Analyse mathématique de quelques modèles en calcul de structures électroniques et homogénéisation

Arnaud Anantharaman

▶ **To cite this version:**

Arnaud Anantharaman. Analyse mathématique de quelques modèles en calcul de structures électroniques et homogénéisation. Mathématiques générales [math.GM]. Université Paris-Est, 2010. Français. NNT : 2010PEST1002 . tel-00558618v2

HAL Id: tel-00558618

https://pastel.hal.science/tel-00558618v2

Submitted on 20 Apr 2011

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.

UNIVERSITÉ PARIS-EST

ÉCOLE DOCTORALE MSTIC

Thèse de doctorat
Mathématiques appliquées

Arnaud Anantharaman

Sujet : Analyse mathématique de quelques modèles en calcul de structures électroniques et homogénéisation

Soutenue le 16 novembre 2010 devant un jury composé de :

Directeur de thèse	**Eric Cancès**
Co-directeur de thèse	**Grégoire Allaire**
Rapporteurs	**Maria Esteban**
	Guillaume Bal
Examinateurs	**Isabelle Terrasse**
	Habib Ammari
	Xavier Blanc
Invité	**Claude Le Bris**

Il y a tant de gens à qui dédier cette thèse...

Avant tout à mes parents Claire et Siva, mon frère Bruno et ma soeur Nalini, pour leur soutien précieux et constant depuis ving-sept ans,

A Sonia la petite nouvelle qui vient agrandir la famille (bienvenue dans le monde !),

A mon grand-père Emile qui aurait été fier de tout ce qui arrive en ce moment,

A Cristina, qui est sans aucun doute la plus belle découverte que j'ai faite pendant cette thèse,

A mes amis sur qui j'ai la chance de pouvoir compter et qui me supportent sans être bien payés en retour,

A ces rencontres au cours de ces trois dernière années qui, positives ou négatives, m'ont fait comprendre tellement de choses,

A Eminem dont la musique rythme ma vie et en accompagne les bons et mauvais moments depuis une décennie.

On a grandi ensemble, on a construit ensemble
Je me remémore les discussions que l'on avait ensemble
Et nos rêves, tu t'en souviens de nos rêves ?
Quand on était dans les hangars, quand on sentait monter la fièvre
Putain c'est loin tout ça, c'est loin
J'ai passé mon adolescence à défoncer des trains
Je ne regrette rien
On a tellement tutoyé de fois le bonheur qu'on pourrait mourir demain
Sans regrets, sans remords
Notre seule erreur était de rêver un peu trop fort
En omettant le rôle important que pouvait jouer le temps
Sur les comportements de chacun, pourtant
On venait tous du même quartier
On avait tous la même culture de cité
Ouais c'était vraiment l'idéal, en effet
On avait vraiment tout pour réussir mais
Tout n'est pas si facile, les destins se séparent, l'amitié c'est fragile
Pour nous la vie ne fut jamais un long fleuve tranquille
Et aujourd'hui encore, tout n'est pas si facile.

NTM, *Tout n'est pas si facile*, Paris sous les bombes, 1995.

Remerciements

Je tiens avant tout à exprimer ma gratitude à Eric Cancès, qui a accepté d'encadrer cette thèse et s'est montré extrêmement disponible tout au long de ces trois ans. Ce travail doit beaucoup à son énergie, son enthousiasme et son optimisme (sans parler de ses idées) qui ont souvent dû se substituer aux miens ! Ma gratitude va pareillement à Grégoire Allaire, dont le cours d'analyse numérique à l'Ecole Polytechnique m'a convaincu de poursuivre les études mathématiques il y a quelques années de cela (avant que je m'aperçoive que le théorème de Lax-Milgram ne pouvait pas tout résoudre), et qui a bien voulu codiriger cette thèse après avoir supervisé mon stage de master.

J'ai une pensée particulière pour Claude Le Bris, que je considère comme mon troisième directeur pour avoir eu la chance de travailler avec lui pendant mes deuxième et troisième années et de bénéficier de son aide de nombreuses heures durant. Le temps qu'il m'a consacré, sa vision d'ensemble de mes domaines d'intérêt, ses perspectives originales sur les problèmes rencontrés et sa propension à « mettre la main à la pâte » m'ont plus d'une fois sorti d'une impasse. Je suis très heureux qu'il soit présent dans le jury en tant que membre invité.

Je suis très reconnaissant à Maria Esteban (que j'ai eu le plaisir de rencontrer lors d'un voyage à Minneapolis et d'une école d'hiver quelque peu originale à Alexandrie) et Guillaume Bal d'avoir accepté d'apporter leur expertise sur les deux sujets différents qui composent ma thèse en en étant les rapporteurs. J'ai plusieurs fois sollicité l'aide d'Habib Ammari et de Xavier Blanc au cours de celle-ci, pour des questions rébarbatives de régularité elliptique, et j'ai toujours trouvé porte ouverte ; je les remercie d'avoir été si disponibles, et de me faire l'honneur de leur présence dans le jury. Je suis également très honoré qu'Isabelle Terrasse, figure tutélaire de ma courte carrière scientifique puisque présente depuis mon stage de master, ait accepté de faire partie du jury.

Enfin, j'ai eu au cours de ces trois années maintes discussions mathématiques fructueuses avec Frédéric Legoll, Mathieu Lewin et Didier Smets. Je tiens à leur adresser mes chaleureux remerciements.

Cette période de doctorat a été rendue très agréable par la très bonne ambiance que j'ai trouvée au Cermics, qui m'a aidé à surmonter les nombreuses phases de disette intellectuelle. J'ai une pensée affectueuse pour mes comparses du bureau B411 - Ronan (et ses cheveux !), Kimiya, Rémi, Virginie -, et pour tous les gens avec qui j'ai partagé de bons moments, dans le désordre le plus complet Ismaïla (seul le crime paie), Yanli, Hanen, David P., Tony, Sébastien, Gabriel (malgré le bizutage permanent à la cantine), Florian, Matthew, Nadia, Jean-Philippe, Alexandre, Régis, et quand je m'aventurais au 3ème étage la sombre équipe des probabilistes - Olivier, Maxence, Abdel, José, Patrick, Raphaël - et celle non moins sombre des numériciens - Julie, Laurent, David D. Un gros merci à Catherine, Martine et Sylvie pour s'être occupé aussi efficacement de tout ce qui était extra-scientifique.

Quand je n'étais pas au Cermics, j'ai pu trouver refuge au centre de recherche d'EADS à Suresnes, terrain de jeu familier pour y avoir effectué mon stage de master. J'ai beaucoup apprécié le temps que j'y ai passé lors de ces quatre années, le mérite en revient aux gens

que j'y ai côtoyés, Michel (qui a eu le malheur d'être mon encadrant de stage et porte donc la responsabilité de tout ce qui s'est ensuivi), Fabien, Vincent (Feuillou !), Vassili, Nabil, Régis, Ariane, Stéphane, Fanny, Pierre, Hichem, Jayant, Eric, Benoît, Anabelle, Jean-Loup, Gilles, Younes. Je souligne que le financement d'EADS a grandement contribué aux excellentes conditions matérielles dont j'ai pu bénéficier pendant ma thèse.

Je profite de cette tribune pour exprimer (pour une fois !) mon immense joie de pouvoir compter sur mes amis qui m'ont aidé à traverser la thèse sans trop de séquelles, je pense en particulier à Sandrine, Thomas, mon groupe de la mort (Alexis, Mehdi, Youssef), mon whity Panzer aka Pierre L., Nico B., Rodolphe, Marion, Paul, Emilie, Cloé, Carole, Pierre, Antoine, Stéphane, Etienne, Cédric, Rémy, Phil, Pierre-Yves, Arthur, Benoît, Marie, Aurélie, Joris, Guillaume, Laurent, Cécile, Olga, Guillem, Diane, Sacha, Matthieu, Romain, Anne-So, Nico D., France, Patricia, Christophe, Guy-Albert.

Je mesure la chance que j'ai d'avoir été constamment soutenu par mes parents dans ma vie professionnelle et extra-professionnelle. Je voudrais leur témoigner ma grande affection et ma profonde admiration. Mon frère et ma soeur ont toujours été présents pour m'encourager (je n'inclus pas les années où j'étais leur souffre-douleur), je ne sais comment les en remercier ; je garde en tête comme moments forts partagés pendant ces trois ans le concert de NTM et une magnifique semaine en Haute-Loire dans une période particulière. Je suis très ému et heureux que ce cycle qui se termine pour moi coïncide avec l'arrivée de ma nièce Sonia, petite chose adorable à qui je souhaite le meilleur, ainsi qu'à son père Charlie qui a eu la bonne idée d'être mon « beauf ».

Cri, tu sais que tu peux compter sur moi ?

Analyse mathématique de quelques modèles en calcul de structures électroniques et homogénéisation

Résumé. Cette thèse comporte deux volets distincts. Le premier, qui fait l'objet du chapitre 2, porte sur les modèles mathématiques en calcul de structures électroniques, et consiste plus particulièrement en l'étude des modèles de type Kohn-Sham avec fonctionnelles d'échange-corrélation LDA et GGA. Nous prouvons, pour un système moléculaire neutre ou chargé positivement, que le modèle Kohn-Sham LDA étendu admet un minimiseur, et que le modèle Kohn-Sham GGA pour un système contenant deux électrons admet un minimiseur. Le second volet de la thèse traite de problématiques diverses en homogénéisation. Dans les chapitres 3 et 4, nous nous intéressons à un modèle de matériau aléatoire dans lequel un matériau périodique est perturbé de manière stochastique. Nous proposons plusieurs approches, certaines rigoureuses et d'autres heuristiques, pour calculer au second ordre en la perturbation le comportement homogénéisé de ce matériau de manière purement déterministe. Les tests numériques effectués montrent que ces approches sont plus efficaces que l'approche stochastique directe. Le chapitre 5 est consacré aux couches limites en homogénéisation périodique, et vise notamment, dans le cadre parabolique, à comprendre comment prendre en compte les conditions aux limites et initiale, et comment corriger en conséquence le développement à deux échelles sur lequel repose classiquement l'homogénéisation, pour obtenir des estimations d'erreur dans des espaces fonctionnels adéquats.

Mots-clés : Equations aux dérivées partielles, Chimie quantique, Modèles de Kohn-Sham, Homogénéisation, Matériaux aléatoires.

Mathematical analysis of some models in electronic structure calculations and homogenization

Abstract. This thesis is divided into two parts. The first part, that coincides with Chapter 2, deals with mathematical models in quantum chemistry, and specifically focuses on Kohn-Sham models with LDA and GGA exchange-correlation functionals. We prove, for a neutral or positively charged system, that the extended Kohn-Sham LDA model admits a minimizer, and that the Kohn-Sham GGA model for a two-electron system admits a minimizer. The second part is concerned with various issues in homogenization. In Chapters 3 and 4, we introduce and study a model in which the material of interest consists of a random perturbation of a periodic material. We propose different approaches, either rigorous or formal, to compute the homogenized behavior of this material up to the second order in the size of the perturbation, in an entirely deterministic way. Numerical experiments show the efficiency of these approaches as compared to the direct stochastic homogenization process. Chapter 5 is devoted to boundary layers in periodic homogenization, in particular in the parabolic setting. It aims at giving a better understanding of how to take into account boundary and initial conditions, and how to correct the two-scale expansion on which homogenization is classically grounded, to obtain fine error estimates.

Keywords : Partial differential equations, Quantum Chemistry, Kohn-Sham models, Homogenization, Random materials.

Articles publiés

- A. Anantharaman, E. Cancès, *Existence of minimizers for Kohn-Sham models in quantum chemistry*, Ann. IHP (C) Nonlinear Analysis, Vol. 26 no. 6 (2009), pp. 2425-2455.

- A. Anantharaman, C. Le Bris, *Homogenization of a weakly randomly perturbed periodic material*, C. R. Acad. Sci. Paris Série I, Vol. 348 (9-10) (2010), pp. 529-534.

Articles soumis pour publication

- A. Anantharaman, C. Le Bris, *A numerical approach related to defect-type theories for some weakly random problems in homogenization*, preprint disponible à http ://arxiv.org/abs/1005.3910, soumis à SIAM Multiscale Modeling & Simulation.

- A. Anantharaman, C. Le Bris, *Elements of mathematical foundations for a numerical approach for weakly random homogenization problems*, preprint disponible à http ://arxiv.org/abs/1005.3922, soumis à Communications in Computational Physics.

Compte-rendus de conférence

- A. Anantharaman, E. Cancès, *Sur les modèles de type Kohn-Sham avec fonctionnelles d'échange-corrélation LDA et GGA*, Comptes-rendus de la 12è Rencontre du Non-Linéaire, IHP Paris NL Pub. (2009), pp. 7-12.

- A. Anantharaman, R. Costaouec, C. Le Bris, F. Legoll, F. Thomines, *Introduction to numerical stochastic homogenization and the related computational challenges : some recent developments*, Lecture Note Series, IMS, National University of Singapore, à paraître.

Communications orales

- *Using concentration-compactness theory to analyze some chemistry models*, CIMPA school "Recents developments in the theory of elliptic PDE's", Alexandrie, 26 janvier - 3 février 2009.

- *Sur les modèles de type Kohn-Sham avec fonctionnelles d'échange-corrélation LDA et GGA*, Rencontre du Non-Linéaire 2009, Paris, 11-13 mars 2009.

- *Homogénéisation et mélange aléatoire de matériaux périodiques*, Journées du GdR MASCOT-NUM, Paris, 18-20 mars 2009.

- *Homogenization of a weakly randomly perturbed periodic material,* 33rd Conference on Stochastic Processes and Applications, Berlin, 27 juin - 31 juillet 2009.

- *Homogenization of a weakly randomly perturbed periodic material,* SIAM Conference on Mathematical Aspects of Material Science, Philadelphie, 23-26 mai 2010.

Table des matières

1 Introduction générale **1**

1.1 Modèles mathématiques en calcul de structures électroniques 1

1.1.1 Modèle de Hartree-Fock 3

1.1.2 Modèles de Kohn-Sham et théorie de la fonctionnelle de la densité 5

1.2 Homogénéisation 10

1.2.1 Problématique industrielle et homogénéisation 10

1.2.2 Homogénéisation périodique 13

1.2.3 Homogénéisation stochastique 16

1.2.4 Matériaux faiblement aléatoires 18

1.2.5 Couches limites en homogénéisation périodique 20

2 Kohn-Sham models in Quantum Chemistry **23**

2.1 Introduction 23

2.2 Mathematical foundations of DFT and Kohn-Sham models 24

2.2.1 Density Functional Theory 24

2.2.2 Kohn-Sham models 27

2.3 Main results 32

2.4 Proofs 35

2.4.1 Preliminary results 36

2.4.2 Proof of Lemma 2.1 37

2.4.3 Proof of Theorem 2.2 40

2.4.4 Proof of Theorem 2.3 45

3 A defect-type weakly random model in homogenization **65**

3.1 Introduction 65

3.2 Some classical results of elliptic homogenization 69

3.2.1 Periodic homogenization 69

3.2.2 Stochastic homogenization 70

3.3 Homogenization of a randomly perturbed periodic material 72

3.3.1 Presentation of the model 72

3.3.2 An ergodic approximation of the homogenized tensor 73

3.3.3 Convergence of the first-order term $A_1^{*,N}$ 76

3.3.4 Convergence of the second-order term $A_2^{*,N}$ 81

3.4 Numerical experiments 90

3.4.1 Methodology 90

3.4.2 Results 93

3.5 Appendix 101

3.5.1 One-dimensional computations 101

3.5.2 Some technical lemmas 104

4 On some approaches for weakly random homogenization 111
4.1 Introduction . . . 111
4.2 A model of a weakly random material and a first approach . . . 113
4.3 A formal approach . . . 121
4.3.1 A new assumption on the image measure . . . 121
4.3.2 An ergodic approximation of the homogenized tensor . . . 124
4.3.3 Convergence of the first-order term . . . 128
4.3.4 Convergence of the second-order term . . . 131
4.4 Numerical experiments . . . 137
4.4.1 Methodology . . . 137
4.4.2 An example of setting for our theory in Section 4.2 (and 4.3) . . . 140
4.4.3 A first example of setting for our formal approach of Section 4.3 . 140
4.4.4 A second example of setting for our formal approach of Section 4.3 146
4.5 Appendix . . . 151
4.5.1 Elements of distribution theory . . . 151
4.5.2 Some technical results . . . 152
4.5.3 The one-dimensional case . . . 155
4.5.4 A proof of the approach of Section 4.3 in a specific setting . . . 161

5 Boundary layers in periodic homogenization 167
5.1 Introduction . . . 167
5.2 General setting and notation . . . 169
5.2.1 Stationary setting . . . 170
5.2.2 Transient setting . . . 172
5.3 Boundary layers in the homogenization of elliptic equations . . . 173
5.3.1 Classical results for Dirichlet boundary conditions . . . 173
5.3.2 Neumann boundary conditions . . . 178
5.4 Boundary layers for parabolic equations . . . 188
5.4.1 Need for an "initial layer" . . . 189
5.4.2 A theoretical boundary+initial layer . . . 191
5.4.3 Initial layer in an "unbounded" domain . . . 193
5.4.4 General case . . . 198
5.4.5 One-dimensional toy model . . . 206
5.5 Appendix: two parabolic regularity results . . . 211

Bibliographie 213

Chapitre 1

Introduction générale

Nous présentons ici les deux grands thèmes de ce travail de thèse. Le premier concerne les propriétés mathématiques de modèles de chimie quantique dits de Kohn-Sham, utilisés en calcul de structures électroniques. Le second, qui fait l'objet d'une collaboration avec EADS IW, traite de deux problématiques distinctes dans le cadre de l'homogénéisation des matériaux composites, et consiste en l'étude d'un modèle perturbatif de matériau aléatoire d'une part, et des couches limites en homogénéisation parabolique d'autre part. Après avoir détaillé les contextes scientifique et le cas échéant applicatif de ces travaux, nous introduisons les résultats qui seront prouvés dans le corps de la thèse.

1.1 Modèles mathématiques en calcul de structures électroniques

On s'intéresse dans cette partie à un système moléculaire comprenant M noyaux atomiques et N électrons. Pour simplifier, on utilise les unités atomiques, ce qui se traduit par

$$m_e = 1, \quad e = 1, \quad \hbar = 1, \quad \frac{1}{4\pi\varepsilon_0} = 1, \tag{1.1}$$

où m_e désigne la masse d'un électron, e sa charge, $\hbar$ la constante de Planck réduite et ε_0 la permittivité diélectrique du vide.

On se place dans tout ce qui suit dans le cadre de l'approximation de Born-Oppenheimer : les noyaux étant beaucoup plus lourds que les électrons, la dynamique des premiers peut être découplée de celle des seconds [20, 4, 41, 60, 82].

Sous cette approximation, les M noyaux sont considérés comme des particules classiques, de positions $\bar{x}_1, \cdots, \bar{x}_M$ dans $\mathbb{R}^3$ et de charges $z_1, \cdots z_M$ dans $\mathbb{N}^*$ en unités atomiques. Les N électrons sont quant à eux représentés dans le formalisme de la physique quantique par une fonction d'onde notée $\psi(x_1, \cdots, x_N)$, où pour tout $i \in [\![1, N]\!]$, x_i est un vecteur de $\mathbb{R}^3$.

Un des problèmes les plus importants dans le calcul des structures électroniques est la détermination de l'état fondamental du système, c'est-à-dire l'état de plus basse énergie. Ce dernier conditionne en effet la plupart des propriétés physiques et chimiques du système. Sous l'approximation de Born-Oppenheimer, cette recherche du fondamental prend la forme d'une double minimisation : les positions des noyaux étant fixées, on calcule la configuration électronique d'énergie minimale, puis on optimise la géométrie des noyaux.

Détaillons ces deux étapes. Pour une configuration atomique $(\bar{x}_1, \cdots, \bar{x}_M)$ donnée, l'évolution des électrons est décrite par le Hamiltonien

$$H^{\{\bar{x}_k\}} = -\sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i} - \sum_{i=1}^{N}\sum_{k=1}^{M} \frac{z_k}{|x_i - \bar{x}_k|} + \sum_{1\leq i<j\leq N} \frac{1}{|x_i - x_j|}. \tag{1.2}$$

Le premier terme de $H^{\{\bar{x}_k\}}$ correspond à l'énergie cinétique des électrons, le second terme à l'attraction coulombienne entre électrons et noyaux et le troisième à la répulsion coulombienne interélectronique. La configuration électronique donnant la plus basse énergie est alors obtenue en calculant

$$U(\bar{x}_1, \cdots, \bar{x}_M) = \inf\left\{\langle\psi, H^{\{\bar{x}_k\}}\psi\rangle, \quad \psi \in \mathcal{H}, \quad \|\psi\|_{L^2(\mathbb{R}^{3N})} = 1\right\}, \tag{1.3}$$

où

$$\mathcal{H} = \bigwedge_{i=1}^{N} H^1(\mathbb{R}^3).$$

Précisons que dans (1.3), et selon les principes de la physique quantique :

- les fonctions d'onde sont normalisées (ceci vient de leur interprétation comme probabilité de présence), d'où la condition $\|\psi\|_{L^2(\mathbb{R}^{3N})} = 1$;
- en vertu du principe d'exclusion de Pauli, l'espace $\bigwedge_{i=1}^{N} H^1(\mathbb{R}^3)$ désigne le sous-ensemble de $H^1(\mathbb{R}^{3N})$ composé des fonctions d'onde antisymétriques par permutation de deux variables, c'est-à-dire

$$\psi(x_{p(1)}, x_{p(2)}, \cdots, x_{p(N)}) = (-1)^{\varepsilon(p)}\psi(x_1, x_2, \cdots, x_N).$$

Les variables de spin ont été omises. Elles ne joueront pas de rôle dans la suite.

Une fois le problème électronique (1.3) résolu, le potentiel effectif dans lequel se déplacent les noyaux est donné par

$$W(\bar{x}_1, \cdots, \bar{x}_M) = U(\bar{x}_1, \cdots, \bar{x}_M) + \sum_{1\leq k<l\leq M} \frac{1}{|\bar{x}_k - \bar{x}_l|}. \tag{1.4}$$

L'état fondamental du système est alors obtenu en minimisant W sur toutes les configurations de noyaux $(\bar{x}_1, \cdots, \bar{x}_M)$ de $\mathbb{R}^{3M}$.

Dans le cadre de cette thèse, nous nous intéresserons uniquement à la résolution du problème électronique (1.3) pour une géométrie de noyaux donnée. Pour simplifier, nous réécrivons ce problème de minimisation de la façon suivante :

$$\inf\left\{\langle\psi, H\psi\rangle, \quad \psi \in \mathcal{H}, \quad \|\psi\|_{L^2(\mathbb{R}^{3N})} = 1\right\} \tag{1.5}$$

avec

$$\mathcal{H} = \bigwedge_{i=1}^{N} H^1(\mathbb{R}^3),$$

$$H = -\sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i} + \sum_{i=1}^{N} V(x_i) + \sum_{1\leq i<j\leq N} \frac{1}{|x_i - x_j|},$$

$$V(x) = -\sum_{k=1}^{M} \frac{z_k}{|x - \bar{x}_k|}.$$

Dans (1.5), les $\bar{x}_k$ jouent le rôle de simples paramètres de $\mathbb{R}^3$.

Une approche numérique directe de (1.5) nécessite de discrétiser $\mathbb{R}^{3N}$. Le coût de calcul qui en résulte est trop élevé pour les systèmes complexes comprenant plus de deux électrons. Pour remédier à ce problème, beaucoup de modèles consistant en des approximations de (1.5) existent dans la littérature. Parmi ceux-ci, on distingue deux grandes classes :

- les modèles reposant sur des méthodes de fonctions d'onde (voir [25] pour une introduction mathématique à ces modèles) : le plus connu d'entre eux, celui de Hartree-Fock, fait l'objet de la section suivante ;
- les modèles issus de la théorie de la fonctionnelle de la densité, et notamment ceux de Kohn-Sham [33, 67] présentés dans la Section 1.1.2.

1.1.1 Modèle de Hartree-Fock

Le modèle de Hartree-Fock repose sur une réduction de l'ensemble de minimisation de (1.5). De manière schématique, le but est de remplacer l'espace $H^1(\mathbb{R}^{3N})$ par le produit $H^1(\mathbb{R}^3) \times \cdots \times H^1(\mathbb{R}^3)$: d'un point de vue numérique, il suffit alors de discrétiser $\mathbb{R}^3$ en lieu et place de $\mathbb{R}^{3N}$.

Le nouvel ensemble de minimisation doit respecter le principe d'exclusion de Pauli et donc l'antisymétrie des fonctions d'onde. Il s'agit de l'espace des déterminants de Slater, c'est-à-dire des fonctions d'onde qui s'écrivent

$$\psi(x_1, \cdots, x_N) = \frac{1}{\sqrt{N!}} \det(\phi_i(x_j)) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(x_1) & \cdots & \phi_1(x_N) \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \cdot & & \cdot \\ \phi_N(x_1) & \cdots & \phi_N(x_N) \end{vmatrix} \tag{1.6}$$

où les ϕ_i sont des fonctions de $H^1(\mathbb{R}^3)$ appelées orbitales moléculaires vérifiant les conditions d'orthonormalité $\displaystyle\int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}$.

On note

$$\mathcal{W}_N = \left\{ \Phi = \{\phi_i\}_{1\leq i\leq N}, \quad \phi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \quad 1 \leq i, j \leq N \right\} \tag{1.7}$$

l'ensemble des configurations de N orbitales moléculaires, et

$$\mathcal{S}_N = \left\{ \psi \in \mathcal{H}, \quad \exists \Phi = \{\phi_i\}_{1\leq i\leq N} \in \mathcal{W}_N, \quad \psi = \frac{1}{\sqrt{N!}} \det(\phi_i(x_j)) \right\} \tag{1.8}$$

l'ensemble des déterminants de Slater.

Le modèle de Hartree-Fock est alors le problème de minimisation

$$\inf\left\{\langle\psi, H\psi\rangle, \quad \psi \in \mathcal{S}_N\right\}. \tag{1.9}$$

Pour une fonction d'onde ψ dans $\mathcal{S}_N$, $\langle\psi, H\psi\rangle$ peut s'écrire en fonction du n-uplet Φ introduit dans (1.7) et (1.8). Plus précisément, on a

$$\langle\psi, H\psi\rangle = \mathcal{E}^{HF}(\Phi), \tag{1.10}$$

où la fonctionnelle d'énergie de Hartree-Fock $\mathcal{E}^{HF}$ est définie par

$$\begin{aligned}\mathcal{E}^{HF}(\Phi) \;=\; & \sum_{i=1}^{N} \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_\Phi\, V + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(x)\,\rho_\Phi(y)}{|x-y|}\,dx\,dy \\ & -\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\tau_\Phi(x,y)^2}{|x-y|}\,dx\,dy,\end{aligned} \tag{1.11}$$

avec $\tau_\Phi(x,y) = \sum_{i=1}^{N} \phi_i(x)\,\phi_i(y)$ et $\rho_\Phi(x) = \tau_\Phi(x,x) = \sum_{i=1}^{N} |\phi_i(x)|^2$.

La fonction ρ_Φ est la densité électronique associée au déterminant de Slater construit à partir de Φ. L'intégrale de cette densité sur $\mathbb{R}^3$ vaut N et permet bien de retrouver le nombre total d'électrons du système. La fonction $\tau_\Phi(x,y)$ définit un opérateur de $L^2(\mathbb{R}^3)$ dans lui-même appelé opérateur densité d'ordre 1, dont le noyau est précisément τ_Φ. Cette terminologie se retrouvera dans la section suivante consacrée aux modèles de Kohn-Sham.

Dans la fonctionnelle d'énergie (1.11), le premier terme correspond à l'énergie cinétique des N électrons. Le second terme représente l'attraction coulombienne exercée par le potentiel V crée par les noyaux, et le troisième l'énergie de Coulomb de la densité ρ_Φ. Ces trois termes admettent une interprétation classique, par opposition au quatrième terme, dit d'échange, d'origine purement quantique car provenant de l'antisymétrie de la fonction d'onde et donc du principe de Pauli.

Au vu de (1.8), (1.9) et (1.10), le modèle de Hartree-Fock peut se réécrire

$$\inf\left\{\mathcal{E}^{HF}(\Phi), \quad \Phi \in \mathcal{W}_N\right\}. \tag{1.12}$$

L'ensemble de minimisation du problème (1.9) étant plus petit que celui du problème initial (1.5), l'énergie fondamentale donnée par (1.9) ou de manière équivalente par (1.12) est plus élevée que celle obtenue par (1.5). La différence est appelée énergie de corrélation.

Soulignons par ailleurs que la restriction de l'ensemble de minimisation sous-jacente à la construction du modèle de Hartree-Fock a une contrepartie : la fonctionnelle d'energie (1.11) n'est pas quadratique en son argument Φ, alors que (1.5) est quadratique en la fonction d'onde ψ.

L'existence d'un minimiseur au problème (1.12) a été montrée pour les système neutres ou chargés positivement, c'est-à-dire pour $Z := \sum_{k=1}^{M} z_k \geq N$ (voir [54] et [58]). La question de l'unicité du minimiseur, ou même celle, moins forte, de l'unicité de la densité associée au minimiseur, est un problème ouvert.

1.1.2 Modèles de Kohn-Sham et théorie de la fonctionnelle de la densité

Le principe de la théorie de la fonctionnelle de la densité (que l'on appellera aussi par son acronyme DFT pour Density Functional Theory), et de tous les modèles qui en découlent, est de décrire le système à l'aide non pas d'une fonction d'onde, mais de la seule densité électronique. Les bénéfices pratiques en sont évidents : on travaille dans $\mathbb{R}^3$ au lieu de $\mathbb{R}^{3N}$.

On définit

$$\mathcal{F}_N = \left\{ \psi \in \mathcal{H},\ \|\psi\|_{L^2(\mathbb{R}^{3N})} = 1 \right\}. \tag{1.13}$$

Le problème de minimisation (1.5) peut se réécrire de la façon suivante, où la dépendance en le potentiel V est explicitée :

$$E(V) = \inf \left\{ \langle \psi, H_V \psi \rangle, \quad \psi \in \mathcal{F}_N \right\}, \tag{1.14}$$

où

$$H_V = H_1 + \sum_{i=1}^{N} V(x_i)$$

et

$$H_1 = -\sum_{i=1}^{N} \frac{1}{2} \Delta_{x_i} + \sum_{1 \leq i < j \leq N} \frac{1}{|x_i - x_j|}. \tag{1.15}$$

A une fonction d'onde ψ dans $\mathcal{F}_N$ est associée la densité électronique

$$\rho_\psi(x) = N \int_{\mathbb{R}^{3(N-1)}} |\psi(x, x_2, \cdots, x_N)|^2 \, dx_2 \cdots dx_N. \tag{1.16}$$

On note

$$\mathcal{I}_N = \{ \rho, \quad \exists \psi \in \mathcal{F}_N, \quad \rho_\psi = \rho \}$$

l'ensemble des densités associées aux fonctions d'ondes admissibles. D'après [53], $\mathcal{I}_N$ peut-être caractérisé de manière équivalente par

$$\mathcal{I}_N = \left\{ \rho \geq 0, \quad \sqrt{\rho} \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \rho = N \right\}. \tag{1.17}$$

Le point de départ de la théorie de la fonctionnelle de la densité est le calcul élémentaire suivant [43, 53] :

$$\begin{aligned} E(V) &= \inf \left\{ \langle \psi, H_V \psi \rangle, \quad \psi \in \mathcal{F}_N \right\} \\ &= \inf \left\{ \inf \{ \langle \psi, H_1 \psi \rangle, \ \ \psi \in \mathcal{F}_N, \ \ \rho_\psi = \rho \} + \int_{\mathbb{R}^3} \rho V, \ \ \rho \in I_N \right\} \\ &= \inf \left\{ F_{LL}(\rho) + \rho V, \ \ \rho \in I_N \right\}, \end{aligned} \tag{1.18}$$

avec

$$F_{LL}(\rho) = \inf \{\langle \psi, H_1 \psi \rangle, \quad \psi \in \mathcal{F}_N, \quad \rho_\psi = \rho\}. \tag{1.19}$$

La fonctionnelle F_{LL} est appelée fonctionnelle de Levy-Lieb. Elle est universelle, au sens où elle ne dépend pas du système moléculaire considéré (ce dernier n'intervenant que dans le potentiel V).

Il est clair d'après (1.18) que la minimisation sur les fonctions d'onde a été remplacée par une minimisation sur la densité électronique. Une autre approche de la DFT est possible, faisant intervenir les opérateurs densité. Nous la détaillons ci-après.

A une fonction d'onde $\psi \in \mathcal{F}_N$, également dénommée état pur, est associé un opérateur densité Γ donné par

$$\Gamma_\psi = |\psi\rangle\langle\psi|.$$

Les états mixtes sont définis comme l'ensemble des combinaisons convexes d'états purs. Ils sont décrits par les opérateurs densité

$$\Gamma = \sum_{i=1}^{+\infty} p_i \, |\psi^i\rangle\langle\psi^i|, \quad 0 \leq p_i \leq 1, \quad \sum_{i=1}^{+\infty} p_i = 1, \quad \psi^i \in \mathcal{F}_N. \tag{1.20}$$

On note $\mathcal{D}_N$ l'ensemble des opérateurs densité admettant la forme (1.20), qui est l'enveloppe convexe de l'ensemble des opérateurs densité associés à des états purs. La densité électronique correspondant à l'opérateur Γ est

$$\rho_\Gamma(x) = \sum_{i=1}^{+\infty} p_i \, \rho_{\psi^i}(x),$$

où ρ_{ψ^i} est la densité associée à l'état pur ψ^i par (1.16).

En désignant par Tr la trace d'un opérateur, il est clair que

$$\mathrm{Tr}\,(\Gamma) = \sum_{i=1}^{+\infty} p_i \|\psi^i\|^2_{L^2(\mathbb{R}^{3N})} = 1, \quad \mathrm{Tr}\,(H_1\Gamma) = \sum_{i=1}^{+\infty} p_i \langle \psi^i, H_1 \psi^i \rangle,$$

$$\mathrm{Tr}\,(H_V\Gamma) = \sum_{i=1}^{+\infty} p_i \langle \psi^i, H_1 \psi^i \rangle + \int_{\mathbb{R}^3} \rho_\Gamma V.$$

On peut montrer que la minimisation sur les états purs (1.14) est équivalente à une minimisation sur les états mixtes, d'où

$$E(V) = \inf \{\mathrm{Tr}\,(H_V\Gamma), \quad \Gamma \in \mathcal{D}_N\}.$$

D'autre part, on a

$$\{\rho, \quad \exists \Gamma \in \mathcal{D}_N, \quad \rho_\Gamma = \rho\} = \mathcal{I}_N.$$

Un calcul analogue à celui de (1.18) nous donne alors

$$E(V) = \inf \left\{ F_L(\rho) + \int_{\mathbb{R}^3} \rho V, \quad \rho \in \mathcal{I}_N \right\}, \tag{1.21}$$

où $F_L(\rho)$ est la fonctionnelle de Lieb définie par

$$F_L(\rho) = \inf \{\mathrm{Tr}\,(H_1\Gamma), \quad \Gamma \in \mathcal{D}_N \text{ de densité } \rho\}\,. \tag{1.22}$$

De même que F_{LL}, F_L ne dépend pas du système moléculaire considéré.

Nous disposons donc, via (1.18) et (1.21), de deux manières de rechercher le fondamental du système en considérant la variable densité électronique plutôt que les fonctions d'onde. L'avantage de la construction reposant sur les états mixtes par rapport à celle fondée sur les états purs, présentement peu évident, apparaîtra clairement lorsque nous introduirons les modèles de Kohn-Sham.

La simplification apportée par la théorie de la fonctionnelle de la densité a une contrepartie : il n'existe pas d'expression explicite des fonctionnelles F_{LL} et F_L définies par (1.19) et (1.22) respectivement. En pratique, on doit donc utiliser des approximations, basées sur des évaluations exactes de ces fonctionnelles pour des systèmes de référence. Beaucoup de modèles existent dans la littérature. Les modèles de type Thomas-Fermi sont fondés sur un système de référence qui est un gaz homogène d'électrons ; ils sont en fait antérieurs à la dérivation de la DFT détaillée ci-dessus. Contenant des difficultés mathématiques que l'on retrouve dans les modèles plus complexes, et donc instructifs d'un point de vue théorique, ils ne sont plus utilisés dans les calculs numériques car trop rudimentaires.

Plus précis que les modèles de type Thomas-Fermi, les modèles dits de Kohn-Sham, dont l'étude fait l'objet du Chapitre 2, ont pour système de référence un système de N électrons sans interaction. Le Hamiltonien H_1 defini par (1.15) est alors remplacé par

$$H_0 = -\sum_{i=1}^{N} \frac{1}{2}\Delta_{x_i}. \tag{1.23}$$

Le Hamiltonien H_0 est utilisé pour obtenir une fonctionnelle d'énergie cinétique. Celle-ci revêt deux expressions différentes suivant que l'on adopte la construction de Levy-Lieb (états purs) ou la construction de Lieb (états mixtes). Dans le premier cas, on introduit la fonctionnelle de Kohn-Sham

$$\widetilde{T}_{KS}(\rho) = \inf \{\langle \psi, H_0\psi\rangle, \quad \psi \in \mathcal{F}_N, \quad \rho_\psi = \rho\}\,. \tag{1.24}$$

La fonctionnelle $\widetilde{T}_{KS}$ n'admet une expression exploitable que si l'infimum dans (1.24) est atteint en une fonction d'onde ψ qui prend la forme d'un déterminant de Slater. Il est prouvé que ce n'est pas toujours le cas [53]. Néanmoins, l'approche pratique est de restreindre la minimisation au sous-ensemble des déterminants de Slater, et de considérer une approximation de $\widetilde{T}_{KS}$ donnée par

$$T_{KS}(\rho) = \inf \left\{ \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla\phi_i|^2, \quad \phi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \quad \sum_{i=1}^{N} \phi_i^2 = \rho \right\}. \tag{1.25}$$

Ce problème de représentation des minimiseurs de (1.24) ne se pose pas si l'on calcule la fonctionnelle d'énergie cinétique en utilisant les états mixtes. La fonctionnelle ainsi obtenue, dite de Janak, s'écrit alors

$$T_J(\rho) = \inf \{\mathrm{Tr}\,(H_0\Gamma), \quad \Gamma \in \mathcal{D}_N, \quad \rho_\Gamma = \rho\}\,, \tag{1.26}$$

et l'on peut montrer rigoureusement l'équivalence entre (1.26) et la formulation suivante :

$$\begin{aligned} T_J(\rho) &= \inf\left\{\frac{1}{2}\sum_{i=1}^{N} n_i \int_{\mathbb{R}^3} |\nabla\phi_i|^2, \quad \phi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \right. \\ &\qquad \left. 0 \le n_i \le 1, \quad \sum_{i=1}^{+\infty} n_i = N, \quad \sum_{i=1}^{+\infty} n_i|\phi_i|^2 = \rho \right\}. \end{aligned} \tag{1.27}$$

Comparant (1.25) et (1.27), la distinction entre états purs et états mixtes (qui sont, rappelons le, combinaisons convexes d'états purs) apparaît clairement.

Nous disposons à présent de deux fonctionnelles censées approcher l'énergie cinétique du système sans interaction. Il est en outre raisonnable d'estimer l'énergie liée à la répulsion interélectronique à l'aide de l'énergie de Coulomb

$$J(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(x)\,\rho(y)}{|x-y|}\,dx\,dy. \tag{1.28}$$

Dans le cas du modèle de Hartree-Fock (1.11), nous avons vu que l'énergie liée au Hamiltonien H_1 se composait de l'énergie cinétique, de l'énergie de Coulomb et d'un troisième terme dit d'échange, la différence entre l'énergie de Hartree-Fock et l'énergie fondamentale exacte étant appelée énergie de corrélation. Cette terminologie se retrouve dans les modèles dits de Kohn-Sham, dans lesquels les erreurs commises sur l'énergie cinétique et la répulsion électronique sont regroupées au sein d'une fonctionnelle appelée fonctionnelle d'échange-corrélation, ainsi définie par

$$E_{xc}(\rho) = F_{LL}(\rho) - T_{KS}(\rho) - J(\rho) \tag{1.29}$$

ou

$$E_{xc}(\rho) = F_L(\rho) - T_J(\rho) - J(\rho), \tag{1.30}$$

selon que l'on adopte la construction de Levy-Lieb ou de Lieb.

Le modèle de Kohn-Sham standard dérive de la formulation de Levy-Lieb et donc de (1.18), (1.19), (1.25), (1.28) et (1.29) :

$$\begin{aligned} E^{KS}(V) &= \inf\left\{\frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i|^2 + \int_{\mathbb{R}^3}\rho V + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy + E_{xc}(\rho), \right. \\ &\qquad \left. \phi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3}\phi_i\phi_j = \delta_{ij}\right\}, \end{aligned} \tag{1.31}$$

où $\rho = \sum_{i=1}^{N} |\phi_i|^2$.

Fondé sur les états mixtes, le modèle de Kohn-Sham étendu provient de (1.21), (1.22), (1.27), (1.28) et (1.30) :

$$E^{EKS}(V) = \inf\left\{\frac{1}{2}\sum_{i=1}^{+\infty} n_i \int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3}\rho V + \frac{1}{2}\iint_{\mathbb{R}^3\times\mathbb{R}^3}\frac{\rho(x)\rho(y)}{|x-y|}\,dx\,dy + E_{xc}(\rho),\right.$$
$$\left.\phi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\phi_i\phi_j = \delta_{ij},\ 0\le n_i\le 1,\ \sum_{i=1}^{+\infty} n_i = N\right\}, \tag{1.32}$$

où $\rho = \sum_{i=1}^{N}|\phi_i|^2$.

Dans le Chapitre 2, nous utiliserons une formulation alternative de (1.32), reposant sur les opérateurs densité d'ordre 1 associés aux états mixtes, définis par

$$\gamma(x,x') = \sum_{i=1}^{+\infty} n_i\phi_i(x)\,\phi_i(x'),\quad \phi_i\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\phi_i\phi_j = \delta_{ij},\ 0\le n_i\le 1,\ \sum_{i=1}^{+\infty} n_i = N.$$

Le modèle de Kohn-Sham étendu peut alors se réécrire

$$E^{EKS}(V) = \inf\left\{\mathrm{Tr}\,\left(-\frac{1}{2}\Delta\gamma\right) + \int_{\mathbb{R}^3}\rho_\gamma V + J(\rho_\gamma) + E_{\mathrm{xc}}(\rho_\gamma)\right.$$
$$\left.\gamma\in\mathcal{S}(L^2(\mathbb{R}^3)),\ 0\le\gamma\le 1,\ \mathrm{Tr}\,(\gamma) = N,\ \mathrm{Tr}\,(-\Delta\gamma)<\infty\right\}, \tag{1.33}$$

où $\rho_\gamma = \gamma(x,x)$, où $\mathcal{S}(L^2(\mathbb{R}^3))$ est l'ensemble des opérateurs auto-adjoints bornés sur $L^2(\mathbb{R}^3)$, et où Tr $(-\Delta\gamma)$ désigne la quantité Tr $(|\nabla|\gamma|\nabla|)$ qui a un sens dans $\mathbb{R}_+\cup\{+\infty\}$ dès que γ est un opérateur auto-adjoint positif. La notation Tr $(-\Delta\gamma)$ est justifiée par le fait que $|\nabla|^2 = -\Delta$.

Comme nous l'avons mentionné plus haut, la fonctionnelle d'échange-corrélation E_{xc} porte les erreurs d'approximation commises sur les autres composantes de l'énergie. C'est sur ce terme que se concentre l'effort de modélisation en calcul de structures électroniques, et c'est l'expression de ce terme qui différencie les modèles de type Kohn-Sham. Dans cette thèse, et plus précisément dans le Chapitre 2, nous nous intéressons aux deux fonctionnelles d'échange-corrélation les plus répandues.

La première est appelée LDA pour Local Density Approximation, et admet la forme générale

$$E_{xc}^{LDA}(\rho) = \int_{\mathbb{R}^3} g(\rho(x))dx. \tag{1.34}$$

La seconde fonctionnelle, appelée GGA pour Generalized Gradient Approximation, est une correction de la précédente faisant intervenir le gradient de la densité, soit

$$E_{xc}^{GGA}(\rho) = \int_{\mathbb{R}^3} h(\rho(x),\nabla\rho(x))dx. \tag{1.35}$$

La fonctionnelle LDA n'est utilisée en pratique qu'avec une seule définition de la fonction g, qui est celle obtenue pour un gaz uniforme d'électrons. Elle a été introduite par Kohn et Sham [47]. A l'inverse, il existe pour la fonctionnelle GGA beaucoup de choix différents de h dans la littérature (voir [48], [70], [12], [69]).

L'objectif du Chapitre 2 de cette thèse, écrit avec E. Cancès, est de montrer l'existence d'un minimiseur pour le modèle de Kohn-Sham étendu (1.33) avec fonctionnelles d'échange-corrélation LDA et GGA. A notre connaissance, le seul résultat relié disponible est la preuve de l'existence d'un minimiseur pour le modèle de Kohn-Sham standard (1.31) avec fonctionnelle LDA, établi par Le Bris dans [49].

Par souci de généralité, nous ne spécifions pas les fonctions g et h dans notre étude. Nous cherchons au contraire à déterminer les hypothèses les plus larges possibles sur g et h sous lesquelles les modèles admettent un minimiseur, afin de pouvoir évaluer le caractère bien posé des divers modèles existants, et de proposer un cadre mathématique rigoureux pour les modèles à venir.

Les résultats principaux du Chapitre 2 sont les théorèmes 2.2 et 2.3. Sous certaines conditions sur g et h, satisfaites en pratique, et pour des structures électroniques neutres ou chargées positivement, nous montrons que le modèle de Kohn-Sham étendu LDA admet un minimiseur, et que le modèle de Kohn-Sham étendu GGA pour les systèmes comprenant deux électrons admet un minimiseur. Dans ce dernier cas, l'hypothèse restrictive sur le nombre d'électrons est due au fait que notre analyse repose sur des résultats de régularité elliptique scalaires dont nous ne savons pas s'ils sont vérifiés pour les systèmes d'équations. Précisons que nous prenons en compte le spin dans cette étude, et que les deux électrons du système sont décrits par une même orbitale moléculaire $\phi \in H^1(\mathbb{R}^3)$.

Les difficultés mathématiques rencontrées dans ce chapitre proviennent essentiellement de la nonlinéarité, de la non convexité et de la non compacité des modèles. L'argument central de nos preuves est le lemme de concentration-compacité de P.-L. Lions [59].

Les résultats du Chapitre 2 ont été publiés dans [5].

1.2 Homogénéisation

1.2.1 Problématique industrielle et homogénéisation

Les matériaux composites sont de nos jours présents partout dans l'industrie, et notamment l'industrie aéronautique. La prochaine génération d'avions civils disposera ainsi d'une voilure et d'un fuselage réalisés principalement à l'aide de ces matériaux. Rappelons que les composites sont des matériaux hétérogènes constitués de deux phases, à savoir une matrice et des inclusions. Lorsque ces deux phases sont arrangées d'une manière astucieuse, le matériau obtenu présente des avantages considérables en comparaison des structures métalliques classiquement utilisées, en terme de poids, de résistance à la fatigue, de robustesse, ... Les économies potentielles, notamment vis-à-vis de la consommation de carburant et de la capacité de transport, sont énormes. Ces choix technologiques ont néanmoins de

très fortes implications relativement aux méthodes de conception et de certification des appareils. Prenons l'exemple d'un avion foudroyé, ce qui constitue un évènement fréquent : alors que l'effet dit de cage de Faraday protège l'avion classique en aluminium, le risque d'endommagement de la structure composite est à prendre en compte.

Cette problématique s'inscrit dans un contexte multi-physique. En effet, bien que l'aspect mécanique et structurel prime dans le dimensionnement des matériaux, d'autres critères ne peuvent être négligés. Sur le plan des échanges thermiques, la généralisation de l'emploi des composites suppose une parfaite maîtrise des températures de service afin de ne pas dégrader la résine qui constitue les matrices organiques. Au niveau électromagnétique, l'interaction de la structure avec les installations de plus en plus complexes d'équipements électroniques, et en particulier la circulation des courants au sein des matériaux, doivent être contrôlées. Enfin, la détermination de la performance acoustique de l'avion et le calcul du bruit externe nécessitent des modélisations plus fines que celles en vigueur actuellement. L'enjeu à venir est d'être capable d'effectuer des arbitrages aboutissant à des solutions optimales vis-à-vis de l'ensemble de ces contraintes.

Chaque composite est conçu à partir d'un arrangement qui lui est propre afin d'en adapter les propriétés. A une matrice et une inclusion données ne sont donc pas associés un matériau mais une famille de matériaux. Caractériser expérimentalement chaque variante d'une même famille est trop coûteux. Il est donc nécessaire de se doter d'outils précis permettant de prédire les caractéristiques comportementales des matériaux en se basant sur un nombre réduit de tests expérimentaux.

Les méthodes de prédiction qui ont cours dans l'industrie aéronautique se fondent sur des approches souvent heuristiques dont le domaine de validité est restreint. Etant de surcroît difficilement adaptables, elles ne permettront pas de traiter aisément les nouvelles générations de matériaux, en particulier les nanomatériaux. En parallèle, il est envisagé d'utiliser de plus en plus massivement les outils de simulation numérique dans les contextes physiques mentionnés ci-dessus. Cependant, une approche basée sur une modélisation et une simulation numérique "exactes" des composites sans traitement préalable ne constitue pas une réponse adéquate. En effet, les hétérogénéités constitutives de ces matériaux ont lieu à une échelle ε beaucoup plus petite que la taille caractéristique du composite, que nous prenons ici égale à 1. En utilisant une méthode numérique standard telle que celle des éléments finis, le maillage devrait être au moins aussi fin que ε pour espérer reproduire convenablement le comportement du matériau. Le nombre de degrés de liberté serait alors de l'ordre de ε^{-d}, où d est la dimension de l'espace de travail, et induirait un coût de calcul trop élevé.

De manière schématique, le but de l'homogénéisation est de remédier à ce problème et de faciliter le traitement des matériaux hétérogènes en les remplaçant par des matériaux homogènes de comportement macroscopique équivalent. Cette définition générale recouvre un ensemble de techniques plus ou moins rigoureuses. Du point de vue mathématique qui sera le nôtre dans cette thèse, l'homogénéisation s'intéresse aux équations aux dérivées partielles dont les coefficients présentent des oscillations à l'échelle microscopique ε introduite ci-dessus.

Considérons ainsi l'équation elliptique modèle

$$-\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f \ \text{ dans } \mathcal{O} \subset \mathbb{R}^d, \tag{1.36}$$

où A_ε est un champ de tenseurs de $\mathbb{R}^{d\times d}$ indexé par le paramètre ε. Cette équation peut par exemple modéliser un problème de mécanique ou de thermique. Dans le premier cas A_ε est le tenseur d'élasticité du matériau, f un chargement et u_ε le déplacement. Dans le second cas A_ε donne la conductivité du matériau, f représente les sources de chaleur et u_ε est la température. L'homogénéisation consiste à prendre la limite $\varepsilon \to 0$ dans (1.36). D'une certaine façon, cela revient à regarder le matériau de très loin pour ne plus voir les hétérogénéités. L'objectif est de trouver un problème limite

$$-\mathrm{div}(A^* \nabla u_0) = f \ \text{ dans } \mathcal{O}, \tag{1.37}$$

où le tenseur A^* définit un matériau dit homogénéisé, et où u_0 est, en un sens à définir, la limite de u_ε. La petite échelle ε ayant disparu dans (1.37), il est beaucoup plus aisé de traiter (1.37) que (1.36) d'un point de vue numérique.

La justification mathématique du passage de (1.36) à (1.37) a été établie par Murat et Tartar dans les années 1970 dans le cadre de la théorie de la H-convergence [81]. Cette théorie généralise la G-convergence de Spagnolo [78], restreinte aux opérateurs symétriques. Nous ne rentrons volontairement pas dans les détails, ni ne précisons les hypothèses nécessaires à cette convergence, et préférons souligner un problème pratique : il n'existe, en général, pas d'expression explicite du tenseur homogénéisé A^*.

On peut cependant obtenir une expression pour A^* sous certaines conditions sur le matériau, par exemple si celui-ci est périodique ou aléatoire stationnaire. Si le premier cas décrit un matériau idéal, le second se prête aux applications industrielles car il permet de prendre en compte les incertitudes inhérentes au processus de fabrication. Dans cette thèse, et plus précisément au sein des Chapitres 3, 4 et 5, nous considérerons toujours des matériaux satisfaisant l'une de ces deux hypothèses.

Nous verrons que le calcul de A^* dans le cas aléatoire stationnaire le plus général est coûteux à mettre en œuvre. Pour proposer des approches efficaces d'un point de vue numérique, nous supposerons que nos matériaux aléatoires sont des perturbations de matériaux périodiques, autrement dit que la quantité d'incertitude présente dans le système est faible. De tels matériaux seront dénommés faiblement aléatoires. Ce faisant, nous espérons modéliser une certaine réalité industrielle.

Enfin, pour ne pas multiplier les difficultés, nous nous intéresserons uniquement à des équations elliptiques linéaires scalaires sous forme divergence telles que (1.36), ou, introduisant la variable temps, à leur équivalent parabolique.

Nous rappelons ci-après les bases de l'homogénéisation dans les contextes périodique et aléatoire stationnaire, puis introduisons les problématiques qui feront l'objet des Chapitres 3, 4 et 5.

1.2.2 Homogénéisation périodique

Soit A un champ de tenseurs de $\mathbb{R}^d$ à valeurs dans $\mathbb{R}^{d\times d}$ tel qu'il existe λ et Λ strictement positifs tels que

$$\forall \xi \in \mathbb{R}^d,\ \text{p.p.t } x \in \mathbb{R}^d,\ \lambda|\xi|^2 \leq A(x)\xi\cdot\xi \ \text{ et } \ |A(x)\xi| \leq \Lambda|\xi|. \tag{1.38}$$

On suppose de plus que A est $\mathbb{Z}^d$-périodique, ce qui signifie que

$$\forall k \in \mathbb{Z}^d,\ \ A(x+k) = A(x)\ \ \text{p.p.t } x \in \mathbb{R}^d.$$

Les hétérogénéités du matériau auquel on s'intéresse ont lieu à l'échelle $\varepsilon > 0$. On définit par conséquent le tenseur A_ε par

$$A_\varepsilon(x) = A\left(\frac{x}{\varepsilon}\right). \tag{1.39}$$

Dans la suite, on appellera variable macroscopique ou lente la variable x, et variable microscopique ou rapide la variable $y = \frac{x}{\varepsilon}$. Ces dénominations sont dues au fait qu'une variation d'ordre 1 sur x entraîne une variation d'ordre $\frac{1}{\varepsilon}$ sur y.

Considérons à présent le problème modèle suivant :

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \quad \text{dans } \mathcal{O}, \\ u_\varepsilon = 0 \ \text{ sur } \partial\mathcal{O}, \end{cases} \tag{1.40}$$

où $\mathcal{O}$ est un ouvert borné de $\mathbb{R}^d$ et f est une fonction de $L^2(\mathcal{O})$.

Le problème (1.40) admet, pour tout $\varepsilon > 0$, une unique solution u_ε dans $H^1_0(\mathcal{O})$. Comme expliqué dans la Section 1.2.1, il est coûteux d'attaquer directement (1.40) numériquement, du fait de la présence de l'échelle microscopique ε qui nécessite un maillage fin. L'homogénéisation consiste en une analyse asymptotique du problème lorsque ε tend vers 0.

En dimension un, on peut aisément résoudre (1.40) explicitement, et obtenir l'expression du problème limite quand $\varepsilon \to 0$. Pour les dimensions supérieures, l'analyse est plus compliquée. Plusieurs techniques existent, comme la méthode de la fonction test oscillante due à Murat et Tartar [81] ou la convergence à deux échelles introduite par Nguetseng et développée par Allaire [1, 65]. Il est possible de retrouver le résultat formellement à l'aide d'un développement à deux échelles. C'est cette dernière approche que nous choisissons. Elle repose sur l'hypothèse, traditionnellement appelée Ansatz, que u_ε s'écrit comme une série entière en ε :

$$u_\varepsilon(x) = u_0(x, \frac{x}{\varepsilon}) + \varepsilon u_1(x, \frac{x}{\varepsilon}) + \varepsilon^2 u_2(x, \frac{x}{\varepsilon}) + ..., \tag{1.41}$$

où pour tout $k \in \mathbb{N}$, la fonction $u_k(x, y)$ est $\mathbb{Z}^d$-périodique en la variable rapide y.

On injecte ensuite (1.41) dans (1.40), en utilisant la règle de dérivation composée

$$\nabla(v(x, \frac{x}{\varepsilon})) = \nabla_x v(x, \frac{x}{\varepsilon}) + \frac{1}{\varepsilon}\nabla_y v(x, \frac{x}{\varepsilon}), \tag{1.42}$$

puis on regroupe les coefficients des différentes puissances de ε. Cela donne un système infini d'équations vérifiées par les fonctions u_k, que l'on résout successivement en supposant que les variables x et $y = \frac{x}{\varepsilon}$ sont indépendantes. La variable x joue le rôle d'un paramètre, et la résolution se fait en y.

Introduisant la cellule unité $Q = [0,1]^d$, l'équation obtenue pour u_0 (correspondant à l'ordre ε^{-2}) s'écrit

$$\begin{cases} -\operatorname{div}_y\left(A(y)\nabla_y u_0(x,y)\right) = 0 \quad \text{dans } Q, \\ y \mapsto u_0(x,y) \;\; \mathbb{Z}^d - \text{périodique.} \end{cases} \tag{1.43}$$

L'équation de (1.43) est en fait posée dans l'espace $\mathbb{R}^d$ tout entier, et le problème est résolu dans l'espace des fonctions de $H^1_{loc}(\mathbb{R}^d)$ qui sont $\mathbb{Z}^d$-périodiques. Nous adopterons par convention la notation (1.43) dans toute la thèse pour souligner que le problème se réduit à un problème posé sur Q.

On déduit aisément de (1.43), par unicité de la solution à une constante additive (fonction de x) près, que u_0 ne dépend pas de la variable microscopique y. Cela est en accord avec l'interprétation de u_0 comme champ homogénéisé n'admettant de variations qu'à l'échelle macroscopique. On écrira donc $u_0(x)$.

La seconde équation, provenant de l'ordre ε^{-1}, relie u_0 et u_1 via

$$\begin{cases} -\operatorname{div}_y\left(A(y)\nabla_y u_1(x,y)\right) = \operatorname{div}_y\left(A(y)\nabla_x u_0(x)\right) \quad \text{dans } Q, \\ y \mapsto u_1(x,y) \;\; \mathbb{Z}^d - \text{périodique.} \end{cases} \tag{1.44}$$

Le second membre de (1.44) peut se réécrire

$$\operatorname{div}_y\left(A(y)\nabla_x u_0(x)\right) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x) \operatorname{div}_y\left(A(y)e_i\right), \tag{1.45}$$

où pour tout $i \in [\![1,d]\!]$, e_i est le i-ème vecteur canonique de $\mathbb{R}^d$.

Utilisant (1.45), la linéarité de (1.44) et l'unicité de la solution de (1.44) à une fonction de x près, il vient que u_1 s'exprime en fonction de u_0 par

$$u_1(x,y) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x) w_i(y) + \tilde{u}_1(x), \tag{1.46}$$

où $\tilde{u}_1(x)$ correspond à l'indétermination en la variable x, et pour tout $i \in [\![1,d]\!]$, $w_i(y)$ est solution du problème dit de cellule

$$\begin{cases} -\operatorname{div}\left(A(y)\nabla w_i(y)\right) = \operatorname{div}(A(y)e_i) \quad \text{dans } Q, \\ w_i \;\; \mathbb{Z}^d - \text{périodique.} \end{cases} \tag{1.47}$$

Les fonctions w_i sont définies à une constante additive près (au vu de (1.46), ces constantes peuvent être intégrées à $\tilde{u}_1$). Intuitivement, le rôle des problèmes de cellule

(dont le nombre est égal à la dimension de l'espace de travail) est de récupérer l'information sur la microstructure, afin de la propager à l'échelle macroscopique.

La dernière équation que nous utiliserons, associée à l'ordre ε^0, s'écrit

$$\begin{cases} -\operatorname{div}_y\left(A(y)\nabla_y u_2(x,y)\right) = \operatorname{div}_x\left(A(y)\nabla_x u_0(x)\right) + \operatorname{div}_x\left(A(y)\nabla_y u_1(x,y)\right) \\ \qquad\qquad + \operatorname{div}_y\left(A(y)\nabla_x u_1(x,y)\right) + f(x) \quad \text{dans } Q, \\ y \mapsto u_2(x,y) \ \ \mathbb{Z}^d - \text{périodique.} \end{cases} \tag{1.48}$$

Le théorème de Lax-Milgram appliqué au problème aux limites (1.48) avec conditions de périodicité montre qu'il y a existence et unicité (à fonction de x près) de la solution u_2 si et seulement si l'intégrale du second membre sur la cellule de périodicité Q est nulle. Notons que cette condition de compatibilité est trivialement vérifiée pour les problèmes précédents (1.43) et (1.44). Elle prend ici la forme

$$\begin{aligned} -\int_Q \left(\operatorname{div}_x\left(A(y)\nabla_x u_0(x)\right) + \operatorname{div}_x\left(A(y)\nabla_y u_1(x,y)\right) + \operatorname{div}_y\left(A(y)\nabla_x u_1(x,y)\right)\right) dy \\ = \int_Q f(x)dy, \end{aligned} \tag{1.49}$$

et quelques simplifications élémentaires mènent à

$$-\operatorname{div}_x \int_Q A(y)\left(\nabla_x u_0(x) + \nabla_y u_1(x,y)\right) dy = f(x). \tag{1.50}$$

Injectant (1.46) dans (1.50), on obtient l'équation suivante sur u_0 :

$$-\operatorname{div}(A^*\nabla u_0) = f, \tag{1.51}$$

où le tenseur A^* est constant et défini par

$$\forall i \in [\![1,d]\!], \ \ A^* e_i = \int_Q A(y)(\nabla w_i(y) + e_i)dy. \tag{1.52}$$

La condition aux limites du problème initial (1.40) doit également être vérifiée par l'approximation d'ordre zéro qu'est u_0. Par conséquent, u_0 est solution du problème dit homogénéisé

$$\begin{cases} -\operatorname{div}\left(A^*\nabla u_0\right) = f \quad \text{dans } \mathcal{O}, \\ u_0 = 0 \ \ \text{sur } \partial\mathcal{O}. \end{cases} \tag{1.53}$$

Comme mentionné précédemment, ces manipulations formelles admettent une justification rigoureuse. La pertinence de (1.53) comme approximation de (1.40) repose sur le résultat suivant :

$$u_\varepsilon \to u_0 \ \ \text{dans } L^2(\mathcal{O}). \tag{1.54}$$

La convergence de u_ε vers u_0 a en fait aussi lieu faiblement dans $H^1(\mathcal{O})$. Pour obtenir une convergence forte dans cet espace, il est nécessaire d'ajouter u_1 défini par (1.46) :

$$u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) \to 0 \ \ \text{dans } H^1(\mathcal{O}). \tag{1.55}$$

Rappelons que u_1 est défini à une fonction $\tilde{u}_1$ de x près, et que les solutions w_i des problèmes de cellule qui constituent u_1 sont elles-mêmes définies à une constante additive près. Ces indéterminations sur u_1 ne jouent aucun rôle dans la convergence (1.55) puisque leur norme dans $H^1(\mathcal{O})$ est d'ordre ε. Elles n'ont bien sûr également aucune influence sur la définition (1.52) de A^*. A ce stade seul le gradient de u_1 par rapport à y a été utilisé.

La fonction u_1 corrige l'approximation du gradient de u_ε par le gradient de u_0. Elle est pour cette raison appelée correcteur d'ordre 1. Par extension, les solutions w_i des problèmes de cellule (1.47) sont parfois également appelées correcteurs. Plus généralement, la fonction u_k de l'Ansatz (1.41) est dénommée correcteur d'ordre k.

Poursuivant la résolution du système d'équations provenant du remplacement de u_ε par la série (1.41) dans (1.40), il est possible de déterminer successivement tous les correcteurs u_k par des arguments similaires à ceux exposés plus haut. Cependant, comme nous le verrons dans la Section 1.2.5 ci-dessous, on ne calcule pas en pratique les correcteurs d'ordre élévé car d'autres termes, dits de couche limite, interviennent dès l'ordre un en ε dans les estimations d'erreur. Dans cette thèse, nous n'irons pas au delà du correcteur d'ordre 2 u_2.

Nous disposons à présent, via (1.54) et (1.55), d'un moyen d'approcher u_ε dans $L^2(\mathcal{O})$ et $H^1(\mathcal{O})$ à l'aide du champ homogénéisé u_0 et du premier correcteur u_1. L'intérêt de la méthode réside dans le fait que d'un point de vue numérique, le calcul de u_0 et u_1 est beaucoup plus simple que la résolution directe du problème initial (1.40). La première étape consiste à calculer le tenseur homogénéisé A^* par la formule (1.52), ce qui nécessite de résoudre les d problèmes de cellule (1.47) posés sur la cellule unité Q. Une fois A^* déterminé, u_0 est donné par la résolution de (1.53) sur $\mathcal{O}$, et u_1 s'obtient "gratuitement" grâce à (1.46) où l'on peut choisir $\tilde{u}_1 = 0$. On doit donc résoudre en tout $d+1$ problèmes aux limites dans lesquels l'échelle microscopique ε a disparu, et qui par conséquent ne requièrent pas l'utilisation d'un maillage fin. Le coût de calcul s'en trouve considérablement réduit.

Le contexte périodique est l'exemple d'homogénéisation le plus simple à mettre en œuvre. Il ne correspond cependant pas à des matériaux concrets, mais au contraire idéalisés. Pour tendre vers plus de généralité, nous présentons dans la section suivante la procédure d'homogénéisation dans un cadre stochastique, qui permet à nouveau d'obtenir une expression explicite pour le tenseur A^*.

1.2.3 Homogénéisation stochastique

Nous introduisons un espace de probabilité $(\Omega, \mathcal{F}, \mathbb{P})$, où $\mathcal{F}$ est une tribu et $\mathbb{P}$ une mesure de probabilité. Le singleton $\omega \in \Omega$ désigne un évènement (ici une réalisation du matériau), et $\mathbb{E}(X)$ l'espérance de la variable aléatoire X.

L'hypothèse cruciale qui fait de l'homogénéisation une méthode pratique dans le cadre stochastique et qui, en quelque sorte, généralise l'hypothèse de périodicité précédente, est la stationnarité. Dans la littérature, la notion de stationnarité principalement rencontrée est *continue*, et implique en particulier que la loi du matériau en deux points x et $x+h$ est

la même pour tout $h \in \mathbb{R}^d$. Dans ce qui suit, et dans les Chapitres 3 et 4, nous emploierons une stationnarité quelque peu différente dite *discrète* : la loi du matériau est la même en deux points x et $x+k$ pour tout $k \in \mathbb{Z}^d$. Ceci nous permet de considérer des matériaux aléatoires ayant une structure périodique sous-jacente.

Cette stationnarité discrète est formalisée de la manière suivante. On suppose que le groupe $(\mathbb{Z}^d, +)$ agit sur Ω, et que cette action, notée τ_k pour $k \in \mathbb{Z}^d$, préserve la mesure $\mathbb{P}$ au sens où

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

On dit alors qu'une fonction $F \in L^1_{loc}(\mathbb{R}^d, L^1(\Omega))$ est stationnaire si

$$\forall k \in \mathbb{Z}^d,\ F(x+k,\omega) = F(x, \tau_k \omega) \ \text{ p.p.t } x \in \mathbb{R}^d \text{ et } \omega \in \Omega. \tag{1.56}$$

On supposera de plus que l'action de groupe est ergodique, soit

$$\forall \mathcal{A} \in \mathcal{F},\ (\forall k \in \mathbb{Z}^d, \mathcal{A} = \tau_k \mathcal{A}) \implies (\mathbb{P}(\mathcal{A}) = 0 \text{ ou } \mathbb{P}(\mathcal{A}) = 1).$$

Intuitivement, l'ergodicité signifie que considérer une réalisation du matériau en tous les points de l'espace revient à considérer toutes les réalisations en un point donné. Sans cette hypothèse, le tenseur homogénéisé A^* obtenu in fine est aléatoire, ce qui complique considérablement la mise en œuvre pratique.

Le cadre de travail aléatoire ayant été précisé, nous considérons un champ de tenseurs A stationnaire, tel que (1.38) est presque sûrement satisfait par $A(\cdot, \omega)$, et introduisons la petite échelle ε en définissant le tenseur A_ε par

$$A_\varepsilon(x,\omega) = A\left(\frac{x}{\varepsilon}, \omega\right). \tag{1.57}$$

Le problème aux limites canonique que nous regardons est l'équivalent aléatoire de (1.40), c'est-à-dire

$$\begin{cases} -\operatorname{div}\left(A_\varepsilon(x,\omega)\nabla u_\varepsilon(x,\omega)\right) = f \quad \text{presque sûrement dans } \mathcal{O}, \\ u_\varepsilon = 0 \ \text{ presque sûrement sur } \partial\mathcal{O}, \end{cases} \tag{1.58}$$

où $\mathcal{O}$ est un ouvert borné de $\mathbb{R}^d$ et f est une fonction de $L^2(\mathcal{O})$.

L'homogénéisation de (1.58) ressemble formellement à celle de (1.40). Les résultats standard d'homogénéisation stochastique [46] impliquent que dans la limite $\varepsilon \to 0$, le problème homogénéisé admet la forme (1.53), où la matrice homogénéisée est à présent définie par

$$\forall i \in [\![1,d]\!],\ \ A^* e_i = \mathbb{E}\left(\int_Q A(y,\omega)(\nabla w_i(y,\omega) + e_i) dy\right), \tag{1.59}$$

et pour tout $i \in [\![1,d]\!]$, w_i est la solution du problème de cellule stochastique

$$\begin{cases} -\operatorname{div}\left(A(y,\omega)(\nabla w_i(y,\omega) + e_i)\right) = 0 \quad \text{presque sûrement dans } \mathbb{R}^d, \\ \nabla w_i \ \text{ stationnaire}, \quad \mathbb{E}\left(\displaystyle\int_Q \nabla w_i\right) = 0. \end{cases} \tag{1.60}$$

Ce problème admet une solution unique à constante près dans l'espace

$$\{w \in L^2_{loc}(\mathbb{R}^d, L^2(\Omega)), \quad \nabla w \in L^2_{unif}(\mathbb{R}^d, L^2(\Omega))\}.$$

La notation L^2_{unif} désigne l'espace des fonctions dont la norme L^2 sur une boule de rayon 1 est bornée indépendamment du centre de cette boule.

Le champ homogénéisé u_0 est déterministe, solution de (1.53) avec A^* donné par (1.59). Le premier correcteur u_1 est stochastique et défini par l'équivalent de (1.46) où les w_i sont à présent solutions de (1.60) et $\tilde{u}_1$ est stochastique. Les convergences (1.54) et (1.55) ont désormais lieu presque sûrement en ω.

Remarquons que pour des fonctions déterministes, la condition de stationnarité (1.56) se réduit à la $\mathbb{Z}^d$-périodicité de la section précédente : l'homogénéisation périodique se déduit donc immédiatement de l'homogénéisation stochastique. Si les deux contextes donnent des formules explicites pour le tenseur homogénéisé A^*, l'implémentation numérique dans le cas stochastique est beaucoup plus compliquée. En effet, contrairement aux problèmes de cellule périodiques (1.47) réductibles à des problèmes posés sur la cellule unité Q, les problèmes de cellule aléatoires (1.60) doivent être résolus en principe sur l'espace $\mathbb{R}^d$ tout entier. En pratique, comme nous le verrons dans les Chapitres 3 et 4, et comme expliqué dans [22], on les résout sur un domaine de grande taille et pour plusieurs réalisations du matériau. Le tenseur A^* est alors obtenu en prenant la moyenne de ces réalisations, et en faisant tendre la taille du domaine vers l'infini. Le coût de calcul lié à une telle approche est très élevé.

La question de l'intérêt de l'homogénéisation par rapport à une résolution directe de (1.58) se pose donc de manière légitime dans ce cadre aléatoire. Nous contournerons ce problème en nous intéressant à des matériaux pour lesquels la part d'aléatoire est faible, ce qui facilite grandement le traitement numérique. Nous présentons ce point de vue dans la section suivante.

1.2.4 Matériaux faiblement aléatoires

Cette partie, développée dans les Chapitres 3 et 4 écrits avec C. Le Bris, repose sur l'hypothèse que les matériaux composites utilisés en pratique ne sont pas totalement désordonnés, et qu'il existe une structure déterministe sous-jacente, laquelle est modifiée par les incertitudes et aléas du processus de fabrication. Autrement dit, nos matériaux aléatoires consistent en des perturbations stochastiques de matériaux déterministes.

Nos travaux s'inscrivent dans la continuité de nombreuses approches perturbatives proposées dans la littérature consacrée à l'homogénéisation, dans des cadres déterministes ou stochastiques. Le prototype de telles approches est de considérer un matériau dont les propriétés sont données par un tenseur A_η de la forme

$$A_\eta = A + \eta C, \tag{1.61}$$

où A et C sont deux tenseurs, et $\eta > 0$ est un petit paramètre interprété comme l'amplitude de la perturbation. On peut alors, suivant une démarche quelque peu similaire à celle de

l'Ansatz (1.41), chercher toutes les quantités d'intérêt, et notamment les correcteurs, sous la forme d'une série entière en η. Injectant l'expression (1.61) dans les problèmes de cellule et identifiant les coefficients des puissances de η, on résout successivement les différents ordres en η, le but étant in fine d'exprimer le tenseur homogénéisé A^*_η comme

$$A^*_\eta = A^* + \eta f_1(A, C) + \eta^2 f_2(A, C) + ..., \tag{1.62}$$

où A^* est le tenseur homogénéisé associé à A et pour $k \in \mathbb{N}^*$, f_k est une fonction de A et C. Bien sûr, une telle approche ne se justifie que si le calcul des premiers ordres en η est plus simple que le calcul direct de A^*_η.

Un exemple de cette démarche dans un contexte déterministe non nécessairement périodique est donné dans [80] sous l'appellation "small amplitude homogenization". Mentionnons de plus les travaux exposés dans [18], dont la philosophie, bien que dépassant le cadre perturbatif, se rapproche de la nôtre. Les auteurs y supposent qu'une structure de référence périodique est déformée par l'action d'un difféomorphisme aléatoire. Plus précisément, ils étudient l'homogénéisation de

$$A(\Phi^{-1}(\frac{x}{\varepsilon}, \omega)), \tag{1.63}$$

où A est un tenseur déterministe $\mathbb{Z}^d$-périodique, et presque sûrement en ω, $\Phi(\cdot, \omega)$ est un difféomorphisme de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Soulignons que le gradient de Φ est supposé stationnaire mais pas la fonction Φ elle-même. Il s'ensuit que le tenseur défini par (1.63) n'est pas stationnaire, et donc que ce modèle ne satisfait pas les hypothèses de la section précédente et n'en constitue pas une application. Les auteurs obtiennent une formule explicite pour le tenseur homogénéisé, dont la mise en œuvre présente les mêmes difficultés pratiques que dans le cadre stationnaire usuel. Ils considèrent alors le cas particulier où Φ est une perturbation de l'identité, soit

$$\Phi(x, \omega) = x + \eta\Psi(x, \omega) + \mathcal{O}(\eta^2), \tag{1.64}$$

et dérivent une formule du type (1.62).

Le point commun aux deux approches ci-dessus est que lorsque η est petit, l'impact de la perturbation sur la structure du matériau est faible. En effet, la perturbation tend vers zéro en norme L^∞ quand η tend vers zéro. Nous souhaitons nous affranchir de cette contrainte, et considérons un tenseur A_η donné par

$$A_\eta = A + b_\eta C, \tag{1.65}$$

où A et C sont deux tenseurs déterministes périodiques, et b_η est un champ scalaire aléatoire stationnaire petit “en moyenne”, mais dont la réalisation peut grandement modifier la structure locale du matériau périodique de référence représenté par A. De manière intuitive, l'idée est de perturber le matériau périodique seulement rarement, mais en contrepartie éventuellement fortement. Les hypothèses de ce modèle sont détaillées dans les Chapitres 3 et 4. Le but est d'obtenir, pour le tenseur homogénéisé A^*_η, une expression de la forme

$$A^*_\eta = A^* + \eta\bar{A}^*_1 + \eta^2\bar{A}^*_2 + o(\eta^2), \tag{1.66}$$

où les corrections $\bar{A}_1^*$ et $\bar{A}_2^*$ sont calculées de manière purement *déterministe* en utilisant des informations statistiques basiques sur b_η (moyenne, variance, corrélation spatiale, ...). Le calcul des coefficients du développement asymptotique (1.66) est alors plus rapide que le calcul direct de A_η^* par la formule (1.59).

Le Chapitre 3 étudie le cas particulier d'une perturbation b_η suivant une loi de Bernoulli, c'est-à-dire prenant uniquement les valeurs 0 et 1. Celle-ci se prête bien à une interprétation du modèle (1.65) comme modèle de défauts. On peut par exemple penser à un composite dont les inclusions sont enlevées de manière aléatoire. Des liens clairs avec les théories de défauts classiques en physique des solides apparaissent. L'approche adoptée dans ce chapitre est heuristique et n'a pas pu être justifiée dans son intégralité, sauf en dimension un où les calculs sont explicites. Ainsi, si des expressions explicites sont obtenues pour les corrections $\bar{A}_1^*$ et $\bar{A}_2^*$ dans (1.66), la validité du développement asymptotique reste un problème ouvert pour nous. Des tests numériques prouvent néanmoins la pertinence et l'efficacité pratique de la méthode.

Le Chapitre 4 généralise les résultats du Chapitre 3 à d'autres lois. Cette extension repose sur un développement de la mesure image de b_η par rapport à η (i.e un développement de la loi de b_η). Par ailleurs, nous proposons dans ce chapitre une approche alternative entièrement rigoureuse, mais dont le domaine d'application nous paraît moins large. Une nouvelle fois, des tests numériques exhaustifs viennent confirmer l'intérêt de ces approches.

Notre modèle perturbatif a fait l'objet d'une publication dans [8]. Les travaux contenus dans le Chapitre 3 ont été soumis pour publication dans *SIAM Multiscale Modeling & Simulation* [6], ceux du Chapitre 4 dans *Communications in Computational Physics* [7].

1.2.5 Couches limites en homogénéisation périodique

Nous abordons dans cette section le problème des couches limites en homogénéisation, qui constitue le sujet du cinquième et dernier chapitre de cette thèse, écrit avec G. Allaire.

Notre intérêt pour cette question trouve une origine pratique dans un dispositif expérimental appelé thermographie infrarouge stimulée (TIS), utilisé pour le contrôle non destructif et la caractérisation des matériaux et structures aéronautiques. La TIS consiste à chauffer rapidement la surface d'un matériau au moyen, par exemple, de lampes flashs, et à mesurer l'élévation de température résultante à l'aide d'une caméra infrarouge. L'analyse du signal obtenu fournit une cartographie thermique du matériau. Les applications en sont diverses : les informations récupérées permettent de détecter des défauts à l'intérieur du matériau, de déterminer certaines propriétés telles que des conductivités thermiques ou des coefficients d'échange entre deux phases hétérogènes, etc.

Simuler numériquement ce procédé dans le cas du contrôle de matériaux composites requiert de disposer de modèles reproduisant fidèlement le comportement du composite à la fois en surface (là où la mesure se fait) et en régime transitoire avant relaxation (la stimulation thermique et la mesure ayant lieu sur une échelle de temps très petite). L'objectif du Chapitre 5 est d'apporter une réponse à cette double exigence pour des matériaux

périodiques, c'est-à-dire dans le contexte de la Section 1.2.2.

Il est naturel de dissocier les difficultés, et de considérer d'abord les phénomènes de frontière en régime stationnaire (ce dernier terme étant pris dans une acception différente de la section précédente, et signifiant "indépendant du temps"). Revenons donc au problème canonique (1.40). Nous avons vu que l'intérêt de l'homogénéisation périodique était de remplacer le calcul coûteux de u_ε par celui, beaucoup plus simple, du champ homogénéisé u_0 et éventuellement des correcteurs. Nous nous intéressons à présent à la qualité d'approximation de u_ε par u_0 et les correcteurs sur le bord $\partial\mathcal{O}$ du domaine, que nous supposerons désormais suffisamment régulier.

A cette fin, la convergence dans $L^2(\mathcal{O})$ donnée par (1.54) est clairement trop faible. En revanche, le théorème de trace dans $H^1(\mathcal{O})$, soit l'injection continue de cet espace dans $L^2(\partial\mathcal{O})$ (et même $H^{1/2}(\mathcal{O})$), implique que la convergence (1.55) dans $H^1(\mathcal{O})$ permet de contrôler l'erreur d'approximation dans $L^2(\partial\mathcal{O})$.

Le résultat fondamental qui précise l'erreur dans (1.55), est le suivant [14] :

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) \right\|_{H^1(\mathcal{O})} \leq C\sqrt{\varepsilon}. \tag{1.67}$$

Le taux de convergence en $\sqrt{\varepsilon}$ donné par (1.67) est optimal. Il est contre-intuitif, car d'après l'Ansatz (1.41), on peut s'attendre à obtenir, formellement,

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) \right\|_{H^1(\mathcal{O})} \simeq \left\| \varepsilon^2 u_2(x, \frac{x}{\varepsilon}) \right\|_{H^1(\mathcal{O})} \leq C\varepsilon. \tag{1.68}$$

L'apparente sous-optimalité de (1.67) par rapport à (1.68) est due au fait que le développement asymptotique (1.41) n'est pas vrai près du bord $\partial\mathcal{O}$. En effet, les correcteurs u_k pour $k \in \mathbb{N}^*$ ne vérifient pas la condition aux limites de Dirichlet de (1.40). A l'ordre un en ε, l'approximation $u_0(x) + \varepsilon u_1(x, \frac{x}{\varepsilon})$ est ainsi égale à $\varepsilon u_1(x, \frac{x}{\varepsilon})$ sur $\partial\mathcal{O}$. Les oscillations du correcteur sur la frontière expliquent la perte d'un facteur $\sqrt{\varepsilon}$ entre (1.68) et (1.67) [2, 14].

Pour améliorer (1.67), il est nécessaire d'ajouter des termes supplémentaires à l'Ansatz afin de le corriger sur la frontière. Ces termes, qui ne "vivent" que près du bord, sont appelés couches limites. Ils n'admettent pas une définition unique, la seule contrainte étant de compenser le correcteur sur $\partial\mathcal{O}$. Néanmoins, une contrainte pratique, et donc un critère discriminant, est que leur calcul et leur implémentation doivent être simples pour ne pas compromettre l'intérêt global de l'approche par homogénéisation par rapport à la résolution directe de (1.40).

Afin de satisfaire à ces exigences pratiques, les nombreux travaux sur les couches limites pour les problèmes elliptiques tels que (1.40) existant dans la littérature supposent une géométrie particulière pour le domaine $\mathcal{O}$: demi-espace dont la frontière intersecte les axes de périodicité avec une pente rationnelle [10, 11, 15, 45, 56], semi-bande ayant la même propriété [66], domaine rectangulaire [2] ou plus récemment polygonal quelconque [37], ou encore domaine dont la frontière est une courbe régulière dans le cas spécifique

d'un milieu stratifé [64].

Nous nous appuyons principalement, dans le Chapitre 5, sur les techniques utilisées dans [2] et [64], qui étudient l'influence au premier ordre des couches limites pour des problèmes d'homogénéisation périodique, et sont en particulier consacrés aux estimations d'erreur du type (1.67). Par simplicité, nous nous restreignons comme dans [2] aux domaines rectangulaires. Les frontières sont alors planes et les couches limites peuvent êtres calculées très simplement numériquement. Tous les travaux mentionnés précédemment correspondant à des conditions aux limites de Dirichlet, nous étudions, dans une première partie, le cas de conditions de Neumann. Nous n'avons pas trouvé de références à ce sujet, mis à part [62] dont l'approche consistant à transformer les conditions de Neumann en conditions de Dirichlet par dualité n'est pas celle que nous souhaitons adopter. L'adaptation de Dirichlet à Neumann se révèle aisée et les techniques employées similaires à celles de [2] et [64].

Notre véritable motivation réside cependant, comme annoncé au début de cette section, dans l'étude des régimes transitoires, et donc des équations paraboliques du type équation de la chaleur dans le cadre de l'homogénéisation périodique. Ceci fait l'objet de la seconde partie du Chapitre 5. L'introduction de la variable temps ajoute en quelque sorte une nouvelle frontière $t = 0$. De même que l'Ansatz (1.41) (généralisé classiquement au contexte parabolique) ne satisfait pas la condition aux limites, il ne vérifie pas la condition initiale. Formellement, le problème est donc identique à celui des couches limites : on doit rajouter un terme corrigeant le développement asymptotique (1.41) à $t = 0$, et ne donnant pas lieu à un surcoût de calcul trop élevé, pour obtenir des estimations d'erreur dans un espace adéquat, en l'occurrence $C([0,T];H^1(\mathcal{O}))$ pour $T > 0$ (voir [23]).

A notre connaissance, la seule étude concernant ce terme est [68]. L'auteur y considère un problème posé sur tout l'espace $\mathbb{R}^d$; l'absence de frontières permet alors de se concentrer uniquement sur la condition initiale, et de proposer une correction que nous appellerons *couche initiale*. A la différence de [68], nous nous intéressons à un problème d'homogénéisation parabolique posé sur un domaine borné. Nous souhaitons utiliser les résultats précités sur les couches limites en régime stationnaire et sur la couche initiale sans frontières pour obtenir des estimations d'erreur dans le cas général. Ceci requiert de comprendre l'interaction entre les couches limites et la couche initiale. Notre résultat principal est le théorème 5.19 qui fournit une estimation d'erreur dans $C([0,T];H^1(\mathcal{O}))$. Malheureusement, ce théorème repose sur des hypothèses de régularité que nous n'avons pu vérifier. En conséquence, nous ne pouvons affirmer qu'il apporte une réponse pertinente. Néanmoins, nous pensons que les travaux contenus dans le Chapitre 5 offrent un panorama exhaustif des difficultés liées aux couches limites et initiale en homogénéisation parabolique, et espérons qu'ils constituent un premier pas vers une compréhension plus fine.

Chapitre 2

Existence of minimizers for Kohn-Sham models in Quantum Chemistry

Sommaire

2.1 Introduction 23
2.2 Mathematical foundations of DFT and Kohn-Sham models . . . 24
2.2.1 Density Functional Theory 24
2.2.2 Kohn-Sham models 27
2.3 Main results 32
2.4 Proofs 35
2.4.1 Preliminary results 36
2.4.2 Proof of Lemma 2.1 37
2.4.3 Proof of Theorem 2.2 40
2.4.4 Proof of Theorem 2.3 45

2.1 Introduction

Density Functional Theory (DFT) is a powerful, widely used method for computing approximations of ground state electronic energies and densities in chemistry, materials science, biology and nanosciences.

According to DFT [43, 53], the electronic ground state energy and density of a given molecular system can be obtained by solving a minimization problem of the form

$$\inf\left\{F(\rho)+\int_{\mathbb{R}^3}\rho V,\ \rho\geq 0,\ \sqrt{\rho}\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\rho=N\right\}$$

where N is the number of electrons in the system, V the electrostatic potential generated by the nuclei, and F some functional of the electronic density ρ, the functional F being universal, in the sense that it does not depend on the molecular system under consideration. Unfortunately, no tractable expression for F is known, which could be used in numerical simulations.

The groundbreaking contribution which turned DFT into a useful tool to perform calculations, is due to Kohn and Sham [47], who introduced the local density approximation (LDA) to DFT. The resulting Kohn-Sham LDA model is still commonly used, in particular in solid state physics. Improvements of this model have then been proposed by many authors, giving rise to Kohn-Sham GGA models [48, 70, 12, 69], GGA being the abbreviation of generalized gradient approximation. While there is basically a unique Kohn-Sham LDA model, there are several Kohn-Sham GGA models, corresponding to different approximations of the so-called exchange-correlation functional. A given GGA model will be known to perform well for some classes of molecular system, and poorly for some other classes. In some cases, the best result will be obtained with LDA. It is to be noticed that each Kohn-Sham model exists in two versions: the standard version, with integer occupation numbers, and the extended version with "fractional" occupation numbers. As explained below, the former one originates from Levy-Lieb's (pure state) contruction of the density functional, while the latter is derived from Lieb's (mixed state) construction.

There are three main mathematical difficulties encountered when studying these models from a theoretical point of view: the nonlinearity, the nonconvexity, and the possible loss of compactness at infinity of the models. To our knowledge, very few results on Kohn-Sham LDA and GGA models exist in the mathematical literature. In fact, we are only aware of a proof of existence of a minimizer for the standard Kohn-Sham LDA model by Le Bris [49]. In this contribution, we prove the existence of a minimizer for the extended Kohn-Sham LDA model, as well as for the two-electron standard and extended Kohn-Sham GGA models, under some conditions on the GGA exchange-correlation functional.

This chapter is organized as follows. First, we provide a detailed presentation of the various Kohn-Sham models, which, despite their importance in physics and chemistry [73], are not very well known in the mathematical community. The mathematical foundations of DFT are recalled in Section 2.2.1, and the derivation of the (standard and extended) Kohn-Sham LDA and GGA models is discussed in Section 2.2.2. We state our main results in Section 2.3, and postpone the proofs until Section 2.4.

We restrict our mathematical analysis to closed-shell, spin-unpolarized models. All our results related to the LDA setting can be easily extended to open-shell, spin-polarized models (i.e. to the local spin-density approximation LSDA). Likewise, we only deal with all electron descriptions, but valence electron models with usual pseudo-potential approximations (norm conserving [84], ultrasoft [85], PAW [19]) can be dealt with in a similar way.

2.2 Mathematical foundations of DFT and Kohn-Sham models

2.2.1 Density Functional Theory

As mentioned previously, DFT aims at calculating electronic ground state energies and densities. Recall that the ground state electronic energy of a molecular system composed of M nuclei of charges z_1, ..., z_M ($z_k \in \mathbb{N} \setminus \{0\}$ in atomic units) and N electrons is the

bottom of the spectrum of the electronic hamiltonian

$$H_N^V = -\frac{1}{2}\sum_{i=1}^{N}\Delta_{\mathbf{r}_i} - \sum_{i=1}^{N}V(\mathbf{r}_i) + \sum_{1\le i<j\le N}\frac{1}{|\mathbf{r}_i-\mathbf{r}_j|} \tag{2.1}$$

where $\mathbf{r}_i$ and $\mathbf{R}_k$ are the positions in $\mathbb{R}^3$ of the i^{th} electron and the k^{th} nucleus respectively, and V is the electrostatic potential generated by the nuclei defined by

$$V(\mathbf{r}) = -\sum_{k=1}^{M}\frac{z_k}{|\mathbf{r}-\mathbf{R}_k|}.$$

The hamiltonian H_N^V acts on electronic wavefunctions $\Psi(\mathbf{r}_1,\sigma_1;\cdots;\mathbf{r}_N,\sigma_N)$, $\sigma_i\in\Sigma := \{|\uparrow\rangle,|\downarrow\rangle\}$ denoting the spin variable of the i^{th} electron, the nuclear coordinates $\{\mathbf{R}_k\}_{1\le k\le M}$ playing the role of parameters. It is convenient to denote by $\mathbb{R}^3_\Sigma := \mathbb{R}^3\times\{|\uparrow\rangle,|\downarrow\rangle\}$ and $\mathbf{x}_i := (\mathbf{r}_i,\sigma_i)$. As electrons are fermions, electronic wavefunctions are antisymmetric with respect to the renumbering of electrons, i.e.

$$\Psi(\mathbf{x}_{p(1)},\cdots,\mathbf{x}_{p(N)}) = \varepsilon(p)\Psi(\mathbf{x}_1,\cdots,\mathbf{x}_N)$$

where $\varepsilon(p)$ is the signature of the permutation p. Note that (in the absence of magnetic fields) $H_N^V\Psi$ is real-valued if Ψ is real-valued. Our purpose being the calculation of the bottom of the spectrum of H_N^V, there is therefore no restriction in considering real-valued wavefunctions only. In other words, H_N^V can be considered here as an operator on the real Hilbert space

$$\mathcal{H}_N = \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3_\Sigma),$$

endowed with the inner product

$$\langle\Psi|\Psi'\rangle_{\mathcal{H}_N} = \int_{(\mathbb{R}^3_\Sigma)^N}\Psi(\mathbf{x}_1,\cdots,\mathbf{x}_N)\,\Psi'(\mathbf{x}_1,\cdots,\mathbf{x}_N)\,d\mathbf{x}_1\cdots d\mathbf{x}_N,$$

where

$$\int_{\mathbb{R}^3_\Sigma} f(\mathbf{x})\,d\mathbf{x} := \sum_{\sigma\in\Sigma}\int_{\mathbb{R}^3} f(\mathbf{r},\sigma)\,d\mathbf{r},$$

and the corresponding norm $\|\cdot\|_{\mathcal{H}_N} = \langle\cdot|\cdot\rangle_{\mathcal{H}_N}^{\frac{1}{2}}$. It is well-known that H_N^V is a self-adjoint operator on $\mathcal{H}_N$ with form domain

$$\mathcal{Q}_N = \bigwedge_{i=1}^{N} H^1(\mathbb{R}^3_\Sigma).$$

Denoting by $Z = \sum_{k=1}^{M} z_k$ the total nuclear charge of the system, it results from the Zhislin-Sigalov theorem [87, 88] that for neutral or positively charged systems ($Z\ge N$), H_N^V has an infinite number of negative eigenvalues below the bottom of its essential spectrum. In particular, the electronic ground state energy $I_N(V)$ is an eigenvalue of H_N^V, and more precisely the lowest one.

In any case, i.e. whatever Z and N, we always have

$$I_N(V) = \inf \left\{ \langle \Psi | H_N^V | \Psi \rangle,\ \Psi \in \mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N} = 1 \right\}. \tag{2.2}$$

Note that it also holds

$$I_N(V) = \inf \left\{ \mathrm{Tr}\,(H_N^V \Gamma),\ \Gamma \in \mathcal{D}_N \right\} \tag{2.3}$$

where $\mathcal{D}_N$ is the set of N-body density matrices defined by

$$\mathcal{D}_N = \{ \Gamma \in \mathcal{S}(\mathcal{H}_N) \mid 0 \le \Gamma \le 1,\ \mathrm{Tr}\,(\Gamma) = 1,\ \mathrm{Tr}\,(-\Delta\Gamma) < \infty \}.$$

In the above expression, $\mathcal{S}(\mathcal{H}_N)$ is the vector space of bounded self-adjoint operators on $\mathcal{H}_N$, and the condition $0 \le \Gamma \le 1$ stands for $0 \le \langle \Psi | \Gamma | \Psi \rangle \le \|\Psi\|^2_{\mathcal{H}_N}$ for all $\Psi \in \mathcal{H}_N$. Note that if H is a bounded-from-below self-adjoint operator on some Hilbert space $\mathcal{H}$, with form domain $\mathcal{Q}$, and if D is a positive trace-class self-adjoint operator on $\mathcal{H}$, $\mathrm{Tr}\,(HD)$ can always be defined in $\mathbb{R}_+ \cup \{+\infty\}$ as $\mathrm{Tr}\,(HD) = \mathrm{Tr}\,((H-a)^{\frac{1}{2}} D (H-a)^{\frac{1}{2}}) + a\mathrm{Tr}\,(D)$ where a is any real number such that $H \ge a$.

From a physical viewpoint, (2.2) and (2.3) mean that the ground state energy can be computed either by minimizing over pure states (characterized by wavefunctions Ψ) or by minimizing over mixed states (characterized by density operators Γ).

With any N-electron density operator $\Gamma \in \mathcal{D}_N$ can be associated the electronic density

$$\rho_\Gamma(\mathbf{r}) = N \sum_{\sigma \in \Sigma} \int_{(\mathbb{R}^3_\Sigma)^{N-1}} \Gamma(\mathbf{r}, \sigma; \mathbf{x}_2, \cdots, \mathbf{x}_N; \mathbf{r}, \sigma; \mathbf{x}_2, \cdots; \mathbf{x}_N)\, d\mathbf{x}_2 \cdots d\mathbf{x}_N$$

(here and below, we use the same notation for an operator and its Green kernel). For an N-electron wavefunction $\Psi \in \mathcal{H}_N$ such that $\|\Psi\|_{\mathcal{H}_N} = 1$, we will denote by $\rho_\Psi := \rho_{|\Psi\rangle\langle\Psi|}$.

Let us now define the interacting free Hamiltonian by

$$H_N^0 = -\frac{1}{2} \sum_{i=1}^N \Delta_{\mathbf{r}_i} + \sum_{1 \le i < j \le N} \frac{1}{|\mathbf{r}_i - \mathbf{r}_j|}. \tag{2.4}$$

It is easy to see that

$$\langle \Psi | H_N^V | \Psi \rangle = \langle \Psi | H_N^0 | \Psi \rangle + \int_{\mathbb{R}^3} \rho_\Psi V \quad \text{and} \quad \mathrm{Tr}\,(H_N^V \Gamma) = \mathrm{Tr}\,(H_N^0 \Gamma) + \int_{\mathbb{R}^3} \rho_\Gamma V.$$

Besides, it can be checked that

$$\begin{aligned} \mathcal{R}_N &= \{ \rho \mid \exists \Psi \in \mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N} = 1,\ \rho_\Psi = \rho \} = \{ \rho \mid \exists \Gamma \in \mathcal{D}_N,\ \rho_\Gamma = \rho \} \\ &= \left\{ \rho \ge 0 \mid \sqrt{\rho} \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \rho = N \right\}. \end{aligned}$$

It therefore follows that

$$I_N(V) = \inf \left\{ F_{\mathrm{LL}}(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \in \mathcal{R}_N \right\} \tag{2.5}$$

$$= \inf \left\{ F_{\mathrm{L}}(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \in \mathcal{R}_N \right\}, \tag{2.6}$$

where Levy-Lieb's and Lieb's density functionals [51, 53] are respectively defined by

$$F_{\mathrm{LL}}(\rho) = \inf\left\{\langle\Psi|H_N^0|\Psi\rangle,\ \Psi\in\mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N}=1,\ \rho_\Psi=\rho\right\} \tag{2.7}$$

$$F_{\mathrm{L}}(\rho) = \inf\left\{\mathrm{Tr}\,(H_N^0\Gamma),\ \Gamma\in\mathcal{D}_N,\ \rho_\Gamma=\rho\right\}. \tag{2.8}$$

Note that the functionals F_{LL} and F_{L} are independent of the nuclear potential V, i.e. they do not depend on the molecular system. They are therefore universal functionals of the density. It is also shown in [53] that F_{L} is the Legendre transform of the function $V\mapsto I_N(V)$. More precisely, it holds

$$F_{\mathrm{L}}(\rho)=\sup\left\{I_N(V)-\int_{\mathbb{R}^3}\rho V,\ V\in L^{\frac{3}{2}}(\mathbb{R}^3)+L^\infty(\mathbb{R}^3)\right\},$$

from which it follows in particular that F_L is convex on the convex set $\mathcal{R}_N$ (and can be extended to a convex functional on $L^1(\mathbb{R}^3)\cap L^3(\mathbb{R}^3)$).

Formulae (2.5) and (2.6) show that, in principle, it is possible to compute the electronic ground state energy (and the corresponding ground state density if it exists) by solving a minimization problem on $\mathcal{R}_N$. At this stage no approximation has been made. But, as neither F_{LL} nor F_{L} can be easily evaluated for the real system of interest (N interacting electrons), approximations are needed to make of the density functional theory a practical tool for computing electronic ground states. Approximations rely on exact, or very accurate, evaluations of the density functional for reference systems "close" to the real system:

- in Thomas-Fermi and related models, the reference system is a homogeneous electron gas;
- in Kohn-Sham models (by far the most commonly used), it is a system of N *non-interacting* electrons.

2.2.2 Kohn-Sham models

For a system of N non-interacting electrons, universal density functionals are obtained as explained in the previous section; it suffices to replace the interacting hamiltonian H_N^0 of the physical system (formula (2.4)) with the hamiltonian of the reference system

$$T_N=-\sum_{i=1}^N\frac{1}{2}\Delta_{\mathbf{r}_i}. \tag{2.9}$$

The analogue of the Levy-Lieb density functional (2.7) then is the Kohn-Sham type kinetic energy functional

$$\widetilde{T}_{\mathrm{KS}}(\rho)=\inf\left\{\langle\Psi|T_N|\Psi\rangle,\ \Psi\in\mathcal{Q}_N,\ \|\Psi\|_{\mathcal{H}_N}=1,\ \rho_\Psi=\rho\right\}, \tag{2.10}$$

while the analogue of the Lieb functional (2.8) is the Janak kinetic energy functional

$$T_{\mathrm{J}}(\rho)=\inf\left\{\mathrm{Tr}\,(T_N\Gamma),\ \Gamma\in\mathcal{D}_N,\ \rho_\Gamma=\rho\right\}.$$

Let Γ be in the above minimization set. The energy $\mathrm{Tr}\,(T_N\Gamma)$ can be rewritten as a function of the one-electron reduced density operator Υ_Γ associated with Γ. Recall that Υ_Γ is the self-adjoint operator on $L^2(\mathbb{R}^3_\Sigma)$ with kernel

$$\Upsilon_\Gamma(\mathbf{x},\mathbf{x}') = N\int_{(\mathbb{R}^3_\Sigma)^{N-1}} \Gamma(\mathbf{x},\mathbf{x}_2,\cdots,\mathbf{x}_N;\mathbf{x}',\mathbf{x}_2,\cdots,\mathbf{x}_N)\,d\mathbf{x}_2\cdots d\mathbf{x}_N.$$

Indeed, a simple calculation yields $\mathrm{Tr}\,(T_N\Gamma) = \mathrm{Tr}\,(-\frac{1}{2}\Delta_{\mathbf{r}}\Upsilon_\Gamma)$, where $\Delta_{\mathbf{r}}$ is the Laplace operator on $L^2(\mathbb{R}^3_\Sigma)$ - acting on the space coordinate $\mathbf{r}$. Besides, it is known (see e.g. [32]) that

$$\{\Upsilon \mid \exists\Gamma\in\mathcal{D}_N,\ \rho_\Gamma=\rho\} = \{\Upsilon\in\mathcal{RD}_N \mid \rho_\Upsilon=\rho\}\,, \tag{2.11}$$

where

$$\mathcal{RD}_N = \left\{\Upsilon\in\mathcal{S}(L^2(\mathbb{R}^3_\Sigma)) \mid 0\le\Upsilon\le 1,\ \mathrm{Tr}\,(\Upsilon)=N\ \mathrm{Tr}\,(-\Delta_{\mathbf{r}}\Upsilon)<\infty\right\}$$

and $\rho_\Upsilon(\mathbf{r}) := \sum_{\sigma\in\Sigma}\Upsilon(\mathbf{r},\sigma;\mathbf{r},\sigma)$. Hence,

$$T_{\mathrm{J}}(\rho) \;=\; \inf\left\{\mathrm{Tr}\,\left(-\frac{1}{2}\Delta_{\mathbf{r}}\Upsilon\right),\ \Upsilon\in\mathcal{RD}_N,\ \rho_\Upsilon=\rho\right\}. \tag{2.12}$$

It is to be noticed that no such simple expression for $\widetilde{T}_{\mathrm{KS}}(\rho)$ is available because one lacks an N-representation result similar to (2.11) for pure state one-particle reduced density operators. In the standard Kohn-Sham model, $\widetilde{T}_{\mathrm{KS}}(\rho)$ is replaced with the Kohn-Sham kinetic energy functional

$$T_{\mathrm{KS}}(\rho) = \inf\{\langle\Psi|T_N|\Psi\rangle,\ \Psi\in\mathcal{Q}_N,\ \Psi \text{ is a Slater determinant},\ \rho_\Psi=\rho\}\,, \tag{2.13}$$

where we recall that a Slater determinant is a wavefunction Ψ of the form

$$\Psi(\mathbf{x}_1,\cdots,\mathbf{x}_N) = \frac{1}{\sqrt{N!}}\det(\phi_i(\mathbf{x}_j)) \quad\text{with}\quad \phi_i\in L^2(\mathbb{R}^3_\Sigma),\quad \int_{\mathbb{R}^3}\phi_i(\mathbf{x})\phi_j(\mathbf{x})\,d\mathbf{x}=\delta_{ij}.$$

It is then easy to check that

$$T_{\mathrm{KS}}(\rho) = \inf\left\{\frac{1}{2}\sum_{i=1}^N\int_{\mathbb{R}^3_\Sigma}|\nabla\phi_i(\mathbf{x})|^2\,d\mathbf{x},\quad \Phi=(\phi_1,\cdots,\phi_N)\in\mathcal{W}_N,\quad \rho_\Phi=\rho\right\}, \tag{2.14}$$

where we have set

$$\mathcal{W}_N = \left\{\Phi=(\phi_1,\cdots,\phi_N)\quad|\quad \phi_i\in H^1(\mathbb{R}^3_\Sigma),\ \int_{\mathbb{R}^3_\Sigma}\phi_i(\mathbf{x})\phi_j(\mathbf{x})\,d\mathbf{x}=\delta_{ij}\right\}$$

and

$$\rho_\Phi(\mathbf{r}) = \sum_{i=1}^N\sum_{\sigma\in\Sigma}|\phi_i(\mathbf{r},\sigma)|^2.$$

Note that for an arbitrary $\rho \in \mathcal{R}_N$, it holds

$$T_{\rm J}(\rho) \leq \widetilde{T}_{\rm KS}(\rho) \leq T_{\rm KS}(\rho).$$

It is not difficult to check that (2.12) always has a minimizer. If one of the minimizers Υ of (2.12) is of rank N, then $\Upsilon = \sum_{i=1}^N |\phi_i\rangle\langle\phi_i|$ with $\Phi = (\phi_1, \cdots, \phi_N) \in \mathcal{W}_N$, Φ being then a minimizer of (2.13) and $T_{\rm KS}(\rho) = T_{\rm J}(\rho)$. Otherwise, $T_{\rm KS}(\rho) > T_{\rm J}(\rho)$.

The density functionals $T_{\rm KS}$ and T_J associated with the non interacting hamiltonian T_N are expected to provide acceptable approximations of the kinetic energy of the real (interacting) system. Likewise, the Coulomb energy

$$J(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\,\rho(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}\,d\mathbf{r}\,d\mathbf{r}'$$

representing the electrostatic energy of a *classical* charge distribution of density ρ is a reasonable guess for the electronic interaction energy in a system of N electrons of density ρ. The errors on both the kinetic energy and the electrostatic interaction are put together in the *exchange-correlation energy* defined as the difference

$$E_{\rm xc}(\rho) = F_{\rm LL}(\rho) - T_{\rm KS}(\rho) - J(\rho), \tag{2.15}$$

or

$$E_{\rm xc}(\rho) = F_{\rm L}(\rho) - T_{\rm J}(\rho) - J(\rho), \tag{2.16}$$

depending on the choices for the interacting and non-interacting density functionals. We finally end up with the so-called Kohn-Sham and extended Kohn-Sham models

$$\begin{aligned} I_N^{\rm KS}(V) \quad = \quad \inf\Bigg\{ & \frac{1}{2}\sum_{i=1}^N \int_{\mathbb{R}^3_\Sigma} |\nabla\phi_i(\mathbf{x})|^2\,d\mathbf{x} + \int_{\mathbb{R}^3}\rho_\Phi V + J(\rho_\Phi) + E_{\rm xc}(\rho_\Phi), \\ & \Phi = (\phi_1, \cdots, \phi_N) \in \mathcal{W}_N \Bigg\}, \end{aligned} \tag{2.17}$$

and

$$I_N^{\rm EKS}(V) = \inf\left\{ {\rm Tr}\,\left(-\frac{1}{2}\Delta_{\mathbf{r}}\Upsilon\right) + \int_{\mathbb{R}^3}\rho_\Upsilon V + J(\rho_\Upsilon) + E_{\rm xc}(\rho_\Upsilon),\ \Upsilon \in \mathcal{RD}_N \right\}. \tag{2.18}$$

Up to now, no approximation has been made, in such a way that for the exact exchange-correlation functionals ((2.15) or (2.16)), $I_N^{\rm KS}(V) = I_N^{\rm EKS}(V) = I_N(V)$ for any molecular system containing N electrons. Unfortunately, there is no tractable expression of $E_{\rm xc}(\rho)$ that can be used in numerical simulations. Before proceeding further, and for the sake of simplicity, we will restrict ourselves to closed-shell, spin-unpolarized, systems. This means that we will only consider molecular systems with an even number of electrons $N = 2N_p$, where N_p is the number of electron pairs in the system, and that we will assume that electrons "go by pairs". In the Kohn-Sham formalism, this means that the set of admissible states reduces to

$$\left\{ \Phi = (\varphi_1\alpha, \varphi_1\beta, \cdots, \varphi_{N_p}\alpha, \varphi_{N_p}\beta) \mid \varphi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}\varphi_i\varphi_j = \delta_{ij} \right\},$$

where $\alpha(|\uparrow\rangle) = 1$, $\alpha(|\downarrow\rangle) = 0$, $\beta(|\uparrow\rangle) = 0$ and $\beta(|\downarrow\rangle) = 1$, yielding the spin-unpolarized (or closed-shell, or restricted) Kohn-Sham model

$$I_N^{\mathrm{RKS}}(V) = \inf\left\{ \sum_{i=1}^{N_p} \int_{\mathbb{R}^3} |\nabla\phi_i|^2 + \int_{\mathbb{R}^3} \rho_{\Phi} V + J(\rho_{\Phi}) + E_{\mathrm{xc}}(\rho_{\Phi}), \right.$$
$$\left. \Phi = (\phi_1, \cdots, \phi_{N_p}) \in (H^1(\mathbb{R}^3))^{N_p}, \quad \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \quad \rho_{\Phi} = 2\sum_{i=1}^{N_p} |\phi_i|^2 \right\}, \quad (2.19)$$

where the factor 2 in the definition of ρ_{Φ} accounts for the spin. Likewise, the constraints on the one-electron reduced density operators originating from the closed-shell approximation read:

$$\Upsilon(\mathbf{r}, |\uparrow\rangle, \mathbf{r}', |\uparrow\rangle) = \Upsilon(\mathbf{r}, |\downarrow\rangle, \mathbf{r}', |\downarrow\rangle) \quad \text{and} \quad \Upsilon(\mathbf{r}, |\uparrow\rangle, \mathbf{r}', |\downarrow\rangle) = \Upsilon(\mathbf{r}, |\downarrow\rangle, \mathbf{r}', |\uparrow\rangle) = 0.$$

Introducing $\gamma(\mathbf{r}, \mathbf{r}') = \Upsilon(\mathbf{r}, |\uparrow\rangle, \mathbf{r}', |\uparrow\rangle)$ and denoting by $\rho_\gamma(\mathbf{r}) = 2\gamma(\mathbf{r}, \mathbf{r})$, we obtain the spin-unpolarized extended Kohn-Sham model

$$I_N^{\mathrm{REKS}}(V) = \inf\left\{\mathcal{E}(\gamma),\ \gamma \in \mathcal{K}_{N_p}\right\}$$

where

$$\mathcal{E}(\gamma) = \mathrm{Tr}\,(-\Delta\gamma) + \int_{\mathbb{R}^3} \rho_\gamma V + J(\rho_\gamma) + E_{\mathrm{xc}}(\rho_\gamma),$$

and

$$\mathcal{K}_{N_p} = \left\{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid 0 \le \gamma \le 1,\ \mathrm{Tr}\,(\gamma) = N_p,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty\right\}.$$

Note that any $\gamma \in \mathcal{K}_{N_p}$ is of the form

$$\gamma = \sum_{i=1}^{+\infty} n_i |\phi_i\rangle\langle\phi_i|$$

with

$$\phi_i \in H^1(\mathbb{R}^3), \quad \int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{ij}, \quad n_i \in [0,1], \quad \sum_{i=1}^{+\infty} n_i = N_p, \quad \sum_{i=1}^{+\infty} n_i \|\nabla\phi_i\|_{L^2}^2 < \infty.$$

In particular,

$$\rho_\gamma(\mathbf{r}) = 2\sum_{i=1}^{+\infty} n_i |\phi_i(\mathbf{r})|^2.$$

Let us also remark that problem (2.19) can be recast in terms of density operators as follows:

$$I_N^{\mathrm{RKS}}(V) = \inf\left\{\mathcal{E}(\gamma), \quad \gamma \in \mathcal{P}_{N_p}\right\}, \quad (2.20)$$

where

$$\mathcal{P}_{N_p} = \left\{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid \gamma^2 = \gamma,\ \mathrm{Tr}\,(\gamma) = N_p,\ \mathrm{Tr}\,(-\Delta\gamma) < \infty\right\}$$

is the set of finite energy rank-N_p orthogonal projectors (note that $\mathcal{K}_{N_p}$ is the convex hull of $\mathcal{P}_{N_p}$). The connection between (2.19) and (2.20) is given by the correspondence

$\gamma = \sum_{i=1}^{N_p} |\phi_i\rangle\langle\phi_i|$, i.e. γ is the orthogonal projector on the vector space spanned by the ϕ_i's. Indeed, as $|\nabla| = (-\Delta)^{\frac{1}{2}}$, it holds

$$\mathrm{Tr}\,(-\Delta\gamma) = \mathrm{Tr}\,(|\nabla|\gamma|\nabla|) = \sum_{i=1}^{N_p} \||\nabla|\phi_i\|_{L^2}^2 = \sum_{i=1}^{N_p} \|\nabla\phi_i\|_{L^2}^2 = \sum_{i=1}^{N_p} \int_{\mathbb{R}^3} |\nabla\phi_i|^2.$$

Let us now address the issue of constructing relevant approximations for $E_{\mathrm{xc}}(\rho)$. In their celebrated article [47], Kohn and Sham proposed to use an approximate exchange-correlation functional of the form

$$E_{\mathrm{xc}}(\rho) = \int_{\mathbb{R}^3} g(\rho(\mathbf{r}))\, d\mathbf{r} \qquad \text{(LDA exchange-correlation functional)} \tag{2.21}$$

where $\rho^{-1}g(\rho)$ is the exchange-correlation energy density for a uniform electron gas with density ρ, yielding the so-called local density approximation (LDA). In practical calculations, it is made use of approximations of the function $\rho \mapsto g(\rho)$ (from $\mathbb{R}_+$ to $\mathbb{R}$) obtained by interpolating asymptotic formulae for the low and high density regimes (see e.g. [33]) and accurate quantum Monte Carlo evaluations of $g(\rho)$ for a small number of values of ρ [29]. Several interpolation formulae are available [72, 71, 86], which provide similar results. In the 80's, refined approximations of E_{xc} have been constructed, which take into account the inhomogeneity of the electronic density in real molecular systems. Generalized gradient approximations (GGA) of the exchange-correlation functional are of the form

$$E_{\mathrm{xc}}(\rho) = \int_{\mathbb{R}^3} h\left(\rho(\mathbf{r}), \frac{1}{2}|\nabla\sqrt{\rho(\mathbf{r})}|^2\right) dx \qquad \text{(GGA exchange-correlation functional).} \tag{2.22}$$

Contrarily to the situation encountered for LDA, the function $(\rho,\kappa) \mapsto g(\rho,\kappa)$ (from $\mathbb{R}_+ \times \mathbb{R}_+$ to $\mathbb{R}$) does not have a univoque definition. Several GGA functionals have been proposed and new ones come up periodically.

Remark 2.1. *We have chosen the form (2.22) for the GGA exchange-correlation functional because it is well suited for the study of spin-unpolarized two electron systems (see Theorem 2.3 below). In the Physics literature, spin-unpolarized LDA and GGA exchange-correlation functionals are rather written as follows*

$$E_{\mathrm{xc}}(\rho) = E_{\mathrm{x}}(\rho) + E_{\mathrm{c}}(\rho)$$

with

$$E_{\mathrm{x}}(\rho) = \int_{\mathbb{R}^3} \rho(\mathbf{r})\, \varepsilon_{\mathrm{x}}(\rho(\mathbf{r}))\, F_{\mathrm{x}}(s_\rho(\mathbf{r}))\, d\mathbf{r}, \tag{2.23}$$

$$E_{\mathrm{c}}(\rho) = \int_{\mathbb{R}^3} \rho(\mathbf{r})\, \left[\varepsilon_{\mathrm{c}}(r_\rho(\mathbf{r})) + H(r_\rho(\mathbf{r}), t_\rho(\mathbf{r}))\right]\, d\mathbf{r}. \tag{2.24}$$

In the above decomposition, E_{x} is the exchange energy, E_{c} is the correlation energy, ε_{x} and ε_{c} are respectively the exchange and correlation energy densities of the homogeneous electron gas, $r_\rho(\mathbf{r}) = \left(\frac{4}{3}\pi\rho(\mathbf{r})\right)^{-\frac{1}{3}}$ is the Wigner-Seitz radius, $s_\rho(\mathbf{r}) = \frac{1}{2(3\pi^2)^{\frac{1}{3}}} \frac{|\nabla\rho(\mathbf{r})|}{\rho(\mathbf{r})^{\frac{4}{3}}}$ is the

(non-dimensional) reduced density gradient, $t_\rho(\mathbf{r}) = \frac{1}{4(3\pi^{-1})^{\frac{1}{6}}} \frac{|\nabla\rho(\mathbf{r})|}{\rho(\mathbf{r})^{\frac{7}{6}}}$ *is the correlation gradient,* $F_{\rm x}$ *is the so-called exchange enhancement factor, and* H *is the gradient contribution to the correlation energy. While* $\varepsilon_{\rm x}$ *has a simple analytical expression, namely*

$$\varepsilon_{\rm x}(\rho) = -\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}} \rho^{\frac{1}{3}},$$

$\varepsilon_{\rm c}$ *has to be approximated (as explained above for the function* g*). For LDA,* $F_{\rm x}$ *is everywhere equal to one and* $H = 0$*. A popular GGA exchange-correlation energy is the PBE functional [69], for which*

$$\begin{aligned} F_{\rm x}(s) &= 1 + \frac{\mu s^2}{1+\mu\nu^{-1}s^2}, \\ H(r,t) &= \theta \ln\left(1 + \frac{\upsilon}{\theta}\, t^2\, \frac{1+A(r)t^2}{1+A(r)t^2+A(r)^2t^4}\right) \quad \textit{with} \quad A(r) = \frac{\upsilon}{\theta}\left(e^{-\varepsilon_{\rm c}(r)/\theta} - 1\right)^{-1}, \end{aligned}$$

the values of the parameters $\mu \simeq 0.21951$*,* $\nu \simeq 0.804$*,* $\theta = \pi^{-2}(1-\ln 2)$ *and* $\upsilon = 3\pi^{-2}\mu$ *following from theoretical arguments.*

2.3 Main results

Let us first set up and comment on the conditions on the LDA and GGA exchange-correlation functionals under which our results hold true:

- the function g in (2.21) is a C^1 function from $\mathbb{R}_+$ to $\mathbb{R}$, twice differentiable and such that

$$g(0) = 0, \tag{2.25}$$
$$g' \leq 0, \tag{2.26}$$
$$\exists 0 < \beta_- \leq \beta_+ < \frac{2}{3} \quad \text{s.t.} \quad \sup_{\rho\in\mathbb{R}_+} \frac{|g'(\rho)|}{\rho^{\beta_-}+\rho^{\beta_+}} < \infty, \tag{2.27}$$
$$\exists 1 \leq \alpha < \frac{3}{2} \quad \text{s.t.} \quad \limsup_{\rho\to 0^+} \frac{g(\rho)}{\rho^\alpha} < 0; \tag{2.28}$$

- the function h in (2.21) is a C^1 function from $\mathbb{R}_+ \times \mathbb{R}_+$ to $\mathbb{R}$, twice differentiable with respect to the second variable, and such that

$$h(0,\kappa) = 0, \ \forall \kappa \in \mathbb{R}_+, \tag{2.29}$$
$$\frac{\partial h}{\partial \rho} \leq 0, \tag{2.30}$$
$$\exists 0 < \beta_- \leq \beta_+ < \frac{2}{3} \quad \text{s.t.} \quad \sup_{(\rho,\kappa)\in\mathbb{R}_+\times\mathbb{R}_+} \frac{\left|\frac{\partial h}{\partial \rho}(\rho,\kappa)\right|}{\rho^{\beta_-}+\rho^{\beta_+}} < \infty, \tag{2.31}$$
$$\exists 1 \leq \alpha < \frac{3}{2} \quad \text{s.t.} \quad \limsup_{(\rho,\kappa)\to(0^+,0^+)} \frac{h(\rho,\kappa)}{\rho^\alpha} < 0, \tag{2.32}$$
$$\exists 0 < a \leq b < \infty \quad \text{s.t.} \quad \forall (\rho,\kappa) \in \mathbb{R}_+\times\mathbb{R}_+, \ \ a \leq 1 + \frac{\partial h}{\partial \kappa}(\rho,\kappa) \leq b, \tag{2.33}$$

$$\forall(\rho,\kappa)\in\mathbb{R}_+\times\mathbb{R}_+,\quad 1+\frac{\partial h}{\partial\kappa}(\rho,\kappa)+2\kappa\frac{\partial^2 h}{\partial\kappa^2}(\rho,\kappa)\geq 0. \tag{2.34}$$

Conditions (2.25)-(2.28) on the LDA exchange-correlation energy are not restrictive. They are obviously fulfilled by the LDA exchange functional ($g_{\rm x}^{\rm LDA}(\rho)=-\frac{3}{4}\left(\frac{3}{\pi}\right)^{\frac{1}{3}}\rho^{\frac{4}{3}}$), and are also satisfied by all the approximate LDA correlation functionals currently used in practice (with $\alpha=\frac{4}{3}$ and $\beta_-=\beta^+=\frac{1}{3}$).

Besides, it is easy to see that the set of functions satisfying assumptions (2.29)-(2.34) is nonempty and nontrivial, meaning that it contains functions really depending on κ and not only LDA-type functions. For instance, c being a sufficiently small positive constant,

$$\tilde{h}(\rho,\kappa)=-c\rho^{\frac{4}{3}}e^{-\kappa/(1+\rho^{\frac{4}{3}})}$$

fulfills all the conditions with $\alpha=\frac{4}{3}$ and $\beta_-=\beta_+=\frac{1}{3}$. We have also checked numerically that assumptions (2.29)-(2.34) are actually satisfied for the PBE exchange-correlation functional (see Remark 2.1), when the LDA correlation energy density $\varepsilon_{\rm c}(r)$ is given by the PZ81 formula [72].

Remark 2.2. *Our results remain true if (2.26) and (2.30) are respectively replaced with the weaker conditions*

$$\exists\frac{1}{3}\leq\beta'_-\leq\beta_+<\frac{2}{3}\quad s.t.\quad \sup_{\rho\in\mathbb{R}_+}\frac{\max(0,g'(\rho))}{\rho^{\beta'_-}+\rho^{\beta_+}}<\infty$$

and

$$\exists\frac{1}{3}\leq\beta'_-\leq\beta_+<\frac{2}{3}\quad s.t.\quad \sup_{(\rho,\kappa)\in\mathbb{R}_+\times\mathbb{R}_+}\frac{\max\left(0,\dfrac{\partial h}{\partial\rho}(\rho,\kappa)\right)}{\rho^{\beta'_-}+\rho^{\beta_+}}<\infty.$$

As usual in the mathematical study of molecular electronic structure models, we embed (2.20) in the family of problems

$$I_\lambda=\inf\left\{\mathcal{E}(\gamma),\ \gamma\in\mathcal{K}_\lambda\right\} \tag{2.35}$$

parametrized by $\lambda\in\mathbb{R}_+$ where

$$\mathcal{K}_\lambda=\left\{\gamma\in\mathcal{S}(L^2(\mathbb{R}^3))\mid 0\leq\gamma\leq 1,\ \mathrm{Tr}\,(\gamma)=\lambda,\ \mathrm{Tr}\,(-\Delta\gamma)<\infty\right\},$$

and introduce the problem at infinity

$$I_\lambda^\infty=\inf\left\{\mathcal{E}^\infty(\gamma),\ \gamma\in\mathcal{K}_\lambda\right\} \tag{2.36}$$

where

$$\mathcal{E}^\infty(\gamma)=\mathrm{Tr}\,(-\Delta\gamma)+J(\rho_\gamma)+E_{\rm xc}(\rho_\gamma).$$

The following results hold true for both the LDA and GGA extended Kohn-Sham models.

Lemma 2.1. *Consider (2.35) and (2.36) with $E_{\rm xc}$ given either by (2.21) or by (2.22) together with the conditions (2.25)-(2.28) or (2.29)-(2.32). Then*

1. *$I_0 = I_0^\infty = 0$ and for all $\lambda > 0$, $-\infty < I_\lambda < I_\lambda^\infty < 0$;*
2. *the functions $\lambda \mapsto I_\lambda$ and $\lambda \mapsto I_\lambda^\infty$ are continuous and decreasing;*
3. *for all $0 < \mu < \lambda$,*

$$I_\lambda \leq I_\mu + I_{\lambda-\mu}^\infty. \tag{2.37}$$

Inequalities (2.37) in Lemma 2.1 are classical concentration-compactness type inequalities [59].

Our main results are the following two theorems.

Theorem 2.2 (Extended KS-LDA model)**.** *Assume that $Z \geq N = 2N_p$ (neutral or positively charged system) and that the function g satisfies (2.25)-(2.28). Then the extended Kohn-Sham LDA model (2.35) with $E_{\rm xc}$ given by (2.21) has a minimizer γ_0. Besides, γ_0 satisfies the self-consistent field equation*

$$\gamma_0 = \chi_{(-\infty,\varepsilon_{\rm F})}(H_{\rho_{\gamma_0}}) + \delta \tag{2.38}$$

for some $\varepsilon_{\rm F} \leq 0$, where

$$H_{\rho_{\gamma_0}} = -\frac{1}{2}\Delta + V + \rho_{\gamma_0} \star |\mathbf{r}|^{-1} + g'(\rho_{\gamma_0}),$$

where $\chi_{(-\infty,\varepsilon_{\rm F})}$ is the characteristic function of the range $(-\infty, \varepsilon_{\rm F})$ and where $\delta \in \mathcal{S}(L^2(\mathbb{R}^3))$ is such that $0 \leq \delta \leq 1$ and $Ran(\delta) \subset Ker(H_{\rho_{\gamma_0}} - \varepsilon_{\rm F})$.

Theorem 2.3 (Extended KS-GGA model for two electron systems)**.** *Assume that $Z \geq N = 2N_p = 2$ (neutral or positively charged system with two electrons) and that the function h satisfies (2.29)-(2.34). Then the extended Kohn-Sham GGA model (2.35) with $E_{\rm xc}$ given by (2.22) has a minimizer γ_0. Besides, $\gamma_0 = |\phi\rangle\langle\phi|$ where ϕ is a minimizer of the standard spin-unpolarized Kohn-Sham problem (2.19) for $N_p = 1$, hence satisfying the Euler equation*

$$-\frac{1}{2}\mathrm{div}\left(\left(1 + \frac{\partial h}{\partial \kappa}(\rho_\phi, |\nabla\phi|^2)\right)\nabla\phi\right) + \left(V + \rho_\phi \star |\mathbf{r}|^{-1} + \frac{\partial h}{\partial \rho}(\rho_\phi, |\nabla\phi|^2)\right)\phi = \varepsilon\phi \tag{2.39}$$

for some $\varepsilon < 0$, where $\rho_\phi = 2\phi^2$. In addition, $\phi \in C^{0,\alpha}(\mathbb{R}^3)$ for some $0 < \alpha < 1$ and decays exponentially fast at infinity. Lastly, ϕ can be chosen non-negative and (ε, ϕ) is the lowest eigenpair of the self-adjoint operator

$$-\frac{1}{2}\mathrm{div}\left(\left(1 + \frac{\partial h}{\partial \kappa}(\rho_\phi, |\nabla\phi|^2)\right)\nabla\cdot\right) + V + \rho_\phi \star |\mathbf{r}|^{-1} + \frac{\partial h}{\partial \rho}(\rho_\phi, |\nabla\phi|^2).$$

We have not been able to extend the results of Theorem 2.3 to the general case of N_p electron pairs. This is mainly due to the fact that the Euler equations for (2.35) with $E_{\rm xc}$ given by (2.22) do not have a simple structure for $N_p \geq 2$ (see remark 2.4 for further details).

Remark 2.3. *Let us explain as of now the usefulness of properties (2.33) and (2.34) in the proof of Theorem 2.3:*

- *(2.33) is necessary to make the operator appearing in the Euler-Lagrange equation (2.39) elliptic;*
- *(2.34) implies that the Kohn-Sham energy functional, considered as a function of ρ and $\kappa = |\nabla\sqrt{\rho}|^2$, is convex w.r.t to κ, and thus ensures some lower semicontinuity property of the gradient terms of the energy for the weak topology of $H^1(\mathbb{R}^3)$.*

Remark 2.4. *(On the difficulties in extending the results of Theorem 2.3 to the general case of $N_p > 1$ electron pairs). Consider the pure-state Kohn-Sham GGA model (2.19) for the sake of simplicity. Under assumptions (2.29) to (2.34), it is easy to see that the equivalent of Lemma 2.10 with $N = 2N_p > 2$ electrons still holds. The main argument is that, using [38, Theorem 2.5], the condition (2.34) still ensures the lower semicontinuity of the energy w.r.t to $|\nabla\sqrt{\rho}|^2$ for the weak topology of $H^1(\mathbb{R}^3;\mathbb{R}^{N_p})$. Therefore, for all $N_p \in \mathbb{N}^*$, if a minimizing sequence $(\mathbf{\Phi}_n)_{n\in\mathbb{N}}$ is compact in $L^2(\mathbb{R}^3;\mathbb{R}^{N_p})$, then its limit is a minimizer of the problem.*

In our proof of compactness in the case $N_p = 1$, we use in a crucial way the properties of the solutions of the Euler equation (2.39), among which boundedness in $L^\infty(\mathbb{R}^3)$ and exponential decay at infinity. When $N_p > 1$, denoting the state vector by $\mathbf{\Phi} = (\phi_1, \cdots, \phi_{N_p})$ and assuming that the energy is differentiable, the Euler-Lagrange optimality conditions turn into the following system: $\forall i \in [\![1, N_p]\!]$,

$$\begin{aligned} &-\frac{1}{2}\mathrm{div}\left(\nabla\phi_i + \frac{\partial h}{\partial\kappa}(\rho_{\mathbf{\Phi}}, \frac{1}{2}|\nabla\sqrt{\rho_{\mathbf{\Phi}}}|^2)\frac{\sum_k \phi_k\nabla\phi_k}{\sum_k \phi_k^2}\phi_i\right) + \frac{1}{2}\frac{\partial h}{\partial\kappa}(\rho_{\mathbf{\Phi}}, \frac{1}{2}|\nabla\sqrt{\rho_{\mathbf{\Phi}}}|^2)\frac{\sum_k \phi_k\nabla\phi_k}{\sum_k \phi_k^2}\cdot\nabla\phi_i \\ &-\frac{1}{2}\frac{\partial h}{\partial\kappa}(\rho_{\mathbf{\Phi}}, \frac{1}{2}|\nabla\sqrt{\rho_{\mathbf{\Phi}}}|^2)\left|\frac{\sum_k \phi_k\nabla\phi_k}{\sum_k \phi_k^2}\right|^2\phi_i + \left(V + \rho_{\mathbf{\Phi}} \star |\mathbf{r}|^{-1} + \frac{\partial h}{\partial\rho}(\rho_{\mathbf{\Phi}}, \frac{1}{2}|\nabla\sqrt{\rho_{\mathbf{\Phi}}}|^2)\right)\phi_i = \varepsilon_i\phi_i. \end{aligned} \tag{2.40}$$

The study of (2.40) is much more involved than that of (2.39). We were not able to prove that solutions of (2.40) still have the required regularity properties and behaviour at infinity, and thus to extend our proof from the scalar case to the vector case.

2.4 Proofs

For clarity, we will use the following notation

$$\begin{aligned} E_{\mathrm{xc}}^{\mathrm{LDA}}(\rho) &= \int_{\mathbb{R}^3} g(\rho(\mathbf{r}))\,d\mathbf{r}, \\ E_{\mathrm{xc}}^{\mathrm{GGA}}(\rho) &= \int_{\mathbb{R}^3} h\left(\rho(\mathbf{r}), \frac{1}{2}|\nabla\sqrt{\rho}(\mathbf{r})|^2\right)d\mathbf{r}, \\ \mathcal{E}^{\mathrm{LDA}}(\gamma) &= \mathrm{Tr}\,(-\Delta\gamma) + \int_{\mathbb{R}^3}\rho_\gamma V + J(\rho_\gamma) + \int_{\mathbb{R}^3} g(\rho_\gamma(\mathbf{r}))\,d\mathbf{r}, \\ \mathcal{E}^{\mathrm{GGA}}(\gamma) &= \mathrm{Tr}\,(-\Delta\gamma) + \int_{\mathbb{R}^3}\rho_\gamma V + J(\rho_\gamma) + \int_{\mathbb{R}^3} h\left(\rho_\gamma(\mathbf{r}), \frac{1}{2}|\nabla\sqrt{\rho_\gamma}(\mathbf{r})|^2\right)d\mathbf{r}. \end{aligned}$$

The notations $E_{\mathrm{xc}}(\rho)$ and $\mathcal{E}(\gamma)$ will refer indifferently to the LDA or the GGA setting.

2.4.1 Preliminary results

Most of the results of this section are elementary, but we provide them for the sake of completeness. Let us denote by $\mathfrak{S}_1$ the vector space of trace-class operators on $L^2(\mathbb{R}^3)$ (see e.g. [74]) and introduce the vector space

$$\mathcal{H} = \{\gamma \in \mathfrak{S}_1 \mid |\nabla|\gamma|\nabla| \in \mathfrak{S}_1\}$$

endowed with the norm $\|\cdot\|_{\mathcal{H}} = \mathrm{Tr}\,(|\cdot|) + \mathrm{Tr}\,(||\nabla|\cdot|\nabla||)$, and the convex set

$$\mathcal{K} = \left\{\gamma \in \mathcal{S}(L^2(\mathbb{R}^3)) \mid 0 \leq \gamma \leq 1,\ \mathrm{Tr}\,(\gamma) < \infty,\ \mathrm{Tr}\,(|\nabla|\gamma|\nabla|) < \infty\right\}.$$

Lemma 2.4. *For all $\gamma \in \mathcal{K}$, $\sqrt{\rho_\gamma} \in H^1(\mathbb{R}^3)$ and the following inequalities hold true*

- *Lower bound on the kinetic energy:*

$$\frac{1}{2}\|\nabla\sqrt{\rho_\gamma}\|_{L^2}^2 \leq \mathrm{Tr}\,(-\Delta\gamma) \tag{2.41}$$

- *Upper bound on the Coulomb energy:*

$$0 \leq J(\rho_\gamma) \leq C(\mathrm{Tr}\,\gamma)^{\frac{3}{2}}(\mathrm{Tr}\,(-\Delta\gamma))^{\frac{1}{2}} \tag{2.42}$$

- *Bounds on the interaction energy between nuclei and electrons:*

$$-4Z(\mathrm{Tr}\,\gamma)^{\frac{1}{2}}(\mathrm{Tr}\,(-\Delta\gamma))^{\frac{1}{2}} \leq \int_{\mathbb{R}^3} \rho_\gamma V \leq 0 \tag{2.43}$$

- *Bounds on the exchange-correlation energy:*

$$-C\left((\mathrm{Tr}\,\gamma)^{1-\frac{\beta_-}{2}}(\mathrm{Tr}\,(-\Delta\gamma))^{\frac{3\beta_-}{2}} + (\mathrm{Tr}\,\gamma)^{1-\frac{\beta_+}{2}}(\mathrm{Tr}\,(-\Delta\gamma))^{\frac{3\beta_+}{2}}\right) \leq E_{\mathrm{xc}}(\rho_\gamma) \leq 0 \tag{2.44}$$

- *Lower bound on the energy:*

$$\mathcal{E}(\gamma) \geq \frac{1}{2}\left((\mathrm{Tr}\,(-\Delta\gamma))^{\frac{1}{2}} - 4Z(\mathrm{Tr}\,\gamma)^{\frac{1}{2}}\right)^2 - 8Z^2\mathrm{Tr}\,\gamma - C\left((\mathrm{Tr}\,\gamma)^{\frac{2-\beta_-}{2-3\beta_-}} + (\mathrm{Tr}\,\gamma)^{\frac{2-\beta_+}{2-3\beta_+}}\right) \tag{2.45}$$

- *Lower bound on the energy at infinity:*

$$\mathcal{E}^\infty(\gamma) \geq \frac{1}{2}\mathrm{Tr}\,(-\Delta\gamma) - C\left((\mathrm{Tr}\,\gamma)^{\frac{2-\beta_-}{2-3\beta_-}} + (\mathrm{Tr}\,\gamma)^{\frac{2-\beta_+}{2-3\beta_+}}\right), \tag{2.46}$$

for a positive constant C independent of γ. In particular, the minimizing sequences of (2.35) and those of (2.36) are bounded in $\mathcal{H}$.

Proof. (2.41) is a straightforward consequence of Cauchy-Schwarz inequality; a proof can be found for instance in [27]. Using Hardy-Littlewood-Sobolev [55], interpolation, and Gagliardo-Nirenberg-Sobolev inequalities, we obtain

$$J(\rho_\gamma) \leq C_1 \|\rho_\gamma\|^2_{L^{\frac{6}{5}}} \leq C_1 \|\rho_\gamma\|_{L^1}^{\frac{3}{2}} \|\rho_\gamma\|_{L^3}^{\frac{1}{2}} \leq C_2 \|\rho_\gamma\|_{L^1}^{\frac{3}{2}} \|\nabla\sqrt{\rho_\gamma}\|_{L^2}.$$

Hence (2.42), using (2.41) and the relation $\|\rho_\gamma\|_{L^1} = 2\mathrm{Tr}\,(\gamma)$. It follows from Cauchy-Schwarz and Hardy inequalities and from the above estimates that

$$\int_{\mathbb{R}^3} \frac{\rho_\gamma}{|\cdot - \mathbf{R}_k|} \leq 2\|\rho_\gamma\|_{L^1}^{\frac{1}{2}} \|\nabla\sqrt{\rho_\gamma}\|_{L^2} \leq 4(\mathrm{Tr}\,\gamma)^{\frac{1}{2}} (\mathrm{Tr}\,(-\Delta\gamma))^{\frac{1}{2}}.$$

Hence (2.43). Conditions (2.25)-(2.27) for LDA and (2.29)-(2.31) for GGA imply that $E_{\mathrm{xc}}(\rho) \leq 0$ and there exists $1 < p_- < p_+ < \frac{5}{3}$ ($p_\pm = 1 + \beta_\pm$) and some constant $C \in \mathbb{R}_+$ such that

$$\forall \rho \in \mathcal{K}, \quad |E_{\mathrm{xc}}(\rho)| \leq C \left(\int_{\mathbb{R}^3} \rho^{p_-} + \int_{\mathbb{R}^3} \rho^{p_+} \right), \tag{2.47}$$

from which we deduce (2.44), using interpolation and Gagliardo-Nirenberg-Sobolev inequalities. Lastly, the estimates (2.45) and (2.46) are straightforward consequences of (2.42)-(2.44). □

Lemma 2.5. *$\mathcal{E}$ and $\mathcal{E}^\infty$ are continuous on $\mathcal{H}$.*

Proof. Let $\gamma \in \mathcal{K}_\lambda$ and consider a sequence $(\gamma_n)_{n\in\mathbb{N}}$ converging to γ strongly in $\mathcal{H}$. It is well-known that ρ_{γ_n} converges to ρ_γ strongly in $L^p(\mathbb{R}^3)$ and $\sqrt{\rho_{\gamma_n}}$ converges to $\sqrt{\rho_\gamma}$ strongly in $H^1(\mathbb{R}^3)$. Since the linear form $\gamma \mapsto \mathrm{Tr}\,(-\Delta\gamma)$ is continuous on $\mathcal{H}$ and the functionals $u \mapsto \int_{\mathbb{R}^3} u^2 V$ and $u \mapsto J(u^2) + E_{\mathrm{xc}}(u^2)$ are continuous on $H^1(\mathbb{R}^3)$, the continuity of $\mathcal{E}$ and $\mathcal{E}^\infty$ on $\mathcal{H}$ immediately follows. □

2.4.2 Proof of Lemma 2.1

Obviously, $I_0 = I_0^\infty = 0$ and $I_\lambda \leq I_\lambda^\infty$ for all $\lambda \in \mathbb{R}_+$.

Let us first prove assertion 3. Let $0 < \mu < \lambda$, $\varepsilon > 0$ and $\gamma \in \mathcal{K}_\mu$ such that $I_\mu \leq \mathcal{E}(\gamma) \leq I_\mu + \varepsilon$. Using Lemma 2.5, the density of finite-rank operators in $\mathcal{H}$ and the density of $C_c^\infty(\mathbb{R}^3)$ in $L^2(\mathbb{R}^3)$, there is no restriction in choosing γ finite-rank and such that $\mathrm{Ran}(\gamma) \subset C_c^\infty(\mathbb{R}^3)$. Likewise, there exists a finite-rank operator $\gamma' \in \mathcal{K}_{\lambda-\mu}$ such that $\mathrm{Ran}(\gamma') \subset C_c^\infty(\mathbb{R}^3)$ and $I_{\lambda-\mu}^\infty \leq \mathcal{E}^\infty(\gamma') \leq I_{\lambda-\mu}^\infty + \varepsilon$.

Let $\mathbf{e}$ be a unit vector of $\mathbb{R}^3$ and τ_a the translation operator on $L^2(\mathbb{R}^3)$ defined by $\tau_a f = f(\cdot - a)$ for all $f \in L^2(\mathbb{R}^3)$. For $n \in \mathbb{N}$, we define $\gamma_n = \gamma + \tau_{n\mathbf{e}}\gamma'\tau_{-n\mathbf{e}}$. It is easy to check that for n large enough, $\gamma_n \in \mathcal{K}_\lambda$ and

$$I_\lambda \leq \mathcal{E}(\gamma_n) \leq \mathcal{E}(\gamma) + \mathcal{E}^\infty(\gamma) + D(\rho_\gamma, \tau_{n\mathbf{e}}\rho_{\gamma'}) \leq I_\mu + I_{\lambda-\mu}^\infty + 3\varepsilon,$$

where $D(\rho, \rho') := \displaystyle\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\mathbf{r})\,\rho'(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|}\, d\mathbf{r}\, d\mathbf{r}'$. Hence (2.37).

Making use of similar arguments, it can also be proved that

$$I_\lambda^\infty \leq I_\mu^\infty + I_{\lambda-\mu}^\infty. \tag{2.48}$$

Let us now consider a function $\phi \in C_c^\infty(\mathbb{R}^3)$ such that $\|\phi\|_{L^2} = 1$. For all $\sigma > 0$ and all $0 \leq \lambda \leq 1$, the density operator $\gamma_{\sigma,\lambda}$ with density matrix $\gamma_{\sigma,\lambda}(\mathbf{r},\mathbf{r}') = \lambda\sigma^3\,\phi(\sigma\mathbf{r})\,\phi(\sigma\mathbf{r}')$ is in $\mathcal{K}_\lambda$. Using (2.28) for LDA and (2.32) for GGA, we obtain that there exists $1 \leq \alpha < \frac{3}{2}$, $c > 0$ and $\sigma_0 > 0$ such that for all $0 \leq \lambda \leq 1$ and all $0 \leq \sigma \leq \sigma_0$,

$$I_\lambda^\infty \leq \mathcal{E}^\infty(\gamma_{\sigma,\lambda}) \leq \lambda\sigma^2 \int_{\mathbb{R}^3} |\nabla\phi|^2 + \lambda^2\sigma J(2|\phi|^2) - c\lambda^\alpha\sigma^{3(\alpha-1)} \int_{\mathbb{R}^3} |\phi|^{2\alpha}.$$

Therefore $I_\lambda^\infty < 0$ for λ positive and small enough. It follows from (2.37) and (2.48) that the functions $\lambda \mapsto I_\lambda$ and $\lambda \mapsto I_\lambda^\infty$ are decreasing, and that for all $\lambda > 0$,

$$-\infty < I_\lambda \leq I_\lambda^\infty < 0.$$

To proceed further, we need the following lemma.

Lemma 2.6. *Let $\lambda > 0$ and $(\gamma_n)_{n\in\mathbb{N}}$ be a minimizing sequence for (2.35). Then the sequence $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ cannot vanish, which means (see [59]) that*

$$\exists R > 0 \quad s.t. \quad \lim_{n\to\infty} \sup_{x\in\mathbb{R}^3} \int_{x+B_R} \rho_{\gamma_n} > 0.$$

The same holds true for the minimizing sequences of (2.36).

Proof. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a minimizing sequence for I_λ. By contradiction, assume that the sequence ρ_{γ_n} vanishes, i.e

$$\forall R > 0, \quad \lim_{n\to\infty} \sup_{x\in\mathbb{R}^3} \int_{x+B_R} \rho_{\gamma_n} = 0.$$

We know from lemma 2.4 that γ_n is bounded in $\mathcal{H}$, and thus that ρ_{γ_n} is bounded in $H^1(\mathbb{R}^3)$. According to lemma I.1 in [59], this and the fact that ρ_{γ_n} is vanishing imply that ρ_{γ_n} converge strongly to 0 in $L^p(\mathbb{R}^3)$ for $1 < p < 3$. In particular, it follows from (2.47) and from the fact that $V \in L^r(\mathbb{R}^3) + L^q(\mathbb{R}^3)$ for some $\frac{3}{2} < r, q < +\infty$, that

$$\lim_{n\to\infty} \int_{\mathbb{R}^3} \rho_{\gamma_n} V + E_{\mathrm{xc}}(\rho_{\gamma_n}) = 0.$$

As

$$\mathcal{E}(\gamma_n) \geq \int_{\mathbb{R}^3} \rho_{\gamma_n} V + E_{\mathrm{xc}}(\rho_{\gamma_n}),$$

we obtain that $I_\lambda \geq 0$. This is in contradiction with the previously proved result stating that $I_\lambda < 0$. Hence $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ cannot vanish. The case of problem (2.36) is easier since the only non-positive term in the energy functional is $E_{\mathrm{xc}}(\rho)$. □

We can now prove that $I_\lambda < I_\lambda^\infty$. For this purpose let us consider a minimizing sequence $(\gamma_n)_{n\in\mathbb{N}}$ for I_λ^∞. We deduce from Lemma 2.6 that there exists $\eta > 0$ and $R > 0$, such that for n large enough, there exists $x_n \in \mathbb{R}^3$ such that

$$\int_{x_n+B_R} \rho_{\gamma_n} \geq \eta.$$

Let us introduce $\widetilde{\gamma}_n = \tau_{\bar{x}_1-x_n}\gamma_n\tau_{x_n-\bar{x}_1}$. Clearly $\widetilde{\gamma}_n \in \mathcal{K}_\lambda$ and $\mathcal{E}(\widetilde{\gamma}_n) \leq \mathcal{E}^\infty(\gamma_n) - \frac{z_1\eta}{R}$. Thus,

$$I_\lambda \leq I_\lambda^\infty - \frac{z_1\eta}{R} < I_\lambda^\infty.$$

It remains to prove that the functions $\lambda \mapsto I_\lambda$ and $\lambda \mapsto I_\lambda^\infty$ are continuous. We will deal here with the former one, the same arguments applying to the latter one. The proof is based on the following lemma.

Lemma 2.7. *Let $(\alpha_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers converging to 1, and $(\rho_k)_{k\in\mathbb{N}}$ a sequence of non-negative densities such that $(\sqrt{\rho_k})_{k\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$. Then*

$$\lim_{k\to\infty} \left(E_{\rm xc}(\alpha_k\rho_k) - E_{\rm xc}(\rho_k)\right) = 0.$$

Proof. In the LDA setting, we deduce from (2.27) that there exists $1 < p_- \leq p_+ < \frac{5}{3}$ and $C \in \mathbb{R}_+$ such that for k large enough

$$\left|E_{\rm xc}^{\rm LDA}(\alpha_k\rho_k) - E_{\rm xc}^{\rm LDA}(\rho_k)\right| \leq C\,|\alpha_k - 1| \int_{\mathbb{R}^3} (\rho_k^{p_-} + \rho_k^{p_+}).$$

In the GGA setting, we obtain from (2.31) and (2.33) that there exists $1 < p_- \leq p_+ < \frac{5}{3}$ and $C \in \mathbb{R}_+$ such that for k large enough

$$\left|E_{\rm xc}^{\rm GGA}(\alpha_k\rho_k) - E_{\rm xc}^{\rm GGA}(\rho_k)\right| \leq C\,|\alpha_k - 1| \int_{\mathbb{R}^3} (\rho_k^{p_-} + \rho_k^{p_+} + |\nabla\sqrt{\rho_k}|^2).$$

As $(\sqrt{\rho_k})_{k\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$, $(\rho_k)_{k\in\mathbb{N}}$ is bounded in $L^p(\mathbb{R}^3)$ for all $1 \leq p \leq 3$ and $(\nabla\sqrt{\rho_k})_{k\in\mathbb{N}}$ is bounded in $(L^2(\mathbb{R}^3))^3$, hence the result. □

We can now complete the proof of Lemma 2.1.

Let $\lambda > 0$, and $(\lambda_k)_{k\in\mathbb{N}}$ be a sequence of positive real numbers converging to λ. Let $\varepsilon > 0$ and $\gamma \in \mathcal{K}_\lambda$ such that

$$I_\lambda \leq \mathcal{E}(\gamma) \leq I_\lambda + \frac{\varepsilon}{2}.$$

For all $k \in \mathbb{N}$, $\gamma_k = \lambda_k\lambda^{-1}\gamma$ is in $\mathcal{K}_{\lambda_k}$ so that $\forall k \in \mathbb{N}, \quad I_{\lambda_k} \leq \mathcal{E}(\gamma_k)$. Besides, it is easy to see that $\mathcal{E}(\gamma_k)$ tends to $\mathcal{E}(\gamma)$ in virtue of Lemma 2.7. Thus $I_{\lambda_k} \leq I_\lambda + \varepsilon$ for k large enough. Now, for each $k \in \mathbb{N}$, we choose $\widetilde{\gamma}_k \in \mathcal{K}_{\lambda_k}$ such that $\mathcal{E}(\widetilde{\gamma}_k) \leq I_{\lambda_k} + \frac{1}{k}$. For all $k \in \mathbb{N}$, we set $\bar{\gamma}_k = \lambda\lambda_k^{-1}\widetilde{\gamma}_k$. As $\bar{\gamma}_k \in \mathcal{K}_\lambda$, it holds $I_\lambda \leq \mathcal{E}(\bar{\gamma}_k)$. We then deduce from Lemma 2.7 that $\lim_{k\to\infty} \left(\mathcal{E}(\widetilde{\gamma}_k) - \mathcal{E}(\bar{\gamma}_k)\right) = 0$, so that for k large enough we get $I_\lambda - \varepsilon \leq I_{\lambda_k}$. This proves the continuity of $\lambda \mapsto I_\lambda$ on $\mathbb{R}_+ \setminus \{0\}$. Lastly, it results from the estimates established in Lemma 2.4 that $\lim_{\lambda\to 0^+} I_\lambda = 0$.

2.4.3 Proof of Theorem 2.2

Let us first prove the following lemma, which relies on classical arguments.

Lemma 2.8. *Let $(\gamma_n)_{n\in\mathbb{N}}$ be a sequence of elements of $\mathcal{K}$, bounded in $\mathcal{H}$, which converges to γ for the weak-$*$ topology of $\mathcal{H}$. If $\lim_{n\to\infty} \mathrm{Tr}\,(\gamma_n) = \mathrm{Tr}\,(\gamma)$, then $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ converges to ρ_γ strongly in $L^p(\mathbb{R}^3)$ for all $1 \le p < 3$ and*

$$\mathcal{E}^{\mathrm{LDA}}(\gamma) \le \liminf_{n\to\infty} \mathcal{E}^{\mathrm{LDA}}(\gamma_n) \quad \textit{and} \quad \mathcal{E}^{\mathrm{LDA},\infty}(\gamma) \le \liminf_{n\to\infty} \mathcal{E}^{\mathrm{LDA},\infty}(\gamma_n).$$

Proof. The fact that $(\gamma_n)_{n\in\mathbb{N}}$ converges to γ for the weak-$*$ topology of $\mathcal{H}$ means that for any compact operator K on $L^2(\mathbb{R}^3)$,

$$\lim_{n\to\infty} \mathrm{Tr}\,(\gamma_n K) = \mathrm{Tr}\,(\gamma K) \quad \text{and} \quad \lim_{n\to\infty} \mathrm{Tr}\,(|\nabla|\gamma_n|\nabla| K) = \mathrm{Tr}\,(|\nabla|\gamma|\nabla| K).$$

For all $W \in C_c^\infty(\mathbb{R}^3)$, the operator $(1+|\nabla|)^{-1} W (1+|\nabla|)^{-1}$ is compact (it is even in the Schatten class $\mathfrak{S}_p$ for all $p > \frac{3}{2}$ in virtue of the Kato-Seiler-Simon inequality [77]), yielding

$$\begin{aligned}\int_{\mathbb{R}^3} \rho_{\gamma_n} W &= 2\,\mathrm{Tr}\,(\gamma_n W) = 2\,\mathrm{Tr}\,((1+|\nabla|)\gamma_n(1+|\nabla|)(1+|\nabla|)^{-1} W (1+|\nabla|)^{-1}) \\ &\underset{n\to\infty}{\to} 2\,\mathrm{Tr}\,((1+|\nabla|)\gamma(1+|\nabla|)(1+|\nabla|)^{-1} W (1+|\nabla|)^{-1}) = 2\,\mathrm{Tr}\,(\gamma W) = \int_{\mathbb{R}^3} \rho_\gamma W.\end{aligned}$$

Hence, $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ converges to ρ_γ in $\mathcal{D}'(\mathbb{R}^3)$. As by (2.41), $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$, it follows that $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ converges to $\sqrt{\rho_\gamma}$ weakly in $H^1(\mathbb{R}^3)$, and strongly in $L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for all $2 \le p < 6$. In particular, $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ converges to $\sqrt{\rho_\gamma}$ weakly in $L^2(\mathbb{R}^3)$. But we also know that

$$\lim_{n\to\infty} \|\sqrt{\rho_{\gamma_n}}\|_{L^2}^2 = \lim_{n\to\infty} \int_{\mathbb{R}^3} \rho_{\gamma_n} = 2 \lim_{n\to\infty} \mathrm{Tr}\,(\gamma_n) = 2\mathrm{Tr}\,(\gamma) = \int_{\mathbb{R}^3} \rho_\gamma = \|\sqrt{\rho_\gamma}\|_{L^2}^2.$$

Therefore, the convergence of $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ to $\sqrt{\rho_\gamma}$ holds strongly in $L^2(\mathbb{R}^3)$. Using Hölder's inequality and the boundedness of $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ in $H^1(\mathbb{R}^3)$, we obtain that $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ converges strongly to $\sqrt{\rho_\gamma}$ in $L^p(\mathbb{R}^3)$ for all $2 \le p < 6$, hence that $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ converges to ρ_γ strongly in $L^p(\mathbb{R}^3)$ for all $1 \le p < 3$. This readily implies

$$\lim_{n\to\infty} \int_{\mathbb{R}^3} \rho_{\gamma_n} V = \int_{\mathbb{R}^3} \rho_\gamma V, \quad \lim_{n\to\infty} J(\rho_{\gamma_n}) = J(\rho_\gamma), \quad \lim_{n\to\infty} E_{\mathrm{xc}}^{\mathrm{LDA}}(\rho_{\gamma_n}) = E_{\mathrm{xc}}^{\mathrm{LDA}}(\rho_\gamma).$$

Lastly, Fatou's theorem for nonnegative trace-class operators yields

$$\mathrm{Tr}\,(|\nabla|\gamma|\nabla|) \le \liminf_{n\to\infty} \mathrm{Tr}\,(|\nabla|\gamma_n|\nabla|).$$

We thus obtain the desired result. □

We will also need the following result.

Lemma 2.9. *Consider $\alpha > 0$ and $\beta > 0$ such that $\alpha + \beta \leq N_p \leq Z/2$. If I_α and I_β^∞ have minimizers, then*

$$I_{\alpha+\beta} < I_\alpha + I_\beta^\infty.$$

Proof. Let γ be a minimizer for I_α. In particular γ satisfies the Euler equation

$$\gamma = 1_{(-\infty,\varepsilon_{\rm F})}(H_{\rho_\gamma}) + \delta$$

for some Fermi level $\varepsilon_{\rm F} \in \mathbb{R}$, where

$$H_{\rho_\gamma} = -\frac{1}{2}\Delta + V + \rho_\gamma \star |\mathbf{r}|^{-1} + g'(\rho_\gamma),$$

and where $0 \leq \delta \leq 1$, $\mathrm{Ran}(\delta) \subset \mathrm{Ker}(H_{\rho_\gamma} - \varepsilon_{\rm F})$. As $V + \rho_\gamma \star |\mathbf{r}|^{-1} + g'(\rho_\gamma)$ is Δ-compact, the essential spectrum of H_{ρ_γ} is $[0, +\infty)$. Besides, H_{ρ_γ} is bounded from below,

$$H_{\rho_\gamma} \leq -\frac{1}{2}\Delta + V + \rho_\gamma \star |\mathbf{r}|^{-1},$$

and we know from [58, Lemma II.1] that as $-\sum_{k=1}^M z_k + \int_{\mathbb{R}^3} \rho_\gamma = -Z + 2\alpha < -Z + 2N_p \leq 0$, the right hand side operator has infinitely many negative eigenvalues of finite multiplicities. Therefore, so has H_{ρ_γ}. Eventually, $\varepsilon_{\rm F} < 0$ and

$$\gamma = \sum_{i=1}^n |\phi_i\rangle\langle\phi_i| + \sum_{i=n+1}^m n_i |\phi_i\rangle\langle\phi_i|$$

where $0 \leq n_i \leq 1$ and where

$$-\frac{1}{2}\Delta\phi_i + V\phi_i + \left(\rho_\gamma \star |\mathbf{r}|^{-1}\right)\phi_i + g'(\rho_\gamma)\,\phi_i = \varepsilon_i\ \phi_i,$$

$\varepsilon_1 < \varepsilon_2 \leq \varepsilon_3 \leq \cdots < 0$ denoting the negative eigenvalues of H_{ρ_γ} including multiplicities (by standard arguments the ground state eigenvalue of H_{ρ_γ} is non-degenerate). It then follows from elementary elliptic regularity results that all the ϕ_i's, hence ρ_γ, are in $H^2(\mathbb{R}^3)$ and therefore vanish at infinity. Using Lemma 2.16, all the ϕ_i decay exponentially fast to zero at infinity.

Now consider γ' a minimizer for I_β^∞. γ' satisfies

$$\gamma' = 1_{(-\infty,\varepsilon'_{\rm F})}(H^\infty_{\rho_{\gamma'}}) + \delta'$$

where

$$H^\infty_{\rho_{\gamma'}} = -\frac{1}{2}\Delta + \rho_{\gamma'} \star |\mathbf{r}|^{-1} + g'(\rho_{\gamma'}),$$

and where $0 \leq \delta' \leq 1$, $\mathrm{Ran}(\delta') \subset \mathrm{Ker}(H^\infty_{\rho_{\gamma'}} - \varepsilon'_{\rm F})$, and $\varepsilon_{\rm F'} \leq 0$.

First consider the case $\varepsilon_{\rm F'} < 0$. Then

$$\gamma' = \sum_{i=1}^{n'} |\phi'_i\rangle\langle\phi'_i| + \sum_{i=n'+1}^{m'} n'_i |\phi'_i\rangle\langle\phi'_i|,$$

all the ϕ_i's being in $C^\infty(\mathbb{R}^3)$ and decaying exponentially fast at infinity. For $n \in \mathbb{N}$ large enough, the operator

$$\gamma_n = \min\left(1, \|\gamma + \tau_{n\mathbf{e}}\gamma'\tau_{-n\mathbf{e}}\|^{-1}\right)(\gamma + \tau_{n\mathbf{e}}\gamma'\tau_{-n\mathbf{e}})$$

then is in $\mathcal{K}$ and $\mathrm{Tr}\,(\gamma_n) \leq (\alpha+\beta)$, which implies $I_{\alpha+\beta} \leq I_{\mathrm{Tr}\,(\gamma_n)}$ due to Lemma 2.1. As both the ϕ_i's and the ϕ_i''s decay exponentially fast to zero, a simple calculation shows that there exists some $\delta > 0$ such that for n large enough

$$\mathcal{E}^{\mathrm{LDA}}(\gamma_n) = \mathcal{E}^{\mathrm{LDA}}(\gamma) + \mathcal{E}^{\mathrm{LDA},\infty}(\gamma') - \frac{2\alpha(Z-2\beta)}{n} + O(e^{-\delta n}) = I_\alpha + I_\beta^\infty - \frac{2\alpha(Z-2\beta)}{n} + O(e^{-\delta n}).$$

Since $2\beta < 2N_p \leq Z$, we have for n large enough

$$I_{\alpha+\beta} \leq I_{\mathrm{Tr}\,(\gamma_n)} \leq \mathcal{E}^{\mathrm{LDA}}(\gamma_n) < I_\alpha + I_\beta^\infty.$$

Now if $\varepsilon_{\mathrm{F}'} = 0$, 0 is an eigenvalue of $H^\infty_{\rho_{\gamma'}}$ and there exists $\psi \in \mathrm{Ker}(H^\infty_{\rho_{\gamma'}}) \subset H^2(\mathbb{R}^3)$ such that $\|\psi\|_{L^2} = 1$ and $\gamma'\psi = \mu\psi$ with $\mu > 0$. For $0 < \eta < \mu$, $\gamma + \eta|\phi_{m+1}\rangle\langle\phi_{m+1}|$ and $\gamma' - \eta|\psi\rangle\langle\psi|$ are in $\mathcal{K}$ and it is easy to see that

$$\mathcal{E}^{\mathrm{LDA}}(\gamma + \eta|\phi_{m+1}\rangle\langle\phi_{m+1}|) = I_\alpha + 2\eta\varepsilon_{m+1} + o(\eta)$$

and

$$\mathcal{E}^{\mathrm{LDA},\infty}(\gamma' - \eta|\psi\rangle\langle\psi|) = I_\beta^\infty + o(\eta).$$

Since $\mathrm{Tr}\,(\gamma + \eta|\phi_{m+1}\rangle\langle\phi_{m+1}|) = \alpha + \eta$ and $\mathrm{Tr}\,(\gamma' - \eta|\psi\rangle\langle\psi|) = \beta - \eta$, we deduce

$$I_{\alpha+\eta} \leq I_\alpha + 2\eta\varepsilon_{m+1} + o(\eta) \quad \text{and} \quad I^\infty_{\beta-\eta} \leq I_\beta^\infty + o(\eta).$$

Then, according to Lemma 2.1, we obtain for η small enough

$$I_{\alpha+\beta} \leq I_{\alpha+\eta} + I^\infty_{\beta-\eta} \leq I_\alpha + I_\beta^\infty + 2\eta\varepsilon_{m+1} + o(\eta) < I_\alpha + I_\beta^\infty.$$

□

We are now in position to prove Theorem 2.2, and even more generally that problem (2.35) with (2.21) has a minimizer for $\lambda \leq N_p$. Let $(\gamma_n)_{n\in\mathbb{N}}$ be a minimizing sequence for I_λ with $\lambda \leq N_p$. We know from Lemma 2.4 that $(\gamma_n)_{n\in\mathbb{N}}$ is bounded in $\mathcal{H}$ and that $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$. Replacing $(\gamma_n)_{n\in\mathbb{N}}$ by a suitable subsequence, we can assume that (γ_n) converges to some $\gamma \in \mathcal{K}$ for the weak-$*$ topology of $\mathcal{H}$ and that $(\sqrt{\rho_{\gamma_n}})_{n\in\mathbb{N}}$ converges to $\sqrt{\rho_\gamma}$ weakly in $H^1(\mathbb{R}^3)$, strongly in $L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for all $2 \leq p < 6$ and almost everywhere.

If $\mathrm{Tr}\,(\gamma) = \lambda$, then $\gamma \in \mathcal{K}_\lambda$ and according to Lemma 2.8,

$$\mathcal{E}^{\mathrm{LDA}}(\gamma) \leq \liminf_{n\to+\infty} \mathcal{E}^{\mathrm{LDA}}(\gamma_n) = I_\lambda$$

yielding that γ is a minimizer of (2.35).

The rest of the proof consists in ruling out the eventuality when $\mathrm{Tr}\,(\gamma) < \lambda$.

Let us first rule out the case $\mathrm{Tr}\,(\gamma) = 0$. By contradiction, assume that $\mathrm{Tr}\,(\gamma) = 0$, which implies $\rho_\gamma = 0$. Then ρ_{γ_n} converges to 0 strongly in $L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for $1 \leq p < 6$, from which we deduce

$$\lim_{n\to+\infty} \int_{\mathbb{R}^3} \rho_{\gamma_n} V = 0.$$

Consequently,

$$I^\infty_\lambda \leq \lim_{n\to+\infty} \mathcal{E}^{\mathrm{LDA},\infty}(\gamma_n) = \lim_{n\to+\infty} \mathcal{E}^{\mathrm{LDA}}(\gamma_n) = I_\lambda$$

which contradicts the first assertion of Lemma 2.1.

Let us now set $\alpha = \mathrm{Tr}\,(\gamma)$ and assume that $0 < \alpha < \lambda$. Following e.g. [35], we consider a quadratic partition of the unity $\xi^2 + \chi^2 = 1$, where ξ is a smooth, radial function, nonincreasing in the radial direction, such that $\xi(0) = 1$, $0 \leq \xi(x) < 1$ if $|x| > 0$, $\xi(x) = 0$ if $|x| \geq 1$, $\|\nabla\xi\|_{L^\infty} \leq 2$ and $\|\nabla(1-\xi^2)^{\frac{1}{2}}\|_{L^\infty} \leq 2$. We then set $\xi_R(\cdot) = \xi\left(\frac{\cdot}{R}\right)$. For all $n \in \mathbb{N}$, $R \mapsto \mathrm{Tr}\,(\xi_R\gamma_n\xi_R)$ is a continuous nondecreasing function which vanishes at $R = 0$ and converges to $\mathrm{Tr}\,(\gamma_n) = \lambda$ when R goes to infinity. Let $R_n > 0$ be such that $\mathrm{Tr}\,(\xi_{R_n}\gamma_n\xi_{R_n}) = \alpha$. The sequence $(R_n)_{n\in\mathbb{N}}$ goes to infinity; otherwise, it would contain a subsequence $(R_{n_k})_{k\in\mathbb{N}}$ converging to a finite value R^*, and we would then get

$$\int_{\mathbb{R}^3} \rho_\gamma(x)\xi^2_{R^*}(x)\,dx = \lim_{k\to\infty} \int_{\mathbb{R}^3} \rho_{\gamma_{n_k}}(x)\xi^2_{R_{n_k}}(x)\,dx = 2\lim_{k\to\infty} \mathrm{Tr}\,(\xi_{R_{n_k}}\gamma_{n_k}\xi_{R_{n_k}}) = 2\alpha = \int_{\mathbb{R}^3} \rho_\gamma(x)\,dx.$$

As $\xi^2_{R^*} < 1$ on $\mathbb{R}^3 \setminus \{0\}$, we reach a contradiction. Consequently, $(R_n)_{n\in\mathbb{N}}$ indeed goes to infinity. Let us now introduce

$$\gamma_{1,n} = \xi_{R_n}\gamma_n\xi_{R_n} \quad \text{and} \quad \gamma_{2,n} = \chi_{R_n}\gamma_n\chi_{R_n}.$$

Note that $\gamma_{1,n}$ and $\gamma_{2,n}$ are trace-class self-adjoint operators on $L^2(\mathbb{R}^3)$ such that $0 \leq \gamma_{j,n} \leq 1$, that $\rho_{\gamma_n} = \rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}$ and that $\mathrm{Tr}\,(\gamma_{1,n}) = \alpha$ while $\mathrm{Tr}\,(\gamma_{2,n}) = \lambda - \alpha$. Besides, using the IMS formula

$$-\Delta = \chi_{R_n}(-\Delta)\chi_{R_n} + \xi_{R_n}(-\Delta)\xi_{R_n} - |\nabla\chi_{R_n}|^2 - |\nabla\xi_{R_n}|^2,$$

it holds

$$\begin{aligned} \mathrm{Tr}\,(-\Delta\gamma_n) &= \mathrm{Tr}\,(-\Delta\gamma_{1,n}) + \mathrm{Tr}\,(-\Delta\gamma_{2,n}) - \mathrm{Tr}\,((|\nabla\chi_{R_n}|^2 + |\nabla\xi_{R_n}|^2)\gamma_n) \\ &\geq \mathrm{Tr}\,(-\Delta\gamma_{1,n}) + \mathrm{Tr}\,(-\Delta\gamma_{2,n}) - \frac{4\lambda}{R_n^2}, \end{aligned} \tag{2.49}$$

from which we infer that both $(\gamma_{1,n})_{n\in\mathbb{N}}$ and $(\gamma_{2,n})_{n\in\mathbb{N}}$ are bounded sequences of $\mathcal{H}$. As for all $\phi \in C^\infty_c(\mathbb{R}^3)$,

$$\begin{aligned} \mathrm{Tr}\,(\gamma_{1,n}(|\phi\rangle\langle\phi|)) &= \mathrm{Tr}\,(\gamma_n(|\xi_{R_n}\phi\rangle\langle\xi_{R_n}\phi|)) \\ &= \mathrm{Tr}\,(\gamma_n(|(\xi_{R_n}-1)\phi\rangle\langle\xi_{R_n}\phi|)) + \mathrm{Tr}\,(\gamma_n(|\phi\rangle\langle(\xi_{R_n}-1)\phi|)) + \mathrm{Tr}\,(\gamma_n(|\phi\rangle\langle\phi|)) \\ &\underset{n\to\infty}{\longrightarrow} \mathrm{Tr}\,(\gamma(|\phi\rangle\langle\phi|)), \end{aligned}$$

we obtain that $(\gamma_{1,n})_{n\in\mathbb{N}}$ converges to γ for the weak-$*$ topology of $\mathcal{H}$.

Since $\mathrm{Tr}\,(\gamma_{1,n}) = \alpha = \mathrm{Tr}\,(\gamma)$ for all n, we deduce from Lemma 2.8 that $(\rho_{\gamma_{1,n}})_{n\in\mathbb{N}}$ converges to ρ_γ strongly in $L^p(\mathbb{R}^3)$ for all $1 \le p < 3$, and that

$$\mathcal{E}^{\mathrm{LDA}}(\gamma) \le \lim_{n\to\infty} \mathcal{E}^{\mathrm{LDA}}(\gamma_{1,n}). \tag{2.50}$$

As a by-product, we also obtain that $(\rho_{\gamma_{2,n}})_{n\in\mathbb{N}}$ converges strongly to zero in $L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for all $1 \le p < 3$ (since $\rho_{\gamma_{2,n}} = \rho_{\gamma_n} - \rho_{\gamma_{1,n}}$ with $(\rho_{\gamma_n})_{n\in\mathbb{N}}$ and $(\rho_{\gamma_{1,n}})_{n\in\mathbb{N}}$ both converging to ρ_γ in $L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for all $1 \le p < 3$). Besides, using again (2.49), it holds

$$\begin{aligned}
\mathcal{E}^{\mathrm{LDA}}(\gamma_n) &= \mathrm{Tr}\,(-\Delta\gamma_n) + \int_{\mathbb{R}^3} \rho_{\gamma_n} V + J(\rho_{\gamma_n}) + \int_{\mathbb{R}^3} g(\rho_{\gamma_n}) \\
&\ge \mathrm{Tr}\,(-\Delta\gamma_{1,n}) + \mathrm{Tr}\,(-\Delta\gamma_{2,n}) + \int_{\mathbb{R}^3} \rho_{\gamma_{1,n}} V + \int_{\mathbb{R}^3} \rho_{\gamma_{2,n}} V \\
&\quad + J(\rho_{\gamma_{1,n}}) + J(\rho_{\gamma_{2,n}}) + \int_{\mathbb{R}^3} g(\rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}) - \frac{4\lambda}{R_n^2} \\
&= \mathcal{E}^{\mathrm{LDA}}(\gamma_{1,n}) + \mathcal{E}^{\mathrm{LDA},\infty}(\gamma_{2,n}) + \int_{\mathbb{R}^3} \rho_{\gamma_{2,n}} V \\
&\quad + \int_{\mathbb{R}^3} (g(\rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}) - g(\rho_{\gamma_{1,n}}) - g(\rho_{\gamma_{2,n}})) - \frac{4\lambda}{R_n^2}.
\end{aligned}$$

For R large enough, one has on the one hand

$$\left| \int_{\mathbb{R}^3} \rho_{\gamma_{2,n}} V \right| \le 2Z \left(\int_{B_R} \rho_{\gamma_{2,n}} \right)^{\frac{1}{2}} \|\nabla \sqrt{\rho_{\gamma_{2,n}}}\|_{L^2} + \frac{2Z(\lambda-\alpha)}{R},$$

and on the other hand

$$\begin{aligned}
&\left| \int_{\mathbb{R}^3} (g(\rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}) - g(\rho_{\gamma_{1,n}}) - g(\rho_{\gamma_{2,n}})) \right| \\
&\qquad \le \int_{B_R} \left| g(\rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}) - g(\rho_{\gamma_{1,n}}) \right| + \int_{B_R} \left| g(\rho_{\gamma_{2,n}}) \right| \\
&\qquad\quad + \int_{B_R^c} \left| g(\rho_{\gamma_{1,n}} + \rho_{\gamma_{2,n}}) - g(\rho_{\gamma_{2,n}}) \right| + \int_{B_R^c} \left| g(\rho_{\gamma_{1,n}}) \right| \\
&\qquad \le C \left(\int_{B_R} (\rho_{\gamma_{2,n}} + \rho_{\gamma_{2,n}}^2) + \|\rho_{\gamma_{1,n}}\|_{L^2} \left(\int_{B_R} \rho_{\gamma_{2,n}}^2 \right)^{\frac{1}{2}} \right) \\
&\qquad\quad + C \left(\int_{B_R} \rho_{\gamma_{2,n}}^{p_-} + \rho_{\gamma_{2,n}}^{p_+} \right) \\
&\qquad\quad + C \left(\int_{B_R^c} (\rho_{\gamma_{1,n}} + \rho_{\gamma_{1,n}}^2) + \|\rho_{\gamma_{2,n}}\|_{L^2} \left(\int_{B_R^c} \rho_{\gamma_{1,n}}^2 \right)^{\frac{1}{2}} \right) \\
&\qquad\quad + C \left(\int_{B_R^c} \rho_{\gamma_{1,n}}^{p_-} + \rho_{\gamma_{1,n}}^{p_+} \right)
\end{aligned}$$

for some constant C independent of R and n. Yet, we know that $(\sqrt{\rho_{\gamma_{1,n}}})_{n\in\mathbb{N}}$ and $(\sqrt{\rho_{\gamma_{2,n}}})_{n\in\mathbb{N}}$ are bounded in $H^1(\mathbb{R}^3)$, that $(\rho_{\gamma_{1,n}})_{n\in\mathbb{N}}$ converges to ρ_γ in $L^p(\mathbb{R}^3)$ for all

$1 \leq p < 3$ and that $(\rho_{\gamma_{2,n}})_{n\in\mathbb{N}}$ converges to 0 in $L^p_{\rm loc}(\mathbb{R}^3)$ for all $1 \leq p < 3$. Consequently, there exists for all $\varepsilon > 0$, some $N \in \mathbb{N}$ such that for all $n \geq N$,

$$\mathcal{E}^{\rm LDA}(\gamma_n) \geq \mathcal{E}^{\rm LDA}(\gamma_{1,n}) + \mathcal{E}^{{\rm LDA},\infty}(\gamma_{2,n}) - \varepsilon \geq I_\alpha + I^\infty_{\lambda-\alpha} - \varepsilon.$$

Letting n go to infinity, ε go to zero, and using (2.37), we obtain that $I_\lambda = I_\alpha + I^\infty_{\lambda-\alpha}$ and that $(\gamma_{1,n})_{n\in\mathbb{N}}$ and $(\gamma_{2,n})_{n\in\mathbb{N}}$ are minimizing sequences for I_α and $I^\infty_{\lambda-\alpha}$ respectively. It also follows from (2.50) that γ is a minimizer for I_α.

Let us now analyze more in details the sequence $(\gamma_{2,n})_{n\in\mathbb{N}}$. As it is a minimizing sequence for $I^\infty_{\lambda-\alpha}$, $(\rho_{\gamma_{2,n}})_{n\in\mathbb{N}}$ cannot vanish, so that there exists $\eta > 0$, $R > 0$ such that for all $n \in \mathbb{N}$, $\int_{y_n+B_R} \rho_{\gamma_{2,n}} \geq \eta$ for some $y_n \in \mathbb{R}^3$. Thus, the sequence $(\tau_{y_n}\gamma_{2,n}\tau_{-y_n})_{n\in\mathbb{N}}$ converges for the weak-$*$ topology of $\mathcal{H}$ to some $\gamma' \in \mathcal{K}$ satisfying ${\rm Tr}\,(\gamma') \geq \eta > 0$. Let $\beta = {\rm Tr}\,(\gamma')$. Reasoning as above, one can easily check that γ' is a minimizer for I^∞_β, and that $I_\lambda = I_\alpha + I^\infty_\beta + I^\infty_{\lambda-\alpha-\beta}$. On the other hand, Lemma 2.9 yields $I_{\alpha+\beta} < I_\alpha + I^\infty_\beta$.

All in all we obtain $I_\lambda > I_{\alpha+\beta} + I^\infty_{\lambda-\alpha-\beta}$, which contradicts Lemma 2.1. The proof is complete.

2.4.4 Proof of Theorem 2.3

For $\phi \in H^1(\mathbb{R}^3)$, we set $\rho_\phi(x) = 2|\phi(x)|^2$ and

$$E(\phi) = \int_{\mathbb{R}^3} |\nabla\phi|^2 + \int_{\mathbb{R}^3} \rho_\phi V + J(\rho_\phi) + E^{\rm GGA}_{\rm xc}(\rho_\phi).$$

For all $\phi \in H^1(\mathbb{R}^3)$ such that $\|\phi\|_{L^2} = 1$, $\gamma_\phi = |\phi\rangle\langle\phi| \in \mathcal{K}_1$ and $\mathcal{E}(\gamma_\phi) = E(\phi)$. Therefore,

$$I_1 \leq \inf\left\{E(\phi),\ \phi \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} |\phi|^2 = 1\right\}.$$

Conversely, for all $\gamma \in \mathcal{K}_1$, $\phi_\gamma = \sqrt{\frac{\rho_\gamma}{2}}$ satisfies $\phi_\gamma \in H^1(\mathbb{R}^3)$, $\|\phi\|_{L^2} = 1$ and

$$\mathcal{E}^{\rm GGA}(\gamma) = \mathcal{E}^{\rm GGA}(|\phi_\gamma\rangle\langle\phi_\gamma|) + {\rm Tr}\,(-\Delta\gamma) - \frac{1}{2}\int_{\mathbb{R}^3} |\nabla\sqrt{\rho_\gamma}|^2 \geq \mathcal{E}^{\rm GGA}(|\phi_\gamma\rangle\langle\phi_\gamma|) = E(\phi_\gamma).$$

Consequently,

$$I_1 = \inf\left\{E(\phi),\ \phi \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} |\phi|^2 = 1\right\} \tag{2.51}$$

and (2.20) has a minimizer for $N_p = 1$, if and only if (2.51) has a minimizer ϕ (γ_ϕ then is a minimizer of (2.20) for $N_p = 1$). We are therefore led to study the minimization problem (2.51). In the GGA setting we are interested in, $E(\phi)$ can be rewritten as

$$E(\phi) = \int_{\mathbb{R}^3} |\nabla\phi|^2 + \int_{\mathbb{R}^3} \rho_\phi V + J(\rho_\phi) + \int_{\mathbb{R}^3} h(\rho_\phi, |\nabla\phi|^2).$$

Conditions (2.29)-(2.33) guarantee that E is Fréchet-differentiable on $H^1(\mathbb{R}^3)$. To see this, it is sufficient to address the exchange-correlation energy, the Fréchet-differentiability of

the other constituents of the energy being classical.

Consider then ϕ and w in $H^1(\mathbb{R}^3)$, and define, for $t \in \mathbb{R}$,

$$\begin{aligned} H(\cdot,t) &= h(\rho_{\phi+tw}, |\nabla(\phi+tw)|^2), \qquad (2.52)\\ K(t) &= \int_{\mathbb{R}^3} H(x,t)dx. \qquad (2.53) \end{aligned}$$

We compute

$$\begin{aligned} \frac{\partial H}{\partial t}(\cdot,t) =& 2\, w(\phi+tw)\frac{\partial h}{\partial \rho}(\rho_{\phi+tw}, |\nabla(\phi+tw)|^2) \\ &+ 2\,\nabla w\cdot(\nabla\phi + t\nabla w)\frac{\partial h}{\partial \kappa}(\rho_{\phi+tw}, |\nabla(\phi+tw)|^2). \end{aligned} \qquad (2.54)$$

It entails from (2.31) and (2.33) that there exists a positive constant C such that

$$\begin{aligned} &\forall t\in\mathbb{R}, \quad \left|\frac{\partial h}{\partial \rho}(\rho_{\phi+tw}, |\nabla(\phi+tw)|^2)\right| \le C(1+\rho_{\phi+tw}^{2/3}),\\ &\forall t\in\mathbb{R}, \quad \left|\frac{\partial h}{\partial \kappa}(\rho_{\phi+tw}, |\nabla(\phi+tw)|^2)\right| \le C. \end{aligned} \qquad (2.55)$$

We derive from (2.54) and (2.55) that there exists a constant C such that for all $t \in]-1,1[$,

$$\left|\frac{\partial H}{\partial t}(\cdot,t)\right| \le C\left((|\phi|+|w|)^2 + (|\phi|+|w|)^{\frac{8}{3}} + (|\nabla\phi|+|\nabla w|)^2\right). \qquad (2.56)$$

The functions ϕ and w being in $H^1(\mathbb{R}^3)$, the right-hand side of (2.56) is in $L^1(\mathbb{R}^3)$.

Using a classical result of differentiation under the integral sign, this shows that K defined by (2.53) is differentiable at $t=0$, with

$$K'(0) = \int_{\mathbb{R}^3} 2\frac{\partial h}{\partial \rho}(\rho_\phi, |\nabla\phi|^2)\phi w + 2\frac{\partial h}{\partial \kappa}(\rho_\phi, |\nabla\phi|^2)\nabla\phi\cdot\nabla w.$$

Consequently E is Gateaux-differentiable and for all $(\phi, w) \in H^1(\mathbb{R}^3)\times H^1(\mathbb{R}^3)$,

$$E'(\phi)\cdot w = 2\left(\frac{1}{2}\int_{\mathbb{R}^3}\left(1+\frac{\partial h}{\partial \kappa}\left(\rho_\phi, |\nabla\phi|^2\right)\right)\nabla\phi\cdot\nabla w + \int_{\mathbb{R}^3}\left(V+\rho_\phi\star|\mathbf{r}|^{-1}+\frac{\partial h}{\partial \rho}\left(\rho_\phi, |\nabla\phi|^2\right)\right)\phi w\right).$$

Since h is a C^1 function from $\mathbb{R}_+\times\mathbb{R}_+$ to $\mathbb{R}$, it is straightforward to see that the function $\phi \to E'(\phi)$ is continuous from $H^1(\mathbb{R}^3)$ to $H^{-1}(\mathbb{R}^3)$.

It is then well known that this implies that E is Fréchet-differentiable on $H^1(\mathbb{R}^3)$.

We now embed (2.51) in the family of problems

$$J_\lambda = \inf\left\{E(\phi),\ \phi\in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3}|\phi|^2 = \lambda\right\} \qquad (2.57)$$

and introduce the problem at infinity

$$J_\lambda^\infty = \inf\left\{E^\infty(\phi),\ \phi \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} |\phi|^2 = \lambda\right\} \tag{2.58}$$

where

$$E^\infty(\phi) = \int_{\mathbb{R}^3} |\nabla\phi|^2 + J(\rho_\phi) + \int_{\mathbb{R}^3} h(\rho_\phi, |\nabla\phi|^2).$$

Note that reasoning as above, one can see that $J_\lambda = I_\lambda$ and $J_\lambda^\infty = I_\lambda^\infty$ for all $0 \leq \lambda \leq 1$ (while these equalities do not *a priori* hold true for $\lambda > 1$).

The rest of this section consists in proving that (2.57) has a minimizer for all $0 \leq \lambda \leq 1$. Let us start with a simple lemma.

Lemma 2.10. *Let $0 \leq \mu \leq 1$ and let $(\phi_n)_{n\in\mathbb{N}}$ be a minimizing sequence for J_μ (resp. for J_μ^∞) which converges to some $\phi \in H^1(\mathbb{R}^3)$ weakly in $H^1(\mathbb{R}^3)$. Assume that $\|\phi\|_{L^2}^2 = \mu$. Then ϕ is a minimizer for J_μ (resp. for J_μ^∞).*

Proof. Let $(\phi_n)_{n\in\mathbb{N}}$ be a minimizing sequence for J_μ which converges to ϕ weakly in $H^1(\mathbb{R}^3)$. For almost all $x \in \mathbb{R}^3$, the function $z \mapsto |z|^2 + h(\rho_\phi(x), |z|^2)$ is convex on $\mathbb{R}^3$ due to (2.34). Besides the function $t \mapsto t + h(\rho_\phi(x), t)$ is Lipschitz on $\mathbb{R}_+$, uniformly in x due to (2.33). It follows that the functional

$$\psi \mapsto \int_{\mathbb{R}^3} \left(|\nabla\psi|^2 + h(\rho_\phi, |\nabla\psi|^2)\right)$$

is convex and continuous on $H^1(\mathbb{R}^3)$. As $(\phi_n)_{n\in\mathbb{N}}$ converges to ϕ weakly in $H^1(\mathbb{R}^3)$, we get

$$\int_{\mathbb{R}^3} \left(|\nabla\phi|^2 + h(\rho_\phi, |\nabla\phi|^2)\right) \leq \liminf_{n\to\infty} \int_{\mathbb{R}^3} \left(|\nabla\phi_n|^2 + h(\rho_\phi, |\nabla\phi_n|^2)\right).$$

Besides, we deduce from (2.31) that

$$\left|\int_{\mathbb{R}^3} \left(h(\rho_{\phi_n}, |\nabla\phi_n|^2) - h(\rho_\phi, |\nabla\phi_n|^2)\right)\right| \leq C\|\phi_n - \phi\|_{L^2},$$

where the constant C only depends on h and on the H^1 bound of $(\phi_n)_{n\in\mathbb{N}}$. As $(\phi_n)_{n\in\mathbb{N}}$ converges to ϕ weakly in $L^2(\mathbb{R}^3)$ and as $\|\phi\|_{L^2} = \|\phi_n\|_{L^2}$ for all $n \in \mathbb{N}$, the convergence of $(\phi_n)_{n\in\mathbb{N}}$ to ϕ holds strongly in $L^2(\mathbb{R}^3)$. Therefore,

$$\begin{aligned}
\int_{\mathbb{R}^3} |\nabla\phi|^2 \quad + \quad & E_{\rm xc}^{\rm GGA}(\rho_\phi) = \int_{\mathbb{R}^3} \left(|\nabla\phi|^2 + h(\rho_\phi, |\nabla\phi|^2)\right) \\
\leq\ & \liminf_{n\to\infty} \int_{\mathbb{R}^3} \left(|\nabla\phi_n|^2 + h(\rho_\phi, |\nabla\phi_n|^2)\right) \\
& + \lim_{n\to\infty} \int_{\mathbb{R}^3} \left(h(\rho_{\phi_n}, |\nabla\phi_n|^2) - h(\rho_\phi, |\nabla\phi_n|^2)\right) \\
=\ & \liminf_{n\to\infty} \int_{\mathbb{R}^3} |\nabla\phi_n|^2 + E_{\rm xc}^{\rm GGA}(\rho_{\phi_n}).
\end{aligned}$$

Finally, as $(\phi_n)_{n\in\mathbb{N}}$ is bounded in H^1 and converges strongly to ϕ in $L^2(\mathbb{R}^3)$, we infer that the convergence holds strongly in $L^p(\mathbb{R}^3)$ for all $2 \leq p < 6$, yielding

$$\lim_{n\to\infty} \int_{\mathbb{R}^3} \rho_{\phi_n} V + J(\rho_{\phi_n}) = \int_{\mathbb{R}^3} \rho_\phi V + J(\rho_\phi).$$

Therefore,

$$E(\phi) \leq \liminf_{n\to\infty} E(\phi_n) = I_\mu.$$

As $\|\phi\|_{L^2}^2 = \mu$, ϕ is a minimizer for J_μ. Obviously, the same arguments can be applied to a minimizing sequence for J_μ^∞. □

Next, we show that the equivalent of Lemma 2.9 in the GGA setting holds.

Lemma 2.11. *Consider $\alpha > 0$ and $\beta > 0$ such that $\alpha + \beta \leq 1$. If J_α and J_β^∞ have minimizers, then*

$$J_{\alpha+\beta} < J_\alpha + J_\beta^\infty.$$

Proof. Let u and v be minimizers for J_α and J_β^∞ respectively. Since $E(\phi) = E(|\phi|)$ for all $\phi \in H^1(\mathbb{R}^3)$, we can assume that u and v are nonnegative. u satisfies the Euler equation

$$-\frac{1}{2}\mathrm{div}\left(\left(1 + \frac{\partial h}{\partial \kappa}(\rho_u, |\nabla u|^2)\right)\nabla u\right) + \left(V + \rho_u \star |\mathbf{r}|^{-1} + \frac{\partial h}{\partial \rho}(\rho_u, |\nabla u|^2)\right) u + \theta_1 u = 0 \tag{2.59}$$

and v satisfies the Euler equation

$$-\frac{1}{2}\mathrm{div}\left(\left(1 + \frac{\partial h}{\partial \kappa}(\rho_v, |\nabla v|^2)\right)\nabla v\right) + \left(\rho_v \star |\mathbf{r}|^{-1} + \frac{\partial h}{\partial \rho}(\rho_v, |\nabla v|^2)\right) v + \theta_2 v = 0 \tag{2.60}$$

where θ_1 and θ_2 are two Lagrange multipliers.

Using properties (2.31) and (2.33) and classical elliptic regularity arguments [39] (see also the proof of Lemma 2.16 below), we obtain that both u and v are in $C^{0,\alpha}(\mathbb{R}^3)$ for some $0 < \alpha < 1$ and vanish at infinity.

Using again (2.31), this implies that $\frac{\partial h}{\partial \rho}(\rho_u, |\nabla u|^2)u$ vanishes at infinity. Since it is a nonpositive function, applying Lemma 2.15 (proved in Appendix) to (2.59) then yields $\theta_1 > 0$.

Moreover, the function $\lambda \mapsto J_\lambda^\infty$ being decreasing on $[0, 1]$, θ_2 is nonnegative.

Let us first assume $\theta_2 > 0$. Applying Lemma 2.16, we then obtain that there exists $\gamma > 0$, $f_1 \in H^1(\mathbb{R}^3)$, $f_2 \in H^1(\mathbb{R}^3)$, $g_1 \in (L^2(\mathbb{R}^3))^3$ and $g_2 \in (L^2(\mathbb{R}^3))^3$ such that

$$u = e^{-\gamma|\cdot|} f_1, \quad v = e^{-\gamma|\cdot|} f_2, \quad \nabla u = e^{-\gamma|\cdot|} g_1, \quad \nabla v = e^{-\gamma|\cdot|} g_2. \tag{2.61}$$

In addition, as $u \geq 0$ and $v \geq 0$, we also have $f_1 \geq 0$ and $f_2 \geq 0$. Let $\mathbf{e}$ be a given unit vector of $\mathbb{R}^3$. For $t > 0$, we set

$$w_t(\mathbf{r}) = \alpha_t \left(u(\mathbf{r}) + v(\mathbf{r} - t\mathbf{e})\right) \quad \text{where} \quad \alpha_t = (\alpha + \beta)^{\frac{1}{2}} \|u + v(\cdot - t\mathbf{e})\|_{L^2}^{-1}.$$

Obviously, $w_t \in H^1(\mathbb{R}^3)$ and $\|w_t\|_{L^2} = \alpha + \beta$, so that

$$E(w_t) \geq J_{\alpha+\beta}. \tag{2.62}$$

Besides,

$$\begin{aligned}\|u+v(\cdot-t\mathbf{e})\|_{L^2}^2 &= \int_{\mathbb{R}^3} u^2+\int_{\mathbb{R}^3} v^2+2\int_{\mathbb{R}^3} f_1(\mathbf{r})\, f_2(\mathbf{r}-t\mathbf{e})\, e^{-\gamma(|\mathbf{r}|+|\mathbf{r}-t\mathbf{e}|)}\, d\mathbf{r}\\ &= \alpha+\beta+2\int_{\mathbb{R}^3} f_1(\mathbf{r})\, f_2(\mathbf{r}-t\mathbf{e})\, e^{-\gamma(|\mathbf{r}|+|\mathbf{r}-t\mathbf{e}|)}\, d\mathbf{r}\\ &= \alpha+\beta+O(e^{-\gamma t}),\end{aligned}$$

yielding

$$\alpha_t = 1+O(e^{-\gamma t}).$$

Likewise, we have

$$\int_{\mathbb{R}^3} |\nabla w_t|^2 = \int_{\mathbb{R}^3} |\nabla u|^2+\int_{\mathbb{R}^3} |\nabla v|^2+O(e^{-\gamma t}), \tag{2.63}$$

$$\int_{\mathbb{R}^3} V|w_t|^2 = \int_{\mathbb{R}^3} V|u|^2+\int_{\mathbb{R}^3} V|v(\cdot-t\mathbf{e})|^2+O(e^{-\gamma t}), \tag{2.64}$$

$$D(\rho_{w_t},\rho_{w_t}) = D(\rho_u,\rho_u)+D(\rho_v,\rho_v)+2D(\rho_u,\rho_{v(\cdot-t\mathbf{e})})+O(e^{-\gamma t}). \tag{2.65}$$

The exchange-correlation term can then be dealt with as follows. Denoting by

$$r_t = \rho_{w_t}-\rho_u-\rho_{v(\cdot-t\mathbf{e})} = 2(\alpha_t^2-1)(|u|^2+|v(\cdot-t\mathbf{e})|^2)+4\alpha_t^2 uv(\cdot-t\mathbf{e})$$

and

$$s_t = |\nabla w_t|^2-|\nabla u|^2-|\nabla v(\cdot-t\mathbf{e})|^2 = (\alpha_t^2-1)(|\nabla u|^2+|\nabla v(\cdot-t\mathbf{e})|^2)+2\alpha_t^2\nabla u\cdot\nabla v(\cdot-t\mathbf{e}),$$

and using (2.31), (2.33), (2.61) and the fact that u and v are bounded in $L^\infty(\mathbb{R}^3)$, we obtain

$$\begin{aligned}&\left|\int_{\mathbb{R}^3} h(\rho_{w_t},|\nabla w_t|^2)-h(\rho_u,|\nabla u|^2)-h(\rho_{v(\cdot-t\mathbf{e})},|\nabla v(\cdot-t\mathbf{e})|^2)\right|\\ \leq\ &\int_{B_{\frac{t}{2}}} \left|h(\rho_u+\rho_{v(\cdot-t\mathbf{e})}+r_t,|\nabla u|^2+|\nabla v(\cdot-t\mathbf{e})|^2+s_t)-h(\rho_u,|\nabla u|^2)\right|\\ +\ &\int_{t\mathbf{e}+B_{\frac{t}{2}}} \left|h(\rho_{v(\cdot-t\mathbf{e})}+\rho_u+r_t,|\nabla v(\cdot-t\mathbf{e})|^2+|\nabla u|^2+s_t)-h(\rho_{v(\cdot-t\mathbf{e})},|\nabla v(\cdot-t\mathbf{e})|^2)\right|\\ +\ &\int_{B_{\frac{t}{2}}} \left|h(\rho_{v(\cdot-t\mathbf{e})},|\nabla v(\cdot-t\mathbf{e})|^2)\right|+\int_{t\mathbf{e}+B_{\frac{t}{2}}} \left|h(\rho_u,|\nabla u|^2)\right|\\ +\ &\int_{\mathbb{R}^3\setminus\left(B_{\frac{t}{2}}\cup(t\mathbf{e}+B_{\frac{t}{2}})\right)} |h(\rho_{w_t},|\nabla w_t|^2)|+h(\rho_u,|\nabla u|^2)|+|h(\rho_{v(\cdot-t\mathbf{e})},|\nabla v(\cdot-t\mathbf{e})|^2)|\\ =\ &O(e^{-\gamma t}).\end{aligned}$$

Combining (2.63)-(2.65) together with the above inequality, we obtain

$$E(w_t)\leq J_\alpha+J_\beta^\infty+\int_{\mathbb{R}^3} V|v(\cdot-t\mathbf{e})|^2+D(\rho_u,\rho_{v(\cdot-t\mathbf{e})})+O(e^{-\gamma t}).$$

Next, using (2.61), we get

$$\begin{aligned}\int_{\mathbb{R}^3} V\rho_{v(\cdot-t\mathbf{e})} + D(\rho_u, \rho_{v(\cdot-t\mathbf{e})}) &= -Zt^{-1}\int_{\mathbb{R}^3}\rho_u + t^{-1}\int_{\mathbb{R}^3}\rho_v\int_{\mathbb{R}^3}\rho_u + o(t^{-1}) \\ &= -2\alpha(Z-2\beta)t^{-1} + o(t^{-1}).\end{aligned}$$

Finally, for t large enough and since $2\beta < 2 \leq Z$,

$$J_{\alpha+\beta} \leq E(w_t) \leq J_\alpha + J_\beta^\infty - 2\alpha(Z-2\beta)t^{-1} + o(t^{-1}) < J_\alpha + J_\beta^\infty.$$

Let us now assume that $\theta_2 = 0$. Using (2.59) and (2.60), we easily get that for $\eta > 0$ small enough,

$$J_{(1+\eta)^2\alpha} \leq E(u+\eta u) = E(u) - \eta\theta_1\alpha + o(\eta) = J_\alpha - \eta\theta_1\alpha + o(\eta)$$

while

$$J^\infty_{(1-2\frac{\alpha}{\beta}\eta)^2\beta} \leq E^\infty(v - 2\frac{\alpha}{\beta}\eta v) = E^\infty(v) + o(\eta) = J_\beta^\infty + o(\eta).$$

Lemma 2.1 then yields

$$J_{(1+\eta)^2\alpha+(1-2\frac{\alpha}{\beta}\eta)^2\beta} \leq J_{(1+\eta)^2\alpha} + J^\infty_{(1-2\frac{\alpha}{\beta}\eta)^2\beta} \leq J_\alpha + J_\beta^\infty - \eta\theta_1\alpha + o(\eta),$$

and for η small enough, it holds $(1+\eta)^2\alpha + (1-2\frac{\alpha}{\beta}\eta)^2\beta \leq \alpha+\beta$ so that

$$J_{\alpha+\beta} \leq J_{(1+\eta)^2\alpha+(1-2\frac{\alpha}{\beta}\eta)^2\beta} \leq J_\alpha + J_\beta^\infty - \eta\theta_1\alpha + o(\eta) < J_\alpha + J_\beta^\infty.$$

□

In order to prove that the minimizing sequences for J_λ (or at least some of them) are indeed precompact in $L^2(\mathbb{R}^3)$ and to apply Lemma 2.10, we will use the concentration-compactness method due to P.-L. Lions [59], for the simpler method used in the LDA setting does not seem to work anymore. Consider an Ekeland sequence $(\phi_n)_{n\in\mathbb{N}}$ for (2.57), that is [34] a sequence $(\phi_n)_{n\in\mathbb{N}}$ such that

$$\forall n \in \mathbb{N}, \quad \phi_n \in H^1(\mathbb{R}^3) \quad \text{and} \quad \int_{\mathbb{R}^3}\phi_n^2 = \lambda, \tag{2.66}$$

$$\lim_{n\to+\infty} E(\phi_n) = J_\lambda, \tag{2.67}$$

$$\lim_{n\to+\infty} E'(\phi_n) + \theta_n\phi_n = 0 \quad \text{in } H^{-1}(\mathbb{R}^3) \tag{2.68}$$

for some sequence $(\theta_n)_{n\in\mathbb{N}}$ of real numbers. As in the proof of Lemma 2.11, we can assume that

$$\forall n \in \mathbb{N}, \quad \phi_n \geq 0 \text{ a.e. on } \mathbb{R}^3 \quad \text{and} \quad \theta_n \geq 0. \tag{2.69}$$

Lastly, up to extracting subsequences, there is no restriction in assuming the following convergences:

$$\phi_n \rightharpoonup \phi \text{ weakly in } H^1(\mathbb{R}^3), \tag{2.70}$$

$$\phi_n \to \phi \text{ strongly in } L^p_{\text{loc}}(\mathbb{R}^3) \text{ for all } 2 \leq p < 6, \tag{2.71}$$

$$\phi_n \to \phi \text{ a.e. in } \mathbb{R}^3, \tag{2.72}$$

$$\theta_n \to \theta \text{ in } \mathbb{R}, \tag{2.73}$$

and it follows from (2.69) that $\phi \geq 0$ a.e. on $\mathbb{R}^3$ and $\theta \geq 0$. Note that the Ekeland condition (2.68) also reads

$$-\frac{1}{2}\mathrm{div}\left(\left(1+\frac{\partial h}{\partial \kappa}\left(\rho_{\phi_n},|\nabla\phi_n|^2\right)\right)\nabla\phi_n\right)+\left(V+\rho_{\phi_n}\star|\mathbf{r}|^{-1}+\frac{\partial h}{\partial \rho}\left(\rho_{\phi_n},|\nabla\phi_n|^2\right)\right)\phi_n+\theta_n\phi_n$$
$$=\eta_n \qquad \text{with} \qquad \eta_n \underset{n\to 0}{\longrightarrow} 0 \text{ in } H^{-1}(\mathbb{R}^3). \tag{2.74}$$

We can apply to the sequence $(\phi_n)_{n\in\mathbb{N}}$ the following version of the concentration-compactness lemma.

Lemma 2.12 (Concentration-compactness lemma [59]). *Let $\lambda > 0$ and $(\phi_n)_{n\in\mathbb{N}}$ be a bounded sequence in $H^1(\mathbb{R}^3)$ such that*

$$\forall n \in \mathbb{N}, \quad \int_{\mathbb{R}^N} \phi_n^2 = \lambda.$$

Then one can extract from $(\phi_n)_{n\in\mathbb{N}}$ a subsequence $(\phi_{n_k})_{k\in\mathbb{N}}$ such that one of the following three conditions holds true:

1. *(Compactness) There exists a sequence $(y_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^3$, such that for all $\varepsilon > 0$, there exists $R > 0$ such that*

$$\forall k \in \mathbb{N}, \quad \int_{y_k+B_R} \phi_{n_k}^2 \geq \lambda - \varepsilon.$$

2. *(Vanishing) For all $R > 0$,*

$$\lim_{k\to\infty} \sup_{y\in\mathbb{R}^3} \int_{y+B_R} \phi_{n_k}^2 = 0.$$

3. *(Dichotomy) There exists $0 < \delta < \lambda$, such that for all $\varepsilon > 0$ there exists*

   - *a sequence $(y_k)_{k\in\mathbb{N}}$ of points of $\mathbb{R}^3$,*
   - *a positive real number R_1 and a sequence of positive real numbers $(R_{2,k})_{k\in\mathbb{N}}$ converging to $+\infty$,*
   - *two sequences $(\phi_{1,k})_{k\in\mathbb{N}}$ and $(\phi_{2,k})_{n\in\mathbb{N}}$ bounded in $H^1(\mathbb{R}^3)$ (uniformly in ε)*

   such that for all k:

$$\begin{cases}
\phi_{n_k} = \phi_{1,k} \quad \text{on } y_k + B_{R_1}, \\
\phi_{n_k} = \phi_{2,k} \quad \text{on } \mathbb{R}^3 \setminus (y_k + B_{R_{2,k}}), \\
\left|\displaystyle\int_{\mathbb{R}^3} \phi_{1,k}^2 - \delta\right| \leq \varepsilon, \quad \left|\displaystyle\int_{\mathbb{R}^3} \phi_{2,k}^2 - (\lambda-\delta)\right| \leq \varepsilon, \\
\displaystyle\lim_{k\to\infty} \mathrm{dist}(\mathrm{Supp}\ \phi_{1,\mathrm{k}}, \mathrm{Supp}\ \phi_{2,\mathrm{k}}) = \infty, \\
\|\phi_{n_k} - (\phi_{1,k}+\phi_{2,k})\|_{L^p(\mathbb{R}^3)} \leq C_p\, \varepsilon^{\frac{6-p}{2p}} \quad \textit{for all } \ 2 \leq p < 6, \\
\|\phi_{n_k}\|_{L^p(y_k+(B_{R_{2,k}}\setminus\overline{B}_{R_1}))} \leq C_p\, \varepsilon^{\frac{6-p}{2p}} \quad \textit{for all } \ 2 \leq p < 6, \\
\displaystyle\liminf_{k\to\infty} \int_{\mathbb{R}^3} \left(|\nabla\phi_{n_k}|^2 - |\nabla\phi_{1,k}|^2 - |\nabla\phi_{2,k}|^2\right) \geq -C\varepsilon,
\end{cases}$$

where the constants C and C_p only depend on the H^1 bound of $(\phi_n)_{n\in\mathbb{N}}$.

We then conclude using the following result.

Lemma 2.13. *Consider $(\phi_n)_{n\in\mathbb{N}}$ satisfying (2.66)-(2.73). Then, using the terminology introduced in the concentration-compactness Lemma in [59],*

1. *if some subsequence $(\phi_{n_k})_{k\in\mathbb{N}}$ of $(\phi_n)_{n\in\mathbb{N}}$ satisfies the compactness condition, then $(\phi_{n_k})_{k\in\mathbb{N}}$ converges to ϕ strongly in $L^p(\mathbb{R}^3)$ for all $2 \leq p < 6$;*
2. *a subsequence of $(\phi_n)_{n\in\mathbb{N}}$ cannot vanish;*
3. *a subsequence of $(\phi_n)_{n\in\mathbb{N}}$ cannot satisfy the dichotomy condition.*

Consequently, $(\phi_n)_{n\in\mathbb{N}}$ converges to ϕ strongly in $L^p(\mathbb{R}^3)$ for all $2 \leq p < 6$. It follows that ϕ is a minimizer to (2.57).

As the explicit form of the functions $\phi_{1,k}$ and $\phi_{2,k}$ arising in Lemma 2.12 will be useful for proving the third assertion of Lemma 2.13, we briefly recall the proof of the former lemma.

Sketch of the proof of Lemma 2.12. The argument is based on the analysis of Levy's concentration function

$$Q_n(R) = \sup_{y\in\mathbb{R}^3} \int_{y+B_R} \phi_n^2.$$

The sequence $(Q_n)_{n\in\mathbb{N}}$ is a sequence of nondecreasing, nonnegative, uniformly bounded functions such that $\lim_{R\to\infty} Q_n(R) = \lambda$.
There exists consequently a subsequence $(Q_{n_k})_{k\in\mathbb{N}}$ and a nondecreasing nonnegative function Q such that $(Q_{n_k})_{k\in\mathbb{N}}$ converges pointwise to Q. We obviously have

$$\lim_{R\to\infty} Q(R) = \delta \in [0, \lambda].$$

The case $\delta = 0$ corresponds to vanishing, while $\delta = \lambda$ corresponds to compactness. We now consider more in details the case when $0 < \delta < \lambda$ (dichotomy). Let ξ, χ be in $C^\infty(\mathbb{R}^3)$ and such that $0 \leq \xi, \chi \leq 1$, $\xi(x) = 1$ if $|x| \leq 1$, $\xi(x) = 0$ if $|x| \geq 2$, $\chi(x) = 0$ if $|x| \leq 1$, $\chi(x) = 1$ if $|x| \geq 2$, $\|\nabla\chi\|_{L^\infty} \leq 2$ and $\|\nabla\xi\|_{L^\infty} \leq 2$. For $R > 0$, we denote by $\xi_R(\cdot) = \xi\left(\frac{\cdot}{R}\right)$ and $\chi_R(\cdot) = \chi\left(\frac{\cdot}{R}\right)$. Let $\varepsilon > 0$ and $R_1 \geq \varepsilon^{-1}$ large enough for $Q(R_1) \geq \delta - \frac{\varepsilon}{2}$ to hold. Then, up to getting rid of the first terms of the sequence, we can assume that for all k, we have $Q_{n_k}(R_1) \geq \delta - \varepsilon$ and $Q_{n_k}(2R_1) \leq \delta + \frac{\varepsilon}{2}$. Furthermore, there exists $y_k \in \mathbb{R}^3$ such that

$$Q_{n_k}(R_1) = \int_{y_k+B_{R_1}} \phi_{n_k}^2$$

and we can choose a sequence $(R_k')_{k\in\mathbb{N}}$ of positive real numbers greater than R_1, converging to infinity, such that $Q_{n_k}(2R_k') \leq \delta + \varepsilon$ for all $k \in \mathbb{N}$. Consider now

$$\phi_{1,k} = \xi_{R_1}(\cdot - y_k)\phi_{n_k} \quad \text{and} \quad \phi_{2,k} = \chi_{R_k'}(\cdot - y_k)\phi_{n_k}.$$

Denoting by $R_{2,k} = 2R'_k$, we clearly have

$$\left|\int_{\mathbb{R}^3} \phi_{1,k}^2 - \delta\right| \leq \varepsilon, \quad \left|\int_{\mathbb{R}^3} \phi_{2,k}^2 - (\lambda - \delta)\right| \leq \varepsilon,$$

$$\int_{y_k+(B_{R_{2,k}}\setminus\overline{B}_{R_1})} \phi_{n_k}^2 = \int_{R_1<|\cdot-y_k|<R_{2,k}} \phi_{n_k}^2 \leq Q_{n_k}(R_{2,k}) - Q_{n_k}(R_1) \leq 2\varepsilon,$$

and

$$\begin{aligned}\int_{\mathbb{R}^3} |\phi_{n_k} - (\phi_{1,k} + \phi_{2,k})|^2 &\leq \int_{\mathbb{R}^3} |1 - \xi_{R_1}(\cdot - y_k) - \chi_{R'_k}(\cdot - y_k)|^2 \phi_{n_k}^2 \\ &\leq \int_{R_1\leq|\cdot-y_k|\leq R_{2,k}} \phi_{n_k}^2 \leq 2\varepsilon.\end{aligned}$$

Similarly, by Hölder and Gagliardo-Nirenberg-Sobolev inequalities, we have for all k and $2 \leq p < 6$ that

$$\|\phi_{n_k} - (\phi_{1,k} + \phi_{2,k})\|_{L^p} \leq \|\phi_{n_k}\|_{L^p(y_k+(B_{R_{2,k}}\setminus\overline{B}_{R_1}))} \leq C_p \varepsilon^{\frac{(6-p)}{2p}}$$

where the constant C_p only depends on p and on the H^1 bound on $(\phi_n)_{n\in\mathbb{N}}$. Finally, we have $\|\nabla\xi_{R_1}\|_{L^\infty} \leq 2R_1^{-1} \leq 2\varepsilon$ and $\|\nabla\chi_{R'_k}\|_{L^\infty} \leq 2(R'_k)^{-1} \leq 2\varepsilon$, so that

$$\left|\int_{\mathbb{R}^3} |\nabla\phi_{1,k}|^2 - \xi_{R_1}^2(\cdot - y_k)|\nabla\phi_{n_k}|^2\right| \leq C\frac{\varepsilon}{2}$$

and

$$\left|\int_{\mathbb{R}^3} |\nabla\phi_{2,k}|^2 - \chi_{R'_k}^2(\cdot - y_k)|\nabla\phi_{n_k}|^2\right| \leq C\frac{\varepsilon}{2}$$

where the constant C only depend on the H^1 bound on $(\phi_n)_{n\in\mathbb{N}}$. Thus

$$\begin{aligned}\int_{\mathbb{R}^3} |\nabla\phi_{n_k}|^2 - |\nabla\phi_{1,k}|^2 - |\nabla\phi_{2,k}|^2 &\geq \int_{\mathbb{R}^3} (1 - \xi_{R_1}^2(\cdot - y_k) - \chi_{R'_k}^2(\cdot - y_k))|\nabla\phi_{n_k}|^2 - C\varepsilon \\ &\geq -C\varepsilon.\end{aligned}$$

□

Proof of the first two assertions of Lemma 2.13. Assume that there exists a sequence $(y_k)_{k\in\mathbb{N}}$ in $\mathbb{R}^3$, such that for all $\varepsilon > 0$, there exists $R > 0$ such that

$$\forall k \in \mathbb{N}, \quad \int_{y_k+B_R} \phi_{n_k}^2 \geq \lambda - \varepsilon.$$

Two situations may be encountered: either $(y_k)_{k\in\mathbb{N}}$ has a converging subsequence, or $\lim_{k\to\infty} |y_k| = \infty$. In the latter case, we would have $\phi = 0$, and therefore

$$\lim_{k\to\infty} \int_{\mathbb{R}^3} \phi_{n_k}^2 V = 0.$$

Hence

$$I_\lambda^\infty \leq \lim_{k\to\infty} E^\infty(\phi_{n_k}) = \lim_{k\to\infty} E(\phi_{n_k}) = I_\lambda,$$

which is in contradiction with the first assertion of Lemma 2.1. Therefore, $(y_k)_{k\in\mathbb{N}}$ has a converging subsequence. It is then easy to see, using the strong convergence of $(\phi_n)_{n\in\mathbb{N}}$ to ϕ in $L^2_{\rm loc}(\mathbb{R}^3)$, that

$$\int_{\mathbb{R}^3} \phi^2 \geq \int_{y+B_R} \phi^2 \geq \lambda - \varepsilon,$$

where y is the limit of some converging subsequence of $(y_k)_{k\in\mathbb{N}}$. This implies that $\|\phi\|_{L^2}^2 = \lambda$, hence that $(\phi_n)_{n\in\mathbb{N}}$ converges to ϕ strongly in $L^2(\mathbb{R}^3)$. As $(\phi_n)_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$, this convergence holds strongly in $L^p(\mathbb{R}^3)$ for all $2 \leq p < 6$.

Assume now that $(\phi_{n_k})_{k\in\mathbb{N}}$ is vanishing. Then we would have $\phi = 0$, an eventuality that has already been excluded. □

Proof of the third assertion of Lemma 2.13. Replacing $(\phi_n)_{n\in\mathbb{N}}$ with a subsequence and using the detailed construction of the dichotomy case given in the proof of Lemma 2.12 above, we can assume that in addition to (2.66)-(2.73), there exist

- $\delta \in]0, \lambda[$,
- a sequence $(y_n)_{n\in\mathbb{N}}$ of points in $\mathbb{R}^3$,
- two increasing sequences of positive real numbers $(R_{1,n})_{n\in\mathbb{N}}$ and $(R_{2,n})_{n\in\mathbb{N}}$ such that

$$\lim_{n\to\infty} R_{1,n} = \infty \quad \text{and} \quad \lim_{n\to\infty} \frac{R_{2,n}}{2} - R_{1,n} = \infty$$

such that the sequences $\phi_{1,n} = \xi_{R_{1,n}}(\cdot - y_n)\phi_n$ and $\phi_{2,n=}\chi_{R_{2,n}/2}(\cdot - y_n)\phi_n$ satisfy

$$\begin{cases} \phi_n = \phi_{1,n} \quad \text{on } y_n + B_{R_{1,n}}, \\ \phi_n = \phi_{2,n} \quad \text{on } \mathbb{R}^3 \setminus (y_n + B_{R_{2,n}}), \\ \lim\limits_{n\to\infty} \displaystyle\int_{\mathbb{R}^3} \phi_{1,n}^2 = \delta, \quad \lim_{n\to\infty} \int_{\mathbb{R}^3} \phi_{2,n}^2 = \lambda - \delta, \\ \lim\limits_{n\to\infty} \|\phi_n - (\phi_{1,n} + \phi_{2,n})\|_{L^p(\mathbb{R}^3)} = 0 \quad \text{for all } 2 \leq p < 6, \\ \lim\limits_{n\to\infty} \|\phi_n\|_{L^p(y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}}))} = 0 \quad \text{for all } \ 2 \leq p < 6, \\ \lim\limits_{n\to\infty} \operatorname{dist}(\operatorname{Supp} \phi_{1,\mathrm{n}}, \operatorname{Supp} \phi_{2,\mathrm{n}}) = \infty, \\ \liminf\limits_{n\to\infty} \displaystyle\int_{\mathbb{R}^3} \left(|\nabla\phi_n|^2 - |\nabla\phi_{1,n}|^2 - |\nabla\phi_{2,n}|^2 \right) \geq 0. \end{cases}$$

Besides, it obviously follows from the construction of the functions $\phi_{1,n}$ and $\phi_{2,n}$ that

$$\forall n \in \mathbb{N}, \quad \phi_{1,n} \geq 0 \quad \text{and} \quad \phi_{2,n} \geq 0 \quad \text{a.e. on } \mathbb{R}^3. \tag{2.75}$$

A straightforward calculation leads to

$$\begin{aligned} E(\phi_n) &= E^\infty(\phi_{1,n}) + \int_{\mathbb{R}^3} \rho_{\phi_{1,n}} V + E^\infty(\phi_{2,n}) + \int_{\mathbb{R}^3} \rho_{\phi_{2,n}} V \\ &\quad + \int_{\mathbb{R}^3} \left(|\nabla \phi_n|^2 - |\nabla \phi_{1,n}|^2 - |\nabla \phi_{2,n}|^2 \right) + \int_{\mathbb{R}^3} \widetilde{\rho}_n V \\ &\quad + D(\rho_{\phi_{1,n}}, \rho_{\phi_{2,n}}) + D(\widetilde{\rho}_n, \rho_{\phi_{1,n}} + \rho_{\phi_{2,n}}) + \frac{1}{2} D(\widetilde{\rho}_n, \widetilde{\rho}_n) \\ &\quad + \int_{\mathbb{R}^3} (h(\rho_{\phi_n}, |\nabla \phi_n|^2) - h(\rho_{\phi_{1,n}}, |\nabla \phi_{1,n}|^2) - h(\rho_{\phi_{2,n}}, |\nabla \phi_{2,n}|^2)), \quad (2.76) \end{aligned}$$

where we have denoted by $\widetilde{\rho}_n = \rho_{\phi_n} - \rho_{\phi_{1,n}} - \rho_{\phi_{2,n}}$. As

$$|\widetilde{\rho}_n| \leq 3 \, \mathbb{1}_{y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}})} \, |\phi_n|^2,$$

where $\mathbb{1}_{y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}})}$ is the characteristic function of $y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}})$, the sequence $(\widetilde{\rho}_n)_{n \in \mathbb{N}}$ goes to zero in $L^p(\mathbb{R}^3)$ for all $1 \leq p < 3$, yielding

$$\int_{\mathbb{R}^3} \widetilde{\rho}_n V + D(\widetilde{\rho}_n, \rho_{\phi_{1,n}} + \rho_{\phi_{2,n}}) + \frac{1}{2} D(\widetilde{\rho}_n, \widetilde{\rho}_n) \underset{n \to \infty}{\longrightarrow} 0.$$

Besides,

$$D(\rho_{\phi_{1,n}}, \rho_{\phi_{2,n}}) \leq 4 \, \mathrm{dist}(\mathrm{Supp}\, \phi_{1,\mathrm{n}}, \mathrm{Supp}\, \phi_{2,\mathrm{n}})^{-1} \, \|\phi_{1,n}\|_{L^2}^2 \, \|\phi_{2,n}\|_{L^2}^2 \underset{n \to \infty}{\longrightarrow} 0$$

and

$$\begin{aligned} &\left| \int_{\mathbb{R}^3} (h(\rho_{\phi_n}, |\nabla \phi_n|^2) - h(\rho_{\phi_{1,n}}, |\nabla \phi_{1,n}|^2) - h(\rho_{\phi_{2,n}}, |\nabla \phi_{2,n}|^2)) \right| \\ &\leq \int_{y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}})} \left| h(\rho_{\phi_n}, |\nabla \phi_n|^2) \right| + \left| h(\rho_{\phi_{1,n}}, |\nabla \phi_{1,n}|^2) \right| + \left| h(\rho_{\phi_{2,n}}, |\nabla \phi_{2,n}|^2) \right| \\ &\leq C \left(\|\rho_{\phi_n}\|_{L^{p_-}(y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}}))}^{p_-} + \|\rho_{\phi_n}\|_{L^{p_+}(y_n + (B_{R_{2,n}} \setminus \overline{B}_{R_{1,n}}))}^{p_+} \right) \underset{n \to \infty}{\longrightarrow} 0 \end{aligned}$$

(recall that $1 < p_\pm = 1 + \beta_\pm < \frac{5}{3}$). Lastly, as $\lim_{n \to \infty} \mathrm{dist}(\mathrm{Supp}\, \phi_{1,\mathrm{n}}, \mathrm{Supp}\, \phi_{2,\mathrm{n}}) = \infty$,

$$\min \left(\left| \int_{\mathbb{R}^3} \rho_{\phi_{1,n}} V \right|, \left| \int_{\mathbb{R}^3} \rho_{\phi_{2,n}} V \right| \right) \underset{n \to \infty}{\longrightarrow} 0.$$

It therefore follows from (2.76) and from the continuity of the functions $\lambda \mapsto J_\lambda$ and $\lambda \mapsto J_\lambda^\infty$ that at least one of the inequalities below holds true

$$J_\lambda \geq J_\delta + J_{\lambda - \delta}^\infty \quad \text{(case 1)} \quad \text{or} \quad J_\lambda \geq J_\delta^\infty + J_{\lambda - \delta} \quad \text{(case 2)}. \quad (2.77)$$

As the opposite inequalities are always satisfied, we obtain

$$J_\lambda = J_\delta + J_{\lambda - \delta}^\infty \quad \text{(case 1)} \quad \text{or} \quad J_\lambda = J_\delta^\infty + J_{\lambda - \delta} \quad \text{(case 2)}, \quad (2.78)$$

and (still up to extraction)

$$\left\{\begin{array}{l}\lim\limits_{n\to\infty} E(\phi_{1,n}) = J_\delta,\\ \lim\limits_{n\to\infty} E^\infty(\phi_{2,n}) = J^\infty_{\lambda-\delta}\end{array}\right. \quad \text{(case 1)} \quad \text{or} \quad \left\{\begin{array}{l}\lim\limits_{n\to\infty} E^\infty(\phi_{1,n}) = J^\infty_\delta,\\ \lim\limits_{n\to\infty} E(\phi_{2,n}) = J_{\lambda-\delta}\end{array}\right. \quad \text{(case 2)}. \tag{2.79}$$

Let us now prove that the sequence $(\psi_n)_{n\in\mathbb{N}}$, where $\psi_n = \phi_n - (\phi_{1,n} + \phi_{2,n})$, goes to zero in $H^1(\mathbb{R}^3)$. For convenience, we rewrite ψ_n as $\psi_n = e_n\phi_n$ where $e_n = 1 - \xi_{R_{1,n}}(\cdot - y_n) - \chi_{R_{2,n}/2}(\cdot - y_n)$ and Ekeland's condition (2.74) as

$$-\text{div}\,(a_n\nabla\phi_n) + V\phi_n + (\rho_{\phi_n} \star |\mathbf{r}|^{-1})\phi_n + V_n^-\phi_n^{1+2\beta_-} + V_n^+\phi_n^{1+2\beta_+} + \theta_n\phi_n = \eta_n \tag{2.80}$$

where

$$\left\{\begin{array}{l} a_n = \dfrac{1}{2}\left(1 + \dfrac{\partial h}{\partial\kappa}(\rho_{\phi_n}, |\nabla\phi_n|^2)\right),\\ V_n^- = 2^{\beta_-}\rho_{\phi_n}^{-\beta_-}\dfrac{\partial h}{\partial\rho}(\rho_{\phi_n}, |\nabla\phi_n|^2)\chi_{\rho_{\phi_n}\le 1},\\ V_n^+ = 2^{\beta_+}\rho_{\phi_n}^{-\beta_+}\dfrac{\partial h}{\partial\rho}(\rho_{\phi_n}, |\nabla\phi_n|^2)\chi_{\rho_{\phi_n}> 1}.\end{array}\right.$$

Due to assumption 2.32, V_n^- and V_n^+ are bounded in $L^\infty(\mathbb{R}^3)$.

The sequence $(V\phi_n + (\rho_{\phi_n}\star|\mathbf{r}|^{-1})\phi_n + V_n^-\phi_n^{1+2\beta_-} + V_n^+\phi_n^{1+2\beta_+} + \theta_n\phi_n)_{n\in\mathbb{N}}$ is bounded in $L^2(\mathbb{R}^3)$, $(\eta_n)_{n\in\mathbb{N}}$ goes to zero in $H^{-1}(\mathbb{R}^3)$, and the sequence $(\psi_n)_{n\in\mathbb{N}}$ is bounded in $H^1(\mathbb{R}^3)$ and goes to zero in $L^2(\mathbb{R}^3)$. We therefore infer from (2.80) that

$$\int_{\mathbb{R}^3} a_n\nabla\phi_n\cdot\nabla\psi_n \underset{n\to\infty}{\longrightarrow} 0.$$

Besides $\nabla\psi_n = e_n\nabla\phi_n + \phi_n\nabla e_n$ with $0\le e_n\le 1$ and $\|\nabla e_n\|_{L^\infty}\to 0$. Thus

$$\int_{\mathbb{R}^3} a_n e_n|\nabla\phi_n|^2 \underset{n\to\infty}{\longrightarrow} 0.$$

As

$$0 < \frac{a}{2} \le a_n = \frac{1}{2}\left(1 + \frac{\partial h}{\partial\kappa}(\rho_{\phi_n}, |\nabla\phi_n|^2)\right) \le \frac{b}{2} < \infty \quad \text{a.e. on } \mathbb{R}^3 \tag{2.81}$$

and $0\le e_n^2\le e_n\le 1$, we finally obtain

$$\int_{\mathbb{R}^3} e_n^2|\nabla\phi_n|^2 \underset{n\to\infty}{\longrightarrow} 0,$$

from which we conclude that $(\nabla\psi_n)_{n\in\mathbb{N}}$ goes to zero in $L^2(\mathbb{R}^3)$. Plugging this information in (2.80) and using the fact that the supports of $\phi_{1,n}$ and $\phi_{2,n}$ are disjoint and go far apart when n goes to infinity, we obtain

$$-\text{div}\,(a_n\nabla\phi_{1,n}) + V\phi_{1,n} + (\rho_{\phi_{1,n}}\star|\mathbf{r}|^{-1})\phi_{1,n} + V_n^-\phi_{1,n}^{1+2\beta_-} + V_n^+\phi_{1,n}^{1+2\beta_+} + \theta_n\phi_{1,n} \overset{H^{-1}}{\underset{n\to\infty}{\longrightarrow}} 0,$$

$$-\text{div}\,(a_n\nabla\phi_{2,n}) + V\phi_{2,n} + (\rho_{\phi_{2,n}}\star|\mathbf{r}|^{-1})\phi_{2,n} + V_n^-\phi_{2,n}^{1+2\beta_-} + V_n^+\phi_{2,n}^{1+2\beta_+} + \theta_n\phi_{2,n} \overset{H^{-1}}{\underset{n\to\infty}{\longrightarrow}} 0.$$

We can now assume that the sequences $(\phi_{1,n})_{n\in\mathbb{N}}$ and $(\phi_{2,n})_{n\in\mathbb{N}}$, which are bounded in $H^1(\mathbb{R}^3)$, respectively converge to ϕ_1 and ϕ_2 weakly in $H^1(\mathbb{R}^3)$, strongly in $L^p_{\text{loc}}(\mathbb{R}^3)$ for

all $2 \leq p < 6$ and a.e. in $\mathbb{R}^3$. In virtue of (2.75), we also have $\phi_1 \geq 0$ and $\phi_2 \geq 0$ a.e. on $\mathbb{R}^3$. To pass to the limit in the above equations, we use a H-convergence result proved in Appendix (Lemma 2.14). The sequence $(a_n)_{n\in\mathbb{N}}$ satisfying (2.81), there exists $a_\infty \in L^\infty(\mathbb{R}^3)$ such that $\frac{a}{2} \leq a_\infty \leq \frac{b^2}{2a}$ and (up to extraction) $a_n I_3 \rightharpoonup_H a_\infty I_3$ (where I_3 is the rank-3 identity matrix). Besides, the sequence $(V_n^\pm)_{n\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}^3)$, so that there exists $V^\pm \in L^\infty(\mathbb{R}^3)$, such that (up to extraction) $(V_n^\pm)_{n\in\mathbb{N}}$ converges to $V^\pm$ for the weak-$*$ topology of $L^\infty(\mathbb{R}^3)$. Hence for $j \in [\![1,2]\!]$ (and up to extraction)

$$\left\{\begin{array}{ll} V\phi_{j,n} \underset{n\to\infty}{\longrightarrow} V\phi_j & \text{strongly in } H^{-1}(\mathbb{R}^3), \\ V_n^\pm \phi_{j,n}^{1+2\beta_\pm} \underset{n\to\infty}{\rightharpoonup} V^\pm \phi_j^{1+2\beta_\pm} & \text{weakly in } L^2_{\rm loc}(\mathbb{R}^3), \\ (\rho_{\phi_{j,n}} \star |\mathbf{r}|^{-1})\phi_{j,n} + \theta_n \phi_{j,n} \underset{n\to\infty}{\longrightarrow} (\rho_{\phi_j} \star |\mathbf{r}|^{-1})\phi_j + \theta\phi_j & \text{strongly in } L^2_{\rm loc}(\mathbb{R}^3). \end{array}\right.$$

We end up, for $j \in [\![1,2]\!]$, with

$$-\text{div}\,(a_\infty \nabla\phi_j) + V\phi_j + (\rho_{\phi_j} \star |\mathbf{r}|^{-1})\phi_j + V^-\phi_j^{1+2\beta_-} + V^+\phi_j^{1+2\beta_+} + \theta\phi_j = 0. \qquad (2.82)$$

Remark 2.5. *The elliptic operator involved in equation (2.80) being monotone, it appears that we could also pass to the limit using Leray-Lions theory instead of H-convergence. Since we are not interested in the very precise structure of the limit equation, we chose not to follow that way.*

By classical elliptic regularity arguments already stated in the proof of Lemma 2.11, both ϕ_1 and ϕ_2 are in $C^{0,\alpha}(\mathbb{R}^3)$ for some $0 < \alpha < 1$ and vanish at infinity. Besides, exactly one of the two functions ϕ_1 and ϕ_2 is different from zero. Indeed, if both ϕ_1 and ϕ_2 were equal to zero, then we would have $\phi = 0$, an eventuality that we have already excluded in the proof of the first two assertions of lemma 2.13. On the other hand, as $\text{dist}(\text{Supp}\,\phi_{1,\text{n}}, \text{Supp}\,\phi_{2,\text{n}}) \to \infty$, at least one of the functions ϕ_1 and ϕ_2 is equal to zero.

We only consider here the case when $\phi_2 = 0$, corresponding to case 1 in (2.77)-(2.79), since the other case can be dealt with the same arguments. A key point of the proof consists in noticing, as in the proof of Lemma 2.11, that applying Lemma 2.15 to (2.82) (note that $W = V^-\phi_1^{\beta_-} + V^+\phi_1^{\beta_+}$ is nonpositive and goes to zero at infinity) yields

$$\theta > 0. \qquad (2.83)$$

Consider now the sequence $(\widetilde{\phi}_{1,n})_{n\in\mathbb{N}}$ defined by $\widetilde{\phi}_{1,n} = \delta^{\frac{1}{2}}\phi_{1,n}\|\phi_{1,n}\|_{L^2}^{-1}$. It is easy to check that

$$\left\{\begin{array}{l} \forall n \in \mathbb{N}, \quad \widetilde{\phi}_{1,n} \in H^1(\mathbb{R}^3), \quad \displaystyle\int_{\mathbb{R}^3} \widetilde{\phi}_{1,n}^2 = \delta \quad \text{and} \quad \widetilde{\phi}_{1,n} \geq 0 \text{ a.e. on } \mathbb{R}^3, \\ \displaystyle\lim_{n\to+\infty} E(\widetilde{\phi}_{1,n}) = J_\delta, \\ -\text{div}\,(a_{1,n}\nabla\widetilde{\phi}_{1,n}) + V\widetilde{\phi}_{1,n} + (\rho_{\widetilde{\phi}_{1,n}} \star |\mathbf{r}|^{-1})\widetilde{\phi}_{1,n} + V_{1,n}^-\widetilde{\phi}_{1,n}^{1+2\beta_-} + V_{1,n}^+\widetilde{\phi}_{1,n}^{1+2\beta_+} + \theta_n\widetilde{\phi}_{1,n} \overset{H^{-1}}{\underset{n\to\infty}{\longrightarrow}} 0, \\ (\widetilde{\phi}_{1,n})_{n\in\mathbb{N}} \text{ converges to } \widetilde{\phi}_1 \neq 0 \text{ weakly in } H^1, \text{ strongly in } L^p_{\rm loc} \text{ for } 2 \leq p < 6 \text{ and a.e. on } \mathbb{R}^3 \end{array}\right.$$

(with in fact $\widetilde{\phi}_1 = \phi$). Likewise, the sequence $((\lambda-\delta)^{\frac{1}{2}}\,\|\phi_{2,n}\|_{L^2}^{-1}\,\phi_{2,n})_{n\in\mathbb{N}}$ being a minimizing sequence for $J^\infty_{\lambda-\delta}$, it cannot vanish. Therefore, there exists $\gamma > 0$, $R > 0$

and a sequence $(x_n)_{n\in\mathbb{N}}$ of points of $\mathbb{R}^3$ such that $\int_{x_n+B_R} |\phi_{2,n}|^2 \geq \gamma$. Then, defining $\widetilde{\phi}_{2,n} = (\lambda-\delta)^{\frac{1}{2}} \|\phi_{2,n}\|_{L^2}^{-1} \phi_{2,n}(\cdot - x_n)$,

$$\begin{cases} \forall n \in \mathbb{N}, \quad \widetilde{\phi}_{2,n} \in H^1(\mathbb{R}^3), \quad \displaystyle\int_{\mathbb{R}^3} \widetilde{\phi}_{2,n}^2 = \lambda - \delta \quad \text{and} \quad \widetilde{\phi}_{2,n} \geq 0 \text{ a.e. on } \mathbb{R}^3, \\ \displaystyle\lim_{n\to+\infty} E^\infty(\widetilde{\phi}_{2,n}) = J^\infty_{\lambda-\delta}, \\ -\text{div}\,(a_{2,n}\nabla\widetilde{\phi}_{2,n}) + (\rho_{\widetilde{\phi}_{2,n}} \star |\mathbf{r}|^{-1})\widetilde{\phi}_{2,n} + V^-_{2,n}\widetilde{\phi}^{1+2\beta_-}_{2,n} + V^+_{2,n}\widetilde{\phi}^{1+2\beta_+}_{2,n} + \theta_n\widetilde{\phi}_{2,n} \overset{H^{-1}}{\underset{n\to\infty}{\longrightarrow}} 0, \\ (\widetilde{\phi}_{2,n})_{n\in\mathbb{N}} \text{ converges to } \widetilde{\phi}_2 \neq 0 \text{ weakly in } H^1, \text{ strongly in } L^p_{\text{loc}} \text{ for } 2 \leq p < 6 \text{ and a.e. on } \mathbb{R}^3. \end{cases}$$

It is important to note that the sequences $(a_{j,n})_{n\in\mathbb{N}}$ and $(V^\pm_{j,n})_{n\in\mathbb{N}}$ for $j \in [\![1,2]\!]$, which we do not detail for their exact expression is not of use, are such that

$$\frac{a}{2} \leq a_{j,n} \leq \frac{b}{2} \quad \text{and} \quad \|V^\pm_{j,n}\|_{L^\infty} \leq 2^{\beta_+} C,$$

where the constants a, b and C are those arising in (2.31) and (2.33).

We can now apply the concentration-compactness lemma to $(\widetilde{\phi}_{1,n})_{n\in\mathbb{N}}$ and to $(\widetilde{\phi}_{2,n})_{n\in\mathbb{N}}$. As these sequences can't vanish, they are either compact or split into subsequences that are either compact or split, and so on. The next step consists in showing that this process necessarily terminates after a finite number of iterations. By contradiction, assume that it is not the case. We could then construct by repeated applications of the concentration-compactness lemma an infinity of sequences $(\widetilde{\psi}_{k,n})_{n\in\mathbb{N}}$, such that for all $k \in \mathbb{N}$

$$\begin{cases} \forall n \in \mathbb{N}, \quad \widetilde{\psi}_{k,n} \in H^1(\mathbb{R}^3), \quad \displaystyle\int_{\mathbb{R}^3} \widetilde{\psi}_{k,n}^2 = \delta_k \quad \text{and} \quad \widetilde{\psi}_{k,n} \geq 0 \text{ a.e. on } \mathbb{R}^3, \\ -\text{div}\,(\widetilde{a}_{k,n}\nabla\widetilde{\psi}_{k,n}) + (\rho_{\widetilde{\psi}_{k,n}} \star |\mathbf{r}|^{-1})\widetilde{\psi}_{k,n} + \widetilde{V}^-_{k,n}\widetilde{\psi}^{1+2\beta_-}_{k,n} + \widetilde{V}^+_{k,n}\widetilde{\psi}^{1+2\beta_+}_{k,n} + \theta_n\widetilde{\psi}_{k,n} \overset{H^{-1}}{\underset{n\to\infty}{\longrightarrow}} 0, \\ (\widetilde{\psi}_{k,n})_{n\in\mathbb{N}} \text{ converges to } \widetilde{\psi}_k \neq 0 \text{ weakly in } H^1, \text{ strongly in } L^p_{\text{loc}} \text{ for } 2 \leq p < 6 \text{ and a.e. on } \mathbb{R}^3, \end{cases}$$

with

$$\sum_{k\in\mathbb{N}} \delta_k \leq \lambda, \tag{2.84}$$

and $\forall k \in \mathbb{N}$, $\forall n \in \mathbb{N}$,

$$\frac{a}{2} \leq \widetilde{a}_{k,n} \leq \frac{b}{2} \quad \text{and} \quad \|\widetilde{V}^\pm_{k,n}\|_{L^\infty} \leq 2^{\beta_+} C.$$

Using Lemma 2.14 to pass to the limit with respect to n in the equation satisfied by $\widetilde{\psi}_{k,n}$, we obtain

$$-\text{div}\,(\widetilde{a}_k\nabla\widetilde{\psi}_k) + (\rho_{\widetilde{\psi}_k} \star |\mathbf{r}|^{-1})\widetilde{\psi}_k + \widetilde{V}^-_k\widetilde{\psi}^{1+2\beta_-}_k + \widetilde{V}^+_k\widetilde{\psi}^{1+2\beta_+}_k + \theta\widetilde{\psi}_k = 0, \tag{2.85}$$

with

$$\frac{a}{2} \leq \widetilde{a}_k \leq \frac{b^2}{2a} \quad \text{and} \quad \|\widetilde{V}^\pm_k\|_{L^\infty} \leq 2^{\beta_+} C.$$

Besides, we infer from (2.84) that $\displaystyle\sum_{k\in\mathbb{N}} \|\widetilde{\psi}_k\|^2_{L^2} \leq \lambda$, hence that

$$\lim_{k\to\infty} \|\widetilde{\psi}_k\|_{L^2} = 0.$$

It then easily follows from (2.85) that

$$\lim_{k\to\infty} \|\mathrm{div}\,(\widetilde{a}_k \nabla \widetilde{\psi}_k)\|_{L^2} = 0.$$

We can now make use of the elliptic regularity result [39] (see also the proof of Lemma 2.16) stating that there exists a constant C, depending only on the positive constants a and b, such that for all $k \in \mathbb{N}$

$$\|\widetilde{\psi}_k\|_{L^\infty} \leq C \left(\|\widetilde{\psi}_k\|_{L^2} + \|\mathrm{div}\,(\widetilde{a}_k \nabla \widetilde{\psi}_k)\|_{L^2} \right),$$

and obtain

$$\lim_{k\to\infty} \|\widetilde{\psi}_k\|_{L^\infty} = 0.$$

Lastly, we deduce from (2.85) that

$$\theta \|\widetilde{\psi}_k\|_{L^2}^2 \leq C \left(\|\widetilde{\psi}_k\|_{L^\infty}^{2\beta_-} + \|\widetilde{\psi}_k\|_{L^\infty}^{2\beta_+} \right) \|\widetilde{\psi}_k\|_{L^2}^2.$$

As $\|\widetilde{\psi}_k\|_{L^2} > 0$ for all $k \in \mathbb{N}$, we obtain that

$$\theta \leq C \left(\|\widetilde{\psi}_k\|_{L^\infty}^{2\beta_-} + \|\widetilde{\psi}_k\|_{L^\infty}^{2\beta_+} \right) \underset{k\to\infty}{\longrightarrow} 0,$$

which obviously contradicts (2.83). We therefore conclude from this analysis that, if dichotomy occurs, $(\phi_n)_{n\in\mathbb{N}}$ splits in a finite number, say K, of compact bits having mass $\delta_k > 0$ with $\sum_{k=1}^K \delta_k = \lambda$. We are now going to prove that this cannot be.

If this was the case, there would exist two sequences $(u_{1,n})_{n\in\mathbb{N}}$ and $(u_{2,n})_{n\in\mathbb{N}}$ such that

$$\begin{cases} \forall n \in \mathbb{N}, \quad u_{1,n} \in H^1(\mathbb{R}^3), \quad \displaystyle\int_{\mathbb{R}^3} |u_{1,n}|^2 = \delta_1, \quad u_1 \geq 0 \text{ a.e. on } \mathbb{R}^3, \\ \displaystyle\lim_{n\to\infty} E(u_{1,n}) = J_{\delta_1} \end{cases}$$

and

$$\begin{cases} \forall n \in \mathbb{N}, \quad u_{2,n} \in H^1(\mathbb{R}^3), \quad \displaystyle\int_{\mathbb{R}^3} |u_{2,n}|^2 = \delta_2, \quad u_2 \geq 0 \text{ a.e. on } \mathbb{R}^3, \\ \displaystyle\lim_{n\to\infty} E^\infty(u_{2,n}) = J_{\delta_2}, \end{cases}$$

and converging weakly in $H^1(\mathbb{R}^3)$ to u_1 and u_2 respectively, with $\|u_1\|_{L^2}^2 = \delta_1$ and $\|u_2\|_{L^2}^2 = \delta_2$ (as the weak limit of $(\phi_n)_{n\in\mathbb{N}}$ in $L^2(\mathbb{R}^3)$ is nonzero, one bit stays at finite distance from the nuclei). It then follows from Lemma 2.10 that u_1 and u_2 are minimizers for J_{δ_1} and $J_{\delta_2}^\infty$, and from Lemma 2.11 that $J_{\delta_1+\delta_2} < J_{\delta_1} + J_{\delta_2}^\infty$.

Applying (2.78) twice, we also have $J_\lambda = J_{\delta_1} + J_{\delta_2}^\infty + J_{\lambda-\delta_1-\delta_2}^\infty$, so that we infer $J_\lambda > J_{\delta_1+\delta_2} + J_{\lambda-\delta_1-\delta_2}^\infty$ which is a contradiction to Lemma 2.1. □

End of the proof of Lemma 2.13. As a consequence of the concentration-compactness lemma and of the first three assertions of Lemma 2.13, the sequence $(\phi_n)_{n\in\mathbb{N}}$ converges to ϕ weakly in $H^1(\mathbb{R}^3)$ and strongly in $L^p(\mathbb{R}^3)$ for all $2 \leq p < 6$. In particular,

$$\int_{\mathbb{R}^3} \phi^2 = \lim_{n\to\infty} \int_{\mathbb{R}^3} \phi_n^2 = \lambda.$$

It follows from Lemma 2.10 that ϕ is a minimizer to (2.57). □

Appendix

In this appendix, we state three technical lemmas, which we make use of in the proof of Theorem 2.3. These lemmas are concerned with second-order elliptic operators of the form $-\mathrm{div}\,(A\nabla\cdot)$. For the sake of generality, we deal with the case when A is a matrix-valued function, although A is a real-valued function in the two-electron GGA model.

For Ω an open subset of $\mathbb{R}^3$ and $0 < \lambda \leq \Lambda < \infty$, we denote by $M^s(\lambda, \Lambda, \Omega)$ the closed convex subset of $L^\infty(\Omega, \mathbb{R}^{3\times 3})$ consisting of the symmetric matrix fields $A \in L^\infty(\Omega, \mathbb{R}^{3\times 3})$ such that for almost all $x \in \Omega$,

$$\lambda \leq A(x) \leq \Lambda.$$

The first lemma is a H-convergence result which allows to pass to the limit in the Ekeland condition (2.74). We shall not give the proof, for it is very similar to the proofs that can be found in the original article by Murat and Tartar [81] . Recall that a sequence $(A_n)_{n\in\mathbb{N}}$ of elements of $M^s(\lambda, \Lambda, \Omega)$ is said to H-converge to some $A \in M^s(\lambda', \Lambda', \Omega)$, which is denoted by $A_n \rightharpoonup_{\mathrm{H}} A$, if for every $\omega \subset\subset \Omega$ the following property holds: $\forall f \in H^{-1}(\omega)$, the sequence $(u_n)_{n\in\mathbb{N}}$ of the elements of $H^1_0(\omega)$ such that $-\mathrm{div}(A_n\nabla u_n) = f|_\omega$ in $H^{-1}(\omega)$, satisfies

$$\begin{cases} u_n \rightharpoonup u \ \text{ weakly in } H^1_0(\omega), \\ A_n\nabla u_n \rightharpoonup A\nabla u \ \text{ weakly in } L^2(\omega) \end{cases}$$

where u is the solution in $H^1_0(\omega)$ to $-\mathrm{div}(A\nabla u) = f|_\omega$. It is known ([81]) that from any bounded sequence $(A_n)_{n\in\mathbb{N}}$ in $M^s(\lambda, \Lambda, \Omega)$ one can extract a subsequence which H-converges to some $A \in M^s(\lambda, \lambda^{-1}\Lambda^2, \Omega)$.

Lemma 2.14. *Let Ω be an open subset of $\mathbb{R}^3$, $0 < \lambda \leq \Lambda < \infty$, $0 < \lambda' \leq \Lambda' < \infty$, and $(A_n)_{n\in\mathbb{N}}$ a sequence of elements of $M^s(\lambda, \Lambda, \Omega)$ which H-converges to some $A \in M^s(\lambda', \Lambda', \Omega)$. Let $(u_n)_{n\in\mathbb{N}}$, $(f_n)_{n\in\mathbb{N}}$ and $(g_n)_{n\in\mathbb{N}}$ be sequences of elements of $H^1(\Omega)$, $H^{-1}(\Omega)$ and $L^2(\Omega)$ respectively, and $u \in H^1(\Omega)$, $f \in H^{-1}(\Omega)$ and $g \in L^2(\Omega)$ such that*

$$\begin{cases} -\mathrm{div}(A_n\nabla u_n) = f_n + g_n \ \text{ in } H^{-1}(\Omega) \text{ for all } n \in \mathbb{N}, \\ u_n \rightharpoonup u \text{ weakly in } H^1(\Omega), \\ f_n \to f \text{ strongly in } H^{-1}(\Omega), \\ g_n \rightharpoonup g \text{ weakly in } L^2(\Omega). \end{cases}$$

Then $-\mathrm{div}\,(A\nabla u) = f + g$ and $A_n\nabla u_n \rightharpoonup A\nabla u$ weakly in $L^2(\Omega)$.

The second lemma is an extension of [58, Lemma II.1] and of a classical result on the ground state of Schrödinger operators [75]. Recall that

$$L^2(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3) = \Big\{ \mathcal{W} \,|\, \forall \varepsilon > 0,\ \exists (\mathcal{W}_2, \mathcal{W}_\infty) \in L^2(\mathbb{R}^3) \times L^\infty(\mathbb{R}^3) \ \text{ s.t. } \|\mathcal{W}_\infty\|_{L^\infty} \leq \varepsilon,\ \mathcal{W} = \mathcal{W}_2 + \mathcal{W}_\infty \Big\}.$$

Lemma 2.15. *Let $0 < \lambda \leq \Lambda < \infty$, $A \in M^s(\lambda, \Lambda, \mathbb{R}^3)$, $W \in L^2(\mathbb{R}^3) + L^\infty_\varepsilon(\mathbb{R}^3)$ such that $W_+ = \max(0, W) \in L^2(\mathbb{R}^3) + L^3(\mathbb{R}^3)$ and μ a positive Radon measure on $\mathbb{R}^3$ such that $\mu(\mathbb{R}^3) < Z = \sum_{k=1}^M z_k$. Then,*

$$H = -\mathrm{div}\,(A\nabla\cdot) + V + \mu \star |\mathbf{r}|^{-1} + W$$

defines a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain

$$D(H) = \left\{ u \in H^1(\mathbb{R}^3) \,|\, \mathrm{div}\,(A\nabla u) \in L^2(\mathbb{R}^3) \right\}.$$

Besides, $D(H)$ is dense in $H^1(\mathbb{R}^3)$ and included in $L^\infty(\mathbb{R}^3) \cap C^{0,\alpha}(\mathbb{R}^3)$ for some $\alpha > 0$, and any function of $D(H)$ vanishes at infinity. In addition,

1. *H is bounded from below, $\sigma_{\mathrm{ess}}(H) \subset [0, \infty)$ and H has an infinite number of negative eigenvalues;*
2. *the lowest eigenvalue μ_1 of H is simple and there exists an eigenvector $u_1 \in D(H)$ of H associated with μ_1 such that $u_1 > 0$ on $\mathbb{R}^3$;*
3. *if $w \in D(H)$ is an eigenvector of H such that $w \geq 0$ on $\mathbb{R}^3$, then there exists $\alpha > 0$ such that $w = \alpha u_1$.*

The third lemma is used to prove that the ground state density of the GGA Kohn-Sham model exhibits exponential decay at infinity (at least for the two electron model considered in this chapter).

Lemma 2.16. *Let $0 < \lambda \leq \Lambda < \infty$, $A \in M^s(\lambda, \Lambda, \mathbb{R}^3)$, $\mathcal{V}$ a function of $L^{\frac{6}{5}}_{\mathrm{loc}}(\mathbb{R}^3)$ which vanishes at infinity, $\theta > 0$ and $u \in H^1(\mathbb{R}^3)$ such that*

$$-\mathrm{div}(A\nabla u) + \mathcal{V}u + \theta u = 0 \quad \text{in } \mathcal{D}'(\mathbb{R}^3).$$

Then there exists $\gamma > 0$ depending on $(\lambda, \Lambda, \theta)$ such that $e^{\gamma|\mathbf{r}|}u \in H^1(\mathbb{R}^3)$.

Proof of Lemma 2.15. The quadratic form q_0 on $L^2(\mathbb{R}^3)$ with domain $D(q_0) = H^1(\mathbb{R}^3)$, defined by

$$\forall (u, v) \in D(q_0) \times D(q_0), \quad q_0(u, v) = \int_{\mathbb{R}^3} A\nabla u \cdot \nabla v,$$

is symmetric and positive. It is also closed since the norm $\sqrt{\|\cdot\|^2_{L^2} + q_0(\cdot)}$ is equivalent to the usual H^1 norm. This implies that q_0 is the quadratic form of a unique self-adjoint operator H_0 on $L^2(\mathbb{R}^3)$, whose domain $D(H_0)$ is dense in $H^1(\mathbb{R}^3)$. It is easy to check that $D(H_0) = \left\{ u \in H^1(\mathbb{R}^3) \mid \mathrm{div}\,(A\nabla u) \in L^2(\mathbb{R}^3) \right\}$ and that $\forall u \in D(H_0)$, $H_0 u = -\mathrm{div}\,(A\nabla u)$. Using classical elliptic regularity results [39], we obtain that there exist two constants $0 < \alpha < 1$ and $C \in \mathbb{R}_+$ (depending on λ and Λ) such that for all regular bounded domains $\Omega \subset\subset \mathbb{R}^3$, and all $v \in H^1(\Omega)$ such that $\mathrm{div}\,(A\nabla v) \in L^2(\Omega)$,

$$\|v\|_{C^{0,\alpha}(\bar{\Omega})} := \sup_{\Omega} |v| + \sup_{(\mathbf{r},\mathbf{r}') \in \Omega \times \Omega} \frac{|v(\mathbf{r}) - v(\mathbf{r}')|}{|\mathbf{r} - \mathbf{r}'|^\alpha} \leq C \left(\|v\|_{L^2(\Omega)} + \|\mathrm{div}\,(A\nabla v)\|_{L^2(\Omega)} \right).$$

It follows that on the one hand, $D(H_0) \hookrightarrow L^\infty(\mathbb{R}^3) \cap C^{0,\alpha}(\mathbb{R}^3)$, with

$$\forall u \in D(H_0), \quad \|u\|_{L^\infty(\mathbb{R}^3)} + \sup_{(\mathbf{r},\mathbf{r}')\in\mathbb{R}^3\times\mathbb{R}^3} \frac{|v(\mathbf{r}) - v(\mathbf{r}')|}{|\mathbf{r}-\mathbf{r}'|^\alpha} \leq C\left(\|u\|_{L^2} + \|H_0 u\|_{L^2}\right), \quad (2.86)$$

and that on the other hand, any $u \in D(H_0)$ vanishes at infinity.

Let us now prove that the multiplication by $\mathcal{W} = V + \mu \star |\mathbf{r}|^{-1} + W$ defines a compact perturbation of H_0. For this purpose, we consider a sequence $(u_n)_{n\in\mathbb{N}}$ of elements of $D(H_0)$ bounded for the norm $\|\cdot\|_{H_0} = (\|\cdot\|_{L^2}^2 + \|H_0 \cdot\|_{L^2}^2)^{\frac{1}{2}}$. Up to extracting a subsequence, we can assume without loss of generality that there exists $u \in D(H_0)$ such that:

$$\begin{cases} u_n \rightharpoonup u \text{ in } H^1(\mathbb{R}^3) \text{ and } L^p(\mathbb{R}^3) \text{ for } 2 \leq p \leq 6, \\ u_n \to u \text{ in } L^p_{loc}(\mathbb{R}^3) \text{ with } 2 \leq p < 6 \text{ and } a.e. \end{cases}$$

Besides, it is then easy to check that the potential $\mathcal{W} = V + \mu \star |\mathbf{r}|^{-1} + W$ belongs to $L^2 + L^\infty_\varepsilon(\mathbb{R}^3)$. Let $\varepsilon > 0$ and $(\mathcal{W}_2, \mathcal{W}_\infty) \in L^2(\mathbb{R}^3) \times L^\infty(\mathbb{R}^3)$ such that $\|\mathcal{W}_\infty\|_{L^\infty} \leq \varepsilon$ and $\mathcal{W} = \mathcal{W}_2 + \mathcal{W}_\infty$. On the one hand, $\|\mathcal{W}_\infty(u_n - u)\|_{L^2} \leq 2\,\varepsilon \sup_{n\in\mathbb{N}} \|u_n\|_{H_0}$, and on the other hand $\lim_{n\to\infty} \|\mathcal{W}_2(u_n - u)\|_{L^2} = 0$. The latter result is obtained from Lebesgue's dominated convergence theorem, using the fact that it follows from (2.86) that $(u_n)_{n\in\mathbb{N}}$ is bounded in $L^\infty(\mathbb{R}^3)$. Consequently,

$$\lim_{n\to\infty} \|\mathcal{W}u_n - \mathcal{W}u\|_{L^2} = 0,$$

which proves that $\mathcal{W}$ is a H_0-compact operator. We can therefore deduce from Weyl's theorem that $H = H_0 + \mathcal{W}$ defines a self-adjoint operator on $L^2(\mathbb{R}^3)$ with domain $D(H) = D(H_0)$, and that $\sigma_{\rm ess}(H) = \sigma_{\rm ess}(H_0)$. As q_0 is positive, $\sigma(H_0) \subset \mathbb{R}_+$ and therefore $\sigma_{\rm ess}(H) \subset \mathbb{R}_+$.

Let us now prove that H has an infinite number of negative eigenvalues which form an increasing sequence converging to zero. First, H is bounded below since for all $v \in D(H)$ such that $\|v\|_{L^2} = 1$,

$$\begin{aligned} \langle v|H|v\rangle &= \int_{\mathbb{R}^3} A\nabla v \cdot \nabla v + \int_{\mathbb{R}^3} \mathcal{W}v^2 \geq \lambda \|\nabla v\|_{L^2}^2 - \|\mathcal{W}_2\|_{L^2} \|\nabla v\|_{L^2}^{\frac{3}{2}} - \varepsilon \\ &\geq -\frac{27}{256}\lambda^{-3}\|\mathcal{W}_2\|^4 - \varepsilon. \end{aligned}$$

In order to prove that H has at least N negative eigenvalues, including multiplicities, first notice that we have

$$H \leq -\Lambda\Delta + V + \mu \star |\mathbf{r}|^{-1} + W_+ \quad (2.87)$$

with $W_+ \in L^2(\mathbb{R}^3) + L^3(\mathbb{R}^3)$. It is proven in [58, Lemma II.1] that the operator in the right hand side of (2.87) has infinitely many eigenvalues including multiplicities. Therefore by the minimax principle, H also has infinitely many negative eigenvalues, including multiplicities.

The lowest eigenvalue of H, which we denote by μ_1, is characterized by

$$\mu_1 = \inf\left\{ \int_{\mathbb{R}^3} A\nabla u \cdot \nabla u + \int_{\mathbb{R}^3} \mathcal{W}|u|^2, \quad u \in H^1(\mathbb{R}^3), \quad \|u\|_{L^2} = 1 \right\}, \quad (2.88)$$

and the minimizers of (2.88) are exactly the set of the normalized eigenvectors of H associated with μ_1. Let u_1 be a minimizer (2.88). As for all $u \in H^1(\mathbb{R}^3)$, $|u| \in H^1(\mathbb{R}^3)$ and $\nabla|u| = \mathrm{sgn}(u)\nabla u$ a.e. on $\mathbb{R}^3$, $|u_1|$ also is a minimizer to (2.88). Up to replacing u_1 with $|u_1|$, there is therefore no restriction in assuming that $u_1 \geq 0$ on $\mathbb{R}^3$. We thus have

$$u_1 \in H^1(\mathbb{R}^3) \cap C^0(\mathbb{R}^3), \quad u_1 \geq 0 \quad \text{and} \quad -\operatorname{div}(A\nabla u_1) + g u_1 = 0$$

with $g = \mathcal{W} - \mu_1 \in L^p_{\mathrm{loc}}(\mathbb{R}^3)$ for some $p > \frac{3}{2}$ (take $p = 2$). A Harnack-type inequality due to Stampacchia [79] then implies that if u_1 has a zero in $\mathbb{R}^3$, then u_1 is identically zero. As $\|u_1\|_{L^2} = 1$, we therefore have $u_1 > 0$ on $\mathbb{R}^3$. Using classical arguments (see e.g. [75]), it is then not difficult to prove that μ_1 is simple. The proof of the third assertion of the Lemma then is straightforward. □

Proof of Lemma 2.16. Consider $R > 0$ large enough to ensure that $\frac{\theta}{2} \leq \mathcal{V}(\mathbf{r}) + \theta \leq \frac{3\theta}{2}$ a.e. on $B_R^c := \mathbb{R}^3 \setminus \overline{B}_R$. It is straightforward to see that u is the unique solution in $H^1(B_R^c)$ to the elliptic boundary problem

$$\begin{cases} -\operatorname{div}(A\nabla v) + \mathcal{V} v + \theta v = 0 & \text{in } B_R^c, \\ v = u & \text{on } \partial B_R. \end{cases}$$

Let $\gamma > 0$, $\tilde{u} = u \exp^{-\gamma(|\cdot| - R)}$ and $w = u - \tilde{u}$. The function w is in $H^1(\mathbb{R}^3)$ and is the unique solution in $H^1(B_R^c)$ to

$$\begin{cases} -\operatorname{div}(A\nabla w) + \mathcal{V} w + \theta w = \operatorname{div}(A\nabla \tilde{u}) - \mathcal{V}\tilde{u} - \theta\tilde{u} & \text{in } B_R^c, \\ w = 0 & \text{on } \partial B_R. \end{cases} \tag{2.89}$$

Let us now introduce the weighted Sobolev space $W_0^\gamma(B_R^c)$ defined by

$$W_0^\gamma(B_R^c) = \left\{ v \in H_0^1(B_R^c) \mid e^{\gamma|\cdot|} v \in H^1(B_R^c) \right\}$$

endowed with the inner product $(v, w)_{W_0^\gamma(B_R^c)} = \displaystyle\int_{B_R^c} e^{\gamma|\mathbf{r}|} (v(\mathbf{r}) w(\mathbf{r}) + \nabla v(\mathbf{r}) \cdot \nabla w(\mathbf{r}))\, d\mathbf{r}$.

Multiplying (2.89) by $\phi e^{2\gamma|\cdot|}$ with $\phi \in \mathcal{D}(B_R^c)$ and integrating by parts, we obtain

$$\begin{aligned} &\int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla w \cdot e^{\gamma|\mathbf{r}|}\nabla\phi + 2\gamma \int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla w \cdot \frac{\mathbf{r}}{|\mathbf{r}|} e^{\gamma|\mathbf{r}|}\phi + \int_{B_R^c} (\mathcal{V} + \theta) e^{\gamma|\mathbf{r}|} w e^{\gamma|\mathbf{r}|}\phi \\ &= -\int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla \tilde{u} \cdot e^{\gamma|\mathbf{r}|}\nabla\phi - 2\gamma \int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla \tilde{u} \cdot \frac{\mathbf{r}}{|\mathbf{r}|} e^{\gamma|\mathbf{r}|}\phi - \int_{B_R^c} (\mathcal{V} + \theta) e^{\gamma|\mathbf{r}|} \tilde{u} e^{\gamma|\mathbf{r}|}\phi. \end{aligned} \tag{2.90}$$

Due to the definitions of $W_0^\gamma(B_R^c)$ and $\tilde{u}$, (2.90) actually holds for $(w, \phi) \in W_0^\gamma(B_R^c) \times W_0^\gamma(B_R^c)$, and it is straightforward to see that (2.90) is a variational formulation equivalent to (2.89). It is also easy to check that the right-hand-side in (2.90) is a continuous form on $W_0^\gamma(B_R^c)$, so that we only have to prove the coercivity of the bilinear form in the left-hand-side of (2.90) to be able to apply Lax-Milgram lemma. We have for $v \in W_0^\gamma(B_R^c)$

$$\begin{aligned} &\int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla v \cdot e^{\gamma|\mathbf{r}|}\nabla v + 2\gamma \int_{B_R^c} A e^{\gamma|\mathbf{r}|}\nabla v \cdot \frac{\mathbf{r}}{|\mathbf{r}|} e^{\gamma|\mathbf{r}|} v + \int_{B_R^c} (\mathcal{V} + \theta) e^{\gamma|\mathbf{r}|} v e^{\gamma|\mathbf{r}|} v \\ &\geq \lambda \left\| e^{\gamma|\mathbf{r}|}\nabla v \right\|^2_{L^2(B_R^c)} - 2\Lambda\gamma \left\| e^{\gamma|\mathbf{r}|}\nabla v \right\|_{L^2(B_R^c)} \left\| e^{\gamma|\mathbf{r}|} v \right\|_{L^2(B_R^c)} + \frac{\theta}{2} \left\| e^{\gamma|\mathbf{r}|} v \right\|^2_{L^2(B_R^c)} \\ &\geq (\lambda - \Lambda\gamma) \left\| e^{\gamma|\mathbf{r}|}\nabla v \right\|^2_{L^2(B_R^c)} + (\frac{\theta}{2} - \Lambda\gamma) \left\| e^{\gamma|\mathbf{r}|} v \right\|^2_{L^2(B_R^c)}. \end{aligned}$$

Thus the bilinear form is clearly coercive if $\gamma < \min(\frac{\lambda}{\Lambda}, \frac{\theta}{2\Lambda})$, and there is a unique w solution of (2.89) in $W_0^\gamma(B_R^c)$ for such a γ. Now since $u = w + \tilde{u}$, it is clear that $e^{\gamma|\cdot|}u \in H^1(B_R^c)$, and then $e^{\gamma|\cdot|}u \in H^1(\mathbb{R}^3)$. □

Acknowledgements.

The authors are grateful to C. Le Bris, M. Lewin and F. Murat for helpful discussions.

Chapitre 3

A numerical approach related to defect-type theories for some weakly random problems in homogenization

Sommaire

3.1 Introduction 65
3.2 Some classical results of elliptic homogenization 69
3.2.1 Periodic homogenization 69
3.2.2 Stochastic homogenization 70
3.3 Homogenization of a randomly perturbed periodic material . . . 72
3.3.1 Presentation of the model 72
3.3.2 An ergodic approximation of the homogenized tensor 73
3.3.3 Convergence of the first-order term $A_1^{*,N}$ 76
3.3.4 Convergence of the second-order term $A_2^{*,N}$ 81
3.4 Numerical experiments 90
3.4.1 Methodology 90
3.4.2 Results 93
3.5 Appendix 101
3.5.1 One-dimensional computations 101
3.5.2 Some technical lemmas 104

3.1 Introduction

Composite materials are increasingly used in industry. For instance, modern aircrafts consist, for more than 50%, of composite materials. Generally speaking, composites are heterogeneous materials obtained by mixing two phases, a matrix and reinforcements (or inclusions). When appropriately designed, these materials outperform traditional materials, notably because they combine robustness and lightness. Their use however raises new challenges. The behavior of these materials under extreme conditions has to be predicted carefully, so as to avoid, in the worst case scenario, separation of the components (think of a plane hit by thunder). While it is possible to create an infinity of composites starting from the same elementary components, it is out of question to actually construct and experimentally test each and every possible combination. Characterizing *a priori* the

properties of a given composite material, not yet synthetized or assembled, is therefore instrumental.

A brute force numerical approach, consisting in directly solving the classical boundary value problems modelling the behavior of the material, is not practical. The heterogeneities indeed often occur at a scale ε much finer than the overall typical lengthscale (say, 1) of the material itself. A finite element mesh would, for example, need to be of size less than ε in order to capture the correct behavior. The number of degrees of freedom would then be proportional to ε^{-d} (where d denotes the dimension of the ambient physical space) and would yield, for ε small, a heavy computational cost one cannot necessarily afford.

The aim of homogenization is to provide a practical alternative to the brute force numerical approach. In a nutshell, homogenization consists in replacing a possibly complicated heterogeneous material with a homogeneous material sharing the same macroscopic properties. It allows for eliminating the fine scale, up to an error which is controlled by ε, the size of this fine scale as compared to the macroscopic size. Homogenization is a well-established theory (see [46] for a comprehensive textbook), which, in a simplified picture, can be seen as averaging partial differential equations that have highly-oscillating coefficients.

Of course, the structure of the material, and more precisely the way the constituents are combined, have a deep influence on the results of the homogenization process. The simplest possible situation is the periodic situation. At the fine scale, a unit cell is repeated in a periodic manner in all directions. Then, in simple cases (say, to be schematic and to fix the ideas, linear well-posed equations), the homogenized material is characterized only using the solution of simple problems on the unit cell, called the cell problems. The role of these cell problems is to encode the information of the micro-scale and convey it to the macro-scale. Related cases, such as locally periodic materials, can be treated similarly.

As Figure 3.1 shows, real life materials are however not often periodic. In particular because of uncertainties and flaws in the industrial process, composites often do not exhibit a perfect periodic structure, even though it was the original plan. A suitable way to account for this is to use random modelling. Although the mathematical theory for homogenization of random materials under classical assumptions (ergodicity and stationarity) is well known, the practice is quite involved. The cell problems are defined over the whole space $\mathbb{R}^d$ and not simply on a "unit" cell. The numerical approximation of such problems using Monte-Carlo type computations is incredibly costly: the cell problems are truncated on a bounded domain, many possible realizations of the materials are considered, averages are performed. Consequently, in the context of random modelling, the benefits of homogenization over the direct attack of the original composite material are arguable.

Our line of thoughts, and the approach we try to advocate here, are based on the following two-fold observation: classical random homogenization is costly but perhaps, in a number of situations, not necessary. A more careful examination of Figure 3.1 indeed shows that albeit not periodic, the material is not totally random. It may probably be

fairly considered as a perturbation of a periodic material. The homogenized behavior should expectedly be close to that of the underlying periodic material, up to a small error depending on the amount of randomness present.

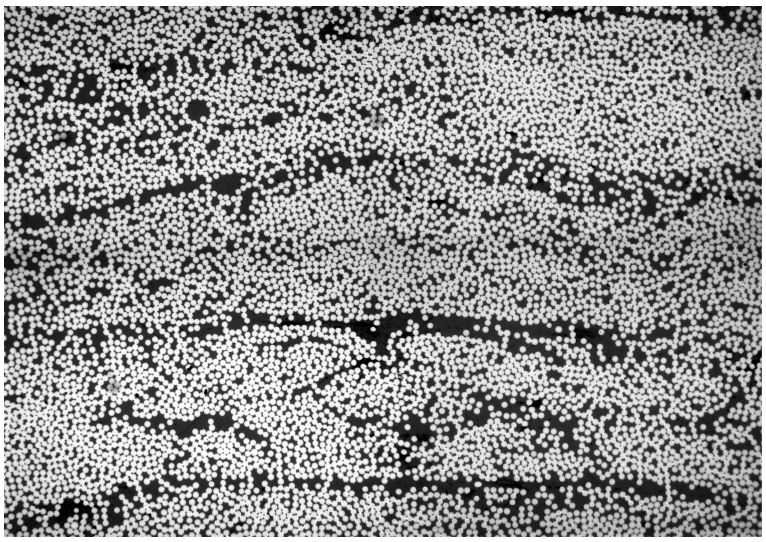

Figure 3.1: Two-dimensional cut of a composite material used in the aeronautics industry, extracted from [83] and reproduced with permission of the author. It is clear that this material is not periodic, yet there is some kind of an underlying periodic arrangement of the fibers.

The aim of this chapter is to give a practical example of theory following the above philosophy. We introduce and study a specific model for such a randomly perturbed periodic material, which we also call a *weakly random material*. More precisely, we are interested in the homogenization of the following elliptic problem

$$\begin{cases} -\operatorname{div}\left((A_{per}(\frac{x}{\varepsilon}) + b_\eta(\frac{x}{\varepsilon},\omega)C_{per}(\frac{x}{\varepsilon}))\nabla u_\varepsilon\right) = f(x) \text{ in } \mathcal{D} \subset \mathbb{R}^d, \\ u_\varepsilon = 0 \text{ on } \partial\mathcal{D}. \end{cases}$$

Here the tensor A_{per} models a reference $\mathbb{Z}^d$-periodic material which is randomly perturbed by the $\mathbb{Z}^d$-periodic tensor C_{per}, the stochastic perturbation being encoded in the stationary ergodic scalar field b_η. In the present chapter, the law of the random variable $b_\eta(x,\cdot)$ is a Bernoulli distribution with parameter η (that is, b_η is equal to 1 with probability η and 0 with probability $1-\eta$. Using an asymptotic analysis in terms of η, we will develop an homogenization theory for $A_\eta\left(\frac{x}{\varepsilon},\omega\right) = A_{per}\left(\frac{x}{\varepsilon}\right) + b_\eta\left(\frac{x}{\varepsilon},\omega\right)C_{per}$ based on the similar theory for $A_{per}\left(\frac{x}{\varepsilon}\right)$.

In short, let us say that the main result of this chapter is to formally derive an expansion

$$A^*_\eta = A^*_{per} + \eta\bar{A}^*_1 + \eta^2\bar{A}^*_2 + o(\eta^2),$$

where A^*_η and A^*_{per} are the homogenized tensors associated with A_η and A_{per} respectively. The first-order and second-order corrections $\bar{A}^*_1$ and $\bar{A}^*_2$ are obtained as limits, when N

goes to infinity, of sequences of tensors $A_1^{*,N}$ and $A_2^{*,N}$ computed on the supercell $[-\frac{N}{2}, \frac{N}{2}]^d$. It is the purpose of Propositions 3.2 and 3.4 to prove the convergence of $A_1^{*,N}$ and $A_2^{*,N}$ respectively. We stress that these corrections are achieved through purely deterministic computations.

The above setting is of course *one* possible setting where we may develop our theory, but not the only one. More general distributions are studied in Chapter 4 (see also [7]). Other forms of random perturbations of periodic problems, in the spirit of [18], could also be addressed. Moreover, we have deliberately considered the simplest possible equation (a scalar, linear second order elliptic equation in divergence form) to avoid any unnecessary technicalities and fundamental difficulties. Other equations could be considered, although it is not currently clear (to us, at least) how general our theory is in this respect.

With the ideas developed here (and originally introduced and further mentioned in [7, 8, 50]) and in Chapter 4, we work in the footsteps of many previous contributors who have considered perturbative approaches in homogenization. In [80] and [3], a deterministic setting in which an asymptotic expansion is assumed on the properties of the material (the latter being not necessarily periodic) is studied under the name "small amplitude homogenization". In [76], the case of a Gaussian perturbation with a small variance is addressed from a mechanical point of view. Our setting here is particular, because our random perturbation has order one in amplitude. It is only *in law* that the perturbation considered is small. The corrections obtained are therefore intrinsically different from those obtained in other settings (including settings we ourselves consider elsewhere, see Chapter 4 and [7]). Also, the present perturbative theory has unanticipated close connections with some classical defect-type theories used in solid state physics.

We emphasize that, contrary to what is presented in Chapter 4 for some other distributions, the theoretical results we obtain below in the Bernoulli case are only formal. We are unfortunately unable to fully justify our manipulations except in the one-dimensional case. Nevertheless, we can prove that the terms we obtain as first-order and second-order corrections are indeed finite and well defined. Our numerical results, on the other hand, show the efficiency of the approach. They somehow constitute a proof of the definite validity of our perturbative approach, although we wish to remain cautious. Note that due to the prohibitive cost of three-dimensional random homogenization problems, our tests are performed in dimension two.

This chapter is organized as follows. For the sake of consistency and the reader's convenience, we start by recalling in Section 3.2 some classical results of periodic and stochastic elliptic homogenization. Then we introduce our perturbative model in Section 3.3, and explain how we obtain the first-order and second-order correction by means of an ergodic approximation. Our elements of proof are exposed in Section 3.3. Our two-dimensional numerical tests are presented in Section 3.4. The Appendix contains explicit computations in the one-dimensional case as well as some useful technical lemmas.

Throughout this chapter, and unless otherwise mentioned, K denotes a constant that depends at most on the ambient dimension d, and on the tensors A_{per} and C_{per}. The

indices i and j denote indices in $[\![1,d]\!]$.

3.2 Some classical results of elliptic homogenization

We recall here some classical well-known results regarding linear elliptic periodic and stochastic homogenization. The reader familiar with homogenization theory can easily skip this section and directly proceed to Section 3.3.

3.2.1 Periodic homogenization

Consider A a $\mathbb{Z}^d$-periodic tensor field from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$, that is

$$\forall k \in \mathbb{Z}^d, A(x+k) = A(x) \text{ almost everywhere in } x \in \mathbb{R}^d.$$

We assume that $A \in L^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$ and A is coercive, which means that there exist $\lambda > 0$ and $\Lambda > 0$ such that

$$\forall \xi \in \mathbb{R}^d, \text{ a.e in } x \in \mathbb{R}^d, \ \lambda|\xi|^2 \leq A(x)\xi \cdot \xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|. \tag{3.1}$$

Consider now a material occupying a bounded domain $\mathcal{O} \subset \mathbb{R}^d$. The constitutive properties of this material are supposed to be periodic, the scale of periodicity being ε, and we assume that these properties are given by the tensor $A_\varepsilon(x) = A\left(\frac{x}{\varepsilon}\right)$.

We consider the following canonical elliptic problem: $f \in L^2(\mathcal{O})$ being given, find $u_\varepsilon \in H^1_0(\mathcal{O})$ solution to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \text{ in } \mathcal{O}, \\ u_\varepsilon = 0 \text{ on } \partial\mathcal{O}. \end{cases} \tag{3.2}$$

A direct numerical handling of (3.2) using finite elements has a heavy computational cost since the scale ε of the heterogeneities requires a fine mesh. The aim of homogenization is to take the limit $\varepsilon \to 0$ in (3.2) so as to replace the heterogeneous material with a homogeneous material. To this end, let us define the periodic cell problems on the unit cell $Q = [-\frac{1}{2}, \frac{1}{2}]^d$ by: $\forall i \in [\![1,d]\!]$,

$$\begin{cases} -\operatorname{div}(A(\nabla w_i + e_i)) = 0 \quad \text{in } Q, \\ w_i \ \ \mathbb{Z}^d - \text{periodic}, \end{cases} \tag{3.3}$$

where e_i is the i-th canonical vector of $\mathbb{R}^d$. Problem (3.3) has a solution unique up to the addition of a constant, in the space of functions in $H^1_{loc}(\mathbb{R}^d)$ that are $\mathbb{Z}^d$-periodic. Note that the number of cell problems is equal to the dimension of the space.

The homogenized tensor A^* is then given by:

$$\forall (i,j) \in [\![1,d]\!]^2, \ \ A^*_{ji} = \int_Q A(\nabla w_i + e_i) \cdot e_j. \tag{3.4}$$

Using (3.3), it also holds

$$A^*_{ji} = \int_Q A(\nabla w_i + e_i) \cdot (\nabla w_j + e_j).$$

Notice that in this periodic setting A^* is a constant matrix.

Finally, let us define the homogenized solution u_0 as the unique solution in $H^1_0(\mathcal{O})$ to

$$\begin{cases} -\operatorname{div}(A^* \nabla u_0) = f \text{ in } \mathcal{O}, \\ u_0 = 0 \text{ on } \partial\mathcal{O}. \end{cases} \tag{3.5}$$

Solving (3.3) and (3.5) is much simpler than directly solving (3.2) for the fine scale ε has disappeared. It is well-known (see [46] for instance) that

$$u_\varepsilon \underset{\varepsilon \to 0}{\to} u_0 \quad \text{in } L^2(\mathcal{O}) \tag{3.6}$$

and

$$u_\varepsilon - u_0 - \varepsilon \sum_{i=1}^{d} w_i \left(\frac{\cdot}{\varepsilon}\right) \frac{\partial u_0}{\partial x_i} \underset{\varepsilon \to 0}{\to} 0 \quad \text{in } H^1(\mathcal{O}). \tag{3.7}$$

The functions w_i are also called the correctors, since they allow for the *strong* convergence in (3.7). Convergences (3.6) and (3.7) show the relevance of the homogenization process: u_ε can be replaced by u_0 or more accurately $u_0 + \varepsilon \sum_{i=1}^{d} w_i \left(\frac{x}{\varepsilon}\right) \frac{\partial u_0}{\partial x_i}(x)$, which are easier to compute.

3.2.2 Stochastic homogenization

Throughout this chapter, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a probability space, $\mathbb{P}$ the probability measure and $\omega \in \Omega$ an event. We denote by $\mathbb{E}(X)$ the expectation of a random variable X.

We assume that the group $(\mathbb{Z}^d, +)$ acts on Ω and denote by $\tau_k, k \in \mathbb{Z}^d$, the group action. We also assume that this action is measure-preserving, that is,

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

and ergodic:

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

We call $F \in L^1_{loc}(\mathbb{R}^d, L^1(\Omega))$ stationary if

$$\forall k \in \mathbb{Z}^d,\ F(x + k, \omega) = F(x, \tau_k \omega) \text{ almost everywhere in } x \in \mathbb{R}^d \text{ and } \omega \in \Omega. \tag{3.8}$$

Notice that the notion of stationarity we use here is *discrete*: the shifts in (3.8) are assumed to be integers. This is related to our wish to connect the random problems considered with some underlying periodic problems. Notice also that for a deterministic F,

stationarity amounts to $\mathbb{Z}^d$-periodicity.

Consider a stationary tensor field $A(x,\omega) \in L^\infty(\mathbb{R}^d \times \Omega, \mathbb{R}^{d\times d})$, such that (3.1) is almost surely satisfied by $A(\cdot,\omega)$, and a material occupying a bounded domain $\mathcal{O} \subset \mathbb{R}^d$ modeled by $A\left(\frac{x}{\varepsilon},\omega\right)$.

We are interested in solving, for a deterministic function f,

$$\begin{cases} -\operatorname{div}\left(A\left(\dfrac{x}{\varepsilon},\omega\right)\nabla u_\varepsilon\right) = f(x) \text{ in } \mathcal{O}, \\ u_\varepsilon = 0 \text{ on } \partial\mathcal{O}. \end{cases} \tag{3.9}$$

In order to describe the behavior of u_ε, we again need to define cell problems. Here they read (see [46]):

$$\begin{cases} -\operatorname{div}\left(A(x,\omega)(\nabla w_i + e_i)\right) = 0 \quad \text{in } \mathbb{R}^d, \\ \nabla w_i \text{ stationary}, \quad \mathbb{E}\left(\displaystyle\int_Q \nabla w_i\right) = 0. \end{cases} \tag{3.10}$$

Problem (3.10) has a solution unique up to the addition of a (possibly random) constant in the space

$$\{w \in L^2_{loc}(\mathbb{R}^d, L^2(\Omega)), \quad \nabla w \in L^2_{unif}(\mathbb{R}^d, L^2(\Omega))\}.$$

We have denoted above by L^2_{unif} the space of functions for which the L^2 norm on a ball of unit size is bounded independently of the center of the ball.

Then we define the homogenized tensor A^* by

$$\forall (i,j) \in [\![1,d]\!]^2, A^*_{ji} = \mathbb{E}\left(\int_Q A(y,\omega)(e_i + \nabla_y w_i(y,\omega)) \cdot e_j dy\right). \tag{3.11}$$

Notice that A^* is deterministic and constant throughout the domain $\mathcal{O}$. The homogenized field u_0, which gives the asymptotic behavior of u_ε (in a sense similar to (3.6) and (3.7)), is also deterministic. It is the unique solution in $H^1_0(\mathcal{O})$ to

$$\begin{cases} -\operatorname{div}\left(A^*\nabla u_0\right) = f \text{ in } \mathcal{O}, \\ u_0 = 0 \text{ on } \partial\mathcal{O}. \end{cases}$$

The computation of the stochastic cell problems (3.10) is not an easy task since the problems are posed in an infinite domain ($\mathbb{R}^d$) with a stationarity condition. As we have seen in the previous paragraph, when the material is periodic, the cell problems (3.10) reduce to the deterministic cell problems (3.3) which are $\mathbb{Z}^d$-periodic and can thus be computed on the unit cell Q. Consequently, when the material under consideration is a stochastic perturbation of a reference periodic material, we expect the computation of the homogenized tensor to be tractable, up to an approximation. This is our motivation for proposing a perturbative approach.

3.3 Homogenization of a randomly perturbed periodic material

3.3.1 Presentation of the model

In the stochastic framework (3.9)-(3.10)-(3.11), we now specifically consider the following tensor field in $\mathbb{R}^d \times \Omega$:

$$A_\eta(x,\omega) = A_{per}(x) + b_\eta(x,\omega)C_{per}(x). \tag{3.12}$$

Here A_{per} and C_{per} are two deterministic $\mathbb{Z}^d$-periodic tensor fields. Intuitively, A_{per} is the reference material perturbed by C_{per}. The random character of the perturbation is encoded in the stationary ergodic scalar field b_η, upon which we assume the expression

$$b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x) B^k_\eta(\omega),$$

where the B^k_η are independent random variables having Bernoulli distribution with parameter η, meaning $B^k_\eta = 0$ with probability $1-\eta$ and $B^k_\eta = 1$ with probability η.

It is clear that as $\eta \to 0$ the perturbation becomes a rare event. However, the realization of this event modifies the microscopic structure of the material since it replaces, in a given cell, A_{per} with $A_{per} + C_{per}$.

We additionally assume that there exist $0 < \alpha \leq \beta$ such that for all $\xi \in \mathbb{R}^d$ and almost all $x \in \mathbb{R}^d$,

$$\alpha|\xi|^2 \leq A_{per}(x)\xi\cdot\xi, \qquad \alpha|\xi|^2 \leq \left(A_{per}+C_{per}\right)(x)\xi\cdot\xi, \tag{3.13}$$

$$|A_{per}(x)\xi| \leq \beta|\xi|, \qquad |\left(A_{per}+C_{per}\right)(x)\xi| \leq \beta|\xi|. \tag{3.14}$$

We can therefore use for every $0 \leq \eta \leq 1$ the stochastic homogenization results recalled in Section 3.2. The cell problems associated with (3.12) read, for $1 \leq i \leq d$,

$$\begin{cases} -\operatorname{div}\left(A_\eta(\nabla w_i^\eta + e_i)\right) = 0 \quad \text{in } \mathbb{R}^d, \\ \nabla w_i^\eta \ \text{ stationary}, \ \ \mathbb{E}\left(\displaystyle\int_Q \nabla w_i^\eta\right) = 0, \end{cases} \tag{3.15}$$

and the homogenized tensor A^*_η is given by

$$A^*_\eta e_i = \mathbb{E}\left(\int_Q A_\eta(\nabla w_i^\eta + e_i)\right), \ \text{ for } 1 \leq i \leq d. \tag{3.16}$$

Throughout the rest of this chapter we denote by w_i^0 the solution to the i-th cell problem (3.3) associated with A_{per}.

Because of the specific form of A_η, and more precisely because A_η converges strongly to A_{per} in $L^2(Q\times\Omega)$ as $\eta \to 0$, it is easy to see that:

Lemma 3.1. *When $\eta \to 0$, $A^*_\eta \to A^*_{per}$.*

Proof. Fix $1 \leq i \leq d$. We start by proving that ∇w_i^η converges strongly in $L^2(Q \times \Omega)$ to ∇w_i^0. Indeed, define $r_i^\eta = w_i^\eta - w_i^0$ solution to

$$\begin{cases} -\operatorname{div}\left(A_\eta \nabla r_i^\eta\right) = \operatorname{div}\left(b_\eta C_{per}\left(\nabla w_i^0 + e_i\right)\right) & \text{in } \mathbb{R}^d, \\ \nabla r_i^\eta \text{ stationary}, \quad \mathbb{E}\left(\displaystyle\int_Q \nabla r_i^\eta\right) = 0. \end{cases} \tag{3.17}$$

Standard cut-off and ergodicity arguments (see e.g the proof of Proposition 3.1 in [18]) show that

$$\begin{aligned} \|\nabla r_i^\eta\|_{L^2(Q\times\Omega)} &\leq \frac{1}{\alpha}\|b_\eta C_{per}\left(\nabla w_i^0 + e_i\right)\|_{L^2(Q\times\Omega)} \\ &= \frac{1}{\alpha}\|B_\eta^0\|_{L^2(\Omega)}\|C_{per}\left(\nabla w_i^0 + e_i\right)\|_{L^2(Q)} \\ &= \frac{1}{\alpha}\sqrt{\eta}\|C_{per}\left(\nabla w_i^0 + e_i\right)\|_{L^2(Q)}, \end{aligned}$$

where α is defined in (3.13), so that $\nabla w_i^\eta \underset{\eta\to 0}{\to} \nabla w_i^0$ in $L^2(Q \times \Omega)$.

Next, it is straightforward to see that A_η converges strongly to A_{per} in $L^2(Q \times \Omega)$. We deduce from these two strong convergences that

$$A^*_\eta e_i = \mathbb{E}\left(\int_Q A_\eta(x,\omega)(\nabla w_i^\eta + e_i)\right) \underset{\eta\to 0}{\to} \int_Q A_{per}(\nabla w_i^0 + e_i) = A^*_{per} e_i.$$

This concludes the proof.

□

Our goal is now to find an asymptotic expansion for A_η with respect to η up to the second order.

3.3.2 An ergodic approximation of the homogenized tensor

We consider a specific realization $\tilde{\omega} \in \Omega$ of the tensor A_η in the truncated domain $I_N = [-\frac{N}{2}, \frac{N}{2}]^d$, with (for simplicity) N an odd integer, and solve the following "supercell" problem:

$$\begin{cases} -\operatorname{div}\left(A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}} + e_i)\right) = 0 \quad \text{in } I_N, \\ w_i^{\eta,N,\tilde{\omega}} \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{3.18}$$

Then an easy adaptation of Theorem 1 of [22], stated in the continuous stationary setting, to our discrete stationary setting, shows that when N goes to infinity,

$$\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x) + e_i)dx \text{ converges to } A^*_\eta e_i \text{ almost surely in } \tilde{\omega} \in \Omega. \tag{3.19}$$

Since $\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x)+e_i)dx$ is the tensor obtained by periodic homogenization of the tensor $A_\eta(x,\tilde{\omega})$ in the supercell I_N, it is also well-known (see [46]) that the following bounds hold for all $(i,j)\in[\![1,d]\!]^2$:

$$\frac{1}{N^d}\left(\int_{I_N} A_\eta^{-1}(x,\tilde{\omega})dx\right)^{-1} e_i\cdot e_j \leq \frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x)+e_i)\cdot e_j dx \leq \frac{1}{N^d}\left(\int_{I_N} A_\eta(x,\tilde{\omega})dx\right) e_i\cdot e_j.$$

As a result, for all N in $2\mathbb{N}+1$, for all $0\leq\eta\leq 1$ and almost all $\tilde{\omega}$ in Ω,

$$\left|\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x)+e_i)\cdot e_j dx\right| \leq \beta, \tag{3.20}$$

where β is defined in (3.14). We then deduce from (3.19), (3.20) and the Lebegue dominated convergence theorem that

$$\forall 1\leq i\leq d,\quad A_\eta^* e_i = \lim_{N\to+\infty}\frac{1}{N^d}\mathbb{E}\left(\int_{I_N} A_\eta(x,\omega)(\nabla w_i^{\eta,N,\omega}(x)+e_i)\right)dx. \tag{3.21}$$

Remark 3.1. *A similar result holds for homogeneous Dirichlet and Neumann boundary conditions instead of periodic conditions in the definition (3.18) of $w_i^{\eta,N,\tilde{\omega}}$ (see [22] for more details).*

Using now the fact that b_η has a Bernoulli distribution in each cell of $\mathbb{Z}^d$, it is a simple matter to count the events and to make (3.21) more precise. We first define the set

$$\mathcal{T}_N = \left\{k\in\mathbb{Z}^d, Q+k\subset I_N\right\} = \left[\!\left[-\frac{N-1}{2},\frac{N-1}{2}\right]\!\right]^d. \tag{3.22}$$

The cardinal of $\mathcal{T}_N$ is of course N^d, and $\bigcup_{k\in\mathcal{T}_N}\{Q+k\} = I_N$.

We then have the following possible values for A_η:

- $A_\eta(x,\tilde{\omega}) = A_{per}$ with probability $(1-\eta)^{N^d}$.
  In this case $w_i^{\eta,N,\tilde{\omega}} = w_i^0$ solves the usual periodic cell problem:
  $$\begin{cases} -\operatorname{div}\left(A_{per}(\nabla w_i^0+e_i)\right)=0 & \text{in } Q,\\ w_i^0\ \mathbb{Z}^d-\text{periodic.}\end{cases}$$

- $A_\eta(x,\tilde{\omega}) = A_{per}+\mathbb{1}_{\{Q+k\}}C_{per}$ for $k\in\mathcal{T}_N$, with probability $\eta(1-\eta)^{N^d-1}$.
  In this case $w_i^{\eta,N,\tilde{\omega}} = w_i^{1,k,N}$ solves the following problem, which we call here a "one defect" supercell problem:
  $$\begin{cases} -\operatorname{div}\left(\left(A_{per}+\mathbb{1}_{\{Q+k\}}C_{per}\right)(\nabla w_i^{1,k,N}+e_i)\right)=0 & \text{in } I_N,\\ w_i^{1,k,N}\ (N\mathbb{Z})^d-\text{periodic.}\end{cases} \tag{3.23}$$

- $A_\eta(x,\tilde{\omega}) = A_{per} + \mathbb{1}_{\{Q+l\}\cup\{Q+m\}}C_{per}$ for $(l,m) \in \mathcal{T}_N$, $l \neq m$, with probability $\eta^2(1-\eta)^{N^d-2}$.

  In this case $w_i^{\eta,N,\tilde{\omega}} = w_i^{2,l,m,N}$ solves the following problem, which we call here a "two defects" supercell problem:

$$\begin{cases} -\operatorname{div}\left(\left(A_{per} + \mathbb{1}_{\{Q+l\}\cup\{Q+m\}}C_{per}\right)(\nabla w_i^{2,l,m,N} + e_i)\right) = 0 \quad \text{in } I_N, \\ w_i^{2,l,m,N} \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{3.24}$$

All the other possible values for A_η, which are of probability less than η^3 and which we will not use in this chapter, can be obtained using similar computations.

An instance of a setting with zero, one and two defects is shown in Figure 3.2 in the two-dimensional case of a material A_{per} consisting of a lattice of inclusions.

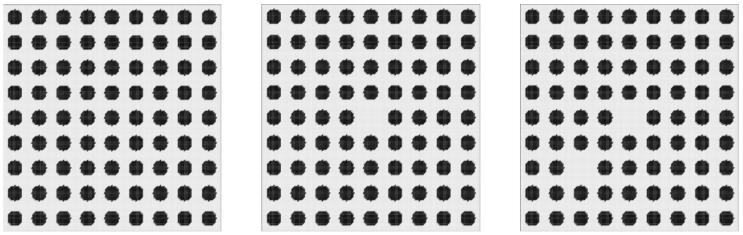

Figure 3.2: From left to right: zero defect, one defect and two defects.

Let us define $A_1^k = A_{per} + \mathbb{1}_{\{Q+k\}}C_{per}$ and $A_2^{l,m} = A_{per} + \mathbb{1}_{\{Q+l\}\cup\{Q+m\}}C_{per}$. Then (3.21) reads

$$A_\eta^* e_i = \lim_{N\to\infty}\left(\frac{(1-\eta)^{N^d}}{N^d}\int_{I_N} A_{per}(\nabla w_i^0 + e_i) + \sum_{k\in\mathcal{T}_N}\frac{\eta(1-\eta)^{N^d-1}}{N^d}\int_{I_N} A_1^k(\nabla w_i^{1,k,N} + e_i)\right.$$
$$\left. + \sum_{l,m\in\mathcal{T}_N, l\neq m}\frac{\eta^2(1-\eta)^{N^d-2}}{2N^d}\int_{I_N} A_2^{l,m}(\nabla w_i^{2,l,m,N} + e_i) + \cdots\right).$$

It is clear, by $(N\mathbb{Z})^d$-periodicity, that $\displaystyle\int_{I_N} A_1^k(\nabla w_i^{1,k,N} + e_i)$ does not depend on the position $k \in \mathcal{T}_N$ of the defect. Likewise, $\displaystyle\int_{I_N} A_2^{l,m}(\nabla w_i^{2,l,m,N} + e_i)$ only depends on the vector $m - l$. Thus we can rewrite

$$A_\eta^* e_i = \lim_{N\to\infty}\left((1-\eta)^{N^d} A_{per}^* e_i + \eta(1-\eta)^{N^d-1}\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\right.$$
$$\left. + \sum_{k\in\mathcal{T}_N\setminus\{0\}}\frac{\eta^2(1-\eta)^{N^d-2}}{2}\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i) + \cdots\right). \tag{3.25}$$

This is of the form

$$\begin{aligned} A_\eta^* &= \lim_{N\to\infty} \sum_{p=0}^{N^d} \eta^p A_p^{*,N} \\ &= \lim_{N\to\infty} \left(A_0^{*,N} + \eta A_1^{*,N} + \eta^2 A_2^{*,N} + o_N(\eta^2) \right), \end{aligned} \tag{3.26}$$

where the remainder $o_N(\eta^2)$ depends on N.

Explicitly expanding the polynomials in η up to the second-order in (3.25), we obtain:

$$A_0^{*,N} = A_{per}^*, \tag{3.27}$$

$$A_1^{*,N} e_i = \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i) - \int_{I_N} A_{per}(\nabla w_i^0 + e_i), \tag{3.28}$$

$$\begin{aligned} A_2^{*,N} e_i = \frac{1}{2} \sum_{k\in\mathcal{T}_N\setminus\{0\}} \Bigg(\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i) - 2\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i) \\ + \int_{I_N} A_{per}(\nabla w_i^0 + e_i) \Bigg) \end{aligned} \tag{3.29}$$

as the first three coefficients in (3.26).

Remark 3.2. *The structure of $A_p^{*,N}$ for $p \in \mathbb{N}$ is obviously related to that of the polynomial $(1-x)^p$.*

Our approach consists in formally exchanging the limits $N \to \infty$ and $\eta \to 0$ in (3.26). In the next section, we show that $A_1^{*,N}$ is a converging sequence when $N \to \infty$. The case of $A_2^{*,N}$, which is also shown to be a converging sequence, is discussed in Section 3.3.4.

We are not able to prove, though, that $A_\eta^* - \lim_{N\to\infty} (A_{per}^* - \eta A_1^{*,N} - \eta^2 A_2^{*,N}) = o(\eta^2)$ with a remainder term $o(\eta^2)$ independent of N.

Remark 3.3. *The expression of $A_1^{*,N}$ (and likewise $A_2^{*,N}$) is reminiscent of standard expressions in solid state theory: each of the two integrals in the definition (3.28) of $A_1^{*,N}$ scales as the volume N^d of the domain I_N, and a priori needs to be renormalized in order to give a finite limit. The difference however has a finite limit without renormalization. In solid state physics, it is common to substract a jellium, that is, a uniform background, and proceed similarly.*

3.3.3 Convergence of the first-order term $A_1^{*,N}$

We study here the convergence, as N goes to infinity, of $A_1^{*,N}$ defined by (3.28), and prove:

Proposition 3.2. *$A_1^{*,N}$ converges to a finite limit $\bar{A}_1^*$ in $\mathbb{R}^{d\times d}$ when $N \to \infty$.*

Proof. We fix (i,j) in $[\![1,d]\!]^2$ and study the convergence of $A_1^{*,N}e_i \cdot e_j$.

Let us define the adjoint problems to the cell problems (3.3):

$$\begin{cases} -\operatorname{div}\left(A_{per}^T(\nabla\tilde{w}_j^0 + e_j)\right) = 0 \quad \text{in } Q, \\ \tilde{w}_j^0 \ \ \mathbb{Z}^d - \text{periodic}, \end{cases} \tag{3.30}$$

where we have denoted by A_{per}^T the transposed matrix of A_{per}. Then using (3.23) and the definition of A_1^0, we have

$$\begin{aligned} \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j &= \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0) \\ &= \int_{I_N} A_{per}(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0) \\ &\quad + \int_Q C_{per}(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0). \end{aligned}$$

Next, using (3.30), we note that

$$\begin{aligned} \int_{I_N} A_{per}(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0) &= \int_{I_N} (\nabla w_i^{1,0,N} + e_i)\cdot A_{per}^T(e_j + \nabla\tilde{w}_j^0) \\ &= \int_{I_N} e_i \cdot A_{per}^T(e_j + \nabla\tilde{w}_j^0) \\ &= N^d \left(A_{per}^T\right)^* e_j \cdot e_i, \end{aligned}$$

and applying (3.4) to the periodic tensor A_{per}^T and noticing that $(A_{per}^T)^* = (A_{per}^*)^T$, we obtain

$$\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j = N^d A_{per}^* e_i \cdot e_j + \int_Q C_{per}(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0). \tag{3.31}$$

Since, by definition,

$$\begin{aligned} A_1^{*,N}e_i &= \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i) - \int_{I_N} A_{per}(\nabla w_i^0 + e_i) \\ &= \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i) - N^d A_{per}^* e_i, \end{aligned}$$

we deduce from (3.31) that

$$A_1^{*,N}e_i \cdot e_j = \int_Q C_{per}(\nabla w_i^{1,0,N} + e_i)\cdot(e_j + \nabla\tilde{w}_j^0). \tag{3.32}$$

We now define

$$q_i^{1,0,N} = w_i^{1,0,N} - w_i^0, \tag{3.33}$$

which solves

$$\begin{cases} -\operatorname{div}\left(A_1^0 \nabla q_i^{1,0,N}\right) = \operatorname{div}(\mathbb{1}_Q C_{per}(\nabla w_i^0 + e_i)) \quad \text{in } I_N, \\ q_i^{1,0,N}\ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{3.34}$$

We deduce from Lemma 3.6 of the Appendix, applied to (3.34), that $\nabla q_i^{1,0,N}$ converges in $L^2_{loc}(\mathbb{R}^d)$, when $N \to +\infty$, to $\nabla q_i^{1,0,\infty}$, where $q_i^{1,0,\infty}$ is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A_1^0 \nabla q_i^{1,0,\infty}\right) = \operatorname{div}(\mathbb{1}_Q C_{per}(\nabla w_i^0 + e_i)) \quad \text{in } \mathbb{R}^d, \\ \nabla q_i^{1,0,\infty} \in L^2(\mathbb{R}^d). \end{cases} \tag{3.35}$$

Defining $w_i^{1,0,\infty} = w_i^0 + q_i^{1,0,\infty}$, it is clear that $\nabla w_i^{1,0,N}$ converges in $L^2(Q)$ to $\nabla w_i^{1,0,\infty}$. It follows from (3.32) that $A_1^{*,N} \underset{N\to+\infty}{\to} \bar{A}_1^*$ with $\bar{A}_1^*$ defined by

$$\forall (i,j) \in [\![1,d]\!]^2, \quad \bar{A}_1^* e_i \cdot e_j \;=\; \int_Q C_{per}(\nabla w_i^{1,0,\infty} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0). \tag{3.36}$$

□

Remark 3.4. *We stress that the expressions above, and in particular (3.36), bear formal resemblance with the results obtained in a deterministic setting in [13], [17], [26], [28]. In these papers, broadly speaking, small inclusions of size ε are put in a medium, and the solution of a given boundary value problem posed in the perturbed medium, say v_ε, is compared to the solution v of the same problem in the perfect medium. An asymptotic expansion for the difference $v_\varepsilon - v$ is derived. The first-order term is written in function of a mathematical object called a polarization tensor. Even though our approach and our applications are different, we can likewise introduce a polarization tensor to define the first-order correction (3.36) and thus make the links between our work and those mentioned above explicit.*

The computation of $\bar{A}_1^*$ requires to solve (3.35) which is defined in $\mathbb{R}^d$, but, in sharp contrast to the stochastic cell problems (3.15), is deterministic and has a right-hand side with compact support in $\mathbb{R}^d$. In practice, problem (3.35) is truncated on I_N. The following result gives insight on the truncation error.

Lemma 3.3. *Assume that $d \geq 3$ and that the unit cell Q contains an inclusion D, the boundary of which has regularity $\mathcal{C}^{1,\mu}$ for some $0 < \mu < 1$, and such that $\operatorname{dist}(D, \partial Q) > 0$. Assume also that A_{per} is Hölder continuous in $\overline{D}$ and in $\overline{Q \backslash D}$. Then there exists a tensor $B_1^{*,N}$, computed on I_N, and a constant K independent of N such that*

$$|B_1^{*,N} - \bar{A}_1^*| \leq K N^{-d}.$$

Proof. **Step 1.**

Fix (i,j) in $[\![1,d]\!]^2$. We first define the adjoint problem for (3.34), namely

$$\begin{cases} -\operatorname{div}\left((A_1^0)^T \nabla \tilde{q}_j^{1,0,N}\right) = \operatorname{div}(\mathbb{1}_Q C_{per}^T(\nabla \tilde{w}_j^0 + e_j)) \quad \text{in } I_N, \\ \tilde{q}_j^{1,0,N}\ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{3.37}$$

Applying Lemma 3.6 to (3.37), we also introduce the limit $\tilde{q}_j^{1,0,\infty}$ of $\tilde{q}_j^{1,0,N}$ when $N \to \infty$. It solves the adjoint problem of (3.35).

Then, using (3.37), we obtain

$$\begin{aligned}\int_Q C_{per}\nabla q_i^{1,0,N}\cdot(e_j+\nabla\tilde{w}_j^0) &= \int_{I_N}\nabla q_i^{1,0,N}\cdot \mathbb{1}_Q C_{per}^T(e_j+\nabla\tilde{w}_j^0)\\ &= -\int_{I_N}\nabla q_i^{1,0,N}\cdot(A_1^0)^T\nabla\tilde{q}_j^{1,0,N}\\ &= -\int_{I_N}A_1^0\nabla q_i^{1,0,N}\cdot\nabla\tilde{q}_j^{1,0,N}.\end{aligned}$$

Consequently, (3.32) and the definition (3.33) of $q_i^{1,0,N}$ yield

$$A_1^{*,N}e_i\cdot e_j = \int_Q C_{per}(\nabla w_i^0+e_i)\cdot(e_j+\nabla\tilde{w}_j^0) - \int_{I_N}A_1^0\nabla q_i^{1,0,N}\cdot\nabla\tilde{q}_j^{1,0,N}. \quad (3.38)$$

We know from Lemma 3.6 applied to (3.34) and (3.37) that the functions $\mathbb{1}_{I_N}\nabla q_i^{1,0,N}$ and $\mathbb{1}_{I_N}\nabla\tilde{q}_j^{1,0,N}$converge strongly in $L^2(\mathbb{R}^d)$ to $\nabla q_i^{1,0,\infty}$ and $\nabla\tilde{q}_j^{1,0,\infty}$ respectively, when $N\to\infty$. Passing to the limit in (3.38) then gives

$$\bar{A}_1^*e_i\cdot e_j = \int_Q C_{per}(\nabla w_i^0+e_i)\cdot(e_j+\nabla\tilde{w}_j^0) - \int_{\mathbb{R}^d}A_1^0\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty}.$$

We now define $v_i^{1,0,N}$ and $\tilde{v}_j^{1,0,N}$ solutions to (3.34) and (3.37) with homogeneous Dirichlet (instead of periodic) boundary conditions on the boundary ∂I_N of I_N, and the tensor $B_1^{*,N}$ by

$$B_1^{*,N}e_i\cdot e_j = \int_Q C_{per}(\nabla w_i^0+e_i)\cdot(e_j+\nabla\tilde{w}_j^0) - \int_{I_N}A_1^0\nabla v_i^{1,0,N}\cdot\nabla\tilde{v}_j^{1,0,N}. \quad (3.39)$$

The proof of Proposition 3.2 is easily adapted to show that $B_1^{*,N}$ converges to A_1^* as N goes to infinity.

Step 2.

We consider

$$\left(B_1^{*,N}-\bar{A}_1^*\right)e_i\cdot e_j = \int_{\mathbb{R}^d}A_1^0\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty} - \int_{I_N}A_1^0\nabla v_i^{1,0,N}\cdot\nabla\tilde{v}_j^{1,0,N},$$

and expand the difference $B_1^{*,N}-\bar{A}_1^*$ as follows:

$$\begin{aligned}\left(B_1^{*,N}-\bar{A}_1^*\right)e_i\cdot e_j = &\int_{\mathbb{R}^d\setminus I_N}A_1^0\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty}\\ &+\left(\int_{I_N}A_1^0\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty} - \int_{I_N}A_1^0\nabla v_i^{1,0,N}\cdot\nabla\tilde{v}_j^{1,0,N}\right).\end{aligned} \quad (3.40)$$

We now show that the two terms in the right-hand side of (3.40) converge to 0 as N^{-d} when $N \to +\infty$.

We first note that the results of Lemma 3.8 of the Appendix, stated for a $\mathbb{Z}^d$-periodic matrix, can be readily extended to address A_1^0 since A_1^0 is equal to A_{per} in $\mathbb{R}^d \backslash Q$.

We deduce from Lemma 3.7 applied to (3.35) that $q_i^{1,0,\infty}$ is defined uniquely up to an additive constant. Moreover, A_{per} being piecewise Hölder continuous, we deduce from Lemma 3.8 that there exists a unique solution to (3.35) which converges to zero at infinity.

Since we only use $\nabla q_i^{1,0,\infty}$ in $\bar{A}_1^*$, we can thus assume without loss of generality that $q_i^{1,0,\infty}$ converges to zero at infinity. Likewise, we assume that $\tilde{q}_j^{1,0,\infty}$ converges to zero at infinity.

We then deduce from Lemma 3.8 that there exists a constant K independent of N such that for $|x| \geq 1$,

$$|q_i^{1,0,\infty}(x)| \leq K|x|^{1-d},\ |\tilde{q}_j^{1,0,\infty}(x)| \leq K|x|^{1-d}, \tag{3.41}$$

$$|\nabla q_i^{1,0,\infty}(x)| \leq K|x|^{-d},\ |\nabla \tilde{q}_j^{1,0,\infty}(x)| \leq K|x|^{-d}, \tag{3.42}$$

$$|v_i^{1,0,N}(x)| \leq K|x|^{1-d},\ |\tilde{v}_j^{1,0,N}(x)| \leq K|x|^{1-d}, \tag{3.43}$$

$$|\nabla v_i^{1,0,N}(x)| \leq K|x|^{-d},\ |\nabla \tilde{v}_j^{1,0,N}(x)| \leq K|x|^{-d}. \tag{3.44}$$

Using (3.42), we have $\|\nabla q_i^{1,0,\infty}\|_{L^2(\mathbb{R}^d \backslash I_N)} \leq KN^{-d/2}$ and $\|\nabla \tilde{q}_j^{1,0,\infty}\|_{L^2(\mathbb{R}^d \backslash I_N)} \leq KN^{-d/2}$, and so

$$\begin{aligned} \left| \int_{\mathbb{R}^d \backslash I_N} A_1^0 \nabla q_i^{1,0,\infty} \cdot \nabla \tilde{q}_j^{1,0,\infty} \right| &\leq \beta \|\nabla q_i^{1,0,\infty}\|_{L^2(\mathbb{R}^d \backslash I_N)} \|\nabla \tilde{q}_j^{1,0,\infty}\|_{L^2(\mathbb{R}^d \backslash I_N)} \\ &\leq KN^{-d}, \end{aligned} \tag{3.45}$$

where β is defined in (3.14).

We now address the second term of the right-hand side of (3.40) and write

$$\begin{aligned} &\int_{I_N} A_1^0 \nabla v_i^{1,0,N} \cdot \nabla \tilde{v}_j^{1,0,N} - \int_{I_N} A_1^0 \nabla q_i^{1,0,\infty} \cdot \nabla \tilde{q}_j^{1,0,\infty} \\ &\quad = \int_{I_N} A_1^0 (\nabla v_i^{1,0,N} - \nabla q_i^{1,0,\infty}) \cdot \nabla \tilde{v}_j^{1,0,N} + \int_{I_N} A_1^0 \nabla q_i^{1,0,\infty} \cdot (\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}). \end{aligned}$$

Since $\operatorname{div}\left(A_1^0(\nabla v_i^{1,0,N} - \nabla q_i^{1,0,\infty})\right) = \operatorname{div}\left((A_1^0)^T(\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty})\right) = 0$ in I_N, and

$\tilde{v}_j^{1,0,N} = 0$ on ∂I_N, we have, using integration by parts,

$$\begin{aligned}\int_{I_N} A_1^0(\nabla v_i^{1,0,N} - \nabla q_i^{1,0,\infty}) \cdot \nabla \tilde{v}_j^{1,0,N} + \int_{I_N} A_1^0 \nabla q_i^{1,0,\infty} \cdot (\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}) \\ = \int_{\partial I_N} A_1^0(\nabla v_i^{1,0,N} - \nabla q_i^{1,0,\infty}) \cdot \nu\, \tilde{v}_j^{1,0,N} \\ + \int_{\partial I_N} (A_1^0)^T(\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}) \cdot \nu\, q_i^{1,0,\infty} \\ = \int_{\partial I_N} (A_1^0)^T(\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}) \cdot \nu\, q_i^{1,0,\infty},\end{aligned}$$

where ν is the unit outward normal vector to ∂I_N.

The estimates (3.41) and (3.44) imply

$$\|q_i^{1,0,\infty}\|_{L^\infty(\partial I_N)} \leq KN^{1-d}, \quad \|(A_1^0)^T(\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}) \cdot \nu\|_{L^\infty(\partial I_N)} \leq KN^{-d},$$

while the measure of the boundary ∂I_N scales as N^{1-d}. Hence

$$\left| \int_{\partial I_N} (A_1^0)^T(\nabla \tilde{v}_j^{1,0,N} - \nabla \tilde{q}_j^{1,0,\infty}) \cdot \nu\, q_i^{1,0,\infty} \right| \leq KN^{-d},$$

and then

$$\left| \int_{I_N} A_1^0 \nabla v_i^{1,0,N} \cdot \nabla \tilde{v}_j^{1,0,N} - \int_{I_N} A_1^0 \nabla q_i^{1,0,\infty} \cdot \nabla \tilde{q}_j^{1,0,\infty} \right| \leq KN^{-d}. \tag{3.46}$$

We conclude by substituting (3.45) and (3.46) into (3.40).

□

Remark 3.5. *We assume $d \geq 3$ and piecewise Hölder regularity on A_{per}, and use Dirichlet boundary conditions in Lemma 3.3, because our proof relies on Lemma 3.8. Note however that the numerical experiments of Section 3.4 show, in dimension $d = 2$, that we again obtain the rate N^{-d} in the convergence of $A_1^{*,N}$ to $\bar{A}_1^*$ for two different A_{per}, one being piecewise Hölder continuous in the sense of Lemma 3.3 and the other not, and with periodic boundary conditions. Moreover, the explicit computations of Proposition 3.5 show that in dimension one, and without any assumption of regularity on A_1^0, the rate of convergence of $A_1^{*,N}$ to $\bar{A}_1^*$ is N^{-1}.*

3.3.4 Convergence of the second-order term $A_2^{*,N}$

We now address the second-order term $A_2^{*,N}$ defined by (3.29).

Proposition 3.4. *$A_2^{*,N}$ converges to a finite limit $\bar{A}_2^*$ in $\mathbb{R}^{d\times d}$ when $N \to \infty$.*

The rest of this section is devoted to the proof of this proposition.

Proof. We fix (i,j) in $[\![1,d]\!]^2$, and proceed in four steps.

Step 1. Rewriting of $A_2^{*,N}$

By $(N\mathbb{Z})^d$-periodicity, we have

$$\forall k \in \mathcal{T}_N, \int_{I_N} A_1^k(\nabla w_i^{1,k,N} + e_i)\cdot e_j = \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j.$$

It follows that

$$\begin{aligned}
&\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i)\cdot e_j - 2\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j + \int_{I_N} A_{per}(\nabla w_i^0 + e_i)\cdot e_j \\
&= \int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i)\cdot e_j - \int_{I_N} A_1^k(\nabla w_i^{1,k,N} + e_i)\cdot e_j \\
&\qquad - \int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j + \int_{I_N} A_{per}(\nabla w_i^0 + e_i)\cdot e_j.
\end{aligned} \tag{3.47}$$

Using (3.31) from the proof of Proposition 3.2, and the similar equalities

$$\int_{I_N} A_1^k(\nabla w_i^{1,k,N} + e_i)\cdot e_j = N^d A_{per}^* e_i \cdot e_j + \int_{Q+k} C_{per}(\nabla w_i^{1,k,N} + e_i)\cdot(\nabla \tilde{w}_j^0 + e_j) \tag{3.48}$$

and

$$\begin{aligned}
\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i)\cdot e_j =& N^d A_{per}^* e_i \cdot e_j \\
&+ \int_{Q\cup\{Q+k\}} C_{per}(\nabla w_i^{2,0,k,N} + e_i)\cdot(\nabla \tilde{w}_j^0 + e_j),
\end{aligned} \tag{3.49}$$

for a defect at position k and two defects at positions 0 and k respectively, and using (3.31), (3.48) and (3.49) in (3.47), we obtain a new expression for the right hand side of (3.47):

$$\begin{aligned}
&\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i)\cdot e_j - 2\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j + \int_{I_N} A_{per}(\nabla w_i^0 + e_i)\cdot e_j \\
&= \int_{Q+k} C_{per}(\nabla w_i^{2,0,k,N} - \nabla w_i^{1,k,N})\cdot(\nabla \tilde{w}_j^0 + e_j) \\
&\quad + \int_{Q} C_{per}(\nabla w_i^{2,0,k,N} - \nabla w_i^{1,0,N})\cdot(\nabla \tilde{w}_j^0 + e_j).
\end{aligned} \tag{3.50}$$

We now define

$$q_i^{2,0,k,N} = w_i^{2,0,k,N} - w_i^{1,0,N} - w_i^{1,k,N} + w_i^0. \tag{3.51}$$

Intuitively, we are comparing the solution $w_i^{2,0,k,N}$ with two defects located at 0 and at $k \in \mathcal{T}_N$ to the sum of the one-defect solutions $w_i^{1,0,N}$ and $w_i^{1,k,N}$ minus the periodic background w_i^0. We expect the difference $q_i^{2,0,k,N}$ to decay sufficiently fast far from the

defects.

The function $q_i^{2,0,k,N}$ solves

$$\begin{cases} -\operatorname{div}\left(A_2^{0,k}\nabla q_i^{2,0,k,N}\right) = \operatorname{div}(\mathbb{1}_{\{Q+k\}}C_{per}\nabla q_i^{1,0,N}) \\ \qquad\qquad + \operatorname{div}(\mathbb{1}_Q C_{per}\nabla q_i^{1,k,N}) \quad \text{in } I_N, \\ q_i^{2,0,k,N}\ (N\mathbb{Z})^d - \text{periodic}, \end{cases} \tag{3.52}$$

with $q_i^{1,k,N} = q_i^{1,0,N}(\cdot - k)$ for $k \in \mathcal{T}_N$. For later use, we also define the adjoint of (3.52) by

$$\begin{cases} -\operatorname{div}\left((A_2^{0,k})^T\nabla \tilde{q}_j^{2,0,k,N}\right) = \operatorname{div}(\mathbb{1}_{\{Q+k\}}C_{per}^T\nabla \tilde{q}_j^{1,0,N}) \\ \qquad\qquad + \operatorname{div}(\mathbb{1}_Q C_{per}^T\nabla \tilde{q}_j^{1,k,N}) \quad \text{in } I_N, \\ \tilde{q}_j^{2,0,k,N}\ (N\mathbb{Z})^d - \text{periodic}. \end{cases} \tag{3.53}$$

Using (3.33) and (3.51), we rewrite (3.50) as follows:

$$\begin{aligned} &\int_{I_N} A_2^{0,k}(\nabla w_i^{2,0,k,N} + e_i)\cdot e_j - 2\int_{I_N} A_1^0(\nabla w_i^{1,0,N} + e_i)\cdot e_j + \int_{I_N} A_{per}(\nabla w_i^0 + e_i)\cdot e_j \\ &= \int_{Q+k} C_{per}\left(\nabla q_i^{1,0,N} + \nabla q_i^{2,0,k,N}\right)\cdot(\nabla \tilde{w}_j^0 + e_j) \\ &\qquad + \int_Q C_{per}\left(\nabla q_i^{1,k,N} + \nabla q_i^{2,0,k,N}\right)\cdot(\nabla \tilde{w}_j^0 + e_j). \end{aligned} \tag{3.54}$$

It entails from (3.29) and (3.54) that

$$\begin{aligned} A_2^{*,N}e_i\cdot e_j = \frac{1}{2}\sum_{k\in\mathcal{T}_N\setminus\{0\}} &\left(\int_{Q+k} C_{per}\left(\nabla q_i^{1,0,N} + \nabla q_i^{2,0,k,N}\right)\cdot(\nabla \tilde{w}_j^0 + e_j)\right. \\ &\left. + \int_Q C_{per}\left(\nabla q_i^{1,k,N} + \nabla q_i^{2,0,k,N}\right)\cdot(\nabla \tilde{w}_j^0 + e_j)\right). \end{aligned} \tag{3.55}$$

Since $q_i^{1,0,N} = q_i^{1,k,N}(\cdot + k)$, and $\tilde{w}_j^0$ is $\mathbb{Z}^d$-periodic, we have

$$\int_Q C_{per}\nabla q_i^{1,k,N}\cdot(\nabla \tilde{w}_j^0 + e_j) = \int_{Q-k} C_{per}\nabla q_i^{1,0,N}\cdot(\nabla \tilde{w}_j^0 + e_j). \tag{3.56}$$

By definition of $\mathcal{T}_N$, we know that

$$\bigcup_{k\in\mathcal{T}_N}\{Q+k\} = \bigcup_{k\in\mathcal{T}_N}\{Q-k\} = I_N,$$

we then have

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q+k} C_{per}\nabla q_i^{1,0,N}\cdot(\nabla \tilde{w}_j^0 + e_j) = \sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q-k} C_{per}\nabla q_i^{1,0,N}\cdot(\nabla \tilde{w}_j^0 + e_j). \tag{3.57}$$

Substituting (3.56) in (3.57), we obtain

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q+k} C_{per}\nabla q_i^{1,0,N}\cdot(\nabla\tilde{w}_j^0+e_j)=\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q} C_{per}\nabla q_i^{1,k,N}\cdot(\nabla\tilde{w}_j^0+e_j), \quad (3.58)$$

and using (3.58) in (3.55), we find that

$$\begin{aligned} A_2^{*,N}e_i\cdot e_j &= \sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q} C_{per}\nabla q_i^{1,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\\ &+\frac{1}{2}\sum_{k\in\mathcal{T}_N\setminus\{0\}}\left(\int_{Q+k} C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)+\int_{Q} C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right). \end{aligned} \quad (3.59)$$

We henceforth denote by

$$D_N=\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{Q} C_{per}\nabla q_i^{1,k,N}\cdot(\nabla\tilde{w}_j^0+e_j), \quad (3.60)$$

and

$$\forall k\in\mathcal{T}_N,\ E_N^k=\int_{Q+k} C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)+\int_{Q} C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j), \quad (3.61)$$

(we omit the dependence on i and j of these quantities to keep the notation light), so that (3.59) writes in the following more concise form:

$$A_2^{*,N}e_i\cdot e_j=D_N+\frac{1}{2}\sum_{k\in\mathcal{T}_N\setminus\{0\}}E_N^k. \quad (3.62)$$

Note that D_N is a "one defect" term and E_N^k a "two defects" term. In the next two steps we are going to prove that D_N and $\sum_{k\in\mathcal{T}_N\setminus\{0\}}E_N^k$ converge to finite limits as $N\to\infty$.

Step 2. Convergence of D_N

Introducing

$$q_i^{1,N}=\sum_{k\in\mathcal{T}_N} q_i^{1,k,N}, \quad (3.63)$$

we rewrite D_N defined by (3.60) as

$$D_N=\int_{Q} C_{per}\nabla q_i^{1,N}\cdot(\nabla\tilde{w}_j^0+e_j)-\int_{Q} C_{per}\nabla q_i^{1,0,N}\cdot(\nabla\tilde{w}_j^0+e_j). \quad (3.64)$$

We will now pass to the limit $N\to+\infty$ in each of the two terms in the right-hand side.

For this purpose, we first obtain from (3.34) that $q_i^{1,k,N}$, which is by definition equal to $q_i^{1,0,N}(\cdot-k)$, is $(N\mathbb{Z})^d$-periodic and satisfies

$$-\mathrm{div}\left(A_{per}\nabla q_i^{1,k,N}\right)=\mathrm{div}(\mathbb{1}_{Q+k}C_{per}(\nabla w_i^0+e_i))+\mathrm{div}\left(\mathbb{1}_{Q+k}C_{per}\nabla q_i^{1,k,N}\right)\quad\text{in } I_N.$$

Then $q_i^{1,N}$ defined by (3.63) is $(N\mathbb{Z})^d$-periodic and satisfies

$$-\mathrm{div}\left(A_{per}\nabla q_i^{1,N}\right) = \mathrm{div}(C_{per}(\nabla w_i^0 + e_i)) + \mathrm{div}(C_{per}\sum_{k\in\mathcal{T}_N} \mathbb{1}_{Q+k}\nabla q_i^{1,k,N}) \quad \text{in } I_N.$$

Actually, since the $q_i^{1,k,N}$ are obtained by a k-shift of $q_i^{1,0,N}$, it follows that their sum $q_i^{1,N}$ as well as $\sum_{k\in\mathcal{T}_N} \mathbb{1}_{Q+k}\nabla q_i^{1,k,N}$ (the latter being extended by $(N\mathbb{Z})^d$-periodicity to the whole space $\mathbb{R}^d$) are $\mathbb{Z}^d$-periodic, so that we can rewrite $q_i^{1,N}$ as the solution (up to an additive constant) to

$$\begin{cases} -\mathrm{div}\left(A_{per}\nabla q_i^{1,N}\right) = \mathrm{div}(C_{per}(\nabla w_i^0 + e_i)) + \mathrm{div}\left(\mathbb{1}_Q C_{per}\nabla q_i^{1,0,N}\right) & \text{in } Q, \\ q_i^{1,N}\ \mathbb{Z}^d - \text{periodic.} \end{cases} \tag{3.65}$$

Since we know from Lemma 3.6 applied to (3.34) that $\nabla q_i^{1,0,N}$ converges in $L^2(Q)$ to $\nabla q_i^{1,0,\infty}$ defined by (3.35), we easily deduce from (3.65) that $\nabla q_i^{1,N}$ converges in $L^2(Q)$ to $\nabla q_i^{1,\infty}$, where $q_i^{1,\infty}$ solves

$$\begin{cases} -\mathrm{div}\left(A_{per}\nabla q_i^{1,\infty}\right) = \mathrm{div}(C_{per}(\nabla w_i^0 + e_i)) + \mathrm{div}\left(\mathbb{1}_Q C_{per}\nabla q_i^{1,0,\infty}\right) & \text{in } Q, \\ q_i^{1,\infty}\ \mathbb{Z}^d - \text{periodic.} \end{cases}$$

Using (3.64), it is clear that

$$D_N \underset{N\to\infty}{\to} \int_Q C_{per}\nabla q_i^{1,\infty}\cdot(\nabla\tilde{w}_j^0 + e_j) - \int_Q C_{per}\nabla q_i^{1,0,\infty}\cdot(\nabla\tilde{w}_j^0 + e_j),$$

which concludes step 2.

Step 3. Convergence of $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$

We first rewrite E_N^k in a more tractable way.

We compute, using integration by parts, (3.37) and the definition of $A_2^{0,k}$,

$$\begin{aligned} E_N^k &= \int_{Q+k} C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0 + e_j) + \int_Q C_{per}\nabla q_i^{2,0,k,N}\cdot(\nabla\tilde{w}_j^0 + e_j) \\ &= -\int_{I_N} q_i^{2,0,k,N}\mathrm{div}\left(C_{per}^T\left(\mathbb{1}_Q(\nabla\tilde{w}_j^0 + e_j) + \mathbb{1}_{Q+k}(\nabla\tilde{w}_j^0 + e_j)\right)\right) \\ &= -\int_{I_N} q_i^{2,0,k,N}\mathrm{div}\left((A_1^0)^T\nabla\tilde{q}_j^{1,0,N} + (A_1^k)^T\nabla\tilde{q}_j^{1,k,N}\right) \\ &= \int_{I_N}\left(A_1^0\nabla q_i^{2,0,k,N}\cdot\nabla\tilde{q}_j^{1,0,N} + A_1^k\nabla q_i^{2,0,k,N}\cdot\nabla\tilde{q}_j^{1,k,N}\right) \\ &= \int_{I_N} A_2^{0,k}\nabla q_i^{2,0,k,N}\cdot\left(\nabla\tilde{q}_j^{1,0,N} + \nabla\tilde{q}_j^{1,k,N}\right) \\ &\qquad - \int_{I_N}\nabla q_i^{2,0,k,N}\cdot C_{per}^T\left(\mathbb{1}_Q\nabla\tilde{q}_j^{1,k,N} + \mathbb{1}_{Q+k}\nabla\tilde{q}_j^{1,0,N}\right). \end{aligned}$$

Using (3.52), (3.53) and integration by parts then yields

$$E_N^k = -\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_Q\nabla q_i^{1,k,N} + \mathbb{1}_{Q+k}\nabla q_i^{1,0,N}\right)\cdot\left(\nabla\tilde{q}_j^{1,0,N} + \nabla\tilde{q}_j^{1,k,N}\right) + \int_{I_N} A_2^{0,k}\nabla q_i^{2,0,k,N}\cdot\nabla\tilde{q}_j^{2,0,k,N}.$$

Defining

$$E_N^{1,k} = -\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_Q\nabla q_i^{1,k,N} + \mathbb{1}_{Q+k}\nabla q_i^{1,0,N}\right)\cdot\left(\nabla\tilde{q}_j^{1,0,N} + \nabla\tilde{q}_j^{1,k,N}\right) \tag{3.66}$$

and

$$E_N^{2,k} = \int_{I_N} A_2^{0,k}\nabla q_i^{2,0,k,N}\cdot\nabla\tilde{q}_j^{2,0,k,N}, \tag{3.67}$$

it holds

$$E_N^k = E_N^{1,k} + E_N^{2,k}.$$

We are going to prove that $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$ and $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ converge to finite limits when N goes to infinity.

Step 3.1. Convergence of $\displaystyle\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$

Since $q_i^{1,k,N} = q_i^{1,0,N}(\cdot - k)$, $\tilde{q}_j^{1,k,N} = \tilde{q}_j^{1,0,N}(\cdot - k)$ and $\mathcal{T}_N = -\mathcal{T}_N$, we start by noting that

$$\begin{aligned}
&\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_Q\nabla q_i^{1,k,N} + \mathbb{1}_{Q+k}\nabla q_i^{1,0,N}\right)\cdot\nabla\tilde{q}_j^{1,k,N}\\
&= \sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_{Q-k}\nabla q_i^{1,0,N} + \mathbb{1}_Q\nabla q_i^{1,0,N}(\cdot + k)\right)\cdot\nabla\tilde{q}_j^{1,0,N}\\
&= \sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_{Q+k}\nabla q_i^{1,0,N} + \mathbb{1}_Q\nabla q_i^{1,0,N}(\cdot - k)\right)\cdot\nabla\tilde{q}_j^{1,0,N}\\
&= \sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_{Q+k}\nabla q_i^{1,0,N} + \mathbb{1}_Q\nabla q_i^{1,k,N}\right)\cdot\nabla\tilde{q}_j^{1,0,N}.
\end{aligned} \tag{3.68}$$

Inserting (3.68) in (3.66) gives

$$\begin{aligned}
\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k} &= -2\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_Q\nabla q_i^{1,k,N} + \mathbb{1}_{Q+k}\nabla q_i^{1,0,N}\right)\cdot\nabla\tilde{q}_j^{1,0,N}\\
&= -2\sum_{k\in\mathcal{T}_N}\int_{\mathbb{R}^d} C_{per}\left(\mathbb{1}_Q\nabla q_i^{1,k,N} + \mathbb{1}_{Q+k}\nabla q_i^{1,0,N}\right)\cdot\nabla\tilde{q}_j^{1,0,N}\\
&\qquad + 4\int_{\mathbb{R}^d} C_{per}\mathbb{1}_Q\nabla q_i^{1,0,N}\cdot\nabla\tilde{q}_j^{1,0,N}\\
&= -2\int_Q C_{per}\nabla q_i^{1,N}\cdot\nabla\tilde{q}_j^{1,0,N} - 2\int_{I_N} C_{per}\nabla q_i^{1,0,N}\cdot\nabla\tilde{q}_j^{1,0,N}\\
&\qquad + 4\int_Q C_{per}\nabla q_i^{1,0,N}\cdot\nabla\tilde{q}_j^{1,0,N},
\end{aligned} \tag{3.69}$$

where we recall that $q_i^{1,N} = \sum_{k\in\mathcal{T}_N} q_i^{1,k,N}$.

We know from Lemma 3.6 applied to (3.34) and (3.37) that $\mathbb{1}_{I_N}\nabla q_i^{1,0,N}$ and $\mathbb{1}_{I_N}\nabla \tilde{q}_j^{1,0,N}$ converge to $\nabla q_i^{1,0,\infty}$ and $\nabla \tilde{q}_j^{1,0,\infty}$ in $L^2(\mathbb{R}^d)$ when $N\to\infty$. Moreover, we have seen in Step 2 that $\nabla q_i^{1,N}$ converges to $\nabla q_i^{1,\infty}$ in $L^2(Q)$.

We can consequently take the limit $N\to\infty$ in (3.69), and obtain

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k} \underset{N\to\infty}{\to} -2\int_Q C_{per}\nabla q_i^{1,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty} - 2\int_{\mathbb{R}^d} C_{per}\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty} + 4\int_Q C_{per}\nabla q_i^{1,0,\infty}\cdot\nabla\tilde{q}_j^{1,0,\infty}. \tag{3.70}$$

Step 3.2. Convergence of $\displaystyle\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$

Following the proof of Lemma 3.6 applied to (3.52) and (3.53), we first note that $\mathbb{1}_{I_N}\nabla q_i^{2,0,k,N}$ and $\mathbb{1}_{I_N}\nabla\tilde{q}_j^{2,0,k,N}$ converge in $L^2(\mathbb{R}^d)$ to $\nabla q_i^{2,0,k,\infty}$ and $\nabla\tilde{q}_j^{2,0,k,\infty}$ respectively, where $q_i^{2,0,k,\infty}$ is in $L^2_{loc}(\mathbb{R}^d)$ and solves

$$\begin{cases} -\operatorname{div}\left(A_2^{0,k}\nabla q_i^{2,0,k,\infty}\right) = \operatorname{div}(\mathbb{1}_{\{Q+k\}}C_{per}\nabla q_i^{1,0,\infty}) \\ \qquad\qquad + \operatorname{div}(\mathbb{1}_Q C_{per}\nabla q_i^{1,k,\infty}) \quad \text{in } \mathbb{R}^d, \\ \nabla q_i^{2,0,k,\infty}\in L^2(\mathbb{R}^d), \end{cases} \tag{3.71}$$

and $\tilde{q}_j^{2,0,k,\infty}$ solves the adjoint problem to (3.71).

Consequently, for each $k\in\mathcal{T}_N\setminus\{0\}$, we deduce from (3.67) that

$$E_N^{2,k} \underset{N\to\infty}{\to} E_\infty^{2,k} := \int_{\mathbb{R}^d} A_2^{0,k}\nabla q_i^{2,0,k,\infty}\cdot\nabla\tilde{q}_j^{2,0,k,\infty}. \tag{3.72}$$

We are going to prove that the series $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ converges to $\displaystyle\sum_{k\in\mathbb{Z}^d\setminus\{0\}} E_\infty^{2,k}$ when $N\to\infty$. For this purpose we first obtain some bounds on $E_N^{2,k}$ and $E_\infty^{2,k}$.

We derive from (3.67) that

$$\left|E_N^{2,k}\right| \le \frac{\beta}{2}\left(\|\nabla q_i^{2,0,k,N}\|^2_{L_2(I_N)} + \|\nabla\tilde{q}_j^{2,0,k,N}\|^2_{L_2(I_N)}\right) \tag{3.73}$$

where β is defined in (3.14).

On the other hand it entails from (3.52) that

$$\int_{I_N} A_2^{0,k}\nabla q_i^{2,0,k,N}\cdot\nabla q_i^{2,0,k,N} = -\int_Q C_{per}\nabla q_i^{1,k,N}\cdot\nabla q_i^{2,0,k,N} - \int_{Q+k} C_{per}\nabla q_i^{1,0,N}\cdot\nabla q_i^{2,0,k,N},$$

whence, using the Cauchy-Schwarz inequality in the right-hand side and the coerciveness constant α defined in (3.13),

$$\|\nabla q_i^{2,0,k,N}\|_{L^2(I_N)} \leq \left(\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}\right)\left(\|\nabla q_i^{1,k,N}\|_{L^2(Q)} + \|\nabla q_i^{1,0,N}\|_{L^2(Q+k)}\right),$$

and then

$$\|\nabla q_i^{2,0,k,N}\|^2_{L^2(I_N)} \leq 2\left(\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}\right)^2\left(\|\nabla q_i^{1,k,N}\|^2_{L^2(Q)} + \|\nabla q_i^{1,0,N}\|^2_{L^2(Q+k)}\right), \quad (3.74)$$

Likewise,

$$\|\nabla \tilde{q}_j^{2,0,k,N}\|^2_{L^2(I_N)} \leq 2\left(\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}\right)^2\left(\|\nabla \tilde{q}_j^{1,k,N}\|^2_{L^2(Q)} + \|\nabla \tilde{q}_j^{1,0,N}\|^2_{L^2(Q+k)}\right). \quad (3.75)$$

Using (3.74) and (3.75) in (3.73), we obtain that there exists a constant C such that for all $k \in \mathcal{T}_N\backslash\{0\}$,

$$\begin{aligned}\left|E_N^{2,k}\right| \leq C\Big(&\|\nabla q_i^{1,k,N}\|^2_{L^2(Q)} + \|\nabla q_i^{1,0,N}\|^2_{L^2(Q+k)} \\ &+ \|\nabla \tilde{q}_j^{1,k,N}\|^2_{L^2(Q)} + \|\nabla \tilde{q}_j^{1,0,N}\|^2_{L^2(Q+k)}\Big).\end{aligned} \quad (3.76)$$

Similar computations yield that for all $k \in \mathbb{Z}^d\backslash\{0\}$,

$$\begin{aligned}\left|E_\infty^{2,k}\right| \leq C\Big(&\|\nabla q_i^{1,k,\infty}\|^2_{L^2(Q)} + \|\nabla q_i^{1,0,\infty}\|^2_{L^2(Q+k)} \\ &+ \|\nabla \tilde{q}_j^{1,k,\infty}\|^2_{L^2(Q)} + \|\nabla \tilde{q}_j^{1,0,\infty}\|^2_{L^2(Q+k)}\Big).\end{aligned} \quad (3.77)$$

Summing (3.77) for all positions $k \in \mathbb{Z}^d\backslash\{0\}$, we find that

$$\sum_{k\in\mathbb{Z}^d\backslash\{0\}}\left|E_\infty^{2,k}\right| \leq C\left(\|\nabla q_i^{1,0,\infty}\|^2_{L^2(\mathbb{R}^d)} + \|\nabla \tilde{q}_j^{1,0,\infty}\|^2_{L^2(\mathbb{R}^d)}\right), \quad (3.78)$$

which proves that the series $\sum_{k\in\mathbb{Z}^d\backslash\{0\}} E_\infty^{2,k}$ is absolutely converging.

Consider now $\varepsilon > 0$. For $(M,N) \in (2\mathbb{N}+1)^2$, $M \leq N$, we compute

$$\begin{aligned}\left|\sum_{k\in\mathcal{T}_N\backslash\{0\}} E_N^{2,k} - \sum_{k\in\mathbb{Z}^d\backslash\{0\}} E_\infty^{2,k}\right| \leq &\sum_{k\in\mathcal{T}_M\backslash\{0\}}\left|E_N^{2,k} - E_\infty^{2,k}\right| + \sum_{k\in\mathcal{T}_N\backslash\mathcal{T}_M}\left|E_N^{2,k}\right| \\ &+ \sum_{k\in\mathbb{Z}^d\backslash\mathcal{T}_M}\left|E_\infty^{2,k}\right|.\end{aligned} \quad (3.79)$$

Summing (3.76) for all $k \in \mathcal{T}_N\backslash\mathcal{T}_M$ and (3.77) for all $k \in \mathbb{Z}^d\backslash\mathcal{T}_M$ yields

$$\sum_{k\in\mathcal{T}_N\backslash\mathcal{T}_M}\left|E_N^{2,k}\right| \leq C\left(\|\nabla q_i^{1,0,N}\|^2_{L^2(I_N\backslash I_M)} + \|\nabla \tilde{q}_j^{1,0,N}\|^2_{L^2(I_N\backslash I_M)}\right) \quad (3.80)$$

and

$$\sum_{k\in\mathbb{Z}^d\setminus\mathcal{T}_M}\left|E_\infty^{2,k}\right| \le C\left(\|\nabla q_i^{1,0,\infty}\|^2_{L^2(\mathbb{R}^d\setminus I_M)} + \|\nabla \tilde{q}_j^{1,0,\infty}\|^2_{L^2(\mathbb{R}^d\setminus I_M)}\right) \tag{3.81}$$

respectively.

We know from Lemma 3.6 applied to (3.34) and (3.37) that $\mathbb{1}_{I_N}\nabla q_i^{1,0,N}$ and $\mathbb{1}_{I_N}\nabla \tilde{q}_j^{1,0,N}$ converge in $L^2(\mathbb{R}^d)$ to $\nabla q_i^{1,0,\infty}$ and $\nabla \tilde{q}_j^{1,0,\infty}$ respectively. It is then straightforward to deduce from (3.80) and (3.81) that there exist M_0 and N_0 such that

$$\sum_{k\in\mathbb{Z}^d\setminus\mathcal{T}_{M_0}}\left|E_\infty^{2,k}\right| \le \varepsilon \tag{3.82}$$

and

$$\forall N \ge N_0, \quad \sum_{k\in\mathcal{T}_N\setminus\mathcal{T}_{M_0}}\left|E_N^{2,k}\right| \le \varepsilon. \tag{3.83}$$

Moreover, (3.72) implies that, choosing N_0 sufficiently large,

$$\forall N \ge N_0, \quad \sum_{k\in\mathcal{T}_{M_0}\setminus\{0\}}\left|E_N^{2,k} - E_\infty^{2,k}\right| \le \varepsilon. \tag{3.84}$$

Inserting (3.83), (3.82) and (3.84) in (3.79), we have shown that for every $\varepsilon > 0$, there exists N_0 such that

$$\forall N \ge N_0, \quad \left|\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k} - \sum_{k\in\mathbb{Z}^d\setminus\{0\}} E_\infty^{2,k}\right| \le 3\varepsilon.$$

This amounts to say that

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k} \underset{N\to\infty}{\to} \sum_{k\in\mathbb{Z}^d\setminus\{0\}} E_\infty^{2,k}.$$

We have thus proved in Steps 3.1 and 3.2 that $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$ and $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ converge when N goes to infinity. Since $E_N^k = E_N^{1,k} + E_N^{2,k}$, we deduce that $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$ converges.

Remark 3.6. *We actually conclude from Step 3.2 a stronger result, namely that the series $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ is absolutely converging. Numerical experiments show that this is not the case for $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$, which can be guessed from the proof.*

Step 4. Conclusion

We have shown in the previous steps that the sequence D_N and the series $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$ converge when $N \to \infty$. Using (3.62), this implies that the sequence $A_2^{*,N}$ converges in $\mathbb{R}^{d\times d}$. □

3.4 Numerical experiments

Our purpose in this section is to assess the approximation of A_η^* by the second-order expansion $A_{per}^* + \eta A_1^{*,N} + \eta^2 A_2^{*,N}$. In order to maintain a reasonable computational cost, we restrict ourselves to the two-dimensional case. We first explain our general methodology and then make precise the specific settings.

3.4.1 Methodology

We will consider two commonly used composite materials as periodic reference materials A_{per}. The first material consists of a constant background reinforced by a periodic lattice of circular inclusions, that is

$$A_{per}(x_1, x_2) = 20 \times Id + 100 \sum_{k \in \mathbb{Z}^2} \mathbb{1}_{B(k,0.3)}(x_1, x_2) \times Id,$$

where $B(k, 0.3)$ is the ball of center k and radius 0.3. The second material is a laminate for which

$$A_{per}(x_1, x_2) = 20 \times Id + 100 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_1 \leq l+1}(x_1, x_2) \times Id.$$

In the case of material 1, the role of the perturbation is, loosely speaking, to randomly eliminate some fibers:

$$C_{per}(x_1, x_2) = -100 \sum_{k \in \mathbb{Z}^2} \mathbb{1}_{B(k,0.3)}(x_1, x_2) \times Id.$$

In the case of material 2, the perturbation consists in a random modification of the lamination direction:

$$C_{per}(x_1, x_2) = -100 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_1 \leq l+1}(x_1, x_2) \times Id + 100 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_2 \leq l+1}(x_1, x_2) \times Id.$$

In both cases, we have chosen the coefficients 20 and 100 in order to have a high contrast between A_{per} and $A_{per} + C_{per}$, and thus for the perturbation to be significant. There is of course nothing specific in the actual values of these coefficients.

These two materials are shown in Figure 3.3.

Our goal is to compare A_η^* with its approximation $A_{per}^* + \eta A_1^{*,N} + \eta^2 A_2^{*,N}$ for each of these two particular settings. A major computational difficulty is the computation of the "exact" matrix A_η^* given by formula (3.16). It ideally requires to solve the stochastic cell problems (3.15) on $\mathbb{R}^d$. To this end we first use ergodicity and formulae (3.18) and (3.21), and actually compute, for a given realization ω and a domain I_N which is here equal to $[0, N]^2$ for convenience, $A_\eta^{*,N}(\omega)$ defined by

$$A_\eta^{*,N}(\omega) e_i = \frac{1}{N^d} \int_{I_N} A_\eta(x, \omega)(\nabla w_i^{\eta,N,\omega}(x) + e_i) dx. \tag{3.85}$$

In a second step, we take averages over the realizations ω.

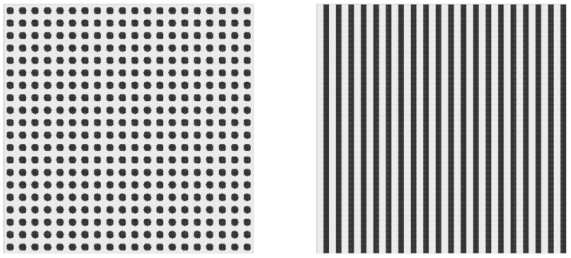

Figure 3.3: Left: a periodic lattice of circular inclusions. Right: a one-dimensional laminate.

For each ω, we use the finite element software FreeFem++ (available at www.freefem.org) to solve the boundary value problems (3.18) and compute the integrals (3.85). We work with standard P1 finite elements on a triangular mesh such that there are 10 degrees of freedom on each edge of the unit cell Q.

We define an approximate value $A_\eta^{*,N}$ as the average of $A_\eta^{*,N}(\omega)$ over 40 realizations ω. Our numerical experiments indeed show that the number 40 is sufficiently large for the convergence of the Monte-Carlo computation. We then let N grow from 5 to 80 by increments of 5. We observe that $A_\eta^{*,N}$ stabilizes at a fixed value around $N = 80$ and thus take $A_\eta^{*,80}$ as the reference value for A_η^* in our subsequent tests.

The next step is to compute the zero-order term A_{per}^*, and the first-order and second-order deterministic corrections $A_1^{*,N}$ and $A_2^{*,N}$. Using the same mesh and finite elements as for our reference computation above, we compute A_{per}^* using (3.3) and (3.4), and for each N we compute $A_1^{*,N}$ and $A_2^{*,N}$ using (3.28) and (3.29). We again let N grow from 5 to 80 by increments of 5 for $A_1^{*,N}$. The computation of $A_2^{*,N}$ being significantly more expensive (note that in (3.29) there is not only an integral over I_N but also a sum over the N^2 cells) we have to limit ourselves to $N = 25$ and approximate the value for N larger than 25 by the value obtained for $N = 25$.

Before presenting our results, we wish to discuss our expectations. Note that there are three distinct sources of error:

- the finite element discretization error;
- the truncation error due to the replacement of $\mathbb{R}^d$ with I_N, in the computation of the stochastic cell problems (3.15) that are replaced with (3.18), as well as in the computation of the integrals (3.85);
- the stochastic error arising from the approximation of the expectation value (3.21) by an empirical mean.

The discretization error originates from the fact that, in practice, we only have access to the finite element approximations of all the functions manipulated here (such as w_i^0, $w_i^{\eta,N,\omega}$,...). Although we have not proved it in the specific context of our work, we believe,

because it is shown in a similar weakly random setting (see [31]), that all the convergences stated here in the infinite-dimensional setting still hold true for the finite-dimensional approximations of the objects. Our numerical results indeed confirm it is the case. In order to eliminate the discretization error from the picture, our practical approach consists in adopting the *same* finite element space for all approximations of the cell and supercell problems, independently of N.

The truncation error is a different issue. For the "exact" computation of A^*_η (we mean not using the second-order expansion (3.26), but (3.85)), we use an empirical mean and a truncation. We know from [22], for a *continuous* notion of stationarity analogous to the *discrete* notion (3.8) we use here, and under mixing conditions which are satisfied in our setting, that the convergence of the truncated approximation to the ideal value holds at a rate $N^{-\kappa}$ with κ a non explicit function of the dimension, the mixing exponent and the coercivity constant of the material. On the other hand, in the second-order expansion (3.26), the zero-order term A^*_{per} is of course free of any truncation error. All that we know for the approximation $A_1^{*,N}$ defined by (3.28) to the first-order correction $\bar{A}^*_1$, is stated in Lemma 3.3 in dimension $d \geq 3$, under Hölder regularity assumptions on A_{per}, and with Dirichlet boundary conditions replacing periodic ones. One of the aims of our experiments is therefore to draw some numerical conclusions on the convergence of this term when these assumptions are not satisfied. Note that the matrices involved in our test materials are clearly discontinuous functions of x. The matrix corresponding to material 1 is piecewise Hölder continuous in the sense of Lemma 3.3, while the matrix corresponding to material 2 is not. As for the second-order approximation $A_2^{*,N}$, we have no insight on the truncation error and we also wish to study its convergence from a numerical point of view.

Finally, we have a practical approach to the stochastic error: besides the empirical mean, we provide, for each N, the minimum and the maximum values of $A_\eta^{*,N}(\omega)$ achieved over the 40 computations.

We now would like to emphasize that the purpose of our numerical tests is not to *prove* that

$$A^*_\eta = A^*_{per} + \eta A_1^{*,N} + \eta^2 A_2^{*,N} + o(\eta^2)$$

for a remainder term $o(\eta^2)$ that is independent of N, of the number of realizations and of the size of the mesh. Establishing experimentally that such an asymptotic holds is too demanding a task. It would indeed require letting η go to 0, which in turn, since we have to observe at least one (and in fact many) event per domain considered, would necessitate a supercell of size N extremely large. We cannot afford such a computational workload.

Using our numerical tests, we only hope here to demonstrate, and we indeed do so, that the second-order expansion is an approximation to A^*_η sufficiently good for all practical purposes, and in particular for η not so small ! We will observe that both $A_1^{*,N}$ and $A_2^{*,N}$ converge to their respective limits faster than $A_\eta^{*,N}$ to A^*_η (which is intuitively expected since the former quantities are deterministic and contain less information). We will also observe that $A^*_{per} + \eta A_1^{*,N}$ is significantly closer to A^*_η than A^*_{per}, thereby motivating the

expansion. The inclusion of the second-order term further improves the situation.

3.4.2 Results

In order to give an idea on how the perturbation affects the materials considered, we first show some typical realizations in Figures 3.4. Our results are presented in Section 3.4.2.1 and Section 3.4.2.2 below. Since these results are qualitatively similar for the two materials, we comment on the results altogether in Section 3.4.2.3.

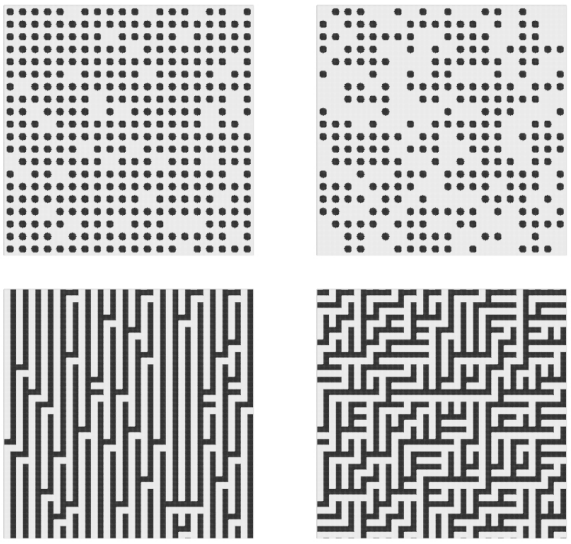

Figure 3.4: Above: two instances of material 1 with $\eta = 0.1$ (left) and $\eta = 0.4$ (right). Below: two instances of material 2 with $\eta = 0.1$ (left) and $\eta = 0.5$ (right).

To present our numerical results, we choose the first diagonal entry $(1,1)$ of all the matrices considered. Other coefficients in the matrices behave qualitatively similarly. As mentioned in the previous section, we illustrate a practical interval of confidence for our Monte-Carlo computation of A^*_η by showing, for each N, the minimum and maximum values of $A^{*,N}_\eta(\omega)$ achieved over the 40 realizations ω.

We will use the following caption in the graphs:
- *periodic*: gives the value of the periodic homogenized tensor A^*_{per};
- *first-order*: gives the value of $A^*_{per} + \eta A^{*,N}_1$;
- *second-order*: gives the value of $A^*_{per} + \eta A^{*,N}_1 + \eta^2 A^{*,N}_2$;
- *stochastic mean, minima and maxima*: respectively give the values of $A^{*,N}_\eta$ and the extrema obtained in the computation of the empirical mean.

Finally, the results are given for *some* specific values of η (not necessarily the same for both materials) which serve the purpose of testing our approach in a diversity of situations, from a "small" to a "not so small" perturbation.

3.4.2.1 Results for material 1

We show the results for $\eta = 0.1$, $\eta = 0.4$ and $\eta = 0.5$ (Figures 3.5, 3.6 and 3.7 respectively).

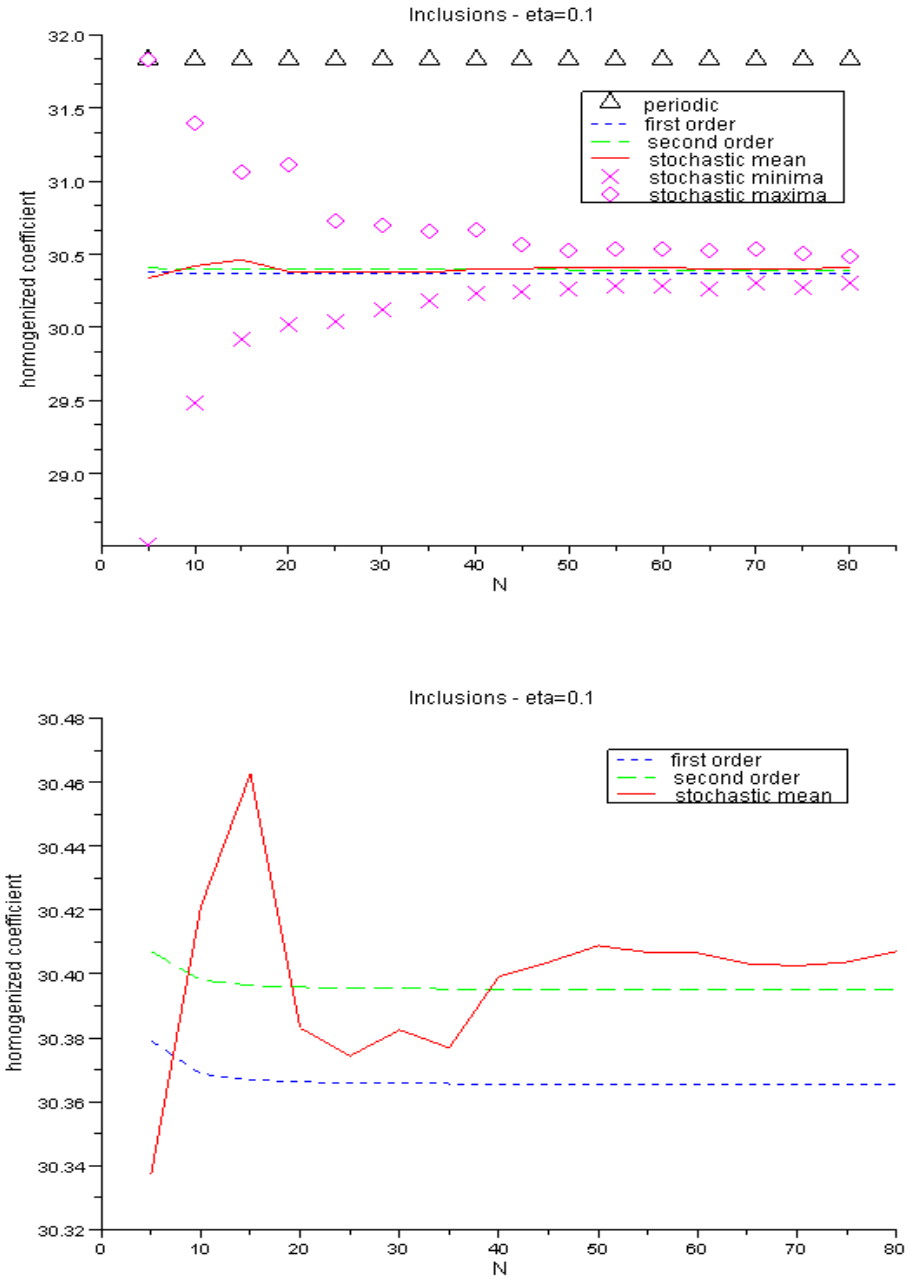

Figure 3.5: Results for material 1 and $\eta = 0.1$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

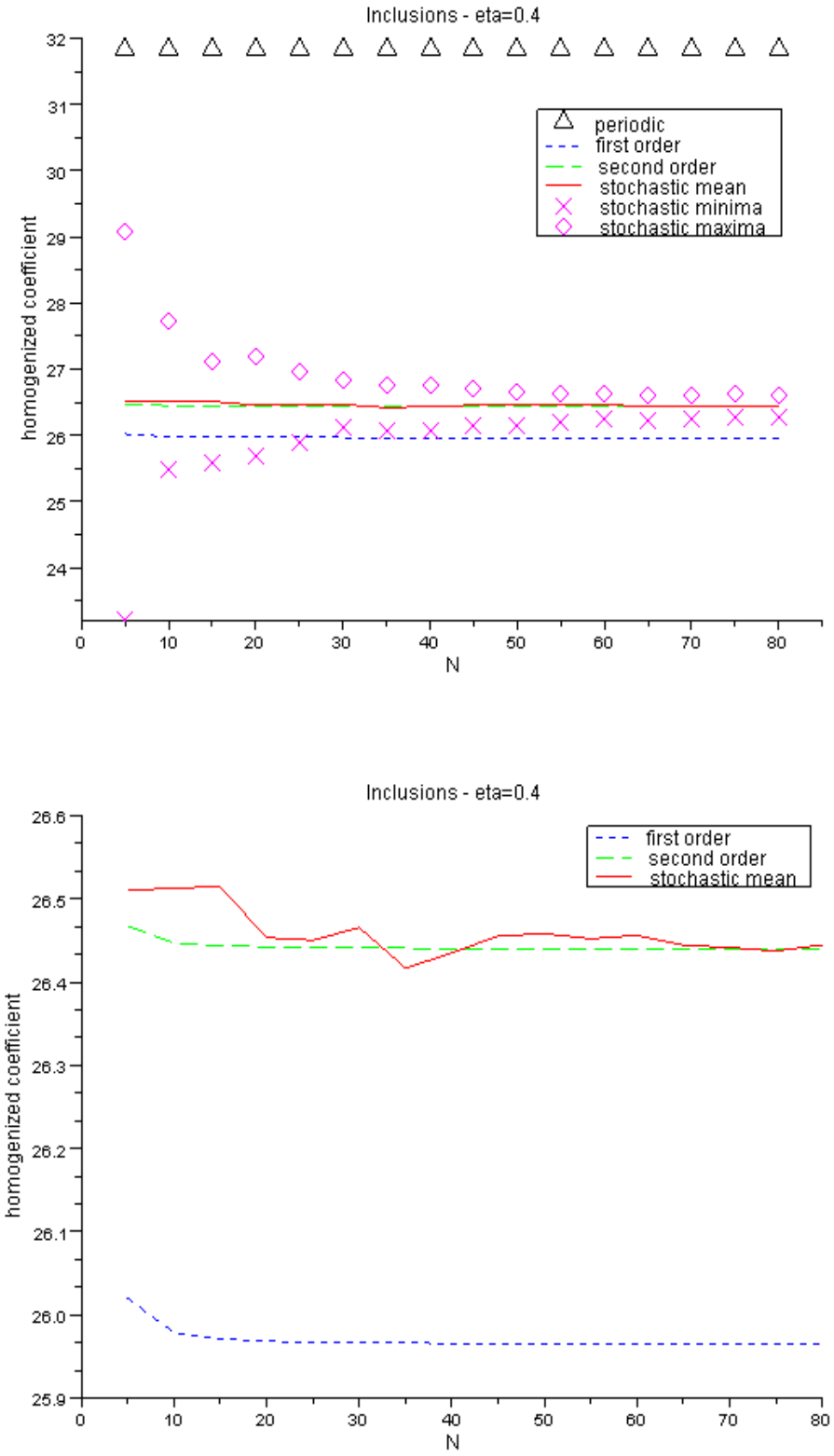

Figure 3.6: Results for material 1 and $\eta = 0.4$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

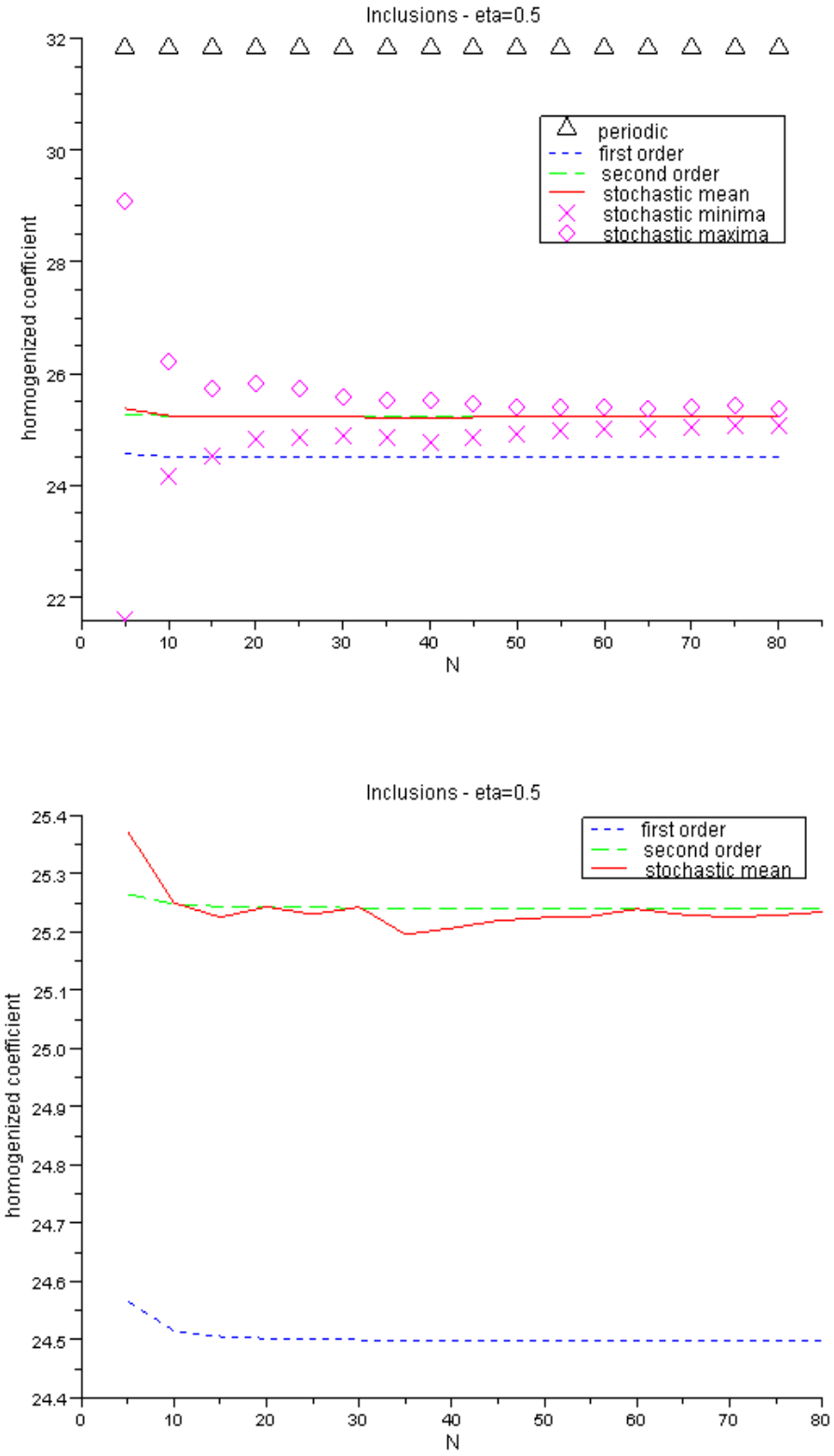

Figure 3.7: Results for material 1 and $\eta = 0.5$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

3.4.2.2 Results for material 2

We now show for material 2 the results for $\eta = 0.1$, $\eta = 0.3$ and $\eta = 0.4$ (Figures 3.8, 3.9 and 3.10 respectively).

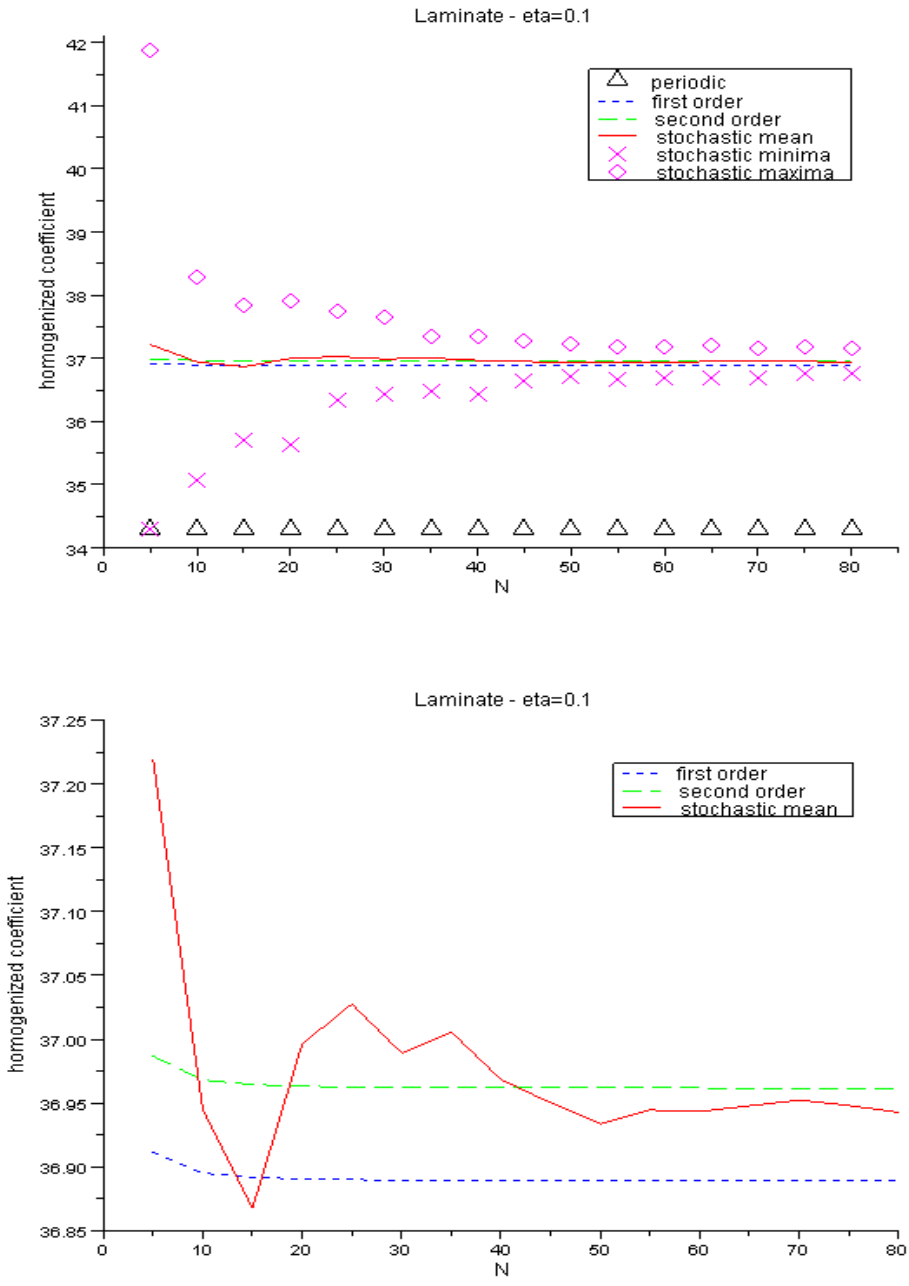

Figure 3.8: Results for material 2 and $\eta = 0.1$. Above: complete results. Below: zoom on $A_\eta^{*,N}$ and the first and second-order corrections.

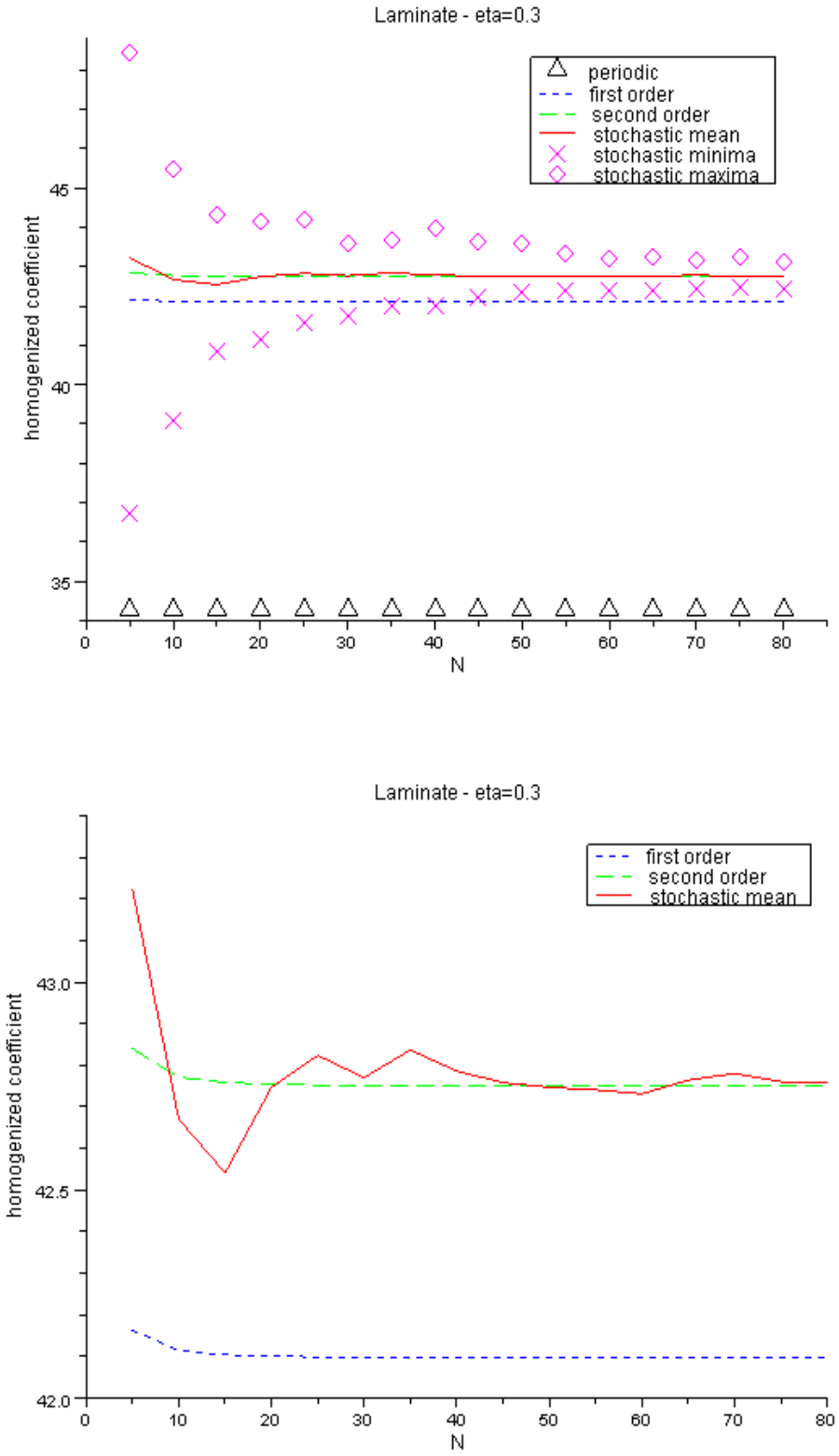

Figure 3.9: Results for material 2 and $\eta = 0.3$. Above: complete results. Below: zoom on $A_\eta^{*,N}$ and the first and second-order corrections.

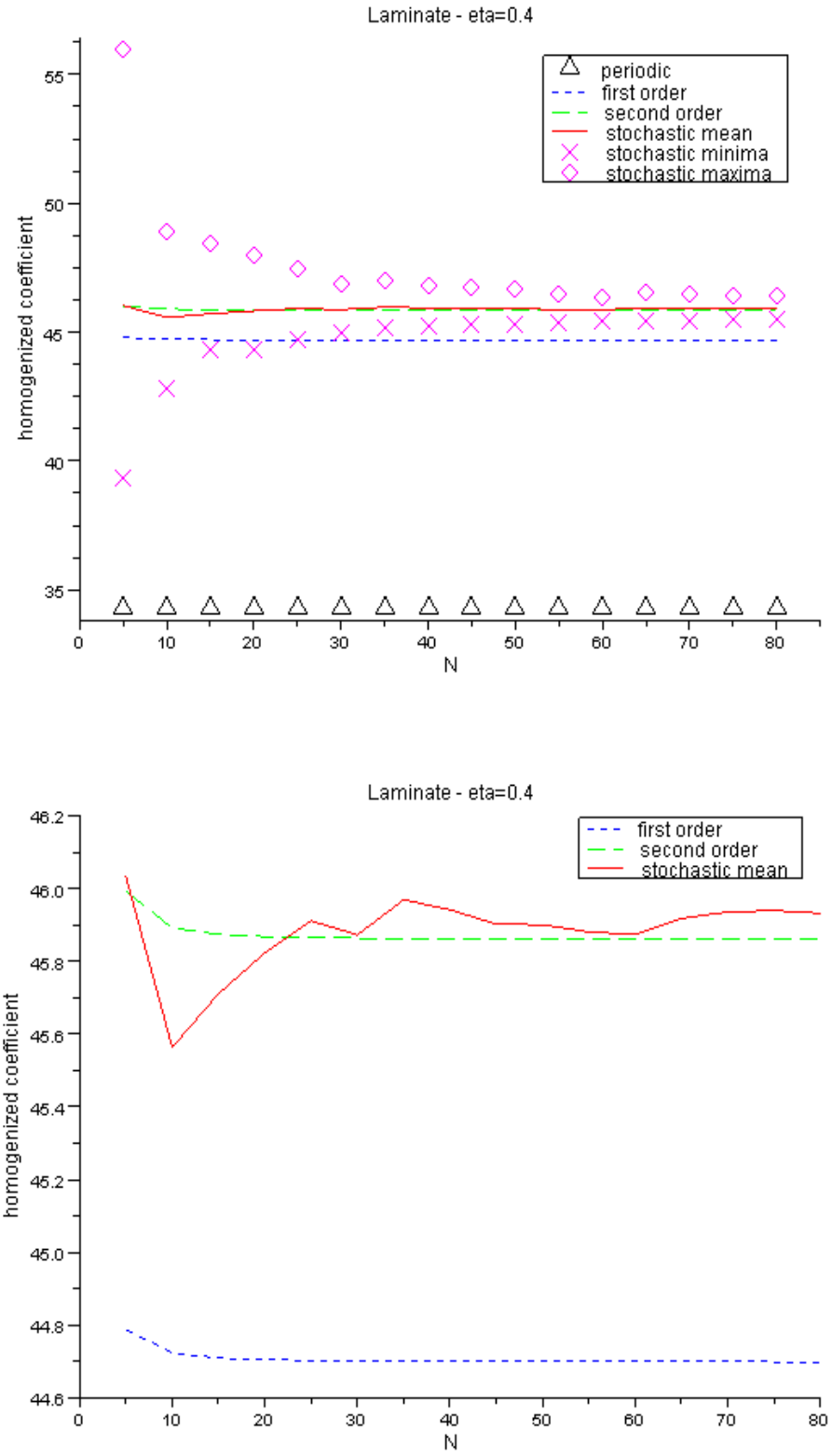

Figure 3.10: Results for material 2 and $\eta = 0.4$. Above: complete results. Below: zoom on $A_\eta^{*,N}$ and the first and second-order corrections.

3.4.2.3 Comments

Notice on the results for both materials (it is especially clear on the close-ups) that the first and second-order corrections $A_1^{*,N}$ and $A_2^{*,N}$ converge very fast in function of N, and in particular, as expected, much faster than the stochastic computation. Convergence of these deterministic computations is actually typically reached for $N = 10$.

Then, for all values of η, it is clear that the first-order correction enables to get substantially closer to A_η^*. The interest of the second-order term is also obvious as η gets larger, and we stress that the results are still excellent for η as large as 0.5, so that our approach is robust.

It is interesting to get some insight on the rate of convergence of the first-order correction, and to see whether the theoretical results of Lemma 3.3 still hold beyond the somewhat restrictive assumptions set in this lemma ($d \geq 3$, piecewise Hölder regularity on A_{per} and Dirichlet boundary conditions on ∂I_N). Recall that d is equal to 2 in our tests, and that $A_1^{*,N}$ is computed with periodic boundary conditions on the supercell I_N. Moreover, while the lattice of inclusions is piecewise Hölder continuous in the sense of Lemma 3.3 (meaning that there is an inclusion stricly contained in the unit cell Q and that the matrix A_{per} is Hölder continuous in each phase), the laminate is not.

We thus plot, for N going from 1 to 20 and for both materials, $\log(|(A_1^{*,N} - A_1^*)e_1 \cdot e_1|)$ in function of $\log(N)$. We recall that A_1^* is numerically given by $A_1^{*,80}$. For both materials the 20 points are arranged in a straight line (Figures 3.11 and 3.12). This leads us to perform a linear regression in order to obtain the slope of the lines. As regards material 1, we find a slope of -2.05 and a coefficient of correlation $R = 0.99$. For material 2, the slope is -1.9 with a coefficient of correlation equal to 0.95. The rate of convergence for both materials is then approximately $\mathcal{O}(N^{-d})$ with $d = 2$, which seems to indicate that the result of Lemma 3.3 still holds true in these circumstances.

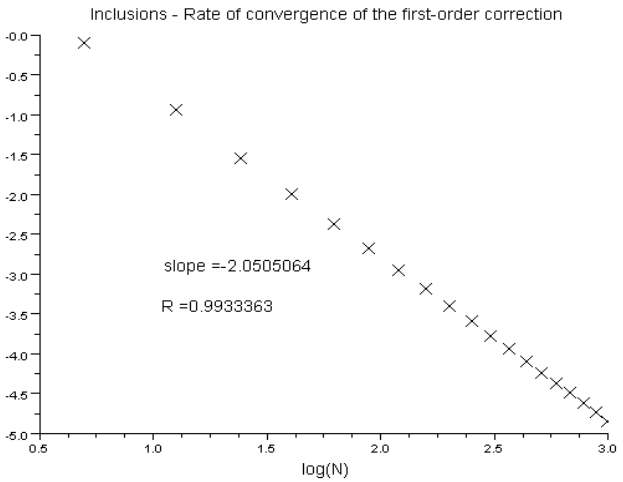

Figure 3.11: Rate of convergence of the first-order correction for material 1.

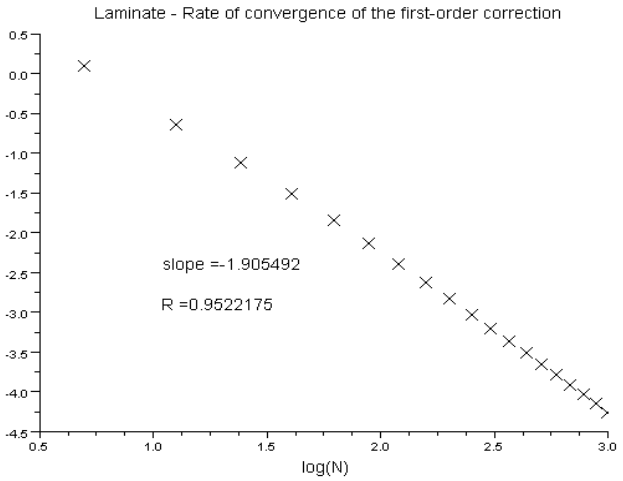

Figure 3.12: Rate of convergence of the first-order correction for material 2.

3.5 Appendix

The purpose of this Appendix is two-fold. In Section 3.5.1 we prove that the approach exposed in Section 3.3, which relies on formal considerations for general dimensions, is rigorous in dimension $d = 1$. In Section 3.5.2, we prove for convenience of the reader some technical results used in Section 3.3.

3.5.1 One-dimensional computations

Although we are aware that homogenization theory is very specific in dimension 1, and can be somehow misleading by its simplicity, it is still important to check that our approach is rigorously founded in this setting. This is the aim of this section.

To stress that we work in dimension one, we use lower-case letters a_{per} and c_{per} instead of A_{per} and C_{per}, respectively, as well as for all the tensors manipulated.

We recall that in dimension one, a^*_{per} and a^*_η are given by the explicit expressions

$$a^*_{per} = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}} \right)^{-1}, \quad a^*_\eta = \left(\mathbb{E} \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per} + b_\eta c_{per}} \right)^{-1}.$$

This enables us to prove the following elementary result which shows that our approach is correct in dimension one:

Proposition 3.5. *In dimension $d = 1$, it holds*

$$a^*_\eta = a^*_{per} + \eta \bar{a}^*_1 + \eta^2 \bar{a}^*_2 + \mathcal{O}(\eta^3),$$

where $\bar{a}^*_1$ and $\bar{a}^*_2$ are the limits as $N \to \infty$ of $a^{*,N}_1$ and $a^{*,N}_2$ defined generally by (3.28) and (3.29) respectively.

Proof. We compute

$$\begin{aligned}(a_\eta^*)^{-1} &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{1-\eta}{a_{per}} + \frac{\eta}{a_{per}+c_{per}}\right) \\ &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}} + \eta \int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{1}{a_{per}+c_{per}} - \frac{1}{a_{per}}\right) \\ &= (a_{per}^*)^{-1} \left(1 - \eta a_{per}^* \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right).\end{aligned}$$

This yields the expansion

$$\begin{aligned}a_\eta^* = a_{per}^* + \eta(a_{per}^*)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} + \eta^2(a_{per}^*)^3 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 \\ +\eta^3(a_{per}^*)^4 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^3 \left(1 - \eta a_{per}^* \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^{-1} \\ = a_{per}^* + \eta(a_{per}^*)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} + \eta^2(a_{per}^*)^3 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 \\ +\eta^3(a_{per}^*)^3 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^3 a_\eta^*.\end{aligned}$$

It follows from (3.20) and (3.21) that the function $\eta \to a_\eta^*$ is bounded on $[0,1]$. Therefore

$$\begin{aligned}a_\eta^* =& a_{per}^* + \eta(a_{per}^*)^2 \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} \\ &+ \eta^2(a_{per}^*)^3 \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 + \mathcal{O}(\eta^3).\end{aligned} \tag{3.86}$$

We now devote the rest of the proof to verifying that the coefficients of η and η^2 in (3.86) are indeed obtained as the limit as $N \to \infty$ of $a_1^{*,N}$ and $a_2^{*,N}$ generally defined by (3.28) and (3.29) respectively, in this particular one-dimensional setting.

The one-defect supercell solution $w^{1,0,N}$ generally defined by (3.23) satisfies here

$$\begin{cases} -\dfrac{d}{dx}\left(a_1^0\left(\dfrac{d}{dx}w^{1,0,N}+1\right)\right) = 0 \quad \text{in }]-\dfrac{N}{2},\dfrac{N}{2}[, \\ w^{1,0,N} \ N-\text{periodic}. \end{cases}$$

We easily compute

$$a_1^0(\frac{d}{dx}w^{1,0,N}+1) = N\left(\int_{-\frac{N}{2}}^{\frac{N}{2}} \frac{1}{a_{per}+\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}c_{per}}\right)^{-1}$$
$$= N\left(N(a_{per}^*)^{-1} - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^{-1}$$
$$= a_{per}^* + \frac{(a_{per}^*)^2}{N}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} + \frac{(a_{per}^*)^3}{N^2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 + o(N^{-2}).$$

Thus $a_1^{*,N}$ defined generally by (3.28) takes here the form

$$a_1^{*,N} = \int_{-\frac{N}{2}}^{\frac{N}{2}} a_1^0(\frac{d}{dx}w^{1,0,N}+1) - Na_{per}^* = (a_{per}^*)^2\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} + o(1),$$

and

$$a_1^{*,N} \underset{N\to\infty}{\to} \bar{a}_1^* = (a_{per}^*)^2\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}. \tag{3.87}$$

Likewise, for $k \in [\![-\frac{N-1}{2}, \frac{N-1}{2}]\!]\backslash\{0\}$,

$$a_2^{0,k}(\frac{d}{dx}w^{2,0,k,N}+1) = N\left(\int_{-\frac{N}{2}}^{\frac{N}{2}} \frac{1}{a_{per}+\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]\cup[k-\frac{1}{2},k+\frac{1}{2}]}c_{per}}\right)^{-1}$$
$$= N\left(N(a_{per}^*)^{-1} - 2\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^{-1}$$
$$= a_{per}^* + 2\frac{(a_{per}^*)^2}{N}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})} + 4\frac{(a_{per}^*)^3}{N^2}\int_{-\frac{1}{2}}^{\frac{1}{2}} \left(\frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 + o(N^{-2}),$$

which is independent of k (and so of the distance between the two defects). Hence, $a_2^{*,N}$ defined generally by (3.29) writes here

$$a_2^{*,N} = (a_{per}^*)^3\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2 + o(1),$$

and

$$a_2^{*,N} \underset{N\to\infty}{\to} \bar{a}_2^* = (a_{per}^*)^3\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{c_{per}}{a_{per}(a_{per}+c_{per})}\right)^2. \tag{3.88}$$

Using (3.86), (3.87) and (3.88), we verify that

$$a_\eta^* = a_{per}^* + \eta\bar{a}_1^* + \eta^2\bar{a}_2^* + \mathcal{O}(\eta^3).$$

□

Remark 3.7. *The fact that the distance between two defects does not play a role in the computation of $a_2^{*,N}$ is of course specific to the one-dimensional setting. As we have seen, this is not true in higher dimensions where the geometry comes into play.*

3.5.2 Some technical lemmas

The second part of this Appendix is different in nature. We prove here three technical lemmas that are useful for our proofs in Section 3.3. These results, or related ones, are probably well known and part of the mathematical literature. We prove them here under specific assumptions for the convenience of the reader and for consistency. We acknowledge several instructive discussions with Xavier Blanc on the content of this section.

We recall that $Q = [-\frac{1}{2}, \frac{1}{2}]^d$ and $I_N = [-\frac{N}{2}, \frac{N}{2}]^d$.

Lemma 3.6. *Consider $f \in L^2(Q)$, and a tensor field A from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$ such that there exist $\lambda > 0$ and $\Lambda > 0$ such that*

$$\forall \xi \in \mathbb{R}^d, \text{ a.e in } x \in \mathbb{R}^d, \ \lambda|\xi|^2 \leq A(x)\xi \cdot \xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|.$$

Consider q^N solution to

$$\begin{cases} -\operatorname{div}\left(A\nabla q^N\right) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } I_N, \\ q^N \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{3.89}$$

Then $\mathbb{1}_{I_N}\nabla q^N$ converges in $L^2(\mathbb{R}^d)$, when N goes to infinity, to ∇q^∞, where q^∞ is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A\nabla q^\infty\right) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } \mathbb{R}^d, \\ \nabla q^\infty \in L^2(\mathbb{R}^d). \end{cases} \tag{3.90}$$

Proof. We first obtain a bound on $\|\nabla q^N\|_{L^2(I_N)}$ and then, by compactness, extract a limit of this sequence.

Multiplying the first line of (3.34) by q^N and integrating by parts yields

$$\int_{I_N} A\nabla q^N \cdot \nabla q^N = -\int_Q f \cdot \nabla q^N, \tag{3.91}$$

from which we deduce

$$\|\nabla q^N\|_{L^2(I_N)} \leq \frac{1}{\lambda}\|f\|_{L^2(Q)}. \tag{3.92}$$

Consider now a bounded domain $\mathcal{D} \subset \mathbb{R}^d$. For N sufficiently large, we have $\mathcal{D} \subset I_N$ and so

$$\|\nabla q^N\|_{L^2(\mathcal{D})} \leq \frac{1}{\lambda}\|f\|_{L^2(Q)}.$$

Thus ∇q^N is bounded in $L^2(\mathcal{D})$ for every bounded subset $\mathcal{D} \subset \mathbb{R}^d$.

Using diagonal extraction and the weak compactness of $L^2_{loc}(\mathbb{R}^d)$, we can classically find a subsequence of ∇q^N such that, without changing the notation for simplicity,

$$\nabla q^N \rightharpoonup h \quad \text{weakly in } L^2_{loc}(\mathbb{R}^d). \tag{3.93}$$

We deduce from (3.92) and (3.93) that for every bounded subset $\mathcal{D} \subset \mathbb{R}^d$,

$$\|h\|_{L^2(\mathcal{D})} \leq \frac{1}{\lambda}\|f\|_{L^2(Q)}.$$

This implies that the vector h is in $L^2(\mathbb{R}^d)$.

We also deduce from (3.93) that for all $(i,j) \in [\![1,d]\!]^2$, $\frac{\partial h_j}{\partial x_i} = \frac{\partial h_i}{\partial x_j}$. This implies that h is the gradient of a function we call q^∞. Since $h \in L^2(\mathbb{R}^d)$, $\nabla q^\infty = h$ is in $L^2(\mathbb{R}^d)$ and q^∞ in $L^2_{loc}(\mathbb{R}^d)$.

Finally, (3.93) yields that ∇q^N converges to ∇q^∞ in $\mathcal{D}'(\mathbb{R}^d)$. We can then pass to the limit $N \to \infty$ in the first line of (3.89) and obtain

$$-\mathrm{div}\,(A\nabla q^\infty) = \mathrm{div}(\mathbb{1}_Q f) \quad \text{in } \mathbb{R}^d$$

in the sense of distributions.

We have proved that ∇q^N converges up to extraction and weakly in $L^2_{loc}(\mathbb{R}^d)$ to ∇q^∞, where q^∞ is in $L^2_{loc}(\mathbb{R}^d)$ and solves

$$\begin{cases} -\mathrm{div}\,(A\nabla q^\infty) = \mathrm{div}(\mathbb{1}_Q f) \quad \text{in } \mathbb{R}^d, \\ \nabla q^\infty \in L^2(\mathbb{R}^d). \end{cases} \tag{3.94}$$

We deduce from Lemma 3.7 thereafter that (3.94) has a solution unique up to an additive constant, so that ∇q^∞ is uniquely defined. A classical compactness argument then yields that the whole sequence ∇q^N converges weakly to ∇q^∞ in $L^2_{loc}(\mathbb{R}^d)$.

It is clear from what precedes that

$$\mathbb{1}_{I_N}\nabla q^N \rightharpoonup \nabla q^\infty \quad \text{weakly in } L^2(\mathbb{R}^d). \tag{3.95}$$

We now prove that the sequence $\mathbb{1}_{I_N}\nabla q^N$ actually converges strongly to ∇q^∞ in $L^2(\mathbb{R}^d)$.

Using a cut-off technique as in the proof of Lemma 3.7 thereafter, we deduce from (3.90) that

$$\int_{\mathbb{R}^d} A\nabla q^\infty \cdot \nabla q^\infty = -\int_Q f \cdot \nabla q^\infty. \tag{3.96}$$

The weak convergence of ∇q^N to ∇q^∞ implies that the right-hand side of (3.91) converges to the right-hand side of (3.96). Consequently,

$$\int_{I_N} A\nabla q^N \cdot \nabla q^N \to \int_{\mathbb{R}^d} A\nabla q^\infty \cdot \nabla q^\infty, \tag{3.97}$$

and, denoting by A_s the symmetric part of A, (3.97) is equivalent to

$$\int_{I_N} A_s\nabla q^N \cdot \nabla q^N \to \int_{\mathbb{R}^d} A_s\nabla q^\infty \cdot \nabla q^\infty. \tag{3.98}$$

A_s is of course a uniformly coercive tensor field, we can thus define its square root $A_s^{1/2}$. It follows from (3.98) that

$$\|A_s^{1/2}\mathbb{1}_{I_N}\nabla q^N\|_{L^2(\mathbb{R}^d)} \to \|A_s^{1/2}\nabla q^\infty\|_{L^2(\mathbb{R}^d)}. \tag{3.99}$$

On the other hand, multiplying (3.95) by $A_s^{1/2}$, we obtain

$$A_s^{1/2}\mathbb{1}_{I_N}\nabla q^N \rightharpoonup A_s^{1/2}\nabla q^\infty \quad \text{weakly in } L^2(\mathbb{R}^d). \tag{3.100}$$

Because of the uniform convexity of $L^2(\mathbb{R}^d)$, it is well known that (3.99) and (3.100) imply

$$A_s^{1/2}\mathbb{1}_{I_N}\nabla q^N \to A_s^{1/2}\nabla q^\infty \quad \text{strongly in } L^2(\mathbb{R}^d). \tag{3.101}$$

Multiplying (3.101) by $A_s^{-1/2}$, we finally have

$$\mathbb{1}_{I_N}\nabla q^N \to \nabla q^\infty \quad \text{strongly in } L^2(\mathbb{R}^d). \tag{3.102}$$

□

Lemma 3.7. *Let A be a tensor field from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$ such that there exist $\lambda > 0$ and $\Lambda > 0$ such that*

$$\forall \xi \in \mathbb{R}^d,\ \text{a.e in } x \in \mathbb{R}^d,\ \lambda|\xi|^2 \le A(x)\xi\cdot\xi \text{ and } |A(x)\xi| \le \Lambda|\xi|.$$

Consider $u \in L^2_{loc}(\mathbb{R}^d)$ solving

$$\begin{cases} -\operatorname{div}(A\nabla u) = 0 \quad \text{in } \mathbb{R}^d, \\ \nabla u \in L^2(\mathbb{R}^d). \end{cases} \tag{3.103}$$

Then u is constant.

Proof. We define a smooth cut-off function $\chi \in C^\infty(\mathbb{R}^d)$ such that $\chi = 1$ in the ball B_R, $\chi = 0$ in $\mathbb{R}^d\backslash B_{2R}$ and $\|\nabla\chi\|_{L^\infty(\mathbb{R}^d)} \le 2/R$.

Multiplying the first line of (3.103) by χu and integrating by parts, we obtain

$$\int_{\mathbb{R}^d} A\nabla u\cdot(\nabla u)\,\chi = -\int_{\mathbb{R}^d} A\nabla u\cdot(\nabla\chi)\,u.$$

Using the Cauchy-Schwarz inequality, this yields

$$\begin{aligned}\int_{B_R}|\nabla u|^2 &\le \frac{\Lambda}{\lambda}\|\nabla\chi\|_{L^\infty(\mathbb{R}^d)}\left(\int_{B_{2R}\backslash B_R}|\nabla u|^2\right)^{1/2}\left(\int_{B_{2R}\backslash B_R}|u|^2\right)^{1/2} \\ &\le \frac{2\Lambda}{R\lambda}\left(\int_{B_{2R}\backslash B_R}|\nabla u|^2\right)^{1/2}\left(\int_{B_{2R}\backslash B_R}|u|^2\right)^{1/2}.\end{aligned} \tag{3.104}$$

Defining

$$u_R = \frac{1}{|B_{2R}\backslash B_R|}\int_{B_{2R}\backslash B_R} u,$$

it is clear that $u - u_R$ is also a solution to (3.103) so that the above computations are valid for $u - u_R$. Since $\nabla(u - u_R) = \nabla u$, we deduce from (3.104) that

$$\int_{B_R} |\nabla u|^2 \leq \frac{2\Lambda}{R\lambda} \left(\int_{B_{2R}\backslash B_R} |\nabla u|^2 \right)^{1/2} \left(\int_{B_{2R}\backslash B_R} |u - u_R|^2 \right)^{1/2}. \tag{3.105}$$

We next apply the Poincaré-Wirtinger inequality to $u - u_R$ on $B_{2R}\backslash B_R$. There exists a constant $C(R)$ which depends only on R such that

$$\int_{B_{2R}\backslash B_R} |u - u_R|^2 \leq C(R) \int_{B_{2R}\backslash B_R} |\nabla u|^2.$$

An easy scaling argument shows that $C(R)$ is equal to R times the Poincaré-Wirtinger constant on $B_2\backslash B_1$, so that there exists a constant C such that

$$\int_{B_{2R}\backslash B_R} |u - u_R|^2 \leq CR \int_{B_{2R}\backslash B_R} |\nabla u|^2. \tag{3.106}$$

We deduce from (3.105) and (3.106) that

$$\int_{B_R} |\nabla u|^2 \leq \frac{2C\Lambda}{\lambda} \int_{B_{2R}\backslash B_R} |\nabla u|^2. \tag{3.107}$$

Since $\nabla u \in L^2(\mathbb{R}^d)$, the left-hand side of (3.107) converges to $\int_{\mathbb{R}^d} |\nabla u|^2$ when $R \to \infty$, and the right-hand side of (3.107) converges to 0. Then $\nabla u = 0$ and u is a constant. □

Lemma 3.8. *For $d \geq 3$, consider a $\mathbb{Z}^d$-periodic tensor field A such that there exist $\lambda > 0$ and $\Lambda > 0$ such that*

$$\forall \xi \in \mathbb{R}^d, \text{ a.e in } x \in \mathbb{R}^d, \ \lambda|\xi|^2 \leq A(x)\xi \cdot \xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|.$$

Assume that the unit cell Q contains an inclusion D, the boundary of which has regularity $\mathcal{C}^{1,\mu}$ for some $0 < \mu < 1$, and such that $\operatorname{dist}(D, \partial Q) > 0$. Assume also that A_{per} is Hölder continuous in $\overline{D}$ and in $\overline{Q\backslash D}$.

Let f be a function in $L^2(Q)$.

There exists a unique solution $u \in L^2_{loc}(\mathbb{R}^d)$ to

$$\begin{cases} -\operatorname{div}(A\nabla u) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } \mathbb{R}^d, \\ \nabla u \in L^2(\mathbb{R}^d), \lim\limits_{|x|\to\infty} u(x) = 0. \end{cases} \tag{3.108}$$

Defining also u_0 the unique solution to

$$\begin{cases} -\operatorname{div}(A\nabla u_0) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } \mathcal{O}, \\ u_0 \in H^1_0(\mathcal{O}), \end{cases} \tag{3.109}$$

where $\mathcal{O}$ is a bounded domain of $\mathbb{R}^d$ containing Q and such that $\mathrm{dist}(\partial\mathcal{O}, Q) > 1$, *there exists a constant K which depends only on λ, Λ, μ, d, f and the Hölder exponents, and* not *on the domain, such that for $|x| \geq 1$, it holds*

$$|u_0(x)| \leq \frac{K}{|x|^{d-1}}, \quad |\nabla u_0(x)| \leq \frac{K}{|x|^d},$$

$$|u(x)| \leq \frac{K}{|x|^{d-1}}, \quad |\nabla u(x)| \leq \frac{K}{|x|^d}.$$

Proof. Let G_0 be the Green kernel associated with A with homogeneous Dirichlet boundary conditions on $\partial\mathcal{O}$, uniquely defined by

$$\begin{cases} -\operatorname{div}(A\nabla G_0(\cdot, y)) = \delta_y \quad \text{in } \mathcal{O}, \\ G_0(\cdot, y) \in W_0^{1,1}(\mathcal{O}), \end{cases}$$

and G be the Green kernel associated with A on $\mathbb{R}^d$, unique solution to

$$\begin{cases} -\operatorname{div}(A\nabla G(\cdot, y)) = \delta_y \quad \text{in } \mathbb{R}^d, \\ G(\cdot, y) \in W^{1,1}_{loc}(\mathbb{R}^d) \cap H^1(\mathbb{R}^d \backslash B(y, 1)). \end{cases}$$

We deduce from arguments stated in [13, Lemma 4.2] and relying on [40, Theorem 3.3], and on [9, Lemma 16] when A is Hölder continuous and [52, Theorem 1.9] when A is piecewise Hölder continuous, that there exists a constant K depending only on λ, Λ, μ, d and the Hölder exponents, and not on the domain, such that

$$\forall (x, y) \in \mathcal{O}, \ \ |\nabla_y G_0(x, y)| \leq \frac{K}{|x-y|^{d-1}}, \quad |\nabla_x \nabla_y G_0(x, y)| \leq \frac{K}{|x-y|^d}, \tag{3.110}$$

$$\forall (x, y) \in \mathbb{R}^d, \ \ |\nabla_y G(x, y)| \leq \frac{K}{|x-y|^{d-1}}, \quad |\nabla_x \nabla_y G(x, y)| \leq \frac{K}{|x-y|^d}. \tag{3.111}$$

It is well known that u_0 solution to (3.109) can be represented as

$$u_0(x) = \int_{\mathcal{O}} G_0(x, y) \operatorname{div}(\mathbb{1}_Q f)(y) dy. \tag{3.112}$$

It is also clear that the function $\tilde{u}$ defined by

$$\tilde{u}(x) = \int_{\mathbb{R}^d} G(x, y) \operatorname{div}(\mathbb{1}_Q f)(y) dy \tag{3.113}$$

is a $H^1_{loc}(\mathbb{R}^d)$ function which satisfies

$$-\operatorname{div}(A\nabla\tilde{u}) = \operatorname{div}(\mathbb{1}_Q f)$$

in the sense of distributions.

Integrating by parts in (3.112) and (3.113) for $x \notin Q$, we have

$$u_0(x) = \int_Q \nabla_y G_0(x, y) \cdot f(y) dy, \ \ \tilde{u}(x) = \int_Q \nabla_y G(x, y) \cdot f(y) dy, \tag{3.114}$$

and then

$$\nabla u_0(x) = \int_Q \nabla_x \nabla_y G_0(x,y) \cdot f(y) dy, \quad \nabla \tilde{u}(x) = \int_Q \nabla_x \nabla_y G(x,y) \cdot f(y) dy. \tag{3.115}$$

Using estimates (3.110) and (3.111) in (3.114) and (3.115) respectively, we find that there exists a constant K depending only on λ, Λ, μ, d, f and the Hölder exponents, and not on the domain, such that for $|x| \geq 1$, we have

$$|u_0(x)| \leq \frac{K}{|x|^{d-1}} \quad \text{and} \quad |\nabla u_0(x)| \leq \frac{K}{|x|^d}, \tag{3.116}$$

$$|\tilde{u}(x)| \leq \frac{K}{|x|^{d-1}} \quad \text{and} \quad |\nabla \tilde{u}(x)| \leq \frac{K}{|x|^d}. \tag{3.117}$$

The function $\tilde{u}$ being in $H^1_{loc}(\mathbb{R}^d)$, we deduce from (3.117) that $\nabla \tilde{u} \in L^2(\mathbb{R}^d)$. Consequently, $\tilde{u}$ solves

$$\begin{cases} -\operatorname{div}(A \nabla \tilde{u}) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } \mathbb{R}^d, \\ \nabla \tilde{u} \in L^2(\mathbb{R}^d). \end{cases} \tag{3.118}$$

We know from Lemma 3.7 that (3.118) has a solution unique up to an additive constant. It follows from (3.117) that $\tilde{u}$ converges to zero at infinity, so that $\tilde{u} = u$ unique solution to (3.108).

□

Acknowledgements.

The authors are grateful to H. Ammari, X. Blanc and F. Legoll for fruitful discussions on the issue of the decay of Green kernels at infinity.

Chapitre 4

Elements of mathematical foundations for a numerical approach for weakly random homogenization problems

Sommaire

4.1 Introduction . . . 111
4.2 A model of a weakly random material and a first approach . . . 113
4.3 A formal approach . . . 121
4.3.1 A new assumption on the image measure . . . 121
4.3.2 An ergodic approximation of the homogenized tensor . . . 124
4.3.3 Convergence of the first-order term . . . 128
4.3.4 Convergence of the second-order term . . . 131
4.4 Numerical experiments . . . 137
4.4.1 Methodology . . . 137
4.4.2 An example of setting for our theory in Section 4.2 (and 4.3) . . . 140
4.4.3 A first example of setting for our formal approach of Section 4.3 . . . 140
4.4.4 A second example of setting for our formal approach of Section 4.3 . . . 146
4.5 Appendix . . . 151
4.5.1 Elements of distribution theory . . . 151
4.5.2 Some technical results . . . 152
4.5.3 The one-dimensional case . . . 155
4.5.4 A proof of the approach of Section 4.3 in a specific setting . . . 161

4.1 Introduction

Our purpose here is to follow up on the study of the weakly random homogenization model presented in Chapter 3. Let us recall, for consistency, that we consider homogenization for the following elliptic problem

$$\begin{cases} -\operatorname{div}\left((A_{per}(\frac{x}{\varepsilon}) + b_\eta(\frac{x}{\varepsilon},\omega)C_{per}(\frac{x}{\varepsilon}))\nabla u_\varepsilon\right) = f(x) \text{ in } \mathcal{D} \subset \mathbb{R}^d, \\ u_\varepsilon = 0 \text{ on } \partial\mathcal{D}, \end{cases} \tag{4.1}$$

where the tensor A_{per} models a reference $\mathbb{Z}^d$-periodic material which is randomly perturbed by the $\mathbb{Z}^d$-periodic tensor C_{per}, the stochastic nature of the problem being encoded in the stationary ergodic scalar field b_η (the latter getting small when η vanishes). We have studied in Chapter 3 (see also [6]) the case of a perturbation that has a Bernoulli law with parameter η, meaning that b_η is equal to 1 with probability η and 0 with probability $1-\eta$. In the present chapter, we address more general laws. The common setting is that all the perturbations we consider are, to some extent, rare events which, although rare, modify the homogenized properties of the material. Our approach is a perturbative approach, and consists in approximating the stochastic homogenization problem for

$$A_\eta(x,\omega) = A_{per}(x) + b_\eta(x,\omega)\, C_{per}$$

using the periodic homogenization problem for A_{per}. In short, let us say that our main contribution is to derive an expansion

$$A^*_\eta = A^*_{per} + \eta \bar{A}^*_1 + \eta^2 \bar{A}^*_2 + o(\eta^2), \tag{4.2}$$

where A^*_η and A^*_{per} are the homogenized tensors associated with A_η and A_{per} respectively, and the first and second-order corrections $\bar{A}^*_1$ and $\bar{A}^*_2$ can be, loosely speaking, computed in terms of the microscopic properties of A_{per} and C_{per} and the statistics of second order of the random field b_η. The formulation has been made precise in Chapter 3, and the changes we introduce in the present chapter are detailed in Sections 4.2 and 4.3 below.

Motivations behind setting (4.1), as well as a review of the mathematical literature on similar issues, can be found in Chapter 3. We complement our study of the perturbative approach introduced in Chapter 3 in two different directions.

In Section 4.2, we rigorously establish an asymptotic expansion of the homogenized tensor in a mathematical setting where our input parameter (the field b_η in (4.1)) enjoys appropriate weak convergence properties, as η vanishes, in a reflexive Banach space, namely a Lebesgue space $L^\infty(\mathcal{D}, L^p(\Omega))$ (with $p > 1$). In such a setting, we are in position to rigorously prove a first order asymptotic expansion (announced in [8] and precisely stated in [8, Théorème 2.1] and Theorem 4.2 below) for the homogenization of A_η, using simple functional analysis techniques very similar to those exposed in [18]. In our Corollaries 4.3 and 4.4, the expansion is pushed to second order under additional assumptions.

Our aim in Section 4.3 is to further extend our formal theory of Chapter 3. Recall that this formal theory, rather than manipulating the random field b_η itself, consists in focusing on its law. We indeed assume that the image measure (the law) corresponding to the perturbation admits an expansion (see (4.48) below) with respect to η in the sense of distributions. While Chapter 3 has only addressed the specific case of a Bernoulli law, we consider here more general laws and proceed with the same formal derivations. These derivations lead to first-order and second-order corrections $\bar{A}^*_1$ and $\bar{A}^*_2$ in (4.2) obtained as limits when $N \to \infty$ of sequences of tensors $A^{*,N}_1$ and $A^{*,N}_2$ computed on the supercell $[-\frac{N}{2}, \frac{N}{2}]^d$. It is the purpose of Propositions 4.7 and 4.8 to prove the convergence of $A^{*,N}_1$ and $A^{*,N}_2$ respectively. As in Chapter 3, our approach in this section exhibits close ties

with classical defect-type theories used in solid state physics.

We emphasize that, in sharp contrast to the exact stochastic homogenization of A_η, the determination of the first and second-order terms in (4.2) relies on entirely deterministic computations, albeit of very different kind, for both approaches of Sections 4.2 and 4.3.

Finally, a comprehensive series of numerical tests in Section 4.4 show, beyond those contained in Chapter 3, that the two approaches exposed here are efficient and quite robust: the computational workload induced by the perturbative approach is light compared to the direct homogenization of A_η, and expansion (4.2) proves to be accurate for not so small perturbations.

We complement the text by a long Appendix. The reader less interested in theoretical issues can easily omit the reading of this Appendix. Besides providing, in Sections 4.5.1 and 4.5.2 and for consistency, some theoretical results useful in the body of the text, the purpose of this Appendix is two-fold. We examine in details in Section 4.5.3 the one-dimensional setting, and we show that, expectedly, all our formal expansions can be made rigorous through explicit computations. We next demonstrate, in Section 4.5.4, that our two modes of derivation coincide in a particular setting appropriate for both the theoretical results of Section 4.2 and the formal results of Section 4.3. This final section therefore provides a *proof* of our formal manipulations of Section 4.3, in a setting – we concede it – that is not the setting the approach was designed to specifically address. Definite conclusions on the theoretical validity of the approach developped in Section 4.3 are yet to be obtained, even though applicability and efficiency are beyond doubt.

Throughout this chapter, and unless otherwise mentioned, C denotes a constant that depends at most on the ambient dimension d, and on the tensors A_{per} and C_{per}. We write $C(\gamma)$ when C depends on γ and possibly on d, A_{per} and C_{per}. The indices i and j denote indices in $[\![1, d]\!]$.

4.2 A model of a weakly random material and a first approach

For consistency, we first recall the general setting introduced in Chapter 3.

Throughout this chapter, $(\Omega, \mathcal{F}, \mathbb{P})$ denotes a probability space with $\mathbb{P}$ the probability measure and $\omega \in \Omega$ an event. We denote by $\mathbb{E}(X)$ the expectation of a random variable X and $Var(X)$ its variance.

We assume that the group $(\mathbb{Z}^d, +)$ acts on Ω and denote by $\tau_k, k \in \mathbb{Z}^d$, the group action. We also assume that this action is measure-preserving, that is,

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

and ergodic:

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

We call $F \in L^1_{loc}(\mathbb{R}^d, L^1(\Omega))$ stationary if

$$\forall k \in \mathbb{Z}^d,\ F(x+k,\omega) = F(x,\tau_k\omega) \ \text{ almost everywhere in } x \in \mathbb{R}^d \text{ and } \omega \in \Omega. \tag{4.3}$$

Notice that if F is deterministic, the notion of stationarity used here reduces to $\mathbb{Z}^d$-periodicity, that is,

$$\forall k \in \mathbb{Z}^d, F(x+k) = F(x) \ \text{ almost everywhere in } x \in \mathbb{R}^d. \tag{4.4}$$

We then consider the tensor field from $\mathbb{R}^d \times \Omega$ to $\mathbb{R}^{d\times d}$:

$$A_\eta(x,\omega) = A_{per}(x) + b_\eta(x,\omega)C_{per}(x), \tag{4.5}$$

where A_{per} and C_{per} are two deterministic $\mathbb{Z}^d$-periodic tensor fields and b_η a stationary ergodic scalar field. The matrix A_{per} models the reference periodic material, perturbed by C_{per}. This perturbation is random, thus the presence of b_η. We refer the reader to [18] for a more detailed presentation of the stationary ergodic setting in a similar weakly random framework.

We make the following assumptions on the random field b_η:

$$\exists M > 0, \forall \eta > 0, \|b_\eta\|_{L^\infty(Q\times\Omega)} \leq M, \tag{4.6}$$

$$\|b_\eta\|_{L^\infty(Q;L^2(\Omega))} \underset{\eta\to 0^+}{\to} 0, \tag{4.7}$$

where Q is the unit cell $[-\frac{1}{2},\frac{1}{2}]^d$.

Assumption (4.7) encodes that the perturbation for small η is a rare event. Still, it is able to significantly modify the local structure of the material when it happens, for we do not require it to be small in $L^\infty(Q\times\Omega)$ as $\eta \to 0$.

We additionally assume that there exist $0 < \alpha \leq \beta$ such that for all $\xi \in \mathbb{R}^d$, for almost all $x \in \mathbb{R}^d$ and for all $s \in [-M, M]$,

$$\alpha|\xi|^2 \leq A_{per}(x)\xi\cdot\xi, \qquad \alpha|\xi|^2 \leq \left(A_{per} + sC_{per}\right)(x)\xi\cdot\xi, \tag{4.8}$$

$$A_{per}(x)\xi| \leq \beta|\xi|, \qquad |\left(A_{per} + sC_{per}\right)(x)\xi| \leq \beta|\xi|. \tag{4.9}$$

We can therefore use the classical stochastic homogenization results (see for instance [46] for a comprehensive review or Chapter 3 for a concise presentation). The cell problems associated with (4.5) read

$$\begin{cases} -\operatorname{div}\left(A_\eta(\nabla w_i^\eta + e_i)\right) = 0 \quad \text{in } \mathbb{R}^d, \\ \nabla w_i^\eta \ \text{ stationary}, \ \ \mathbb{E}\left(\displaystyle\int_Q \nabla w_i^\eta\right) = 0. \end{cases} \tag{4.10}$$

Problem (4.10) has a solution unique up to the addition of a random constant in

$$\{w \in L^2_{loc}(\mathbb{R}^d, L^2(\Omega)), \quad \nabla w \in L^2_{unif}(\mathbb{R}^d, L^2(\Omega))\}.$$

The function w_i^η is called the i-th corrector or cell solution.

The homogenized tensor A_η^* is given by

$$\forall i \in [\![1,d]\!],\ \ A_\eta^* e_i = \mathbb{E}\left(\int_Q A_\eta(\nabla w_i^\eta + e_i)\right). \tag{4.11}$$

Throughout the rest of this chapter we will denote by w_i^0 the i-th cell solution associated with A_{per}, defined up to an additive constant in the space of $H^1_{loc}(\mathbb{R}^d)$ functions that are $\mathbb{Z}^d$-periodic by

$$\begin{cases} -\operatorname{div}\left(A_{per}(\nabla w_i^0 + e_i)\right) = 0 \quad \text{in } Q, \\ w_i^0 \ \ \mathbb{Z}^d - \text{periodic.} \end{cases} \tag{4.12}$$

The periodic homogenized tensor is then given by

$$\forall i \in [\![1,d]\!],\ \ A_{per}^* e_i = \int_Q A_{per}(\nabla w_i^0 + e_i). \tag{4.13}$$

Due to the specific form of A_η, the following zero-order result can be easily proved. The proof is actually the same as that of Lemma 3.1 of Chapter 3, which relies on the fact that $\|b_\eta\|_{L^\infty(Q;L^2(\Omega))}$ converges to 0 as η tends to 0.

Lemma 4.1. *When $\eta \to 0$, $A_\eta^* \to A_{per}^*$.*

Our goal is to find an asymptotic expansion for A_η with respect to η, and a first answer is given by the following theorem announced as Théorème 1 in [8]:

Theorem 4.2 (Théorème 1, [8])**.** *Assume that b_η satisfies (4.6) and (4.7), and denote by $m_\eta = \|b_\eta\|_{L^\infty(Q;L^2(\Omega))}$. There exists a subsequence of η, still denoted η for the sake of simplicity, such that $\frac{b_\eta}{m_\eta}$ converges weakly-* in $L^\infty(Q;L^2(\Omega))$ to a limit field denoted by $\bar{b}_0$ when $n \to 0$. Then*

- *for all $i \in [\![1,d]\!]$, the following expansion*

$$\nabla w_i^\eta = \nabla w_i^0 + m_\eta \nabla v_i^0 + o(m_\eta) \tag{4.14}$$

holds weakly in $L^2(Q;L^2(\Omega))$, where w_i^0 is the solution to the i-th periodic cell problem and v_i^0 is solution to

$$\begin{cases} -\operatorname{div}(A_{per}\nabla v_i^0) = \operatorname{div}\left(\bar{b}_0 C_{per}(\nabla w^0 + e_i)\right) \quad \text{in } \mathbb{R}^d, \\ \nabla v_i^0 \ \text{ stationary, } \mathbb{E}\left(\int_Q \nabla v_i^0\right) = 0. \end{cases} \tag{4.15}$$

- *A_η^* can be expanded up to first order as*

$$A_\eta^* = A_{per}^* + m_\eta \tilde{A}_1^* + o(m_\eta), \tag{4.16}$$

where

$$\forall i \in [\![1,d]\!],\ \tilde{A}_1^* e_i = \int_Q \mathbb{E}(\bar{b}_0) C_{per}(\nabla w_i^0 + e_i) + \int_Q A_{per}\nabla \mathbb{E}(v_i^0). \tag{4.17}$$

Proof. We fix $i \in [\![1,d]\!]$ and define $v_i^\eta = \frac{w_i^\eta - w_i^0}{m_\eta}$. v_i^η is solution to

$$\begin{cases} -\operatorname{div}\left(A_\eta \nabla v_i^\eta\right) = \operatorname{div}\left(\dfrac{b_\eta}{m_\eta} C_{per}\left(\nabla w_i^0 + e_i\right)\right) & \text{in } \mathbb{R}^d, \\ \nabla v_i^\eta \text{ stationary, } \mathbb{E}\left(\displaystyle\int_Q \nabla v_i^\eta\right) = 0. \end{cases} \tag{4.18}$$

Using an argument similar to that used in the proof of Lemma 3.1 in Chapter 3, we have

$$\forall \eta > 0, \ \|\nabla v_i^\eta\|_{L^2(Q\times\Omega)} \leq \frac{1}{\alpha} \|C_{per}\left(\nabla w_i^0 + e_i\right)\|_{L^2(Q)}.$$

where α is defined in (4.8).

The sequence ∇v_i^η is bounded in $L^2(Q\times\Omega)$ and therefore, up to extraction, weakly converges in $L^2(Q\times\Omega)$ to some limit which is necessarily a gradient and which we denote ∇v_i^0. Since b_η converges strongly to 0 in $L^2(Q\times\Omega)$, $b_\eta \nabla v_i^\eta$ converges to 0 in $\mathcal{D}'(Q\times\Omega)$. It is then easy to pass to the limit $\eta \to 0$ in (4.18) and to deduce that v_i^0 is solution to

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla v_i^0\right) = \operatorname{div}\left(\bar{b}_0 C_{per}\left(\nabla w_i^0 + e_i\right)\right) & \text{in } \mathbb{R}^d, \\ \nabla v_i^0 \text{ stationary, } \mathbb{E}\left(\displaystyle\int_Q \nabla v_i^0\right) = 0. \end{cases}$$

Thus $\frac{\nabla w_i^\eta - \nabla w_i^0}{m_\eta}$ converges, up to extraction, weakly to ∇v_i^0 in $L^2(Q\times\Omega)$. This amounts to say that we have the following first-order expansion:

$$\nabla w_i^\eta = \nabla w_i^0 + m_\eta \nabla v_i^0 + o(m_\eta) \quad \text{in } L^2(Q\times\Omega) \text{ weak.}$$

Inserting this expansion in (4.11), we obtain

$$A_\eta^* e_i \ = \ A_{per}^* e_i + m_\eta \int_Q \mathbb{E}(\bar{b}_0) C_{per}(\nabla w_i^0 + e_i) + m_\eta \int_Q A_{per} \nabla \mathbb{E}(v_i^0) + o(m_\eta),$$

which concludes the proof.

□

Remark 4.1. *Notice that taking the expectation of both sides of (4.15), $\mathbb{E}(v_i^0)$ is actually the $\mathbb{Z}^d$-periodic function that is the unique solution (up to an additive constant) to*

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla \mathbb{E}(v_i^0)\right) = \operatorname{div}\left(\mathbb{E}(\bar{b}_0) C_{per}\left(\nabla w_i^0 + e_i\right)\right) & \text{in } Q, \\ \mathbb{E}(v_i^0) \ \ \mathbb{Z}^d - \text{periodic.} \end{cases} \tag{4.19}$$

The computation of A_η^* up to the first order in m_η only requires solving $2d$ deterministic problems, namely (4.12) and (4.19), in the unit cell Q.

In fact, the situation is even more advantageous when A_{per} is a symmetric matrix, as shown by our next remark.

Remark 4.2. *Defining the adjoint problems to the cell problems (4.12),*

$$\begin{cases} -\operatorname{div}\left(A_{per}^T(\nabla\tilde{w}_i^0+e_i)\right)=0 \quad \text{in } Q,\\ \tilde{w}_i^0 \ \ \mathbb{Z}^d-\text{periodic}, \end{cases} \tag{4.20}$$

where we have denoted by A_{per}^T the transposed matrix of A_{per}, allows to write the first-order correction (4.17) in a slightly different form. Indeed, multiplying (4.19) by $\tilde{w}_j^0$ and integrating by parts, we obtain

$$\int_Q A_{per}\nabla\mathbb{E}(v_i^0)\cdot\nabla\tilde{w}_j^0=-\int_Q\mathbb{E}(\bar{b}_0)C_{per}\left(\nabla w_i^0+e_i\right)\cdot\nabla\tilde{w}_j^0.$$

Likewise, multiplying (4.20) by $\nabla\mathbb{E}(v_i^0)$ and integrating by parts yields

$$\int_Q A_{per}\nabla\mathbb{E}(v_i^0)\cdot\left(\nabla\tilde{w}_j^0+e_j\right)=0.$$

Combining these equalities gives

$$\int_Q A_{per}\nabla\mathbb{E}(v_i^0)\cdot e_j=\int_Q\mathbb{E}(\bar{b}_0)C_{per}\left(\nabla w_i^0+e_i\right)\cdot\nabla\tilde{w}_j^0,$$

and thus (4.17) may be equivalently phrased as

$$\forall(i,j)\in[\![1,d]\!]^2,\ \tilde{A}_1^*e_i\cdot e_j \ = \ \int_Q\mathbb{E}(\bar{b}_0)C_{per}(\nabla w_i^0+e_i)\cdot(\nabla\tilde{w}_j^0+e_j). \tag{4.21}$$

When A_{per} is symmetric, $\tilde{w}_j^0=w_j^0$, and solving the periodic cell problems (4.12) suffices to determine A_η^* up to the first order in m_η.

Pushing expansion (4.16) to second order requires more information on b_η:

Corollary 4.3. *Assume in addition to (4.6) and (4.7) that*

$$b_\eta=\eta\bar{b}_0+\eta^2\bar{r}_0+o(\eta^2) \quad \text{weakly}-^* \text{ in } L^\infty(Q;L^2(\Omega)). \tag{4.22}$$

Then

- *for all $i\in[\![1,d]\!]$, the following expansion*

$$\nabla w_i^\eta=\nabla w_i^0+\eta\nabla v_i^0+\eta^2\nabla z_i^0+o(\eta^2) \tag{4.23}$$

holds weakly in $L^2(Q;L^2(\Omega))$, where z_i^0 is solution to

$$\begin{cases} -\operatorname{div}(A_{per}\nabla z_i^0)=\operatorname{div}\left(\bar{r}_0C_{per}(\nabla w_i^0+e_i)\right)+\operatorname{div}\left(\bar{b}_0C_{per}\nabla v_i^0\right) \quad \text{in } \mathbb{R}^d,\\ \nabla z_i^0 \ \text{ stationary, } \mathbb{E}\left(\displaystyle\int_Q\nabla z_i^0\right)=0. \end{cases} \tag{4.24}$$

- *A_η^* can be expanded up to second order as*

$$A_\eta^* = A_{per}^* + \eta \tilde{A}_1^* + \eta^2 \tilde{A}_2^* + o(\eta^2), \tag{4.25}$$

where $\tilde{A}_1^*$ is defined by (4.17) and for all $i \in [\![1,d]\!]$,

$$\tilde{A}_2^* e_i = \int_Q \mathbb{E}(\bar{r}_0) C_{per}(\nabla w_i^0 + e_i) + \eta^2 \int_Q C_{per}\mathbb{E}(\bar{b}_0 \nabla v_i^0) + \int_Q A_{per} \nabla \mathbb{E}(z_i^0), \tag{4.26}$$

or equivalently, for all $(i,j) \in [\![1,d]\!]^2$,

$$\tilde{A}_2^* e_i \cdot e_j = \int_Q \mathbb{E}(\bar{r}_0) C_{per}(\nabla w_i^0 + e_i) \cdot (\nabla \tilde{w}_j{}^0 + e_j) + \int_Q C_{per}\mathbb{E}(\bar{b}_0 \nabla v_i^0) \cdot (\nabla \tilde{w}_j{}^0 + e_j). \tag{4.27}$$

Proof. The proof follows the same pattern as that of Theorem 4.2. The computation of the second order relies on the fact that (4.22) implies that $\frac{b_\eta}{\eta}$ converges strongly to $\bar{b}_0$ in $L^\infty(Q; L^2(\Omega))$, whereas the convergence was weak in Theorem 4.2. Likewise, the expansion of the cell solution, namely (4.23), implies that $\frac{\nabla w_i^\eta - \nabla w_i^0}{\eta}$ converges strongly to ∇v_i^0 in $L^2(Q; L^2(\Omega))$. We then obtain (4.25) and (4.26) by inserting (4.23) in (4.11), and deduce (4.27) from (4.26) as in Remark 4.2. □

The computation of A_η^* up to the order η^2 is much more intricate than that up to the order η, for it requires determining $\mathbb{E}(\bar{b}_0 \nabla v_i^0)$. Computing the periodic deterministic function $\mathbb{E}(v_i^0)$ solution to the simpler problem (4.19) is not sufficient in general. We have to determine the stationary random field v_i^0 solution to (4.15) in $\mathbb{R}^d$.

It turns out that in a particular, practically relevant setting, we may still avoid solving the random problem (4.15). This setting presents the additional advantage to provide insight on the influence of spatial correlation.

Corollary 4.4. *Assume that b_η is uniform in each cell of $\mathbb{Z}^d$, and writes*

$$b_\eta(x,\omega) = \sum_{k \in \mathbb{Z}^d} \mathbb{1}_{Q+k}(x) B_\eta(\tau_k \omega), \tag{4.28}$$

where B_η satisfies

$$\forall \eta > 0, \|B_\eta\|_{L^\infty(\Omega)} \leq M, \tag{4.29}$$

$$B_\eta = \eta \bar{B}_0 + \eta^2 \bar{R}_0 + o(\eta^2) \quad \text{weakly in } L^2(\Omega). \tag{4.30}$$

Assume also that

$$\sum_{k \in \mathbb{Z}^d} |cov(\bar{B}_0, \bar{B}_0(\tau_k \cdot))| < \infty. \tag{4.31}$$

Then the second-order term (4.27) can be rewritten

$$\begin{aligned} \tilde{A}_2^* e_i \cdot e_j =& \mathbb{E}(\bar{R}_0) \int_Q C_{per}(\nabla w_i^0 + e_i) \cdot (\nabla \tilde{w}_j{}^0 + e_j) + Var(\bar{B}_0) \int_Q C_{per} \nabla t_i \cdot (\nabla \tilde{w}_j^0 + e_j) \\ &+ (\mathbb{E}(\bar{B}_0))^2 \int_Q C_{per} \nabla s_i \cdot (\nabla \tilde{w}_j^0 + e_j) \\ &+ \sum_{k \in \mathbb{Z}^d \setminus \{0\}} cov(\bar{B}_0, \bar{B}_0(\tau_k \cdot)) \int_Q C_{per} \nabla t_i(\cdot - k) \cdot (\nabla \tilde{w}_j^0 + e_j), \end{aligned} \tag{4.32}$$

where t_i is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla t_i\right) = \operatorname{div}\left(C_{per}\mathbb{1}_Q\left(\nabla w_i^0 + e_i\right)\right) & \text{in } \mathbb{R}^d, \\ \nabla t_i \in L^2(\mathbb{R}^d), \end{cases} \tag{4.33}$$

and s_i solves

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla s_i\right) = \operatorname{div}\left(C_{per}\left(\nabla w_i^0 + e_i\right)\right) & \text{in } Q, \\ s_i \ \ \mathbb{Z}^d - \text{periodic}. \end{cases} \tag{4.34}$$

Proof. We notice first that the specific form (4.28) of b_η considered implies that $\bar{b}_0$ and $\bar{r}_0$ defined in (4.22) here write

$$\bar{b}_0(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)\bar{B}_0(\tau_k\omega), \tag{4.35}$$

$$\bar{r}_0(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)\bar{R}_0(\tau_k\omega). \tag{4.36}$$

The rest of the proof mainly consists in showing that in this particular setting, ∇v_i^0 and the product $\bar{b}_0\nabla v_i^0$ can be written using the deterministic functions t_i and s_i. The existence of t_i and its uniqueness up to an additive constant come from Lemma 4.15, and Lemma 3.7 in Chapter 3 respectively.

We start by proving that the sum

$$\sum_{k\in\mathbb{Z}^d} \left(\bar{B}_0(\tau_k\omega) - \mathbb{E}(\bar{B}_0)\right)\nabla t_i(x-k) \tag{4.37}$$

is a convergent series in $L^2(Q\times\Omega)$.

To this end, we compute the norm of the remainder of this series:

$$\begin{aligned}
&\left\| \sum_{|k|\geq N} \left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right)\nabla t_i(\cdot - k) \right\|^2_{L^2(Q\times\Omega)} \\
&= \sum_{|k|\geq N}\sum_{|l|\geq N} cov(\bar{B}_0(\tau_k\cdot), \bar{B}_0(\tau_l\cdot)) \int_Q \nabla t_i(\cdot - k)\nabla t_i(\cdot - l) \\
&\leq \frac{1}{2}\sum_{|k|\geq N}\sum_{|l|\geq N} |cov(\bar{B}_0(\tau_k\cdot), \bar{B}_0(\tau_l\cdot))|(\|\nabla t_i(\cdot - k)\|^2_{L^2(Q)} + \|\nabla t_i(\cdot - l)\|^2_{L^2(Q)}) \\
&\leq \sum_{|k|\geq N}\sum_{|l|\geq N} |cov(\bar{B}_0(\tau_k\cdot), \bar{B}_0(\tau_l\cdot))|\, \|\nabla t_i(\cdot - k)\|^2_{L^2(Q)} \\
&\leq \sum_{|k|\geq N} \left(\|\nabla t_i(\cdot - k)\|^2_{L^2(Q)} \sum_{|l|\geq N} |cov(\bar{B}_0(\tau_k\cdot), \bar{B}_0(\tau_l\cdot))| \right) \\
&\leq \sum_{|k|\geq N} \left(\|\nabla t_i(\cdot - k)\|^2_{L^2(Q)} \sum_{|l|\geq N} |cov(\bar{B}_0, \bar{B}_0(\tau_{l-k}\cdot))| \right)
\end{aligned}$$

$$\leq \sum_{|k|\geq N} \|\nabla t_i(\cdot - k)\|^2_{L^2(Q)} \sum_{k\in\mathbb{Z}^d} |cov(\bar{B}_0, \bar{B}_0(\tau_k\cdot))|.$$

Using (4.31), we obtain

$$\left\| \sum_{|k|\geq N} \left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right) \nabla t_i(\cdot - k) \right\|^2_{L^2(Q\times\Omega)} \leq C \sum_{|k|\geq N} \|\nabla t_i(\cdot - k)\|^2_{L^2(Q)}. \tag{4.38}$$

Since $\nabla t_i \in L^2(\mathbb{R}^d)$, the right-hand side of (4.38) converges to zero when N goes to infinity.

Consequently, (4.37) defines a vector T in $L^2(Q\times\Omega)$. It is clear from (4.37) that $\frac{\partial T_p}{\partial x_n} = \frac{\partial T_n}{\partial x_p}$ for all $(n,p) \in [\![1,d]\!]^2$. Thus T is a gradient, and there exists a function $\tilde{v}_i$ such that

$$\nabla\tilde{v}_i = T + \mathbb{E}(\bar{B}_0)\nabla s_i = \sum_{k\in\mathbb{Z}^d} \left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right) \nabla t_i(x-k) + \mathbb{E}(\bar{B}_0)\nabla s_i. \tag{4.39}$$

Since s_i is $\mathbb{Z}^d$-periodic, we deduce from (4.39) that

$$\nabla\tilde{v}_i \text{ is stationary and } \mathbb{E}\left(\int_Q \nabla\tilde{v}_i\right) = 0. \tag{4.40}$$

We then compute, using (4.33) and (4.34),

$$\begin{aligned}
-\mathrm{div}(A_{per}\nabla\tilde{v}_i) &= \sum_{k\in\mathbb{Z}^d} -\mathrm{div}(A_{per}\nabla t_i(\cdot-k)) \left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right) \\
&\quad -\mathrm{div}(A_{per}\nabla s_i)\mathbb{E}(\bar{B}_0) \\
&= \sum_{k\in\mathbb{Z}^d} \mathrm{div}\left(C_{per}\mathbb{1}_{Q+k}\left(\nabla w_i^0 + e_i\right)\right)\left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right) \\
&\quad +\mathrm{div}\left(C_{per}\left(\nabla w_i^0 + e_i\right)\right)\mathbb{E}(\bar{B}_0) \\
&= \sum_{k\in\mathbb{Z}^d} \mathrm{div}\left(C_{per}\mathbb{1}_{Q+k}\,\bar{B}_0(\tau_k\cdot)\left(\nabla w_i^0 + e_i\right)\right).
\end{aligned} \tag{4.41}$$

Because of (4.35), (4.41) implies

$$-\mathrm{div}(A_{per}\nabla\tilde{v}_i) = \mathrm{div}\left(\bar{b}_0 C_{per}(\nabla w^0 + e_i)\right). \tag{4.42}$$

It follows from (4.40) and (4.42) that $\tilde{v}_i$ solves (4.15). As (4.15) has a solution unique up to the addition of a random constant, we obtain

$$\nabla v_i^0 = \nabla\tilde{v}_i = \sum_{k\in\mathbb{Z}^d} \left(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)\right) \nabla t_i(x-k) + \mathbb{E}(\bar{B}_0)\nabla s_i. \tag{4.43}$$

We deduce from (4.35) and (4.43) that

$$\begin{aligned}
\mathbb{E}(\bar{b}_0\nabla v_i^0) &= \sum_{k\in\mathbb{Z}^d}\sum_{l\in\mathbb{Z}^d} \mathbb{1}_{Q+l}\,\mathbb{E}(\bar{B}_0(\tau_l\cdot)(\bar{B}_0(\tau_k\cdot) - \mathbb{E}(\bar{B}_0)))\nabla t_i(\cdot - k) \\
&\quad +(\mathbb{E}(\bar{B}_0))^2 \sum_{l\in\mathbb{Z}^d} \mathbb{1}_{Q+l}\nabla s_i \\
&= \sum_{k\in\mathbb{Z}^d}\sum_{l\in\mathbb{Z}^d} \mathbb{1}_{Q+l}\,cov(\bar{B}_0(\tau_k\cdot), \bar{B}_0(\tau_l\cdot))\nabla t_i(\cdot - k) + (\mathbb{E}(\bar{B}_0))^2 \sum_{l\in\mathbb{Z}^d} \mathbb{1}_{Q+l}\nabla s_i,
\end{aligned}$$

and then that

$$\begin{aligned}\mathbb{1}_Q\mathbb{E}(\bar{b}_0\nabla v_i^0) =& Var(\bar{B}_0)\nabla t_i + \sum_{k\in\mathbb{Z}^d\setminus\{0\}} cov(\bar{B}_0(\cdot),\bar{B}_0(\tau_k\cdot))\nabla t_i(\cdot-k)\\ &+ (\mathbb{E}(\bar{B}_0))^2\nabla s_i.\end{aligned} \tag{4.44}$$

We conclude by inserting (4.36) and (4.44) in (4.27).

□

Theorem 4.2 (and its two corollaries) are only of interest if $\mathbb{E}(\bar{b}_0) \neq 0$. Indeed, if $\mathbb{E}(\bar{b}_0) = 0$ it only states that $A_\eta^* = A_{per}^* + o(m_\eta)$.

The prototypical case where Theorem 4.2 does not provide valuable information is the case studied in Chapter 3: $b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)B_\eta^k(\omega)$, where the B_η^k are independent identically distributed variables that have Bernoulli law with parameter η, i.e are equal to 1 with probability η and to 0 with probability $1-\eta$. Then, using the notation of Theorem 4.2, $b_\eta^2 = b_\eta$, $m_\eta = \sqrt{\eta}$ and $\bar{b}_0 = 0$, and we only get $A_\eta^* = A_{per}^* + o(\sqrt{\eta})$ (while Section 3.5.1 of the Appendix of Chapter 3 shows that there exists a tensor $\bar{A}_1^*$ such that $A_\eta^* = A_{per}^* + \eta\bar{A}_1^* + o(\eta)$ at least in dimension one). Omitting the dependence on the space variables since b_η is uniform in each cell of $\mathbb{Z}^d$ in this particular setting, a suitable functional space F on Ω to obtain a non trivial weak limit of $\frac{b_\eta}{\|b_\eta\|_F}$ would be $L^1(\Omega)$ for the norm of each B_η^k in $L^1(\Omega)$ is equal to η. The Dunford-Petti weak compactness criterion in that space is however not satisfied by $\frac{b_\eta}{\|b_\eta\|_{L^1(\Omega)}}$. The reason is of course that $\frac{b_\eta}{\|b_\eta\|_{L^1(\Omega)}}$ converges in the set of bounded measures to a Dirac mass. The techniques used in the proof of Theorem 4.2 and its two corollaries thus do not work in this setting.

The above considerations somehow suggest that an alternative viewpoint might be useful. Because of (4.7), the image measure dP_η^x of $b_\eta(x,\cdot)$ converges to a Dirac mass in the sense of distributions. Our alternate approach, related to our work in Chapter 3, consists in working out an expansion of the image measure (or of the law), rather than an expansion of the random variable. As in Chapter 3, our manipulations are mostly formal. Some rigorous foundations, in specific settings, are provided in the Appendix.

4.3 A formal approach

4.3.1 A new assumption on the image measure

For simplicity, we assume as in Corollary 4.4 that b_η is uniform in each cell of $\mathbb{Z}^d$, and is of the form

$$b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x)B_\eta^k(\omega), \tag{4.45}$$

where the B_η^k are independent identically distributed random variables, the distribution of which is given by a "mother variable" B_η. For convenience we slightly modify (4.29)

and require

$$\exists \varepsilon > 0, \forall \eta > 0, \|B_\eta\|_{L^\infty(\Omega)} \leq M - \varepsilon \tag{4.46}$$

$$\|B_\eta\|_{L^2(\Omega)} \underset{\eta \to 0}{\to} 0\,. \tag{4.47}$$

Assumption (4.46) is a technical assumption which implies in particular that for every $\eta > 0$, the image measure dP_η of B_η is a distribution with compact support contained in the *open* set $]-M, M[$. Of course the specific values of M and ε have no particular significance. Throughout the sequel we denote by $\mathcal{E}'(]-M, M[)$ the space of distributions on $\mathbb{R}$ with compact support in $]-M, M[$, and by $\langle T, \varphi \rangle$ the action of a distribution $T \in \mathcal{E}'(]-M, M[)$ on a test function $\varphi \in \mathcal{C}^\infty(]-M, M[)$. Basic elements of distribution theory are recalled in Section 4.5.1 of the Appendix, for convenience of the reader not familiar with technical issues.

Because of assumption (4.47) and Lebesgue dominated convergence theorem, it is clear that for every $\varphi \in \mathcal{C}^\infty(]-M, M[)$,

$$\mathbb{E}(\varphi(B_\eta)) \underset{\eta \to 0^+}{\to} \varphi(0).$$

Since $\mathbb{E}(\varphi(B_\eta)) = \langle dP_\eta, \varphi \rangle$ and $\varphi(0) = \langle \delta_0, \varphi \rangle$ where δ_0 is the Dirac mass at 0, dP_η converges to δ_0 in $\mathcal{E}'(]-M, M[)$.

This leads us to assume that dP_η satisfies

$$dP_\eta = \delta_0 + \eta d\bar{P}_1 + \eta^2 d\bar{P}_2 + o(\eta^2) \quad \text{in } \mathcal{E}'(]-M, M[), \tag{4.48}$$

which is equivalent to

$$\forall \varphi \in \mathcal{C}^\infty(]-M, M[),\ \mathbb{E}(\varphi(B_\eta)) = \langle dP_\eta, \varphi \rangle = \varphi(0) + \eta \langle d\bar{P}_1, \varphi \rangle + \eta^2 \langle d\bar{P}_2, \varphi \rangle + o(\eta^2).$$

Of course $d\bar{P}_1$ and $d\bar{P}_2$ also have a compact support contained in $]-M, M[$: for every test function φ with compact support in $\mathbb{R} \backslash [-M + \varepsilon, M - \varepsilon]$, it holds for all $\eta > 0$

$$\langle dP_\eta, \varphi \rangle = \mathbb{E}(\varphi(B_\eta)) = 0 = \eta \langle d\bar{P}_1, \varphi \rangle + \eta^2 \langle d\bar{P}_2, \varphi \rangle + o(\eta^2),$$

which yields $\langle d\bar{P}_1, \varphi \rangle = \langle d\bar{P}_2, \varphi \rangle = 0$. Then the supports of $d\bar{P}_1$ and $d\bar{P}_2$ are contained in $[-M + \varepsilon, M - \varepsilon] \subset]-M, M[$.

Denoting by $M' = M - \varepsilon/2$, we deduce from Proposition 4.13 of the Appendix that there exists a constant $C > 0$ and integers p_1 and p_2 (namely the orders of $d\bar{P}_1$ and $d\bar{P}_2$ respectively) such that

$$\forall \varphi \in C^\infty(]-M, M[),\ |\langle d\bar{P}_1, \varphi \rangle| \leq C \sup_{s \in [-M', M']} \sup_{0 \leq n \leq p_1} \left| \frac{d^n}{ds^n} \varphi(s) \right|, \tag{4.49}$$

$$\forall \varphi \in C^\infty(]-M, M[),\ |\langle d\bar{P}_2, \varphi \rangle| \leq C \sup_{x \in [-M', M']} \sup_{0 \leq n \leq p_2} \left| \frac{d^n}{ds^n} \varphi(s) \right|. \tag{4.50}$$

Let us now give some additional motivations underlying assumption (4.48).

The first motivation is related to our work presented in Chapter 3 in which B_η has Bernoulli law with parameter η, meaning that it is equal to 1 with probability η and 0 with probability $1-\eta$. Then the image measure dP_η is equal to $\delta_0 + \eta(\delta_1 - \delta_0)$, so that it satisfies (4.48) exactly at order 1 with $d\bar{P}_1 = \delta_1 - \delta_0$.

The second motivation comes from the following result, which shows that there is an easy way, used in our numerical experiments, to build perturbations satisfying (4.48).

Lemma 4.5. *Consider B a random variable in $L^3(\Omega)$. Let K be a positive real, and define $B_\eta = \eta B \mathbb{1}_{|\eta B| \leq K}$. Then B_η, which obviously satisfies (4.46) and (4.47), also satisfies (4.48) with*

$$dP_\eta = \delta_0 - \eta \mathbb{E}(B)\delta_0' + \frac{\eta^2}{2}\mathbb{E}(B^2)\delta_0'' + \mathcal{O}(\eta^3) \text{ in } \mathcal{E}'(\mathbb{R}). \tag{4.51}$$

Proof. Let us denote by dP the image measure of B, and consider $\varphi \in \mathcal{D}(\mathbb{R})$ (i.e $\varphi \in \mathcal{C}^\infty(\mathbb{R})$ and has compact support). Then

$$\begin{aligned} \langle dP_\eta, \varphi \rangle &= \int_{|\eta s| \leq K} \varphi(\eta s) dP + \varphi(0) \int_{|\eta s| \geq K} dP & (4.52) \\ &= \int_{\mathbb{R}} \varphi(\eta s) dP + \int_{|\eta s| \geq K} (\varphi(0) - \varphi(\eta s)) dP. & (4.53) \end{aligned}$$

Since B is in $L^3(\Omega)$,

$$\int_{|\eta s| \geq K} dP = \mathcal{O}(\eta^3),$$

and thus, φ being a bounded function,

$$\langle dP_\eta, \varphi \rangle = \int_{\mathbb{R}} \varphi(\eta s) dP + \mathcal{O}(\eta^3).$$

Then, since $\varphi \in \mathcal{D}(\mathbb{R})$, there exists $C > 0$ such that

$$\forall s \in \mathbb{R}, \quad \left| \frac{\varphi(\eta s) - \varphi(0) - \eta s \varphi'(0) - \frac{\eta^2}{2} s^2 \varphi''(0)}{\eta^2} \right| \leq C\eta |s|^3.$$

Again using $B \in L^3(\Omega)$, this implies that

$$\int_{\mathbb{R}} \left(\frac{\varphi(\eta s) - \varphi(0) - \eta s \varphi'(0) - \frac{\eta^2}{2} s^2 \varphi''(0)}{\eta^2} \right) dP \underset{\eta \to 0}{\to} 0$$

which is just a rewriting of (4.51) since $\int dP = 1$, $\int s dP = \mathbb{E}(B)$ and $\int s^2 dP = \mathbb{E}(B^2)$. □

Before exposing our approach in this new setting, we prove the following elementary result which we will often use in the sequel:

Lemma 4.6. *It holds* $\langle d\bar{P}_1, 1\rangle = 0$ *and* $\langle d\bar{P}_2, 1\rangle = 0$.

Proof. It holds on the one hand $\langle dP_\eta, 1\rangle = 1$ since dP_η is a probability measure, and on the other hand

$$\begin{aligned}\langle dP_\eta, 1\rangle &= \langle \delta_0, 1\rangle + \eta\langle d\bar{P}_1, 1\rangle + \eta^2\langle d\bar{P}_2, 1\rangle + o(\eta^2)\\ &= 1 + \eta\langle d\bar{P}_1, 1\rangle + \eta^2\langle d\bar{P}_2, 1\rangle + o(\eta^2),\end{aligned}$$

so that the conclusion follows.

□

4.3.2 An ergodic approximation of the homogenized tensor

Let us consider a specific realization $\tilde{\omega} \in \Omega$ of A_η in $I_N = [-\frac{N}{2}, \frac{N}{2}]^d$, N being for simplicity an odd integer, and solve the following "supercell" problem:

$$\begin{cases} -\operatorname{div}\left(A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}} + e_i)\right) = 0 \quad \text{in } I_N,\\ w_i^{\eta,N,\tilde{\omega}}\ (N\mathbb{Z})^d - \text{periodic.}\end{cases} \tag{4.54}$$

Then we have

$$\forall i \in [\![1,d]\!],\ \ A_\eta^* e_i = \lim_{N\to+\infty} \frac{1}{N^d}\mathbb{E}\left(\int_{I_N} A_\eta(x,\omega)(\nabla w_i^{\eta,N,\omega}(x) + e_i)\right)dx. \tag{4.55}$$

The proof of (4.55) is given in Chapter 3. We only outline it here for convenience. We know from Theorem 1 in [22] that

$$\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x) + e_i)\,dx \text{ converges to } A_\eta^* e_i \text{ almost surely in } \tilde{\omega}\in\Omega. \tag{4.56}$$

Since $\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x) + e_i)dx$ is the periodic homogenization of $A_\eta(x,\tilde{\omega})$ on I_N, it is also well known that for all $(i,j) \in [\![1,d]\!]^2$,

$$\begin{aligned}\frac{1}{N^d}\left(\int_{I_N} A_\eta^{-1}(x,\tilde{\omega})dx\right)^{-1} e_i\cdot e_j \le \frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x) + e_i)\cdot e_j dx\\ \le \frac{1}{N^d}\left(\int_{I_N} A_\eta(x,\tilde{\omega})dx\right)e_i\cdot e_j,\end{aligned} \tag{4.57}$$

so that for all $N \in 2\mathbb{N}+1$, for all $\eta > 0$ and for almost all $\tilde{\omega}\in\Omega$,

$$\left|\frac{1}{N^d}\int_{I_N} A_\eta(x,\tilde{\omega})(\nabla w_i^{\eta,N,\tilde{\omega}}(x) + e_i)\cdot e_j dx\right| \le \beta, \tag{4.58}$$

where β is defined by (4.9). Using (4.58) and the Lebesgue dominated convergence theorem, we can take the expectation in (4.56) and get (4.55).

Remark 4.3. *The same result holds for homogeneous Dirichlet and Neumann boundary conditions instead of periodic conditions in the definition of* $w_i^{\eta,N,\tilde{\omega}}$ *(see [22] for more details).*

For convenience, we label the unit cells of I_N from 1 to N^d. The k-th cell is denoted by Q_k, for $1 \leq k \leq N^d$. A given realization $A_\eta(x, \tilde{\omega})$ can then be rewritten

$$A_\eta(x, \tilde{\omega}) = A_{per}(x) + \sum_{k=1}^{N^d} \mathbb{1}_{Q_k}(x) s_k C_{per}(x),$$

with $s_k = B_\eta^k(\tilde{\omega})$ for all $k \in [\![1, N^d]\!]$. The $B_\eta^k(\tilde{\omega})$ being independent random variables, the joint probability of the N^d-uplet $(s_1, \cdots, s_{N^d})$ is simply the product $\displaystyle\prod_{k=1}^{N^d} dP_\eta(s_k)$.

Remark 4.4. *The approach exposed in the sequel also works, with minor changes, for random variables which are not independent but correlated with a finite length of correlation. We present it in the independent setting for simplicity.*

We now define $A^{s_1,\cdots,s_{N^d}} = A_{per} + \displaystyle\sum_{k=1}^{N^d} \mathbb{1}_{Q_k} s_k C_{per}$ for $(s_1, \cdots, s_{N^d}) \in [-M, M]^{N^d}$. We denote by $w_i^{s_1,\cdots,s_{N^d}}$ the solution of the i-th cell problem for the periodic homogenization of $A^{s_1,\cdots,s_{N^d}}$ on I_N, that is

$$\begin{cases} -\operatorname{div}\left(A^{s_1,\cdots,s_{N^d}}(\nabla w_i^{s_1,\cdots,s_{N^d}} + e_i)\right) = 0 \quad \text{in } I_N, \\ w_i^{s_1,\cdots,s_{N^d}} \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.59}$$

Then, defining

$$A_\eta^{*,N} e_i = \frac{1}{N^d} \mathbb{E}\left(\int_{I_N} A_\eta(x, \omega)(\nabla w_i^{\eta,N,\omega}(x) + e_i)\right) dx, \tag{4.60}$$

we have

$$A_\eta^{*,N} e_i = \frac{1}{N^d} \int_{\mathbb{R}^{N^d}} \left(\int_{I_N} A^{s_1,\cdots,s_{N^d}}(\nabla w_i^{s_1,\cdots,s_{N^d}} + e_i)\right) \prod_{k=1}^{N^d} dP_\eta(s_k). \tag{4.61}$$

It is proved in Lemma 4.14 of the Appendix that $\nabla w_i^{s_1,\cdots,s_{N^d}}$ is a $\mathcal{C}^\infty$ function of $(s_1, \cdots, s_{N^d})$ in $]-M, M[^{N^d}$. Thus, since $d\bar{P}_1$ and $d\bar{P}_2$ have compact support in $]-M, M[$ (as well as δ_0 of course), we can make these distributions act on $A^{s_1,\cdots,s_{N^d}}$ and $\nabla w_i^{s_1,\cdots,s_{N^d}}$ as functions of $(s_1, \cdots, s_{N^d})$.

It follows from (4.48) that

$$\begin{aligned} \prod_{k=1}^{N^d} dP_\eta(s_k) = &\prod_{k=1}^{N^d} \delta_0(s_k) + \eta \sum_{l=1}^{N^d} d\bar{P}_1(s_l) \prod_{k=1, k\neq l}^{N^d} \delta_0(s_k) \\ &+ \frac{\eta^2}{2} \sum_{l=1}^{N^d} \sum_{m=1}^{N^d} d\bar{P}_1(s_l) d\bar{P}_1(s_m) \prod_{k=1, k\neq\{l,m\}}^{N^d} \delta_0(s_k) \\ &+ \eta^2 \sum_{l=1}^{N^d} d\bar{P}_2(s_l) \prod_{k=1, k\neq l}^{N^d} \delta_0(s_k) + o_N(\eta^2) \ \text{ in } \mathcal{E}'(]-M, M[^{N^d}). \end{aligned} \tag{4.62}$$

We stress that the remainder $o_N(\eta^2)$ in (4.62) depends on N, hence the notation.

Moreover the products (4.62) are to be understood as tensorized products: we work in $\mathcal{E}'(]-M,M[)\otimes_1\mathcal{E}'(]-M,M[)\otimes_2\cdots\otimes_{N^d-1}\mathcal{E}'(]-M,M[)\subset\mathcal{E}'(]-M,M[^{N^d})$.

Inserting (4.62) in (4.61), we obtain the following second-order expansion

$$A_\eta^{*,N}=A_0^{*,N}+\eta A_1^{*,N}+\eta^2A_2^{*,N}+o_N(\eta^2). \tag{4.63}$$

Before making the first three orders in (4.63) precise, note that (4.55), (4.60) and (4.63) imply

$$A_\eta^*=\lim_{N\to\infty}\left(A_0^{*,N}+\eta A_1^{*,N}+\eta^2A_2^{*,N}+o_N(\eta^2)\right) \tag{4.64}$$

In the sequel we exchange in (4.64) the limit in N and the series in η in order to guess a second-order expansion of A_η^* depending only on η. Since we are not able to justify this permutation, our approach is formal.

We now detail the first three orders in (4.63).

First, we notice that for $i\in[\![1,d]\!]$,

$$\begin{aligned}
A_0^{*,N}e_i &= \frac{1}{N^d}\left\langle\prod_{k=1}^{N^d}\delta_0(s_k),\int_{I_N}A^{s_1,\cdots,s_{N^d}}(\nabla w_i^{s_1,\cdots,s_{N^d}}+e_i)\right\rangle\\
&= \frac{1}{N^d}\int_{I_N}A^{0,\cdots,0}(\nabla w_i^{0,\cdots,0}+e_i)\\
&= \frac{1}{N^d}\int_{I_N}A_{per}\left(\nabla w_i^0+e_i\right)\\
&= A_{per}^*e_i,
\end{aligned}$$

which obviously gives the zero-order term expected for A_η^*. Then

$$A_1^{*,N}e_i=\frac{1}{N^d}\sum_{l=1}^{N^d}\left\langle d\bar{P}_1(s_l)\prod_{k=1,k\neq l}^{N^d}\delta_0(s_k),\int_{I_N}A^{s_1,\cdots,s_{N^d}}(\nabla w_i^{s_1,\cdots,s_{N^d}}+e_i)\right\rangle. \tag{4.65}$$

It is easy to see that, by $(N\mathbb{Z})^d$-periodicity of $w_i^{s_1,\cdots,s_{N^d}}$,

$$\left\langle d\bar{P}_1(s_l)\prod_{k=1,k\neq l}^{N^d}\delta_0(s_k),\int_{I_N}A^{s_1,\cdots,s_{N^d}}(\nabla w_i^{s_1,\cdots,s_{N^d}}+e_i)\right\rangle$$

does not depend on l. The expression (4.65) can then be rewritten

$$A_1^{*,N}e_i = \left\langle d\bar{P}_1(s),\int_{I_N}A^{s,0\cdots,0}(\nabla w_i^{s,0,\cdots,0}+e_i)\right\rangle. \tag{4.66}$$

We change the notations for convenience, and define, for $s \in [-M, M]$,

$$A_1^{s,0} = A^{s,0\cdots,0} = A_{per} + s\mathbb{1}_Q C_{per}, \tag{4.67}$$

and $w_i^{1,s,0,N} = w_i^{s,0,\cdots,0}$ solution to

$$\begin{cases} -\operatorname{div}\left(A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i)\right) = 0 \quad \text{in } I_N, \\ w_i^{1,s,0,N}\ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.68}$$

The matrix $A_1^{s,0}$ corresponds to the periodic material with a defect of amplitude s located in Q (i.e at a position $0 \in \mathbb{Z}^d$ in I_N), and $w_i^{1,s,0,N}$ is the i-th cell solution for the periodic homogenization of $A_1^{s,0}$ in I_N. Since $w_i^{1,s,0,N} = w_i^{s,0,\cdots,0}$, it is of course a C^∞ function of $s \in]-M, M[$.

With these notations, we find that

$$A_1^{*,N} e_i = \left\langle d\bar{P}_1(s), \int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \right\rangle. \tag{4.69}$$

For the second-order term, we first define the set

$$\mathcal{T}_N = \left\{ k \in \mathbb{Z}^d, Q + k \subset I_N \right\} = \left[\!\!\left[-\frac{N-1}{2}, \frac{N-1}{2} \right]\!\!\right]^d. \tag{4.70}$$

The cardinal of $\mathcal{T}_N$ is of course N^d, and $\bigcup_{k \in \mathcal{T}_N} \{Q + k\} = I_N$.

For $(s,t) \in [-M, M]^2$ and $k \in \mathcal{T}_N$, we define

$$A_2^{s,t,0,k} = A_{per} + s\mathbb{1}_Q C_{per} + t\mathbb{1}_{Q+k} C_{per}, \tag{4.71}$$

and $w_i^{2,s,t,0,k,N}$ solution to

$$\begin{cases} -\operatorname{div}\left(A_2^{s,t,0,k}(\nabla w_i^{2,s,t,0,k,N} + e_i)\right) = 0 \quad \text{in } I_N, \\ w_i^{2,s,t,0,k,N}\ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.72}$$

The matrix $A_2^{s,t,0,k}$ corresponds to the periodic material with two defects of amplitude s and t located in Q and $Q + k$ (i.e at positions $0 \in \mathbb{Z}^d$ and $k \in \mathbb{Z}^d$ in I_N) respectively. The function $w_i^{2,s,t,0,k,N}$ is the i-th cell solution for the periodic homogenization of $A_2^{s,t,0,k}$ in I_N. It is a $\mathcal{C}^\infty$ function of $(s,t) \in]-M, M[^2$.

Then computations similar to that presented for the first order yield

$$\begin{aligned} A_2^{*,N} e_i = \frac{1}{2} \sum_{k \in \mathcal{T}_N \setminus \{0\}} & \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} A_2^{s,t,0,k}(\nabla w_i^{2,s,t,0,k,N} + e_i) \right\rangle \\ & + \left\langle d\bar{P}_2(s), \int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \right\rangle. \end{aligned} \tag{4.73}$$

A setting with zero, one and two defects is shown in Figure 3.2 of Chapter 3 in the two-dimensional case of a reference material A_{per} consisting of a periodic lattice of circular inclusions.

Remark 4.5. *It is illustrative to consider the particular case where the random variable B_η has a Bernoulli law. This is the case treated in Chapter 3. Then, expansion (4.48) holds exactly with $d\bar{P}_1 = \delta_1 - \delta_0$. The distribution $d\bar{P}_2$ and all other terms of higher order identically vanish. The expressions (4.69) and (4.73) then coincide with (3.28) and (3.29) in Chapter 3.*

In the next section we prove that $A_1^{*,N}$ converges to a finite limit when $N \to \infty$. The case of the second-order term $A_2^{*,N}$, which is also proved to converge, is discussed in Section 4.3.4.

4.3.3 Convergence of the first-order term

We study here the convergence as N goes to infinity of $A_1^{*,N}$ defined by (4.69).

Proposition 4.7. *The sequence $A_1^{*,N}$ converges in $\mathbb{R}^{d\times d}$ to a finite limit $\bar{A}_1^*$ when $N \to \infty$.*

Proof. We fix $(i,j) \in [\![1,d]\!]^2$ and study the convergence of $A_1^{*,N} e_i \cdot e_j$.

Using (4.68) and the adjoint problems defined by (4.20), we first obtain, for all $s \in [-M, M]$,

$$\int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \cdot e_j = \int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0).$$

Then, letting the distribution $d\bar{P}_1$ act on the left and right-hand sides, and using (4.69), we find that

$$A_1^{*,N} e_i \cdot e_j = \left\langle d\bar{P}_1(s), \int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \tag{4.74}$$

Because of the definition of $A_1^{s,0}$,

$$\begin{aligned} \int_{I_N} A_1^{s,0}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) = &\int_{I_N} A_{per}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \\ &+ \int_Q sC_{per}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0). \end{aligned} \tag{4.75}$$

Next, using (4.20),

$$\begin{aligned} \int_{I_N} A_{per}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) &= \int_{I_N} (\nabla w_i^{1,s,0,N} + e_i) \cdot A_{per}^T (e_j + \nabla \tilde{w}_j^0) \\ &= \int_{I_N} e_i \cdot A_{per}^T (e_j + \nabla \tilde{w}_j^0). \end{aligned} \tag{4.76}$$

We know from Lemma 4.6 that $\langle d\bar{P}_1, 1 \rangle = 0$. Thus

$$\left\langle d\bar{P}_1(s), \int_{I_N} e_i \cdot A_{per}^T (e_j + \nabla \tilde{w}_j^0) \right\rangle = 0. \tag{4.77}$$

Collecting (4.74), (4.75), (4.76) and (4.77), we get

$$A_1^{*,N} e_i \cdot e_j = \left\langle d\bar{P}_1(s), \int_Q sC_{per}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \tag{4.78}$$

We now define

$$q_i^{1,s,0,N} = w_i^{1,s,0,N} - w_i^0. \tag{4.79}$$

$q_i^{1,s,0,N}$ solves

$$\begin{cases} -\operatorname{div}\left(A_1^{s,0} \nabla q_i^{1,s,0,N}\right) = \operatorname{div}(s \mathbb{1}_Q C_{per}(\nabla w_i^0 + e_i)) \quad \text{in } I_N, \\ q_i^{1,s,0,N} \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.80}$$

Using (4.79) in (4.78), we rewrite

$$\begin{aligned} A_1^{*,N} e_i \cdot e_j \ &= \ \left\langle s d\bar{P}_1(s), \int_Q C_{per}(\nabla w_i^0 + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ &\quad + \left\langle d\bar{P}_1(s), \int_Q sC_{per}(\nabla q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned} \tag{4.81}$$

The rest of the proof consists in showing that

$$\left\langle d\bar{P}_1(s), \int_Q sC_{per}(\nabla q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle,$$

which is of course equal to

$$\left\langle s d\bar{P}_1(s), \int_Q C_{per}(\nabla q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle,$$

converges to a finite limit when $N \to \infty$.

More precisely, defining

$$\forall s \in [-M, M], \ \forall N \in 2\mathbb{N} + 1, \ \ f^N(s) = \int_Q C_{per}(\nabla q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0),$$

we will prove that the sequence f^N and its derivatives converge uniformly, when N goes to infinity, to a limit function f^∞ and its derivatives.

Applying Lemma 4.15 of the Appendix to (4.80), we obtain that for all $s \in [-M, M]$, $\nabla q_i^{1,s,0,N}$ converges in $L^2(Q)$, when $N \to \infty$, to $\nabla q_i^{1,s,0,\infty}$, where $q_i^{1,s,0,\infty}$ is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A_1^{s,0} \nabla q_i^{1,s,0,\infty}\right) = \operatorname{div}(s \mathbb{1}_Q C_{per}(\nabla w_i^0 + e_i)) \quad \text{in } \mathbb{R}^d, \\ \nabla q_i^{1,s,0,\infty} \in L^2(\mathbb{R}^d). \end{cases} \tag{4.82}$$

Moreover, arguing as in the proof of Lemma 4.15 (given in Chapter 3, see Lemma 3.6), it is easy to see that for all $n \in \mathbb{N}$ and all $s \in [-M, M]$, $\nabla \partial_s^n q_i^{1,s,0,N}$ converges in $L^2(Q)$ to $\nabla \partial_s^n q_i^{1,s,0,\infty}$.

We then define f^∞ by

$$\forall s \in [-M, M], \quad f^\infty(s) = \int_Q C_{per}(\nabla q_i^{1,s,0,\infty} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0).$$

Because of (4.121) and (4.122) in Lemma 4.16 of the Appendix, and using a classical result of differentiation under the integral sign, it is clear that

$$\forall n \in \mathbb{N}, \forall s \in]-M, M[, \ \frac{d^n}{ds^n} f^N(s) = \int_Q C_{per}(\nabla \partial_s^n q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0),$$

and

$$\forall n \in \mathbb{N}, \forall s \in]-M, M[, \ \frac{d^n}{ds^n} f^\infty(s) = \int_Q C_{per}(\nabla \partial_s^n q_i^{1,s,0,\infty} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0).$$

The convergence of $\nabla \partial_s^n q_i^{1,s,0,N}$ to $\nabla \partial_s^n q_i^{1,s,0,\infty}$ in $L^2(Q)$ for every $n \in \mathbb{N}$ thus yields

$$\forall n \in \mathbb{N}, \forall s \in]-M, M[, \lim_{N \to +\infty} \frac{d^n}{ds^n} f^N(s) = \frac{d^n}{ds^n} f^\infty(s). \tag{4.83}$$

On the other hand, we deduce from Lemma 4.17 that there exists a constant $C(p_1, M)$ (recall that p_1 is the order of $d\bar{P}_1(s)$) such that for all $n \in [\![0, p_1]\!]$,

$$\forall (s, s') \in]-M, M[^2, \forall N \in 2\mathbb{N}+1, \ \left|\frac{d^n}{ds^n} f^N(s) - \frac{d^n}{ds^n} f^N(s')\right| \leq C(p_1, M)|s - s'|. \tag{4.84}$$

It is straightforward to see that (4.83) and (4.84) imply that

$$\forall 0 \leq n \leq p_1, \ \frac{d^n}{ds^n} f^N \text{ converges uniformly to } \frac{d^n}{ds^n} f^\infty \text{ in }]-M, M[. \tag{4.85}$$

It follows from (4.49) and (4.85) that

$$\langle s d\bar{P}_1(s), f^N(s) \rangle \to \langle s d\bar{P}_1(s), f^\infty(s) \rangle,$$

and then

$$\begin{aligned} &\left\langle d\bar{P}_1(s), \int_Q s C_{per}(\nabla q_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ &\qquad \underset{N \to \infty}{\to} \left\langle d\bar{P}_1(s), \int_Q s C_{per}(\nabla q_i^{1,s,0,\infty} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned} \tag{4.86}$$

Collecting (4.81) and (4.86), we conclude that $A_1^{*,N}$ converges to a limit tensor $\bar{A}_1^*$ defined by

$$\begin{aligned} \forall (i,j) \in [\![1,d]\!]^2, \quad \bar{A}_1^* e_i \cdot e_j &= \left\langle s d\bar{P}_1(s), \int_Q C_{per}\left(\nabla w_i^0 + e_i\right) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ &\quad + \left\langle d\bar{P}_1(s), \int_Q s C_{per}\left(\nabla q_i^{1,s,0,\infty} + e_i\right) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned} \tag{4.87}$$

□

4.3.4 Convergence of the second-order term

We study here the second-order term of (4.63), namely $A_2^{*,N}$ given by (4.73).

Proposition 4.8. *$A_2^{*,N}$ converges to a finite limit $\bar{A}_2^*$ in $\mathbb{R}^{d\times d}$ when $N \to \infty$.*

Proof. We fix $(i,j) \in [\![1,d]\!]^2$.

In (4.73), the term $\left\langle d\bar{P}_2(s), \int_{I_N} A_1^{s,0}\left(\nabla w_i^{1,s,0,N} + e_i\right)\right\rangle$ is shown to converge to a finite limit exactly as in the proof of convergence of $A_1^{*,N}$ given in Proposition 4.7. Therefore it is sufficient to focus on the term

$$B_2^{*,N} e_i := \frac{1}{2} \sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} A_2^{s,t,0,k}(\nabla w_i^{2,s,t,0,k,N} + e_i) \right\rangle. \tag{4.88}$$

In order to show that $B_2^{*,N}$ and then $A_2^{*,N}$ converge, we proceed in four steps.

Step 1. Rewriting of $B_2^{*,N}$

Using the adjoint problems, we obtain, exactly as in the proof of (4.78) in Proposition 4.7:

$$\begin{aligned} &B_2^{*,N} e_i \cdot e_j = \\ &\frac{1}{2} \sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per}\left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right)\left(\nabla w_i^{2,s,t,0,k,N} + e_i\right)\cdot\left(\nabla \tilde{w}_j^0 + e_j\right)\right\rangle. \end{aligned} \tag{4.89}$$

Let us define

$$q_i^{2,s,t,0,k,N} = w_i^{2,s,t,0,k,N} - w_i^{1,s,0,N} - w_i^{1,t,k,N} + w_i^0, \tag{4.90}$$

where $w_i^{1,t,k,N} = w_i^{1,t,0,N}(\cdot - k)$.

The role of $q_i^{2,s,t,0,k,N}$ is to compare the cell solution $w_i^{2,s,t,0,k,N}$ with two defects located at 0 and at $k \in \mathcal{T}_N$ to the sum of the one-defect solutions $w_i^{1,s,0,N}$ and $w_i^{1,t,k,N}$ minus the periodic background w_i^0. We expect this difference to decay sufficiently fast far from the defects.

Note that $q_i^{2,s,t,0,k,N}$ solves

$$\begin{cases} -\operatorname{div}\left(A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\right) = \operatorname{div}(s\mathbb{1}_Q C_{per}\nabla q_i^{1,t,k,N}) \\ \qquad\qquad + \operatorname{div}(t\mathbb{1}_{Q+k} C_{per}\nabla q_i^{1,s,0,N}) \quad \text{in } I_N, \\ q_i^{2,s,t,0,k,N} \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.91}$$

It follows from (4.79) and (4.90) that

$$w_i^{2,s,t,0,k,N} = w_i^0 + q_i^{1,s,0,N} + q_i^{1,t,k,N} + q_i^{2,s,t,0,k,N}, \tag{4.92}$$

where $q_i^{1,t,k,N} = q_i^{1,t,0,N}(\cdot - k)$. Inserting (4.92) in (4.89), we obtain

$$\begin{aligned} &B_2^{*,N} e_i \cdot e_j = \\ &\frac{1}{2} \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per} \left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right) (\nabla w_i^0 + e_i) \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle \\ &+ \frac{1}{2} \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per} \left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right) (\nabla q_i^{1,s,0,N} + \nabla q_i^{1,t,k,N}) \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle \\ &+ \frac{1}{2} \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per} \left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right) \nabla q_i^{2,s,t,0,k,N} \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle. \end{aligned} \tag{4.93}$$

We know from Lemma 4.6 that $\langle d\bar{P}_1, 1\rangle = 0$. This implies that

$$\left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per} \left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right) (\nabla w_i^0 + e_i) \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle = 0. \tag{4.94}$$

Still using $\langle d\bar{P}_1, 1\rangle = 0$, an easy computation yields

$$\begin{aligned} &\left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_{I_N} C_{per} \left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right) (\nabla q_i^{1,s,0,N} + \nabla q_i^{1,t,k,N}) \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle \\ &\qquad = \langle s d\bar{P}_1(s), 1\rangle \left\langle d\bar{P}_1(s), \int_Q C_{per} \nabla q_i^{1,s,k,N} \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle \\ &\qquad\quad + \langle s d\bar{P}_1(s), 1\rangle \left\langle d\bar{P}_1(s), \int_{Q+k} C_{per} \nabla q_i^{1,s,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j) \right\rangle. \end{aligned} \tag{4.95}$$

Since $q_i^{1,s,k,N} = q_i^{1,s,0,N}(\cdot - k)$, and $\tilde{w}_j^0$ is $\mathbb{Z}^d$-periodic, we have

$$\int_Q C_{per} \nabla q_i^{1,s,k,N} \cdot (\nabla \tilde{w}_j^0 + e_j) = \int_{Q-k} C_{per} \nabla q_i^{1,s,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j). \tag{4.96}$$

By definition of $\mathcal{T}_N$, we know that

$$\bigcup_{k \in \mathcal{T}_N} \{Q + k\} = \bigcup_{k \in \mathcal{T}_N} \{Q - k\} = I_N,$$

so that

$$\sum_{k \in \mathcal{T}_N \setminus \{0\}} \int_{Q+k} C_{per} \nabla q_i^{1,s,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j) = \sum_{k \in \mathcal{T}_N \setminus \{0\}} \int_{Q-k} C_{per} \nabla q_i^{1,s,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j). \tag{4.97}$$

Substituting (4.96) in (4.97), we have

$$\sum_{k \in \mathcal{T}_N \setminus \{0\}} \int_{Q+k} C_{per} \nabla q_i^{1,s,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j) = \sum_{k \in \mathcal{T}_N \setminus \{0\}} \int_{Q} C_{per} \nabla q_i^{1,s,k,N} \cdot (\nabla \tilde{w}_j^0 + e_j). \tag{4.98}$$

Using (4.94), (4.95) and (4.98) in (4.93), we finally obtain the more convenient expression:

$$\begin{aligned} B_2^{*,N} e_i \cdot e_j = \langle s d\bar{P}_1(s), 1\rangle \sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle \\ + \frac{1}{2}\sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} (s\mathbb{1}_Q C_{per} + t\mathbb{1}_{Q+k}C_{per})\,\nabla q_i^{2,s,t,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle. \end{aligned} \tag{4.99}$$

Defining

$$D_N = \sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle, \tag{4.100}$$

and, for all $k\in\mathcal{T}_N\setminus\{0\}$,

$$E_N^k = \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} (s\mathbb{1}_Q C_{per} + t\mathbb{1}_{Q+k}C_{per})\,\nabla q_i^{2,s,t,0,k,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle, \tag{4.101}$$

we have

$$B_2^{*,N} e_i \cdot e_j = \langle s d\bar{P}_1(s), 1\rangle\, D_N + \frac{1}{2}\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k. \tag{4.102}$$

Intuitively, D_N is a "one defect" term and E_N^k a "two defects" term. In the next two steps we are going to prove that D_N and $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$ converge to finite limits as $N\to\infty$.

Step 2. Convergence of D_N

It readily follows from (4.100) that

$$\begin{aligned} D_N = \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle \\ - \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,0,N}\cdot(\nabla\tilde{w}_j^0+e_j)\right\rangle, \end{aligned}$$

where we have denoted by $q_i^{1,s,N} = \sum_{k\in\mathcal{T}_N} q_i^{1,s,k,N}$.

As detailed in Chapter 3, since $q_i^{1,s,k,N}$ is obtained by a k-shift of $q_i^{1,s,0,N}$, we notice that $q_i^{1,s,N}$ is a $\mathbb{Z}^d$-periodic function, unique solution up to the addition of a constant to

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla q_i^{1,s,N}\right) = \operatorname{div}(sC_{per}(\nabla w_i^0+e_i)) \\ \qquad\qquad + \operatorname{div}\left(s\mathbb{1}_Q C_{per}\nabla q_i^{1,s,0,N}\right) \quad \text{in } Q, \\ q_i^{1,s,N}\ \mathbb{Z}^d-\text{periodic}. \end{cases} \tag{4.103}$$

We know from Lemma 4.15 that for every $s \in [-M, M]$, $\nabla q_i^{1,s,0,N}$ converges in $L^2(Q)$ to $\nabla q_i^{1,s,0,\infty}$ defined by (4.82). This readily implies that for every $s \in [-M, M]$, $\nabla q_i^{1,s,N}$ converges in $L^2(Q)$ to $\nabla q_i^{1,s,\infty}$, where $q_i^{1,s,\infty}$ solves

$$\begin{cases} -\operatorname{div}\left(A_{per}\nabla q_i^{1,s,\infty}\right) = \operatorname{div}(sC_{per}(\nabla w_i^0 + e_i)) \\ \qquad\qquad + \operatorname{div}\left(s\mathbb{1}_Q C_{per}\nabla q_i^{1,s,0,\infty}\right) \quad \text{in } Q, \\ q_i^{1,s,\infty}\ \mathbb{Z}^d - \text{periodic.} \end{cases} \tag{4.104}$$

Arguing exactly as in the proof of Proposition 4.7, we finally find that

$$D_N \underset{N\to\infty}{\to} \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,\infty}\cdot(\nabla\tilde{w}_j^0 + e_j)\right\rangle \\ - \left\langle d\bar{P}_1(s), \int_Q C_{per}\nabla q_i^{1,s,0,\infty}\cdot(\nabla\tilde{w}_j^0 + e_j)\right\rangle.$$

Step 3. Convergence of $\displaystyle\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$

We first define the adjoint problems to (4.80) and (4.91) respectively by

$$\begin{cases} -\operatorname{div}\left((A_1^{s,0})^T\nabla\tilde{q}_j^{1,s,0,N}\right) = \operatorname{div}(s\mathbb{1}_Q C_{per}^T(\nabla\tilde{w}_j^0 + e_i)) \quad \text{in } I_N, \\ \tilde{q}_j^{1,s,0,N}\ (N\mathbb{Z})^d - \text{periodic,} \end{cases} \tag{4.105}$$

$$\begin{cases} -\operatorname{div}\left((A_2^{s,t,0,k})^T\nabla\tilde{q}_j^{2,s,t,0,k,N}\right) = \operatorname{div}(s\mathbb{1}_Q C_{per}^T\nabla\tilde{q}_j^{1,t,k,N}) \\ \qquad\qquad + \operatorname{div}(t\mathbb{1}_{Q+k} C_{per}^T\nabla\tilde{q}_j^{1,s,0,N}) \quad \text{in } I_N, \\ \tilde{q}_j^{2,s,t,0,k,N}\ (N\mathbb{Z})^d - \text{periodic,} \end{cases} \tag{4.106}$$

with $\tilde{q}_j^{1,t,k,N} = \tilde{q}_j^{1,t,0,N}(\cdot - k)$. In what follows we also use the notation $A_1^{t,k} := A_1^{t,0}(\cdot - k)$.

Now, using integration by parts, (4.105) and the definition of $A_2^{s,t,0,k}$, we compute

$$\begin{aligned} &\int_{I_N} C_{per}\left(s\mathbb{1}_Q + t\mathbb{1}_{Q+k}\right)\nabla q_i^{2,s,t,0,k,N}\cdot(\nabla\tilde{w}_j^0 + e_j) \\ &= -\int_{I_N} q_i^{2,s,t,0,k,N}\operatorname{div}\left(C_{per}^T\left(s\mathbb{1}_Q(\nabla\tilde{w}_j^0 + e_j) + t\mathbb{1}_{Q+k}(\nabla\tilde{w}_j^0 + e_j)\right)\right) \\ &= -\int_{I_N} q_i^{2,s,t,0,k,N}\operatorname{div}\left((A_1^{s,0})^T\nabla\tilde{q}_j^{1,s,0,N} + (A_1^{t,k})^T\nabla\tilde{q}_j^{1,t,k,N}\right) \\ &= \int_{I_N}\left(A_1^{s,0}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{1,s,0,N} + A_1^{t,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{1,t,k,N}\right) \\ &= \int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\left(\nabla\tilde{q}_j^{1,s,0,N} + \nabla\tilde{q}_j^{1,t,k,N}\right) \\ &\quad - \int_{I_N}\nabla q_i^{2,s,t,0,k,N}\cdot C_{per}^T\left(t\mathbb{1}_{Q+k}\nabla\tilde{q}_j^{1,s,0,N} + s\mathbb{1}_Q\nabla\tilde{q}_j^{1,t,k,N}\right). \end{aligned}$$

Using then (4.91), (4.106) and integration by parts, we obtain

$$\begin{aligned}
&\int_{I_N} C_{per}\,(s\mathbb{1}_Q + t\mathbb{1}_{Q+k})\,\nabla q_i^{2,s,t,0,k,N}\cdot(\nabla\tilde{w}_j^0 + e_j)\\
&= -\int_{I_N} C_{per}\left(s\mathbb{1}_Q\nabla q_i^{1,t,k,N} + t\mathbb{1}_{Q+k}\nabla q_i^{1,s,0,N}\right)\cdot\left(\nabla\tilde{q}_j^{1,s,0,N} + \nabla\tilde{q}_j^{1,t,k,N}\right)\\
&\quad + \int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,N}.
\end{aligned}$$

Thus E_N^k defined by (4.101) can be rewritten

$$\begin{aligned}
E_N^k &= -\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} C_{per}\left(s\mathbb{1}_Q\nabla q_i^{1,t,k,N} + t\mathbb{1}_{Q+k}\nabla q_i^{1,s,0,N}\right)\cdot\left(\nabla\tilde{q}_j^{1,s,0,N} + \nabla\tilde{q}_j^{1,t,k,N}\right)\right\rangle\\
&\quad + \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,N}\right\rangle. \qquad (4.107)
\end{aligned}$$

For convenience we part E_N^k into $E_N^{1,k} + E_N^{2,k}$ where

$$E_N^{1,k} = -\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} C_{per}\left(s\mathbb{1}_Q\nabla q_i^{1,t,k,N} + t\mathbb{1}_{Q+k}\nabla q_i^{1,s,0,N}\right)\cdot\left(\nabla\tilde{q}_j^{1,s,0,N} + \nabla\tilde{q}_j^{1,t,k,N}\right)\right\rangle, \qquad (4.108)$$

and

$$E_N^{2,k} = \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,N}\right\rangle. \qquad (4.109)$$

We are going to prove that the series $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$ and $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ are converging when N goes to infinity.

Step 3.1. Convergence of $\displaystyle\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$

Using the fact that $q_i^{1,t,k,N} = q_i^{1,t,0,N}(\cdot - k)$ and $\tilde{q}_j^{1,t,k,N} = \tilde{q}_j^{1,t,0,N}(\cdot - k)$ in (4.108), a simple calculation shows that

$$\begin{aligned}
\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k} &= -2\sum_{k\in\mathcal{T}_N\setminus\{0\}}\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q C_{per}s\nabla q_i^{1,t,k,N}\cdot\nabla\tilde{q}_j^{1,s,0,N}\right\rangle\\
&\quad -2\langle td\bar{P}_1(t), 1\rangle\sum_{k\in\mathcal{T}_N\setminus\{0\}}\left\langle d\bar{P}_1(s), \int_{Q+k} C_{per}s\nabla q_i^{1,s,0,N}\cdot\nabla\tilde{q}_j^{1,s,0,N}\right\rangle.
\end{aligned}$$

Then, recalling that $q_i^{1,t,N} = \sum_{k\in\mathcal{T}_N} q_i^{1,t,k,N}$, it comes

$$\begin{aligned}
\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k} &= -2\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q C_{per}s\nabla q_i^{1,t,N}\cdot\nabla\tilde{q}_j^{1,s,0,N}\right\rangle\\
&\quad + 2\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q C_{per}s\nabla q_i^{1,t,0,N}\cdot\nabla\tilde{q}_j^{1,s,0,N}\right\rangle \qquad (4.110)\\
&\quad - 2\langle td\bar{P}_1(t), 1\rangle\left\langle d\bar{P}_1(s), \int_{I_N\setminus Q} C_{per}s\nabla q_i^{1,s,0,N}\cdot\nabla\tilde{q}_j^{1,s,0,N}\right\rangle.
\end{aligned}$$

The proof of Proposition 4.7 is then easily adapted to show that

$$\begin{aligned}\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k} \underset{N\to\infty}{\to} & -2\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q C_{per}s\nabla q_i^{1,t,\infty}\cdot\nabla\tilde{q}_j^{1,s,0,\infty}\right\rangle \\ & +2\left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q C_{per}s\nabla q_i^{1,t,0,\infty}\cdot\nabla\tilde{q}_j^{1,s,0,\infty}\right\rangle \\ & -2\langle td\bar{P}_1(t),1\rangle\left\langle d\bar{P}_1(s), \int_{\mathbb{R}^d\setminus Q} C_{per}s\nabla q_i^{1,s,0,\infty}\cdot\nabla\tilde{q}_j^{1,s,0,\infty}\right\rangle.\end{aligned} \tag{4.111}$$

Step 3.2. Convergence of $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$

We do not give here the detailed proof of convergence for it consists of long technical computations. The core of the proof is essentially the same as that of Proposition 3.4 in Chapter 3. We content ourselves with presenting the main ingredients.

Following the proof of Lemma 4.15 (given in Chapter 3, see Lemma 3.6) applied to (4.91) and (4.106), we first note that for all $(s,t)\in[-M,M]^2$, $\mathbb{1}_{I_N}\nabla q_i^{2,s,t,0,k,N}$ and $\mathbb{1}_{I_N}\nabla\tilde{q}_j^{2,s,t,0,k,N}$ converge in $L^2(\mathbb{R}^d)$ to $\nabla q_i^{2,s,t,0,k,\infty}$ and $\nabla\tilde{q}_j^{2,s,t,0,k,\infty}$ respectively, where $q_i^{2,s,t,0,k,\infty}$ is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,\infty}\right) = \operatorname{div}(s\mathbb{1}_Q C_{per}\nabla q_i^{1,t,k,\infty}) \\ \qquad\qquad + \operatorname{div}(t\mathbb{1}_{Q+k}C_{per}\nabla q_i^{1,s,0,\infty}) \quad \text{in } \mathbb{R}^d, \\ \nabla q_i^{2,s,t,0,k,\infty} \in L^2(\mathbb{R}^d), \end{cases} \tag{4.112}$$

and $\tilde{q}_j^{2,s,t,0,k,\infty}$ solves the adjoint problem to (4.112).

This implies that for all $(s,t)\in[-M,M]^2$,

$$\int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,N} \underset{N\to\infty}{\to} \int_{\mathbb{R}^d} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,\infty}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,\infty}. \tag{4.113}$$

Convergence (4.113) can be differentiated indefinitely with respect to s and t. Arguing as in the proof of Proposition 4.7 (which requires first to adapt Lemma 4.17 to address the "two defects" setting), we actually prove that all these convergences are uniform in $[-M,M]^2$.

It then follows from (4.49) that

$$\begin{aligned} E_N^{2,k} &= \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{I_N} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,N}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,N}\right\rangle \\ &\underset{N\to\infty}{\to} E_\infty^{2,k} := \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_{\mathbb{R}^d} A_2^{s,t,0,k}\nabla q_i^{2,s,t,0,k,\infty}\cdot\nabla\tilde{q}_j^{2,s,t,0,k,\infty}\right\rangle. \end{aligned} \tag{4.114}$$

Using (4.131) in Lemma 4.18, we easily prove that the series $\sum_{k\in\mathbb{Z}^d\setminus\{0\}} E_\infty^{2,k}$ is absolutely converging.

We finally conclude, as in the proof of Proposition 4.8 in Chapter 3, and using (4.128) in Lemma 4.18, that

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_\infty^{2,k} \underset{N\to\infty}{\to} \sum_{k\in\mathbb{Z}^d\setminus\{0\}} E_\infty^{2,k}.$$

We have thus shown in Steps 3.1 and 3.2 that $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{1,k}$ and $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^{2,k}$ converge when N goes to infinity. Since $E_N^k = E_N^{1,k} + E_N^{2,k}$ we deduce that $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$ converges.

Step 4. Conclusion

We have shown in the previous steps that the sequence D_N and the series $\sum_{k\in\mathcal{T}_N\setminus\{0\}} E_N^k$ converge when $N\to\infty$. Using (4.102), this implies that $B_2^{*,N}$ converges in $\mathbb{R}^{d\times d}$ and then that $A_2^{*,N}$ converges in $\mathbb{R}^{d\times d}$.

□

4.4 Numerical experiments

The purpose of this section is to assess the numerical relevance of the approaches of Sections 4.2 and 4.3. To this end we build and homogenize stochastic composite materials using laws that satisfy the assumptions of these sections. Our motivations are not strictly identical for the two approaches. In contrast to the first approach which relies on a rigorous proof, our second approach is formal and we thus need to demonstrate its correctness experimentally (note that the tests performed in Chapter 3 in the Bernoulli case are already to be considered as a component of the validation of the approach). We wish to check that the expansions derived in Sections 4.2 and 4.3 provide an accurate and efficient approximation to the direct stochastic computation. Due to the prohibitive cost of three-dimensional random homogenization problems, we restrict ourselves to the two-dimensional setting. We first explain our general methodology, which is the same as that presented in Chapter 3, and then make precise the specific settings.

4.4.1 Methodology

We mainly consider as in Chapter 3 a reference material A_{per} that consists of a constant background reinforced by a periodic lattice of circular inclusions, that is

$$A_{per}(x_1,x_2) = 20\times Id + 100\sum_{k\in\mathbb{Z}^2} \mathbb{1}_{B(k,0.3)}(x_1,x_2)\times Id,$$

where $B(k,0.3)$ is the ball of center k and radius 1. Loosely speaking, the role of the perturbation is to randomly eliminate some fibers:

$$C_{per}(x_1,x_2) = -100\sum_{k\in\mathbb{Z}^2} \mathbb{1}_{B(k,0.3)}(x_1,x_2)\times Id.$$

We will also, in our last test, consider a laminate

$$A_{per}(x_1, x_2) = 5 + 10 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_1 \leq l+1}(x_1, x_2),$$

with the perturbation yielding an error in the lamination direction:

$$C_{per}(x_1, x_2) = 10 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_2 \leq l+1}(x_1, x_2) \times Id - 10 \sum_{l \in \mathbb{Z}} \mathbb{1}_{l \leq x_1 \leq l+1}(x_1, x_2) \times Id.$$

For both materials (shown in Figure 3.3 of Chapter 3), we have chosen the values of the coefficients in order to have a high contrast between A_{per} and $A_{per} + C_{per}$ and thus for the perturbation to have an important impact on the microscopic structure. The specific values of these coefficients has no other significance.

We will consider different perturbations b_η, all of them satisfying (4.45) with the B_η^k independent and identically distributed.

Our goal is to compare A_η^* with its approximation $A_{per}^* + \eta A_1^{*,N} + \eta^2 A_2^{*,N}$. A major computational difficulty is the computation of the "exact" matrix A_η^* given by formula (4.11). It ideally requires to solve the stochastic cell problems (4.10) on $\mathbb{R}^d$. To this end we first use ergodicity and formula (4.55), and actually compute, for a given realization ω and a domain I_N chosen here to be $[0, N]^2$ for convenience, $A_\eta^{*,N}(\omega)$ defined by

$$A_\eta^{*,N}(\omega) e_i = \frac{1}{N^d} \int_{I_N} A_\eta(x, \omega)(\nabla w_i^{\eta,N,\omega}(x) + e_i) dx. \tag{4.115}$$

In a second step, we take averages over the realizations ω.

For each ω, we use the finite element software FreeFem++ (available at www.freefem.org) to solve the boundary value problems (4.54) and compute the integrals (4.115). We work with standard P1 finite elements on a triangular mesh such that there are 10 degrees of freedom on each edge of the unit cell Q.

We define an approximate value $A_\eta^{*,N}$ as the average of $A_\eta^{*,N}(\omega)$ over 40 realizations ω. Our numerical experiments indeed show that the number 40 is sufficiently large for the convergence of the Monte-Carlo computation. We then let N grow from 5 to 80 by steps of 5. We observe that $A_\eta^{*,N}$ stabilizes at a fixed value around $N = 80$ and thus take $A_\eta^{*,80}$ as the reference value for A_η^* in our subsequent tests.

The next step is to compute the zero-order term A_{per}^*, and the first-order and second-order deterministic corrections. Using the same mesh and finite elements as for our reference computation above, we compute A_{per}^* using (4.12) and (4.13). The computation of the next orders depends on the setting:

- in the setting of Section 4.2, the first-order correction is given by (4.17) in Theorem 4.2 and is thus independent of N; since b_η is of the form (4.28), we use formula (4.32) in Corollary 4.4 for the second-order correction which depends on N through the term t_i defined on $\mathbb{R}^d$ by (4.33), and which has to be approximated on I_N; we let N grow from 5 to 80 by steps of 5;

- in the setting of Section 4.3, the corrections $A_1^{*,N}$ and $A_2^{*,N}$ are respectively given by (4.69) and (4.73); we let N grow from 5 to 80 by steps of 5 for $A_1^{*,N}$; the computation of $A_2^{*,N}$ being far more expensive (there is not only an integral over I_N but also a sum over the N^2 cells in (4.73)), we have to limit ourselves to $N = 25$ and approximate the value for N larger than 25 by the value obtained for $N = 25$.

We stress that there are three distinct sources of error in these computations:

- the finite element discretization error;
- the truncation error due to the replacement of $\mathbb{R}^d$ with I_N, in the computation of the stochastic cell problems (4.10) that are replaced with (4.54), as well as in the computation of the integrals (4.115);
- the stochastic error arising from the approximation of the expectation value by an empirical mean.

Detailed comments on these various errors and the way we deal with them are provided in Chapter 3. We just emphasize, in the setting of Section 4.3, that it is not our purpose to prove through our tests that

$$A_\eta^* = A_{per}^* + \eta \bar{A}_1^* + \eta^2 \bar{A}_2^* + o(\eta^2)$$

with a $o(\eta^2)$ which would be independent of N, of the number of realizations and of the size of the mesh. We only wish to demonstrate that the second-order expansion is an approximation to A_η^* sufficiently good for all practical purposes. We will observe that both $A_1^{*,N}$ and $A_2^{*,N}$ converge to their respective limits faster than $A_\eta^{*,N}$ to A_η^* (which is expected since the former quantities are deterministic and contain less information). We will also observe that $A_{per}^* + \eta A_1^{*,N}$ is closer to A_η^* than A_{per}^* and that the inclusion of the second order improves the situation for $A_{per}^* + \eta A_1^{*,N} + \eta^2 A_2^{*,N}$ is even closer.

To present our numerical results, we choose the first diagonal entry $(1,1)$ of all the matrices considered. Other coefficients in the matrices behave qualitatively similarly. We illustrate a practical interval of confidence for our Monte-Carlo computation of A_η^* by showing, for each N, the minimum and maximum values of $A_\eta^{*,N}(\omega)$ achieved over the 40 realizations ω.

We will use the following caption in the graphs:

- *periodic*: gives the value of the periodic homogenized tensor A_{per}^*;
- *first-order*: gives the value of the first-order expansion;
- *second-order*: gives the value of the second-order expansion;
- *stochastic mean, minima and maxima*: respectively give the values of $A_\eta^{*,N}$ and the extrema obtained in the computation of the empirical mean.

Finally, the results are given for various values of η which serve the purpose of testing our approach in a diversity of situations, and in particular for perturbations that are "not so small".

4.4.2 An example of setting for our theory in Section 4.2 (and 4.3)

Consider $B_\eta = \eta\, G\, \mathbb{1}_{0\leq \eta G\leq 1}$ where G is a normalized centered Gaussian random variable. It is easy to check that

$$B_\eta = \eta G \mathbb{1}_{0\leq G\leq +\infty} + o(\eta^2) \quad \text{in } L^2(\Omega),$$

so that Corollary 4.4 of Section 4.2 applies. Alternatively, we can use Lemma 4.5, which gives

$$dP_\eta \;=\; \delta_0 - \eta\frac{1}{\sqrt{2}}\delta_0' + \frac{\eta^2}{4}\delta_0'' + o(\eta^2) \text{ in } \mathcal{E}'(\mathbb{R}),$$

to perform our formal approach. We verify in Section 4.5.4 of the Appendix that both approaches yield the same results up to second order.

We show results for the lattice of inclusions and for $\eta = 0.1$ and $\eta = 0.2$ (Figures 4.1 and 4.2 respectively).

The results are very satisfying for both values of η. The first-order correction, which does not depend on N according to Theorem 4.2, enables to get substantially closer to A^*_η. Moreover, it is clear (especially from the close-ups) that the second-order correction $A_2^{*,N}$ converges very fast (convergence is already reached at $N = 5$), and in particular much faster than the stochastic computation $A_\eta^{*,N}$. It also provides excellent accuracy.

4.4.3 A first example of setting for our formal approach of Section 4.3

Consider R_η a random variable having Bernoulli law with parameter η, and G a normalized centered Gaussian random variable independent of R_η. We define the product random variable $B_\eta = R_\eta \times \eta G \mathbb{1}_{|\eta G|\leq 1}$. Then

$$\begin{aligned}
\mathbb{E}(\varphi(B_\eta)) &= \mathbb{E}(\varphi(R_\eta \times \eta G \mathbb{1}_{|\eta G|\leq 1})) \\
&= \eta \mathbb{E}(\varphi(\eta G \mathbb{1}_{|\eta G|\leq 1})) + (1-\eta)\varphi(0) \\
&= \eta(\varphi(0) + \eta\mathbb{E}(G)\varphi'(0) + \frac{\eta^2}{2}\varphi''(0) + o(\eta^2)) + (1-\eta)\varphi(0) \\
&= \varphi(0) + \frac{\eta^3}{2}\varphi''(0) + o(\eta^3).
\end{aligned}$$

This implies

$$dP_\eta = \delta_0 + \frac{\eta^3}{2}\delta_0'' + o(\eta^3) \quad \text{in } \mathcal{E}'(\mathbb{R}). \tag{4.116}$$

In this case we only consider the first-order correction since the dominant order in (4.116) is already tiny. We present the results in the case of the lattice of inclusions, for $\eta = 0.2$, $\eta = 0.3$ and $\eta = 0.5$ (Figures 4.3, 4.4, 4.5 respectively).

Once again, our approach converges rapidly and allows for an accurate approximate value of A^*_η even for η as large as 0.5.

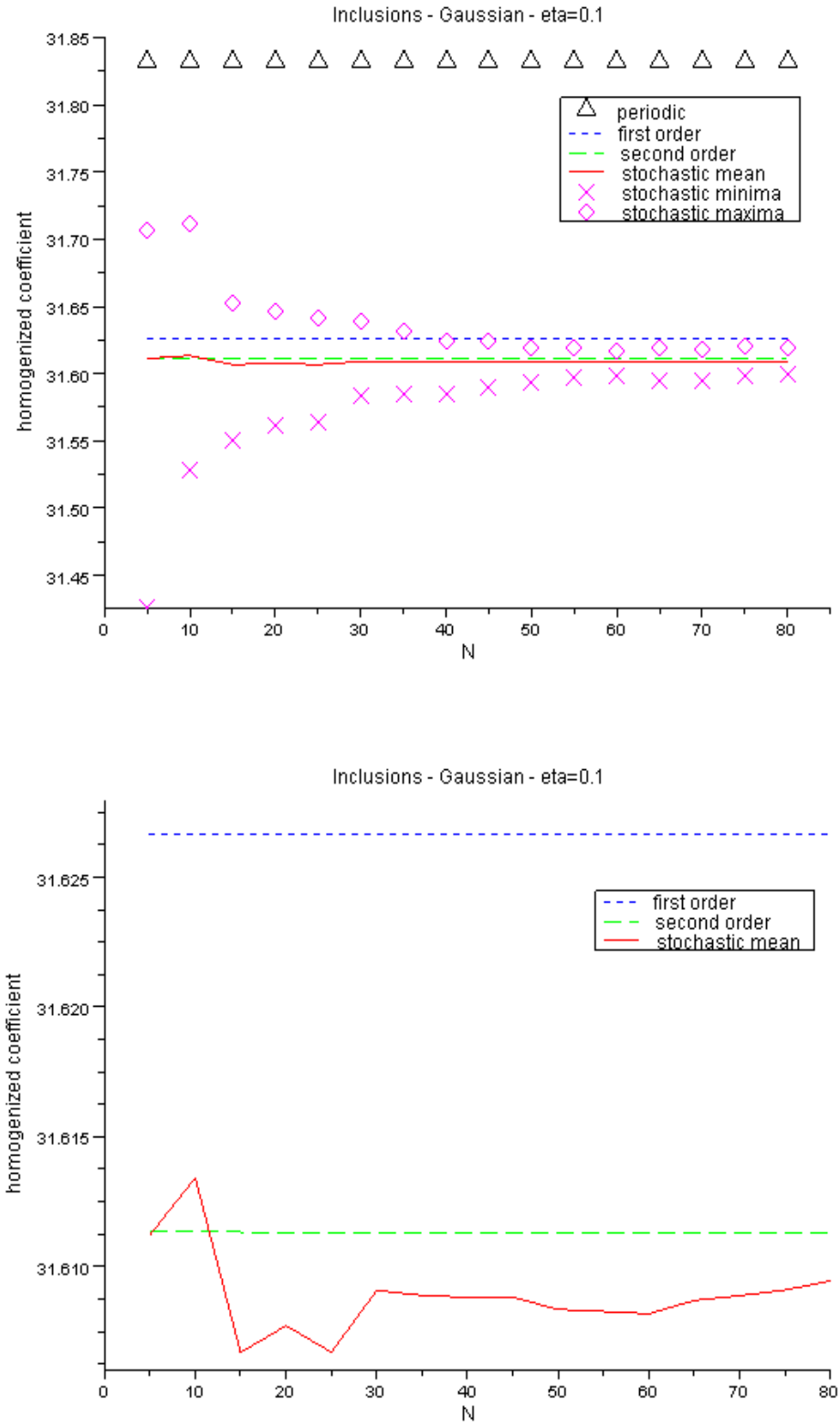

Figure 4.1: Inclusions - results for a Gaussian perturbation and $\eta = 0.1$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

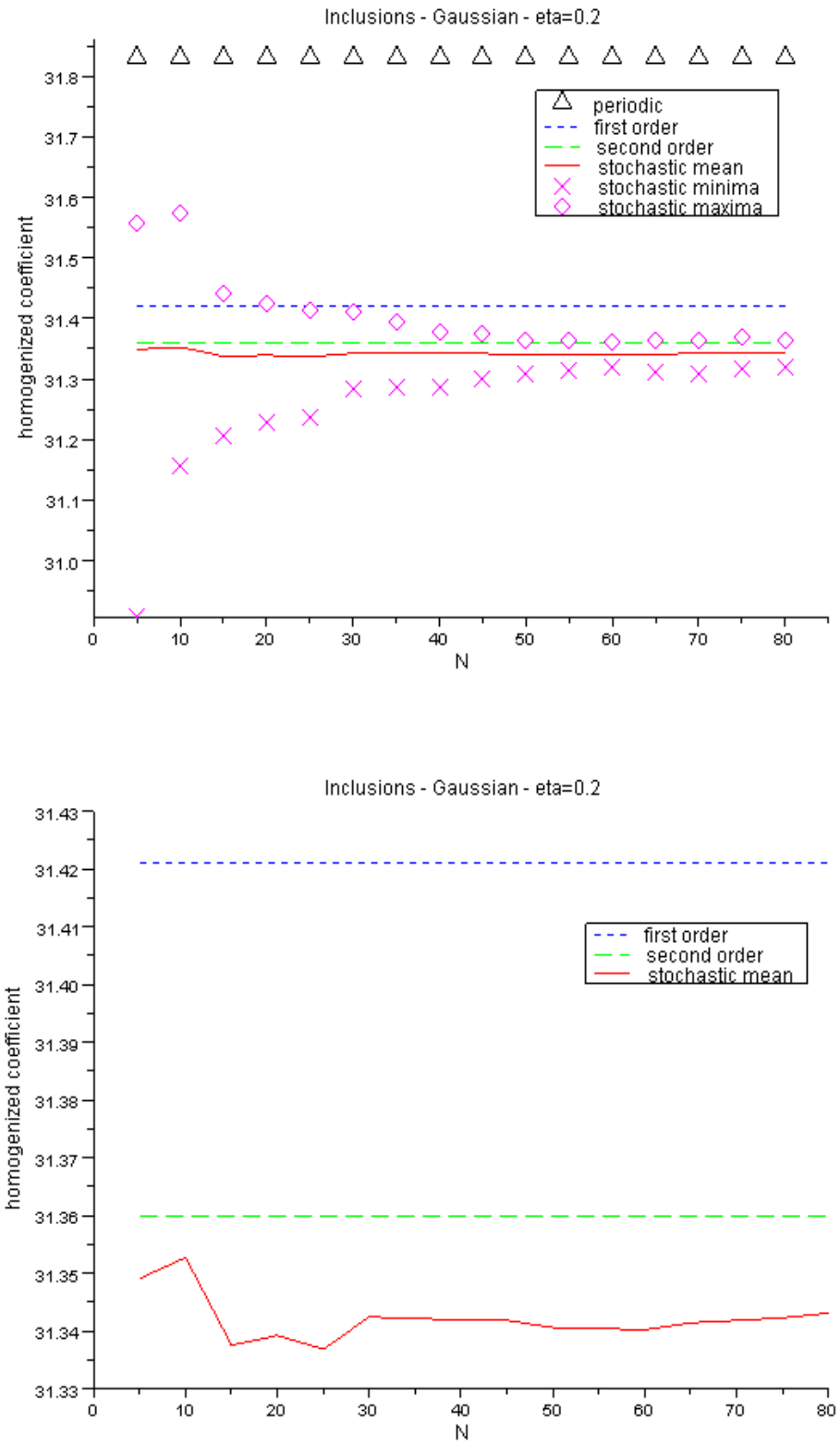

Figure 4.2: Inclusions - Results for a Gaussian perturbation and $\eta = 0.2$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

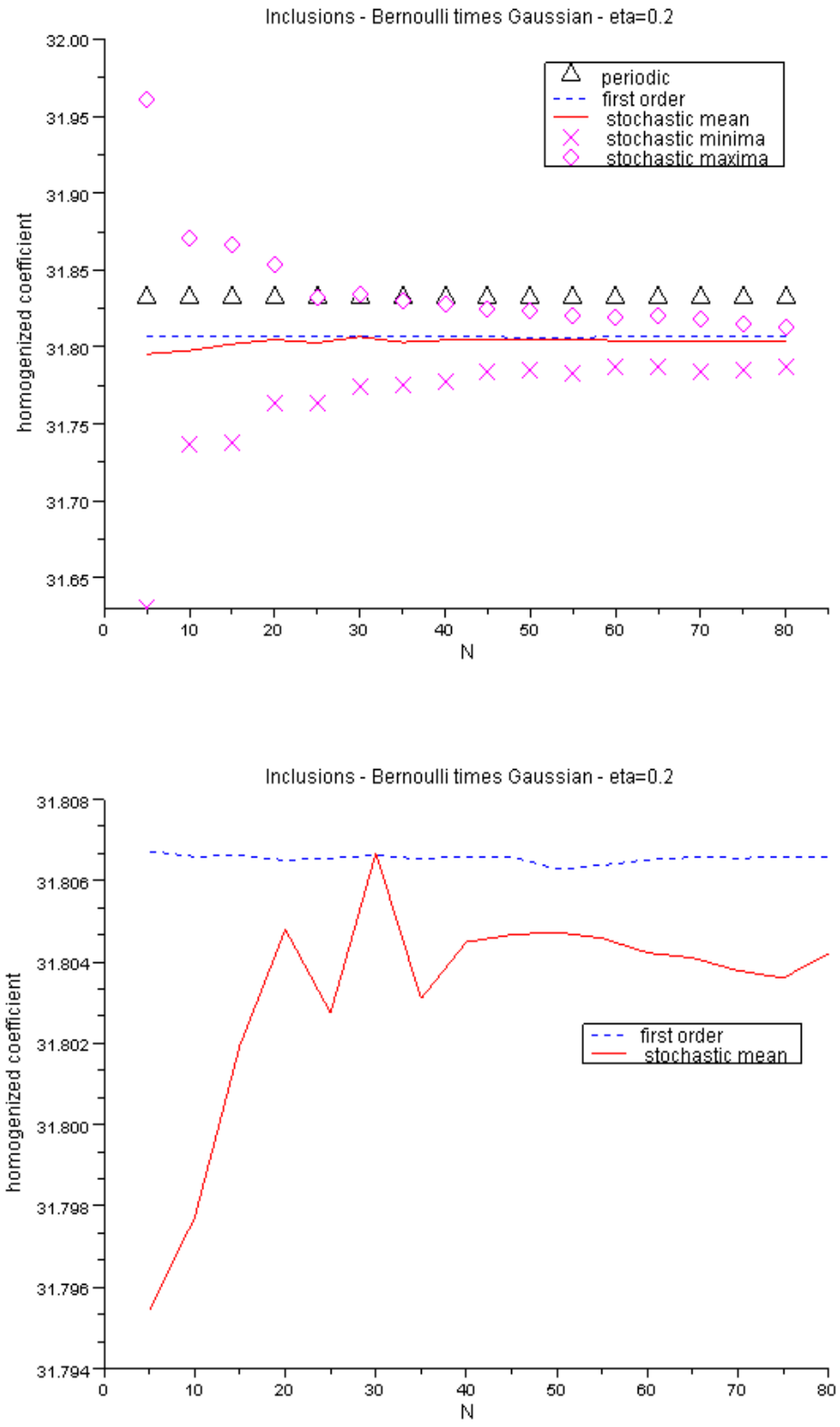

Figure 4.3: Inclusions - results for perturbation (4.116) and $\eta = 0.1$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first-order correction.

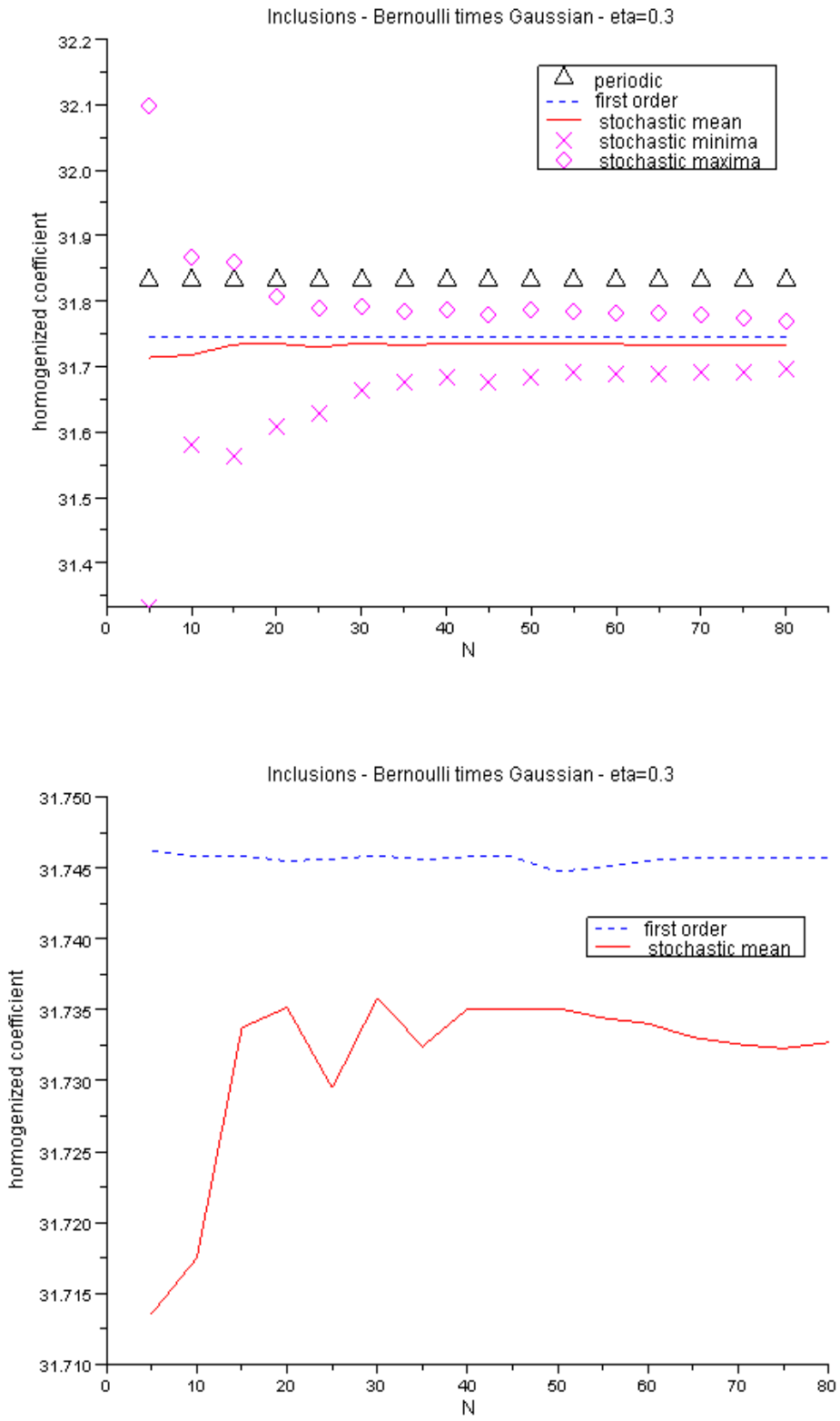

Figure 4.4: Inclusions - results for perturbation (4.116) and $\eta = 0.3$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first-order correction.

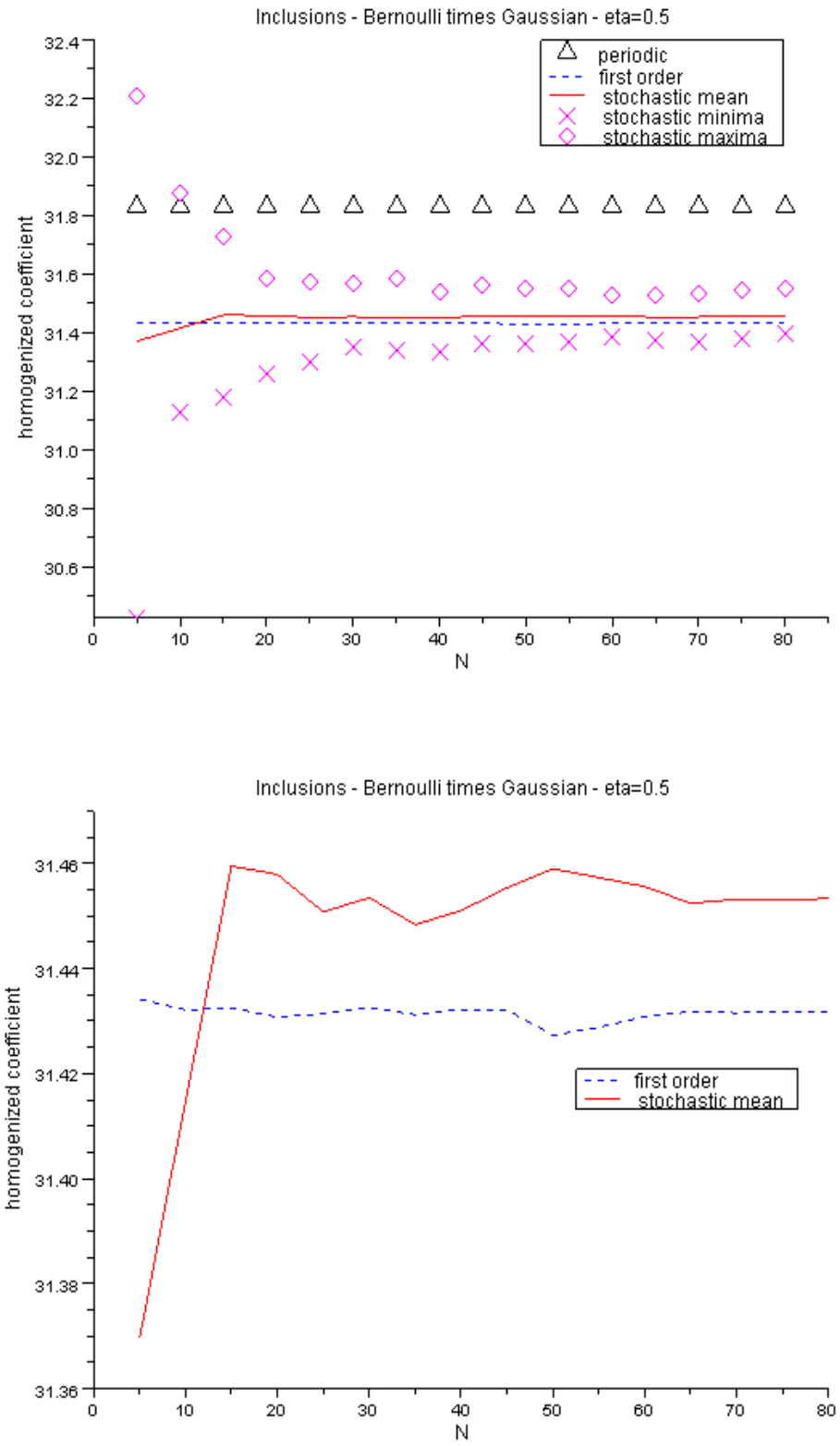

Figure 4.5: Inclusions - results for perturbation (4.116) and $\eta = 0.5$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first-order correction.

4.4.4 A second example of setting for our formal approach of Section 4.3

Consider R_η a random variable having Bernoulli law with parameter η, and U a uniform variable on $[0,1]$ independent of R_η. We define $B_\eta = R_\eta - \eta U$. Then

$$\begin{aligned}
\mathbb{E}(\varphi(B_\eta)) &= \mathbb{E}(\varphi(R_\eta - \eta U)) \\
&= \eta\mathbb{E}(\varphi(1-\eta U)) + (1-\eta)\mathbb{E}(\varphi(-\eta U)) \\
&= \eta\left(\varphi(1) - \eta\mathbb{E}(U)\varphi'(1) + o(\eta)\right) \\
&\quad +(1-\eta)\left(\varphi(0) - \eta\mathbb{E}(U)\varphi'(0) + \frac{\eta^2}{2}\mathbb{E}(U^2)\varphi''(0) + o(\eta^2)\right) \\
&= \varphi(0) + \eta\left(-\mathbb{E}(U)\varphi'(0) + \varphi(1) - \varphi(0)\right) \\
&\quad +\eta^2\left(-\mathbb{E}(U)(\varphi'(1) - \varphi'(0)) + \frac{1}{2}\mathbb{E}(U^2)\varphi''(0)\right) + o(\eta^2),
\end{aligned}$$

so that

$$\begin{aligned}
dP_\eta =& \delta_0 + \eta\left(-\mathbb{E}(U)\delta_0' + \delta_1 - \delta_0\right) \\
&+ \eta^2\left(-\mathbb{E}(U)(\delta_1' - \delta_0') + \frac{1}{2}\mathbb{E}(U^2)\delta''(0)\right) + o(\eta^2) \quad \text{in } \mathcal{E}'(\mathbb{R}).
\end{aligned} \tag{4.117}$$

Notice that this complex case is a mixture of Sections 4.2 and 4.3. The first-order perturbation is of course only the sum of the first-order perturbations for a Bernoulli law (Section 4.3 and Chapter 3) and a uniform law (Section 4.2). The interaction of these laws at order 2, and notably the δ_1' term, is much more involved and requires the computation of the cross derivatives of $w_i^{2,s,t,0,k,N}$ with respect to s and t at $s=0$ and $t=1$.

We give the results in the case of the inclusions and for $\eta = 0.05$, $\eta = 0.1$ and $\eta = 0.2$ (Figures 4.6, 4.7, 4.8, respectively).

For $\eta = 0.05$ and $\eta = 0.1$, the results display the same features as in our previous tests and are very good. The case $\eta = 0.2$ is instructive: the second-order expansion significantly departs from the "exact" value provided by the direct stochastic computation. Our interpretation is that, far from contradicting the validity of our expansion in the limit of small η, it shows the limitations of the approach. The value $\eta = 0.2$ is too large for the expansion to be accurate in the case of a lattice of inclusions with a high contrast between the inclusions and the surrounding phase.

Interestingly, a value of η twice as large (0.4) provides a very accurate approximation for another material, as shown by our final test performed on the laminate (Figure 4.9).

Our approach has limitations and deteriorates, like any asymptotic approach, for large values of η. The threshold is case dependent. The approach is however generically robust.

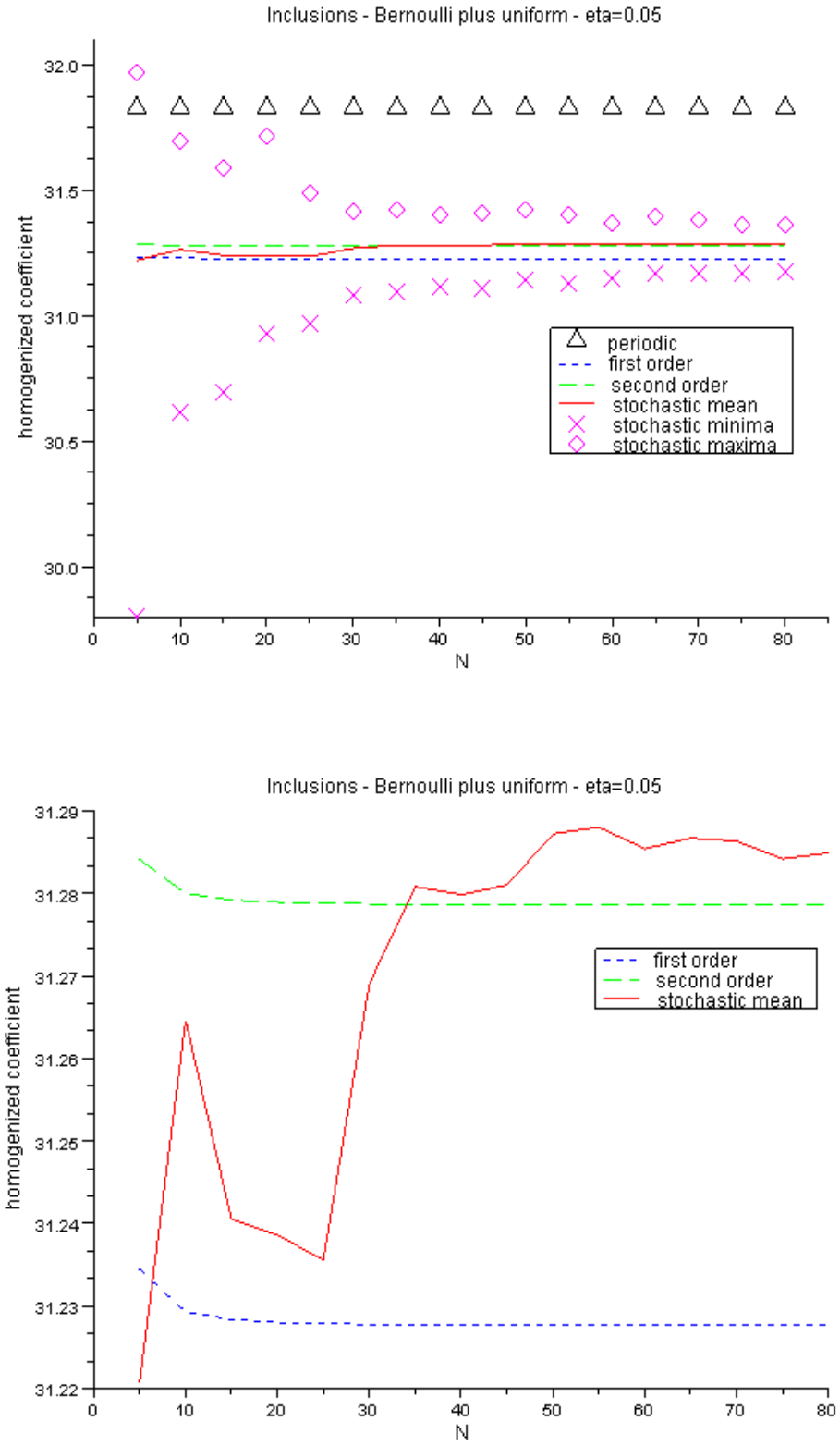

Figure 4.6: Inclusions - results for perturbation (4.117) and $\eta = 0.05$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

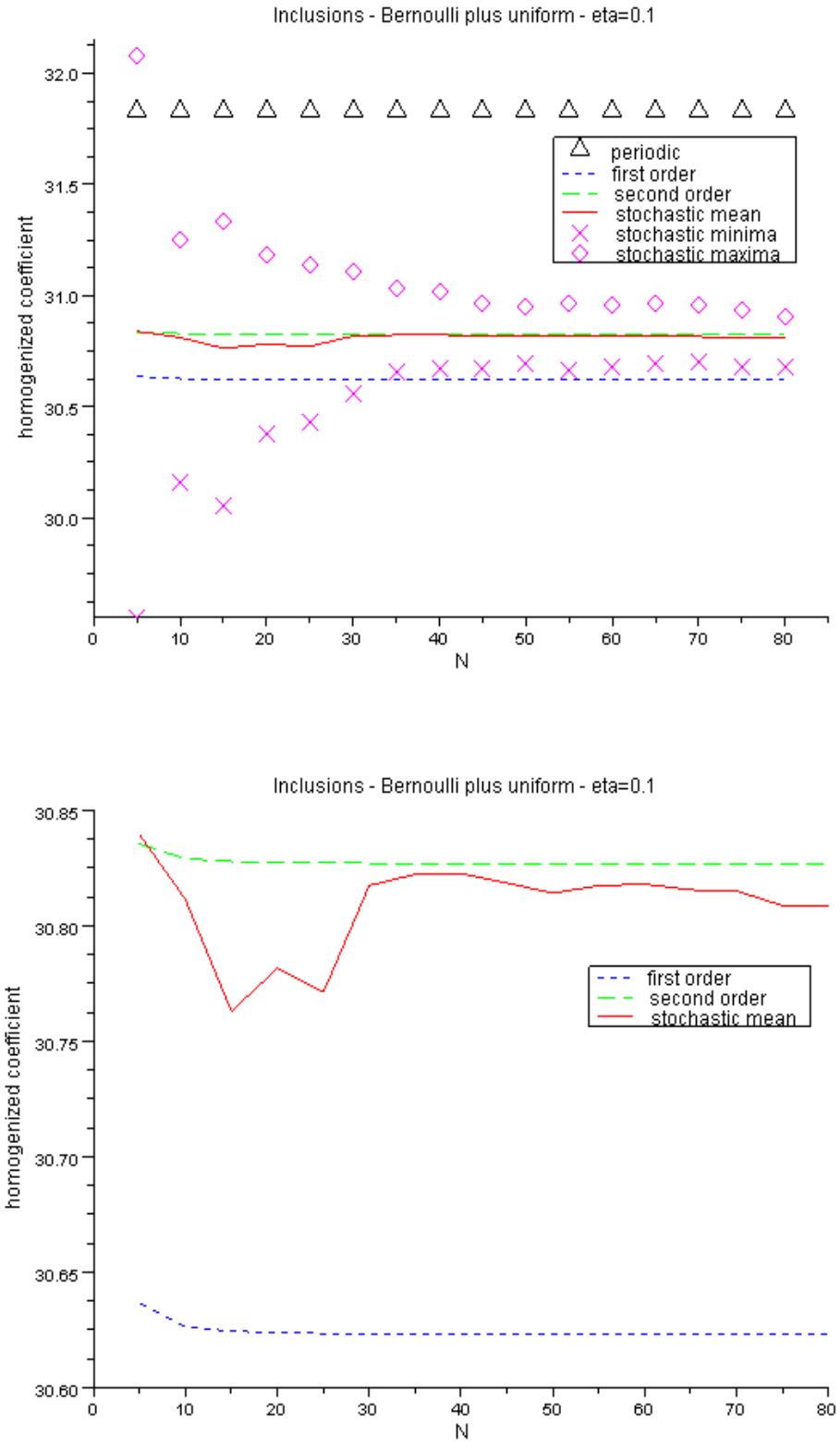

Figure 4.7: Inclusions - results for perturbation (4.117) and $\eta = 0.1$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

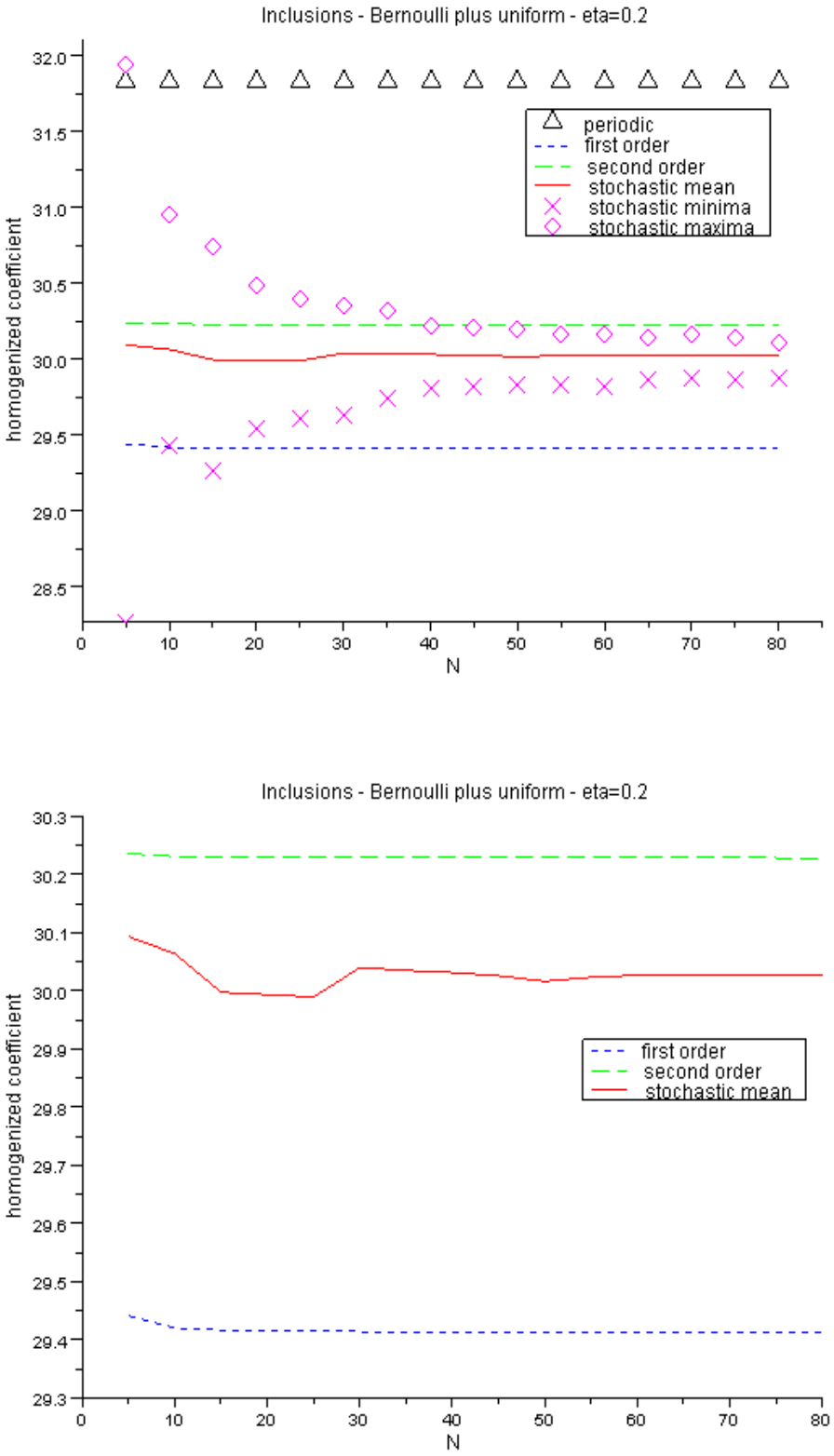

Figure 4.8: Inclusions - results for perturbation (4.117) and $\eta = 0.2$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

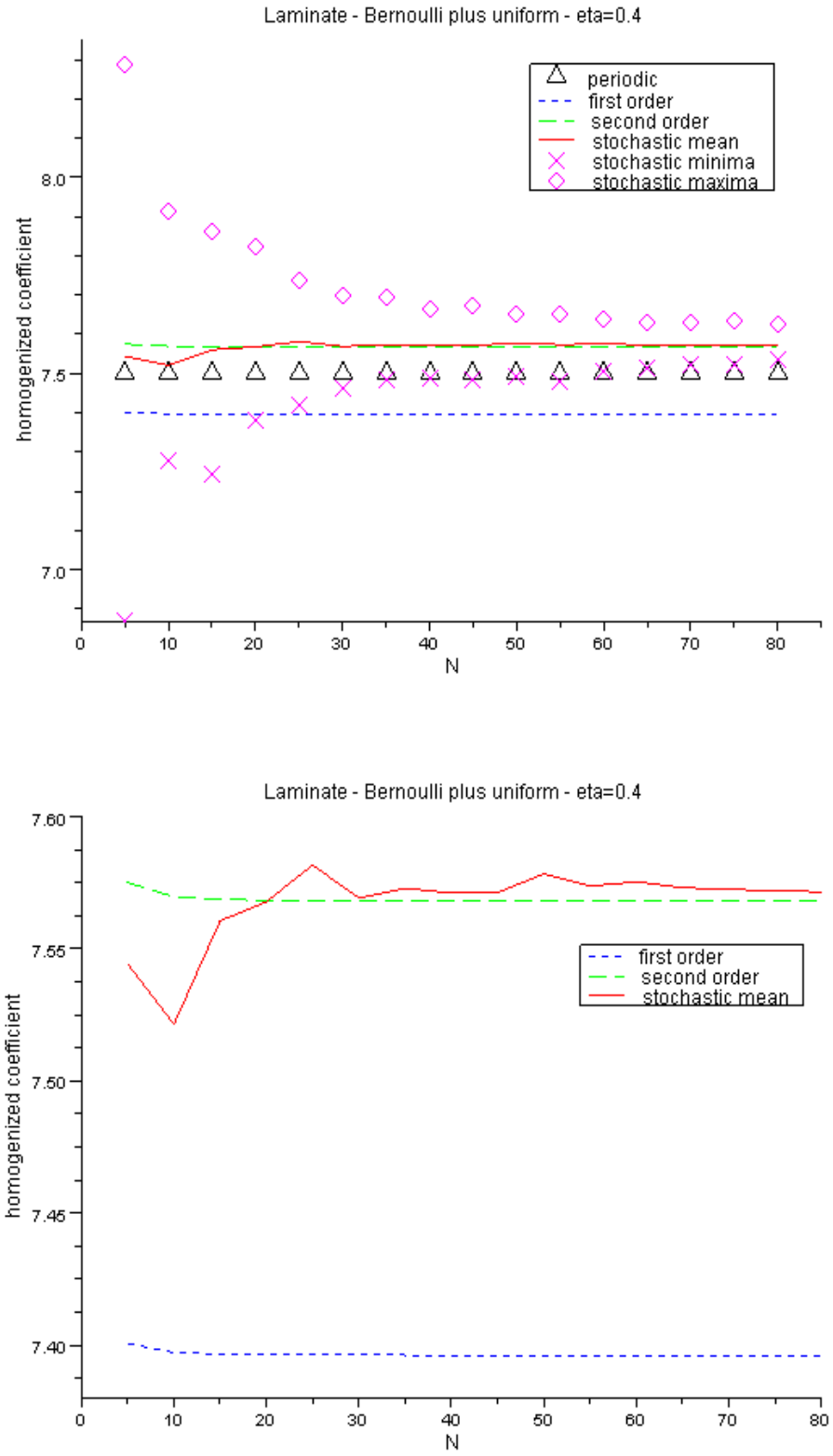

Figure 4.9: Laminate - results for perturbation (4.117) and $\eta = 0.4$. Above: complete results. Below: close-up on $A_\eta^{*,N}$ and the first and second-order corrections.

4.5 Appendix

The objectives of this Appendix are diverse. We first quickly recall some elements of distribution theory. We then prove technical results used in Section 4.3. Next we show that the approach formally derived in Section 4.3 is rigorous in dimension one. Finally we prove that this approach is also rigorous, in general dimensions, in a specific setting close to those of Theorem 4.2 and Corollary 4.4.

4.5.1 Elements of distribution theory

We recall here some basic definitions and results of distribution theory for convenience of the reader. See [44] for a comprehensive presentation.

In this section, $\mathcal{O}$ denotes an open set in $\mathbb{R}$.

Definition 4.9. *We denote by $\mathcal{D}(\mathcal{O})$ the space of infinitely differentiable functions on $\mathcal{O}$ having compact support in $\mathcal{O}$.*

Definition 4.10. *T is a distribution on $\mathcal{O}$ if T is a linear form on $\mathcal{D}(\mathcal{O})$ satisfying the following continuity property: for every compact $K \subset \mathcal{O}$, there exists an integer p and a constant C such that for all $\varphi \in \mathcal{D}(\mathcal{O})$ having compact support in K,*

$$|\langle T, \varphi \rangle| \leq C \sup_{x \in K, 0 \leq n \leq p} \left| \frac{d^n}{dx^n} \varphi(x) \right|. \tag{4.118}$$

The space of distributions on $\mathcal{O}$ is denoted by $\mathcal{D}'(\mathcal{O})$.

If the integer p in (4.118) can be chosen independently of K, the distribution T is said to have a finite order. The smallest possible value for p is called the order of T.

Definition 4.11. *A distribution $T \in \mathcal{D}'(\mathcal{O})$ is said to have compact support if there exists a compact set $K \subset \mathcal{O}$ such that for all $\varphi \in \mathcal{D}(\mathcal{O})$ having compact support in $\mathcal{O} \backslash K$, $\langle T, \varphi \rangle = 0$.*

The support of T is defined as the smallest compact set K which satisfies the above assertion.

The space of distributions on $\mathcal{O}$ having compact support is denoted by $\mathcal{E}'(\mathcal{O})$.

Proposition 4.12. *If $T \in \mathcal{E}'(\mathcal{O})$, its action on $\mathcal{D}(\mathcal{O})$ can be naturally extended to $\mathcal{C}^\infty(\mathcal{O})$. Denoting by K a compact neighborhood of the support of T, and by χ a cut-off function in $\mathcal{D}(\mathcal{O})$ equal to 1 on K, we define*

$$\forall \varphi \in \mathcal{C}^\infty(\mathcal{O}),\ \langle T, \varphi \rangle := \langle T, \chi \varphi \rangle.$$

This definition does not depend on K and χ.

Proposition 4.13. *If a distribution T is in $\mathcal{E}'(\mathcal{O})$, it has a finite order. Denoting by p its order and by K a compact neighborhood of the support of T, there exists a constant $C > 0$ such that:*

$$\forall \varphi \in C^\infty(\mathcal{O}),\ |\langle T, \varphi \rangle| \leq C \sup_{x \in K, 0 \leq n \leq p} \left| \frac{d^n}{dx^n} \varphi(x) \right|.$$

4.5.2 Some technical results

This section is devoted to the proof of technical lemmas used in Section 4.3. Loosely speaking, these lemmas all deal with the variation of the supercell correctors defined by (4.59), (4.68), and (4.72) with respect to the amplitudes of the defects.

Lemma 4.14. *Let $\tilde{H}^1_{per}(I_N)$ be the set of $(N\mathbb{Z})^d$-periodic functions in $H^1_{loc}(\mathbb{R}^d)$ with zero mean on I_N. The function*

$$F :]-M,M[^{N^d} \ni (s_1,\cdots,s_{N^d}) \mapsto \bar{w}_i^{s_1,\cdots,s_{N^d}} \in \tilde{H}^1_{per}(I_N),$$

where $\bar{w}_i^{s_1,\cdots,s_{N^d}} = w_i^{s_1,\cdots,s_{N^d}} - \int_{I_N} w_i^{s_1,\cdots,s_{N^d}}$ and $w_i^{s_1,\cdots,s_{N^d}}$ is defined by (4.59), is C^∞.

Proof. For $(s_1,\cdots,s_{N^d}) \in [-M,M]^{N^d}$, $\bar{w}_i^{s_1,\cdots,s_{N^d}}$ is the unique solution to

$$\begin{cases} -\operatorname{div}\left(A^{s_1,\cdots,s_{N^d}}(\nabla \bar{w}_i^{s_1,\cdots,s_{N^d}} + e_i)\right) = 0 \quad \text{in } I_N, \\ \bar{w}_i^{s_1,\cdots,s_{N^d}} \ (N\mathbb{Z})^d - \text{periodic}, \quad \displaystyle\int_{I_N} \bar{w}_i^{s_1,\cdots,s_{N^d}} = 0, \end{cases}$$

so that F is well defined.
Let us now define $G :]-M,M[^{N^d} \times \tilde{H}^1_{per}(I_N) \to H^{-1}(I_N)$ by

$$G(s_1,\cdots,s_{N^d},w) = -\operatorname{div}\left(A^{s_1,\cdots,s_{N^d}}(\nabla w + e_i)\right),$$

so that $F(s_1,\cdots,s_{N^d}) = \bar{w}_i^{s_1,\cdots,s_{N^d}}$ is the unique solution to

$$G(s_1,\cdots,s_{N^d},F(s_1,\cdots,s_{N^d})) = 0.$$

It is easy to see that G is a C^1 function, and that

$$\forall h \in \tilde{H}^1_{per}(I_N), \quad \partial_w G(s_1,\cdots,s_{N^d},w)\cdot h = -\operatorname{div}\left(A^{s_1,\cdots,s_{N^d}}\nabla h\right),$$

where $\partial_w G(s_1,\cdots,s_{N^d},w)$ is the first derivative of G with respect to w at $(s_1,\cdots,s_{N^d},w)$.

The Lax-Milgram theorem and the coercivity of $A^{s_1,\cdots,s_{N^d}}$ show that $\partial_w G(s_1,\cdots,s_{N^d},w)$ is an isomorphism. We can therefore apply the inverse function theorem and deduce that F is C^1, with $\partial_{s_l} F$ the unique solution to

$$\begin{cases} -\operatorname{div}\left(A^{s_1,\cdots,s_{N^d}}(\nabla \partial_{s_l} F)\right) = \operatorname{div}\left(\mathbb{1}_{Q_l} C_{per}(\nabla F + e_i)\right) \quad \text{in } I_N, \\ \partial_{s_l} F \ (N\mathbb{Z})^d - \text{periodic}, \quad \displaystyle\int_{I_N} \partial_{s_l} F = 0. \end{cases}$$

Arguing by induction, we obtain that F is a C^∞ function.

□

For consistency, we state next a lemma proved in Chapter 3 (as Lemma 3.6).

Lemma 4.15. *Consider $f \in L^2(Q)$, and a tensor field A from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$ such that there exist $\lambda > 0$ and $\Lambda > 0$ such that*

$$\forall \xi \in \mathbb{R}^d, \text{ a.e in } x \in \mathbb{R}^d, \ \lambda|\xi|^2 \leq A(x)\xi \cdot \xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|.$$

Consider q^N solution to

$$\begin{cases} -\operatorname{div}\left(A\nabla q^N\right) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } I_N, \\ q^N \ (N\mathbb{Z})^d - \text{periodic.} \end{cases} \tag{4.119}$$

Then $\mathbb{1}_{I_N}\nabla q^N$ converges in $L^2(\mathbb{R}^d)$, when N goes to infinity, to ∇q^∞, where q^∞ is a $L^2_{loc}(\mathbb{R}^d)$ function solving

$$\begin{cases} -\operatorname{div}\left(A\nabla q^\infty\right) = \operatorname{div}(\mathbb{1}_Q f) & \text{in } \mathbb{R}^d, \\ \nabla q^\infty \in L^2(\mathbb{R}^d). \end{cases} \tag{4.120}$$

Lemma 4.16. *Consider $q_i^{1,s,0,N}$ and $q_i^{1,s,0,\infty}$ solutions to (4.80) and (4.82) respectively, and $n \in \mathbb{N}$. There exists a constant $C(n,M)$, such that*

$$\begin{aligned} \forall s \in]-M, M[, \ \forall N \in 2\mathbb{N}+1, \ \ \|\nabla \partial_s^n q_i^{1,s,0,N}\|_{L^2(I_N)} \\ \leq C(n,M)\|\nabla w_i^0 + e_i\|_{L^2(Q)}, \end{aligned} \tag{4.121}$$

$$\forall s \in]-M, M[, \ \ \|\nabla \partial_s^n q_i^{1,s,0,\infty}\|_{L^2(\mathbb{R}^d)} \leq C(n,M)\|\nabla w_i^0 + e_i\|_{L^2(Q)}. \tag{4.122}$$

Proof. Multiplying the first line of (4.80) by $q_i^{1,s,0,N}$ and integrating by parts, we find that

$$\|\nabla q_i^{1,s,0,N}\|_{L^2(I_N)} \leq M\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}\|\nabla w_i^0 + e_i\|_{L^2(Q)}, \tag{4.123}$$

where α is defined by (4.8).

Thus (4.121) is true for $n = 0$ with $C(0,M) = M\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}$.

Next, the first derivative $\partial_s q_i^{1,s,0,N}$ is solution to

$$\begin{cases} -\operatorname{div}(A_1^{s,0}\nabla \partial_s q_i^{1,s,0,N}) = \operatorname{div}\left(\mathbb{1}_Q C_{per}(\nabla w_i^0 + e_i)\right) \\ \qquad\qquad + \operatorname{div}\left(\mathbb{1}_Q C_{per}\nabla q_i^{1,s,0,N}\right) \quad \text{in } I_N, \\ \partial_s q_i^{1,s,0,N} \ \ (N\mathbb{Z})^d - \text{periodic,} \end{cases} \tag{4.124}$$

from which we deduce

$$\|\nabla \partial_s q_i^{1,s,0,N}\|_{L^2(I_N)} \leq \frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}\left(\|\nabla w_i^0 + e_i\|_{L^2(Q)} + \|\nabla q_i^{1,s,0,N}\|_{L^2(Q)}\right)$$

and, using (4.123),

$$\|\nabla \partial_s q_i^{1,s,0,N}\|_{L^2(I_N)} \leq \frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}(M+1)\|\nabla w_i^0 + e_i\|_{L^2(Q)}.$$

Thus (4.121) is true for $n = 1$ with $C(1, M) = (M + 1)\frac{\|C_{per}\|_{L^\infty(Q)}}{\alpha}$.

Finally, we have for $n \geq 2$

$$\begin{cases} -\operatorname{div}(A_1^{s,0}\nabla\partial_s^n q_i^{1,s,0,N}) = n \operatorname{div}\left(\mathbb{1}_Q C_{per}\nabla\partial_s^{n-1} q_i^{1,s,0,N}\right) & \text{in } I_N, \\ \nabla\partial_s^n q_i^{1,s,0,N} \ \ (N\mathbb{Z})^d - \text{periodic}, \end{cases} \tag{4.125}$$

so that an easy induction proves (4.121). The proof of (4.122) is identical. □

The following result is an immediate consequence of Lemma 4.16.

Lemma 4.17. *Consider $q_i^{1,s,0,N}$ and $q_i^{1,s,0,\infty}$ solutions to (4.80) and (4.82) respectively. For every $n \in \mathbb{N}$, there exists a constant $C(n, M)$ such that for all $(s, s') \in]-M, M[^2$,*

$$\forall N \in 2\mathbb{N}+1, \ \ \|\nabla\partial_s^n q_i^{1,s,0,N} - \nabla\partial_s^n q_i^{1,s',0,N}\|_{L^2(I_N)} \leq C(n, M)\|\nabla w_i^0 + e_i\|_{L^2(Q)}|s - s'|, \tag{4.126}$$

$$\|\nabla\partial_s^n q_i^{1,s,0,\infty} - \nabla\partial_s^n q_i^{1,s',0,\infty}\|_{L^2(\mathbb{R}^d)} \leq C(n, M)\|\nabla w_i^0 + e_i\|_{L^2(Q)}|s - s'|. \tag{4.127}$$

Lemma 4.18. *Consider $q_i^{2,s,t,0,k,N}$ and $q_i^{2,s,t,0,k,\infty}$ solutions to (4.91) and (4.112) respectively, and $(p, r) \in \mathbb{N}^2$. There exists a constant $C(p, r, M)$ such that for all $(s, t) \in]-M, M[^2$,*

$$\|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,N}\|_{L^2(I_N)} \leq C(p, r, M)\left(\sum_{0\leq m\leq p}\|\nabla\partial_s^m q_i^{1,s,0,N}\|_{L^2(Q+k)} + \sum_{0\leq n\leq r}\|\nabla\partial_t^n q_i^{1,t,k,N}\|_{L^2(Q)}\right), \tag{4.128}$$

$$\|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,\infty}\|_{L^2(I_N)} \leq C(p, r, M)\left(\sum_{0\leq m\leq p}\|\nabla\partial_s^m q_i^{1,s,0,\infty}\|_{L^2(Q+k)} + \sum_{0\leq n\leq r}\|\nabla\partial_t^n q_i^{1,t,k,\infty}\|_{L^2(Q)}\right), \tag{4.129}$$

and

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}}\|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,N}\|^2_{L^2(I_N)} \leq C(p, r, M)\|\nabla w_i^0 + e_i\|^2_{L^2(Q)}, \tag{4.130}$$

$$\sum_{k\in\mathbb{Z}^d\setminus\{0\}}\|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,\infty}\|^2_{L^2(\mathbb{R}^d)} \leq C(p, r, M)\|\nabla w_i^0 + e_i\|^2_{L^2(Q)}. \tag{4.131}$$

Proof. The proof of (4.128) is identical to that of (4.121) in Lemma 4.16.

Summing (4.128) over all $k \in \mathcal{T}_N \backslash \{0\}$, we obtain

$$\begin{aligned}&\sum_{k\in\mathcal{T}_N\backslash\{0\}} \|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,N}\|^2_{L^2(I_N)}\\ &\leq C(p,r,M)\sum_{k\in\mathcal{T}_N}\left(\sum_{0\leq m\leq p}\|\nabla\partial_s^m q_i^{1,s,0,N}\|^2_{L^2(Q+k)} + \sum_{0\leq n\leq r}\|\nabla\partial_t^n q_i^{1,t,k,N}\|^2_{L^2(Q)}\right).\end{aligned} \tag{4.132}$$

Next, we have

$$\sum_{k\in\mathcal{T}_N}\|\nabla\partial_s^m q_i^{1,s,0,N}\|^2_{L^2(Q+k)} = \|\nabla\partial_s^m q_i^{1,s,0,N}\|^2_{L^2(I_N)}, \tag{4.133}$$

and since $q_i^{1,t,k,N} = q_i^{1,t,0,N}(\cdot - k)$,

$$\sum_{k\in\mathcal{T}_N}\|\nabla\partial_t^n q_i^{1,t,k,N}\|^2_{L^2(Q)} = \|\nabla\partial_t^n q_i^{1,t,0,N}\|^2_{L^2(I_N)}. \tag{4.134}$$

Thus (4.132), (4.133) and (4.134) yield

$$\begin{aligned}&\sum_{k\in\mathcal{T}_N\backslash\{0\}} \|\nabla\partial_s^p\partial_t^r q_i^{2,s,t,0,k,N}\|^2_{L^2(I_N)}\\ &\leq C(p,r,M)\left(\sum_{0\leq m\leq p}\|\nabla\partial_s^m q_i^{1,s,0,N}\|^2_{L^2(I_N)} + \sum_{0\leq n\leq r}\|\nabla\partial_t^n q_i^{1,t,0,N}\|^2_{L^2(I_N)}\right).\end{aligned} \tag{4.135}$$

We finally obtain (4.130) using (4.121) in (4.135). The proofs of (4.129) and (4.131) are similar.

□

4.5.3 The one-dimensional case

We address here the one-dimensional context. All the computations are explicit, for the settings of Sections 4.2 and 4.3. To stress the fact that we deal with scalar quantities, we use lower-case letters for the tensors. Note also that in this section $Q = [-\frac{1}{2}, \frac{1}{2}]$ and $I_N = [-\frac{N}{2}, \frac{N}{2}]$.

4.5.3.1 An extension of Theorem 4.2

The following theorem extends the result of Theorem 4.2, stated in $L^\infty(Q; L^2(\Omega))$, to $L^\infty(Q; L^p(\Omega))$ for any $p \in]1, \infty]$:

Theorem 4.19 (one-dimensional setting). *Assume that $d = 1$, that b_η satisfies (4.6) and $m_\eta := \|b_\eta\|_{L^\infty([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))} \underset{\eta\to 0}{\to} 0$ for some $p > 1$. There exists a subsequence of η, still denoted η for simplicity, such that $\frac{b_\eta}{m_\eta}$ converges weakly-* in $L^\infty([-\frac{1}{2}, \frac{1}{2}]; L^p(\Omega))$ to a limit field denoted by $\bar{b}_0$ when $\eta \to 0$. Then*

- *the expansion*

$$\frac{d}{dx}w^\eta = \frac{d}{dx}w^0 + m_\eta \frac{d}{dx}v^0 + o(m_\eta) \tag{4.136}$$

holds weakly in $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$, *where* w^0 *is the periodic corrector and* v^0 *solves*

$$\begin{cases} -\dfrac{d}{dx}(a_{per}\dfrac{d}{dx}v^0) = \dfrac{d}{dx}\left(\bar{b}_0 c_{per}(\dfrac{d}{dx}w^0+1)\right) & \text{in } \mathbb{R}, \\ \dfrac{d}{dx}v^0 \text{ stationary, } \mathbb{E}\left(\displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{d}{dx}v^0\right) = 0. \end{cases} \tag{4.137}$$

- a_η^* *reads*

$$a_\eta^* = a_{per}^* + m_\eta \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{E}(\bar{b}_0) c_{per}(\frac{d}{dx}w^0+1) + m_\eta \int_{-\frac{1}{2}}^{\frac{1}{2}} a_{per}\frac{d}{dx}\mathbb{E}(v^0) + o(m_\eta).$$

Proof. The periodic and stochastic correctors can be computed explicitly. They are respectively given by

$$\frac{d}{dx}w^0 = \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} a_{per}^{-1}\right)^{-1} a_{per}^{-1} - 1 \text{ and } \frac{d}{dx}w^\eta = \left(\mathbb{E}\left(\int_{-\frac{1}{2}}^{\frac{1}{2}} a_\eta^{-1}\right)\right)^{-1} a_\eta^{-1} - 1.$$

Note that w^0 is in $W^{1,\infty}(-\frac{1}{2},\frac{1}{2})$.

We define $v^\eta = \dfrac{w^\eta - w^0}{m_\eta}$. It solves

$$\begin{cases} -\dfrac{d}{dx}(a_\eta \dfrac{d}{dx}v^\eta) = \dfrac{d}{dx}\left(\dfrac{b_\eta}{\eta} c_{per}(\dfrac{d}{dx}w^0+1)\right) & \text{in } \mathbb{R}, \\ \dfrac{d}{dx}v^\eta \text{ stationary, } \mathbb{E}\left(\displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{d}{dx}v^\eta\right) = 0. \end{cases} \tag{4.138}$$

We deduce from (4.138) that

$$a_\eta \frac{d}{dx}v^\eta = \frac{b_\eta}{m_\eta} c_{per}(\frac{d}{dx}w^0+1) + k_\eta, \tag{4.139}$$

where k_η depends only on ω. Since k_η is by construction stationary ergodic, it is constant, and we compute from (4.138) and (4.139):

$$k_\eta = -\frac{1}{m_\eta}\left(\mathbb{E}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_\eta}\right)^{-1} \times \left(\mathbb{E}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{b_\eta}{a_\eta} c_{per}(\frac{d}{dx}w^0+1)\right).$$

Since w^0 is in $W^{1,\infty}(-\frac{1}{2},\frac{1}{2})$, a_η is coercive and c_{per} is bounded, it holds

$$\begin{aligned} |k_\eta| &\leq C\frac{\|b_\eta\|_{L^1([-\frac{1}{2},\frac{1}{2}]\times\Omega)}}{m_\eta} \\ &\leq C\frac{\|b_\eta\|_{L^1([-\frac{1}{2},\frac{1}{2}]\times\Omega)}}{\|b_\eta\|_{L^\infty([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))}}. \end{aligned}$$

This implies that k_η is a bounded function of η whatever $p \geq 1$ and thus, using (4.139), that $\frac{d}{dx}v^\eta$ is bounded in $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$ for all $p \geq 1$. As a result, for $p > 1$, $\frac{d}{dx}v^\eta$ converges weakly and up to extraction in $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$ to a limit we denote $\frac{d}{dx}v^0$.

The random field b_η tends to 0 in $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$. Since it is bounded in $L^\infty([-\frac{1}{2},\frac{1}{2}]\times\Omega)$, it converges to 0 in $L^2([-\frac{1}{2},\frac{1}{2}];L^r(\Omega))$ for all $r > p$. By Hölder inequality it also converges to 0 in $L^2([-\frac{1}{2},\frac{1}{2}];L^r(\Omega))$ for all $1 < r < p$. Thus it converges to 0 in $L^2([-\frac{1}{2},\frac{1}{2}];L^q(\Omega))$ where $q = \frac{p}{p-1}$.

The space $L^2([-\frac{1}{2},\frac{1}{2}];L^q(\Omega))$ being the dual of $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$, we obtain that $b_\eta\, c_{per}\, \frac{d}{dx}v^\eta$ tends to 0 in $\mathcal{D}'([-\frac{1}{2},\frac{1}{2}]\times\Omega)$. We can then take the limit $\eta \to 0$ in (4.138) and obtain that v^0 is solution to (4.137).

We have thus proved that $\frac{1}{m_\eta}\left(\frac{d}{dx}w^\eta - \frac{d}{dx}w^0\right)$ converges, up to extraction, weakly to $\frac{d}{dx}v^0$ in $L^2([-\frac{1}{2},\frac{1}{2}];L^p(\Omega))$, which is equivalent to (4.136).

The second assertion of Theorem 4.19 is obtained by inserting (4.136) into the expression (4.11) of a^*_η. □

Note that the proof of Theorem 4.19 depends crucially on the fact that we are able to solve explicitly the cell problems.

Theorem 4.19 allows for a better intuitive understanding of Theorem 4.2. In dimension one, the homogenized coefficient is explicitly given by

$$a^*_\eta = \left(\mathbb{E}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per} + b_\eta c_{per}}\right)^{-1},$$

which, when $b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}} \mathbb{1}_{[k,k+1]}(x) B_\eta(\tau_k\omega)$, may be rewritten as the *formal* series

$$\frac{1}{a^*_\eta} = \sum_{k=0}^{\infty}(-1)^k \mathbb{E}((B_\eta)^k)\int_{-\frac{1}{2}}^{\frac{1}{2}}\left(\frac{c_{per}}{a_{per}}\right)^k a_{per}^{-1}. \tag{4.140}$$

Assume now that there exists $p > 1$ such that $\|B_\eta\|_{L^p(\Omega)} \to 0$ when $\eta \to 0$ and $\frac{B_\eta}{\|B_\eta\|_{L^p(\Omega)}}$ converges weakly in $L^p(\Omega)$ to some $\bar{B}_0$ with $\mathbb{E}(\bar{B}_0) \neq 0$. We have in particular

$$\frac{\mathbb{E}(B_\eta)}{\|B_\eta\|_{L^p(\Omega)}} \to \mathbb{E}(\bar{B}_0) \neq 0,$$

which, since $\mathbb{E}(|B_\eta|^p) \to 0$, implies

$$\mathbb{E}(|B_\eta|^p) = \underset{\eta\to 0^+}{o}\left(\mathbb{E}(B_\eta)\right). \tag{4.141}$$

We now claim that, without loss of generality and up to an extraction in η, we may take $p = 2$ in (4.141). Indeed, if $p < 2$, then since B_η is bounded in $L^\infty(\Omega)$, (4.141)

implies $\mathbb{E}(|B_\eta|^2) = \underset{\eta\to 0^+}{o}(\mathbb{E}(B_\eta))$. On the other hand, if $p > 2$, we consider the normalized sequence $\frac{B_\eta}{\|B_\eta\|_{L^2(\Omega)}}$ in $L^2(\Omega)$. Up to extraction, it weakly converges to $\bar{B}_2 \in L^2(\Omega)$. Since

$$\frac{\mathbb{E}(B_\eta)}{\|B_\eta\|_{L^p(\Omega))}} = \frac{\mathbb{E}(B_\eta)}{\|B_\eta\|_{L^2(\Omega)}} \frac{\|B_\eta\|_{L^2(\Omega)}}{\|B_\eta\|_{L^p(\Omega)}}$$

where the left hand side converges to $\mathbb{E}(\bar{B}_0) \neq 0$ and $\frac{\|B_\eta\|_{L^2(\Omega)}}{\|B_\eta\|_{L^p(\Omega)}}$ is bounded by 1 by Hölder's inequality, $\mathbb{E}(\bar{B}_2) = \lim_{\eta\to 0} \dfrac{\mathbb{E}(B_\eta)}{\|B_\eta\|_{L^2(\Omega))}} \neq 0$ and (4.141) is satisfied with $p = 2$.

We then take $p = 2$. Since $\mathbb{E}(|B_\eta|^2) = \underset{\eta\to 0^+}{o}(\mathbb{E}(B_\eta))$ and B_η is bounded in $L^\infty(\Omega)$, $\mathbb{E}(|B_\eta|^k) = \underset{\eta\to 0^+}{o}(\mathbb{E}(B_\eta))$ for all $k \geq 2$.

This intuitively expresses that all orders higher than or equal to 2 are negligible as compared to the first-order term in the series (4.140), and thus that a kind of "separation of scales" is satisfied. This is of course formal since one has to check that the remainder term consisting of the *sum* of all terms of order higher than or equal to 2 is $o(\mathbb{E}(B_\eta))$, so that

$$\begin{aligned} a^*_\eta &= \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}}\right)^{-1} + \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}}\right)^{-2} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{E}(B_\eta)\frac{c_{per}}{a_{per}}\right) + o(\mathbb{E}(B_\eta)) \\ &= \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}}\right)^{-1} + m_\eta \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}}\right)^{-2} \left(\int_{-\frac{1}{2}}^{\frac{1}{2}} \mathbb{E}(\bar{B}_0)\frac{c_{per}}{a_{per}}\right) + o(\mathbb{E}(B_\eta)). \end{aligned}$$

But this is the purpose of the proofs of Theorems 4.2 and 4.19, using another viewpoint, to show this is indeed the case.

4.5.3.2 The setting of Section 4.3 in dimension one

We now prove that our approach of Section 4.3 is rigorous in dimension one.

Lemma 4.20. *In dimension $d = 1$, it holds*

$$a^*_\eta = a^*_{per} + \eta\bar{a}^*_1 + \eta^2\bar{a}^*_2 + o(\eta^2),$$

where $\bar{a}^*_1$ and $\bar{a}^*_2$ are the limits as $N \to \infty$ of $a^{*,N}_1$ and $a^{*,N}_2$ defined generally by (4.69) and (4.73) respectively.

Proof. Recall that in dimension one, a^*_η is given by the simple explicit expression

$$a^*_\eta = \left(\mathbb{E}\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per} + b_\eta c_{per}}\right)^{-1} = \left\langle dP_\eta(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per} + sc_{per}}\right\rangle^{-1}.$$

The proof thus consists in inserting expansion (4.48) in this explicit expression and identifying successively the first three dominant orders.

Using (4.48), we write

$$\begin{aligned}(a_\eta^*)^{-1} &= \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}} + \eta \left\langle d\bar{P}_1(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle + \eta^2 \left\langle d\bar{P}_2(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle \\ &\quad +o(\eta^2) \\ &= (a_{per}^*)^{-1} \left(1 + \eta a_{per}^* \left\langle d\bar{P}_1(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle + \eta^2 a_{per}^* \left\langle d\bar{P}_2(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle\right) \\ &\quad +o(\eta^2).\end{aligned}$$

This yields the expansion

$$\begin{aligned}a_\eta^* =& a_{per}^* - \eta (a_{per}^*)^2 \left\langle d\bar{P}_1(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle \\ &+ \eta^2 (a_{per}^*)^3 \left\langle d\bar{P}_1(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle^2 \\ &- \eta^2 (a_{per}^*)^2 \left\langle d\bar{P}_2(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}} \right\rangle + o(\eta^2).\end{aligned} \tag{4.142}$$

We now devote the rest of the proof to verifying that the coefficients of η and η^2 in (4.142) are indeed obtained as the limit as $N \to \infty$ of $a_1^{*,N}$ and $a_2^{*,N}$ defined generally by (4.69) and (4.73) respectively, in this particular one-dimensional setting.

The function $w^{1,s,0,N}$ generally defined by (4.68) satisfies here

$$\begin{cases} -\frac{d}{dx}\left(a_1^{s,0}(\frac{d}{dx}w_i^{1,s,0,N}+1)\right) = 0 \quad \text{in }]-\frac{N}{2},\frac{N}{2}[, \\ w_i^{1,s,0,N}\ N-\text{periodic}. \end{cases} \tag{4.143}$$

We easily compute using (4.143):

$$\begin{aligned}a_1^{s,0}(\frac{d}{dx}w^{1,s,0,N}+1) &= N\left(\int_{-\frac{N}{2}}^{\frac{N}{2}} \frac{1}{a_{per}+s\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}c_{per}}\right)^{-1} \\ &= N\left(N(a_{per}^*)^{-1} - f(s)\right)^{-1} \\ &= a_{per}^* + \frac{(a_{per}^*)^2}{N} f(s) + \frac{(a_{per}^*)^3}{N^2} f(s)^2 + o(N^{-2}),\end{aligned}$$

where $f(s) = \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{sc_{per}}{a_{per}(a_{per}+sc_{per})}$.

Thus $a_1^{*,N}$ defined generally by (4.69) takes here the form

$$\begin{aligned}a_1^{*,N} &= \left\langle d\bar{P}_1(s), \int_{-\frac{N}{2}}^{\frac{N}{2}} a_1^{s,0}(\frac{d}{dx}w^{1,s,0,N}+1) \right\rangle \\ &= N a_{per}^* \left\langle d\bar{P}_1(s), 1\right\rangle + (a_{per}^*)^2 \left\langle d\bar{P}_1(s), f(s)\right\rangle + o(1).\end{aligned}$$

We know from Lemma 4.6 that $\left\langle d\bar{P}_1(s), 1\right\rangle = 0$, whence

$$a_1^{*,N} \underset{N\to\infty}{\to} \bar{a}_1^* = (a_{per}^*)^2 \left\langle d\bar{P}_1(s), f(s)\right\rangle. \tag{4.144}$$

Likewise, we compute from (4.72), for $k \in [\![-\frac{N-1}{2}, \frac{N-1}{2}]\!]\backslash\{0\}$,

$$\begin{aligned} a_2^{s,t,0,k}(\frac{d}{dx}w^{2,s,t,0,k,N}+1) &= N\left(\int_{-\frac{N}{2}}^{\frac{N}{2}} \frac{1}{a_{per}+s\mathbb{1}_{[-\frac{1}{2},\frac{1}{2}]}c_{per}+t\mathbb{1}_{[k-\frac{1}{2},k+\frac{1}{2}]}c_{per}}\right)^{-1} \\ &= N\left(N(a_{per}^*)^{-1} - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{sc_{per}}{a_{per}(a_{per}+sc_{per})} - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{tc_{per}}{a_{per}(a_{per}+tc_{per})}\right)^{-1} \\ &= N\left(N(a_{per}^*)^{-1} - f(s) - f(t)\right)^{-1}. \end{aligned}$$

Then

$$a_2^{s,t,0,k}(\frac{d}{dx}w^{2,s,t,0,k,N}+1) = a_{per}^* + \frac{(a_{per}^*)^2}{N}(f(s)+f(t)) + \frac{(a_{per}^*)^3}{N^2}(f(s)+f(t))^2 + o(N^{-2}).$$

Notice that this expression is independent of k (and so of the distance between the two defects), so that $a_2^{*,N}$ defined by (4.73) here reads

$$\begin{aligned} a_2^{*,N} =& \frac{N(N-1)}{2}\left\langle d\bar{P}_1(s)d\bar{P}_1(t), a_{per}^* + \frac{(a_{per}^*)^2}{N}(f(s)+f(t)) + \frac{(a_{per}^*)^3}{N^2}(f(s)+f(t))^2\right\rangle \\ &+ N\left\langle d\bar{P}_2(s), a_{per}^* + \frac{(a_{per}^*)^2}{N}f(s)\right\rangle + o(1). \end{aligned} \tag{4.145}$$

Since we know from Lemma 4.6 that $\left\langle d\bar{P}_1(s), 1\right\rangle = 0$ and $\left\langle d\bar{P}_2(s), 1\right\rangle = 0$, (4.145) reduces to

$$a_2^{*,N} = (a_{per}^*)^3 \left\langle d\bar{P}_1(s), f(s)\right\rangle^2 + (a_{per}^*)^2 \left\langle d\bar{P}_2(s), f(s)\right\rangle + o(1).$$

Thus

$$a_2^{*,N} \underset{N\to\infty}{\to} \bar{a}_2^* = (a_{per}^*)^3 \left\langle d\bar{P}_1(s), f(s)\right\rangle^2 + (a_{per}^*)^2 \left\langle d\bar{P}_2(s), f(s)\right\rangle. \tag{4.146}$$

Finally, since $f(s) = \displaystyle\int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}} - \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}}$, and $\left\langle d\bar{P}_1(s), 1\right\rangle = \left\langle d\bar{P}_2(s), 1\right\rangle = 0$, we have

$$\left\langle d\bar{P}_1(s), f(s)\right\rangle = -\left\langle d\bar{P}_1(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}}\right\rangle, \tag{4.147}$$

and

$$\left\langle d\bar{P}_2(s), f(s)\right\rangle = -\left\langle d\bar{P}_2(s), \int_{-\frac{1}{2}}^{\frac{1}{2}} \frac{1}{a_{per}+sc_{per}}\right\rangle. \tag{4.148}$$

In view of (4.142), (4.144), (4.146), (4.147) and (4.148), we have proved

$$a_\eta^* = a_{per}^* + \eta\bar{a}_1^* + \eta^2\bar{a}_2^* + o(\eta^2).$$

□

4.5.4 A proof of the approach of Section 4.3 in a specific setting

The purpose of this final section is to prove that the formal approach of Section 4.3 is rigorous in a setting related to that of Corollary 4.4.

More precisely, we assume that the random field b_η satisfies the assumptions of Corollary 4.4. These assumptions do not imply that the image measure dP_η satisfies assumption (4.48) which is at the heart of the approach of Section 4.3, so that we have to impose that dP_η additionally satisfies (4.48). The following preliminary result then gives the necessary form of the expansion of the image measure dP_η.

Lemma 4.21. *Assume that b_η satisfies*

$$b_\eta(x,\omega) = \sum_{k\in\mathbb{Z}^d} \mathbb{1}_{Q+k}(x) B_\eta^k(\omega), \tag{4.149}$$

where the B_η^k are i.i.d random variables, the distribution of which is given by a "mother variable" B_η satisfying

$$\forall \eta > 0, \|B_\eta\|_{L^\infty(\Omega)} \leq M, \tag{4.150}$$

$$B_\eta = \eta \bar{B}_0 + \eta^2 \bar{R}_0 + o(\eta^2) \quad \text{weakly in } L^2(\Omega). \tag{4.151}$$

Assume further that the image measure dP_η of B_η satisfies (4.48). Then

$$dP_\eta = \delta_0 - \eta \mathbb{E}(\bar{B}_0)\delta_0' + \frac{\eta^2}{2}\mathbb{E}(\bar{B}_0^2)\delta_0'' - \eta^2 \mathbb{E}(\bar{R}_0)\delta_0' + o(\eta^2) \text{ in } \mathcal{E}'(\mathbb{R}). \tag{4.152}$$

Proof. Firstly, notice that $\frac{B_\eta}{\eta}$ converges strongly to $\bar{B}_0$ in $L^2(\Omega)$ because of (4.151). Now consider $\varphi \in \mathcal{D}(\mathbb{R})$. We have on the one hand

$$\mathbb{E}\left(\frac{B_\eta^2}{\eta^2}\varphi(B_\eta)\right) \to \mathbb{E}\left(\bar{B}_0^2\right)\varphi(0),$$

and on the other hand

$$\mathbb{E}\left(B_\eta^2 \varphi(B_\eta)\right) = \eta\langle s^2 d\bar{P}_1, \varphi\rangle + \eta^2 \langle s^2 d\bar{P}_2, \varphi\rangle + o(\eta^2).$$

Thus $s^2 d\bar{P}_1 = 0$ and $s^2 d\bar{P}_2 = \mathbb{E}\left(\bar{B}_0^2\right)\delta_0$ in $\mathcal{D}'(\mathbb{R})$. It is then well known that there exist γ_1, κ_1, γ_2, κ_2 in $\mathbb{R}$ such that

$$d\bar{P}_1 = \gamma_1\delta_0 + \kappa_1\delta_0' \quad \text{and} \quad d\bar{P}_2 = \gamma_2\delta_0 + \kappa_2\delta_0' + \frac{\mathbb{E}\left(\bar{B}_0^2\right)}{2}\delta_0''.$$

Lemma 4.6 implies $\gamma_1 = \gamma_2 = 0$. Then, we have

$$\mathbb{E}(B_\eta) = \eta\mathbb{E}(\bar{B}_0) + \eta^2\mathbb{E}(\bar{R}_0) + o(\eta^2)$$

and also

$$\mathbb{E}(B_\eta) = \eta\langle s d\bar{P}_1, 1\rangle + \eta^2\langle s d\bar{P}_2, 1\rangle + o(\eta^2).$$

Thus $\langle s d\bar{P}_1, 1\rangle = \mathbb{E}(\bar{B}_0)$ and $\langle s d\bar{P}_2, 1\rangle = \mathbb{E}\left(\bar{R}_0\right)$, from which we deduce $\kappa_1 = -\mathbb{E}(\bar{B}_0)$ and $\kappa_2 = -\mathbb{E}(\bar{R}_0)$.

□

Theorem 4.2 and Corollary 4.4 rigorously yield the second-order expansion

$$A^*_\eta = A^*_{per} + \eta \tilde{A}^*_1 + \eta^2 \tilde{A}^*_2 + o(\eta^2)$$

with $\tilde{A}^*_1$ and $\tilde{A}^*_2$ respectively defined by (4.17) and (4.26).

On the other hand, using (4.152), Section 4.3 yields the formal expansion

$$A^*_\eta = A^*_{per} + \eta \bar{A}^*_1 + \eta^2 \bar{A}^*_2 + o(\eta^2),$$

where $\bar{A}^*_1$ is the limit of the sequence $A^{*,N}_1$ defined by (4.69) or equivalently by (4.74), and $\bar{A}^*_2$ the limit of the sequence $A^{*,N}_2$ defined by (4.73).

The rest of this section is devoted to verifying that $\bar{A}^*_1$ coincides with $\tilde{A}^*_1$ and $\bar{A}^*_2$ coincides with $\tilde{A}^*_2$ in the specific setting of Lemma 4.21.

4.5.4.1 First-order term

Using (4.152), (4.74) reads

$$A^{*,N}_1 e_i \cdot e_j = -\mathbb{E}(\bar{B}_0) \left\langle \delta_0'(s), \int_Q s C_{per}(\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle,$$

and we compute

$$A^{*,N}_1 = \mathbb{E}(\bar{B}_0) \int_Q C_{per}(\nabla w_i^{1,0,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0).$$

Setting $s = 0$ in (4.68), it is clear that $w_i^{1,0,0,N}$ is equal to the periodic corrector w_i^0. Then

$$A^{*,N}_1 = \mathbb{E}(\bar{B}_0) \int_Q C_{per}(\nabla w_i^0 + e_i) \cdot (e_j + \nabla \tilde{w}_j^0). \tag{4.153}$$

Clearly $A^{*,N}_1$ does not depend on N and its limit is then

$$\bar{A}^*_1 = \mathbb{E}(\bar{B}_0) \int_Q C_{per}(\nabla w_i^0 + e_i) \cdot (e_j + \nabla \tilde{w}_j^0). \tag{4.154}$$

We recognize in the right-hand side of (4.154) the first-order coefficient in (4.21), which we know from Remark 4.2 is equivalent to (4.17). Theorem 4.2 therefore shows that the first-order expansion

$$A^*_\eta = A^*_{per} + \eta \bar{A}^*_1 + o(\eta)$$

is correct with the values of the coefficients given by our formal approach of Section 4.3.

We now proceed similarly with the second-order coefficient.

4.5.4.2 Second-order term

Using the adjoint cell problems (4.20) in (4.73) as in the proof of Proposition 4.7, let us first rewrite

$$\begin{aligned} A_2^{*,N} e_i \cdot e_j = & \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_Q sC_{per} \nabla w_i^{2,s,t,0,k,N} \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ & + \left\langle d\bar{P}_2(s), \int_Q sC_{per} (\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned} \tag{4.155}$$

Inserting (4.152) in (4.155), we start by focusing on

$$\begin{aligned} & \left\langle d\bar{P}_2(s), \int_Q sC_{per} (\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle = \\ & \frac{1}{2} \left(\mathbb{E}(\bar{B}_0) \right)^2 \left\langle \delta_0''(s), \int_Q sC_{per} (\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ & - \mathbb{E}(\bar{R}_0) \left\langle \delta_0'(s), \int_Q sC_{per} (\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned}$$

Denoting by $\partial_s w_i^{1,0,0,N}$, the first derivative of $w_i^{1,s,0,N}$ evaluated at $s = 0$, we compute

$$\begin{aligned} & \left\langle d\bar{P}_2(s), \int_Q sC_{per} (\nabla w_i^{1,s,0,N} + e_i) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \\ & = \mathbb{E}(\bar{B}_0^2) \int_Q C_{per} \nabla \partial_s w_i^{1,0,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j) + \mathbb{E}(\bar{R}_0) \int_Q C_{per} (\nabla w_i^{1,0,0,N} + e_i) \cdot (\nabla \tilde{w}_j^0 + e_j) \\ & = \mathbb{E}(\bar{B}_0^2) \int_Q C_{per} \nabla \partial_s w_i^{1,0,0,N} \cdot (\nabla \tilde{w}_j^0 + e_j) + \mathbb{E}(\bar{R}_0) \int_Q C_{per} (\nabla w_i^0 + e_i) \cdot (\nabla \tilde{w}_j^0 + e_j). \end{aligned}$$

It follows from (4.68) that $\partial_s w_i^{1,0,0,N}$ solves

$$\begin{cases} -\operatorname{div}(A_{per} \nabla \partial_s w_i^{1,0,0,N}) = \operatorname{div}\left(\mathbb{1}_Q C_{per} (\nabla w_i^0 + e_i) \right) & \text{in } I_N, \\ \partial_s w_i^{1,0,0,N} \ (N\mathbb{Z})^d - \text{periodic}. \end{cases} \tag{4.156}$$

Applying Lemma 4.15 to (4.156), we deduce that $\nabla \partial_s w_i^{1,0,0,N}$ converges in $L^2(Q)$, when $N \to \infty$, to ∇t_i defined by (4.33) in Corollary 4.4. Consequently,

$$\begin{aligned} & \left\langle d\bar{P}_2(s), \int_Q sC_{per} \left(\nabla w_i^{1,s,0,N} + e_i \right) \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \underset{N \to \infty}{\to} \\ & \mathbb{E}(\bar{B}_0^2) \int_Q C_{per} \nabla t_i \cdot (\nabla \tilde{w}_j^0 + e_j) + \mathbb{E}(\bar{R}_0) \int_Q C_{per} \left(\nabla w_i^0 + e_i \right) \cdot (\nabla \tilde{w}_j^0 + e_j). \end{aligned} \tag{4.157}$$

Next, we address

$$\begin{aligned} & \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle d\bar{P}_1(s) d\bar{P}_1(t), \int_Q sC_{per} \nabla w_i^{2,s,t,0,k,N} \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle \quad = \\ & \mathbb{E}(\bar{B}_0)^2 \sum_{k \in \mathcal{T}_N \setminus \{0\}} \left\langle \delta_0'(s) \delta_0'(t), \int_Q sC_{per} \nabla w_i^{2,s,t,0,k,N} \cdot (e_j + \nabla \tilde{w}_j^0) \right\rangle. \end{aligned}$$

Denoting by $\partial_t w_i^{2,0,0,0,k,N}$ the first derivative of $w_i^{2,s,t,0,k,N}$ with respect to t evaluated at $s=t=0$, we have

$$\sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q sC_{per}\nabla w_i^{2,s,t,0,k,N}\cdot(e_j+\nabla\tilde{w}_j^0)\right\rangle \;=\; \mathbb{E}(\bar{B}_0)^2\sum_{k\in\mathcal{T}_N\setminus\{0\}}\int_Q C_{per}\nabla\partial_t w_i^{2,0,0,0,k,N}\cdot(e_j+\nabla\tilde{w}_j^0). \tag{4.158}$$

It follows from (4.72) that $\partial_t w_i^{2,0,0,0,k,N}$ solves

$$\begin{cases} -\operatorname{div}(A_{per}\nabla\partial_t w_i^{2,0,0,0,k,N}) = \operatorname{div}\left(\mathbb{1}_{Q+k}C_{per}(\nabla w_i^0+e_i)\right) & \text{in } I_N,\\ \partial_t w_i^{2,0,0,0,k,N}\ (N\mathbb{Z})^d-\text{periodic.}\end{cases} \tag{4.159}$$

Defining $d_i^N=\sum_{k\in\mathcal{T}_N}\partial_t w_i^{2,0,0,0,k,N}$, it is easy to see that d_i^N is a $\mathbb{Z}^d$-periodic function that solves

$$\begin{cases} -\operatorname{div}(A_{per}\nabla d_i^N) = \operatorname{div}\left(C_{per}(\nabla w_i^0+e_i)\right) & \text{in } Q,\\ d_i^N\ \mathbb{Z}^d-\text{periodic.}\end{cases} \tag{4.160}$$

Since problem (4.160) has a unique solution up to an additive constant, $\nabla d_i^N=\nabla s_i$ where s_i is defined by (4.34) in Corollary 4.4.

Finally, comparing (4.156) to (4.159) for $k=0$, we find that $\nabla\partial_t w_i^{2,0,0,0,0,N}$ is equal to $\nabla\partial_s w_i^{1,0,0,N}$ and then also converges in $L^2(Q)$ to ∇t_i when $N\to\infty$.

Then, starting from (4.158),

$$\begin{aligned}
&\sum_{k\in\mathcal{T}_N\setminus\{0\}} \left\langle d\bar{P}_1(s)d\bar{P}_1(t), \int_Q sC_{per}\nabla w_i^{2,s,t,0,k,N}\cdot(e_j+\nabla\tilde{w}_j^0)\right\rangle\\
&= \left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\sum_{k\in\mathcal{T}_N}\nabla\partial_t w_i^{2,0,0,0,k,N}\cdot(e_j+\nabla\tilde{w}_j^0)\\
&\quad -\left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\nabla\partial_t w_i^{2,0,0,0,0,N}\cdot(e_j+\nabla\tilde{w}_j^0)\\
&= \left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\nabla s_i\cdot(e_j+\nabla\tilde{w}_j^0)-\left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\nabla\partial_t w_i^{2,0,0,0,0,N}\cdot(e_j+\nabla\tilde{w}_j^0)\\
&\underset{N\to\infty}{\to} \left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\nabla s_i\cdot(e_j+\nabla\tilde{w}_j^0)-\left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q C_{per}\nabla t_i\cdot(e_j+\nabla\tilde{w}_j^0).
\end{aligned} \tag{4.161}$$

It entails from (4.155), (4.157) and (4.161) that $A_2^{*,N}$ converges to a limit $\bar{A}_2^*$ defined by

$$\begin{aligned}
\bar{A}_2^* e_i\cdot e_j \;=\; &\mathbb{E}(\bar{R}_0)\int_Q C_{per}\left(\nabla w_i^0+e_i\right)\cdot(\nabla\tilde{w}_j^0+e_j)+Var(\bar{B}_0)\int_Q C_{per}\nabla t_i\cdot(\nabla\tilde{w}_j^0+e_j)\\
&+\left(\mathbb{E}(\bar{B}_0)\right)^2\int_Q\nabla s_i\cdot(e_j+\nabla\tilde{w}_j^0).
\end{aligned}$$

The matrix $\bar{A}_2^*$ obtained is equal to the second-order term given by (4.32) in Corollary 4.4 since we deal with independent random variables in each cell of $\mathbb{Z}^d$. Thus the second-order expansion

$$A_\eta^* = A_{per}^* + \eta \bar{A}_1^* + \eta^2 \bar{A}_2^* + o(\eta^2)$$

derived from the formal approach of Section 4.3 is correct in this specific setting.

Chapitre 5

Boundary layers in periodic homogenization

Sommaire

5.1 Introduction 167
5.2 General setting and notation 169
5.2.1 Stationary setting 170
5.2.2 Transient setting 172
5.3 Boundary layers in the homogenization of elliptic equations . . . 173
5.3.1 Classical results for Dirichlet boundary conditions 173
5.3.2 Neumann boundary conditions 178
5.4 Boundary layers for parabolic equations 188
5.4.1 Need for an "initial layer" 189
5.4.2 A theoretical boundary+initial layer 191
5.4.3 Initial layer in an "unbounded" domain 193
5.4.4 General case 198
5.4.5 One-dimensional toy model 206
5.5 Appendix : two parabolic regularity results 211

5.1 Introduction

We are interested in this chapter in the issue of boundary layers for elliptic and above all parabolic periodic homogenization problems.

Generally speaking, the aim of periodic homogenization is to address rapidly oscillating partial differential equations of the form

$$-\mathrm{div}\left(A(\frac{x}{\varepsilon})\nabla u_\varepsilon\right) = f \ \text{ in } \Omega \subset \mathbb{R}^d, \tag{5.1}$$

where A is a $\mathbb{Z}^d$-periodic matrix field. The small parameter ε represents the lengthscale of the heterogeneities in the domain Ω.

From a numerical point of view, solving directly (5.1) is involved. For instance, a standard finite element approach would require the use of a mesh of size at least as fine as ε, and the resulting computational cost would be very high. The homogenization process

consists in taking the limit $\varepsilon \to 0$ in equation (5.1) in order to obtain an "averaged" or homogenized field u_0, close to u_ε in some sense, and easier to compute.

In the elliptic periodic setting, homogenization is classically grounded on the assumption, called an "Ansatz", that u_ε can be written as the two-scale expansion

$$u_\varepsilon(x) = u_0(x) + \varepsilon u_1(x, \frac{x}{\varepsilon}) + \varepsilon^2 u_2(x, \frac{x}{\varepsilon}) + ... \tag{5.2}$$

In (5.2), u_0 is the homogenized field introduced above, and the functions u_1 and u_2 and more generally u_k for $k \in \mathbb{N}^*$ are called the correctors for they allow to refine the approximation of u_ε by u_0. The correctors are periodic with respect to the so-called fast variable $y = \frac{x}{\varepsilon}$. Substituting (5.2) in (5.1), we obtain the equations satisfied by u_0 and the correctors [14]. These manipulations, which are a priori only formal, can be justified by several means [1, 14, 81].

The issue of boundary layers in homogenization originates from the following well-known fact: if the domain Ω in (5.1) is not $\mathbb{R}^d$, i.e if there are boundaries, then the Ansatz (5.2) is not correct near $\partial\Omega$. Intuitively, this comes from the fact that as such, this Ansatz is written independently of any boundary condition on $\partial\Omega$, so that it generally violates any boundary condition that we may impose on $\partial\Omega$, and can thus only hold in the interior of the domain.

Concretely, this implies that the $H^1(\Omega)$-norm of the error $u_\varepsilon - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon})$ is not of order ε, contrary to what might be expected from (5.2) since the $H^1(\Omega)$-norm of the following term $\varepsilon^2 u_2(x, \frac{x}{\varepsilon})$ is of order ε. This statement is made precise in Theorem 5.1 in Section 5.3.1, excerpted from [14], in the case of homogeneous Dirichlet boundary conditions on $\partial\Omega$.

One of the purposes of this chapter is then to propose, in some classical homogenization settings, boundary layers to improve the approximation of u_ε by $u_0 + \varepsilon u_1$. We will not go further than order one in ε, the difficulties being the same at higher orders. Our ultimate goal is to address the parabolic setting, that, contrary to the elliptic one, is not well documented in the literature (we are only aware of [68] in an unbounded domain).

As will appear clear to the reader by Section 5.3.1, it is straightforward to find and add terms in (5.2) that yield fine error estimates. However it is imperative that these terms be computationally tractable, otherwise the homogenization approach would not be more advantageous than directly solving the initial oscillating problem (5.1). So as to fulfill this practical requirement, a simplifying assumption will be to consider only rectangular domains, as in [2]. Though this assumption is restrictive, and can be alleviated at the expense of much greater complexity (general polygonal domains are studied in [37], and curved domains are considered in [64] in the case of a layered medium), it will allow us not to address all difficulties simultaneously, and in particular, in the parabolic setting, to focus merely on the new issues coming from the introduction of the time variable while relying on a fully understood elliptic background.

The approach that we follow in this chapter is two-fold: we first deal with the elliptic case, and then use the knowledge acquired to tackle the parabolic case. In Section 5.2, we introduce general notation and recall classical facts about periodic homogenization. Section 5.3 focuses on the elliptic setting. In Section 5.3.1, we cite some well-known results (taken from [14], [2], [9], [62] and [64]) concerning boundary layers in the case of Dirichlet boundary conditions. Section 5.3.2 consists of an adaptation of these results to the case of Neumann boundary conditions: although this adaptation is rather straightforward, it is, as far as we know, not written anywhere in the literature. The parabolic setting is the object of Section 5.4. After two introductory sections, Section 5.4.3, that is greatly inspired by [68], addresses the specific case when the domain Ω is equipped with periodic boundary conditions, which allows us not to consider boundary layers and to concentrate solely on the new boundary $t = 0$ and the subsequent need for what we call an initial layer. Finally, Section 5.4.4 aims at discussing the general parabolic setting and notably at understanding the interaction between the boundary layers and the initial layer.

We emphasize that the results of Section 5.4.4 in the general parabolic setting are unfortunately not conclusive yet. Although we are able to propose a candidate for a complete "boundary+initial" layer, it relies on regularity assumptions that we could not prove to hold. Nonetheless, we reckon that it is a first step in the search for tractable layers in parabolic homogenization.

One last word is of order here. All the results found in the homogenization literature dealing with boundary layers rely on assumptions of smoothness of the homogenized solution u_0 and of the matrix field A. However, one observes, reading the many articles aiming at alleviating those assumptions, that the boundary layer terms proposed never depend on the smoothness of u_0 and A; only the technical complexity of the proofs actually depends on it. Since our aim is to find relevant boundary and initial layers, we choose here not to focus on regularity issues, and we will always assume that u_0 and A are smooth enough for our purposes without entering into details.

Throughout this chapter, C denotes a generic constant which does not depend on ε, i.e on the size of the heterogeneities.

5.2 General setting and notation

In the sequel, A denotes a $\mathbb{Z}^d$-periodic tensor field from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$:

$$\forall k \in \mathbb{Z}^d, \quad A(x+k) = A(x) \text{ almost everywhere in } x \in \mathbb{R}^d.$$

We assume that $A \in L^\infty(\mathbb{R}^d, \mathbb{R}^{d\times d})$ and that there exist $\lambda > 0$ and $\Lambda > 0$ such that

$$\forall \xi \in \mathbb{R}^d, \text{ a.e in } x \in \mathbb{R}^d, \ \lambda|\xi|^2 \leq A(x)\xi \cdot \xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|. \tag{5.3}$$

We consider a material occupying a bounded regular open set $\Omega \subset \mathbb{R}^d$. The properties of the material are given by the tensor field $A_\varepsilon(x) = A\left(\frac{x}{\varepsilon}\right)$. The material is subject to a force $f \in L^2(\Omega)$.

For $i \in [\![1,d]\!]$, we denote by w_i the i-th cell solution, that solves the cell problem

$$\begin{cases} -\operatorname{div}\left(A(\nabla w_i + e_i)\right) = 0 \quad \text{in } Q, \\ w_i \ \ \mathbb{Z}^d - \text{periodic}, \end{cases} \tag{5.4}$$

where Q is the unit cell $[0,1]^d$ and e_i the i-th canonical vector of $\mathbb{R}^d$. Note that for every i, w_i is uniquely defined up to an additive constant.

The homogenized tensor A^* is then given by

$$\forall (i,j) \in [\![1,d]\!]^2, \ \ A^*_{ji} = \int_Q A(\nabla w_i + e_i) \cdot e_j. \tag{5.5}$$

We will consider both stationary and transient settings for different boundary conditions. Remark that in this chapter, "stationary" does not have the same meaning as in the previous chapters at all: it is here synonymous with steady-state.

5.2.1 Stationary setting

The stationary setting is the elliptic equation

$$\begin{cases} -\operatorname{div}\left(A_\varepsilon \nabla u_\varepsilon\right) = f \quad \text{in } \Omega, \\ \text{boundary condition} \quad \text{on } \partial\Omega. \end{cases} \tag{5.6}$$

It is well known that the boundary condition has no influence on the homogenization process away from the boundary (i.e in the core of the material). We will thus specify it later when dealing with boundary layers.

It is classical to look for u_ε in Ω as the following two-scale expansion, called Ansatz,

$$u_\varepsilon(x) = u_0(x) + \varepsilon u_1(x, \frac{x}{\varepsilon}) + \varepsilon^2 u_2(x, \frac{x}{\varepsilon}) + ... \tag{5.7}$$

where each function $u_k(x,y)$ for $k \in \mathbb{N}^*$ is $\mathbb{Z}^d$-periodic with respect to the so-called fast variable $y = \frac{x}{\varepsilon}$.

The function u_0 depends only on the slow variable x and is called the homogenized solution. The function u_k for $k \in \mathbb{N}^*$ is called the k-th corrector.

Inserting (5.7) in (5.6), and identifying different powers of ε, we obtain a cascade of equations solved by the functions u_k.

To detail these equations, we use for convenience the notation introduced in [14] and [2], and define the operator L_ε by

$$L_\varepsilon \phi = -\operatorname{div}(A_\varepsilon \nabla \phi). \tag{5.8}$$

Then we may write

$$L_\varepsilon = \frac{1}{\varepsilon^2} L_0 + \frac{1}{\varepsilon} L_1 + L_2, \tag{5.9}$$

where, in terms of the fast variable y and the slow variable x, we have

$$\begin{aligned} L_0 &= -\mathrm{div}_y(A(y)\nabla_y), && (5.10)\\ L_1 &= -\mathrm{div}_y(A(y)\nabla_x) - \mathrm{div}_x(A(y)\nabla_y), && (5.11)\\ L_2 &= -\mathrm{div}_x(A(y)\nabla_x). && (5.12) \end{aligned}$$

Taking the variables x and y as independent, equation (5.6) becomes equivalent to the system

$$\begin{aligned} &L_0u_0 = 0,\\ &L_1u_0 + L_0u_1 = 0,\\ &L_2u_0 + L_1u_1 + L_0u_2 = 0,\\ &L_2u_1 + L_1u_2 + L_0u_3 = 0,\\ &..... \end{aligned} \qquad (5.13)$$

This heuristic computation can be rigorously justified [1, 14, 81].

The homogenized field u_0 solves

$$\begin{cases} -\mathrm{div}\,(A^*\nabla u_0) = f & \text{in } \Omega,\\ \text{boundary condition} & \text{on } \partial\Omega. \end{cases} \qquad (5.14)$$

The first corrector u_1 is given by

$$u_1(x,y) = \sum_{i=1}^{d} w_i(y)\frac{\partial u_0}{\partial x_i}(x) + \tilde{u}_1(x). \qquad (5.15)$$

It is defined up to a function $\tilde{u}_1$ of x that is determined by solving the fourth equation of system (5.13), and that will not play a role in this chapter since we will not go further than order one in ε, hence we can consider it to be 0.

The second corrector u_2 writes

$$u_2(x,y) = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}(y)\frac{\partial^2 u_0}{\partial x_i \partial x_j}(x) + \sum_{i=1}^{d} w_i(y)\frac{\partial \tilde{u}_1}{\partial x_i}(x) + \tilde{u}_2(x), \qquad (5.16)$$

where for $(i,j) \in [\![1,d]\!]^2$, w_{ij} solves another cell problem

$$\begin{cases} -\mathrm{div}\,(A\nabla w_{ij}) = b_{ij} - \displaystyle\int_Q b_{ij} & \text{in } Q,\\ w_{ij} \;\; \mathbb{Z}^d - \text{periodic}, \end{cases} \qquad (5.17)$$

with

$$b_{ij} = A_{ij} + A\nabla w_j \cdot e_i + \mathrm{div}(Ae_i w_j).$$

It is defined up to a function $\tilde{u}_2$ of x that only plays a role in the computation of higher order correctors, and that we can consequently choose to be 0 in the sequel.

5.2.2 Transient setting

The transient setting is the parabolic equation posed in $\Omega \times (0,T)$ for some $T > 0$:

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \quad \text{in } \Omega \times (0,T), \\ \text{boundary condition} \quad \text{on } \partial\Omega \times (0,T), \\ \text{initial condition} \quad \text{in } \Omega. \end{cases} \tag{5.18}$$

Once again the boundary and initial conditions do not have to be specified if we are only interested in finding the homogenized equation corresponding to the first line of (5.18). We will make them precise when addressing the issue of boundary and initial layers in Section 5.4.

As in the stationary case, we write an Ansatz

$$u_\varepsilon(x,t) = u_0(x,t) + \varepsilon u_1(x, \frac{x}{\varepsilon}, t) + \varepsilon^2 u_2(x, \frac{x}{\varepsilon}, t) + ... \tag{5.19}$$

where each function $u_k(x,y,t)$ for $k \in \mathbb{N}^*$ is $\mathbb{Z}^d$-periodic with respect to the fast space variable y.

For later use we define the fast time variable $\tau = \frac{t}{\varepsilon^2}$. Note that the so-called parabolic scaling $\frac{1}{\varepsilon^2}$ for τ, as compared to the scaling $\frac{1}{\varepsilon}$ for y, is intuitive since there is one derivative in time and two derivatives in space in (5.18).

We also redefine the operator L_ε introduced in the stationary setting (5.8) by

$$L_\varepsilon \phi = \frac{\partial \phi}{\partial t} - \operatorname{div}(A_\varepsilon \nabla \phi),$$

and again write a decomposition

$$L_\varepsilon = \frac{1}{\varepsilon^2} L_0 + \frac{1}{\varepsilon} L_1 + L_2, \tag{5.20}$$

with

$$\begin{aligned} L_0 &= \frac{\partial}{\partial \tau} - \operatorname{div}_y(A(y)\nabla_y), \\ L_1 &= -\operatorname{div}_y(A(y)\nabla_x) - \operatorname{div}_x(A(y)\nabla_y), \\ L_2 &= \frac{\partial}{\partial t} - \operatorname{div}_x(A(y)\nabla_x). \end{aligned} \tag{5.21}$$

Inserting (5.19) in (5.18), we find heuristically that the functions u_0, u_1 and u_2 satisfy (5.13) with L_0, L_1 and L_2 given by (5.21). This can be justified rigorously [14].

It is then well known (see [14]) that the homogenized solution u_0 solves

$$\begin{cases} \dfrac{\partial u_0}{\partial t} - \operatorname{div}(A^* \nabla u_0) = f \quad \text{in } \Omega \times (0,T), \\ \text{boundary condition} \quad \text{on } \partial\Omega \times (0,T), \\ \text{initial condition} \quad \text{in } \Omega \times \{0\}. \end{cases} \tag{5.22}$$

The first and second-order correctors u_1 and u_2 are given by

$$u_1(x,y,t) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x,t) w_i(y) + \tilde{u}_1(x,t), \tag{5.23}$$

and

$$u_2(x,y,t) = \sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}(y) \frac{\partial^2 u_0}{\partial x_i \partial x_j}(x,t) + \sum_{i=1}^{d} w_i(y) \frac{\partial \tilde{u}_1}{\partial x_i}(x,t) + \tilde{u}_2(x,t). \tag{5.24}$$

As in the stationary setting, the functions $\tilde{u}_1$ and $\tilde{u}_2$ only come into play if we are interested in higher order correctors, and may be taken as 0 in what follows.

Note that the fast time variable τ does not play a role in the expansion (5.19). We will use it to build the initial layer in Sections 5.4.3 and 5.4.4.

Until now, boundary and initial conditions have not been taken into account. We have only looked at homogenization far from the boundary and for "large" times. In the next section, we concentrate on the issue of boundary conditions in the elliptic setting, while Section 5.4 addresses boundary and initial conditions in the parabolic case.

5.3 Boundary layers in the homogenization of elliptic equations

The aim of this section is two-fold. We first recall some well-known results in the case of Dirichlet boundary conditions, and then adapt them to address Neumann boundary conditions.

We will often for simplicity (and when there is no possible confusion) write u_1 and u_2 instead of using the full notation $u_1(x, \frac{x}{\varepsilon})$ and $u_2(x, \frac{x}{\varepsilon})$. The same holds for all functions depending on the slow variable x and the fast variable $\frac{x}{\varepsilon}$ once they have been defined.

5.3.1 Classical results for Dirichlet boundary conditions

We consider in a first step the case of the elliptic equation (5.6) with homogeneous Dirichlet boundary conditions, which is well documented in the homogenization literature. Here u_ε and u_0 solve

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \quad \text{in } \Omega, \\ u_\varepsilon = 0 \ \text{ on } \partial\Omega, \end{cases} \tag{5.25}$$

and

$$\begin{cases} -\operatorname{div}(A^* \nabla u_0) = f \quad \text{in } \Omega, \\ u_0 = 0 \ \text{ on } \partial\Omega, \end{cases} \tag{5.26}$$

respectively, where we recall that Ω is a bounded regular open set in $\mathbb{R}^d$.

All the results presented in this section can be readily deduced from (sometimes more complicated) results in [2], [9], [14], [62] and [64]. We shall not give here the proofs for similar proofs for Neumann boundary conditions will be presented in the next section.

The starting point in our study of boundary layers is the following well-known theorem, the proof of which can be found in [14].

Theorem 5.1. *Assume that $u_0 \in W^{2,\infty}(\Omega)$. Then*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) \right\|_{H^1(\Omega)} \leq C\sqrt{\varepsilon}, \tag{5.27}$$

where u_1 is given by (5.15).

The rate $\sqrt{\varepsilon}$ in (5.27) is somewhat surprising, for Ansatz (5.2) hints at a remainder $\varepsilon^2 u_2(x, \frac{x}{\varepsilon})$ of order ε in $H^1(\Omega)$. However, as mentioned in the introduction to this chapter, Ansatz (5.2) is not correct near $\partial\Omega$ because the correctors do not vanish on the boundary and therefore violate the boundary condition in (5.25). Therefore we introduce the new Ansatz

$$u_\varepsilon(x) = u_0(x) + \sum_{k=1}^{\infty} \varepsilon^k \left(u_k\left(x, \frac{x}{\varepsilon}\right) + u_k^{bl,\varepsilon}(x) \right), \tag{5.28}$$

where the $u_k^{bl,\varepsilon}$ are designed to guarantee that the coefficients of all powers of ε in (5.28) satisfy the homogeneous Dirichlet boundary conditions, hence

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_k^{bl,\varepsilon}) = 0 \ \text{ in } \Omega, \\ u_k^{bl,\varepsilon}(x) = -u_k\left(x, \frac{x}{\varepsilon}\right) \ \text{ on } \partial\Omega. \end{cases} \tag{5.29}$$

We will only address the first boundary layer $u_1^{bl,\varepsilon}$ that compensates for the first corrector on $\partial\Omega$, the computations for higher orders being identical. We have the following improvement over (5.27), found for instance in [62]:

Theorem 5.2. *Assume that u_0 is smooth. Then*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) - \varepsilon u_1^{bl,\varepsilon}(x) \right\|_{H_0^1(\Omega)} \leq C\varepsilon.$$

For completeness, let us stress that the boundary layer does not play a role far from the boundary, as shown by the following result [9, 2]:

Theorem 5.3. *Assume that u_0 is smooth. Then, for any open set $\omega \subset\subset \Omega$ (i.e compactly embedded in Ω), there exists a constant C, depending on ω but not on ε, such that*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) \right\|_{H^1(\omega)} \leq C\varepsilon.$$

Remark 5.1. *The link between Theorems 5.1, 5.2 and 5.3 can be explained as follows: the H^1-norm of $u_1^{bl,\varepsilon}$ blows up like $1/\sqrt{\varepsilon}$ in Ω (this estimate is optimal, see [2]), whereas it is bounded independently of ε in any open subset $\omega \subset\subset \Omega$.*

The introduction of the boundary layer $u_1^{bl,\varepsilon}$ allows to recover a precision of order ε in the whole domain Ω. However it is clear from (5.29) that the computation of $u_1^{bl,\varepsilon}$ is as intricate as that of u_ε, so that it is not of practical interest. The challenge is then to propose tractable boundary layers ensuring a precision of order $\mathcal{O}(\varepsilon)$. This has been done when the domain is a half-space, the boundary of which intersects the axes of periodicity in an angle with rational slope [10, 11, 15, 45, 56], when it is a half strip satisfying the same property [66], when the domain is rectangular [2], for a curved domain in the specific case of a laminate [64], and recently in the case of general polygonal domains [37]. In the sequel we consider for simplicity a rectangular domain as in [2] and we use the same notation.

We assume in the rest of this section that $\Omega = (0,1)^d$, and that the sequence of ε satisfies $\frac{1}{\varepsilon} \in \mathbb{N}$, so that Ω always contains an integer number of cells. We denote by $\Gamma_1 = (0,1)^{d-1} \times \{0\}$, $\Gamma_2 = (0,1)^{d-1} \times \{1\}$, $\Gamma_\# = \partial\Omega \backslash \Gamma_1 \cup \Gamma_2$, $x' = (x_1, x_2, \dots x_{d-1})$ (see Figure 5.1).

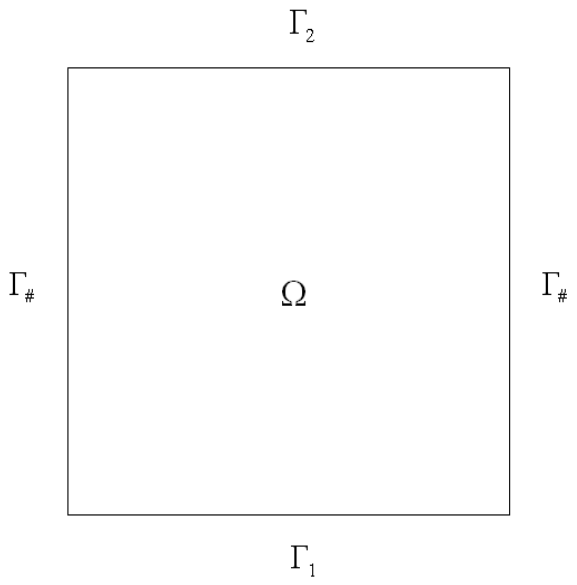

Figure 5.1: Boundaries for the domain of work $\Omega = (0,1)^d$.

For convenience we consider Dirichlet boundary conditions only on Γ_1 and Γ_2, and impose periodic boundary conditions on the rest of the boundary $\Gamma_\#$. This implies that there are no boundary layer terms due to $\Gamma_\#$, and so no interaction between adjacent edges of Ω, which would require to introduce specific boundary layers in the corners.

Let us define the Sobolev space

$$H_D(\Omega) = \{v \in H^1(\Omega),\ \ v = 0 \text{ on } \Gamma_1 \cup \Gamma_2,\ \ x' \mapsto v(x', x_d)\ \mathbb{Z}^{d-1} - \text{periodic}\}, \qquad (5.30)$$

equipped with the $H^1(\Omega)$ norm.

The function u_ε is then the unique solution in $H_D(\Omega)$ to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \ \text{ in } \Omega, \\ u_\varepsilon = 0 \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto u_\varepsilon(x', x_d) \ \ \mathbb{Z}^{d-1} - \text{periodic}. \end{cases} \qquad (5.31)$$

The role of the first boundary layer is to compensate for the first corrector

$$u_1(x, \frac{x}{\varepsilon}) = \sum_{i=1}^{n} \frac{\partial u_0}{\partial x_i}(x) w_i(\frac{x}{\varepsilon})$$

on Γ_1 and Γ_2 (note that we have taken $\tilde{u}_1 = 0$ in (5.15)). The linear structure of u_1 implies that we can associate to each cell solution w_i and each boundary Γ_j a boundary layer term.

For this purpose, we define $Q' = (0,1)^{d-1}$, $G^1 = Q' \times (0, +\infty)$, $\Gamma = Q' \times \{0\}$, $\partial G^1_\# = \partial Q' \times (0, +\infty)$, $G^2 = Q' \times (-\infty, 0)$, and $\partial G^2_\# = \partial Q' \times (-\infty, 0)$ (see Figure 5.2).

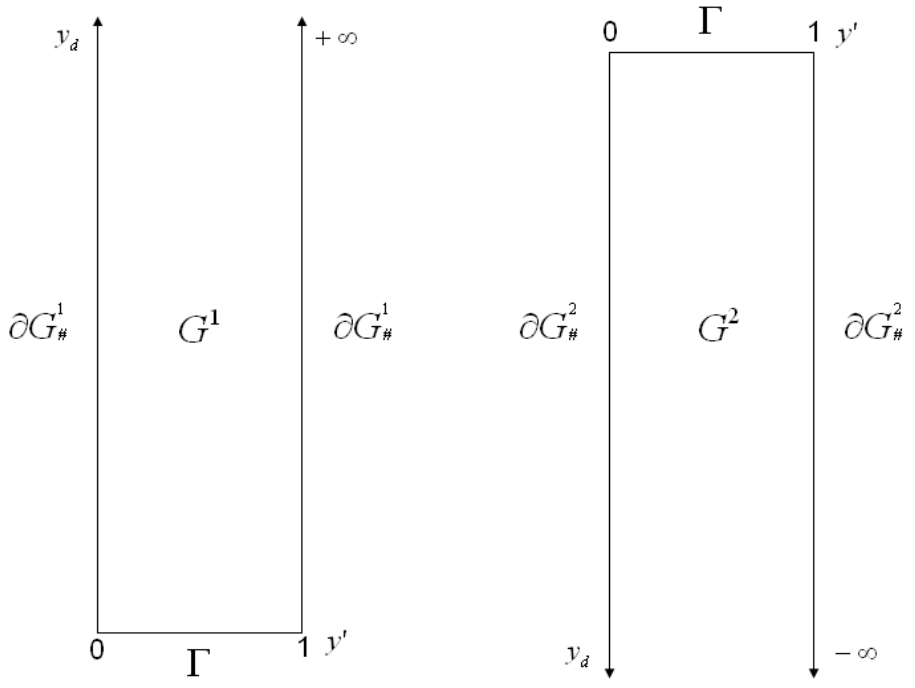

Figure 5.2: Semi-infinite strips G^1 and G^2.

For $j \in [\![1,2]\!]$ and $i \in [\![1,d]\!]$, we denote by $\psi_D^{i,j}$ the boundary layer term aiming at compensating for w_i on Γ_j, solution to

$$\begin{cases} -\operatorname{div}(A\nabla\psi_D^{i,j}) = 0 \ \text{ in } G^j, \\ \psi_D^{i,j} = -w_i \ \text{ on } \Gamma, \\ y' \mapsto \psi_D^{i,j}(y', y_d) \ \ \mathbb{Z}^{d-1} - \text{periodic.} \end{cases} \tag{5.32}$$

The following lemma, the proof of which can be found in [11] and [56], gives crucial properties of the functions $\psi_D^{i,j}$, namely exponential decay far from the boundary.

Lemma 5.4. *For all $j \in [\![1,2]\!]$ and $i \in [\![1,d]\!]$, there exists a unique solution $\psi_D^{i,j}$ of (5.32) in $H^1_{loc}(G^j)$. Moreover, there exist an exponent $\gamma > 0$ and a unique real constant $d^{i,j}$ such that*

$$e^{\gamma|y_d|}(\psi_D^{i,j} - d^{i,j}) \in L^2(G^j), \ \ e^{\gamma|y_d|}\nabla\psi_D^{i,j} \in L^2(G^j).$$

Due to the linearity of the structure of u_1, we define the global Dirichlet boundary layer in Ω by

$$u_D^{bl}(x, \frac{x}{\varepsilon}) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x) \left(\chi^1(x)\psi_D^{i,1}(\frac{x}{\varepsilon}) + \chi^2(x)\psi_D^{i,2}\left(\frac{x'}{\varepsilon}, \frac{1-x_d}{\varepsilon}\right)\right) \tag{5.33}$$

where, for $j \in [\![1,2]\!]$, χ^j is a smooth cut-off function equal to 1 on Γ^j and 0 on the opposite boundary. We then have the following result [64]:

Theorem 5.5. *Consider u_ε solution to (5.31) and assume that u_0 is smooth. Then*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) - \varepsilon u_D^{bl}(x, \frac{x}{\varepsilon}) \right\|_{H^1(\Omega)} \leq C\varepsilon.$$

The boundary layer u_D^{bl} therefore yields the same precision as $u_1^{bl,\varepsilon}$ defined by (5.29). It is however, contrary to $u_1^{bl,\varepsilon}$, tractable from a numerical point of view. Indeed, although one has in theory to solve $2d$ problems (5.32) defined on half strips to compute u_D^{bl}, the exponential decay given by Lemma 5.4 implies that we may, up to an error of order smaller than ε, truncate those strips so that they contain a small number of cells. In practice, it is sufficient to work with five to ten cells. As an illustration of this, we show in Figure 5.3 an instance of computation of the functions $\psi_D^{i,1}$ on a truncation of the strip G^1 composed of five cells, for a material consisting of a periodic lattice of circular inclusions in dimension two. It is clear from the isovalues of the functions in Figure 5.3 that the boundary layer only lives very close to the boundary, and that one merely needs a small number of cells to determine it entirely.

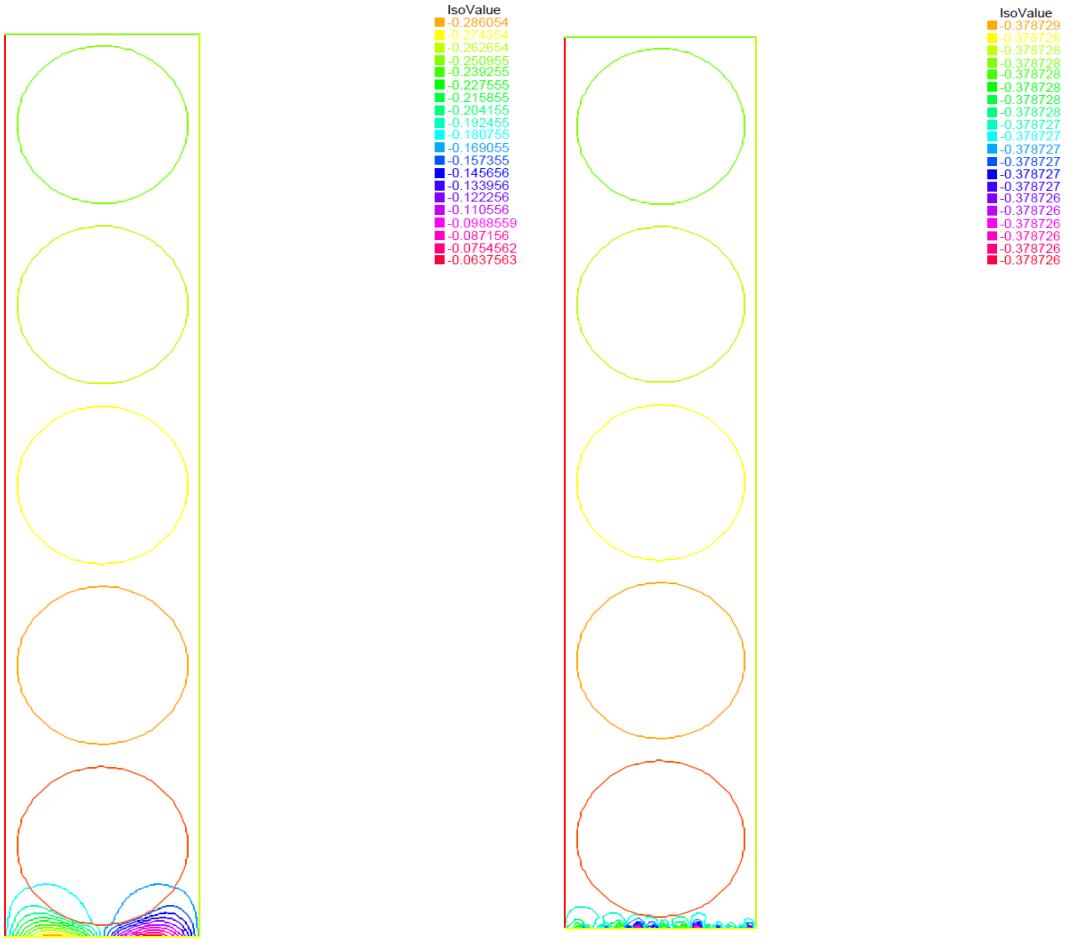

Figure 5.3: An instance of computation of the Dirichlet boundary layer for a periodic lattice of inclusions in dimension 2. Left: isovalues of the function $\psi_D^{1,1}$ associated with w_1 on G^1. Right: isovalues of the function $\psi_D^{2,1}$ associated with w_2 on G^1.

The goal of the next section is to adapt the approach exposed in the Dirichlet setting to address Neumann boundary conditions. Although this is quite a straightforward adaptation, we have not found it in the literature. The only related work we are aware of is [62], where the Neumann boundary conditions are transformed by a duality argument into Dirichlet boundary conditions, which is not the path we want to follow.

5.3.2 Neumann boundary conditions

Let us temporarily go back to a general regular bounded domain Ω, and consider $f \in L^2(\Omega)$ and $g \in L^2(\partial\Omega)$ satisfying the Neumann compatibility condition $\int_\Omega f + \int_{\partial\Omega} g = 0$, and u_ε solution to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \ \text{ in } \Omega, \\ A_\varepsilon \nabla u_\varepsilon \cdot n = g \ \text{ on } \partial\Omega, \end{cases} \tag{5.34}$$

where n is the outward unit normal vector on $\partial\Omega$.

It is well known that the corresponding homogenized problem reads

$$\begin{cases} -\operatorname{div}(A^*\nabla u_0) = f \ \text{ in } \Omega, \\ A^*\nabla u_0 \cdot n = g \ \text{ on } \partial\Omega. \end{cases} \tag{5.35}$$

Ansatz (5.2) induces a flux on $\partial\Omega$ that is formally equal to

$$A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1\left(x, \frac{x}{\varepsilon}\right)\right) \cdot n + \mathcal{O}(\varepsilon),$$

and that thus differs from the flux of the exact solution $A_\varepsilon \nabla u_\varepsilon \cdot n$.

Consequently, we have to add boundary layers in order to satisfy the Neumann boundary condition. As in the Dirichlet setting, it is possible to propose at order one in ε an intuitive boundary layer $v_1^{bl,\varepsilon}$ solution to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla v_1^{bl,\varepsilon}) = f + \operatorname{div}\left(A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1(x, \frac{x}{\varepsilon})\right)\right) \ \text{ in } \Omega, \\ A_\varepsilon \nabla v_1^{bl,\varepsilon} \cdot n = A_\varepsilon \nabla u_\varepsilon(x) \cdot n - A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1(x, \frac{x}{\varepsilon})\right) \cdot n \ \text{ on } \partial\Omega. \end{cases} \tag{5.36}$$

It is clear in (5.36) that the boundary condition compensates for the difference between the flux of the exact solution and the flux of the first-order expansion, up to an error of order ε. Note moreover that contrary to (5.29), the source term in (5.36) is not zero. It is actually chosen in order to satisfy the Neumann compatibility condition, and is also of order ε. It entails that $v_1^{bl,\varepsilon}$ is well defined (up to the addition of a constant).

A proof identical to that of Theorem 5.2 (see [62] or [2]) yields the following error estimate:

Theorem 5.6. *Assume that u_0 is smooth. It holds*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) - \varepsilon v_1^{bl,\varepsilon}(x) \right\|_{H^1(\Omega)/\mathbb{R}} \leq C\varepsilon,$$

where

$$\forall v \in H^1(\Omega), \ \ \|v\|_{H^1(\Omega)/\mathbb{R}} = \|\nabla v\|_{L^2(\Omega)}.$$

Even though the boundary layer $v_1^{bl,\varepsilon}$ allows to improve estimate (5.27), it is only of theoretical interest since its computation is as involved as that of u_ε.

To obtain practically relevant boundary layers, we consider in the rest of this section the case of the domain $\Omega = (0,1)^d$ containing an integer number of cells (i.e $1/\varepsilon \in \mathbb{N}$), and use the notation of Figures 5.1 and 5.2.

Defining the Sobolev space

$$H_N(\Omega) = \{v \in H^1(\Omega), \ \ x' \mapsto v(x', x_d) \ \ \mathbb{Z}^{d-1} - \text{periodic}\} \tag{5.37}$$

equipped with the $H^1(\Omega)$ norm, u_ε is now the unique solution in $H_N(\Omega)/\mathbb{R}$ to

$$\begin{cases} -\operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \ \text{ in } \Omega, \\ A_\varepsilon \nabla u_\varepsilon \cdot n = g \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto u_\varepsilon(x', x_d) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \end{cases} \tag{5.38}$$

where f and g satisfy the compatibility condition $\displaystyle\int_\Omega f + \int_{\Gamma_1 \cup \Gamma_2} g = 0$.

At order one in ε, the boundary layer has to compensate for the flux discrepancy $A_\varepsilon \nabla u_\varepsilon(x) \cdot n - A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1\left(x, \frac{x}{\varepsilon}\right)\right) \cdot n$ on $\Gamma_1 \cup \Gamma_2$. Notice that using the expression (5.15) of the first corrector and the fact that the original problem and the homogenized problem satisfy the same Neumann condition g, we find that for $x \in \Gamma_1 \cup \Gamma_2$,

$$\begin{aligned} A_\varepsilon \nabla u_\varepsilon(x) \cdot n - A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1(x, \frac{x}{\varepsilon})\right) \cdot n & \\ = g(x) - A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1(x, \frac{x}{\varepsilon})\right) \cdot n & \\ = A^* \nabla u_0(x) \cdot n - A_\varepsilon \left(\nabla u_0(x) + \nabla_y u_1(x, \frac{x}{\varepsilon})\right) \cdot n & \\ = \sum_{i=1}^n \frac{\partial u_0}{\partial x_i}(x) \left(A^* e_i \cdot n - A_\varepsilon \left(e_i + \nabla w_i(\frac{x}{\varepsilon})\right) \cdot n\right). & \end{aligned} \tag{5.39}$$

The linear structure of (5.39) with respect to the dimension implies that it is possible to write the global Neumann boundary layer as a sum of d terms, the i-th term aiming at compensating for $A^* e_i \cdot n - A_\varepsilon \left(e_i + \nabla_y w_i(\frac{x}{\varepsilon})\right) \cdot n$ on $\Gamma_1 \cup \Gamma_2$. For all $i \in [\![1, d]\!]$ and all $j \in [\![1, 2]\!]$ we denote by $\psi_N^{i,j}$ the term associated with the i-th dimension on Γ_j, solution to

$$\begin{cases} -\operatorname{div}\left(A \nabla \psi_N^{i,j}\right) = 0 \ \text{ in } G^j, \\ A \nabla \psi_N^{i,j} \cdot n = A^* e_i \cdot n - A(e_i + \nabla w_i) \cdot n \ \text{ on } \Gamma, \\ y' \mapsto \psi_N^{i,j}(y', \cdot) \ \ \mathbb{Z}^{d-1} - \text{periodic}. \end{cases} \tag{5.40}$$

Problem (5.40) is well posed in $H^1_{loc}(G^j)/\mathbb{R}$ if and only if the Neumann compatibility condition is satisfied, i.e if and only if

$$\int_\Gamma A^* e_i \cdot n = \int_\Gamma A(e_i + \nabla w_i) \cdot n. \tag{5.41}$$

We check thereafter that this is the case. For clarity, we call n_Γ the outward unit normal vector on Γ, which is a constant vector. Using (5.5), we have

$$\begin{aligned} \int_\Gamma A^* e_i \cdot n_\Gamma &= A^* e_i \cdot n_\Gamma \\ &= \left(\int_Q A(e_i + \nabla w_i)\right) \cdot n_\Gamma. \end{aligned} \tag{5.42}$$

Note that n_Γ is equal to $-\nabla y_d$ if we work in the domain G^1 and to ∇y_d if we work in G^2. It is then equal to $(-1)^j \nabla y_d$ in G^j. Consequently,

$$\left(\int_Q A(e_i + \nabla w_i)\right) \cdot n_\Gamma = (-1)^j \int_Q A(e_i + \nabla w_i) \cdot \nabla y_d \tag{5.43}$$

and an integration by parts in the right-hand side of (5.43) shows that

$$\begin{aligned} &\left(\int_Q A(e_i + \nabla_y w_i)\right) \cdot n_\Gamma \\ &= (-1)^j \left(-\int_Q \mathrm{div}(A(e_i + \nabla_y w_i) y_d + \int_{\partial Q} A(e_i + \nabla_y w_i) \cdot n\, y_d\right). \end{aligned} \tag{5.44}$$

The first term in the right-hand side of (5.44) is zero because of (5.4). Using the $\mathbb{Z}^d$-periodicity of A and w_i, we have on the other hand

$$(-1)^j \int_{\partial Q} A(e_i + \nabla_y w_i) \cdot n\, y_d = \int_\Gamma A(e_i + \nabla_y w_i) \cdot n_\Gamma. \tag{5.45}$$

Collecting (5.42), (5.44) and (5.45), we conclude that (5.41) holds. Thus the functions $\psi_N^{i,j}$ are well defined (up to an additive constant).

The functions $\psi_N^{i,j}$ defined by (5.40) behave exactly as the functions $\psi_D^{i,j}$ defined by (5.32), as shown by the following result, the proof of which is similar to that of Lemma 5.4 (see [11] or [56]):

Lemma 5.7. *For all $j \in [\![1,2]\!]$ and $i \in [\![1,d]\!]$, there exists a unique $\psi_N^{i,j}$ solution to (5.40) in $H^1_{loc}(G^j)/\mathbb{R}$. Moreover, there exist an exponent $\gamma > 0$ and a unique real constant $\tilde{d}^{i,j}$ such that*

$$e^{\gamma|y_d|}(\psi_N^{i,j} - \tilde{d}^{i,j}) \in L^2(G^j), \;\; e^{\gamma|y_d|}\nabla\psi_N^{i,j} \in L^2(G^j).$$

We then propose the global Neumann boundary layer

$$u_N^{bl}(x, \frac{x}{\varepsilon}) \;\; = \;\; \sum_{i=1}^d \frac{\partial u_0}{\partial x_i}(x) \left(\psi_N^{i,1}(\frac{x}{\varepsilon}) + \psi_N^{i,2}(\frac{x'}{\varepsilon}, \frac{1-x_d}{\varepsilon})\right). \tag{5.46}$$

Due to the exponential decay of the functions $\psi_N^{i,j}$ given by Lemma 5.7, u_N^{bl} is tractable from a numerical point of view. The qualitative behavior of the boundary layer is the same as in the Dirichlet setting, and it suffices in practice to compute the functions $\psi_N^{i,j}$ on truncated strips consisting of a small number of cells. Figure 5.4 shows an instance of computation of the functions $\psi_N^{i,1}$ on a truncation of the strip G^1 composed of five cells, for a periodic lattice of circular inclusions in dimension two.

The rest of this section is devoted to the proof of the following error estimate that shows the relevance of u_N^{bl}. It relies on the same techniques used in [2] and [64].

Theorem 5.8. *Consider u_ε solution to (5.38) and assume that u_0 and A are smooth. Then*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) - \varepsilon u_N^{bl}(x, \frac{x}{\varepsilon}) \right\|_{H^1(\Omega)/\mathbb{R}} \leq C\varepsilon.$$

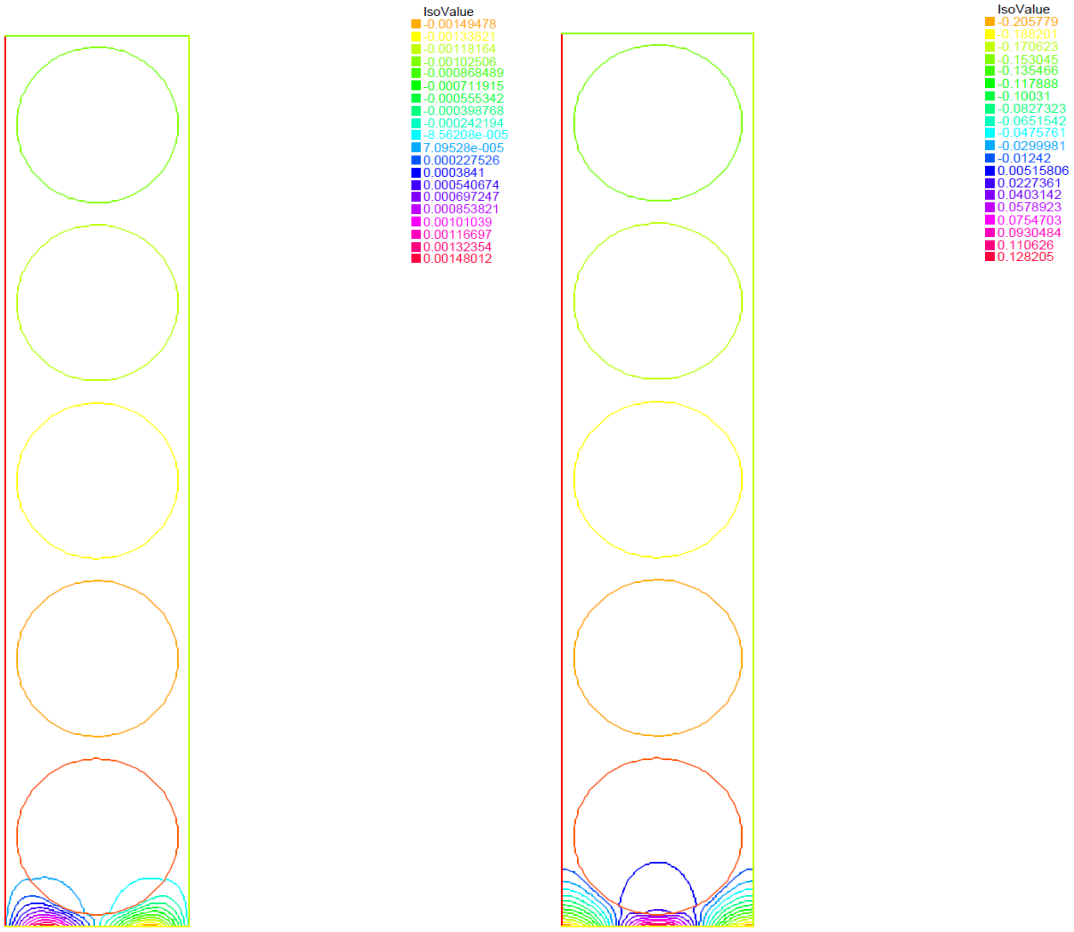

Figure 5.4: An instance of computation of the Neumann boundary layer for a periodic lattice of inclusions in dimension 2. Left: isovalues of the function $\psi_N^{1,1}$ associated with w_1 on G^1. Right: isovalues of the function $\psi_N^{2,1}$ associated with w_2 on G^1.

Proof. We define the remainder

$$r_\varepsilon(x) = u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}) - \varepsilon u_N^{bl}(x, \frac{x}{\varepsilon}),$$

and

$$v_\varepsilon = u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}).$$

Our goal is to prove that there exists a constant C such that, for all $\varepsilon > 0$,

$$\|\nabla r_\varepsilon\|_{L^2(\Omega)} \leq C\varepsilon.$$

By definition of r_ε and v_ε, we have, for all $\phi \in H_N(\Omega)$,

$$\int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi = \int_\Omega A_\varepsilon \nabla v_\varepsilon \cdot \nabla \phi - \varepsilon \int_\Omega A_\varepsilon \nabla u_N^{bl} \cdot \nabla \phi. \tag{5.47}$$

We proceed in three steps.

Step 1.

We first address the second term of the right-hand side of (5.47). Using (5.46), we have

$$\begin{aligned}\varepsilon \int_\Omega A_\varepsilon \nabla \left(u_N^{bl}(x, \frac{x}{\varepsilon}) \right) \cdot \nabla \phi =& \varepsilon \sum_{i=1}^d \int_\Omega A_\varepsilon \nabla \left(\frac{\partial u_0}{\partial x_i}(x) \psi_N^{1,i}(\frac{x}{\varepsilon}) \right) \cdot \nabla \phi \\ &+ \varepsilon \sum_{i=1}^d \int_\Omega A_\varepsilon \nabla \left(\frac{\partial u_0}{\partial x_i}(x) \psi_N^{2,i}(\frac{x'}{\varepsilon}, \frac{1-x_d}{\varepsilon}) \right) \cdot \nabla \phi.\end{aligned} \tag{5.48}$$

It suffices to focus on $\varepsilon \displaystyle\int_\Omega A_\varepsilon \nabla \left(\frac{\partial u_0}{\partial x_1}(x) \chi^1(x) \psi_N^{1,1}(\frac{x}{\varepsilon}) \right) \cdot \nabla \phi$ in (5.48), the other terms being dealt with in a similar manner.

The function u_0 is smooth, and our regularity assumptions on A yield that $\psi_N^{1,1}$ is in $L^\infty(G^1)$. This implies that

$$\begin{aligned}\varepsilon \int_\Omega A_\varepsilon \nabla \left(\frac{\partial u_0}{\partial x_1}(x) \psi_N^{1,1}(\frac{x}{\varepsilon}) \right) \cdot \nabla \phi =& \int_\Omega A_\varepsilon \frac{\partial u_0}{\partial x_1}(x) \nabla (\varepsilon \psi_N^{1,1}(\frac{x}{\varepsilon})) \cdot \nabla \phi \\ &+ \mathcal{O}(\varepsilon) \|\nabla \phi\|_{L^2(\Omega)} \\ =& \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \nabla \left(\phi \frac{\partial u_0}{\partial x_1} \right) \\ &- \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \phi \, \nabla \frac{\partial u_0}{\partial x_1} \\ &+ \mathcal{O}(\varepsilon) \|\nabla \phi\|_{L^2(\Omega)}.\end{aligned} \tag{5.49}$$

Integrating by parts in the right-hand side of (5.49) and using (5.40), we find that

$$\begin{aligned}\varepsilon \int_\Omega A_\varepsilon \nabla \left(\frac{\partial u_0}{\partial x_1}(x) \psi_N^{1,1}(\frac{x}{\varepsilon}) \right) \cdot \nabla \phi =& \int_\Omega -\frac{1}{\varepsilon} \operatorname{div}_y \left(A \nabla_y \psi_N^{1,1} \right) (\frac{x}{\varepsilon}) \frac{\partial u_0}{\partial x_1} \phi \\ &+ \int_{\Gamma_1 \cup \Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n \, \phi \, \frac{\partial u_0}{\partial x_1} \\ &- \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \phi \, \nabla \frac{\partial u_0}{\partial x_1} \\ &+ \mathcal{O}(\varepsilon) \|\nabla \phi\|_{L^2(\Omega)} \\ =& \int_{\Gamma_1 \cup \Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n \, \phi \, \frac{\partial u_0}{\partial x_1} \\ &- \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \phi \, \nabla \frac{\partial u_0}{\partial x_1} \\ &+ \mathcal{O}(\varepsilon) \|\nabla \phi\|_{L^2(\Omega)}.\end{aligned} \tag{5.50}$$

We will first prove that in (5.50), $\displaystyle\int_{\Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n \, \phi \, \frac{\partial u_0}{\partial x_1}$ is negligible compared to ε, which amounts to say that the boundary layer term associated with Γ_1 is negligible

on Γ_2.

By a classical duality argument, it holds

$$\left|\int_{\Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \phi\, \frac{\partial u_0}{\partial x_1}\right| \leq \left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\right\|_{H^{-1/2}(\Gamma_2)} \left\|\phi\, \frac{\partial u_0}{\partial x_1}\right\|_{H^{1/2}(\Gamma_2)}. \tag{5.51}$$

Using the definition of $H^{1/2}(\Omega)$ and the smoothness of u_0, we deduce from (5.51) that

$$\left|\int_{\Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \phi\, \frac{\partial u_0}{\partial x_1}\right| \leq C \left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\right\|_{H^{-1/2}(\Gamma_2)} \|\phi\|_{H^1(\Omega)}. \tag{5.52}$$

Let us then define $\Omega_1 = (0,1)^{d-1} \times (0, \frac{1}{2})$, $\Omega_2 = (0,1)^{d-1} \times (\frac{1}{2}, 1)$ and $\chi \in \mathcal{C}^\infty(\overline{\Omega}) \cap H(\Omega)$ equal to 1 on Γ_2 and to 0 in Ω_1. We have

$$\left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\right\|_{H^{-1/2}(\Gamma_2)} = \left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \chi\right\|_{H^{-1/2}(\partial\Omega_2)},$$

and because of a trace theorem in $H^{div}(\Omega_2)$,

$$\begin{aligned}\left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \chi\right\|_{H^{-1/2}(\partial\Omega_2)} \leq C \Bigg(&\left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\, \chi\right\|_{L^2(\Omega_2)} \\ &+ \left\|\mathrm{div}\left(A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\, \chi\right)\right\|_{L^2(\Omega_2)}\Bigg).\end{aligned} \tag{5.53}$$

Using (5.40), we compute

$$\begin{aligned}\mathrm{div}\left(A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\, \chi\right) &= \mathrm{div}\left(A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\right) \chi + A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \nabla\chi \\ &= A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \nabla\chi.\end{aligned} \tag{5.54}$$

It follows from (5.53), (5.54) and the boundedness of A that

$$\begin{aligned}\left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \chi\right\|_{H^{-1/2}(\partial\Omega_2)} &\leq C \Bigg(\left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\, \chi\right\|_{L^2(\Omega_2)} \\ &\quad + \left\|A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \nabla\chi\right\|_{L^2(\Omega_2)}\Bigg) \\ &\leq C \left\|\nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\right\|_{L^2(\Omega_2)}.\end{aligned} \tag{5.55}$$

The exponential decay given by Lemma 5.7 yields

$$\left\|\nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\right\|_{L^2(\Omega_2)} \leq C e^{-\frac{\gamma}{\varepsilon}} \tag{5.56}$$

for some $\gamma > 0$. We deduce from (5.52), (5.55) and (5.56) that

$$\left|\int_{\Gamma_2} A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot n\, \phi\, \frac{\partial u_0}{\partial x_1}\right| \leq C e^{-\frac{\gamma}{\varepsilon}} \|\phi\|_{H^1(\Omega)}. \tag{5.57}$$

Next, we address the second term in the right-hand side of (5.50) and following [2], we part the domain Ω as

$$\Omega = \bigcup_{k=1}^{\varepsilon^{-1}} C_\varepsilon^k$$

with $C_\varepsilon^k = \{(x', x_d) \in \Omega, (k-1)\varepsilon \leq x_d \leq k\varepsilon\}$.

Moreover we denote by $D_\varepsilon^k = \bigcup_{l=1}^k C_\varepsilon^l = [0,1]^{n-1} \times [0, k\varepsilon]$.

Since u_0 is smooth and A is bounded,

$$\begin{aligned} \left| \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \phi \, \nabla \frac{\partial u_0}{\partial x_1} \right| &\leq C \int_\Omega \left| \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \right| \, |\phi| \\ &\leq C \sum_{k=1}^{\varepsilon^{-1}} \int_{C_\varepsilon^k} \left| \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \right| |\phi|, \end{aligned} \tag{5.58}$$

and using the Cauchy-Schwarz inequality in (5.58),

$$\left| \int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \cdot \phi \, \nabla \frac{\partial u_0}{\partial x_1} \right| \leq \sum_{k=1}^{\varepsilon^{-1}} \left\| \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \right\|_{L^2(C_\varepsilon^k)} \|\phi\|_{L^2(D_\varepsilon^k)}. \tag{5.59}$$

For $x \in D_\varepsilon^k$, we write

$$\phi(x) = \phi(x', 0) + \int_{z=0}^{x_d} \frac{\partial \phi}{\partial z}(x', z) dz,$$

whence

$$\begin{aligned} \phi(x)^2 &\leq 2\phi(x', 0)^2 + 2 \left(\int_{z=0}^{x_d} \frac{\partial \phi}{\partial x_d}(x', z) dz \right)^2 \\ &\leq 2\phi(x', 0)^2 + 2x_d \int_{z=0}^{x_d} \left| \frac{\partial \phi}{\partial x_d} \right|^2 (x', z) dz \\ &\leq 2\phi(x', 0)^2 + 2k\varepsilon \int_{z=0}^{k\varepsilon} |\nabla \phi|^2 (x', z) dz. \end{aligned} \tag{5.60}$$

Integrating (5.60) on D_ε^k shows that

$$\begin{aligned} \|\phi\|_{L^2(D_\varepsilon^k)}^2 &\leq 2k\varepsilon \|\phi\|_{L^2(\Gamma_1)}^2 + 2(k\varepsilon)^2 \|\nabla \phi\|_{L^2(D_\varepsilon^k)}^2 \\ &\leq 2k\varepsilon \|\phi\|_{L^2(\Gamma_1)}^2 + 2(k\varepsilon)^2 \|\nabla \phi\|_{L^2(\Omega)}^2. \end{aligned} \tag{5.61}$$

On the other hand, using Lemma 5.7, we find that there exist constants $C > 0$ and $\gamma > 0$ such that

$$\forall \varepsilon > 0, \ \forall k \in [\![1, \varepsilon^{-1}]\!], \left\| \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon}) \right\|_{L^2(C_\varepsilon^k)} \leq C\sqrt{\varepsilon} e^{-\gamma k}. \tag{5.62}$$

Substituting (5.61) and (5.62) in (5.59) yields

$$\begin{aligned}\left|\int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\cdot\phi\,\nabla\frac{\partial u_0}{\partial x_1}\right| &\le C\sum_{k=1}^{\varepsilon^{-1}}\sqrt{\varepsilon}e^{-\gamma k}(\sqrt{k\varepsilon}\|\phi\|_{L^2(\Gamma_1)}+k\varepsilon\|\nabla\phi\|_{L^2(\Omega)})\\ &\le C\varepsilon(\|\phi\|_{L^2(\Gamma_1)}+\sqrt{\varepsilon}\|\nabla\phi\|_{L^2(\Omega)}).\end{aligned} \tag{5.63}$$

A trace theorem in $H^1(\Omega)$ then implies

$$\left|\int_\Omega A_\varepsilon \nabla_y \psi_N^{1,1}(\frac{x}{\varepsilon})\cdot\phi\,\nabla\frac{\partial u_0}{\partial x_1}\right| \le C\varepsilon\|\phi\|_{H^1(\Omega)}. \tag{5.64}$$

Collecting (5.50), (5.57) and (5.64), we have thus proved that

$$\int_\Omega A_\varepsilon\frac{\partial u_0}{\partial x_1}(x)\nabla(\varepsilon\psi_N^{1,1}(\frac{x}{\varepsilon}))\cdot\nabla\phi = \int_{\Gamma_1} A_\varepsilon\nabla_y\psi_N^{1,1}(\frac{x}{\varepsilon})\cdot n\;\phi\,\frac{\partial u_0}{\partial x_1}+\mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}.$$

Arguing similarly for other boundary layer terms, we derive from (5.48) that

$$\begin{aligned}\varepsilon\int_\Omega A_\varepsilon\nabla\left(u_N^{bl}(x,\frac{x}{\varepsilon})\right)\cdot\nabla\phi &= \sum_{i=1}^d\int_{\Gamma_1}A_\varepsilon\nabla_y\psi_N^{i,1}(\frac{x}{\varepsilon})\cdot n\;\phi\,\frac{\partial u_0}{\partial x_i}\\ &+\sum_{i=1}^d\int_{\Gamma_2}A_\varepsilon\nabla_y\psi_N^{i,2}(\frac{x'}{\varepsilon},\frac{1-x_d}{\varepsilon})\cdot n\;\phi\,\frac{\partial u_0}{\partial x_i}\\ &+\mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}.\end{aligned} \tag{5.65}$$

Step 2.

We now address the term $\displaystyle\int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi$ in (5.47). By definition of v_ε, we have

$$\begin{aligned}&\int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi =\\ &\int_\Omega A_\varepsilon\left(\nabla u_\varepsilon-\nabla u_0(x)-\nabla_y u_1(x,\frac{x}{\varepsilon})-\varepsilon\nabla_x u_1(x,\frac{x}{\varepsilon})\right)\cdot\nabla\phi.\end{aligned} \tag{5.66}$$

Using the smoothness of u_0 and the definition (5.16) of u_2, it is straightforward to see that there exists a constant C such that for all ε,

$$\varepsilon\left|\int_\Omega\nabla_y u_2(x,\frac{x}{\varepsilon})\cdot\nabla\phi\right| \le C\varepsilon\|\phi\|_{H^1(\Omega)}. \tag{5.67}$$

Inserting (5.67) in (5.66), we write

$$\begin{aligned}&\int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi =\\ &\int_\Omega A_\varepsilon\left(\nabla u_\varepsilon-\nabla_x u_0(x)-\nabla_y u_1(x,\frac{x}{\varepsilon})-\varepsilon\nabla_x u_1(x,\frac{x}{\varepsilon})-\varepsilon\nabla_y u_2(x,\frac{x}{\varepsilon})\right)\cdot\nabla\phi\\ &+\mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}.\end{aligned} \tag{5.68}$$

Since u_0, u_1 and u_2 satisfy (5.13), it is straightforward to see that

$$\begin{aligned} -\mathrm{div}\left(A_\varepsilon\left(\nabla u_\varepsilon - \nabla u_0(x) - \nabla_y u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_x u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_y u_2(x,\frac{x}{\varepsilon})\right)\right) \\ = -\varepsilon\ \mathrm{div}_x\left(A_\varepsilon\nabla_y u_2 + A_\varepsilon\nabla_x u_1\right)(x,\frac{x}{\varepsilon}). \end{aligned} \tag{5.69}$$

It follows from the smoothness of u_0 and the definitions (5.15) and (5.16) of u_1 and u_2 that $\mathrm{div}_x\left(A_\varepsilon\nabla_y u_2 + A_\varepsilon\nabla_x u_1\right)(x,\frac{x}{\varepsilon})$ is bounded in $L^2(\Omega)$ independently of ε.

Consequently, integrating by parts in the right-hand side of (5.68) and using (5.69), we get

$$\begin{aligned} \int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi &= -\varepsilon\int_\Omega \mathrm{div}_x\left(A_\varepsilon\nabla_y u_2 + A_\varepsilon\nabla_x u_1\right)(x,\frac{x}{\varepsilon})\,\phi \\ &+ \int_{\Gamma_1\cup\Gamma_2} A_\varepsilon\left(\nabla u_\varepsilon - \nabla u_0(x) - \nabla_y u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_x u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_y u_2(x,\frac{x}{\varepsilon})\right)\cdot n\,\phi \\ &+ \mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)} \\ &= \int_{\Gamma_1\cup\Gamma_2} A_\varepsilon\left(\nabla u_\varepsilon - \nabla u_0(x) - \nabla_y u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_x u_1(x,\frac{x}{\varepsilon}) - \varepsilon\nabla_y u_2(x,\frac{x}{\varepsilon})\right)\cdot n\,\phi \\ &+ \mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}. \end{aligned} \tag{5.70}$$

We deduce from the definitions of u_1 and u_2 and the smoothness of u_0 and A that there exists a constant C such that

$$\left\|A_\varepsilon\left(\nabla_y u_2 + \nabla_x u_1\right)(x,\frac{x}{\varepsilon})\cdot n\right\|_{L^\infty(\Gamma_1\cup\Gamma 2)} \le C.$$

The previous inequality and a trace theorem in $H^1(\Omega)$ give

$$\begin{aligned} \left|\int_{\Gamma_1\cup\Gamma_2} A_\varepsilon\left(\nabla_y u_2 + \nabla_x u_1\right)(x,\frac{x}{\varepsilon})\cdot n\ \phi\right| &\le C\|\phi\|_{L^2(\Gamma_1\cup\Gamma_2)} \\ &\le C\|\phi\|_{H^1(\Omega)}. \end{aligned} \tag{5.71}$$

It entails from (5.70) and (5.71) that

$$\int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi = \int_{\Gamma_1\cup\Gamma_2} A_\varepsilon\left(\nabla u_\varepsilon - \nabla u_0(x) - \nabla_y u_1(x,\frac{x}{\varepsilon})\right)\cdot n\,\phi + \mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}, \tag{5.72}$$

and then from (5.39) that

$$\begin{aligned} \int_\Omega A_\varepsilon\nabla v_\varepsilon\cdot\nabla\phi = \sum_{i=1}^d \int_{\Gamma_1\cup\Gamma_2}\left(A^* e_i - A_\varepsilon(e_i + \nabla_y w_i(\frac{x}{\varepsilon}))\right)\cdot n\,\frac{\partial u_0}{\partial x_i}\,\phi \\ +\mathcal{O}(\varepsilon)\|\phi\|_{H^1(\Omega)}. \end{aligned} \tag{5.73}$$

Step 3.

Collecting (5.65) and (5.73) yields

$$\begin{aligned}\int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi = \sum_{i=1}^n \int_{\Gamma_1} \Big(A^* e_i - A_\varepsilon (e_i + \nabla_y w_i(\frac{x}{\varepsilon}) + \nabla_y \psi_N^{i,1}(\frac{x}{\varepsilon})) \Big) \cdot n \, \frac{\partial u_0}{\partial x_i} \, \phi \\ + \sum_{i=1}^n \int_{\Gamma_2} \Big(A^* e_i - A_\varepsilon (e_i + \nabla_y w_i(\frac{x}{\varepsilon}) + \nabla_y \psi_N^{i,2}(\frac{x'}{\varepsilon}, \frac{1-x_d}{\varepsilon})) \Big) \cdot n \, \frac{\partial u_0}{\partial x_i} \, \phi \\ + \mathcal{O}(\varepsilon) \|\phi\|_{H^1(\Omega)}.\end{aligned} \tag{5.74}$$

The definition (5.40) of the functions $\psi_N^{i,j}$ implies that the boundary integrals in (5.74) are all equal to zero, so that we are left with

$$\int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi = \mathcal{O}(\varepsilon) \|\phi\|_{H^1(\Omega)},$$

which is equivalent to

$$\left| \int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi \right| \leq C\varepsilon \|\phi\|_{H^1(\Omega)}.$$

Replacing ϕ by $\phi + c$ for any real constant c, we obtain

$$\forall c \in \mathbb{R}, \quad \left| \int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi \right| \leq C\varepsilon \|\phi + c\|_{H^1(\Omega)}$$

and then

$$\left| \int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi \right| \leq C\varepsilon \inf_{c \in \mathbb{R}} \|\phi + c\|_{H^1(\Omega)}. \tag{5.75}$$

Thanks to Deny-Lions' theorem, we deduce from (5.75) that

$$\left| \int_\Omega A_\varepsilon \nabla r_\varepsilon \cdot \nabla \phi \right| \leq C\varepsilon \|\phi\|_{H^1(\Omega)/\mathbb{R}}.$$

This being true for any $\phi \in H_N(\Omega)$, we conclude that

$$\|r_\varepsilon\|_{H^1(\Omega)/\mathbb{R}} \leq C\varepsilon.$$

□

5.4 Boundary layers for parabolic equations

Here we deal with the parabolic setting presented in Section 5.2.2. We first explain why it is necessary to add to Ansatz (5.19) a new term, that we call an "initial layer", to account for the initial condition at $t = 0$. Interestingly, this initial condition somehow plays the role of a new boundary and leads to issues similar to those encountered in the previous section. Then we consider, following [68], a specific parabolic problem in which there are no boundaries, so as to focus only on the initial layer and to understand how to design

it in an efficient and practical way. Finally we gather all the results about boundary and initial layers to address the general parabolic case. Unfortunately our results in this latter case are not conclusive, for they rely on regularity assumptions the validity of which we were not able to evaluate.

Throughout this section we assume that A is a *symmetric* tensor field. We will also often for simplicity (and when there is no possible confusion) write u_1 and u_2 instead of using the full notation $u_1(x, \frac{x}{\varepsilon}, t)$ and $u_2(x, \frac{x}{\varepsilon}, t)$. The same holds for all functions depending on the slow space variable x, the fast space variable $y = \frac{x}{\varepsilon}$, the slow time variable t and the fast time variable $\tau = \frac{t}{\varepsilon^2}$, once they have been defined.

5.4.1 Need for an "initial layer"

Let Ω be a general bounded regular open set of $\mathbb{R}^d$.

Consider, for $T > 0$, $f \in L^2(\Omega \times (0,T))$, $g \in L^2(\Omega)$ and u_ε solution to

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f \ \text{ in } \Omega \times (0,T), \\ u_\varepsilon = 0 \ \text{ on } \partial\Omega \times (0,T), \\ u_\varepsilon(\cdot, 0) = g \ \text{ in } \Omega. \end{cases} \tag{5.76}$$

It follows from Lemma 5.21 of the Appendix that for every $\varepsilon > 0$ there exists a unique u_ε solution to (5.76) in $\mathcal{C}([0,T]; L^2(\Omega)) \cap L^2(]0,T[; H_0^1(\Omega))$ and that u_ε is bounded independently of ε in this space. Consequently, u_ε converges weakly in $L^\infty(]0,T[; L^2(\Omega)) \cap L^2(]0,T[; H_0^1(\Omega))$ to a homogenized limit u_0 which solves (see [14])

$$\begin{cases} \dfrac{\partial u_0}{\partial t} - \operatorname{div}(A^* \nabla u_0) = f \ \text{ in } \Omega \times (0,T), \\ u_0 = 0 \ \text{ on } \partial\Omega \times (0,T), \\ u_0(\cdot, 0) = g \ \text{ in } \Omega. \end{cases} \tag{5.77}$$

It is further proved in [14] and [23] that u_ε converges to u_0 in

$$L^2(]0,T[; L^2(\Omega)) \cap \mathcal{C}([0,T]; H^{-1}(\Omega)),$$

and that adding the first corrector (5.23) yields the stronger convergence result:

$$u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) \underset{\varepsilon \to 0}{\to} 0 \ \text{ in } L^\infty(]0,T[; L^2(\Omega)) \cap L^2(]0,T[; H^1(\Omega)). \tag{5.78}$$

Under assumptions of smoothness on u_0, and following a proof identical to that of Theorem 5.1, it is easy to quantify the error estimate in the convergence (5.78) and to obtain:

$$\|u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t)\|_{\mathcal{C}([0,T]; L^2(\Omega)) \cap L^2(]0,T[; H^1(\Omega))} \leq C\sqrt{\varepsilon}. \tag{5.79}$$

The order $\sqrt{\varepsilon}$ in (5.79) comes, as in the stationary setting of Section 5.3, from the fact that the first corrector does not satisfy the Dirichlet boundary condition imposed on

$\partial\Omega$ in (5.76). To improve estimate (5.79), we have to add boundary layers. Those built in Sections 5.3.1 (for Dirichlet boundary conditions) and 5.3.2 (for Neumann boundary conditions) work, and allow to replace $\sqrt{\varepsilon}$ with ε.

More precisely, all the results presented in the stationary case can be readily adapted to address the parabolic setting: it suffices to formally replace the space $H^1(\Omega)$ by $\mathcal{C}([0,T];L^2(\Omega)) \cap L^2(]0,T[;H^1(\Omega))$.

However, in order to model real life experiments such as the pulsed-infrared thermography described in Section 2 of Chapter 1, we have to work in a space in which the traces of the functions on $\partial\Omega$ are defined for every t in $[0,T]$. To this end the space $\mathcal{C}([0,T];L^2(\Omega)) \cap L^2(]0,T[;H^1(\Omega))$ does not provide enough regularity, and the purpose of all that follows is to replace it with another classical space in the analysis of parabolic equations, namely $C([0,T];H^1(\Omega))$.

The use of $C([0,T];H^1(\Omega))$ yet raises a new difficulty, due to the fact that the Ansatz (5.19) does not satisfy the initial condition in (5.76). Indeed, at first order in ε, the difference

$$u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t)$$

is equal to $-\varepsilon u_1(x,\frac{x}{\varepsilon},0)$ at $t=0$. In view of Lemma 5.21, this is not an issue if we are to work in $C([0,T];L^2(\Omega)) \cap L^2(]0,T[;H^1_0(\Omega))$, since the estimates in this space rely on the $L^2(\Omega)$-norm of the initial condition, here of order ε. On the contrary, this matters if we choose to work in $C([0,T];H^1(\Omega))$: in this case, Lemma 5.22 yields that we use the $H^1(\Omega)$-norm of the initial condition, that is of order 1.

It is then necessary to add what we call an "initial layer" to compensate for the first corrector at $t=0$. Formally, this resembles very much what has been exposed in the elliptic setting for boundary layers, the difference being that the boundary is now $t=0$. To our knowledge, the only work available on this initial layer is [68]. The latter addresses the case of an infinite domain, which allows the author not to consider boundary layers and in particular not to deal with the interaction between the boundary layers and the initial layer. Our aim in this section is to extend the results of [68] and to give an ensemble picture of the problem of boundary and initial layers in parabolic homogenization.

As mentioned in the introduction to this chapter, we are not interested in finding the most general regularity assumptions under which our results hold, since we observe in the homogenization literature that only the proofs and not the intrinsic results do depend on these assumptions. Another point of view on this topic is that our approach has to be correct at least in regular settings. Thus, we will always assume that we have sufficient regularity, and in particular that u_0 is sufficiently smooth for our purposes. A convenient assumption for all that follows is e.g $u_0 \in \mathcal{C}^3(\Omega \times [0,T])$.

Remark 5.1. *The assumption of regularity on u_0 is not as restrictive as it may seem. Indeed, u_0 is solution to (5.77) which is the heat equation with a constant tensor A^*. Therefore the regularity of u_0 only depends on the regularity of the boundary and initial conditions.*

5.4.2 A theoretical boundary+initial layer

Here we verify that, as announced in the previous section, compensating for the first corrector on the boundary $\partial\Omega$ and at $t = 0$ yields an error estimate of order ε in the space $C([0,T];H^1(\Omega))$.

An intuitive way to proceed is to define a boundary+initial layer $u_1^{bil,\varepsilon}$ by

$$\begin{cases} \dfrac{\partial u_1^{bil,\varepsilon}}{\partial t} - \mathrm{div}(A_\varepsilon \nabla u_1^{bil,\varepsilon}) = 0 \ \text{ in } \Omega\times(0,T), \\ u_1^{bil,\varepsilon}(x,t) = -u_1(x,\dfrac{x}{\varepsilon},t) \ \text{ on } \partial\Omega\times(0,T), \\ u_1^{bil,\varepsilon}(x,0) = -u_1(x,\dfrac{x}{\varepsilon},0) \ \text{ in } \Omega. \end{cases} \tag{5.80}$$

Adding $u_1^{bil,\varepsilon}$ to Ansatz (5.19) yields the expected result:

Theorem 5.9. *Assume that u_0 is smooth. Then it holds*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t) - \varepsilon u_1^{bil,\varepsilon}(x,t) \right\|_{C([0,T];H^1(\Omega))} \le C\varepsilon.$$

Proof. We define the remainder $r_\varepsilon = \frac{1}{\varepsilon}(u_\varepsilon - u_0 - \varepsilon u_1 - \varepsilon u_1^{bil,\varepsilon})$. Using the operators (5.21) and the system (5.13), we see that r_ε is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon}{\partial t} - \mathrm{div}(A_\varepsilon \nabla r_\varepsilon) = \dfrac{1}{\varepsilon} L_0 u_2 - L_2 u_1 \ \text{ in } \Omega\times(0,T), \\ r_\varepsilon = 0 \ \text{ on } \partial\Omega\times(0,T), \\ r_\varepsilon(\cdot,0) = 0 \ \text{ in } \Omega. \end{cases} \tag{5.81}$$

We next verify that the right-hand side of (5.81) is bounded in the functional spaces used in Lemma 5.22 of the Appendix.

Since u_0 is smooth, and because of the definition (5.23) of u_1, $L_2 u_1 = \frac{\partial u_1}{\partial t} - \mathrm{div}_x A \nabla_x u_1$ is bounded in $L^2(]0,T[;L^2(\Omega))$. Then, we write

$$\frac{1}{\varepsilon} L_0 u_2 = -\frac{1}{\varepsilon}\mathrm{div}_y(A_\varepsilon \nabla_y u_2) \tag{5.82}$$

$$= -\mathrm{div}(A_\varepsilon \nabla_y u_2) + \mathrm{div}_x(A_\varepsilon \nabla_y u_2). \tag{5.83}$$

The smoothness of u_0 and the definition (5.24) of u_2 imply that

- $\mathrm{div}_x(A_\varepsilon \nabla_y u_2)(x,\frac{x}{\varepsilon},t)$ is bounded in $L^2(]0,T[;L^2(\Omega))$;
- $A_\varepsilon \nabla_y u_2(x,\frac{x}{\varepsilon},t)$ is bounded in $L^\infty(]0,T[;L^2(\Omega))$;
- $\partial_t\left(A_\varepsilon \nabla_y u_2(x,\frac{x}{\varepsilon},t)\right)$ is bounded in $L^1(]0,T[;L^2(\Omega))$.

Consequently all the terms of the right-hand side of (5.81) are bounded in the functional spaces of Lemma 5.22. It follows from Lemma 5.22 that r_ε is bounded in $C([0,T];H^1(\Omega))$, which concludes the proof.

□

For completeness, we check that $u_1^{bil,\varepsilon}$ is not useful if we are not interested in initial instants and points close to the boundary. This is the object of the following result:

Theorem 5.10. *Assume that u_0 is smooth. Consider $0 < \kappa < T$ and an open set $\omega \subset\subset \Omega$. Then*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) \right\|_{C([\kappa,T];H^1(\omega))} \leq C\varepsilon.$$

Proof. The proof is elementary. We write

$$\begin{aligned}
&\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) \right\|_{C([\kappa,T];H^1(\omega))} \\
&\quad \leq \left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) - \varepsilon u_1^{bil,\varepsilon}(x,t) \right\|_{C([\kappa,T];H^1(\omega))} + \varepsilon \left\| u_1^{bil,\varepsilon} \right\|_{C([\kappa,T];H^1(\omega))} \\
&\quad \leq \left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) - \varepsilon u_1^{bil,\varepsilon}(x,t) \right\|_{C([0,T];H^1(\Omega))} + \varepsilon \left\| u_1^{bil,\varepsilon} \right\|_{C([\kappa,T];H^1(\omega))}
\end{aligned}$$

and conclude by using Theorem 5.9 and Lemma 5.11 therafter. □

Lemma 5.11. *Assume that u_0 is smooth. Consider $0 < \kappa < T$ and an open set $\omega \subset\subset \Omega$. It holds*

$$\left\| u_1^{bil,\varepsilon} \right\|_{\mathcal{C}([0,T];L^2(\omega)) \cap L^2(]0,T[;H^1(\omega))} \leq C, \tag{5.84}$$

and

$$\left\| u_1^{bil,\varepsilon} \right\|_{C([\kappa,T];H^1(\omega))} \leq C. \tag{5.85}$$

Proof. Classical regularity results imply that the cell solutions w_i belong to $L^\infty(Q)$. The latter and the smoothness of u_0 yield that the first corrector satisfies

$$\| u_1(x, \frac{x}{\varepsilon}, t) \|_{\mathcal{C}([0,T];L^\infty(\Omega)} \leq C. \tag{5.86}$$

Then, thanks to the bound (5.86) and a weak maximum principle applied to (5.80), we have

$$\| u_1^{bil,\varepsilon} \|_{\mathcal{C}([0,T];L^\infty(\Omega)} \leq C. \tag{5.87}$$

Next, let ϕ be a smooth function in $\mathcal{D}(\Omega)$ such that $\phi = 1$ in ω. Multiplying the first equation of (5.80) by $u_1^{bil,\varepsilon}\phi^2$ and integrating by parts on $\Omega \times (0,t)$ for some $0 < t \leq T$ gives

$$\begin{aligned}
&\frac{1}{2} \int_\Omega (u_1^{bil,\varepsilon})^2(\cdot,t)\phi^2 + \int_\Omega \int_0^t A_\varepsilon \nabla u_1^{bil,\varepsilon} \cdot (\nabla u_1^{bil,\varepsilon})\phi^2 \\
&= \frac{1}{2} \int_\Omega (u_1(x, \frac{x}{\varepsilon}, 0))^2 \phi^2 - \int_\Omega \int_0^t A_\varepsilon (\nabla u_1^{bil,\varepsilon})\phi \cdot u_1^{bil,\varepsilon} \nabla \phi.
\end{aligned} \tag{5.88}$$

We deduce from the smoothness of u_0 that the $L^2(\Omega)$-norm of the initial condition $u_1(x, \frac{x}{\varepsilon}, 0)$, and so the third term of (5.88), are bounded independently of ε. Using the Cauchy-Schwarz inequality and (5.87) in the fourth term of (5.88), and the uniform coerciveness of A_ε in the second term, it is then straightforward to obtain (5.84) from (5.88).

Consider now ψ a smooth function of the time variable such that $\psi = 1$ in (κ, T). Multiplying the first equation of (5.80) by $\frac{\partial u_1^{bil,\varepsilon}}{\partial t}\phi^2\psi^2$ and integrating by parts on $\Omega\times(0,t)$ for some $\kappa \leq t \leq T$, we find that

$$\begin{aligned}&\int_0^t\int_\Omega |\frac{\partial u_1^{bil,\varepsilon}}{\partial t}|^2\phi^2\psi^2 + \frac{1}{2}\int_\Omega A_\varepsilon \nabla u_1^{bil,\varepsilon}(\cdot,t)\cdot\nabla u_1^{bil,\varepsilon}(\cdot,t)\phi^2 \\ &= \int_0^t\int_\Omega A_\varepsilon\nabla u_1^{bil,\varepsilon}\cdot(\nabla u_1^{bil,\varepsilon})\psi\frac{\partial\psi}{\partial t}\phi^2 - \int_0^t\int_\Omega A_\varepsilon\nabla u_1^{bil,\varepsilon}\cdot(\nabla\phi^2)\frac{\partial u_1^{bil,\varepsilon}}{\partial t}\psi^2.\end{aligned} \tag{5.89}$$

Using the Cauchy-Schwarz inequality and (5.84) in both terms of the right-hand side of (5.89), as well as the uniform coerciveness of A_ε in the second term of the left-hand side, we obtain (5.85). □

Remark 5.2. *The weak maximum principle is instrumental in obtaining (5.87). If we were to work not in a scalar setting but with systems of equations (for instance in the context of elasticity), then the compactness method of Avellaneda and Lin [9] may be used, under additional regularity assumptions on A.*

The computation of the boundary+initial layer $u_1^{bil,\varepsilon}$ is as involved as that of u_ε. It is therefore necessary to find a tractable alternative. To this end, we proceed in two steps:

- in the next section, we get rid of the boundaries so as to gain insight on how to design a practical initial layer;
- in Section 5.4.4, we address simultaneously the initial and boundary layers.

5.4.3 Initial layer in an "unbounded" domain

The results given in this section can be derived of those of [68]. We nonetheless believe it is useful to state and prove them here in perhaps a clearer and more concise fashion, all the more so since we will significantly use the same arguments in the next sections.

So as to get rid of the boundaries, we consider here $\Omega = (0,1)^d$ equipped with fully periodic boundary conditions, and thus define u_ε as the solution to

$$\begin{cases}\dfrac{\partial u_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon\nabla u_\varepsilon) = f \ \text{ in } \Omega\times(0,T), \\ x\mapsto u_\varepsilon(x,\cdot) \ \ \mathbb{Z}^d-\text{periodic}, \\ u_\varepsilon(\cdot,0) = g \ \text{ in } \Omega,\end{cases} \tag{5.90}$$

where g is a $\mathbb{Z}^d$-periodic function. We also suppose that the sequence of ε is such that Ω contains an integer number of cells (i.e $\frac{1}{\varepsilon}\in\mathbb{N}$).

The first thing to note is that, because of the periodic boundary conditions, there is no need for boundary layers in the homogenization of (5.90), hence the following estimate in $C([0,T];L^2(\Omega))\cap L^2(]0,T[;H^1(\Omega))$:

Theorem 5.12. *Consider u_ε solution to (5.90), and assume that u_0 is smooth. Then it holds*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) \right\|_{C([0,T];L^2(\Omega))\cap L^2(]0,T[;H^1(\Omega))} \leq C\varepsilon.$$

Proof. We define $r_\varepsilon(x,t) = \frac{1}{\varepsilon}(u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1\left(x, \frac{x}{\varepsilon}, t\right))$. Using system (5.13), and after some rewriting already detailed in the proof of Theorem 5.80, we see that r_ε is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon) = -\operatorname{div}\left(A\nabla_y u_2\right) + \operatorname{div}_x A\nabla_y u_2 \\ \qquad\qquad + \operatorname{div}_x A\nabla_x u_1 - \dfrac{\partial u_1}{\partial t} \ \text{ in } \Omega\times(0,T), \\ x \mapsto r_\varepsilon(x,\cdot) \ \ \mathbb{Z}^d - \text{periodic}, \\ r_\varepsilon(x,0) = -u_1(x, \dfrac{x}{\varepsilon}, 0) \ \text{ in } \Omega. \end{cases} \tag{5.91}$$

The smoothness of u_0 and the definitions (5.23) and (5.24) of u_1 and u_2 imply that

- $A\nabla_y u_2(x, \frac{x}{\varepsilon}, t)$ is bounded in $L^2(]0,T[;L^2(\Omega))$;
- $\operatorname{div}_x A\nabla_y u_2(x, \frac{x}{\varepsilon}, t)$ and $\operatorname{div}_x A\nabla_x u_1(x, \frac{x}{\varepsilon}, t)$ are bounded in $L^1(]0,T[;L^2(\Omega))$;
- $\frac{\partial u_1}{\partial t}(x, \frac{x}{\varepsilon}, t)$ is bounded in $L^1(]0,T[;L^2(\Omega))$;
- $u_1(x, \frac{x}{\varepsilon}, 0)$ is bounded in $L^2(\Omega)$.

Thus all the terms in the right-hand side of (5.91) are bounded in the functional spaces used in Lemma 5.21. It entails from Lemma 5.21 (easily adapted to handle periodic boundary conditions instead of Dirichlet boundary conditions on $\partial\Omega$) that r_ε is bounded in $C([0,T];L^2(\Omega))\cap L^2(]0,T[;H^1(\Omega))$, which concludes the proof.

□

To obtain error estimates in the space $C([0,T];H^1(\Omega))$ rather than in $C([0,T];L^2(\Omega))\cap L^2(]0,T[;H^1(\Omega))$, and since there are no boundaries here, we only have to compensate for the first corrector at $t=0$, in a more tractable way than (5.80).

For this purpose, we will once again use the linear structure of u_1. We follow [68] and define parabolic cell problems by: $\forall i \in [\![1,d]\!]$,

$$\begin{cases} \dfrac{\partial z_i}{\partial \tau} - \operatorname{div}(A\nabla z_i) = 0 \ \text{ in } Q\times\mathbb{R}_+^*, \\ y \mapsto z_i(y,\cdot) \ \ \mathbb{Z}^d - \text{periodic}, \\ z_i(y,0) = -w_i \ \text{ in } Q. \end{cases} \tag{5.92}$$

Recall that each w_i is defined up to the addition of a constant, and therefore so is each z_i.

We start by proving the following elementary properties of the functions z_i.

Lemma 5.13. *For all $i \in [\![1,d]\!]$, there exists a unique solution z_i to (5.92) such that:*

$$z_i \in \mathcal{C}(\mathbb{R}_+; L^2(Q)) \cap L^\infty(\mathbb{R}_+; L^2(Q)), \tag{5.93}$$

$$z_i + \int_Q w_i \in L^2(\mathbb{R}_+; L^2(Q)), \tag{5.94}$$

$$\nabla z_i \in L^2(\mathbb{R}_+; L^2(Q)). \tag{5.95}$$

Moreover, z_i satisfies

$$\frac{\partial z_i}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(Q)) \cap L^1(\mathbb{R}_+; L^2(Q)). \tag{5.96}$$

Proof. For simplicity, we drop indices in this proof and replace z_i and w_i with z and w respectively.

Let us call $(\lambda_k, a_k)_{k\in\mathbb{N}}$ the eigenpairs of the operator $-\mathrm{div}(A\nabla\cdot)$ on Q with periodic boundary conditions. It is well known that $(a_k)_{k\in\mathbb{N}}$ is an orthonormal basis of $L^2(Q)$ and that we can arrange these pairs in such a way that the sequence λ_k is nondecreasing and goes to infinity when k goes to infinity. Moreover we have $(\lambda_0, a_0) = (0, 1)$ and $\lambda_1 > 0$.

The periodic function w can be expanded in the basis $(a_k)_{k\in\mathbb{N}}$ as

$$w(y) = \sum_{k\in\mathbb{N}} c_k a_k(y) \tag{5.97}$$

with $c_k = \displaystyle\int_Q w(y)a_k(y)dy$ for all $k \in \mathbb{N}$. Note that (5.97) implies

$$\|w\|^2_{L^2(Q)} = \sum_{k\in\mathbb{N}} c_k^2. \tag{5.98}$$

It is classical that there is a unique solution z to (5.92) such that (5.93), (5.94) and (5.95) hold, which writes

$$z(y,\tau) = -\sum_{k\in\mathbb{N}} c_k e^{-\lambda_k\tau} a_k(y). \tag{5.99}$$

The sequel is devoted to proving (5.96).

For a given τ, it follows from (5.99) that

$$\left\|\frac{\partial z}{\partial \tau}(\cdot,\tau)\right\|^2_{L^2(Q)} = \sum_{k\in\mathbb{N}^*} \lambda_k^2 c_k^2 e^{-2\lambda_k\tau}. \tag{5.100}$$

Then

$$\left\|\frac{\partial z}{\partial \tau}\right\|^2_{L^2(\mathbb{R}_+;L^2(Q))} = \sum_{k\in\mathbb{N}^*} \frac{1}{2} c_k^2 \lambda_k = \frac{1}{2}\int_Q A\nabla w \cdot \nabla w < +\infty.$$

Therefore

$$\frac{\partial z}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(Q)), \tag{5.101}$$

and as an immediate consequence

$$\frac{\partial z}{\partial \tau} \in L^1(]0,1[; L^2(Q)). \tag{5.102}$$

Next, we rewrite (5.100) as

$$\left\| \frac{\partial z}{\partial \tau}(\cdot,\tau) \right\|^2_{L^2(Q)} = \sum_{k\in\mathbb{N}^*} \frac{c_k^2}{\lambda_k} \lambda_k^3 e^{-2\lambda_k \tau}. \tag{5.103}$$

For $\tau > 0$, the function $x \in \mathbb{R}_+ \mapsto x^3 e^{-\tau x}$ reaches its maximum at $x = \frac{1}{\tau}$. Hence we deduce from (5.103) that

$$\begin{aligned} \left\| \frac{\partial z}{\partial \tau}(\cdot,\tau) \right\|^2_{L^2(Q)} &\leq \frac{e^{-1}}{8\tau^3} \sum_{k\in\mathbb{N}^*} \frac{c_k^2}{\lambda_k} \\ &\leq \frac{C}{\lambda_1 \tau^3} \sum_{k\in\mathbb{N}^*} c_k^2 \\ &\leq \frac{C}{\lambda_1 \tau^3} \|w\|^2_{L^2(Q)}, \end{aligned} \tag{5.104}$$

so that

$$\left\| \frac{\partial z}{\partial \tau}(\cdot,\tau) \right\|_{L^2(Q)} \leq \frac{C}{\tau^{3/2}} \|w\|_{L^2(Q)}. \tag{5.105}$$

Since $\frac{1}{\tau^{3/2}} \in L^1(]1,+\infty[)$, (5.105) implies

$$\frac{\partial z}{\partial \tau} \in L^1(]1,+\infty[; L^2(Q)). \tag{5.106}$$

We conclude from (5.101), (5.102) and (5.106) that $\frac{\partial z}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(Q)) \cap L^1(\mathbb{R}_+; L^2(Q))$, which proves (5.96). □

We can now define our candidate as an initial layer by

$$u_z(x, \frac{x}{\varepsilon}, \frac{t}{\varepsilon^2}) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x,0) z_i(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2}). \tag{5.107}$$

The following result shows that u_z is a relevant initial layer. The proof we give is different from that found in [68].

Theorem 5.14. *Consider u_ε solution to (5.90) and assume that u_0 is smooth. Then*

$$\left\| u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) - \varepsilon u_z(x, \frac{x}{\varepsilon}, \frac{t}{\varepsilon^2}) \right\|_{C([0,T];H^1(\Omega))} \leq C\varepsilon.$$

Proof. We define the remainder

$$r_\varepsilon = \varepsilon^{-1}\left(u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t) - \varepsilon u_z(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})\right).$$

Our aim is to prove that r_ε is bounded in $C([0,T];H^1(\Omega))$.

Remark that the definition (5.107) of u_z and the smoothness of u_0 imply that $r_\varepsilon = 0$ at $t = 0$. Using the parabolic operators (5.21) and the system (5.13), the function r_ε is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon}{\partial t} - \mathrm{div} A_\varepsilon \nabla r_\varepsilon = \varepsilon^{-1} L_0 u_2 - L_2 u_1 - \varepsilon^{-1} L_1 u_z - L_2 u_z \ \text{ in } \Omega \times (0,T), \\ x \mapsto r_\varepsilon(x,\cdot) \ \ \mathbb{Z}^d - \text{periodic}, \\ r_\varepsilon = 0 \ \text{ at } t = 0. \end{cases}$$

We decompose $r_\varepsilon = r_\varepsilon^1 + r_\varepsilon^2$ with r_ε^1 solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^1}{\partial t} - \mathrm{div} A_\varepsilon \nabla r_\varepsilon^1 = \varepsilon^{-1} L_0 u_2 - L_2 u_1 \ \text{ in } \Omega \times (0,T), \\ x \mapsto r_\varepsilon^1(x,\cdot) \ \ \mathbb{Z}^d - \text{periodic}, \\ r_\varepsilon^1 = 0 \ \text{ at } t = 0, \end{cases} \tag{5.108}$$

and r_ε^2 solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^2}{\partial t} - \mathrm{div} A_\varepsilon \nabla r_\varepsilon^2 = -\varepsilon^{-1} L_1 u_z - L_2 u_z \ \text{ in } \Omega \times (0,T), \\ x \mapsto r_\varepsilon^2(x,\cdot) \ \ \mathbb{Z}^d - \text{periodic}, \\ r_\varepsilon^2 = 0 \ \text{ at } t = 0. \end{cases} \tag{5.109}$$

Following the proof of Theorem 5.9, it is straightforward to see that r_ε^1 is bounded in $C([0,T];H^1(\Omega))$. The main difficulty lies in the term r_ε^2, to which we devote the rest of this proof. The latter consists in showing that the right-hand side of (5.109) is bounded in the functional spaces of Lemma 5.22.

For this purpose, we rewrite

$$\begin{aligned} -\varepsilon^{-1} L_1 u_z - L_2 u_z &= \varepsilon^{-1}\mathrm{div}_x A \nabla_y u_z + \varepsilon^{-1}\mathrm{div}_y A \nabla_x u_z + \mathrm{div}_x\left(A\nabla_x u_z\right) \\ &= \varepsilon^{-1}\mathrm{div}_x A \nabla_y u_z + \mathrm{div}(A\nabla_x u_z). \end{aligned} \tag{5.110}$$

We first deal with $\varepsilon^{-1}\mathrm{div}_x A\nabla_y u_z$ in (5.110). We know from (5.95) in Lemma 5.13 that $\nabla z_i(y,\tau) \in L^2(\mathbb{R}_+;L^2(Q))$. Note then that by $\mathbb{Z}^d$- periodicity and a scaling argument, we have, for all $i \in [\![1,d]\!]$,

$$\begin{aligned} \frac{1}{\varepsilon}\|\nabla_y z_i(\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})\|_{L^2(]0,T[;L^2(\Omega))} &= \|\nabla z_i\|_{L^2(]0,\frac{T}{\varepsilon^2}[;L^2(Q))} \\ &\le \|\nabla z_i\|_{L^2(\mathbb{R}_+;L^2(Q))}. \end{aligned} \tag{5.111}$$

It follows from (5.111) that $\frac{1}{\varepsilon}\nabla_y z_i(\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})$ is bounded in $L^2(]0,T[;L^2(\Omega))$. In view of the definition (5.107) of u_z, and using additionally the smoothness of u_0, this implies that $\varepsilon^{-1}\mathrm{div}_x A\nabla_y u_z(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})$ is bounded in $L^2(]0,T[;L^2(\Omega))$.

We now focus on $\mathrm{div}(A\nabla_x u_z)$ in (5.110). According to Lemma 5.22, we have to prove that $A\nabla_x u_z(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})$ is bounded in $C([0,T];L^2(\Omega))$ and that $\frac{\partial}{\partial t}(A\nabla_x u_z\left(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2}\right))$ is bounded in $L^1(]0,T[;L^2(\Omega))$. The former easily comes from the smoothness of u_0 and (5.93) in Lemma 5.13. As for the latter, by $\mathbb{Z}^d$- periodicity and a scaling argument, we note that

$$\begin{aligned}\frac{1}{\varepsilon^2}\|\frac{\partial z_i}{\partial\tau}(\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})\|_{L^1(]0,T[;L^2(\Omega))} &= \|\frac{\partial z_i}{\partial\tau}\|_{L^1(]0,\frac{T}{\varepsilon^2}[;L^2(Q))}\\ &\le \|\frac{\partial z_i}{\partial\tau}\|_{L^1(\mathbb{R}_+;L^2(Q))}.\end{aligned} \tag{5.112}$$

We deduce from (5.112) and (5.96) in Lemma 5.13 that $\frac{1}{\varepsilon^2}\frac{\partial z_i}{\partial\tau}(\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})$ is bounded in the space $L^1(]0,T[;L^2(\Omega))$. Using on the other hand the smoothness of u_0, we obtain that $\frac{\partial}{\partial t}(A\nabla_x u_z\left(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2}\right))$ is bounded in $L^1(]0,T[;L^2(\Omega))$.

Consequently, all the terms in the right-hand side of (5.109) are bounded in the functional spaces of Lemma 5.22. We can then conclude from Lemma 5.22 that r_ε^2 is bounded in $C([0,T];H^1(\Omega))$.

It follows from the decomposition $r_\varepsilon = r_\varepsilon^1 + r_\varepsilon^2$ that r_ε is bounded in $C([0,T];H^1(\Omega))$, which is the desired result. □

According to (5.99), the functions z_i decay exponentially with respect to τ. In practice, it is sufficient to compute them only for small τ and to assume that they vanish afterwards. These computations are easy since they take place in the unit cell Q.

Once the parabolic cell solutions z_i have been computed, the initial layer u_z is obtained in a straightforward manner from (5.107). Recalling that $\tau = \frac{t}{\varepsilon^2}$, it has only to be added to the Ansatz (5.19) for small t, for its contribution becomes negligible after initial instants. It is then much more convenient for practical purposes than the layer (5.80).

Now that we have an efficient method to compute an initial layer, and also, from Section 5.3.1, tractable boundary layers for stationary problems, we seek in the next section to address the full issue of boundary+initial layers.

5.4.4 General case

In this final section we wish to use our knowledge of the stationary setting and of the parabolic setting without boundaries to handle simultaneously the initial condition and the boundary conditions and to find relevant boundary+initial layers.

In the sequel, we use the notation of Section 5.3.1 and more precisely of Figures 5.1 and 5.2. We assume that $\Omega = (0,1)^d$, and work in the space $H_D(\Omega)$ defined by (5.30).

This means that u_ε is solution to

$$\begin{cases} \dfrac{\partial u_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla u_\varepsilon) = f(x,t) \ \text{ in } \Omega \times (0,T), \\ u_\varepsilon = 0 \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto u_\varepsilon(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ u_\varepsilon(\cdot, 0) = g \ \text{ in } \Omega, \end{cases} \tag{5.113}$$

with $g \in H_D(\Omega)$.

Let us recall that in the previous section we have obtained an initial layer u_z defined by (5.107). On the other hand, in the stationary setting of Section 5.3.1, we have obtained a boundary layer u_D^{bl} defined by (5.33). We need to slightly modify the latter to account for the introduction of the time variable, and define the new Dirichlet boundary layer by

$$u_D^{bl}(x, \frac{x}{\varepsilon}, t) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x,t) \left(\chi^1(x) \psi_D^{i,1}(\frac{x}{\varepsilon}) + \chi^2(x) \psi_D^{i,2}(\frac{x'}{\varepsilon}, \frac{1-x_d}{\varepsilon}) \right), \tag{5.114}$$

where for all $i \in [\![1,d]\!]$ and all $j \in [\![1,2]\!]$, $\psi_D^{i,j}$ is defined by (5.32) and χ^j is the same cut-off function as in (5.33).

We have seen that at order one in ε, the need for an initial layer comes from the oscillation of the first corrector at $t = 0$. Before going further, we just check that when there is no initial oscillation of the corrector, the boundary layer u_D^{bl} suffices to obtain the desired error estimate in $C([0,T]; H^1(\Omega))$.

Theorem 5.15. *Consider u_ε solution to (5.113) with $g = 0$. Assume that the homogenized solution u_0 is smooth. Then*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x, \frac{x}{\varepsilon}, t) - \varepsilon u_D^{bl}(x, \frac{x}{\varepsilon}, t) \right\|_{C([0,T];H^1(\Omega))} \leq C\varepsilon.$$

Proof. Let us define the remainder $r_\varepsilon = \varepsilon^{-1}(u_\varepsilon - u_0 - \varepsilon u_1 - \varepsilon u_D^{bl})$.

Since $u_\varepsilon = 0$ at $t = 0$, the same is true for u_0. The smoothness of u_0 then implies that u_1 and u_D^{bl} are equal to 0 at $t = 0$. It follows that $r_\varepsilon = 0$ at $t = 0$.

Using the operators (5.21) and the system (5.13), we find that r_ε is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon) = \dfrac{1}{\varepsilon} L_0 u_2 - L_2 u_1 \\ \qquad\qquad - \dfrac{\partial}{\partial t} u_D^{bl}(x, \dfrac{x}{\varepsilon}, t) + \operatorname{div}\left(A_\varepsilon \nabla \left(u_D^{bl}(x, \dfrac{x}{\varepsilon}, t) \right) \right) \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon = 0 \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon(\cdot, 0) = 0 \ \text{ in } \Omega. \end{cases} \tag{5.115}$$

We decompose $r_\varepsilon = r_\varepsilon^1 + r_\varepsilon^2$, with r_ε^1 solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^1}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon^1) = \dfrac{1}{\varepsilon} L_0 u_2 - L_2 u_1 - \dfrac{\partial}{\partial t} u_D^{bl} \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon^1 = 0 \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon^1(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon^1(\cdot, 0) = 0 \ \text{ in } \Omega, \end{cases} \tag{5.116}$$

and r_ε^2 solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^2}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon^2(x,t)) = \operatorname{div}\left(A_\varepsilon \nabla \left(u_D^{bl}(x, \dfrac{x}{\varepsilon}, t)\right)\right) \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon^2 = 0 \ \text{ on } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon^2(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon^2(\cdot, 0) = 0 \ \text{ in } \Omega. \end{cases} \tag{5.117}$$

Using Lemma 5.22, we show exactly as in the proof of Theorem 5.9 that r_ε^1 is bounded in $C([0,T]; H^1(\Omega))$.

In order to deal with r_ε^2, we need to proceed a bit differently. For $t \in [0,T]$, multiplying (5.117) by $\frac{\partial r_\varepsilon^2}{\partial t}$ and integrating by parts, we obtain the energy equality

$$\begin{aligned} &\int_\Omega \int_0^t \left| \frac{\partial r_\varepsilon^2}{\partial t} \right|^2 + \frac{1}{2} \int_\Omega A_\varepsilon \nabla r_\varepsilon^2(\cdot,t) \cdot \nabla r_\varepsilon^2(\cdot,t) \\ &= \int_\Omega \int_0^t A_\varepsilon \frac{\partial \nabla \left(u_D^{bl}(x, \frac{x}{\varepsilon}, s)\right)}{\partial t} \cdot \nabla r_\varepsilon^2(x,s) - \int_\Omega A_\varepsilon \nabla \left(u_D^{bl}(x, \frac{x}{\varepsilon}, t)\right) \cdot \nabla r_\varepsilon^2(\cdot,t). \end{aligned} \tag{5.118}$$

The same arguments as those used in the study of boundary layers in the stationary setting, either in the proof of Theorem 5.5 (see [2] or [64]) or in the proof of Theorem 5.8, apply. In particular, we derive an expression similar to formula (5.65) where the boundary integrals are zero for we here deal with homogeneous Dirichlet boundary conditions. This leads to

$$\left| \int_\Omega A_\varepsilon \nabla \left(u_D^{bl}(x, \frac{x}{\varepsilon}, t)\right) \cdot \nabla \phi \right| \leq C \|\nabla \phi\|_{L^2(\Omega)}. \tag{5.119}$$

Hence, taking $\phi = r_\varepsilon^2$ in (5.119), we get

$$\left| \int_\Omega A_\varepsilon \nabla \left(u_D^{bl}(x, \frac{x}{\varepsilon}, t)\right) \cdot \nabla r_\varepsilon^2(x,t) \right| \leq C \|\nabla r_\varepsilon^2(\cdot,t)\|_{L^2(\Omega)}. \tag{5.120}$$

Likewise, we have

$$\left| \int_\Omega \int_0^t A_\varepsilon \frac{\partial \nabla \left(u_D^{bl}(x, \frac{x}{\varepsilon}, s)\right)}{\partial t} \cdot \nabla r_\varepsilon^2(x,s) \right| \leq C \|\nabla r_\varepsilon^2\|_{C([0,T];L^2(\Omega))}. \tag{5.121}$$

Using (5.120) and (5.121) in the right-hand side of (5.118), and the uniform coerciveness of A_ε in the left-hand side, we find that $\|r_\varepsilon^2\|_{C([0,T];H^1(\Omega))} \leq C\sqrt{\varepsilon}$ and a fortiori that r_ε^2 is bounded in $C([0,T]; H^1(\Omega))$, which concludes the proof since $r_\varepsilon = r_\varepsilon^1 + r_\varepsilon^2$. □

Let us sum up where we have come to. Theorem 5.14 in Section 5.4.3 gives an initial layer for parabolic problems with no boundary layers. Theorem 5.15 above gives a boundary layer for parabolic problems with no initial layer. For the sake of comprehensiveness, we emphasize that these two results are readily adapted, in the same fashion as in Theorem 5.10, to address problem (5.113) when we get rid of boundaries and of initial instants respectively, as stated thereafter:

Theorem 5.16. *Consider u_ε solution to (5.113), and an open subset $\omega \subset\subset \Omega$. Assume that the homogenized solution u_0 is smooth. Then*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t) - \varepsilon u_z(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2}) \right\|_{\mathcal{C}([0,T];H^1(\omega))} \leq C\varepsilon.$$

Theorem 5.17. *Consider u_ε solution to (5.113), and $0 < \kappa < T$. Assume that the homogenized solution u_0 is smooth. Then*

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t) - \varepsilon u_D^{bl}(x,\frac{x}{\varepsilon},t) \right\|_{\mathcal{C}([\kappa,T];H^1(\Omega))} \leq C\varepsilon.$$

We now want to find the general boundary+initial layer. An intuitive way to proceed is to add the initial layer (5.107) and the boundary layer (5.114), since the former is needed when there is no boundary and the latter is needed when there is no initial oscillation. However it is clearly not sufficient to add them, for the boundary layer violates the initial condition, while the initial layer violates the Dirichlet boundary condition on $\Gamma_1 \cup \Gamma_2$. Therefore we have to add a new term that corrects the unwanted effects induced by both layers, and that compensates for u_z on $\partial\Omega$ and for u_D^{bl} at $t = 0$.

Due to the linear structure of the boundary and initial layers, we define this new term by

$$u_{Dz}(x,\frac{x}{\varepsilon},t,\frac{t}{\varepsilon^2}) = \sum_{i=1}^{d} \frac{\partial u_0}{\partial x_i}(x,t) \left(p_i^1(\frac{x}{\varepsilon},\frac{t}{\varepsilon^2})\chi^1(x) + p_i^2(\frac{x'}{\varepsilon},\frac{1-x_d}{\varepsilon},\frac{t}{\varepsilon^2})\chi^2(x) \right) \quad (5.122)$$

where for all $i \in [\![1,d]\!]$ and all $j \in [\![1,2]\!]$, $\psi_D^{i,j}$ is defined by (5.32), p_i^j solves

$$\begin{cases} \dfrac{\partial p_i^j}{\partial \tau} - \mathrm{div} A\nabla p_i^j = 0 \ \text{ in } G^j \times \mathbb{R}_+^*, \\ p_i^j(y,\tau) = -z_i(y,\tau) \ \text{ on } \Gamma, \\ y' \mapsto p_i^j(y',y_d,\tau) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ p_i^j(\cdot,0) = -\psi_D^{i,j} \ \text{ in } G^j, \end{cases} \quad (5.123)$$

and χ^j is the same cut-off function as in (5.114).

Note that the initial and boundary conditions in (5.123) are compatible at $y_d = 0$ and $\tau = 0$, because $\psi_D^{i,j}$ is equal to w_i on the edge Γ and z_i is equal to w_i at $\tau = 0$. Consequently we can expect some regularity of p_i^j. On the other hand, since we know from (5.99) that z_i decays exponentially in function of τ to a constant , and from Lemma 5.4 that $\psi_D^{i,j}$ decays exponentially in function of y_d to a constant, we expect some integrability in space and time of the derivatives of p_i^j. We actually have the following existence result:

Lemma 5.18. *For all $i \in [\![1,d]\!]$ and all $j \in [\![1,2]\!]$, there exists a unique solution p_i^j to (5.123) such that*

$$\frac{\partial p_i^j}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(G^j)), \tag{5.124}$$

$$\nabla p_i^j \in L^\infty(\mathbb{R}_+; L^2(G^j)) \cap C(\mathbb{R}_+; L^2(G^j)). \tag{5.125}$$

Moreover,

$$p_i^j \in L^\infty(\mathbb{R}_+; L^\infty(G^j)). \tag{5.126}$$

Proof. Let ψ be a smooth test function depending only on y_d such that $\psi(0) = 1$ and ψ vanishes for $|y_d| \geq 1$. For all $i \in [\![1,d]\!]$ and all $j \in [\![1,2]\!]$, we look at the following problem:

$$\begin{cases} \dfrac{\partial q_i^j}{\partial \tau} - \mathrm{div} A \nabla q_i^j = -\mathrm{div}(A\, z_i \nabla \psi) - A \nabla z_i \cdot \nabla \psi \ \text{ in } G^j \times \mathbb{R}_+^*, \\ q_i^j(y,\tau) = 0 \ \text{ on } \Gamma, \\ y' \mapsto q_i^j(y', y_d, \tau) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ q_i^j(\cdot, 0) = -\psi_D^{i,j} + \psi w_i \ \text{ in } G^j. \end{cases} \tag{5.127}$$

The cell solutions w_i being defined up to the addition of a constant, we can assume without loss of generality that their average over the unit cell Q is 0. It follows from (5.97) and (5.99) that z_i and ∇z_i decay exponentially to 0 with respect to τ. Applying Lemma 5.22 to (5.127) for T as large as we want, we find that there exists a unique solution q_i^j to (5.127) such that

$$\begin{aligned} &\frac{\partial q_i^j}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(G^j)), \\ &\nabla q_i^j \in L^\infty(\mathbb{R}_+; L^2(G^j)) \cap C(\mathbb{R}_+; L^2(G^j)). \end{aligned} \tag{5.128}$$

On the other hand, it obviously holds by definition of ψ and due to the exponential decay of z_i that

$$\begin{aligned} &\psi \frac{\partial z_i}{\partial \tau} \in L^2(\mathbb{R}_+; L^2(G^j)), \\ &\nabla(\psi z_i) \in L^\infty(\mathbb{R}_+; L^2(G^j)) \cap C(\mathbb{R}_+; L^2(G^j)). \end{aligned} \tag{5.129}$$

We then define

$$p_i^j = q_i^j - \psi z_i. \tag{5.130}$$

In view of (5.128) and (5.129), p_i^j satisfies (5.124) and (5.125), and an easy calculation shows that it is solution to (5.123). The uniqueness of p_i^j follows from Lemma 5.22.

Finally a weak maximum principle applied to (5.123) yields (5.126).

□

Remark 5.3. *If we choose to normalize the cell solutions so that their average over the unit cell Q is zero, then the boundary condition of (5.123), that is z_i, converges to zero when τ goes to infinity, hence p_i^j goes to zero when τ goes to infinity. In general, p_i^j converges to $\int_Q w_i$.*

The integrability given by (5.124) and (5.125) is not sufficient for our purposes. Indeed, our main result, which consists of an error estimate obtained by adding the layer u_{Dz} defined by (5.122) to the Ansatz, relies on stronger integrability assumptions:

Theorem 5.19. *Consider u_ε solution to (5.113). Assume that u_0 is smooth, and that for all $i \in [\![1,d]\!]$ and all $j \in [\![1,2]\!]$, p_i^j satisfies*

$$\frac{\partial p_i^j}{\partial \tau} \in L^1(\mathbb{R}_+; L^2(G^j)), \tag{5.131}$$

$$\nabla p_i^j \in L^2(\mathbb{R}_+; L^2(G^j)). \tag{5.132}$$

Then it holds

$$\left\| u_\varepsilon(x,t) - u_0(x,t) - \varepsilon u_1(x,\frac{x}{\varepsilon},t) - \varepsilon u_D^{bl}(x,\frac{x}{\varepsilon},t) - \varepsilon u_z(x,\frac{x}{\varepsilon},\frac{t}{\varepsilon^2}) - \varepsilon u_{Dz}(x,\frac{x}{\varepsilon},t,\frac{t}{\varepsilon^2}) \right\|_{C([0,T];H^1(\Omega))} \leq C\varepsilon. \tag{5.133}$$

Proof. Let us introduce the remainder $r_\varepsilon = \frac{1}{\varepsilon}\left(u_\varepsilon - u_0 - \varepsilon u_1 - \varepsilon u_D^{bl} - \varepsilon u_z - \varepsilon u_{Dz}\right)$. Using the operators (5.21) and the system (5.13), we find that r_ε is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon) = \dfrac{1}{\varepsilon} L_0 u_2 - L_2(u_1 + u_D^{bl} + u_z + u_{Dz}) \\ \qquad\qquad - \dfrac{1}{\varepsilon} L_1(u_D^{bl} + u_z + u_{Dz}) \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon = 0 \ \text{ in } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon(\cdot, 0) = 0 \ \text{ in } \Omega. \end{cases} \tag{5.134}$$

We then decompose $r_\varepsilon = r_\varepsilon^1 + r_\varepsilon^2$ where r_ε^1 is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^1}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon^1) = \dfrac{1}{\varepsilon} L_0 u_2 - L_2(u_1 + u_D^{bl} + u_z) - \dfrac{1}{\varepsilon} L_1(u_D^{bl} + u_z) \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon^1 = 0 \ \text{ in } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon^1(x', x_d, t) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon^1(\cdot, 0) = 0 \ \text{ in } \Omega, \end{cases} \tag{5.135}$$

and r_ε^2 is solution to

$$\begin{cases} \dfrac{\partial r_\varepsilon^2}{\partial t} - \operatorname{div}(A_\varepsilon \nabla r_\varepsilon^2) = -L_2 u_{Dz} - \dfrac{1}{\varepsilon} L_1 u_{Dz} \ \text{ in } \Omega \times (0,T), \\ r_\varepsilon^2 = 0 \ \text{ in } \Gamma_1 \cup \Gamma_2, \\ x' \mapsto r_\varepsilon^2(x', x_d, \cdot) \ \ \mathbb{Z}^{d-1} - \text{periodic}, \\ r_\varepsilon^2(\cdot, 0) = 0 \ \text{ in } \Omega. \end{cases} \tag{5.136}$$

It follows from the same arguments as in the proofs of Theorem 5.14 (regarding the initial layer u_z) and Theorem 5.15 (concerning the boundary layer u_D^{bl}) that r_ε^1 is bounded

in $C([0,T];H^1(\Omega))$. Therefore we only have to consider r_ε^2.

The rest of the proof is devoted to verifying that the terms in the right-hand side of (5.136) are bounded in the functional spaces of Lemma 5.22. To this end we will use (5.126) and assumptions (5.131) and (5.132).

First, we rewrite

$$\begin{aligned} -L_2 u_{Dz} - \frac{1}{\varepsilon} L_1 u_{Dz} &= \varepsilon^{-1}\mathrm{div}_x A\nabla_y u_{Dz} + \varepsilon^{-1}\mathrm{div}_y A\nabla_x u_{Dz} + \mathrm{div}_x A\nabla_x u_{Dz} - \frac{\partial}{\partial t} u_{Dz} \\ &= \varepsilon^{-1}\mathrm{div}_x A\nabla_y u_{Dz} + \mathrm{div}(A\nabla_x u_{Dz}) - \frac{\partial}{\partial t} u_{Dz}. \end{aligned} \tag{5.137}$$

It is clear from the definition (5.122) of u_{Dz}, the smoothness of u_0 and (5.126) that $\frac{\partial}{\partial t} u_{Dz}(x, \frac{x}{\varepsilon}, t, \frac{t}{\varepsilon^2})$ in (5.137) is bounded in $L^2(]0,T[;L^2(\Omega))$.

Next, we deal with the term $\varepsilon^{-1}\mathrm{div}_x A\nabla_y u_{Dz}$ in (5.137). Using $\mathbb{Z}^{d-1}$- periodicity and a scaling argument, we obtain, for all $i \in [\![1,d]\!]$,

$$\frac{1}{\varepsilon}\|\nabla_y p_i^1(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2})\|_{L^2(]0,T[;L^2(\Omega))} \leq \sqrt{\varepsilon}\,\|\nabla_y p_i^1(y,\tau)\|_{L^2(\mathbb{R}_+;L^2(G^1))}. \tag{5.138}$$

It follows from (5.138) and assumption (5.132) that $\frac{1}{\varepsilon}\nabla_y p_i^1(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2})$ is bounded in the space $L^2(]0,T[;L^2(\Omega))$. The same is obviously true for $\frac{1}{\varepsilon}\nabla_y p_i^2(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2})$. In view of (5.122), and using additionally the smoothness of u_0, this implies that $\varepsilon^{-1}\mathrm{div}_x A\nabla_y u_{Dz}(x, \frac{x}{\varepsilon}, t, \frac{t}{\varepsilon^2})$ is bounded in $L^2(]0,T[;L^2(\Omega))$.

Finally, we focus on $\mathrm{div}(A\nabla_x u_{Dz})$ in (5.137). According to Lemma 5.22, we have to prove that $A\nabla_x u_{Dz}(x, \frac{x}{\varepsilon}, t, \frac{t}{\varepsilon^2})$ is bounded in $C([0,T];L^2(\Omega))$ and $\frac{\partial}{\partial t}(A\nabla_x u_{Dz}\left(x, \frac{x}{\varepsilon}, t, \frac{t}{\varepsilon^2}\right))$ is bounded in $L^1(]0,T[;L^2(\Omega))$. The former easily comes from the smoothness of u_0 and (5.126). Regarding the latter, we compute, for all $i \in [\![1,d]\!]$,

$$\frac{1}{\varepsilon^2}\|\frac{\partial p_i^1}{\partial \tau}(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2})\|_{L^1(]0,T[;L^2(\Omega))} \leq \sqrt{\varepsilon}\,\|\frac{\partial p_i^1}{\partial \tau}\|_{L^1(\mathbb{R}_+;L^2(G^1))}. \tag{5.139}$$

Collecting (5.139) and assumption (5.131) yields that $\frac{1}{\varepsilon^2}\frac{\partial p_i^1}{\partial \tau}(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2})$ is bounded in the space $L^1(]0,T[;L^2(\Omega))$. The same holds for p_i^2. We then deduce from the definition (5.122) of u_{Dz}, the smoothness of u_0 and (5.126) that $\frac{\partial}{\partial t}(A\nabla_x u_{Dz}\left(x, \frac{x}{\varepsilon}, t, \frac{t}{\varepsilon^2}\right))$ is bounded in $L^1(]0,T[;L^2(\Omega))$.

We have thus proved that all terms in the right-hand side of (5.136) are bounded in the functional spaces of Lemma 5.22, which allows us to conclude, using the latter theorem, that r_ε^2 is bounded in $C([0,T];H^1(\Omega))$.

Finally, r_ε being equal to $r_\varepsilon^1 + r_\varepsilon^2$, it is bounded in $C([0,T];H^1(\Omega))$, which terminates the proof. □

Theorem 5.19 is only relevant from a practical point of view if assumptions (5.131) and (5.132) are satisfied. These assumptions ensure that the decay with respect to space

and time of the derivatives of the function p_i^j solution to (5.123) is sufficiently strong. As mentioned previously, they seem reasonable since the boundary condition z_i in (5.123) has exponential decay in time and the initial condition $\psi_D^{i,j}$ has exponential decay in space. Unfortunately, we are not able to prove that these assumptions hold.

Let us explain the main difficulty. According to (5.99), the function z_i converges when $\tau \to \infty$ to a constant which is $-\int_Q w_i$; on the other hand $\psi_D^{i,j}$ converges when $|y_d| \to +\infty$ to the constant $d^{i,j}$ defined in Lemma 5.4. These constants depend linearly on w_i since z_i and $\psi_D^{i,j}$ depend linearly on w_i. If we normalize w_i so that its average over Q is 0, then the boundary condition z_i goes to 0 when $\tau \to \infty$, but the initial condition $\psi_D^{i,j}$ is not integrable for a priori $d^{i,j} \neq 0$, and we cannot use Lemma 5.21 to obtain (5.132). The converse is true if we normalize w_i to make $\psi_D^{i,j}$ integrable. Thus it is not possible to choose the cell solutions in order to have nice properties on z_i and $\psi_D^{i,j}$ simultaneously. Somehow, the initial layer and the boundary layers are not compatible at infinity, i.e in the limits $\tau \to \infty$ and $|y_d| \to +\infty$.

Of course, it is possible to use $\psi_D^{i,j} - d^{i,j}$ instead of $\psi_D^{i,j}$ as boundary layer, and $z_i + \int_Q w_i$ instead of z_i as initial layer (the estimate of Theorems 5.5 and 5.14 still hold). This removes the issue of compatibility at infinity, but at the expense of having boundary and initial layers that are not compatible anymore at $\tau = 0$ and $y_d = 0$. As a result the proof of Lemma 5.18 does not apply anymore, and we lose properties (5.124) and (5.125) and the regularity of the function p_i^j.

We stress however that assumptions (5.131) and (5.132) are sufficient for our purposes but not necessary. Indeed, it readily follows from the proof of Theorem 5.19 that we only need the left-hand sides of (5.138) and (5.139) to be bounded independently of ε for the error estimate (5.133) to hold. Thus assumptions (5.131) and (5.132) can be replaced with less demanding assumptions. This is the object of the following Corollary.

Corollary 5.20. *Assume that u_0 is smooth, and that there exists a constant $C > 0$ such that for all $i \in [\![1, d]\!]$, all $j \in [\![1, 2]\!]$ and all $\varepsilon > 0$, p_i^j satisfies*

$$\frac{1}{\varepsilon^2} \| \frac{\partial p_i^j}{\partial \tau}(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2}) \|_{L^1(]0,T[;L^2(\Omega))} \leq C, \tag{5.140}$$

$$\frac{1}{\varepsilon} \| \nabla_y p_i^j(\frac{x}{\varepsilon}, \frac{t}{\varepsilon^2}) \|_{L^2(]0,T[;L^2(\Omega))} \leq C. \tag{5.141}$$

Then estimate (5.133) holds.

Albeit less demanding, assumptions (5.140) and (5.141) also seem less natural to us. Besides, we are not able to prove that they are satisfied either.

So as to gain some insight on the relevance of assumptions (5.131) and (5.132) of Theorem 5.19, and assumptions (5.140) and (5.141) of Corollary 5.20, we consider in the next section a one-dimensional setting allowing some explicit computations. Since problem (5.123) is posed on a strip, we believe that the one-dimensional case is representative of what happens in higher dimensions.

5.4.5 One-dimensional toy model

Let us consider problem (5.113) with $\Omega = (0,1)$. There is only one cell solution to (5.4), which we denote w. The boundary $\partial\Omega$ only consists of the two points $\{0\}$ and $\{1\}$.

It is then straightforward to see that the function p associated to the boundary $x = 0$, generally defined by (5.123), is here solution to

$$\begin{cases} \dfrac{\partial p}{\partial \tau} - \mathrm{div} A \nabla p = 0 \ \text{ in } \mathbb{R}_+^* \times \mathbb{R}_+^*, \\ p(0,\tau) = -z(0,\tau) \ \text{ in } \mathbb{R}_+^*, \\ p(\cdot,0) = w(0) \ \text{ in } \mathbb{R}_+^*, \end{cases} \tag{5.142}$$

with z defined by (5.92).

Denoting by $\tilde{p} = p - w(0)$, we have

$$\begin{cases} \dfrac{\partial \tilde{p}}{\partial \tau} - \mathrm{div} A \nabla \tilde{p} = 0 \ \text{ in } \mathbb{R}_+^* \times \mathbb{R}_+^*, \\ \tilde{p}(0,\tau) = -w(0) - z(0,\tau) \ \text{ in } \mathbb{R}_+^*, \\ \tilde{p}(\cdot,0) = 0 \ \text{ in } \mathbb{R}_+^*. \end{cases} \tag{5.143}$$

Obviously, p satisfies assumptions (5.131) and (5.132) of Theorem 5.19, or assumptions (5.140) and (5.141) of Corollary 5.20, if and only if $\tilde{p}$ satisfies them. In the sequel we choose to work with $\tilde{p}$ for convenience.

Using the expansions (5.97) and (5.99), we obtain the following form for the boundary condition of (5.143):

$$-w(0) - z(0,\tau) = -\sum_{k=1}^{\infty} c_k (1 - e^{-\lambda_k \tau}) a_k(0). \tag{5.144}$$

The boundary condition (5.144) is thus the sum of a constant and a function that is exponentially decreasing with respect to τ. Note that it vanishes at $\tau = 0$, as a result it is compatible with the initial condition of (5.143).

Remark 5.4. *The constant in (5.144) is generally nonzero. It is equal to zero if and only if $w(0)$ is equal to the limit of $-z(0,\tau)$ when $\tau \to \infty$, that is $\int_0^1 w$, which is generally not true. This is related to the fact, explained in the general d-dimensional case, that the initial layer and the boundary layers are not compatible at infinity. However, there classically exists some $b \in [0,1]$ such that $w(b) = \int_0^1 w$, so that if we were to consider $\tilde{A}(y) = A(y+b)$ instead of A, we would obtain a purely exponentially decreasing function of τ as boundary condition. This trick works only in dimension 1.*

We would like to determine if the function $\tilde{p}$ solution to (5.143) satisfies the assumptions of Theorem 5.19 and Corollary 5.20. Rather than directly tackling (5.143), which still

remains quite involved, we consider the following toy model with constant coefficients:

$$\begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = 0 \ \text{ in } \mathbb{R}_+^* \times \mathbb{R}_+^*, \\ u(0,t) = c(t) \ \text{ in } \mathbb{R}_+^*, \\ u(\cdot,0) = 0 \ \text{ in } \mathbb{R}_+^*, \end{cases} \tag{5.145}$$

for some function $c(t)$ writing as the sum of a constant and an exponentially decreasing function of t.

Our motivation in studying (5.145) is to understand the behavior of solutions to problems of the same kind as (5.143) (namely heat equation with specific boundary and initial conditions). The aim of the computations below is to determine if u verifies the assumptions of Theorem 5.19, that is

$$\frac{\partial u}{\partial t} \in L^1(\mathbb{R}_+; L^2(\mathbb{R}_+)) \tag{5.146}$$

and

$$\frac{\partial u}{\partial x} \in L^2(\mathbb{R}_+; L^2(\mathbb{R}_+)), \tag{5.147}$$

or if, a minima, u verifies the assumptions of Corollary 5.20.

Problem (5.145) can be solved explicitly by means of a Fourier analysis. For this purpose, we first extend it to $\mathbb{R}$. Therefore we define v on $\mathbb{R} \times \mathbb{R}_+^*$ by

$$v(x,t) = u(x,t)\mathbb{1}_{x>0} - u(-x,t)\mathbb{1}_{x<0}. \tag{5.148}$$

Then

$$\frac{\partial v}{\partial x}(x,t) = \frac{\partial u}{\partial x}(x,t)\mathbb{1}_{x>0} + \frac{\partial u}{\partial x}(-x,t)\mathbb{1}_{x<0} + 2c(t)\delta_0, \tag{5.149}$$

$$\frac{\partial v}{\partial t}(x,t) = \frac{\partial u}{\partial t}(x,t)\mathbb{1}_{x>0} + \frac{\partial u}{\partial t}(-x,t)\mathbb{1}_{x<0}, \tag{5.150}$$

and

$$\frac{\partial^2 v}{\partial x^2}(x,t) = \frac{\partial^2 u}{\partial x^2}(x,t)\mathbb{1}_{x>0} + \frac{\partial^2 u}{\partial x^2}(-x,t)\mathbb{1}_{x<0} + 2c(t)\delta_0',$$

so that v is solution to the equation

$$\begin{cases} \dfrac{\partial v}{\partial t} - \dfrac{\partial^2 v}{\partial x^2} = -2c(t)\delta_0' \ \text{ in } \mathbb{R} \times \mathbb{R}_+^*, \\ v(x,0) = 0 \ \text{ in } \mathbb{R}_+^*. \end{cases}$$

The Fourier transform $\hat{v}$ of v, defined by

$$\hat{v}(\xi,t) = \int_{\mathbb{R}} v(x,t) e^{-i\xi x} dx,$$

then solves

$$\begin{cases} \dfrac{\partial \hat{v}}{\partial t} + \xi^2 \hat{v} = -2ic(t)\xi \ \text{ in } \mathbb{R} \times \mathbb{R}_+^*, \\ \hat{v}(\xi, 0) = 0 \ \text{ in } \mathbb{R}_+^*, \end{cases} \tag{5.151}$$

and we deduce from (5.151) that

$$\hat{v}(\xi, t) = -2i\xi e^{-\xi^2 t} \int_0^t c(s) e^{\xi^2 s} ds. \tag{5.152}$$

It is clear from (5.149) and (5.150) that u satisfies (5.146) and (5.147) if and only if v satisfies

$$\frac{\partial v}{\partial t} \in L^1(\mathbb{R}_+; L^2(\mathbb{R})) \tag{5.153}$$

and

$$\frac{\partial v}{\partial x} - 2c(t)\delta_0 \in L^2(\mathbb{R}_+; L^2(\mathbb{R})), \tag{5.154}$$

hence, taking the Fourier transform of (5.153) and (5.154), if and only if

$$\frac{\partial \hat{v}}{\partial t} \in L^1(\mathbb{R}_+; L^2(\mathbb{R}))$$

and

$$\widehat{\frac{\partial v}{\partial x}} - 2c(t) \in L^2(\mathbb{R}_+; L^2(\mathbb{R})).$$

We then compute from (5.152) that

$$\frac{\partial \hat{v}}{\partial t}(\xi, t) = 2i\xi^3 e^{-\xi^2 t} \int_0^t c(s) e^{\xi^2 s} ds - \frac{2i\xi}{\sqrt{2\pi}} c(t) \tag{5.155}$$

and

$$\left(\widehat{\frac{\partial v}{\partial x}} - 2c \right)(\xi, t) = -2e^{-\xi^2 t} \int_0^t c'(s) e^{\xi^2 s} ds - c(0) e^{-\xi^2 t}. \tag{5.156}$$

We thereafter consider two different expressions for the function $c(t)$, aiming at emulating (5.144).

a) Case $\mathbf{c(t) = 1 - e^{-t}}$

As the sum of a constant and an exponentially decreasing function of t, $c(t)$ is designed to reproduce the behavior of (5.144). It also satisfies the compatibility condition with the initial condition $c(0) = 0$. Inserting this specific expression of $c(t)$ in (5.155) and (5.156) yields

$$\frac{\partial \hat{v}}{\partial t}(\xi, t) = 2i\xi \frac{e^{-\xi^2 t} - e^{-t}}{\xi^2 - 1}, \tag{5.157}$$

and

$$\left(\widehat{\frac{\partial v}{\partial x}} - 2c\right)(\xi, t) = 2\frac{e^{-\xi^2 t} - e^{-t}}{\xi^2 - 1}. \tag{5.158}$$

Computing the $L^2(\mathbb{R})$-norm of (5.157) and (5.158), it is straightforward to see that

$$\left\| \frac{\partial \hat{v}}{\partial t}(\cdot, t) \right\|_{L^2(\mathbb{R})} \underset{t \to +\infty}{\sim} \frac{C}{t^{3/4}} \tag{5.159}$$

and

$$\left\| \widehat{\frac{\partial v}{\partial x}}(\cdot, t) - 2c(t) \right\|_{L^2(\mathbb{R})} \underset{t \to +\infty}{\sim} \frac{C}{t^{1/4}}. \tag{5.160}$$

Clearly then,

$$\frac{\partial \hat{v}}{\partial t} \notin L^1(\mathbb{R}_+; L^2(\mathbb{R})),$$

and

$$\widehat{\frac{\partial v}{\partial x}} - 2c(t) \notin L^2(\mathbb{R}_+; L^2(\mathbb{R})).$$

As a consequence

$$\frac{\partial u}{\partial t} \notin L^1(\mathbb{R}_+; L^2(\mathbb{R}_+)),$$

and

$$\frac{\partial u}{\partial x} \notin L^2(\mathbb{R}_+; L^2(\mathbb{R}_+)).$$

This shows that the assumptions of Theorem 5.19 are not satisfied by u solution to (5.145) with $c(t) = 1 - e^{-t}$.

However, it is easy to deduce from estimates (5.159) and (5.160) that the bounds (5.140) and (5.141) hold for u. Thus u verifies the assumptions of Corollary 5.20.

b) Case $\mathbf{c(t) = e^{-t} - e^{-2t}}$.

Following Remark 5.4, $c(t)$ here mimics the boundary condition (5.144) that we obtain when the corrector w is such that $w(0) = \int_0^1 w$, i.e when the initial layer and the boundary layers are compatible at infinity: it is the sum of exponentially decreasing functions of t. It also satisfies the initial compatibility condition $c(0) = 0$.

With this specific function $c(t)$, we find that

$$\frac{\partial \hat{v}}{\partial t}(\xi, t) = 2i\xi \left(\frac{e^{-\xi^2 t} - e^{-t}}{\xi^2 - 1} - 2\frac{e^{-\xi^2 t} - e^{-2t}}{\xi^2 - 2} \right), \tag{5.161}$$

and

$$\left(\widehat{\frac{\partial v}{\partial x}} - 2c\right)(\xi,t) = 2\left(\frac{e^{-\xi^2 t} - e^{-t}}{\xi^2 - 1} - 2\frac{e^{-\xi^2 t} - e^{-2t}}{\xi^2 - 2}\right). \tag{5.162}$$

Computing the $L^2(\mathbb{R})$-norm of (5.161) and (5.162), we get the following behavior for large times:

$$\left\|\frac{\partial \hat{v}}{\partial t}(\cdot,t)\right\|_{L^2(\mathbb{R})} \underset{t\to+\infty}{\sim} \frac{C}{t^{7/4}}$$

and

$$\left\|\widehat{\frac{\partial v}{\partial x}}(\cdot,t) - 2c(t)\right\|_{L^2(\mathbb{R})} \underset{t\to+\infty}{\sim} \frac{C}{t^{5/4}}.$$

Therefore,

$$\frac{\partial \hat{v}}{\partial t} \in L^1(\mathbb{R}_+; L^2(\mathbb{R})), \tag{5.163}$$

and

$$\widehat{\frac{\partial v}{\partial x}} - 2c(t) \in L^2(\mathbb{R}_+; L^2(\mathbb{R})). \tag{5.164}$$

It follows from (5.163) and (5.164) that

$$\frac{\partial u}{\partial t} \in L^1(\mathbb{R}_+; L^2(\mathbb{R}_+))$$

and

$$\frac{\partial u}{\partial x} \in L^2(\mathbb{R}_+; L^2(\mathbb{R}_+)).$$

The function u solution to (5.145) with $c(t) = e^{-t} - e^{-2t}$ thus satisfies the assumptions of Theorem 5.19 and a fortiori those of Corollary 5.20.

These two examples show that assumptions (5.131) and (5.132) of Theorem 5.19 are generally not satisfied in the case of a problem (5.145) resembling the original problem (5.123), though they may hold if by chance the constant in the boundary condition is zero and then if only exponentially decreasing functions remain.

Note however that these one-dimensional computations, albeit instructive, do not allow to draw definitive conclusions on the validity of assumptions (5.131) and (5.132), for the toy problem (5.145) is not equivalent to the original problem (5.123). In particular, we have injected and used in (5.145) less information than we originally had in (5.123): in the latter setting, the boundary and initial conditions depend on the matrix A via (5.32) and (5.92), whereas it is not the case in (5.145). More precisely, if A was the identity matrix in (5.123), then the functions z and $\psi_D^{i,j}$ would necessarily be constants, hence p_i^j would also be a constant and assumptions (5.131) and (5.132) would be trivially verified. This

is obviously not taken into account in (5.145), hence we do not have a counter-example.

On the other hand, the computations above give us hope that assumptions (5.140) and (5.141) of Corollary 5.20 may generally hold.

So as to go further, we would have to define a more suitable toy model with an operator $-\mathrm{div}(A\nabla\cdot)$ instead of the Laplace operator, and with boundary and initial conditions related to A. Then the Fourier analysis performed above would have to be replaced with a Bloch waves analysis. Our attempts to do so have been unfruitful so far.

In conclusion, the validity of the assumptions of Theorem 5.19 and Corollary 5.20 remains an open problem to us.

5.5 Appendix: two parabolic regularity results

In this section we recall two standard parabolic regularity results. We shall not give the proofs and refer the reader to [57] for details.

In the sequel A denotes a *symmetric* tensor field from $\mathbb{R}^d$ to $\mathbb{R}^{d\times d}$ such that there exist $\lambda > 0$ and $\Lambda > 0$ such that

$$\forall \xi \in \mathbb{R}^d,\ \text{a.e in } x \in \mathbb{R}^d,\ \lambda|\xi|^2 \leq A(x)\xi\cdot\xi \text{ and } |A(x)\xi| \leq \Lambda|\xi|. \tag{5.165}$$

We consider the generic parabolic problem

$$\begin{cases} \dfrac{\partial u}{\partial t} - \mathrm{div} A\nabla u = f + \mathrm{div} g \ \text{ in } \Omega\times(0,T), \\ u = 0 \ \text{ on } \partial\Omega\times(0,T), \\ u(\cdot,0) = h \ \text{ in } \Omega, \end{cases} \tag{5.166}$$

where Ω is a regular open set in $\mathbb{R}^d$, not necessarily bounded, and the regularity assumptions on f, g and h will be detailed in the statements of the results below.

Lemma 5.21. *Assume that $h \in L^2(\Omega)$, $g \in L^2(]0,T[;L^2(\Omega))$ and $f \in L^1(]0,T[;L^2(\Omega))$.*

There exists a unique solution u to (5.166) in $C([0,T];L^2(\Omega)) \cap L^2(]0,T[;H^1(\Omega))$. Moreover, there exists a constant C such that

$$\|u\|_{L^\infty(]0,T[;L^2(\Omega))} + \|\nabla u\|_{L^2(]0,T[;L^2(\Omega))} \leq C\left(\|f\|_{L^1(]0,T[;L^2(\Omega))} + \|g\|_{L^2(]0,T[;L^2(\Omega))} + \|h\|_{L^2(\Omega)}\right).$$

The constant C only depends on λ and Λ defined in (5.165).

Lemma 5.22. *Assume that* $h \in L^2_{loc}(\Omega)$, $\nabla h \in L^2(\Omega)$, $h_{|\partial\Omega} = 0$, $g \in C([0,T];L^2(\Omega))$, $\frac{\partial g}{\partial t} \in L^1(]0,T[;L^2(\Omega))$ *and* $f \in L^2(]0,T[;L^2(\Omega))$.

There exists a unique solution u *to (5.166) such that* $\nabla u \in C([0,T];L^2(\Omega))$ *and* $\frac{\partial u}{\partial t} \in L^2(]0,T[;L^2(\Omega))$. *Moreover there exists a constant* C *such that*

$$\left\|\frac{\partial u}{\partial t}\right\|_{L^2(]0,T[;L^2(\Omega))} + \|\nabla u\|_{C([0,T];L^2(\Omega))} \leq C\left(\|f\|_{L^2(]0,T[;L^2(\Omega))} + \left\|\frac{\partial g}{\partial t}\right\|_{L^1(]0,T[;L^2(\Omega))} + \|g\|_{C([0,T];L^2(\Omega))} + \|\nabla h\|_{L^2(\Omega)}\right).$$

The constant C *only depends on* λ *and* Λ *defined in (5.165).*

Remark 5.5. *(On the assumptions of Lemma 5.22). Note that if* Ω *has a boundary, then the assumption* $\nabla h \in L^2(\Omega)$ *implies that* $h \in L^2_{loc}(\Omega)$ *because of Poincaré inequality. If* Ω *is the whole space* $\mathbb{R}^d$, *then an adequate definition of the working space (Beppo-levi or Deny-Lions) also guarantees that* $h \in L^2_{loc}(\Omega)$ *if* $\nabla h \in L^2(\Omega)$.

Bibliographie

[1] G. Allaire, *Homogenization and two-scale convergence*, SIAM J. Math. Anal. 23 (6) (1992), pp. 1482-1518.

[2] G. Allaire and M. Amar, *Boundary layer tails in periodic homogenization*, ESAIM: Control Optim. and Calc. Var. 4 (1999), pp. 209-243.

[3] G. Allaire and S. Gutierrez, *Optimal design in small amplitude homogenization*, ESAIM: M2AN, Vol. 41 no.3 (2007), pp. 543-574.

[4] L. Ambrosio, G. Friesecke and J. Giannoulis, *Passage from quantum to classical molecular dynamics in the presence of Coulomb interactions*, to appear in Comm. PDE, http://arxiv.org/abs/0907.1205.

[5] A. Anantharaman and E. Cancès, *Existence of minimizers for Kohn-Sham models in quantum chemistry*, Ann. IHP (C) Nonlinear Analysis, Vol. 26 no. 6 (2009), pp. 2425-2455.

[6] A. Anantharaman and C. Le Bris, *A numerical approach related to defect-type theories for some weakly random problems in homogenization*, preprint available at http://arxiv.org/abs/1005.3910.

[7] A. Anantharaman and C. Le Bris, *Elements of mathematical foundations for a numerical approach for weakly random homogenization problems*, preprint available at http://arxiv.org/abs/1005.3922.

[8] A. Anantharaman and C. Le Bris, *Homogenization of a weakly randomly perturbed periodic material*, C. R. Acad. Sci. Paris Série I, Vol. 348 (9-10) (2010), pp. 529-534.

[9] M. Avellaneda and F.-H. Lin, *Compactness methods in the theory of homogenization*, Communications on Pure and Applied Mathematics, Vol. XL (1987), pp. 803-847.

[10] I. Babuska, *Solution of interface problems by homogenization I, II, III*, SIAM J. Math Anal. 7 (1976) pp. 603-634 and 635-645; 8 (1977), pp. 923-937.

[11] N. Bakhvalov and G. Panasenko, *Homogenization: averaging processes in periodic media,* Mathematics and its applications, vol.36, Kluwer Academic Publishers, Dordrecht (1990).

[12] A.D. Becke, *Density-functional exchange-energy approximation with correct asymptotic behavior*, Phys. Rev. A 38 (1988), pp. 3098-3100.

[13] M. F. Ben Hassen and E. Bonnetier, *An asymptotic formula for the voltage potential in a perturbed ε-periodic composite medium containing misplaced inclusions of size ε*, Proceedings of the Royal Society of Edinburgh, 136A (2006), pp. 669-700.

[14] A. Bensoussan, J.-L. Lions and G. Papanicolaou, *Asymptotic Methods in Periodic Media*, North Holland (1978).

[15] A. Bensoussan, J.-L. Lions and G. Papanicolaou, *Boundary layer analysis in homogenization of diffusion equations with Dirichlet conditions on the half space*, Proc. Internat. Symposium SDE, K. Ito Ed. J. Wiley, New York (1978), pp. 21-40.

[16] A. Bensoussan, J.-L. Lions and G. Papanicolaou, *Boundary layers and homogenization of transport processes*, Publ. Res. Inst. Math. Sci. 15 (1979), pp 53-157.

[17] E. Beretta, A. Mukherjee and M. Vogelius, *Asymptotic formulas for steady state voltage potentials in the presence of conductivity imperfections of small area*, Z. angew. Math. Phys. 52 (2001), pp. 543-572.

[18] X. Blanc, C. Le Bris and P.-L. Lions, *Stochastic homogenization and random lattices*, J. Math. Pures Appl., 88 (2007), pp. 34-63.

[19] P.E. Blöchl, *Projector augmented-wave method*, Phys. Rev. B 50 (1994), pp. 17953-17979.

[20] M. Born and R. Oppenheimer, *Zur Quantentheorie der Molekeln*, Ann. Phys. (Leipzig) 84 (1927), pp. 457-484.

[21] A. Bourgeat and E. Marusic-Paloka, *Non-linear effects for flow in periodically constricted channel caused by high injection rate*, Mathematical Models and Methods in Applied Sciences 8 (1998), pp. 379-405.

[22] A. Bourgeat and A. Piatnitski, *Approximations of effective coefficients in stochastic homogenization*, Annales de l'Institut Henri Poincaré (B) Probabilités et Statistiques, 40 no. 2 (2004), pp. 153-165.

[23] S. Brahim-Otsmane, G. Francfort and F. Murat, *Correctors for the homogenization of the wave and heat equations*, J. Maths. Pures Appl., 9 (1992), pp. 197-231.

[24] R. Brizzi and J.-P. Chalot, *Homogénéisation de frontière*, PhD Thesis, Université de Nice (1978).

[25] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris and Y. Maday, *Computational quantum chemistry: a primer*, Handbook of numerical analysis, Volume X: special volume: computational chemistry, Ph. Ciarlet and C. Le Bris eds, North Holland (2003).

[26] Y. Capdeboscq and M. S. Vogelius, *A general representation formula for boundary voltage perturbations caused by internal conductivity inhomogeneities of low volume fraction*, ESAIM: M2AN, Vol. 37, No 1 (2003), pp.159-173.

[27] I. Catto, C. Le Bris and P.-L. Lions, *On the thermodynamic limit for Hartree-Fock type models*, Ann. I. H. Poincaré - AN 18, 6 (2001), pp. 687-760.

[28] D. J. Cedio-Fengya, S. Moskow and M. S. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Continuous dependence and computational reconstruction*, Inverse Problems 14 (1998), pp. 553-595.

[29] D.M. Ceperley and B.J. Alder, *Ground state of the electron gas by a stochastic method*, Phys. Rev. Lett. 45 (1980), pp. 566-569.

[30] G. Chechkin, A. Friedman and A. Piatnitski, *The boundary-value problem in domains with very rapidly oscillating boundary*, INRIA Report 3062 (1996).

[31] R. Costaouec, C. Le Bris and F. Legoll, *Numerical approximation of a class of problems in stochastic homogenization*, C. R. Acad. Sci. Paris Série I, Vol. 348 (1-2) (2010), pp. 99-103.

[32] E.R. Davidson, *Reduced density matrices in quantum chemistry*, Academic Press, New York (1976).

[33] R.M. Dreizler and E.K.U. Gross, *Density functional theory*, Springer (1990).

[34] I. Ekeland, *Nonconvex minimization problems*, Bull. Am. Math. Soc. 1 (1979), pp. 443-474.

[35] R. L. Frank, E. H. Lieb, R. Seiringer and H. Siedentop, *Müller's exchange-correlation energy in density-matrix-functional theory*, Phys. Rev. A 76 (2007) 052517.

[36] A. Friedman, B. Hu and Y. Liu, *A boundary value problem for the Poisson equation with multi-scale oscillating boundary*, J. Diff. Eq. 137 (1997), pp. 54-93.

[37] D. Gérard-Varet and N. Masmoudi, *Homogenization in polygonal domains*, to appear in Journal of the European Mathematical society.

[38] M. Giaquinta, *Multiple integrals in the calculus of variations and nonlinear elliptic systems*, Princeton. Univ. Press, Princeton (1983).

[39] D. Gilbarg and N.S. Trudinger, *Elliptic partial differential equations of second order*, 3rd edition, Springer (1998).

[40] M. Grüter and K-O. Widman, *The Green function for uniformly elliptic equations*, manuscripta math., 37 (1982), pp. 303-342.

[41] G.A. Hagedorn and A. Joye, *A time-dependent Born-Oppenheimer approximation with exponentially small error estimates*, Comm. Math. Phys. 223 (3) (2001), pp. 583-626.

[42] W.J. Hehre, L. Radom, P.v.R. Schleyer and J.A. Pople, *Ab initio molecular orbital theory*, Wiley (1986).

[43] P. Hohenberg and W. Kohn, *Inhomogeneous electron gas*, Phys. Rev. 136 (1964), pp. 864-871.

[44] L. Hörmander, *The analysis of linear partial differential operators I: Distribution Theory and Fourier Analysis*, Grundl. Math. Wissenschaft. 256, Springer (1983).

[45] W. Jäger and A. Mikelic, *On the boundary conditions at the contact interface between a porous medium and a free fluid*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. 23 (1996), pp. 403-465.

[46] V. V. Jikov, S. M. Kozlov and O. A. Oleinik, *Homogenization of Differential Operators and Integral Functionals*, Springer Verlag (1994).

[47] W. Kohn and L. J. Sham, *Self-consistent equations including exchange and correlation effects*, Phys. Rev. 140 (1965) A1133.

[48] D.C. Langreth and J.P. Perdew, *Theory of nonuniform electronic systems. I. Analysis of the gradient approximation and a generalization that works*, Phys. Rev. B 21 (1980), pp. 5469-5493.

[49] C. Le Bris, *Quelques problèmes mathématiques en chimie quantique moléculaire*, Thèse de l'Ecole Polytechnique (1993).

[50] C. Le Bris, *Some numerical approaches for "weakly" random homogenization*, to appear in Proceedings of the ENUMATH 2009 Conference, Springer (2010).

[51] M. Levy, *Universal variational functionals of electron densities, first order density matrices, and natural spin-orbitals and solution of the V-representability problem*, Proc. Natl. Acad. Sci. USA 76 (1979), pp. 6062-6065.

[52] Y. Li and L. Nirenberg, *Estimates for elliptic systems from composite material*, Communications on Pure and Applied Mathematics, Vol. LVI (2003), pp. 892-925.

[53] E.H. Lieb, *Density Functional for Coulomb systems*, Int. J. Quant. Chem. 24 (1983), pp. 243-277.

[54] E.H. Lieb and B. Simon, *The Hartree-Fock theory for Coulomb systems*, Commun. Math. Phys. 53 (1977), pp. 185-194.

[55] E.H. Lieb and M. Loss. *Analysis*, Second Edition. Graduate Studies in Mathematics, Vol. 14. American Mathematical Society, Providence, Rhode Island (2001).

[56] J.-L. Lions, *Some methods in the mathematical analysis of systems and their control,* Science Press, Beijing, Gordon and Breach, New York (1981).

[57] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, volumes 1, 2 et 3, Dunod (1968).

[58] P.-L. Lions, *Solutions of Hartree-Fock equations for Coulomb systems*, Comm. Math. Phys. 109 (1987), pp. 33-97.

[59] P.-L. Lions, *The concentration-compactness method in the Calculus of Variations. The locally compact case. Part. I:* Anal. non-linéaire, Ann. IHP 1 (1984), p. 109-145. *Part. II:* Anal. non-linéaire, Ann. IHP 1 (1984), pp. 223-283.

[60] P.-L. Lions and T. Paul, *Sur les mesures de Wigner*, Rev. Mat. Iberoamericana 9 (1993), pp. 553-618.

[61] R. McWeeny, *Methods of molecular quantum mechanics*, 2nd edition, Academic Press (1992).

[62] S. Moskow and M. Vogelius, *First order corrections to the homogenized eigenvalues or a periodic composite medium: The case of Neumann boundary conditions*, Indiana Univ. Math. J., to appear.

[63] F. Murat, *Compacité par compensation*, Ann. Scuola Norm. Sup. Pisa. Cl. Sci. 5 (4) (1978), pp. 485-507.

[64] M. Neuss-Radu, *The boundary behavior of a composite material*, ESAIM: Mathematical Modelling and Numerical Analysis 35, no. 3, (2001), pp. 407-435.

[65] G. Nguetseng, *A general convergence result for a functional related to the theory of homogenization*, SIAM J. Math. Anal. 20 (3) (1989), pp. 608-623.

[66] O. Oleinik, A. Shamaev and G. Yosifian, *Mathematical problems in elasticity and homogenization*, North Holland, Amsterdam (1992).

[67] R.G. Parr and W. Yang, *Density-functional theory of atoms and molecules*, Oxford University Press (1989).

[68] A.L. Piatnitski, *A parabolic equation with rapidly oscillating coefficients*, Mosc. Univ. Math. Bull. 35, no.3, (1980), pp. 35-42.

[69] J.P. Perdew, K. Burke and M. Ernzerhof, *Generalized gradient approximation made simple*, Phys. Rev. Lett. 77 (1996), pp. 3865-3868.

[70] J.P. Perdew and Y. Wang, *Accurate and simple density functional for the electronic exchange energy: Generalized gradient approximation*, Phys. Rev. B 33 (1986), pp. 8800-8802.

[71] J.P. Perdew and Y. Wang, *Accurate and simple analytic representation of the electron-gas correlation energy*, Phys. Rev. B 45 (1992), pp. 13244-13249.

[72] J.P. Perdew and A. Zunger, *Self-interaction correction to density-functional approximations for many-electron systems*, Phys. Rev. B 23 (1981), pp. 5048-5079.

[73] S. Redner, *Citation statistics from 110 years of Physical Review*, Physics Today 49 (2005), pp. 49-54.

[74] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol I, Functional Analysis, 2nd edition, Academic Press, New York (1980).

[75] M. Reed and B. Simon, *Methods of Modern Mathematical Physics*, Vol IV, Analysis of Operators, Academic Press, New York (1978).

[76] S. Sakata, F. Ashida, T. Kojima and M. Zako, *Three-dimensional stochastic analysis using a perturbation-based homogenization method for elastic properties of composite material considering microscopic uncertainty*, International Journal of Solids and Structures, 45 (2008), pp. 894-907.

[77] B. Simon, *Trace Ideals and their Applications.* Vol 35 of *London Mathematical Society Lecture Notes Series*, Cambridge University Press, (1979).

[78] S. Spagnolo, *Convergence in energy for elliptic operators*, Proc. Third Symp. Numer. Solut. Partial. Diff. Equat., Academic Press, New York (1976), pp. 469-498.

[79] G. Stampacchia, *Le problème de Dirichlet pour les équations elliptiques du second ordre à coefficients discontinus*, Ann. Inst. Fourier 15 (1965), pp. 189-257.

[80] L. Tartar, *H-measures, a new approach for studying homogenisation, oscillations and concentration effects in partial differential equations*, Proceedings of the Royal Society of Edinburgh, 115A (1990), pp. 193-230.

[81] L. Tartar, *Homogénéisation et compacité par compensation*, Cours Peccot au College de France (1977). F. Murat, *H-convergence*, Séminaire d'Analyse Fonctionnelle et Numérique de l'Université d'Alger (1978). F. Murat and L. Tartar, *H-convergence*, In: *Mathematical Modelling of Composites Materials*, A. Cherkaev and R.V. Kohn. (eds.), Progress in Nonlinear Differential Equations and their Applications, Birkhäuser (1997).

[82] S. Teufel, *Adiabatic perturbation theory in quantum dynamics*, Lecture Notes in Mathematics 1821, Springer (2003).

[83] M. Thomas, *Propriétés thermiques de matériaux composites: caractérisation expérimentale et approche microstructurale*, Thèse de l'Université de Nantes, Laboratoire de Thermocinétique, CNRS-UMR 6607 (2008).

[84] N. Troullier and J.L. Martins, *Efficient pseudopotentials for plane wave calculations*, Phys. Rev. B 43 (1991), pp. 1993-2006.

[85] D. Vanderbilt, *Soft self-consistent pseudopotentials in a generalized eigenvalue formalism*, Phys. Rev. B41 (1990), pp. 7892-7895.

[86] S.H. Vosko, L. Wilk and M. Nusair, *Accurate spin-dependent electron liquid correlation energy for local spin density calculations: a critical analysis*, Can. J. Phys. 58 (1980), pp. 1200-1211.

[87] G.M Zhislin, *A study of the spectrum of the Schrödinger operator for a system of several particles*, Trudy Moskov. Mat. Obsc. 9 (1960), pp. 81-120.

[88] G.M. Zhislin and A.G. Sigalov, *The spectrum of the energy operator for atoms with fixed nuclei on subspaces corresponding to irreducible representations of the group of permutations*, Izv. Akad. Nauk SSSR Ser. Mat. 29 (1965), pp. 835-860.