1)$ ces variables aléatoires qui représentent la « source » de l'aléa. Un cas particulier de fonctions stationnaires qui reviendra souvent dans la thèse est le suivant :

$$\chi(x,\omega) \ := \sum_{k\in\mathbb{Z}^d} g(X_k(\omega))\mathbb{1}_{k+Q}(x), \tag{1.23}$$

où $g : (0,1) \mapsto \mathbb{R}$ est une fonction bornée, et $Q = (0,1)^d$. Le champ aléatoire χ est constant sur chaque cellule $k+Q$, où il prend la valeur $g(X_k)$. Sous une hypothèse adéquate sur X_k (par exemple s'ils sont indépendants), le champ χ est stationnaire. Ainsi le cas $W(x,\omega,\xi) = f(\xi)\chi(x,\omega)$ est un exemple de W satisfaisant les hypothèses (1.18)–(1.19) si f est continue.

Notons $X_k^{\text{ant}} := 1 - X_k$ la réalisation *antithétique* à X_k. On pose $\chi^{\text{ant}}(x,\omega) := \sum_{k\in\mathbb{Z}^d} g(X_k^{\text{ant}}(\omega))\mathbb{1}_{k+Q}(x)$, et $W^{\text{ant}}(x,\omega,\xi) := f(\xi)\chi^{\text{ant}}(x,\omega)$. Le champ W^{ant} est distribué suivant la même loi que W. On définit

$$W_N^{\star,\text{ant}}(\omega,\xi) \ := \inf\left\{\frac{1}{N^d}\int_{Q_N} W^{\text{ant}}\left(x,\omega,p+\nabla w(x)\right), \quad w \in W^{1,p}_{\text{per}}(Q_N)\right\},$$

puis on introduit l'approximation de $W^\star$ suivante :

$$\widetilde{W}_N^\star(\omega,\xi) \ := \frac{1}{2}\left(W_N^\star(\omega,\xi) + W_N^{\star,\text{ant}}(\omega,\xi)\right). \tag{1.24}$$

Puisque $W_N^{\star,\text{ant}}(\cdot,\xi)$ a la même loi que $W_N^\star(\cdot,\xi)$, l'espérance de $\widetilde{W}_N^\star(\cdot,\xi)$ est identique à celle de $W_N^\star(\cdot,\xi)$, donc l'erreur systématique définie dans (1.11) n'est pas modifiée par le changement d'estimateur. L'erreur statistique est, elle, modifiée, car la variance de $\widetilde{W}_N^\star$ vient remplacer celle de $W_N^\star$.

La réduction de variance est alors démontrée grâce à des propriétés de monotonie de $W_N^\star$ par rapport à ses variables d'aléa. On peut en effet utiliser un résultat du type suivant pour conclure.

Lemme 1.10 ([Liu08]). *Soit $f : \mathbb{R} \mapsto \mathbb{R}$ une fonction* croissante. *Soit $X : \Omega \mapsto \mathbb{R}$ une variable aléatoire distribuée selon la loi uniforme $\mathcal{U}[0,1]$. Alors*

$$\mathbb{V}\text{ar}\left(\frac{1}{2}\left(f(X) + f(1-X)\right)\right) \leq \frac{1}{2}\mathbb{V}\text{ar}\left(f(X)\right).$$

Réduction de l'erreur statistique

La stratégie de calcul présentée dans la Section 1.2.4 revient ici à résoudre $2M$ problèmes du type (1.22), et de calculer l'approximation aléatoire

$$\mathcal{I}_{MC}^M(\omega) \ := \frac{1}{2M}\sum_{j=1}^{2M} W_{N,j}^\star(\omega),$$

où les $W^\star_{N,j}$ sont des copies indépendantes et identiquement distribuées de $W^\star_N$. La stratégie des variables antithétiques consiste à mettre en oeuvre cette méthode pour $\tilde{W}^\star_N$: on considère l'approximation aléatoire

$$\mathcal{I}^M_{\text{ant}}(\omega) \ := \frac{1}{M}\sum_{j=1}^{M} \tilde{W}^\star_{N,j}(\omega).$$

Puisque chaque calcul de $\tilde{W}^\star_N$ nécessite deux résolutions de (1.22), le coût calcul de ces deux méthodes est identique.

Le résultat principal du **Chapitre** 2 est la Proposition 2.1, où nous démontrons que la variance de l'estimateur antithétique $\mathcal{I}^M_{\text{ant}}$ est plus petite que celle de l'estimateur standard $\mathcal{I}^M_{MC}$. Pour ce faire, nous établissons dans (2.13) que la variance de $\widetilde{W}^\star_N$ est plus petite que la moitié de la variance de $W^\star_N$. L'estimateur $\mathcal{I}^M_{\text{ant}}$ ayant le même coût que $\mathcal{I}^M_{MC}$ et une variance plus petite, il est donc plus performant.

Nous montrons également, dans le cas unidimensionnel et sous des hypothèses plus contraignantes, que nous réduisons la variance des approximations de $\xi \cdot \partial_\xi W^\star_N(\cdot,\xi)$ (respectivement $\xi \cdot \partial^2_\xi W^\star_N(\cdot,\xi)\xi$) en considérant $\xi \cdot \partial_\xi \widetilde{W}^\star_N(\cdot,\xi)$ (respectivement $\xi \cdot \partial^2_\xi \widetilde{W}^\star_N(\cdot,\xi)\xi$). Ces quantités sont identiques dans le cas où W prend la forme $W(x,\omega,\xi) = \frac{1}{2}\xi \cdot A(x,\omega)\xi$ ce qui revient au cas linéaire traité dans [BCLBL12a, BCLBL12b, CLBL10].

Résultats numériques en dimension deux

Nous illustrons ces résultat théoriques par des expériences numériques en dimension $d = 2$. En particulier, nous observons numériquement la réduction de variance sur $\partial_\xi W^\star_N$ et $\partial^2_{\xi\xi} W^\star_N$, bien que nous n'ayons pas prouvé ceci en dimension $d \geq 2$.

1.3.2 Une approche par variables de contrôle basée sur une théorie perturbative pour la réduction de variance en homogénéisation stochastique.

Dans le **Chapitre** 3, nous introduisons une méthode de variable de contrôle pour obtenir une approximation de $A^\star$ plus performante que $\mathcal{I}^M_{MC}$ (défini par (1.12)), car de plus faible variance. Considérons le cas où

$$A \ := A_\eta(x,\omega) = A_{\text{per}}(x) + b_\eta(x,\omega)(C_{\text{per}} - A_{\text{per}})(x) \tag{1.25}$$

est une perturbation à l'ordre η de A_{per}. Nous dérivons une approximation Y^2_η précise à l'ordre 2 en loi de $A^{\star,N}_\eta$. Cependant, nous n'utilisons pas $Y^2_\eta(\omega)$ dans le régime $\eta \ll 1$ pour approcher directement $A^{\star,N}(\omega)$, nous l'utilisons dans le cas où η n'est pas petit, en tant que variable de contrôle. De plus, nous utilisons la méthode de base réduite de [LBT12] pour calculer efficacement $Y^2_\eta(\omega)$.

Méthode des variables de contrôle

Présentons tout d'abord la méthode des variables de contrôle.

Il s'agit d'une méthode standard pour réduire l'erreur sur l'estimation de l'espérance d'une variable aléatoire. Cette méthode est utile lorsqu'est disponible une variable réduite, d'espérance connue, bien corrélée à la variable d'intérêt. Soit $X \in L^2(\Omega)$ la variable aléatoire dont on cherche à estimer l'espérance, et $Y \in L^2(\Omega)$ la variable de contrôle. Soit $\rho \in \mathbb{R}$, on pose :

$$X_\rho(\omega) \ := X(\omega) - \rho(Y(\omega) - \mathbb{E}[Y]). \tag{1.26}$$

La méthode des variables de contrôle consiste à approcher l'espérance de X_ρ par méthode de Monte Carlo :

$$\mathcal{I}_{CV}^{M,\rho}(\omega) := \frac{1}{M}\sum_{k=1}^{M} X_\rho^k(\omega), \tag{1.27}$$

où les X_ρ^k sont des copies indépendantes de X_ρ. Puisque ρ n'est pas aléatoire, il est évident que $\mathbb{E}[X_\rho] = \mathbb{E}[X]$, ainsi $\mathcal{I}_{CV}^{M,\rho}$ est un estimateur convergent quand M tend vers l'infini de $\mathbb{E}[X]$. Observons également que

$$\mathbb{V}\mathrm{ar}X_\rho = \mathbb{V}\mathrm{ar}X - 2\rho\mathbb{C}\mathrm{ov}[X,Y] + \rho^2\mathbb{V}\mathrm{ar}Y,$$

où $\mathbb{C}\mathrm{ov}[a,b] := \mathbb{E}\left[(a-\mathbb{E}a)(b-\mathbb{E}b)\right]$ est la covariance de deux variables aléatoires. Puisque $\mathbb{V}\mathrm{ar}X_\rho$ est une forme quadratique par rapport à ρ, elle admet un minimum $\rho^\star = \frac{\mathbb{C}\mathrm{ov}[X,Y]}{\mathbb{V}\mathrm{ar}Y}$. Pour cette valeur de ρ, nous obtenons ainsi $\mathbb{V}\mathrm{ar}X_{\rho^\star} = \mathbb{V}\mathrm{ar}X\left(1 - \frac{\mathbb{C}\mathrm{ov}[X,Y]^2}{\mathbb{V}\mathrm{ar}X\mathbb{V}\mathrm{ar}Y}\right)$. L'estimateur $\mathcal{I}_{CV}^{M,\rho^\star}$ est d'autant plus performant que sa variance est petite, ce qui est le cas si X et Y sont fortement corrélés.

On estime en pratique $\rho^\star$ par des moyennes empiriques. Nous discutons le choix de Y (qui dépend du problème) ci-après.

Construction d'une expansion en loi de $A_\eta^{\star,N}$

Soit A_η de la forme (1.25). Nous supposons de plus que b_η prend la forme

$$b_\eta(x,\omega) := \sum_{k\in\mathbb{Z}^d} B_k^\eta(\omega)\mathbb{1}_{k+Q}(x), \tag{1.28}$$

où B_k^η sont des variables de Bernoulli indépendantes et identiquement distribuées de paramètre η : $\mathbb{P}[B_k^\eta = 1] = \eta$ et $\mathbb{P}[B_k^\eta = 0] = 1-\eta$. Ainsi, si $\eta \ll 1$, les défauts, représentés par $B_k = 1$, sont rares, mais leurs effets sont importants car ils changent à l'ordre 1 la valeur locale de la matrice A. On définit la matrice $A_1^k := A_{\mathrm{per}} + (C_{\mathrm{per}} - A_{\mathrm{per}})\mathbb{1}_{k+Q}$, correspondant au milieu A_{per} perturbé par un défaut localisé en $k \in \mathbb{Z}^d \cap Q_N$, et le correcteur $w_p^{1,k,N}$ associé à un défaut localisé en $k \in \mathbb{Z}^d \cap Q_N$, solution Q_N-périodique de l'équation

$$-\operatorname{div}\left[A_1^k(p + \nabla w_p^{1,k,N})\right] = 0 \quad \text{sur } Q_N. \tag{1.29}$$

On montre que la quantité $\bar{A}_{1\mathrm{def}}^{k,N} := \int_{Q_N} A_1^k(p+\nabla w_p^{1,k,N}) - \int_{Q_N} A_{\mathrm{per}}(p+\nabla w_{\mathrm{per}})$, qui représente la variation de la réponse de l'environnement de référence perturbée par un défaut localisé en $k+Q$, joue un rôle important. Ainsi, dans le Lemme 3.4, on montre que

$$\forall\varphi \in C(\mathbb{R}), \quad \mathbb{E}\left[\varphi(A_\eta^{\star,N})\right] = \mathbb{E}\left[\varphi\left(A_{\mathrm{per}}^\star + \sum_{k\in\mathbb{Z}^d\cap Q_N} B_k^\eta \bar{A}_{1\mathrm{def}}^{k,N}\right)\right] + O(\eta^2). \tag{1.30}$$

Cela suggère d'utiliser $Y_\eta^1(\omega) := A_{\mathrm{per}}^\star + \sum_{k\in\mathbb{Z}^d\cap Q_N} B_k^\eta(\omega)\bar{A}_{1\mathrm{def}}^{k,N}$ comme variable de contrôle. Pour un certain $\rho \in \mathbb{R}$, la variable aléatoire suivante :

$$X_\rho(\omega) := A_\eta^{\star,N}(\omega) - \rho(Y_\eta^1(\omega) - \mathbb{E}Y_\eta^1), \tag{1.31}$$

d'espérance $\mathbb{E}[A^{\star,N}]$, a une variance significativement plus petite que celle de $A^{\star,N}$.

Nous avons amélioré (1.31) en considérant plusieurs contrôles, $Y_\eta^1(\omega)$, le contrôle lié à l'expansion poussée à l'ordre deux $Y_\eta^2(\omega)$ (considérant la perturbation de A_{per} par un *couple* de défauts). Nous utilisons ainsi la variable aléatoire

$$X_{\rho_1,\rho_2}(\omega) = A^{\star,N}(\omega) - \rho_1(Y_\eta^1(\omega) - \mathbb{E}[Y_\eta^1]) - \rho_2(Y_\eta^2(\omega) - \mathbb{E}[Y_\eta^2]),$$

où ρ_1 et ρ_2 sont choisis afin de minimiser la variance de X_{ρ_1,ρ_2}.

Calcul par bases réduites de la partie déterministe du modèle

Il est crucial pour la méthode que le calcul des Y_η^2 soit très peu coûteux. Nous avons recours à la méthode de base réduite développée dans [LBT12], afin d'avoir un coût de calcul essentiellement indépendant de N.

Ainsi, le calcul de la variable de contrôle n'est pas coûteux en pratique (le surcoût représente 13% de temps de calcul dans nos simulations), tandis que la réduction de variance est très significative.

Analyse de la méthode en dimension 1

Nous montrons que la variance de $A^{\star,N}$ est d'ordre N^{-1}, alors que la variance de $X^1 := A^{\star,N} - \rho_N^1(Y_\eta^1 - \mathbb{E}[Y_\eta^1])$ (où ρ_N^1 est le paramètre optimal) est d'ordre N^{-2} (voir Proposition 3.11), puis nous montrons que la variance de $X^2 := X_{\rho_1^N,\rho_2^N}$ (où ρ_1^N, ρ_2^N sont les paramètres optimaux) est d'ordre N^{-3} (voir Proposition 3.13).

Résultats numériques en dimension 2

Nous testons notre méthode, toujours dans un cas non perturbatif, où le paramètre de la loi de Bernoulli est $\eta = 0.5$ avec $A_{\text{per}} = 3Id$ et $C_{\text{per}} = 23Id$ et nous observons une réduction de variance de l'ordre de 40 dans ce cas, indépendamment de la taille de la supercellule.

1.3.3 Structures quasi-aléatoires spéciales : une approche de sélection pour l'homogénéisation stochastique.

Dans le **Chapitre** 4, nous introduisons une troisième méthode de réduction de variance. Il s'agit d'une approche de sélection. Nous ne calculons la solution de (1.9) que pour certaines réalisations de $A(x,\omega)$ sur Q_N. Nous sélectionnons celles qui vérifient au mieux des propriétés normalement satisfaites seulement asymptotiquement quand N tend vers l'infini. L'approche de sélection que nous avons développée s'inspire d'une méthode venant de la physique du solide. L'approche consiste à sélectionner des réalisations qui vérifient au mieux certaines propriétés. Par exemple, pour un matériau composé de deux phases A et B, on sélectionne les réalisations de sorte que la fraction volumique de chaque phase dans la supercellule soit exactement la proportion de chaque matériau (propriété qui n'est généralement réalisée que dans la limite des grandes supercellules Q_N). Dans un alliage biphasique, on sélectionne ainsi les réalisations selon la fraction volumique, et selon d'autres quantités (proportion d'atomes B voisins des atomes A, proportions de chaîne d'atomes BB voisines de A, etc.).

Dérivation des conditions

Pour *dériver* les conditions utilisées pour la sélection des réalisations, nous considérons ici un modèle faiblement aléatoire. Toutefois, ce régime perturbatif n'est utile que pour

construire les critères de sélection, et la méthode est testée ensuite dans un cas non perturbatif.

Nous étudions un cas où la matrice A prend la forme

$$A := A_\eta(x,\omega) = C_0 + \eta\chi(x,\omega)C_1(x), \tag{1.32}$$

où C_0 est constante, C_1 est périodique, et $-1 \leq \chi \leq 1$ est stationnaire et prend la forme

$$\chi(x,\omega) := \sum_{k\in\mathbb{Z}^d} X_k(\omega)\mathbb{1}_{k+Q}(x). \tag{1.33}$$

Ainsi, la source d'aléa du problème est entièrement contenue dans le champ aléatoire χ, lequel est aléatoire à travers les variables aléatoires scalaires $(X_k)_{k\in\mathbb{Z}^d}$. Dans l'article [BCLBL12b], les auteurs établissent que la matrice homogénéisée $A^\star_\eta$ vérifie le développement limité suivant :

$$A^\star_\eta = C_0 + \eta A^\star_1 + \eta^2 A^\star_2 + o(\eta^2), \tag{1.34}$$

où le coefficient $A^\star_1$ est simplement $\mathbb{E}\int_Q \chi C_1$, et où $A^\star_2$ peut être calculé en résolvant dans $\{u \in L^2_{\text{loc}}(\mathbb{R}^d), \quad \nabla u \in L^2(\mathbb{R}^d)^d\}$ le problème suivant

$$-\operatorname{div} C_0\nabla\phi_1 = \operatorname{div} \mathbb{1}_Q C_1 p \quad \text{dans } \mathbb{R}^d. \tag{1.35}$$

De même, on peut montrer que la matrice homogénéisée apparente $A^{\star,N}_\eta(\omega)$ vérifie un développement limité similaire

$$A^{\star,N}_\eta(\omega) = C_0 + \eta A^{\star,N}_1(\omega) + \eta^2 A^{\star,N}_2(\omega) + o(\eta^2), \tag{1.36}$$

où le coefficient $A^{\star,N}_1(\omega)$ vaut $\frac{1}{|Q_N|}\int_{Q_N}\chi(x,\omega)C_1(x)dx$, et où $A^{\star,N}_2(\omega)$ peut être calculé en résolvant la version tronquée de (1.35) :

$$-\operatorname{div} C_0\nabla\phi^N_1 = \operatorname{div} \mathbb{1}_Q C_1 p \quad \text{dans } Q_N, \text{ et } \phi^N_1 \text{ est } Q_N\text{-périodique}. \tag{1.37}$$

On dit qu'un environnement $\omega \in \Omega$ satisfait la condition d'ordre k ($k = 1, 2$ que l'on peut généraliser) si l'égalité $A^{\star,N}_k(\omega) = A^\star_k$ est vérifiée. Si un certain ω satisfait les conditions jusqu'à l'ordre k, nous avons alors

$$A^\star_\eta(\omega) = A^{\star,N}_\eta(\omega) + o(\eta^k),$$

où le reste est uniforme par rapport à N et ω. Jusqu'à l'ordre 2, dans cet exemple, les conditions que nous imposons sont faciles à évaluer.

En toute généralité, la condition à l'ordre 1 s'écrit

$$\frac{1}{|Q_N|}\int_{Q_N}\chi(x,\omega)C_1(x,\omega)p\,dx = \mathbb{E}\int_Q \chi C_1 p. \tag{1.38}$$

Lorsque de plus C_1 est constante, et si χ prend la forme (1.33) alors cette condition se réécrit

$$\frac{1}{N^d}\sum_{|k|_\infty\leq N} X_k(\omega) = \mathbb{E}[X_0]. \tag{1.39}$$

En utilisant seulement cette sélection au premier ordre, nous obtenons l'algorithme suivant :

Algorithme 1 (Sélection-Monte Carlo).
Cet algorithme requiert une tolérance $\mathrm{tol} \geq 0$.
Pour $m = 1, \ldots, M$,

1. *Générer un environnement aléatoire* ω_m.
2. *Si* $\left| \frac{1}{N^d} \sum_{|k|_\infty \leq N} X_k - \mathbb{E}[X_0] \right| > \mathrm{tol}$, *retourner à l'étape 1.*
3. *Résoudre le problème du correcteur tronqué (1.9).*
4. *Calculer* $A^{\star,N}(\omega_m)$ *par (1.10).*

Calculer l'approximation de $\mathbb{E}[A^\star]$ *suivante :* $\mathcal{I}_{SQS}^M := \frac{1}{M} \sum_{m=1}^{M} A_N^\star(\omega_m)$.

Cet algorithme diffère de l'algorithme usuel par l'étape 2. Le problème du correcteur (1.9) n'est résolu que pour certaines réalisations.

En général, le coût des méthodes de sélection est dominé par l'étape de sélection elle-même, car un grand nombre de tirages est nécessaire pour produire *une* réalisation convenable, et le coût d'un tirage est élevé. Cependant, nous ne sommes pas ici dans un tel cas : le coût de résolution de (1.9) domine complètement le coût des autres étapes. Ainsi, il est pertinent de bien choisir les $\omega \in \Omega$ pour lesquels on résout (1.9), même si cette sélection s'effectue en rejetant beaucoup de réalisations.

Contrôle de l'erreur

Seules certaines réalisations sélectionnées participent à l'estimation de $A^\star$, ce qui se traduit mathématiquement de la façon suivante : au lieu d'échantillonner $A^{\star,N}(\omega)$ selon la mesure de probabilité de référence $\mathbb{P}$, on l'échantillonne suivant une mesure *conditionnelle*. L'algorithme ainsi proposé correspond exactement à une méthode de Monte-Carlo standard pour l'estimation de $\mathbb{E}^{\mu_{\mathrm{cond}}}[A^{\star,N}]$, où μ_{cond} est la mesure de référence conditionnée à satisfaire les critères (par exemple (1.39)). Afin d'analyser l'erreur à l'instar de (1.11), nous introduisons $\mathbb{E}^{\mu_{\mathrm{cond}}}[A^{\star,N}]$ au lieu de $\mathbb{E}[A^{\star,N}]$ et nous obtenons la décomposition

$$A^\star - A^{\star,N}(\omega) \quad = \quad A^\star - \mathbb{E}^{\mu_{\mathrm{cond}}}[A^{\star,N}] \quad + \quad \mathbb{E}^{\mu_{\mathrm{cond}}}[A^{\star,N}] - A^{\star,N}(\omega). \tag{1.40}$$

La première partie de l'erreur est une erreur de biais. Comme mentionné dans la Section 1.2.5, on peut s'attendre à ce qu'elle ne soit pas dominante, mais l'analyse complète du taux de convergence de cette erreur est une question non résolue. Sous des hypothèses assez générales, nous montrons dans le Théorème 4.8 qu'elle tend vers 0. Cela assure la convergence de l'approche. Nous vérifions numériquement que l'erreur de biais n'est pas plus grande que dans le cas Monte Carlo usuel, voire même est diminuée en utilisant notre méthode.

La seconde erreur est bien sûr de nature statistique. Cependant, elle est contrôlée par la variance de $A^{\star,N}$ sous la mesure conditionnelle, dont on s'attend à ce qu'elle soit plus petite. Nous montrons, dans certains cas simplifiés (en particulier en dimension un d'espace) que c'est le cas : la variance est réduite.

Expérimentations numériques en dimension deux

En pratique, les conditions que doivent satisfaire les réalisations peuvent ne pas être *exactement* satisfaites, et nous autorisons une certaine déviation comme dans l'Algorithme 1 en autorisant un niveau d'erreur pour la satisfaction du critère, ou alors en faisant un grand nombre de tirages, et en sélectionnant les meilleurs des échantillons. Nous vérifions numériquement, dans un cas non perturbatif et en dimension $d = 2$, la robustesse de notre méthode : même en présence d'une certaine erreur dans la condition que nous imposons, notre approche permet de significativement réduire l'erreur (1.40).

Toujours dans un cas non perturbatif et en dimension $d = 2$, nous observons numériquement une bonne réduction de variance ainsi qu'une réduction de l'erreur totale.

1.3.4 Un problème d'identification de paramètres en homogénéisation stochastique.

Dans le **Chapitre** 5, nous abordons un problème inverse dans un cadre d'homogénéisation aléatoire. Le cadre d'étude est celui d'équations aux dérivées partielles *discrètes*. Toutefois il serait possible d'effectuer la même démarche dans le cadre d'équations aux dérivées partielles continues considéré jusqu'à présent.

Modélisation d'un réseau de pores

Le problème de physique motivant notre étude est un problème de transport d'ions dans un réseau de pores au sein d'argiles. La physique complexe décrivant ce milieu rend la modélisation complète de la géométrie du milieu difficile. Une modélisation usuellement adoptée est celle, simpliste, d'un réseau structuré (le réseau $\mathbb{Z}^d$) dont les noeuds sont les pores et les arêtes les canaux. Le problème d'intérêt est le suivant : calibrer les paramètres de la loi de la taille des canaux (à l'échelle microscopique), à l'aide de simples expériences hydrauliques sur une espèce non chargée (qui correspond à un calcul de perméabilité effective donc d'homogénéisation), pour diminuer le nombre de paramètres du modèle de transport ionique.

Équations aux dérivées partielles discrètes

Nous étudions l'équation

$$\nabla_\varepsilon^\star\big[A(x/\varepsilon,\omega)\nabla_\varepsilon u_\varepsilon(x,\omega)\big] = f(x) \quad \text{dans } \mathcal{D}\cap\varepsilon\mathbb{Z}^d, \qquad u_\varepsilon(x,\omega) = 0 \text{ dans } (\mathbb{R}^d\setminus\mathcal{D})\cap\varepsilon\mathbb{Z}^d, \tag{1.41}$$

contrepartie discrète de (1.2). L'opérateur de différences finies $\nabla_\varepsilon \ : \ell^2(\varepsilon\mathbb{Z}^d) \mapsto \ell^2(\varepsilon\mathbb{Z}^d)^d$ est défini par

$$\nabla_\varepsilon g : x \mapsto \frac{1}{\varepsilon}\begin{pmatrix} g(x+\varepsilon e_1) - g(x) \\ \vdots \\ g(x+\varepsilon e_d) - g(x)\end{pmatrix}.$$

L'opérateur $\nabla_\varepsilon^\star \ : \ell^2(\varepsilon\mathbb{Z}^d)^d \mapsto \ell^2(\varepsilon\mathbb{Z}^d)$ (dont l'équivalent au niveau continu est $-\operatorname{div}$) est défini comme l'adjoint dans $\ell^2(\varepsilon\mathbb{Z}^d)$ de ∇_ε. C'est pourquoi il correspond à une différence finie dont le décentrage est inverse :

$$-\nabla_\varepsilon^\star G : x \mapsto \sum_{i=1}^d \frac{G_i(x) - G_i(x-\varepsilon e_i)}{\varepsilon}.$$

Estimation de paramètres

On suppose que la loi de probabilité de A est paramétrée par un certain $\theta \in \Theta$. Ce paramètre, de faible dimension (de dimension 2 dans notre travail) est supposé inconnu. Ainsi, la loi de probabilité de A est supposée mal connue, et nous cherchons à estimer θ, au moyen de quantités qui sont facilement accessibles à l'expérience : des quantités macroscopiques. Nous utilisons pour ce faire deux quantités (et leur nombre est réminiscent de la dimension de l'espace des paramètres à identifier : il est nécessaire d'avoir au moins deux quantités connues pour en estimer deux), qui sont le premier coefficient de la matrice homogénéisée $[A^\star]_{1,1}$ et la variance relative de la matrice homogénéisée apparente

$$\mathbb{V}\mathrm{arR}\left[\left(A^{\star,N}\right)_{1,1}\right] := \frac{\mathbb{V}\mathrm{ar}\left[\left(A^{\star,N}\right)_{1,1}\right]}{\left(\mathbb{E}\left[\left(A^{\star,N}\right)_{1,1}\right]\right)^2}. \tag{1.42}$$

Comme le problème d'intérêt est isotrope, nous ne nous intéressons qu'à un seul coefficient diagonal de $A^\star$.

Nous supposons que ces deux quantités sont connues, par exemple grâce à une expérience physique, et nous cherchons alors à retrouver le paramètre $\theta \in \Theta$ associé. Nous supposons en outre une forme particulière de dépendance de A par rapport à θ : nous supposons, pour des raisons physiques, que $A = aId$, où a suit une loi de Weibull à deux paramètres $\theta := (\lambda, k) \in (\mathbb{R}_+^\star)^2$. Rappelons que ces variables aléatoires sont positives, et leur densité s'écrit

$$\forall r > 0, \qquad f(r; k, \lambda) = \frac{k}{\lambda}\left(\frac{r}{\lambda}\right)^{k-1} \exp\left(-(r/\lambda)^k\right),$$

ce qui correspond à la distribution cumulée

$$F(r; k, \lambda) = \int_0^r f(s; k, \lambda)\, ds = 1 - \exp\left(-(r/\lambda)^k\right).$$

Nous formulons ce problème *inverse* sous la forme d'un problème d'optimisation aux moindres carrés. On introduit

$$F_{N,M}(\theta, \omega) \ := \left(\frac{\overline{K}_M^{\star,N}(\theta,\omega)}{K_{\mathrm{obs}}^{\star,N}} - 1\right)^2 + \left(\frac{S_M^N(\theta,\omega)}{S_{\mathrm{obs}}^N} - 1\right)^2, \tag{1.43}$$

où $\overline{K}_M^{\star,N}(\theta,\omega) \ := \dfrac{1}{M}\displaystyle\sum_{k=1}^M [A_k^{\star,N}(\theta,\omega)]_{1,1}$ est une approximation de $e_1 \cdot \mathbb{E}[A^{\star,N}(\theta,\cdot)]e_1$, et $S_M^N(\theta,\omega)$ est une approximation de la variance relative (1.42). Les quantités $K_{\mathrm{obs}}^{\star,N}$ et S_{obs}^N sont les données de notre problème, et on suppose qu'il existe un paramètre déterministe θ_{obs} tel que $K_{\mathrm{obs}}^{\star,N} = e_1 \cdot \mathbb{E}[A^{\star,N}(\theta_{\mathrm{obs}},\cdot)]e_1$ et de même pour S_{obs}^N. Nous cherchons donc à minimiser $F_{N,M}$ par rapport à θ.

Lorsque N tend vers l'infini, la fonctionnelle $F_{N,M}(\theta,\omega)$ admet une limite $F_\infty(\theta)$. Nous démontrons, en une dimension d'espace, que le problème d'optimisation $\inf_{\theta\in\Theta} F_\infty(\theta)$ est bien posé, et admet pour unique solution θ_{obs}.

Résolution numérique

Pour résoudre ce problème inverse, nous devons *évaluer* de façon répétée $F_{N,M}(\theta,\omega)$, ainsi que ses dérivées, pour différentes valeurs de $\theta \in \Theta$. C'est là la partie principale du temps

de calcul, et c'est pourquoi nous avons recours à un algorithme d'optimisation efficace pour atteindre une faible erreur en peu d'itérations, minimisant ainsi le nombre d'appels de la fonctionnelle $F_{N,M}$.

Afin de résoudre le problème d'optimisation (1.43), nous avons recours à un algorithme de Newton. Nous pouvons en effet calculer les dérivées de $A^{\star,N}(\theta,\omega)$ (et donc de $F_{N,M}$) par rapport à θ jusqu'à l'ordre 2 sans recourir à des différences finies et ceci à coût comparable à la simple *évaluation* de $F_{N,M}$ (voir paragraphe suivant).

Nous obtenons ainsi en quelques itérations une bonne approximation d'un minimiseur local du problème (1.43), qu'on note $\theta^\star(\omega)$. Puisque $F_{N,M}$ est une variable aléatoire, $\theta^\star$ est aussi une variable aléatoire. Nous vérifions numériquement que la variance de $\theta^\star$ est comparable à la variance de $\overline{K}_M^{\star,N}$ et S_M^N. Ainsi l'erreur statistique supplémentaire introduite par la résolution du problème inverse n'est pas trop importante. Nous vérifions en outre que $\mathbb{E}[\theta^\star]$ est une bonne approximation de θ_{obs}.

Les expériences numériques menées dans le **Chapitre** 5 ont été conduites en une dimension d'espace, puis généralisées en dimension deux.

Calcul des dérivées de $A^{\star,N}(\theta,\omega)$ par rapport au paramètre $\theta = (k,\lambda)$

Pour simplifier, nous supposons que A est symétrique. De même, nous utilisons les notations des équations aux dérivées partielles *continues* mais les calculs s'étendent au cadre discret.

Au vu de (1.10) et de la formulation variationnelle du problème (1.9), nous pouvons symétriser davantage l'équation définissant $A^{\star,N}(\omega)$:

$$\forall p,q \in \mathbb{R}^d, \quad q \cdot A^{\star,N} p = \frac{1}{|Q_N|} \int_{Q_N} (q + \nabla w_q^N) \cdot A(p + \nabla w_p^N).$$

En utilisant les équations d'Euler-Lagrange, il vient, pour la dérivée par rapport à k,

$$\forall p,q \in \mathbb{R}^d, \quad q \cdot \partial_k A^{\star,N} p = \frac{1}{|Q_N|} \int_{Q_N} (q + \nabla w_q^N) \cdot \partial_k A(p + \nabla w_p^N)$$

et de même pour la dérivée par rapport à λ.

On considère maintenant les dérivées secondes de $A^{\star,N}$. Soient τ_1 et τ_2 deux directions de dérivation ($\tau_1 = k$ ou λ et de même pour τ_2). En dérivant l'équation ci-dessus par rapport à τ_2, il vient :

$$\begin{aligned} q \cdot \partial^2_{\tau_1\tau_2} A^{\star,N} p &= \frac{2}{|Q_N|} \int_{Q_N} \nabla \partial_{\tau_2} w_q^N \cdot \partial_{\tau_1} A(p + \nabla w_p^N) \\ &+ \frac{1}{|Q_N|} \int_{Q_N} (q + \nabla w_q^N) \cdot \partial^2_{\tau_1\tau_2} A(p + \nabla w_p^N). \end{aligned}$$

La dérivée par rapport au paramètre τ (i.e. k ou λ) du correcteur w_p^N est aisée à calculer : en dérivant la formulation variationnelle satisfaite par w_p^N, on obtient que, pour toute fonction test $\varphi \in H^1_{\text{per}}(Q_N)$:

$$\int_{Q_N} \nabla\varphi \cdot \partial_\tau A(p + \nabla w_p^N) + \int_{Q_N} \nabla\varphi \cdot A \nabla \partial_\tau w_p^N = 0, \tag{1.44}$$

ce qui est un problème bien posé pour $\partial_\tau w_p^N$.

Pour conclure, le calcul des dérivées premières et secondes de $A^{\star,N}$ par rapport au paramètre θ ne nécessite que la résolution d'un nombre d'équations (1.44) égal à $\dim(\Theta)$.

1.4 Perspectives

L'homogénéisation aléatoire est une théorie couvrant des cas très différents et parfois très difficile. Prise dans toute sa généralité, elle est peu accessible à l'expérience numérique. En effet, la méthode standard d'estimation de $A^\star$ est inefficace car on explore l'intégralité de l'espace d'aléa, dont la diversité reste importante lorsque la taille de la supercellule Q_N est finie.

Pour réduire la complexité du problème, une approche consiste à développer $A^\star$ par rapport à un petit paramètre $\eta \ll 1$ (contraste, proportion de défauts, ...) présent dans la définition du champ A. On aboutit alors à des approches très peu coûteuses, mais dont le domaine de validité est *a priori* limité au régime perturbatif.

Une approche possible, dans laquelle les **Chapitre** 3 et **Chapitre** 4 s'inscrivent, est d'utiliser ces approches perturbatives comme un *préconditionnement*. Ceci permet d'aboutir à des méthodes efficaces dans un régime de paramètres bien plus grand que le régime perturbatif $\eta \ll 1$.

1.4.1 Comparaison des méthodes de réduction de variance

Dans cette thèse, nous avons exploré trois méthodes de réduction de variance. Nous présentons à présent une description synthétique de ces résultats.

La méthode des variables antithétique explorée dans le **Chapitre** 2 présente l'avantage d'être peu invasive et n'est pas coûteuse. Nous avons observé en contrepartie une réduction de variance limité, de l'ordre de 10 (20 dans les meilleurs cas, et 4 dans les pires). Nous renvoyons à la Section 2.3 pour des résultats détaillés.

La méthode des variables de contrôle introduite dans le **Chapitre** 3 nécessite un bon modèle approché, dont la construction peut être complexe. Cependant, une fois cette étape franchie, la réduction de variance peut être bien meilleure : de l'ordre de 40 dans les cas considérés (voir la Section 3.5 pour des résultats détaillés).

La méthode de sélection introduite dans le **Chapitre** 4 nécessite non seulement un bon critère de sélection, mais aussi un algorithme de sélection (qui toutefois peut-être choisi simple). Cette approche a donné les résultats les plus intéressants : une réduction de variance de l'ordre de 200 dans les cas considérés (voir la Section 5.4.2 pour des résultats détaillés).

1.4.2 Extension à la stationnarité continue des variables de contrôle

Une extension possible du **Chapitre** 3 serait de généraliser la modélisation pour couvrir des cas de stationnarité *continue*. On remplace dans ce cas l'équation (1.3) de la Définition 1.1 par (1.5). Sous cette définition, la loi d'une variable aléatoire est invariante par toute translation et pas seulement celles d'un multiple entier de la cellule Q. Cette hypothèse modélise en particulier des milieux peu structurés, où l'emplacement d'inclusions ou de défauts, par exemple circulaires et de rayons aléatoires, est choisi au hasard sans tenir compte d'un réseau sous-jacent.

Dans le contexte du **Chapitre** 3, la méthode d'ordre 1, assez générique, peut s'étendre sans complications méthodologiques : la variable de contrôle d'ordre 1 reste la proportion volumique d'inclusions. Pour mettre en oeuvre la méthode d'ordre 2, qui seule prend en compte la géométrie, une difficulté nouvelle survient. La variable de contrôle à l'ordre 2 fait intervenir des configurations à deux défauts, lesquelles sont maintenant présentes en

un nombre *infini* (au lieu de $N^d - 1$), puisque les défauts ne sont plus positionnés sur le réseau $\mathbb{Z}^d \cap Q_N$.

Afin de généraliser rigoureusement la méthode, il faut adapter à ce nouveau cadre le développement faiblement aléatoire introduit par [ALB12]. Il faut aussi adapter à ce nouveau cadre la méthode des bases réduites proposée dans [LBT12]. Enfin, si tous ces éléments sont réunis, il faudra adapter le développement en loi introduit dans le **Chapitre** 3, lequel pourrait alors être utilisé comme variable de contrôle.

1.4.3 Utilisation d'autres variables de contrôle

Une autre extension possible serait d'utiliser d'autres variables de contrôle. Pour estimer la matrice homogénéisée $A^\star$, on fait parfois appel à des méthodes de bornes (bornes de Hashin et Shtrikman). On commence par montrer que le calcul de $A^{\star,N}$ est équivalent à la résolution d'un problème variationnel, dit problème de Hashin–Shtrikman :

$$\inf_{\tau \in L^2(Q_N)} \mathcal{E}_N(\tau, \omega).$$

Les bornes de Hashin–Shtrikman sont obtenues en choisissant des fonctions τ particulières.

Une possibilité serait d'utiliser ces bornes (qui sont des quantités aléatoires) sur $A^{\star,N}$ en tant que variables de contrôle.

1.4.4 Sélection pour des problèmes nonlinéaires

Il serait intéressant d'étendre la méthode présentée dans le **Chapitre** 4 à des cas nonlinéaires. Par exemple, on pourrait considérer le cas (1.13) issu du **Chapitre** 2. Les équations d'Euler-Lagrange associées s'écrivent :

$$-\operatorname{div}[\partial_\xi W(x/\varepsilon, \omega, \nabla u_\varepsilon)] = f.$$

L'équation linéarisée s'écrit sous la forme

$$-\operatorname{div}[A(x/\varepsilon, \omega)\nabla u_\varepsilon] = f.$$

Supposons par simplicité que nous connaissons complètement le coefficient homogénéisé $A^\star$ associé au problème linéaire, réputé plus facile à résoudre. Alors, il serait pertinent de sélectionner suivant le critère $A^{\star,N}(\omega) = A^\star$ les réalisations pour lesquelles on va effectivement calculer $W_N^\star$ défini par (1.22).

L'efficacité de l'approche sera sans doute liée au rapport du coût entre résoudre le problème du correcteur dans le cas linéaire ou bien dans le cas nonlinéaire. Notons de plus que, dans le cas nonlinéaire, on peut être intéressé par *tout* le champ $\xi \mapsto W^\star(\xi)$, et pas seulement l'évaluation de $W^\star(\xi)$ pour un vecteur ξ particulier. Ainsi, le calcul de $A^\star$ serait vu comme un unique calcul préalable, avant le calcul de $W^\star(\xi)$ pour de nombreux ξ différents.

1.4.5 Convergence du problème inverse en dimension supérieure

Dans la Section précédente, nous avons présenté l'algorithme proposé dans le **Chapitre** 5 en dimension d'espace quelconque. Toutefois sa convergence n'est traitée que dans le cas unidimensionnel. Il serait donc intéressant d'introduire une méthode pour prouver la convergence de notre approximation obtenue par optimisation aux moindres carrés en plusieurs dimensions d'espace.

Dans une autre direction, il serait aussi intéressant de prouver un résultat d'homogénéisation sous des hypothèses sur A plus faibles que l'uniforme coercivité et l'existence d'une borne uniforme en x et ω. En effet, cette hypothèse n'est pas satisfaite dans le cas que nous étudions (une variable aléatoire distribuée suivant la loi de Weibull peut prendre des valeurs aussi grandes et aussi proches de 0 que voulu). De plus, nous montrons en dimension un d'espace qu'elle peut être relaxée (cf. le Theorem 5.13). La généralisation de ce résultat en dimension quelconque reste ouverte.

1.4.6 Perspectives générales dans le champ de recherche

Nous concluons cette introduction par un résumé plus général. Comment s'inscrivent les travaux de la thèse dans le champ de recherche ?

Tout d'abord, ces différentes approches montrent qu'il est possible de mettre en oeuvre des méthodes de réduction de variance en homogénéisation stochastique. Ces méthodes sont souvent efficaces, et ce d'autant plus que la méthode est sophistiquée.

Seuls des cas simples ont été abordés dans cette thèse : nous avons abordé des équations *linéaires elliptiques*, un problème variationnel *convexe*, une équation aux dérivées partielles *discrète*. Les méthodes introduites doivent être considérés comme des preuves de concept, car il est possible d'étendre les techniques à des cas plus difficiles. En outre, il est sans doute possible d'améliorer qualitativement ces techniques.

En conclusion, on pourrait imaginer à terme l'homogénéisation stochastique *par la pratique*, rendue possible par des approches auxiliaires dont pourraient faire partie les méthodes présentées. Cela permettrait d'améliorer des calculs actuellement réalisés par homogénéisation *périodique*, et ainsi d'avoir un modèle plus riche et prédictif.

Chapter 2

Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem

Ce **Chapitre** reprend l'intégralité d'un article écrit en collaboration avec Frédéric Legoll et accepté dans Discrete and Continuous Dynamical Systems - Series S [LM15b].

Dans un cadre d'homogénéisation stochastique, nous utilisons la méthode des variables antithétiques pour obtenir un estimateur de moindre variance, donc plus précis, de quantités d'intérêt. La spécificité de notre travail par rapport aux travaux antérieurs est que le modèle est *nonlinéaire*. Nous démontrons dans certains cas que la technique des variables antithétique peut encore être utilisée, et nos expérimentations numériques en dimension 2 démontrent son efficacité.

Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem

Frédéric Legoll[1,3] and William Minvielle[2,3]

legoll@lami.enpc.fr, william.minvielle@cermics.enpc.fr

[1] Laboratoire Navier, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal,
77455 Marne-La-Vallée Cedex 2, France,

[2] CERMICS, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2, France;

[3] INRIA Rocquencourt, MATHERIALS research-team, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France.

Abstract. *We consider a nonlinear convex stochastic homogenization problem, in a stationary setting. In practice, the deterministic homogenized energy density can only be approximated by a random apparent energy density, obtained by solving the corrector problem on a truncated domain.*

We show that the technique of antithetic variables can be used to reduce the variance of the computed quantities, and thereby decrease the computational cost at equal accuracy. This leads to an efficient approach for approximating expectations of the apparent homogenized energy density and of related quantities.

The efficiency of the approach is numerically illustrated on several test cases. Some elements of analysis are also provided.

2.1 Introduction

In this article, we consider some theoretical and numerical questions related to variance reduction techniques for some nonlinear convex stochastic homogenization problems. In short, we show here that a technique based on antithetic variables can be used in that context, provide some elements of analysis, and demonstrate numerically the efficiency of that approach on several test cases. This work is a follow-up of the articles [BCLBL12b, BCLBL12b, CLBL10] where the same questions are considered for a *linear* elliptic equation in divergence form.

The stochastic homogenization problem we consider here writes as follows. Let $\mathcal{D}$ be an open bounded domain of $\mathbb{R}^d$ and $2 \leq p < \infty$. We consider the highly oscillatory problem

$$\inf \left\{ \int_{\mathcal{D}} W\left(\frac{x}{\varepsilon}, \omega, \nabla u(x)\right) dx - \int_{\mathcal{D}} f(x)u(x)dx, \quad u \in W_0^{1,p}(\mathcal{D}) \right\} \tag{2.1}$$

for some f and some random smooth field W, which is stationary in a sense made precise below, and satisfies some convexity and growth conditions such that, for any $\varepsilon > 0$,

problem (2.1) is well-posed. See Section 2.1.1 below for a precise description of the mathematical setting, which has been introduced in [DMM86a, DMM86b]. A classical example that motivated this framework is when

$$W(y,\omega,\xi) = \frac{1}{p} a(y,\omega) |\xi|^p ,$$

where a is stationary (see e.g. [DMM86a, page 382]).

In (2.1), ε denotes a supposedly small, positive constant that models the smallest possible scale present in the problem. For ε small, it is extremely expensive, in practice, to directly attack (2.1) with a numerical discretization. A useful practical approach is to *first* approximate (2.1) by its associated homogenized problem, which reads

$$\inf \left\{ \int_{\mathcal{D}} W^\star (\nabla u(x))\, dx - \int_{\mathcal{D}} f(x)u(x)dx, \quad u \in W_0^{1,p}(\mathcal{D}) \right\}, \tag{2.2}$$

and *next* numerically solve the latter problem. The two-fold advantage of (2.2) as compared to (2.1) is that *it is deterministic* and *it does not involve the small scale* ε.

This simplification comes at a price. The homogenized energy density $W^\star$ in (2.2) is given by an integral involving a so-called corrector function, solution to a nonlinear problem (see (2.8) below for a precise formula). As most often in stochastic homogenization, this corrector problem is set on the *entire* space $\mathbb{R}^d$. In practice, approximations are therefore in order. A standard approach (see e.g. [BP04] in the linear setting) is to generate realizations of the energy density W over a finite, supposedly large volume at the microscale, that we denote Q_N, and approach the homogenized energy density by some empirical means using approximate correctors computed on Q_N. Although the *exact* homogenized density $W^\star$ is deterministic, its practical approximation is random, due to the truncation procedure. It is then natural to generate several realizations. However, efficiently averaging over these realizations requires to understand how variance affects the result. This is the purpose of the present article to investigate some questions in this direction, both from the theoretical and numerical standpoints.

Before proceeding and for the sake of consistency, we now present the framework of nonlinear stochastic homogenization we adopt, and make precise the questions we consider.

2.1.1 Homogenization theoretical setting

To begin with, we introduce the basic setting of stochastic homogenization we employ. We refer to [ES08] for a general, numerically oriented presentation, and to [BLP78, CD99, JKO94] for classical textbooks. We also refer to [LB10] and the review article [ACLB+12] (and the extensive bibliography contained therein) for a presentation of our particular setting. Throughout this article, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ the expectation value of any random variable $X \in L^1(\Omega, d\mathbb{P})$. We next fix $d \in \mathbb{N}^\star$ (the ambient physical dimension), and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, that is, for all $k \in \mathbb{Z}^d$ and all $A \in \mathcal{F}$, $\mathbb{P}(\tau_k A) = \mathbb{P}(A)$. We assume that the action τ is *ergodic*, that is, if $A \in \mathcal{F}$ is such that $\tau_k A = A$ for any $k \in \mathbb{Z}^d$, then $\mathbb{P}(A) = 0$ or 1. In addition, we define the following notion of stationarity (see [ACLB+12, Section 2.2]): a function $F \in L^1_{\text{loc}}\left(\mathbb{R}^d, L^1(\Omega)\right)$ is said to be *stationary* if, for all $k \in \mathbb{Z}^d$,

$$F(y+k,\omega) = F(y,\tau_k\omega) \text{ almost everywhere and almost surely.} \tag{2.3}$$

In this setting, the ergodic theorem [Kre85, Shi84, Tem72] can be stated as follows: *Let* $F \in L^\infty\left(\mathbb{R}^d, L^1(\Omega)\right)$ *be a stationary random variable in the above sense. For* $k = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d$*, we set* $|k|_\infty = \sup_{1\le i\le d} |k_i|$*. Then*

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

This implies (denoting by Q *the unit cube in* $\mathbb{R}^d$*) that*

$$F\left(\frac{x}{\varepsilon}, \omega\right) \underset{\varepsilon\to 0}{\overset{*}{\rightharpoonup}} \mathbb{E}\left(\int_Q F(y,\cdot)dy\right) \quad \text{in } L^\infty(\mathbb{R}^d), \text{ almost surely.}$$

The purpose of the above setting is simply to formalize that, even though realizations may vary, the function F at point $y \in \mathbb{R}^d$ and the function F at point $y + k$, $k \in \mathbb{Z}^d$, share the same law. In the homogenization context we now turn to, this means that the local, microscopic environment (encoded in the energy density W) is everywhere the same *on average*. From this, homogenized, macroscopic properties will follow.

We now describe more precisely the multiscale random problem (2.1). The domain $\mathcal{D}$ is a regular (in the sense its boundaries are Lipschitz-continuous) bounded domain of $\mathbb{R}^d$. The right-hand side function f belongs to $L^{p'}(\mathcal{D})$, with $1/p + 1/p' = 1$ (hence f is indeed in the dual space of $L^p(\mathcal{D})$). For any $\xi \in \mathbb{R}^d$, the random field $y, \omega \mapsto W(y, \omega, \xi)$ is assumed stationary in the sense (2.3). We assume that it is continuous (and even C^3) with respect to the ξ variable, and that it is measurable with respect to the y argument. We also assume that there exists $c_2 \ge c_1 > 0$ such that

$$\forall y \in \mathbb{R}^d, \quad \forall \omega \in \Omega, \quad \forall \xi \in \mathbb{R}^d, \quad c_1|\xi|^p \le W(y, \omega, \xi) \le c_2(1 + |\xi|^p). \tag{2.4}$$

Furthermore, we assume henceforth that W is *strictly convex* with respect to the argument ξ, in the sense that

$$\forall \eta \in \mathbb{R}^d, \quad \forall \xi \in \mathbb{R}^d, \quad \eta^T \partial_\xi^2 W(y, \omega, \xi)\eta > 0 \ \text{ a.e. and a.s.,} \tag{2.5}$$

where $\partial_\xi^2 W \in \mathbb{R}^{d\times d}$ is the Hessian matrix of $\xi \mapsto W(y, \omega, \xi)$. A more demanding assumption is that W is α-*convex* with respect to the argument ξ, in the sense that there exists $\alpha > 0$ such that

$$\forall \eta \in \mathbb{R}^d, \quad \forall \xi \in \mathbb{R}^d, \quad \eta^T \partial_\xi^2 W(y, \omega, \xi)\eta \ge \alpha\, \eta^T \eta \ \text{ a.e. and a.s.} \tag{2.6}$$

Unless otherwise stated, we only assume (2.5) in the sequel. When needed, we will explicitly assume (2.6).

Under (2.4) and (2.5), the variational problem (2.1) is well-posed. In addition, the homogenized limit of (2.1) has been identified in [DMM86a, DMM86b] (see also [GN11, Theorem 3.1]): the unique solution $u^\varepsilon(\cdot, \omega)$ to (2.1) converges (weakly in $W^{1,p}(\mathcal{D})$ and strongly in $L^p(\mathcal{D})$, almost surely) to some deterministic function $u^\star \in W^{1,p}(\mathcal{D})$, solution to (2.2), where the homogenized energy density $W^\star$ is given, for any $\xi \in \mathbb{R}^d$, by

$$W^\star(\xi) = \lim_{N\to\infty} \inf\left\{\frac{1}{|Q_N|}\int_{Q_N} W\left(y, \omega, \xi + \nabla w(y)\right) dy, \quad w \in W_0^{1,p}(Q_N)\right\} \tag{2.7}$$

where $Q_N = (-N, N)^d$. The convergence in (2.7) holds almost surely.

We show in Appendix 2.4 below that

$$W^\star(\xi) = \lim_{N\to\infty} \inf \left\{ \frac{1}{|Q_N|} \int_{Q_N} W\left(y, \omega, \xi + \nabla w(y)\right) dy, \quad w \in W^{1,p}_{\#}(Q_N) \right\} \tag{2.8}$$

where $W^{1,p}_{\#}(Q_N)$ denotes the set of functions that belong to $W^{1,p}_{\mathrm{loc}}(\mathbb{R}^d)$ and are Q_N-periodic (the only difference between (2.7) and (2.8) is thus the boundary conditions that we consider). The convergence in (2.8) again holds almost surely.

In the sequel, we work on the basis of (2.8), namely using periodic boundary conditions. We could as well, up to slight modifications, work with homogeneous Dirichlet boundary conditions, on the basis of (2.7). We choose periodic boundary conditions as they have been shown, in practice, to provide more accurate numerical results (see e.g. [KFG+03]).

2.1.2 The questions we consider

In practice, we cannot compute $W^\star(\xi)$, and have to restrict ourselves to finite size domains. We therefore introduce

$$W^\star_N(\omega, \xi) := \inf \left\{ \frac{1}{|Q_N|} \int_{Q_N} W\left(y, \omega, \xi + \nabla w(y)\right) dy, \quad w \in W^{1,p}_{\#}(Q_N) \right\} \tag{2.9}$$

and readily see from (2.8) that

$$W^\star(\xi) = \lim_{N\to\infty} W^\star_N(\omega, \xi) \text{ a.s.}$$

As briefly explained above, although $W^\star$ itself is a deterministic object, its practical approximation $W^\star_N$ is random. It is only in the limit of infinitely large domains Q_N that the deterministic value is attained. This is a standard situation in stochastic homogenization.

Many studies have been recently devoted (at least in the linear case) to establishing sharp estimates on the convergence of the random apparent homogenized quantities (computed on Q_N) to the exact deterministic homogenized quantities. We refer e.g. to [BP04, GO12] and to the comprehensive discussion of [BCLBL12b, Section 1.2]. We take here the problem from a slightly different perspective. We observe that the error

$$W^\star(\xi) - W^\star_N(\omega, \xi) = \Big(W^\star(\xi) - \mathbb{E}\left[W^\star_N(\cdot, \xi)\right] \Big) + \Big(\mathbb{E}\left[W^\star_N(\cdot, \xi)\right] - W^\star_N(\omega, \xi) \Big)$$

is the sum of a systematic error (the first term in the above right-hand side) and of a statistical error (the second term in the above right-hand side). We *focus here on the statistical error*, and propose approaches to reduce the confidence interval of empirical means approximating $\mathbb{E}\left[W^\star_N(\cdot, \xi)\right]$ (or similar quantities), for a given truncated domain Q_N.

Recall that a standard technique to compute an approximation of $\mathbb{E}\left[W^\star_N(\cdot, \xi)\right]$ is to consider several independent and identically distributed realizations of the energy density W, solve for each of them the corrector problem (2.9) (thereby obtaining several i.i.d. values $W^{\star,m}_N(\omega, \xi)$), and proceed following a Monte Carlo approach:

$$\mathbb{E}\left[W^\star_N(\cdot, \xi)\right] \approx I_{2M} := \frac{1}{2M} \sum_{m=1}^{2M} W^{\star,m}_N(\omega, \xi).$$

In view of the Central Limit Theorem, we know that our quantity of interest $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right]$ lies in the confidence interval

$$\left[I_{2M} - 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[W_N^\star(\cdot,\xi)\right]}}{\sqrt{2M}}, I_{2M} + 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[W_N^\star(\cdot,\xi)\right]}}{\sqrt{2M}}\right]$$

with a probability equal to 95 %.

In this article, we show that, using a well known variance reduction technique, the technique of *antithetic variables* [Liu08, page 27], we can design a practical approach that, for finite N and any vector ξ, allows to compute a better approximation of $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right]$ (and likewise for similar homogenized quantities). Otherwise stated, for an equal computational cost, the approach provides a more accurate (i.e. with a smaller confidence interval) approximation. We thereby extend to this nonlinear convex setting the results of [BCLBL12b, BCLBL12b, CLBL10] obtained in the linear case.

Our article is articulated as follows. In Section 2.2.1, we describe the proposed approach, and state our main results. The ingredients to prove these results are collected in Sections 2.2.2, 2.2.3 and 2.2.4. The actual proof of our main results is performed in Section 2.2.5. We make there several structural assumptions on the form of the energy density W to obtain these variance reduction results. In Section 2.2.6, we describe a general class of energy densities W for which our assumptions are indeed satisfied. We next turn in Section 2.3 to some illustrative numerical examples, where we demonstrate the efficiency of the approach, even in cases where the theoretical analysis is incomplete.

2.2 Description of the proposed approach and main results

2.2.1 Statement of our main results

This section is devoted to the presentation and the analysis of our approach. We first focus on estimating the expectation $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right]$ of the apparent homogenized energy density (see Section 2.2.1). Our variance reduction result, Proposition 2.1, shows that the technique of antithetic variables is indeed efficient. As often the case, it is difficult to *quantitatively* assess how efficient the approach is, and this will be the purpose of the numerical tests described in Section 2.3 to address this question.

We then turn to the estimation of the first (and next second) derivatives of $W_N^\star(\cdot,\xi)$ with respect to ξ. These quantities naturally appear when one solves the convex homogenized problem (2.2) (approximating $W^\star$ by $W_N^\star(\omega,\cdot)$), e.g. using a Newton algorithm. For these two quantities, our result is restricted to the one-dimensional setting. See Section 2.2.1 and Proposition 2.2 for the first derivative, and Section 2.2.1 and Proposition 2.5 for the second derivative.

Sections 2.2.2, 2.2.3, 2.2.4 and 2.2.5 are devoted to the proof of the results stated here. In Section 2.2.6, we discuss an explicit class of energy densities W that falls into our framework.

Variance reduction on the homogenized energy density

In this section, we make the following two *structure assumptions* on the rapidly oscillating field W of (2.1). First, we assume that, for any N, there exists an integer n (possibly

$n = |Q_N|$, but not necessarily) and a function $\mathcal{A}$, defined on $Q_N \times \mathbb{R}^n \times \mathbb{R}^d$, such that the field $W(y, \omega, \xi)$ writes

$$\forall y \in Q_N,\ \forall \xi \in \mathbb{R}^d, \quad W(y, \omega, \xi) = \mathcal{A}\left(y, X_1(\omega), \ldots, X_n(\omega), \xi\right) \quad \text{a.s.}, \tag{2.10}$$

where $\{X_k(\omega)\}_{1 \leq k \leq n}$ are independent scalar random variables, which are all distributed according to the uniform law $\mathcal{U}[0, 1]$. In general, the function $\mathcal{A}$, as well as the number n of independent, identically distributed variables involved in (2.10), depend on N, the size of Q_N, although this dependency is not made explicit in (2.10).

Second, we assume that the function $\mathcal{A}$ in (2.10) is such that, for all $y \in Q_N$ and all $\xi \in \mathbb{R}^d$, the map

$$(x_1, \ldots, x_n) \in \mathbb{R}^n \mapsto \mathcal{A}(y, x_1, \ldots, x_n, \xi) \tag{2.11}$$

is non-decreasing with respect to each of its arguments.

Proposition 2.1. *We assume (2.10)–(2.11). Let $W_N^\star(\omega, \xi)$ be the approximated homogenized energy density field defined by (2.9). We define on Q_N the field*

$$W^{\text{ant}}(y, \omega, \xi) := \mathcal{A}(y, 1 - X_1(\omega), \ldots, 1 - X_n(\omega), \xi),$$

antithetic to W defined by (2.10). We associate to this field the approximate homogenized energy density field $W_N^{\text{ant},\star}(\omega, \xi)$, defined by (2.9) (replacing W by W^{ant}). Set

$$\widetilde{W}_N^\star(\omega, \xi) := \frac{1}{2}\left(W_N^\star(\omega, \xi) + W_N^{\text{ant},\star}(\omega, \xi)\right). \tag{2.12}$$

Then, for any $\xi \in \mathbb{R}^d$,

$$\mathbb{E}\left[\widetilde{W}_N^\star(\cdot, \xi)\right] = \mathbb{E}\left[W_N^\star(\cdot, \xi)\right] \quad \textit{and} \quad \mathbb{V}\text{ar}\left[\widetilde{W}_N^\star(\cdot, \xi)\right] \leq \frac{1}{2}\mathbb{V}\text{ar}\left[W_N^\star(\cdot, \xi)\right]. \tag{2.13}$$

Otherwise stated, $\widetilde{W}_N^\star(\omega, \xi)$ is a random variable which has the same expectation as $W_N^\star(\omega, \xi)$, and its variance is smaller than half of that of $W_N^\star(\omega, \xi)$.

As mentioned above, this result generalizes [BCLBL12b, Proposition 2.1] to the non-linear convex variational setting considered here.

Before proceeding, we briefly explain the usefulness of the above result for variance reduction techniques. Assume we want to compute the expectation of $W_N^\star(\omega, \xi)$, for some fixed vector $\xi \in \mathbb{R}^d$. Following the classical Monte-Carlo method recalled in Section 2.1.2, we estimate $\mathbb{E}\left[W_N^\star(\cdot, \xi)\right]$ by its empirical mean. To this end, we consider $2M$ independent, identically distributed copies $\{W_m(y, \omega, \xi)\}_{1 \leq m \leq 2M}$ of the random field $W(y, \omega, \xi)$ on Q_N. To each copy W_m, we associate the approximate homogenized energy density $W_N^{\star,m}(\omega, \xi)$ defined by (2.9). We next introduce the empirical mean

$$I_{2M} = \frac{1}{2M} \sum_{m=1}^{2M} W_N^{\star,m}(\omega, \xi), \tag{2.14}$$

and consider that, in practice, the mean $\mathbb{E}\left[W_N^\star(\cdot, \xi)\right]$ is equal to the estimator I_{2M} within an approximate margin of error $1.96 \dfrac{\sqrt{\mathbb{V}\text{ar}\left[W_N^\star(\cdot, \xi)\right]}}{\sqrt{2M}}$.

Alternate to considering (2.14), we may consider

$$\widetilde{I}_{2M} = \frac{1}{M} \sum_{m=1}^{M} \widetilde{W}_N^{\star,m}(\omega, \xi), \tag{2.15}$$

where $\widetilde{W}_N^{\star,m}$ is defined by (2.12). Again, in practice, the mean $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right] = \mathbb{E}\left[\widetilde{W}_N^\star(\cdot,\xi)\right]$ is equal to $\widetilde{I}_{2M}$ within an approximate margin of error $1.96\dfrac{\sqrt{\mathbb{V}\mathrm{ar}\left[\widetilde{W}_N^\star(\cdot,\xi)\right]}}{\sqrt{M}}$. Observe now that both estimators (2.14) and (2.15) are of equal cost, since they require the same number $2M$ of corrector problems to be solved. The accuracy of the latter is better if and only if $\mathbb{V}\mathrm{ar}\left[\widetilde{W}_N^\star(\cdot,\xi)\right] \leq \frac{1}{2}\mathbb{V}\mathrm{ar}\left[W_N^\star(\cdot,\xi)\right]$, which is exactly the bound (2.13) of Proposition 2.1.

Variance reduction on the first derivative of the homogenized energy density

Restricting ourselves to the one-dimensional setting, we now state a variance reduction result for the estimation of $\mathbb{E}\left[\xi\partial_\xi W_N^\star(\cdot,\xi)\right]$. Note that, to distinguish derivatives with respect to y from derivatives with respect to ξ, we keep the notation $\partial_\xi W$, even though we are in the one-dimensional situation.

We again make the structure assumption (2.10), and observe that it implies that

$$\forall y \in (-N,N),\ \forall \xi \in \mathbb{R}, \quad \xi\partial_\xi W(y,\omega,\xi) = \mathcal{A}_1\left(y, X_1(\omega), \ldots, X_n(\omega), \xi\right) \quad \text{a.s.},$$

where $\{X_k(\omega)\}_{1\leq k\leq n}$ are scalar i.i.d. random variables, which are all distributed according to the uniform law $\mathcal{U}[0,1]$, and where the function $\mathcal{A}_1$, defined on $(-N,N)\times\mathbb{R}^n\times\mathbb{R}$, is given by

$$\mathcal{A}_1(y,x,\xi) = \xi\partial_\xi\mathcal{A}(y,x,\xi). \tag{2.16}$$

In addition, we assume that, for all $y \in (-N,N)$ and all $\xi \in \mathbb{R}$, the map

$$(x_1,\ldots,x_n) \in \mathbb{R}^n \mapsto \mathcal{A}_1(y,x_1,\ldots,x_n,\xi) \tag{2.17}$$

is non-decreasing with respect to each of its arguments.

We recall that the function $\xi \mapsto W(y,\omega,\xi)$ is strictly convex (see assumption (2.5)) and satisfies (2.4). It therefore has a unique minimizer $\xi_0(y,\omega)$. In the sequel, we consider energy densities such that this minimizer is independent of y and ω (see Remark 2.3 below). Without loss of generality, we can assume that $\xi_0 = 0$. We thus consider energy densities W such that

$$\xi \mapsto W(y,\omega,\xi) \text{ attains its minimum at } \xi = 0, \text{ a.e. and a.s.} \tag{2.18}$$

Proposition 2.2. *Let $d = 1$, and assume (2.10), (2.16), (2.17) and (2.18). We introduce*

$$\widetilde{\xi\partial_\xi W_N^\star}(\omega,\xi) := \frac{1}{2}\left(\xi\partial_\xi W_N^{\mathrm{ant},\star}(\omega,\xi) + \xi\partial_\xi W_N^\star(\omega,\xi)\right), \tag{2.19}$$

where $W_N^{\mathrm{ant},\star}(\omega,\xi)$ and $W_N^\star(\omega,\xi)$ are defined as in Proposition 2.1. Then, for any $\xi \in \mathbb{R}$,

$$\mathbb{E}\left[\widetilde{\xi\partial_\xi W_N^\star}(\cdot,\xi)\right] = \mathbb{E}\left[\xi\partial_\xi W_N^\star(\cdot,\xi)\right] \quad \textit{and} \quad \mathbb{V}\mathrm{ar}\left[\widetilde{\xi\partial_\xi W_N^\star}(\cdot,\xi)\right] \leq \frac{1}{2}\mathbb{V}\mathrm{ar}\left[\xi\partial_\xi W_N^\star(\cdot,\xi)\right]. \tag{2.20}$$

Remark 2.3. *Our work is motivated by the modeling of hyperelastic materials experiencing large deformations. In this context, Assumption (2.18) amounts to assuming that there exists a* natural configuration *of the material. This assumption is well-known in the material science community. Many materials, but not all, indeed satisfy such an assumption. Note also that similar assumptions can be found in different but related contexts, such as homogenization of the Hamilton-Jacobi equation (see e.g. [ACS14, Equation (2.9)]).*

Remark 2.4. *The fact that we consider the quantity of interest $\mathbb{E}[\xi\partial_\xi W_N^\star]$ is reminiscent of our work in the linear case, that is when $W(y,\omega,\xi) = \frac{1}{2}\xi^T A(y,\omega)\xi$ where A is a coercive bounded stationary symmetric matrix. In that case (see [BCLBL12b, BCLBL12b, CLBL10]), we* proved *variance reduction for the scalar quantities $\xi^T A_N^\star(\omega)\xi$ for any vector $\xi \in \mathbb{R}^d$ (note that variance reduction was also* observed *for other quantities). In the nonlinear setting studied in this article, we consider quantities of interest that, if the problem turns out to be linear, are equal to $\xi^T A_N^\star(\omega)\xi$ for some vector ξ. More general quantities of interest are considered in the numerical tests reported on in Section 2.3.*

Variance reduction on the second derivative of the homogenized energy density

Considering again the one-dimensional setting as in Section 2.2.1, we eventually state a variance reduction result for the estimation of $\mathbb{E}\left[\partial_\xi^2 W_N^\star(\cdot,\xi)\right]$.

Recall that, for any y and ω, the map $\xi \mapsto \partial_\xi W(y,\omega,\xi)$ is increasing. We can therefore introduce its reciprocal function $\zeta \mapsto \psi(y,\omega,\zeta)$, which is also increasing.

We again make the structure assumption (2.10), and observe that it implies that, for any $y \in (-N,N)$ and any $\zeta \in \mathbb{R}$,

$$\partial_\xi^2 W\left(y,\omega,\psi(y,\omega,\zeta)\right) = \mathcal{A}_2(y, X_1(\omega),\ldots,X_n(\omega),\zeta) \quad \text{a.s.},$$

where $\{X_k(\omega)\}_{1\le k\le n}$ are scalar i.i.d. random variables, which are all distributed according to the uniform law $\mathcal{U}[0,1]$, and where the function $\mathcal{A}_2$, defined on $(-N,N)\times\mathbb{R}^n\times\mathbb{R}$, is given by

$$\mathcal{A}_2(y,x,\zeta) = \partial_\xi^2 \mathcal{A}\left(y,x,\left[\partial_\xi \mathcal{A}(y,x,\cdot)\right]^{-1}(\zeta)\right), \tag{2.21}$$

where $\zeta \mapsto \left[\partial_\xi \mathcal{A}(y,x,\cdot)\right]^{-1}(\zeta)$ is the reciprocal function of $\xi \mapsto \partial_\xi \mathcal{A}(y,x,\xi)$.

In addition, we assume that, for all $y \in (-N,N)$ and all $\zeta \in \mathbb{R}$, the map

$$(x_1,\ldots,x_n) \in \mathbb{R}^n \mapsto \mathcal{A}_2(y,x_1,\ldots,x_n,\zeta) \tag{2.22}$$

is non-decreasing with respect to each of its arguments.

Proposition 2.5. *Let $d = 1$, and assume (2.10), (2.16), (2.17), (2.21) and (2.22). We also assume that (2.18) holds, and that*

$$\begin{gathered}\xi \mapsto \partial_\xi^2 W\left(y,\omega,\xi\right) \text{ is non decreasing for } \xi \ge 0 \\ \text{and non increasing for } \xi \le 0, \text{ a.e. and a.s.}\end{gathered} \tag{2.23}$$

We introduce

$$\widetilde{\partial_\xi^2 W_N^\star}(\omega,\xi) := \frac{1}{2}\left(\partial_\xi^2 W_N^{\mathrm{ant},\star}(\omega,\xi) + \partial_\xi^2 W_N^\star(\omega,\xi)\right),$$

where $W_N^{\mathrm{ant},\star}(\omega,\xi)$ and $W_N^\star(\omega,\xi)$ are defined as in Proposition 2.1. Then, for any $\xi \in \mathbb{R}$,

$$\mathbb{E}\left[\widetilde{\partial_\xi^2 W_N^\star}(\cdot,\xi)\right] = \mathbb{E}\left[\partial_\xi^2 W_N^\star(\cdot,\xi)\right] \quad \text{and} \quad \mathbb{V}\mathrm{ar}\left[\widetilde{\partial_\xi^2 W_N^\star}(\cdot,\xi)\right] \le \frac{1}{2}\mathbb{V}\mathrm{ar}\left[\partial_\xi^2 W_N^\star(\cdot,\xi)\right]. \tag{2.24}$$

The density $W(y,\omega,\xi) = a(y,\omega)|\xi|^p$, where a is positive and bounded away from zero and $p \ge 2$, typically satisfies the assumption (2.23).

2.2.2 Classical results on antithetic variables

We first recall the following lemma, and refer e.g. to [BCLBL12b, Lemma 2.1] for a proof. This result is crucial for our proof of variance reduction using the technique of antithetic variables, performed in Section 2.2.5.

Lemma 2.6 ([Liu08], page 27)**.** *Let f and g be two real-valued functions defined on $\mathbb{R}^n$, which are non-decreasing with respect to each of their arguments. Consider $X = (X_1, \dots, X_n)$ a vector of random variables, which are all independent from one another. Then*

$$\mathbb{C}\mathrm{ov}(f(X), g(X)) \geq 0. \tag{2.25}$$

The following result is a simple consequence of the above lemma (see [BCLBL12b, Corollary 2.3] for a proof).

Corollary 2.7 ([Liu08])**.** *Let f be a function defined on $\mathbb{R}^n$, which is non-decreasing with respect to each of its arguments. Consider $X = (X_1, \dots, X_n)$ a vector of random variables, which are all independent from one another, and distributed according to the uniform law $\mathcal{U}[0,1]$. Then*

$$\mathbb{V}\mathrm{ar}\left(\frac{1}{2}\left(f(X) + f(1-X)\right)\right) \leq \frac{1}{2}\mathbb{V}\mathrm{ar}\left(f(X)\right),$$

where we denote $1 - X = (1 - X_1, \dots, 1 - X_n) \in \mathbb{R}^n$.

Proof. Choosing $g(x_1, \dots, x_n) = -f(1 - x_1, \dots, 1 - x_n)$ in Lemma 2.6, we obtain that

$$\mathbb{C}\mathrm{ov}(f(X), f(1-X)) = \mathbb{C}\mathrm{ov}(f(X_1, \dots, X_n), f(1 - X_1, \dots, 1 - X_n)) \leq 0.$$

We next observe that

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\frac{1}{2}\left(f(X) + f(1-X)\right)\right) &= \frac{1}{2}\mathbb{V}\mathrm{ar}(f(X)) + \frac{1}{2}\mathbb{C}\mathrm{ov}\left(f(X), f(1-X)\right) \\ &\leq \frac{1}{2}\mathbb{V}\mathrm{ar}(f(X)),\end{aligned}$$

where we have used that $\mathbb{V}\mathrm{ar}(f(X)) = \mathbb{V}\mathrm{ar}(f(1-X))$. □

2.2.3 Derivatives of the corrector and of the homogenized energy density

We now introduce the correctors as the solutions to (2.9):

$$w^N(\cdot, \omega, \xi) := \operatorname{arginf}\left\{\int_{Q_N} W(\cdot, \omega, \xi + \nabla v), \quad v \in W^{1,p}_{\#}(Q_N), \quad \int_{Q_N} v = 0\right\}.$$

In this section, we derive some useful expressions for the derivatives with respect to ξ of w^N and of $W^\star_N$.

The first order optimality condition in (2.9) reads

$$\forall h \in W^{1,p}_{\#}(Q_N), \quad \int_{Q_N} (\nabla h)^T \, \partial_\xi W\left(\cdot, \omega, \xi + \nabla w^N\right) = 0. \tag{2.26}$$

We deduce from that condition that

$$\partial_\xi W^\star_N(\omega, \xi) = \frac{1}{|Q_N|}\int_{Q_N} \partial_\xi W\left(\cdot, \omega, \xi + \nabla w^N\right), \tag{2.27}$$

and we note that we do not need to know $\partial_\xi w^N$ to compute $\partial_\xi W_N^\star$. Computing the derivative of this equality with respect to ξ, we obtain that

$$\partial_\xi^2 W_N^\star(\omega,\xi) = \frac{1}{|Q_N|}\int_{Q_N} \left(\mathrm{Id} + \partial_\xi \nabla w^N\right) \partial_\xi^2 W\left(\cdot,\omega,\xi+\nabla w^N\right) \tag{2.28}$$

with the convention that $\left[\partial_\xi \nabla w^N\right]_{jk} = \dfrac{\partial^2 w^N}{\partial \xi_j \partial y_k}$ for $1 \le j,k \le d$. We can actually obtain a somewhat more symmetric expression. Computing the derivative of (2.26) with respect to ξ, we indeed see that

$$\forall h \in W^{1,p}_\#(Q_N), \quad \int_{Q_N} \left(\mathrm{Id} + \partial_\xi \nabla w^N\right) \partial_\xi^2 W\left(\cdot,\omega,\xi+\nabla w^N\right) \nabla h = 0. \tag{2.29}$$

We then infer from (2.28) and (2.29) that

$$\partial_\xi^2 W_N^\star(\omega,\xi) = \frac{1}{|Q_N|}\int_{Q_N} \left(\mathrm{Id} + \partial_\xi \nabla w^N\right) \partial_\xi^2 W\left(\cdot,\omega,\xi+\nabla w^N\right) \left(\mathrm{Id} + \partial_\xi \nabla w^N\right)^T. \tag{2.30}$$

Remark 2.8. *Using the same kind of arguments, we see that the function* $g_j = \dfrac{\partial w^N}{\partial \xi_j} \in W^{1,p}_\#(Q_N)$ *is solution to the variational formulation*

$$\forall h \in W^{1,p}_\#(Q_N), \quad \int_{Q_N} (\nabla h)^T \partial_\xi^2 W\left(\cdot,\omega,\xi+\nabla w^N\right) \nabla g_j = -\sum_{i=1}^d \int_{Q_N} \frac{\partial h}{\partial y_i} \frac{\partial^2 W}{\partial \xi_j \partial \xi_i}\left(\cdot,\omega,\xi+\nabla w^N\right). \tag{2.31}$$

Suppose that W is α-convex (i.e. satisfies (2.6)). Then problem (2.31) is well-posed and allows to uniquely determine (up to an additive constant) g_j, by solving a linear elliptic partial differential equation.

Combined with (2.30), this remark provides a practical way to compute $\partial_\xi^2 W_N^\star(\omega,\xi)$ without using any finite difference approximation in ξ.

We finally note that, in view of (2.27), we have

$$\xi \cdot \partial_\xi W_N^\star(\omega,\xi) = \frac{1}{|Q_N|}\int_{Q_N} \xi \cdot \partial_\xi W\left(\cdot,\omega,\xi+\nabla w^N\right). \tag{2.32}$$

Likewise, in view of (2.30), we see that

$$\xi^T \partial_\xi^2 W_N^\star(\omega,\xi)\xi = \frac{1}{|Q_N|}\int_{Q_N} \left[\xi + \nabla\left(\xi\cdot\partial_\xi w^N\right)\right]^T \partial_\xi^2 W\left(\cdot,\omega,\xi+\nabla w^N\right) \left[\xi + \nabla\left(\xi\cdot\partial_\xi w^N\right)\right]. \tag{2.33}$$

2.2.4 Monotonicity properties

Our goal in this section is to establish monotonicity properties for the homogenization process. Such properties are indeed useful to apply Corollary 2.7 and therefore prove variance reduction.

To simplify the notation, we assume in this section that we are in a *periodic* setting. For any $\xi \in \mathbb{R}^d$, the function $y \mapsto W(y,\xi)$ is supposed to be Q-periodic (with $Q = (0,1)^d$),

to satisfy the growth condition (2.4) and to be strictly convex with respect to ξ. The associated homogenized energy density is then given by

$$W^\star(\xi) = \inf \left\{ \int_Q W\left(y, \xi + \nabla w(y)\right) dy, \quad w \in W^{1,p}_\#(Q), \quad \int_Q w = 0 \right\}. \tag{2.34}$$

We first show a monotonicity property on the homogenized energy density in Section 2.2.4. Next, restricting ourselves to the one-dimensional setting, we show monotonicity properties for the first and the second derivative of the homogenized energy density (see respectively Sections 2.2.4 and 2.2.4).

On the homogenized energy density

The following result is an extension to the nonlinear setting of a well-known result in the linear setting (see [Tar97, page 12]).

Lemma 2.9. *Suppose that the fields W_1 and W_2 satisfy*

$$\forall \xi \in \mathbb{R}^d, \quad W_2(y,\xi) \geq W_1(y,\xi) \text{ a.e. on } Q. \tag{2.35}$$

We denote $W_1^\star$ and $W_2^\star$ the corresponding homogenized energy densities, defined by (2.34). We then have

$$\forall \xi \in \mathbb{R}^d, \quad W_2^\star(\xi) \geq W_1^\star(\xi). \tag{2.36}$$

Proof. Fix $\xi \in \mathbb{R}^d$. For any $v \in W^{1,p}_\#(Q)$ with $\displaystyle\int_Q v = 0$, we have that

$$W_1^\star(\xi) \leq \int_Q W_1\left(y, \xi + \nabla v(y)\right) dy \leq \int_Q W_2\left(y, \xi + \nabla v(y)\right) dy.$$

Taking the infimum over v, we obtain the claimed result. □

Remark 2.10. *Consider the case of an energy density that is positively homogeneous of degree p with respect to its variable ξ, that is such that $W(y,\lambda\xi) = |\lambda|^p W(y,\xi)$ for any $y \in \mathbb{R}^d$, $\xi \in \mathbb{R}^d$ and $\lambda \in \mathbb{R}$. A typical example is $W(y,\xi) = \dfrac{1}{p} a(y)|\xi|^p$. We then have, for any y and ξ, that*

$$\xi \cdot \partial_\xi W(y,\xi) = pW(y,\xi) \quad \text{and} \quad \xi^T \partial_\xi^2 W(y,\xi)\xi = p(p-1)W(y,\xi). \tag{2.37}$$

Using successively (2.32), (2.26) and (2.37), we obtain that

$$\begin{aligned} \xi \cdot \partial_\xi W^\star(\xi) &= \int_Q \xi \cdot \partial_\xi W\left(\cdot, \xi + \nabla w\right) \\ &= \int_Q (\xi + \nabla w) \cdot \partial_\xi W\left(\cdot, \xi + \nabla w\right) \\ &= p \int_Q W\left(\cdot, \xi + \nabla w\right) \\ &= pW^\star(\xi), \end{aligned} \tag{2.38}$$

where w is the corrector, solution to (2.34).

We next observe that, for any $\lambda \in \mathbb{R}$ and any $\xi \in \mathbb{R}^d$, we have $w(\cdot, \lambda\xi) = \lambda w(\cdot, \xi)$. Thus, for any y, the map $\xi \mapsto w(y, \xi)$ is homogeneous of degree one, and therefore $\xi \cdot \partial_\xi w = w$. We thus infer from (2.33), using (2.37), that

$$\begin{aligned} \xi^T \partial_\xi^2 W^\star(\xi)\xi &= \int_Q [\xi + \nabla w]^T \partial_\xi^2 W(\cdot, \xi + \nabla w)[\xi + \nabla w] \\ &= p(p-1) \int_Q W(\cdot, \xi + \nabla w) \\ &= p(p-1) W^\star(\xi). \end{aligned} \tag{2.39}$$

Consider now two fields W_1 and W_2 that are positively homogeneous of degree p with respect to the variable ξ and satisfy (2.35). Then we deduce from (2.36), (2.38) and (2.39) that, for all $\xi \in \mathbb{R}^d$,

$$\xi \cdot \partial_\xi W_2^\star(\xi) \geq \xi \cdot \partial_\xi W_1^\star(\xi) \quad \text{and} \quad \xi^T \partial_\xi^2 W_2^\star(\xi)\xi \geq \xi^T \partial_\xi^2 W_1^\star(\xi)\xi.$$

On the first derivative of the homogenized energy density

We now establish a monotonicity result on the derivative of $W^\star(\xi)$, in the one-dimensional setting.

As in Section 2.2.1 (see (2.18)), we consider energy densities W such that

$$\xi \mapsto W(y, \xi) \text{ attains its minimum at } \xi = 0 \text{ for almost all } y \in Q. \tag{2.40}$$

Lemma 2.11. *Let $d = 1$, and consider two energy densities W_1 and W_2 satisfying (2.40), and such that*

$$\forall \xi \in \mathbb{R}, \quad \xi \partial_\xi W_2(y, \xi) \geq \xi \partial_\xi W_1(y, \xi) \text{ a.e. on } (0, 1). \tag{2.41}$$

We denote $W_1^\star$ and $W_2^\star$ the corresponding homogenized energy densities, defined by (2.34). We then have

$$\forall \xi \in \mathbb{R}, \quad \xi \partial_\xi W_2^\star(\xi) \geq \xi \partial_\xi W_1^\star(\xi). \tag{2.42}$$

Proof. We first claim that

$$\partial_\xi W^\star(\xi) \text{ has the same sign as } \xi. \tag{2.43}$$

To prove this, we note that the corrector equation reads (see (2.26))

$$\frac{d}{dy}\left[\partial_\xi W\left(y, \xi + \frac{dw}{dy}(y, \xi)\right)\right] = 0 \quad \text{on } (0, 1), \quad w(\cdot, \xi) \text{ is 1-periodic.}$$

We therefore see that $\partial_\xi W\left(y, \xi + \frac{dw}{dy}(y, \xi)\right)$ is independent of y, and using (2.27), we obtain that

$$\partial_\xi W\left(y, \xi + \frac{dw}{dy}(y, \xi)\right) = \partial_\xi W^\star(\xi) \quad \text{on } (0, 1).$$

Let $\xi \mapsto \psi(y, \xi)$ be the reciprocal function of $\xi \mapsto \partial_\xi W(y, \xi)$, which exists and is increasing thanks to the strict convexity of $\xi \mapsto W(y, \xi)$. We deduce from the above equation, after integration over $(0, 1)$, that

$$\xi = \int_0^1 \psi\left(y, \partial_\xi W^\star(\xi)\right) dy. \tag{2.44}$$

We are now in position to prove (2.43). Indeed, we first note that (2.40), that reads $\partial_\xi W(y, \xi = 0) = 0$, implies that $\psi(y, 0) = 0$. If $\partial_\xi W^\star(\xi) \geq 0$, then $\psi(y, \partial_\xi W^\star(\xi)) \geq$

$\psi(y,0) = 0$, hence, integrating over $(0,1)$ and using (2.44), we obtain $\xi \geq 0$. Likewise, $\partial_\xi W^\star(\xi) \leq 0$ implies that $\xi \leq 0$. The claim (2.43) is proved.

To proceed, we see that the assumption (2.41) equivalently reads, using the reciprocal functions,

$$\forall \zeta \in \mathbb{R}, \quad \zeta\psi_2(y,\zeta) \leq \zeta\psi_1(y,\zeta) \quad \text{a.e. on } (0,1). \tag{2.45}$$

We now prove (2.42) by contradiction. Assume that $\xi\partial_\xi W_2^\star(\xi) < \xi\partial_\xi W_1^\star(\xi)$ for some $\xi \in \mathbb{R}$. Without loss of generality, we can assume that $\xi > 0$, and therefore $\partial_\xi W_2^\star(\xi) < \partial_\xi W_1^\star(\xi)$. Using (2.43), we additionally have $0 < \partial_\xi W_2^\star(\xi)$. Using that $\zeta \mapsto \psi_2(y,\zeta)$ is increasing and (2.45) with $\zeta = \partial_\xi W_1^\star(\xi) > 0$, we have

$$\psi_2\left(y, \partial_\xi W_2^\star(\xi)\right) < \psi_2\left(y, \partial_\xi W_1^\star(\xi)\right) \leq \psi_1\left(y, \partial_\xi W_1^\star(\xi)\right).$$

Integrating over $(0,1)$ and using (2.44) yields

$$\xi = \int_0^1 \psi_2\left(y, \partial_\xi W_2^\star(\xi)\right) dy < \int_0^1 \psi_1\left(y, \partial_\xi W_1^\star(\xi)\right) dy = \xi,$$

and we reach a contradiction. This concludes the proof. □

On the second derivative of the homogenized energy density

We next turn to monotonicity properties of the second derivative of the homogenized energy density. As in Section 2.2.4, we consider energy densities satisfying (2.40). In the spirit of (2.23), we additionally request that, almost everywhere in $(0,1)$,

$$\begin{gathered} \xi \mapsto \partial_\xi^2 W\left(y,\xi\right) \text{ is non decreasing for } \xi \geq 0 \\ \text{and non increasing for } \xi \leq 0. \end{gathered} \tag{2.46}$$

Lemma 2.12. *Let $d = 1$, and consider two energy densities W_1 and W_2 satisfying (2.40), (2.41), (2.46) and such that*

$$\forall \zeta \in \mathbb{R}, \quad \partial_\xi^2 W_2\left(y, \psi_2(y,\zeta)\right) \geq \partial_\xi^2 W_1\left(y, \psi_1(y,\zeta)\right) \ \text{ a.e. on } (0,1). \tag{2.47}$$

We denote $W_1^\star$ and $W_2^\star$ the corresponding homogenized energy densities, defined by (2.34). We then have

$$\forall \xi \in \mathbb{R}, \quad \partial_\xi^2 W_2^\star(\xi) \geq \partial_\xi^2 W_1^\star(\xi). \tag{2.48}$$

We recall that $\zeta \mapsto \psi(y,\zeta)$ is the reciprocal function of $\xi \mapsto \partial_\xi W(y,\xi)$.

Proof. We first compute the derivative of (2.44) and obtain

$$\frac{1}{\partial_\xi^2 W^\star(\xi)} = \int_0^1 \frac{dy}{\partial_\xi^2 W\left[y, \psi\left(y, \partial_\xi W^\star(\xi)\right)\right]}. \tag{2.49}$$

It is sufficient to prove (2.48) for $\xi > 0$. Using (2.42) and the fact that ψ_1 and $\partial_\xi^2 W_1$ are non-decreasing with respect to their second argument, we have

$$\partial_\xi^2 W_1\left(y, \psi_1\left(y, \partial_\xi W_1^\star(\xi)\right)\right) \leq \partial_\xi^2 W_1\left(y, \psi_1\left(y, \partial_\xi W_2^\star(\xi)\right)\right).$$

Using (2.47) for $\zeta = \partial_\xi W_2^\star(\xi)$, we deduce that

$$\partial_\xi^2 W_1\left(y, \psi_1\left(y, \partial_\xi W_1^\star(\xi)\right)\right) \leq \partial_\xi^2 W_2\left(y, \psi_2\left(y, \partial_\xi W_2^\star(\xi)\right)\right).$$

In view of (2.49), this inequality readily implies (2.48) for $\xi > 0$. This concludes the proof. □

2.2.5 Proof of Propositions 2.1, 2.2 and 2.5

Now that we have collected all the necessary ingredients, we are in position to prove our main results.

Variance reduction on the homogenized energy density

Proof of Proposition 2.1. As $1 - X_k(\omega)$ and $X_k(\omega)$ share the same law, so do the fields W and W^{ant} on Q_N. Hence, the homogenized fields $W_N^\star(\omega,\xi)$ and $W_N^{\text{ant},\star}(\omega,\xi)$ share the same law, and we obtain the first assertion of (2.13).

We now choose a vector $\xi \in \mathbb{R}^d$, and denote by $\mathcal{P}_N^\xi$ the operator that associates to a given Q_N-periodic energy density the homogenized energy density evaluted at ξ. We see from (2.9) that $W_N^\star(\omega,\xi)$ is the effective energy density (evaluated at ξ) obtained by periodic homogenization of $W_{|y\in Q_N}$:

$$W_N^\star(\omega,\xi) = \mathcal{P}_N^\xi\left[W(\cdot,\omega,\cdot)_{|y\in Q_N}\right] \quad \text{a.s.} \tag{2.50}$$

Using the function $\mathcal{A}$ of (2.10), we introduce the map

$$\begin{array}{rcl} f:\mathbb{R}^n & \to & \mathbb{R} \\ x & \mapsto & \mathcal{P}_N^\xi\left[\mathcal{A}(\cdot,x,\cdot)\right], \end{array}$$

see that $f(X(\omega)) = W_N^\star(\omega,\xi)$ and that, using the definition (2.12) of $\widetilde{W}_N^\star(\omega,\xi)$, we have

$$\frac{1}{2}\left(f(X(\omega)) + f(1-X(\omega))\right) = \frac{1}{2}\left(W_N^\star(\omega,\xi) + W_N^{\text{ant},\star}(\omega,\xi)\right) = \widetilde{W}_N^\star(\omega,\xi). \tag{2.51}$$

We have used above the notation $1 - X = (1-X_1,\dots,1-X_n) \in \mathbb{R}^n$ introduced in Corollary 2.7.

We know from Assumption (2.11) that, for any $y \in Q_N$ and any $\zeta \in \mathbb{R}^d$, the function $\mathcal{A}(y,\cdot,\zeta)$ is non-decreasing with respect to each of its arguments. In view of Lemma 2.9, we obtain that f is non-decreasing.

We are thus in position to use Corollary 2.7, which yields

$$\mathbb{V}\text{ar}\left(\frac{1}{2}\left(f(X)+f(1-X)\right)\right) \le \frac{1}{2}\mathbb{V}\text{ar}\left(f(X)\right).$$

Using (2.51), we obtain

$$\mathbb{V}\text{ar}\left(\widetilde{W}_N^\star(\cdot,\xi)\right) = \mathbb{V}\text{ar}\left[\frac{1}{2}\left(f(X)+f(1-X)\right)\right] \le \frac{1}{2}\mathbb{V}\text{ar}\left(f(X)\right) = \frac{1}{2}\mathbb{V}\text{ar}\left(W_N^\star(\cdot,\xi)\right),$$

which concludes the proof of the second assertion of (2.13) and of Proposition 2.1. □

Remark 2.13. *Following Remark 2.10, consider a positively homogeneous energy density W. We have shown there that $\xi\cdot\partial_\xi W_N^\star(\omega,\xi)$ and $\xi^T\partial_\xi^2 W_N^\star(\omega,\xi)\xi$ are equal (up to a deterministic multiplicative constant) to $W_N^\star(\omega,\xi)$. Thus, under Assumptions (2.10)–(2.11), variance reduction holds for these two outputs as well.*

Variance reduction on the first derivative of the homogenized energy density

Proof of Proposition 2.2. The proof follows the same lines as that of Proposition 2.1.

As $1 - X_k(\omega)$ and $X_k(\omega)$ share the same law, so do the fields W and W^{ant} on Q_N. Hence, the quantities $\xi\partial_\xi W_N^\star(\omega,\xi)$ and $\xi\partial_\xi W_N^{\text{ant},\star}(\omega,\xi)$ share the same law, which implies the first assertion of (2.20).

To prove the second assertion, we again make use, as in the proof of Proposition 2.1, of the operator $\mathcal{P}_N^\xi$ that associates to a given Q_N-periodic energy density the homogenized energy density evaluted at ξ (here, $Q_N = (-N,N)$). Expression (2.50) holds. Choosing a vector $\xi \in \mathbb{R}$, we introduce the function

$$\begin{aligned} f : \mathbb{R}^n &\to \mathbb{R} \\ x &\mapsto \xi\partial_\xi\left[\mathcal{P}_N^\xi\left(\mathcal{A}(\cdot,x,\cdot)\right)\right], \end{aligned}$$

which obviously satisfies $f(X(\omega)) = \xi\partial_\xi W_N^\star(\omega,\xi)$. Using the definition (2.19) of $\widetilde{\xi\partial_\xi W_N^\star}$, we have

$$\frac{1}{2}\left[f(X(\omega)) + f(1-X(\omega))\right] = \frac{1}{2}\left[\xi\partial_\xi W_N^\star(\omega,\xi) + \xi\partial_\xi W_N^{\text{ant},\star}(\omega,\xi)\right] = \widetilde{\xi\partial_\xi W_N^\star}(\omega,\xi). \quad (2.52)$$

Using (2.16) and (2.17), we infer from Lemma 2.11 that f is non-decreasing.

Using Corollary 2.7, we write that $\mathbb{V}\text{ar}\left(\frac{1}{2}\left(f(X)+f(1-X)\right)\right) \leq \frac{1}{2}\mathbb{V}\text{ar}\left(f(X)\right)$. In view of (2.52), we recast this inequality as

$$\mathbb{V}\text{ar}\left[\widetilde{\xi\partial_\xi W_N^\star}(\cdot,\xi)\right] \leq \frac{1}{2}\mathbb{V}\text{ar}\left(\xi\partial_\xi W_N^\star(\cdot,\xi)\right),$$

and therefore obtain the second assertion of (2.20). This concludes the proof of Proposition 2.2. □

Variance reduction on the second derivative of the homogenized energy density

Proof of Proposition 2.5. The proof follows the same lines as the proof of Proposition 2.2. Using Assumptions (2.17) and (2.22), we see that Assumptions (2.41) and (2.47) of Lemma 2.12 are satisfied. The monotonicity result of Lemma 2.12 next allows to use Corollary 2.7, which implies (2.24). □

2.2.6 Examples satisfying our structure assumptions

Before proceeding to the numerical tests, we give here some specific examples of fields W that satisfy the above assumptions. We consider the case

$$W(y,\omega,\xi) = a(y,\omega)\frac{|\xi|^p}{p} + c(y,\omega)\frac{|\xi|^2}{2}, \quad p \geq 2, \quad (2.53)$$

with $c(y,\omega) \geq 0$ and $a(y,\omega) \geq a_- > 0$ a.e. and a.s., and provide sufficient conditions on the scalar fields a and c for the structure assumptions (2.10), (2.11), (2.17) and (2.22) to be satisfied. Note that (2.18) and (2.23) are already fullfilled.

Consider two families $(a_k(\omega))_{k\in\mathbb{Z}^d}$ and $(c_k(\omega))_{k\in\mathbb{Z}^d}$ of independent, identically distributed random variables, and assume that

$$a(y,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(y)a_k(\omega), \quad c(y,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(y)c_k(\omega), \quad (2.54)$$

where $Q = (0,1)^d$ and $Q+k$ is the cube Q translated by the vector $k \in \mathbb{Z}^d$. The scalar field $a(y,\omega)$ is therefore constant in each cube $Q+k$ with i.i.d. values $a_k(\omega)$, and likewise for $c(y,\omega)$.

We assume that there exist $\alpha > 0$ and $\beta < \infty$ such that, for all $k \in \mathbb{Z}^d$, $0 < \alpha \leq a_k(\omega) \leq \beta < +\infty$ and $0 \leq c_k(\omega) \leq \beta < +\infty$ almost surely. Consequently, (2.4) holds.

Introduce now the cumulative distribution functions $P_a(x) = \nu_a(-\infty, x)$, where ν_a is the common probability measure of all the a_k, and next the non-decreasing functions $f_a(x) = \inf\{z; P_a(x) \geq z\}$. Then, for any random variable $X^a(\omega)$ uniformly distributed in $[0,1]$, the random variable $f_a(X^a(\omega))$ is distributed according to the measure ν_a. As a consequence, we can recast (2.54) in the form

$$a(y,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(y) f_a(X_k^a(\omega)),$$

where $(X_k^a(\omega))_{k\in\mathbb{Z}^d}$ is a family of independent random variables that are all uniformly distributed in $[0,1]$, and f_a is non-decreasing. We can proceed likewise for the variables c_k. This yields an example where (2.10), (2.11) and (2.17) hold. In particular, the function $\mathcal{A}$ of (2.10) reads

$$\mathcal{A}(y, x_a, x_c, \xi) = \frac{|\xi|^p}{p} \sum_{k\in I_N} \mathbf{1}_{Q+k}(y) f_a(x_k^a) + \frac{|\xi|^2}{2} \sum_{k\in I_N} \mathbf{1}_{Q+k}(y) f_c(x_k^c),$$

where $I_N = \left\{k \in \mathbb{Z}^d \text{ s.t. } Q+k \subset Q_N\right\}$ and $x_a = \{x_k^a\}_{k\in I_N}$. As shown in [BCLBL12b], more general fields $a(y,\omega)$ (where random variables may be correlated) also fall into this framework.

In what follows, we prove that, under assumptions (2.53) and (2.54), and if $p \leq 3$, the structure assumption (2.22) holds. Without loss of generality, we may assume that $y \in (0,1)$, and write that

$$\forall y \in (0,1), \quad \mathcal{A}(x_a, x_c, \xi) = \bar{a}\frac{|\xi|^p}{p} + \bar{c}\frac{|\xi|^2}{2},$$

with $\bar{a} = f_a(x_0^a)$ and $\bar{c} = f_c(x_0^c)$. By a slight abuse of notation, we keep implicit the dependency with respect to y, work with $\bar{a}$ and $\bar{c}$ rather than x_a and x_c, and write

$$\mathcal{A}(\bar{a}, \bar{c}, \xi) = \bar{a}\frac{|\xi|^p}{p} + \bar{c}\frac{|\xi|^2}{2}.$$

We compute

$$\partial_\xi \mathcal{A}(\bar{a}, \bar{c}, \xi) = \bar{a}|\xi|^{p-2}\xi + \bar{c}\xi$$

and denote $\zeta \mapsto g(\bar{a}, \bar{c}, \zeta)$ the reciprocal to the function $\xi \mapsto \partial_\xi \mathcal{A}(\bar{a}, \bar{c}, \xi)$:

$$\zeta = \bar{a}\ |g(\bar{a}, \bar{c}, \zeta)|^{p-2}\, g(\bar{a}, \bar{c}, \zeta) + \bar{c} g(\bar{a}, \bar{c}, \zeta).$$

The function $\mathcal{A}_2$ of (2.22) then reads

$$\mathcal{A}_2(\bar{a}, \bar{c}, \zeta) = (p-1)\bar{a}\,|g(\bar{a}, \bar{c}, \zeta)|^{p-2} + \bar{c}.$$

We are left with showing that $\mathcal{A}_2$ is non-decreasing with respect to $\bar{a}$ and $\bar{c}$.

A first remark is that since $g(\bar{a}, \bar{c}, \zeta)$ has the same sign as ζ (recall that $\bar{a} > 0$ and $\bar{c} \geq 0$), we may as well restrict ourselves to $\zeta > 0$ and $g(\bar{a}, \bar{c}, \zeta) > 0$. We hence have

$$\begin{aligned} \mathcal{A}_2(\bar{a}, \bar{c}, \zeta) &= (p-1)\bar{a} g(\bar{a}, \bar{c}, \zeta)^{p-2} + \bar{c}, \\ \zeta &= \bar{a} g(\bar{a}, \bar{c}, \zeta)^{p-1} + \bar{c} g(\bar{a}, \bar{c}, \zeta). \end{aligned} \tag{2.55}$$

We first compute the derivative of $\mathcal{A}_2$ with respect to $\bar{a}$:

$$\frac{\partial \mathcal{A}_2}{\partial \bar{a}} = (p-1)g(\bar{a},\bar{c},\zeta)^{p-2} + (p-1)(p-2)\bar{a}\, g(\bar{a},\bar{c},\zeta)^{p-3}\frac{\partial g}{\partial \bar{a}}.$$

Using (2.55) to compute $\dfrac{\partial g}{\partial \bar{a}}$, we obtain that

$$\left(\bar{c} + (p-1)\bar{a}g^{p-2}\right)\frac{\partial \mathcal{A}_2}{\partial \bar{a}} = (p-1)\bar{c}g^{p-2} + \bar{a}(p-1)g^{2p-4},$$

and since $p > 1$, $\bar{a} > 0$ and $g > 0$, we deduce that $\dfrac{\partial \mathcal{A}_2}{\partial \bar{a}} \geq 0$.

We next compute the derivative of $\mathcal{A}_2$ with respect to $\bar{c}$. Using again (2.55) to compute $\dfrac{\partial g}{\partial \bar{c}}$, we obtain that

$$\left(\bar{c} + (p-1)\bar{a}g^{p-2}\right)\frac{\partial \mathcal{A}_2}{\partial \bar{c}} = \bar{c} - (p-1)(p-3)\bar{a}g^{p-2}.$$

Recall that $\bar{a} > 0$, $\bar{c} \geq 0$, $p > 1$ and $g > 0$. We have assumed that $p \leq 3$, and therefore deduce from the above relation that $\dfrac{\partial \mathcal{A}_2}{\partial \bar{c}} \geq 0$. The structure assumption (2.22) hence holds in that case.

Remark 2.14. *The argument above also shows that the case*

$$W(y,\omega,\xi) = a(y,\omega)\frac{|\xi|^p}{p},$$

along with assumption (2.54), falls into our framework for any $p \geq 2$.

It is likely that other settings, such as

$$W(y,\omega,\xi) = \left(a(y,\omega) + c(y,\omega)\right)\frac{|\xi|^p}{p} + c(y,\omega)\frac{|\xi|^2}{2},$$

along with assumption (2.54), where a_k and c_k are all independent random variables, also fall into our framework. We do not pursue in this direction here.

2.3 Numerical results

Our numerical experiments are presented in Section 2.3.2, and discussed in details in the subsequent sections. In Section 2.3.1, we first discuss the algorithm we used to solve the variational problem (2.9) that defines the apparent homogenized energy density.

2.3.1 Newton algorithm to solve the truncated corrector problem

As mentioned above, the corrector problem (2.9) is a convex minimization problem, which has been well studied in the literature (see e.g. [BL93, Cho89, GM75, LT94]). We explain here how we proceed in practice to solve this problem, assuming that W is not only strictly convex, but actually α-convex (i.e. satisfies (2.6)).

To simplify our exposition, we use the notation of the Q-periodic case, where the corrector problem is (2.34). We introduce some basis functions $\{\varphi_i\}_{i\in I}$ (e.g. finite element

functions) where $\varphi_i \in W^{1,p}_{\#}(Q)$, and the finite dimensional space $V_h = \mathrm{Span}\,\{\varphi_i, i \in I\}$. Consider the functional

$$J(w) = J\left(\{w_i\}_{i\in I}\right) = \int_Q W(y, \xi + \nabla w(y))\, dy$$

defined on V_h, with

$$w(y) = \sum_{i\in I} w_i\ \varphi_i(y),$$

and the variational problem

$$\inf_{v_h \in V_h} J(v_h). \tag{2.56}$$

This problem has a unique solution (denoted $w_h \in V_h$) up to the addition of a constant. The quantity ∇w_h is well-defined, and is the finite-dimensional approximation of ∇w, where w is the solution to (2.34).

In practice, problem (2.56) is solved using a Newton algorithm. We see that

$$\frac{\partial J}{\partial w_j}(w) = D_w(\varphi_j) \quad \text{and} \quad \frac{\partial^2 J}{\partial w_j \partial w_k}(w) = H_w(\varphi_j, \varphi_k)$$

where

$$D_w(\varphi) = \int_Q \nabla\varphi(y) \cdot \partial_\xi W(y, \xi + \nabla w(y))\, dy$$

and

$$H_w(\varphi, \psi) = \int_Q (\nabla\varphi(y))^T\, \partial^2_\xi W(y, \xi + \nabla w(y))\, \nabla\psi(y)\, dy.$$

The Newton algorithm consists in defining $w_h^{m+1} \in V_h$ from $w_h^m \in V_h$ by the following linear elliptic problem: find $w_h^{m+1} \in V_h$ such that

$$\forall \theta \in V_h, \quad H_{w_h^m}(w_h^{m+1} - w_h^m, \theta) = -D_{w_h^m}(\theta).$$

Again, w_h^{m+1} is uniquely defined up to the addition of a constant.

The finite-dimensional problem (2.56) is α-convex, and W is smooth with respect to ξ: the Newton algorithm hence locally converges (quadratically), and $\lim_{m\to\infty} \nabla w_h^m = \nabla w_h$.

In practice, we consider a sequence $\mathcal{T}_h$ of meshes on Q, and set $V_h = \mathbb{P}^1_h(Q) = \{v_h \in C(Q) \text{ s.t. } \forall T \in \mathcal{T}_h, v_h \text{ is affine on } T\}$. By classical finite element results, we know that $\lim_{h\to 0} \|\nabla w_h - \nabla w\|_{L^p(Q)} = 0$ (see e.g. [BL93] and also [AV12, Tho97]).

2.3.2 Overview of numerical results

We have considered three test-cases of the form (2.53)–(2.54), namely

$$W(y, \omega, \xi) = a(y, \omega)\frac{|\xi|^p}{p} + c(y, \omega)\frac{|\xi|^2}{2}$$
$$\text{with } a(y, \omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(y) a_k(\omega) \quad \text{and} \quad c(y, \omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(y) c_k(\omega),$$

with $p = 4$, in dimension $d = 2$. The random variables a_k follow a Bernoulli distribution: $\mathbb{P}(a_k = \alpha) = \mathbb{P}(a_k = \beta) = 1/2$, with $\alpha = 3$ and $\beta = 23$. The value of the field c is chosen as follows:

- Test Case 1: in this first test case, $c(y,\omega) = 0$. The problem is thus strictly convex but not α-convex. In addition, the energy density is positively homogeneous of degree p, hence Remarks 2.10 and 2.13 apply.

- Test Case 2: the second test case corresponds to $c(y,\omega) = 1$. The problem is then α-convex, and highly oscillatory only in its non-harmonic component.

- Test Case 3: for the third test case, we work with $c(y,\omega)$ chosen according to (2.54), where $\mathbb{P}(c_k = \gamma) = \mathbb{P}(c_k = \delta) = 1/2$, with $\gamma = 1$ and $\delta = 3$. The problem is thus highly oscillatory both in its non-harmonic and its harmonic components.

We take the meshsize $h = 0.2$. The Newton algorithm is initialized with the solution w_0 to

$$-\mathrm{div}\,[(a(y,\omega) + c(y,\omega))(\xi + \nabla w_0)] = 0 \quad \text{in } Q_N, \quad w_0 \text{ is } Q_N\text{-periodic},$$

and the iterations stop when $\dfrac{\|w_h^{n+1} - w_h^n\|_{W^{1,p}}}{\|w_h^n\|_{W^{1,p}}} \leq \texttt{tol}$. If $\texttt{tol}$ is chosen too large, then (2.56) is inaccurately solved, and the variance reduction is not very good. For our numerical tests, we set $\texttt{tol} = 10^{-5}$: the discrete problem (2.56) is accurately solved, while only a limited number of iterations (in practice, around 5 iterations) are needed.

For the numerical tests, we adopt the convention that $Q_N = (-N,N)^2$. For each Q_N, the standard Monte Carlo results have been obtained using $2M = 100$ realizations (from which we build the empirical estimator (2.14)). For the antithetic variable approach, we have also solved $2M$ corrector problems, from which we build the empirical estimator (2.15). Therefore, in all what follows, we compare the accuracy of the Monte Carlo approach (MC) and the Antithetic Variable approach (AV) at *equal computational cost*.

2.3.3 Test Case 1

In this test case, the energy density is positively homogeneous. We therefore know, from Proposition 2.1 and Remark 2.13, that our approach yields estimations of the expectation of $W_N^\star(\omega,\xi)$, $\xi \cdot \partial_\xi W_N^\star(\omega,\xi)$ and $\xi^T \partial_\xi^2 W_N^\star(\omega,\xi)\xi$ with a smaller variance than the standard Monte Carlo approach. Our aim here is to *quantify* the efficiency gain. Note also that we have *not* taken into account, in our implementation, the fact that $W_N^\star(\omega,\xi)$, $\xi \cdot \partial_\xi W_N^\star(\omega,\xi)$ and $\xi^T \partial_\xi^2 W_N^\star(\omega,\xi)\xi$ are here proportional to one another.

To begin with, we show on Figure 2.1 the estimation by empirical means (along with a 95 % confidence interval) of three quantities (the homogenized energy density, its first derivative with respect to ξ_1 and its second derivative with respect to ξ_1 and ξ_2; we refer to [LM13] for more comprehensive numerical results). We observe that the variance of all quantities decreases when the size of Q_N increases, and that confidence intervals obtained with the antithetic variable approach are smaller than those obtained with a standard Monte Carlo approach, for an equal computational cost.

We now turn to a more quantitative analysis of the variance. Figure 2.2 shows the variances

$$V_{\mathrm{MC}} = \frac{1}{2}\mathbb{V}\mathrm{ar}\,[W_N^\star(\cdot,\xi)] \quad \text{and} \quad V_{\mathrm{AV}} = \mathbb{V}\mathrm{ar}\left[\widetilde{W}_N^\star(\cdot,\xi)\right] \tag{2.57}$$

as a function of N (note the factor $1/2$ in the definition of V_{MC}, consistent with (2.13), (2.14) and (2.15)). We observe that the variance of any of our quantities of interest (obtained either with the Monte Carlo approach or the Antithetic Variable approach) decreases at the rate $1/|Q_N|$ as N increases (as expected if one could use the Central Limit Theorem).

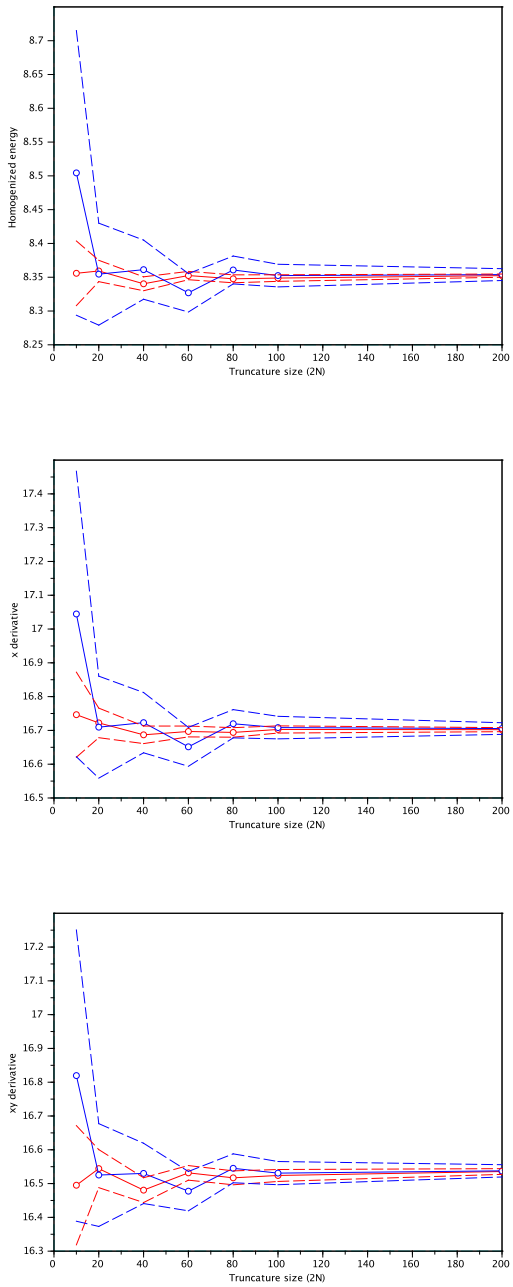

Figure 2.1 – Test-Case 1: Homogenized quantities as a function of $2N$, for the vector $\xi = (1,1)^T$ (Blue: Monte Carlo results; Red: Antithetic variable approach; Dashed lines: 95% confidence interval, equating the cost of the two approaches). From top to bottom: estimation of $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right]$, $\mathbb{E}\left[\partial_{\xi_1} W_N^\star(\cdot,\xi)\right]$ and $\mathbb{E}\left[\partial^2_{\xi_1\xi_2} W_N^\star(\cdot,\xi)\right]$.

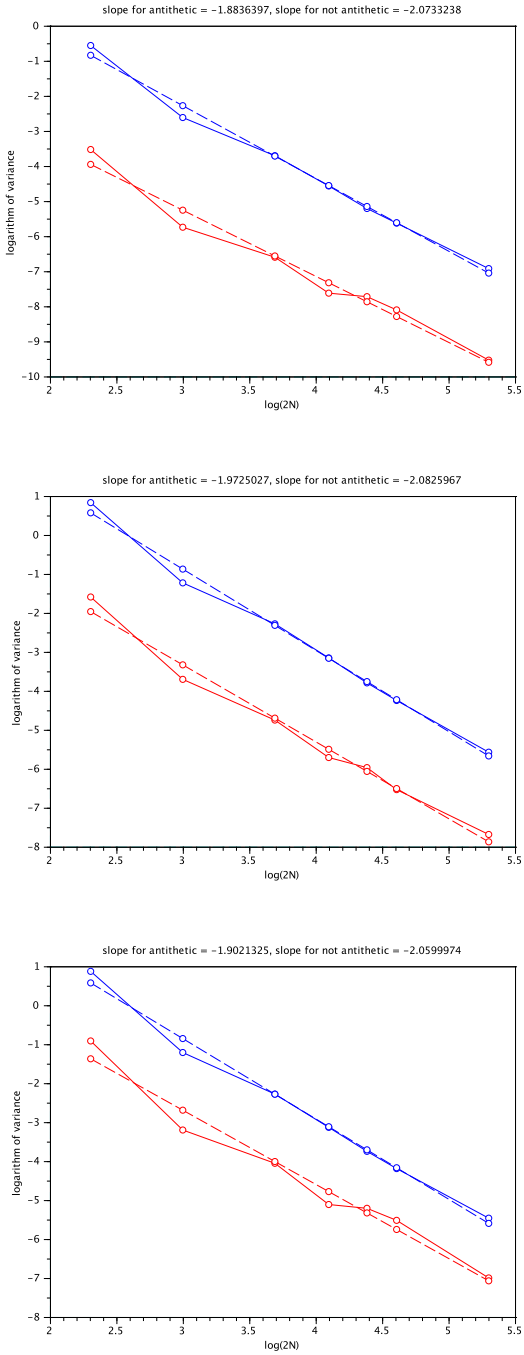

Figure 2.2 – Test-Case 1: Variances (2.57) of the same quantities of interest as on Figure 2.1, as a function of $2N$ (Blue: Monte Carlo approach; Red: Antithetic Variable approach; Natural logarithm plot). Solid line: actual results. Dashed line: linear regression fit.

We also observe that the variance obtained with our approach is systematically smaller than the Monte Carlo variance, in the sense that $V_{\rm AV} \leq V_{\rm MC}$.

We next report on Table 2.1 the variance reduction ratio

$$R = \frac{V_{\rm MC}}{V_{\rm AV}} = \frac{\mathbb{V}{\rm ar}\left[W_N^\star(\cdot,\xi)\right]}{2\mathbb{V}{\rm ar}\left[\widetilde{W}_N^\star(\cdot,\xi)\right]}, \tag{2.58}$$

which measures the gain in computational cost at equal accuracy, or the square of the accuracy gain at equal computational cost. We report this ratio for several quantities of interest. Although this ratio somewhat varies with N, we observe that it is of the order of 10 for all quantities of interest, except for $\partial^2_{\xi_1\xi_2} W_N^\star$, for which it is always larger than 4. In particular, even if N is not large (because we cannot afford to work on a large domain Q_N), we still observe variance reduction.

$2N$	$W_N^\star$	$\partial_{\xi_1} W_N^\star$	$\partial_{\xi_2} W_N^\star$	$\partial^2_{\xi_1\xi_1} W_N^\star$	$\partial^2_{\xi_1\xi_2} W_N^\star$	$\partial^2_{\xi_2\xi_2} W_N^\star$	$\xi\cdot\partial_\xi W_N^\star$	$\xi^T\partial^2_\xi W_N^\star \xi$
10	19.41	11.26	13.86	9.846	5.966	13.34	19.39	19.41
20	22.82	11.89	13.03	9.865	7.306	9.096	22.77	22.83
40	18.08	11.82	9.816	9.576	5.904	8.831	18.03	18.11
60	21.26	12.89	12.98	10.57	7.247	10.73	21.24	21.28
80	12.36	8.798	9.050	10.05	4.316	8.454	12.31	12.37
100	11.88	9.856	8.412	11.10	3.775	10.24	11.82	11.88
200	13.60	8.261	11.52	8.057	4.636	12.62	13.54	13.61

Table 2.1 – Test-Case 1: Variance reduction ratios (2.58).

Remark 2.15. *Similar variance reduction ratios are obtained in the case when the corrector problem is supplemented with homogeneous Dirichlet boundary conditions on the boundary on Q_N (in the spirit of (2.7)), rather than periodic boundary conditions as used here following (2.9) (results not shown).*

2.3.4 Test Case 2

We now consider a test-case for which the energy density is not positively homogeneous. From our results of Section 2.2.1, we know that our approach yields variance reduction for the estimation of $\mathbb{E}\left[W_N^\star(\cdot,\xi)\right]$. Our aim here is two-fold: we first quantify the efficiency gain, and we next verify (and this will indeed be the case) that we also obtain a gain in efficiency for quantities of interest (such as the first or second derivatives of $W_N^\star(\omega,\xi)$ with respect to ξ) for which we do not have theoretical results in the two-dimensional case.

We show on Figure 2.3 the variances (2.57) of the same quantities of interest as on Figures 2.1 and 2.2 (obtained either with the Monte Carlo approach or the Antithetic Variable approach). As for the previous test-case, we observe that all variances decrease at the rate $1/|Q_N|$ as N increases. In addition, we observe that the variance obtained with our approach is systematically smaller than the Monte Carlo variance, in the sense that $V_{\rm AV} \leq V_{\rm MC}$.

On Table 2.2, we report the variance reduction ratios (2.58) (with the same convention as in Table 2.1). We observe an efficiency gain of more than 10 for all quantities of interest, except again the cross derivative $\partial^2_{\xi_1\xi_2} W_N^\star$, for which the gain is smaller, and of the order of 4.

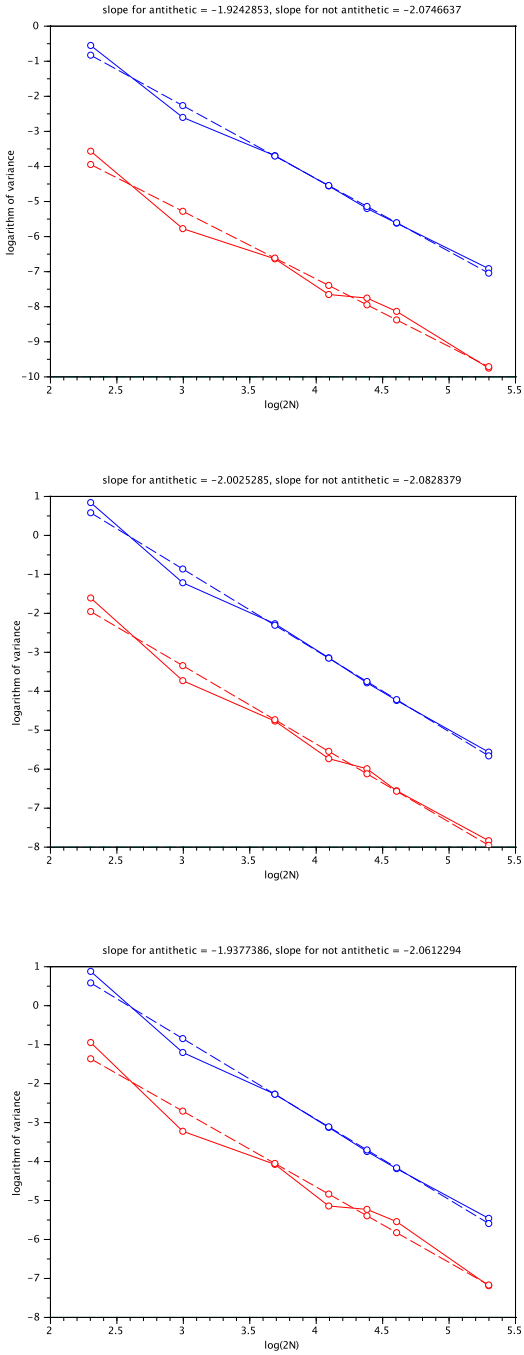

Figure 2.3 – Test-Case 2: Variances (2.57) as a function of $2N$ (Blue: Monte Carlo approach; Red: Antithetic Variable approach; Natural logarithm plot). Solid line: actual results. Dashed line: linear regression fit. The quantities of interest are the same as on Figure 2.1.

$2N$	$W_N^\star$	$\partial_{\xi_1} W_N^\star$	$\partial_{\xi_2} W_N^\star$	$\partial^2_{\xi_1\xi_1} W_N^\star$	$\partial^2_{\xi_1\xi_2} W_N^\star$	$\partial^2_{\xi_2\xi_2} W_N^\star$	$\xi \cdot \partial_\xi W_N^\star$	$\xi^T \partial^2_\xi W_N^\star \xi$
10	20.38	11.57	14.14	9.940	6.206	13.28	19.89	19.57
20	23.86	12.34	13.32	9.993	7.548	9.265	23.33	23.00
40	18.94	12.16	10.16	9.726	6.060	8.902	18.50	18.24
60	22.11	13.30	13.35	10.73	7.513	10.88	21.68	21.41
80	12.89	9.080	9.295	10.09	4.420	8.598	12.61	12.45
100	12.37	10.17	8.635	11.21	3.896	10.24	12.12	11.96
200	17.07	9.708	9.864	7.731	5.631	8.284	16.71	16.49

Table 2.2 – Test-Case 2: Variance reduction ratios (2.58).

2.3.5 Test Case 3

We eventually turn to our final test-case, where both coefficients a and c do depend on the space variable.

We show on Figure 2.4 the variances (2.57). Again, we observe that they all decrease at the rate $1/|Q_N|$ as N increases, and that the variance obtained with our approach is systematically smaller than the Monte Carlo variance.

On Table 2.3, we report the variance reduction ratios (2.58) (with the same convention as in Table 2.1). Results are quantitatively similar to the ones obtained on Table 2.2: we do observe a robust variance reduction, even in cases for which theoretical support is still currently missing.

$2N$	$W_N^\star$	$\partial_{\xi_1} W_N^\star$	$\partial_{\xi_2} W_N^\star$	$\partial^2_{\xi_1\xi_1} W_N^\star$	$\partial^2_{\xi_1\xi_2} W_N^\star$	$\partial^2_{\xi_2\xi_2} W_N^\star$	$\xi \cdot \partial_\xi W_N^\star$	$\xi^T \partial^2_\xi W_N^\star \xi$
10	14.26	12.69	10.00	12.38	8.333	10.65	14.76	19.37
20	10.82	8.166	7.669	8.304	7.730	8.827	11.29	18.11
40	7.014	7.077	5.613	10.28	6.776	7.310	7.731	14.32
60	10.45	10.84	8.666	11.72	8.896	9.524	11.82	19.01
80	6.961	5.880	7.250	8.800	4.646	8.996	7.522	11.10
100	8.543	6.780	7.970	8.873	4.669	10.26	8.798	11.66
200	7.589	7.362	6.816	9.457	5.373	9.328	8.391	13.14

Table 2.3 – Test Case 3: Variance reduction ratios (2.58).

Acknowledgments

The work of FL and WM is partially supported by ONR under Grant N00014-09-1-0470. WM gratefully acknowledges the support from Labex MMCD (Multi-Scale Modelling & Experimentation of Materials for Sustainable Construction) under contract ANR-11-LABX-0022. We also wish to thank Claude Le Bris for enlightning discussions.

2.4 Appendix: Proof of (2.8)

In all what follows, $\xi \in \mathbb{R}^d$ is fixed. We introduce the quantity

$$W_N^{\star,D}(\omega,\xi) := \inf\left\{ \frac{1}{|Q_N|} \int_{Q_N} W\left(y,\omega,\xi+\nabla w(y)\right)\, dy,\ w \in W_0^{1,p}(Q_N) \right\}$$

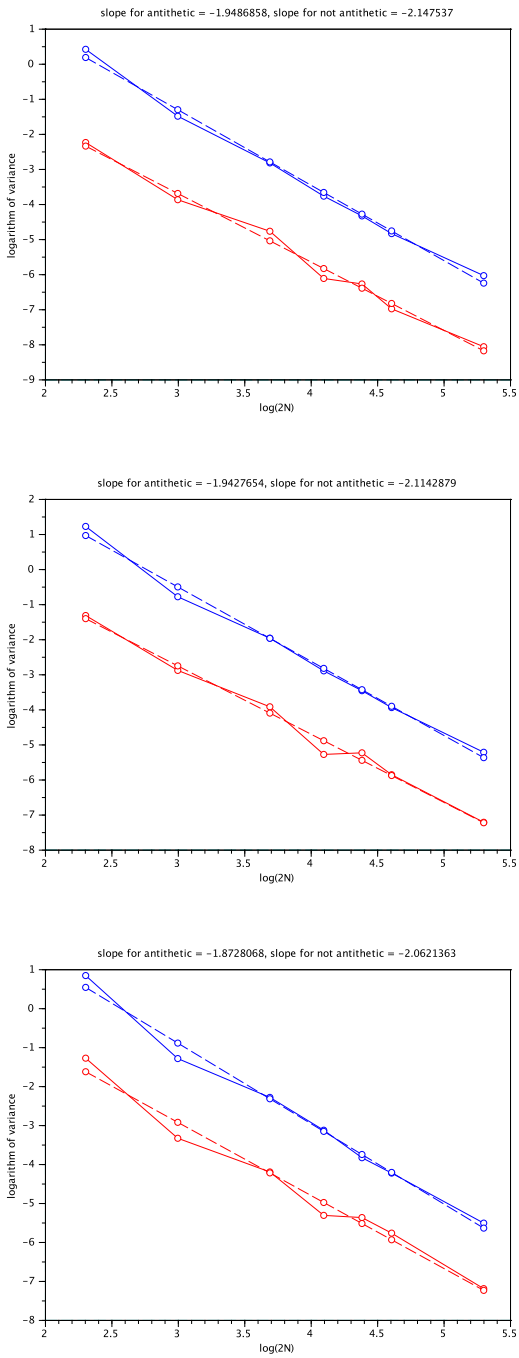

Figure 2.4 – Test-Case 3: Variances (2.57) as a function of $2N$ (Blue: Monte Carlo approach; Red: Antithetic Variable approach; Natural logarithm plot). Solid line: actual results. Dashed line: linear regression fit. The quantities of interest are the same as on Figure 2.1.

where we emphasize in the notation that we work with homogeneous Dirichlet boundary conditions. We know from [DMM86a, DMM86b] (see also [GN11, Theorem 3.1]) that $W_N^{\star,D}(\omega,\xi)$ almost surely converges to a deterministic limit, and that this limit is the homogenized energy density:

$$W^\star(\xi) = \lim_{N\to\infty} W_N^{\star,D}(\omega,\xi) \quad \text{a.s.} \tag{2.59}$$

We now minimize the same functional in the space of functions satisfying periodic boundary conditions rather than homogeneous Dirichlet boundary conditions: following (2.9), we introduce

$$W_N^\star(\omega,\xi) := \inf\left\{\frac{1}{|Q_N|}\int_{Q_N} W\left(y,\omega,\xi+\nabla w(y)\right)\,dy, \quad w\in W^{1,p}_\#(Q_N)\right\}. \tag{2.60}$$

We see that $W_0^{1,p}(Q_N)\subset W^{1,p}_\#(Q_N)$, hence $W_N^\star(\omega,\xi)\le W_N^{\star,D}(\omega,\xi)$. Passing to the limit $N\to\infty$ and using (2.59), we have

$$\limsup_{N\to\infty} W_N^\star(\omega,\xi) \le W^\star(\xi) \quad \text{a.s.} \tag{2.61}$$

We prove below the following result, claimed in (2.8):

Lemma 2.16. *Under the assumptions of Section 2.1.1, we have*

$$\lim_{N\to\infty} W_N^\star(\omega,\xi) = W^\star(\xi) \quad \textit{a.s.} \tag{2.62}$$

The proof goes as follows. Let w_N be the minimizer of (2.60) of mean zero:

$$W_N^\star(\omega,\xi) = \frac{1}{|Q_N|}\int_{Q_N} W\left(y,\omega,\xi+\nabla w_N(y,\omega)\right)\,dy, \qquad \int_{Q_N} w_N(y,\omega)=0.$$

We introduce

$$w_0^N(x,\omega) := \frac{1}{N} w^N(Nx,\omega)$$

and recognize that

$$w_0^N(\cdot,\omega) = \arg\inf\left\{\int_Q W\left(Nx,\omega,\xi+\nabla\theta(x)\right)\,dx,\ \theta\in W^{1,p}_\#(Q),\ \int_Q \theta = 0\right\} \tag{2.63}$$

and

$$W_N^\star(\omega,\xi) = \int_Q W\left(Nx,\omega,\xi+\nabla w_N^0(x,\omega)\right)\,dx. \tag{2.64}$$

Using the test function $\theta\equiv 0$ in (2.63) and using (2.4), we obtain

$$c_1\int_Q |\xi+\nabla w_0^N|^p \le W_N^\star(\omega,\xi) \le \int_Q W(Nx,\omega,\xi) \le c_2(1+|\xi|^p).$$

Hence there exists a constant C independent of N and ω such that $\|\nabla w_0^N(\cdot,\omega)\|_{L^p(Q)}\le C$. A standard Poincaré-Wirtinger argument yields $\|w_0^N(\cdot,\omega)\|_{W^{1,p}(Q)}\le C$. Up to the extraction of a subsequence, we thus have

$$w_0^N(\cdot,\omega) \rightharpoonup w_0^\infty(\cdot,\omega) \quad \text{in} \quad W^{1,p}(Q) \quad \text{almost surely}$$

where, almost surely, $w_0^\infty(\cdot,\omega) \in W^{1,p}_\#(Q)$ and satisfies $\displaystyle\int_Q w_0^\infty(\cdot,\omega) = 0$.

We are now in position to apply [JKO94, Lemma 15.3 p. 444], from which we deduce that

$$\liminf_{N\to\infty} \int_Q W(N\cdot,\omega,\xi+\nabla w_0^N) \geq \int_Q W^\star(\xi+\nabla w_0^\infty). \tag{2.65}$$

Using Jensen inequality and (2.64), we infer from (2.65) that

$$\liminf_{N\to\infty} W_N^\star(\omega,\xi) \geq W^\star\left(\int_Q \xi+\nabla w_0^\infty\right) = W^\star(\xi). \tag{2.66}$$

Collecting (2.61) and (2.66), we get

$$W^\star(\xi) \leq \liminf_{N\to\infty} W_N^\star(\omega,\xi) \leq \limsup_{N\to\infty} W_N^\star(\omega,\xi) \leq W^\star(\xi).$$

This concludes the proof of Lemma 2.16.

Chapter 3

A control variate approach based on a defect-type theory for variance reduction in stochastic homogenization

Ce **Chapitre** reprend l'intégralité d'un article publié dans le journal *Multiscale Modeling & Simulation* [LM15a], et écrit en collaboration avec Frédéric Legoll.

Par la méthode des variables de contrôle, nous construisons un estimateur de la matrice homogénéisée (pour un problème d'homogénéisation aléatoire) de moindre variance. Pour construire notre estimateur, nous exploitons certains éléments d'une théorie perturbative, que nous utilisons en tant que variable de contrôle, et non dans le régime perturbatif. Nous utilisons pour calculer notre variable de contrôle une méthode de bases réduites. L'efficacité de l'approche est illustrée par des expérimentations numériques en dimension 2. De plus, l'approche est analysée dans certains cas simples.

A control variate approach based on a defect-type theory for variance reduction in stochastic homogenization

Frédéric Legoll[1,3] and William Minvielle[2,3]

legoll@lami.enpc.fr, william.minvielle@cermics.enpc.fr

[1] Laboratoire Navier, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal,
77455 Marne-La-Vallée Cedex 2, France,

[2] CERMICS, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2, France;

[3] INRIA Rocquencourt, MATHERIALS research-team, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France.

Abstract. *We consider a variance reduction approach for the stochastic homogenization of divergence form linear elliptic problems. Although the exact homogenized coefficients are deterministic, their practical approximations are random. We introduce a control variate technique to reduce the variance of the computed approximations of the homogenized coefficients. Our approach is based on a surrogate model inspired by a defect-type theory, where a perfect periodic material is perturbed by rare defects. This model has been introduced in [ALB10] in the context of weakly random models. In this work, we address the fully random case, and show that the perturbative approaches proposed in [ALB10, ALB11] can be turned into an efficient control variable.*

We theoretically demonstrate the efficiency of our approach in simple cases. We next provide illustrating numerical results and compare our approach with other variance reduction strategies. We also show how to use the Reduced Basis approach proposed in [LBT12] so that the cost of building the surrogate model remains limited.

3.1 Introduction

In this work, we introduce a variance reduction approach based on the control variate technique for the homogenization of the following stochastic, elliptic, linear problem:

$$-\operatorname{div}\left(A\left(\frac{x}{\varepsilon},\omega\right)\nabla u^\varepsilon\right) = f \quad \text{in } \mathcal{D}, \qquad u^\varepsilon(\cdot,\omega) = 0 \quad \text{on } \partial\mathcal{D}, \tag{3.1}$$

set on a bounded domain $\mathcal{D}$ in $\mathbb{R}^d$, where f is a deterministic function in $L^2(\mathcal{D})$. The random matrix A is assumed to be uniformly elliptic, bounded and stationary in a sense made precise below.

It is well-known that, in the limit when ε goes to 0, the above problem converges to the homogenized problem

$$-\operatorname{div}\left(A^\star\nabla u^\star\right) = f \quad \text{in } \mathcal{D}, \qquad u^\star = 0 \quad \text{on } \partial\mathcal{D}, \tag{3.2}$$

where the homogenized matrix $A^\star$ is deterministic, and given by an expectation of an integral involving the so-called corrector function, that solves a random auxiliary problem set

on the *entire* space. In practice, the corrector problem is approximated by a problem set on a *bounded* domain Q_N (see Section 3.1.2 below for details). A by-product of this truncation procedure is that the *deterministic* matrix $A^\star$ is in practice approximated by a *random*, apparent homogenized matrix $A^\star_N(\omega)$. Randomness therefore comes again into the picture. In this work, we introduce a variance reduction approach to obtain practical approximations of $A^\star$ with a smaller variance. Our approach is a control variate technique, which is based on a surrogate random model, simple enough to allow for easier computations, and close enough to the reference model to eventually improve the accuracy.

We mention that, in our previous works [BCLBL12b, BCLBL12a, CLBL10], we have already proposed variance reduction approaches to compute better approximations of $A^\star$. We used there the technique of antithetic variables, which is a generic variance reduction approach. In addition, we have shown in [LM15b] that this technique carries over to nonlinear stochastic homogenization problems, when the problem at hand is formulated as a variational convex problem. In this work, we return to the linear equation (3.1), and design an approach based on the control variate technique, where a surrogate model is used to improve the computational efficiency. Our approach here is therefore much more specific to the problem at hand than the antithetic variable approaches proposed previously. We therefore expect this technique to provide better results. This is indeed the case, as discussed along the numerical examples of Section 3.5.1.

Generally speaking, control variate approaches are based on using surrogate models as a kind of preconditioner (see Section 3.1.3 below for more details). In this work, the surrogate model that we use is inspired by a defect-type model, introduced in [ALB10, ALB12, ALB11] in the context of weakly random models. The model considered there is that of a perfect periodic material perturbed by rare defects. These defects may introduce a significant change in the local properties of the random matrix $A(x,\omega)$. However they only occur with a small probability η. In that setting, when η is small, the authors of [ALB10, ALB12, ALB11] have shown that a good approximation of the homogenized properties can be obtained by only solving deterministic problems rather than random problems, as usually required in stochastic homogenization. In this work, we build our surrogate model upon the ideas of [ALB10, ALB12, ALB11]. However, we address the regime when η is not small, hence perturbative approaches are not accurate enough.

Our article is organized as follows. In the sequel of this introduction, we present in more details some basic elements of stochastic homogenization, situate the questions under consideration in a more general setting, and introduce the control variate approach in a general setting (see Section 3.1.3). In Section 3.2, we recall the weakly stochastic model introduced in [ALB10, ALB12, ALB11].

Next, in Section 3.3, we describe how to use this weakly stochastic model to build surrogate models that can be used in the "fully random" (non perturbative) regime. We introduce two control variate approaches. The first approach (see Section 3.3.1) is based on a first-order weakly stochastic approach, where defects are considered as *isolated* from one another. The second one (see Section 3.3.2) is based on a second-order weakly stochastic approach, where *pairs of defects* are considered. The main qualitative difference between these two control variate approaches is that the second one takes into account the geometry, whereas the first one essentially only depends on $\displaystyle\int_{Q_N} A(x,\omega)\,dx$. It is well known that, in dimension $d \geq 2$, geometry – i.e. the way different materials are located one with respect to the other – matters in the homogenization process. The fact that our second approach takes into account the geometry is thus a very interesting feature.

We next collect in Section 3.4 some elements of theoretical analysis. We first consider

the one-dimensional case (Section 3.4.1) and provide there a complete analysis of our approach (see Propositions 3.11 and 3.13). We show that the variance of the apparent homogenized coefficient scales as N^{-1} (where N is the size of the large domain on which, in practice, the corrector problem is solved), while it is decreased to N^{-2} (resp. N^{-3}) when using our first-order (resp. second-order) control variate approach. In Section 3.4.2, we next turn to the multi-dimensional case. Our main result is Lemma 3.14.

Section 3.5 is devoted to numerical experiments. We quantitatively demonstrate the efficiency of our approach on two test cases in Sections 3.5.1 and 3.5.2. As pointed out above, our second approach is based on considering pairs of defects. In order to keep limited the offline cost associated to building the surrogate model, we show in Section 3.5.3 that it is possible to use the Reduced Basis approach introduced in [LBT12]: the precomputation cost is then dramatically decreased, while the gain in variance with respect to a Monte Carlo approach remains similar.

3.1.1 Homogenization theoretical setting

To begin with, we introduce the basic setting of stochastic homogenization we employ. We refer to [PV81] for some seminal contribution, to [ES08] for a general, numerically oriented presentation, and to [BLP78, CD99, JKO94] for classical textbooks. We also refer to [LB10] and the review article [ACLB+12] (and the extensive bibliography contained therein) for a presentation of our particular setting. Throughout this article, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ the expectation of any random variable $X \in L^1(\Omega, d\mathbb{P})$. We next fix $d \in \mathbb{N}^\star$ (the ambient physical dimension), and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, that is, for all $k \in \mathbb{Z}^d$ and all $A \in \mathcal{F}$, $\mathbb{P}(\tau_k A) = \mathbb{P}(A)$. We assume that the action τ is *ergodic*, that is, if $A \in \mathcal{F}$ is such that $\tau_k A = A$ for any $k \in \mathbb{Z}^d$, then $\mathbb{P}(A) = 0$ or 1. In addition, we define the following notion of stationarity (see [BLBL06, BLBL07]): a function $F \in L^1_{\text{loc}}\left(\mathbb{R}^d, L^1(\Omega)\right)$ is *stationary* if

$$\forall k \in \mathbb{Z}^d, \quad F(x+k, \omega) = F(x, \tau_k \omega) \quad \text{a.e. in } x \text{ and a.s.} \tag{3.3}$$

In this setting, the ergodic theorem [Kre85, Shi84, Tem72] can be stated as follows: *Let $F \in L^\infty\left(\mathbb{R}^d, L^1(\Omega)\right)$ be a stationary random variable in the above sense. For $k = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d$, we set $|k|_\infty = \sup_{1\le i\le d} |k_i|$. Then*

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

This implies (denoting by Q the unit cube in $\mathbb{R}^d$) that

$$F\left(\frac{x}{\varepsilon}, \omega\right) \underset{\varepsilon\to 0}{\overset{*}{\rightharpoonup}} \mathbb{E}\left(\int_Q F(x, \cdot) dx\right) \quad \textit{in } L^\infty(\mathbb{R}^d), \ \textit{almost surely.}$$

Besides technicalities, the purpose of the above setting is simply to formalize that, even though realizations may vary, the function F at point $x \in \mathbb{R}^d$ and the function F at point $x + k$, $k \in \mathbb{Z}^d$, share the same law. In the homogenization context we now turn to, this means that the local, microscopic environment (encoded in the matrix field A in (3.1)) is everywhere the same *on average*. From this, homogenized, macroscopic properties will follow. In addition, and this is evident reading the above setting, the

microscopic environment has a relation to an underlying *periodic* structure (thus the integer shifts k in (3.3)).

We consider problem (3.1), where $\mathcal{D}$ is an open, bounded domain of $\mathbb{R}^d$ and where $f \in L^2(\mathcal{D})$ is deterministic. The random matrix A is assumed stationary in the sense of (3.3). We also assume that A is bounded and that, in the sense of quadratic forms, A is positive and almost surely bounded away from zero: there exist deterministic constants c and C such that, almost surely,

$$\|A(\cdot,\omega)\|_{L^\infty(\mathbb{R}^d)} \leq C \quad \text{and} \quad \forall \xi \in \mathbb{R}^d, \quad \xi^T A(x,\omega)\xi \geq c\xi^T\xi \quad \text{a.e.} \tag{3.4}$$

In this specific setting, the solution $u^\varepsilon(\cdot,\omega)$ to (3.1) converges (when ε goes to 0) to the solution $u^\star$ to the homogenized problem (3.2) almost surely, weakly in $H^1(\mathcal{D})$ and strongly in $L^2(\mathcal{D})$. The homogenized matrix $A^\star$ that appears in (3.2) reads

$$\forall p \in \mathbb{R}^d, \quad A^\star p = \mathbb{E}\left[\int_Q A(x,\cdot)\,(\nabla w_p(x,\cdot) + p)\,dx\right], \qquad Q = (0,1)^d, \tag{3.5}$$

where, for any vector $p \in \mathbb{R}^d$, the *corrector* w_p is the solution (unique up to the addition of a random constant) to the following corrector problem:

$$\begin{cases} -\operatorname{div}\left[A(\nabla w_p + p)\right] = 0 \quad \text{in } \mathbb{R}^d \ \text{ a.s.}, \\ \nabla w_p \text{ is stationary in the sense of (3.3)}, \qquad \displaystyle\int_Q \mathbb{E}(\nabla w_p) = 0. \end{cases} \tag{3.6}$$

3.1.2 Practical approximation of the homogenized matrix

The corrector problem (3.6) is set on the entire space $\mathbb{R}^d$, and is therefore challenging to solve. Approximations are in order. In practice, the deterministic matrix $A^\star$ is approximated by the random matrix $A^\star_N(\omega)$ defined by

$$\forall p \in \mathbb{R}^d, \quad A^\star_N(\omega)\,p = \frac{1}{|Q_N|}\int_{Q_N} A(x,\omega)\left(p + \nabla w_p^N(x,\omega)\right)\,dx, \tag{3.7}$$

which is obtained by solving the corrector problem on a *truncated* domain, say the cube $Q_N = (-N/2, N/2)^d$:

$$-\operatorname{div}\left(A(\cdot,\omega)\left(p + \nabla w_p^N(\cdot,\omega)\right)\right) = 0, \qquad w_p^N(\cdot,\omega) \text{ is } Q_N\text{-periodic}. \tag{3.8}$$

As briefly explained above, although $A^\star$ itself is a deterministic object, its practical approximation $A^\star_N$ is random. It is only in the limit of infinitely large domains Q_N that the deterministic value is attained. Indeed, as shown in [BP04], we have

$$\lim_{N\to\infty} A^\star_N(\omega) = A^\star \quad \text{almost surely.}$$

Many studies have been recently devoted to establishing sharp estimates on the convergence of the random apparent homogenized quantities (computed on Q_N) to the exact deterministic homogenized quantities. We refer e.g. to [BP04, GNO14, Nol14, Yur86] and to the comprehensive discussion of [BCLBL12b, Section 1.2]. We take here the problem from a slightly different perspective. We observe that the error

$$A^\star - A^\star_N(\omega) = \left(A^\star - \mathbb{E}\left[A^\star_N\right]\right) + \left(\mathbb{E}\left[A^\star_N\right] - A^\star_N(\omega)\right)$$

is the sum of a systematic error and of a statistical error (the first and second terms in the above right-hand side, respectively). We focus here on the *statistical error*, and propose approaches to reduce the confidence interval of empirical means approximating $\mathbb{E}[A_N^\star]$, for a given truncated domain Q_N. Optimal estimates on the variance of $A_N^\star$ have been established in [Nol14, Theorem 1.3 and Proposition 1.4]. For a setting slightly different from ours (namely for homogenization problems set on random *lattices*), optimal estimates on the systematic and statistical errors have been established in [GNO14, Theorem 2]. The authors noted there that "the systematic error is much smaller than the statistical error", in the sense that the latter decays with a slower rate with respect to N than the former. For large values of N, the statistical error (that we address in this work) is therefore dominating over the systematic error.

A standard technique to compute an approximation of $\mathbb{E}\left[(A_N^\star)_{ij}\right]$ (for any entry ij) is to consider M independent and identically distributed realizations of the field A, solve for each of them the corrector problem (3.8) (thereby obtaining i.i.d. realizations $A_N^{\star,m}(\omega)$) and proceed following a Monte Carlo approach:

$$\mathbb{E}\left[(A_N^\star)_{ij}\right] \approx I_M^{\mathrm{MC}} := \frac{1}{M}\sum_{m=1}^{M}\left(A_N^{\star,m}(\omega)\right)_{ij}. \tag{3.9}$$

In view of the Central Limit Theorem, we know that our quantity of interest $\mathbb{E}\left[(A_N^\star)_{ij}\right]$ asymptotically lies in the confidence interval

$$\left[I_M^{\mathrm{MC}} - 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[(A_N^\star)_{ij}\right]}}{\sqrt{M}}, I_M^{\mathrm{MC}} + 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[(A_N^\star)_{ij}\right]}}{\sqrt{M}}\right]$$

with a probability equal to 95 %.

In this article, we show that, using a control variate approach, we can design a practical approach that, for any finite N, allows to compute a better approximation of $\mathbb{E}\left[(A_N^\star)_{ij}\right]$ than I_M^{MC}. Otherwise stated, for an equal computational cost, we obtain a more accurate (i.e. with a smaller confidence interval) approximation.

3.1.3 Control variate approach

Before presenting our specific approach, we describe here the control variate approach in a general context (see [Fis96, page 277]). Consider a general probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a scalar random variable $X \in L^2(\Omega, \mathbb{R})$. Our aim is to compute its expectation $\mathbb{E}(X)$. In the sequel, we will use that approach for the random variable $(A_N^\star(\omega))_{ij}$, for any entry $1 \le i, j \le d$.

As always, a first possibility is to resort to M i.i.d. realizations of X, denoted $X^m(\omega)$ for $1 \le m \le M$. The expectation is then approximated by the Monte Carlo empirical mean

$$I_M^{\mathrm{MC}} := \frac{1}{M}\sum_{m=1}^{M} X^m(\omega)$$

and we know that, with a probability equal to 95 %, $\mathbb{E}[X]$ asymptotically lies in the confidence interval

$$\left[I_M^{\mathrm{MC}} - 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}[X]}}{\sqrt{M}}, I_M^{\mathrm{MC}} + 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}[X]}}{\sqrt{M}}\right]. \tag{3.10}$$

To reduce the variance of the estimation, consider now a random variable $Y \in L^2(\Omega, \mathbb{R})$, the expectation of which is analytically known. Then, for any scalar deterministic parameter ρ to be fixed later, we consider the *controlled variable*

$$D_\rho(\omega) = X(\omega) - \rho\Big(Y(\omega) - \mathbb{E}[Y]\Big). \tag{3.11}$$

Since $\mathbb{E}[Y]$ is known exactly, sampling realizations of D_ρ amounts to sampling realizations of X and Y. We obviously have $\mathbb{E}[D_\rho] = \mathbb{E}[X]$. To approximate $\mathbb{E}[X]$, the control variate approach consists in performing a standard Monte Carlo approximation on D_ρ. We hence consider M i.i.d. realizations of D_ρ, denoted $D_\rho^m(\omega)$, introduce the empirical mean

$$I_M^{\mathrm{CV}} := \frac{1}{M}\sum_{m=1}^{M} D_\rho^m(\omega)$$

and write that, with a probability equal to 95 %, $\mathbb{E}[D_\rho] = \mathbb{E}\left[X\right]$ asymptotically lies in the confidence interval

$$\left[I_M^{\mathrm{CV}} - 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[D_\rho\right]}}{\sqrt{M}}, I_M^{\mathrm{CV}} + 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[D_\rho\right]}}{\sqrt{M}}\right]. \tag{3.12}$$

If ρ and Y are such that $\mathbb{V}\mathrm{ar}\left[D_\rho\right] < \mathbb{V}\mathrm{ar}\left[X\right]$, then the width of the above confidence interval is smaller than that of (3.10), and hence we have built a more accurate approximation of $\mathbb{E}\left[X\right]$.

We now detail how to choose ρ and Y in (3.11). Suppose for now that Y is given. We wish to pick ρ such that the variance of D_ρ is minimal. Writing that

$$\mathbb{V}\mathrm{ar}[D_\rho] = \mathbb{V}\mathrm{ar}[X] - 2\rho\mathbb{C}\mathrm{ov}[X, Y] + \rho^2\mathbb{V}\mathrm{ar}[Y],$$

we see that the optimal value of ρ reads

$$\rho^\star = \mathrm{argmin}\ \mathbb{V}\mathrm{ar}[D_\rho] = \frac{\mathbb{C}\mathrm{ov}[X, Y]}{\mathbb{V}\mathrm{ar}[Y]}. \tag{3.13}$$

For this choice, we have, using the Cauchy-Schwarz inequality,

$$\mathbb{V}\mathrm{ar}[D_{\rho^\star}] = \mathbb{V}\mathrm{ar}[X]\left(1 - \frac{(\mathbb{C}\mathrm{ov}[X, Y])^2}{\mathbb{V}\mathrm{ar}[X]\mathbb{V}\mathrm{ar}[Y]}\right) \leq \mathbb{V}\mathrm{ar}[X].$$

We thus observe that, for any choice of Y, we can choose ρ such that the variance of D_ρ is indeed smaller than that of X. Of course, the ratio of variances $\dfrac{\mathbb{V}\mathrm{ar}[D_{\rho^\star}]}{\mathbb{V}\mathrm{ar}[X]}$, which is directly related to the gain in accuracy, depends on Y, and more precisely on the value of $\dfrac{(\mathbb{C}\mathrm{ov}[X, Y])^2}{\mathbb{V}\mathrm{ar}[X]\mathbb{V}\mathrm{ar}[Y]}$. The larger the correlation between X and Y, the better. In contrast to the choice of ρ, the choice of Y is problem dependent. In addition, the control variable Y needs to be *random*.

Remark 3.1. *In practice, we do not have access to the optimal value (3.13), which involves exact expectations. One possibility (which is the one we adopt in this work) is to replace (3.13) by the empirical estimator*

$$\rho^\star \approx \frac{\sum_{m=1}^{M}(X^m(\omega) - \mu_M(X))\,(Y^m(\omega) - \mathbb{E}[Y])}{\sum_{m=1}^{M}(Y^m(\omega) - \mathbb{E}[Y])^2},$$

where $\mu_M(X) = \frac{1}{M}\sum_{m=1}^{M} X^m(\omega)$. *This choice corresponds to minimizing with respect to* ρ *the empirical variance of* D_ρ *defined as* $\frac{1}{M}\sum_{m=1}^{M}\big(D_\rho^m(\omega) - \mu_M(X)\big)^2$, *where* $D_\rho^m(\omega) = X^m(\omega) - \rho\Big(Y^m(\omega) - \mathbb{E}[Y]\Big)$.

3.2 A weakly random setting: rare defects in a periodic structure

As pointed out above, the surrogate model that we use to build our controlled variable is inspired by a defect-type model, introduced in [ALB10, ALB12, ALB11] in the context of weakly random models, and that we describe now.

3.2.1 Presentation of the model

Assume that, in (3.1), the random matrix A is of the form

$$A(x,\omega) = A_\eta(x,\omega) = A_{\rm per}(x) + b_\eta(x,\omega)\Big(C_{\rm per}(x) - A_{\rm per}(x)\Big) \tag{3.14}$$

where $A_{\rm per}$ and $C_{\rm per}$ are $\mathbb{Z}^d$-periodic matrices that are bounded and positive in the sense of (3.4), and

$$b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbf{1}_{Q+k}(x) B_k^\eta(\omega), \tag{3.15}$$

where $(B_k^\eta)_{k\in\mathbb{Z}^d}$ are i.i.d. scalar random variables. The matrix A is indeed stationary in the sense of (3.3). We furthermore assume that B_k^η follows a Bernoulli law of parameter $\eta \in (0,1)$:

$$\mathbb{P}(B_k^\eta = 1) = \eta, \quad \mathbb{P}(B_k^\eta = 0) = 1-\eta. \tag{3.16}$$

The matrix $A(x,\omega)$ then satisfies assumption (3.4).

In each cell $Q+k$, the field A is equal to $A_{\rm per}$ with the probability $1-\eta$, and equal to $C_{\rm per}$ with the probability η. When η is small, then (3.14)–(3.15)–(3.16) models a periodic material (described by $A_{\rm per}$) that is randomly perturbed (and then described by $C_{\rm per}$). The perturbation is rare when η is small (therefore the material is described by $A_{\rm per}$ "most of the time"), and thus it can be considered as a defect. However, the perturbation is not small in L^∞ norm: $\|C_{\rm per} - A_{\rm per}\|_{L^\infty}$ is not assumed to be small. We refer to [ALB11] for practical examples motivating this framework.

On Fig. 3.1, we show two realizations of the field $A_\eta(x,\omega)$ (on the domain Q_N for $N = 20$) for some specific choices of $A_{\rm per}$ and $C_{\rm per}$ (see [ALB11, Fig. 4.2] for more details). On the right part of that figure, we set $\eta = 0.4$, which is close to the value $\eta = 1/2$, when defects are as frequent as non-defects.

Note that specifying $A_\eta(x,\omega)$ on Q_N simply amounts to specifying the values of $B_k^\eta(\omega)$ for all k such that $k + Q \subset Q_N$.

The above setting is actually quite general. Consider for instance a classical test-case, the random checkerboard case:

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

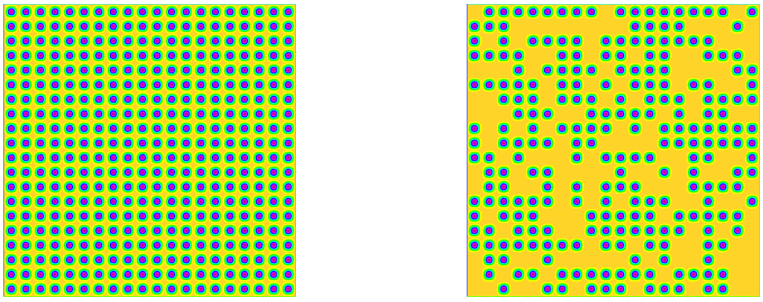

Figure 3.1 – Two instances of material (3.14). Left ($\eta = 0$): perfect material with circular inclusions located on a periodic network. Right ($\eta = 0.4$): perturbed material (each inclusion is deleted with a probability equal to 0.4). Courtesy A. Anantharaman and C. Le Bris.

where X_k are i.i.d. random variables satisfying $\mathbb{P}(X_k = \alpha) = \mathbb{P}(X_k = \beta) = 1/2$. This model falls into the framework (3.14)–(3.15)–(3.16) with

$$A_{\rm per} = \alpha\,{\rm Id}, \quad C_{\rm per} = \beta\,{\rm Id}, \quad \eta = 1/2.$$

An alternate choice (corresponding to choosing a different reference periodic materials) is

$$A_{\rm per} = \beta\,{\rm Id}, \quad C_{\rm per} = \alpha\,{\rm Id}, \quad \eta = 1/2.$$

In this work, we restrict our attention to the case (3.14)–(3.15)–(3.16), i.e. when B_k^η are i.i.d. Bernoulli random variables. This is the case specifically studied in [ALB11]. See [ALB10, ALB12] for more general settings.

3.2.2 Weakly-random homogenization result

Consider the model (3.14)–(3.15)–(3.16). The random variable $B_k^\eta(\omega)$ can take only two values, 0 or 1. Therefore, on the domain Q_N, there are only a finite number of realizations of $A_\eta(x,\omega)$. The realizations with the highest probability are as follows.

With probability $(1-\eta)^{|Q_N|}$, there are no defects in Q_N, and the realization actually corresponds to the perfect periodic situation. We introduce the periodic corrector w_p^0, solution to

$$-\,{\rm div}\left(A_{\rm per}\left(p + \nabla w_p^0\right)\right) = 0, \qquad w_p^0 \text{ is } Q\text{-periodic}, \tag{3.17}$$

and the associated matrix $A_{\rm per}^\star$, obtained by periodic homogenization:

$$\forall p \in \mathbb{R}^d, \quad A_{\rm per}^\star\, p = \int_Q A_{\rm per}\left(p + \nabla w_p^0\right). \tag{3.18}$$

With probability $\eta(1-\eta)^{|Q_N|-1}$, there is a unique defect in Q_N, located, say, in the cell $k+Q$ (see Fig. 3.2). Let us define

$$A_1^k = A_{\rm per} + 1_{k+Q}\Big(C_{\rm per} - A_{\rm per}\Big), \tag{3.19}$$

the associated corrector $w_p^{1,k,N}$, solution to

$$-\,{\rm div}\left(A_1^k\left(p + \nabla w_p^{1,k,N}\right)\right) = 0, \qquad w_p^{1,k,N} \text{ is } Q_N\text{-periodic}, \tag{3.20}$$

and the homogenized matrix $A^\star_{1,k,N}$, given by

$$\forall p \in \mathbb{R}^d, \quad A^\star_{1,k,N}\, p = \frac{1}{|Q_N|} \int_{Q_N} A_1^k \left(p + \nabla w_p^{1,k,N}\right). \tag{3.21}$$

With probability $\eta^2(1-\eta)^{|Q_N|-2}$, there are two defects in Q_N, located, say, in the cells $k+Q$ and $l+Q$ (see Fig. 3.2). Let us define

$$A_2^{k,l} = A_{\rm per} + \left(1_{k+Q} + 1_{l+Q}\right)\left(C_{\rm per} - A_{\rm per}\right), \tag{3.22}$$

the associated corrector $w_p^{2,k,l,N}$, solution to

$$-\operatorname{div}\left(A_2^{k,l}\left(p + \nabla w_p^{2,k,l,N}\right)\right) = 0, \qquad w_p^{2,k,l,N} \text{ is } Q_N\text{-periodic}, \tag{3.23}$$

and the homogenized matrix $A^\star_{2,k,l,N}$, given by

$$\forall p \in \mathbb{R}^d, \quad A^\star_{2,k,l,N}\, p = \frac{1}{|Q_N|} \int_{Q_N} A_2^{k,l} \left(p + \nabla w_p^{2,k,l,N}\right). \tag{3.24}$$

All the other configurations (with three defects or more) have a smaller probability.

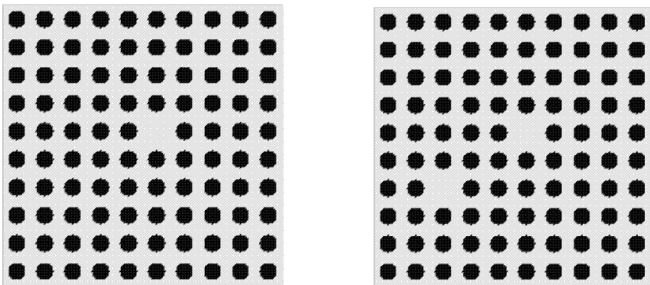

Figure 3.2 – Left: material modelled by A_1^k, with a single defect. Right: material modelled by $A_2^{k,l}$, with two defects (Courtesy A. Anantharaman and C. Le Bris).

Let us define

$$\mathcal{I}_N := \left\{k \in \mathbb{Z}^d;\ Q + k \subset Q_N\right\}.$$

As shown in [ALB11], we then have the following result:

Proposition 3.2 ([ALB11], Section 3.2)**.** *Let $A^\star_{\eta,N}(\omega)$ be the apparent homogenized matrix defined by (3.7), where $A \equiv A_\eta$ is given by (3.14)–(3.15)–(3.16). Then*

$$\mathbb{E}\left[A^\star_{\eta,N}\right] = A^\star_{\rm per} + \eta \overline{A}_1^N + \eta^2 \overline{A}_2^N + O_N(\eta^3), \tag{3.25}$$

where $O_N(\eta^3)$ is a quantity of the order of η^3 with a prefactor that may depend on N, $A^\star_{\rm per}$ is given by (3.18) and

$$\begin{aligned}
\overline{A}_1^N &= \sum_{k \in \mathcal{I}_N} \left(A^\star_{1,k,N} - A^\star_{\rm per}\right), \\
\overline{A}_2^N &= \frac{1}{2} \sum_{k,l \in \mathcal{I}_N, k \neq l} \left(A^\star_{2,k,l,N} - A^\star_{1,k,N} - A^\star_{1,l,N} + A^\star_{\rm per}\right).
\end{aligned}$$

We note that

$$\overline{A}_1^N = \sum_{k\in\mathcal{I}_N} \overline{A}_{1\,\mathrm{def}}^{k,N}, \quad \text{and} \quad \overline{A}_2^N = \frac{1}{2}\sum_{k\in\mathcal{I}_N}\sum_{l\in\mathcal{I}_N, l\neq k} \overline{A}_{2\,\mathrm{def}}^{k,l,N}, \tag{3.26}$$

where $\overline{A}_{1\,\mathrm{def}}^{k,N}$ (resp. $\overline{A}_{2\,\mathrm{def}}^{k,l,N}$) is the marginal contribution to the homogenized matrix from a configuration with a single defect in $k+Q$ (resp. two defects in $k+Q$ and $l+Q$):

$$\overline{A}_{1\,\mathrm{def}}^{k,N} = A^\star_{1,k,N} - A^\star_{\mathrm{per}}, \tag{3.27}$$

$$\overline{A}_{2\,\mathrm{def}}^{k,l,N} = A^\star_{2,k,l,N} - A^\star_{1,k,N} - A^\star_{1,l,N} + A^\star_{\mathrm{per}}. \tag{3.28}$$

Remark 3.3. *Passing to the limit $N\to\infty$ in (3.25) is not easy. We refer to [ALB11, Section 3.2] and [Mou15].*

When η is small, the advantage of (3.25) over the approach recalled in Section 3.1.2 is evident. Rather than solving the random problem (3.8) (for several realizations of A_η), it is enough to solve the deterministic problems (3.17), (3.20) and (3.23) to infer an accurate approximation of $\mathbb{E}\left[A^\star_{\eta,N}\right]$. We refer to [ALB11] for illustrative numerical results.

Furthermore, due to periodic boundary conditions (3.20), that are reminiscent of the periodic boundary conditions in (3.8), we have that

$$A^\star_{1,k,N} \text{ does not depend on } k. \tag{3.29}$$

Likewise, $A^\star_{2,k,l,N}$ depends only on $k-l$. Thus, there is only *one* problem (3.20) to be solved (say for $k=0$). Likewise, there are $|\mathcal{I}_N|-1$ problems (3.23) to be solved (say for $k=0$ and $l\neq 0$), and not $|\mathcal{I}_N|\,(|\mathcal{I}_N|-1)$. Noticing that (3.23) is a problem parameterized by l, the authors of [LBT12] have shown how to use a Reduced Basis approach to further speed-up the computation of $\overline{A}_2^N$. In practice, one can still obtain a good approximation of $\overline{A}_2^N$ without solving all the $|\mathcal{I}_N|-1$ problems (3.23). We return to this specific question in Section 3.5.3.

3.3 Control variate approaches for stochastic homogenization

We now introduce, for the model (3.14)–(3.15)–(3.16), a control variate approach. Our aim is now to address the regime when η is not close to 0 or 1 (the approximation (3.25) is therefore not accurate enough). Recall also that, in view of the discussion at the end of Section 3.1.3, we need a *random* surrogate model to build our controlled variable. In what follows, we first build an approximate model based on configurations with a single defect (see Section 3.3.1), and next turn to building a better approximate model that also uses configurations with two defects (see Section 3.3.2). As will be seen below, this second approximate model not only depends on the quantity of defects, but also on their geometry, that is on where the defects are located in Q_N.

3.3.1 A first-order model

Introduce

$$A_1^{\eta,N}(\omega) = \sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)\,\overline{A}_{1\,\mathrm{def}}^{k,N}, \tag{3.30}$$

where $\overline{A}_{1\,\mathrm{def}}^{k,N}$, defined by (3.27), is the marginal contribution to the homogenized matrix coming the configuration with a single defect located in $k+Q$. In view of (3.26), we notice that

$$\mathbb{E}\left[A_1^{\eta,N}\right] = \sum_{k\in\mathcal{I}_N} \mathbb{E}\left[B_k^\eta\right] \overline{A}_{1\,\mathrm{def}}^{k,N} = \eta \sum_{k\in\mathcal{I}_N} \overline{A}_{1\,\mathrm{def}}^{k,N} = \eta \overline{A}_1^N,$$

which is the first order correction in the expansion (3.25). When η is small, the *expectation* of $A_{\mathrm{per}}^\star + A_1^{\eta,N}(\omega)$ is a good approximation of the expectation of $A_{\eta,N}^\star(\omega)$, accurate up to an error of the order of η^2. The following observation provides additional motivation for our choice (3.30). It turns out that the *law* of the random variable $A_{\mathrm{per}}^\star + A_1^{\eta,N}(\omega)$ is a good approximation of that of $A_{\eta,N}^\star(\omega)$:

Lemma 3.4. *For any deterministic and continuous function φ, we have*

$$\mathbb{E}\left[\varphi\left(A_{\eta,N}^\star\right)\right] = \mathbb{E}\left[\varphi\left(A_{\mathrm{per}}^\star + A_1^{\eta,N}\right)\right] + O_N(\eta^2).$$

The proof of Lemma 3.4 is postponed until Section 3.4.2.

We thus think that $A_{\mathrm{per}}^\star + A_1^{\eta,N}(\omega)$ is a good surrogate model for $A_{\eta,N}^\star(\omega)$. As shown by Lemma 3.4, this is the case when $\eta \ll 1$, which is however not the regime we address. One-dimensional computations presented in Section 3.4.1 and numerical observations reported in Section 3.5 (for two-dimensional test-cases) confirm that it is indeed the case, even when η is not small.

Following (3.11), we now introduce our controlled variable as

$$\begin{aligned} D_\rho^{1,\eta}(\omega) &= A_{\eta,N}^\star(\omega) - \rho\left(A_{\mathrm{per}}^\star + A_1^{\eta,N}(\omega) - \mathbb{E}\left[A_{\mathrm{per}}^\star + A_1^{\eta,N}\right]\right) \\ &= A_{\eta,N}^\star(\omega) - \rho\left(A_1^{\eta,N}(\omega) - \eta\overline{A}_1^N\right). \end{aligned} \tag{3.31}$$

In view of (3.30), (3.27) and (3.29), we recast (3.31) as

$$D_\rho^{1,\eta}(\omega) = A_{\eta,N}^\star(\omega) - \rho\left[\left(\sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)\right) - \eta\left|\mathcal{I}_N\right|\right] \overline{A}_{1\,\mathrm{def}}^{0,N}. \tag{3.32}$$

Remark 3.5. *Note that, in (3.32), $A_{\eta,N}^\star(\omega)$ and $\sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)$ are correlated. Indeed, in practice, we start by drawing a realization of the random variables $B_\eta^k(\omega)$ for all $k \in \mathcal{I}_N$. This determines first $\sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)$, and second the field $A(x,\omega)$ on Q_N, from which we compute the associated $A_{\eta,N}^\star(\omega)$ following (3.7)–(3.8).*

Computing M realizations of $D_\rho^{1,\eta}(\omega)$ therefore amounts to:

- offline stage: determine $\overline{A}_{1\,\mathrm{def}}^{0,N}$ by solving the problem (3.17)–(3.18) on Q and solving only once the problem (3.20)–(3.21) on Q_N (say for $k=0$).
- online stage: solve M corrector problems (3.7)–(3.8) on Q_N (for M i.i.d. realizations of A on Q_N), and evaluate $D_\rho^{1,\eta}(\omega)$ according to (3.32).

Let $\mathcal{C}_N$ be the cost to solve a single corrector problem on Q_N. The Monte Carlo empirical estimator and the Control Variate empirical estimator, defined respectively by

$$I_M^{\mathrm{MC}} = \frac{1}{M}\sum_{m=1}^M A_{\eta,N}^{\star,m}(\omega) \quad \text{and} \quad I_M^{\mathrm{CV}} = \frac{1}{M}\sum_{m=1}^M D_\rho^{1,\eta,m}(\omega)$$

therefore share the same cost ($M\,\mathcal{C}_N$ for the former, $(1+M)\,\mathcal{C}_N$ for the latter). To minimize the variance of $D_\rho^{1,\eta}$, the parameter ρ in (3.31) is chosen following (3.13).

Notice that, in the above construction, we have considered as reference configuration the defect-free material, i.e. that for $\eta = 0$. Since, in the regime we focus on, η is not small, there is no reason to favor the defect-free configuration ($\eta = 0$) rather than the full defect configuration ($\eta = 1$), which corresponds to the periodic matrix $C_{\rm per}$. We therefore introduce (compare with (3.27))

$$\overline{C}^{k,N}_{1\,\rm def} = C^\star_{1,k,N} - C^\star_{\rm per},$$

where $C^\star_{1,k,N}$ is the homogenized matrix corresponding to a unique defect with respect to the periodic configuration $C_{\rm per}$ (compare with (3.19), (3.20) and (3.21)):

$$\forall p \in \mathbb{R}^d, \quad C^\star_{1,k,N}\,p = \frac{1}{|Q_N|}\int_{Q_N} C_1^k\left(p + \nabla v_p^{1,k,N}\right), \tag{3.33}$$

where, for any p, the corrector $v_p^{1,k,N}$ is a solution to

$$-\operatorname{div}\left(C_1^k\left(p + \nabla v_p^{1,k,N}\right)\right) = 0, \qquad v_p^{1,k,N} \text{ is } Q_N\text{-periodic},$$

where $C_1^k = C_{\rm per} - 1_{k+Q}\Big(C_{\rm per} - A_{\rm per}\Big)$. In the spirit of (3.32), we introduce the controlled variable

$$\widehat{D}^{1,\eta}_{\widehat{\rho}}(\omega) = A^\star_{\eta,N}(\omega) - \widehat{\rho}\left[\left(\sum_{k\in\mathcal{I}_N}\left(1 - B_k^\eta(\omega)\right)\right) - (1-\eta)\,|\mathcal{I}_N|\right]\overline{C}^{0,N}_{1\,\rm def},$$

that we recast as

$$\widehat{D}^{1,\eta}_{\widehat{\rho}}(\omega) = A^\star_{\eta,N}(\omega) + \widehat{\rho}\left[\left(\sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)\right) - \eta\,|\mathcal{I}_N|\right]\overline{C}^{0,N}_{1\,\rm def}.$$

Consider now any entry $1 \le i, j \le d$ of the homogenized matrix. Assuming that our control variate model is non trivial (i.e. that $\left[\overline{A}^{0,N}_{1\,\rm def}\right]_{ij} \neq 0$), we see that, for any deterministic $\widehat{\rho}$, there exists a deterministic parameter ρ such that $\left[\widehat{D}^{1,\eta}_{\widehat{\rho}}(\omega)\right]_{ij} = \left[D^{1,\eta}_\rho(\omega)\right]_{ij}$ a.s. Working with the controlled variable $D^{1,\eta}_\rho(\omega)$ is hence equivalent to working with the controlled variable $\widehat{D}^{1,\eta}_{\widehat{\rho}}(\omega)$. In the sequel, we only consider the former.

Remark 3.6. *The situation is different in the second order model, where taking $A_{\rm per}$ or $C_{\rm per}$ as reference is not equivalent. See Section 3.3.2 below.*

Remark 3.7. *In view of (3.32), we see that our first order control variable only depends on $\sum_{k\in\mathcal{I}_N} B_k^\eta(\omega)$, which is the number of defects in the material. This approach can thus be extended to any two-phase materials, say of the type $A(x,\omega) = A_1 + \chi(x,\omega)A_2$, where χ is stationary and equal to 0 or 1. In this case, the control variable reads $\int_{Q_N}\chi(x,\omega)\,dx$. We refer to [BL] for works in that direction.*

3.3.2 A second-order model

We now introduce a model that not only takes into account the contributions from single defects (through $\overline{A}^{k,N}_{1\,\mathrm{def}}$, see (3.30)) but also contributions from pairs of defects. To that aim, we introduce

$$A_2^{\eta,N}(\omega) = \frac{1}{2} \sum_{k\in\mathcal{I}_N} \sum_{l\in\mathcal{I}_N,\, l\neq k} B_k^\eta(\omega)\, B_l^\eta(\omega)\, \overline{A}^{k,l,N}_{2\,\mathrm{def}}\,, \tag{3.34}$$

where $\overline{A}^{k,l,N}_{2\,\mathrm{def}}$, defined by (3.28), is the marginal contribution to the homogenized matrix associated to the configuration with two defects located in $k+Q$ and $l+Q$. In view of (3.26), we notice that

$$\mathbb{E}\left[A_2^{\eta,N}\right] = \frac{1}{2} \sum_{k\in\mathcal{I}_N} \sum_{l\in\mathcal{I}_N,\, l\neq k} \mathbb{E}\left[B_k^\eta\, B_l^\eta\right] \overline{A}^{k,l,N}_{2\,\mathrm{def}} = \frac{\eta^2}{2} \sum_{k\in\mathcal{I}_N} \sum_{l\in\mathcal{I}_N,\, l\neq k} \overline{A}^{k,l,N}_{2\,\mathrm{def}} = \eta^2 \overline{A}_2^N\,,$$

which is the second order correction in the expansion (3.25). When η is small, the *expectation* of $A^\star_{\mathrm{per}} + A_1^{\eta,N}(\omega) + A_2^{\eta,N}(\omega)$ is a good approximation of the expectation of $A^\star_{\eta,N}(\omega)$, accurate up to an error of the order of η^3. Furthermore, we have the following result (compare with Lemma 3.4), the proof of which follows the same lines as that of Lemma 3.4 and is therefore omitted:

Lemma 3.8. *For any deterministic and continuous function φ, we have*

$$\mathbb{E}\left[\varphi\left(A^\star_{\eta,N}\right)\right] = \mathbb{E}\left[\varphi\left(A^\star_{\mathrm{per}} + A_1^{\eta,N} + A_2^{\eta,N}\right)\right] + O_N(\eta^3).$$

In a way similar to (3.31), we now introduce our second-order controlled variable as

$$D^{2,\eta}_{\rho_1,\rho_2}(\omega) = A^\star_{\eta,N}(\omega) - \rho_1\left(A_1^{\eta,N}(\omega) - \eta\overline{A}_1^N\right) - \rho_2\left(A_2^{\eta,N}(\omega) - \eta^2\overline{A}_2^N\right). \tag{3.35}$$

We have introduced two deterministic parameters ρ_1 and ρ_2, which need not be equal. For any choice of these parameters, we have $\mathbb{E}\left[D^{2,\eta}_{\rho_1,\rho_2}\right] = \mathbb{E}\left[A^\star_{\eta,N}\right]$.

To evaluate (3.35), we first have to precompute the deterministic matrices

$$\overline{A}^{k,N}_{1\,\mathrm{def}} = \overline{A}^{0,N}_{1\,\mathrm{def}} \quad \text{and} \quad \overline{A}^{k,l,N}_{2\,\mathrm{def}} = \overline{A}^{0,l-k,N}_{2\,\mathrm{def}}\,.$$

Computing M realizations of $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$ therefore amounts to:

- offline stage: (i) determine $\overline{A}^{0,N}_{1\,\mathrm{def}}$ by solving the problem (3.17)–(3.18) on Q and by solving only once the problem (3.20)–(3.21) on Q_N (say for $k=0$); (ii) determine $\overline{A}^{0,l,N}_{2\,\mathrm{def}}$ by solving $|\mathcal{I}_N|-1$ problems (3.23)–(3.24) on Q_N (for $k=0$ and $l\in\mathcal{I}_N$, $l\neq 0$).
- online stage: solve M corrector problems (3.7)–(3.8) on Q_N (for M i.i.d. realizations of A on Q_N), and evaluate $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$ according to (3.35).

Questions related to the cost for evaluating $\overline{A}^{0,l,N}_{2\,\mathrm{def}}$ are discussed at the end of this section.

As pointed out in Section 3.3.1, in our regime of interest, there is no reason to favor the defect-free configuration rather than the full defect configuration, which corresponds to the periodic matrix C_{per}. We have shown there that there is no use to introduce the terms

representing the first order correction with respect to $C_{\rm per}$. We therefore solely introduce the second order correction (compare with (3.28)):

$$\overline{C}^{k,l,N}_{2\,\rm def} = C^\star_{2,k,l,N} - C^\star_{1,k,N} - C^\star_{1,l,N} + C^\star_{\rm per}, \tag{3.36}$$

where $C^\star_{1,k,N}$ is defined by (3.33) and $C^\star_{2,k,l,N}$ is defined by (compare with (3.22), (3.23) and (3.24)):

$$\forall p \in \mathbb{R}^d, \quad C^\star_{2,k,l,N}\, p = \frac{1}{|Q_N|} \int_{Q_N} C_2^{k,l} \left(p + \nabla v_p^{2,k,l,N} \right), \tag{3.37}$$

where, for any $p \in \mathbb{R}^d$, the corrector $v_p^{2,k,l,N}$ is a solution to

$$-\operatorname{div}\left(C_2^{k,l} \left(p + \nabla v_p^{2,k,l,N} \right) \right) = 0, \qquad v_p^{2,k,l,N} \text{ is } Q_N\text{-periodic},$$

where $C_2^{k,l} = C_{\rm per} - \left(1_{k+Q} + 1_{l+Q} \right) \left(C_{\rm per} - A_{\rm per} \right)$. As in (3.34), we introduce

$$C_2^{\eta,N}(\omega) = \frac{1}{2} \sum_{k \in \mathcal{I}_N} \sum_{l \in \mathcal{I}_N,\, l \neq k} \left(1 - B_k^\eta(\omega) \right) \left(1 - B_l^\eta(\omega) \right) \overline{C}^{k,l,N}_{2\,\rm def}, \tag{3.38}$$

where $\overline{C}^{k,l,N}_{2\,\rm def}$ is defined by (3.36), and its expectation reads

$$\begin{aligned} \overline{C}_2^{\eta,N} := \mathbb{E}\left[C_2^{\eta,N} \right] &= \frac{1}{2} \sum_{k \in \mathcal{I}_N} \sum_{l \in \mathcal{I}_N,\, l \neq k} \mathbb{E}\left[(1 - B_k^\eta)\, (1 - B_l^\eta) \right] \overline{C}^{k,l,N}_{2\,\rm def} \\ &= \frac{1}{2} \sum_{k \in \mathcal{I}_N} \sum_{l \in \mathcal{I}_N,\, l \neq k} (1-\eta)^2\, \overline{C}^{k,l,N}_{2\,\rm def}. \end{aligned}$$

We eventually introduce the controlled variable (compare with (3.35))

$$\begin{aligned} D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega) = A^\star_{\eta,N}(\omega) &- \rho_1 \left(A_1^{\eta,N}(\omega) - \eta \overline{A}_1^N \right) \\ &- \rho_2 \left(A_2^{\eta,N}(\omega) - \eta^2 \overline{A}_2^N \right) - \rho_3 \left(C_2^{\eta,N}(\omega) - \overline{C}_2^{\eta,N} \right). \end{aligned} \tag{3.39}$$

Consider now a specific entry $1 \leq i,j \leq d$ of the homogenized matrix. The control variate approach consists in approximating $\mathbb{E}\left[\left(A^\star_{\eta,N} \right)_{ij} \right]$ by considering a Monte Carlo estimator for $\mathbb{E}\left[\left(D^{3,\eta}_{\rho_1,\rho_2,\rho_3} \right)_{ij} \right]$. The deterministic parameters ρ_1, ρ_2 and ρ_3 are chosen to minimize the variance of $\left(D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega) \right)_{ij}$. They are thus the solution of the following 3×3 linear system (we drop the subscript i,j for conciseness):

$$\begin{aligned} \mathbb{V}\mathrm{ar}[A_1^{\eta,N}]\rho_1 + \mathbb{C}\mathrm{ov}[A_1^{\eta,N}, A_2^{\eta,N}]\rho_2 + \mathbb{C}\mathrm{ov}[A_1^{\eta,N}, C_2^{\eta,N}]\rho_3 &= \mathbb{C}\mathrm{ov}[A^\star_{\eta,N}, A_1^{\eta,N}] \\ \mathbb{C}\mathrm{ov}[A_2^{\eta,N}, A_1^{\eta,N}]\rho_1 + \mathbb{V}\mathrm{ar}[A_2^{\eta,N}]\rho_2 + \mathbb{C}\mathrm{ov}[A_2^{\eta,N}, C_2^{\eta,N}]\rho_3 &= \mathbb{C}\mathrm{ov}[A^\star_{\eta,N}, A_2^{\eta,N}] \\ \mathbb{C}\mathrm{ov}[C_2^{\eta,N}, A_1^{\eta,N}]\rho_1 + \mathbb{C}\mathrm{ov}[C_2^{\eta,N}, A_2^{\eta,N}]\rho_2 + \mathbb{V}\mathrm{ar}[C_2^{\eta,N}]\rho_3 &= \mathbb{C}\mathrm{ov}[A^\star_{\eta,N}, C_2^{\eta,N}] \end{aligned} \tag{3.40}$$

depending on the covariances between the entries ij of $A^\star_{\eta,N}$, $A_1^{\eta,N}$, $A_2^{\eta,N}$ and $C_2^{\eta,N}$. In practice, these covariances are approximated by empirical estimators (see Remark 3.1).

In practice, computing the matrices $\overline{A}^{0,l,N}_{2\,\rm def}$ (and likewise $\overline{C}^{0,l,N}_{2\,\rm def}$) is rather expensive (because each problem is set on the large domain Q_N, and the number of these problems

increases when N increases). It is therefore useful to approximate them using the Reduced Basis strategy introduced in [LBT12], which dramatically decreases the computational cost. The procedure is essentially as follows. We first solve the single defect problem (3.20) for $k = 0$, and solve (3.23) for a limited number of locations of the defect pairs, say $k = 0$ and l close to k. On the basis of these computations, we are then in position to obtain very efficient approximations of the matrices $\overline{A}_{2\ \mathrm{def}}^{0,l,N}$ for all $l \in \mathcal{I}_N$, $l \neq 0$. Evaluating (3.34) is thus inexpensive. Thus, up to a limited offline cost (i.e. the cost for solving the few problems (3.23) that we have to consider), the Monte Carlo empirical estimator and the Control Variate empirical estimator, defined respectively by

$$I_M^{\mathrm{MC}} = \frac{1}{M} \sum_{m=1}^{M} A_{\eta,N}^{\star,m}(\omega) \quad \text{and} \quad I_M^{\mathrm{CV}} := \frac{1}{M} \sum_{m=1}^{M} D_{\rho_1,\rho_2,\rho_3}^{3,\eta,m}(\omega)$$

share the same cost. We refer to Section 3.5.3 for numerical experiments using this procedure.

Remark 3.9. *In sharp contrast to the first order control variable, the second order control variable not only depends on the number of defects in the materials, i.e.* $\sum_{k\in\mathcal{I}_N} B_k^{\eta}(\omega)$*, but also on their location. The specific geometry of the materials, which is ignored in (3.32), is taken into account in (3.39).*

3.4 Elements of theoretical analysis

This section is devoted to establishing estimates on the gain provided by our approach. We proceed in two directions. First, in Section 3.4.1, we consider the one-dimensional case. Our main results are Propositions 3.11 and 3.13. We consider the large N regime, and estimate the variance (in terms of N) of $A_{\eta,N}^{\star}$, the controlled variables $D_{\rho}^{1,\eta}$ defined by (3.31) and $D_{\rho_1,\rho_2,\rho_3}^{3,\eta}$ defined by (3.39). We show that they are of the order of N^{-1}, N^{-2} and N^{-3}, respectively. Note that, in this section, we do not assume η to be close to 0 or 1, i.e. we are in a fully random case.

In Section 3.4.2, we turn to the multi-dimensional case. Our main result is Lemma 3.14. We consider the regime when η is small, and estimate the variance (in terms of η) of $A_{\eta,N}^{\star}$ and of the controlled variables $D_{\rho}^{1,\eta}$ defined by (3.31) and $D_{\rho_1,\rho_2}^{2,\eta}$ defined by (3.35). We show that the control variate approach using the first order (resp. second order) surrogate model allows to decrease the variance from $O(\eta)$ to $O(\eta^2)$ (resp. from $O(\eta)$ to $O(\eta^3)$).

Still in the regime $\eta \ll 1$, we show in Section 3.4.2 that, for an equal computational cost, the weakly stochastic approach proposed in [ALB11] (which directly compute $\mathbb{E}(A_{\eta,N}^{\star})$ as in series in powers of η) is more accurate than the control variate approach proposed in this work. The regime of interest for our approach is therefore when η is neither close to 0 nor to 1. This is the regime we consider in the numerical experiments of Section 3.5.

3.4.1 One-dimensional case

In the one-dimensional case, we know that

$$A_{\eta,N}^{\star}(\omega) = \left(\frac{1}{N} \int_0^N \frac{1}{A_\eta(x,\omega)} \right)^{-1},$$

where, for ease of notation, we set $Q_N = (0, N)$ rather than $Q_N = (-N/2, N/2)$ as before. In view of (3.14)–(3.15)–(3.16), we thus have

$$\frac{1}{A^\star_{\eta,N}(\omega)} = \frac{1}{N}\sum_{k=0}^{N-1}\int_k^{k+1} \frac{dx}{A_{\rm per}(x) + B^\eta_k(\omega)\Big(C_{\rm per}(x) - A_{\rm per}(x)\Big)}.$$

Introducing the functions

$$f(x) = \frac{1}{x} \quad \text{and} \quad \phi(b) = \int_0^1 \frac{dx}{A_{\rm per}(x) + b\Big(C_{\rm per}(x) - A_{\rm per}(x)\Big)},$$

we thus see that

$$A^\star_{\eta,N}(\omega) = f\left(\frac{1}{N}\sum_{k=0}^{N-1}\phi(B^\eta_k(\omega))\right).$$

Since $B^\eta_k(\omega)$ are equal to 0 or 1, we can write $\phi(B^\eta_k(\omega)) = \phi(0) + B^\eta_k(\omega)(\phi(1) - \phi(0))$, and thus

$$A^\star_{\eta,N}(\omega) = g\left(\frac{1}{N}\sum_{k=0}^{N-1} B^\eta_k(\omega)\right) \tag{3.41}$$

where the smooth function g is defined by $g(b) = f\Big(\phi(0) + b(\phi(1) - \phi(0))\Big)$.

First order model

In view of (3.31), (3.30) and (3.27), the first-order surrogate model is given by $A^\star_{\rm per} + A^{\eta,N}_1(\omega)$, with

$$A^{\eta,N}_1(\omega) = \sum_{k=0}^{N-1} B^\eta_k(\omega)\,\overline{A}^{k,N}_{1\,\rm def} = \overline{A}^{0,N}_{1\,\rm def}\sum_{k=0}^{N-1} B^\eta_k(\omega). \tag{3.42}$$

We first state the following general result, the proof of which is postponed until Section 3.4.1.

Lemma 3.10. *Let*

$$X(\omega) = g\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right)$$

where $B_k(\omega)$ are i.i.d. random variables valued in $[0,1]$ and g is a function in $C^3(\mathbb{R})$. Then

$$\mathbb{V}{\rm ar}(X) = \frac{(g'(\eta))^2\,\sigma^2}{N} + O\left(\frac{1}{N^2}\right) \tag{3.43}$$

with $\eta = \mathbb{E}(B_0)$ and $\sigma = \sqrt{\mathbb{V}{\rm ar}(B_0)}$.

For any ρ, introduce

$$D_\rho(\omega) = X(\omega) - \rho\Big(Y_1(\omega) - \mathbb{E}\left[Y_1\right]\Big) \quad \textit{where} \quad Y_1(\omega) = \sum_{k=0}^{N-1} B_k(\omega). \tag{3.44}$$

There exists a constant C independent of N and some deterministic parameter ρ_N such that

$$\mathbb{V}{\rm ar}\left(D_{\rho_N}\right) \le \frac{C}{N^2}. \tag{3.45}$$

The following proposition, of direct interest to us, directly falls from the above lemma.

Proposition 3.11. *Consider the model (3.14)–(3.15)–(3.16). Let $A^\star_{\eta,N}$ be the apparent homogenized matrix defined by (3.7)–(3.8) and $D^{1,\eta}_\rho$ be the first-order controlled variable defined by (3.31). In the one-dimensional case, we have*

$$\mathbb{V}\mathrm{ar}(A^\star_{\eta,N}) = \frac{C}{N} + O\left(\frac{1}{N^2}\right) \tag{3.46}$$

and, for the optimal value of the deterministic parameter ρ,

$$\min_\rho \mathbb{V}\mathrm{ar}\Big(D^{1,\eta}_\rho\Big) = \mathbb{V}\mathrm{ar}\Big(D^{1,\eta}_{\rho^\star}\Big) = O\left(\frac{1}{N^2}\right). \tag{3.47}$$

Using the control variate approach based on the first-order model, the variance is thus improved by at least one order in terms of N. Note in particular that, in the above results, we have not assumed η to be small.

Proof of Proposition 3.11. The proof of (3.46) falls from (3.41) and (3.43). We now prove (3.47). In view of (3.31), (3.41), (3.42) and (3.44), we see that

$$D^{1,\eta}_\rho = X(\omega) - \rho \overline{A}^{0,N}_{1\,\mathrm{def}}\Big(Y_1(\omega) - \mathbb{E}\left[Y_1\right]\Big).$$

Using (3.45), we thus have

$$\min_\rho \mathbb{V}\mathrm{ar}\Big(D^{1,\eta}_\rho\Big) \leq \mathbb{V}\mathrm{ar}\Big(D_{\rho_N}\Big) \leq \frac{C}{N^2},$$

which concludes the proof of Proposition 3.11. □

Second order model

In view of (3.39), (3.30), (3.34) and (3.38), the second-order controlled variable reads

$$\begin{aligned} D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega) = A^\star_{\eta,N}(\omega) &- \rho_1 \overline{A}^{0,N}_{1\,\mathrm{def}} \sum_{k=0}^{N-1} \Big(B^\eta_k(\omega) - \eta\Big) \\ &- \rho_2 \overline{A}^{0,1,N}_{2\,\mathrm{def}} \sum_{k\neq l}^{N-1} \Big(B^\eta_k(\omega) B^\eta_l(\omega) - \eta^2\Big) \\ &- \rho_3 \overline{C}^{0,1,N}_{2\,\mathrm{def}} \sum_{k\neq l}^{N-1} \Big((1 - B^\eta_k(\omega))(1 - B^\eta_l(\omega)) - (1-\eta)^2\Big) \end{aligned}$$

where we have used (3.29) and the fact that, in the one-dimensional case, $\overline{A}^{k,l,N}_{2\,\mathrm{def}}$ and $\overline{C}^{k,l,N}_{2\,\mathrm{def}}$ are independent of k and l. We hence obtain that

$$D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega) = A^\star_{\eta,N}(\omega) - \overline{\rho_1} \sum_{k=0}^{N-1} \Big(B^\eta_k(\omega) - \eta\Big) - \overline{\rho_2} \sum_{k\neq l}^{N-1} \Big(B^\eta_k(\omega) B^\eta_l(\omega) - \eta^2\Big) \tag{3.48}$$

with

$$\overline{\rho_1} = \rho_1 \overline{A}^{0,N}_{1\,\mathrm{def}} - 2(N-1)\rho_3 \overline{C}^{0,1,N}_{2\,\mathrm{def}}, \qquad \overline{\rho_2} = \rho_2 \overline{A}^{0,1,N}_{2\,\mathrm{def}} + \rho_3 \overline{C}^{0,1,N}_{2\,\mathrm{def}}.$$

We first state the following general result, the proof of which is postponed until Section 3.4.1.

Lemma 3.12. *Let*

$$X(\omega) = g\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right)$$

where g is a function in $C^3(\mathbb{R})$ and $B_k(\omega)$ are i.i.d. random variables taking values in $\{0,1\}$. Let Y_1 be defined by (3.44) and Y_2 be defined by

$$Y_2(\omega) = \sum_{k=0}^{N-1}\sum_{l=0, l\neq k}^{N-1} B_k(\omega)B_l(\omega). \tag{3.49}$$

There exists a constant C independent of N and some deterministic parameters $\overline{\rho_1}$ and $\overline{\rho_2}$ (that depend on N) such that

$$\mathbb{V}\mathrm{ar}\Big(\overline{D}_{\overline{\rho_1},\overline{\rho_2}}\Big) \leq \frac{C}{N^3} \tag{3.50}$$

where $\overline{D}_{\overline{\rho_1},\overline{\rho_2}}(\omega) = X(\omega) - \overline{\rho_1}\Big(Y_1(\omega) - \mathbb{E}(Y_1)\Big) - \overline{\rho_2}\Big(Y_2(\omega) - \mathbb{E}(Y_2)\Big)$.

The following proposition directly falls from the above lemma.

Proposition 3.13. *Consider the model (3.14)–(3.15)–(3.16). Let $A^\star_{\eta,N}$ be the apparent homogenized matrix defined by (3.7)–(3.8) and $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$ be the second-order controlled variable defined by (3.39). In the one-dimensional case, for the optimal value of the deterministic parameters ρ_1, ρ_2 and ρ_3, we have*

$$\min_{\rho_1,\rho_2,\rho_3} \mathbb{V}\mathrm{ar}\Big(D^{3,\eta}_{\rho_1,\rho_2,\rho_3}\Big) = O\left(\frac{1}{N^3}\right). \tag{3.51}$$

We recall that

$$\mathbb{V}\mathrm{ar}\Big(A^\star_{\eta,N}\Big) = \frac{C}{N} + O\left(\frac{1}{N^2}\right).$$

Thus, using the control variate approach based on the second-order model, the variance is improved by at least two orders in terms of N. This result is to be compared with Proposition 3.11.

Proof of Proposition 3.13. In view of (3.48), (3.41), (3.44) and (3.49), we see that

$$D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega) = X(\omega) - \overline{\rho_1}\Big(Y_1(\omega) - \mathbb{E}\left[Y_1\right]\Big) - \overline{\rho_2}\Big(Y_2(\omega) - \mathbb{E}\left[Y_2\right]\Big).$$

Using (3.50), we thus have

$$\min_{\rho_1,\rho_2,\rho_3} \mathbb{V}\mathrm{ar}\Big(D^{3,\eta}_{\rho_1,\rho_2,\rho_3}\Big) \leq \mathbb{V}\mathrm{ar}\Big(\overline{D}_{\overline{\rho_1},\overline{\rho_2}}\Big) \leq \frac{C}{N^3},$$

which concludes the proof of Proposition 3.13. □

Proofs of Lemmas 3.10 and 3.12

Proof of Lemma 3.10. Introducing the centered random variables

$$d_k(\omega) = B_k(\omega) - \eta$$

and a smooth function h on $[0,1]$, we write

$$\begin{aligned} h\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right) &= h\left(\eta + \frac{1}{N}\sum_{k=0}^{N-1} d_k(\omega)\right) \\ &= h(\eta) + \frac{h'(\eta)}{N}\sum_{k=0}^{N-1} d_k(\omega) + \frac{h''(\eta)}{2}\left(\frac{1}{N}\sum_{k=0}^{N-1} d_k(\omega)\right)^2 \\ &\quad + \frac{h'''(\theta_3^N(\omega))}{6}\left(\frac{1}{N}\sum_{k=0}^{N-1} d_k(\omega)\right)^3 \end{aligned} \tag{3.52}$$

for some $\theta_3^N(\omega) \in [0,1]$. Recall now that any i.i.d. variables d_k with mean value zero satisfy the following bounds:

$$\forall p \in \mathbb{N}^\star,\ \exists C_p > 0, \quad \left|\mathbb{E}\left[\left(\frac{1}{N}\sum_{k=0}^{N-1} d_k\right)^p\right]\right| \leq \begin{cases} \dfrac{C_p}{N^{p/2}} & \text{if } p \text{ is even;} \\ \dfrac{C_p}{N^{(p+1)/2}} & \text{if } p \text{ is odd.} \end{cases} \tag{3.53}$$

This is proved by developing the power p of the sum, and then using the fact that the variables are i.i.d and have mean value zero. Taking expectations in (3.52), we thus deduce that

$$\mathbb{E}\left[h\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right)\right] = h(\eta) + \frac{h''(\eta)}{2N}\sigma^2 + O\left(\frac{1}{N^2}\right),$$

where $\sigma^2 = \mathbb{E}[d_0^2] = \mathbb{V}\mathrm{ar}(B_0)$. Choosing $h(x) = g(x)$ and $h(x) = (g(x))^2$, we obtain (3.43).

We next turn to proving (3.45). As in (3.52), we have

$$\begin{aligned} X(\omega) &= g\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right) \\ &= g(\eta) + \frac{g'(\eta)}{N}\sum_{k=0}^{N-1} d_k(\omega) + \frac{g''(\theta_2^N(\omega))}{2} S_N(\omega) \\ &= g(\eta) + \frac{g'(\eta)}{N}\Big(Y_1(\omega) - \mathbb{E}\left[Y_1\right]\Big) + \frac{g''(\theta_2^N(\omega))}{2} S_N(\omega) \end{aligned}$$

for some $\theta_2^N(\omega) \in [0,1]$, where $S_N(\omega) = \left(\frac{1}{N}\sum_{k=0}^{N-1} d_k(\omega)\right)^2$. Set $\rho_N = \frac{g'(\eta)}{N}$. Then

$$D_{\rho_N}(\omega) = X(\omega) - \rho_N\Big(Y_1(\omega) - \mathbb{E}\left[Y_1\right]\Big) = g(\eta) + \frac{g''(\theta_2^N(\omega))}{2} S_N(\omega).$$

Using (3.53), we thus obtain that

$$\mathbb{V}\mathrm{ar}\left(D_{\rho_N}\right) \leq \mathbb{E}\left[\left(\frac{g''(\theta_2^N(\omega))}{2} S_N(\omega)\right)^2\right] \leq C\mathbb{E}\left[(S_N)^2\right] \leq \frac{C}{N^2}$$

which is the claimed bound (3.45). This concludes the proof of Lemma 3.10. □

Proof of Lemma 3.12. We follow the same lines as in the proof of Lemma 3.10. Introducing the centered random variables

$$d_k(\omega) = B_k(\omega) - \eta,$$

we write, as in (3.52), that

$$\begin{aligned} X(\omega) &= g\left(\frac{1}{N}\sum_{k=0}^{N-1} B_k(\omega)\right) \\ &= g(\eta) + \frac{g'(\eta)}{N}\sum_{k=0}^{N-1} d_k(\omega) + \frac{g''(\eta)}{2N^2}\left(\sum_{k=0}^{N-1} d_k(\omega)\right)^2 + \frac{g'''(\theta_3^N(\omega))}{6} S_N(\omega) \end{aligned} \tag{3.54}$$

for some $\theta_3^N(\omega) \in [0,1]$, where $S_N(\omega) = \left(\frac{1}{N}\sum_{k=0}^{N-1} d_k(\omega)\right)^3$. We now recall that $\sum_{k=0}^{N-1} d_k(\omega) = Y_1 - \mathbb{E}(Y_1)$. Furthermore, we compute that

$$\left(\sum_{k=0}^{N-1} d_k(\omega)\right)^2 = N^2\eta^2 + (1-2N\eta)Y_1(\omega) + Y_2(\omega).$$

We thus recast (3.54) as

$$X(\omega) = \mathcal{C} + \frac{g'(\eta)}{N} Y_1(\omega) + \frac{g''(\eta)}{2N^2}\Big((1-2N\eta)Y_1(\omega) + Y_2(\omega)\Big) + \frac{g'''(\theta_3^N(\omega))}{6} S_N(\omega)$$

where $\mathcal{C}$ is a deterministic quantity.

Set $\overline{\rho_1} = \frac{g'(\eta)}{N} + \frac{g''(\eta)}{2N^2}(1-2N\eta)$ and $\overline{\rho_2} = \frac{g''(\eta)}{2N^2}$. Then

$$\overline{D}_{\overline{\rho_1},\overline{\rho_2}}(\omega) = X(\omega) - \overline{\rho_1}\Big(Y_1(\omega) - \mathbb{E}(Y_1)\Big) - \overline{\rho_2}\Big(Y_2(\omega) - \mathbb{E}(Y_2)\Big) = \mathcal{C} + \frac{g'''(\theta_3^N(\omega))}{6} S_N(\omega).$$

Using (3.53), we thus obtain that

$$\mathbb{V}\mathrm{ar}\left(\overline{D}_{\overline{\rho_1},\overline{\rho_2}}\right) \le \mathbb{E}\left[\left(\frac{g'''(\theta_3^N(\omega))}{6} S_N(\omega)\right)^2\right] \le C\mathbb{E}\left[(S_N)^2\right] \le \frac{C}{N^3}$$

which is the claimed bound (3.50). This concludes the proof of Lemma 3.12. □

3.4.2 Multi-dimensional case

Proof of Lemma 3.4

The proof follows the same lines as that of (3.25). It falls by enumerating the possible configurations according to the number of defects they include. We thus have, following Section 3.2.2,

$$\mathbb{E}\left[\varphi\left(A^\star_{\eta,N}\right)\right] = (1-\eta)^{|Q_N|}\varphi\left(A^\star_{\mathrm{per}}\right) + \sum_{k\in\mathcal{I}_N} \eta(1-\eta)^{|Q_N|-1}\varphi\left(A^\star_{1,k,N}\right) + O_N(\eta^2). \tag{3.55}$$

On the other hand, using (3.30) and (3.27), we write

$$\begin{aligned} &\mathbb{E}\left[\varphi\left(A^\star_{\mathrm{per}} + A_1^{\eta,N}\right)\right] \\ =\ & (1-\eta)^{|Q_N|}\varphi\left(A^\star_{\mathrm{per}}\right) + \sum_{k\in\mathcal{I}_N} \eta(1-\eta)^{|Q_N|-1}\varphi\left(A^\star_{\mathrm{per}} + \overline{A}^{k,N}_{1\,\mathrm{def}}\right) + O_N(\eta^2) \\ =\ & (1-\eta)^{|Q_N|}\varphi\left(A^\star_{\mathrm{per}}\right) + \sum_{k\in\mathcal{I}_N} \eta(1-\eta)^{|Q_N|-1}\varphi\left(A^\star_{1,k,N}\right) + O_N(\eta^2). \end{aligned}$$

We deduce from the above relation and (3.55) the claimed result.

Estimates of the variances as a function of η

Lemmas 3.4 and 3.8 show that our surrogate model is a good approximation (in terms of its law) of the random variable $A^\star_{\eta,N}$. The lemma below shows, again in the regime $\eta \ll 1$, that variance is indeed decreased.

Consider any entry ij of the homogenized matrix. The estimation of $\mathbb{E}\left[\left(A^\star_{\eta,N}\right)_{ij}\right]$ can be done by a Monte Carlo empirical mean on $\left(A^\star_{\eta,N}(\omega)\right)_{ij}$, $\left(D^{1,\eta}_{\rho}(\omega)\right)_{ij}$ (see Section 3.3.1) or $\left(D^{2,\eta}_{\rho_1,\rho_2}(\omega)\right)_{ij}$ (see Section 3.3.2).

Lemma 3.14. *For any entry ij of the homogenized matrix, we have*

$$\mathbb{V}\mathrm{ar}\left[\left(A^\star_{\eta,N}\right)_{ij}\right] = \eta C^0_N + O_N(\eta^2), \tag{3.56}$$

$$\mathbb{V}\mathrm{ar}\left[\left(D^{1,\eta}_{\rho=1}\right)_{ij}\right] = O_N(\eta^2), \tag{3.57}$$

$$\mathbb{V}\mathrm{ar}\left[\left(D^{2,\eta}_{\rho_1=\rho_2=1}\right)_{ij}\right] = O_N(\eta^3), \tag{3.58}$$

where C^0_N is a positive constant.

In practice, we would not necessarily work with $\rho = 1$, but with the optimal parameter $\rho^\star$. A direct consequence of (3.57) is of course that

$$\mathbb{V}\mathrm{ar}\left[\left(D^{1,\eta}_{\rho^\star}\right)_{ij}\right] = \inf_\rho \mathbb{V}\mathrm{ar}\left[\left(D^{1,\eta}_{\rho}\right)_{ij}\right] = O_N(\eta^2).$$

Remark 3.15. *Even though the variance of $D^{1,\eta}_{\rho^\star}$ is much smaller than that of $A^\star_{\eta,N}$, we will see in Section 3.4.2 below that, in the regime $\eta \ll 1$, the weakly stochastic approximation described in Section 3.2.2 is even more efficient.*

Proof. We infer from (3.55) and (3.29) that, for any function φ,

$$\mathbb{E}\left[\varphi\left(A^\star_{\eta,N}\right)\right] = \varphi\left(A^\star_{\mathrm{per}}\right) + \eta|Q_N|\Big(\varphi\left(A^\star_{1,0,N}\right) - \varphi\left(A^\star_{\mathrm{per}}\right)\Big) + O_N(\eta^2).$$

Taking $\varphi(M) = M_{ij}$ and $\varphi(M) = M^2_{ij}$, we obtain (3.56).

We next turn to proving (3.57). For any function φ, we write, using (3.31) and (3.27), that

$$\begin{aligned}
&\mathbb{E}\left[\varphi\left(D^{1,\eta}_{\rho=1}\right)\right] \\
= \;& (1-\eta)^{|Q_N|}\varphi\left(A^\star_{\mathrm{per}} + \eta\overline{A}^N_1\right) \\
&\quad + \sum_{k\in\mathcal{I}_N} \eta(1-\eta)^{|Q_N|-1}\varphi\left(A^\star_{1,k,N} - \overline{A}^{k,N}_{1\,\mathrm{def}} + \eta\overline{A}^N_1\right) + O_N(\eta^2) \\
= \;& (1-\eta)^{|Q_N|}\varphi\left(A^\star_{\mathrm{per}} + \eta\overline{A}^N_1\right) + \sum_{k\in\mathcal{I}_N} \eta(1-\eta)^{|Q_N|-1}\varphi\left(A^\star_{\mathrm{per}} + \eta\overline{A}^N_1\right) + O_N(\eta^2) \\
= \;& \varphi\left(A^\star_{\mathrm{per}} + \eta\overline{A}^N_1\right) + O_N(\eta^2).
\end{aligned}$$

Taking $\varphi(M) = M_{ij}$ and $\varphi(M) = M^2_{ij}$, we obtain (3.57). The proof of (3.58) follows the same lines. □

Comparison to a weakly stochastic approach

In the regime $\eta \ll 1$, we have three approaches at our disposal to estimate $\mathbb{E}\left[A^\star_{\eta,N}\right]$: the standard Monte Carlo approach, the control variate approach, and the weakly stochastic approach described in Section 3.2.2. We compare here their efficiency. Let $\mathcal{C}_N$ be the cost to solve a single corrector problem on Q_N.

The standard Monte Carlo approach amounts to writing

$$\mathbb{E}\left[A^\star_{\eta,N}\right] \approx \frac{1}{M}\sum_{m=1}^{M} A^{\star,m}_{\eta,N}(\omega).$$

In the above approximation, the error on the entry ij is controlled by $\sqrt{\mathbb{V}\mathrm{ar}\left[\left(A^\star_{\eta,N}\right)_{ij}\right]/M}$. In view of (3.56), it is thus of the order of $\sqrt{\eta/M}$. The cost is $M\,\mathcal{C}_N$.

The control variate approach (say using the first order surrogate model) amounts to writing

$$\mathbb{E}\left[A^\star_{\eta,N}\right] \approx \frac{1}{M}\sum_{m=1}^{M} D^{1,\eta,m}_{\rho}(\omega),$$

where $D^{1,\eta}_{\rho}(\omega)$ is defined by (3.31). The error is of the order of $\sqrt{\eta^2/M}$ in view of (3.57). The cost is that of solving M corrector problems and that of determining $\overline{A}^{0,N}_{1\,\mathrm{def}}$, namely $(1+M)\,\mathcal{C}_N$.

Using the same kind of information as in the above control variate approach, the weakly stochastic approximation (3.25) reads

$$\mathbb{E}\left[A^\star_{\eta,N}\right] \approx A^\star_{\mathrm{per}} + \eta\overline{A}^{N}_{1}.$$

The error is of the order of η^2. The cost is that of determining $\overline{A}^{0,N}_{1\,\mathrm{def}}$, i.e. $\mathcal{C}_N$.

Obviously, the control variate approach is always more efficient than the Monte Carlo approach. However, to reach the same accuracy as the weakly stochastic approach, one would need to take $M = \eta^{-2}$ realizations, leading to a cost much larger than with the weakly stochastic approach. The same observation holds when using the control variate approach using the second order surrogate model. Therefore, in the regime $\eta \ll 1$, the weakly stochastic approach (3.25) is the most efficient one.

3.5 Numerical results

We consider the so-called random checkerboard case, in dimension $d = 2$ (see Fig. 3.3). It falls into the framework (3.14)–(3.15)–(3.16) with

$$A_{\mathrm{per}}(x) = \alpha\mathrm{Id}_2 \quad \text{and} \quad C_{\mathrm{per}}(x) = \beta\mathrm{Id}_2. \tag{3.59}$$

In what follows, we choose $\alpha = 3$ and $\beta = 23$ (in Section 3.5.1) or $\beta = 103$ (in Section 3.5.2). All variances are estimated on the basis of $M = 100$ independent realizations.

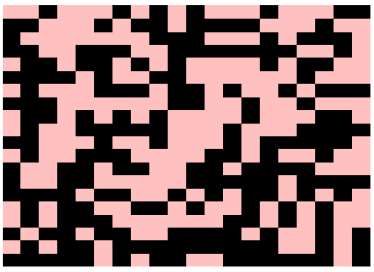

Figure 3.3 – A typical realization of the checkerboard test-case with $\eta = 1/2$.

3.5.1 Low contrast test-case

We choose here $(\alpha, \beta) = (3, 23)$. The motivation for this choice is that we already considered this test-case in [BCLBL12b, BCLBL12a, CLBL10] when introducing an antithetic variable approach. We are thus in position to compare the results obtained here with our previous results.

On Fig. 3.4, we plot as a function of $\eta \in (0, 1)$ three quantities:

- the first entry of the matrix $\mathbb{E}\left[A^\star_{\eta,N}\right]$ (obtained in practice by an expensive Monte Carlo estimation);
- the weakly stochastic approximation (3.25), which is an approximation of $\mathbb{E}\left[A^\star_{\eta,N}\right]$ with an error of the order of $O_N(\eta^3)$;
- the weakly stochastic approximation obtained in the regime $(1-\eta) \ll 1$, which is an approximation of $\mathbb{E}\left[A^\star_{\eta,N}\right]$ with an error of the order of $O_N\big((1-\eta)^3\big)$.

In all cases, we work with $N = 10$, and the following observations are also valid for larger values of N. We see on Fig. 3.4 that, when $\eta \leq 0.4$, the deterministic expansion (3.25) is a very accurate approximation of $\mathbb{E}\left[\left(A^\star_{\eta,N}\right)_{11}\right]$. This approximation is inexpensive to compute. The same observation holds in the regime $\eta \geq 0.7$, where the deterministic expansion around $\eta = 1$ provides a satisfying approximation. However, we note that none of the two weakly stochastic expansions are accurate when $0.4 \leq \eta \leq 0.7$. In that regime, one has to compute $\mathbb{E}\left[\left(A^\star_{\eta,N}\right)_{11}\right]$ by considering several realizations of (3.7)–(3.8). In that regime, considering a variance reduction approach is useful.

In the regime we have identified, we show on Fig. 3.5 the ratios of variance

$$R_{\eta,N} = \frac{\mathbb{V}\mathrm{ar}\Big(\left[A^\star_{\eta,N}\right]_{11}\Big)}{\mathbb{V}\mathrm{ar}(D)}, \tag{3.60}$$

where D is either the first-order controlled variable $D^{1,\eta}_{\rho}(\omega)$ defined by (3.31), or the second-order controlled variable $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$ defined by (3.35), or the controlled variable $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$ defined by (3.39). The parameter ρ (resp. (ρ_1, ρ_2) and (ρ_1, ρ_2, ρ_3)) is chosen to minimize the variance of the estimator. In this section, we exactly compute (up to finite element errors) the quantities $\overline{A}^{k,l,N}_{2\,\mathrm{def}}$ needed to build the controlled variables (3.35) and (3.39). In

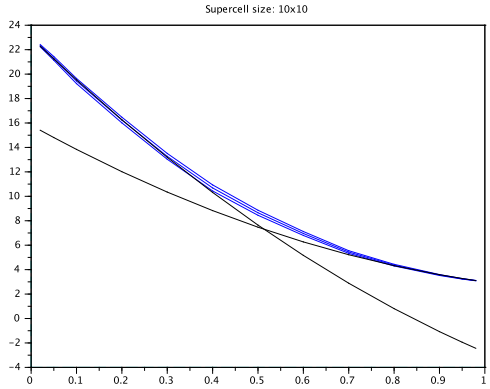

Figure 3.4 – $\mathbb{E}\left[\left(A^{\star}_{\eta,N}\right)_{11}\right]$ as a function of η, for $N = 10$. Black curves: weakly stochastic approximations. Blue curve: Monte Carlo standard estimator.

Section 3.5.3 below, we approximate them using a Reduced Basis approach. We postpone until that section the discussion on computational costs and only focus here on accuracy.

Remark 3.16. *The second-order controlled variable $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$ defined by (3.35) is built by considering $A_{\rm per}$ as the reference. One could alternatively build a second-order controlled variable considering $C_{\rm per}$ as the reference. Numerical results obtained with such a controlled variable are similar to those obtained with $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$ (results not shown).*

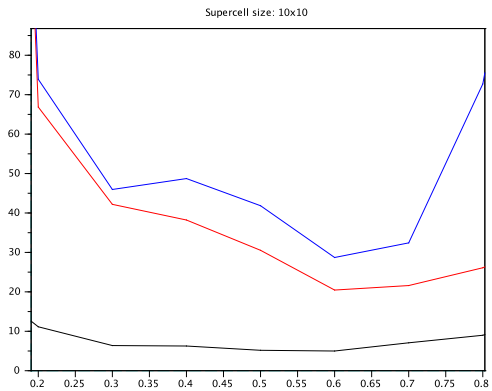

Figure 3.5 – Ratio $R_{\eta,N}$ defined by (3.60) as a function of η ($N = 10$). Black curve: controlled variable $D^{1,\eta}_{\rho}(\omega)$. Red curve: controlled variable $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$. Blue curve: controlled variable $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$.

We observe on Fig. 3.5 that, for $\eta = 1/2$, the approach using the first-order controlled variable (3.31) provides a variance reduction ratio (3.60) close to 6. This gain is close to the gain obtained using an antithetic variable approach (see [CLBL10, Table 2]). In contrast, when using the controlled variable (3.39) taking into account first order and second order corrections with respect to both the cases $\eta = 0$ and $\eta = 1$, we obtain a gain close to 40.

We now monitor how the gain depends on the size of the domain Q_N. To that aim, we

show on Table 3.1 the ratio (3.60) as a function of N, for $\eta = 1/2$. We observe that the gain is essentially independent of N.

	$N = 6$	$N = 10$	$N = 20$	$N = 30$	$N = 50$
First order	7.57	5.18	6.55	8.51	7.34
Second order	35.9	41.8	37.6	35.6	40.4

Table 3.1 – Ratio $R_{\eta,N}$ defined by (3.60) as a function of N ($\eta = 1/2$). First order: controlled variable $D^{1,\eta}_{\rho}(\omega)$. Second order: controlled variable $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$.

Remark 3.17. *In the one-dimensional case, we have shown that the variance ratio is proportional to N or N^2 (see Propositions 3.11 and 3.13). In the two-dimensional case, we do not observe such an excellent behavior for our approach. The gain rather seems to be independent of N (see also Fig. 3.9). Nevertheless, the variance ratio is significantly higher than 1, making the approach definitely superior to the standard Monte Carlo approach.*

3.5.2 High contrast test-case

We now turn to a test-case with a larger contrast and set $(\alpha, \beta) = (3, 103)$ in (3.59). On Fig. 3.6, we plot as a function of $\eta \in (0, 1)$ the same three quantities as on Fig. 3.4 (again with $N = 10$). We again see that, when $0.3 \leq \eta \leq 0.7$, none of the two weakly stochastic expansions are accurate. This is the regime we focus on.

We also show on Fig. 3.6 the ratios of variance (3.60) for the same three control variate approaches as on Fig. 3.5. We observe that, for $\eta = 1/2$, the approach using the controlled variable (3.39) provides a gain close to 6.7. This gain is smaller than in the case of Section 3.5.1 (the contrast is now larger), but still significant. As in the low-contrast test-case, the gain is essentially independent of N, as shown in Table 3.2.

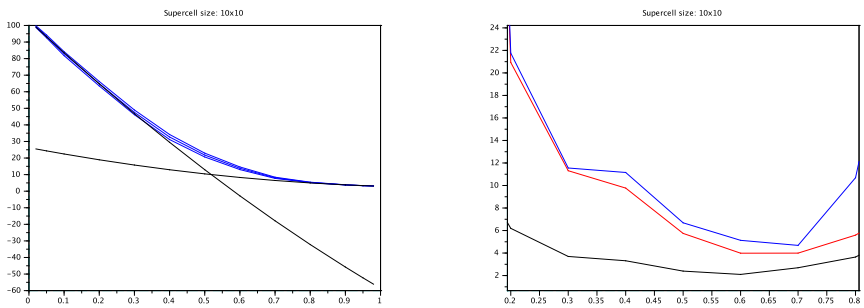

Figure 3.6 – Left: $\mathbb{E}\left[\left(A^{\star}_{\eta,N}\right)_{11}\right]$ as a function of η, for $N = 10$. Blue curve: standard Monte Carlo estimator. Black curves: weakly stochastic approximations. Right: Ratio $R_{\eta,N}$ defined by (3.60) as a function of η ($N = 10$). Black curve: controlled variable $D^{1,\eta}_{\rho}(\omega)$. Red curve: controlled variable $D^{2,\eta}_{\rho_1,\rho_2}(\omega)$. Blue curve: controlled variable $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$.

3.5.3 Using a Reduced Basis (RB) approach

In Sections 3.5.1 and 3.5.2, we have used the second-order surrogate model (3.39), which takes into account the contributions from pairs of defects located at any site k and l, namely

	$N = 10$	$N = 30$	$N = 50$
First order	2.40	3.62	3.87
Second order	6.69	6.32	5.82

Table 3.2 – Ratio $R_{\eta,N}$ defined by (3.60) as a function of N ($\eta = 1/2$). First order: controlled variable $D^{1,\eta}_{\rho}(\omega)$. Second order: controlled variable $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$.

$A^\star_{2,k,l,N}$ defined by (3.24) and $C^\star_{2,k,l,N}$ defined by (3.37). These quantities are deterministic, and computed beforehand. However, in practice, computing these quantities is expensive, because we have to consider all possible configurations of pairs of defects.

This high computational cost can be decreased by using the Reduced Basis (RB) approach proposed in [LBT12]. This approach amounts to solving the one-defect problem (3.20), and *a few* two-defects problems (3.23), for $k = 0$ and l in some set $\mathcal{N}_N \subset \mathcal{I}_N \setminus \{0\}$ (in practice, we solve (3.23) for some l close to k). Then, it turns out that the solutions to the other two-defects problems, i.e. $w_p^{2,k,l,N}$ for $k = 0$ and $l \notin \mathcal{N}_N$, can be well-approximated on the basis of $w_p^{1,0,N}$ and $\left\{w_p^{2,k,l,N}\right\}_{k=0,\, l\in\mathcal{N}_N}$.

In the sequel, we consider the low-contrast test-case (i.e. $(\alpha, \beta) = (3, 23)$ in (3.59)), set $\eta = 1/2$, and use this RB approach in order to decrease the offline cost of our control variate approach.

Robustness with respect to the RB basis set

First, we evaluate the robustness of the gain in variance when we approximate the quantities $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$ by the above RB approach, in contrast to computing them exactly (i.e., up to a small Finite Element error). To do so, we fix N and monitor the variance ratio for the sets $\mathcal{N}_N$ shown on Fig. 3.7. Results are given in Table 3.3. We see that the gain in variance is independent of the set $\mathcal{N}_N$: we can use the RB approach with a very small set of configurations for which the correctors $w_p^{2,k,l,N}$ are exactly computed (thereby dramatically decreasing the offline computational cost), and still retain an excellent variance reduction.

	$N = 6$	$N = 20$
$\mathcal{N}_N = \mathcal{I}_N \setminus \{0\}$	35.9	37.6
Card $\mathcal{N}_N = 20$	36.1	37.6
Card $\mathcal{N}_N = 12$	35.7	37.0
Card $\mathcal{N}_N = 8$	36.6	36.5
Card $\mathcal{N}_N = 4$	36.6	37.6

Table 3.3 – Ratio $R_{\eta,N}$ defined by (3.60) for two values of N ($\eta = 1/2$), using the second order model $D^{3,\eta}_{\rho_1,\rho_2,\rho_3}(\omega)$ defined by (3.39). The first line corresponds to the reference computation of $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$. The subsequent lines correspond to using a RB approach to compute $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$, with a decreasing set $\mathcal{N}_N$.

Following the above idea, we have also tested the approach when we set $\overline{A}^{k,l,N}_{2\,\mathrm{def}} = \overline{C}^{k,l,N}_{2\,\mathrm{def}} = \mathrm{Id}$ in (3.34) and (3.38) for any $k \neq l$ (which amounts to setting $A^\star_{2,k,l,N} = \mathrm{Id} + 2A^\star_{1,0,N} - A^\star_{\mathrm{per}}$, see (3.28)). We do not expect (and this is indeed the case) to obtain

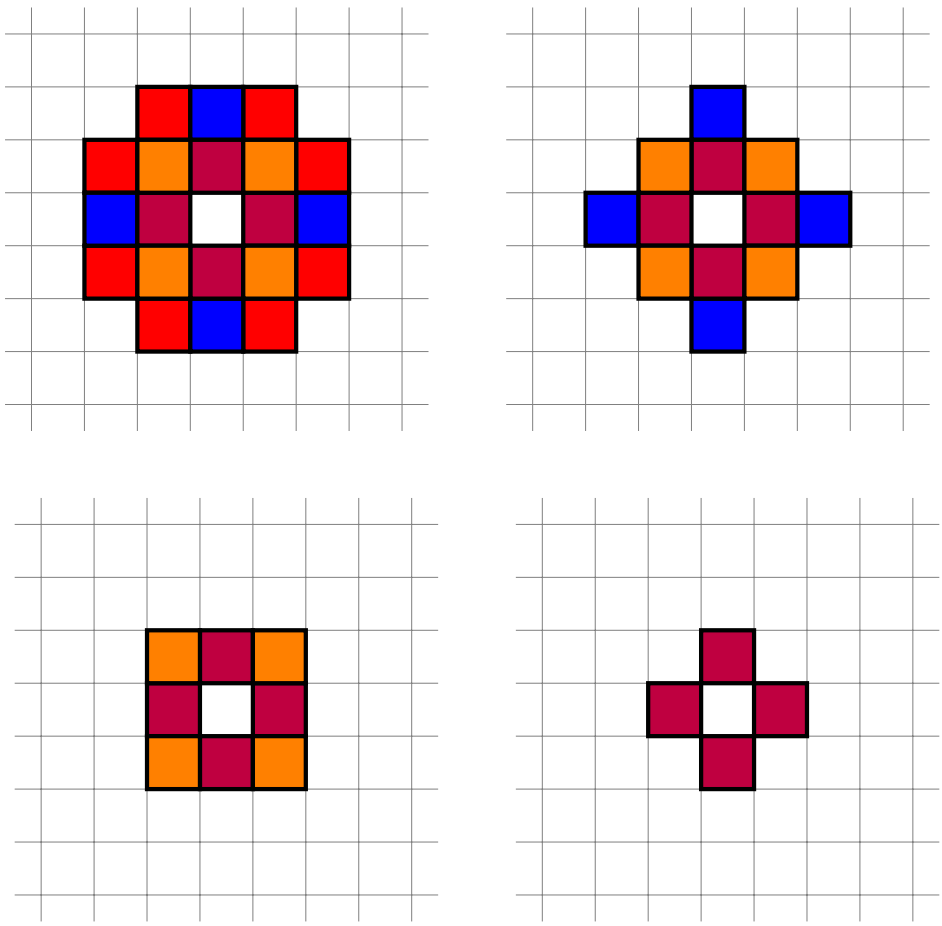

Figure 3.7 – Sets $\mathcal{N}_N$ of position of second defect that we consider to build the RB basis set (the first defect is always in the central white cell). Top left: Card $\mathcal{N}_N = 20$. Top right: Card $\mathcal{N}_N = 12$. Bottom left: Card $\mathcal{N}_N = 8$. Bottom right: Card $\mathcal{N}_N = 4$.

good results. The controlled variable reads

$$D^{3,\eta,\text{approx}}_{\rho_1,\rho_2,\rho_3}(\omega) = A^\star_{\eta,N}(\omega) - \rho_1 \left(A_1^{\eta,N}(\omega) - \eta \overline{A}_1^N \right) - \frac{\rho_2}{2} \sum_{k \neq l \in \mathcal{I}_N} \left(B_k^\eta(\omega) B_l^\eta(\omega) - \mathbb{E}\left[B_k^\eta B_l^\eta \right] \right) - \frac{\rho_3}{2} \sum_{k \neq l \in \mathcal{I}_N} \left((1 - B_k^\eta(\omega))(1 - B_l^\eta(\omega)) - \mathbb{E}\left[(1 - B_k^\eta)(1 - B_l^\eta) \right] \right) \quad (3.61)$$

instead of (3.39). Computing the second order surrogate model is then extremely cheap, and as expensive as computing the first order surrogate model: one only has to solve the one-defect problem (3.20). In that case, for $N = 20$ and $\eta = 1/2$, the variance ratio is equal to 6.96, which is extremely close to the variance ratio obtained by simply using the first order model (see Table 3.1), which is equal to 6.55. Considering the last two lines in (3.61) therefore does not improve the efficiency.

The above results show that it is not needed to compute with a high accuracy the quantities $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$ to obtain a significant variance reduction. Using a RB approach with a very small set $\mathcal{N}_N$ is sufficient and the gain (in terms of variance reduction) is essentially the same as that if $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$ are exactly computed. However, even though the approach is quite flexible, it still requires approximations of $A^\star_{2,k,l,N}$ and $C^\star_{2,k,l,N}$ with a reasonable accuracy. Otherwise, the efficiency significantly drops down, as shown by our last test.

Results as a function of N

We now fix the RB basis set corresponding to Card $\mathcal{N}_N = 12$ on Fig. 3.7, and compare the Monte Carlo results with our control variate results, using the controlled variable (3.39). To evaluate the Monte Carlo estimator

$$I_M^{\text{MC}} = \frac{1}{M} \sum_{m=1}^{M} A^{\star,m}_{\eta,N}(\omega),$$

we need to solve M corrector problems. In contrast, to evaluate the Control Variate estimator

$$I_M^{\text{CV}} := \frac{1}{M} \sum_{m=1}^{M} D^{3,\eta,m}_{\rho_1,\rho_2,\rho_3}(\omega),$$

we need to solve first the problem (3.20) and the problems (3.23) for $k = 0$ and $l \in \mathcal{N}_N$, and second M corrector problems. Let $\mathcal{C}_N$ be the cost to solve a single corrector problem on Q_N. Then the Monte Carlo cost is $M\,\mathcal{C}_N$, the Control Variate offline cost is $(1+\mathcal{N}_N)\,\mathcal{C}_N = 13\mathcal{C}_N$, and its online cost is $M\,\mathcal{C}_N$. In the sequel, we work with $M = 100$, therefore the Control Variate cost is just 13% higher than the Monte Carlo cost.

First, we plot on Fig. 3.8 the confidence intervals obtained for the Monte Carlo approach and the Control Variate approach based on (3.39). The latter confidence interval width is dramatically smaller than the former.

We next show on Fig. 3.9 the variance ratios (3.60). They somewhat vary with N. Recall that these ratios are computed on the basis of $M = 100$ i.i.d. realizations. From one set of i.i.d. realizations to another, results may slightly vary, although qualitative conclusions remain alike. For the first order method based on (3.31), the variance ratio

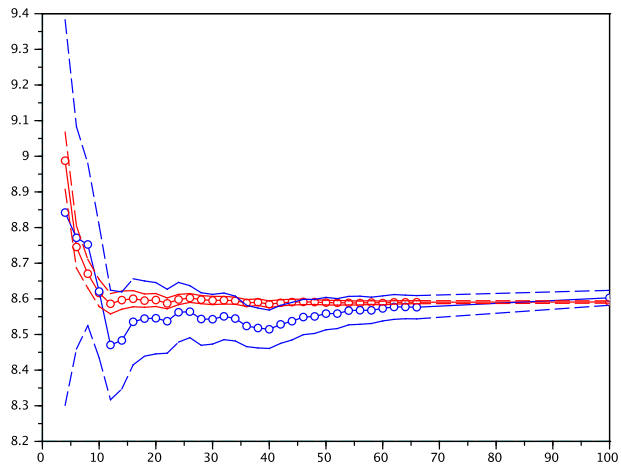

Figure 3.8 – Estimation of $\mathbb{E}\left(\left[A^{\star}_{\eta,N}\right]_{11}\right)$ as a function of N. Blue: standard Monte-Carlo estimator. Red: Control Variate estimator based on (3.39). In both cases, estimators are built using $M = 100$ i.i.d. realizations.

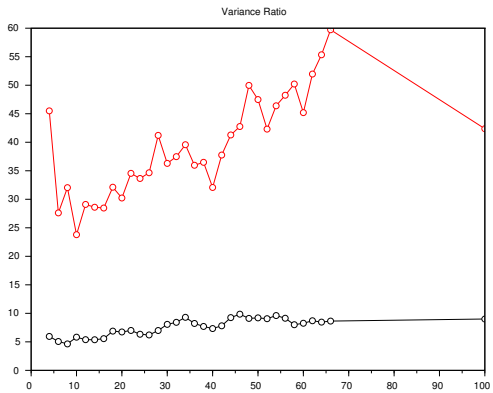

Figure 3.9 – Variance ratio (3.60) as a function of N. Black curve: using the first order controlled variable (3.31). Red curve: using the second order controlled variable (3.39). We have considered all values $N \in \{4, 6, \ldots, 66\}$ as well as $N = 100$.

is between 5 and 10, whereas it is around 30 or more for the second order method based on (3.39).

We plot on Fig. 3.10 the optimal values of ρ_1, ρ_2 and ρ_3, solution to (3.40). None of these parameters is close to 0: all random variables $A_1^{\eta,N}(\omega)$, $A_2^{\eta,N}(\omega)$ and $C_2^{\eta,N}(\omega)$ are useful in (3.39) to decrease the variance.

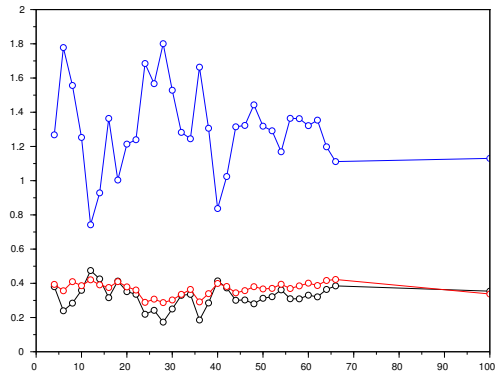

Figure 3.10 – Optimal values of ρ_1 (black), ρ_2 (red) and ρ_3 (blue) for the controlled variable (3.39) as a function of N. We have considered all values $N \in \{4, 6, \ldots, 66\}$ as well as $N = 100$.

On Fig. 3.11, we eventually plot the complete errors, that is

$$e_{N,M}^{\mathrm{MC}} = \left| \frac{1}{M} \sum_{m=1}^{M} A_{\eta,N}^{\star,m}(\omega) - A_\eta^\star \right|, \quad e_{N,M}^{\mathrm{CV}} = \left| \frac{1}{M} \sum_{m=1}^{M} D_{\rho_1,\rho_2,\rho_3}^{3,\eta,m}(\omega) - A_\eta^\star \right|, \tag{3.62}$$

where the exact value $A_\eta^\star$ is actually approximated using M_{ref} realizations on a large domain $Q_{N_{\mathrm{ref}}}$. These errors are a sum of:

- the bias error $\mathbb{E}\left[A_{\eta,N}^\star\right] - A_\eta^\star$,
- the statistical error, which scales as $\sqrt{\mathbb{V}\mathrm{ar}\left(A_{\eta,N}^\star\right)/M}$ for the Monte-Carlo approach and $\sqrt{\mathbb{V}\mathrm{ar}\left(D_{\rho_1,\rho_2,\rho_3}^{3,\eta}\right)/M}$ for the Control Variate approach.

When $d \geq 3$, the variance of $A_{\eta,N}^\star$ has been shown to scale as N^{-d} in [Nol14, Theorem 1.3 and Proposition 1.4]. For homogenization problems set on random *lattices*, optimal estimates on the above two errors have been established in [GNO14, Theorem 2] for any $d \geq 2$: the former scales $N^{-d}(\ln N)^d$ while $\mathbb{V}\mathrm{ar}\left(A_{\eta,N}^\star\right)$ scales as N^{-d}.

In the standard Monte Carlo approach, for large values of N, we expect the statistical error to dominate, and thus the error to be of the order of $N^{-d/2}$. This is indeed what we observe on the blue curve of Fig. 3.11. For the Control Variate approach, we observe that the error decreases as N^{-d} (see red curve of Fig. 3.11). This is consistent with the fact that, for the values of N we consider, the statistical error has been dramatically decreased and is now smaller than the bias error.

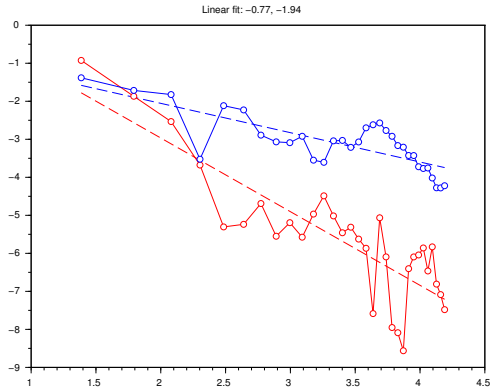

Figure 3.11 – Errors (3.62) as a function of N ($M = 100$; log-log plot). Blue curve (with slope -0.77): Monte-Carlo approach. Red curve (with slope -1.94): Control Variate approach using (3.39). The reference value has been computed using $N_{\text{ref}} = 100$ and $M_{\text{ref}} = 100$.

Acknowledgments

The work of FL and WM is partially supported by ONR under Grant N00014-12-1-0383 and EOARD under grant FA8655-13-1-3061. WM gratefully acknowledges the support from Labex MMCD (Multi-Scale Modelling & Experimentation of Materials for Sustainable Construction) under contract ANR-11-LABX-0022. We also wish to thank Claude Le Bris and Xavier Blanc for enlightning discussions.

Chapter 4

Special Quasirandom Structures: a selection approach for stochastic homogenization

Ce **Chapitre** reprend l'intégralité d'un manuscrit écrit en collaboration avec Claude Le Bris et Frédéric Legoll.

Nous adaptons et étudions une approche de réduction de variance au contexte de l'homogénéisation stochastique. L'approche initiale, utilisée en physique du solide (voir [vPDFN10, WFBZ90, ZWFB90]) consiste à sélectionner les réalisations aléatoires qui satisfont au mieux certaines propriétés statistiques (comme la proportion volumique atteignant le taux de présence de chaque phase dans un matériau biphasique) que l'on attend qu'asymptotiquement.

Nous étudions l'approche théoriquement dans des cas simplifiés (uni dimensionnel, perturbatif en dimension plus grande) et nous démontrons son efficacité dans des cas plus généraux numériquement.

Special Quasirandom Structures: a selection approach for stochastic homogenization

Claude Le Bris[1,3], Frédéric Legoll[2,3] and William Minvielle[1,3]

legoll@lami.enpc.fr, {lebris, william.minvielle}@cermics.enpc.fr

[1] CERMICS, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2, France;

[2] Laboratoire Navier, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal,
77455 Marne-La-Vallée Cedex 2, France ;

[3] INRIA Rocquencourt, MATHERIALS research-team, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France.

Abstract. We adapt and study a variance reduction approach for the homogenization of elliptic equations in divergence form. The approach, borrowed from atomistic simulations and solid-state science [vPDFN10, WFBZ90, ZWFB90], consists in selecting random realizations that best satisfy some statistical properties (such as the volume fraction of each phase in a composite material) usually only obtained asymptotically.

We study the approach theoretically in some simplified settings (one-dimensional setting, perturbative setting in higher dimensions), and numerically demonstrate its efficiency in more general cases.

4.1 Introduction

4.1.1 Overview

In this article, we adapt, theoretically study and numerically test a specific variance reduction approach for the numerical homogenization of an elliptic equation with heterogeneous coefficients.

The equation we consider is the following scalar elliptic equation in divergence form

$$-\operatorname{div}\left(A\left(\frac{\cdot}{\varepsilon},\omega\right)\nabla u^\varepsilon(\cdot,\omega)\right)=f \quad \text{in } \mathcal{D}, \qquad u^\varepsilon(\cdot,\omega)=0 \quad \text{on } \partial\mathcal{D}, \tag{4.1}$$

set on a bounded regular domain $\mathcal{D}$ in $\mathbb{R}^d$ (for some $d \geq 1$), with a function $f \in H^{-1}(\mathcal{D})$ in the right-hand side. The field A is a fixed matrix-valued random field. It is assumed to be uniformly elliptic, uniformly bounded and stationary in a discrete sense. All this is made precise in Section 4.1.2. Since the coefficient ε in (4.1) is assumed small, the coefficient $A\left(\frac{\cdot}{\varepsilon},\omega\right)$ is oscillatory and (4.1) is challenging to solve numerically. On the other hand, the problem is theoretically well understood, as is recalled below.

In the numerical practice, the traditional approach to approximate the solution $u^\varepsilon(\cdot,\omega)$ to (4.1) is to consider (for any $p \in \mathbb{R}^d$), and solve, the so-called corrector problem

$$\begin{cases} -\operatorname{div}\left[A(p+\nabla w_p)\right]=0 \quad \text{in } \mathbb{R}^d \ \text{ almost surely,} \\ \displaystyle\int_Q \mathbb{E}(\nabla w_p)=0, \quad \nabla w_p \text{ is stationary in the sense of (4.5) below,} \end{cases} \tag{4.2}$$

associated to (4.1). The solution to (4.2) gives the deterministic and constant coefficient $A^\star$ of the homogenized equation that in turn serves for the approximation of (4.1). We refer to Section 4.1.2 below for details.

Since (4.2) is a problem set on the entire space $\mathbb{R}^d$, it is necessary to truncate it on a bounded domain, and to complement it with appropriate boundary conditions. In practice, it is standard to consider the problem

$$-\operatorname{div}\left(A(\cdot,\omega)\left(p+\nabla w_p^N(\cdot,\omega)\right)\right)=0, \quad w_p^N(\cdot,\omega) \text{ is } Q_N\text{-periodic}, \tag{4.3}$$

where, say, $Q_N=(0,N)^d$. The deterministic homogenized matrix $A^\star$ is then approximated by the random variable $A_N^\star(\omega)$ defined by

$$\forall p\in\mathbb{R}^d, \quad A_N^\star(\omega)\,p=\frac{1}{|Q_N|}\int_{Q_N}A(\cdot,\omega)(p+\nabla w_p^N(\cdot,\omega)). \tag{4.4}$$

This approximate homogenized coefficient $A_N^\star(\omega)$ is then evaluated using the Monte-Carlo method. Random realizations of the environment, namely the matrix coefficient $A(y,\omega)$, are considered within the truncated domain Q_N. For each of these environments, (4.3) is solved, and $A_N^\star(\omega)$ is computed using (4.4). The homogenized coefficient $A^\star$ is eventually approximated as an empirical mean over several realizations of $A_N^\star(\omega)$. More details are given below in Section 4.1.3.

The purpose of this article is to reduce the variance of the approximation of $A^\star$.

For this purpose, we borrow a variance reduction approach originally introduced in a completely different context, namely that of atomistic simulations for microscopic solid state science. In the series of articles [vPDFN10, WFBZ90, ZWFB90], an approach is indeed described that selects some particular random realizations of the environment, based on some selection criteria derived from asymptotic properties. Intuitively, the approach aims at considering only realizations that, for N fixed, *already* satisfy properties that are usually only obtained in the asymptotic limit $N\to\infty$. The approach carries the name SQS, abbreviation of *Special Quasirandom Structures*.

We aim at adapting this approach to our context, at studying it theoretically in some simple situations, and testing it numerically in more general situations.

For the sake of completeness, we mention that we have already studied the theoretical properties and the practical performance of several variance reduction methods for numerical random homogenization in some previous works of ours. The classical approach of *antithetic variables*, an approach that is quite generic and does not require nor exploit knowledge of the specific structure of the random problem at hand, has been considered in [BCLBL12a, BCLBL12b, CLBL10, LM15b]. The significantly more elaborate (and thus efficient) approach of *control variates* is the subject of [LM15a]. That approach requires a better knowledge of the problem considered, and is not always amenable to fully generic situations.

Our article is articulated as follows.

In the remainder of this introductory section, we present the basics of the theoretical setting (in Section 4.1.2) and of the numerical approximation method (in Section 4.1.3) for the homogenization of the random equation (4.1).

In Section 4.2, we introduce the variance reduction approach we consider. For pedagogic purposes, we first briefly expose the approach in the context of solid state physics it has originally been introduced in. This is the purpose of Section 4.2.1. In Section 4.2.2, we formally derive the specifics of our variance reduction approach using a perturbative

setting. This formal derivation provides the motivation for the general so-called SQS conditions that we use in the sequel of the work. Section 4.2.3 presents how we compute these conditions in practice. Section 4.2.4 contains the pseudo-code of our approach, along with some comments.

The theoretical analysis of the approach is the substance of Section 4.3. We begin by proving, in a fairly general situation (in any ambient dimension), that the approximation provided by our approach (at least the simplest variant of our approach) converges to the homogenized coefficient $A^\star$ when the truncated domain converges to the whole space (see Theorem 4.8 in Section 4.3.1). Next, in Section 4.3.2, we investigate more thoroughly particular and simple situations (such as the one-dimensional setting), where we can indeed completely analyze our approach and actually prove its efficiency.

Our final Section 5.4 contains numerical tests. First, it is often necessary to enforce the desired conditions up to some tolerance (see Remark 4.3 below). In Section 4.4.1, we investigate how this tolerance affects the quality of the approximation and the efficiency of the approach. We observe here that the approach is robust in this respect.

In Section 4.4.2, we illustrate on a prototypical situation the efficiency of our approach. The systematic error is not deteriorated by the approach (it might even be reduced), while the variance is reduced by several orders of magnitude. Such an efficiency is achieved at almost no additional cost with respect to the classical Monte Carlo algorithm.

4.1.2 Theoretical setting

To begin with, we introduce the basic setting of stochastic homogenization. We refer to the seminal works [Koz79, PV81], to [ES08] for a general, numerically oriented presentation and to [BLP78, CD99, JKO94] for classical textbooks. We also refer to [LB10] and the review article [ACLB+12] (and the extensive bibliography therein) for a presentation of our particular setting.

Throughout this article, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ the expectation of any random variable $X \in L^1(\Omega, d\mathbb{P})$. We next fix $d \in \mathbb{N}^\star$ (the ambient physical dimension), and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, that is, for all $k \in \mathbb{Z}^d$ and all $E \in \mathcal{F}$, $\mathbb{P}(\tau_k E) = \mathbb{P}(E)$. We assume that the action τ is *ergodic*, that is, if $E \in \mathcal{F}$ is such that $\tau_k E = E$ for any $k \in \mathbb{Z}^d$, then $\mathbb{P}(E) = 0$ or 1. In addition, we define the following notion of stationarity (see [LB10]): a function $F \in L^1_{\text{loc}}\left(\mathbb{R}^d, L^1(\Omega)\right)$ is *stationary* if

$$\forall k \in \mathbb{Z}^d, \quad F(x+k, \omega) = F(x, \tau_k \omega) \quad \text{a.e. in } x \text{ and a.s.} \tag{4.5}$$

In this setting, the ergodic theorem [Shi84] can be stated as follows:

Let $F \in L^\infty\left(\mathbb{R}^d, L^1(\Omega)\right)$ be a stationary random variable in the above sense. For $k = (k_1, k_2, \ldots, k_d) \in \mathbb{Z}^d$, we set $|k|_\infty = \max_{1\le i\le d} |k_i|$. Then

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

This implies (denoting by $Q = (0,1)^d$ the unit cube in $\mathbb{R}^d$) that

$$F\left(\frac{x}{\varepsilon}, \omega\right) \underset{\varepsilon\to 0}{\overset{\star}{\rightharpoonup}} \mathbb{E}\left(\int_Q F(x, \cdot) dx\right) \quad \textit{in } L^\infty(\mathbb{R}^d), \textit{ almost surely.}$$

Besides technicalities, the purpose of the above setting is simply to formalize that, even though realizations may vary, the function F at point $x \in \mathbb{R}^d$ and the function F at point $x + k$, $k \in \mathbb{Z}^d$, share the same law. In the homogenization context, this means that the local, microscopic environment (encoded in the matrix field A in (4.1)) is everywhere the same *on average*. From this, homogenized, macroscopic properties follow.

We consider problem (4.1), which we recall here for convenience:

$$-\operatorname{div}\left(A\left(\frac{\cdot}{\varepsilon},\omega\right)\nabla u^\varepsilon(\cdot,\omega)\right) = f \quad \text{in } \mathcal{D}, \qquad u^\varepsilon(\cdot,\omega) = 0 \quad \text{on } \partial\mathcal{D}.$$

The random matrix A is assumed stationary in the sense of (4.5). We also assume that A is bounded and coercive, that is, there exist two scalars $0 < c \leq C < \infty$ such that, almost surely,

$$\|A(\cdot,\omega)\|_{L^\infty(\mathbb{R}^d)} \leq C \quad \text{and} \quad \forall \xi \in \mathbb{R}^d, \quad \xi^T A(x,\omega)\xi \geq c\,\xi^T\xi \quad \text{a.e.}$$

In this specific setting, the solution $u^\varepsilon(\cdot,\omega)$ to (4.1) almost surely converges (when ε goes to 0) to the solution $u^\star$ to the homogenized problem

$$-\operatorname{div}\left(A^\star \nabla u^\star\right) = f \quad \text{in } \mathcal{D}, \qquad u^\star = 0 \quad \text{on } \partial\mathcal{D}. \tag{4.6}$$

The convergence of $u^\varepsilon(\cdot,\omega)$ to $u^\star$ holds weakly in $H^1(\mathcal{D})$ and strongly in $L^2(\mathcal{D})$.

The homogenized matrix $A^\star$ in (4.6) is deterministic, and given by an expectation of an integral involving the so-called corrector function, that solves a random auxiliary problem set on the *entire* space. It is given by

$$\forall p \in \mathbb{R}^d, \quad A^\star\, p = \mathbb{E}\left[\int_Q A(x,\cdot)\,\left(p + \nabla w_p(x,\cdot)\right)dx\right], \tag{4.7}$$

where we recall that $Q = (0,1)^d$ and where, for any vector $p \in \mathbb{R}^d$, the *corrector* w_p is the unique solution (up to the addition of a random constant) in $L^2(\Omega; L^2_{\text{loc}}(\mathbb{R}^d))$ with gradient in $L^2(\Omega; L^2_{\text{unif}}(\mathbb{R}^d))^d$ of the corrector problem (4.2). We have used the notation $L^2_{\text{unif}}(\mathbb{R}^d)$ for the *uniform* L^2 space, that is the space of functions for which, say, the L^2 norm on a ball of unit size is bounded from above independently of the center of the ball.

4.1.3 Numerical approximation of the homogenized matrix

As briefly mentioned above, the corrector problem (4.2) is set on the *entire* space $\mathbb{R}^d$, and is therefore challenging to solve. Approximations are in order. In practice, the deterministic matrix $A^\star$ is approximated by the random matrix $A^\star_N(\omega)$ defined by (4.4), which is obtained by solving the corrector problem (4.3) on a *truncated* domain, say the cube $Q_N = (0,N)^d$. Although $A^\star$ itself is a deterministic object, its practical approximation $A^\star_N$ is random. It is only in the limit of infinitely large domains Q_N that the deterministic value is attained. As shown in [BP04], we indeed have

$$\lim_{N\to\infty} A^\star_N(\omega) = A^\star \quad \text{almost surely.} \tag{4.8}$$

As usual, the error $A^\star - A^\star_N(\omega)$ may be expanded as

$$A^\star - A^\star_N(\omega) = \Big(A^\star - \mathbb{E}\left[A^\star_N\right]\Big) + \Big(\mathbb{E}\left[A^\star_N\right] - A^\star_N(\omega)\Big), \tag{4.9}$$

that is the sum of a *systematic* error and of a *statistical* error (the first and second terms in the above right-hand side, respectively).

A standard technique to compute an approximation of $\mathbb{E}\left[A_N^\star\right]$ is to consider M independent and identically distributed realizations of the field A, solve for each of them the corrector problem (4.3) (thereby obtaining i.i.d. realizations $A_N^{\star,m}(\omega)$, for $1 \leq m \leq M$) and compute the Monte Carlo approximation

$$\mathbb{E}\left[\left(A_N^\star\right)_{ij}\right] \approx I_M^{\mathrm{MC}}(\omega) := \frac{1}{M}\sum_{m=1}^{M}\left(A_N^{\star,m}(\omega)\right)_{ij} \tag{4.10}$$

for any $1 \leq i, j \leq d$. In view of the Central Limit Theorem, we know that $\mathbb{E}\left[\left(A_N^\star\right)_{ij}\right]$ asymptotically lies within the confidence interval

$$\left[I_M^{\mathrm{MC}} - 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[\left(A_N^\star\right)_{ij}\right]}}{\sqrt{M}}, I_M^{\mathrm{MC}} + 1.96\frac{\sqrt{\mathbb{V}\mathrm{ar}\left[\left(A_N^\star\right)_{ij}\right]}}{\sqrt{M}}\right]$$

with a probability equal to 95 %.

For simplicity, and because this is overwhelmingly the case in the numerical practice, we have considered in (4.3) *periodic* boundary conditions. These are the conditions we adopt throughout our study. It is to be remarked, however, that other boundary conditions may be employed. Likewise, other slightly modified forms of equation (4.3) may be considered. The specific choice of approximation technique is motivated by considerations about the decrease of the systematic error in (4.9). Several recent mathematical studies have clarified this issue. In addition, in the particular case of periodic boundary conditions (4.3), it has been recently established in [GNO15, Theorem 2] that the statistical error in (4.9) decays like $N^{-d/2}$ while the systematic error in (4.9) scales as $N^{-d}(\log N)^d$. Both estimates have been established for the *discrete variant* of the problem. A similar decay of the statistical error has also been established for the continuous case we consider in the present article (see [GO15, Theorem 1] and [Nol14, Theorem 1.3 and Proposition 1.4]).

4.2 Variance reduction approach

4.2.1 Original formulation of the SQS approach

The variance reduction approach we elaborate upon in this article has been originally introduced for a slightly different purpose in atomistic solid-state science [vPDFN10, WFBZ90, ZWFB90].

In order to convey to the reader the intuition of the original approach, we consider here a simple one-dimensional setting, which nevertheless includes all the difficulties of a generic problem. We consider a linear chain of atomistic sites of two species A and B which interact by an interaction potential V_{AA}, V_{AB} and V_{BB} with obvious notation. For simplicity we consider only nearest neighbour interaction. The atomic sites are occupied by a single species randomly chosen between A and B. A typical random configuration of the "material" therefore reads as an infinite sequence of the type $\cdots ABBAAABBAAAA \cdots$

In order to compute the energy per unit particle of that atomistic system, one has to consider all possible such infinite sequences, and for each of them its normalized energy

$$\lim_{N\to\infty}\frac{1}{2N+1}\sum_{i=-N}^{N} V_{X_{i+1}X_i}, \tag{4.11}$$

where X_i denotes the species present at the i-th site for that particular configuration ($X_i \equiv A$ or B). The "energy" of the system is then defined as the *expectation* of (4.11) over all possible configurations. Other quantities than (4.11) may be considered, or may be simultaneously considered.

In practice, one considers a presumably extremely large, finite N, truncates the infinite sequence over the finite length $2N+1$, and compute

$$\frac{1}{2N+1}\sum_{i=-N}^{N} V_{X_{i+1}X_i}$$

for many (say M, where M is also presumably large) configurations.

The approach introduced in [vPDFN10, WFBZ90, ZWFB90] consists in *selecting* specific configurations $(X_i)_{-N\leq i\leq N}$ of atomic sites that satisfy statistical properties usually obtained only in the limit of infinitely large N.

The first such statistical property is the volume fraction, namely the proportion of species (A, B) present on average. If the sites are all occupied randomly with probability $1/2$ of A and $1/2$ of B (and assuming that all these random variables are independent), then obviously the volume fraction of A is $1/2$ and so is that of B. Then, one only consider truncated sequences $(X_i)_{-N\leq i\leq N}$ that *exactly* reproduce that volume fraction.

Similarly, again for such an evenly distributed proportion of A and B, the energy of the entire infinite system evidently reads as

$$\mathcal{E} = \frac{1}{4}\left[V_{AA} + 2V_{AB} + V_{BB}\right]$$

(recall that we only consider nearest-neighbour interactions). Thus, one only considers truncated sequences $(X_i)_{-N\leq i\leq N}$ which, in addition to exhibiting the exact volume fraction, have an average energy $\frac{1}{2N+1}\sum_{i=-N}^{N} V_{X_{i+1}X_i}$ which is *equal* to $\mathcal{E}$. And so on and so forth for other quantities of interest.

Mathematically, this *selection* of suitable configurations among all the possible configurations classically considered in a Monte-Carlo sample amounts to replacing the computation of an expectation by that of a *conditional expectation*.

The simplistic model we have just considered for pedagogic purposes can of course be replaced by more elaborate models, with more sophisticated quantities to compute, and more demanding statistical quantities to condition the computations with. The bottom line of the approach remains the same, and we adapt it to design a variance reduction approach for numerical random homogenization.

In the next section, we derive the appropriate conditions, which we call the SQS conditions, for our specific context.

4.2.2 Formal derivation of the SQS conditions using a perturbative setting

The purpose of this Section is to formally derive the SQS conditions that we use in the sequel. Such conditions can be easily intuitively understood. We however believe it is interesting to (formally) *derive* them in a particular case. The case we proceed with is a perturbative setting (although, we emphasize it, the conditions will be employed even in the full general, not necessarily perturbative, setting).

We assume throughout this section that the matrix valued coefficient A in (4.1) reads as

$$A_\eta(x,\omega) = C_0(x,\omega) + \eta\ \chi(x,\omega)C_1(x,\omega) \tag{4.12}$$

for some presumably small scalar coefficient η, where

- C_0 and C_1 are two stationary, uniformly bounded matrix fields,
- $C_0 - C_1$ and $C_0 + C_1$ are coercive,
- χ is a stationary scalar field with values in $[-1,1]$.

Under these assumptions, for any $\eta \in (-1,1)$, the matrix A_η is stationary, bounded and coercive. Intuitively, when η is small, A_η is a perturbation of the matrix-valued field $C_0(x,\omega)$.

Remark 4.1. *The expression (4.12) models e.g. a two-phase composite material, where the phases are modelled by the coefficients C_0 and C_1, while χ is the indicator function of the first phase.*

Let $p \in \mathbb{R}^d$. The corrector problem (4.2) reads, in this particular setting, as

$$\begin{cases} -\operatorname{div}\left[(C_0 + \eta\chi C_1)(p + \nabla w_\eta)\right] = 0 \quad \text{in } \mathbb{R}^d, \\ \mathbb{E}\displaystyle\int_Q \nabla w_\eta = 0, \quad \nabla w_\eta \text{ is stationary in the sense of (4.5)}, \end{cases} \tag{4.13}$$

and the homogenized matrix (4.7) is given by

$$\forall p \in \mathbb{R}^d, \quad A^\star_\eta\, p = \mathbb{E}\int_Q A_\eta(p + \nabla w_\eta). \tag{4.14}$$

Note that, for the sake of clarity, we omit to write the dependency of w_η with respect to p.

The truncated version of (4.13) on the supercell Q_N is

$$\begin{cases} -\operatorname{div}\left[(C_0 + \eta\chi C_1)(p + \nabla w^N_\eta)\right] = 0 \quad \text{in } Q_N, \\ w^N_\eta \text{ is } Q_N\text{-periodic}, \end{cases} \tag{4.15}$$

and we approach the homogenized matrix (4.14) by

$$\forall p \in \mathbb{R}^d, \quad A^{\star,N}_\eta(\omega)\, p = \frac{1}{|Q_N|}\int_{Q_N} A_\eta(\cdot,\omega)\left(p + \nabla w^N_\eta(\cdot,\omega)\right). \tag{4.16}$$

Expansion in powers of η

As η goes to 0, we may now expand $A^{\star,N}_\eta$ and $A^\star_\eta$ in powers of η. This expansion is classical (see for instance [BCLBL12b, Cos12]). We only provide it here for the sake of consistency. The corrector expands as

$$\nabla w_\eta = \nabla w_0 + \eta\nabla u_1 + \eta^2\nabla u_2 + o(\eta^2). \tag{4.17}$$

This expansion holds in $L^2(\Omega; L^2_{\text{unif}}(\mathbb{R}^d))$. The functions w_0, u_1 and u_2 appearing in the expansion are respectively defined by the following systems of equations:

$$\begin{cases} -\operatorname{div}\left[C_0(p + \nabla w_0)\right] = 0 \quad \text{in } \mathbb{R}^d, \\ \mathbb{E}\displaystyle\int_Q \nabla w_0 = 0, \quad \nabla w_0 \text{ is stationary}, \end{cases} \tag{4.18}$$

$$\begin{cases} -\operatorname{div}[C_0\nabla u_1] = \operatorname{div}[\chi C_1(p+\nabla w_0)] \quad \text{in } \mathbb{R}^d, \\ \mathbb{E}\displaystyle\int_Q \nabla u_1 = 0, \quad \nabla u_1 \text{ is stationary}, \end{cases} \tag{4.19}$$

and

$$\begin{cases} -\operatorname{div}[C_0\nabla u_2] = \operatorname{div}[\chi C_1\nabla u_1] \quad \text{in } \mathbb{R}^d, \\ \mathbb{E}\displaystyle\int_Q \nabla u_2 = 0, \quad \nabla u_2 \text{ is stationary}. \end{cases}$$

Inserting the expansion (4.12) of A_η and (4.17) of w_η in (4.14), we obtain

$$A_\eta^\star = A_0^\star + \eta A_1^\star + \eta^2 A_2^\star + o(\eta^2), \tag{4.20}$$

with, for any $p \in \mathbb{R}^d$,

$$\begin{aligned} A_0^\star p &= \mathbb{E}\left[\int_Q C_0(p+\nabla w_0)\right], \\ A_1^\star p &= \mathbb{E}\left[\int_Q \chi C_1(p+\nabla w_0)\right] + \mathbb{E}\left[\int_Q C_0\nabla u_1\right], \\ A_2^\star p &= \mathbb{E}\left[\int_Q \chi C_1\nabla u_1\right] + \mathbb{E}\left[\int_Q C_0\nabla u_2\right]. \end{aligned} \tag{4.21}$$

Likewise, we expand w_η^N as

$$\nabla w_\eta^N = \nabla w_0^N + \eta\nabla u_1^N + \eta^2\nabla u_2^N + o(\eta^2),$$

with

$$\begin{cases} -\operatorname{div}\left[C_0(p+\nabla w_0^N)\right] = 0 \quad \text{in } Q_N, \\ w_0^N \text{ is } Q_N\text{-periodic}, \end{cases} \tag{4.22}$$

$$\begin{cases} -\operatorname{div}\left[C_0\nabla u_1^N\right] = \operatorname{div}\left[\chi C_1(p+\nabla w_0^N)\right] \quad \text{in } Q_N, \\ u_1^N \text{ is } Q_N\text{-periodic}, \end{cases} \tag{4.23}$$

and

$$\begin{cases} -\operatorname{div}\left[C_0\nabla u_2^N\right] = \operatorname{div}\left[\chi C_1\nabla u_1^N\right] \quad \text{in } Q_N, \\ u_2^N \text{ is } Q_N\text{-periodic}. \end{cases}$$

The homogenized matrix $A_\eta^{\star,N}(\omega)$ therefore satisfies

$$\left|A_\eta^{\star,N}(\omega) - \left[A_0^{\star,N}(\omega) + \eta A_1^{\star,N}(\omega) + \eta^2 A_2^{\star,N}(\omega)\right]\right| \le C\eta^3, \tag{4.24}$$

where C is independent of η, N and ω, and where the matrices $A_0^{\star,N}(\omega)$, $A_1^{\star,N}(\omega)$ and $A_2^{\star,N}(\omega)$ are defined by

$$\begin{aligned} A_0^{\star,N}(\omega)\,p &= \frac{1}{|Q_N|}\int_{Q_N} C_0(p+\nabla w_0^N), \\ A_1^{\star,N}(\omega)\,p &= \frac{1}{|Q_N|}\int_{Q_N} \chi C_1(p+\nabla w_0^N) + \frac{1}{|Q_N|}\int_{Q_N} C_0\nabla u_1^N, \\ A_2^{\star,N}(\omega)\,p &= \frac{1}{|Q_N|}\int_{Q_N} \chi C_1\nabla u_1^N + \frac{1}{|Q_N|}\int_{Q_N} C_0\nabla u_2^N. \end{aligned} \tag{4.25}$$

SQS conditions

In line with the motivation we have mentioned above in Section 4.1.3, we are now in position to introduce the conditions that we use to *select* particular configurations of the environment within Q_N for which we compute the solution to (4.15), and, in turn, compute the approximation (4.16) of $A^\star_\eta$. Our conditions are based upon the comparison of (4.21) and (4.25).

Definition 4.2. *For finite fixed N, we say that an environment $\omega \in \Omega$ satisfies the SQS condition of*

- *order 0 if $A_0^{\star,N}(\omega) = A_0^\star$, that is to say, for any $p \in \mathbb{R}^d$,*

$$\frac{1}{|Q_N|}\int_{Q_N} C_0(\cdot,\omega)(p+\nabla w_0^N(\cdot,\omega)) = \mathbb{E}\left[\int_Q C_0(p+\nabla w_0)\right], \tag{4.26}$$

- *order 1 if $A_1^{\star,N}(\omega) = A_1^\star$, that is to say, for any $p \in \mathbb{R}^d$,*

$$\frac{1}{|Q_N|}\int_{Q_N}\left[\chi(\cdot,\omega)C_1(\cdot,\omega)(p+\nabla w_0^N(\cdot,\omega)) + C_0(\cdot,\omega)\nabla u_1^N(\cdot,\omega)\right] = \mathbb{E}\left[\int_Q \chi C_1(p+\nabla w_0) + C_0\nabla u_1\right], \tag{4.27}$$

- *order 2 if $A_2^{\star,N}(\omega) = A_2^\star$, that is to say, for any $p \in \mathbb{R}^d$,*

$$\frac{1}{|Q_N|}\int_{Q_N}\left[\chi(\cdot,\omega)C_1(\cdot,\omega)\nabla u_1^N(\cdot,\omega) + C_0(\cdot,\omega)\nabla u_2^N(\cdot,\omega)\right] = \mathbb{E}\left[\int_Q \chi C_1\nabla u_1 + C_0\nabla u_2\right]. \tag{4.28}$$

Remark 4.3. *In full generality, we* do not *claim that there exist environments that satisfy these conditions. This might be the case that no such environment exists. One may for instance simply remark that a random variable that takes value -1 and $+1$ both with probability $1/2$ never has value zero, which is its expectation! In some situations, we therefore have to* relax *the above conditions (see Section 4.2.4 below), but we temporarily leave these technicalities aside and* assume *that suitable environments exist.*

Consider now the two expansions (4.20) and (4.24). It is immediate to see, by subtraction, that

$$A_\eta^{\star,N}(\omega) - A_\eta^\star = (A_0^{\star,N}(\omega) - A_0^\star) + \eta(A_1^{\star,N}(\omega) - A_1^\star) + \eta^2(A_2^{\star,N}(\omega) - A_2^\star) + o(\eta^2).$$

Therefore it is readily seen that, if the configuration ω satisfies the SQS conditions of Definition 4.2 up to the order k included ($k = 0$, 1, 2 in our definition, but clearly one could consider higher order conditions derived likewise), then

$$A_\eta^{\star,N}(\omega) - A_\eta^\star = o(\eta^k), \tag{4.29}$$

where the constant in the right-hand side is independent of η, N and ω. Taking the expectation over such configurations therefore formally provides a more accurate approximation of $A^\star_\eta$.

Now that we have derived the conditions (4.26)–(4.27)–(4.28) (which we henceforth call the *SQS conditions*) in the perturbative setting, we will actually use them in the non perturbative setting, namely for a similar two-phase composite material, but with η *not* small. Of course, a property like (4.29) cannot be expected any longer since the homogenized matrix $A^\star$ is no longer a polynomial in a small coefficient that encodes a perturbation. Nevertheless, it can be expected that selecting the configurations using these conditions may improve the approximation, in particular by reducing the variance. We show in Sections 4.3 and 5.4 that it is indeed the case, theoretically and experimentally.

For the time being, we need to make a *practical* observation. The right-hand side of conditions (4.26)–(4.27)–(4.28) need to be evaluated in order to practically encode the SQS conditions. In principle, the computation of those right-hand sides are exact expectations, that can only be determined using an asymptotic limit, and are therefore almost as challenging to compute in practice as $A^\star$ itself.

We therefore need to restrict the generality of our setting (4.12) and consider cases where those right-hand sides are indeed amenable to a simple, inexpensive computation. This is the purpose of the next section.

4.2.3 Practical evaluation of the SQS conditions

In order to make our approach practical, we need, as mentioned above, to consider settings where the expectations present in the right-hand sides of (4.26)–(4.27)–(4.28) may be computed effectively.

Condition of order 0

We first consider (4.26) and its right-hand side

$$\mathbb{E}\left[\int_Q C_0(x,\cdot)(p+\nabla w_0(x,\cdot))dx\right]. \tag{4.30}$$

A natural assumption, which already covers a large portion of practically relevant situations, is

$$C_0(x,\omega) = C_0(x) \quad \text{is a deterministic, } \mathbb{Z}^d\text{-periodic matrix.} \tag{4.31}$$

The computation of (4.30) is then inexpensive since the solution w_0 to (4.18) is in fact the deterministic solution to

$$-\operatorname{div}\left[C_0(p+\nabla w_0)\right] = 0 \ \text{ in } \mathbb{R}^d, \quad w_0 \text{ is } \mathbb{Z}^d\text{-periodic},$$

which is unique up to the addition of a constant.

In addition, when N is an integer (and when the approximation chosen for (4.2) is the *periodic* approximation (4.3), as is indeed the case throughout this work), the solution to (4.22) is $w_0^N \equiv w_0$ (up to an additive constant), and hence the condition (4.26) is systematically satisfied.

We henceforth assume that (4.31) holds, that N is an integer, and we proceed with the periodic approximation (4.3).

Condition of order 1

We next consider the SQS condition (4.27). One possible assumption to make that condition practical is

$$C_0(x,\omega) = C_0 \quad \text{is a deterministic, } \textit{constant} \text{ matrix.} \tag{4.32}$$

Since $\nabla w_0 = 0$, the right-hand side of (4.27) reads

$$\mathbb{E}\left[\int_Q \chi C_1(p+\nabla w_0) + C_0 \nabla u_1\right] = \int_Q \mathbb{E}\left[\chi C_1\right] p + C_0 \mathbb{E}\int_Q \nabla u_1,$$

where the rightmost term vanishes in view of (4.19) and where the first term of the right-hand side may be computed using only characteristic properties of the environment considered. The condition (4.27) thus reads

$$\frac{1}{|Q_N|}\int_{Q_N} \chi(\cdot,\omega) C_1(\cdot,\omega) = \mathbb{E}\left[\int_Q \chi C_1\right]. \tag{4.33}$$

For instance, in a two-phase composite material mixing two *constant* and *deterministic* matrices C_0 and C_1, we have

$$\mathbb{E}\left[\int_Q \chi C_1\right] = \mathbb{E}\left[\int_Q \chi\right] C_1.$$

This quantity obviously only depends on the *volume fraction* of the two phases (recall (4.12)). Proceeding likewise with the left-hand side of the condition (4.27), we see that this condition reads

$$\frac{1}{|Q_N|}\int_{Q_N} \chi(x,\omega)dx = \mathbb{E}\left[\int_Q \chi\right].$$

Interestingly (and not unexpectedly), we notice here that this condition on the volume fraction agrees with the condition we used to consider in the simple atomistic system of Section 4.2.1.

Condition of order 2

We next proceed with condition (4.28). In addition to (4.32), we assume that

$$C_1(x,\omega) = C_1(x) \quad \text{is a deterministic, } \mathbb{Z}^d\text{-periodic matrix,} \tag{4.34}$$

and that

$$\chi(y,\omega) = \sum_{k\in\mathbb{Z}^d} X_k(\omega)\mathbb{1}_{Q+k}(y), \tag{4.35}$$

where X_k are identically distributed scalar random variables taking their values in $[-1,1]$. We also assume that

$$\mathcal{C} = \sum_{k\in\mathbb{Z}^d} |\mathbb{C}\mathrm{ov}(X_0, X_k)| < \infty, \tag{4.36}$$

which is obviously satisfied if X_k are independent one from each other.

We then have the following result, which will be useful to make condition (4.28) practical. Its proof is postponed until Appendix 4.5.

Lemma 4.4. *Under the assumptions (4.32), (4.34), (4.35) and (4.36), the solution u_1 to (4.19) satisfies*

$$\nabla u_1(y,\omega) = \mathbb{E}[X_0]\nabla\overline{u_1}(y) + \sum_{k\in\mathbb{Z}^d}\Big(X_k(\omega) - \mathbb{E}[X_k]\Big)\nabla\phi_1(y-k), \tag{4.37}$$

where ϕ_1 is the (unique up to the addition of a constant) solution in $\{v \in L^2_{\mathrm{loc}}(\mathbb{R}^d),\ \ \nabla v \in (L^2(\mathbb{R}^d))^d\}$ to

$$-\operatorname{div}\left[C_0\nabla\phi_1\right] = \operatorname{div}\left[\mathbb{1}_Q C_1 p\right] \quad \text{in } \mathbb{R}^d \tag{4.38}$$

and $\overline{u_1}$ is the (unique up to the addition of a constant) solution to

$$-\operatorname{div}\left[C_0\nabla\overline{u_1}\right] = \operatorname{div}\Big[C_1 p\Big] \quad \text{in } \mathbb{R}^d, \quad \overline{u_1} \text{ is } \mathbb{Z}^d\text{-periodic.} \tag{4.39}$$

The sum in (4.37) is a convergent series in $L^2(Q\times\Omega)$.

Using simpler arguments, we see that the solution u_1^N to (4.23) satisfies

$$\nabla u_1^N(y,\omega) = \mathbb{E}[X_0]\nabla\overline{u_1}(y) + \sum_{k\in\mathbb{Z}^d\cap Q_N}\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)\nabla\phi_1^N(y-k), \tag{4.40}$$

where $\overline{u_1}$ is defined by (4.39) and ϕ_1^N is the (unique up to the addition of a constant) solution to

$$-\operatorname{div}\left[C_0\nabla\phi_1^N\right] = \operatorname{div}\left[\mathbb{1}_Q C_1 p\right] \quad \text{in } Q_N, \quad \phi_1^N \text{ is } Q_N\text{-periodic.} \tag{4.41}$$

In practice, we can easily obtain an accurate approximation of ϕ_1 since the right-hand side of (4.38) has compact support. Truncating (4.38) over a sufficiently large bounded domain (with homogeneous Dirichlet boundary conditions) provides such an accurate approximation. Given (4.32), the right-hand side of Condition (4.28) rewrites $\mathbb{E}\left[\int_Q \chi C_1\nabla u_1\right]$ since $\mathbb{E}\left[\int_Q \nabla u_2 = 0\right]$. In view of (4.37), this quantity is in turn expanded as

$$\begin{aligned}
&\mathbb{E}\left[\int_Q \chi C_1\nabla u_1\right]\\
=\;& (\mathbb{E}[X_0])^2\int_Q C_1\nabla\overline{u_1} + \sum_{k\in\mathbb{Z}^d}\mathbb{E}\left[\int_Q X_0\Big(X_k-\mathbb{E}[X_k]\Big)C_1(y)\nabla\phi_1(y-k)dy\right]\\
=\;& (\mathbb{E}[X_0])^2\int_Q C_1\nabla\overline{u_1}\\
&+\sum_{k\in\mathbb{Z}^d}\mathbb{E}\left[\int_Q \Big(X_0-\mathbb{E}[X_0]\Big)\Big(X_k-\mathbb{E}[X_k]\Big)C_1(y)\nabla\phi_1(y-k)dy\right],
\end{aligned} \tag{4.42}$$

where, as mentioned above, $\nabla\phi_1$ can be easily and accurately computed, while the series in $k\in\mathbb{Z}^d$ may be truncated in an efficient manner because of the rapid decay at infinity of $\nabla\phi_1$ (see [BCLBL12b, Lemma 3.1]).

We correspondingly expand the left-hand side of (4.28). The second term vanishes,

while the first term reads, in view of (4.40),

$$\begin{aligned}
&\frac{1}{|Q_N|}\int_{Q_N}\chi(y,\omega)C_1(y)\nabla u_1^N(y,\omega)\,dy\\
=&\sum_{j\in\mathbb{Z}^d\cap Q_N}\frac{1}{|Q_N|}\int_{Q_N}X_j(\omega)\mathbb{1}_{Q+j}(y)C_1(y)\mathbb{E}[X_0]\nabla\overline{u_1}(y)dy\\
+&\sum_{k,j\in\mathbb{Z}^d\cap Q_N}\frac{1}{|Q_N|}\int_{Q_N}X_j(\omega)\mathbb{1}_{Q+j}(y)C_1(y)\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)\nabla\phi_1^N(y-k)dy\\
=&(\mathbb{E}[X_0])^2\int_Q C_1(y)\nabla\overline{u_1}(y)dy\\
+&\mathbb{E}[X_0]\left(\frac{1}{|Q_N|}\sum_{j\in\mathbb{Z}^d\cap Q_N}\Big(X_j(\omega)-\mathbb{E}[X_j]\Big)\right)\int_Q C_1(y)\nabla\overline{u_1}(y)dy\\
+&\mathbb{E}[X_0]\sum_{k\in\mathbb{Z}^d\cap Q_N}\frac{1}{|Q_N|}\int_{Q_N}C_1(y)\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)\nabla\phi_1^N(y-k)dy\\
+&\sum_{k,j\in\mathbb{Z}^d\cap Q_N}\frac{1}{|Q_N|}\int_{Q+j}\Big(X_j(\omega)-\mathbb{E}[X_j]\Big)C_1(y)\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)\nabla\phi_1^N(y-k)dy.
\end{aligned}\tag{4.43}$$

In this particular (however still very generic) setting, we infer from (4.42) and (4.43) that Condition (4.28) reads as

$$\begin{aligned}
&\frac{1}{|Q_N|}\sum_{k,j\in Q_N\cap\mathbb{Z}^d}\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)\Big(X_j(\omega)-\mathbb{E}[X_j]\Big)I_{k,j}^N\\
&\qquad+\frac{1}{|Q_N|}\mathbb{E}[X_0]\sum_{k\in Q_N\cap\mathbb{Z}^d}\Big(X_k(\omega)-\mathbb{E}[X_k]\Big)I_k^N=\sum_{k\in\mathbb{Z}^d}\mathbb{C}\mathrm{ov}(X_0,X_k)I_k^\infty,
\end{aligned}\tag{4.44}$$

where

$$I_k^\infty = \int_{Q+k}C_1(y)\nabla\phi_1(y),\tag{4.45}$$

$$I_{k,j}^N = \int_{Q+j}C_1(y)\nabla\phi_1^N(y-k)dy,\tag{4.46}$$

$$I_k^N = \int_{Q_N}C_1(y)\nabla\phi_1^N(y-k)dy+\int_Q C_1(y)\nabla\overline{u_1}(y)dy.\tag{4.47}$$

Summary

In the prototypical case where

$$A(x,\omega)=C_0+\chi(x,\omega)C_1(x),$$

where C_0 is constant, C_1 is $\mathbb{Z}^d$ periodic and χ takes the form (4.35) (and where we consider the periodic approximation (4.3) of (4.2)), we have that:

- The condition (4.26) (SQS condition of order 0) is systematically fulfilled.
- In view of (4.33) and (4.35), the condition (4.27) (SQS condition of order 1) rewrites as

$$\frac{1}{|Q_N|}\sum_{k\in\mathbb{Z}^d\cap Q_N}X_k(\omega)=\mathbb{E}\left[X_0\right].\tag{4.48}$$

- In view of (4.44), the condition (4.28) (SQS condition of order 2) writes as

$$\frac{1}{|Q_N|}\sum_{k,j\in Q_N\cap\mathbb{Z}^d}\overline{X}_k(\omega)\overline{X}_j(\omega)I^N_{k,j} + \frac{1}{|Q_N|}\mathbb{E}[X_0]\sum_{k\in Q_N\cap\mathbb{Z}^d}\overline{X}_k(\omega)I^N_k = \sum_{k\in\mathbb{Z}^d}\mathbb{C}\mathrm{ov}(X_0,X_k)I^\infty_k, \quad (4.49)$$

  where $\overline{X}_k(\omega) = X_k(\omega) - \mathbb{E}[X_k]$.

The conditions (4.48) and (4.49) are henceforth called the SQS 1 and SQS 2 conditions, respectively.

Remark 4.5. *If (4.48) is satisfied, then the coefficient I^N_k in (4.49) can be replaced by*

$$\overline{I}^N_k = \int_{Q_N} C_1(y)\nabla\phi^N_1(y-k)dy$$

and there is no need to compute $\overline{u_1}$.

4.2.4 Selection Monte Carlo sampling

We are now in position to describe the selection Monte Carlo sampling we employ. We recall that the classical Monte Carlo sampling reads as follows:

Algorithm 1 (Classical Monte Carlo).
For $m = 1,\dots,M$,

1. *Generate a random environment ω_m.*
2. *Solve the truncated corrector problem (4.3).*
3. *Compute $A^\star_N(\omega_m)$.*

Compute the estimator $\mathcal{I}^M_{MC} = \dfrac{1}{M}\displaystyle\sum_{m=1}^M A^\star_N(\omega_m)$ for $A^\star$.

In contrast, our selection Monte Carlo sampling algorithm, in the particular case described in Section 4.2.3, reads as follows:

Algorithm 2.
The algorithm requires a tolerance tol > 0, *fixed by the user.*

1. **Offline stage**
   (a) *Solve the equation (4.38).*
   (b) *Compute $(I^\infty_k)_{k\in\mathbb{Z}^d}$ defined by (4.45).*
   (c) *Compute the right-hand side of the SQS conditions (4.48) and (4.49).*
   (d) *Solve the equations (4.39) and (4.41).*
   (e) *Compute $(I^N_{k,j})_{k,j\in\mathbb{Z}^d\cap Q_N}$ and $(I^N_k)_{k\in\mathbb{Z}^d\cap Q_N}$ defined by (4.46) and (4.47).*
2. **Online stage**
   For $m = 1,\dots,M$,

(a) *Generate a random environment* ω_m.

(b) *Using* $I^N_{k,j}$ *and* I^N_k*, compute the left-hand sides of (4.48) and (4.49).*

(c) *If the left-hand sides differ from the right-hand sides by more than* tol*, return to Step 2a.*

(d) *Solve the truncated corrector problem (4.3).*

(e) *Compute* $A^\star_N(\omega_m)$.

Compute the estimator $\mathcal{I}^M_{SQS} = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} A^\star_N(\omega_m)$ *for* $A^\star$.

Remark 4.6. *As pointed out above, the series in* $k \in \mathbb{Z}^d$ *in the right-hand side of (4.49) may be truncated in an efficient manner because of the rapid decay at infinity of* $\nabla\phi_1$*. Therefore only a few factors* I^∞_k *have to be computed at Step 1b.*

Remark 4.7. *When several SQS conditions (in practice SQS 1 and SQS 2) have to be simultaneously satisfied, we simply add them up using some weighting parameter. We have not observed any particular sensitivity of our numerical results (collected in Section 5.4 below) with respect to the adjustment of this parameter, provided it remains not too close to 0 and 1.*

We have already mentioned that, in many situations, there might not be *any* random environments that satisfy some, or all, of the SQS conditions (4.26)–(4.27)–(4.28) we wish to enforce. Therefore, some adaptation is in order, and we have used in Algorithm 2 a tolerance parameter tol > 0 for the SQS conditions to be satisfied.

However, if these conditions are enforced within some given tolerance as in Algorithm 2, the following issue arises. Since the motivation for precisely considering the SQS conditions is that they are fulfilled *asymptotically*, the larger the truncated computational domain we consider (that is, the larger N), the less restrictive the conditions are, and therefore the less effective the variance reduction is likely to be. To circumvent this difficulty, a first possibility is to consider a tolerance that decreases when the size of Q_N increases. We consider this variant in our theoretical study of Section 4.3.2 below (see formula (4.69)). More precisely, we require in Proposition 4.14 that

$$\text{the SQS condition is satisfied with the tolerance } \frac{\lambda}{\sqrt{|Q_N|}}$$

for some λ. In practice, implementing such a threshold is not an easy matter, as the rate and the constants need to be adequately adjusted. In order to avoid such technicalities, we prefer to take a slightly different perspective, the purpose of which is to always select a *fixed proportion* of the original sample of the $\mathcal{M}$ environments drawn. Practically, we pick the M configurations that best satisfy the SQS conditions, for some M fixed (given $\mathcal{M}$).

The practical algorithm we employ is therefore as follows:

Algorithm 3 (**Selection Monte Carlo sampling**).
The algorithm requires a number of trials $\mathcal{M}$*, fixed by the user.*

1. **Offline stage 1**: *same as the offline stage of Algorithm 2.*

2. **Offline stage 2: selection step**
   For $m = 1, \dots, \mathcal{M}$,

(a) *Generate a random environment* ω_m.

(b) *Using* $I^N_{k,j}$ *and* I^N_k, *compute the left-hand sides of (4.48) and (4.49).*

(c) *Compute the error* error_m *between the left-hand sides and the right-hand sides of (4.48) and (4.49).*

Sort the random environments $(\omega_m)_{1\leq m\leq \mathcal{M}}$ *according to* error_m. *Keep the* M *best realizations, and reject the others.*

3. **Online stage: resolution**
For $m = 1, \ldots, M$,

(a) *Solve the truncated corrector problem (4.3).*

(b) *Compute* $A^\star_N(\omega_m)$.

Compute the estimator $\mathcal{I}^M_{SQS} = \dfrac{1}{M}\displaystyle\sum_{m=1}^{M} A^\star_N(\omega_m)$ *for* $A^\star$.

We wish to make a couple of comments about this selection Monte Carlo approach.

In full generality, the cost of Monte Carlo approaches is usually dominated by the cost of draws, and therefore selection algorithms are targeted to reject as few draws as possible.

In the present context, where boundary value problems such as (4.3) are to be solved repeatedly, the cost of draws for the environment is negligible in front of the cost of the solution procedure for such boundary value problems. Likewise, evaluating the quantities present in e.g. (4.49) is not expensive. Therefore, the purpose of the selection mechanism is to limit the number of boundary value problems to be solved, even though this comes at the (tiny) price of rejecting many environments. This also explains why we employ a simplistic rejection procedure for the selection, while in other situations of Monte Carlo samplings, one would invest in a more clever selection procedure.

A second observation is that, as potentially for any selection procedure, our selection introduces a bias (i.e. a modification of the systematic error in (4.9)). The point is to ensure that the gain in variance superseeds the bias introduced by the variance reduction approach.

Our next section addresses some theoretical aspects of our approach.

4.3 Elements of theoretical analysis

This section contains some elements of analysis that we are able to provide. We begin with a (somewhat) general result of convergence, and next, in some simplified cases, study our approach more thoroughly.

4.3.1 Proof of convergence of the approach

Formally, our approach consists in replacing an empirical average provided by the classical Monte Carlo approach to compute $\mathbb{E}[A^\star_N]$ by an empirical average *restricted* to some environments within Q_N satisfying some additional condition(s) (see Section 4.2.4). We work at a fixed size N of the truncation domain Q_N and recall that $A^\star_N(\omega)$ is defined by (4.4). Mathematically, our approach amounts to considering conditional expectations of the type $\mathbb{E}[A^\star_N \mid \text{SQS}]$, where SQS encodes that one, or several, of the conditions summarized in (4.48)–(4.49) are satisfied.

The least we can expect from our approach is that it converges to the correct limit when $N \to \infty$, namely $A^\star$, as in (4.8).

The theorem we now state establishes this fact. In order to prove it, we need to make some assumptions on our setting (see the details below), and also to make specific the SQS conditions we use. In Theorem 4.8 below, we specifically use the SQS 1 condition, in the form (4.48).

In order to state a result as general as possible, we therefore consider a condition that reads $\frac{1}{|Q_N|} \sum_{k \in \mathbb{Z}^d \cap Q_N} f(X_k) = \mathbb{E}[f(X_0)]$ for some function f. In practice, our specific SQS 1 condition (4.48) corresponds to the choice $f(x) = x$.

Theorem 4.8. *Let $(X_k)_{k \in \mathbb{Z}^d}$ be a sequence of independent and identically distributed scalar random variables following a common law μ. We assume that μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$, and that, for any $k \in \mathbb{Z}^d$, $X_k(\omega) \in [-1, 1]$ almost surely. We consider the stationary random field*

$$A(y, \omega) = C_0 + \sum_{k \in \mathbb{Z}^d} X_k(\omega)\, \mathbb{1}_{Q+k}(y)\, C_1(y),$$

where C_0 is constant and C_1 is $\mathbb{Z}^d$-periodic and bounded. We also assume that $C_0 + C_1(y)$ and $C_0 - C_1(y)$ are uniformly coercive, and that C_0 and C_1 are symmetric.

Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a measurable function with compact level sets. We assume that f is not constant. Then we have

$$\mathbb{E}\left[A_N^\star \;\middle|\; \frac{1}{|Q_N|} \sum_{k \in Q_N \cap \mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)] \right] \xrightarrow[N \to \infty]{} A^\star, \tag{4.50}$$

where $A_N^\star(\omega)$ is defined by (4.4) and $A^\star$ is defined by (4.7).

Some remarks are in order.

Remark 4.9. *As is the case throughout this article, we have considered the* periodic *approximation (4.3) of (4.2). The proof of Theorem 4.8 actually carries over to the case of Neumann or Dirichlet boundary conditions, or any alternate truncation problem that provides some $A^{\star,N}(\omega)$ such that $A^{\star,N}_{\mathrm{Neu}}(\omega) \le A^{\star,N}(\omega) \le A^{\star,N}_{\mathrm{Dir}}(\omega)$ (see additional details in [Min15, Appendix]).*

Remark 4.10. *The assumptions regarding independence of the X_k, absolute continuity of their common law with respect to the Lebesgue measure and compactness of the level sets of f are necessary for technical reasons, since we need to apply a general result from [BO14]. See below for details.*

The proof of Theorem 4.8 is based on the following result, which is a particular case of a more general result due to C. Bernardin and S. Olla (see [BO14, Theorem B.2.2]):

Theorem 4.11 (C. Bernardin and S. Olla, [BO14])**.** *Consider n scalar random variables $X_1, \ldots, X_n$, that are independent and that all share the same probability distribution $\mu(x)\,dx$ on $\mathbb{R}$. Consider a measurable function $f : \mathbb{R} \mapsto \mathbb{R}$, which is assumed to be not constant and to have compact level sets. Let $f_0 = \mathbb{E}[f(X_1)] = \int_{\mathbb{R}} f(x)\,\mu(x)\,dx$. Consider also a bounded and continuous function $F : \mathbb{R}^k \mapsto \mathbb{R}$. Then*

$$\lim_{n \to \infty} \mathbb{E}\left[F(X_1, \ldots, X_k) \;\middle|\; \frac{1}{n} \sum_{i=1}^{n} f(X_i) = f_0 \right] = \mathbb{E}\left[F(X_1, \ldots, X_k) \right]. \tag{4.51}$$

Note that, when $n \to \infty$, the quantity $\frac{1}{n}\sum_{i=1}^{n} f(X_i)$ almost surely converges to f_0. Theorem 4.11 shows that conditioning on the manifold $\frac{1}{n}\sum_{i=1}^{n} f(X_i) = f_0$ does not change the value (when $n \to \infty$) of the expectation of a function F of a *finite* number k of random variables.

In our context, the variable X_i is the value of the field A on the cell $Q + i$. The conditioning in the left-hand side of (4.51) is identical to the conditioning in the left-hand side of (4.50).

The difference between Theorem 4.11 and our result lies in the quantity of which we compute the expectation. In our case, this quantity is $A^\star_N(\omega)$, which is (asymptotically when $N \to \infty$) a function of *all* the variables X_i and not only of a *finite* number of them. We hence cannot directly use Theorem 4.11. The proof of our result essentially amounts to introducing an upper bound and a lower bound on $A^\star_N(\omega)$ that both read as a sum of functions that depend on a *finite* number of random variables (see e.g. (4.54) below). We will then be in position to apply Theorem 4.11 on these functions.

Proof of Theorem 4.8. We fix some $p \in \mathbb{R}^d$. For the sake of clarity, the approximate homogenized matrix $A^\star_N(\omega)$ defined by (4.4) is here denoted $A^{\star,N}_{\rm per}(\omega)$, to emphasize that we have considered periodic boundary conditions. Since the matrix A is symmetric, we have

$$p^T A^{\star,N}_{\rm per}(\omega) p = \inf \left\{ \mathcal{J}_{Q_N}(v,\omega), \quad v \in H^1_{\rm per}(Q_N) \right\},$$

where

$$\mathcal{J}_{Q_N}(v,\omega) = \frac{1}{|Q_N|} \int_{Q_N} (p + \nabla v)^T A(\cdot,\omega)(p + \nabla v).$$

We have considered in (4.3) periodic boundary conditions. As is well-known, other boundary conditions can be used, and these alternate approximations will be useful for the proof.

Step 1: Upper bound. We first introduce an approximation of $A^\star$ using a truncated corrector problem complemented with homogeneous Dirichlet boundary conditions. We consider the problem

$$\left\{ \begin{array}{c} -\operatorname{div}\Big(A(\cdot,\omega)\Big(p + \nabla w^N_{p,\rm Dir}(\cdot,\omega)\Big)\Big) = 0 \ \text{ in } Q_N, \\ w^N_{p,\rm Dir}(\cdot,\omega) = 0 \ \text{ on } \partial Q_N, \end{array} \right.$$

which yields an approximation of $A^\star$ that we denote $A^{\star,N}_{\rm Dir}(\omega)$ and which is defined by

$$\forall p \in \mathbb{R}^d, \quad A^{\star,N}_{\rm Dir}(\omega) p = \frac{1}{|Q_N|} \int_{Q_N} A(\cdot,\omega)(p + \nabla w^N_{p,\rm Dir}(\cdot,\omega)).$$

As shown in [BP04], we know that

$$\lim_{N\to\infty} A^{\star,N}_{\rm Dir}(\omega) = A^\star \quad \text{a.s.} \tag{4.52}$$

Since A is symmetric, we have

$$p^T A^{\star,N}_{\rm Dir}(\omega) p = \inf \left\{ \mathcal{J}_{Q_N}(v,\omega), \quad v \in H^1_0(Q_N) \right\}.$$

The matrix $A^{\star,N}_{\mathrm{Dir}}(\omega)$ is always larger (in the sense of symmetric matrices) than $A^{\star,N}_{\mathrm{per}}(\omega)$. Indeed, let $v \in H^1_0(Q_N)$, and consider its Q_N-periodic extension $\widetilde{v}$. Then this function belongs to $H^1_{\mathrm{per}}(Q_N)$. We hence have that

$$p^T A^{\star,N}_{\mathrm{per}}(\omega)p \leq \mathcal{J}_{Q_N}(\widetilde{v},\omega) = \mathcal{J}_{Q_N}(v,\omega).$$

Minimizing over $v \in H^1_0(Q_N)$, we get that

$$p^T A^{\star,N}_{\mathrm{per}}(\omega)p \leq p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \quad \text{a.s.} \tag{4.53}$$

Just as $A^{\star,N}_{\mathrm{per}}(\omega)$, the matrix $A^{\star,N}_{\mathrm{Dir}}(\omega)$ depends on all the random variables $X_i(\omega)$, $i \in Q_N \cap \mathbb{Z}^d$. But, thanks to the use of homogeneous Dirichlet boundary conditions, it can be bounded from above by a sum of matrices that depend only on a finite number of random variables. To show this, we proceed as follows.

For any positive integers N and R, we introduce the integer part M of N/R. Then Q_N can decomposed into a set of cubes of size R^d, up to some boundary layer $B_{N,R}$:

$$Q_N = \Big(\cup_{j\in\mathbb{Z}^d,\, |j|\leq M} R\,j + Q_R \Big) \cup B_{N,R}.$$

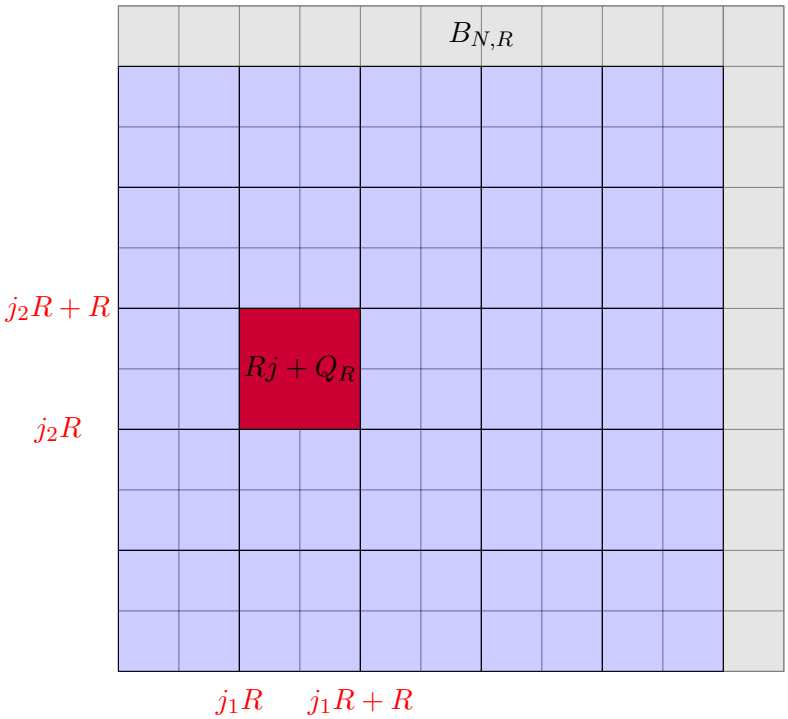

Figure 4.1 – The supercell Q_N (here represented for $N = 11$) is split into cells of size R^d (here $R = 2$), up to some boundary layer $B_{N,R}$.

For any $j \in \mathbb{Z}^d$, $|j| \leq M$, consider a function $v_j \in H^1_0(R\,j + Q_R)$. We now define the function v on Q_N as:

- for any $x \in R\,j + Q_R$, we set $v(x) = v_j(x)$;
- if $x \in B_{N,R}$, we set $v(x) = 0$.

The function v belongs to $H^1_0(Q_N)$. We hence write that

$$p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \leq \mathcal{J}_{Q_N}(v,\omega) = \frac{|Q_R|}{|Q_N|} \sum_{j\in\mathbb{Z}^d,\, |j|\leq M} \mathcal{J}_{Rj+Q_R}(v_j,\omega).$$

Minimizing over the functions $v_j \in H^1_0(Rj+Q_R)$, we hence get that

$$p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \leq \frac{|Q_R|}{|Q_N|} \sum_{j\in\mathbb{Z}^d,\, |j|\leq M} Y_j(\omega) \quad \text{a.s.} \tag{4.54}$$

where

$$Y_j(\omega) = \inf\left\{\mathcal{J}_{Rj+Q_R}(v,\omega), \quad v \in H^1_0(Rj+Q_R)\right\}.$$

Since A is stationary, we note that all the random variables $Y_j(\omega)$ share the same law. Moreover, we observe that $Y_0(\omega) = p^T A^{\star,R}_{\mathrm{Dir}}(\omega)p$, which is the approximation of the homogenized matrix using Dirichlet boundary conditions on Q_R.

We now take the conditional expectation of (4.54), and use the fact that the variables Y_j all share the same law:

$$\mathbb{E}\left[p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right]$$
$$\leq \frac{R^d\, M^d}{N^d}\, \mathbb{E}\left[p^T A^{\star,R}_{\mathrm{Dir}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right].$$

We now observe that $p^T A^{\star,R}_{\mathrm{Dir}}(\omega)p$ only depends on a *finite* number of random variables, namely only on $X_k(\omega)$ with $k \in Q_R \cap \mathbb{Z}^d$. We are thus in position to use Theorem 4.11, which yields the limit of the above right-hand side when $N \to \infty$. Hence, for any fixed R, we have

$$\limsup_{N\to\infty} \mathbb{E}\left[p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \leq \mathbb{E}\left[p^T A^{\star,R}_{\mathrm{Dir}}(\omega)p\right].$$

Letting R go to ∞ in the above bound and using (4.52), we obtain that

$$\limsup_{N\to\infty} \mathbb{E}\left[p^T A^{\star,N}_{\mathrm{Dir}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \leq p^T A^\star p.$$

Using (4.53), we deduce that

$$\forall p \in \mathbb{R}^d, \quad \limsup_{N\to\infty} p^T U_N p \leq p^T A^\star p, \tag{4.55}$$

where

$$U_N = \mathbb{E}\left[A^{\star,N}_{\mathrm{per}}(\omega) \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right]. \tag{4.56}$$

Step 2: Lower bound. We now introduce an approximation of $A^\star$ using a truncated problem complemented with homogeneous Neumann boundary conditions. We consider the problem

$$\left\{\begin{array}{l} -\operatorname{div}\Big(A(\cdot,\omega)\left(p + \nabla w^N_{p,\mathrm{Neu}}(\cdot,\omega)\right)\Big) = 0 \;\text{ in } Q_N, \\ n^T A(\cdot,\omega)(p + \nabla w^N_{p,\mathrm{Neu}}(\cdot,\omega)) = n^T p \;\text{ on } \partial Q_N, \end{array}\right. \tag{4.57}$$

which yields an approximation of $A^\star$ that we denote $A^{\star,N}_{\rm Neu}(\omega)$ and which is defined by

$$A^{\star,N}_{\rm Neu}(\omega) = \left(S^{\star,N}_{\rm Neu}(\omega)\right)^{-1}, \tag{4.58}$$

where $S^{\star,N}_{\rm Neu}(\omega)$ is defined by

$$\forall p \in \mathbb{R}^d, \quad S^{\star,N}_{\rm Neu}(\omega)p = \frac{1}{|Q_N|}\int_{Q_N} p + \nabla w^N_{p,\rm Neu}(\cdot,\omega). \tag{4.59}$$

See Remark 4.12 below for some heuristic justification of (4.58)–(4.59).

As recalled in [Min15, Appendix], we have that

$$\lim_{N\to\infty} A^{\star,N}_{\rm Neu}(\omega) = A^\star \quad \text{a.s.} \tag{4.60}$$

and

$$p^T A^{\star,N}_{\rm Neu}(\omega)p \le p^T A^{\star,N}_{\rm per}(\omega)p \quad \text{a.s.} \tag{4.61}$$

In addition, we have the following variational characterization:

$$p^T S^{\star,N}_{\rm Neu}(\omega)p = \inf\left\{\mathcal{E}_{Q_N}(\sigma,\omega), \quad \sigma \in V(Q_N)\right\}, \tag{4.62}$$

where

$$\mathcal{E}_{Q_N}(\sigma,\omega) = \frac{1}{|Q_N|}\int_{Q_N} (p+\sigma)^T A^{-1}(\cdot,\omega)(p+\sigma)$$

and

$$V(Q_N) = \left\{\sigma \in (L^2(Q_N))^d, \quad \operatorname{div}\sigma = 0 \text{ in } Q_N, \quad n^T\sigma = 0 \text{ on } \partial Q_N\right\}.$$

The matrix $S^{\star,N}_{\rm Neu}(\omega)$ (and hence the matrix $A^{\star,N}_{\rm Neu}(\omega)$) depends on all the variables $X_i(\omega)$, $i \in Q_N \cap \mathbb{Z}^d$. However, thanks to the characterization (4.62), it can be bounded from above by a sum of matrices that depend only on a finite number of random variables.

To show this, we proceed as in Step 1 of the proof. For any positive integers N and R, we introduce the integer part M of N/R, and decompose Q_N into a set of cubes of size R^d, up to some boundary layer $B_{N,R}$ (see Figure 4.1):

$$Q_N = \left(\cup_{j\in\mathbb{Z}^d,\,|j|\le M} R\,j + Q_R\right) \cup B_{N,R}.$$

For any $j \in \mathbb{Z}^d$, $|j| \le M$, consider a function $\sigma_j \in V(R\,j + Q_R)$. We now define the function σ on Q_N as:

- for any $x \in R\,j + Q_R$, we set $\sigma(x) = \sigma_j(x)$;
- if $x \in B_{N,R}$, we set $\sigma(x) = 0$.

We claim that $\sigma \in V(Q_N)$. We indeed first have that $\sigma \in (L^2(Q_N))^d$. We next consider $\varphi \in C^\infty_0(Q_N)$ and compute that

$$\begin{aligned}
\langle \operatorname{div}\sigma, \varphi\rangle &= -\langle \sigma, \nabla\varphi\rangle \\
&= -\sum_{j\in\mathbb{Z}^d,\,|j|\le M} \int_{R\,j+Q_R} \sigma_j \cdot \nabla\varphi \\
&= -\sum_{j\in\mathbb{Z}^d,\,|j|\le M} \int_{\partial(R\,j+Q_R)} n_j^T\sigma_j\ \varphi \\
&= 0,
\end{aligned}$$

where n_j is the outward normal to the domain $Rj + Q_R$. We hence have checked that $\sigma \in V(Q_N)$.

We next write that

$$p^T S^{\star,N}_{\mathrm{Neu}}(\omega)p \leq \mathcal{E}_{Q_N}(\sigma,\omega) = \frac{|Q_R|}{|Q_N|} \sum_{j\in\mathbb{Z}^d,\, |j|\leq M} \mathcal{E}_{Rj+Q_R}(\sigma_j,\omega).$$

Minimizing over the functions $\sigma_j \in V(Rj + Q_R)$, we hence get that

$$p^T S^{\star,N}_{\mathrm{Neu}}(\omega)p \leq \frac{|Q_R|}{|Q_N|} \sum_{j\in\mathbb{Z}^d,\, |j|\leq M} Z_j(\omega) \quad \text{a.s.} \tag{4.63}$$

where

$$Z_j(\omega) = \inf\left\{\mathcal{E}_{Rj+Q_R}(\sigma,\omega), \quad \sigma \in V(Rj+Q_R)\right\}.$$

Since A is stationary, we note that all the random variables $Z_j(\omega)$ share the same law. Moreover, we observe that $Z_0(\omega) = p^T S^{\star,R}_{\mathrm{Neu}}(\omega)p$.

We now take the conditional expectation of (4.63), and use the fact that the variables Z_j all share the same law:

$$\begin{aligned}\mathbb{E}\left[p^T S^{\star,N}_{\mathrm{Neu}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \\ \leq \frac{R^d M^d}{N^d}\mathbb{E}\left[p^T S^{\star,R}_{\mathrm{Neu}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right].\end{aligned}$$

We observe that $p^T S^{\star,R}_{\mathrm{Neu}}(\omega)p$ only depends on a *finite* number of random variables, namely only on X_k with $k \in Q_R \cap \mathbb{Z}^d$. We are thus in position to use Theorem 4.11, which yields the limit of the above right-hand side when $N \to \infty$. Hence, for any fixed R, we have

$$\limsup_{N\to\infty} \mathbb{E}\left[p^T S^{\star,N}_{\mathrm{Neu}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \leq \mathbb{E}\left[p^T S^{\star,R}_{\mathrm{Neu}}(\omega)p\right].$$

Letting R go to ∞ in the above bound and using (4.58) and (4.60), we obtain that

$$\limsup_{N\to\infty} \mathbb{E}\left[p^T S^{\star,N}_{\mathrm{Neu}}(\omega)p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \leq p^T (A^\star)^{-1} p.$$

Using (4.58) and (4.61), we deduce that

$$\limsup_{N\to\infty} \mathbb{E}\left[p^T \left(A^{\star,N}_{\mathrm{per}}(\omega)\right)^{-1} p \;\middle|\; \frac{1}{|Q_N|}\sum_{k\in Q_N\cap\mathbb{Z}^d} f(X_k) = \mathbb{E}[f(X_0)]\right] \leq p^T (A^\star)^{-1} p.$$

Using Jensen inequality, we infer from the above bound that

$$\forall p \in \mathbb{R}^d, \quad \limsup_{N\to\infty} p^T \left(U_N\right)^{-1} p \leq p^T (A^\star)^{-1} p, \tag{4.64}$$

where the matrix U_N is defined by (4.56).

Step 3: Conclusion. We eventually show that (4.55) and (4.64) imply that U_N converges to $A^\star$ when $N \to \infty$.

From the assumptions on A, we know that there exists $0 < a_- \leq a_+ < \infty$ such that, for any N and almost surely, $a_- \leq A^{\star,N}_{\rm per}(\omega) \leq a_+$. Hence, for any N, the symmetric matrix U_N satisfies $a_- \leq U_N \leq a_+$. We can thus extract a subsequence $U_{\varphi(N)}$ that converges to some symmetrix matrix B. Let us show that $B = A^\star$.

Let $p \in \mathbb{R}^d$. We first observe that, by definition,

$$\limsup_{k\to\infty} \; p^T U_k p \geq \lim_{k\to\infty} \; p^T U_{\varphi(k)} p = p^T B p.$$

We thus infer from (4.55) that

$$\forall p \in \mathbb{R}^d, \quad p^T B p \leq p^T A^\star p. \tag{4.65}$$

We now proceed likewise with U_k^{-1}. We observe that,

$$\limsup_{k\to\infty} \; p^T U_k^{-1} p \geq \lim_{k\to\infty} \; p^T U_{\varphi(k)}^{-1} p = p^T B^{-1} p.$$

We thus infer from (4.64) that

$$\forall p \in \mathbb{R}^d, \quad p^T B^{-1} p \leq p^T (A^\star)^{-1} p. \tag{4.66}$$

Collecting (4.65) and (4.66), we deduce that $B = A^\star$.

The sequence U_N is bounded, and we have shown that any converging subsequence converges to $A^\star$. This implies that U_N converges to $A^\star$ when $N \to \infty$, which is exactly the result (4.50). This concludes the proof of Theorem 4.8. □

Remark 4.12. *In view of (4.57), we can check that*

$$\frac{1}{|Q_N|} \int_{Q_N} A(\cdot,\omega) \left(p + \nabla w^N_{p,\rm Neu}(\cdot,\omega)\right) = p.$$

The definition (4.58)–(4.59) can hence be understood as

$$\Big\langle A(\cdot,\omega) \left(p + \nabla w^N_{p,\rm Neu}(\cdot,\omega)\right) \Big\rangle = A^{\star,N}_{\rm Neu}(\omega) \; \Big\langle p + \nabla w^N_{p,\rm Neu}(\cdot,\omega) \Big\rangle,$$

where $\langle \cdot \rangle = |Q_N|^{-1} \displaystyle\int_{Q_N} \cdot$ *is the average on* Q_N.

4.3.2 Complete analysis in some simple cases

In this section, we aim at improving the convergence result (4.50) of the previous section by quantifying both the statistical and systematic errors, in order to assess the efficiency of our approach. We are only able to proceed in simple situations where all the quantities are indeed accessible using analytic calculations. These two situations are examined in Sections 4.3.2 and 4.3.2 respectively. For the sake of brevity, and because the proofs are not very enlightening and are not likely to carry over to more general cases, we do not provide the proofs of our claims here. We refer to [Min15] where they are presented in details.

We establish below that our approach preserves the *rate* of decay of the standard Monte Carlo sampling both for the systematic and the statistical error (and thus, in particular, the systematic error remains, in rate, smaller than the statistical error). Furthermore, the *prefactor* in the statistical error is significantly reduced by our approach.

"Zero-dimensional" homogenization

As simplest possible situation, we consider a function $f : \mathbb{R} \mapsto \mathbb{R}$ and the random variables $(X_i)_{1\le i\le n}$. We assume that these random variables are independent and that they are all centered Gaussian random variables with unit variance. We also assume that $f \in C^1(\mathbb{R})$ and that $\mathbb{E}\left[|f(X_1)| + |f'(X_1)|\right] < \infty$. Note that it is not surprising to make some smoothness assumptions on f as we are here after *rates* of convergence, and not only a convergence result as in Section 4.3.1.

We set

$$\xi : x \mapsto \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Assume we want to compute $\mathbb{E}[f(X_1)]$. A classical Monte Carlo approach would approximate this by the limit of the empirical mean $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i(\omega))$. In this particular instance, the simplest version of our variance reduction approach instead considers $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i(\omega))$ for realizations $X(\omega)$ that satisfy $\xi(X(\omega)) = 0$.

In this simple case, the bias of the classical approach is actually identically zero: of course, $\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]$ does not depend on n. The statistical error is controlled by the Central Limit Theorem and is asymptotically of order $\sqrt{\frac{\mathbb{V}\mathrm{ar}[f(X_1)]}{n}}$.

Proposition 4.13. *Under the assumptions of this section, the bias of the selection method is of order $1/n$. More specifically,*

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) \;\middle|\; \xi(X) = 0\right] - \mathbb{E}[f(X_1)] = -\frac{1}{2n}\mathbb{E}[f'(X_1)] + O\left(\frac{1}{n^2}\right). \tag{4.67}$$

The variance of the selection method is reduced by a factor independent of n. More specifically,

$$\frac{\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) \;\middle|\; \xi(X) = 0\right]}{\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]} = 1 - \frac{(\mathbb{E}[f'(X_1)])^2}{\mathbb{V}\mathrm{ar}[f(X_1)]} + O\left(\frac{1}{n}\right). \tag{4.68}$$

In view of (4.67)–(4.68), we observe that, in spite of introducing a bias of order $O(1/n)$, our approach reduces the statistical error from $\frac{\lambda_{\mathrm{MC}}}{\sqrt{n}}$ to $\frac{\lambda_{\mathrm{SQS}}}{\sqrt{n}}$ (with $\lambda_{\mathrm{SQS}} < \lambda_{\mathrm{MC}}$), and therefore, for sufficiently large n, reduces the total error.

The following result covers the case where we insert a non-zero tolerance in Algorithm 2.

Proposition 4.14. *Under the assumptions of this section, consider the selection method where we condition on the realizations such that $\frac{z_0}{\sqrt{n}} \le \xi(X(\omega)) \le \frac{z_1}{\sqrt{n}}$, for some z_0 and $z_1 > z_0$ in $\mathbb{R}$. Then, for any choice of z_0 and $z_1 > z_0$, the variance of the selection method is reduced by a factor independent of n:*

$$\frac{\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) \;\middle|\; \frac{z_0}{\sqrt{n}} \le \xi(X) \le \frac{z_1}{\sqrt{n}}\right]}{\mathbb{V}\mathrm{ar}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right]} = 1 - (1-C)\frac{(\mathbb{E}[f'(X_1)])^2}{\mathbb{V}\mathrm{ar}[f(X_1)]} + O\left(\frac{1}{n}\right), \tag{4.69}$$

where $C = \mathbb{V}\mathrm{ar}\left[X_1 \;\middle|\; z_0 \leq X_1 \leq z_1\right]$.

The conditioning $z_0/\sqrt{n} \leq \xi(X) \leq z_1/\sqrt{n}$ is deliberately chosen in order to match the rate of the Central Limit Theorem. It corresponds to the selection of a fixed *proportion* of samples (as in Algorithm 3 when $\mathcal{M}$ is proportional to M). Note that $C > 0$, hence the variance is less reduced than when conditioning at $\xi(X) = 0$ (which is the case considered in Proposition 4.13). Note also that the variance is reduced (with respect to the standard Monte Carlo sampling) if, and only if, $1 - C \geq 0$. We are yet unable to conclude that this is the case in general. We simply note that, when $z_1 = -z_0 > 0$, then $C = 1$, yielding no gain.

One-dimensional homogenization

In the one-dimensional case, the homogenization of a random field $a : (y, \omega) \mapsto \sum_{i \in \mathbb{Z}} g(X_i(\omega)) \mathbb{1}_{(i,i+1)}(y)$ (where g is valued, say, in $[a_-, a_+]$ with $a_- > 0$) is a simple harmonic average. It is readily seen that

$$a_N^\star(\omega) = \left(\frac{1}{N}\sum_{i=1}^N \frac{1}{g(X_k)}\right)^{-1} = \varphi\left(\frac{1}{N}\sum_{i=1}^N \frac{1}{g(X_k)}\right) \quad \text{with } \varphi(x) = 1/x.$$

Formally, the problem is thus analogous to that of the previous section, for a certain $\varphi : \mathbb{R} \mapsto \mathbb{R}$ instead of $\varphi = \mathrm{Id}$. Therefore, it is sufficient to prove consistency and variance reduction for quantities of the form $\varphi\left(\frac{1}{N}\sum_{i=1}^N f(X_i)\right)$.

Proposition 4.15. *Consider a smooth function $\varphi : \mathbb{R} \mapsto \mathbb{R}$. Under the assumptions of this section, the bias of the standard method and that of the selection method respectively are*

$$\mathbb{E}\left[\varphi\left(\frac{1}{N}\sum_{i=1}^N f(X_i)\right)\right] - \varphi(f_0) = \frac{\varphi''(f_0)}{2N}\mathbb{V}\mathrm{ar}[f(X_1)] + O\left(\frac{1}{N^2}\right) \tag{4.70}$$

and

$$\begin{aligned}\mathbb{E}&\left[\varphi\left(\frac{1}{N}\sum_{i=1}^N f(X_i)\right) \;\middle|\; \xi(X) = 0\right] - \varphi(f_0) \\ &= \frac{\varphi''(f_0)}{2N}\left(\mathbb{V}\mathrm{ar}[f(X_1)] - (\mathbb{E}[f'(X_1)])^2\right) - \frac{\varphi'(f_0)}{2N}\mathbb{E}[X_1 f'(X_1)] + o\left(\frac{1}{N}\right),\end{aligned} \tag{4.71}$$

with $f_0 = \mathbb{E}[f(X_1)]$.

The variance of the selection method is reduced by a factor independent of N:

$$\frac{\mathbb{V}\mathrm{ar}\left[\varphi\left(\frac{1}{N}\sum_{i=1}^N f(X_i)\right) \;\middle|\; \xi(X) = 0\right]}{\mathbb{V}\mathrm{ar}\left[\varphi\left(\frac{1}{N}\sum_{i=1}^N f(X_i)\right)\right]} = 1 - \frac{(\mathbb{E}[f'(X_1)])^2}{\mathbb{V}\mathrm{ar}[f(X_1)]} + o(1). \tag{4.72}$$

To keep things simple, we do not investigate whether a more general result, accounting for some tolerance in the manner our condition is fulfilled (in the spirit of Proposition 4.14), holds here.

Proposition 4.15 shows that the bias is unchanged in rate, while the prefactor for the variance is reduced. Since the variance only decays at the rate $1/\sqrt{N}$ while the bias decays

at the rate $1/N$, we see that our approach indeed reduces the total error for sufficiently large N.

In the numerical practice (mimicking in this one-dimensional setting what is actually performed for higher dimensional settings – although it is in some sense unnecessary here), we generate several, independent realizations of the N-tuples $(X_i)_{1\leq i\leq N}$ corresponding to as many draws of environments within the "cube" Q_N. In the classical Monte Carlo approach, we keep all such N-tuples. In our approach, we only consider those that satisfy an additional criterion.

An empirical mean (aimed at approximating $A^\star$) is then computed. The systematic error and the statistical error of the latter approximation are precisely related to the errors estimated in (4.70)–(4.71)–(4.72) respectively. Thus a theoretical assessment of our practical approach.

4.4 Numerical experiments

We first present in this section some numerical experiments that show the robustness of our variance reduction approach with respect to the tolerance with which we enforce the SQS conditions (see Section 4.4.1). We next turn to studying the performance of our approach in Section 4.4.2.

We consider the test-case when A reads as in (4.12), that is

$$A_\eta(x,\omega) = C_0(x,\omega) + \eta\ \chi(x,\omega)C_1(x,\omega),$$

with $\eta = 0.5$, $C_0 = C_1 = \mathrm{Id}$, and χ is of the form (4.35), that is

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

The random variables X_k are i.i.d. and follow a Bernoulli law of parameter $1/2$ valued in $\{-1,+1\}$. The contrast (i.e. the ratio of the largest value of A divided by its minimum value) is equal to 3. The influence of the contrast on the efficiency of our approach is investigated at the end of Section 4.4.2 (see Table 4.1). We consider there much larger values of the contrast (however all smaller than 20).

In what follows, we only consider Algorithm 3, where we take $M = 100$ and $\mathcal{M} = 2000$ (thus an acceptance rate of 5%).

In this setting, the SQS 1 condition as stated in (4.48) is satisfied if and only if the numbers of cells within which $X_k(\omega) = 1$ is equal to the number of cells within which $X_k(\omega) = -1$. We enforce this by randomly selecting $|Q_N|/2$ cells within the $|Q_N|$ cells that are in Q_N, and setting $X_k = 1$ on these cells and $X_k = -1$ on the others.

In all our tests, we have kept the computational time fixed, or almost fixed, since the additional time needed by the selection step (namely Steps 1 and 2 of Algorithm 3) is roughly 5% of the total original computational time.

4.4.1 Robustness of the approach

As pointed out above, the SQS 2 condition as stated in (4.49) is only enforced in Algorithm 3 up to some tolerance. In this section, we experimentally investigate how this tolerance affects the quality of the approximation and the efficiency of the approach. To mimick the difficulty associated with the SQS 2 condition, we have also performed some

tests where we only enforce the SQS 2 condition up to some tolerance, and not exactly. The results of our numerical tests are displayed in Figures 4.2 through 4.5.

Figures 4.2 and 4.3 show the sensitivity of the variance reduction ratio upon the first order condition (4.48). Following Algorithm 3, we have sorted the realizations with respect to the error in (4.48). On Figure 4.2, the left-most dot displays the ratio $V_{\text{SQS 1}}/V_{\text{MC}}$ between the empirical variance $V_{\text{SQS 1}}$ among the best $M = 100$ realizations and the reference Monte Carlo variance V_{MC}. The second dot shows the ratio between the empirical variance among the next best $M = 100$ realizations and the reference Monte Carlo variance. We next proceed with the subsequent groups of $M = 100$ realizations. We work with $\mathcal{M} = 2000$, hence there are 20 groups of 100 realizations, and hence 20 dots on Figure 4.2. On Figure 4.3, we display the same ratio of variances, but the x axis provides (for each group of $M = 100$ realizations) the maximum error with which the first order condition (4.48) is fulfilled (rather than the index of that group, as on Figure 4.2). Hence, the first group (left-most dot) corresponds to a vanishing error, the second group corresponds to an error between 0 and tol, the third group corresponds to an error between tol and 2 tol, and so on and so forth.

Figures 4.4 and 4.5 show the sensitivity upon the second order condition (4.49). Here, the first order condition (4.48) is directly embedded into the random environment generator, so that every realization that is considered actually satisfies (4.48). Following Algorithm 3, we have sorted the realizations with respect to the error in (4.49). We present the results on Figures 4.4 and 4.5 following the same procedure as for Figures 4.2 and 4.3. We plot the ratio $\dfrac{V_{\text{SQS 2}}}{V_{\text{exact SQS 1}}}$ between the variance $V_{\text{SQS 2}}$ among the $M = 100$ realizations that exactly satisfy the SQS 1 condition and best satisfy the SQS 2 condition on the one hand, and, on the other hand, the variance $V_{\text{exact SQS 1}}$ of the realizations that exactly satisfy the SQS 1 condition.

We conclude that our approach is robust in this respect. Even if the SQS conditions (4.48)–(4.49) are not *exactly* satisfied, but only with some small tolerance, we obtain a significant variance reduction.

We next investigate how much the two SQS conditions (4.48) and (4.49) are independent (in a probabilistic sense) one from the other. Results are shown on Figure 4.6. The rightmost histogram in black is the distribution of the criterion SQS 1 (namely, the left-hand side of (4.48)) among all realizations (we have considered $M = 100$ samples). The left histogram in black is the distribution of the criterion SQS 2 (namely, the left-hand side of (4.49)) among all realizations (we have used the same samples for both black histograms). Finally, the left histogram in blue is the conditional distribution of the criterion SQS 2 among $M = 100$ realizations that exactly satisfy the criterion SQS 1. These two histograms on the left of the figure are sufficiently close to each other to state that conditioning with respect to SQS 1 does not change the distribution of the SQS 2 criterion. Therefore, enforcing first the condition (4.48) does not make challenging the subsequent selection of realizations that best satisfy (4.49).

4.4.2 Efficiency of the approach

In this section, we investigate how the efficiency of our approach depends (i) on the size of the truncated domain Q_N and (ii) on the contrast in A.

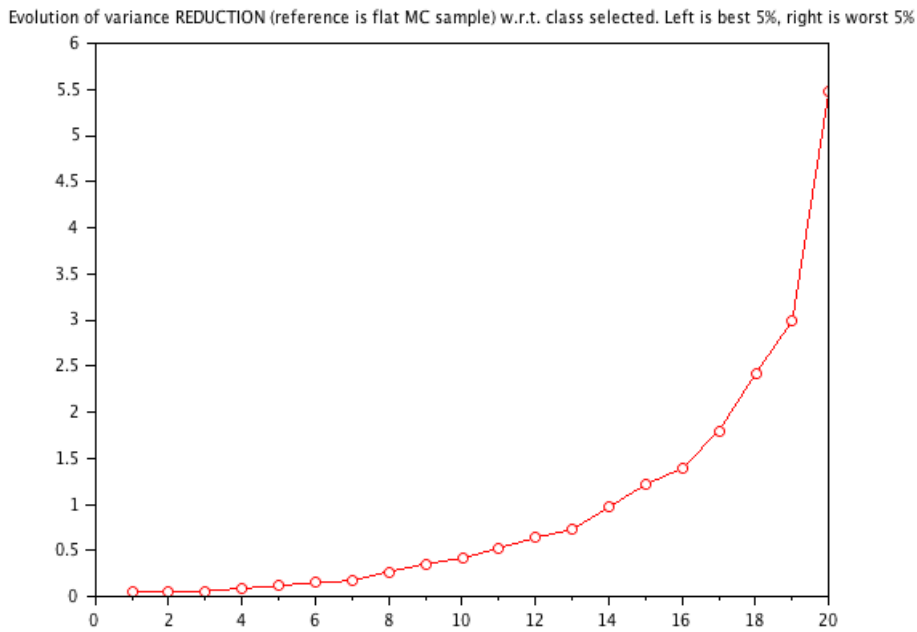

Figure 4.2 – Variance ratio $V_{\text{SQS }1}/V_{\text{MC}}$ for the 20 groups of realizations (sorted according to their SQS 1 error)

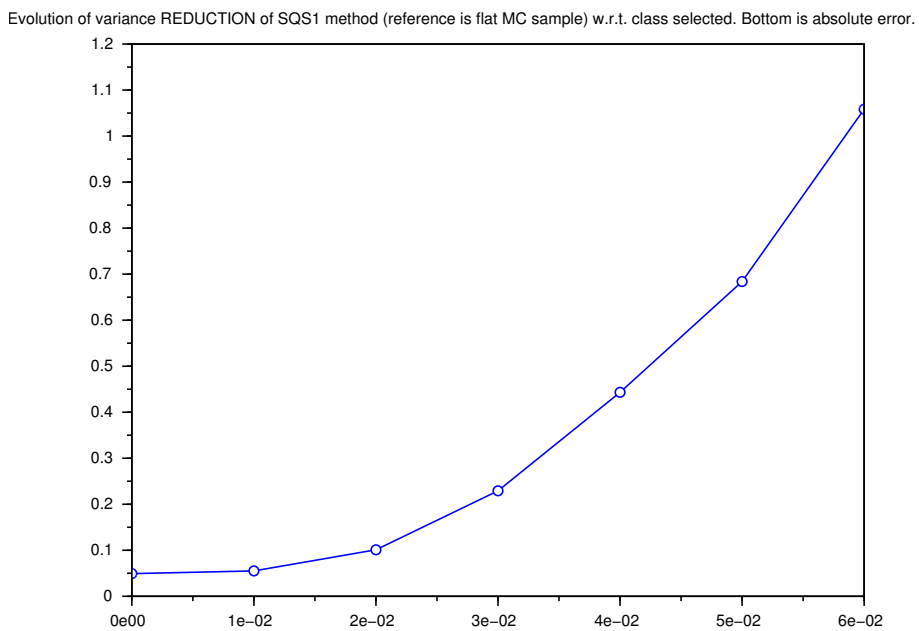

Figure 4.3 – Variance ratio $V_{\text{SQS }1}/V_{\text{MC}}$ as a function of the error in (4.48)

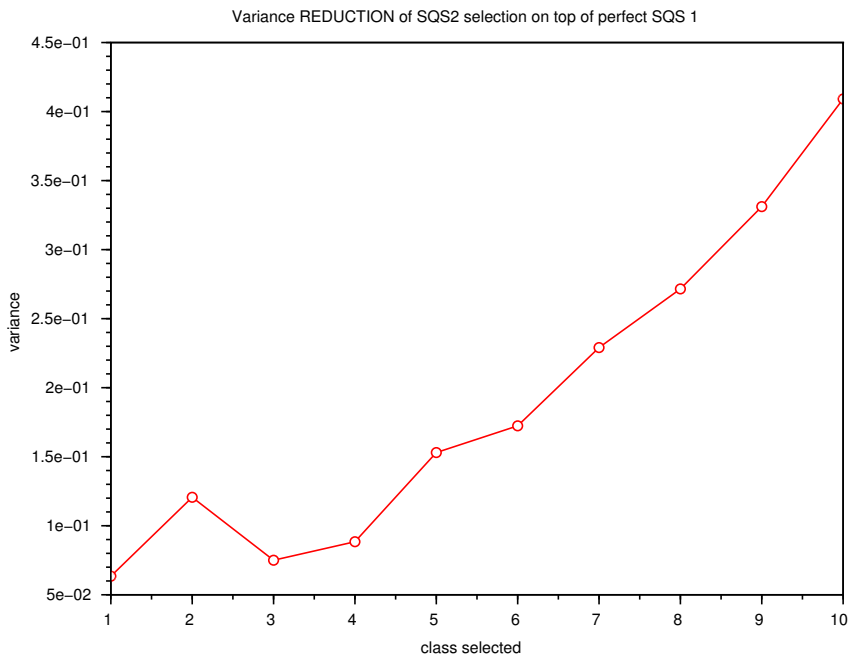

Figure 4.4 – Variance ratio $\frac{V_{\text{SQS 2}}}{V_{\text{exact SQS 1}}}$ for the different groups of realizations (sorted according to their SQS 2 error; the SQS 1 condition is exactly satisfied). Only the 10 best groups are shown.

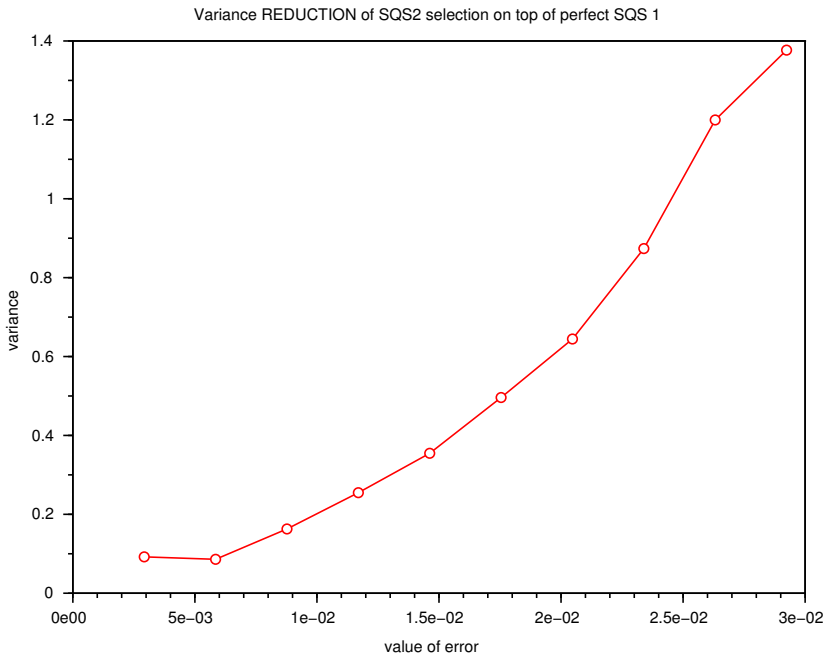

Figure 4.5 – Variance ratio $\dfrac{V_{\text{SQS 2}}}{V_{\text{exact SQS 1}}}$ as a function of the error in (4.49) (the condition (4.48) is exactly satisfied). Only the 10 best groups are shown.

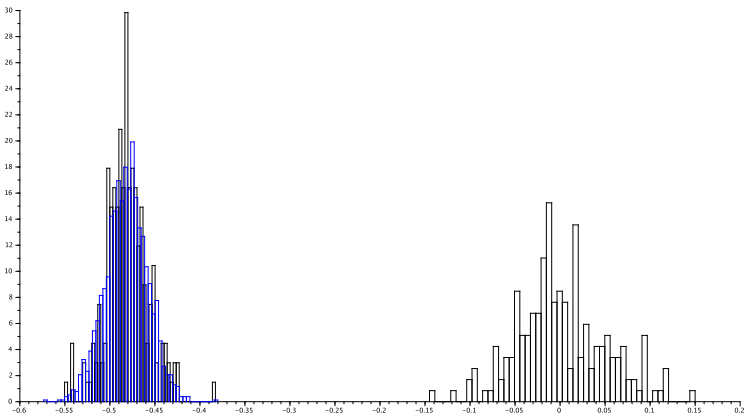

Figure 4.6 – Left: Empirical probability distribution function of the SQS 2 criterion (black histogram: no conditioning; blue histogram: the samples exactly satisfy the SQS 1 criterion). Right: empirical probability distribution of the SQS 1 criterion.

Efficiency with respect to N

Figure 4.7 shows the different estimators of the first entry $[A^\star]_{11}$ of the homogenized matrix and their respective confidence intervals. The black curve is the standard Monte Carlo estimator defined by (4.10). The variance is large. In red, we display the estimator obtained by selecting realizations that exactly satisfy the SQS 1 condition. The variance is much smaller, leading in turn to a narrower confidence interval. In blue we display the estimator obtained with realizations satisfying exactly the SQS 1 condition and selected according to the SQS 2 condition (see Algorithm 3). The variance is much smaller than when using the SQS 1 approach, even when the supercell size N is small.

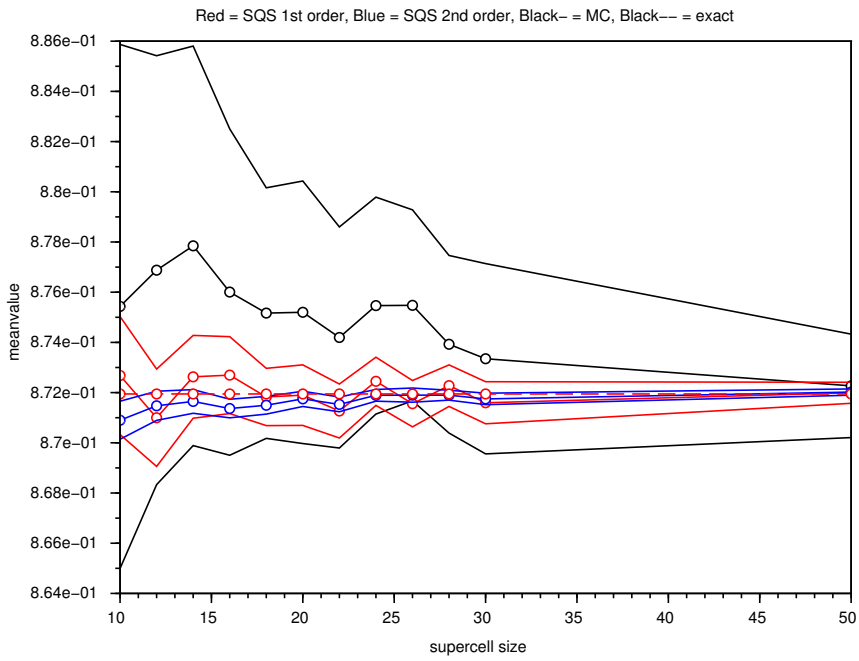

Figure 4.7 – Estimators of $[A^\star]_{11}$ (along with confidence intervals) as a function of the supercell side length N. Black curve: Monte Carlo method. Red curve: SQS 1 method. Blue curve: SQS 2 method (see text).

Figure 4.8 shows a representation of the total error as a function of the supercell size. The reference result is defined as follows. In addition to the fact that we work on a finite domain Q_N and with a finite number of samples, there is of course a slight finite element error due to the finiteness of the meshsize used to solve (4.3). This is why we do not take as reference the exact homogenized matrix, which is here known and equal to $A^\star = \sqrt{(1+\eta)(1-\eta)}\,\mathrm{Id}$. We rather take as reference the empirical expectation of $A^\star_{N_{\mathrm{ref}}}(\omega)$ over $\mathcal{M}_{\mathrm{ref}} = 2000$ random realizations exactly satisfying the SQS 1 condition, and for the largest supercell size we have considered, that is $N_{\mathrm{ref}} = 50$. The reference value is thus defined as

$$A^\star_{\mathrm{ref}} = \mathbb{E}\left[A^\star_{N_{\mathrm{ref}}} \mid \text{SQS 1 condition is exactly satisfied}\right],$$

which is in practice computed as an empirical mean over $\mathcal{M}_{\mathrm{ref}} = 2000$ realizations.

On the black curve, we see the total error of the standard Monte Carlo method, defined as

$$\text{total error} = \left| \frac{1}{M} \sum_{m=1}^{M} A_N^\star(\omega_m) - A_{\text{ref}}^\star \right|.$$

The two other curves show the same quantity, where the M environments considered now satisfy exactly the SQS 1 condition (red curve) and additionally the SQS 2 condition (blue curve). We observe that our two methods (SQS 1 and SQS 2) yield a total error roughly 7 times smaller than the Monte Carlo error. Note that we are not able to distinguish between our two methods because the reference value $A_{\text{ref}}^\star$ is computed with parameters N_{ref} and $\mathcal{M}_{\text{ref}}$ not large enough for that purpose.

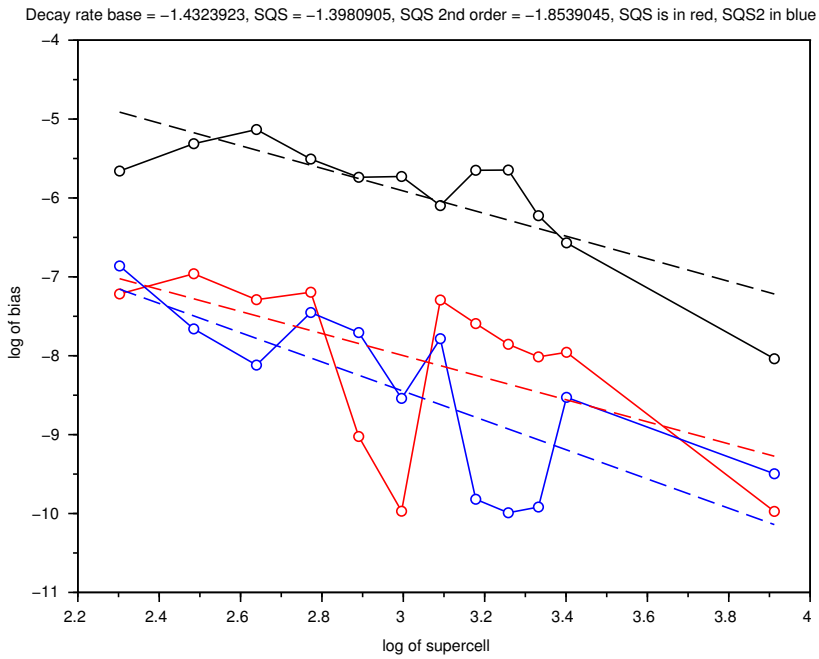

Figure 4.8 – log-log plot of the total error as a function of the supercell side length N (natural logarithm). Black curve: Monte Carlo method. Red curve: SQS 1 method. Blue curve: SQS 2 method (see text).

Figure 4.9 shows the empirical variance of the different estimators of $[A^\star]_{11}$ as a function of the supercell size. The black curve is the standard Monte Carlo estimator defined by (4.10). In red, we display the estimator obtained by selecting realizations that exactly satisfy the SQS 1 condition. In blue, we display the estimator obtained with realizations exactly satisfying the SQS 1 condition and selected according to the SQS 2 condition (see Algorithm 3). We observe that, each time we consider an additionnal SQS condition, the empirical variance of the estimator is significantly reduced (even if this SQS condition is not exactly enforced; recall that we consider here only the 5 % best samples in terms of the SQS 2 condition, but that we are unable to enforce it exactly). On our test-case, enforcing the SQS 1 condition leads to a variance 20 times smaller than that of the standard Monte Carlo approach, while additionnaly enforcing the SQS 2 condition leads to an additional variance reduction of a factor of 10.

We also observe on Figure 4.9 that all variances decay as $\lambda/|Q_N|$, where

$$\lambda_{\text{SQS 2}} < \lambda_{\text{exact SQS 1}} < \lambda_{\text{MC}}.$$

This corroborates in higher dimension the behaviour predicted in Section 4.3.2. In particular, the gain in variance does not decrease when the supercell becomes larger.

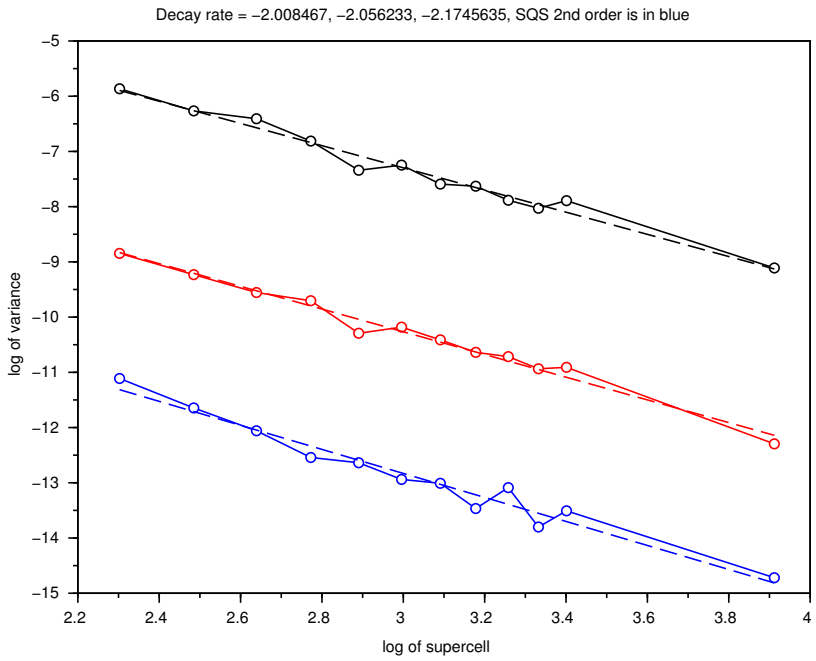

Figure 4.9 – log-log plot of the variance as a function of the supercell side length N (natural logarithm). Black curve: Monte Carlo method. Red curve: SQS 1 method. Blue curve: SQS 2 method.

Efficiency with respect to the contrast

We eventually investigate how the contrast in the field A affects the gain in variance. Results are shown on Table 4.1. We observe that the gain decreases when the contrast increases. Note that this is also the case with the antithetic variable and the control variate techniques that we have previously studied (see [BCLBL12a, LM15b, LM15a]).

However, our SQS 2 approach still yields a significant gain of a factor of 10 when the contrast is equal to 20.

Conclusion

We may summarize the numerical results by saying that, on the set of two-dimensional test-cases we have considered, the variance is significantly reduced by the approach presented in this article.

Although definite conclusions are yet to be obtained for more challenging, possibly three-dimensional, test-cases, and possibly for more elaborate equations, the results obtained are promising. The robustness and versatility of the approach give us hope for its general, efficient applicability.

Contrast	V_{MC}	$V_{\text{exact SQS 1}}$	$V_{\text{SQS 2}}$	$\frac{V_{\mathrm{MC}}}{V_{\text{exact SQS 1}}}$	$\frac{V_{\mathrm{MC}}}{V_{\text{SQS 2}}}$
1.22	0.0000273	5.801e-08	6.858e-10	470	39821
1.50	0.0001097	0.0000009	1.585e-08	118	6921
1.86	0.0002488	0.0000047	0.0000001	52.6	1996
2.33	0.0004478	0.0000151	0.0000006	29.5	720
3.00	0.0007118	0.0000379	0.0000024	18.8	296
4.00	0.0010496	0.0000814	0.0000080	12.8	131
5.67	0.0014769	0.0001600	0.0000244	9.23	60.5
9.00	0.0020289	0.0003021	0.0000739	6.71	27.4
19.0	0.0028330	0.0006061	0.0002554	4.67	11.1

Table 4.1 – For different values of the contrast, we show the Monte Carlo variance (column #2), the variance of the SQS 1 method (column #3) and the variance of the SQS 2 method (column #4). We next show the variance ratio $V_{\mathrm{MC}}/V_{\text{exact SQS 1}}$ for the SQS 1 approach (column #5) and the variance ratio $V_{\mathrm{MC}}/V_{\text{SQS 2}}$ for the SQS 2 approach (column #6). The supercell size is fixed at $N = 20$.

Acknowledgements

The first two authors would like to thank E. Cancès for introducing them to the SQS approach in the context of solid state physics, pointing out to them references [vPDFN10, WFBZ90, ZWFB90], as well as for several stimulating discussions in the early stages of this work. The authors also thank X. Blanc for his suggestions on a previous version of this article.

This work was partially supported by ONR under Grant N00014-12-1-0383 and by EOARD under Grant FA8655-13-1-3061. This work has benefited from a French government grant managed by ANR within the frame of the national program Investments for the Future ANR-11-LABX-022-01.

4.5 Proof of Lemma 4.4

We follow the arguments of the proof of [BCLBL12b, Lemma 3.2].

The existence and uniqueness (up to the addition of a constant) of ϕ_1 solution to (4.38) is established in [BCLBL12b, Lemma 3.1]. We next point out that (4.39) admits a unique (up to the addition of a constant) solution in $H^1_{\mathrm{per}}(Q)$. It is a simple consequence of the Lax-Milgram lemma.

We now prove that the sum in (4.37) is a convergent series in $L^2(Q \times \Omega)$. For this purpose, we compute the norm of the remainder of the series, using the notation $\overline{X}_k(\omega) =$

$X_k(\omega) - \mathbb{E}[X_k]$:

$$\begin{aligned}
&\left\| \sum_{|k|\geq N+1} \overline{X}_k \nabla\phi_1(\cdot - k) \right\|^2_{L^2(Q\times\Omega)} \\
= &\sum_{|k|\geq N+1} \sum_{|\ell|\geq N+1} \mathbb{E}\left[\overline{X}_k \overline{X}_\ell\right] \int_Q \nabla\phi_1(y-k)\cdot\nabla\phi(y-\ell)\,dy \\
\leq &\sum_{|k|\geq N+1} \sum_{|\ell|\geq N+1} |\mathbb{C}\mathrm{ov}(X_k, X_\ell)| \ \|\nabla\phi_1\|_{L^2(Q-k)} \ \|\nabla\phi_1\|_{L^2(Q-\ell)} \\
\leq &\sum_{|k|\geq N+1} \sum_{|\ell|\geq N+1} |\mathbb{C}\mathrm{ov}(X_k, X_\ell)| \ \|\nabla\phi_1\|^2_{L^2(Q-k)},
\end{aligned}$$

where we have used at the last line the discrete Cauchy-Schwarz inequality between $|\mathbb{C}\mathrm{ov}(X_k, X_\ell)|^{1/2} \ \|\nabla\phi_1\|_{L^2(Q-k)}$ and $|\mathbb{C}\mathrm{ov}(X_k, X_\ell)|^{1/2} \ \|\nabla\phi_1\|_{L^2(Q-\ell)}$. We next write, using the stationarity of X_k and (4.36), that

$$\begin{aligned}
&\left\| \sum_{|k|\geq N+1} \overline{X}_k \nabla\phi_1(\cdot - k) \right\|^2_{L^2(Q\times\Omega)} \\
\leq &\sum_{|k|\geq N+1} \|\nabla\phi_1\|^2_{L^2(Q-k)} \sum_{|\ell|\geq N+1} |\mathbb{C}\mathrm{ov}(X_k, X_\ell)| \\
\leq &\ \mathcal{C} \sum_{|k|\geq N+1} \|\nabla\phi_1\|^2_{L^2(Q-k)}.
\end{aligned}$$

The above right-hand side converges to 0 as $N \to \infty$ since $\nabla\phi_1 \in \left(L^2(\mathbb{R}^d)\right)^d$.

Hence, the right-hand side of (4.37) defines a function $T \in \left(L^2(Q\times\Omega)\right)^d$. As $\partial_i T_j = \partial_j T_i$, there exists a function $\widetilde{u}_1$ such that

$$\nabla\widetilde{u}_1 = T = \mathbb{E}[X_0]\nabla\overline{u_1} + \sum_{k\in\mathbb{Z}^d} \Big(X_k(\omega) - \mathbb{E}[X_k]\Big)\nabla\phi_1(\cdot - k).$$

As $\overline{u_1}$ is $\mathbb{Z}^d$-periodic, we infer from the above equality that

$$\nabla\widetilde{u}_1 \text{ is stationary and } \int_Q \mathbb{E}(\nabla\widetilde{u}_1) = 0. \tag{4.73}$$

Next, we compute

$$C_0\nabla\widetilde{u}_1 = \mathbb{E}[X_0]C_0\nabla\overline{u_1} + \sum_{k\in\mathbb{Z}^d} \Big(X_k(\omega) - \mathbb{E}[X_k]\Big)C_0\nabla\phi_1(\cdot - k).$$

Taking the divergence of this equation and using (4.32) and (4.34), we thus find that, in the distribution sense,

$$\begin{aligned}
-\operatorname{div}\left[C_0\nabla\widetilde{u}_1\right] &= \sum_{k\in\mathbb{Z}^d} -\Big(X_k(\omega) - \mathbb{E}[X_k]\Big) \operatorname{div}\left[C_0\nabla\phi_1(\cdot - k)\right] \\
&\quad -\mathbb{E}[X_0] \operatorname{div}\left[C_0\nabla\overline{u_1}\right] \\
&= \sum_{k\in\mathbb{Z}^d} \Big(X_k(\omega) - \mathbb{E}[X_k]\Big) \operatorname{div}\left[\mathbb{1}_{Q+k}C_1 p\right] \\
&\quad +\mathbb{E}[X_0] \operatorname{div}\left[C_1 p\right] \\
&= \operatorname{div}\left[\chi(\cdot,\omega)C_1 p\right].
\end{aligned} \tag{4.74}$$

Collecting (4.73) and (4.74), we see that $\widetilde{u}_1$ solves (4.19). As the solution to this equation is unique up to the addition of a (possibly random) constant $C(\omega)$, we obtain that $\widetilde{u}_1 = u_1 + C(\omega)$, hence proving (4.37).

Chapter 5

A parameter identification problem in stochastic homogenization

Ce **Chapitre** reprend l'intégralité d'un article écrit en collaboration avec Frédéric Legoll, Marielle Simon et Amaël Obliger et accepté dans ESAIM ProcS [LMOS15].

Dans un cadre d'homogénéisation stochastique discrète, nous cherchons à identifier des paramètres de la loi de probabilité du milieu microscopique, à l'aide de quantités macroscopiques. Nous formulons ce problème à la manière d'un problème d'optimisation aux moindres carrés, et nous démontrons en dimension 1 d'espace qu'il est asymptotiquement bien posé. Ce résultat théorique est complété par des expérimentations numériques.

A parameter identification problem in stochastic homogenization

Frédéric Legoll[1,2], William Minvielle[3,2], Amaël Obliger[4,5] and Marielle Simon[6]

legoll@lami.enpc.fr, minvielw@cermics.enpc.fr, amael.obliger@upmc.fr, marielle.simon@ens-lyon.fr

[1] Laboratoire Navier, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal,
77455 Marne-La-Vallée Cedex 2, France ;

[2] INRIA Rocquencourt, MATHERIALS research-team, Domaine de Voluceau, B.P. 105, 78153 Le Chesnay Cedex, France ;

[3] CERMICS, École Nationale des Ponts et Chaussées, Université Paris-Est, 6 et 8 avenue Blaise Pascal, 77455 Marne-La-Vallée Cedex 2, France ;

[4] Sorbonne Universités, UPMC Univ. Paris 06, UMR 8234 PHENIX, 75005 Paris, France ;

[5] ANDRA, Parc de la Croix-Blanche, 1-7, rue Jean-Monnet, 92298 Châtenay-Malabry, France

[6] UMPA, UMR-CNRS 5669, ENS Lyon, 46 allée d'Italie, 69007 Lyon, France.

Abstract. *In porous media physics, calibrating model parameters through experiments is a challenge. This process is plagued with errors that come from modelling, measurement and computation of the macroscopic observables through random homogenization – the forward problem – as well as errors coming from the parameters fitting procedure – the inverse problem. In this work, we address these issues by considering a least-square formulation to identify parameters of the microscopic model on the basis on macroscopic observables, including homogenized coefficients. In particular, we discuss the selection of the macroscopic observables which we need to know in order to uniquely determine these parameters. To gain a better intuition and explore the problem without a too high computational load, we mostly focus on the one-dimensional case. We show that the Newton algorithm can be efficiently used to robustly determine optimal parameters, even if some small statistical noise is present in the system.*

Résumé. *En physique des milieux poreux, calibrer certains paramètres d'un modèle microscopique sur la base d'expériences donnant accès à des grandeurs macroscopiques est un enjeu majeur. Cette démarche est entachée d'erreurs de modèle, de mesure et de calculs dans la procédure d'homogénéisation: le problème direct est biaisé. La résolution du problème inverse, lorsqu'il s'agit d'estimer les paramètres à partir des observations, engendre aussi des erreurs. Nous considérons ici une formulation "moindres carrés" du problème, cherchant à minimiser l'erreur entre les quantités macroscopiques observées et celles calculées via l'homogénéisation aléatoire. Nous discutons en particulier de la nature des informations macroscopiques nécessaires pour déterminer de manière univoque les paramètres de la densité de probabilité des propriétés microscopiques. Afin d'explorer plus facilement cette question, nous nous intéressons ici essentiellement au cas unidimensionel. Nous montrons que le problème peut être résolu de manière efficace par l'algorithme de Newton, même en présence d'un petit bruit statistique.*

5.1 Introduction

Modelling porous media is a challenge, in particular because the geometry of such materials can be extremely complex. Rock samples are often described as a pile of layers of solid phase which do not permit flows, creating voids in-between layers that are connected by channels, the size and shape of which is difficult to describe (and to observe experimentally, although, in rare cases, imaging methods such as micro-tomography can be used). Besides these issues related to the description of the geometry of the media, another difficulty is to properly model the physical phenomena occuring in the flow. To circumvent these difficulties, a possible approach consists in completely forgetting the exact geometry of the system except for a few parameters (e.g. the size of the channels), and consider that the channels form a simple network, often taken to be $\mathbb{Z}^d$. This results in the so-called pore-network models (PNM), initially introduced by Fatt in the 1950s [Fat56] and which have been widely used since then. The void space of a rock (its porosity) is described by a pore network connected by channels. In this framework, the geometry of pores and channels is idealized. Some microscopic properties are assigned to network elements (e.g. the conductance of the channels) and rules are defined to compute the upscaled (homogenized) properties on the basis of this microscopic description. In turn, these upscaled properties can be compared to the available experimental data. The aim is to construct a microscopic network with the same effective properties as those of a real representative sample of rock.

In this work, we follow this approach, and assume that pores are located at the vertices of a simple lattice. Physical properties are described by some random field at the microscopic scale. We focus on monophasic transport phenomena in porous media, where the sample of rock is mainly characterized by its permeability. These phenomena are described by the Darcy law, where the local flux of water is assumed to be proportional to the local pressure gradient, and the microscopic properties of interest are the conductances of the channels. In the pore network model, conductances are solely assigned to channels, and it is assumed that pores do not contribute to the flow. Following Darcy's equation, the microscopic pressure field is computed in the network by ensuring mass conservation at each pore. The equation to solve is therefore a *discrete* linear elliptic equation in divergence form, with random coefficients (see (5.13)–(5.14) below for a more detailed physical description, and (5.4) for a more mathematical description).

The conductances of the channels (i.e. their microscopic permeabilities) depend on their size. Therefore the construction of the network starts by randomly attributing a size to each channel. In practice, this channel size distribution can be inferred from experiments such as mercury porosimetry: we denote it by $\mathcal{L}_{\text{exp}}$. Several issues of different nature arise in this procedure. As a consequence, it turns out that the effective properties (e.g. macroscopic permeability) that are computed for a pore network with channel sizes distributed according to $\mathcal{L}_{\text{exp}}$ are different from the experimental effective properties. The extraction procedure, which provides a channel size distribution, is thus somewhat slightly inconsistent. The main goal of this work consists in improving that distribution, when starting from the experimental initial guess, in order to eventually achieve a better agreement between measured and computed effective properties.

From a more mathematical standpoint, the question can be phrased in the following terms. Consider a second-order divergence-form operator whose coefficients are random. If the distribution of the coefficients is stationary and ergodic, then (under some additional technical assumptions) this random operator can be replaced, over large scales, by an effective operator with constant homogenized coefficients (see Theorem 5.5 below). Random homogenization theory actually provides formulas to compute the homogenized quanti-

ties. We thus have at our disposal a procedure to compute macroscopic quantities if we know the microscopic quantities, and to solve the so-called *forward problem*. However, in practice, given a heterogeneous materials, it is a difficult question to decide on the law of the microscopic physical properties. On the other hand, macroscopic quantities are more easily accessible. It is thus of interest to consider the *inverse problem*, and try to extract some information on the properties of the materials at the microscopic scale on the basis of macroscopic quantities.

In the same spirit, if one makes assumptions on the microscopic law, then macroscopic quantities can be computed, and for instance compared to experimental values. In view of the possible discrepancy between the two, one could question or revisit the assumptions made at the microscopic scale.

Of course, homogenization is an averaging process, which filters out many features of the microscopic coefficients. There is thus no hope to recover a full information about the microstructure (in our case, the *probability distribution* of the conductances) from the only knowledge of macroscopic quantities. We adopt here a more restricted objective. We assume a functional form for the distribution of the microscopic conductances (namely, a Weibull distribution). Our aim is to recover the *parameters* (denoted here θ) of that microscopic law of the basis of macroscopic quantities.

We point out that our approach is not specific to Weibull laws, and that it could be used for other distribution laws with parameters θ. What we need is that the random field $A(x,\omega)$ used at the microscopic scale can be written as

$$A(x,\omega) = \mathcal{F}\Big(u(x,\omega),\theta\Big)$$

where $u(x,\omega)$ is a field of random variables that are uniformly distributed and $\mathcal{F}$ smoothly depends on the parameters θ (see (5.16) in our particular case). Computing the derivatives of the microscopic random field $A(x,\omega)$ (and next of the macroscopic, homogenized quantities) with respect to θ is then easy. Our motivation for choosing Weibull laws comes from physical reasons: based on experimental results, it appears to be a reasonable choice.

Likewise, our approach is not specific to *discrete* elliptic equations. It could also be applied for problems modelled by *continuous* elliptic partial differential equations (PDEs) with random, highly oscillatory coefficients. Here, we consider discrete equations because the pore network model, which is naturally written in terms of discrete equations, is commonly used for such materials.

The question of recovering the unknown parameters θ of the microscopic distribution from homogenized (and more generally macroscopic) quantities belongs to the wide family of *inverse problems*. In this work, a major point of interest is the selection of the macroscopic quantities which we need to know in order to uniquely determine the parameters θ. This point is discussed in Section 5.3.2. We refer to [NPS12] for a review article on inverse problems in a multiscale context.

The article is organised as follows. In Section 5.2, we recall some elements of stochastic homogenization and describe the physical problem that motivates this work (including the choice of Weibull laws). We conclude that section with results specific to the one-dimensional case. In particular, random variables distributed according to a Weibull law are not isolated from 0 or $+\infty$, and thus the microstructure does not satisfy the classical assumption of ellipticity, namely (5.3) below. We show in Section 5.2.4 that, in the one-dimensional case, homogenization still holds under a weaker assumption, that in turn is satisfied by Weibull random variables.

Next, in Section 5.3, we introduce our parameter fitting problem, formulated as least-square optimization. We first consider the general (multi-dimensional) case before turning to the one-dimensional case. In that latter case, we discuss the macroscopic quantities that are needed to uniquely determine the parameters θ. More precisely, Weibull laws have two parameters, and the knowledge of a single homogenized quantity (namely the macroscopic permeability) is, as expected, insufficient to determine the two unknown parameters. We show there (in the one-dimensional case) that, if we additionally specify the relative variance of the effective macroscopic permeability, then we are in position to uniquely determine the two parameters of the microscopic Weibull law.

Section 5.4 is dedicated to numerical results, again in the one-dimensional case. We show that the Newton algorithm can be efficiently used in the current least-square optimization setting. In particular, in practice, the exact homogenized coefficients cannot be computed, and only a random approximation of them is available. We monitor here how this randomness propagates to the optimal parameters. The extension of this work to the two-dimensional case will be addressed in a future work [LMOS].

5.2 Discrete homogenization theory

For the sake of completeness, we recall first, in Sections 5.2.1 and 5.2.2, some elements of homogenization for discrete elliptic equations with random coefficients. We refer to [Kün83, Koz87] for seminal contributions on this topic. For homogenization of elliptic partial differential equations (PDEs), we refer to the seminal work [PV81], to the textbooks [BLP78, CD99, JKO94], to [ES08] for a general, numerically oriented presentation, and to the review article [ACLB+12].

Next, in Section 5.2.3, we describe the physical background that motivates this work. We eventually turn in Section 5.2.4 to the one-dimensional case, where explicit formulae can be obtained.

5.2.1 Homogenization result

We first recall some definitions useful for stochastic homogenization, before turning to the specific case of discrete elliptic equations.

Throughout this article, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and we denote by $\mathbb{E}(X) = \int_\Omega X(\omega) d\mathbb{P}(\omega)$ the expectation value of any random variable $X \in L^1(\Omega, d\mathbb{P})$. We next fix $d \in \mathbb{N}^\star$ (the ambient physical dimension), and assume that the group $(\mathbb{Z}^d, +)$ acts on Ω. We denote by $(\tau_k)_{k\in\mathbb{Z}^d}$ this action, and assume that it preserves the measure $\mathbb{P}$, that is, for all $k \in \mathbb{Z}^d$ and all $B \in \mathcal{F}$, $\mathbb{P}(\tau_k B) = \mathbb{P}(B)$. We assume that the action τ is *ergodic*, that is, if $B \in \mathcal{F}$ is such that $\tau_k B = B$ for any $k \in \mathbb{Z}^d$, then $\mathbb{P}(B) = 0$ or 1. In addition, we introduce the following notion of stationarity:

Definition 5.1. *We say that a function $\psi : \mathbb{Z}^d \times \Omega \to \mathbb{R}$ is* stationary *if*

$$\forall x, z \in \mathbb{Z}^d, \quad \psi(x + z, \omega) = \psi(x, \tau_z \omega) \quad a.s. \tag{5.1}$$

We now focus on the case of discrete elliptic equations. We view $\mathbb{Z}^d$ as a lattice, whose unit vectors are denoted by e_i, $i \in \{1, \dots, d\}$. Each vertex $x \in \mathbb{Z}^d$ of the lattice is connected to $2d$ other vertices: $x \pm e_i$, $i \in \{1, \dots, d\}$. We write $x \sim y$ if x and y are neighbours (i.e. connected), and $e = (x, y)$ the corresponding (non-oriented) edge. For any vertex $x \in \mathbb{Z}^d$ and any direction $1 \le i \le d$, we denote by $a_i(x, \omega) \in (0, \infty)$ the random conductance of the

edge $(x, x+e_i)$. We next introduce the diagonal matrix A defined for any vertex $x \in \mathbb{Z}^d$ by

$$A(x,\omega) = \operatorname{diag}\Big(a_1(x,\omega), \dots, a_d(x,\omega)\Big). \tag{5.2}$$

We assume that, for any direction i, the conductances $\{a_i(x,\cdot)\}_{x\in\mathbb{Z}^d}$ form an i.i.d. sequence of random variables. The matrix A is therefore stationary.

We introduce the following assumption:

ASSUMPTION 5.2 (Ellipticity – boundedness condition)**.** *There exist two positive deterministic constants c and C such that the matrix A defined by (5.2) satisfies*

$$\forall \xi \in \mathbb{R}^d, \quad \forall x \in \mathbb{Z}^d, \quad c|\xi|^2 \le \xi \cdot A(x,\omega)\xi \le C|\xi|^2 \quad a.s. \tag{5.3}$$

In view of (5.2), note that this simply means that $0 < c \le a_j(x,\omega) \le C$ almost surely, for any $1 \le j \le d$ and any $x \in \mathbb{Z}^d$.

We next introduce discrete differential operators on the lattice $\mathbb{Z}^d$.

Definition 5.3. *For a function $g : \mathbb{Z}^d \to \mathbb{R}$, the gradient $\nabla g : \mathbb{Z}^d \to \mathbb{R}^d$ is defined by*

$$(\nabla g)(x) = \begin{pmatrix} g(x+e_1) - g(x) \\ \vdots \\ g(x+e_d) - g(x) \end{pmatrix}.$$

For a function $G = (G_1, \dots, G_d) : \mathbb{Z}^d \to \mathbb{R}^d$, the function $\nabla^\star G : \mathbb{Z}^d \to \mathbb{R}$ is defined by

$$-(\nabla^\star G)(x) = \sum_{i=1}^d \big(G_i(x) - G_i(x-e_i)\big).$$

We think of $\nabla^\star G$ as the negative divergence of G. The operator $\nabla^\star$ is the ℓ^2 transpose of ∇ in the following sense: for any compactly supported functions $g : \mathbb{Z}^d \to \mathbb{R}$ and $G : \mathbb{Z}^d \to \mathbb{R}^d$,

$$\sum_{x\in\mathbb{Z}^d} g(x)\nabla^\star G(x) = \sum_{x\in\mathbb{Z}^d} \nabla g(x) \cdot G(x).$$

Hereafter, the notation $a \cdot b$ stands for the usual scalar product in $\mathbb{R}^d$.

We additionally define rescaled discrete differential operators as follows:

Definition 5.4. *For a function $g : \varepsilon\mathbb{Z}^d \to \mathbb{R}$, the gradient $\nabla_\varepsilon g : \varepsilon\mathbb{Z}^d \to \mathbb{R}^d$ is defined by*

$$(\nabla_\varepsilon g)(x) = \frac{1}{\varepsilon}\begin{pmatrix} g(x+\varepsilon e_1) - g(x) \\ \vdots \\ g(x+\varepsilon e_d) - g(x) \end{pmatrix}.$$

For a function $G = (G_1, \dots, G_d) : \varepsilon\mathbb{Z}^d \to \mathbb{R}^d$, the function $\nabla_\varepsilon^\star G : \varepsilon\mathbb{Z}^d \to \mathbb{R}$ is defined by

$$-(\nabla_\varepsilon^\star G)(x) = \sum_{i=1}^d \frac{G_i(x) - G_i(x-\varepsilon e_i)}{\varepsilon}.$$

The following homogenization result holds (we refer to [Kün83, Theorems 3 and 4] for a proof):

Theorem 5.5. *Let $\mathcal{D}$ be a bounded domain of $\mathbb{R}^d$ and $f \in C^0(\overline{\mathcal{D}})$. Let A be the random stationary matrix field given by (5.2). We assume that (5.3) holds. Let $u_\varepsilon \in \ell^2(\varepsilon\mathbb{Z}^d;\mathbb{R})$ be the unique solution to*

$$\nabla_\varepsilon^\star\big[A(x/\varepsilon,\omega)\nabla_\varepsilon u_\varepsilon(x,\omega)\big] = f(x) \quad \text{in } \mathcal{D}\cap\varepsilon\mathbb{Z}^d, \qquad u_\varepsilon(x,\omega) = 0 \text{ in } (\mathbb{R}^d\setminus\mathcal{D})\cap\varepsilon\mathbb{Z}^d. \tag{5.4}$$

When ε goes to 0, $u_\varepsilon(\cdot,\omega)$ converges to some homogenized function $u^\star$.

For any $\xi \in \mathbb{R}^d$, introduce the corrector *φ_ξ in the direction ξ as the unique solution (defined on $\mathbb{Z}^d\times\Omega$) to*

$$\begin{cases} -\nabla^\star\left[A(\cdot,\omega)\big(\xi+\nabla\varphi_\xi(\cdot,\omega)\big)\right] = 0 \text{ in } \mathbb{Z}^d, \text{ a.s.,} \\ \nabla\varphi_\xi \text{ is stationary in the sense of (5.1),} \\ \forall x\in\mathbb{Z}^d, \quad \mathbb{E}[\nabla\varphi_\xi(x,\cdot)] = 0, \\ \varphi_\xi(0,\omega) = 0 \text{ a.s.} \end{cases} \tag{5.5}$$

Introduce next the constant matrix $A^\star$ defined by

$$\forall\xi\in\mathbb{R}^d, \quad A^\star\xi = \mathbb{E}\big[A(x,\cdot)(\xi+\nabla\varphi_\xi(x,\cdot))\big] \tag{5.6}$$

and the unique solution $u^\star \in H_0^1(\mathcal{D})$ to the (continuous) partial differential equation

$$-\mathrm{div}\big[A^\star\widehat{\nabla}u^\star\big] = f \qquad \text{in } \mathcal{D},$$

where $\widehat{\nabla}$ and div *are the usual (continuous) gradient and divergence differential operators.*

Then, we have the (strong) convergence $u_\varepsilon \xrightarrow[\varepsilon\to0]{} u^\star$, in the sense that

$$\varepsilon^d \sum_{x\in\mathcal{D}\cap\varepsilon\mathbb{Z}^d} |u_\varepsilon(x,\omega)-u^\star(x)|^2 \xrightarrow[\varepsilon\to0]{} 0 \quad \text{almost surely.} \tag{5.7}$$

Note that, in the right-hand side of (5.6), the vector $A(\xi+\nabla\varphi_\xi)$ is stationary, and therefore the expectation may be evaluated at any $x\in\mathbb{Z}^d$. Note also that, in general, φ_ξ itself is not stationary, as the one-dimensional case shows. Only its gradient is.

Remark 5.6. *We can define, on $\mathcal{D}$, the function*

$$\widetilde{u}_\varepsilon(x,\omega) = \sum_{k\in\varepsilon\mathbb{Z}^d\cap\mathcal{D}} u_\varepsilon(k,\omega)\ \mathbb{1}_{k+\varepsilon Q}(x), \qquad \text{where } Q=(0,1)^d.$$

Then (5.7) implies that $\widetilde{u}_\varepsilon(\cdot,\omega) \xrightarrow[\varepsilon\to0]{} u^\star$ in $L^2(\mathcal{D})$ almost surely.

5.2.2 Approximation on finite boxes

The corrector problem (5.5) is untractable in practice, since it is posed in the entire lattice $\mathbb{Z}^d$. Approximations are therefore in order. The standard procedure amounts to considering finite boxes (see e.g. [BP04]). For a positive integer N, we denote by Λ_N the finite box $\{0,\dots,N\}^d$ and by $\mathcal{E}_N$ the set of edges in Λ_N (see Figure 5.1).

The *truncated corrector* φ_ξ^N defined on $\Lambda_N\times\Omega$ is the unique solution to

$$\begin{cases} -\nabla^\star\Big[A(\cdot,\omega)\big(\xi+\nabla\varphi_\xi^N(\cdot,\omega)\big)\Big] = 0 \text{ in } \Lambda_N, \text{ a.s.,} \\ \varphi_\xi^N(\cdot,\omega) \text{ is } \Lambda_N\text{-periodic,} \\ \varphi_\xi^N(0,\omega) = 0 \text{ a.s.} \end{cases} \tag{5.8}$$

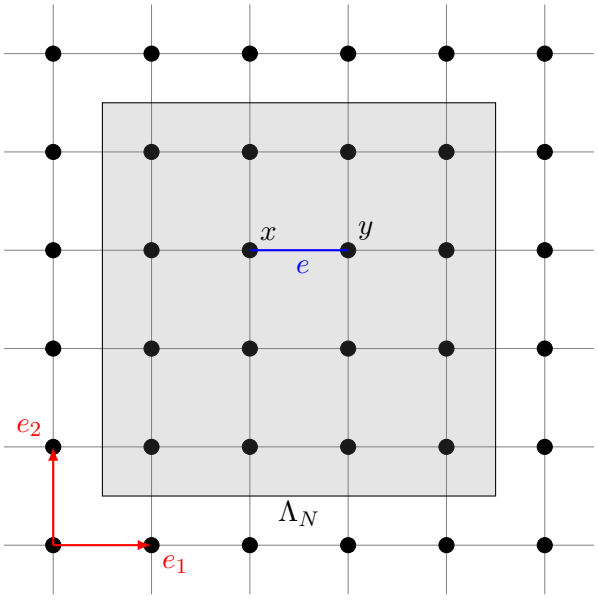

Figure 5.1 – Finite box Λ_N in $\mathbb{Z}^2$

The homogenized matrix $A^\star$, which is deterministic, is then approximated by the matrix $A^\star_N$ defined by

$$\forall \xi \in \mathbb{R}^d, \quad A^\star_N(\omega)\xi = \frac{1}{|\Lambda_N|} \sum_{x \in \Lambda_N} A(x,\omega)\big(\xi + \nabla \varphi^N_\xi(x,\omega)\big). \tag{5.9}$$

Because of truncation, the practical approximation $A^\star_N$ is random. In the large N limit, the deterministic value is attained, thanks to ergodicity. More precisely, $A^\star_N(\omega)$ converges almost surely towards $A^\star$ as N goes to infinity, thanks to the ergodic theorem.

Remark 5.7. *In (5.8), we have complemented the elliptic equation in Λ_N with* periodic *boundary conditions. Other choices could be made, such as imposing homogeneous Dirichlet boundary conditions: $\varphi^N_\xi(\cdot,\omega) = 0$ on $\partial\Lambda_N$ (see e.g. [BP04] for a similar discussion in the case of continuous PDEs). In the numerical experiments of Section 5.4, we only use periodic boundary conditions, following (5.8).*

In practice, we work on a finite box Λ_N, on which the apparent homogenized matrix $A^\star_N$ is random. It is therefore natural to introduce M i.i.d. realizations of the random field $A(x,\omega)$ and solve (5.8)–(5.9) for each of them, thereby obtaining i.i.d. realizations $A^{\star,m}_N(\omega)$, $1 \le m \le M$. We next introduce the empirical mean

$$\overline{A}^\star_{N,M}(\omega) = \frac{1}{M} \sum_{m=1}^{M} A^{\star,m}_N(\omega) \tag{5.10}$$

which is, according to the Central Limit Theorem, a converging approximation of $\mathbb{E}\left[A^\star_N\right]$. We have that

$$\overline{A}^\star_{N,M}(\omega) \xrightarrow[M\to\infty]{} \mathbb{E}\left[A^\star_N\right] \quad \text{a.s.}$$

In addition, for any entry $1 \le i,j \le d$ of the matrix, we have that, with a probability of 95 %,

$$\left| \left(\overline{A}^\star_{N,M}(\omega)\right)_{ij} - \mathbb{E}\left[(A^\star_N)_{ij}\right] \right| \le 1.96 \sqrt{\frac{\mathbb{V}\mathrm{ar}\ \left(A^\star_N\right)_{ij}}{M}}.$$

The error when approximating $A^\star$ by $\overline{A}^\star_{N,M}$ can be written as the sum of two contributions,

$$A^\star - \overline{A}^\star_{N,M} = \Big(A^\star - \mathbb{E}[A^\star_N]\Big) + \Big(\mathbb{E}[A^\star_N] - \overline{A}^\star_{N,M}\Big). \tag{5.11}$$

The second term in the right-hand side of (5.11) is the *statistical* error. The first term is the *systematic* error, due to the fact that, for any finite N, $\mathbb{E}[A^\star_N] \neq A^\star$. The dominated convergence theorem ensures that this error vanishes as $N \to \infty$. Many studies have been recently devoted to proving sharp estimates on the rate of this convergence, following the seminal works [Yur86, BP04]. In [GNO14, Theorem 2], the authors show (for any dimension $d \geq 2$) that the systematic error is of order $N^{-d}\ln^d(N)$ when using periodic boundary conditions (namely, solving (5.8)), and that $\mathbb{V}\mathrm{ar}(A^\star_N)$ scales as N^{-d}. Likewise, in the case of *continuous* PDEs, when $d \geq 3$, optimal estimates on $\mathbb{V}\mathrm{ar}(A^\star_N)$ have been established in [Nol14, Theorem 1.3 and Proposition 1.4].

Remark 5.8. *The estimator $\overline{A}^\star_{N,M}$ only agrees with $\mathbb{E}\left[A^\star_N\right]$ in the limit of an asymptotically large number M of realizations. Note that variance reduction approaches have been introduced in this context (see e.g. [BCLBL12a, BCLBL12b, CLBL10], and [LM15b] for the extension to a nonlinear setting) to obtain approximations of $\mathbb{E}\left[A^\star_N\right]$ in a more efficient manner than by using $\overline{A}^\star_{N,M}$.*

In the sequel, we will identify the parameters of the microscopic probability distribution on the basis of two types of macroscopic quantities:

1. the homogenized permeability, which is in practice approximated by $\overline{A}^\star_{N,M}$;
2. the relative variance of any entry $\left(A^\star_N(\omega)\right)_{ij}$, defined by

$$\mathbb{V}\mathrm{arR}\left[\left(A^\star_N\right)_{ij}\right] := \frac{\mathbb{V}\mathrm{ar}\left[\left(A^\star_N\right)_{ij}\right]}{\left(\mathbb{E}\left[\left(A^\star_N\right)_{ij}\right]\right)^2},$$

which is in practice approximated by

$$S_{N,M} = \frac{1}{(\overline{A}^\star_{N,M})^2_{ij}}\left(\frac{1}{M}\sum_{m=1}^{M}\left(\left(A^{\star,m}_N(\omega)\right)_{ij} - \left(\overline{A}^\star_{N,M}\right)_{ij}\right)^2\right). \tag{5.12}$$

5.2.3 Physical problem

We describe here the physical background which inspires this work. As pointed out above, from a physical viewpoint, understanding the microscopic properties of charged porous media is of great importance. Such materials have elaborate geometries that make direct computations very challenging. To circumvent this issue, we use here the Pore Network Model (PNM), which involves a simplified model of the geometry. In the PNM model, pores are located at the vertices of the lattice $\mathbb{Z}^d$. Neighbouring pores are connected by channels, which allow water to flow. Each channel $(x, x+e_i)$ is endowed with its random conductance $a_i(x,\omega) > 0$, the probability distribution of which is discussed below.

Experiments provide measures on the *macroscopic permeability* K_{obs}, which is modelled as a homogenized coefficient $K^\star$. In practice, as explained in Section 5.2.2, the homogenized coefficient can only be approximated through a computation on a large box. Assuming that the conductance field $a_i(x,\omega)$ is given for any direction $1 \leq i \leq d$ and any vertex x on the finite lattice Λ_N, the PNM model consists in computing the pressure field $P(x,\omega)$

by solving the conservation equations (i.e., Darcy law) in the network. This leads to the following linear system:

$$\forall x \in \Lambda_N, \quad \sum_{y \sim x} \tilde{a}(x,y,\omega)\big(P(y,\omega) - P(x,\omega)\big) = 0, \tag{5.13}$$

where $\tilde{a}(x,y,\omega)$ is the conductance of the non-oriented edge (x,y). Some boundary conditions need to be imposed to make this problem well-posed, they are discussed below. We next see, by definition of a_i, that

$$\begin{aligned} & \sum_{y \sim x} \tilde{a}(x,y,\omega)\big(P(y,\omega) - P(x,\omega)\big) \\ = & \sum_{i=1}^{d} a_i(x,\omega)\big(P(x+e_i,\omega) - P(x,\omega)\big) + \sum_{i=1}^{d} a_i(x-e_i,\omega)\big(P(x-e_i,\omega) - P(x,\omega)\big) \\ = & \nabla^\star\big[A(\cdot,\omega)\nabla P(\cdot,\omega)\big](x), \end{aligned} \tag{5.14}$$

where the matrix A is defined in terms of $\{a_i\}_{i=1}^d$ by (5.2).

We now describe (in the two-dimensional case, for the sake of simplicity) the boundary conditions imposed on (5.13). They are designed to mimic experimental conditions. We first recall that the large box reads $\Lambda_N = \{0,\dots,N\}^2$. The pressure field is assumed to be periodic in the vertical direction, whereas a macroscopic gradient is imposed in the horizontal direction as follows. Imagine that all vertices with coordinates $(0,\cdot)$ are connected to one fixed vertex denoted by O, representing a pressure reservoir at pressure P_O. Likewise, all vertices with coordinates $(N,\cdot)$ are connected to one fixed vertex denoted by I at pressure P_I (see Figure 5.2). Then, the boundary conditions write

$$\text{for all } j \in \{0,\dots,N\}, \quad P(0,j) = P_O \ \text{ and } \ P(N,j) = P_I.$$

Once (5.13) is solved with the above boundary conditions, the *macroscopic permeability* $K_N^\star$ is defined by

$$K_N^\star(\omega) := \frac{N}{P_O - P_I}\frac{1}{|\Lambda_N|}\sum_{x \in \Lambda_N} \tilde{a}(x,x+e_1,\omega)\big(P(x,\omega) - P(x+e_1,\omega)\big). \tag{5.15}$$

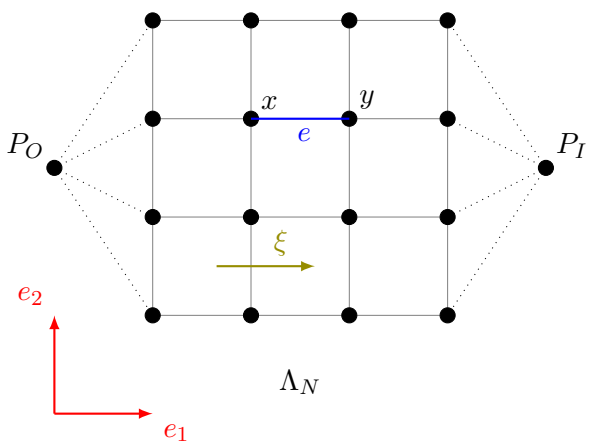

Figure 5.2 – Finite lattice with boundary conditions

Let us now show that Equations (5.13)–(5.15) are actually the same as Equations (5.8)–(5.9) written above. By linearity of (5.13)–(5.15), we can assume that $P_O = 0$ and $P_I = N$ without loss of generality. Let P be a solution to (5.13). We introduce φ_{e_1} such that

$$P(x,\omega) = x \cdot e_1 + \varphi_{e_1}(x,\omega).$$

In view of (5.13) and (5.14), we see that $\varphi_{e_1}(\cdot,\omega)$ is solution to

$$\forall x \in \Lambda_N, \quad \nabla^\star\big[A(\cdot,\omega)(e_1 + \nabla\varphi_{e_1}(\cdot,\omega)\big](x) = 0$$

with $\varphi_{e_1}((0,j),\omega) = \varphi_{e_1}((N,j),\omega) = 0$ for any j and $\varphi_{e_1}(\cdot,\omega)$ is periodic in the vertical direction. Up to the choice of boundary conditions, we thus recognize (5.8) for $\xi = e_1$. We also infer from (5.15) that

$$\begin{aligned}
K_N^\star(\omega) &= \frac{N}{P_O - P_I}\frac{1}{|\Lambda_N|}\sum_{x\in\Lambda_N} \tilde{a}(x,x+e_1,\omega)\big(P(x,\omega) - P(x+e_1,\omega)\big) \\
&= \frac{-1}{|\Lambda_N|}\sum_{x\in\Lambda_N} a_1(x,\omega)\big(\varphi_{e_1}(x,\omega) - \varphi_{e_1}(x+e_1,\omega) - e_1\cdot e_1\big) \\
&= \frac{1}{|\Lambda_N|}\sum_{x\in\Lambda_N} a_1(x,\omega)\ e_1^T\ \big(e_1 + \nabla\varphi_{e_1}(x,\omega)\big) \\
&= e_1^T A_N^\star(\omega) e_1,
\end{aligned}$$

where $A_N^\star(\omega)$ is defined by (5.9), and where we have used (5.2) in the last line. Thus, up to the choice of boundary conditions in the corrector problem, the formulation (5.13)–(5.15) is identical to the formulation (5.8)–(5.9).

We eventually discuss the choice of the probability distribution for the conductances. Based on experimental results, it is reasonable to assume the following:

Assumption 5.9. *We assume that the radius r of the channels are i.i.d. random variables distributed according to a Weibull law of parameter $\theta := (\lambda,k) \in (\mathbb{R}_+^\star)^2$, that we denote $\mathcal{W}(\lambda,k)$. We recall that such random variables are positive, with a probability density that reads (see Figure 5.3)*

$$\forall r > 0, \qquad f(r;k,\lambda) = \frac{k}{\lambda}\left(\frac{r}{\lambda}\right)^{k-1}\exp\Big(-(r/\lambda)^k\Big),$$

corresponding to the cumulative distribution function

$$F(r;k,\lambda) = \int_0^r f(s;k,\lambda)\,ds = 1 - \exp\Big(-(r/\lambda)^k\Big).$$

Note that the radius of all channels (independently of their direction $1 \le i \le d$) share the same probability distribution.

In practice, a Weibull distribution is generated as follows. Let $u(\omega)$ be a random variable uniformly distributed in $[0,1]$. Then

$$r(\omega) = \lambda\Big[-\ln(1-u(\omega))\Big]^{1/k}$$

is distributed according to the Weibull law of parameter (λ,k).

Physical arguments lead to the fact that the conductance $a_i(x,\omega)$ of any channel $(x, x+e_i)$ is directly related to its radius $r(x, x+e_i, \omega)$. Hereafter, we assume that

$$a_i(x,\omega) = C_0\, r^4(x, x+e_i, \omega) = C_0\, \lambda^4 \Big[-\ln(1-u_i(x,\omega)) \Big]^{4/k}, \tag{5.16}$$

where C_0 is a constant (for instance, for a Poiseuille flow, $C_0 = \pi/(8\eta)$ where η is the fluid viscosity and $u_i(x,\omega)$ is uniformly distributed in $[0,1]$). For the sake of simplicity, we take $C_0 = 1$ in the sequel. Therefore, we assume that

The conductances $\{a_i(x,\omega)\}_{x\in\mathbb{Z}^d,\, 1\le i\le d}$ form an i.i.d. sequence of random variables that are distributed according to the Weibull law of parameter $(\lambda^4, k/4)$. (5.17)

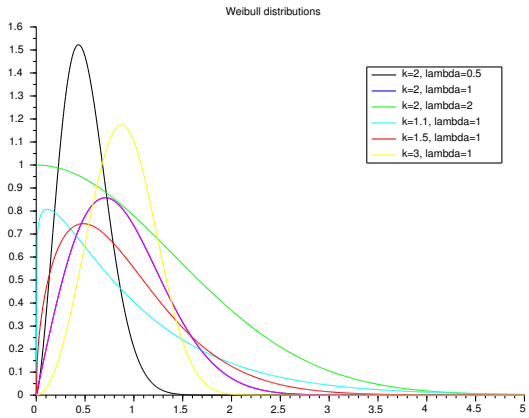

Figure 5.3 – Examples of Weibull distributions.

Remark 5.10. *Note that the Weibull distribution is isolated neither from 0 nor from ∞. The above model therefore does not satisfy the ellipticity condition (5.3). First, we show in Section 5.2.4 below that, in the one-dimensional case, the assumption (5.3) is not necessary, and that homogenization holds under a weaker assumption. Second, we refer to [Bis11] for similar studies (again under assumptions weaker than (5.3)) in higher-dimensional cases.*

Remark 5.11. *The numerical tests of Section 5.4 are performed with the above model, and thus aim at identifying the two parameters λ and k. We however note that nothing in our approach is specific to this particular model using Weibull laws. This choice is only motivated by physical reasons.*

Since the conductances $a_i(x,\omega)$ are all i.i.d. (for any $1 \le i \le d$ and any $x \in \mathbb{Z}^d$), the problem is invariant by any rotation of angle $\pi/2$. The homogenized matrix $A^\star$ is therefore proportional to the identity matrix Id_d, and reads

$$A^\star = K^\star\, \mathrm{Id}_d$$

where $K^\star \in (0,\infty)$ is the homogenized permeability. We can also write that

$$K^\star = e_1 \cdot A^\star e_1.$$

In practice, we only have access to $A^\star_N(\omega)$, which is a symmetric matrix (but is a priori not proportionnal to the identity matrix). All directions are statistically identical, hence we only focus in the sequel on

$$K^\star_N(\omega) := e_1 \cdot A^\star_N(\omega) e_1. \tag{5.18}$$

5.2.4 The one dimensional case

The purpose of this section is two-fold. First, we recall explicit formulas for the homogenized quantities in terms of the microscopic field $A(x,\omega)$. We derive these formulas assuming that (5.3) holds. Second, we show that we can relax Assumption (5.3) and still state a homogenization result.

Explicit formulas in the elliptic case (5.3)

In the one-dimensional case, the problem (5.8)–(5.9) can be analytically solved. We have

$$A_N^\star(\omega) = \left(\frac{1}{N}\sum_{x\in\Lambda_N}\frac{1}{A(x,\omega)}\right)^{-1} \qquad \text{for almost all } \omega. \tag{5.19}$$

Likewise, the problem (5.5)–(5.6) can also be solved, yielding the formula

$$A^\star = \left(\mathbb{E}\left[\frac{1}{A(x,\cdot)}\right]\right)^{-1}, \tag{5.20}$$

which can be evaluated at any $x \in \mathbb{Z}$ due to the stationarity of A.

We note that, as soon as $A(x,\omega) > 0$ a.s. for any $x \in \mathbb{Z}$ and $A^{-1}(x,\cdot) \in L^1(\Omega)$ (this latter condition being independent of x), formulas (5.19) and (5.20) are well-defined. The aim of the next section is to recall that, in the one-dimensional case, these assumptions are enough for homogenization to hold.

Relaxing Assumption (5.3)

In this section, we show that the following assumption is enough for homogenization to hold:

Assumption 5.12. *We assume that the coefficient A is almost surely positive and finite and satisfies*

$$A^{-1}(x,\cdot) \in L^1(\Omega). \tag{5.21}$$

Of course, by stationarity, if (5.21) is satisfied for some $x \in \mathbb{Z}$, then it is satisfied for all $x \in \mathbb{Z}$.

Theorem 5.13. *Let $\mathcal{D}$ be a bounded domain of $\mathbb{R}$, $f \in C^0(\overline{\mathcal{D}})$ and A be a random stationary scalar field (defined on $\mathbb{Z}\times\Omega$) that satisfies (5.21). Let $u_\varepsilon \in \ell^2(\varepsilon\mathbb{Z};\mathbb{R})$ be the unique solution to*

$$\nabla_\varepsilon^\star\big[A(x/\varepsilon,\omega)\nabla_\varepsilon u_\varepsilon(x,\omega)\big] = f(x) \quad \text{in } \mathcal{D}\cap\varepsilon\mathbb{Z}, \qquad u_\varepsilon(x,\omega) = 0 \text{ in } (\mathbb{R}\setminus\mathcal{D})\cap\varepsilon\mathbb{Z}, \tag{5.22}$$

and let $u^\star \in H_0^1(\mathcal{D})$ be the unique solution to the (continuous) boundary value problem

$$-\big[A^\star(u^\star)'\big]' = f \qquad \text{in } \mathcal{D}, \tag{5.23}$$

where $A^\star$ is defined by (5.20).

Then, when $\varepsilon \to 0$, $u_\varepsilon(\cdot,\omega)$ converges to the homogenized solution $u^\star$, in the sense that

$$\varepsilon\sum_{x\in\mathcal{D}\cap\varepsilon\mathbb{Z}}|u_\varepsilon(x,\omega)-u^\star(x)|^2 \xrightarrow[\varepsilon\to 0]{} 0 \quad \text{almost surely.} \tag{5.24}$$

Note that (5.22) is almost surely well-posed. Indeed, since A is stationary and $0 < A(0,\omega) < \infty$ almost surely, we have that, almost surely, $0 < A(x,\omega) < \infty$ for any $x \in \mathbb{Z}$. For those ω, problem (5.22) is well-posed. Likewise, since A is almost surely finite (resp. $A^{-1}(0,\cdot) \in L^1(\Omega)$), we have that $A^\star < \infty$ (resp. $A^\star > 0$) and hence (5.23) is well-posed.

Proof. The proof proceeds by truncation of the coefficient A in the neighbourhood of 0 and $+\infty$. For the sake of simplicity, we take $\mathcal{D} = (0,1)$. For any $m \in \mathbb{N}^\star$, we introduce the coefficient A_m defined on $\mathbb{Z} \times \Omega$ by

$$A_m(x,\omega) := \begin{cases} \dfrac{1}{m} & \text{if} \quad 0 < A(x,\omega) < \dfrac{1}{m}, \\ A(x,\omega) & \text{if} \quad \dfrac{1}{m} \le A(x,\omega) \le m, \\ m & \text{if} \quad A(x,\omega) > m. \end{cases}$$

We set

$$A_m^\star = \left(\mathbb{E}\left[\frac{1}{A_m(0,\cdot)} \right] \right)^{-1}.$$

For almost all ω (i.e. those such that $A(0,\omega) > 0$), we have

$$\lim_{m\to\infty} \frac{1}{A_m(0,\omega)} = \frac{1}{A(0,\omega)},$$
$$\forall m \in \mathbb{N}^\star, \quad 0 < \frac{1}{A_m(0,\omega)} \le 1 + \frac{1}{A(0,\omega)},$$

where the right-hand side of the above second line belongs to $L^1(\Omega)$, in view of the assumption (5.21). Therefore, the dominated convergence theorem implies that

$$\lim_{m\to\infty} A_m^\star = A^\star. \tag{5.25}$$

Let $u_\varepsilon^m \in \ell^2(\varepsilon\mathbb{Z};\mathbb{R})$ be the unique solution to

$$\nabla_\varepsilon^\star\big[A_m(x/\varepsilon,\omega)\nabla_\varepsilon u_\varepsilon^m(x,\omega)\big] = f(x) \ \text{ in } (0,1)\cap\varepsilon\mathbb{Z}, \qquad u_\varepsilon^m(x,\omega) = 0 \text{ in } (\mathbb{R}\setminus(0,1))\cap\varepsilon\mathbb{Z}, \tag{5.26}$$

and let $u_m^\star \in H_0^1(0,1)$ be the unique solution to the (continuous) boundary value problem

$$-\big[A_m^\star (u_m^\star)'\big]' = f \qquad \text{in } (0,1).$$

We write

$$\|u_\varepsilon(\cdot,\omega) - u^\star\|_{\ell^2_\varepsilon} \le \|u_\varepsilon(\cdot,\omega) - u_\varepsilon^m(\cdot,\omega)\|_{\ell^2_\varepsilon} + \|u_\varepsilon^m(\cdot,\omega) - u_m^\star\|_{\ell^2_\varepsilon} + \|u_m^\star - u^\star\|_{\ell^2_\varepsilon} \tag{5.27}$$

where, for any function v,

$$\|v\|_{\ell^2_\varepsilon} := \sqrt{\varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} v^2(x)}.$$

We successively study the three terms of the right-hand side of (5.27).

First, we have

$$\lim_{\varepsilon\to 0} \|u_m^\star - u^\star\|_{\ell^2_\varepsilon} = \|u_m^\star - u^\star\|_{L^2(0,1)},$$

and the convergence (5.25) implies that

$$\lim_{m\to\infty}\lim_{\varepsilon\to 0} \|u_m^\star - u^\star\|_{\ell^2_\varepsilon} = 0. \tag{5.28}$$

Second, the coefficient A_m satisfies the ellipticity condition (5.3), so we infer from Theorem 5.5 that, for any $m \in \mathbb{N}^\star$,

$$\lim_{\varepsilon\to 0} \|u_\varepsilon^m(\cdot,\omega) - u_m^\star\|_{\ell^2_\varepsilon} = 0 \quad \text{a.s.} \tag{5.29}$$

We eventually turn to the first term of the right-hand side of (5.27). Let

$$F_\varepsilon(x) = \varepsilon \sum_{y\in(0,x]\cap\varepsilon\mathbb{Z}} f(y),$$

which satisfies, for any x, $|F_\varepsilon(x)| \le \|f\|_{L^\infty}$. Integrating once the equations (5.22) and (5.26), we can show that there exist two random variables $C_\varepsilon(\omega)$ and $C_\varepsilon^m(\omega)$, independent of x, such that

$$A_m\left(\frac{x}{\varepsilon},\omega\right)\nabla_\varepsilon u_\varepsilon^m(x,\omega) = -F_\varepsilon(x) + C_\varepsilon^m(\omega), \tag{5.30}$$

$$A\left(\frac{x}{\varepsilon},\omega\right)\nabla_\varepsilon u_\varepsilon(x,\omega) = -F_\varepsilon(x) + C_\varepsilon(\omega). \tag{5.31}$$

Using the boundary conditions on u_ε^m and u_ε, we get

$$C_\varepsilon(\omega) = \frac{\mathcal{N}_\varepsilon(\omega)}{\mathcal{D}_\varepsilon(\omega)} \quad \text{and} \quad C_\varepsilon^m(\omega) = \frac{\mathcal{N}_\varepsilon^m(\omega)}{\mathcal{D}_\varepsilon^m(\omega)}$$

where

$$\mathcal{D}_\varepsilon(\omega) = \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} A\left(\frac{x}{\varepsilon},\omega\right)^{-1}, \qquad \mathcal{N}_\varepsilon(\omega) = \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} A\left(\frac{x}{\varepsilon},\omega\right)^{-1} F_\varepsilon(x),$$

$$\mathcal{D}_\varepsilon^m(\omega) = \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} A_m\left(\frac{x}{\varepsilon},\omega\right)^{-1}, \qquad \mathcal{N}_\varepsilon^m(\omega) = \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} A_m\left(\frac{x}{\varepsilon},\omega\right)^{-1} F_\varepsilon(x).$$

All these quantities are well-defined for almost all ω. We claim that

$$\lim_{m\to\infty} \limsup_{\varepsilon\to 0} |C_\varepsilon(\omega) - C_\varepsilon^m(\omega)| = 0 \quad \text{a.s.} \tag{5.32}$$

To prove this claim, we start by writing that

$$C_\varepsilon(\omega) - C_\varepsilon^m(\omega) = \frac{\mathcal{N}_\varepsilon(\omega) - \mathcal{N}_\varepsilon^m(\omega)}{\mathcal{D}_\varepsilon(\omega)} + \frac{\mathcal{N}_\varepsilon^m(\omega)}{\mathcal{D}_\varepsilon^m(\omega)\mathcal{D}_\varepsilon(\omega)}\left(\mathcal{D}_\varepsilon^m(\omega) - \mathcal{D}_\varepsilon(\omega)\right). \tag{5.33}$$

Introduce

$$b_m(x,\omega) = \left|\frac{1}{A_m(x,\omega)} - \frac{1}{A(x,\omega)}\right| \quad \text{and} \quad \mathcal{B}_\varepsilon^m(\omega) = \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} b_m\left(\frac{x}{\varepsilon},\omega\right).$$

For any $m \in \mathbb{N}^\star$, we get

$$|\mathcal{N}_\varepsilon^m(\omega) - \mathcal{N}_\varepsilon(\omega)| \le \|f\|_{L^\infty}\, \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} b_m\left(\frac{x}{\varepsilon},\omega\right) = \|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega), \tag{5.34}$$

$$|\mathcal{D}_\varepsilon^m(\omega) - \mathcal{D}_\varepsilon(\omega)| \le \varepsilon \sum_{x\in(0,1)\cap\varepsilon\mathbb{Z}} b_m\left(\frac{x}{\varepsilon},\omega\right) = \mathcal{B}_\varepsilon^m(\omega), \tag{5.35}$$

$$|\mathcal{N}_\varepsilon^m(\omega)| \le \|f\|_{L^\infty}\, \mathcal{D}_\varepsilon^m(\omega). \tag{5.36}$$

Using the ergodic theorem for the stationary functions A^{-1}, A_m^{-1} and b_m, we have that, for any $m \in \mathbb{N}^\star$, almost surely,

$$\lim_{\varepsilon\to 0} \mathcal{D}_\varepsilon(\omega) = \frac{1}{A^\star}, \qquad \lim_{\varepsilon\to 0} \mathcal{D}_\varepsilon^m(\omega) = \frac{1}{A_m^\star}, \qquad \lim_{\varepsilon\to 0} \mathcal{B}_\varepsilon^m(\omega) = \mathcal{B}_\star^m := \mathbb{E}\left[\left|\frac{1}{A_m(0,\cdot)} - \frac{1}{A(0,\cdot)}\right|\right]. \tag{5.37}$$

We introduce

$$\Omega_{\rm conv} = \left\{\omega \in \Omega;\ \lim_{\varepsilon\to 0} \mathcal{D}_\varepsilon(\omega) = \frac{1}{A^\star} \ \text{ and } \forall m \in \mathbb{N}^\star,\ \lim_{\varepsilon\to 0} \mathcal{D}_\varepsilon^m(\omega) = \frac{1}{A_m^\star},\ \lim_{\varepsilon\to 0} \mathcal{B}_\varepsilon^m(\omega) = \mathcal{B}_\star^m\right\}$$

and we deduce that $\mathbb{P}(\Omega_{\rm conv}) = 1$.

Let $\omega \in \Omega_{\rm conv}$. In view of (5.37), we know that there exists $\varepsilon_0^m(\omega)$ such that, for any $\varepsilon < \varepsilon_0^m(\omega)$, we have

$$\frac{1}{2A^\star} \le \mathcal{D}_\varepsilon(\omega), \quad \frac{1}{2A_m^\star} \le \mathcal{D}_\varepsilon^m(\omega) \le \frac{3}{2A_m^\star}. \tag{5.38}$$

We thus infer from (5.33), (5.38), (5.34), (5.36) and (5.35) that, for any $\omega \in \Omega_{\rm conv}$, any $m \in \mathbb{N}^\star$ and any $\varepsilon < \varepsilon_0^m(\omega)$, we have

$$\begin{aligned} |C_\varepsilon(\omega) - C_\varepsilon^m(\omega)| &\le 2A^\star\, |\mathcal{N}_\varepsilon(\omega) - \mathcal{N}_\varepsilon^m(\omega)| + 4A^\star A_m^\star\, |\mathcal{N}_\varepsilon^m(\omega)|\ |\mathcal{D}_\varepsilon^m(\omega) - \mathcal{D}_\varepsilon(\omega)| \\ &\le 2A^\star\, \|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega) + 4A^\star A_m^\star\, \|f\|_{L^\infty}\, \mathcal{D}_\varepsilon^m(\omega)\, \mathcal{B}_\varepsilon^m(\omega) \\ &\le 2A^\star\, \|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega) + 6A^\star\, \|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega). \end{aligned} \tag{5.39}$$

Hence, for any $\omega \in \Omega_{\rm conv}$ and any $m \in \mathbb{N}^\star$, we have

$$\limsup_{\varepsilon\to 0} |C_\varepsilon(\omega) - C_\varepsilon^m(\omega)| \le 8A^\star\, \|f\|_{L^\infty}\, \mathcal{B}_m^\star.$$

The dominated convergence theorem implies that $\lim_{m\to\infty} \mathcal{B}_m^\star = 0$, hence, for any $\omega \in \Omega_{\rm conv}$, we have

$$\lim_{m\to\infty} \limsup_{\varepsilon\to 0} |C_\varepsilon(\omega) - C_\varepsilon^m(\omega)| = 0.$$

Since $\mathbb{P}(\Omega_{\rm conv}) = 1$, we have proved the claim (5.32).

We now proceed and deduce from (5.30) and (5.31) that

$$\begin{aligned} u_\varepsilon^m(z,\omega) &= \varepsilon \sum_{x\in(0,z)\cap\varepsilon\mathbb{Z}} A_m\left(\frac{x}{\varepsilon},\omega\right)^{-1} (C_\varepsilon^m(\omega) - F_\varepsilon(x)), \\ u_\varepsilon(z,\omega) &= \varepsilon \sum_{x\in(0,z)\cap\varepsilon\mathbb{Z}} A\left(\frac{x}{\varepsilon},\omega\right)^{-1} (C_\varepsilon(\omega) - F_\varepsilon(x)), \end{aligned}$$

hence

$$|u_\varepsilon^m(z,\omega) - u_\varepsilon(z,\omega)| \le |C_\varepsilon^m(\omega) - C_\varepsilon(\omega)|\ \mathcal{D}_\varepsilon^m(\omega) + (|C_\varepsilon(\omega)| + \|f\|_{L^\infty})\ \mathcal{B}_\varepsilon^m(\omega).$$

Using that $|C_\varepsilon(\omega)| \le \|f\|_{L^\infty}$, we deduce that

$$\|u_\varepsilon^m(\cdot,\omega) - u_\varepsilon(\cdot,\omega)\|_{\ell_\varepsilon^2} \le |C_\varepsilon^m(\omega) - C_\varepsilon(\omega)|\ \mathcal{D}_\varepsilon^m(\omega) + 2\|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega).$$

For any $\omega \in \Omega_{\rm conv}$, any $m \in \mathbb{N}^\star$ and any $\varepsilon < \varepsilon_0^m(\omega)$, using (5.39) and (5.38), we obtain that

$$\|u_\varepsilon^m(\cdot,\omega) - u_\varepsilon(\cdot,\omega)\|_{\ell_\varepsilon^2} \le 8A^\star\, \|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega)\, \frac{3}{2A_m^\star} + 2\|f\|_{L^\infty}\, \mathcal{B}_\varepsilon^m(\omega),$$

hence, for any $\omega \in \Omega_{\rm conv}$ and any $m \in \mathbb{N}^\star$,

$$\limsup_{\varepsilon\to 0} \|u_\varepsilon^m(\cdot,\omega) - u_\varepsilon(\cdot,\omega)\|_{\ell^2_\varepsilon} \leq 8A^\star \ \|f\|_{L^\infty} \ \mathcal{B}_\star^m \ \frac{3}{2A_m^\star} + 2\|f\|_{L^\infty} \ \mathcal{B}_\star^m,$$

and thus, almost surely,

$$\lim_{m\to\infty} \limsup_{\varepsilon\to 0} \|u_\varepsilon^m(\cdot,\omega) - u_\varepsilon(\cdot,\omega)\|_{\ell^2_\varepsilon} = 0. \tag{5.40}$$

Collecting (5.27), (5.28), (5.29) and (5.40), we obtain that

$$\limsup_{\varepsilon\to 0} \|u_\varepsilon(\cdot,\omega) - u^\star\|_{\ell^2_\varepsilon} = 0 \quad \text{a.s.},$$

which is the convergence (5.24). □

The case of Weibull laws

Following Section 5.2.3, assume that the conductances are given by (5.17), i.e. are distributed according to the Weibull law of parameter $(\lambda^4, k/4)$. For any $k > 0$, Assumption (5.3) is not satisfied. However, when $k > 4$, Assumption (5.21) is satisfied: in view of Theorem 5.13, homogenization holds and the homogenized coefficient is given by

$$A^\star = \frac{\lambda^4}{\Gamma(1-4/k)}, \tag{5.41}$$

where Γ is the Euler Gamma function defined for any $z > 0$ by

$$\Gamma(z) = \int_0^\infty t^{z-1} \exp(-t)\, dt. \tag{5.42}$$

The variance of $A_N^\star$ is finite if and only if $k > 8$. In the sequel, we work in the range $k > 8$.

5.3 A parameter fitting problem

We now describe the problem we consider, first in the general case (Section 5.3.1), next in the one-dimensional case (Section 5.3.2). In that latter section, we also motivate our choice of macroscopic quantities from which we fit the parameters of the Weibull law.

5.3.1 General case

We assume that we are given two observed quantities, the first coefficient of the macroscopic permeability matrix (see (5.18))

$$K_N^{\star,\rm obs}(\omega) = e_1 \cdot A_N^{\star,\rm obs}(\omega) e_1$$

and its relative variance $S_N^{\rm obs}$ for some parameter $\theta_{\rm obs} = (\lambda_{\rm obs}, k_{\rm obs})$ of the Weibull law. Note that the relative variance crucially depends on the size N^d of the finite box on which it is measured (in contrast to the apparent permeability, which converges to a finite value when $N \to \infty$). We assume here that we know this size. In practice, these three quantities, N, $K_N^{\star,\rm obs}$ and $S_N^{\rm obs}$, can be obtained by physical experiments. We therefore assume that there exists $\theta_{\rm obs}$ and N such that

$$\mathbb{E}[K_N^\star(\cdot,\theta_{\rm obs})] = K_N^{\star,\rm obs}, \qquad \mathbb{V}\mathrm{ar}\mathrm{R}[K_N^\star(\cdot,\theta_{\rm obs})] = S_N^{\rm obs}, \tag{5.43}$$

where, we recall,

$$\mathbb{V}\mathrm{arR}[K_N^\star(\cdot,\theta_{\mathrm{obs}})] = \frac{\mathbb{V}\mathrm{ar}[K_N^\star(\cdot,\theta_{\mathrm{obs}})]}{\left(\mathbb{E}[K_N^\star(\cdot,\theta_{\mathrm{obs}})]\right)^2}.$$

Given N, $K_N^{\star,\mathrm{obs}}$ and S_N^{obs}, our aim is to recover (an approximation of) θ_{obs}. To that aim, we consider the function

$$F_{N,M}: \quad \begin{cases} (\mathbb{R}_+^\star)^2 & \to \quad \mathbb{R}_+ \\ \theta & \mapsto \left(\dfrac{\overline{K}_{N,M}^\star(\theta)}{K_N^{\star,\mathrm{obs}}} - 1\right)^2 + \left(\dfrac{S_{N,M}(\theta)}{S_N^{\mathrm{obs}}} - 1\right)^2 \end{cases} \tag{5.44}$$

which penalizes the sum of the (relative) errors between

- on the one hand, $\overline{K}_{N,M}^\star(\theta)$ (which is an empirical estimator of $\mathbb{E}\left[K_N^\star(\cdot,\theta)\right]$ when M is large, see (5.10) and (5.18)) and $K_N^{\star,\mathrm{obs}}$
- and, on the other hand, $S_{N,M}(\theta)$ (which is an empirical estimator of the relative variance of $K_N^\star(\omega,\theta)$ when M is large, see (5.12) and (5.18)) and S_N^{obs}.

Of course, different weights could be assigned to the error on the permeability and the error on its relative variance. We eventually cast our parameter fitting problem in the form of the optimization problem

$$\inf_{\theta=(\lambda,k)\in(0,\infty)\times\mathcal{K}} F_{N,M}(\theta),$$

where $\mathcal{K} \subset (0,\infty)$ is the admissible set of parameters k such that homogenization holds (even if Assumption (5.3) is not satisfied for any $k > 0$) and the variance of $K_N^\star$ is also well-defined. In the one-dimensional case we focus on in this article, $\mathcal{K} = (8,\infty)$.

Note that $F_{N,M}(\theta)$ is random, as it depends on the realizations used to evaluate $\overline{K}_{N,M}^\star(\theta)$ and $S_{N,M}(\theta)$ (see (5.10) and (5.12)). For any θ, in the limit when $M \to \infty$, $F_{N,M}(\theta)$ converges almost surely to the deterministic limit

$$F_N(\theta) = \left(\frac{\mathbb{E}\left[K_N^\star(\cdot,\theta)\right]}{K_N^{\star,\mathrm{obs}}} - 1\right)^2 + \left(\frac{\mathbb{V}\mathrm{arR}[K_N^\star(\cdot,\theta)]}{S_N^{\mathrm{obs}}} - 1\right)^2.$$

Under Assumption (5.43), we have

$$F_N(\theta) = \left(\frac{\mathbb{E}\left[K_N^\star(\cdot,\theta)\right]}{\mathbb{E}\left[K_N^\star(\cdot,\theta_{\mathrm{obs}})\right]} - 1\right)^2 + \left(\frac{\mathbb{V}\mathrm{arR}[K_N^\star(\cdot,\theta)]}{\mathbb{V}\mathrm{arR}[K_N^\star(\cdot,\theta_{\mathrm{obs}})]} - 1\right)^2.$$

5.3.2 The one-dimensional case

Recall that, in that case, $K_N^\star = A_N^\star$ (see (5.18)).

Theoretical result

We first identify the limit when $N \to \infty$ of F_N, which we recall reads

$$F_N(\theta) = \left(\frac{\mathbb{E}\left[A_N^\star(\cdot,\theta)\right]}{\mathbb{E}\left[A_N^\star(\cdot,\theta_{\mathrm{obs}})\right]} - 1\right)^2 + \left(\frac{\mathbb{V}\mathrm{arR}[A_N^\star(\cdot,\theta)]}{\mathbb{V}\mathrm{arR}[A_N^\star(\cdot,\theta_{\mathrm{obs}})]} - 1\right)^2. \tag{5.45}$$

Obviously, the first term above converges to

$$\left(\frac{A^\star(\theta)}{A^\star(\theta_{\rm obs})} - 1\right)^2,$$

with, following (5.41),

$$A^\star = \frac{\lambda^4}{\Gamma(1-4/k)}$$

where Γ is the Euler Gamma function.

For the second term of (5.45), it is clear that $\mathbb{V}\mathrm{arR}[A_N^\star(\cdot,\theta)]$ vanishes in the limit $N \to \infty$, since $A_N^\star(\cdot,\theta)$ converges almost surely to a deterministic limit. Furthermore, equation (5.19) implies that

$$\mathbb{V}\mathrm{ar}[A_N^\star] = \frac{(A^\star)^4}{N}\,\mathbb{V}\mathrm{ar}\left[\frac{1}{A(0,\cdot)}\right] + o\left(\frac{1}{N}\right). \tag{5.46}$$

The conductances are distributed according to a Weibull law (see (5.17)), therefore

$$\mathbb{V}\mathrm{ar}[A_N^\star] = \frac{\lambda^{16}}{N\,\Gamma(1-4/k)^4}\left(\frac{\Gamma(1-8/k)}{\lambda^8} - \frac{\Gamma(1-4/k)^2}{\lambda^8}\right) + o\left(\frac{1}{N}\right),$$

hence the relative variance reads

$$\mathbb{V}\mathrm{arR}[A_N^\star] = \frac{1}{N}\left(\frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2} - 1\right) + o\left(\frac{1}{N}\right), \tag{5.47}$$

which implies that

$$\lim_{N\to\infty} \frac{\mathbb{V}\mathrm{arR}[A_N^\star(\cdot,\theta)]}{\mathbb{V}\mathrm{arR}[A_N^\star(\cdot,\theta_{\rm obs})]} = \frac{\frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2} - 1}{\frac{\Gamma(1-8/k_{\rm obs})}{\Gamma(1-4/k_{\rm obs})^2} - 1}.$$

In the one-dimensional case, we are thus able to identify the limit as $N \to \infty$ of $F_N(\theta)$, which reads

$$F_\infty^{\rm 1D}(\theta) := \lim_{N\to\infty} F_N(\theta) = \left(\frac{\lambda^4}{\lambda_{\rm obs}^4}\frac{\Gamma(1-4/k_{\rm obs})}{\Gamma(1-4/k)} - 1\right)^2 + \left(\frac{\frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2} - 1}{\frac{\Gamma(1-8/k_{\rm obs})}{\Gamma(1-4/k_{\rm obs})^2} - 1} - 1\right)^2. \tag{5.48}$$

Obviously, this function is minimal (and vanishes) when $\theta = \theta_{\rm obs}$. It turns out that this minimizer is the unique minimizer, as shown below.

Lemma 5.14. *The function $F_\infty^{\rm 1D}$ defined by (5.48) has a unique minimizer, which is $\theta_{\rm obs}$.*

Homogenization is an averaging process, which filters out many features of the microscopic coefficient A. These features cannot be recovered from the knowledge of macroscopic quantities. The above lemma shows (in the one-dimensional case) that, if one assumes a given form for the probability distribution of A (here, a Weibull distribution), then one is able to recover the two parameters of that law on the basis of two macroscopic quantities, the permeability and its relative variance.

It is also obvious from (5.41) that knowing the macroscopic permeability is not enough to uniquely determine the two parameters λ and k of the Weibull law. Additional information is needed. Our choice of considering the relative variance of the permeability is motivated by the following observation. This quantity, in the one-dimensional case, only depends (at first order in N) on k and does not depend on λ, as can be seen on (5.47).

Knowing this quantity is therefore very useful to estimate the parameter k. Once k has been identified, knowing the macroscopic permeability yields, using (5.41), an estimation of the parameter λ.

Of course, it is likely that the knowledge of quantities of interest alternate to the relative variance of the permeability may also prove useful to determine the unknown parameters. Note also that such alternate relevant quantities should be "different enough" from the homogenized permeability to indeed bring new information. We do not pursue in that direction.

We plot on Figure 5.4 the function $\theta \mapsto F_\infty^{1\mathrm{D}}(\theta)$ for $\lambda_{\mathrm{obs}} = 1$ and $k_{\mathrm{obs}} = 15$. We observe that the function is not degenerated at its minimum, in the sense that its Hessian matrix at θ_{obs} is positive definite, with eigenvalues equal to 16 and 0.04. We thus expect that a standard algorithm (such as the Newton algorithm) will be able to converge to the minimizer of $F_\infty^{1\mathrm{D}}$. This is indeed the case, as shown in Section 5.4.

Proof of Lemma 5.14. The proof consists of three steps: in Step 1, we recall (and prove for the sake of completeness) that $\ln \Gamma$ is a convex function. In Step 2, we prove that the function

$$\zeta : k \mapsto \frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2}$$

is monotone (hence injective). We conclude in Step 3.

Step 1. From (5.42), we compute that, for any $z > 0$,

$$\begin{aligned} \Gamma'(z) &= \int_0^\infty (\ln t)\, t^{z-1} \exp(-t)\, dt, \\ \Gamma''(z) &= \int_0^\infty (\ln t)^2\, t^{z-1} \exp(-t)\, dt, \end{aligned}$$

therefore $\Gamma''(z) > 0$ and Γ is positive and convex on $(0, \infty)$. In addition, we have

$$(\ln \Gamma)''(z) = \frac{\Gamma(z)\Gamma''(z) - (\Gamma'(z))^2}{\Gamma^2(z)}$$

which is positive, in view of the Cauchy-Schwartz inequality:

$$(\Gamma'(z))^2 = \left(\int_0^\infty \left((\ln t) \sqrt{t^{z-1} \exp(-t)} \right) \sqrt{t^{z-1} \exp(-t)}\, dt \right)^2 < \Gamma''(z)\Gamma(z).$$

Therefore, $\ln \Gamma$ is a strictly convex function.

Step 2. We define the function

$$\zeta : k \mapsto \frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2},$$

the derivative of which reads

$$\zeta'(k) = \frac{\Gamma(1-8/k)}{\Gamma(1-4/k)^2} \left(\frac{8}{k^2} \frac{\Gamma'(1-8/k)}{\Gamma(1-8/k)} - \frac{8}{k^2} \frac{\Gamma'(1-4/k)}{\Gamma(1-4/k)} \right).$$

For any $k > 0$, we have that $1 - 8/k < 1 - 4/k$. As a consequence of $\ln \Gamma$ being strictly convex, we have that its derivative is increasing, therefore

$$\frac{\Gamma'(1-8/k)}{\Gamma(1-8/k)} < \frac{\Gamma'(1-4/k)}{\Gamma(1-4/k)}.$$

We can now conclude that $\zeta'(k) < 0$, hence ζ is decreasing.

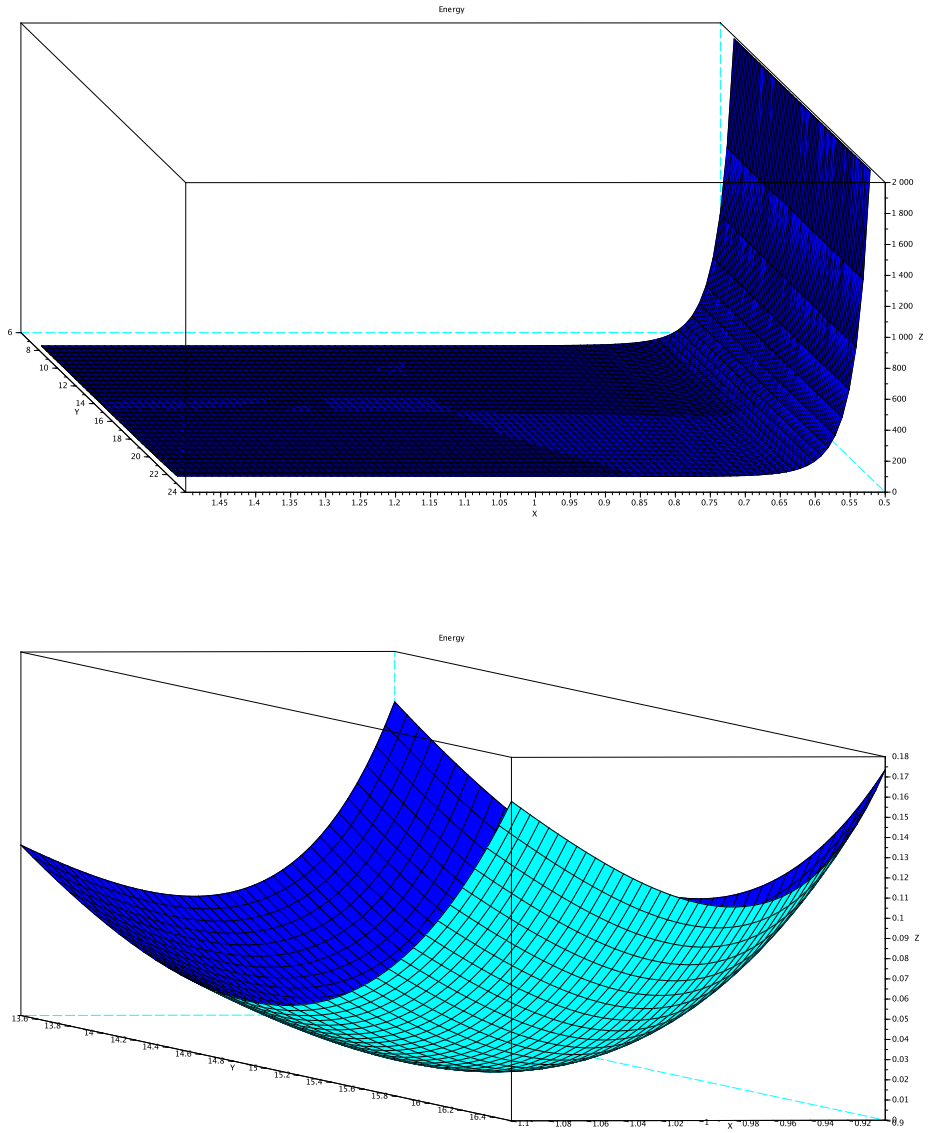

Figure 5.4 – Plot of $\theta \mapsto F_{\infty}^{1\mathrm{D}}(\theta)$ for $\lambda_{\mathrm{obs}} = 1$ and $k_{\mathrm{obs}} = 15$. The bottom plot is a zoom of the top plot around the minimizer.

Step 3. By definition of ζ, we have that

$$F_\infty^{1\mathrm{D}}(\theta) = \left(\frac{\lambda^4}{\lambda_{\mathrm{obs}}^4}\frac{\Gamma(1-4/k_{\mathrm{obs}})}{\Gamma(1-4/k)} - 1\right)^2 + \left(\frac{\zeta(k)-1}{\zeta(k_{\mathrm{obs}})-1} - 1\right)^2.$$

We obviously have that θ_{obs} is a minimizer of $F_\infty^{1\mathrm{D}}$, with $F_\infty^{1\mathrm{D}}(\theta_{\mathrm{obs}}) = 0$. Conversely, let θ be a minimizer of $F_\infty^{1\mathrm{D}}$. We thus have $F_\infty^{1\mathrm{D}}(\theta) = 0$, which implies that $\zeta(k) = \zeta(k_{\mathrm{obs}})$. The function ζ being monotone, this implies that $k = k_{\mathrm{obs}}$. Since the first term in $F_\infty^{1\mathrm{D}}(\theta)$ also vanishes, we obtain that $\lambda = \lambda_{\mathrm{obs}}$ as well. This concludes the proof. □

Practical situation

In the general (i.e. multi-dimensional) case, we have introduced in (5.44) the function $F_{N,M}$ that we wish to minimize. Turning next to the one-dimensional case, we have theoretically identified its limit when $M \to \infty$ and $N \to \infty$. In practice, we cannot take any of these limits, and have thus to work with $F_{N,M}$ defined by

$$F_{N,M}(\theta) = \left(\frac{\overline{K}^\star_{N,M}(\theta)}{K_N^{\star,\mathrm{obs}}} - 1\right)^2 + \left(\frac{S_{N,M}(\theta)}{S_N^{\mathrm{obs}}} - 1\right)^2.$$

In view of (5.46), we see that, when $M \to \infty$ and $N \to \infty$, the relative variance $S_{N,M}(\theta)$ is close to

$$S_{N,M}(\theta) \approx \frac{(A^\star)^2}{N}\,\mathbb{V}\mathrm{ar}\left[\frac{1}{A(0,\cdot)}\right] = \frac{\mathbb{E}[W^{-8}] - \left(\mathbb{E}[W^{-4}]\right)^2}{N\,\left(\mathbb{E}[W^{-4}]\right)^2} = \frac{\mathbb{E}[W^{-8}]}{N\,\left(\mathbb{E}[W^{-4}]\right)^2} - \frac{1}{N},$$

where W is a random variable distributed according to the Weibull law $\mathcal{W}(1,k)$. Likewise,

$$\overline{K}^\star_{N,M}(\theta) \approx A^\star = \frac{\lambda^4}{\mathbb{E}[W^{-4}]}.$$

Let $\{u_i(\omega)\}_{i=1}^N$ be a sequence of i.i.d. random variables uniformly distributed in $[0,1]$. We define

$$w_i(k,\omega) := (-\ln(1-u_i(\omega)))^{-1/k}, \tag{5.49}$$

so that $\{1/w_i(k,\omega)\}_{i=1}^N$ are i.i.d. random variables distributed according to $\mathcal{W}(1,k)$. In the sequel, we approximate the function to minimize by

$$\widetilde{F}_N^{1\mathrm{D}}(\theta,\omega) = \left(\frac{\lambda^4}{K_N^{\star,\mathrm{obs}}}\left[\frac{1}{N}\sum_{i=1}^N w_i^4(k,\omega)\right]^{-1} - 1\right)^2 + \left(\frac{1}{S_N^{\mathrm{obs}}}\left[\frac{\sum_{i=1}^N w_i^8(k,\omega)}{\left(\sum_{i=1}^N w_i^4(k,\omega)\right)^2} - \frac{1}{N}\right] - 1\right)^2. \tag{5.50}$$

This function is consistent in the sense that it almost surely converges, when $N \to \infty$, to the exact function (5.48). On the other hand, $\widetilde{F}_N^{1\mathrm{D}}(\theta,\omega)$ is random, and thus somewhat mimics the difficulties that one would encounter in the multi-dimensional case when working with $F_{N,M}(\theta)$.

5.4 Numerical results

We briefly explain in Section 5.4.1 how in practice we minimize the function (5.50), before turning in Section 5.4.2 to our numerical results. As pointed out in the introduction, we only consider here the one-dimensional case, and postpone the study of two-dimensional examples to the future work [LMOS].

5.4.1 Optimization algorithm

We provide in Appendix 5.5 expressions for the first and second derivatives of the function $\widetilde{F}_N^{1\mathrm{D}}(\theta,\omega)$ defined by (5.50) with respect to $\theta = (\lambda, k)$. We are thus in position to use the Newton algorithm, and compute a sequence θ_j according to

$$\theta_{j+1} = \theta_j - \mu_j \left[\mathcal{H}\left(\widetilde{F}_N^{1\mathrm{D}}\right)(\theta_j)\right]^{-1} \nabla \widetilde{F}_N^{1\mathrm{D}}(\theta_j), \tag{5.51}$$

where $\mathcal{H}\left(\widetilde{F}_N^{1\mathrm{D}}\right) \in \mathbb{R}^{2\times 2}$ is the Hessian matrix of $\widetilde{F}_N^{1\mathrm{D}}$ and $\nabla \widetilde{F}_N^{1\mathrm{D}} \in \mathbb{R}^2$ is the gradient of $\widetilde{F}_N^{1\mathrm{D}}$ (for the sake of simplicity, we keep implicit the dependence with respect to ω). In turn, $\mu_j > 0$ is the step-size by which we move. To choose μ_j, we have used a line-search algorithm (along the descent direction prescribed by the Newton algorithm) using Goldstein (respectively Armijo) rule to increase (respectively decrease) the step-size.

We note that the function $\theta \mapsto F_\infty^{1\mathrm{D}}(\theta)$ is not convex. It is possible to find some θ such that the Hessian matrix $\mathcal{H}\left(F_\infty^{1\mathrm{D}}\right)(\theta)$ is not positive definite, but rather has (at least) one negative eigenvalue. We thus cannot expect the function $\theta \mapsto \widetilde{F}_N^{1\mathrm{D}}(\theta)$ to be convex (even for large values of N), and the Newton algorithm to be globally convergent. We are therefore careful to start the Newton iterations from an initial guess θ_0 (given by physical experiments) that we hope to be close enough to the minimizer of $\widetilde{F}_N^{1\mathrm{D}}$.

5.4.2 Numerical results

In all what follows, we set $N = 10^5$.

Robustness of the algorithm with respect to the initial guess

Our first numerical test is a simple one, to check whether the Newton algorithm (5.51) is indeed able to minimize the function $\theta \mapsto \widetilde{F}_N^{1\mathrm{D}}(\theta,\omega)$. We pick once for all one realization of the i.i.d. random variables $\{u_i(\omega)\}_{1\le i\le N}$ (which, we recall, are uniformly distributed in $[0,1]$). We then build $\{w_i(k,\omega)\}_{1\le i\le N}$ according to (5.49) and consider the function $\theta \mapsto \widetilde{F}_N^{1\mathrm{D}}(\theta,\omega)$ defined by (5.50), where the observed quantities are defined by

$$K_N^{\star,\mathrm{obs}} = \lambda_{\mathrm{obs}}^4 \left[\frac{1}{N}\sum_{i=1}^N w_i^4(k_{\mathrm{obs}},\omega)\right]^{-1}, \qquad S_N^{\mathrm{obs}} = \frac{\sum_{i=1}^N w_i^8(k_{\mathrm{obs}},\omega)}{\left(\sum_{i=1}^N w_i^4(k_{\mathrm{obs}},\omega)\right)^2} - \frac{1}{N},$$

with $\lambda_{\mathrm{obs}} = 1$ and $k_{\mathrm{obs}} = 15$. The function $\theta \mapsto \widetilde{F}_N^{1\mathrm{D}}(\theta,\omega)$ obviously vanishes at $\theta_{\mathrm{obs}} = (\lambda_{\mathrm{obs}}, k_{\mathrm{obs}})$.

We run the Newton algorithm (5.51) starting from several initial guesses θ_0, and check that it indeed always converges to θ_{obs} in a limited number of iterations. We also observe that, for some initial guesses, using an adaptive step-size μ_j as in (5.51) is critical: in contrast, if one uses the step-size $\mu_j = 1$, then the algorithm may not converge, or converges after a much larger number of iterations.

Robustness with respect to statistical noise

For our second test, we proceed as follows. We first set $\theta_{\mathrm{ref}} = (\lambda_{\mathrm{ref}}, k_{\mathrm{ref}}) = (1, 15)$ and pick one realization of the i.i.d. random variables $\{u_i(\overline{\omega})\}_{1\le i\le N}$ (which, we recall, are uniformly

distributed in $[0,1]$). We then build $\{w_i(k_{\rm ref},\overline{\omega})\}_{1\le i\le N}$ according to (5.49) and define the macroscopic observed quantities as

$$K_N^{\star,\rm obs} = \lambda_{\rm ref}^4 \left[\frac{1}{N}\sum_{i=1}^N w_i^4(k_{\rm ref},\overline{\omega})\right]^{-1}, \qquad S_N^{\rm obs} = \frac{\sum_{i=1}^N w_i^8(k_{\rm ref},\overline{\omega})}{\left(\sum_{i=1}^N w_i^4(k_{\rm ref},\overline{\omega})\right)^2} - \frac{1}{N}. \tag{5.52}$$

We now fix the initial guess $\theta_0 = (1.1, 16.5)$ (10% off the reference value $\theta_{\rm ref}$) and set $M = 500$. For any $1 \le m \le M$, we perform the following procedure:

- we draw a realization of N i.i.d. random variables $\{u_i(\omega_m)\}_{1\le i\le N}$ which is independent of the realization $\{u_i(\omega_{m'})\}_{1\le i\le N}$ for any $m' \ne m$, and independent of the realization $\{u_i(\overline{\omega})\}_{1\le i\le N}$ used to compute $K_N^{\star,\rm obs}$ and $S_N^{\rm obs}$ in (5.52);
- using $\{u_i(\omega_m)\}_{1\le i\le N}$, we build $w_i(k,\omega_m)$ according to (5.49) and we consider the function $\theta \mapsto \widetilde{F}_N^{\rm 1D}(\theta,\omega_m)$ defined by (5.50), i.e.

$$\widetilde{F}_N^{\rm 1D}(\theta,\omega_m) = \left(\frac{\lambda^4}{K_N^{\star,\rm obs}}\left[\frac{1}{N}\sum_{i=1}^N w_i^4(k,\omega_m)\right]^{-1} - 1\right)^2 + \left(\frac{1}{S_N^{\rm obs}}\left[\frac{\sum_{i=1}^N w_i^8(k,\omega_m)}{\left(\sum_{i=1}^N w_i^4(k,\omega_m)\right)^2} - \frac{1}{N}\right] - 1\right)^2.$$

  Recall that the macroscopic observed quantities are independent of ω_m.
- we run the Newton algorithm (5.51) to minimize the function $\theta \mapsto \widetilde{F}_N^{\rm 1D}(\theta,\omega_m)$. The optimal parameter found by the algorithm depends on ω_m and is denoted $\theta_{\rm opt}(\omega_m)$. Since the realization ω_m is different from the reference realization $\overline{\omega}$, we have in general $\theta_{\rm opt}(\omega_m) \ne \theta_{\rm ref}$.

We show on Figure 5.5 the histogram of the optimal parameters $\theta_{\rm opt}(\omega_m)$ for $1 \le m \le M$. We see that these histograms are centered close to the reference value ($k_{\rm ref}$, resp. $\lambda_{\rm ref}$). There is however a small bias, i.e. $\mathbb{E}\left(\theta_{\rm opt}\right) \ne \theta_{\rm ref}$. We also observe that the width of these histograms (related to the variance of $k_{\rm opt}$ and $\lambda_{\rm opt}$) is quite small.

Remark 5.15. *Of course, the variance of $k_{\rm opt}$ and $\lambda_{\rm opt}$ is related to N. In the limit $N \to \infty$, the function $\widetilde{F}_N^{\rm 1D}(\theta,\omega)$ almost surely converges to the deterministic limit $F_\infty^{\rm 1D}(\theta)$ defined by (5.48), and we thus expect $k_{\rm opt}$ and $\lambda_{\rm opt}$ to almost surely converge to a deterministic limit. But this is not the regime we are interested in, since in practice (in the two-dimensional case), we have to work with the* random *function $F_{N,M}$.*

We next compare the variance of $\theta_{\rm opt}$ with the amount of randomness introduced in the function $\widetilde{F}_N^{\rm 1D}(\cdot,\omega)$ defined by (5.50). By construction,

$$\widetilde{F}_N^{\rm 1D}(\theta,\omega) = \left(\frac{K_N^\star(\theta,\omega)}{K_N^{\star,\rm obs}} - 1\right)^2 + \left(\frac{S_N(k,\omega)}{S_N^{\rm obs}} - 1\right)^2$$

with

$$S_N(k,\omega) = \left[\frac{\sum_{i=1}^N w_i^8(k,\omega)}{\left(\sum_{i=1}^N w_i^4(k,\omega)\right)^2} - \frac{1}{N}\right],$$

which is an approximation of the relative variance of $K_N^\star(\theta,\omega)$. We show on Figure 5.6 the histograms, for $1 \le m \le M$, of $K_N^\star(\theta_0,\omega_m)$ and of $S_N(k_0,\omega_m)$, for the initial guess parameter $\theta_0 = (1.1, 16.5)$.

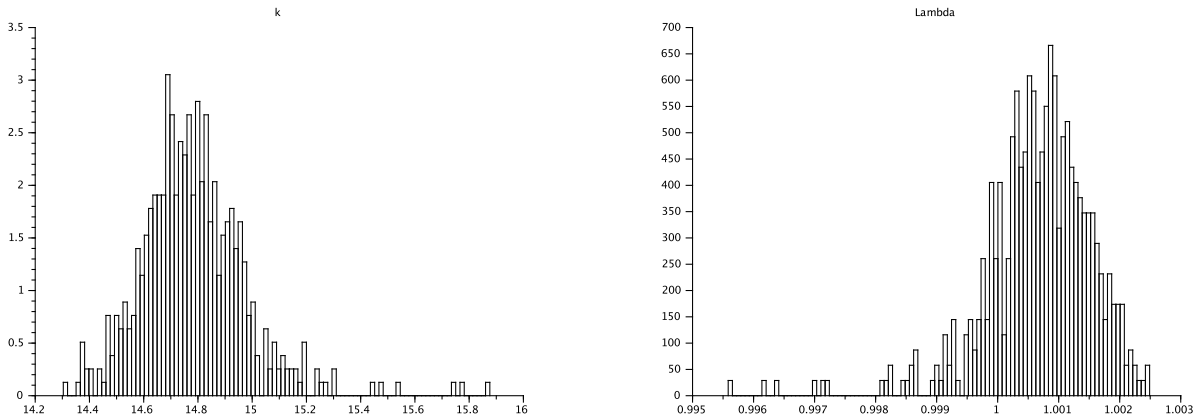

Figure 5.5 – Top: distribution of $k_{\text{opt}}(\omega)$. Bottom: distribution of $\lambda_{\text{opt}}(\omega)$.

On this test-case, we compute that $\mathbb{V}\text{ar}[\lambda_{\text{opt}}] \approx 7.9\ 10^{-7}$ and $\mathbb{V}\text{ar}[k_{\text{opt}}] \approx 3.8\ 10^{-2}$, thus

$$\mathbb{V}\text{arR}[\lambda_{\text{opt}}] \approx 7.9\ 10^{-7} \quad \text{and} \quad \mathbb{V}\text{arR}[k_{\text{opt}}] \approx 1.7\ 10^{-4}.$$

On the other hand, $A^\star(\theta_0) \approx 1.2$, $\mathbb{V}\text{ar}\left[A_N^\star(\theta_0)\right] \approx 2.0\ 10^{-6}$ and $\mathbb{V}\text{ar}\left[S_N(k_0)\right] \approx 4.5\ 10^{-15}$, thus

$$\mathbb{V}\text{arR}\left[K_N^\star(\theta_0)\right] \approx 1.4\ 10^{-6} \quad \text{and} \quad \mathbb{V}\text{arR}\left[S_N(k_0)\right] \approx 10^{-3}.$$

We thus observe that the relative variance of the optimal parameters is roughly of the same order of magnitude as the relative variance introduced in the function to minimize. Given the amount of noise present in the system, our procedure robustly identifies the optimal parameters of the microscopic distribution.

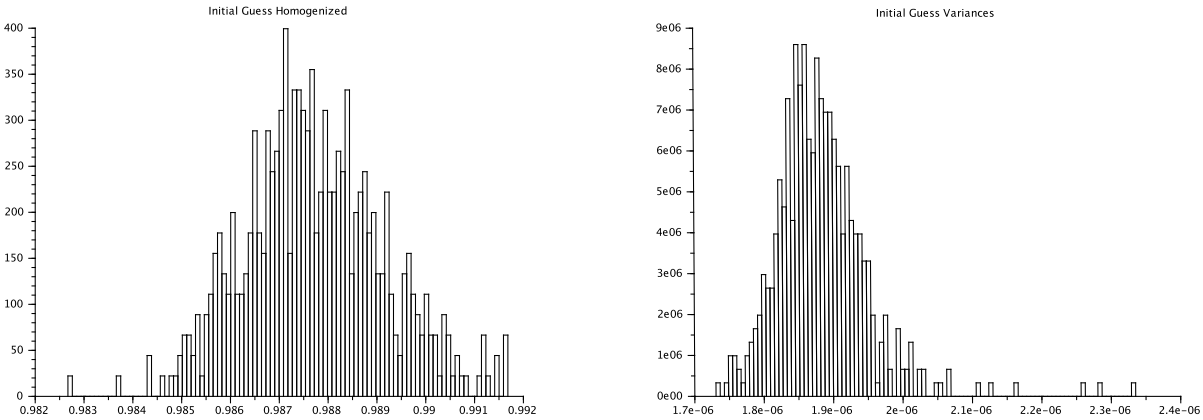

Figure 5.6 – Top: distribution of $K_N^\star(\theta_0, \omega)$. Bottom: distribution of $S_N(k_0, \omega)$.

1

5.5 Appendix: Computation of the derivatives of (5.50)

We introduce

$$f(\lambda,k) := \lambda^4 \left(\sum_{i=1}^N w_i^4(k) \right)^{-1}$$

and

$$g(k) := \left(\sum_{i=1}^N w_i^8(k) \right) \left(\sum_{i=1}^N w_i^4(k) \right)^{-2}$$

where $w_i(k)$ is defined by (5.49), and recast the function (5.50) as

$$\widetilde{F}_N^{\mathrm{1D}}(\theta) = \left(\frac{N}{K_N^{\star,\mathrm{obs}}}\, f(\lambda,k) - 1 \right)^2 + \left(\frac{1}{S_N^{\mathrm{obs}}} \left(g(k) - \frac{1}{N} \right) - 1 \right)^2,$$

where we have kept implicit the dependence with respect to ω. Computing the derivatives of $\widetilde{F}_N^{\mathrm{1D}}$ therefore amounts to computing those of f and g.

A tedious but straightforward computation leads to the following expressions:

$$\begin{aligned}
g'(k) &= \frac{8}{k} \left(\sum_i w_i^4 \right)^{-2} \left[\frac{\sum_i w_i^8}{\sum_i w_i^4} \sum_i \ln(w_i) w_i^4 - \sum_i \ln(w_i) w_i^8 \right], \\
g''(k) &= \frac{16}{k^2} \left(\sum_i w_i^4 \right)^{-2} \left[\sum_i \ln(w_i) w_i^8 (1 + 4 \ln w_i) - \frac{\sum_i w_i^8}{\sum_i w_i^4} \sum_i \ln(w_i) w_i^4 (1 + 2 \ln w_i) \right] \\
&\quad - \frac{32}{k^2} \left(\sum_i w_i^4 \right)^{-3} \left[4 \left(\sum_i w_i^8 \ln(w_i) \right) \left(\sum_i w_i^4 \ln(w_i) \right) - 3 \left(\sum_i w_i^4 \ln(w_i) \right)^2 \frac{\sum_i w_i^8}{\sum_i w_i^4} \right],
\end{aligned}$$

whereas

$$\begin{aligned}
\partial_\lambda f &= \frac{4}{\lambda} f(\lambda,k), \\
\partial_{\lambda\lambda}^2 f &= \frac{12}{\lambda^2} f(\lambda,k), \\
\partial_k f &= \frac{4\lambda^4}{k} \left(\sum_i \ln(w_i) w_i^4 \right) \left(\sum_i w_i^4 \right)^{-2}, \\
\partial_{kk}^2 f &= -\frac{8\lambda^4}{k^2} \left(\sum_i w_i^4 \right)^{-2} \left[\sum_i \ln(w_i) w_i^4 (1 + 2 \ln w_i) \right] + \frac{32\lambda^4}{k^2} \left(\sum_i w_i^4 \right)^{-3} \left(\sum_i \ln(w_i) w_i^4 \right)^2, \\
\partial_{\lambda k}^2 f &= \frac{16\lambda^3}{k} \left(\sum_i \ln(w_i) w_i^4 \right) \left(\sum_i w_i^4 \right)^{-2} = \frac{4}{\lambda} \partial_k f.
\end{aligned}$$

Appendix A

Neumann, Dirichlet and periodic approximations of the corrector problem

A.1 Introduction

The purpose of this appendix, useful for **Chapter** 4, is to present three classical approximations of the random corrector problem (1.8), using a problem set on a truncated domain complemented by periodic, Dirichlet or Neumann boundary conditions. These three approximations are $A^{\star,N}_{\mathrm{Per}}$ (see (A.4)–(A.7)), $A^{\star,N}_{\mathrm{Dir}}$ (see (A.2)–(A.3)) and $A^{\star,N}_{\mathrm{Neu}}$ (see (A.9)–(A.10)) of $A^\star$.

For the sake of completeness, we also provide a complete proof of the fact that

$$A^{\star,N}_{\mathrm{Neu}}(\omega) \leq A^{\star,N}_{\mathrm{Per}}(\omega) \leq A^{\star,N}_{\mathrm{Dir}}(\omega), \quad \text{almost surely.} \tag{A.1}$$

In section A.2, we introduce these three approximations $A^{\star,N}_{\mathrm{Dir}}$, $A^{\star,N}_{\mathrm{Neu}}$ and $A^{\star,N}_{\mathrm{Per}}$, and we recall the corresponding convergence results when N tends to infinity. Throughout this section, A need not be symmetric. In section A.3, we introduce a dual formulation in terms of flux. In the case when A is symmetric, it allows us to reformulate the corrector problem with Neumann and periodic boundary conditions as variational problems on the flux. In section A.4, we compare these variational problems to infer the ordering claimed in (A.1) in the case when the matrix A (and hence each of the approximations $A^{\star,N}_{\mathrm{Dir}}$, $A^{\star,N}_{\mathrm{Neu}}$ and $A^{\star,N}_{\mathrm{Per}}$) is symmetric.

A.2 Approximations of the corrector

In the celebrated article [BP04], A. Bourgeat and A. Piatnitski introduce three approximations of (1.8). They consider the corrector equation on the bounded domain $Q_N = (0, N)^d$, complemented with either homogeneous Dirichlet boundary conditions, periodic boundary conditions, or Neumann boundary conditions. We discuss the Dirichlet approximation in Section A.2.1, the periodic approximation in Section A.2.2, and the Neumann approximation in Section A.2.3. The two first approximations are carefully considered and the third one is rapidly processed, but it turns out the Neumann approximation is more subtle.

Throughout this appendix, we assume that A is uniformly elliptic, bounded and stationary, that is to say:

- *(boundedness)*: The field A is bounded almost everywhere and almost surely by some constant $\|A\|_{L^\infty(\Omega\times\mathbb{R}^d)}$.
- *(ellipticity)*: There exists some constant $\alpha > 0$ such that, for almost all $x \in \mathbb{R}^d$, all $p \in \mathbb{R}^d$ and almost surely in $\omega \in \Omega$,
$$\alpha|p|^2 \leq p \cdot A(x,\omega)p.$$
- *(stationnarity)* : A is stationary in the sense of (1.3).

A.2.1 The Dirichlet approximation

Consider the problem
$$\left\{ \begin{array}{c} -\operatorname{div}\left[A(\cdot,\omega)\left(p+\nabla w^{N,p}_{\mathrm{Dir}}(\cdot,\omega)\right)\right] = 0 \text{ in } Q_N, \\ w^{N,p}_{\mathrm{Dir}}(\cdot,\omega) = 0 \text{ on } \partial Q_N. \end{array} \right. \tag{A.2}$$

This problem is well-posed almost surely in $\omega \in \Omega$. The corresponding apparent homogenized matrix $A^{\star,N}_{\mathrm{Dir}}(\omega)$ is defined by
$$\forall p \in \mathbb{R}^d, \quad A^{\star,N}_{\mathrm{Dir}}(\omega)p = \frac{1}{|Q_N|}\int_{Q_N} A(x,\omega)(p+\nabla w^{N,p}_{\mathrm{Dir}}(x,\omega))dx. \tag{A.3}$$
It is shown in [BP04, Theorem 2] that
$$\lim_{N\to\infty} A^{\star,N}_{\mathrm{Dir}}(\omega) = A^\star \text{ a.s.}$$

A.2.2 The periodic approximation

Consider the problem
$$\left\{ \begin{array}{c} -\operatorname{div}\left[\tilde{A}^N_{\mathrm{Per}}(\cdot,\omega)\left(p+\nabla w^{N,p}_{\mathrm{Per}}(\cdot,\omega)\right)\right] = 0 \text{ in } \mathbb{R}^d, \\ w^{N,p}_{\mathrm{Per}}(\cdot,\omega) \text{ is } Q_N\text{-periodic}, \end{array} \right. \tag{A.4}$$
where $\tilde{A}^N_{\mathrm{Per}}$ is the Q_N periodic extension of A, and the first equality is to be understood in the periodic distributional sense (alternatively, one may consider Equation (A.4) on the torus $\mathbb{R}^d/Q_N$, the point being that, for instance, periodically repeated linear functions are not admissible as test functions nor solutions). This problem is well-posed up to the addition of a (possibly random) constant. The variational formulation of (A.4) is: find $w^{N,p}_{\mathrm{Per}} \in V_{\mathrm{per}}(Q_N)$ such that
$$\forall \psi \in V_{\mathrm{per}}(Q_N), \quad \int_{Q_N} \nabla\psi \cdot A(x,\omega)\left(p+\nabla w^{N,p}_{\mathrm{Per}}(x,\omega)\right)dx = 0, \tag{A.5}$$
where $\cdot$ denotes the usual $\mathbb{R}^d$ scalar product and where
$$V_{\mathrm{per}}(Q_N) = \left\{ v \in H^1_{\mathrm{loc}}(\mathbb{R}^d),\ v \text{ is } Q_N\text{-periodic},\ \int_{Q_N} v = 0 \right\}. \tag{A.6}$$
The corresponding apparent homogenized matrix $A^{\star,N}_{\mathrm{Per}}(\omega)$ is defined by
$$\forall p \in \mathbb{R}^d, \quad A^{\star,N}_{\mathrm{Per}}(\omega)p = \frac{1}{|Q_N|}\int_{Q_N} A(x,\omega)(p+\nabla w^{N,p}_{\mathrm{Per}}(x,\omega))dx. \tag{A.7}$$
It is shown in [BP04, Theorem 1] that
$$\lim_{N\to\infty} A^{\star,N}_{\mathrm{Per}}(\omega) = A^\star \text{ almost surely.} \tag{A.8}$$

A.2.3 Neumann approximation

Theorem A.1. *Consider the Neumann corrector, that is*

$$\begin{cases} -\operatorname{div} A(\cdot,\omega)(p+\nabla w^{N,p}_{\mathrm{Neu}}(\cdot,\omega)) = 0 \text{ in } Q_N, \\ A(\cdot,\omega)(p+\nabla w^{N,p}_{\mathrm{Neu}}(\cdot,\omega))\cdot n = p\cdot n \text{ on } \partial Q_N, \end{cases} \tag{A.9}$$

the solution of which is unique in $H^1(Q_N)$ up to the addition of a random constant. We define the random variable $S^{\star,N}_{\mathrm{Neu}}(\omega)$ and the approximation $A^{\star,N}_{\mathrm{Neu}}(\omega)$ of $A^\star$ by

$$S^{\star,N}_{\mathrm{Neu}}(\omega)p := \frac{1}{|Q_N|}\int_{Q_N}\left(p+\nabla w^{N,p}_{\mathrm{Neu}}(x,\omega)\right)dx, \quad A^{\star,N}_{\mathrm{Neu}}(\omega) := S^{\star,N}_{\mathrm{Neu}}(\omega)^{-1}. \tag{A.10}$$

Then,

$$\lim_{N\to\infty} A^{\star,N}_{\mathrm{Neu}}(\omega) = A^\star \text{ almost surely.} \tag{A.11}$$

Proof. There exists a solution to the problem (A.9) since the compatibility condition $\int_{\partial Q_N} p\cdot n = 0$ is satisfied. In the sequel, we impose the condition

$$\int_{Q_N}\left(p\cdot x + w^{N,p}_{\mathrm{Neu}}(x,\omega)\right)dx = 0 \tag{A.12}$$

to enforce uniqueness. The equation (A.9) is equivalent to the variational formulation:

$$\forall\varphi\in H^1(Q_N),\quad \int_{Q_N}\nabla\varphi\cdot A(x,\omega)(p+\nabla w^{N,p}_{\mathrm{Neu}}(x,\omega))dx = \int_{\partial Q_N}\varphi\, p\cdot n.$$

To prove the convergence of $S^{\star,N}_{\mathrm{Neu}}$, we use the same type of arguments as in [BP04] (where there is a typo concerning which quantity converges towards $A^\star$).

We first rescale the equation to the domain Q. Let $\tilde{w}^{N,p}_{\mathrm{Neu}}(\cdot,\omega) := \frac{1}{N}w^{N,p}_{\mathrm{Neu}}(N\cdot,\omega)$ be the rescaled corrector. It satisfies the equation

$$\begin{cases} -\operatorname{div} A_N(\cdot,\omega)(p+\nabla \tilde{w}^{N,p}_{\mathrm{Neu}}(\cdot,\omega)) = 0 \text{ in } Q, \\ A_N(\cdot,\omega)(p+\nabla \tilde{w}^{N,p}_{\mathrm{Neu}}(\cdot,\omega))\cdot n = p\cdot n \text{ on } \partial Q, \end{cases} \tag{A.13}$$

where $A_N(\cdot,\omega) := A(N\cdot,\omega)$. The variational formulation of (A.13) is given by

$$\forall\varphi\in H^1(Q),\quad \int_Q \nabla\varphi(x)\cdot A_N(x,\omega)(p+\nabla\tilde{w}^{N,p}_{\mathrm{Neu}}(x,\omega))dx = \int_{\partial Q}\varphi\, p\cdot n. \tag{A.14}$$

Taking $\varphi(x,\omega) = p\cdot x + \tilde{w}^{N,p}_{\mathrm{Neu}}(x,\omega)$, we obtain a $H^1(Q)$ estimate on $\tilde{w}^{N,p}_{\mathrm{Neu}}$:

$$\begin{aligned}\|p+\nabla\tilde{w}^{N,p}_{\mathrm{Neu}}\|^2_{L^2(Q)} &\le \frac{1}{a_-}\int_Q (p+\nabla\tilde{w}^{N,p}_{\mathrm{Neu}})\cdot A_N(p+\nabla\tilde{w}^{N,p}_{\mathrm{Neu}}) \\ &= \frac{1}{a_-}\int_{\partial Q}(p\cdot x+\tilde{w}^{N,p}_{\mathrm{Neu}})p\cdot n \le \frac{C_1C_2}{a_-}\|p\cdot x+\tilde{w}^{N,p}_{\mathrm{Neu}}\|_{H^1(Q)},\end{aligned}$$

where a_- is the coercivity constant of A, C_1 is continuity constant of the trace operator from $H^1(Q)$ into $H^{1/2}(\partial Q)$ and $C_2 = \|p\cdot n\|_{L^2(\partial Q)}$. We next combine this estimate with the Poincaré-Wirtinger inequality (recall (A.12)), to obtain that the function $x\mapsto p\cdot x+\tilde{w}^{N,p}_{\mathrm{Neu}}(x,\omega)$ is uniformly bounded in $H^1(Q)$.

Therefore, up to extraction, the function $x \mapsto p \cdot x + \tilde{w}^{N,p}_{\text{Neu}}(x,\omega)$ converges weakly in $H^1(Q)$ and almost surely to some $v^\star(\cdot,\omega) \in H^1(Q)$. In view of [JKO94, p.14, Convergence of arbitrary solutions], $A_N(\cdot,\omega)(p + \nabla \tilde{w}^{N,p}_{\text{Neu}}(\cdot,\omega))$ converges weakly in $L^2(Q)$ and almost surely towards $A^\star \nabla v^\star(\cdot,\omega)$.

We are now in position to pass to the limit in (A.14). Let $\varphi \in H^1(Q)$,

$$0 = \int_Q \nabla\varphi \cdot A_N(\cdot,\omega)(p + \nabla \tilde{w}^{N,p}_{\text{Neu}}(\cdot,\omega)) - \int_{\partial Q} \varphi\, p \cdot n \xrightarrow[N\to\infty]{} \int_Q \nabla\varphi \cdot A^\star \nabla v^\star(\cdot,\omega) - \int_{\partial Q} \varphi\, p \cdot n.$$

Therefore, $v^\star$ satisfies the variational formulation

$$\forall \varphi \in H^1(Q),\ \int_Q \nabla\varphi \cdot A^\star \nabla v^\star(\cdot,\omega) = \int_{\partial Q} \varphi\, p \cdot n, \tag{A.15}$$

complemented with the constraint $\displaystyle\int_Q v^\star(\cdot,\omega) = 0$ (this is a trivial consequence of (A.12) and weak $L^2(Q)$ limit). This equation, whose strong formulation is

$$\begin{cases} -\operatorname{div} A^\star \nabla v^\star = 0 \text{ in } Q, \\ A^\star \nabla v^\star \cdot n = p \cdot n \text{ on } \partial Q, \end{cases} \tag{A.16}$$

has a unique solution $v^\star(x) = x \cdot (A^\star)^{-1} p - \displaystyle\int_Q x \cdot (A^\star)^{-1} p$ (this solution is deterministic).

We now use the weak limit in $H^1(Q)$. The convergence

$$S^{\star,N}_{\text{Neu}}(\omega) = \frac{1}{|Q_N|} \int_{Q_N} p + \nabla \tilde{w}^{N,p}_{\text{Neu}}(\cdot,\omega) = \int_Q p + \nabla \tilde{w}^{N,p}_{\text{Neu}}(\cdot,\omega) \xrightarrow[N\to\infty]{} \int_Q \nabla v^\star = (A^\star)^{-1} p,$$

holds almost surely. This concludes the proof.

□

A.3 Flux formulation

Assume now that A is symmetric. Then it turns out that the Neumann corrector problem (A.9) is related to a problem where the unknown is $\sigma = A\nabla u$. This reformulation is useful in order to establish (A.1). We first introduce the appropriate functional spaces.

A.3.1 The H^{div} space

For any open set $\mathcal{D}$, we consider the space

$$H^{\text{div}}(\mathcal{D}) = \left\{ \sigma \in L^2(\mathcal{D})^d, \quad \operatorname{div} \sigma \in L^2(\mathcal{D}) \right\},$$

which is an Hilbert space for the scalar product

$$(\sigma, t)_{H^{\text{div}}(\mathcal{D})} = (\sigma, t)_{L^2(\mathcal{D})^d} + (\operatorname{div} \sigma, \operatorname{div} t)_{L^2(\mathcal{D})}.$$

Assume that the boundary of $\mathcal{D}$ is sufficiently smooth (say e.g. Lipschitz smooth) so that the Green formula holds for $C^1(\overline{\mathcal{D}})$ functions:

$$\forall \sigma \in C^1(\overline{\mathcal{D}})^d,\ \forall \varphi \in C^1(\overline{\mathcal{D}}), \quad \int_{\partial\mathcal{D}} \varphi(\sigma \cdot n) = \int_{\mathcal{D}} \nabla\varphi \cdot \sigma + \int_{\mathcal{D}} \varphi \operatorname{div} \sigma.$$

it is well known that, for any $\sigma \in H^{\mathrm{div}}(\mathcal{D})$, we can define the trace of $\sigma \cdot n$ on $\partial\mathcal{D}$ by extension of the above formula (although σ itself does not have a well-defined trace on $\partial\mathcal{D}$). For any $\varphi \in H^1(\mathcal{D})$, we set

$$\langle \sigma \cdot n, \varphi \rangle = \int_{\mathcal{D}} \nabla\varphi \cdot \sigma + \int_{\mathcal{D}} \varphi \operatorname{div} \sigma. \tag{A.17}$$

A.3.2 The periodic $H^{\mathrm{div}}_{\mathrm{per}}$ space

We recall that $Q = (0,1)^d$ and $Q_N = (0,N)^d$. We now define the well-known space of periodic H^{div} functions as

$$H^{\mathrm{div}}_{\mathrm{per}}(Q_N) = \left\{ \sigma \in L^2_{\mathrm{per}}(Q_N)^d, \quad \operatorname{div}\sigma \in L^2_{\mathrm{loc}}(\mathbb{R}^d) \right\},$$

where we recall that

$$L^2_{\mathrm{per}}(Q_N) = \left\{ v \in L^2_{\mathrm{loc}}(\mathbb{R}^d), \quad v \text{ is } Q_N\text{-periodic} \right\}.$$

In $H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$, we request that $\operatorname{div}\sigma$ to be in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$, and not only in $L^2(Q_N)$ (see Remark A.2 below). The space $H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$, endowed with its natural scalar product $(\cdot,\cdot)_{H^{\mathrm{div}}(Q_N)}$ is a Hilbert space.

Remark A.2. *The space $H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$ is a distinct space from*

$$\left\{ \sigma \in L^2_{\mathrm{per}}(Q_N)^d, \quad \operatorname{div}\sigma \in L^2(Q_N) \right\}.$$

Indeed, in the above space, the normal trace $\sigma \cdot n$ may jump across ∂Q_N. In contrast, in $H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$, we have that $\operatorname{div}\sigma \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$, hence the normal trace is well-defined on ∂Q_N.

A.3.3 Flux formulation of the periodic problem (A.4)

For any $\sigma \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$ and $p \in \mathbb{R}^d$, we consider the energy functional

$$\mathcal{E}_p(\sigma,\omega) = \frac{1}{|Q_N|} \int_{Q_N} (\sigma + p) \cdot A^{-1}(\cdot,\omega)(\sigma + p). \tag{A.18}$$

We also consider the variational problem

$$\mathcal{S}^{\star,N}_{\mathrm{Per}}(p,\omega) = \inf \left\{ \mathcal{E}_p(\sigma,\omega), \quad \sigma \in H^{\mathrm{div}}_{\mathrm{per}}(Q_N), \ \int_{Q_N} \sigma = 0, \ \operatorname{div}\sigma = 0 \text{ in } \mathbb{R}^d \right\}. \tag{A.19}$$

Let

$$\mathcal{W} = \left\{ \sigma \in H^{\mathrm{div}}_{\mathrm{per}}(Q_N), \ \int_{Q_N} \sigma = 0, \ \operatorname{div}\sigma = 0 \text{ in } \mathbb{R}^d \right\}.$$

The following result is proved in [JKO94, Equation (1.64)]. We provide its proof for consistency.

Lemma A.3. *The variational problem (A.19) is well posed and there exists a matrix $S^{\star,N}_{\mathrm{Per}}(\omega)$ such that $\mathcal{S}^{\star}_{\mathrm{Per}}(p,\omega) = p \cdot S^{\star,N}_{\mathrm{Per}}(\omega)p$. Additionally, assume that A is symmetric. Then the Euler-Lagrange equations of problem (A.19) read*

$$\left\{ \begin{array}{c} \textit{Find } \sigma \in \mathcal{W} \textit{ such that} \\ \textit{for any } h \in \mathcal{W}, \textit{ we have } \displaystyle\int_{Q_N} h \cdot A^{-1}(\cdot,\omega)(p+\sigma) = 0. \end{array} \right. \tag{A.20}$$

Finally, it holds that $S^{\star,N}_{\mathrm{Per}}(\omega) = A^{\star,N}_{\mathrm{Per}}(\omega)^{-1}$.

Proof. The space $\mathcal{W}$ is a Hilbert space for the same norm as $H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$. On $\mathcal{W}$, the norm is simply the $L^2(Q_N)$ norm. Problem (A.19) consists in minimizing a strongly convex quadratic functional over $\mathcal{W}$. The functional is coercive according to the $L^2(Q_N)$ norm. We can thus apply the Lax-Milgram theorem, which implies that the problem is well-posed. From the linearity of the Euler-Lagrange equation, we deduce that $\mathcal{S}^{\star,N}_{\mathrm{Per}}(p,\omega) = p\cdot S^{\star,N}_{\mathrm{Per}}(\omega)p$ for some matrix $S^{\star,N}_{\mathrm{Per}}(\omega)$.

It remains to show that $S^{\star,N}_{\mathrm{Per}}(\omega) = A^{\star,N}_{\mathrm{Per}}(\omega)^{-1}$. Let $p \in \mathbb{R}^d$ and set $\xi(\omega) = A^{\star,N}_{\mathrm{Per}}(\omega)^{-1}p$. We denote by $w^{N,\xi}_{\mathrm{Per}}$ the periodic corrector function (defined up to the addition of a random constant) in the direction ξ, namely the solution to (A.4), that is

$$-\operatorname{div}\left[\tilde{A}^N_{\mathrm{Per}}\left(\nabla w^{N,\xi}_{\mathrm{Per}} + \xi\right)\right] = 0 \text{ in } \mathbb{R}^d, \quad w^{N,\xi}_{\mathrm{Per}}(\cdot,\omega) \text{ is } Q_N\text{-periodic},$$

where we recall that $\tilde{A}^N_{\mathrm{Per}}$ is the Q_N periodic extension of A.

We set $\sigma_\xi = \tilde{A}^N_{\mathrm{Per}}\left(\xi + \nabla w^{N,\xi}_{\mathrm{Per}}\right) - p$ and claim that $\sigma_\xi(\cdot,\omega) \in \mathcal{W}$. We indeed have that $\operatorname{div}\sigma_\xi = 0$ in $\mathbb{R}^d$, while

$$\frac{1}{|Q_N|}\int_{Q_N} \sigma_\xi(x,\omega)dx = A^{\star,N}_{\mathrm{Per}}(\omega)\xi(\omega) - p = 0.$$

Furthermore, $\sigma_\xi(\cdot,\omega) \in L^2_{\mathrm{loc}}(\mathbb{R}^d)$ and is Q_N periodic, while $\operatorname{div}\sigma_\xi$ is in $L^2_{\mathrm{loc}}(\mathbb{R}^d)$. We hence indeed have that $\sigma_\xi(\cdot,\omega) \in \mathcal{W}$.

We next show that σ_ξ satisfies the Euler-Lagrange equation (A.20). Let $h \in \mathcal{W}$. Then,

$$\int_{Q_N} h\cdot A^{-1}(p+\sigma_\xi) = \int_{Q_N} h\cdot\left(\xi + \nabla w^{N,\xi}_{\mathrm{Per}}\right) = \int_{Q_N} h\cdot\nabla w^{N,\xi}_{\mathrm{Per}},$$

using that the mean of h vanishes. We next integrate by part and use that $\operatorname{div} h = 0$ on $\mathbb{R}^d$, yielding

$$\int_{Q_N} h\cdot A^{-1}(p+\sigma_\xi) = \int_{\partial Q_N} (h\cdot n)\, w^{N,\xi}_{\mathrm{Per}}.$$

We now observe that $w^{N,\xi}_{\mathrm{Per}}$ has a well-defined trace on ∂Q_N (since $w^{N,\xi}_{\mathrm{Per}} \in H^1_{\mathrm{loc}}(\mathbb{R}^d)$). Likewise, $h\cdot n$ has also a well-defined trace since $h \in H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$. Since h and $w^{N,\xi}_{\mathrm{Per}}$ are periodic, the above right-hand side vanishes. Thus, σ_ξ satisfies the Euler-Lagrange equation (A.20), and is hence the unique minimizer of (A.19). As a consequence,

$$\begin{aligned} p\cdot S^{\star,N}_{\mathrm{Per}}p &= \frac{1}{|Q_N|}\int_{Q_N}(\sigma_\xi + p)\cdot A^{-1}(\sigma_\xi + p) \\ &= \frac{1}{|Q_N|}\int_{Q_N}(\xi + \nabla w^{N,\xi}_{\mathrm{Per}})\cdot A(\xi + \nabla w^{N,\xi}_{\mathrm{Per}}) = \xi\cdot A^{\star,N}_{\mathrm{Per}}\xi = p\cdot(A^{\star,N}_{\mathrm{Per}})^{-1}p. \end{aligned}$$

Since p is arbitrary, this concludes the proof. □

A.3.4 Flux formulation of the Neumann problem (A.9)

For any vector $p \in \mathbb{R}^d$, we define

$$\mathcal{S}^{\star,N}_{\mathrm{Neu}}(p,\omega) = \inf\left\{\mathcal{E}_p(\sigma,\omega),\ \sigma \in H^{\mathrm{div}}(Q_N),\ \sigma\cdot n = 0 \text{ on } \partial Q_N,\ \operatorname{div}\sigma = 0 \text{ in } Q_N\right\}, \tag{A.21}$$

where $H^{\mathrm{div}}(Q_N)$ is defined in Section A.3.1 and $\mathcal{E}_p(\sigma,\omega)$ is defined by (A.18).

Lemma A.4. *Problem (A.21) is well-posed.*

Additionally, assume that A is symmetric. Then, the solution σ_N of this problem is given by $\sigma_N = A(p+\nabla w_{\mathrm{Neu}}^{N,p}) - p$, where $w_{\mathrm{Neu}}^{N,p}$ is the solution to (A.9). In addition, it holds that $\mathcal{S}_{\mathrm{Neu}}^{\star,N}(p,\omega) = p \cdot S_{\mathrm{Neu}}^{\star,N}(\omega)p$, where $S_{\mathrm{Neu}}^{\star,N}(\omega)$ is defined by (A.10).

Proof. The space

$$\mathcal{V} = \left\{\sigma \in H^{\mathrm{div}}(Q_N),\ \sigma \cdot n = 0 \text{ on } \partial Q_N,\ \operatorname{div}\sigma = 0 \text{ in } Q_N\right\}$$

is closed subspace of $H^{\mathrm{div}}(Q_N)$ and therefore a Hilbert space. In $\mathcal{V}$, the H^{div} norm reduces to the L^2 norm. The functional $\mathcal{E}_p(\cdot,\omega)$ is hence coercive on $\mathcal{V}$. It is also continuous and strongly convex, therefore the problem (A.21) is well posed. Let $\sigma_N^p(\cdot,\omega) \in \mathcal{V}$ be its unique solution.
In the symmetric case, the Euler-Lagrange equation of (A.21) reads

$$\forall \psi \in \mathcal{V},\ \int_{Q_N} \psi \cdot A^{-1}(\sigma_N^p + p) = 0. \tag{A.22}$$

We next show that $\sigma_N^p = A(p + \nabla w_{\mathrm{Neu}}^{N,p}) - p$. Let $\psi \in \mathcal{V}$. We see that

$$\int_{Q_N} \psi \cdot A^{-1}(A(p + \nabla w_{\mathrm{Neu}}^{N,p}) - p + p) = \int_{Q_N} \psi \cdot \nabla(p \cdot x + w_{\mathrm{Neu}}^{N,p}) = 0,$$

where the latter equality is obtained by integration by parts. Thus, $A(p + \nabla w_{\mathrm{Neu}}^{N,p}) - p$ satisfies the Euler-Lagrange equations associated to problem (A.21). Therefore it is the solution. In addition,

$$\begin{aligned}
\mathcal{S}_{\mathrm{Neu}}^{\star,N}(p,\omega) &= \frac{1}{|Q_N|}\int_{Q_N} \left((\sigma_N^p + p) \cdot A^{-1}A(p + \nabla w_{\mathrm{Neu}}^{N,p})\right)(x,\omega)dx \\
&= \frac{1}{|Q_N|}\int_{Q_N} p \cdot (p + \nabla w_{\mathrm{Neu}}^{N,p}(\cdot,\omega)),
\end{aligned}$$

because $\displaystyle\int_{Q_N} \sigma_N^p \cdot (p + \nabla w_{\mathrm{Neu}}^{N,p}) = 0$ by integration by parts. By definition of $S_{\mathrm{Neu}}^{\star,N}(\omega)$ (see (A.22)) we have $\mathcal{S}_{\mathrm{Neu}}^{\star,N}(p,\omega) = p \cdot S^{\star,N}(\omega)p$. This concludes the proof. □

A.4 Comparison of the approximations

We now establish (A.1).

A.4.1 Comparison of the Dirichlet and periodic approximations

We have the following result:

Theorem A.5. *We assume that A is a symmetric matrix. Then $A_{\mathrm{Dir}}^{\star,N}(\omega)$ and $A_{\mathrm{Per}}^{\star,N}(\omega)$ are symmetric matrices and, in the sense of symmetric matrices, we have, for any N,*

$$A_{\mathrm{Per}}^{\star,N}(\omega) \leq A_{\mathrm{Dir}}^{\star,N}(\omega) \text{ almost surely.} \tag{A.23}$$

Proof. The symmetry of $A^{\star,N}_{\mathrm{Dir}}(\omega)$ and $A^{\star,N}_{\mathrm{Per}}(\omega)$ is a classical result. We now prove (A.23). Let $p \in \mathbb{R}^d$. Using the symmetry of A, we see that

$$p \cdot A^{\star,N}_{\mathrm{Dir}}(\omega)p = \inf\left\{\frac{1}{|Q_N|}\int_{Q_N}(p+\nabla v)\cdot A(\cdot,\omega)(p+\nabla v), \quad v \in H^1_0(Q_N)\right\}$$

while

$$p \cdot A^{\star,N}_{\mathrm{Per}}(\omega)p = \inf\left\{\frac{1}{|Q_N|}\int_{Q_N}(p+\nabla v)\cdot A(\cdot,\omega)(p+\nabla v), \quad v \in V_{\mathrm{per}}(Q_N)\right\}, \tag{A.24}$$

where $V_{\mathrm{per}}(Q_N)$ is defined by (A.6). For any $v \in H^1(Q_N)$, let

$$J(v,\omega) = \frac{1}{|Q_N|}\int_{Q_N}(p+\nabla v)\cdot A(\cdot,\omega)(p+\nabla v).$$

Let $v \in H^1_0(Q_N)$. Consider the Q_N-periodic extension $\widetilde{v}$ of v, and set

$$\overline{v} = \widetilde{v} - \frac{1}{|Q_N|}\int_{Q_N}\widetilde{v}.$$

By construction, we see that $\overline{v} \in H^1_{\mathrm{loc}}(\mathbb{R}^d)$ and that $\overline{v}$ is Q_N periodic. In addition, the mean of $\overline{v}$ on Q_N vanishes. Hence $\overline{v} \in V_{\mathrm{per}}(Q_N)$ and $J(\overline{v}) = J(v)$. Thus

$$p \cdot A^{\star,N}_{\mathrm{Per}}(\omega)p \le J(\overline{v},\omega) = J(v,\omega).$$

Taking the infimum over $v \in H^1_0(Q_N)$, we deduce that

$$p \cdot A^{\star,N}_{\mathrm{Per}}(\omega)p \le p \cdot A^{\star,N}_{\mathrm{Dir}}(\omega)p$$

which is valid for any p. This concludes the proof. □

A.4.2 Comparison of the Neumann and periodic approximations

We have the following result:

Theorem A.6. *We assume that A is a symmetric matrix. Then $A^{\star,N}_{\mathrm{Neu}}(\omega)$ is a symmetric matrix and, in the sense of symmetric matrices, we have, for any N,*

$$A^{\star,N}_{\mathrm{Neu}}(\omega) \le A^{\star,N}_{\mathrm{Per}}(\omega) \text{ almost surely.} \tag{A.25}$$

Proof. We deduce the properties on $A^{\star,N}$ from the properties of $S^{\star,N}$ that we now establish.

The symmetry of $S^{\star,N}_{\mathrm{Neu}}(\omega)$ relies on the equality $\sigma^p_N = A(p+\nabla w^{N,p}_{\mathrm{Neu}})-p$ from Lemma A.4. Indeed, for all $\xi, p \in \mathbb{R}^d$,

$$\xi \cdot S^{\star,N}_{\mathrm{Neu}}p = \frac{1}{|Q_N|}\int_{Q_N}\xi\cdot(p+\nabla w^{N,p}_{\mathrm{Neu}}) = \frac{1}{|Q_N|}\int_{Q_N}\xi\cdot A^{-1}(p+\sigma^p_N).$$

In view of the Euler-Lagrange equations of (A.21), namely (A.22), we have $\int_{Q_N}\sigma^\xi_N\cdot A^{-1}(p+\sigma^p_N) = 0$. Therefore,

$$\xi \cdot S^{\star,N}_{\mathrm{Neu}}p = \frac{1}{N^d}\int_{Q_N}(\xi+\sigma^\xi_N)\cdot A^{-1}(p+\sigma^p_N),$$

and thus $S^{\star,N}_{\mathrm{Neu}}(\omega)$ is symmetric.

We now prove the bound $S^{\star,N}_{\mathrm{Per}}(\omega) \leq S^{\star,N}_{\mathrm{Neu}}(\omega)$ almost surely. Let $p \in \mathbb{R}^d$. We have that

$$p \cdot S^{\star,N}_{\mathrm{Neu}}(\omega)p = \inf \{\mathcal{E}_p(\sigma,\omega), \quad \sigma \in \mathcal{V}\}$$

while

$$p \cdot S^{\star,N}_{\mathrm{Per}}(\omega)p = \inf \{\mathcal{E}_p(\sigma,\omega), \quad \sigma \in \mathcal{W}\}.$$

Let $\sigma \in \mathcal{V}$. Consider the Q_N-periodic extension $\overline{\sigma}$ of σ. We already have that $\overline{\sigma} \in L^2_{\mathrm{per}}(Q_N)$. Let us show that $\operatorname{div}\overline{\sigma} = 0$ in $\mathcal{D}'(\mathbb{R}^d)$. Let $\varphi \in \mathcal{D}(\mathbb{R}^d)$. Then

$$\langle \operatorname{div}\overline{\sigma}, \varphi\rangle = -\langle \overline{\sigma}, \nabla\varphi\rangle = \sum_{k\in\mathbb{Z}^d} \int_{Q_N+Nk} \overline{\sigma}\cdot\nabla\varphi = \sum_{k\in\mathbb{Z}^d} \int_{Q_N} \overline{\sigma}\cdot\nabla\varphi(\cdot + Nk)$$

where the sum in k is actually finite. Let $\psi = \sum_{k\in\mathbb{Z}^d} \varphi(\cdot + Nk)$. We thus have

$$\langle \operatorname{div}\overline{\sigma}, \varphi\rangle = \int_{Q_N} \sigma\cdot\nabla\psi = -\int_{Q_N} \psi \operatorname{div}\sigma + \langle \sigma\cdot n, \psi\rangle$$

and both terms vanish because $\operatorname{div}\sigma = 0$ on Q_N and $\sigma\cdot n = 0$ on ∂Q_N for any $\sigma \in \mathcal{V}$.

We hence have that $\overline{\sigma} \in H^{\mathrm{div}}_{\mathrm{per}}(Q_N)$ and $\operatorname{div}\overline{\sigma} = 0$ in $\mathcal{D}'(\mathbb{R}^d)$. We eventually compute that

$$e_i \cdot \int_{Q_N} \overline{\sigma} = e_i \cdot \int_{Q_N} \sigma = \int_{\partial Q_N} x_i \sigma\cdot n - \int_{Q_N} x_i \operatorname{div}\sigma = 0,$$

using again that $\operatorname{div}\sigma = 0$ on Q_N and $\sigma\cdot n = 0$ on ∂Q_N. Hence $\overline{\sigma} \in \mathcal{W}$, and $\mathcal{E}_p(\overline{\sigma},\omega) = \mathcal{E}_p(\sigma,\omega)$. Thus

$$p \cdot S^{\star,N}_{\mathrm{Per}}(\omega)p \leq \mathcal{E}_p(\overline{\sigma},\omega) = \mathcal{E}_p(\sigma,\omega).$$

Taking the infimum over $\sigma \in \mathcal{V}$, we deduce that

$$p \cdot S^{\star,N}_{\mathrm{Per}}(\omega)p \leq p \cdot S^{\star,N}_{\mathrm{Neu}}(\omega)p$$

which is valid for any p. This concludes the proof of (A.24). □

Bibliography

[ACLB+12] A. Anantharaman, R. Costaouec, C. Le Bris, F. Legoll, and F. Thomines, *Introduction to numerical stochastic homogenization and the related computational challenges: some recent developments*, Multiscale modeling and analysis for materials simulation, Lect. Notes Ser. Inst. Math. Sci. Natl. Univ. Singap., vol. 22, World Sci. Publ., Hackensack, NJ, 2012, pp. 197–272. MR 2895600

[ACS14] S. N. Armstrong, P. Cardaliaguet, and P. E. Souganidis, *Error estimates and convergence rates for the stochastic homogenization of Hamilton-Jacobi equations*, J. Amer. Math. Soc. **27** (2014), no. 2, 479–540. MR 3164987

[ALB10] A. Anantharaman and C. Le Bris, *Homogénéisation d'un matériau périodique faiblement perturbé aléatoirement*, C. R. Math. Acad. Sci. Paris **348** (2010), no. 9-10, 529–534. MR 2645167 (2011g:35023)

[ALB11] ———, *A numerical approach related to defect-type theories for some weakly random problems in homogenization*, Multiscale Modeling & Simulation **9** (2011), no. 2, 513–544.

[ALB12] ———, *Elements of mathematical foundations for numerical approaches for weakly random homogenization problems*, Commun. Comput. Phys. **11** (2012), no. 4, 1103–1143. MR 2864078 (2012k:35028)

[AV12] A. Abdulle and G. Vilmart, *A priori error estimates for finite element methods with numerical quadrature for nonmonotone nonlinear elliptic problems*, Numer. Math. **121** (2012), no. 3, 397–431. MR 2929073

[Bal76] John M Ball, *Convexity conditions and existence theorems in nonlinear elasticity*, Archive for rational mechanics and Analysis **63** (1976), no. 4, 337–403.

[BCLBL12a] X. Blanc, R. Costaouec, C. Le Bris, and F. Legoll, *Variance reduction in stochastic homogenization: the technique of antithetic variables*, Numerical Analysis of Multiscale Computations, Springer, 2012, pp. 47–70.

[BCLBL12b] ———, *Variance reduction in stochastic homogenization using antithetic variables*, Markov Processes and Related Fields **18** (2012), no. 1, 31–66.

[Bis11] M. Biskup, *Recent progress on the random conductance model*, Probab. Surv. **8** (2011), 294–373. MR 2861133

[BL] M. Bornert and F. Legoll, *in prep.*

[BL93] J. W. Barrett and W. B. Liu, *Finite element approximation of the p-Laplacian*, Math. Comp. **61** (1993), no. 204, 523–537. MR 1192966 (94c:65129)

[BLBL06] X. Blanc, C. Le Bris, and P.-L. Lions, *Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques*, C. R. Math. Acad. Sci. Paris **343** (2006), no. 11-12, 717–724. MR 2284699 (2009c:60195)

[BLBL07] ———, *Stochastic homogenization and random lattices*, Journal de mathématiques pures et appliquées **88** (2007), no. 1, 34–63.

[BLP78] A. Bensoussan, J.-L. Lions, and G. Papanicolaou, *Asymptotic methods in periodic structures*, Studies in Math. Appl **5** (1978).

[BO14] C. Bernardin and S. Olla, *Thermodynamics and non-equilibrium macroscopic dynamics of chains of anharmonic oscillators*, Lecture Notes available at https://www.ceremade.dauphine.fr/~olla (2014).

[BP04] A. Bourgeat and A. Piatnitski, *Approximations of effective coefficients in stochastic homogenization*, Annales de l'Institut Henri Poincare (B) Probability and Statistics, vol. 40, Elsevier, 2004, pp. 153–165.

[CD99] D. Cioranescu and P. Donato, *An introduction to homogenization*, Oxford Lecture Series in Mathematics and its Applications, vol. 17, The Clarendon Press, Oxford University Press, New York, 1999. MR 1765047 (2001j:35019)

[Cho89] S.-S. Chow, *Finite element error estimates for nonlinear elliptic equations of monotone type*, Numer. Math. **54** (1989), no. 4, 373–393. MR 972416 (90a:65235)

[CLBL10] R. Costaouec, C. Le Bris, and F. Legoll, *Variance reduction in stochastic homogenization: proof of concept, using antithetic variables*, SeMA Journal **50** (2010), no. 1, 9–26.

[Cos12] R. Costaouec, *Asymptotic expansion of the homogenized matrix in two weakly stochastic homogenization settings*, Applied Mathematics Research eXpress **2012** (2012), no. 1, 76–104.

[DMM86a] G. Dal Maso and L. Modica, *Nonlinear stochastic homogenization*, Ann. Mat. Pura Appl. (4) **144** (1986), 347–389. MR 870884 (88h:49025)

[DMM86b] ———, *Nonlinear stochastic homogenization and ergodic theory*, J. Reine Angew. Math. **368** (1986), 28–42. MR 850613 (88k:28021)

[ES08] B. Engquist and P. E. Souganidis, *Asymptotic and numerical homogenization*, Acta Numerica **17** (2008), 147–190.

[Fat56] I. Fatt, *The network model of porous media.*, Petrol. Trans. AIME **207** (1956), 144–159.

[Fis96] G. S. Fishman, *Monte Carlo*, Springer Series in Operations Research, Springer-Verlag, New York, 1996, Concepts, algorithms, and applications. MR 1392474 (97g:65019)

[GM75] R. Glowinski and A. Marroco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires*, ESAIM: Mathematical Modelling and Numerical Analysis-Modélisation Mathématique et Analyse Numérique **9** (1975), no. R2, 41–76.

[GN11] A. Gloria and S. Neukamm, *Commutability of homogenization and linearization at identity in finite elasticity and applications*, Ann. Inst. H. Poincaré Anal. Non Linéaire **28** (2011), no. 6, 941–964. MR 2859933 (2012j:35017)

[GNO14] A. Gloria, S. Neukamm, and F. Otto, *An optimal quantitative two-scale expansion in stochastic homogenization of discrete elliptic equations*, ESAIM Math. Model. Numer. Anal. **48** (2014), no. 2, 325–346. MR 3177848

[GNO15] ———, *Quantification of ergodicity in stochastic homogenization: optimal bounds via spectral gap on Glauber dynamics*, Invent. Math. **199** (2015), no. 2, 455–515. MR 3302119

[GO12] A. Gloria and F. Otto, *An optimal error estimate in stochastic homogenization of discrete elliptic equations*, Ann. Appl. Probab. **22** (2012), no. 1, 1–28. MR 2932541

[GO15] ———, *Quantitative estimates on the periodic approximation of the corrector in stochastic homogenization*, ESAIM: Proc. **48** (2015), 80–97.

[JKO94] V. V. Jikov, S. M. Kozlov, and O. A. Oleĭnik, *Homogenization of differential operators and integral functionals*, Springer-Verlag, Berlin, 1994. MR 1329546 (96h:35003b)

[KFG+03] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin, *Determination of the size of the representative volume element for random composites: statistical and numerical approach*, International Journal of solids and structures **40** (2003), no. 13, 3647–3679.

[Koz79] S. M. Kozlov, *Averaging of random operators*, Matematicheskii Sbornik **151** (1979), no. 2, 188–202.

[Koz87] ———, *Averaging of difference schemes*, Mathematics of the USSR-Sbornik **57** (1987), no. 2, 351.

[Kre85] U. Krengel, *Ergodic theorems*, de Gruyter Studies in Mathematics, vol. 6, Walter de Gruyter & Co., Berlin, 1985, With a supplement by Antoine Brunel. MR 797411 (87i:28001)

[Kün83] R. Künnemann, *The diffusion limit for reversible jump processes onz d with ergodic random bond conductivities*, Communications in Mathematical Physics **90** (1983), no. 1, 27–68.

[LB10] C. Le Bris, *Some numerical approaches for weakly random homogenization*, Numerical Mathematics and Advanced Applications 2009, Springer, 2010, pp. 29–45.

[LBLM] C. Le Bris, F. Legoll, and W. Minvielle, *Special quasirandom structures: a selection approach for stochastic homogenization.*

[LBT12] C. Le Bris and F. Thomines, *A reduced basis approach for some weakly stochastic multiscale problems*, Chinese Annals of Mathematics, Series B **33** (2012), no. 5, 657–672.

[Liu08] J. S. Liu, *Monte Carlo strategies in scientific computing*, Springer Series in Statistics, Springer, New York, 2008. MR 2401592 (2010b:65013)

[LM13] F. Legoll and W. Minvielle, *Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem*, arXiv preprint arXiv:1302.0038 (2013).

[LM15a] ———, *A control variate approach based on a defect-type theory for variance reduction in stochastic homogenization*, Multiscale Modeling & Simulation **13** (2015), no. 2, 519–550.

[LM15b] ———, *Variance reduction using antithetic variables for a nonlinear convex stochastic homogenization problem*, Discrete and Continuous Dynamical Systems - Series S **8** (2015), no. 1, 1–27.

[LMOS] F. Legoll, W. Minvielle, A. Obliger, and M. Simon, *In prep.*

[LMOS15] ———, *A parameter identification problem in stochastic homogenization*, ESAIM: Proc. **48** (2015), 190–214.

[LT94] P. Le Tallec, *Numerical methods for nonlinear three-dimensional elasticity*, Handbook of numerical analysis, Vol. III, Handb. Numer. Anal., III, North-Holland, Amsterdam, 1994, pp. 465–622. MR 1307410 (96b:73093)

[Min15] W. Minvielle, *Thèse de l'Université Paris Est.*

[Mou15] J.-C. Mourrat, *First-order expansion of homogenized coefficients under bernoulli perturbations*, Journal de Mathématiques Pures et Appliquées **103** (2015), no. 1, 68–101.

[Nol14] J. Nolen, *Normal approximation for a random elliptic equation*, Probability Theory and Related Fields **159** (2014), no. 3-4, 661–700.

[NPS12] J. Nolen, G. A. Pavliotis, and A. M. Stuart, *Multiscale modelling and inverse problems*, Numerical Analysis of Multiscale Problems, Springer, 2012, pp. 1–34.

[PV81] G. C. Papanicolaou and S. R. S. Varadhan, *Boundary value problems with rapidly oscillating random coefficients*, Random fields, Vol. I, II (Esztergom, 1979), Colloq. Math. Soc. János Bolyai, vol. 27, North-Holland, Amsterdam-New York, 1981, pp. 835–873. MR 712714 (84k:58233)

[Shi84] A. N. Shiryayev, *Probability*, Graduate Texts in Mathematics, vol. 95, Springer-Verlag, New York, 1984. MR 737192 (85a:60007)

[Tar97] L. Tartar, *Estimations of homogenized coefficients [MR0540123 (80i:35010)]*, Topics in the mathematical modelling of composite materials, Progr. Nonlinear Differential Equations Appl., vol. 31, Birkhäuser Boston, Boston, MA, 1997, pp. 9–20. MR 1493038

[Tem72] A. A. Tempel'man, *Ergodic theorems for general dynamical systems*, Trudy Moskov. Mat. Obšč. **26** (1972), 95–132. MR 0374388 (51 #10588)

[Tho97] V. Thomée, *Galerkin finite element methods for parabolic problems*, Springer Series in Computational Mathematics, vol. 25, Springer-Verlag, Berlin, 1997. MR 1479170 (98m:65007)

[vPDFN10] J. von Pezold, A. Dick, M. Friák, and J. Neugebauer, *Generation and performance of special quasirandom structures for studying the elastic properties of random alloys: Application to Al-Ti*, Physical Review B **81** (2010), no. 9, 094203.

[WFBZ90] S.-H. Wei, L. G. Ferreira, J. E. Bernard, and A. Zunger, *Electronic properties of random alloys: Special quasirandom structures*, Physical Review B **42** (1990), no. 15, 9622.

[Yur86] V. V. Yurinskii, *Averaging of symmetric diffusion in random medium*, Siberian Mathematical Journal **27** (1986), no. 4, 603–613.

[ZWFB90] A. Zunger, S.-H. Wei, L. G. Ferreira, and J. E. Bernard, *Special quasirandom structures*, Physical Review Letters **65** (1990), no. 3, 353.