Numerical homogenization: survey, new results, and perspectives

Antoine Gloria

ESAIM: PROCEEDINGS, September 2012, Vol. 37, p. 50-116
E. Cancès and S. Labbé, Editors

NUMERICAL HOMOGENIZATION: SURVEY, NEW RESULTS, AND PERSPECTIVES

Antoine Gloria[1]

Abstract. These notes give a state of the art of numerical homogenization methods for linear elliptic equations. The guideline of these notes is analysis. Most of the numerical homogenization methods can be seen as (more or less different) discretizations of the same family of continuous approximate problems, which H-converges to the homogenized problem. Likewise numerical correctors may also be interpreted as approximations of Tartar's correctors. Hence the convergence analysis of these methods relies on the H-convergence theory. When one is interested in convergence rates, the story is different. In particular one first needs to make additional structure assumptions on the heterogeneities (say periodicity for instance). In that case, a crucial tool is the spectral interpretation of the corrector equation by Papanicolaou and Varadhan. Spectral analysis does not only allow to obtain convergence rates, but also to devise efficient new approximation methods. For both qualitative and quantitative properties, the development and the analysis of numerical homogenization methods rely on seminal concepts of the homogenization theory. These notes contain some new results.

Résumé. Ces notes de cours dressent un état de l'art des méthodes d'homogénéisation numérique pour les équations elliptiques linéaires. Le fil conducteur choisi est l'analyse. La plupart des méthodes d'homogénéisation numérique s'interprète comme des discrétisations (plus ou moins différentes) d'une même famille de problèmes continus approchés qui H-converge vers le problème homogénéisé. De même, le concept de correcteur numérique s'interprète comme une approximation des correcteurs introduits par Tartar. Ainsi l'analyse de convergence repose essentiellement sur la théorie de la H-convergence. Si on s'intéresse aux estimations quantitatives d'erreur, il faut faire des hypothèses supplémentaires de structure sur les hétérogénéités (périodicité par exemple). Dans ce cas, un outil important est l'interprétation spectrale de l'équation du correcteur introduite par Papanicolaou et Varadhan, qui permet non seulement de démontrer des résultats quantitatifs, mais aussi de développer des méthodes numériques efficaces. Qu'il s'agisse de propriétés qualitatives ou quantitatives, le développement et l'analyse de méthodes d'homogénéisation numérique reposent sur des concepts fondateurs de la théorie de l'homogénéisation. Ces notes contiennent quelques résultats nouveaux.

Contents

[1] Project-team SIMPAF & Laboratoire Paul Painlevé UMR 8524
INRIA Lille - Nord Europe & Université Lille 1
Villeneuve d'Ascq, France
antoine.gloria@inria.fr

1. Introduction

1.1. Motivation

Numerical homogenization methods (see [20, 42, 43], [6], [25], [17–19] e.g.) are designed to solve partial differential equations for which the operator is strongly heterogeneous spatially. Such problems arise in many applications such as diffusion in porous media or composite materials. We refer the reader to the bibliography for details on the fields of application. By numerical homogenization, we mean that we compute not only an "averaged" solution of the highly heterogeneous problem, but also the local fluctuations, which may be important in many applications. We focus in this article on the prototypical problem of a scalar linear elliptic equation: for some $1 \gg \varepsilon_0 > 0$, find $u_{\varepsilon_0} \in H^1_0(D)$ such that

$$-\nabla \cdot A_{\varepsilon_0} \nabla u_{\varepsilon_0} = f \qquad \text{in } D, \tag{1.1}$$

on a Lipschitz domain D for some $f \in H^{-1}(D)$. Here, the spatial dependence of the operator is encoded in the function A_{ε_0}, whose frequencies are assumed to be of order ε_0^{-1}. Academic cases are of the form: $A_{\varepsilon_0}(x) = A(x/\varepsilon_0)$, with A periodic, quasi-periodic or stationary in an ergodic stochastic setting. More realistic models can be of the form: $A_{\varepsilon_0}(x) = A(x, x/\varepsilon_0)$, where $A(x, \cdot)$ may be periodic, quasi-periodic or stationary for all $x \in D$, provided some suitable cross-regularity holds (see [4] e.g.). In all these examples, ε_0 refers to the *actual* lengthscale of the heterogeneities. A frontal approach to solve (1.2) numerically would require to discretize (1.2) with a meshsize smaller than ε_0, which is prohibitive in practice. The aim of numerical homogenization is to

benefit from the structure of A_{ε_0} to design more efficient methods, avoiding the use of fine mesh on the whole domain D.

A first abstract step consists in imbedding (1.1) into a whole family of problems parametrized by $\varepsilon > 0$:

$$-\nabla \cdot A_\varepsilon \nabla u_\varepsilon \ = \ f \qquad \text{in } D, \tag{1.2}$$

and which coincides with (1.1) for $\varepsilon = \varepsilon_0$. Since ε_0 is small, a natural strategy is to pass to the limit as $\varepsilon \to 0$ in (1.2), and solve the limiting problem at "$\varepsilon = 0$" instead of (1.1). Of course, from a practical point of view we are given A_{ε_0} and not necessarily $\{A_\varepsilon\}_{\varepsilon>0}$, and although ε_0 is small, it is not zero and one should somehow remember this scale in the approximation. These two concerns are addressed formally in this introduction. We shall first quickly recall the main results of the H-convergence theory, and then show how to deduce a numerical homogenization procedure from this theory using a "self-consistent" approach.

The rest of this survey is organized as follows. In the second section we introduce more rigorously these numerical homogenization methods, and provide a qualitative convergence analysis using H-convergence. In Section 3 we display a quantitative convergence analysis of the periodic case, putting in evidence the so-called resonance error. We then present two standard ways to reduce the resonance error, windowing and oversampling, and extend the qualitative convergence analysis to these cases. In Section 4 we turn to a more efficient way to deal with the resonance error, based on the introduction of a zero-order term in the corrector equation. We describe the approach on the approximation of homogenized coefficients, and prove the convergence of the method using spectral analysis — which requires the matrix A_ε to be symmetric. In Section 5 we combine the numerical homogenization methods with the regularization approach to reduce the resonance error, and introduce another numerical corrector which approximates better the local fluctuations of the solution. We then quickly mention in Section 6 two other numerical homogenization methods based on unfolding and on harmonic coordinates, as well as a small collection of examples illustrating how the strategies developed so far for the linear case can be adapted to the nonlinear case, and why new ideas are definitely needed.

A a general rule, most of the qualitative convergence results are proved in detail. For quantitative results, the proofs are often omitted and precise references are given.

1.2. **Self-consistent approach**

In this paragraph we present a formal approach to compute an approximation of the solution of (1.2) which takes advantage of the scale separation of A_ε. The justification of this approach will be given in Section 2.

The starting point of the self-consistent approach is the assumption that the solution u_ε to (1.2) displays the same scale separation as A_ε in the sense that it can be decomposed as

$$u_\varepsilon \ = \ \overline{u}_\varepsilon + \check{u}_\varepsilon,$$

where

- $\overline{u}_\varepsilon$ is a low frequency part (say with frequencies of order 1),
- $\check{u}_\varepsilon$ is a high frequency part (frequencies of order ε^{-1}) possibly modulated by a factor with frequencies of order 1.

In particular, we assume that for every open bounded subdomain T of D

$$\fint_T \check{u}_\varepsilon \lesssim \varepsilon.$$

Note that the fact that the solution u_ε to (1.2) displays the same scale separation as A_ε may not be true in general. This decomposition can be made explicit in the periodic case using the two-scale expansion [8]

$$u_\varepsilon(x) \ = \ u_0(x) + \varepsilon u_1(x, \frac{x}{\varepsilon}) + o(\varepsilon),$$

which yields $\overline{u}_\varepsilon = u_0(x)$ and $\check{u}_\varepsilon = \varepsilon u_1(x, \frac{x}{\varepsilon}) + o(\varepsilon)$.

The self-consistent approach consists in deriving equations for $\overline{u}_\varepsilon$ and $\check{u}_\varepsilon$ as consequences of the fact that $\overline{u}_\varepsilon + \check{u}_\varepsilon$ is a solution to (1.2). For the reasoning, we let $\overline{u}_{\varepsilon,H}$ be an approximation of $\overline{u}_\varepsilon$ in some $P1$-FE space $V_H \subset H^1_0(D)$ associated with a triangulation $\{T_k\}_k$ of D of meshsize $H \gg \varepsilon$, and f_H be an approximation of f in the associated $P0$-FE space. We rephrase the question as: What is the link between $\overline{u}_{\varepsilon,H}$ and A_ε ? By assumption, $\overline{u}_{\varepsilon,H} + \check{u}_\varepsilon$ is a good approximation of u_ε in $H^1_0(D)$. In addition, the assumption on $\check{u}_\varepsilon$ reads on each element T_k

$$\int_{T_k} \check{u}_\varepsilon \;=\; |T_k| O(\varepsilon),$$

which we will assume to be zero for *simplicity* of the argument. Inserting "$u_\varepsilon = \overline{u}_{\varepsilon,H} + \check{u}_\varepsilon$" in (1.2), we obtain that $\check{u}_\varepsilon$ satisfies for all $v_\varepsilon \in H^1_0(T_k) \cap L^2_0(T_k)$ (that is the functions of $H^1_0(T_k)$ with zero average on T_k)

$$\begin{aligned}\int_{T_k} \nabla v_\varepsilon \cdot A_\varepsilon \nabla \check{u}_\varepsilon &= -\int_{T_k} \nabla v_\varepsilon \cdot A_\varepsilon \nabla \overline{u}_{\varepsilon,H} + \int_{\partial T_k} v_\varepsilon n \cdot A_\varepsilon \nabla(\overline{u}_{\varepsilon,H} + \check{u}_\varepsilon) + \int_{T_k} f_H v_\varepsilon \\ &= -\int_{T_k} \nabla v_\varepsilon \cdot A_\varepsilon \nabla \overline{u}_{\varepsilon,H}\end{aligned}$$

since f_H is constant on T_k and $\int_{T_k} v_\varepsilon = 0$. Hence, we have

- one equation for $\check{u}_\varepsilon$ on each element T_k,
- compatibility conditions for $\check{u}_\varepsilon$ on ∂T_k (continuity at the interfaces).

This couples the scales ε and H at the interfaces

We then make a *local closure assumption*, and impose $\check{u}_\varepsilon = 0$ on ∂T_k — which implies that $\check{u}_\varepsilon \in H^1_0(T_k) \cap L^2_0(T_k)$. This decouples ε and H. Indeed, on every element T_k, by the Lax-Milgram theorem, $\check{u}_\varepsilon \in H^1_0(T_k) \cap L^2_0(T_k)$ is the unique solution to: For all $v_\varepsilon \in H^1_0(T_k) \cap L^2_0(T_k)$

$$\int_{T_k} \nabla v_\varepsilon \cdot A_\varepsilon(\nabla \overline{u}_{\varepsilon,H} + \nabla \check{u}_\varepsilon) \;=\; 0.$$

Define $\psi^k_{\varepsilon,i} \in H^1_0(T_k) \cap L^2_0(T_k)$ for every k and every $\{\mathbf{e}_i\}_{1\le i\le d}$ (the canonical basis of $\mathbb{R}^d$) as the unique weak solution to: for all $\chi \in H^1_0(T_k) \cap L^2_0(T_k)$,

$$\int_{T_k} \nabla \chi \cdot A_\varepsilon(\mathbf{e}_i + \nabla \psi^k_{\varepsilon,i}) \;=\; 0.$$

Then by linearity and using that $\nabla \overline{u}_{\varepsilon,H}$ is piecewise-constant:

$$\check{u}_\varepsilon = \sum_k \sum_{i=1}^d (\partial_i \overline{u}_{\varepsilon,H})_{|T_k} \psi^k_{\varepsilon,i} \in H^1_0(D).$$

Hence, setting $\Psi_\varepsilon = \sum_k 1_{T_k}(\psi^k_{\varepsilon,1}, \cdots, \psi^k_{\varepsilon,d}) \in H^1_0(D, \mathbb{R}^d)$, we have for all $\phi_H \in V_H$ and all $\check{\phi} \in H^1_0(D)$ such that $\check{\phi}_{|T_k} \in H^1_0(T_k) \cap L^2_0(T_k)$ for all k,

$$\int_D \nabla(\phi_H + \check{\phi}) \cdot A_\varepsilon(\mathrm{Id} + \nabla \Psi_\varepsilon) \nabla \overline{u}_{\varepsilon,H} \;=\; \int_D (\phi_H + \check{\phi}) f_H \;=\; \int_D \phi_H f_H.$$

This allows us to obtain a closed equation for $\overline{u}_{\varepsilon,H}$ by taking $\check{\phi} \equiv 0$: For all $\phi_H \in V_H$,

$$\sum_k (\nabla \phi_H)_{|T_k} \left[\int_{T_k} A_\varepsilon(\mathrm{Id} + \nabla \Psi_\varepsilon)\right] (\nabla \overline{u}_{\varepsilon,H})_{|T_k} \;=\; \int_D \phi_H f,$$

that is

$$\int_D \nabla\phi_H \cdot A^*_{\varepsilon,H} \nabla \overline{u}_{\varepsilon,H} \;=\; \int_D \phi_H f$$

with $A^*_{\varepsilon,H} \;=\; \sum_k 1_{|T_k} \int_{T_k} A_\varepsilon(\mathrm{Id} + \nabla\Psi_\varepsilon)$. So defined, $A^*_{\varepsilon,H}$ is expected to have frequencies of order 1 (or say H^{-1} at worst), but not ε^{-1}. We then finally obtain

$$u_\varepsilon \simeq \overline{u}_{\varepsilon,H} + \check{u}_\varepsilon = \overline{u}_{\varepsilon,H} + \sum_k \sum_{i=1}^d (\partial_i \overline{u}_{\varepsilon,H})_{|T_k} \psi^k_{\varepsilon,i}.$$

There are at least two ways to exploit this chain of arguments in practice at ε fixed. We start with the "direct approach", which appears for instance in [17,18,25]. The method is as follows: first approximate Ψ^k_ε by $\Psi^k_{\varepsilon,h}$ for all k using a FE method and a fine mesh of T_k (with meshsize $h \ll \varepsilon$), then construct the associated approximation of $A^*_{\varepsilon,H}$

$$A^*_{\varepsilon,H,h} \;:=\; \sum_k 1_{|T_k} \int_{T_k} A_\varepsilon(\mathrm{Id} + \nabla\Psi^k_{\varepsilon,h}),$$

define $\overline{u}_{\varepsilon,H,h}$ as the unique solution in V_H to

$$\text{For all } \phi_H \in V_H, \qquad \int_D \nabla\phi_H \cdot A^*_{\varepsilon,H,h} \nabla \overline{u}_{\varepsilon,H,h} \;=\; \int_D \phi_H f, \tag{1.3}$$

and finally reconstruct an approximation of u_ε via

$$u_\varepsilon \simeq \overline{u}_{\varepsilon,H,h} + \sum_k \sum_{i=1}^d (\partial_i \overline{u}_{\varepsilon,H,h})_{|T_k} \psi^k_{\varepsilon,i,h}. \tag{1.4}$$

This method amounts to approximating the low frequency part $\overline{u}_\varepsilon$ of u_ε, and to reconstructing the high frequency part $\check{u}_\varepsilon$ afterwards. Hence the equation is changed (A_ε is replaced by $A^*_{\varepsilon,H,h}$), but the finite element space is the coarse space V_H (with $H \gg \varepsilon$).

We now turn another point of view, which we call the "dual approach". This method was introduced in [42,43]. The starting point is the observation that the formula (1.4) amounts to looking for an approximation of u_ε in the space

$$V_{\varepsilon,H,h} \;=\; \left\{ \phi_H + \sum_k \sum_{i=1}^d (\partial_i \phi_H)_{|T_k} \psi^k_{\varepsilon,i,h} \;\middle|\; \phi_H \in V_H \right\},$$

whose elements displays frequencies of order ε^{-1} (via $\Psi_{\varepsilon,h}$) but whose dimension is precisely that of the coarse space V_H. We make use of the following notation: for all $\phi_{\varepsilon,H,h} \in V_{\varepsilon,H,h}$, we denote by ϕ_H the unique element of V_H such that $\phi_{\varepsilon,H,h} = \phi_H + \sum_k \sum_{i=1}^d (\partial_i \phi_H)_{|T_k} \psi^k_{\varepsilon,i,h}$ (this identification will be used for $\tilde{u}_{\varepsilon,H,h} \in V_{\varepsilon,H,h}$ and $\tilde{u}_H \in V_H$ as well). The streamline of the dual approach is to keep the equation unchanged but replace the finite element space V_H by $V_{\varepsilon,H,h}$. The approximation $\tilde{u}_{\varepsilon,H,h} = \tilde{u}_H + \sum_k \sum_{i=1}^d (\partial_i \tilde{u}_H)_{|T_k} \psi^k_{\varepsilon,i,h}$ of u_ε is then given by the unique solution in $V_{\varepsilon,H,h}$ to

$$\text{For all } \phi_{\varepsilon,H,h} \in V_{\varepsilon,H,h}, \qquad \int_D \nabla\phi_{\varepsilon,H,h} \cdot A_\varepsilon \nabla \tilde{u}_{\varepsilon,H,h} \;=\; \int_D \phi_{\varepsilon,H,h} f_H.$$

We claim that we have $\tilde{u}_{\varepsilon,H,h} \equiv \overline{u}_{\varepsilon,H,h} + \sum_k \sum_{i=1}^d (\partial_i \overline{u}_{\varepsilon,H,h})_{|T_k} \psi^k_{\varepsilon,i,h}$, where $\overline{u}_{\varepsilon,H,h} \in V_H$ is the solution to (1.3) obtained by the direct approach. Indeed, since $\Psi_{\varepsilon,h}$ has zero average on T_k and f_H is constant on T_k for all k,

$\int_D \phi_{\varepsilon,H,h} f_H = \int_D \phi_H f_H$. Likewise, since $\nabla\phi_H$ is constant on each T_k,

$$\int_D \nabla\phi_{\varepsilon,H,h} \cdot A_\varepsilon \nabla\tilde{u}_{\varepsilon,H,h} \;=\; \int_D \nabla\phi_H \cdot A_\varepsilon \nabla\tilde{u}_{\varepsilon,H,h} \;=\; \int_D \nabla\phi_H \cdot A^*_{\varepsilon,H,h} \nabla\tilde{u}_H.$$

Hence the this equation coincides with (1.3) so that $\tilde{u}_H = \overline{u}_{\varepsilon,H,h}$, which proves the claim.

This elementary construction using a formal self-consistent approach has allowed us to introduce two classes of numerical homogenization methods (the direct and dual approaches), which coincide in this particular case. In order to analyze the convergence of these methods, we'll have to show that $A^*_{\varepsilon,H}$ tends to some meaningful quantity as ε vanishes. This is where the H-convergence theory comes into the picture. Before we turn to the core of the survey and present a rigorous theory to analyze the methods obtained by the self-consistent approach, we complete this introduction with a short review of important results of the H-convergence theory.

1.3. **H-convergence**

Let D be a bounded open Lipchitz subset of $\mathbb{R}^d$, let $\beta \geq \alpha > 0$, let $\mathcal{M}_d$ be the set of real $d \times d$ matrices, and $\{\mathbf{e}_i\}_{i\in\{1,\dots,d\}}$ denote the canonical basis of $\mathbb{R}^d$. We denote by $\mathcal{M}_{\alpha\beta}(D)$ the set of measurable functions A from D to $\mathcal{M}_d$, such that for all $\xi \in \mathbb{R}^d$ and for almost every $x \in D$,

$$|A(x)\xi| \;\leq\; \beta|\xi|, \qquad \alpha|\xi|^2 \;\leq\; \xi \cdot A(x)\xi.$$

The notion of H-convergence, introduced by Tartar [57] and developed by Murat and Tartar [50,51], is defined as:

Definition 1. *A sequence A_ε in $\mathcal{M}_{\alpha\beta}(D)$ H-converges to some $A_0 \in \mathcal{M}_{\alpha'\beta'}(D)$ for some $\beta' \geq \alpha' > 0$ if for every function $f \in H^{-1}(D)$, the weak solution $u_\varepsilon \in H^1_0(D)$ to*

$$-\nabla \cdot A_\varepsilon \nabla u_\varepsilon \;=\; f \tag{1.5}$$

is such that

$$u_\varepsilon \;\rightharpoonup\; u_0 \quad \textit{weakly in } H^1_0(D), \tag{1.6}$$

$$A_\varepsilon \nabla u_\varepsilon \;\rightharpoonup\; A_0 \nabla u_0 \quad \textit{weakly in } L^2(D,\mathbb{R}^d), \tag{1.7}$$

where u_0 is the unique weak solution in $H^1_0(D)$ to

$$-\nabla \cdot A_0 \nabla u_0 \;=\; f. \tag{1.8}$$

This definition makes sense due to the following four properties.

Lemma 1. (1) *(uniqueness) The H-limit of a H-converging sequence $A_\varepsilon \in \mathcal{M}_{\alpha\beta}$ is unique.*

(2) *(locality) Let A_ε and B_ε be two sequences in $\mathcal{M}_{\alpha\beta}(D)$ which H-converge to some A_0 and B_0, respectively. If for some $\Gamma \subset D$, the sequences A_ε and B_ε coincide on Γ for all ε, then A_0 and B_0 coincide on Γ as well.*

(3) *(compactness) Let A_ε be a sequence in $\mathcal{M}_{\alpha\beta}(D)$. Then there exists $A_0 \in \mathcal{M}_{\alpha,\beta^2/\alpha}(D)$, such that A_ε H-converges to A_0 up to extraction.*

(4) *(Urysohn property) A sequence A_ε of $\mathcal{M}_{\alpha\beta}(D)$ H-converges if and only if all its H-converging subsequences have the same limit.*

The definition of H-converges ensures that the weak solution u_ε to (1.5) converges weakly to the weak solution u_0 to (1.8) in $H^1_0(D)$. In particular, ∇u_ε does not necessarily converge strongly to ∇u_0 in $L^2(D,\mathbb{R}^d)$. The defect of strong convergence can be compensated by the introduction of a corrector field.

Definition 2. *Let A_ε be a sequence of $\mathcal{M}_{\alpha\beta}(D)$ which H-converges to some A_0. For all $\varepsilon > 0$, we define the corrector matrix $C_\varepsilon \in L^2(D, \mathcal{M}_d)$ by: for all $i, j \in \{1, \dots, d\}$,*

$$(C_\varepsilon)_{ij} = \frac{\partial w_\varepsilon^j}{\partial y_i},$$

where w_ε^j is the weak solution in $H_0^1(D)$ to

$$-\nabla \cdot A_\varepsilon \nabla w_\varepsilon^j = -\nabla \cdot (A_0 \mathbf{e}_j). \tag{1.9}$$

By definition of H-convergence, $A_\varepsilon \nabla w_\varepsilon^j \rightharpoonup A_0 \nabla w_0^j$ weakly in $L^2(D, \mathbb{R}^d)$, and $w_\varepsilon^j \rightharpoonup w_0^j$ weakly in $H^1(D)$ where w_0^j is the unique weak solution in $H_0^1(D)$ to

$$-\nabla \cdot A_0 \nabla w_0^j = -\nabla \cdot (A_0 \mathbf{e}_j).$$

This implies that $\nabla w_0^j \equiv \mathbf{e}_j$, and therefore,

$$C_\varepsilon \rightharpoonup \mathrm{Id} \quad \text{weakly in } L^2(D, \mathcal{M}_d),$$

where Id denotes the identity matrix. In addition H-convergence implies that, denoting by u_ε and u_0 the weak solutions of (1.5) and (1.8),

$$\nabla u_\varepsilon - C_\varepsilon \nabla u_0 \rightharpoonup 0 \quad \text{weakly in } L^1(D, \mathbb{R}^d).$$

We indeed have much better:

Theorem 1. *Suppose that A_ε H-converges to A_0, and let u_ε and u_0 be the weak solutions of (1.5) and (1.8). Let C_ε be given by Definition 2. Then*

$$\nabla u_\varepsilon - C_\varepsilon \nabla u_0 \to 0 \quad \textit{strongly in } L^1(D, \mathbb{R}^d).$$

In addition, if C_ε is bounded in $L^r(D, \mathcal{M}_d)$ for some $2 \le r \le \infty$, and $\nabla u_0 \in L^s(D, \mathbb{R}^d)$ for some $2 \le s < \infty$, then

$$\nabla u_\varepsilon - C_\varepsilon \nabla u_0 \to 0 \quad \textit{strongly in } L^t(D, \mathbb{R}^d)$$

where $t = \min\left\{2, \dfrac{rs}{r+s}\right\}$.

The proof of these results essentially rely on the celebrated div-curl lemma, which will be useful for the numerical analysis as well.

Lemma 2 (div-curl lemma)**.** *Let u_ε and v_ε be two bounded sequences in $L^2(D, \mathbb{R}^d)$, which converge weakly in $L^2(D, \mathbb{R}^d)$ to some u_0 and v_0. If $\nabla \cdot u_\varepsilon$ is compact in $H^{-1}(D)$, and if $\nabla \times v_\varepsilon$ is bounded in $L^2(D, \mathbb{R}^{d \times d})$, where*

$$[\nabla \times v_\varepsilon]_{ij} := \partial_j [v_\varepsilon]_i - \partial_i [v_\varepsilon]_j,$$

then the product $u_\varepsilon \cdot v_\varepsilon$ converges to $u_0 \cdot v_0$ in the sense of distributions.

In view of these results, a natural candidate for the limit of $A^*_{\varepsilon,H}$ as ε and H go to zero is A_0. In the following section we show how H-convergence can be used to prove the convergence of the self-consistent approach.

Throughout the text, we'll make use of the following notation

- d is the space dimension ;
- D is a bounded open Lipschitz domain of $\mathbb{R}^d$;
- $\mathcal{M}_d$ denotes the set of d-dimensional real square matrices ;

- $\mathcal{M}_{\alpha\beta}$ denotes the set of d-dimensional real square matrices which are α-elliptic and β-continuous ;
- $\mathcal{M}_{\alpha\beta}^{\rm sym}$ is the subset of those symmetric matrices of $\mathcal{M}_{\alpha\beta}$;
- for all ρ, $Q_\rho = (-\rho/2, \rho/2)^d$, and we use the short hand notation $Q = Q_1 = (-1/2, 1/2)^d$;
- for all $x \in \mathbb{R}^d$, T_x denotes the translation by x and for every measurable subset B of $\mathbb{R}^d$, $T_x B = \{x + y : y \in B\}$;
- for all $x \in \mathbb{R}^d$, and all $\rho > 0$, $Q_\rho(x) := T_x Q_\rho$;
- $H^1_{\rm per}(Q)$ denotes the closure of smooth Q-periodic function with zero average in the Hilbert space $H^1(Q)$;
- for all $1 \leq p \leq \infty$, $W^{1,p}(D)$ denotes the Sobolev space of p-integrable functions whose distributional derivatives are p-integrable functions ;
- for all $1 \leq p \leq \infty$, $W^{1,p}_0(D)$ denotes the subspace of functions $W^{1,p}(D)$ which vanishe on ∂D in the sense of traces ;
- $\langle \cdot \rangle$ is the ensemble average, that is the periodic average in the periodic case, and the expectation in the random case ;
- ${\rm var}\,[\cdot]$ is the variance associated with the ensemble average ;
- $\lesssim$ and $\gtrsim$ stand for $\leq$ and $\geq$ up to a multiplicative constant which only depends on the dimension d and the coercivity constants (denoted by α, β in the text) if not otherwise stated;
- when both $\lesssim$ and $\gtrsim$ hold, we simply write $\sim$;
- we use $\gg$ instead of $\gtrsim$ when the multiplicative constant is (much) larger than 1;
- $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$ denotes the canonical basis of $\mathbb{R}^d$.

2. Analytical framework by H-convergence

In this section we present an analytical framework to analyze the convergence of numerical homogenization methods in the case of linear elliptic equations in divergence form. These results are proved using a simplified version of the string of arguments used in [27] to treat the case of general multiple integrals. In addition they cover the case of non symmetric matrices (which was not treated in [27]).

2.1. General framework

Let $A_\varepsilon \in \mathcal{M}_{\alpha\beta}(D)$ be a H-convergent sequence whose limit is denoted by $A_{\rm hom} \in \mathcal{M}_{\alpha,\beta^2/\alpha}(D)$. Unlike what we've presented in the self-consistent approach, we focus here on a continuous approximation, and shall only later on discretize the equations. We begin with the definition of a local approximation of $A_{\rm hom}$ on domains of size $\rho > 0$.

Definition 3. *For all $\rho > 0$ and $\varepsilon > 0$, we denote by $A_{\rho,\varepsilon}$ the element of $\mathcal{M}_{\alpha,\beta^2/\alpha}(D)$ defined by: for all $i, j \in \{1, \ldots, d\}$ and for $x \in D$,*

$$[A_{\rho,\varepsilon}(x)]_{ij} := \fint_{Q_\rho \cap T_{-x} D} \mathbf{e}_j \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y))dy, \tag{2.1}$$

where $v_i^{\rho,\varepsilon}(x,\cdot)$ is the unique weak solution in $H^1_0(Q_\rho \cap T_{-x}D)$ to

$$-\nabla \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y)) = 0 \qquad \text{in } Q_\rho \cap T_{-x}D. \tag{2.2}$$

These approximations $A_{\rho,\varepsilon}$ of $A_{\rm hom}$ are similar to the coefficients $A^*_{\varepsilon,H}$ of the self-consistent approach. The fact that $A_{\rho,\varepsilon} \in \mathcal{M}_{\alpha,\beta^2/\alpha}(D)$ is proved as follows.

The weak formulation of (2.2) tested with function $v_i^{\rho,\varepsilon}$ yields

$$\int_{Q_\rho \cap T_{-x}D} (\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y))dy = \int_{Q_\rho \cap T_{-x}D} \mathbf{e}_i \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y))dy.$$

Since $A_\varepsilon \in \mathcal{M}_{\alpha\beta}(D)$, by Cauchy-Schwarz inequality this turns into

$$\alpha \int_{Q_\rho \cap T_{-x}D} |\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y)|^2 dy \leq \beta |Q_\rho \cap T_{-x}D|^{1/2} \left(\int_{Q_\rho \cap T_{-x}D} |\mathbf{e}_i + \nabla_y v_i^{\rho,\varepsilon}(x,y)|^2 dy \right)^{1/2}, \tag{2.3}$$

from which the upper bound follows using the defining equation (2.1). We turn to the lower bound. For all $\xi \in \mathbb{R}^d$, we let $v_\xi^{\rho,\varepsilon}(x,\cdot)$ be the weak solution in $H_0^1(Q_\rho \cap T_{-x}D)$ to

$$-\nabla \cdot A_\varepsilon(x+y)(\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y)) = 0 \qquad \text{in } Q_\rho \cap T_{-x}D.$$

Using the lower bound on A_ε and Jensen's inequality, we have for all $\xi \in \mathbb{R}^d$ with $|\xi| = 1$

$$\begin{aligned} &\int_{Q_\rho \cap T_{-x}D} (\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(x+y)(\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y)) dy \\ &\geq \alpha \int_{Q_\rho \cap T_{-x}D} |\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y))|^2 dy \\ &\geq \alpha |Q_\rho \cap T_{-x}D|, \end{aligned} \tag{2.4}$$

which is the desired lower bound since

$$\xi \cdot A_{\rho,\varepsilon}(x)\xi = \fint_{Q_\rho \cap T_{-x}D} (\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(x+y)(\xi + \nabla_y v_\xi^{\rho,\varepsilon}(x,y)) dy.$$

If $\{A_\varepsilon\}$ is a family of symmetric matrices, (2.2) is the Euler-Lagrange equation associated with the following equivalent definition of (2.1): for all $\xi \in \mathbb{R}^d$,

$$\xi \cdot A_{\rho,\varepsilon}(x)\xi := \inf \left\{ \fint_{Q_\rho \cap T_{-x}D} (\xi + \nabla v(y)) \cdot A_\varepsilon(x+y)(\xi + \nabla v(y)) dy, v \in H_0^1(Q_\rho \cap T_{-x}D) \right\}.$$

The main result of this section is the following theorem.

Theorem 2. *Let A_ε and $A_{\rho,\varepsilon}$ be as in Definition 3, then for all $\rho > 0$ there exists $A_{\rho,\mathrm{hom}} \in \mathcal{M}_{\alpha,\beta^2/\alpha}(D)$ such that for almost every $x \in D$,*

$$\lim_{\varepsilon \to 0} A_{\rho,\varepsilon}(x) = A_{\rho,\mathrm{hom}}(x), \tag{2.5}$$

$$\lim_{\rho \to 0} A_{\rho,\mathrm{hom}}(x) = A_{\mathrm{hom}}(x). \tag{2.6}$$

As a direct corollary we have

Corollary 1. *Let A_ε, $A_{\rho,\varepsilon}$ and $A_{\rho,\mathrm{hom}}$ be as in Theorem 2, and $f \in H^{-1}(D)$. Then, the weak solution $u_{\rho,\varepsilon} \in H_0^1(D)$ to*

$$-\nabla \cdot A_{\rho,\varepsilon} \nabla u_{\rho,\varepsilon} = f$$

satisfies

$$\lim_{\rho \to 0} \lim_{\varepsilon \to 0} \|u_{\rho,\varepsilon} - u_{\mathrm{hom}}\|_{H^1(D)} = 0, \tag{2.7}$$

where $u_{\mathrm{hom}} \in H_0^1(D)$ is the weak solution to

$$-\nabla \cdot A_{\mathrm{hom}} \nabla u_{\mathrm{hom}} = f.$$

As a consequence of H-convergence we also have that

$$\lim_{\rho\to 0}\lim_{\varepsilon\to 0}\|u_{\rho,\varepsilon}-u_\varepsilon\|_{L^2(D)}=0,$$

Let us point out that without any further assumption on A_ε, one cannot get quantitative convergence rates for (2.7). A trivial example is provided by a constant family: $A_\varepsilon := A_{\rm hom}$ for all $\varepsilon>0$. In this case, if $A_{\rm hom}:\mathbb{R}^d\to\mathcal{M}_d$ is Lipschitz continuous, then the convergence rate in (2.7) is $O(\rho)$.

Remark 1. Corollary 1 also holds with general Dirichlet, Neumann, mixed Dirichlet-Neumann boundary conditions.

In the following subsection we shall complete Corollary 1 with a corrector result in order to approximate correctly ∇u_ε in $L^2(D,\mathcal{M}_d)$.

The proof of Theorem 2 relies on three ingredients:

- the definition of H-convergence for (2.5),
- the approximate continuity of integrable functions (see (2.9) below),
- the continuous dependence of solutions to linear elliptic problems with respect to the coefficients of the operator stated in the following lemma.

Lemma 3. *Let $A\in\mathcal{M}_{\alpha\beta}(D)$, $(A_\rho)_{\rho>0}\in\mathcal{M}_{\alpha,\beta^2/\alpha}(D)$ and $f,(f_\rho)_{\rho>0}\in H^{-1}(D)$ be such that $A_\rho\to A$ pointwise in D, and $f_\rho\to f$ in $H^{-1}(D)$ as ρ goes to zero. Then the unique weak solution $u_\rho\in H^1_0(D)$ to*

$$-\nabla\cdot A_\rho\nabla u_\rho=f_\rho$$

converges in $H^1(D)$ to the unique weak solution u in $H^1_0(D)$ to

$$-\nabla\cdot A\nabla u=f.$$

We first prove Theorem 2 and Corollary 1, and then turn to the proof of Lemma 3.

Proof of Theorem 2. Let $x\in D$ and $\rho>0$, and consider problem (2.2). By the locality and definition of H-convergence (see property (2) of Lemma 1 and Definition 1),

$$\begin{array}{l} v_i^{\rho,\varepsilon}(x,\cdot)\rightharpoonup v_i^{\rho,\rm hom}(x,\cdot)\quad\text{in } H^1_0(Q_\rho\cap T_{-x}D),\\ A_\varepsilon(x+\cdot)(\mathbf{e}_i+\nabla_y v_i^{\rho,\varepsilon}(x,\cdot))\rightharpoonup A_{\rm hom}(x+\cdot)(\mathbf{e}_i+\nabla_y v_i^{\rho,\rm hom}(x,\cdot))\quad\text{in } L^2(Q_\rho\cap T_{-x}D,\mathbb{R}^d),\end{array}\tag{2.8}$$

where $v_i^{\rho,\rm hom}(x,\cdot)$ is the unique solution in $H^1_0(Q_\rho\cap T_{-x}D)$ to

$$-\nabla\cdot A_{\rm hom}(x+y)(\mathbf{e}_i+\nabla_y v_i^{\rho,\rm hom}(x,y))=0.$$

Hence, setting

$$[A_{\rho,\rm hom}(x)]_{ij}:=\fint_{Q_\rho\cap T_{-x}D}\mathbf{e}_j\cdot A_{\rm hom}(x+y)(\mathbf{e}_i+\nabla_y v_i^{\rho,\rm hom}(x,y))dy,$$

(2.8) implies the the claim (2.5).

To prove (2.6), we appeal to Lemma 3. To this aim, we note that for all ρ small enough, $Q_\rho(x)\subset D$, so that after a change of variables

$$[A_{\rho,\rm hom}(x)]_{ij}=\fint_Q\mathbf{e}_j\cdot A_{\rm hom}(x+\rho y)(\mathbf{e}_i+\nabla_y v_i^{\rho,\rm hom}(x,\rho y))dy.$$

By the continuity of translations in $L^1(D)$ (see [56] or [24] for instance), since $A_{\text{hom}} \in L^1(D)$, for all $y \in Q = (-1/2, 1/2)^d$ and $B \subset\subset D$ we have

$$\int_B |A_{\text{hom}}(x+\rho y) - A_{\text{hom}}(x)| dx \overset{\rho\to 0}{\to} 0.$$

Integrating over Q and using Fubini's theorem, one obtains

$$\int_B \left(\int_Q |A_{\text{hom}}(x+\rho y) - A_{\text{hom}}(x)| dy \right) dx \overset{\rho\to 0}{\to} 0. \tag{2.9}$$

Consequently, for almost every $x \in B$, and almost every $y \in Q$,

$$A_{\text{hom}}(x+\rho y) \overset{\rho\to 0}{\to} A_{\text{hom}}(x). \tag{2.10}$$

Let now $x \in B$ be such a point, and let then $w_i^\rho \in H_0^1(Q)$ be solutions for $i \in \{1, \dots, d\}$ to

$$-\nabla_y \cdot A_{\text{hom}}(x+\rho y) \nabla_y w_i^\rho(y) = \nabla_y \cdot A_{\text{hom}}(x+\rho y) \mathbf{e}_i.$$

Estimate (2.10) implies that the assumptions of Lemma 3 are satisfied, so that $w_i^\rho \to w_i$ in $H^1(Q)$, where w_i is the unique weak solution in $H_0^1(Q)$ to

$$-\nabla_y \cdot A_{\text{hom}}(x) \nabla_y w_i(y) = \nabla_y \cdot A_{\text{hom}}(x) \mathbf{e}_i = 0. \tag{2.11}$$

Hence, for all $i, j \in \{1, \dots, d\}$

$$\begin{aligned} [A_{\rho,\text{hom}}(x)]_{ij} &= \int_{C(0,1)} \mathbf{e}_j \cdot A_{\text{hom}}(x+\rho y)(\nabla_y w_i^\rho(y) + \mathbf{e}_i) dy \\ &\overset{\rho\to 0}{\to} \int_{C(0,1)} \mathbf{e}_j \cdot A_{\text{hom}}(x)(\nabla_y w_i(y) + \mathbf{e}_i) dy \\ &= [A_{\text{hom}}(x)]_{ij} \end{aligned}$$

since $w_i = 0$ is the trivial solution to (2.11). This concludes the proof of the theorem. □

We now prove Corollary 1.

Proof of Corollary 1. Let $u_{\rho,\text{hom}}$ be the unique weak solution in $H_0^1(D)$ to

$$-\nabla \cdot A_{\rho,\text{hom}} \nabla u_{\rho,\text{hom}} = f.$$

Due to (2.5), Lemma 3 implies

$$\lim_{\varepsilon\to 0} \|u_{\rho,\text{hom}} - u_{\rho,\varepsilon}\|_{H^1(D)} = 0. \tag{2.12}$$

Similarly, from (2.6) we get

$$\lim_{\rho\to 0} \|u_{\rho,\text{hom}} - u_{\text{hom}}\|_{H^1(D)} = 0. \tag{2.13}$$

The claim follows from the combination of (2.12) and (2.13) □

We conclude this subsection by the proof of Lemma 3.

Proof of Lemma 3. Let us substract the weak forms of the two equations tested against the admissible test-function $u_\rho - u \in H^1_0(D)$. This yields

$$\int_D \nabla(u_\rho - u) \cdot (A_\rho \nabla u_\rho - A\nabla u) = \Big(f_\rho - f, u_\rho - u\Big)_{H^{-1}(D),H^1_0(D)},$$

where $\Big(\cdot,\cdot\Big)_{H^{-1}(D),H^1_0(D)}$ denotes the duality product between $H^{-1}(D)$ and $H^1_0(D)$, that we rewrite in the form

$$\int_D \nabla(u_\rho - u) \cdot A_\rho \nabla(u_\rho - u) = -\int_D \nabla(u_\rho - u) \cdot (A_\rho - A)\nabla u + \Big(f_\rho - f, u_\rho - u\Big)_{H^{-1}(D),H^1_0(D)}.$$

Using the uniform coercivity of $A_\rho \in \mathcal{M}_{\alpha,\beta^2/\alpha}(D)$ and Cauchy-Schwarz inequality, this turns into

$$\alpha\|\nabla(u_\rho - u)\|^2_{L^2(D)} \leq \left(\int_D \nabla u \cdot (A_\rho - A)\nabla u\right)^{1/2} \left(\int_D \nabla(u_\rho - u) \cdot (A_\rho - A)\nabla(u_\rho - u)\right)^{1/2} + \|f_\rho - f\|_{H^{-1}(D)}\|u_\rho - u\|_{H^1_0(D)}.$$

The first factor of the first term of the r. h. s. vanishes as $\rho \to 0$ by the Lebesgue dominated convergence theorem since $A_\rho \to A$ pointwise and $0 \leq \nabla u \cdot (A_\rho - A)\nabla u \leq (\beta + \beta^2/\alpha)|\nabla u|^2$. The first factor of the second term of the r. h. s. vanishes as $\rho \to 0$ as well. Since the other terms are bounded using an a priori estimate and Poincaré's inequality, the claim follows. □

2.2. **Numerical corrector**

Definition 4. *Let $H > 0$, $I_H \in \mathbb{N}$, and let $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D in disjoint subdomains of diameter of order H. We define a family (M_H) of approximations of the identity on $L^2(D)$ associated with $Q_{H,i}$: for every $w \in L^2(D)$ and $H > 0$,*

$$M_H(w) = \sum_{i=1}^{I_H} \left(\fint_{Q_{H,i}} w\right) 1_{Q_{H,i}}.$$

With the notation of Corollary 1, we define the numerical corrector $\gamma^{H,i}_{\rho,\varepsilon}$ associated with $u_{\rho,\varepsilon}$ on $Q_{H,i}$ as the unique weak solution in $H^1_0(Q_{H,i})$ to

$$-\nabla \cdot A_\varepsilon\Big(M_H(\nabla u_{\rho,\varepsilon}) + \nabla\gamma^{H,i}_{\rho,\varepsilon}\Big) = 0, \tag{2.14}$$

we set

$$\nabla u^{H,i}_{\rho,\varepsilon} := M_H(\nabla u_{\rho,\varepsilon})|_{Q_{H,i}} + \nabla\gamma^{H,i}_{\rho,\varepsilon}$$

for all $1 \leq i \leq I_H$, and define the corrector as

$$C^H_{\rho,\varepsilon} = \sum_{i=1}^{I_H} \nabla u^{H,i}_{\rho,\varepsilon} 1_{Q_{H,i}}.$$

We then have the following corrector result:

Theorem 3. *Under the assumptions of Corollary 1, the corrector of Definition 4 satisfies*

$$\lim_{\rho,H\to 0} \lim_{\varepsilon\to 0} \Big\|\nabla u_\varepsilon - C^H_{\rho,\varepsilon}\Big\|_{L^p(D)} = 0, \tag{2.15}$$

for all exponents p such that

- $1 \leq p \leq 2$ *if* A_ε *is a family of symmetric matrices,*
- $1 \leq p < 2$ *if* A_ε *is not a family of symmetric matrices.*

In addition, if the r. h. s. $f \in H^{-1}(D)$ *of equation* (1.5) *belongs to* $W^{-1,q}(D)$ *for some* $q > 2$*, then one can take* $p = 2$ *in* (2.15) *even if* A_ε *is not symmetric.*

Remark 2. The order of the limits in H and ρ in (2.15) is not important, and we may take, e.g., $H = \rho \to 0$. However, we have to first let ε go to zero.

The proof of Theorem 3 is rather long and technical. The main idea is to use Tartar's correctors of on each element $Q_{H,i}$ of the partition of D, pass to the limit in ε first, and then in H. Yet this would require us to know $\nabla u_{\rm hom}$ a priori — which we don't. Hence one has to approximate Tartar's correctors themselves using $\nabla u_{\rho,\varepsilon}$ in place of $\nabla u_{\rm hom}$. As we shall see, the proof relies on two main arguments:

- the div-curl lemma to prove the convergence of Tartar's correctors (and of their variants),
- the convergence $\nabla u_{\rho,\varepsilon} \to \nabla u_{\rm hom}$ in $L^2(D,\mathbb{R}^d)$ to prove that the approximations of Tartar's correctors do not spoil the corrector result.

Proof. We recall that u_ε, $u_{\rm hom}$, $u_{\rho,\varepsilon}$, and $u_{\rho,{\rm hom}}$ are solutions in $H^1_0(D)$ to

$$\begin{aligned} -\nabla\cdot A_\varepsilon\nabla u_\varepsilon &= f, && (2.16)\\ -\nabla\cdot A_{\rm hom}\nabla u_{\rm hom} &= f, && (2.17)\\ -\nabla\cdot A_{\rho,\varepsilon}\nabla u_{\rho,\varepsilon} &= f,\\ -\nabla\cdot A_{\rho,{\rm hom}}\nabla u_{\rho,{\rm hom}} &= f, \end{aligned}$$

that for all $1 \leq i \leq I_H$, $\gamma^{H,i}_{\rho,\varepsilon}$ is solution in $H^1_0(Q_{H,i})$ to

$$-\nabla\cdot A_\varepsilon(M_H(\nabla u_{\rho,\varepsilon}) + \nabla\gamma^{H,i}_{\rho,\varepsilon}) = 0, \qquad (2.18)$$

and that $\nabla u^{H,i}_{\rho,\varepsilon} = M_H(\nabla u_{\rho,\varepsilon})|_{Q_{H,i}} + \nabla\gamma^{H,i}_{\rho,\varepsilon}$. Likewise, for all $1 \leq i \leq I_H$ we introduce $\gamma^{H,i}_{\rm hom}$, $\gamma^{H,i}_\varepsilon$, and $\gamma^{H,i}_{\rho,{\rm hom}}$ solutions in $H^1_0(Q_{H,i})$ to

$$\begin{aligned} -\nabla\cdot A_{\rm hom}(M_H(\nabla u_{\rm hom}) + \nabla\gamma^{H,i}_{\rm hom}) &= 0, && (2.19)\\ -\nabla\cdot A_\varepsilon(M_H(\nabla u_\varepsilon) + \nabla\gamma^{H,i}_\varepsilon) &= 0, && (2.20)\\ -\nabla\cdot A_{\rm hom}(M_H(\nabla u_{\rho,{\rm hom}}) + \nabla\gamma^{H,i}_{\rho,{\rm hom}}) &= 0; && (2.21) \end{aligned}$$

and we set

$$\begin{aligned} \nabla u^{H,i}_{\rm hom} &= M_H(\nabla u_{\rm hom})|_{Q_{H,i}} + \nabla\gamma^{H,i}_{\rm hom}, && (2.22)\\ \nabla u^{H,i}_\varepsilon &= M_H(\nabla u_\varepsilon)|_{Q_{H,i}} + \nabla\gamma^{H,i}_\varepsilon, && (2.23)\\ \nabla u^{H,i}_{\rho,{\rm hom}} &= M_H(\nabla u_{\rho,{\rm hom}})|_{Q_{H,i}} + \nabla\gamma^{H,i}_{\rho,{\rm hom}}. && (2.24) \end{aligned}$$

We finally define variants of Tartar's corrector:

$$\begin{aligned} C^H_{\rho,\varepsilon} &= \sum_{i=1}^{I_H} \nabla u^{H,i}_{\rho,\varepsilon} 1_{Q_{H,i}}, \\ C^H_{\rm hom} &= \sum_{i=1}^{I_H} \nabla u^{H,i}_{\rm hom} 1_{Q_{H,i}}, \\ C^H_{\varepsilon} &= \sum_{i=1}^{I_H} \nabla u^{H,i}_{\varepsilon} 1_{Q_{H,i}}, \\ C^H_{\rho,{\rm hom}} &= \sum_{i=1}^{I_H} \nabla u^{H,i}_{\rho,{\rm hom}} 1_{Q_{H,i}}. \end{aligned}$$

We have by the triangle inequality

$$\int_D |\nabla u_\varepsilon - C^H_{\rho,\varepsilon}|^p \lesssim \int_D |\nabla u_\varepsilon - C^H_\varepsilon|^p + \int_D |C^H_\varepsilon - C^H_{\rho,\varepsilon}|^p,$$

and we shall show that

$$\lim_{H\to 0} \limsup_{\varepsilon\to 0} \int_D |\nabla u_\varepsilon - C^H_\varepsilon|^p = 0, \tag{2.25}$$

$$\lim_{\rho\to 0} \limsup_{\varepsilon\to 0} \int_D |C^H_\varepsilon - C^H_{\rho,\varepsilon}|^p = 0, \quad \text{uniformly in } H, \tag{2.26}$$

for all $1 \le p < 2$.

Step 1. Proof of (2.25).

Using the uniform ellipticity of A_ε we write

$$\alpha \int_D |\nabla u_\varepsilon - C^H_\varepsilon|^p \le \int_D \Big[(\nabla u_\varepsilon - C^H_\varepsilon) \cdot A_\varepsilon (\nabla u_\varepsilon - C^H_\varepsilon) \Big]^{p/2}. \tag{2.27}$$

By H-convergence we know that $A_\varepsilon(\nabla u_\varepsilon - C^H_\varepsilon)$ and $(\nabla u_\varepsilon - C^H_\varepsilon)$ converge weakly in $L^2(D,\mathbb{R}^d)$ to $A_{\rm hom}(\nabla u_{\rm hom} - C^H_{\rm hom})$ and $(\nabla u_{\rm hom} - C^H_{\rm hom})$, respectively. This does *not* imply that the limit of the product converges to the product of the limits, and we need to appeal to compensated compactness. We'd like to pass to the limit $\varepsilon \to 0$ in this estimate for $p = 2$. Unfortunately, in general, the integrand only converges in the sense of distributions, not pointwise — so that one cannot take the characteristic function of D as a test function. Yet the result will hold true for any $1 \le p < 2$ — the proof of which is slightly technical.

We let $\varphi \in C^\infty_0(D,[0,1])$ be a (non-negative) function such that $\varphi \in C^\infty_0(Q_{H,i})$ for all $1 \le i \le I_H$, and set $D_\varphi := \{x \in D : \varphi(x) \neq 1\}$. We have by definition of φ and by Hölder's inequality with exponents

$(2/p, 2/(2-p))$ for $p < 2$

$$\begin{aligned}
&\alpha \int_D |\nabla u_\varepsilon - C_\varepsilon^H|^p \\
&\quad \leq \int_D \Big[(\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\varphi\Big]^{p/2} + \int_{D_\varphi} \Big[(\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\Big]^{p/2} \\
&\quad \leq \Big[\int_D (\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\varphi\Big]^{p/2} |D|^{(2-p)/2} \\
&\qquad\qquad + \Big[\int_{D_\varphi} (\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\Big]^{p/2} |D_\varphi|^{(2-p)/2}. \qquad (2.28)
\end{aligned}$$

We begin with the second term of the r. h. s. which we control by

$$\Big[\int_{D_\varphi} (\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\Big]^{p/2} |D_\varphi|^{(2-p)/2} \lesssim (\|\nabla u_\varepsilon\|_{L^2(D)}^p + \|C_\varepsilon^H\|_{L^2(D)}^p)|D_\varphi|^{(2-p)/2}.$$

An a priori estimate combined with Poincaré's inequality on $H_0^1(D)$ yields

$$\|\nabla u_\varepsilon\|_{L^2(D)} \lesssim \|f\|_{H^{-1}(D)}.$$

Likewise, for all $1 \leq i \leq I_H$, by (2.20) & (2.23),

$$\|\nabla u_\varepsilon^{H,i}\|_{L^2(Q_{H,i})}^2 \lesssim \Big| \fint_{Q_{H,i}} \nabla u_\varepsilon \Big|^2 |Q_{H,i}| \leq \int_{Q_{H,i}} |\nabla u_\varepsilon|^2,$$

so that

$$\|C_\varepsilon^H\|_{L^2(D)}^2 = \|\sum_{i=1}^{I_H} \nabla u_\varepsilon^{H,i} 1_{Q_{H,i}}\|_{L^2(D)}^2 \lesssim \sum_{i=1}^{I_H} \int_{Q_{H,i}} |\nabla u_\varepsilon|^2 = \|\nabla u_\varepsilon\|_{L^2(D)}^2 \leq \|f\|_{H^{-1}(D)}^2. \qquad (2.29)$$

Hence,

$$\Big[\int_{D_\varphi} (\nabla u_\varepsilon - C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon - C_\varepsilon^H)\Big]^{p/2} |D_\varphi|^{(2-p)/2} \lesssim \|f\|_{H^{-1}(D)}^p |D_\varphi|^{(2-p)/2}. \qquad (2.30)$$

We now turn to the first term. We apply the div-curl lemma on each $Q_{H,i}$. On the one hand, by H-convergence, $\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i}$ is curl free and converges weakly in $L^2(Q_{H,i}, \mathbb{R}^d)$ to $\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i}$. On the other hand, by H-convergence, $A_\varepsilon(\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i})$ converges weakly in $L^2(Q_{H,i}, \mathbb{R}^d)$ to $A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})$, and its divergence is bounded by $2\|f\|_{H^{-1}(D)}$ in $H^{-1}(Q_{H,i})$. Hence, the product $(\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i})\cdot A_\varepsilon(\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i})$ converges to $(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\cdot A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})$ in the sense of distributions on $Q_{H,i}$. This implies in particular by definition of φ that

$$\int_{Q_{H,i}} (\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i})\cdot A_\varepsilon(\nabla u_\varepsilon - \nabla u_\varepsilon^{H,i})\varphi \overset{\varepsilon\to 0}{\to} \int_{Q_{H,i}} (\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\cdot A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\varphi.$$

Since the integrand $(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\cdot A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})$ is non-negative and $\varphi \leq 1$,

$$\int_{Q_{H,i}} (\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\cdot A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\varphi \leq \int_{Q_{H,i}} (\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i})\cdot A_{\rm hom}(\nabla u_{\rm hom} - \nabla u_{\rm hom}^{H,i}),$$

so that by (2.28), (2.30), and the definition of $C^H_{\rm hom}$,

$$\limsup_{\varepsilon\to 0}\int_D |\nabla u_\varepsilon - C^H_\varepsilon|^p \lesssim \|f\|^p_{H^{-1}(D)}|D_\varphi|^{(2-p)/2} + \bigg[\int_D (\nabla u_{\rm hom} - C^H_{\rm hom})\cdot A_{\rm hom}(\nabla u_{\rm hom} - C^H_{\rm hom})\bigg]^{p/2}.$$

Since φ is arbitrary, $|D_\varphi|$ can be chosen as small as desired, and the above estimate turns into

$$\limsup_{\varepsilon\to 0}\int_D |\nabla u_\varepsilon - C^H_\varepsilon|^p \lesssim \bigg[\int_D (\nabla u_{\rm hom} - C^H_{\rm hom})\cdot A_{\rm hom}(\nabla u_{\rm hom} - C^H_{\rm hom})\bigg]^{p/2}. \tag{2.31}$$

It remains to prove that the r. h. s. goes to zero as H vanishes. To this aim we use the approximation M_H of identity. In particular, by the triangle inequality

$$\begin{aligned}
&\int_D (\nabla u_{\rm hom} - C^H_{\rm hom})\cdot A_{\rm hom}(\nabla u_{\rm hom} - C^H_{\rm hom})\\
&\quad= \int_D (\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom}) + M_H(\nabla u_{\rm hom}) - C^H_{\rm hom})\cdot A_{\rm hom}(\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom}) + M_H(\nabla u_{\rm hom}) - C^H_{\rm hom})\\
&\quad\le \int_D \Big(\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom})\Big)\cdot A_{\rm hom}\Big(\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom})\Big)\\
&\qquad+ \int_D \Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\cdot A_{\rm hom}\Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\\
&\qquad+ \int_D \Big|\Big(\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom})\Big)\cdot A_{\rm hom}\Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\Big|\\
&\qquad+ \int_D \Big|\Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\cdot A_{\rm hom}\Big(\nabla u_{\rm hom} - M_H(\nabla u_{\rm hom})\Big)\Big|.
\end{aligned}\tag{2.32}$$

On the one hand, by the triangle inequality, M_H is a contraction on $L^2(D)$, so that

$$\|M_H(\nabla u_{\rm hom})\|_{L^2(D)} \le \|\nabla u_{\rm hom}\|_{L^2(D)} \lesssim \|f\|_{H^{-1}(D)},$$

and on the other hand it follows from (2.19) & (2.22) (the proof is similar to (2.29)) that

$$\|C^H_{\rm hom}\|_{L^2(D)} \lesssim \|f\|_{H^{-1}(D)}. \tag{2.33}$$

Hence, the first, third, and last terms of (2.32) vanish as $H\to 0$.

We therefore focus on the second term:

$$\begin{aligned}
&\int_D \Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\cdot A_{\rm hom}\Big(M_H(\nabla u_{\rm hom}) - C^H_{\rm hom}\Big)\\
&\qquad\qquad= \sum_{i=1}^{I_H}\int_{Q_{H,i}} (M_H(\nabla u_{\rm hom}) - \nabla u^{H,i}_{\rm hom})\cdot A_{\rm hom}(M_H(\nabla u_{\rm hom}) - \nabla u^{H,i}_{\rm hom}).
\end{aligned}$$

By (2.19) & (2.22),

$$\int_{Q_{H,i}} (M_H(\nabla u_{\rm hom}) - \nabla u^{H,i}_{\rm hom})\cdot A_{\rm hom}\nabla u^{H,i}_{\rm hom} = 0.$$

for all $1 \leq i \leq I_H$ so that

$$\begin{aligned}
&\int_{Q_{H,i}} (M_H(\nabla u_{\rm hom}) - \nabla u_{\rm hom}^{H,i}) \cdot A_{\rm hom}(M_H(\nabla u_{\rm hom}) - \nabla u_{\rm hom}^{H,i}) \\
&\quad = \int_{Q_{H,i}} (M_H(\nabla u_{\rm hom}) - \nabla u_{\rm hom}^{H,i}) \cdot A_{\rm hom} M_H(\nabla u_{\rm hom}) \\
&\quad = \int_{Q_{H,i}} M_H(\nabla u_{\rm hom}) \cdot A_{\rm hom} M_H(\nabla u_{\rm hom}) - \int_{Q_{H,i}} \nabla u_{\rm hom}^{H,i} \cdot \Big(\fint_{Q_{H,i}} A_{\rm hom} \Big) M_H(\nabla u_{\rm hom}) \\
&\qquad\quad + \int_{Q_{H,i}} \nabla u_{\rm hom}^{H,i} \cdot \Big(\fint_{Q_{H,i}} A_{\rm hom} - A_{\rm hom} \Big) M_H(\nabla u_{\rm hom}) \\
&\quad = \int_{Q_{H,i}} \nabla u_{\rm hom}^{H,i} \cdot \Big(\fint_{Q_{H,i}} A_{\rm hom} - A_{\rm hom} \Big) M_H(\nabla u_{\rm hom}),
\end{aligned}$$

using that $\fint_{Q_{H,i}} \nabla u_{\rm hom}^{H,i} = M_H(\nabla u_{\rm hom})|_{Q_{H,i}}$. Hence we need to prove that

$$\lim_{H \to 0} \int_D C_{\rm hom}^H \cdot \Big(M_H(A_{\rm hom}) - A_{\rm hom} \Big) M_H(\nabla u_{\rm hom}) \, = \, 0. \tag{2.34}$$

Since M_H converges to Id in $L^2(D)$, the second factor of the integrand converges to zero in $L^2(D)$. The result essentially follows from the Lebesgue dominated convergence theorem, although one needs to take care of the first factor of the integrand which depends on H as well. We conclude as follows. Let $\{u_{{\rm hom},\lambda}\}_\lambda$ be a sequence of λ-Lipschitz functions which converges to $u_{\rm hom}$ in $H^1(D)$ as λ goes to infinity. By definition of M_H, $M_H(\nabla u_{{\rm hom},\lambda})$ is essentially bounded by λ for all $H > 0$. We then write

$$\begin{aligned}
&\int_D C_{\rm hom}^H \cdot \Big(M_H(A_{\rm hom}) - A_{\rm hom} \Big) M_H(\nabla u_{\rm hom}) \\
&\qquad\qquad = \int_D C_{\rm hom}^H \cdot \Big(M_H(A_{\rm hom}) - A_{\rm hom} \Big) M_H(\nabla u_{{\rm hom},\lambda}) \\
&\qquad\qquad\qquad + \int_D C_{\rm hom}^H \cdot \Big(M_H(A_{\rm hom}) - A_{\rm hom} \Big) M_H(\nabla u_{\rm hom} - \nabla u_{{\rm hom},\lambda}).
\end{aligned}$$

The first term of the r. h. s. converges to zero as H vanishes by the Cauchy-Schwarz inequality since $|M_H(\nabla u_{{\rm hom},\lambda})| \leq \lambda$. The second term of the r. h. s. is bounded by a constant times $\|f\|_{H^{-1}(D)} \|\nabla u_{\rm hom} - \nabla u_{{\rm hom},\lambda}\|_{L^2(D)}$ using (2.33) and that M_H is a contraction on $L^2(D)$. Hence it converges to zero uniformly in H as $\lambda \to \infty$. We have thus proved (2.34), and therefore using (2.32) that

$$\lim_{H \to 0} \int_D (\nabla u_{\rm hom} - C_{\rm hom}^H) \cdot A_{\rm hom} (\nabla u_{\rm hom} - C_{\rm hom}^H) \, = \, 0,$$

and finally using (2.31) that

$$\lim_{H \to 0} \limsup_{\varepsilon \to 0} \int_D |\nabla u_\varepsilon - C_\varepsilon^H|^p \, = \, 0,$$

as desired.

Step 2. Proof of (2.26).

By the uniform ellipticity of A_ε,

$$\alpha \int_D |C_\varepsilon^H - C_{\rho,\varepsilon}^H|^p \, \leq \, \int_D \Big[(C_{\rho,\varepsilon}^H - C_\varepsilon^H) \cdot A_\varepsilon (C_{\rho,\varepsilon}^H - C_\varepsilon^H) \Big]^{p/2}. \tag{2.35}$$

The same string of arguments leading to (2.31) in Step 1 (using compensated compactness) allows to pass to the limsup in ε in (2.35), and yields

$$\limsup_{\varepsilon\to 0}\int_D |C^H_\varepsilon - C^H_{\rho,\varepsilon}|^p \lesssim \left[\int_D (C^H_{\rho,\mathrm{hom}} - C^H_{\mathrm{hom}})\cdot A_{\mathrm{hom}}(C^H_{\rho,\mathrm{hom}} - C^H_{\mathrm{hom}})\right]^{p/2}. \tag{2.36}$$

Using then equations (2.19) & (2.22) and (2.21) & (2.24) on each $Q_{H,i}$ which yield

$$\begin{aligned}
&\int_{Q_{H,i}} (C^H_{\rho,\mathrm{hom}} - C^H_{\mathrm{hom}})\cdot A_{\mathrm{hom}}(C^H_{\rho,\mathrm{hom}} - C^H_{\mathrm{hom}})\\
&\quad= \int_{Q_{H,i}} (\nabla u^{H,i}_{\rho,\mathrm{hom}} - \nabla u^{H,i}_{\mathrm{hom}})\cdot A_{\mathrm{hom}}(\nabla u^{H,i}_{\rho,\mathrm{hom}} - \nabla u^{H,i}_{\mathrm{hom}})\\
&\quad= \int_{Q_{H,i}} (M_H(\nabla u_{\mathrm{hom}}) - M_H(\nabla u_{\rho,\mathrm{hom}}))\cdot A_{\mathrm{hom}}(\nabla u^{H,i}_{\rho,\mathrm{hom}} - \nabla u^{H,i}_{\mathrm{hom}})\\
&\quad= \int_{Q_{H,i}} (M_H(\nabla u_{\mathrm{hom}}) - M_H(\nabla u_{\rho,\mathrm{hom}}))\cdot A_{\mathrm{hom}}(C^H_{\rho,\mathrm{hom}} - C^H_{\mathrm{hom}}),
\end{aligned}$$

and using the a priori estimates

$$\|C^H_\rho\|_{L^2(D)}, \|C^H_{\rho,\mathrm{hom}}\|_{L^2(D)} \lesssim \|f\|_{H^{-1}(D)}$$

(whose proofs are similar to (2.29)), (2.36) turns into

$$\begin{aligned}
\limsup_{\varepsilon\to 0}\int_D |C^H_\varepsilon - C^H_{\rho,\varepsilon}|^p &\lesssim \left[\Big|\int_D (M_H(\nabla u_{\mathrm{hom}}) - M_H(\nabla u_{\rho,\mathrm{hom}}))\cdot A_{\mathrm{hom}}(C^H_{\mathrm{hom}} - C^H_{\rho,\mathrm{hom}})\Big|\right]^{p/2}\\
&\lesssim \|M_H(\nabla u_{\mathrm{hom}} - \nabla u_{\rho,\mathrm{hom}})\|^{p/2}_{L^2(D)}(\|C^H_\rho\|^{p/2}_{L^2(D)} + \|C^H_{\rho,\mathrm{hom}}\|^{p/2}_{L^2(D)})\\
&\lesssim \|\nabla u_{\mathrm{hom}} - \nabla u_{\rho,\mathrm{hom}}\|^{p/2}_{L^2(D)}\|f\|^{p/2}_{H^{-1}(D)},
\end{aligned}$$

which converges to zero as $\rho\to 0$ by Corollary 1.

Step 3. Extensions.

The starting point is as in Step 1:

$$\begin{aligned}
\int_D |\nabla u_\varepsilon - C^H_{\rho,\varepsilon}|^2 &\lesssim \int_D (\nabla u_\varepsilon - C^H_{\rho,\varepsilon})\cdot A_\varepsilon(\nabla u_\varepsilon - C^H_{\rho,\varepsilon})\\
&\le 2\int_D (\nabla u_\varepsilon - C^H_\varepsilon)\cdot A_\varepsilon(\nabla u_\varepsilon - C^H_\varepsilon) + 2\int_D (C^H_\varepsilon - C^H_{\rho,\varepsilon})\cdot A_\varepsilon(C^H_\varepsilon - C^H_{\rho,\varepsilon}).
\end{aligned}$$

By symmetry we then have

$$\int_D (\nabla u_\varepsilon - C^H_\varepsilon)\cdot A_\varepsilon(\nabla u_\varepsilon - C^H_\varepsilon) = \int_D \nabla u_\varepsilon\cdot A_\varepsilon\nabla u_\varepsilon + \sum_{i=1}^{I_H}\int_D \nabla u^{H,i}_\varepsilon\cdot A_\varepsilon\nabla u^{H,i}_\varepsilon - 2\sum_{i=1}^{I_H}\int_{Q_{H,i}} \nabla u^{H,i}_\varepsilon\cdot A_\varepsilon\nabla u_\varepsilon. \tag{2.37}$$

Extending $u^{H,i}_\varepsilon$ by zero on $D\setminus Q_{H,i}$, we obtain a function in $H^1_0(D)$, and the last term of this identity turns into

$$\sum_{i=1}^{I_H}\int_{Q_{H,i}} \nabla u^{H,i}_\varepsilon\cdot A_\varepsilon\nabla u_\varepsilon = \sum_{i=1}^{I_H}\left(f, u^{H,i}_\varepsilon\right)_{H^{-1}(D),H^1_0(D)}$$

using the weak form of the defining equation (2.16). Since $u_\varepsilon^{H,i}$ converges weakly to $u_{\rm hom}^{H,i}$ in $H_0^1(Q_{H,i})$ (and therefore in $H_0^1(D)$), we obtain

$$\lim_{\varepsilon\to 0}\sum_{i=1}^{I_H}\int_{Q_{H,i}}\nabla u_\varepsilon^{H,i}\cdot A_\varepsilon\nabla u_\varepsilon \;=\; \sum_{i=1}^{I_H}\left(f,u_{\rm hom}^{H,i}\right)_{H^{-1}(D),H_0^1(D)},$$

and therefore

$$\lim_{\varepsilon\to 0}\sum_{i=1}^{I_H}\int_{Q_{H,i}}\nabla u_\varepsilon^{H,i}\cdot A_\varepsilon\nabla u_\varepsilon \;=\; \sum_{i=1}^{I_H}\int_{Q_{H,i}}\nabla u_{\rm hom}^{H,i}\cdot A_{\rm hom}\nabla u_{\rm hom},$$

using (2.17). On the other hand, H-convergence implies the convergence of the energy, so that

$$\lim_{\varepsilon\to 0}\int_D\nabla u_\varepsilon\cdot A_\varepsilon\nabla u_\varepsilon \;=\; \int_D\nabla u_{\rm hom}\cdot A_{\rm hom}\nabla u_{\rm hom},$$

and for all $1\le i\le I_H$,

$$\lim_{\varepsilon\to 0}\int_{Q_{H,i}}\nabla u_\varepsilon^{H,i}\cdot A_\varepsilon\nabla u_\varepsilon^{H,i} \;=\; \int_{Q_{H,i}}\nabla u_{\rm hom}^{H,i}\cdot A_{\rm hom}\nabla u_{\rm hom}^{H,i}$$

(the proof of which is as above, starting from (2.16) & (2.17) and (2.20) & (2.19)).

Putting things back together yields the desired identity

$$\begin{aligned}\limsup_{\varepsilon\to 0}\int_D|\nabla u_\varepsilon-C_\varepsilon^H|^2 \lesssim \limsup_{\varepsilon\to 0}\int_D(\nabla u_\varepsilon-C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon-C_\varepsilon^H)\\ =\int_D(\nabla u_{\rm hom}-C_{\rm hom}^H)\cdot A_{\rm hom}(\nabla u_{\rm hom}-C_{\rm hom}^H).\end{aligned}$$

Likewise,

$$\begin{aligned}\limsup_{\varepsilon\to 0}\int_{Q_{H,i}}|C_\varepsilon^H-C_{\rho,\varepsilon}^H|^2 \lesssim \limsup_{\varepsilon\to 0}\int_D(C_\varepsilon^H-C_{\rho,\varepsilon}^H)\cdot A_\varepsilon(C_\varepsilon^H-C_{\rho,\varepsilon}^H)\\ =\int_D(C_{\rm hom}^H-C_{\rho,{\rm hom}}^H)\cdot A_{\rm hom}(C_{\rm hom}^H-C_{\rho,{\rm hom}}^H).\end{aligned}$$

We then finish the proof as in Step 2, which yields the result for $p=2$.

When the matrix A_ε is not symmetric and the r. h. s. f is in $W^{-1,q}(D)$ for some $q>2$, we appeal to Meyers' estimates [49], which yield the higher integrability result for all $\rho,\varepsilon>0$:

$$\|\nabla u_{\rm hom}\|_{L^q(D)},\|\nabla u_\varepsilon\|_{L^q(D)},\|\nabla u_{\rho,\varepsilon}\|_{L^q(D)},\|\nabla u_{\rho,{\rm hom}}\|_{L^q(D)}\;\lesssim\;\|f\|_{W^{-1,q}(D)},$$

provided $q-2$ is small enough (this exponent only depends on the ellipticity constants α and β). This allows us to take $p=2$ in (2.27) and replace (2.28) by

$$\begin{aligned}&\alpha\int_D|\nabla u_\varepsilon-C_\varepsilon^H|^2\\ &\quad\le\quad\int_D(\nabla u_\varepsilon-C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon-C_\varepsilon^H)\varphi+\int_{D_\varphi}(\nabla u_\varepsilon-C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon-C_\varepsilon^H)\\ &\quad\le\quad\int_D(\nabla u_\varepsilon-C_\varepsilon^H)\cdot A_\varepsilon(\nabla u_\varepsilon-C_\varepsilon^H)\varphi+\left[\int_D|\nabla u_\varepsilon-C_\varepsilon^H|^q\right]^{2/q}|D_\varphi|^{(q-2)/q},\end{aligned}$$

the last term of which can be controlled using Meyers' estimate on the corrector as well. We then conclude the proof as in Step 2, which yields the desired result for $p = 2$ in the non-symmetric case. □

Remark 3. There is some flexibility for the choice of the boundary conditions in the definition of the correctors. In particular, the conclusions of Theorem 3 still hold if the homogeneous Dirichlet boundary conditions in (2.14) are replaced either by homogeneous Neumann boundary conditions, or even a zero average condition (that is $⨍_{Q_{H,i}} \nabla\gamma^{H,i}_{\rho,\varepsilon} = 0$).

The proof for nonsymmetric A_ε remains essentially unchanged. Only the proof for symmetric A_ε has to be slightly adapted to treat the other boundary conditions. In particular, the right way to write the double product in (2.37) is, for Neumann boundary conditions,

$$\begin{aligned}
\int_{Q_{H,i}} \nabla u_\varepsilon \cdot A_\varepsilon \nabla u^{H,i}_\varepsilon &= \int_{Q_{H,i}} \nabla u_\varepsilon \cdot A_\varepsilon M_H(\nabla u_\varepsilon) = \Big(⨍_{Q_{H,i}} \nabla u_\varepsilon\Big) \cdot \int_{Q_{H,i}} A_\varepsilon \nabla u_\varepsilon \\
&\overset{\varepsilon\to 0}{\longrightarrow} \Big(⨍_{Q_{H,i}} \nabla u_{\rm hom}\Big) \cdot \int_{Q_{H,i}} A_{\rm hom} \nabla u_{\rm hom} = \int_{Q_{H,i}} \nabla u_{\rm hom} \cdot A_{\rm hom} M_H(\nabla u_{\rm hom}) \\
&= \int_{Q_{H,i}} \nabla u_{\rm hom} \cdot A_{\rm hom} \nabla u^{H,i}_{\rm hom},
\end{aligned}$$

and, for the zero average condition,

$$\begin{aligned}
\int_{Q_{H,i}} \nabla u_\varepsilon \cdot A_\varepsilon \nabla u^{H,i}_\varepsilon &= \int_{Q_{H,i}} \Big(⨍_{Q_{H,i}} \nabla u_\varepsilon\Big) \cdot A_\varepsilon \nabla u^{H,i}_\varepsilon + \int_{Q_{H,i}} \Big(\nabla u_\varepsilon - ⨍_{Q_{H,i}} \nabla u_\varepsilon\Big) \cdot A_\varepsilon \nabla u^{H,i}_\varepsilon \\
&= \Big(⨍_{Q_{H,i}} \nabla u_\varepsilon\Big) \cdot \int_{Q_{H,i}} A_\varepsilon \nabla u^{H,i}_\varepsilon \\
&\overset{\varepsilon\to 0}{\longrightarrow} \Big(⨍_{Q_{H,i}} \nabla u_{\rm hom}\Big) \cdot \int_{Q_{H,i}} A_{\rm hom} \nabla u_{\rm hom} = \int_{Q_{H,i}} \nabla u_{\rm hom} \cdot A_{\rm hom} M_H(\nabla u_{\rm hom}) \\
&= \int_{Q_{H,i}} \nabla u_{\rm hom} \cdot A_{\rm hom} \nabla u^{H,i}_{\rm hom}.
\end{aligned}$$

2.3. **Direct approach**

We are now in position to introduce rigorously the direct approach, which consists in approximating $u_{\rho,\varepsilon}$ for some $\rho > \varepsilon$ on the one hand, and then construct a numerical corrector on the other hand. We present the method for a Galerkin approach. Let $\{V_H\}$ be a suitable sequence of finite-dimensional subspaces of $H^1_0(D)$. We then denote by $u^H_{\rho,\varepsilon} \in V_H$ the unique weak solution in V_H to

$$-\nabla \cdot A_{\rho,\varepsilon} \nabla u^H_{\rho,\varepsilon} \;=\; f,$$

where $A_{\rho,\varepsilon}$ is defined by (2.1). From Theorem 2, Corollary 1 and standard approximation arguments, we deduce

$$\limsup_{H,\rho\to 0} \lim_{\varepsilon\to 0} \|u^H_{\rho,\varepsilon} - u_{\rm hom}\|_{H^1(D)} \;=\; 0, \tag{2.38}$$

where the limits in H and ρ commute.

In practice, the matrix $A_{\rho,\varepsilon}$ is itself approximated. In particular, for all $x \in D$ and $h > 0$, denoting by $V_h(x)$ a finite dimensional subspace of $H^1_0(Q_\rho \cap T_{-x}D)$, we define an approximation $A^h_{\rho,\varepsilon}$ of $A_{\rho,\varepsilon}$ at point x by: for all $i, j \in \{1, \dots, d\}$

$$[A^h_{\rho,\varepsilon}(x)]_{ij} := ⨍_{Q_\rho \cap T_{-x}D} \mathbf{e}_j \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v^{\rho,\varepsilon,h}_i(x,y))dy,$$

where $v_i^{\rho,\varepsilon,h}(x,\cdot)$ is the unique weak solution in $V_h(x)$ to

$$-\nabla\cdot A_\varepsilon(x+y)(\mathbf{e}_i+\nabla_y v_i^{\rho,\varepsilon,h}(x,y))=0 \qquad \text{in } Q_\rho\cap T_{-x}D.$$

We then define for all $\varepsilon>0$, $\rho>\varepsilon$, $H>0$ and $h<\rho$ the weak solution $u_{\rho,\varepsilon}^{H,h}$ in V_H to

$$-\nabla\cdot A_{\rho,\varepsilon}^h\nabla u_{\rho,\varepsilon}^{H,h} \;=\; f.$$

A further approximation argument then yields

$$\lim_{H,\rho\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0}\|u_{\rho,\varepsilon}^{H,h}-u_{\hom}\|_{H^1(D)} \;=\; 0. \tag{2.39}$$

Note that the practical implementation of the method makes use of a quadrature rule on D so that $A_{\rho,\varepsilon}^h$ only has to be calculated at the quadrature points of D.

Let us give the argument for (2.39). The first two limits yield by convergence of the Galerkin method and H-convergence

$$\lim_{\varepsilon\to 0}\lim_{h\to 0}\|u_{\rho,\varepsilon}^{H,h}-u_{\hom}\|_{H^1(D)} \;=\; \|u_{\rho,\hom}^{H}-u_{\hom}\|_{H^1(D)}.$$

By the triangle inequality

$$\|u_{\rho,\hom}^{H}-u_{\hom}\|_{H^1(D)} \;\le\; \|u_{\rho,\hom}^{H}-u_{\rho,\hom}\|_{H^1(D)}+\|u_{\rho,\hom}-u_{\hom}\|_{H^1(D)}.$$

The first term of the r. h. s. goes to zero as $H\to 0$ by convergence of the Galerkin approximation. We need to understand how the convergence depends on ρ. From Céa's lemma and Poincaré's inequality, we have

$$\|u_{\rho,\hom}^{H}-u_{\rho,\hom}\|_{H^1(D)} \;\lesssim\; \inf_{v_H\in V_H}\|v_H-u_{\rho,\hom}\|_{H^1(D)},$$

which, using the triangle inequality, turns into

$$\|u_{\rho,\hom}^{H}-u_{\rho,\hom}\|_{H^1(D)} \;\lesssim\; \|u_{\rho,\hom}-u_{\hom}\|_{H^1(D)}+\inf_{v_H\in V_H}\|v_H-u_{\hom}\|_{H^1(D)}.$$

We thus have

$$\|u_{\rho,\hom}^{H}-u_{\hom}\|_{H^1(D)} \;\lesssim\; \|u_{\rho,\hom}-u_{\hom}\|_{H^1(D)}+\inf_{v_H\in V_H}\|v_H-u_{\hom}\|_{H^1(D)},$$

which vanishes as ρ and H go to zero (independently, as desired). This shows (2.39).

We may then turn to the numerical corrector result. As in Definition 4 we let $I_H\in\mathbb{N}$, and $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D in disjoint subdomains of diameter of order H. For all $h>0$ and $i\in[\![1,I_H]\!]$ we let $V_{H,i,h}$ be a Galerkin subspace of $H_0^1(Q_{H,i})$. We define the numerical correctors $\gamma_{\rho,\varepsilon}^{H,h,i}$ associated with $u_{\rho,\varepsilon}^{H,h}$ on $Q_{H,i}$ as the unique weak solution in $V_{H,i,h}$ to

$$-\nabla\cdot A_\varepsilon\Big(M_H(\nabla u_{\rho,\varepsilon}^{H,h})+\nabla\gamma_{\rho,\varepsilon}^{H,h,i}\Big) \;=\; 0,$$

we set

$$\nabla v_{\rho,\varepsilon}^{H,h,i} := M_H(\nabla u_{\rho,\varepsilon}^{H,h})|_{Q_{H,i}}+\nabla\gamma_{\rho,\varepsilon}^{H,h,i}$$

for all $1\le i\le I_H$, and finally define

$$C_{\rho,\varepsilon}^{H,h} \;=\; \sum_{i=1}^{I_H}\nabla v_{\rho,\varepsilon}^{H,h,i}1_{Q_{H,i}}.$$

We then have the following numerical corrector result:

$$\lim_{\rho,H\to 0}\limsup_{\varepsilon\to 0}\lim_{h\to 0}\left\|\nabla u_\varepsilon - C^{H,h}_{\rho,\varepsilon}\right\|_{L^p(D)} = 0, \tag{2.40}$$

for all exponents p such that

- $1 \le p \le 2$ if A_ε is a family of symmetric matrices,
- $1 \le p < 2$ if A_ε is not a family of symmetric matrices.

In addition, if the r. h. s. $f \in H^{-1}(D)$ of equation (1.5) belongs to $W^{-1,q}(D)$ for some $q > 2$, then one can take $p = 2$ in (2.40) even if A_ε is not symmetric.

This result is essentially Theorem 3, although there is an additional approximation argument needed since $M_H(\nabla u^H_{\rho,\varepsilon}) \neq M_H(\nabla u_{\rho,\varepsilon})$. It is enough to note that

$$\lim_{H,\rho\to 0}\limsup_{\varepsilon\to 0}\|M_H(\nabla u^H_{\rho,\varepsilon}) - M_H(\nabla u_{\rho,\varepsilon})\|_{L^2(D)} \le \limsup_{H,\rho\to 0}\lim_{\varepsilon\to 0}\|\nabla u^H_{\rho,\varepsilon} - \nabla u_{\rho,\varepsilon}\|_{L^2(D)} = 0$$

by (2.38) & (2.39) to conclude.

In this subsection we have proved the convergence of the direct approach to numerical homogenization in the framework of H-convergence. This provides a convergence analysis for the so-called Heterogeneous Multiscale Method (HMM) applied to homogenization problems, as introduced by E et. al. in [17, 18]. It also makes rigorous the numerical corrector approach by Arbogast [6].

2.4. **Dual approach**

As we have already seen, the dual approach consists in approximating u_ε in some adapted Galerkin subspace of $H^1_0(D)$ rather than approximating the H-limit of A_ε first. The Multiscale Finite Element (MsFEM) basis is constructed as follows. For all $H > 0$, let $\mathcal{T}_H$ be a regular mesh of D by tetrahedra of diameter of order H, and let V_H be the associated $P1$-finite element subspace of $H^1_0(D)$. We denote by I_H and J_H the number of tetrehedra in $\mathcal{T}_H$ and the dimension of V_H respectively, and we let $\{\psi_{H,i}\}_{1\le i\le J_H}$ be the associated hat functions generating V_H. For all $\varepsilon > 0$ and all $0 < h < \varepsilon$ we define multiscale hat functions $\{\psi_{H,\varepsilon,h,i}\}_{1\le i\le J_H}$ by their restrictions on the tetrahedra T_H of $\mathcal{T}_H$. In particular, for every tetrahedron T_H of $\mathcal{T}_H$, we let $V_h(T_H)$ be a Galerkin subspace of $H^1_0(T_H)$ and let $\gamma_{H,\varepsilon,h,i}|_{T_H}$ be the unique weak solution in $V_h(T_H)$ to

$$-\nabla\cdot A_\varepsilon(\nabla\psi_{H,i}|_{T_H} + \nabla\gamma_{H,\varepsilon,h,i}) = 0 \qquad \text{in } T_H.$$

and set $\psi_{H,\varepsilon,h,i}|_{T_H} := (\psi_{H,i} + \gamma_{H,\varepsilon,h,i})|_{T_H}$. So defined, the multiscale hat functions $\{\psi_{H,\varepsilon,h,i}\}_{1\le i\le J_H}$ belong to $H^1_0(D)$ and for all $i \in \{1,\dots,J_H\}$, $\psi_{H,\varepsilon,h,i}$ has the same support and the same nodal values as $\psi_{H,i}$. Hence, the multiscale finite element space $V_{H,\varepsilon,h}$ spanned by the multiscale hat functions $\{\psi_{H,\varepsilon,h,i}\}_{1\le i\le J_H}$ is a subspace of $H^1_0(D)$ of dimension J_H.

The approximation $u_{H,\varepsilon,h}$ of u_ε is then defined as the unique weak solution in $V_{H,\varepsilon,h}$ to

$$-\nabla\cdot A_\varepsilon\nabla u_{H,\varepsilon,h} = f. \tag{2.41}$$

We then have the following convergence result:

$$\lim_{H\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0}\|u_\varepsilon - u_{H,\varepsilon,h}\|_{W^{1,p}(D)} = 0 \tag{2.42}$$

for all exponents p such that

- $1 \le p \le 2$ if A_ε is a family of symmetric matrices,
- $1 \le p < 2$ if A_ε is not a family of symmetric matrices.

In addition, if the r. h. s. $f \in H^{-1}(D)$ of equation (1.5) belongs to $W^{-1,q}(D)$ for some $q > 2$, then one can take $p = 2$ in (2.42) even if A_ε is not symmetric.

The proof of (2.42) consists in two steps. Let us recall there is a one-to-one mapping $\mathcal{M}^{H,\varepsilon,h}_{\rm MsFEM}$ from V_H to $V_{H,\varepsilon,h}$. In particular, with every $v_H = \sum_{i=1}^{J_H} \nu_{H,i}\psi_{H,i} \in V_H$ we associate the multiscale finite element function $v_{H,\varepsilon,h} = \mathcal{M}^{H,\varepsilon,h}_{\rm MsFEM}(v_H) := \sum_{i=1}^{J_H} \nu_{H,\varepsilon,h,i}\psi_{H,\varepsilon,h,i} \in V_{H,\varepsilon,h}$ (and vice-versa). We may characterize this mapping using corrector fields. In particular, for every tetrahedron T_H of $\mathcal{T}_H$ and every $j \in \{1,\dots,d\}$ we let $\phi^j_{H,\varepsilon,h}|_{T_H}$ be the unique weak solution in $V_h(T_H)$ to

$$-\nabla \cdot A_\varepsilon(\mathbf{e}_j + \nabla\phi^j_{H,\varepsilon,h}) \,=\, 0 \qquad \text{in } T_H,$$

and we set $\Phi_{H,\varepsilon,h} := (\phi^1_{H,\varepsilon,h},\dots,\phi^d_{H,\varepsilon,h})$. By definition, $\Phi_{H,\varepsilon,h} \in H^1_0(D)$, and we have for all $v_H \in V_H$

$$\mathcal{M}^{H,\varepsilon,h}_{\rm MsFEM}(v_H) \,=\, v_H + \nabla v_H \cdot \Phi_{H,\varepsilon,h}.$$

We denote by $\mathfrak{u}_{H,\varepsilon,h}$ the function of V_H associated with the weak solution $u_{H,\varepsilon,h} \in V_{H,\varepsilon,h}$ of (2.41) through the one-to-one mapping $\mathfrak{u}_{H,\varepsilon,h} = \big(\mathcal{M}^{H,\varepsilon,h}_{\rm MsFEM}\big)^{-1}(u_{H,\varepsilon,h})$. We shall first prove that

$$\lim_{H\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0} \|\mathfrak{u}_{H,\varepsilon,h} - u_{\rm hom}\|_{H^1(D)} \,=\, 0. \tag{2.43}$$

To this aim, we write the weak formulation of (2.41) as follows: for all $v_{H,\varepsilon,h} \in V_{H,\varepsilon,h}$,

$$\int_D \nabla v_{H,\varepsilon,h} \cdot A_\varepsilon \nabla u_{H,\varepsilon,h} \,=\, \langle f, v_{H,\varepsilon,h}\rangle_{H^{-1}(D),H^1_0(D)}\,. \tag{2.44}$$

Let us focus on the l. h. s. of (2.44), use the characterization of the mapping $\mathcal{M}^{H,\varepsilon,h}_{\rm MsFEM}$ from V_H to $V_{H,\varepsilon,h}$ and that functions of V_H are locally affine on $\mathcal{T}_H$, and that $\Phi_{H,\varepsilon,h}$ vanishes on ∂T^i_H for all $1 \le i \le I_H$:

$$\begin{aligned}
&\int_D \nabla v_{H,\varepsilon,h} \cdot A_\varepsilon \nabla u_{H,\varepsilon,h} \\
&\quad= \int_D (\nabla v_H + \nabla v_H \cdot \nabla\Phi_{H,\varepsilon,h}) \cdot A_\varepsilon(\nabla \mathfrak{u}_{H,\varepsilon,h} + \nabla \mathfrak{u}_{H,\varepsilon,h} \cdot \nabla\Phi_{H,\varepsilon,h}) \\
&\quad= \int_D \nabla v_H \cdot \Big[(\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h})A_\varepsilon(\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h})\Big]\nabla \mathfrak{u}_{H,\varepsilon,h} \\
&\quad= \sum_{i=1}^{I_H} |T^i_H|(\nabla v_H)|_{T^i_H}\Big[\fint_{T^i_H} (\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h})A_\varepsilon(\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h})\Big](\nabla \mathfrak{u}_{H,\varepsilon,h})|_{T^i_H} \\
&\quad= \int_D \nabla v_H \cdot A_{H,\varepsilon,h}\nabla \mathfrak{u}_{H,\varepsilon,h},
\end{aligned}$$

where $A_{H,\varepsilon,h}$ is the piecewise constant matrix defined by

$$A_{H,\varepsilon,h} \,=\, \sum_{i=1}^{I_H} \fint_{T^i_H} (\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h}) \cdot A_\varepsilon(\mathrm{Id} + \nabla\Phi_{H,\varepsilon,h})1_{T^i_H}.$$

We then focus on the r. h. s. of (2.44), and assume without loss of generality that $f \in L^\infty(D)$ (the general case can be dealt with by approximation), so that

$$\langle f, v_{H,\varepsilon,h}\rangle_{H^{-1}(D),H^1_0(D)} \,=\, \int_D f v_{H,\varepsilon,h} \,=\, \int_D f v_H + \int_D \nabla v_H \cdot (f\Phi_{H,\varepsilon,h}).$$

We then define $f_{H,\varepsilon,h}$ as

$$f_{H,\varepsilon,h} := f - \nabla\cdot(f\Phi_{H,\varepsilon,h}),$$

and note that by assumption on f we have $f_{H,\varepsilon,h}\in H^{-1}(D)$, so that the equation for $u_{H,\varepsilon,h}\in V_{H,\varepsilon,h}$ turns into an equation for $\mathfrak{u}_{H,\varepsilon,h}\in V_H$: for all $v_H\in V_H$,

$$\int_D \nabla v_H\cdot A_{H,\varepsilon,h}\nabla\mathfrak{u}_{H,\varepsilon,h} = \langle f_{H,\varepsilon,h}, v_H\rangle_{H^{-1}(D),H^1_0(D)}.$$

By H-convergence, for all $H>0$, the sequence $\Phi_{H,\varepsilon} := \lim_{h\to 0}\Phi_{H,\varepsilon,h}$ converges weakly to 0 in $H^1(D)$ as ε vanishes, so that for all $H>0$

$$\lim_{\varepsilon\to 0}\lim_{h\to 0}\|\Phi_{H,\varepsilon,h}\|_{L^2(D)} = 0. \tag{2.45}$$

We are in position to prove that

$$\lim_{H\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0}\|\mathfrak{u}_{H,\varepsilon,h}-u_{\rm hom}\|_{H^1(D)} = 0.$$

Let denote by $u_{H,\rm hom}$ the weak solution in V_H to

$$-\nabla\cdot A_{\rm hom}\nabla u_{H,\rm hom} = f,$$

and by $u^{H,\varepsilon,h}_{\rm hom}$ the weak solution in V_H to

$$-\nabla\cdot A_{\rm hom}\nabla u^{H,\varepsilon,h}_{\rm hom} = f_{H,\varepsilon,h},$$

Then, by the triangle inequality

$$\begin{aligned}\|\nabla\mathfrak{u}_{H,\varepsilon,h}-\nabla u_{\rm hom}\|_{L^2(D)} \le{}& \|\nabla\mathfrak{u}_{H,\varepsilon,h}-\nabla u^{H,\varepsilon,h}_{\rm hom}\|_{L^2(D)}\\ &+\|\nabla u_{H,\rm hom}-\nabla u_{\rm hom}\|_{L^2(D)}+\|\nabla u^{H,\varepsilon,h}_{\rm hom}-\nabla u_{H,\rm hom}\|_{L^2(D)}.\end{aligned} \tag{2.46}$$

We treat the three terms of the r. h. s. separately and start with the first one. The function $\mathfrak{u}_{H,\varepsilon,h}-u^{H,\varepsilon,h}_{\rm hom}$ is the weak solution in V_H to

$$-\nabla\cdot A_{H,\varepsilon,h}\nabla(\mathfrak{u}_{H,\varepsilon,h}-u_{H,\rm hom}) = -\nabla\cdot(A_{\rm hom}-A_{H,\varepsilon,h})\nabla u^{H,\varepsilon,h}_{\rm hom},$$

so that

$$\begin{aligned}\|\nabla\mathfrak{u}_{H,\varepsilon,h}-\nabla u^{H,\varepsilon,h}_{\rm hom}\|_{L^2(D)} &\lesssim \|(A_{\rm hom}-A_{H,\varepsilon,h})\nabla u^{H,\varepsilon,h}_{\rm hom}\|_{L^2(D)}\\ &\le \|(A_{\rm hom}-A_{H,\varepsilon,h})\nabla u_{\rm hom}\|_{L^2(D)}+2\beta\|\nabla u_{\rm hom}-\nabla u_{H,\rm hom}\|_{L^2(D)}+2\beta\|\nabla u_{H,\rm hom}-\nabla u^{H,\varepsilon,h}_{\rm hom}\|_{L^2(D)}.\end{aligned}$$

The first term of the r. h. s. converges to zero as ε and H go to zero by the dominated convergence theorem and using the fact that the following convergence holds pointwise, as in Subsection 2.1,

$$\lim_{H\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0}A_{H,\varepsilon,h} = A_{\rm hom}.$$

The second term coincides with the second term of the r. h. s. of (2.46), and vanishes as $H\to 0$ by convergence of the Galerkin method for the homogenized equation. We now treat the last and third term, which coincides with the third term of the r. h. s. of (2.46). We recall that $u_{H,\rm hom}-u^{H,\varepsilon,h}_{\rm hom}$ is the weak solution in V_H to

$$-\nabla\cdot A_{\rm hom}\nabla(u_{H,\rm hom}-u^{H,\varepsilon,h}_{\rm hom}) = f-f_{H,\varepsilon,h} = -\nabla\cdot(f\Phi_{H,\varepsilon,h}),$$

so that

$$\|\nabla u_{H,\mathrm{hom}} - \nabla u_{\mathrm{hom}}^{H,\varepsilon,h}\|_{L^2(D)} \lesssim \|f\Phi_{H,\varepsilon,h}\|_{L^2(D)}$$

and therefore using (2.45) and the assumption that $f \in L^\infty(D)$, for all $H > 0$,

$$\lim_{\varepsilon\to 0}\lim_{h\to 0} \|\nabla u_{H,\mathrm{hom}} - \nabla u_{\mathrm{hom}}^{H,\varepsilon,h}\|_{L^2(D)} = 0.$$

We have thus proved (2.43).

It remains to note that $\nabla u_{H,\varepsilon,h}$ is the corrector associated with $\nabla \mathfrak{u}_{H,\varepsilon,h}$ and with the partition $\mathcal{T}_H$ of D. Hence, from (2.43) and the same string of arguments as for the direct approach, we deduce that

$$\lim_{H\to 0}\lim_{\varepsilon\to 0}\lim_{h\to 0} \|\nabla u_\varepsilon - \nabla u_{H,\varepsilon,h}\|_{L^p(D)} = 0,$$

for all exponents p such that

- $1 \le p \le 2$ if A_ε is a family of symmetric matrices,
- $1 \le p < 2$ if A_ε is not a family of symmetric matrices.

This implies the desired convergence result (2.42) by Poincaré's inequality for $u_\varepsilon - u_{H,\varepsilon,h} \in W_0^{1,p}(D)$:

Instead of a Galerkin approximation $u_{H,\varepsilon,h}$ of u_ε, we could have considered a Petrov-Galerkin approximation of u_ε (in which case the test-functions are in V_H, not in $V_{H,\varepsilon,h}$). The convergence proof is indeed simpler (one does not need to introduce f_H). This variant will be used in the next section.

3. Resonance, windowing, and oversampling

In the previous section we have introduced an analytical framework and proved the convergence of some numerical homogenization methods within the framework of H-convergence. Quantitative convergence rates further depend on the class of heterogeneities considered. In this section, we provide convergence rates for the simplest heterogeneities possible, that is we assume the coefficients A_ε to be ε-periodic. This allows us to give a complete numerical analysis of the methods, and identify the limiting term in the error. This term is the so-called resonance error. It is related to the boundary conditions used for the corrector. We shall then recall a standard way to reduce the resonance error (windowing and oversampling), check it does indeed reduce the error in the case of periodic structures, and then adapt the analytical framework of Section 2 to include windowing and oversampling.

3.1. Numerical analysis of the periodic case and the resonance error

In this subsection we assume that $A_\varepsilon = A(\cdot/\varepsilon)$ where A is a symmetric $Q = (-1/2, 1/2)^d$-periodic matrix. In this case, the homogenized matrix A_{hom} is symmetric, does not depend on the macroscopic space variable, and is charaterized by: for all $\xi \in \mathbb{R}^d$,

$$\xi \cdot A_{\mathrm{hom}}\xi = \int_Q (\xi + \nabla\phi)\cdot A(\xi + \nabla\phi),$$

where $\phi \in H^1_\#(Q)$ is the unique Q-periodic weak solution to the corrector equation

$$-\nabla \cdot A(\xi + \nabla\phi) = 0.$$

Furthermore, we let $f \in L^\infty(D)$ and D be smooth enough so that by elliptic regularity, the solution $u_{\mathrm{hom}} \in H^1_0(D)$ to the homogenized problem is indeed of class $H^2(D)$.

Then, it is proved in [1, 17, 18] that for the direct approach we have for $\rho \sim H$, $P1$-finite elements for both the macroscopic and the microscopic variables,

$$\|u_{H,\varepsilon}^{H,h} - u_{\text{hom}}\|_{H^1(D)} \lesssim H + \frac{\varepsilon}{H} + \left(\frac{h}{\varepsilon}\right)^2, \tag{3.1}$$

$$\|\nabla u_\varepsilon - \sum_{i=1}^{I_H} \nabla v_{H,\varepsilon}^{H,h,i} 1_{Q_{H,i}}\|_{L^2(D)} \lesssim H + \sqrt{\frac{\varepsilon}{H}} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}. \tag{3.2}$$

For the dual approach, it is shown in [42, 43] that

$$\|u_{H,\varepsilon,h} - u_\varepsilon\|_{H^1(D)} \lesssim H + \sqrt{\frac{\varepsilon}{H}} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}. \tag{3.3}$$

We do not display the proofs of these estimates, and refer the reader to the original papers. Note however that:

- the term $O(H)$ comes from the discretization of D, as it is standard for $P1$-finite elements,
- the term $O(h/\varepsilon)$ in (3.2) and (3.3) comes from the discretizations of the mesh elements T_H and Q_H with a meshsize h such that $h \ll \varepsilon$,
- the term $\sqrt{\varepsilon}$ is a theoretical limit due to boundary layers in the neigborhood of ∂D in periodic homogenization,
- the third term in (3.1) may be surprising at a first glance. This part of the error is indeed driven by the finite element error in the computation of the corrector, which is squared when computing the approximation of the homogenized matrix — due to symmetry.

The limiting term is however the term involving $\frac{\varepsilon}{H}$, which is the inverse of a measure of the number of ε-rescaled periodic cells contained in Q_H or T_H (which is of order $(H/\varepsilon)^d$). The larger $\frac{H}{\varepsilon}$, the more expensive the method, so that there is a trade-off cost/accuracy between H and ε. This error is even more important at the level of the numerical corrector since its square-root appears in (3.2) and (3.3).

There are two sources of this error

- The homogeneous Dirichlet boundary conditions used in (2.2) are not consistent with the periodic boundary conditions of the corrector ϕ,
- The average (2.1) defining $A_{\rho,\varepsilon}$ is not consistent with the average defining A_{hom} since $Q_\rho \cap T_{-x}D$ is not a multiple of periodic cells in general.

A first idea to reduce the error due to boundary conditions on (2.2) consists in imposing the boundary conditions far from the domain of interest, hoping the error is localized on a neighborhood of the boundary. This is indeed the case, as shown on a half plane by Bensoussan, Lions, and Papanicolaou [7]. In particular, this is efficient at the level of the corrector, but not at the level of the homogenized coefficients. To illustrate this, we display the results of two academic series of tests. In the first series of tests, we compare the homogenized coefficients A_{hom} to two different approximations: A_R, which is defined as

$$\xi \cdot A_R \xi := \fint_{Q_R} (\xi + \nabla \phi_R) \cdot A(\xi + \nabla \phi_R)$$

where $Q_R = (-R/2, R/2)^d$ for $R \in \mathbb{N}$, and ϕ_R is the unique solution in $H_0^1(Q_R)$ to

$$-\nabla \cdot A(\xi + \nabla \phi_R) = 0,$$

and $\tilde{A}_R$ defined as

$$\xi \cdot \tilde{A}_R \xi := \fint_Q (\xi + \nabla \phi_R) \cdot A(\xi + \nabla \phi_R).$$

TABLE 1. Error on the approximated homogenized coefficients (performed with [26, *FreeFEM*], smooth periodic coefficients, $P2$-finite elements, and 100 elements per periodic cell).

Number R of periodic cells per dimension	$\lvert A_{\rm hom} - A_R\rvert$			$\lvert A_{\rm hom} - \tilde{A}_R\rvert$		
	Error	Rate of convergence	Prefactor (rate=1)	Error	Rate of convergence	Prefactor (rate=1)
1	0.157	-	0.157	0.157	-	0.157
2	0.0845	0.895	0.169	0.0210	2.90	0.0420
4	0.0433	0.963	0.173	0.0118	0.835	0.0471
8	0.0219	0.983	0.175	0.00597	0.979	0.0478
12	0.0146	1.01	0.175	0.00397	1.00	0.0476
16	0.0110	0.965	0.176	0.00299	0.985	0.0478
20	0.00876	1.03	0.175	0.00239	1.00	0.0478

TABLE 2. L^2-norm of the error on the corrector (performed with [26, *FreeFEM*], smooth periodic coefficients, $P2$-finite elements, and 100 elements per periodic cell).

Number R of periodic cells per dimension	$\mathrm{E}_1(R)$			$\mathrm{E}_2(R)$		
	Error	Rate of convergence	Prefactor (rate=0.5)	Error	Rate of convergence	Prefactor (rate=1)
1	0.210	-	0.210	0.210	-	0.210
2	0.156	0.425	0.221	0.0116	0.893	0.0232
4	0.113	0.468	0.226	0.00361	1.684	0.0144
8	0.0808	0.484	0.229	0.00181	0.988	0.0145
12	0.0662	0.491	0.229	0.00121	1.00	0.0145
16	0.0574	0.496	0.230	0.000910	0.992	0.0146
20	0.0515	0.492	0.230	0.000726	1.02	0.0145

In particular, $\tilde{A}_R$ is the best approximation of $A_{\rm hom}$ one can devise using ϕ_R since the center of the computation domain is the place where the effect of the Dirichlet boundary conditions is expected to be the smallest. As can be seen on Table 1, as expected $|A_{\rm hom} - \tilde{A}_R| \leq |A_{\rm hom} - A_R|$. Yet the rate of convergence in R is -1 in both cases (which corresponds to the term ε/H in (3.1).

In Table 2, we do not compare homogenized coefficients, but rather the correctors themselves, and define two errors:

$$\begin{aligned} \mathrm{E}_1(R) &:= \left(\fint_{Q_R} |\nabla\phi_R - \nabla\phi|^2 \right)^{1/2}, \\ \mathrm{E}_2(R) &:= \left(\fint_{Q} |\nabla\phi_R - \nabla\phi|^2 \right)^{1/2}. \end{aligned}$$

This time, not only the prefactor of the error changes, but also the convergence rate which passes from 1/2 to 1, which will improve both (3.2) and (3.3), as we shall see below.

In order to reduce the second source of error (the fact that the average (2.1) may be calculated on a domain which is not a multiple of the periodic cell), one may use different averaging functions than simply the indicator function which are such that they approximate the mean of a periodic function at a higher order — we call such functions masks, and this approach filtering, see Definition 7 in Section 4.

To fix the vocabulary,

- windowing amounts to approximating the corrector on a larger domain than needed to reduce the effect of spurious boundary conditions,
- filtering amounts to approximating averages by integrals with a weighted measure (using the so-called masks).

3.2. **Windowing and filtering in the direct approach**

For the direct approach, the use of windowing is straightforward to describe. Let $\delta > 1$ be fixed, and $\eta : Q \to \mathbb{R}^+$ be an integrable function of mass one (that is, an averaging mask on Q). For all $r > 0$, we define $\eta_r : Q_r \to \mathbb{R}^+$ by $\eta_r(y) = r^{-d}\eta(y/r)$, which is an averaging mask on Q_r. Then, for all $x \in D$ and $h > 0$, denoting by $V_h(x)$ a finite dimensional subspace of $H^1_0(Q_{\delta\rho} \cap T_{-x}D)$, we define an approximation $A^{\delta,h}_{\rho,\varepsilon}$ of $A_{\rho,\varepsilon}$ at point x by: for all $i, j \in \{1, \dots, d\}$

$$[A^{\delta,h}_{\rho,\varepsilon}(x)]_{ij} := \int_{Q_\rho \cap T_{-x}D} (\mathbf{e}_j + \nabla_y v_j^{\delta\rho,\varepsilon,h}(x,y))) \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\delta\rho,\varepsilon,h}(x,y))\eta_\rho(y)dy,$$

where for all $k \in \{1, \dots, d\}$, $v_k^{\delta\rho,\varepsilon,h}(x,\cdot)$ is the unique weak solution in $V_h(x)$ to

$$-\nabla \cdot A_\varepsilon(x+y)(\mathbf{e}_k + \nabla_y v_k^{\delta\rho,\varepsilon,h}(x,y)) = 0 \qquad \text{in } Q_{\delta\rho} \cap T_{-x}D.$$

We define for all $\varepsilon > 0$, $\rho > \varepsilon$, $H > 0$ and $h < \rho$ the weak solution $u^{\delta,H,h}_{\rho,\varepsilon}$ in V_H to

$$-\nabla \cdot A^{\delta,h}_{\rho,\varepsilon} \nabla u^{\delta,H,h}_{\rho,\varepsilon} \;=\; f.$$

We may then turn to the numerical corrector. As in Definition 4 we let $I_H \in \mathbb{N}$, and $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D into disjoint subdomains of diameter of order H. We further set $Q^\delta_{H,i} := \{x \in D \,|\, d(x, Q_{H,i}) < \delta H\}$, which is an enlarged version of $Q_{H,i}$. For all $h > 0$ and $i \in [\![1, I_H]\!]$ we let $V^\delta_{H,i,h}$ be a Galerkin subspace of $H^1_0(Q^\delta_{H,i})$. We define the numerical correctors $\gamma^{\delta,H,h,i}_{\rho,\varepsilon}$ associated with $u^{\delta,H,h}_{\rho,\varepsilon}$ as the unique weak solution in $V^\delta_{H,i,h}$ to

$$-\nabla \cdot A_\varepsilon\left(M_H(\nabla u^{\delta,H,h}_{\rho,\varepsilon}) + \nabla\gamma^{\delta,H,h,i}_{\rho,\varepsilon}\right) \;=\; 0,$$

we set

$$\nabla v^{\delta,H,h,i}_{\rho,\varepsilon} := M_H(\nabla u^{\delta,H,h}_{\rho,\varepsilon})|_{Q_{H,i}} + (\nabla\gamma^{\delta,H,h,i}_{\rho,\varepsilon})|_{Q_{H,i}}$$

for all $1 \le i \le I_H$, and we define

$$C^{\delta,H}_{\rho,\varepsilon} \;= \sum_{i=1}^{I_H} \nabla v^{\delta,H,h,i}_{\rho,\varepsilon} 1_{Q_{H,i}}.$$

In the case of periodic coefficients, E, Ming and Zhang [18] have essentially proved that (3.1) and (3.2) are replaced by

$$\|u^{\delta,H,h}_{H,\varepsilon} - u_{\hom}\|_{H^1(D)} \;\lesssim\; H + \frac{\varepsilon}{H} + \left(\frac{h}{\varepsilon}\right)^2, \tag{3.4}$$

$$\|\nabla u_\varepsilon - C^{\delta,H,h}_{H,\varepsilon}\|_{L^2(D)} \;\lesssim\; H + \frac{\varepsilon}{H} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}, \tag{3.5}$$

where the multiplicative constant depends on δ. This is an improvement for the reconstruction of the fine scale features since $\sqrt{\frac{\varepsilon}{H}}$ in (3.2) is replaced by $\frac{\varepsilon}{H}$ in (3.5).

Yet, as already mentioned and further studied in [59], the estimate (3.4) is not better than (3.1) in terms of convergence rates. In addition, there are examples in [29, 59] (using reasonable averaging masks) for which

the prefactor in (3.4) is larger than in (3.1). It is therefore not clear whether the use of a mask yields better results than simply taking the average in general. A notable exception is provided in [10], where the mask is not used only as a post-processing to compute the average, but already in the definition of the bilinear form associated with the weak form of the corrector equation. A formal two-scale expansion shows that using this modified corrector equation allows to replace the error corresponding to the second term in (3.1) and (3.4) by $\left(\frac{\varepsilon}{H}\right)^2$ — which is definitely better. We do not give more details on this method since we shall present an even more efficient approach in Section 4.

3.3. **Oversampling in the dual approach**

For the dual approach, windowing is usually called oversampling. We construct a new Multiscale Finite Element (MsFEM) basis as follows. Let $\delta > 1$ be fixed. For all $H > 0$, let $\mathcal{T}_H$ be a regular mesh of D by tetrahedra of diameter of order H, and let V_H be the associated $P1$-finite element subspace of $H^1_0(D)$. We denote by I_H and J_H the number of tetrehedra in $\mathcal{T}_H$ and the dimension of V_H respectively, and we let $\{\psi_{H,i}\}_{1\leq i\leq J_H}$ be the associated hat functions generating V_H. For all $\varepsilon > 0$ and all $0 < h < \varepsilon$ we define multiscale hat functions $\{\psi^\delta_{H,\varepsilon,h,i}\}_{1\leq i\leq J_H}$ by their restrictions on the tetrahedra $(T_k)_{1\leq k\leq I_H}$ of $\mathcal{T}_H$. In particular, for every tetrahedron T_k of $\mathcal{T}_H$, we define $T^\delta_k := \{x \in D \,|\, d(x, T_k) < (\delta - 1)H\}$, denote by $V_h(T^\delta_k)$ a Galerkin subspace of $H^1_0(T^\delta_k)$ and let $\gamma^{\delta,k}_{H,\varepsilon,h,i}$ be the unique weak solution in $V_h(T^\delta_k)$ to

$$-\nabla \cdot A_\varepsilon(\nabla\psi_{H,i}|_{T_k} + \nabla\gamma^{\delta,k}_{H,\varepsilon,h,i}) \,=\, 0 \qquad \text{in } T^\delta_k,$$

and set $\psi^\delta_{H,\varepsilon,h,i}|_{T_k} := (\psi_{H,i} + \gamma^{\delta,k}_{H,\varepsilon,h,i})|_{T_k}$. So defined, the multiscale hat functions $\{\psi^\delta_{H,\varepsilon,h,i}\}_{1\leq i\leq J_H}$ do not belong to $H^1_0(D)$, and we use the notation $\nabla_H\psi^\delta_{H,\varepsilon,h,i}$ for the square integrable function of $L^2(D)$ defined on each element T_k by

$$\nabla_H\psi^\delta_{H,\varepsilon,h,i}|_{T_k} := (\nabla\psi_{H,i} + \nabla\gamma^{\delta,k}_{H,\varepsilon,h,i})|_{T_k}.$$

Note that $\nabla_H\psi^\delta_{H,\varepsilon,h,i}$ is not the distributional derivative of $\psi^\delta_{H,\varepsilon,h,i}$ (it is the absolutely continuous part of this derivative with respect to the Lebesgue measure). Hence, the multiscale finite element space $V^\delta_{H,\varepsilon,h}$ spanned by the multiscale hat functions $\{\psi^\delta_{H,\varepsilon,h,i}\}_{1\leq i\leq J_H}$ is not a subspace of $H^1_0(D)$ (note that it has dimension J_H).

The approximation $u^\delta_{H,\varepsilon,h} \in V^\delta_{H,\varepsilon,h}$ of u_ε can be defined using either a (discontinuous) Galerkin, or using a Petrov-(discontinuous) Galerkin method (for which the test-functions are in V_H, not in $V^\delta_{H,\varepsilon,h}$). We focus on the Petrov-(discontinuous) Galerkin method, and define $u^\delta_{H,\varepsilon,h} \in V^\delta_{H,\varepsilon,h}$ as the unique solution (this statement has to be proved) to: For all $v_H \in V_H$,

$$\int_D \nabla v_H \cdot A_\varepsilon \nabla_H u^\delta_{H,\varepsilon,h} \,=\, \langle f, v_H\rangle_{H^{-1}(D),H^1_0(D)}\,. \tag{3.6}$$

In the case of periodic coefficients, both for the (discontinuous) Galerkin and Petrov-Galerkin method (see Efendiev and Hou [22] and Hou, Wu, and Zhang [44]), (3.3) is replaced by the improved estimate

$$\|u^\delta_{H,\varepsilon,h} - u_\varepsilon\|_{H^1_H(D)} \lesssim H + \frac{\varepsilon}{H} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}, \tag{3.7}$$

where the $\|\cdot\|_{H^1_H(D)}$ is a notation for the broken norm

$$\|v\|^2_{H^1_H(D)} \,=\, \|v\|^2_{L^2(D)} + \|\nabla_H v\|^2_{L^2(D,\mathbb{R}^d)}.$$

In [22], two contributions to the resonance error are identified and are proved to be of the same order. In [44], it is showed that one of these two contributions disappears by using the Petrov-Galerkin formulation. Hence,

the multiplicative constant in (3.7) is expected to be smaller for the Petrov-Galerkin formulation than for the Galerkin formulation — and this is confirmed by numerical tests (the resonance error may even completely disappear in some specific situations, see [44]).

So far we've seen that windowing and oversampling are efficient in the periodic case to reduce the resonance error for the corrector. In the following paragraph, we show that all the convergence results proved for general H-convergent coefficients hold as well with windowing and oversampling.

3.4. **Analytical framework**

In this paragraph, we shall generalize Theorems 2 and 3 to the case of windowing. The proofs we present here are variants of the ones of [28] (which treat the case of nonlinear operators).

We begin with the definition of a local approximation of $A_{\rm hom}$ on domains of size $\rho > 0$.

Definition 5. *Let $\delta > 1$, and let $\eta : Q_\delta \to [0, \delta]$ be a measurable function of mass one (called a mask), such that $\inf_Q \eta \geq 1/\delta$, and for all $\rho > 0$, let $\eta_\rho : C(x, \delta\rho) \to [0, \rho^{-d}\delta]$ be the rescaled version $\eta_\rho : y \mapsto \rho^{-d}\eta(y/\rho)$ of η. For all $\rho > 0$ and $\varepsilon > 0$, we denote by $A^\delta_{\rho,\varepsilon} : D \to \mathcal{M}_d$ the function defined by: for all $i, j \in \{1, \dots, d\}$ and for $x \in D$,*

$$[A^\delta_{\rho,\varepsilon}(x)]_{ij} := \Big(\int_{Q_{\delta\rho} \cap T_{-x}D} \eta_\rho \Big)^{-1} \int_{Q_{\delta\rho} \cap T_{-x}D} (\mathbf{e}_j + \nabla_y v_j^{\delta\rho,\varepsilon}(x, y)) \cdot A_\varepsilon(x + y)(\mathbf{e}_i + \nabla_y v_i^{\delta\rho,\varepsilon}(x, y))\eta_\rho(y)dy, \quad (3.8)$$

where for all $k \in \{1, \dots, k\}$, $v_k^{\delta\rho,\varepsilon}(x, \cdot)$ is the unique weak solution in $H^1_0(Q_{\delta\rho} \cap T_{-x}D)$ to

$$-\nabla \cdot A_\varepsilon(x + y)(\mathbf{e}_k + \nabla_y v_k^{\delta\rho,\varepsilon}(x, y)) = 0 \qquad \text{in } Q_{\delta\rho} \cap T_{-x}D. \quad (3.9)$$

Unless η is a constant function, it is not clear a priori whether $A^\delta_{\rho,\varepsilon}$ is a coercive matrix. This is indeed the case under mild conditions, which are stated in the main result of this section:

Theorem 4. *Let D be smooth. Let A_ε be a H-convergent sequence, and $A^\delta_{\rho,\varepsilon}$ be as in Definition 5. Then there exist $\delta > 1$ small enough and $\beta' \geq \alpha' > 0$ such that for all $\rho > 0$ and $\varepsilon > 0$, $A^\delta_{\rho,\varepsilon} \in \mathcal{M}_{\alpha'\beta'}(D)$. In addition, for all $\rho > 0$, there exists $A^\delta_{\rho,\rm hom} \in \mathcal{M}_{\alpha'\beta'}(D)$ such that for almost every $x \in D$,*

$$\lim_{\varepsilon\to 0} A^\delta_{\rho,\varepsilon}(x) = A^\delta_{\rho,\rm hom}(x), \quad (3.10)$$

$$\lim_{\rho\to 0} A^\delta_{\rho,\rm hom}(x) = A_{\rm hom}(x). \quad (3.11)$$

As a direct corollary we have

Corollary 2. *Let A_ε, $A^\delta_{\rho,\varepsilon}$ and $A^\delta_{\rho,\rm hom}$ be as in Theorem 4, and $f \in H^{-1}(D)$. Then, the weak solution $u^\delta_{\rho,\varepsilon} \in H^1_0(D)$ to*

$$-\nabla \cdot A^\delta_{\rho,\varepsilon} \nabla u^\delta_{\rho,\varepsilon} = f$$

satisfies

$$\lim_{\rho\to 0} \lim_{\varepsilon\to 0} \|u^\delta_{\rho,\varepsilon} - u_{\rm hom}\|_{H^1(D)} = 0, \quad (3.12)$$

where $u_{\rm hom} \in H^1_0(D)$ is the weak solution to

$$-\nabla \cdot A_{\rm hom} \nabla u_{\rm hom} = f.$$

As a consequence of H-convergence we also have that

$$\lim_{\rho\to 0} \lim_{\varepsilon\to 0} \|u^\delta_{\rho,\varepsilon} - u_\varepsilon\|_{L^2(D)} = 0,$$

Before we proceed with the proofs, let us give the associated corrector result.

Definition 6. *Let $H > 0$, $I_H \in \mathbb{N}$, and let $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D in disjoint subdomains of diameter of order H. We define a family (M_H) of approximations of the identity on $L^2(D)$ associated with $Q_{H,i}$: for every $w \in L^2(D)$ and $H > 0$,*

$$M_H(w) = \sum_{i=1}^{I_H} \left(\fint_{Q_{H,i}} w \right) 1_{Q_{H,i}}.$$

Let $\delta > 1$ be as in Theorem 4, and for all $i \in [\![1, I_H]\!]$, set

$$Q^\delta_{H,i} := \{x \in D \,|\, d(x, Q_{H,i}) < (\delta - 1)H\}.$$

With the notation of Corollary 2, we define the numerical correctors $\gamma^{\delta,H,i}_{\rho,\varepsilon}$ associated with $u^\delta_{\rho,\varepsilon}$ as the unique weak solution in $H^1_0(Q^\delta_{H,i})$ to

$$-\nabla \cdot A_\varepsilon \left(M_H(\nabla u^\delta_{\rho,\varepsilon}) + \nabla \gamma^{\delta,H,i}_{\rho,\varepsilon} \right) = 0, \tag{3.13}$$

we set

$$\nabla u^{\delta,H,i}_{\rho,\varepsilon} := M_H(\nabla u^\delta_{\rho,\varepsilon})|_{Q_{H,i}} + (\nabla \gamma^{\delta,H,i}_{\rho,\varepsilon})|_{Q_{H,i}}$$

for all $1 \le i \le I_H$, and we define

$$C^{\delta,H}_{\rho,\varepsilon} := \sum_{i=1}^{I_H} \nabla u^{\delta,H,i}_{\rho,\varepsilon} 1_{Q_{H,i}}.$$

We then have the following corrector result:

Theorem 5. *Under the assumptions of Corollary 2, the corrector of Definition 6 satisfies*

$$\lim_{\rho,H\to 0} \lim_{\varepsilon\to 0} \left\| \nabla u_\varepsilon - C^{\delta,H}_{\rho,\varepsilon} \right\|_{L^p(D)} = 0, \tag{3.14}$$

for all exponents p such that

- *$1 \le p \le 2$ if A_ε is a family of symmetric matrices,*
- *$1 \le p < 2$ if A_ε is not a family of symmetric matrices.*

In addition, if the r. h. s. $f \in H^{-1}(D)$ of equation (1.5) belongs to $W^{-1,q}(D)$ for some $q > 2$, then one can take $p = 2$ in (3.14) even if A_ε is not symmetric.

We begin with the proof of Theorem 4, which has the same structure as the proof of Theorem 2.

Proof of Theorem 4. We first prove that $A^\delta_{\rho,\varepsilon} \in \mathcal{M}_{\alpha'\beta'}(D)$ for some $\delta > 1$ small enough. By regularity of the domain D, for all $0 < \gamma < 1$, there exists $\delta > 1$ such that for all $x \in D$,

$$\frac{|Q_\rho \cap T_{-x}D|}{|Q_{\delta\rho} \cap T_{-x}D|} \ge \gamma. \tag{3.15}$$

By definition of η, non-negativity of the integrand, and using that $\eta_\rho \le \rho^{-d}\delta$ and $\inf_Q \eta \ge \rho^{-d}/\delta$ by assumption, one has for all $\xi \in \mathbb{R}^d$ such that $|\xi| = 1$

$$\begin{aligned} \xi \cdot A^\delta_{\rho,\varepsilon}\xi &= \Big(\int_{Q_{\delta\rho}\cap T_{-x}D} \eta_\rho \Big)^{-1} \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla v^{\delta\rho,\varepsilon}_\xi(x,y)) \cdot A_\varepsilon(x+y)(\xi + \nabla v^{\delta\rho,\varepsilon}_\xi(x,y))\eta_\rho(y)dy \\ &\ge \frac{1}{\delta^2 |Q_{\delta\rho} \cap T_{-x}D|} \int_{Q_\rho \cap T_{-x}D} (\xi + \nabla v^{\delta\rho,\varepsilon}_\xi(x,y)) \cdot A_\varepsilon(x+y)(\xi + \nabla v^{\delta\rho,\varepsilon}_\xi(x,y))dy, \end{aligned}$$

where $v_\xi^{\delta\rho,\mathrm{hom}}(x,\cdot)$ is the unique solution in $H_0^1(Q_{\delta\rho}\cap T_{-x}D)$ to

$$-\nabla\cdot A_{\mathrm{hom}}(x+y)(\xi+\nabla_y v_\xi^{\delta\rho,\mathrm{hom}}(x,y))=0.$$

By Meyers' estimate and the regularity of D, there exist $q>2$ and $\sigma>0$ such that for all $x\in D$ and all $\xi\in\mathbb{R}^d$ with $|\xi|=1$, $v_\xi^{\delta\rho,\mathrm{hom}}(x,\cdot)\in W_0^{1,q}(Q_{\delta\rho}\cap T_{-x}D)$ and $\|\xi+\nabla v_\xi^{\delta\rho,\mathrm{hom}}(x,\cdot)\|^p_{L^p(Q_{\delta\rho}\cap T_{-x}D)}\le\sigma|Q_{\delta\rho}\cap T_{-x}D|$. Hence, by Hölder's inequality, the fact that $|\xi|=1$, the argument leading to (2.4) (coercivity of A_ε and Jensen's inequality), and (3.15), the estimate above turns into

$$\begin{aligned}
\xi\cdot A^\delta_{\rho,\varepsilon}\xi &\ge \frac{1}{\delta^2|Q_\rho\cap T_{-x}D|}\int_{Q_{\delta\rho}\cap T_{-x}D}(\xi+\nabla v_\xi^{\delta\rho,\varepsilon}(x,y))\cdot A_\varepsilon(x+y)(\xi+\nabla v_\xi^{\delta\rho,\varepsilon}(x,y))dy\\
&\quad-\frac{1}{\delta^2|Q_{\delta\rho}\cap T_{-x}D|}\int_{(Q_{\delta\rho}\setminus Q_\rho)\cap T_{-x}D}(\xi+\nabla v_\xi^{\delta\rho,\varepsilon}(x,y))\cdot A_\varepsilon(x+y)(\xi+\nabla v_\xi^{\delta\rho,\varepsilon}(x,y))dy\\
&\ge \frac{\alpha}{\delta^2}-\frac{\beta}{\delta^2|Q_{\delta\rho}\cap T_{-x}D|}|(Q_{\delta\rho}\setminus Q_\rho)\cap T_{-x}D|^{(q-2)/q}(\sigma|Q_{\delta\rho}\cap T_{-x}D|)^{2/q}\\
&= \frac{1}{\delta^2}(\alpha-\beta(1-\gamma)^{(q-2)/q}\sigma^{2/q}).
\end{aligned}$$

Since q and σ only depend on α,β and D, there exists $0<\gamma<1$ and therefore some $\delta>1$ such that

$$\xi\cdot A^\delta_{\rho,\varepsilon}\xi \ge \frac{\alpha}{2},$$

as desired. The upper bound is proved as in (2.3).

Let $x\in D$ and $\rho>0$, and consider problem (2.2). By the locality and definition of H-convergence (see property (2) of Lemma 1 and Definition 1), for all $k\in\{1,\dots,d\}$,

$$\begin{aligned}
&v_k^{\delta\rho,\varepsilon}(x,\cdot)\rightharpoonup v_k^{\delta\rho,\mathrm{hom}}(x,\cdot)\quad\text{in } H_0^1(Q_{\delta\rho}\cap T_{-x}D),\\
&A_\varepsilon(x+\cdot)(\mathbf{e}_k+\nabla_y v_k^{\delta\rho,\varepsilon}(x,\cdot))\rightharpoonup A_{\mathrm{hom}}(x+\cdot)(\mathbf{e}_k+\nabla_y v_k^{\delta\rho,\mathrm{hom}}(x,\cdot))\quad\text{in } L^2(Q_{\delta\rho}\cap T_{-x}D,\mathbb{R}^d),
\end{aligned}\tag{3.16}$$

where $v_k^{\delta\rho,\mathrm{hom}}(x,\cdot)$ is the unique solution in $H_0^1(Q_{\delta\rho}\cap T_{-x}D)$ to

$$-\nabla\cdot A_{\mathrm{hom}}(x+y)(\mathbf{e}_k+\nabla_y v_k^{\delta\rho,\mathrm{hom}}(x,y))=0.$$

Hence, setting

$$[A^\delta_{\rho,\mathrm{hom}}(x)]_{ij}:=\Big(\int_{Q_{\delta\rho}\cap T_{-x}D}\eta_\rho\Big)^{-1}\int_{Q_{\delta\rho}\cap T_{-x}D}(\mathbf{e}_j+\nabla_y v_j^{\delta\rho,\mathrm{hom}}(x,y))\cdot A_{\mathrm{hom}}(x+y)(\mathbf{e}_i+\nabla_y v_i^{\delta\rho,\mathrm{hom}}(x,y))\eta_\rho(y)dy,$$

(3.10) follows from (3.16) by the div-curl lemma and Meyers' estimates. Indeed, $\mathbf{e}_j+\nabla_y v_j^{\delta\rho,\varepsilon}(x,y)$ is a gradient, and therefore curl-free, whereas $A_{\mathrm{hom}}(x+y)(\mathbf{e}_i+\nabla_y v_i^{\delta\rho,\varepsilon}(x,y))$ is divergence free by definition, so that their product converges in the sense of distributions to the product of the limits as $\varepsilon\to0$. Yet, by Meyers' estimates, $y\mapsto\nabla_y v_k^{\delta\rho,\varepsilon}(x,y)$ is in $L^q(Q_{\delta\rho}\cap T_{-x}D)$ so that $y\mapsto(\mathbf{e}_j+\nabla_y v_j^{\delta\rho,\varepsilon}(x,y))\cdot A_\varepsilon(x+y)(\mathbf{e}_i+\nabla_y v_i^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)$ is bounded in $L^{q/2}(Q_{\delta\rho}\cap T_{-x}D)$ with $q/2>1$, which upgrades the convergence in the sense of distributions to a weak convergence in $L^{q/2}(Q_{\delta\rho}\cap T_{-x}D)$ (by the Banach-Alaoglu theorem), and yields the desired result (3.10).

To prove (3.11), we appeal to Lemma 3. To this aim, we note that for all ρ small enough, $Q_{\delta\rho}\subset T_{-x}D$, so that

$$[A^\delta_{\rho,\mathrm{hom}}(x)]_{ij}=\int_{Q_\delta}(\mathbf{e}_j+\nabla_y v_j^{\delta\rho,\mathrm{hom}}(x,\rho y))\cdot A_{\mathrm{hom}}(x+\rho y)(\mathbf{e}_i+\nabla_y v_i^{\delta\rho,\mathrm{hom}}(x,\rho y))\eta(y)dy.$$

By the continuity of translations in $L^1(D)$, since $A_{\rm hom} \in L^1(D)$, for all $y \in Q_\delta$ and $B \subset\subset D$ we have

$$\int_B |A_{\rm hom}(x+\rho y) - A_{\rm hom}(x)| dx \overset{\rho\to 0}{\to} 0.$$

Integrating over Q_δ and using Fubini's theorem, one obtains

$$\int_B \left(\int_{Q_\delta} |A_{\rm hom}(x+\rho y) - A_{\rm hom}(x)| dy \right) dx \overset{\rho\to 0}{\to} 0. \tag{3.17}$$

Consequently, for almost every $x \in B$, and almost every $y \in Q_\delta$,

$$A_{\rm hom}(x+\rho y) \overset{\rho\to 0}{\to} A_{\rm hom}(x). \tag{3.18}$$

Let now $x \in B$ be such a point, and let then $w_k^\rho \in H^1_0(Q_\delta)$ be solutions for $k \in \{1,\dots,d\}$ to

$$-\nabla_y \cdot A_{\rm hom}(x+\rho y)\nabla_y w_k^\rho(y) = \nabla_y \cdot A_{\rm hom}(x+\rho y)\mathbf{e}_k.$$

Estimate (3.18) implies that the assumptions of Lemma 3 are satisfied, so that $w_k^\rho \to w_k$ in $H^1(Q_\delta)$, where w_k is the unique weak solution in $H^1_0(Q_\delta)$ to

$$-\nabla_y \cdot A_{\rm hom}(x)\nabla_y w_k(y) = \nabla_y \cdot A_{\rm hom}(x)\mathbf{e}_k = 0. \tag{3.19}$$

Hence, for all $i,j \in \{1,\dots,d\}$

$$\begin{aligned}
[A^\delta_{\rho,\rm hom}(x)]_{ij} &= \int_{Q_\delta} (\mathbf{e}_j + \nabla_y w_j^\rho(y)) \cdot A_{\rm hom}(x+\rho y)(\mathbf{e}_i + \nabla_y w_i^\rho(y))\eta(y)dy \\
&\overset{\rho\to 0}{\to} \int_{Q_\delta} (\mathbf{e}_j + \nabla_y w_j(y)) \cdot A_{\rm hom}(x)(\mathbf{e}_i + \nabla_y w_i(y))\eta(y)dy \\
&= [A_{\rm hom}(x)]_{ij}
\end{aligned}$$

since $w_k \equiv 0$ is the trivial solution to (3.19). This concludes the proof of the theorem. □

The adaptation of the proof of Theorem 3 to cover Theorem 5 is straightforward, and we leave the details to the reader.

To deduce the convergence of the direct approach with windowing and filtering from Theorems 4 and 5, one proceeds as in Subsection 2.3. The proof of convergence of the Petrov-Galerkin version of the dual approach with oversampling is slightly different than in Subsection 2.4. The key observation is that this "multiscale" Petrov-discontinuous Galerkin method can be interpreted as a simple Galerkin method for an approximate homogenized problem, which makes the analysis much more elementary than expected.

Let us recall there is a one-to-one mapping $\mathcal{M}^{\delta,H,\varepsilon,h}_{\rm MsFEM}$ from V_H to $V^\delta_{H,\varepsilon,h}$. The Petrov-discontinuous Galerlin method is as follows: Find $u^\delta_{H,\varepsilon,h} \in V^\delta_{H,\varepsilon,h}$ such that for all $v_H \in V_H$,

$$\int_D \nabla v_H \cdot A_\varepsilon \nabla_H u^\delta_{H,\varepsilon,h} \;=\; \langle f, v_H \rangle_{H^{-1}(D),H^1_0(D)}\,.$$

Recall that $V^\delta_{H,\varepsilon,h} \not\in H^1(D)$, and that for all $v^\delta_{H,\varepsilon,h} = \sum_{i=1}^{J_H} \nu_{H,i}\psi^\delta_{H,\varepsilon,h,i}$, $\nabla_H u^\delta_{H,\varepsilon,h}$ denotes the broken gradient $\sum_{i=1}^{J_H} \nu_{H,i}\nabla\psi^\delta_{H,\varepsilon,h,i}$. We now use the one-to-one mapping $\mathcal{M}^{\delta,H,\varepsilon,h}_{\rm MsFEM}$ from V_H to $V^\delta_{H,\varepsilon,h}$, so that one may write

$$\nabla u^\delta_{H,\varepsilon,h} \;=\; \mathcal{M}^{\delta,H,\varepsilon,h}_{\rm MsFEM}(\mathfrak{u}^\delta_{H,\varepsilon,h}) \;:=\; \nabla \mathfrak{u}^\delta_{H,\varepsilon,h} + \nabla \mathfrak{u}^\delta_{H,\varepsilon,h} \cdot \nabla \Phi^\delta_{H,\varepsilon,h}.$$

for some $\mathfrak{u}^\delta_{H,\varepsilon,h} \in V_H$. This turns the equation for $u^\delta_{H,\varepsilon,h}$ into an equation for $\mathfrak{u}^\delta_{H,\varepsilon,h}$: Find $\mathfrak{u}^\delta_{H,\varepsilon,h} \in V_H$ such that for all $v_H \in V_H$,

$$\int_D \nabla v_H \cdot A^\delta_{H,\varepsilon,h} \nabla \mathfrak{u}^\delta_{H,\varepsilon,h} \;=\; \langle f, v_H \rangle_{H^{-1}(D),H^1_0(D)}\,,$$

where $A^\delta_{H,\varepsilon,h}$ is defined as

$$A^\delta_{H,\varepsilon,h} := \sum_{i=1}^{I_H} \fint_{T^i_H} A_\varepsilon(\mathrm{Id} + \nabla \Phi^\delta_{H,\varepsilon,h}) 1_{T^i_H}\,.$$

For this equation to be well-posed we need $A^\delta_{H,\varepsilon,h}$ to be coercive. Indeed, there exists some $\delta > 1$ small enough such that for all $H, \varepsilon, h > 0$, $A^\delta_{H,\varepsilon,h}$ is uniformly coercive on D. The proof of this property is similar to the proof of the corresponding result in Theorem 4, and relies on Meyers' estimates. We leave the details to the reader.

We shall now show that

$$\lim_{H\to 0} \lim_{\varepsilon \to 0} \lim_{h \to 0} \|\mathfrak{u}^\delta_{H,\varepsilon,h} - u_{\mathrm{hom}}\|_{H^1(D)} \;=\; 0. \tag{3.20}$$

The limit $h \to 0$ is standard, and we let $A^\delta_{H,\varepsilon}$ and $\mathfrak{u}^\delta_{H,\varepsilon}$ denote the limits of $A^\delta_{H,\varepsilon,h}$ and $\mathfrak{u}^\delta_{H,\varepsilon,h}$. Furthermore we denote by $u_{H,\mathrm{hom}}$ the weak solution in V_H to

$$-\nabla \cdot A_{\mathrm{hom}} \nabla u_{H,\mathrm{hom}} \;=\; f.$$

Then, by the triangle inequality

$$\|\nabla \mathfrak{u}^\delta_{H,\varepsilon} - \nabla u_{\mathrm{hom}}\|_{L^2(D)} \;\le\; \|\nabla \mathfrak{u}^\delta_{H,\varepsilon} - \nabla u_{H,\mathrm{hom}}\|_{L^2(D)} + \|\nabla u_{H,\mathrm{hom}} - \nabla u_{\mathrm{hom}}\|_{L^2(D)}.$$

We then treat the two terms of the r. h. s. separately and start with the first one. The function $\mathfrak{u}^\delta_{H,\varepsilon} - u_{H,\mathrm{hom}}$ is the weak solution in V_H to

$$-\nabla \cdot A^\delta_{H,\varepsilon} \nabla (\mathfrak{u}^\delta_{H,\varepsilon} - u_{H,\mathrm{hom}}) \;=\; -\nabla \cdot (A_{\mathrm{hom}} - A^\delta_{H,\varepsilon}) \nabla u_{H,\mathrm{hom}},$$

so that

$$\begin{aligned}\|\nabla \mathfrak{u}^\delta_{H,\varepsilon} - \nabla u_{H,\mathrm{hom}}\|_{L^2(D)} \;&\lesssim\; \|(A_{\mathrm{hom}} - A^\delta_{H,\varepsilon}) \nabla u_{H,\mathrm{hom}}\|_{L^2(D)} \\ &\le\; \|(A_{\mathrm{hom}} - A^\delta_{H,\varepsilon}) \nabla u_{\mathrm{hom}}\|_{L^2(D)} + 2\beta \|\nabla u_{H,\mathrm{hom}} - \nabla u_{\mathrm{hom}}\|_{L^2(D)}.\end{aligned}$$

The first term of the r. h. s. converges to zero as ε and H go to zero by the dominated convergence theorem and using the fact that the following convergence holds pointwise, as above,

$$\lim_{H\to 0} \lim_{\varepsilon \to 0} A^\delta_{H,\varepsilon} = A_{\mathrm{hom}}.$$

The second term is standard and vanishes due to the convergence of the Galerkin method:

$$\lim_{H\to 0} \|\nabla u_{H,\mathrm{hom}} - \nabla u_{\mathrm{hom}}\|_{L^2(D)} \;=\; 0.$$

We have thus proved (3.20). The rest of the proof is as in Subsection 2.4.

The case of the (discontinuous) Galerkin version of the dual approach can be dealt with the same way, although care has to be taken for the right hand side (the same way as in Subsection 2.4 where some f_H is introduced)

Comments are in order. In this subsection we have shown the convergence of the direct and dual approaches of numerical homogenization when combined with windowing, filtering, and oversampling. A crucial question

in the analysis (and in practice) is under which conditions the approximate homogenized matrices $A^{\delta}_{H,\varepsilon,h}$ are coercive. We've shown this is the case for δ close enough to 1. Other conditions may ensure this as well (such as H small and $A_{\rm hom}$ smooth). Yet this can be an issue for the numerical practice.

It is also worth mentioning that the analysis of convergence of the Petrov-discontinuous Galerkin method does not require any stabilization. This is rather unusual since discontinuous Galerkin methods normally require stabilization to converge to the right solution. The reason for this is that although the method is written in terms of a Petrov-discontinuous Galerkin formulation, it is equivalent (using the one-to-one mapping $\mathcal{M}^{\delta,H,\varepsilon,h}_{\rm MsFEM}$) to a simple conforming Galerkin method on a different equation. The convergence analysis for the latter is then standard, and implies the convergence of the former — this structure is unusual and peculiar to the homogenization problem under consideration.

Here we have not considered other boundary conditions for the numerical correctors than homogeneous Dirichlet boundary conditions. This owes to the fact that windowing aims at reducing the effect of boundary conditions, so that the precise choice of the boundary conditions shouldn't affect the convergence result (indeed, the results hold as well with the boundary conditions of Remark 3).

A last and important remark is the following. In (3.8), the quantity which is averaged is the energy density $(\mathbf{e}_j + \nabla_y v_j^{\delta\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(y)(\mathbf{e}_i + \nabla_y v_i^{\delta\rho,\varepsilon}(x,y))$. Yet other choices than this one are possible, and $\mathbf{e}_j \cdot A_\varepsilon(y)(\mathbf{e}_i + \nabla_y v_i^{\delta\rho,\varepsilon}(x,y))$ would do the job as well. As opposed to the approach without windowing, these two choices give rise to two different approximations, although both converge within the analytical framework (the proof adapt in a straightforward way). Of course, for symmetric diffusion coefficients, it makes "more sense" to use a symmetric formula. For non-symmetric diffusion coefficients however, there is a more subtle way to generalize this formula. Indeed a property which should be preserved at the level of the approximation is the fact that homogenizing the adjoint problem is equivalent to taking the adjoint of the homogenized problem (in other words, we always have $(A_{\rm hom})^T = (A^T)_{\rm hom}$). This property is only satisfied at the level of the approximation formula provided the quantity to be averaged is $(\mathbf{e}_j + \nabla_y \overline{v}_j^{\delta\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(x+y)(\mathbf{e}_i + \nabla_y v_i^{\delta\rho,\varepsilon}(x,y))$, where $\overline{v}_j^{\delta\rho,\varepsilon}$ is the unique weak solution in $H^1_0(Q_{\delta\rho} \cap T_{-x}D)$ to

$$-\nabla \cdot A_\varepsilon^T(x+y)(\mathbf{e}_k + \nabla_y \overline{v}_k^{\delta\rho,\varepsilon}(x,y)) = 0 \qquad \text{in } Q_{\delta\rho} \cap T_{-x}D.$$

This property will be important in the next section.

4. Reduction of the resonance error by zero-order regularization

We propose now a refinement of the method of windowing to reduce more efficiently the resonance error. The method is based on the introduction of a zero-order term in the corrector equation, and the use of a suitable averaging mask. In this section we describe the method on a prototypical problem, and shall combine it with numerical homogenization methods in the following section. We assume that the homogenized matrix $A_{\rm hom}$ is given by the asymptotic formula: for all $i,j \in \{1,\dots,d\}$,

$$\mathbf{e}_j \cdot A_{\rm hom}\mathbf{e}_i \;=\; \lim_{R\to\infty} \fint_{Q_R} (\mathbf{e}_j + \nabla\phi^j) \cdot A(\mathbf{e}_i + \nabla\phi^i), \tag{4.1}$$

where for all $k \in \{1,\dots,d\}$, ϕ^k is the weak solution to the corrector equation on $\mathbb{R}^d$

$$-\nabla \cdot A(\mathbf{e}_k + \nabla\phi^k) \;=\; 0. \tag{4.2}$$

If the matrix A is for instance periodic, quasi-periodic, or more generally stationary ergodic, this corrector equation is well-posed (uniqueness follows from the condition of sublinearity at infinity) and the definition of homogenized coefficients makes sense.

In this form, the numerical homogenization methods presented so far essentially consist in replacing ϕ^k by ϕ^k_R, unique weak solution in $H^1_0(Q_R)$ to (4.2) on Q_R for some large $R > 0$, and in approximating $A_{\rm hom}$ by

$$\mathbf{e}_j \cdot A_{R,L}\mathbf{e}_i \,=\, \int_{Q_R} (\mathbf{e}_j + \nabla\phi^j_R) \cdot A(\mathbf{e}_i + \nabla\phi^i_R)\eta_L \tag{4.3}$$

for some averaging mask of support Q_L of size $R/2 \le L \le R$.

As already mentioned, the error in the periodic case is:

$$|A_{\rm hom} - A_{R,L}| \,\lesssim\, 1/R. \tag{4.4}$$

The objective of this section is to obtain a better convergence rate.

4.1. **Description of the method, and analysis in the periodic case**

The starting point of the method is the following observation: solving (4.2) on a bounded domain Q_R requires to introduce artificial boundary conditions (namely here homogeneous Dirichlet boundary conditions). The associated error then propagates from the boundary ∂Q_R to the inside of Q_R due to the poor decay of the Green's function associated with the operator $-\nabla \cdot A\nabla$ (which is algebraic), so that windowing "alone" is not that efficient. We need to find a way to localize the error to a boundary layer in the neighborhood of ∂Q_R.

This can be achieved by adding an absorbing term in the equation, namely a zero-order term. For all $T > 0$ we consider the "regularized" corrector equation on $\mathbb{R}^d$: for all $k \in \{1, \dots, d\}$,

$$T^{-1}\phi^k_T - \nabla \cdot A(\mathbf{e}_k + \nabla\phi^k_T) \,=\, 0. \tag{4.5}$$

The (unique) weak solution to this equation ϕ^k_T is much easier to approximate than ϕ^k because the Green's function associated with the operator $T^{-1} - \nabla \cdot A\nabla$ decays exponentially fast in terms of distance measured in units of $\sqrt{T}$. In particular, the difference on Q_L between ϕ^k_T and a solution computed on a bounded domain Q_R is essentially of infinite order in terms of $\frac{R-L}{\sqrt{T}}$ (for $R \ge L$). In order to deal with non-symmetric matrices efficiently as well, we introduce the adjoint problem and consider the regularized adjoint corrector equation on $\mathbb{R}^d$:

$$T^{-1}\overline{\phi}^k_T - \nabla \cdot A^T(\mathbf{e}_k + \nabla\overline{\phi}^k_T) \,=\, 0. \tag{4.6}$$

The use of the adjoint problem indeed follows Tartar's seminal ideas, whereas the introduction of the zero order term is somehow inspired by the analysis of the corrector equation in the stochastic case.

The approximation of $A_{\rm hom}$ we shall consider is then given for all $i, j \in \{1, \dots, d\}$ by

$$\mathbf{e}_j \cdot A_{T,R,L}\mathbf{e}_i \,:=\, \int_{Q_R} (\mathbf{e}_j + \nabla\overline{\phi}^j_{T,R}) \cdot A(\mathbf{e}_i + \nabla\phi^i_{T,R})\eta_L, \tag{4.7}$$

where η_L is a suitable averaging mask, and $\phi^k_{T,R}, \overline{\phi}^k_{T,R} \in H^1_0(Q_R)$ are the weak solutions in Q_R to (4.5) and (4.6), respectively.

Before we turn to the analysis of the periodic case, let us make the notion of averaging mask more precise.

Definition 7. *A function $\eta : [-1/2, 1/2] \to \mathbb{R}^+$ is said to be a filter of order $p \ge 0$ if*

(i) $\eta \in C^p([-1/2, 1/2]) \cap W^{p+1,\infty}((-1/2, 1/2))$,

(ii) $\int_{-1/2}^{1/2} \eta(x)dx = 1$,

(iii) $\eta^{(k)}(-1/2) = \eta^{(k)}(1/2) = 0$ *for all* $k \in \{0, \dots, p-1\}$.

The associated mask $\eta_L : [-L/2, L/2]^d \to \mathbb{R}^+$ *in dimension* $d \geq 1$ *is then defined for all* $L > 0$ *by*

$$\eta_L(x) := L^{-d} \prod_{i=1}^{d} \eta(L^{-1} x_i),$$

where $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$.

In the periodic case, we then have the following quantitative convergence result:

Theorem 6. *Let* $d \geq 2$, $A \in \mathcal{M}_{\alpha\beta}$ *be* Q*-periodic,* η *be a filter of order* $p \geq 0$, *and* $A_{\rm hom}$ *and* $A_{T,R,L}$ *be the homogenized matrix and its approximation* (4.7) *respectively, where* $R^2 \gtrsim T \gtrsim R$, $R \geq L \sim R \sim R - L$. *Then, there exists* $c > 0$ *depending only on* α, β *and* d *such that we have*

$$|A_{T,R,L} - A_{\rm hom}| \lesssim L^{-(p+1)} + T^{-2} + T^{1/4} \exp\left(-c \frac{R-L}{\sqrt{T}}\right). \tag{4.8}$$

Remark 4. This result is an extension of [29, Theorem 3.1] which allows to cover the case of non-symmetric diffusion coefficients as well. If in (4.7) we had used the (primal) corrector instead of the corrector associated with the adjoint problem, the term T^{-2} would be replaced by T^{-1} for non-symmetric matrices.

The starting point of the proof is the decomposition of the error into three contributions:

$$|A_{T,R,L} - A_{\rm hom}| \leq |A_{T,R,L} - A_{T,L}| + |A_{T,L} - A_T| + |A_T - A_{\rm hom}|,$$

where

$$\mathbf{e}_j \cdot A_{T,L} \mathbf{e}_i := \int_{\mathbb{R}^d} (\mathbf{e}_j + \nabla \overline{\phi}_T^j) \cdot A(\mathbf{e}_i + \nabla \phi_T^i) \eta_L,$$

and

$$\begin{aligned} \mathbf{e}_j \cdot A_T \mathbf{e}_i &:= \lim_{R \to \infty} \fint_{Q_R} (\mathbf{e}_j + \nabla \overline{\phi}_T^j) \cdot A(\mathbf{e}_i + \nabla \phi_T^i) \\ &= \fint_{Q} (\mathbf{e}_j + \nabla \overline{\phi}_T^j) \cdot A(\mathbf{e}_i + \nabla \phi_T^i), \end{aligned} \tag{4.9}$$

by periodicity of A, ϕ_T^i and $\overline{\phi}_T^j$.

The first contribution measures the error due to the use of boundary conditions to approximate ϕ_T^k and $\overline{\phi}_T^k$ on a bounded domain. As already mentioned, this term is exponentially small due to the decay of the Green's function associated with the Helmholtz operators $T^{-1} - \nabla \cdot A \nabla$ and $T^{-1} - \nabla \cdot A^T \nabla$ on the whole space (and the maximum principle). The second contribution measures the error made in approximating the average of the periodic function $(\mathbf{e}_j + \nabla \overline{\phi}_T^j) \cdot A(\mathbf{e}_i + \nabla \phi_T^i)$ (recall that ϕ_T^k, $\overline{\phi}_T^k$, ϕ^k, and $\overline{\phi}^k$ are periodic) by the average on Q_R with the mask η_L. By Fourier analysis, one can show that this approximation error decays as in $L^{-(p+1)}$, where p is the order of the filter, so that this contribution is not the limiting one for p sufficiently large. Both errors are analyzed in detail in [29].

The last contribution is the so-called systematic error, which is a consequence of the fact that we have modified the corrector equation by a zero order term, and therefore introduced a bias which is controlled by the parameter T. This is the only place which changes in the proof with respect to the symmetric case treated in [29]. We shall make use of the weak forms of the adjoint corrector and corrector equations: for all $v \in H^1_\#(Q)$

$$\int_Q \nabla v \cdot A^T (\mathbf{e}_j + \nabla \overline{\phi}^j) = 0, \tag{4.10}$$

$$\int_Q \nabla v \cdot A (\mathbf{e}_i + \nabla \phi^i) = 0. \tag{4.11}$$

By definition of A_T and $A_{\rm hom}$ and using (4.11) for $v=\phi^j$ and then $v=\overline{\phi}^j$, we obtain

$$\begin{aligned}
\mathbf{e}_j\cdot(A_T-A_{\rm hom})\mathbf{e}_i &= \int_Q(\mathbf{e}_j+\nabla\overline{\phi}^j_T)\cdot A(\mathbf{e}_i+\nabla\phi^i_T)-\int_Q(\mathbf{e}_j+\nabla\phi^j)\cdot A(\mathbf{e}_i+\nabla\phi^i)\\
&= \int_Q(\mathbf{e}_j+\nabla\overline{\phi}^j_T)\cdot A(\mathbf{e}_i+\nabla\phi^i_T)-\int_Q(\mathbf{e}_j+\nabla\overline{\phi}^j)\cdot A(\mathbf{e}_i+\nabla\phi^i)\\
&= \int_Q(\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j)\cdot A(\mathbf{e}_i+\nabla\phi^i_T)+\int_Q(\mathbf{e}_j+\nabla\overline{\phi}^j)\cdot A(\nabla\phi^i_T-\nabla\phi^i)\\
&= \int_Q(\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j)\cdot A(\mathbf{e}_i+\nabla\phi^i_T)+\int_Q(\nabla\phi^i_T-\nabla\phi^i)\cdot A^T(\mathbf{e}_j+\nabla\overline{\phi}^j).
\end{aligned}$$

Using then (4.10) with $v=\phi^i_T-\phi^i$, and (4.11) with $v=\overline{\phi}^j_T-\overline{\phi}^j$, this turns into

$$\begin{aligned}
\mathbf{e}_j\cdot(A_T-A_{\rm hom})\mathbf{e}_i &= \int_Q(\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j)\cdot A(\mathbf{e}_i+\nabla\phi^i_T)-\int_Q(\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j)\cdot A(\mathbf{e}_i+\nabla\phi^i)\\
&= \int_Q(\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j)\cdot A(\nabla\phi^i_T-\nabla\phi^i).
\end{aligned}$$

Hence, since $A\in\mathcal{M}_{\alpha\beta}$, this turns into

$$|A_T-A_{\rm hom}|\lesssim\|\nabla\overline{\phi}^j_T-\nabla\overline{\phi}^j\|_{L^2(D)}\|\nabla\phi^i_T-\nabla\phi^i\|_{L^2(D)},$$

and we need to estimate the convergence of $\nabla\phi^i_T$ to $\nabla\phi^i$ (the second term is similar). Since $\phi^i_T-\phi^i$ is the unique weak solution in $H^1_\#(Q)$ to

$$T^{-1}(\phi^i_T-\phi^i)-\nabla\cdot A\nabla(\phi^i_T-\phi^i)=-T^{-1}\phi^i,$$

an a priori estimate combined with Poincaré's inequality in $H^1_\#(Q)$ allows to conclude that

$$\|\nabla(\phi^i_T-\phi^i)\|_{L^2(Q)}\lesssim T^{-1}.$$

This finaly implies the desired estimate

$$|A_T-A_{\rm hom}|\lesssim T^{-2}. \tag{4.12}$$

Let us give a simple application of this estimate: For $p\geq 3$, the rate in (4.8) is controlled by the last two terms. In particular the last term requires T to be such that $L\gg\sqrt{T}$. A possible choice is then given by

- $T=L^2(\ln L)^{-4}$,
- $R=3L/2$,

for which (4.8) reads:

$$|A_{T,R,L}-A_{\rm hom}|\lesssim R^{-4}\ln^8 R. \tag{4.13}$$

Whereas the convergence rate in estimate (4.4) is of order 1, the regularization approach yields a convergence rate in (4.13) up to order 4^-.

4.2. **Spectral analysis for symmetric coefficients and consistency in the stationary ergodic case**

The aim of this subsection is to show that the regularization method introduced in the previous subsection is consistent at the level of the homogenized coefficients for a large class of heterogeneities, and not only for the periodic case. Unlike the consistency results we have proved so far, one cannot consider here any H-convergent

sequence. We shall indeed prove consistency for the restricted class of stationary ergodic coefficients (which is already quite large, and includes periodic and quasiperiodic coefficients, as well as standard examples of random inclusions etc.).

This is where the second fundamental tool of the homogenization theory pops up: spectral theory. Because of the nature of the tools we shall use (essentially the spectral theorem), the argument we present here only covers the case of *symmetric* diffusion coefficients (for which we then have $\overline{\phi}_T \equiv \phi_T$), as opposed to Theorem 6.

In order for this survey to be as self-contained as possible, we quickly recall standard qualitative results in stochastic homogenization of linear elliptic equations. We refer the reader to the original papers [55] by Papanicolaou and Varadhan, and [47] by Kozlov for details (see also the monography [45]).

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and denote by $\langle\cdot\rangle$ the associated expectation. We shall say that the family of mappings $(\theta_z)_{z\in\mathbb{R}^d}$ from Ω to Ω is a strongly continuous measure-preserving ergodic translation group if:

- $(\theta_z)_{z\in\mathbb{R}^d}$ has the group property: $\theta_0 = \mathrm{Id}$ (the identity mapping), and for all $x, y \in \mathbb{R}^d$, $\theta_{x+y} = \theta_x \circ \theta_y$;
- $(\theta_z)_{z\in\mathbb{R}^d}$ preserves the measure: for all $x \in \mathbb{R}^d$, and every measurable set $F \in \mathcal{F}$, $\theta_x F$ is measurable and $\mathbb{P}(\theta_x F) = \mathbb{P}(F)$;
- $(\theta_z)_{z\in\mathbb{R}^d}$ is strongly continuous: for any measurable function f on Ω, the function $(\omega, x) \mapsto f(\theta_x\omega)$ defined on $\Omega \times \mathbb{R}^d$ is measurable (with the Lebesgue measure on $\mathbb{R}^d$);
- $(\theta_z)_{z\in\mathbb{R}^d}$ is ergodic: for all $F \in \mathcal{F}$, if for all $x \in \mathbb{R}^d$, $\theta_x F \subset F$, then $\mathbb{P}(F) \in \{0, 1\}$.

Let $0 < \alpha \leq \beta < \infty$, and let $A \in L^2(\Omega, \mathcal{M}_{\alpha\beta})$. We define a stationary extension of A (still denoted by A) on $\mathbb{R}^d \times \Omega$ as follows:

$$A : (x, \omega) \mapsto A(x, \omega) \,=\, A(\theta_x\omega).$$

Homogenization theory ensures that the solution operator associated with $-\nabla \cdot A(x/\varepsilon, \omega)\nabla$ converges as $\varepsilon > 0$ vanishes to the solution operator of $-\nabla \cdot A_{\mathrm{hom}}\nabla$ for $\mathbb{P}$-almost every ω, where A_{hom} is a deterministic elliptic matrix characterized as follows. For all $\xi, \zeta \in \mathbb{R}^d$, and $\mathbb{P}$-almost every ω,

$$\begin{aligned} \xi \cdot A_{\mathrm{hom}}\zeta &= \lim_{R\to\infty} \fint_{Q_R} (\xi + \nabla\phi^\xi(x, \omega)) \cdot A(x, \omega)(\zeta + \nabla\phi^\zeta(x, \omega))dx \\ &= \left\langle (\xi + \nabla\phi^\xi) \cdot A(\zeta + \nabla\phi^\zeta) \right\rangle, \end{aligned}$$

where $\phi^\xi : \mathbb{R}^d \times \Omega \to \mathbb{R}$ is Borel measurable, is such that $\phi^\xi(0, \cdot) \equiv 0$, $\nabla\phi^\xi$ is stationary, $\left\langle \nabla\phi^\xi \right\rangle = 0$, and $\phi^\xi(\cdot, \omega) \in H^1_{\mathrm{loc}}(\mathbb{R}^d)$ is almost surely a distributional solution to the corrector equation

$$-\nabla \cdot A(x, \omega)(\xi + \nabla\phi^\xi(x, \omega)) \,=\, 0 \text{ in } \mathbb{R}^d, \tag{4.14}$$

and likewise for ϕ^ζ.

The proof of existence and uniqueness of these correctors is obtained by regularization, and we consider for all $T > 0$ the stationary solution ϕ^ξ_T with zero expectation to the equation

$$T^{-1}\phi^\xi_T(x, \omega) - \nabla \cdot A(x, \omega)(\xi + \nabla\phi^\xi_T(x, \omega)) \,=\, 0 \text{ in } \mathbb{R}^d.$$

This equation has an equivalent form in the probability space, to which we can apply the Lax-Milgram theorem. This formulation requires a bit of functional analysis: the stochastic counterpart of ∇_i (for $i \in \{1, \ldots, d\}$) is denoted by D_i and defined by

$$\mathrm{D}_i f(\omega) \,=\, \lim_{h\to 0} \frac{f(\theta_{h\mathbf{e}_i}\omega) - f(\omega)}{h}.$$

These are the infinitesimal generators of the d one-parameter strongly continuous unitary groups on $L^2(\Omega)$ defined by the translations in each of the d directions. These operators commute and are closed and densely

defined on $L^2(\Omega)$. We denote by $\mathcal{H}(\Omega)$ the domain of $\mathrm{D} = (\mathrm{D}_1, \dots, \mathrm{D}_d)$. This subset of $L^2(\Omega)$ is a Hilbert space for the norm

$$\|f\|_{\mathcal{H}}^2 \,=\, \left\langle |\mathrm{D}f|^2 \right\rangle + \left\langle f^2 \right\rangle .$$

Since the groups are unitary, the operators are skew-adjoint so that we have the "integration by parts" formula: for all $f, g \in \mathcal{H}(\Omega)$

$$\langle f \mathrm{D}_i g \rangle \,=\, - \langle g \mathrm{D}_i f \rangle .$$

The equivalent form of the regularized corrector equation is as follows:

$$T^{-1}\phi_T^\xi - \mathrm{D} \cdot A(\xi + \mathrm{D}\phi_T^\xi) \,=\, 0, \tag{4.15}$$

which admits a unique weak solution in $\phi_T^\xi \in \mathcal{H}(\Omega)$, that is such that for all $\psi \in \mathcal{H}(\Omega)$,

$$\left\langle T^{-1}\phi_T^\xi \psi + \mathrm{D}\psi \cdot A(\xi + \mathrm{D}\phi_T^\xi) \right\rangle \,=\, 0. \tag{4.16}$$

One may prove using the integration by parts formula that $\mathrm{D}\phi_T^\xi$ is bounded in $L^2(\Omega)$ and converges weakly (up to extraction) in $L^2(\Omega)$ to some solution Φ^ξ, which is a gradient. Using then the spectral representation of the translation group we may prove uniqueness of the corrector ϕ^ξ (which is such that $\nabla\phi^\xi = \Phi^\xi$), see [55].

Up to here, we have not required A to be symmetric. Let $\mathcal{M}_{\alpha\beta}^{\mathrm{sym}}$ denote the subset of symmetric matrices of $\mathcal{M}_{\alpha\beta}$, and set

$$\mathcal{A}_{\alpha\beta}^{\mathrm{sym}} = L^\infty(\mathbb{R}^d, \mathcal{M}_{\alpha\beta}^{\mathrm{sym}}). \tag{4.17}$$

In the rest of this section, we shall consider that $A \in L^2(\Omega, \mathcal{M}_{\alpha\beta}^{\mathrm{sym}})$ so that one can appeal to spectral theory. Note that the stationary extension of A belongs to $L^2(\Omega, \mathcal{A}_{\alpha\beta}^{\mathrm{sym}})$, and that A_{hom} is also symmetric.

Let $\mathcal{L} = -\mathrm{D} \cdot A\mathrm{D}$ be the operator defined on $\mathcal{H}(\Omega)$ as a quadratic form. We still denote by $\mathcal{L}$ its Friedrichs extension in $L^2(\Omega)$. This operator is a nonnegative selfadjoint operator. By the spectral theorem it admits a spectral resolution

$$\mathcal{L} \,=\, \int_{\mathbb{R}^+} \lambda G(d\lambda).$$

Recall that this allows one to define a spectral calculus, namely for all suitable function $g : \mathbb{R}^+ \to \mathbb{R}$, one may define the operator $g(\mathcal{L})$ by

$$g(\mathcal{L}) \,:=\, \int_{\mathbb{R}^+} g(\lambda) G(d\lambda),$$

as one would do for symmetric matrices. We further denote by $\mathfrak{d} := -\mathrm{D} \cdot A\xi$ the local drift in direction ξ, and shall consider the projection $e_{\mathfrak{d}} = \langle \mathfrak{d} G \mathfrak{d} \rangle$ of the spectral measure G onto the local drift. Since $\phi = \mathcal{L}^{-1}\mathfrak{d}$ and $\langle \nabla\psi \cdot A\nabla\chi \rangle = \langle \psi \mathcal{L} \chi \rangle$ by integration by parts for all $\psi, \chi \in \mathcal{H}(\Omega)$, one has formally

$$\langle \nabla\phi \cdot A\nabla\phi \rangle \,=\, \left\langle (\mathcal{L}^{-1}\mathfrak{d}) \mathcal{L} (\mathcal{L}^{-1}\mathfrak{d}) \right\rangle .$$

This identity can be proved using the regularized corrector ϕ_T and passing to the limit as $T \to \infty$. We then apply spectral calculus to $\mathcal{L}$ and the function

$$g : \lambda \,\mapsto\, \frac{1}{\lambda},$$

which yields the following spectral identity:

$$\begin{aligned}\xi\cdot A_{\rm hom}\xi &= \langle(\xi+\nabla\phi)\cdot A(\xi+\nabla\phi)\rangle\\ &= \langle\xi\cdot A\xi\rangle+\langle\nabla\phi\cdot A\nabla\phi\rangle-2\langle\nabla\phi\cdot A\xi\rangle\\ &= \langle\xi\cdot A\xi\rangle-\langle\nabla\phi\cdot A\nabla\phi\rangle\\ &= \langle\xi\cdot A\xi\rangle-\langle\mathfrak{d}g(\mathcal{L})\mathfrak{d}\rangle\\ &= \langle\xi\cdot A\xi\rangle-\int_{\mathbb{R}^+}\frac{1}{\lambda}de_{\mathfrak{d}}(\lambda),\end{aligned}\tag{4.18}$$

where we have used the weak form of the corrector equation tested with ϕ itself

$$\langle\nabla\phi\cdot A(\xi+\nabla\phi)\rangle=0,$$

which is a consequence of the analysis in [55]. This formula will be crucial for the analysis of the regularization method:

With the help of this framework, one may prove the consistency of the regularization approach.

Theorem 7. *Let $d\geq 2$, $A\in\mathcal{A}^{\rm sym}_{\alpha\beta}$ be a stationary random field, η be some mask, and $A_{\rm hom}$ and $A_{T,R,L}$ be the homogenized matrix and its approximation* (4.7)*, respectively. Then, almost surely,*

$$\lim_{T\to\infty}\lim_{R\geq L\to\infty}|A_{T,R,L}-A_{\rm hom}|=0.$$

Let $\xi\in\mathbb{R}^d$ with $|\xi|=1$ be fixed. By the triangle inequality we have:

$$|\xi\cdot A_{T,R,L}\xi-\xi\cdot A_{\rm hom}\xi|\leq|\xi\cdot A_{T,R,L}\xi-\xi\cdot A_{T,L}\xi|+|\xi\cdot A_{T,L}\xi-\xi\cdot A_T\xi|+|\xi\cdot A_T\xi-\xi\cdot A_{\rm hom}\xi|,\tag{4.19}$$

where

$$\xi\cdot A_{T,L}\xi = \int_{\mathbb{R}^d}(\xi+\nabla\phi_T^\xi)\cdot A(\xi+\nabla\phi_T^\xi)\eta_L,\tag{4.20}$$

$$\xi\cdot A_T\xi = \Big\langle(\xi+\nabla\phi_T^\xi)\cdot A(\xi+\nabla\phi_T^\xi)\Big\rangle.\tag{4.21}$$

The first term $|\xi\cdot A_{T,R,L}\xi-\xi\cdot A_{T,L}\xi|$ of the r. h. s. of (4.19) is a measure of the error due to boundary conditions and scales like

$$|A_{T,R,L}-A_{T,L}|\lesssim T^{3/4}\exp\Big(-c\frac{R-L}{\sqrt{T}}\Big),\tag{4.22}$$

see [30, Proposition 2.8 and Remark 2.9]. In particular, it vanishes in the limit $(R-L)/\sqrt{T}\to\infty$.

The second term $|\xi\cdot A_{T,L}\xi-\xi\cdot A_T\xi|$ is a measure of the error between an average in physical space and the expectation. By the ergodic theorem, this vanishes almost surely as $L\to\infty$.

The last term $|\xi\cdot A_T\xi-\xi\cdot A_{\rm hom}\xi|$ is the systematic error. We shall use spectral theory to show that it converges to zero as $T\to\infty$. As for (4.18), using the regularized corrector equation (4.16), the spectral theorem indeed yields

$$\begin{aligned}\xi\cdot A_T\xi &= \langle\xi\cdot A\xi\rangle+\langle\nabla\phi_T\cdot A\nabla\phi_T\rangle+2\langle\nabla\phi_T\cdot A\xi\rangle\\ &= \langle\xi\cdot A\xi\rangle+\langle\nabla\phi_T\cdot A\nabla\phi_T\rangle-2\langle\nabla\phi_T\cdot A\nabla\phi_T\rangle-2T^{-1}\langle\phi_T^2\rangle\\ &= \langle\xi\cdot A\xi\rangle-\int_{\mathbb{R}^+}\frac{\lambda}{(T^{-1}+\lambda)^2}de_{\mathfrak{d}}(\lambda)-2T^{-1}\int_{\mathbb{R}^+}\frac{1}{(T^{-1}+\lambda)^2}de_{\mathfrak{d}}(\lambda)\\ &= \langle\xi\cdot A\xi\rangle-\int_{\mathbb{R}^+}\frac{2T^{-1}+\lambda}{(T^{-1}+\lambda)^2}de_{\mathfrak{d}}(\lambda),\end{aligned}$$

so that

$$\xi \cdot (A_T - A_{\rm hom})\xi \,=\, T^{-2} \int_{\mathbb{R}^+} \frac{1}{\lambda (T^{-1}+\lambda)^2} de_{\mathfrak{d}}(\lambda), \tag{4.23}$$

which vanishes as $T \to \infty$ by the dominated convergence theorem (since λ^{-1} is integrable for $de_{\mathfrak{d}}$):

$$\lim_{T\to\infty} |A_T - A_{\rm hom}| \,=\, 0. \tag{4.24}$$

This concludes the proof of Theorem 7.

As a by-product of this analysis we also have consistency of the regularized corrector almost surely:

$$\lim_{T\to\infty} \lim_{R\geq L\to\infty} \fint_{Q_L} |\nabla \phi^\xi_{T,R} - \nabla \phi^\xi|^2 \,=\, 0. \tag{4.25}$$

On the one hand, by ergodicity,

$$\lim_{R\to\infty} \fint_{Q_R} |\nabla \phi_{T,R} - \nabla \phi|^2 = \Big\langle |\nabla \phi^\xi_T - \nabla \phi^\xi|^2 \Big\rangle \lesssim \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\nabla \phi^\xi_T - \nabla \phi^\xi) \Big\rangle .$$

On the other hand, using the symmetry of A (in the third identity) and the weak form of the corrector equation for ϕ^ξ (in the fourth identity), that is for all $\mathrm{D}\psi \in L^2(\Omega)$,

$$\big\langle \mathrm{D}\psi \cdot A(\xi + \mathrm{D}\phi^\xi) \big\rangle \,=\, 0,$$

we obtain

$$\begin{aligned}
\xi \cdot A_T \xi - \xi \cdot A_{\rm hom} \xi &= \Big\langle (\xi + \nabla \phi^\xi_T) \cdot A (\xi + \nabla \phi^\xi_T) \Big\rangle - \big\langle (\xi + \nabla \phi^\xi) \cdot A (\xi + \nabla \phi^\xi) \big\rangle \\
&= \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\xi + \nabla \phi^\xi_T) \Big\rangle + \Big\langle (\xi + \nabla \phi^\xi) \cdot A (\nabla \phi^\xi_T - \nabla \phi^\xi) \Big\rangle \\
&= \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\xi + \nabla \phi^\xi_T) \Big\rangle + \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\xi + \nabla \phi^\xi) \Big\rangle \\
&= \Big\langle (\xi + \nabla \phi^\xi) \cdot A (\nabla \phi^\xi_T - \nabla \phi^\xi) \Big\rangle - \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\xi + \nabla \phi^\xi) \Big\rangle \\
&= \Big\langle (\nabla \phi^\xi_T - \nabla \phi^\xi) \cdot A (\nabla \phi^\xi_T - \nabla \phi^\xi) \Big\rangle .
\end{aligned} \tag{4.26}$$

The combination of this inequality and this identity with (4.24) yields (4.25).

Remark 5. With some additional work we can even consider some "diagonal" extraction in a weaker norm. In particular, in the regime $R^2 \gtrsim T \gtrsim R$, $R \geq L \sim R \sim R - L$, we have

$$\lim_{T,R,L\to\infty} \langle |A_{T,R,L} - A_{\rm hom}| \rangle \,=\, 0.$$

4.3. **Convergence rates in the stochastic case with finite correlation-length**

In the case of stationary random symmetric coefficients with finite correlation-length c_l, that is coefficients A such that for all $x, y \in \mathbb{R}^d$, $A(x)$ and $A(y)$ are independent random fields if $|x - y| \geq c_l$, we may even give quantitative results for the regularization method.

This covers for instance the case of a random checkerboard (say when the colors are independent and identically distributed random variables) or the case of random inclusions whose centers are distributed according to a Poisson point process (or hard-core Poisson point process to avoid overlapping inclusions).

In this case, Otto and the author proved the following convergence rates for all $d = 2, 3, 4$ in [37]:

$$\left\langle |A_{T,R,L} - A_{\rm hom}|^2 \right\rangle \lesssim \left| \begin{array}{rcl} d = 2 & : & L^{-2} \ln^q T, \\ 2 < d \leq 4 & : & L^{-d}, \end{array} \right. \tag{4.27}$$

for $R = 2L, T = L \ln^2 L$, and some $q > 0$ depending only on α, β. In particular, these estimates are expected to be optimal (they have the central limit theorem scaling) up to the logarithmic correction for $d = 2$. They improve the estimates by Bourgeat and Piatnitski in [12] and by E, Ming, and Zhang in [18], which essentially show (for more general statistics on the coefficients however) that there exists some non explicit exponent $\gamma > 0$ such that

$$\left\langle |A_{T,R,L} - A_{\rm hom}|^2 \right\rangle \lesssim L^{-\gamma} \tag{4.28}$$

provided $d > 2$. The core ingredient in their proof is the very insightful contribution by Yurinskiĭ [60]. However, as explained in the introduction of [38], even if all the steps of Yurinskiĭ's work yielded the expected optimal exponents, the method to pass from the results of [60] to (4.28) does not allow to obtain the optimal convergence rate, so that necessarily $\gamma < d$.

The proof of (4.27) is rather long, and go beyond the scope of this survey. Let us also mention that in the case of discrete elliptic equations on $\mathbb{Z}^d$ with random conductivities, the picture is even more complete, and we refer the reader to the series of papers [30, 35, 36, 38, 39].

It is also worth noting that the scaling of (4.27) is better than the scaling of (4.4).

4.4. **Improving the convergence rate by Richardson extrapolation**

In (4.27), the result is only stated up to $d = 4$ because it does not hold for $d > 4$. Let us discard the error due to boundary conditions, which is controlled by (4.22) and can be made decay at any order in L (provided $R - L \gg \sqrt{T}$). We focus instead on the error term $\left\langle |A_{T,L} - A_{\rm hom}|^2 \right\rangle$, which splits into two contributions: the random error ($A_{T,L}$ fluctuates around its expectation $\langle A_{T,L} \rangle$, which by stationarity is nothing but A_T) and a systematic error (the difference between the expectation of $\langle A_{T,L} \rangle = A_T$ and $A_{\rm hom}$). As proved in [37], the random error scales as the central limit theorem in any dimension

$$\left\langle |A_{T,L} - A_T|^2 \right\rangle \lesssim \left| \begin{array}{rcl} d = 2 & : & L^{-2} \ln^q L, \\ d > 2 & : & L^{-d}, \end{array} \right.$$

for some $q > 0$ depending only on α, β, provided $L \leq T$ (which is compatible with the requirement $R - L \gg \sqrt{T}$). The scaling of the systematic error is however different:

$$|A_T - A_{\rm hom}| \lesssim \left| \begin{array}{rcl} d = 2 & : & T^{-1} \ln^q T \\ d = 3 & : & T^{-3/2}, \\ d = 4 & : & T^{-2} \ln T, \\ d > 4 & : & T^{-2}, \end{array} \right. \tag{4.29}$$

so that the choice $T = L \ln^2 L$ (which is convenient to control the error due to boundary conditions via the exponential decay of the Green's function of the Helmholtz equation) only yields the central limit theorem scaling in (4.27) up to $d = 4$.

Recall that for stationary random fields A, we have the spectral formula (4.23) for the systematic error:

$$\xi \cdot (A_T - A_{\rm hom}) \xi \, = \, T^{-2} \int_{\mathbb{R}^+} \frac{1}{\lambda (T^{-1} + \lambda)^2} de_{\mathfrak{d}}(\lambda).$$

Hence, the best one can hope for the convergence of A_T to $A_{\rm hom}$ is indeed T^{-2}, which holds provided $\lambda \mapsto \lambda^{-3}$ is $de_{\mathfrak{d}}$ integrable. What (4.29) implies is that $\lambda \mapsto \lambda^{-3}$ is $de_{\mathfrak{d}}$ integrable for $d > 4$ in the case of stationary

coefficients with finite correlation length. In the periodic case, since there is a spectral gap, there exists $\mu > 0$ such that $e_{\mathfrak{d}}((0,\mu)) = 0$, and $\lambda \mapsto \lambda^{-3}$ is $de_{\mathfrak{d}}$ integrable in any dimension, so that (4.23) implies (4.12) (provided the matrix is symmetric). In particular, using the approximation A_T of $A_{\rm hom}$, one cannot benefit from the fact that $\lambda \mapsto \lambda^{-k}$ may be $de_{\mathfrak{d}}$ integrable for some $k > 3$. In order to benefit from this, one needs to introduce other approximation formulas of $A_{\rm hom}$ than A_T.

Ideally we are looking for approximations $A_{T,k}$ of $A_{\rm hom}$ which are such that

$$|A_{T,k} - A_{\rm hom}| \lesssim T^{-2k} \int_{\mathbb{R}^+} \frac{1}{\lambda(T^{-1}+\lambda)^{2k}} de_{\mathfrak{d}}(\lambda). \tag{4.30}$$

In [34], Mourrat and the author defined such a family of approximations of $A_{\rm hom}$ by induction in terms of their spectral representations. Yet, this approach is only useful if these approximations admit an explicit counterpart in physical space. The formulas introduced in [34] do indeed have the following representation:

$$\xi \cdot A_{T,k}\xi \;=\; \langle (\xi + \nabla\phi_T)\cdot A(\xi+\nabla\phi_T)\rangle + T^{-1}\sum_{i,j=0}^{k-1} \gamma_{ij}\left\langle \phi_{2^{-i}T}\phi_{2^{-j}T}\right\rangle,$$

where the coefficients γ_{ij} are defined by induction. To be concrete, the first three approximations of $A_{\rm hom}$ are given by:

$$\begin{aligned}
\xi \cdot A_{T,1}\xi &= \langle (\xi + \nabla\phi_T)\cdot A(\xi+\nabla\phi_T)\rangle,\\
\xi \cdot A_{T,2}\xi &= \langle (\xi + \nabla\phi_T)\cdot A(\xi+\nabla\phi_T)\rangle - 3T^{-1}\left\langle \phi_T^2\right\rangle - 2T^{-1}\left\langle \phi_{T/2}^2\right\rangle + 5T^{-1}\left\langle \phi_T\phi_{T/2}\right\rangle,\\
\xi \cdot A_{T,3}\xi &= \langle (\xi + \nabla\phi_T)\cdot A(\xi+\nabla\phi_T)\rangle - \frac{55}{9}T^{-1}\left\langle \phi_T^2\right\rangle - 8T^{-1}\left\langle \phi_{T/2}^2\right\rangle - \frac{4}{9}T^{-1}\left\langle \phi_{T/4}^2\right\rangle\\
&\qquad + \frac{41}{3}T^{-1}\left\langle \phi_T\phi_{T/2}\right\rangle - \frac{22}{9}T^{-1}\left\langle \phi_T\phi_{T/4}\right\rangle + \frac{10}{3}T^{-1}\left\langle \phi_{T/2}\phi_{T/4}\right\rangle.
\end{aligned}$$

For general stationary ergodic coefficients, the approximation formulas $A_{T,k}$ are consistent with $A_{\rm hom}$ in the sense that

$$\lim_{T\to\infty} A_{T,k} \;=\; A_{\rm hom},$$

which can be proved using the Lebesgue dominated convergence theorem as for the convergence of A_T in Theorem 7. Note that in order to approximate $A_{T,k}$ in practice one needs to solve the regularized corrector equation for k different zero order terms of magnitude $2^{k-i}T^{-1}$ for $i = 1,\dots k-1$. The associated computable approximations $A_{T,k,R,L}$ of $A_{\rm hom}$ are given by

$$\xi \cdot A_{T,k,R,L}\xi := \int_{Q_R} (\xi+\nabla\phi_{T,R})\cdot A(\xi+\nabla\phi_{T,R})\eta_L + T^{-1}\sum_{i,j=0}^{k-1}\gamma_{ij}\int_{Q_R}\phi_{2^{-i}T,R}\phi_{2^{-j}T,R}\eta_L,$$

where η_L is a suitable averaging mask, and $\phi_{T^{-i},R} \in H^1_0(Q_R)$ are the weak solutions in Q_R to (4.5) with the corresponding magnitude of the zero order term.

In the case of (symmetric) periodic coefficients, one may conclude using the spectral gap and (4.30) that for all $k \geq 1$,

$$|A_{T,k} - A_{\rm hom}| \lesssim T^{-2k}.$$

As proved in [34], (4.8) is then replaced for all $k \geq 1$ by

$$|A_{T,k,R,L} - A_{\rm hom}| \lesssim L^{-(p+1)} + T^{-2k} + T^{1/4}\exp\left(-c2^k\frac{R-L}{\sqrt{T}}\right), \tag{4.31}$$

provided η is a mask of order p. This allows to drastically reduce the resonance error at the level of the homogenized coefficients.

In the case of a discrete elliptic equation on $\mathbb{Z}^d$ with i. i. d. bond conductivities, Neukamm, Otto and the author proved in [36] (see also [34]) that for all $d \geq 2$, there exists $k_d \geq 1$ such that for all $k \geq k_d$ we have

$$\left\langle |A_{T,k,R,L} - A_{\rm hom}|^2 \right\rangle \lesssim L^{-d} \tag{4.32}$$

provided $R = 2L$ and $T = L \ln^2 L$ — which the central limit theorem scaling in any dimension. It is to be expected that the result will hold for continuous equations as well, which would extend the results (4.27) of [37] to dimensions larger than 4.

In this subsection, we have introduced a family of approximation formulas using spectral calculus, which is proved to drastically reduce the resonance error at the level of the homogenized coefficients. Yet, the use of spectral calculus requires the matrix A to be symmetric and it is not clear how to generalize the approximation formulas $A_{T,k}$ to non-symmetric coefficients.

To this aim, we give now another interpretation of these approximation formulas, which will allow us to introduce a second class of approximation formulas which readily generalizes to the non-symmetric case. In terms of their spectral representation, $A_{T,k}$ can be seen as extrapolations of $A_{\rm hom}$. In particular, it may make sense to try to extrapolate directly in physical space, and we consider Richardson extrapolations defined by the following induction rule: $\tilde{A}_{T,1} = A_T$, and for all $k \geq 1$,

$$\tilde{A}_{T,k+1} := \frac{1}{2^k - 1}(2^k \tilde{A}_{2T,k} - \tilde{A}_{T,k}).$$

As shown in [35], in terms of spectral representation, these extrapolation formulas satisfy

$$|\tilde{A}_{T,k} - A_{\rm hom}| \lesssim T^{-(k+1)} \int_{\mathbb{R}^+} \frac{1}{\lambda(T^{-1} + \lambda)^{k+1}} de_{\mathfrak{d}}(\lambda).$$

so that in the case of symmetric periodic matrices, we have using the spectral gap of the elliptic operator:

$$|\tilde{A}_{T,k} - A_{\rm hom}| \lesssim T^{-(k+1)}. \tag{4.33}$$

The extrapolation formulas obtained by Richardson extrapolation are particularly interesting because it also makes sense to use them in the nonsymmetric case (where A_T is defined by the asymptotic formula (4.9) using the corrector of the adjoint problem). As shown in [32], one can bypass the use of spectral theory in the periodic case, and directly prove by PDE arguments that (4.33) holds as well for nonsymmetric matrices. In the stochastic case however, it is not clear how and even whether these convergence rates can be understood without spectral theory.

This idea of using Richardson extrapolation is very fruitful and will be applied to the corrector itself in Subsection 5.4.

4.5. **Numerical tests**

In this subsection, we quickly illustrate the efficiency of the regularization method and the sharpness of the analysis. We start with some periodic examples to check the sharpness of the analysis and then turn to more challenging cases such as some quasi-periodic and stochastic cases.

4.5.1. *Discrete periodic example*

To check the validity of Theorem 6 (and of its generalization (4.31)) in the regime of R large, we have considered the example of a discrete elliptic equation:

$$-\nabla^* \cdot A(\xi + \nabla\phi) = 0 \qquad \text{in } \mathbb{Z}^2, \tag{4.34}$$

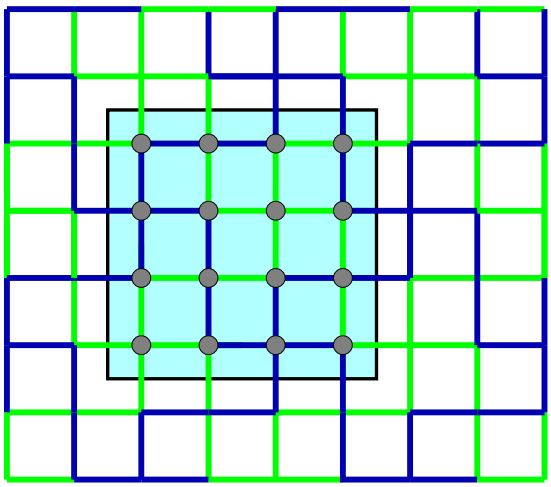

FIGURE 1. Periodic cell in the discrete case

where for all $u : \mathbb{Z}^2 \to \mathbb{R}$,

$$\nabla u(x) := \begin{bmatrix} u(x+\mathbf{e}_1) - u(x) \\ u(x+\mathbf{e}_2) - u(x) \end{bmatrix}, \ \nabla^* u(x) := \begin{bmatrix} u(x) - u(x-\mathbf{e}_1) \\ u(x) - u(x-\mathbf{e}_2) \end{bmatrix},$$

and

$$A(x) := \operatorname{diag}\left[a(x, x+\mathbf{e}_1), a(x, x+\mathbf{e}_2)\right].$$

The matrix A is $[0,4)^2$-periodic, and sketched on a periodic cell on Figure 1. In the example considered, $a(x, x+\mathbf{e}_1)$ and $a(x, x+\mathbf{e}_2)$ represent the conductivities 1 or 100 of the horizontal edge $[x, x+\mathbf{e}_1]$ and the vertical edge $[x, x+\mathbf{e}_2]$ respectively, according to the colors on Figure 1. The homogenization theory for such discrete elliptic operators is similar to the continuous case (see for instance [58] in two dimensions, and [2] in the general case). Since the periodic pattern is invariant by the rotation $\mathcal{R}$ of angle $\pi/2$, the homogenized matrix satisfies $\mathcal{R} A_{\text{hom}} = A_{\text{hom}}$. In dimension $d = 2$, this implies that A_{hom} is a multiple of the identity. It can be evaluated numerically (note that we do not make any other error than the machine precision). Its numerical value is $A_{\text{hom}} = 26.240099009901\ldots$. We have considered the first two approximations formulas $A_{T,1,R,L}$ and $A_{T,2,R,L}$ of A_{hom}. In all the cases treated, we've taken $L = R/3$. For the approximation $A_{T,1,R,L}$, we have tested the following parameters:

- Four values for the zero-order term: $T = \infty$ (no zero-order term), $T \sim R$, $T \sim R^{3/2}$, and $T \sim R^{7/4}$;
- Two different filters: orders $p = 0$ (no filter) and $p = \infty$.

For the approximation $A_{T,2,R,L}$, we have tested the following parameters:

- One value of the zero-order term: $T \sim R^{3/2}$;
- Filter of infinite order $p = \infty$.

The theoretical predictions in terms of convergence rate of $A_{T,k,R,L}$ to A_{hom} in function of R are gathered and compared to the results of numerical tests in Table 3. More details are also given on Figures 2–5, where the overall error

$$\text{Error}(k, T, R) := |A_{\text{hom}} - A_{T,k,R,L}|$$

is plotted in log scale in function of R. Let us quickly comment on the values of T in Figures 2–5. For the four dependences of T upon R, we have chosen the prefactors so that their values roughly coincide for $R = 25$ (that

TABLE 3. Order of convergence: predictions and numerical results.

	$T=\infty$		$T\sim R$		$T\sim R^{3/2}$		$T\sim R^{7/4}$	
k=1	pred.	test	pred.	test	pred.	test	pred.	test
p = 0	1	1	1	1	1	1	1	1
p= ∞	1	1	2	2	3	3.1	3.5	3.4
k=2					pred.	test		
p= ∞					6	5.2		

is for 25 periodic cells per dimension):

$$\begin{aligned} T &= R/25, \\ T &= (4R)^{3/2}/1000, \\ T &= (4R)^{7/4}/5000, \\ T &= (4R)^2/(25\ln^4(4R)). \end{aligned}$$

The numerical result confirm the analysis, and perfectly illustrate the specific influences of the three parameters k, p and T.

4.5.2. *Continuous periodic example*

We consider the following matrix $A:\mathbb{R}^2\to\mathcal{M}^{\rm sym}_{\alpha\beta}$,

$$A(x) = \left(\frac{2+1.8\sin(2\pi x_1)}{2+1.8\cos(2\pi x_2)}+\frac{2+\sin(2\pi x_2)}{2+1.8\cos(2\pi x_1)}\right)\mathrm{Id}, \tag{4.35}$$

used as benchmark tests in [44]. In this case, $\alpha\simeq 0.35$, $\beta\simeq 20.5$, and $A_{\rm hom}\simeq 2.75\,\mathrm{Id}$. We take $L=R/3$, $T=R/10$ and a filter of order 2. The global error $|A_{T,1,R,L}-A_{\rm hom}|$ and the error without zero order term and without filtering are plotted on Figures 6 & 7. Without zero-order term, the convergence rate is R^{-1} as expected, and the use of a filtering method reduces the prefactor but does not change the rate. With the zero-order term and the filtering method, the apparent convergence rate is R^{-3} (note that the asymptotic theoretical rate R^{-2} is not attained yet), which coincides with the convergence rate associated with filters of order 2. This is in agreement with the tests in the discrete case, and confirms the analysis.

4.5.3. *Continuous quasiperiodic example*

We consider the following coefficients:

$$A(x) = \begin{pmatrix} 4+\cos(2\pi(x_1+x_2))+\cos(2\pi\sqrt{2}(x_1+x_2)) & 0 \\ 0 & 6+\sin^2(2\pi x_1)+\sin^2(2\pi\sqrt{2}x_1)\end{pmatrix}. \tag{4.36}$$

In this case, the homogenized coefficients are not easy to compute. They can only be extrapolated. We have taken for the approximation of the homogenized coefficients (that we call coefficient of reference) the output of the computation with $k=1$, $T=R/100$ and $R=52$. Although this may introduce a bias in favour of the proposed strategy, it can be checked a posteriori: the method without zero-order term and without filtering is expected to converge at a rate R^{-1}. This is effectively what we observe on Figure 8 using this coefficient of reference. Instead, if we use as a reference the output of the computation for $R=52$ without zero-order term nor filtering, then we observe a super-linear convergence which is artificial (see Figure 8). With the proposed method, as can be seen on Figure 9, the rate of convergence seems to be much better (the slope of the straight line is -5). The reason for this fact is that there should be a spectral gap in this case as well.

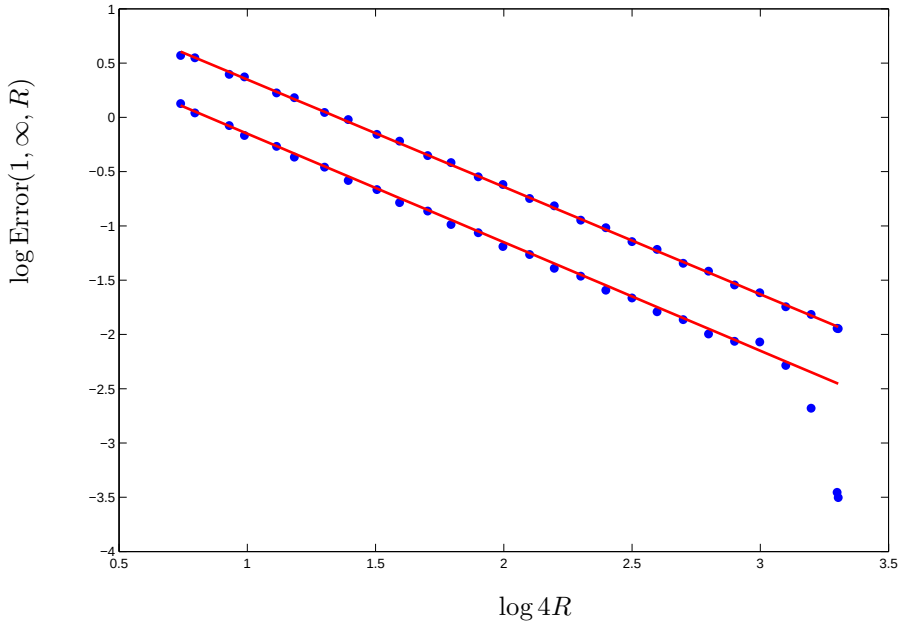

FIGURE 2. Absolute error in log scale without zero order term, no filter (slope -1), infinite order filter (slope -1, better prefactor).

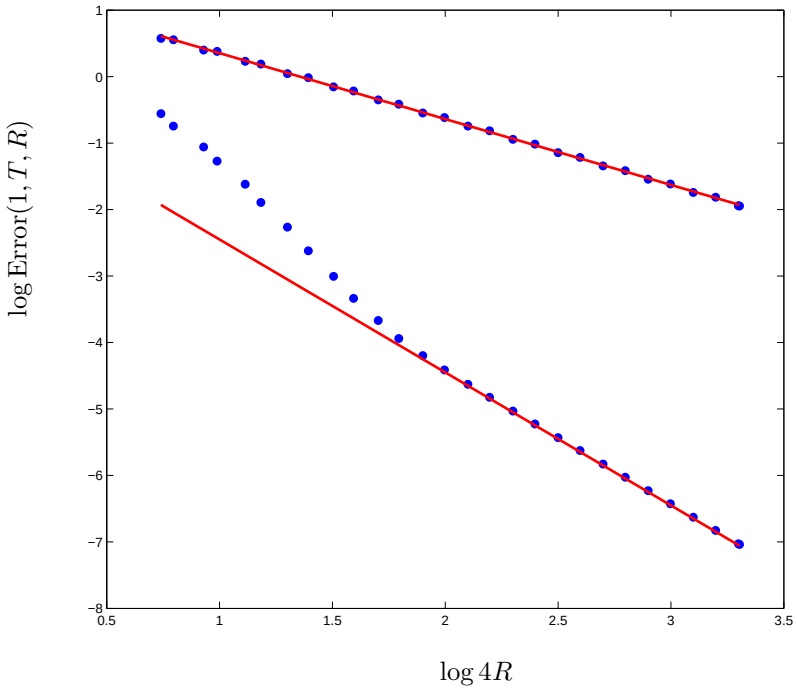

FIGURE 3. Absolute error in log scale for $T = R/25$, no filter (slope -1), infinite order filter (slope -2).

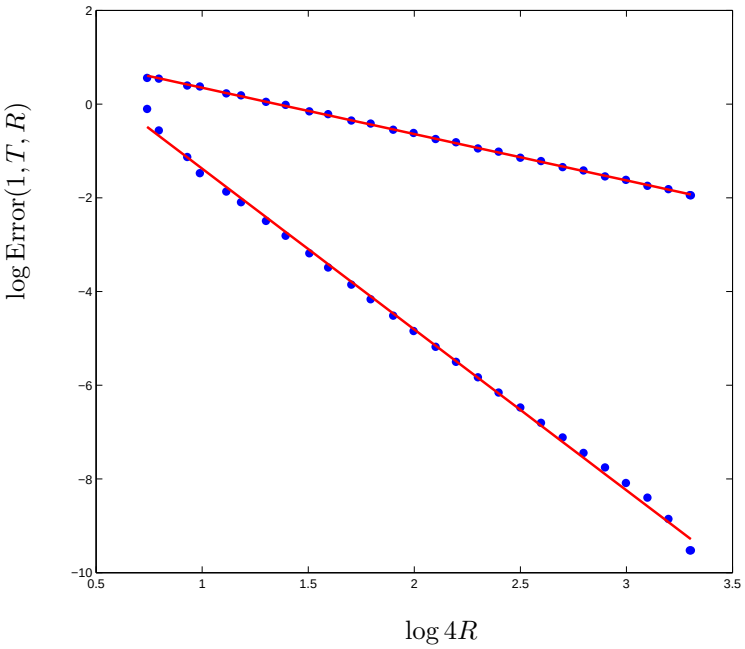

FIGURE 4. Absolute error in log scale for $T = (4R)^{7/4}/5000$, no filter (slope -1), infinite order filter (slope -3.4).

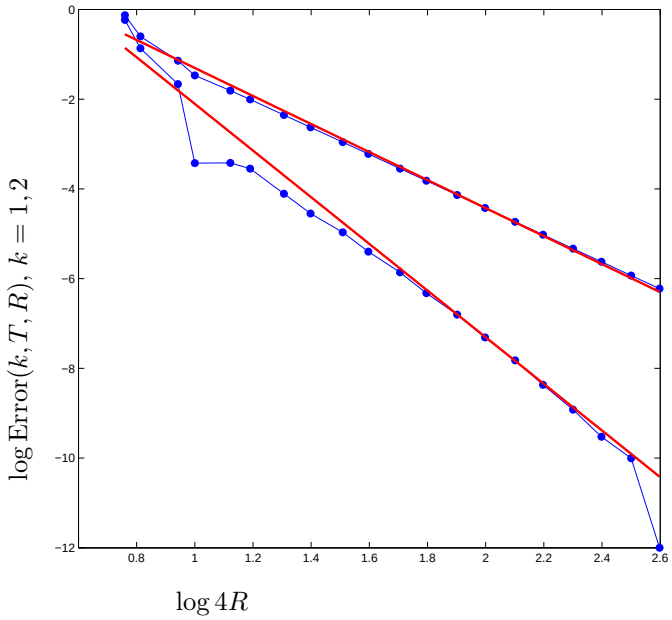

FIGURE 5. Absolute error in log scale for $T = (4R)^{3/2}/1000$, $A_{T,0,R,L}$ (slope -3.1) and $A_{T,1,R,L}$ (slope -5.2), filter of infinite order.

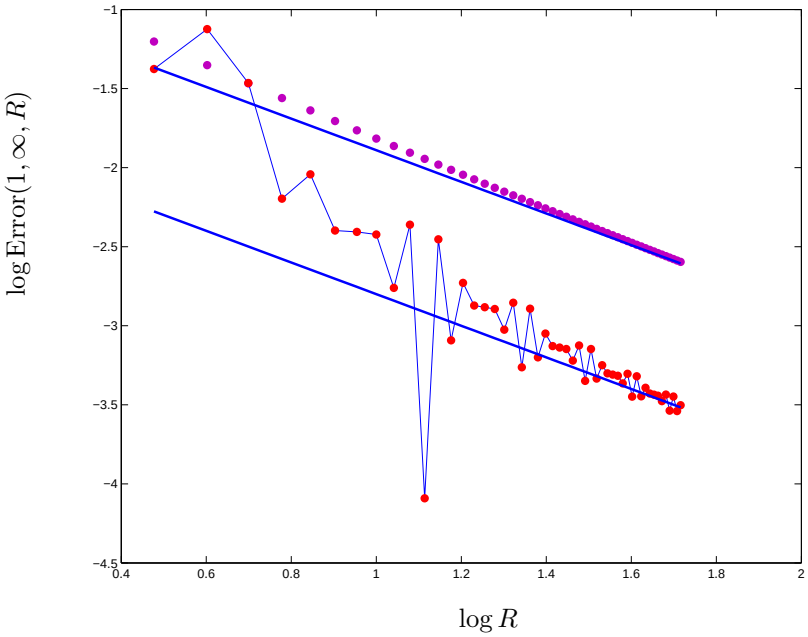

FIGURE 6. Error in log scale for (4.35) in function of the number of cells per dimension $R \in [3, 52]$ without zero-order term, with and without filtering: Slope -1 in both cases.

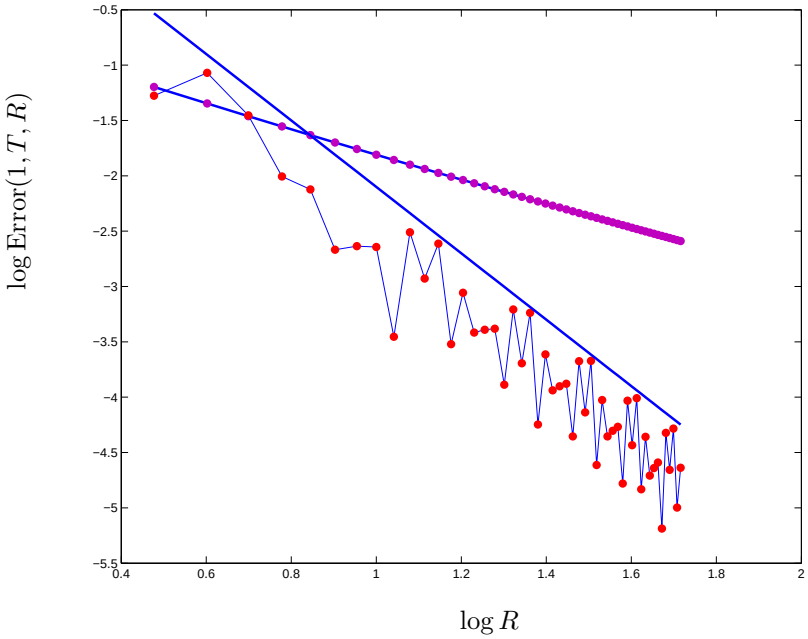

FIGURE 7. Error in log scale for (4.35) in function of the number of cells per dimension $R \in [3, 52]$ with a zero-order term $T = R/10$, with and without filtering: Slopes -1 and -3.

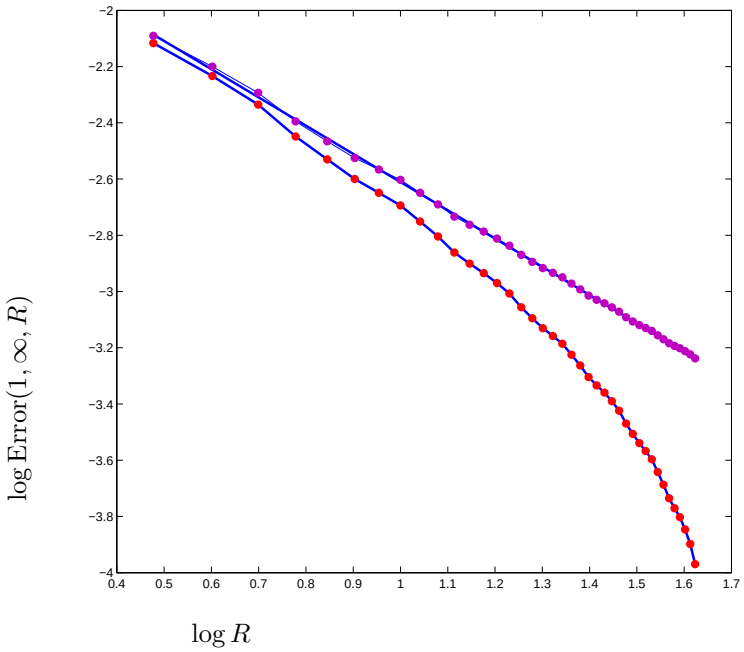

FIGURE 8. Error in log scale for (4.36) in function of the number of cells per dimension $R \in [3, 42]$ without zero-order term and without filtering, for the two different coefficients of reference: Slope -1 and artificial super-linear convergence.

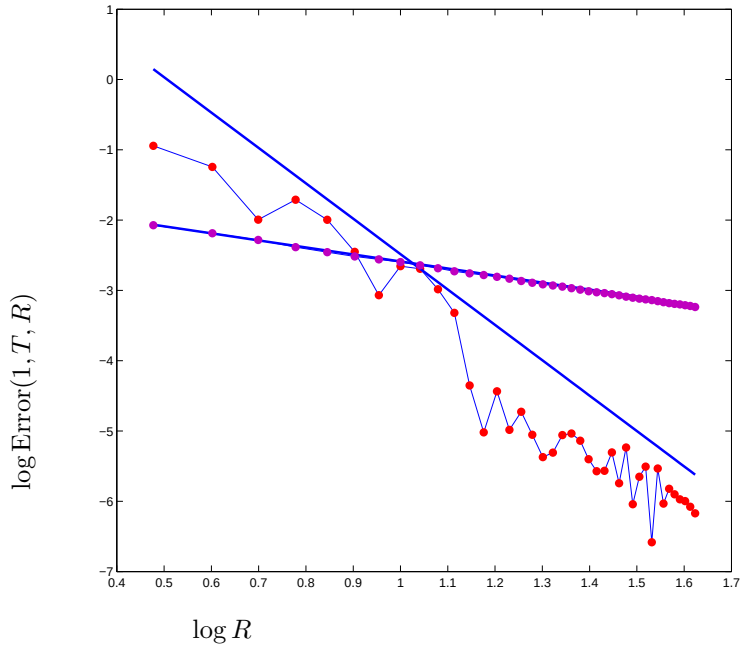

FIGURE 9. Error in log scale for (4.36) in function of the number of cells per dimension $R \in [3, 42]$ with a zero-order term $T = R/100$, with and without filtering: Slopes -1 and -5.

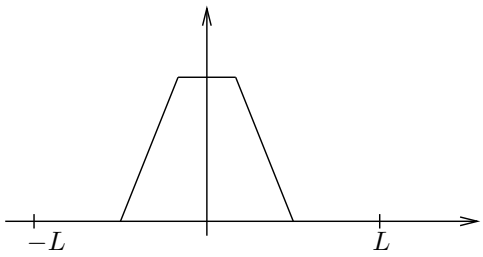

FIGURE 10. Filter for the discrete stochastic example.

4.5.4. *Discrete stochastic example*

To check the validity of (4.27), we consider the discrete case, in the form of equation (4.34). This time we assume the entries a of the symmetric matrix A to be random, independent from one another, and identically distributed (i. i. d.). More precisely, we consider a Bernoulli law (which amounts to topping a coin) with values $\alpha > 0$ et $\beta \geq \alpha$ with probability 1/2. In this case the Dykhne formula holds for $d = 2$ so that the homogenized matrix is explicit: $A_{\text{hom}} = \sqrt{\alpha\beta}\text{Id}$ (see [30, Appendix A]). For the numerical tests, we take $\alpha = 1$, $\beta = 9$ (so that $A_{\text{hom}} = 3\text{Id}$), and

$$\left\{ \begin{array}{l} T \,=\, L + 3\,, \\ R \,=\, \left(1 + 0.1\dfrac{\ln^2(L)}{\sqrt{L}}\right)(L+1)\,. \end{array} \right.$$

The mask is the following:

$$\mu_L(x) = \tilde{\mu}_L(x_1)\tilde{\mu}_L(x_2),$$

where $x = (x_1, x_2) \in \mathbb{Z}^2$, and $\tilde{\mu}_L : \mathbb{Z} \to \mathbb{R}^+$ is given by

$$\tilde{\mu}_L(t) = \gamma_L \left\{ \begin{array}{lll} L/2 \leq & |t| & : \quad 0, \\ L/6 \leq & |t| \leq L/2 & : \quad 3/2(1 - 2|t|/L), \\ & |t| \leq L/6 & : \quad 1, \end{array} \right.$$

and γ_L is such that $\int_{\mathbb{Z}} \tilde{\mu}_L(t)dt = 1$, see Figure 10. For a uniform sampling of $\log L$, $L \in [20, 4000]$, we approximate the expectation

$$\left\langle \left(\int_{\mathbb{Z}^d} (\mathbf{e}_1 + \nabla\phi^1_{T,R}(x)) \cdot A(x)(\mathbf{e}_1 + \nabla\phi^1_{T,R}(x))\mu_L(x)dx - 3 \right)^2 \right\rangle$$

by an empirical average over $r(L)$ realizations, and define the error by

$$\text{Error}(L) := \sqrt{\frac{1}{r(L)} \sum_{j=1}^{r(L)} \left(\int_{\mathbb{Z}^d} (\mathbf{e}_1 + \nabla\phi^{1,j}_{T,R}(x)) \cdot A_j(x)(\mathbf{e}_1 + \nabla\phi^{1,j}_{T,R}(x))\mu_L(x)dx - 3 \right)^2}, \tag{4.37}$$

where $\{\phi^{1,j}_{T,R}\}$ are solutions of the regularized corrector equation on Q_R for $r(L)$ different realizations A_j of the coefficients A. In order to be sure that the error computed using the emprirical average $A^N_{T,R,L}$ of $A_{T,R,L}$ over N independent realization is close to the true error, we have taken $r(L)$ large enough so that the empirical variance of $A^{r(L)}_{T,R,L}$ is much smaller than Error(L).

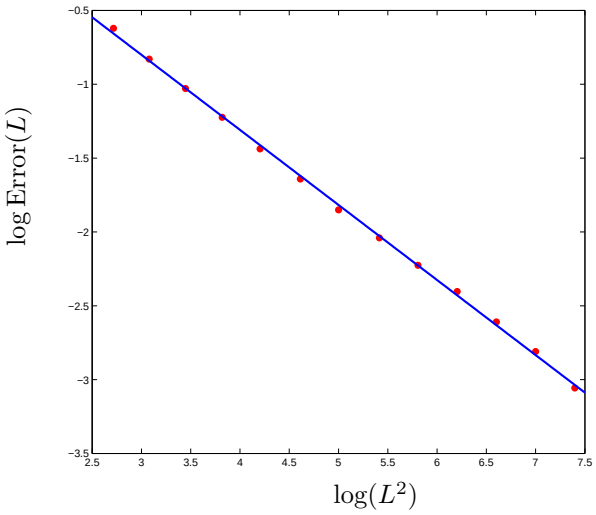

FIGURE 11. Error (4.37) in log scale in function of L^2 (slope $-1/2$).

The error is plotted in function of L^2 in logarithmic scale on Figure 11. The dots (which indicate calculations) are in very good agreement with the straight line of slope $-1/2$ corresponding to the decay provided by (4.32) (the exponent $-1/2$ is the central limit theorem scaling):

$$\left\langle |A_{T,R,L} - A_{\rm hom}|^2 \right\rangle^{1/2} \lesssim (L^2)^{-1/2}.$$

4.6. **Comments on the periodization method**

A very popular method in numerical homogenization is the so-called periodization method, where homogeneous Dirichlet boundary conditions are replaced by periodic boundary conditions. The numerical tests by Yue and E in [59] tend to show that periodic boundary conditions perform usually better than Dirichlet boundary conditions. The method described there reads: for all $L > 0$, an approximation of $A_{\rm hom}$ is given by, for all $ij, \in \{1, \dots, d\}$,

$$\mathbf{e}_j \cdot A_{L,\#} \mathbf{e}_i := \fint_{Q_L} (\mathbf{e}_j + \nabla \phi^j_{L,\#}) \cdot A(\mathbf{e}_i + \nabla \phi^i_{L,\#}),$$

where for all $k \in \{1, \dots, d\}$, $\phi^k_{L,\#}$ is the unique weak *periodic* solution (with zero mean) to

$$-\nabla \cdot A(\mathbf{e}_k + \nabla \phi^k_{L,\#}) = 0 \qquad \text{in } Q_L.$$

The aim of this subsection is to make two remarks on this numerical observation:

- in the case of random coefficients, the periodization method yields optimal results (without the zero order term regularization) provided the periodization is made at the level of the distribution of the random coefficients and not at the level of a realization,
- in the case of periodic (or quasiperiodic) coefficents or random coefficients with correlations, the periodization method can only be performed at the level of the realizations, and does not always reach optimal convergence rate (unlike the zero order regularization method).

Let us make these comments more precise by treating a couple of examples.

In the case of a discrete elliptic equation with i. i. d. coefficients, it is shown in [35] that

$$\left\langle |A_{L,\#} - A_{\rm hom}|^2 \right\rangle \lesssim L^{-d},$$

which is the optimal (central limit theorem) scaling. The analysis of [35] indeed yields more details: the error splits into a random error and a systematic error,

$$\left\langle |A_{L,\#} - A_{\rm hom}|^2 \right\rangle = {\rm var}\left[A_{L,\#}\right] + |\left\langle A_{L,\#} \right\rangle - A_{\rm hom}|^2,$$

whose scalings are for $d \geq 2$:

$$\begin{aligned} {\rm var}\left[A_{L,\#}\right] &\lesssim L^{-d}, \\ |\left\langle A_{L,\#} \right\rangle - A_{\rm hom}|^2 &\lesssim L^{-2d} \ln^d L. \end{aligned}$$

In particular, the crucial point in the proof of the estimate of the systematic error is the compatibility of the periodicity and the randomness — which can be easily coupled due to the product structure of the probability space in the i. i. d. case, see also [35] for more general statistics. It is not clear that in the presence of correlations, the right "periodic approximation" can be constructed, and it is to be expected (although it is not proved) that if this is not done properly the systematic error may scale as $|\left\langle A_{L,\#} \right\rangle - A_{\rm hom}| \sim L^{-1}$ in any dimension (which is the scaling of a boundary effect, as in (4.4)).

The analysis also shows that the scaling of the overall error is the same if the L^d degrees of freedom are distributed over several independent computations. For all $N \in \mathbb{N}$ let us define $A^N_{L,\#}$ as the empirical average of N independent realizations of $A_{L,\#}$, then we have

$$\left\langle |A^N_{L,\#} - A_{\rm hom}|^2 \right\rangle \lesssim N^{-1} L^{-d} + L^{-2d} \ln^d L,$$

so that for the same convergence rate, it may be very advantageous in terms of computational cost to run several realizations on small domains rather than one realization on a large domain (this can indeed be made quantitative using this estimate). This result thus sets on mathematical grounds a method advocated by the mechanical engineering community, see [46].

Let us illustrate how to treat correlations correctly in the continuous case on the example of a homogeneous material perturbed by unit spherical inclusions whose centers are distributed according to a Poisson point process in $\mathbb{R}^d$. The naive "periodic" approach would be to consider a domain Q_L, generate a realization of a Poisson point process in Q_L, construct the associated diffusion matrix by adding spherical inclusions in Q_L centered at the Poisson points, and solve the corrector equation on Q_L with periodic boundary conditions. An intuitive way to see that this periodization is not compatible with the statistics of the random diffusion matrix on $\mathbb{R}^d$ is that it creates new inclusion shapes in the picture (inclusions which intersect the boundary ∂Q_L are simply cut), so that stationarity is broken (the boundary of Q_L is clearly identified and there is no translation invariance any longer). The right way to proceed is to periodize the underlying point process, see Q_L as the torus $\mathbb{T}_L$, simulate a realization of a Poisson point process on $\mathbb{T}_L$ (which is actually the same as on the cube), and then construct a diffusion matrix *not on the cube but on the torus* by adding spherical inclusions centered at the Poisson points. This way, spherical inclusions are not cut, and the statistics of the periodized random diffusion matrix and of the original random diffusion matrix are compatible in the sense that they are both translation invariant. The outputs of these two procedures are sketched on Figure 12. In other terms, there are two operators to be considered: the periodization operator π_L (seen as acting both on point sets of $\mathbb{R}^d$ and on $L^\infty(\mathbb{R}^d, \mathcal{A}_{\alpha\beta})$) and the operator $\mathcal{I}$ which acts on point sets and takes values in $L^\infty(\mathbb{R}^d, \mathcal{A}_{\alpha\beta})$ (it adds inclusions centered at the points of the point set). What Figure 12 illustrate is that given a Poisson process $\mathcal{P}$, one has $\pi_L \circ \mathcal{I}(\mathcal{P}) \neq \mathcal{I} \circ \pi_L(\mathcal{P})$ — the latter being preferable to obtain optimal error estimates.

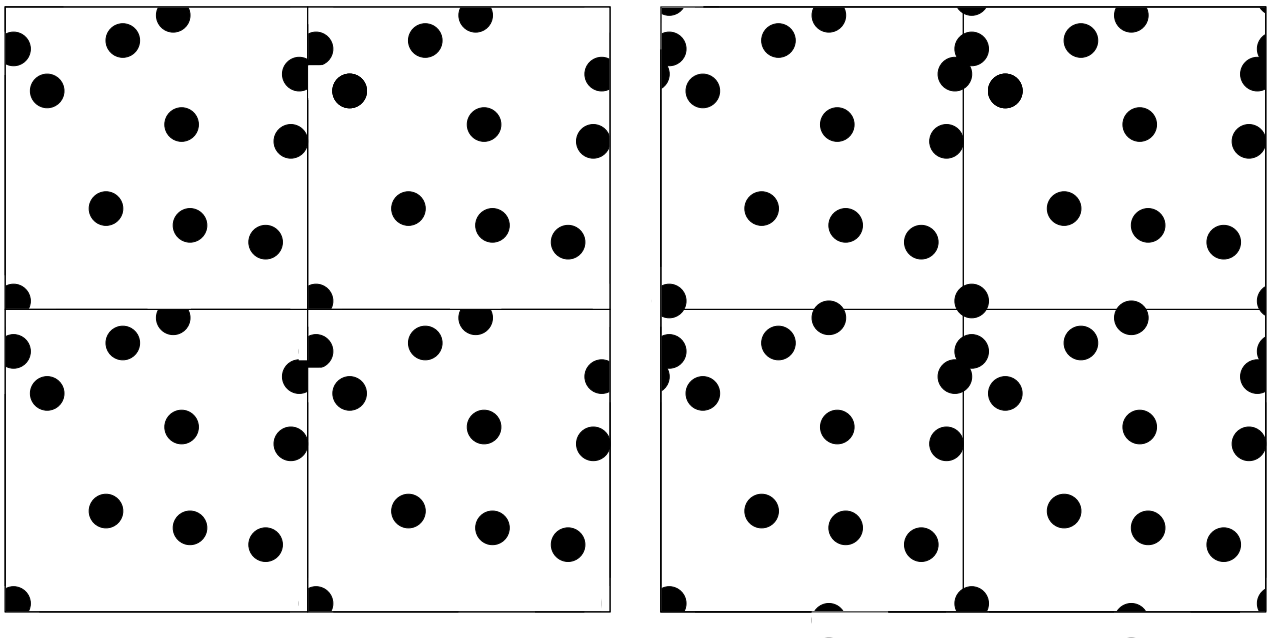

FIGURE 12. Naive periodization (left) versus compatible periodization (right)

The worst case for the naive periodization method is in the periodic case itself. Let A be a Q-periodic matrix, and for all $L \in \mathbb{R}^+$ let $A_{L,\#}$ be defined for all $\xi \in \mathbb{R}^d$ by

$$\xi \cdot A_{L,\#}\xi = \inf\{\fint_{Q_L} (\xi + \nabla\phi) \cdot A(\xi + \nabla\phi), \phi \in H^1_\#(Q_L)\}.$$

Then, if $L \in \mathbb{N}$, $A_{L,\#} = A_{\rm hom}$. Yet we have for all $k \in \mathbb{N}$

$$\sup_{k \le L < k+1} |A_{L,\#} - A_{\rm hom}| \sim 1/k,$$

which shows that generically the convergence rate of $|A_{L,\#} - A_{\rm hom}|$ to zero is not better than in the case when Dirichlet boundary conditions are used on ∂Q_L (yet, numerical tests show that the prefactor is better with periodic boundary conditons).

In the case of a discrete elliptic equation on $\mathbb{Z}^d$ with random conductivities with finite correlation-length, numerical examples in [23] show that the naive periodization indeed yields a systematic error which scales as a surface effect.

5. Numerical homogenization with zero-order regularization

In this section, we show how to combine the regularization method with the direct and dual approaches to numerical homogenization. As in the previous sections we begin with the analytical framework. There is a gap in terms of generality between the results we shall prove here and the results of Sections 2 and 3: our analysis does not apply to general H-converging sequences, and we shall limit ourselves to the case of *symmetric* diffusion matrices which are *locally stationary and ergodic*.

This section contains the main new results of this survey, and a more thorough analysis will be given in [32]. Our main achievement here is the introduction of a technique which is consistent in general, and allows to rid numerical homogenization methods of the resonance error completely in the periodic case.

5.1. **Analytical framework**

Let $A_\varepsilon : D \times \Omega \to \mathcal{M}^{\mathrm{sym}}_{\alpha\beta}$ be a family of random symmetric diffusion coefficients parametrized by $\varepsilon > 0$. We make the assumption that A_ε is locally stationary and ergodic (and assume some "cross-regularity", see below).

Hypothesis 1. There exists a random Carathéodory function (that is continuous in the first variable and measurable in the second variable) $\tilde{A} : D \times \mathbb{R}^d \times \Omega \to \mathcal{M}^{\mathrm{sym}}_{\alpha\beta}$, and a constant $\kappa > 0$ such that

- For all $x \in D$, the random field $\tilde{A}(x, \cdot, \cdot)$ is stationary on $\mathbb{R}^d \times \Omega$, and ergodic;
- *(cross-regularity)* $\tilde{A}$ is κ-Lipschitz in the first variable: for all $x, y \in D$, for almost every $z \in \mathbb{R}^d$, and for almost every realization $\omega \in \Omega$,

$$|\tilde{A}(x, z, \omega) - \tilde{A}(y, z, \omega)| \leq \kappa |x - y|;$$

- For all $x \in D$, for and all $\varepsilon > 0$, and for almost every realization $\omega \in \Omega$,

$$A_\varepsilon(x, \omega) := \tilde{A}(x, \frac{x}{\varepsilon}, \omega)$$

almost surely.

It is not difficult to prove that A_ε H-converges on D to some deterministic diffusion function $A_{\mathrm{hom}} : D \to \mathcal{M}^{\mathrm{sym}}_{\alpha,\beta^2/\alpha}$ almost surely, which is κ-Lipschitz and characterized for all $x \in D$ and $\xi \in \mathbb{R}^d$ by

$$\xi \cdot A_{\mathrm{hom}}(x)\xi = \left\langle (\xi + \nabla\phi(x, 0, \cdot)) \cdot \tilde{A}(x, 0, \cdot)(\xi + \nabla\phi(x, 0, \cdot)) \right\rangle,$$

where $\phi(x, \cdot, \cdot)$ is the corrector in direction ξ associated with the stationary coefficients $\tilde{A}(x, \cdot, \cdot)$ (x is treated as a parameter). The proof is standard: by H-compacteness, for almost every realization, A_ε H-converges up to extraction to some limit A^*. Using the locality of H-convergence and the fact that $\tilde{A}$ is uniformly Lipschitz in the first variable, one concludes that $A^*(x) = A_{\mathrm{hom}}(x)$ almost surely. Note that if we weaken the cross-regularity assumption from Lipschitz continuity on D to continuity on $\overline{D}$, the result holds true as well — although the proof of the homogenization result is much more subtle, see [11, Theorem 4.1.1].

The combination of the regularization approach with the analytical framework of Section 2 yields the following local approximation of A_{hom} in the symmetric case.

Definition 8. *Let $\delta > 1$, and let $\eta : Q_\delta \to [0, \delta]$ be a measurable mask of mass one, such that $\inf_Q \eta \geq 1/\delta$, and for all $\rho > 0$, let $\eta_\rho : Q_{\delta\rho} \to \mathbb{R}^+$ be the rescaled version $\eta_\rho : y \mapsto \rho^{-d}\eta(y/\rho)$ of η. For all $\rho > 0$, $T > 0$ and $\varepsilon > 0$, we denote by $A^\delta_{T,\rho,\varepsilon} : D \to \mathcal{M}_d$ the function defined by: for all $i, j \in \{1, \ldots, d\}$ and for $x \in D$,*

$$\xi \cdot A^\delta_{T,\rho,\varepsilon}(x)\xi := \int_{Q_{\delta\rho} \cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x, y)) \cdot A_\varepsilon(x + y)(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x, y))\eta_\rho(y)dy, \tag{5.1}$$

where $v_T^{\delta\rho,\varepsilon}(x, \cdot)$ is the unique weak solution in $H^1_0(Q_{\delta\rho} \cap T_{-x}D)$ to

$$(T\varepsilon^2)^{-1} v_T^{\delta\rho,\varepsilon}(x, y) - \nabla \cdot A_\varepsilon(x + y)(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x, y)) = 0 \qquad \text{in } Q_{\delta\rho} \cap T_{-x}D. \tag{5.2}$$

For an adaptation of this definition to the non-symmetric case (for which the convergence analysis is not yet clear beyond the periodic case) one may use the solution to the adjoint corrector equation, as in (4.7).

The scaling of the zero-order term in (5.2) is natural, as can be seen on the periodic case: if $A_\varepsilon(y) = A(y/\varepsilon)$ with A periodic, the change of variable $y = \varepsilon z$ turns (5.2) into the equation

$$T^{-1}w(z) - \nabla \cdot A(x+z)(\xi + \nabla w(z)) = 0 \qquad \text{in } Q_{\varepsilon^{-1}\delta\rho} \cap (T_{-x}D)/\varepsilon,$$

which is of the same form as (4.5).

Under Hypothesis 1, we have the following convergence result:

Theorem 8. *Let D be smooth. For all $\varepsilon > 0$, $\rho > 0$, $\delta > 1$, and $T > 0$, let A_ε satisfy Hypothesis 1, and $A^\delta_{T,\rho,\varepsilon}$ be as in Definition 8. Then there exist $\delta > 1$ small enough and $\beta' \geq \alpha' > 0$ such that for all $\varepsilon > 0$, $\rho > 0$ and $T > 0$, $A^\delta_{T,\rho,\varepsilon} \in \mathcal{M}^{\mathrm{sym}}_{\alpha'\beta'}(D)$. In addition, for all $x \in D$,*

$$\lim_{T\to\infty,\rho\to 0} \lim_{\varepsilon\to 0} A^\delta_{T,\rho,\varepsilon}(x) \quad = \quad A_{\mathrm{hom}}(x), \tag{5.3}$$

almost surely. The limit in T in (5.3) is uniform in ρ.

From this theorem, one may directly deduce the convergence of the regularizing method:

Corollary 3. *Under the assumption of Theorem 8, for all $f \in H^{-1}(D)$, the weak solution $u^\delta_{T,\rho,\varepsilon} \in H^1_0(D)$ to*

$$-\nabla \cdot A^\delta_{T,\rho,\varepsilon} \nabla u^\delta_{T,\rho,\varepsilon} = f$$

satisfies

$$\lim_{T\to\infty,\rho\to 0} \lim_{\varepsilon\to 0} \|u^\delta_{\rho,\varepsilon} - u_{\mathrm{hom}}\|_{H^1(D)} = 0 \tag{5.4}$$

almost surely, where $u_{\mathrm{hom}} \in H^1_0(D)$ is the weak solution to

$$-\nabla \cdot A_{\mathrm{hom}} \nabla u_{\mathrm{hom}} = f.$$

As a consequence of H-convergence we also have that

$$\lim_{T\to\infty,\rho\to 0} \lim_{\varepsilon\to 0} \|u^\delta_{\rho,\varepsilon} - u_\varepsilon\|_{L^2(D)} = 0$$

almost surely, where u_ε is the unique weak solution in $H^1_0(D)$ to

$$-\nabla \cdot A_\varepsilon \nabla u_\varepsilon \;=\; f$$

Before we proceed with the proofs, let us give the associated corrector result.

Definition 9. *Let $H > 0$, $I_H \in \mathbb{N}$, and let $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D in disjoint subdomains of diameter of order H. We define a family (M_H) of approximations of the identity on $L^2(D)$ associated with $Q_{H,i}$: for every $w \in L^2(D)$ and $H > 0$,*

$$M_H(w) = \sum_{i=1}^{I_H} \left(\fint_{Q_{H,i}} w \right) 1_{Q_{H,i}}.$$

Let $\delta > 1$ be as in Theorem 8, and for all $i \in [\![1, I_H]\!]$, set

$$Q^\delta_{H,i} := \{x \in D \,|\, d(x, Q_{H,i}) < (\delta - 1)H\}.$$

With the notation of Corollary 3, we define the numerical correctors $\gamma^{\delta,H,i}_{T,\rho,\varepsilon}$ *associated with* $u^{\delta}_{T,\rho,\varepsilon}$ *as the unique weak solution in* $H^1_0(Q^{\delta}_{H,i})$ *to*

$$(T\varepsilon^2)^{-1}\gamma^{\delta,H,i}_{\rho,\varepsilon} - \nabla\cdot A_\varepsilon\Big(M_H(\nabla u^{\delta}_{T,\rho,\varepsilon}) + \nabla\gamma^{\delta,H,i}_{\rho,\varepsilon}\Big) \,=\, 0, \tag{5.5}$$

we set

$$\nabla u^{\delta,H,i}_{T,\rho,\varepsilon} \,:=\, M_H(\nabla u^{\delta}_{T,\rho,\varepsilon})|_{Q_{H,i}} + (\nabla\gamma^{\delta,H,i}_{\rho,\varepsilon})|_{Q_{H,i}} \in L^2(Q_{H,i})$$

for all $1\le i\le I_H$*, and we define*

$$C^{\delta,H}_{T,\rho,\varepsilon} \,=\, \sum_{i=1}^{I_H} \nabla u^{\delta,H,i}_{T,\rho,\varepsilon} 1_{Q_{H,i}}.$$

We then have the following corrector result:

Theorem 9. *Under the assumptions of Corollary 3, the corrector of Definition 9 satisfies*

$$\lim_{T\to\infty,\rho,H\to 0}\lim_{\varepsilon\to 0}\Big\|\nabla u_\varepsilon - C^{\delta,H}_{T,\rho,\varepsilon}\Big\|_{L^2(D)} = 0 \tag{5.6}$$

almost surely.

We only prove Theorem 8, and quickly show how to adapt the proof of Theorem 3 to deal with the regularization.

Proof of Theorem 8. The proof that $A^{\delta}_{T,\rho,\varepsilon}\in\mathcal{M}_{\alpha'\beta'}(D)$ for some $\delta>1$ small enough is the same as for Theorem 2 since the argument only relies on Meyers' estimates, which hold uniformly with respect to the zero-order term. We split the rest of the proof into three steps.

Step 1. From locally ergodic to ergodic.

In this step we introduce a proxy $\tilde{A}^{\delta}_{T,\rho,\varepsilon}(x)$ for $A^{\delta}_{T,\rho,\varepsilon}(x)$ which is uniformly close to $A^{\delta}_{T,\rho,\varepsilon}(x)$ in ρ, and for which we can apply our analysis of Section 4. For all $x\in D$, we define $\tilde{A}^{\delta}_{T,\rho,\varepsilon}(x)$ by: For all $\xi\in\mathbb{R}^d$,

$$\xi\cdot\tilde{A}^{\delta}_{T,\rho,\varepsilon}(x)\xi := \int_{Q_{\delta\rho}\cap T_{-x}D}(\xi+\nabla_y\tilde{v}^{\delta\rho,\varepsilon}_T(x,y))\cdot\tilde{A}(x,\frac{x+y}{\varepsilon})(\xi+\nabla_y\tilde{v}^{\delta\rho,\varepsilon}_T(x,y))\eta_\rho dy,$$

where $\tilde{v}^{\delta\rho,\varepsilon}_T(x,\cdot)$ is the unique weak solution in $H^1_0(Q_{\delta\rho}\cap T_{-x}D)$ to

$$(T\varepsilon^2)^{-1}\tilde{v}^{\delta\rho,\varepsilon}_T(x,y) - \nabla\cdot\tilde{A}(x,\frac{x+y}{\varepsilon})(\xi+\nabla_y\tilde{v}^{\delta\rho,\varepsilon}_T(x,y)) = 0 \qquad \text{in } Q_{\delta\rho}\cap T_{-x}D.$$

We shall prove that

$$|A^{\delta}_{T,\rho,\varepsilon}(x) - \tilde{A}^{\delta}_{T,\rho,\varepsilon}(x)| \,\lesssim\, \rho \tag{5.7}$$

uniformly in x, T, ε, and the randomness.

We first write the difference as: for all $\xi \in \mathbb{R}^d$ with $|\xi| = 1$,

$$\begin{aligned}
&\xi \cdot (A^\delta_{T,\rho,\varepsilon}(x) - \tilde{A}^\delta_{T,\rho,\varepsilon}(x))\xi \\
&= \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)) \cdot A_\varepsilon(x+y)(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\quad - \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&= \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\quad - \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\quad + \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)) \cdot (A_\varepsilon(x+y) - \tilde{A}(x, \frac{x+y}{\varepsilon}))(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy.
\end{aligned}$$

Using that

$$|A_\varepsilon(x+y) - \tilde{A}(x, \frac{x+y}{\varepsilon})| = |\tilde{A}(x+y, \frac{x+y}{\varepsilon}) - \tilde{A}(x, \frac{x+y}{\varepsilon})| \le \kappa|y| \tag{5.8}$$

the third term of the r. h. s. is easily estimated:

$$\begin{aligned}
&\int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)) \cdot (A_\varepsilon(x+y) - \tilde{A}(x, \frac{x+y}{\varepsilon}))(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\qquad\qquad \le \kappa\rho \int_{Q_{\delta\rho}\cap T_{-x}D} |\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)|^2 \eta_\rho(y)dy \lesssim \rho
\end{aligned} \tag{5.9}$$

by definition of η_ρ and an elementary a priori estimate on $v_T^{\delta\rho,\varepsilon}$ (using that $|\xi| = 1$). We now turn to the first two terms, which we write in the form

$$\begin{aligned}
&\int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\quad - \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&= \int_{Q_{\delta\rho}\cap T_{-x}D} (\nabla_y v_T^{\delta\rho,\varepsilon}(x,y) - \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\xi + \nabla_y v_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy \\
&\quad + \int_{Q_{\delta\rho}\cap T_{-x}D} (\xi + \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \cdot \tilde{A}(x, \frac{x+y}{\varepsilon})(\nabla_y v_T^{\delta\rho,\varepsilon}(x,y) - \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y))\eta_\rho(y)dy
\end{aligned}$$

so that using the same a priori estimate as above, it enough to prove that

$$\int_{Q_{\delta\rho}\cap T_{-x}D} |\nabla_y v_T^{\delta\rho,\varepsilon}(x,y) - \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)|^2 \eta_\rho(y)dy \lesssim \rho^2 \tag{5.10}$$

in order to deduce (5.7) from (5.9). This estimate directly follows from the Lipschitz bound (5.8) and an a priori estimate on the equation satisfied by $v_T^{\delta\rho,\varepsilon}(x,\cdot) - \tilde{v}_T^{\delta\rho,\varepsilon}(x,\cdot)$:

$$\begin{aligned}
&(T\varepsilon^2)^{-1}(v_T^{\delta\rho,\varepsilon}(x,y) - \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) - \nabla_y \tilde{A}(x, \frac{x+y}{\varepsilon})\nabla_y(v_T^{\delta\rho,\varepsilon}(x,y) - \nabla_y \tilde{v}_T^{\delta\rho,\varepsilon}(x,y)) \\
&\qquad\qquad = \nabla_y \cdot (A_\varepsilon(x+y) - \tilde{A}(x, \frac{x+y}{\varepsilon}))\nabla_y v_T^{\delta\rho,\varepsilon}(x,y).
\end{aligned}$$

Step 2. Limit as $\varepsilon \to 0$.

The change of variables $z = y/\varepsilon$ allows us to interprete $\tilde{A}^{\delta}_{T,\rho,\varepsilon}(x)$ as the approximation "$A_{T,R,L}(x)$" of $A_{\rm hom}(x)$ defined in (4.7) (with $\tilde{A}(x,\cdot)$ in place of $A(\cdot)$), with the parameters

$$R = \delta\frac{\rho}{\varepsilon}, \qquad L = \frac{\rho}{\varepsilon}.$$

Following Subsection 4.2, we define $A_{T,L}(x)$ and $A_T(x)$ by (4.20) and (4.21) (with $\tilde{A}(x,\cdot)$ in place of $A(\cdot)$). By the triangle inequality,

$$|A_{T,R,L}(x) - A_T(x)| \le |A_{T,R,L}(x) - A_{T,L}(x)| + |A_{T,L}(x) - A_T(x)|.$$

By ergodicity,

$$\lim_{L\to\infty} |A_{T,L}(x) - A_T(x)| = 0$$

almost surely, so that the almost-sure convergence

$$\lim_{\varepsilon\to 0} \tilde{A}^{\delta}_{T,\rho,\varepsilon}(x) = \lim_{L,R-L\to\infty} A_{T,R,L}(x) = A_T(x) \tag{5.11}$$

follows from (4.22) for all $x \in D$. Note that the limit does not depend on ρ, as expected by stationarity of $\tilde{A}(x,y)$ in y.

Step 3. Limit in T and conclusion.

We finally use spectral analysis in the form of (4.24), which yields for all $x \in D$,

$$\lim_{T\to\infty} A_T(x) = A_{\rm hom}(x).$$

We are in position to conclude: the combination of (5.7), (5.11), and (4.24) yields for all $x \in D$

$$\limsup_{T\to\infty,\rho\to 0} \limsup_{\varepsilon\to 0} |A^{\delta}_{T,\rho,\varepsilon}(x) - A_{\rm hom}(x)| = 0,$$

as desired. □

The proof of Theorem 9 closely follows the proof of Theorem 3. There is yet one slight difference since the numerical corrector of Theorem 9 is constructed using the zero-order regularization of the corrector equation. This additionnal difficulty can be taken care of by using (5.10) and (4.25). We leave the details to the reader.

Remark 6. In this section we have only considered the approximation $A_{T,k}$ of $A_{\rm hom}$ for $k = 1$. We could of course use approximations of higher order $k > 1$. All the convergence results proved above hold true for higher order approximations as well.

5.2. **Direct and dual approaches**

The direct approach associated with the analytical framework is a straightforward modification of the direct approach with oversampling, where the corrector equation is regularized by the zero-order term as in (5.2). Likewise, the Petrov-discontinuous Galerkin version of the dual approach of Subsection 2 can be adapted in a straightforward way by adding a zero-order term, as in (5.2). The extensions of Theorems 8 and 9 to cover the direct and dual approaches can be done as in Section 2, and we leave the details to reader as well.

The more interesting question regards the practical use of the method, and therefore the choice of the strength of the zero-order term. The coefficient in front of the zero-order term in (5.2) is of the form $(T\varepsilon^2)^{-1}$, and one needs some "automatic" rule to choose T and ε. The quantity ε represents the local correlation length of A_ε (it is the period of A_ε in the periodic case, it is the correlation length in the stochastic case with finite correlation

length, and in more general cases when there is a scale separation in A_ε this length can typically be identified using a wavelet transform). The choice of T is easier since it only depends on the ellipticity ratio β/α (recall that $A_\varepsilon \in \mathcal{M}^{\rm sym}_{\alpha\beta}$) which drives the value of the coefficient c in the argument of the exponential decay in (4.22). Hence the same value of T can be used for all the coefficients A_ε of some class $\mathcal{M}^{\rm sym}_{\alpha\beta}$. This choice of ε and T in actual computations may be a subtle issue.

To complete the convergence analysis, we provide in the following subsection the numerical analysis of the locally periodic case.

5.3. **Numerical analysis of the locally periodic case**

In this subsection we consider the case of a diffusion function $\tilde{A} : D \times \mathbb{R}^d \to \mathcal{M}^{\rm sym}_{\alpha\beta}$ such that for all $x \in D$, $\tilde{A}(x,\cdot)$ is measurable and Q-periodic, and such that $\tilde{A}$ is uniformly Lipschitz in the first variable. Defined for all $\varepsilon > 0$ and $x \in D$ by $A_\varepsilon(x) := \tilde{A}(x, x/\varepsilon)$, A_ε satisfies Hypothesis 1, so that the regularized versions of the direct and dual approaches converge.

This qualitative convergence result can be turned quantitative, and the combination of the analysis leading to Theorem 6 with [1,17] and [44] yields for the direct approach (with obvious notation):

$$\|u^{\delta,H,h}_{T,H,\varepsilon} - u_{\rm hom}\|_{H^1(D)} \lesssim H + \left(\frac{\varepsilon}{H}\right)^{4^-} + \left(\frac{h}{\varepsilon}\right)^2, \tag{5.12}$$

$$\|\nabla u_\varepsilon - C^{\delta,H,h}_{T,H,\varepsilon}\|_{L^2(D)} \lesssim H + \left(\frac{\varepsilon}{H}\right)^{2^-} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}, \tag{5.13}$$

and for the dual approach (with obvious notation)

$$\|u^{\delta}_{T,H,\varepsilon,h} - u_\varepsilon\|_{H^1(D)} \lesssim H + \left(\frac{\varepsilon}{H}\right)^{2^-} + \frac{h}{\varepsilon} + \sqrt{\varepsilon}, \tag{5.14}$$

where 4^- (resp. 2^-) stands for any exponent strictly less than 4 (resp. 2). These scalings are obtained by taking

$$T := \left(\frac{H}{\varepsilon}\right)^2 \ln^{-2}\frac{H}{\varepsilon}$$

in the estimates.

These rates are to be compared to (3.4), (3.5), and (3.7). In the three cases, the regularization method yields better convergence rates. The effect of the regularization on the corrector estimates (5.13) and (5.14) are much less impressive than for (5.12). One can do better.

5.4. **Richardson extrapolation for the numerical corrector**

In this subsection we show how to take advantage of the Richardson extrapolation technique to reduce the resonance error on the corrector itself. Let us go back to the spectral representation (4.18) of the homogenized coefficients in the stationary ergodic case. We have seen in (4.26) that

$$\xi \cdot (A_T - A_{\rm hom})\xi = \left\langle (\nabla\phi^\xi_T - \nabla\phi_T) \cdot A(\nabla\phi^\xi_T - \nabla\phi^\xi) \right\rangle,$$

where ϕ^ξ and ϕ^ξ_T are the corrector and regularized corrector in direction ξ. In Subsection 4.4, we have also seen that the approximation A_T of $A_{\rm hom}$ could be made more accurate by using extrapolation. The same strategy holds for the corrector as well.

Let A be a symmetric stationary random field, and let $\xi \in \mathbb{R}^d$ with $|\xi| = 1$ be fixed. We present the first step of the extrapolation, and we define $\phi_{2,T}$ as:

$$\phi_{2,T} := 2\phi_{2T} - \phi_T,$$

where ϕ_{2T} and ϕ_T are the unique stationary solutions with vanishing expectation to

$$\tau^{-1}\phi_\tau - \nabla \cdot A(\xi + \nabla\phi_\tau) = 0,$$

for $\tau = 2T$ and $\tau = T$, respectively. Recall that $e_{\mathfrak{d}}$ is the projection of the spectral measure of $\mathcal{L} = -\mathrm{D}\cdot A\mathrm{D}$ on the local drift $\mathfrak{d} = -\mathrm{D}\cdot A\xi$. The spectral representation of the operators $(T^{-1}+\mathcal{L})^{-1}$, $((2T)^{-1}+\mathcal{L})^{-1}$, $\mathcal{L}^{-1}$ and $\mathcal{L}$ being $\lambda \mapsto (T^{-1}+\lambda)^{-1}$, $\lambda \mapsto ((2T)^{-1}+\lambda)^{-1}$, $\lambda \mapsto \lambda^{-1}$, and $\lambda \mapsto \lambda$, we have

$$\begin{aligned}\langle \nabla(\phi - \phi_{2,T}) \cdot A\nabla(\phi - \phi_{2,T})\rangle &= \int_0^\infty \lambda\left(\frac{1}{\lambda} - 2\frac{1}{T^{-1}/2+\lambda} + \frac{1}{T^{-1}+\lambda}\right)^2 de_{\mathfrak{d}}(\lambda)\\ &= \frac{T^{-4}}{4}\int_0^\infty \frac{1}{\lambda(T^{-1}+\lambda)^2(T^{-1}/2+\lambda)^2} de_{\mathfrak{d}}(\lambda).\end{aligned}$$

This implies the consistency of the method for general stationary ergodic random fields A. Indeed, letting $T \to \infty$ and using the Lebesgue dominated convergence theorem, as for (4.24), we have

$$\lim_{T\to\infty} \left\langle |\nabla(\phi - \phi_{2,T})|^2 \right\rangle = 0.$$

In addition, provided $\lambda \mapsto \lambda^{-5}$ is $de_{\mathfrak{d}}$ integrable,

$$\|\nabla(\phi - \phi_{2,T})\|_{L^2(Q)} \lesssim T^{-2}.$$

This is in particular the case when A is periodic, since then the operator $\mathcal{L}$ has a spectral gap (there is a Poincaré inequality on $H^1_\#(Q)$), so that the spectral integral is bounded independently of T (the integration interval is indeed isolated from zero, and λ^{-1} is $de_{\mathfrak{d}}(\lambda)$ on $\mathbb{R}^+$). We have therefore gained one power of T^{-1} with respect to our previous estimate $\|\nabla(\phi - \phi_T)\|_{L^2(Q)} \lesssim T^{-1}$. We can iterate this procedure and obtain any convergence rate in T for the symmetric periodic case. This approach also generalizes to the nonsymmetric periodic case using only PDE arguments (and not spectral theory), and we refer the reader to [32] for details.

Let use this new approximation of the corrector in the direct approach. As in Definition 4 we let $I_H \in \mathbb{N}$, and $\{Q_{H,i}\}_{i\in[\![1,I_H]\!]}$ be a partition of D in disjoint subdomains of diameter of order H. We further set $Q^\delta_{H,i} := \{x \in D \,|\, d(x, Q_{H,i}) < (\delta-1)H\}$, which is an enlarged version of $Q_{H,i}$. For all $h > 0$ and $i \in [\![1, I_H]\!]$ we let $V^\delta_{H,i,h}$ be a Galerkin subspace of $H^1_0(Q^\delta_{H,i})$. For all $\tau > 0$, we define the numerical correctors $\gamma^{\delta,H,h,i}_{\tau,\rho,\varepsilon}$ associated with $u^{\delta,H,h}_{T,\rho,\varepsilon}$ as the unique weak solution in $V^\delta_{H,i,h}$ to

$$(\tau\varepsilon^2)^{-1}\gamma^{\delta,H,h,i}_{\tau,\rho,\varepsilon} - \nabla\cdot A_\varepsilon\left(M_H(\nabla u^{\delta,H,h}_{T,\rho,\varepsilon}) + \nabla\gamma^{\delta,H,h,i}_{\tau,\rho,\varepsilon}\right) = 0,$$

define $\gamma^{\delta,H,h,i}_{2,T,\rho,\varepsilon} := 2\gamma^{\delta,H,h,i}_{2T,\rho,\varepsilon} - \gamma^{\delta,H,h,i}_{T,\rho,\varepsilon}$, set

$$\nabla v^{\delta,H,h,i}_{2,T,\rho,\varepsilon} := M_H(\nabla u^{\delta,H,h}_{T,\rho,\varepsilon})|_{Q_{H,i}} + (\nabla\gamma^{\delta,H,h,i}_{2,T,\rho,\varepsilon})|_{Q_{H,i}}$$

for all $1 \le i \le I_H$, and finally consider

$$C^{\delta,H,h}_{2,T,H,\varepsilon} := \sum_{i=1}^{I_H} \nabla v^{\delta,H,h,i}_{2,T,H,\varepsilon} 1_{Q_{H,i}}.$$

In the periodic case, we then have:

$$\begin{aligned}\|u^{\delta,H,h}_{T,H,\varepsilon} - u_{\rm hom}\|_{H^1(D)} &\lesssim H + \left(\frac{\varepsilon}{H}\right)^{4^-} + \left(\frac{h}{\varepsilon}\right)^2,\\ \|\nabla u_\varepsilon - C^{\delta,H,h}_{2,T,H,\varepsilon}\|_{L^2(D)} &\lesssim H + \left(\frac{\varepsilon}{H}\right)^{4^-} + \frac{h}{\varepsilon} + \sqrt{\varepsilon},\end{aligned}$$

provided

$$T := \left(\frac{H}{\varepsilon}\right)^2 \ln^{-2}\frac{H}{\varepsilon}.$$

Doing so, we've upgraded the convergence rate for the corrector. These theoretical results are quite spectacular, and one readily convinces oneself that any finite order of convergence can be reached by iterating the Richardson procedure: the resonance error has disappeared.

6. Other approaches and perspectives

6.1. **Other approaches**

In this survey we have essentially focused on the so-called HMM (direct approach) and so-called MsFEM (dual approach), which are two popular numerical homogenization methods. The convergence analysis of these methods can be made using the same analytical framework — the difference between the two approaches arising during the discretization. Another more global approach has been introduced by Schwab and Hoang in [41], inspired by

- two-scale convergence [4] and periodic unfolding [14],
- sparse tensor-products approximation [40].

The consistency of this approach can be proved using our analytical framework as well. With the notation of Section 2.1, their starting point is indeed the fact that the equation for $u_{\rho,\varepsilon}$ takes the equivalent form (provided A_ε is extended on a neigborhood of D): Find $u_{\rho,\varepsilon} \in H^1_0(D)$ and $U_{\rho,\varepsilon} \in L^2(D, H^1_0(Q))$ such that for all $v \in H^1_0(D)$ and all $V \in L^2(D, H^1_0(Q))$,

$$\int_D \int_Q (\nabla_x v(x) + \nabla_y V(x,y)) \cdot A_\varepsilon(x+\rho y)(\nabla_x u_{\rho,\varepsilon}(x) + \nabla_y U_{\rho,\varepsilon}(x,y))dxdy \;=\; \int_D f(x)v(x)dx.$$

This point of view allows them to look for a efficient approximation of the space $L^2(D, H^1_0(Q))$, and implement sparse-tensor product strategies.

Another approach advocated by Owhadi and Zhang in [53], whose origin is not the homogenization theory properly speaking, is based on the notion of harmonic coordinates. The starting point of their work is that although solutions $u \in H^1_0(D)$ to the equation

$$-\nabla \cdot A\nabla u \;=\; f$$

are usually not more regular than H^1 in the Euclidian coordinates, they are H^2 in the so-called harmonic coordinates provided $f \in L^2(D)$. These harmonic coordinates are defined as $x \mapsto x + \Gamma(x) \cdot x$, where $\Gamma = (\gamma_1, \ldots, \gamma_d)$ and γ_k is the unique weak solution in $H^1_0(D)$ to

$$-\nabla \cdot A(\mathbf{e}_k + \nabla\gamma_k) = 0$$

for all $k \in \{1, \ldots, k\}$. The issue is then to compute (approximations of) these harmonic coordinates. An interesting link between harmonic coordinates and the MsFEM has been made by Allaire and Brizzi in [5], who consider as multiscale finite element space the functions $x \mapsto v_H(x + \Gamma_H(x))$, where Γ_H is a local approximation

of the harmonic coordinates Γ on each simplex of the macroscopic tesselation of D. The combination of harmonic coordinates with the regularization method was recently studied by Owhadi and Zhang in [54].

6.2. **Beyond the linear case**

The reader may wonder what remains of all this when we replace the linear equation by a nonlinear equation, say elliptic monotone. Then part of the results translates quite naturally to the nonlinear case, namely the analytical framework (see [20, 21] for the description of the MsFEM method for nonlinear problems and error estimates in the periodic case, and [27,28] for the extension of the analytical framework of Section 2 to monotone elliptic operators and nonconvex integral functionals). Yet most of the quantitative results break down. It is also not clear how the regularization method behaves in that case.

In addition, from a practical point of view, things get much worse. In the linear case, one could have the excuse that the multiscale finite element basis or the local homogenized coefficients and local correctors only had to be computed once, and could be used to solve the equation for a wide variety of boundary conditions and right hand sides. This is definitely an advantage when dealing with optimal control for instance. In the nonlinear case, this does not hold any longer, and the local basis or local values of the operator have to be computed on the fly. This is not doable in practice.

For nonlinear problems, even periodic homogenization is not that easy since the object to reconstruct is a homogenized nonlinear map (and not a single matrix). The same fact holds for stochastic homogenization (with the additional difficulty that the corrector equation is posed on the whole space). A possible strategy is to compute approximations of the nonlinear map at some sampling points, and try to reconstruct an analytical approximation of this map by solving an inverse problem. This analytical approximation should at least satisfy similar properties as the homogenized nonlinear map (say convexity, isotropy, etc.). This strategy has been used in [16] to approximate the homogenized energy density associated with a discrete model for rubber introduced in [33] and whose homogenization limit has been etablished in [3]. There, the homogenized energy density is known to be quasiconvex, isotropic, and minimal at identity. An important issue is then to identify a suitable (explicit) subset of such functions in order to approximate the inverse problem. The constraint being nonlinear and the error landscape being quite rough, deterministic optimization methods usually fail to converge. In [16], the use of an evolutionary algorithm has been quite efficient.

An intermediate case between nonlinear problems and the linear equations considered throughout this survey is the case of parametrized linear equations. Such an example naturally appears in [31], where a coupled system of elliptic/parabolic equations is studied (for application to nuclear waste storage). Indeed, although the two equations are linear and the coefficients periodic, the nonlinear coupling condition makes the homogenized coefficient of the parabolic equation depend locally on the gradient of the solution of the homogenized elliptic equation. We are thus lead to solve a family of corrector equations of the form: Find $\phi^\zeta \in H^1_\#(Q)$ solution to

$$-\nabla \cdot A(\zeta)(\xi + \nabla\phi^\zeta) \,=\, 0 \qquad \text{in } Q,$$

parametrized by $\zeta \in \mathbb{R}^d$. This is an ideal framework for the reduced basis method [48], whose paradigm is that the family $\{\phi^\zeta, \zeta \in \mathbb{R}^d\}$ *may* be a thin subset of $H^1_\#(Q)$, so that it could be well approximated by a suitable low dimensional subspaces of $H^1_\#(Q)$ (see [9,15] for the case when $A(\zeta) = A_0 + \zeta A_1$, with ζ in some compact set of $\mathbb{R}$). This method has already been used in the context of homogenization by Boyaval in [13]. As usual with the reduced basis method, most of the work has to be done on the choice of the estimator, and on the fast-assembly method. In [31] we have used an estimator which is specific to homogenization problems, and appealed to the fast Fourier transform to assemble rigidity matrices efficiently.

6.3. **What next ?**

In the nonlinear case, much remains to be done. The approach which consists in constructing analytical proxies for the homogenized operators raise quite either unaddressed or unsolved challenging questions in approximation theory.

Even in the linear case, things are far from complete yet. In particular, as emphasized by Nolen, Papanicolaou, and Pironneau in [52], the numerical homogenization Grail is an adaptive method which would take the best of all worlds: use efficient domain decomposition techniques when needed, and exploit homogenization when possible.

References

[1] A. Abdulle. On a priori error analysis of fully discrete heterogeneous multiscale FEM. *Multiscale Model. Simul.*, 4:447–459, 2005.

[2] R. Alicandro and M. Cicalese. A general integral representation result for the continuum limits of discrete energies with superlinear growth. *SIAM J. Math. Anal.*, 36(1):1–37, 2004.

[3] R. Alicandro, M. Cicalese, and A. Gloria. Integral representation results for energies defined on stochastic lattices and application to nonlinear elasticity. *Arch. Ration. Mech. Anal.*, 200(3):881–943, 2011.

[4] G. Allaire. Homogenization and two-scale convergence. *SIAM J. Math. Anal.*, 23:1482–1518, 1992.

[5] G. Allaire and R. Brizzi. A multiscale finite element method for numerical homogenization. *Multiscale Model. Simul.*, 4:790–812, 2005.

[6] T. Arbogast. Numerical subgrid upscaling of two-phase flow in porous media. In *Numerical treatment of multiphase flows in porous media (Beijing, 1999)*, volume 552 of *Lecture Notes in Phys.*, pages 35–49. Springer, Berlin, 2000.

[7] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. Boundary layer analysis in homogenization of diffusion equations with Dirichlet conditions in the half space. Wiley, 1976.

[8] A. Bensoussan, J.-L. Lions, and G. Papanicolaou. *Asymptotic analysis for periodic structures*, volume 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1978.

[9] P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, and P. Wojtaszczyk. Convergence rates for greedy algorithms in reduced basis methods. *SIAM J. Math. Anal.*, 43(3):1457–1472, 2011.

[10] X. Blanc and C. Le Bris. Improving on computation of homogenized coefficients in the periodic and quasi-periodic settings. *Netw. Heterog. Media*, 5(1):1–29, 2010.

[11] A. Bourgeat, A. Mikelić, and S. Wright. Stochastic two-scale convergence in the mean and applications. *J. Reine Angew. Math.*, 456:19–51, 1994.

[12] A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. I. H. Poincaré Probab. Stat.*, 40(2):153–165, 2004.

[13] S. Boyaval. Reduced-basis approach for homogenization beyond the periodic setting. *Multiscale Model. Simul.*, 7(1):466–494, 2008.

[14] D. Cioranescu, A. Damlamian, and G. Griso. Periodic unfolding and homogenization. *C. R. Math. Acad. Sci. Paris*, 335(1):99–104, 2002.

[15] A. Cohen, R. Devore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl. (Singap.)*, 9(1):11–47, 2011.

[16] M. De Buhan, A. Gloria, P. Le Tallec, and M. Vidrascu. Reconstruction of a constitutive law for rubber from in silico experiments. In preparation.

[17] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden. Heterogeneous multiscale methods: A review. *Commun. Comput. Phys.*, 2:367–450, 2007.

[18] W. E, P.B. Ming, and P.W. Zhang. Analysis of the heterogeneous multiscale method for elliptic homogenization problems. *J. Amer. Math. Soc.*, 18:121–156, 2005.

[19] Weinan E. *Principles of multiscale modeling*. Cambridge University Press, Cambridge, 2011.

[20] Y. Efendiev and T. Y. Hou. *Multiscale finite element methods*, volume 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York, 2009. Theory and applications.

[21] Y.R. Efendiev, T.Y. Hou, and V. Ginting. Multiscale finite element methods for nonlinear problems and their applications. *Com. Math. Sc.*, 2(4):553–589, 2004.

[22] Y.R. Efendiev, T.Y. Hou, and X.H. Wu. Convergence of a nonconforming multiscale finite element method. *SIAM J. Num. Anal.*, 37:888–910, 2000.

[23] A.-C. Egloffe, A. Gloria, J.-C. Mourrat, and T. N. Nguyen. Theoretical and numerical investigation of convergence rates in stochastic homogenization: corrector equation and random walk in random environment. In prepration.

[24] L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.

[25] F. Féyel. Multiscale FE^2 elastoviscoplastic analysis of composite structures. *Comp. Mat. Sci.*, 16:344–354, 1999.
[26] FreeFEM. http://www.freefem.org/.
[27] A. Gloria. An analytical framework for the numerical homogenization of monotone elliptic operators and quasiconvex energies. *Multiscale Model. Simul.*, 5(3):996–1043, 2006.
[28] A. Gloria. An analytical framework for numerical homogenization - Part II: windowing and oversampling. *Multiscale Model. Simul.*, 7(1):275–293, 2008.
[29] A. Gloria. Reduction of the resonance error - Part 1: Approximation of homogenized coefficients. *Math. Models Methods Appl. Sci.*, 21(8):1601–1630, 2011.
[30] A. Gloria. Numerical approximation of effective coefficients in stochastic homogenization of discrete elliptic equations. *M2AN Math. Model. Numer. Anal.*, 46(1):1–38, 2012.
[31] A. Gloria, T. Goudon, and S. Krell. Numerical homogenization of a nonlinearly coupled elliptic-parabolic system, reduced basis method, and application to nuclear waste storage. 2012. Preprint, http://hal.archives-ouvertes.fr/hal-00674519.
[32] A. Gloria and Z. Habibi. Reduction of the resonance error - Part 2: Approximation of correctors and spectral theory. In preparation.
[33] A. Gloria, P. Le Tallec, and M. Vidrascu. Comparison of network-based models for rubber. 2012. Preprint, http://hal.archives-ouvertes.fr/hal-00673406.
[34] A. Gloria and J.-C. Mourrat. Spectral measure and approximation of homogenized coefficients. *Probab. Theory. Relat. Fields.* DOI 10.1007/s00440-011-0370-7.
[35] A. Gloria, S. Neukamm, and F. Otto. Approximation of effective coefficients by periodization in stochastic homogenization. In preparation.
[36] A. Gloria, S. Neukamm, and F. Otto. Quantification of ergodicity in stochastic homogenization: optimal bounds via spectral gap on Glauber dynamics. In preparation.
[37] A. Gloria and F. Otto. Optimal quantitative estimates in stochastic homogenization of linear elliptic equations. In preparation.
[38] A. Gloria and F. Otto. An optimal variance estimate in stochastic homogenization of discrete elliptic equations. *Ann. Probab.*, 39(3):779–856, 2011.
[39] A. Gloria and F. Otto. An optimal error estimate in stochastic homogenization of discrete elliptic equations. *Ann. Appl. Probab.*, 22(1):1–28, 2012.
[40] M. Griebel and S. Knapek. Optimized tensor-product approximation spaces. *Constr. Approx.*, 16(4):525–540, 2000.
[41] V.H. Hoang and Ch. Schwab. High-dimensional finite elements for elliptic problems with multiple scales. *Multiscale Model. Simul.*, 3(1):168–194, 2005.
[42] T.Y. Hou and X.H Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134:169–189, 1997.
[43] T.Y. Hou, X.H. Wu, and Z.Q. Cai. Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comput.*, 68:913–943, 1999.
[44] T.Y. Hou, X.H. Wu, and Y. Zhang. Removing the cell resonance error in the multiscale finite element method via a Petrov-Galerkin formulation. *Comm. in Math. Sci.*, 2(2):185–205, 2004.
[45] V.V. Jikov, S.M. Kozlov, and O.A. Oleinik. *Homogenization of Differential Operators and Integral Functionals*. Springer-Verlag, Berlin, 1994.
[46] T. Kanit, S. Forest, I. Galliet, V. Mounoury, and D. Jeulin. Determination of the size of the representative volume element for random composites: statistical and numerical approach. *Int. J. Sol. Struct.*, 40:3647–3679, 2003.
[47] S.M. Kozlov. The averaging of random operators. *Mat. Sb. (N.S.)*, 109(151)(2):188–202, 327, 1979.
[48] Y. Maday, A. T. Patera, and G. Turinici. Global a priori convergence theory for reduced-basis approximations of single-parameter symmetric coercive elliptic partial differential equations. *C. R. Math. Acad. Sci. Paris*, 335(3):289–294, 2002.
[49] N. Meyers. An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (3)*, 17(3):189–206, 1963.
[50] F. Murat. H-convergence. Séminaire d'Analyse fonctionnelle et numérique, Univ. Alger, multigraphié, 1978.
[51] F. Murat and L. Tartar. H-convergence. In A.V. Cherkaev and R.V. Kohn, editors, *Topics in the Mathematical Modelling of Composites Materials*, volume 31 of *Progress in nonlinear differential equations and their applications*, pages 21–44. Birkhäuser, 1997.
[52] J. Nolen, G. Papanicolaou, and O. Pironneau. A framework for adaptive multiscale methods for elliptic problems. *Multiscale Model. Simul.*, 7(1):171–196, 2008.
[53] H. Owhadi and L. Zhang. Metric-based upscaling. *Comm. Pure Appl. Math.*, 60(5):675–723, 2007.
[54] H. Owhadi and L. Zhang. Localized bases for finite dimensional homogenization approximations with non-separated scales and high-contrast. *Multiscale Model. Simul.*, 9(4):1373–1398, 2011.
[55] G.C. Papanicolaou and S.R.S. Varadhan. Boundary value problems with rapidly oscillating random coefficients. In *Random fields, Vol. I, II (Esztergom, 1979)*, volume 27 of *Colloq. Math. Soc. János Bolyai*, pages 835–873. North-Holland, Amsterdam, 1981.
[56] E. M. Stein. *Singular integrals and differentiability properties of functions*, volume 30 of *Princeton Mathematical Series*. Princeton University Press, Princeton, NJ, 1970.

[57] L. Tartar. Cours Peccot du Collège de France. 1977.
[58] M. Vogelius. A homogenization result for planar, polygonal networks. *RAIRO Modél. Math. Anal. Numér.*, 25(4):483–514, 1991.
[59] X. Yue and W. E. The local microscale problem in the multiscale modeling of strongly heterogeneous media: effects of boundary conditions and cell size. *J. Comput. Phys.*, 222(2):556–572, 2007.
[60] V. V. Yurinskiĭ. Averaging of symmetric diffusion in random medium. *Sibirskii Matematicheskii Zhurnal*, 27(4):167–180, 1986.