ESAIM: COCV 24 (2018) 1345–1380
https://doi.org/10.1051/cocv/2017061

ESAIM: Control, Optimisation and Calculus of Variations
www.esaim-cocv.org

ON THE BEST CONSTANT MATRIX APPROXIMATING AN OSCILLATORY MATRIX-VALUED COEFFICIENT IN DIVERGENCE-FORM OPERATORS

Claude Le Bris[1,2], Frédéric Legoll[2,3,*] and Simon Lemaire[1,2]

Abstract. We approximate an elliptic problem with oscillatory coefficients using a problem of the same type, but with constant coefficients. We deliberately take an engineering perspective, where the information on the oscillatory coefficients in the equation can be incomplete. A theoretical foundation of the approach in the limit of infinitely small oscillations of the coefficients is provided, using the classical theory of homogenization. We present a comprehensive study of the implementation aspects of our method, and a set of numerical tests and comparisons that show the potential practical interest of the approach. The approach detailed in this article improves on an earlier version briefly presented in [C. Le Bris, F. Legoll and K. Li, *C.R. Acad. Sci. Paris, Série I* **351** (2013) 265–270].

Mathematics Subject Classification. 35J, 35B27, 74Q15.

Received December 17, 2016. Accepted September 5, 2017.

1. Introduction

1.1. Context

Consider the simple, linear, elliptic equation

$$-\mathrm{div}(A_\varepsilon \nabla u_\varepsilon) = f \quad \text{in} \quad \mathcal{D}, \qquad u_\varepsilon = 0 \quad \text{on} \quad \partial\mathcal{D}, \tag{1.1}$$

in divergence-form, where $\mathcal{D} \subset \mathbb{R}^d$, $d \geq 1$, is an open, bounded domain which delimits what we hereafter call 'the physical medium', and where A_ε is a possibly random oscillatory matrix-valued coefficient. We suppose that all the requirements are satisfied so that problem (1.1) is well-posed. In particular, we assume that A_ε is bounded and bounded away from zero uniformly in ε. Our assumptions will be detailed in Section 2.1 below. The subscript ε encodes the characteristic scale of variation of the matrix field A_ε. For instance, one may think of the case $A_\varepsilon(\boldsymbol{x}) = A^{\mathrm{per}}(\boldsymbol{x}/\varepsilon)$ for a fixed $\mathbb{Z}^d$-periodic matrix field A^{per}, although all what follows is not restricted to that particular case.

It is well-known that, for ε small (comparatively to the size of $\mathcal{D}$), and not necessarily infinitesimally small, the direct computation of the solution to (1.1) is expensive since, in order to capture the oscillatory behavior of

Keywords and phrases. Elliptic PDEs, Oscillatory coefficients, Homogenization, Coarse-graining.

[1] École des Ponts ParisTech, CERMICS, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France.

[2] Inria Paris, MATHERIALS project-team, 2 rue Simone Iff, CS 42112, 75589 Paris Cedex 12, France.

[3] École des Ponts ParisTech, Laboratoire Navier – UMR 8205, 6 et 8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France.

* Corresponding author: frederic.legoll@enpc.fr

A_ε and u_ε, one has to discretize the domain $\mathcal{D}$ with a meshsize $h \ll \varepsilon$. The computation becomes prohibitively expensive in a multi-query context where the solution $u_\varepsilon(f)$ is needed for a large number of right-hand sides f (think, *e.g.*, of a time-dependent model where (1.1), or a similar equation, should be solved at each time step t^n with a right-hand side $f(t^n)$, or of an optimization loop with f as an unknown variable, where (1.1) would encode a distributed constraint). Alternatives to the direct computation of u_ε exist. Depending on the value of ε, the situation is schematically as follows.

- For $\varepsilon < \overline{\varepsilon}$, where $\overline{\varepsilon}$ is a given, medium-dependent threshold (typically $\overline{\varepsilon} \approx \text{size}(\mathcal{D})/10$), one can consider that homogenization theory [3,13,19] provides a suitable framework to address problem (1.1). That theory ensures the existence of a limit problem for infinitely small oscillations of the coefficient A_ε. The limit problem reads

$$-\text{div}(A_\star \boldsymbol{\nabla} u_\star) = f \quad \text{in} \quad \mathcal{D}, \qquad u_\star = 0 \quad \text{on} \quad \partial\mathcal{D}. \tag{1.2}$$

  The matrix-valued coefficient $A_\star$ is (i) non-oscillatory, (ii) independent of f, and (iii) given by an abstract definition that can become more or less explicit, depending on the assumptions concerning the structure of A_ε (and the probabilistic setting in the random case). The solution to the homogenized problem (1.2) can be considered an accurate L^2-approximation of the oscillatory solution to (1.1) as soon as the size ε of the oscillations of A_ε is sufficiently small.

  There are several cases for which the abstract definition giving $A_\star$ can be made explicit. The simplest examples are (i) periodic coefficients of the form $A_\varepsilon(\boldsymbol{x}) = A^{\text{per}}(\boldsymbol{x}/\varepsilon)$, with A^{per} a $\mathbb{Z}^d$-periodic matrix field, and (ii) stationary ergodic coefficients of the form $A_\varepsilon(\boldsymbol{x},\omega) = A^{\text{sto}}(\boldsymbol{x}/\varepsilon,\omega)$, with A^{sto} a (continuous or discrete) stationary matrix field. In both cases, one can prove that $A_\star$ is a deterministic constant (*i.e.* independent of $\boldsymbol{x}$) matrix, for which a simple explicit expression is available. Whenever a corrector (in the terminology of homogenization theory, see [3, 13, 19] and (2.7)–(2.8) below) exists, it is in addition possible to reconstruct an H^1-approximation of the solution to (1.1), using the solutions to the corrector problem and to the homogenized problem (1.2).

  Practically, whenever an explicit definition is available for $A_\star$, one can compute an approximation of the oscillatory solution to (1.1) by solving the non-oscillatory problem (1.2). The advantage is obviously that the latter can be solved on a coarse mesh. The cost of the method then lies in the offline computation of $A_\star$.
- For $\varepsilon \geq \overline{\varepsilon}$, the size of the oscillations is too large to consider that homogenization theory provides a suitable framework to approximate problem (1.1), and one may use, in order to efficiently compute an approximation of u_ε, dedicated numerical approaches.

  Classical examples include the Variational Multiscale Method (VMM) introduced by Hughes *et al.* [12], and the Multiscale Finite Element Method (MsFEM) introduced by Hou and Wu [11] (see also the textbook [9]). We also refer to the more recent works by Målqvist and Peterseim [17] (on the Local Orthogonal Decomposition (LOD) method), or Kornhuber and Yserentant [14], on localization and subspace decomposition. Many more examples of approaches are available in the literature.

  The MsFEM approach (as well as the LOD approach) is essentially based on an offline/online decomposition of the computations. In the first step, local problems are solved at the microscale, in order to compute oscillatory basis functions. Each basis function is obtained by solving an oscillatory problem posed on a macro-element or on a patch of macro-elements. These oscillatory problems do not depend on the right-hand side f, and are independent one from another. In the second step, the global problem, which depends on the right-hand side f, is solved. The second step is performed, *e.g.*, by considering a Galerkin approximation on the multiscale discrete space built in the offline step. The original online cost of solving an oscillatory problem on a fine mesh (using a discrete space at one single fine scale) is reduced to solving an oscillatory problem on a coarse mesh consisting of macro-elements (using a multiscale discrete space).

  These methods provide an H^1-approximation of the oscillatory solution u_ε. Note that they are (*a priori*) applicable without any restriction on the structure of A_ε, and are also applicable, and indeed applied, in the regime $\varepsilon < \overline{\varepsilon}$. Note also that, in the stochastic setting, the computations must be performed ω by ω, for "each" realization ω of the random environment.

The finite element Heterogeneous Multiscale Method (HMM) introduced by E and Engquist [8] is another popular multiscale technique. It is however based on a different perspective. Its aim is to compute an approximation of the coarse solution $u_\star$ by means of local averages of the oscillatory coefficient A_ε.

One way or another, all these approaches rely on the knowledge of the coefficient A_ε. It turns out that there are several contexts where such a knowledge is incomplete, or sometimes merely unavailable. From an engineering perspective (think, *e.g.*, of experiments in Mechanics), there are numerous prototypical situations where the response $u_\varepsilon(f)$ can be measured for *some* loadings f, but where A_ε is *not* completely known. In these situations, it is thus not possible to use homogenization theory, nor to proceed with any MsFEM-type approach or with the similar approaches mentioned above.

We have discussed above two possibilities to address multiscale problems such as (1.1), using either the homogenization theory or dedicated numerical approaches. Restricting our discussion to homogenization theory, we can identify three limitations, quite different in nature, to the practical application of the theory:

- First, homogenization theory has been developed in order to address the case of infinitely small oscillations of the coefficients, and is hence not appropriate for media such that $\varepsilon \geq \bar{\varepsilon}$. In practice, one may for instance want to evaluate the effective coefficients (such as the Poisson ratio and the Young modulus for problems in Mechanics) of a medium for which $\varepsilon \geq \bar{\varepsilon}$. It is always possible (if an explicit definition is available) to compute $A_\star$, considering on purpose the (fictitious) limit of infinitely small oscillations, but there is no reason for that $A_\star$ to be an accurate approximation of the medium it is supposed to describe.
- Assume that an explicit expression is available for $A_\star$. A practical limitation is that, in most cases except for the somewhat ideal case of periodic coefficients (with a known period), the computation of $A_\star$ by classical methods is expensive. For instance, in the stochastic setting, the computation of $A_\star$ requires to solve, many times, a corrector problem set on a truncated approximation of an asymptotically infinitely large domain. This is especially challenging in the stationary ergodic case with long-range correlations. Note that equivalent limitations appear for MsFEM-type or similar approaches in the stochastic setting.
- Another evident limitation shows up when one examines the homogenized limit of (1.1) for a coefficient A_ε such that no explicit expression is available for $A_\star$ (although A_ε is well-known, and although the homogenized limit of (1.1) is known to read as (1.2)). This case might occur as soon as A_ε is not the rescaling $A(\cdot/\varepsilon)$ of a simple (periodic, quasi-periodic, random stationary, ...) function A.

Finding a pathway alternate to standard approaches is thus a practically relevant question. Given our discussion above, we are interested in approaches valid for the different regimes of ε, which make no use of the knowledge on the coefficient A_ε, but only use some (measurable) responses of the medium (obtained for certain given solicitations). Questions similar in spirit, but different in practice, have been addressed two decades ago by Durlofsky in [7]. They are similar in spirit because the point is to define an effective coefficient only using outputs of the system. They are however different in practice because the effective matrix is defined by upscaling, and hence the approach of [7] is local. This approach is indeed based on considering, in a representative elementary volume, some particular problems (with zero loading and suitable boundary conditions), for which the solutions in the case of homogeneous coefficients are affine and write as independent of these homogeneous coefficients. Considering d choices of such problems (that is, d choices of boundary conditions), and postulating the equality of the fluxes respectively resulting from the original oscillatory and homogeneous equivalent problems, one determines the coefficients of an "effective" matrix. Several variants exist in the literature, as well as many other approaches.

The original approach we introduce in this article improves on an earlier version briefly presented in [16]. Our approach is global, in the sense that it uses the responses of the system in the *whole* domain $\mathcal{D}$. Note of course that it can be used locally as an upscaling technique, for instance in problems featuring a prohibitively large number of degrees of freedom.

In passing, we note that our approach provides, at least in some settings, a characterization of the homogenized matrix which is an alternative to the standard characterization of homogenization theory (see Prop. 3.2 below). To the best of our knowledge, this characterization has never been made explicit in the literature.

Throughout this article, we restrict ourselves to cases when problem (1.1) admits (possibly up to some extraction) a homogenized limit that reads as problem (1.2), where the homogenized matrix coefficient

$$A_\star \text{ is } \textit{deterministic} \text{ and } \textit{constant}.$$

This restrictive assumption on the class of $A_\star$ (and thus on the structure of the coefficient A_ε in (1.1), and on the probabilistic setting in the random case) is useful for our theoretical justifications, but not mandatory for the approach to be applicable (see Sect. 1.3 below).

1.2. Presentation of our approach

We now sketch, for a coefficient A_ε that we take for simplicity deterministic, the idea underlying our approach. Let $\mathcal{S}$ denote the set of real-valued $d \times d$ positive-definite symmetric matrices.

For any constant matrix $\overline{A} \in \mathcal{S}$, consider generically the problem with *constant* coefficients

$$-\mathrm{div}(\overline{A}\boldsymbol{\nabla}\overline{u}) = f \quad \text{in} \quad \mathcal{D}, \qquad \overline{u} = 0 \quad \text{on} \quad \partial\mathcal{D}. \tag{1.3}$$

We investigate, for any value of the parameter ε, how we may define a constant matrix $\overline{A}_\varepsilon \in \mathcal{S}$ such that the solution $\overline{u}_\varepsilon$ to problem (1.3) with matrix $\overline{A} = \overline{A}_\varepsilon$ best approximates the solution u_ε to (1.1). Note that, since $\overline{A}_\varepsilon$ is constant, its skew-symmetric part plays no role in (1.3). We hence cannot hope for characterizing the skew-symmetric part of $\overline{A}_\varepsilon$. Without loss of generality, we henceforth make the additional assumption that the homogenized matrix $A_\star$ is *symmetric* and that we seek a best (constant) symmetric matrix. Should $A_\star$ not be symmetric, it is replaced in the sequel by its symmetric part. In [16], the constant matrix $\overline{A}_\varepsilon$ is defined as a minimizer of

$$\inf_{\overline{A}\in\mathcal{S}} \sup_{f\in L^2(\mathcal{D}),\, \|f\|_{L^2(\mathcal{D})}=1} \|u_\varepsilon(f) - \overline{u}(f)\|^2_{L^2(\mathcal{D})}, \tag{1.4}$$

where we have emphasized the dependency upon the right-hand side f of the solutions to (1.1) and (1.3). The use of a L^2 norm in (1.4) (and not of *e.g.* a H^1 norm) is reminiscent of the fact that, for sufficiently small ε, we wish the best constant matrix to be close to $A_\star$, and that u_ε converges to $u_\star$ in the L^2 norm but not in the H^1 norm.

Note that problem (1.4) is only based on the knowledge of the outputs $u_\varepsilon(f)$ (that could be, *e.g.*, experimentally measured), and *not* on that of A_ε itself. Note also that, in practice, we cannot maximize upon all right-hand sides f in $L^2(\mathcal{D})$ (with unit norm). We therefore have to replace the supremum in (1.4) by a maximization upon a finite-dimensional set of right-hand sides, which we will have to select thoughtfully (see Sect. 3.1.1).

In this article, we keep the same type of characterization for $\overline{A}_\varepsilon$ as in [16] (that is, through an inf-sup problem), but we use a slightly different cost function than in (1.4). The constant matrix $\overline{A}_\varepsilon$ is here defined as a minimizer of

$$\inf_{\overline{A}\in\mathcal{S}} \sup_{f\in L^2(\mathcal{D}),\, \|f\|_{L^2(\mathcal{D})}=1} \left\|(-\Delta)^{-1}\left(\mathrm{div}(\overline{A}\boldsymbol{\nabla}u_\varepsilon(f)) + f\right)\right\|^2_{L^2(\mathcal{D})}, \tag{1.5}$$

where $(-\Delta)^{-1}$ is the inverse laplacian operator supplied with homogeneous Dirichlet boundary conditions: for any $g \in H^{-1}(\mathcal{D})$, $z = (-\Delta)^{-1}g$ is the unique solution in $H^1_0(\mathcal{D})$ to

$$-\Delta z = g \quad \text{in} \quad \mathcal{D}, \qquad z = 0 \quad \text{on} \quad \partial\mathcal{D}.$$

The cost function of (1.5) is related to the one of (1.4) through the application, inside the L^2 norm of the latter, of the zero-order differential operator $(-\Delta)^{-1}\left(\mathrm{div}(\overline{A}\boldsymbol{\nabla}\cdot)\right)$. Note that, in sharp contrast with (1.4), the function $\left\|(-\Delta)^{-1}\left(\mathrm{div}(\overline{A}\boldsymbol{\nabla}u_\varepsilon(f)) + f\right)\right\|^2_{L^2(\mathcal{D})}$ used in (1.5) is a polynomial function of degree 2 in terms of $\overline{A}$, a property which brings stability and significantly speeds up the computations. The specific choice (1.5) has been suggested to us by Albert Cohen (Université Pierre et Marie Curie).

Remark 1.1. The reason to choose $f \in L^2(\mathcal{D})$ in (1.5), rather than $f \in H^{-1}(\mathcal{D})$, is discussed in Remark 3.1 below.

Several criteria can be considered to assess the quality and the usefulness of our approach:

(i) *asymptotic consistency:* does the sequence $\left\{\overline{A}_\varepsilon\right\}_{\varepsilon>0}$ of best matrices, defined as minimizers of (1.5), converge, when ε goes to 0, to the homogenized matrix $A_\star$? If this is indeed the case, the approach provides an approximation for the homogenized matrix alternate to standard homogenization (note, in particular, that our approach does not require solving a corrector problem).

(ii) *efficiency:* practically, is this best matrix $\overline{A}_\varepsilon$ efficiently computable? In particular, how many right-hand sides does its computation really require?

(iii) L^2*-approximation:* for any fixed ε, not necessarily small, how well does the solution $\overline{u}_\varepsilon$ to (1.3) with matrix $\overline{A}_\varepsilon$ approximate the reference solution u_ε to (1.1) in the L^2 norm?

(iv) H^1*-approximation:* using $\overline{A}_\varepsilon$, is it possible to reconstruct (if possible for a marginal additional cost) an accurate approximation of u_ε in the H^1 norm? Recall that in homogenization theory, a corrector problem must be solved to compute the homogenized matrix, but once this is performed, one can reconstruct an H^1-approximation of u_ε using the solution of the latter problem at no additional cost.

1.3. Outline and perspectives

The article is organized as follows. To begin with, we introduce in Section 2 the assumptions we will make throughout the article, and we recall the basics of homogenization. We formalize our approach in Section 3. We establish an asymptotic consistency result (thereby positively answering to Question (i) above, see Prop. 3.2), and we explain how the best matrix we compute can be used to construct an approximation in the H^1 norm of the oscillatory solution (hence addressing Question (iv) above). We also detail how to approximate the infinite-dimensional space $\{f \in L^2(\mathcal{D}), \quad \|f\|_{L^2(\mathcal{D})} = 1\}$ present in (1.5) by a finite-dimensional space of the form $\mathrm{Span}\{f_p,\ 1 \le p \le P\}$ for some appropriate functions f_p (see (3.5) below). In Section 4, we explain how the problem of finding the best constant matrix can be efficiently solved in practice (thereby answering to Question (ii)).

Finally, in Section 5, we present, as a practical answer to Questions (i), (ii), (iii) and (iv), a number of representative numerical experiments, both in the periodic and stationary ergodic settings, and we provide some comparison with the classical homogenization approach. We show in particular that choosing a small number P of right-hand sides (in practice, we often set $P = d(d+1)/2$) is sufficient for our approach to provide accurate results.

We emphasize that the aim of the numerical experiments described in Section 5 is different in the periodic setting and in the stochastic setting. In the former case, computing the homogenized matrix is inexpensive, and thus we cannot hope for our approach (which requires solving highly oscillatory equations) to outperform the classical homogenization approach in terms of efficiency. The periodic setting is hence to be considered as a validation setting.

The situation is entirely different in the stochastic setting, which is much more challenging. In that setting, our approach can compete as far as Questions (i), (ii), (iii) and (iv), are concerned. We show that, for an essentially identical computational cost compared to the standard homogenization approach, our approach allows us to compute a more accurate approximation of the solution u_ε to the highly oscillatory equation, both in L^2 and in H^1 norms.

More importantly, the reader should bear in mind that our approach targets practical situations where the information on the oscillatory coefficients in the equation may be incomplete. The comparison with standard homogenization approaches which is performed in Section 5 is hence somewhat unfair for our approach, as the former approaches need a complete knowledge of the coefficient A_ε, whereas ours does not.

There are several possible follow-ups for this work:

- First, one can perform a detailed study of the robustness of the approach with respect to imprecise data, assuming for instance that we only have access locally to *coarse* averages of the outputs $u_\varepsilon(f)$ or $\nabla u_\varepsilon(f)$.
- Second, the extension to nonlinear equations may be studied, where the oscillatory problem is formulated as the optimization problem

$$\inf\left\{\int_{\mathcal{D}} K\left(\frac{\boldsymbol{x}}{\varepsilon},\nabla u(\boldsymbol{x})\right)\mathrm{d}\boldsymbol{x}-\int_{\mathcal{D}} f(\boldsymbol{x})u(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},\quad u\in W_0^{1,p}(\mathcal{D})\right\},$$

  where the function $\boldsymbol{\xi}\in\mathbb{R}^d\mapsto K(\cdot,\boldsymbol{\xi})$ is strictly convex. In a multi-query context, our approach (and this is also true for other approaches) is even more interesting for nonlinear equations than for linear ones. Indeed, however large the parameter ε is, solving a nonlinear oscillatory equation for a large number of right-hand sides is prohibitively expensive. In contrast, in the linear case, as soon as the LU decomposition of the stiffness matrix can be computed and stored, *i.e.* as soon as ε is not too small, the cost for computing several solutions becomes almost equal to the cost for computing one. The computational workload thus remains affordable. This is not the case in a nonlinear context.
- Third, the approach may be extended to homogenized matrices that are not constant. Indeed, as soon as some additional information is available on $A_\star$, one could adequately modify the search space for $\overline{A}$ in (1.4) or (1.5). For instance, the case of a slowly varying matrix $A_\star(\boldsymbol{x})$, depending upon $\boldsymbol{x}\in\mathcal{D}$ in a sense to be made precise, can be considered. Following a suggestion by Albert Cohen, it may also be possible to balance the dimension of the space in which $\overline{A}$ is searched with the amount of noise present in the problem (which is related to the value of ε) and the number of fine-scale solutions that are available (here the dimension P of the space (3.5) introduced below).

2. Preliminaries

We describe the stationary ergodic setting we adopt. This setting includes, as a particular case, the periodic case. For a more detailed presentation of the particular stochastic setting we here consider, we refer to the theoretically-oriented articles [4, 5], to the numerically-oriented articles [6, 15], and to the review article [2] (as well as to the extensive bibliography contained therein). For more insight on stochastic homogenization in general, we refer the reader to the seminal contribution [18], to [10] for a numerically-oriented presentation, as well as to the classical textbooks [3, 13]. The reader familiar with that theory may easily skip this section and directly proceed to Section 3.

2.1. Assumptions

Recall that $\mathcal{D}$ denotes an open, bounded subset of $\mathbb{R}^d$, $d\geq 1$. Let $(\Omega,\mathcal{Z},\mathbb{P})$ be a probability space, on which we assume an ergodic structure, and let $\mathbb{E}(X)=\displaystyle\int_\Omega X(\omega)\,\mathrm{d}\mathbb{P}(\omega)$ be the expectation of any random variable $X\in L^1(\Omega,\mathrm{d}\mathbb{P})$. We consider problem (1.1), which reads, in the stochastic setting, as

$$-\mathrm{div}(A_\varepsilon(\cdot,\omega)\nabla u_\varepsilon(\cdot,\omega))=f\quad\text{a.s. in}\quad\mathcal{D},\qquad u_\varepsilon(\cdot,\omega)=0\quad\text{a.s. on}\quad\partial\mathcal{D},\tag{2.1}$$

where the function $f\in L^2(\mathcal{D})$ is independent of ε and deterministic (see Rem. 3.1 below for a discussion on the choice of taking f in $L^2(\mathcal{D})$).

We assume that

$$A_\varepsilon(\boldsymbol{x},\omega)=A^{\mathrm{sto}}(\boldsymbol{x}/\varepsilon,\omega),\tag{2.2}$$

where A^{sto} is such that there exist deterministic real numbers $\alpha,\beta>0$ such that

$$A^{\mathrm{sto}}(\cdot,\omega)\in L^\infty(\mathbb{R}^d;\mathcal{S}_{\alpha,\beta})\quad\text{almost surely},\tag{2.3}$$

with

$$\mathcal{S}_{\alpha,\beta} = \left\{ M \in \mathbb{R}^{d\times d}, \quad M \text{ is symmetric}, \quad \alpha\,|\boldsymbol{\xi}|^2 \leq \boldsymbol{\xi}^T M \boldsymbol{\xi} \leq \beta\,|\boldsymbol{\xi}|^2 \quad \text{for any} \quad \boldsymbol{\xi} \in \mathbb{R}^d \right\}.$$

In addition, we assume that $A^{\rm sto}$ is a discrete stationary matrix field. A complete description of the discrete stationary ergodic setting we here consider can be found, *e.g.*, in the review article [2], (Sect. 2.2). For brevity, we only mention here that the purpose of this setting is to formalize the fact that, even though realizations may vary, the matrix $A^{\rm sto}$ at point $\boldsymbol{y} \in \mathbb{R}^d$ and the matrix $A^{\rm sto}$ at point $\boldsymbol{y}+\boldsymbol{k}$, $\boldsymbol{k} \in \mathbb{Z}^d$, share the same probability law. The local, microscopic environment (encoded in the oscillatory matrix field $A_\varepsilon(\boldsymbol{x},\omega) = A^{\rm sto}(\boldsymbol{x}/\varepsilon,\omega)$) has a $\varepsilon\mathbb{Z}^d$-periodic structure *on average*.

Assumption (2.3) ensures the existence and uniqueness of the solution to (2.1) in $H^1_0(\mathcal{D})$, almost surely. Furthermore, almost surely, the solution $u_\varepsilon(\cdot,\omega)$ to (2.1) converges (strongly in $L^2(\mathcal{D})$ and weakly in $H^1(\mathcal{D})$) to some $u_\star \in H^1_0(\mathcal{D})$ solution to (1.2), where the homogenized matrix $A_\star$ is *deterministic*, *constant* and belongs to $\mathcal{S}_{\alpha,\beta}$. As is well-known, $A_\star$ is independent of the right-hand side f in (2.1).

Remark 2.1. The above discussion is not restricted to the *discrete* stationary setting. We could as well have considered the *continuous* stationary setting, where the probability law of $A^{\rm sto}(\boldsymbol{y},\omega)$ does not depend on $\boldsymbol{y}$.

Remark 2.2. The form of the homogenized equation (1.2) is in this context identical to that of the original equation (1.1). This is not a general fact. Although definite conclusions are yet to be obtained, there are all reasons to believe that the practical approach we introduce in this article carries over to cases where the homogenized equation is of a different form.

The periodic setting is a particular case of the above discrete stationary setting, when A is independent of ω. This amounts to assuming that

$$A_\varepsilon(\boldsymbol{x}) = A^{\rm per}(\boldsymbol{x}/\varepsilon), \tag{2.4}$$

with $A^{\rm per}$ a $\mathbb{Z}^d$-periodic matrix field such that

$$A^{\rm per} \in L^\infty(\mathbb{R}^d;\mathcal{S}_{\alpha,\beta}). \tag{2.5}$$

2.2. Classical homogenization approach

We briefly recall here the basics of homogenization. We focus the presentation on the stationary ergodic setting. The easy adaptation to the periodic setting is briefly commented upon.

Let $Q = (0,1)^d$. In the discrete stationary ergodic setting, the (deterministic, constant and symmetric) homogenized matrix $A_\star$ reads, for all $1 \leq i,j \leq d$, as

$$[A_\star]_{i,j} = \mathbb{E}\left(\int_Q (\boldsymbol{e}_i + \boldsymbol{\nabla} w_{\boldsymbol{e}_i}(\boldsymbol{y},\cdot))^T \; A^{\rm sto}(\boldsymbol{y},\cdot) \; \left(\boldsymbol{e}_j + \boldsymbol{\nabla} w_{\boldsymbol{e}_j}(\boldsymbol{y},\cdot)\right) \, {\rm d}\boldsymbol{y} \right), \tag{2.6}$$

where $(\boldsymbol{e}_1,\ldots,\boldsymbol{e}_d)$ denotes the canonical basis of $\mathbb{R}^d$, and where, for any $\boldsymbol{p} \in \mathbb{R}^d$, $w_{\boldsymbol{p}}$ is the solution (unique up to the addition of a random constant) to the so-called *corrector equation*

$$\begin{cases} -{\rm div}\left(A^{\rm sto}(\cdot,\omega)(\boldsymbol{p} + \boldsymbol{\nabla} w_{\boldsymbol{p}}(\cdot,\omega))\right) = 0 \quad \text{a.s. in} \quad \mathbb{R}^d, \\ \boldsymbol{\nabla} w_{\boldsymbol{p}} \text{ is stationary}, \qquad \mathbb{E}\left(\displaystyle\int_Q \boldsymbol{\nabla} w_{\boldsymbol{p}}(\boldsymbol{y},\cdot)\,{\rm d}\boldsymbol{y} \right) = 0. \end{cases} \tag{2.7}$$

In the periodic case $A_\varepsilon(\boldsymbol{x}) = A^{\rm per}(\boldsymbol{x}/\varepsilon)$, the corrector equation reads as

$$\begin{cases} -{\rm div}\left(A^{\rm per}(\boldsymbol{p} + \boldsymbol{\nabla} w_{\boldsymbol{p}})\right) = 0 \quad \text{in} \quad \mathbb{R}^d, \\ w_{\boldsymbol{p}} \text{ is } \mathbb{Z}^d\text{-periodic}, \end{cases} \tag{2.8}$$

and the homogenized matrix $A_\star$ is given by

$$[A_\star]_{i,j} = \int_Q (\boldsymbol{e}_i + \boldsymbol{\nabla} w_{\boldsymbol{e}_i}(\boldsymbol{y}))^T \; A^{\rm per}(\boldsymbol{y}) \; (\boldsymbol{e}_j + \boldsymbol{\nabla} w_{\boldsymbol{e}_j}(\boldsymbol{y})) \; {\rm d}\boldsymbol{y}.$$

In sharp contrast with the periodic case where, precisely by periodicity, it is sufficient to solve the corrector equation (2.8) on the unit cell Q, the corrector equation (2.7) must be solved in the discrete stationary ergodic setting on the entire space $\mathbb{R}^d$. As pointed out in the introduction, this is computationally challenging. In practice, one often considers a truncated corrector equation posed, for an integer $N \neq 0$, on a large domain $Q^N = (-N, N)^d$:

$$-{\rm div}\left(A^{\rm sto}(\cdot,\omega)(\boldsymbol{p} + \boldsymbol{\nabla} w^N_{\boldsymbol{p}}(\cdot,\omega))\right) = 0 \quad \text{a.s. in} \quad Q^N, \qquad w^N_{\boldsymbol{p}}(\cdot,\omega) \quad \text{is a.s.} \quad Q^N - \text{periodic}. \tag{2.9}$$

The random matrix $A^N_\star(\omega)$, approximation of the deterministic homogenized matrix $A_\star$ given by (2.6), is defined, for all $1 \leq i, j \leq d$, by

$$\left[A^N_\star(\omega)\right]_{i,j} = \frac{1}{|Q^N|} \int_{Q^N} \left(\boldsymbol{e}_i + \boldsymbol{\nabla} w^N_{\boldsymbol{e}_i}(\boldsymbol{y},\omega)\right)^T \; A^{\rm sto}(\boldsymbol{y},\omega) \; \left(\boldsymbol{e}_j + \boldsymbol{\nabla} w^N_{\boldsymbol{e}_j}(\boldsymbol{y},\omega)\right) {\rm d}\boldsymbol{y}. \tag{2.10}$$

Almost surely, it converges, in the limit of infinitely large domains Q^N, *i.e.* when $N \to +\infty$, to the (deterministic) matrix $A_\star$ (see [6]). Since $A^N_\star(\omega)$ is random, it is natural to consider M independent and identically distributed (i.i.d.) realizations of the field $A^{\rm sto}$, say $\{A^{\rm sto}(\cdot,\omega_m)\}_{1\leq m\leq M}$, solve (2.9) and compute (2.10) for each of them, and define

$$A^{N,M}_\star = \frac{1}{M} \sum_{m=1}^{M} A^N_\star(\omega_m) \tag{2.11}$$

as a practical approximation to $A_\star$. Owing to the strong law of large numbers, we have that $\lim\limits_{N\to\infty} \lim\limits_{M\to\infty} A^{N,M}_\star = A_\star$ almost surely.

3. Formalization of our approach

The approach we introduce below applies, up to minor changes, to both the periodic and the stationary ergodic settings. We however recall from Section 1.3 that only the stochastic setting (and more difficult cases) is practically relevant for our approach. For simplicity and clarity, we first present the full study of the approach in the periodic setting (see Sects. 3.1, 3.2 and 3.3). We next discuss its extension to the stationary ergodic setting in Section 3.4.

3.1. Infsup formulation

As exposed in the introduction and expressed in formula (1.5), we are going to seek a constant, symmetric, positive-definite matrix $\overline{A}_\varepsilon$, so that problem (1.3) with matrix $\overline{A}_\varepsilon$ best approximates problem (1.1). To do so, we consider the problem introduced in (1.5), that is

$$I_\varepsilon = \inf_{\overline{A}\in\mathcal{S}} \sup_{f\in L^2_{\rm n}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f), \tag{3.1}$$

where $L^2_{\rm n}(\mathcal{D}) = \{f \in L^2(\mathcal{D}), \quad \|f\|_{L^2(\mathcal{D})} = 1\}$ and where, for any $\overline{A} \in \mathbb{R}^{d\times d}_{\rm sym}$ (the space of $d \times d$ real symmetric matrices) and any $f \in L^2(\mathcal{D})$,

$$\Phi_\varepsilon(\overline{A}, f) = \left\|(-\Delta)^{-1}\left({\rm div}(\overline{A}\boldsymbol{\nabla} u_\varepsilon(f)) + f\right)\right\|^2_{L^2(\mathcal{D})}. \tag{3.2}$$

Note that formula (3.2) is well-defined since ${\rm div}(\overline{A}\boldsymbol{\nabla} u_\varepsilon(f))$ clearly belongs to $H^{-1}(\mathcal{D})$ for all $\overline{A} \in \mathbb{R}^{d\times d}_{\rm sym}$ and $f \in L^2(\mathcal{D})$. We observe, as briefly mentioned in Section 1.2, that the cost function $\Phi_\varepsilon(\cdot, f)$ depends quadratically

upon $\overline{A}$. From a computational viewpoint, in an iterative algorithm solving (1.5) or (3.1) that successively optimizes on f and $\overline{A}$, minimizing Φ_ε with respect to $\overline{A}$ for a fixed $f \in L^2_{\rm n}(\mathcal{D})$ thus reduces to the simple inversion of a small linear system with $d(d+1)/2$ unknowns (see Sect. 4.3.3). This is in sharp contrast with our former formulation (1.4). Of course, in both formulations (1.4) or (1.5), for ε fixed, it is not guaranteed that our numerical algorithm captures the value I_ε defined by (3.1). It only captures an approximation of it.

For both approaches (1.4) and (1.5), one can prove an asymptotic consistency result for the sequence $\{\overline{A}_\varepsilon\}_{\varepsilon>0}$: see Proposition 3.2 below in the case of (1.5) and [16] in the case of (1.4). As the proof is essentially identical for both approaches, we only detail it for the present choice (1.5) (see Appendix A below) and briefly point out to the case (1.4) considered in [16] in Remark 3.3 below.

In order to gain further insight, and before stating the asymptotic consistency result, we first study, separately and for a fixed value of ε, the maximization and minimization problems involved in (3.1).

3.1.1. The sup *problem*

We show here that, for any fixed $\overline{A} \in \mathcal{S}$, the maximization problem over f that is involved in (3.1), namely $\sup_{f\in L^2_{\rm n}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f)$, is attained, and discuss how it can be solved in practice.

Let $\overline{A} \in \mathcal{S}$ be given. We introduce the notation

$$\Delta_{\overline{A}} = \mathrm{div}(\overline{A}\boldsymbol{\nabla}\cdot),$$

and let $(-\Delta_{\overline{A}})^{-1}$ be the operator defined by: for any $g \in H^{-1}(\mathcal{D})$, $z = (-\Delta_{\overline{A}})^{-1} g$ is the unique solution in $H^1_0(\mathcal{D})$ to

$$-\mathrm{div}(\overline{A}\boldsymbol{\nabla} z) = g \quad \text{in} \quad \mathcal{D}, \qquad z = 0 \quad \text{on} \quad \partial\mathcal{D}.$$

We denote by $\mathrm{L}_\varepsilon^{-1}$ the linear, compact and positive-definite operator from $L^2(\mathcal{D})$ to $L^2(\mathcal{D})$ such that, for any $f \in L^2(\mathcal{D})$, $\mathrm{L}_\varepsilon^{-1} f = u_\varepsilon(f)$, where $u_\varepsilon(f)$ is the unique solution in $H^1_0(\mathcal{D})$ to (1.1). Starting from (3.2), it can be easily shown that

$$\Phi_\varepsilon(\overline{A}, f) = \int_{\mathcal{D}} \mathcal{H}_\varepsilon^{\overline{A}}(f)\, f, \tag{3.3}$$

where

$$\mathcal{H}_\varepsilon^{\overline{A}}(f) = \left(\left(\mathrm{L}_\varepsilon^{-1}\right)^\star \Delta_{\overline{A}} (-\Delta)^{-1} + (-\Delta)^{-1} \right) \left((-\Delta)^{-1} \Delta_{\overline{A}} \mathrm{L}_\varepsilon^{-1} + (-\Delta)^{-1} \right) f \tag{3.4}$$

is a compact, self-adjoint and positive semi-definite linear operator from $L^2(\mathcal{D})$ to $L^2(\mathcal{D})$. The eigenvalues of $\mathcal{H}_\varepsilon^{\overline{A}}$ are thus nonnegative real numbers forming a sequence that converges to zero. We denote by $\lambda_{\varepsilon,\mathrm{m}}^{\overline{A}}$ and $f_{\varepsilon,\mathrm{m}}^{\overline{A}}$ the largest eigenvalue of $\mathcal{H}_\varepsilon^{\overline{A}}$ and an associated normalized eigenvector, respectively. In view of (3.3), we have

$$\sup_{f\in L^2_{\rm n}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f) = \lambda_{\varepsilon,\mathrm{m}}^{\overline{A}}$$

and the supremum is attained at $f_{\varepsilon,\mathrm{m}}^{\overline{A}}$, which is hence a solution to the sup problem involved in (3.1).

In practice, instead of looking for the largest eigenvalue (and the associated eigenvector) of $\mathcal{H}_\varepsilon^{\overline{A}}$ in the infinite-dimensional space $L^2_{\rm n}(\mathcal{D})$, our approach consists in approximating this space $L^2_{\rm n}(\mathcal{D})$ by a finite-dimensional subspace of the form

$$V_{\rm n}^P(\mathcal{D}) = \left\{ f \in L^2_{\rm n}(\mathcal{D}) \text{ s.t. there exists } \boldsymbol{c} = \{c_p\}_{1\le p\le P} \in \mathbb{R}^P, \quad |\boldsymbol{c}|^2 = 1, \quad f = \sum_{p=1}^P c_p f_p \right\}, \tag{3.5}$$

where $(f_1, \ldots, f_P)$ is an orthonormal family of functions in $L^2(\mathcal{D})$.

We discuss the choice of the dimension P and of the family of functions $\{f_p\}_{1\le p\le P}$. First of all, in the light of Lemma A.2 below (see also Sect. 3.2), it seems in order to choose the dimension of $V_{\rm n}^P(\mathcal{D})$ such that $P \ge d(d+1)/2$.

We now proceed, considering the regime ε small. Let $\overline{A} \neq A_\star$ be fixed. Homogenization theory states that, for ε sufficiently small, the operator $\mathrm{L}_\varepsilon^{-1}$ (considered as an operator from $L^2(\mathcal{D})$ to $L^2(\mathcal{D})$) is close to the operator $(-\Delta_{A_\star})^{-1}$. Thus the operator $\mathcal{H}_\varepsilon^{\overline{A}}$ defined by (3.4) is expected to be well-approximated by

$$\mathcal{H}_\star^{\overline{A}} = \left((-\Delta_{A_\star})^{-1}\,\Delta_{\overline{A}}\,(-\Delta)^{-1} + (-\Delta)^{-1}\right)\left((-\Delta)^{-1}\,\Delta_{\overline{A}}\,(-\Delta_{A_\star})^{-1} + (-\Delta)^{-1}\right). \tag{3.6}$$

Up to the extraction of a subsequence, the eigenvector $f_{\varepsilon,\mathrm{m}}^{\overline{A}}$ we are seeking thus satisfies, by homogenization theory on eigenvalue problems,

$$\lim_{\varepsilon\to 0} \left\| f_{\varepsilon,\mathrm{m}}^{\overline{A}} - f_{\star,\mathrm{m}}^{\overline{A}} \right\|_{L^2(\mathcal{D})} = 0,$$

where $f_{\star,\mathrm{m}}^{\overline{A}}$ is a normalized eigenvector associated with the largest eigenvalue of $\mathcal{H}_\star^{\overline{A}}$. In view of the expression (3.6) of the limit operator, it seems natural to choose for the family of functions $\{f_p\}_{1\leq p\leq P}$ the first P (normalized) eigenvectors of the laplacian operator in the domain $\mathcal{D}$. For small values of ε, say $\varepsilon < \bar{\varepsilon}$, we show that considering $P = d(d+1)/2$ functions f_p is sufficient. This threshold $d(d+1)/2$ is at least intuitive thinking at the case of a constant symmetric matrix $\overline{A}$ and the set of equations $\displaystyle\sum_{1\leq i,j\leq d} -\overline{A}_{i,j}\,\partial_{ij}u_p = f_p$. In order to determine the $d(d+1)/2$ coefficients $\overline{A}_{i,j}$, the correct number of right-hand sides f_p to consider is $d(d+1)/2$. The fact that it is indeed sufficient is made precise in the proof of Proposition 3.2 below (see in particular Lem. A.2) and in Remark 3.5 below.

When the parameter ε takes larger values, say $\varepsilon \geq \bar{\varepsilon}$, the operator $\mathcal{H}_\varepsilon^{\overline{A}}$ cannot be anymore approximated by the operator (3.6) (with constant coefficients), and it may thus be necessary in that case to consider a larger number $P > d(d+1)/2$ of functions. We refer to Section 5 for concrete examples.

Remark 3.1. We discuss here why we have chosen to work with right-hand sides f of the equation (*e.g.* (2.1)) in $L^2(\mathcal{D})$ rather than in $H^{-1}(\mathcal{D})$. We have here considered $\displaystyle\sup_{f\in L^2(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{L^2(\mathcal{D})}}$, and we could have considered $\displaystyle\sup_{f\in H^{-1}(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{H^{-1}(\mathcal{D})}}$.

Since $L^2(\mathcal{D}) \subset H^{-1}(\mathcal{D})$, we of course have $\displaystyle\sup_{f\in H^{-1}(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{H^{-1}(\mathcal{D})}} \geq \sup_{f\in L^2(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{H^{-1}(\mathcal{D})}}$. Using the density of $L^2(\mathcal{D})$ in $H^{-1}(\mathcal{D})$ and the continuity of $\Phi_\varepsilon(\overline{A},\cdot)$ in $H^{-1}(\mathcal{D})$, we actually get

$$\sup_{f\in H^{-1}(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{H^{-1}(\mathcal{D})}} = \sup_{f\in L^2(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{H^{-1}(\mathcal{D})}}. \tag{3.7}$$

The right-hand side of (3.7) is of course different from the quantity $\displaystyle\sup_{f\in L^2(\mathcal{D})} \frac{\Phi_\varepsilon(\overline{A},f)}{\|f\|^2_{L^2(\mathcal{D})}}$, which we have considered in this article. Our choice is motivated by the fact that it is easier *in practice* to manipulate functions of unit L^2-norm. From the *theoretical* viewpoint, similar results would have been obtained with the left-hand side of (3.7).

3.1.2. The inf *problem*

We discuss here how to efficiently solve the minimization problem over $\overline{A}$ that is involved in (3.1), namely

$$\inf_{\overline{A}\in\mathcal{S}} \Phi_\varepsilon(\overline{A},f). \tag{3.8}$$

Let $f \in L^2_{\mathrm{n}}(\mathcal{D})$ be fixed. It can be easily shown, starting from (3.2) and using the linearity of both the divergence and inverse laplacian operators, that

$$\Phi_\varepsilon(\overline{A},f) = \frac{1}{2}\sum_{1\leq i,j,k,l\leq d} [\mathbb{B}_\varepsilon(f)]_{i,j,k,l}\,\overline{A}_{i,j}\,\overline{A}_{k,l} - \sum_{1\leq i,j\leq d} [B_\varepsilon(f)]_{i,j}\,\overline{A}_{i,j} + b(f), \tag{3.9}$$

where the fourth-order tensor $\mathbb{B}_\varepsilon(f)$, the matrix $B_\varepsilon(f)$ and the scalar $b(f)$, which all depend on f, are given, for integers $1 \leq i, j, k, l \leq d$, by

$$\begin{aligned}
[\mathbb{B}_\varepsilon(f)]_{i,j,k,l} &= 2 \int_{\mathcal{D}} \left[(-\Delta)^{-1}(\partial_{ij} u_\varepsilon(f))\right] \left[(-\Delta)^{-1}(\partial_{kl} u_\varepsilon(f))\right], \\
[B_\varepsilon(f)]_{i,j} &= -2 \int_{\mathcal{D}} \left[(-\Delta)^{-1}(\partial_{ij} u_\varepsilon(f))\right] \left[(-\Delta)^{-1} f\right], \\
b(f) &= \left\| (-\Delta)^{-1} f \right\|^2_{L^2(\mathcal{D})}.
\end{aligned}$$

Practically, the inf problem (3.8) (with fixed f) is solved on the whole set $\mathbb{R}^{d\times d}_{\mathrm{sym}}$ of symmetric matrices, instead of considering the subset $\mathcal{S}$ of positive-definite symmetric matrices. Under this simplification, solving the inf problem (3.8) amounts to considering the linear system

$$\forall\, 1 \leq i, j \leq d, \qquad \sum_{1\leq k,l\leq d} [\mathbb{B}_\varepsilon(f)]_{i,j,k,l}\ \overline{A}_{k,l} = [B_\varepsilon(f)]_{i,j}\,. \tag{3.10}$$

This system is low-dimensional and inexpensive to solve. In our numerical experiments, we have observed that the problem (3.10) always has a unique solution in $\mathbb{R}^{d\times d}_{\mathrm{sym}}$, for all the functions f that our algorithm explores. In addition, this solution is in $\mathcal{S}$.

3.2. Asymptotic consistency

We study here problem (3.1) in the limit of a vanishing parameter ε. We introduce the notation

$$\Phi_\varepsilon(\overline{A}) = \sup_{f\in L^2_{\mathrm{n}}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f). \tag{3.11}$$

Note that Φ_ε is nonnegative. Consequently, for any ε, problem (3.1) admits a quasi-minimizer, namely a matrix $\overline{A}^\flat_\varepsilon \in \mathcal{S}$ such that

$$I_\varepsilon \leq \Phi_\varepsilon(\overline{A}^\flat_\varepsilon) \leq I_\varepsilon + \varepsilon \leq \Phi_\varepsilon(\overline{A}) + \varepsilon \qquad \text{for any } \overline{A} \in \mathcal{S}. \tag{3.12}$$

The following proposition holds.

Proposition 3.2 (Asymptotic consistency, periodic case)**.** *Consider problem* (3.1)*, that is*

$$I_\varepsilon = \inf_{\overline{A}\in\mathcal{S}} \sup_{f\in L^2_{\mathrm{n}}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f).$$

In the periodic setting, namely under the assumptions (2.4) *and* (2.5)*, the following convergence holds*:

$$\lim_{\varepsilon\to 0} I_\varepsilon = 0. \tag{3.13}$$

Furthermore, for any sequence $\{\overline{A}^\flat_\varepsilon \in \mathcal{S}\}_{\varepsilon>0}$ *of quasi-minimizers of* (3.1)*, we have*

$$\lim_{\varepsilon\to 0} \overline{A}^\flat_\varepsilon = A_\star. \tag{3.14}$$

The proof of these results, which is postponed until Appendix A, relies on two facts:

(1) The homogenized matrix $A_\star \in \mathcal{S}_{\alpha,\beta} \subset \mathcal{S}$ can be used as a test-matrix in (3.12). In view of Lemma A.1 below, it satisfies $\lim_{\varepsilon\to 0} \Phi_\varepsilon(A_\star) = 0$, which directly implies (3.13);

(2) We show in Lemma A.2 below that there exist $d(d+1)/2$ right-hand sides $f_{\star,k} \in L^2_{\rm n}(\mathcal{D})$ such that the knowledge of $f_{\star,k}$ and of $u_{\star,k}$ solution to (1.2) with right-hand side $f_{\star,k}$, $1 \leq k \leq d\,(d+1)/2$, is sufficient to uniquely reconstruct the constant symmetric matrix $A_\star$. The proof of (3.14) relies on this argument and on (3.13). We denote

$$\mathcal{F} = \{f_{\star,k}, \quad 1 \leq k \leq d(d+1)/2\} \tag{3.15}$$

this set.

We do not know whether, for ε fixed, the infimum in (3.1) is attained, unless ε is sufficiently small (see Rem. A.4 in Appendix A.2 below). We will proceed throughout the article manipulating quasi-minimizers in the sense of (3.12).

Remark 3.3. The analysis of the approach (1.4) introduced in [16] relies on the same arguments as the approach introduced here: Lemma A.2, and the equivalent of Lemma A.1 for the functional considered in [16], that is $\lim_{\varepsilon\to 0} \Psi_\varepsilon(A_\star) = 0$, where, for any $\overline{A} \in \mathcal{S}$,

$$\Psi_\varepsilon(\overline{A}) = \sup_{f\in L^2_{\rm n}(\mathcal{D})} \|u_\varepsilon(f) - \overline{u}(f)\|^2_{L^2(\mathcal{D})}.$$

Remark 3.4. Note that the assumptions (2.4) and (2.5) are not necessary to prove the results (3.13) and (3.14). All that needs to be assumed is that the sequence of matrices $\{A_\varepsilon\}_{\varepsilon>0}$ converges, in the sense of homogenization, to a *constant* and *symmetric* homogenized matrix $A_\star$. In that vein, we will see in Section 3.4 below that the conclusions of Proposition 3.2 carry over to the specific stochastic case we consider there.

Remark 3.5. Consider the set $\mathcal{F}$ defined by (3.15), and let

$$I^{\max}_\varepsilon = \inf_{\overline{A}\in\mathcal{S}} \max_{f\in\mathcal{F}} \Phi_\varepsilon(\overline{A}, f). \tag{3.16}$$

This problem is, in principle, easier to solve than (3.1), as we replaced the supremum over $f \in L^2_{\rm n}(\mathcal{D})$ by a maximization over the finite set $\mathcal{F}$. Let $\Phi^{\max}_\varepsilon(\overline{A}) = \max_{f\in\mathcal{F}} \Phi_\varepsilon(\overline{A}, f)$. For any quasi-minimizer $\overline{A}^{\max,\flat}_\varepsilon \in \mathcal{S}$ of (3.16), we have

$$I^{\max}_\varepsilon \leq \Phi^{\max}_\varepsilon(\overline{A}^{\max,\flat}_\varepsilon) \leq I^{\max}_\varepsilon + \varepsilon \leq \Phi^{\max}_\varepsilon(A_\star) + \varepsilon \leq \Phi_\varepsilon(A_\star) + \varepsilon.$$

Since $\lim_{\varepsilon\to 0} \Phi_\varepsilon(A_\star) = 0$, we get that $\lim_{\varepsilon\to 0} I^{\max}_\varepsilon = 0$. In addition, one can show that $\overline{A}^{\max,\flat}_\varepsilon \to A_\star$ as $\varepsilon \to 0$ (we refer to Rem. A.3 below for details). Similarly to (3.1), the approach (3.16) is therefore asymptotically consistent. Note however that, in practice, the functions of the set $\mathcal{F}$ defined by (3.15) are unknown.

We note that Proposition 3.2 provides, in the setting described in Section 2.1, a characterization of the homogenized matrix which is an alternative to the standard characterization of homogenization theory. To the best of our knowledge, this characterization has never been made explicit in the literature.

3.3. Approximation of u_ε in the H^1 norm

As a consequence of Proposition 3.2, we note that $\overline{u}_\varepsilon$, solution to (1.3) with matrix $\overline{A}_\varepsilon$, is an accurate approximation of u_ε in the L^2 norm, but not in the H^1 norm. Indeed, when ε goes to zero, $\overline{A}_\varepsilon$ converges to $A_\star$. Hence, for ε sufficiently small, $\overline{u}_\varepsilon$ is an accurate H^1-approximation of $u_\star$ solution to (1.2). In addition, from homogenization theory, we know that $u_\star$ is an accurate L^2-approximation of u_ε. This implies that $\lim_{\varepsilon\to 0} \|\overline{u}_\varepsilon - u_\varepsilon\|_{L^2(\mathcal{D})} = 0$.

Note also that $u_\star$ and u_ε are not close to each other in the H^1 norm, and hence $\overline{u}_\varepsilon$ is not an accurate approximation of u_ε in the H^1 norm. We present here an approach to reconstruct such an approximation.

In many settings of homogenization theory (and in particular in the periodic setting we consider here), once the corrector problems are solved to compute the homogenized matrix, one can consider the two-scale expansion (truncated at the first-order)

$$u_\varepsilon^{1,\boldsymbol{\theta}}(\boldsymbol{x}) = u_\star(\boldsymbol{x}) + \varepsilon \sum_{i=1}^{d} w_{\boldsymbol{e}_i}^{\theta_i}(\boldsymbol{x}/\varepsilon)\, \partial_i u_\star(\boldsymbol{x}), \tag{3.17}$$

where $w_{\boldsymbol{e}_i}^{\theta_i}$ is the unique solution with mean value $\theta_i \in \mathbb{R}$ to the periodic corrector equation (2.8) for $\boldsymbol{p} = \boldsymbol{e}_i$. It is well-known that this two-scale expansion approximates u_ε in the H^1 norm, in the sense that, under some regularity assumptions (see *e.g.* [1]), we have

$$\left\| u_\varepsilon - u_\varepsilon^{1,\boldsymbol{\theta}} \right\|_{H^1(\mathcal{D})} \leq C\, \sqrt{\varepsilon} \tag{3.18}$$

for a constant C independent of ε.

Remark 3.6. From the theoretical perspective, the mean value $\boldsymbol{\theta}$ of the correctors is irrelevant, and the estimate (3.18) holds for any fixed $\boldsymbol{\theta}$. From the numerical perspective, the error $\left\| u_\varepsilon - u_\varepsilon^{1,\boldsymbol{\theta}} \right\|_{H^1(\mathcal{D})}$ slightly depends on $\boldsymbol{\theta}$, in particular when ε is not asymptotically small. In view of the numerical tests described in Section 5 below (see *e.g.* (5.10)), we keep track of this parameter.

Computing the gradient of (3.17), we deduce from (3.18) that

$$\boldsymbol{\nabla} u_\varepsilon = C_\varepsilon\, \boldsymbol{\nabla} u_\star + \text{h.o.t.}, \tag{3.19}$$

where the $d \times d$ matrix C_ε is given by

$$[C_\varepsilon]_{i,i} = 1 + \partial_i w_{\boldsymbol{e}_i}(\cdot/\varepsilon), \qquad [C_\varepsilon]_{i,j} = \partial_i w_{\boldsymbol{e}_j}(\cdot/\varepsilon) \quad \text{if } j \neq i. \tag{3.20}$$

Our idea for constructing an approximation of $\boldsymbol{\nabla} u_\varepsilon$ is to mimick formula (3.19) and seek an approximation under the form $\overline{C}_\varepsilon \boldsymbol{\nabla} \overline{u}_\varepsilon$. Once the best matrix $\overline{A}_\varepsilon$ has been computed, we compute a surrogate $\overline{C}_\varepsilon$ of C_ε by solving the least-squares problem

$$\inf_{\overline{C} \in (L^2(\mathcal{D}))^{d\times d}} \quad \sum_{r=1}^{R} \left\| \boldsymbol{\nabla} u_\varepsilon(f_r) - \overline{C}\ \boldsymbol{\nabla} \overline{u}_\varepsilon(f_r) \right\|_{L^2(\mathcal{D})^d}^2 \tag{3.21}$$

for a given number R of right-hand sides.

In practice, the right-hand sides f_r selected for (3.21) are the first R basis functions of the space $V_{\mathrm{n}}^P(\mathcal{D})$ defined by (3.5), with R such that

$$R \leq P.$$

This choice makes the H^1-reconstruction an inexpensive post-processing procedure once the best matrix is computed, as we already have at our disposal $u_\varepsilon(f_r)$ for $1 \leq r \leq R$.

Remark 3.7. In our numerical experiments, we have observed that the surrogate $\overline{C}_\varepsilon$ that we construct is indeed oscillatory, and essentially periodic when A_ε is periodic. This is expected since $\overline{C}_\varepsilon$ is meant to be an approximation of C_ε.

In practice, we independently identify each row of $\overline{C}_\varepsilon$, by considering (for any $1 \leq i \leq d$) the least-squares problem

$$\inf_{\overline{c}^i \in (L^2(\mathcal{D}))^d} \quad \sum_{r=1}^{R} \left\| \partial_i u_\varepsilon(f_r) - \overline{c}^i \cdot \boldsymbol{\nabla} \overline{u}_\varepsilon(f_r) \right\|_{L^2(\mathcal{D})}^2.$$

We next define the matrix $\overline{C}_\varepsilon$ by $\left[\overline{C}_\varepsilon\right]_{i,j} = \left[\overline{c}^i_\varepsilon\right]_j$. In our numerical experiments, the functions u_ε and $\overline{u}_\varepsilon$ are approximated by $u_{\varepsilon,h}$ and $\overline{u}_{\varepsilon,h}$ using a $\mathbb{P}^1$ Finite Element Method, and $\overline{c}^i_\varepsilon$ is searched as a piecewise constant function. The value of $\overline{c}^i_\varepsilon$ on an element T is defined by the problem

$$\inf_{\overline{\boldsymbol{c}}^i_T \in \mathbb{R}^d} \sum_{r=1}^{R} \left| \left[\partial_i u_{\varepsilon,h}(f_r)\right]_{|T} - \overline{\boldsymbol{c}}^i_T \cdot \left[\boldsymbol{\nabla} \overline{u}_{\varepsilon,h}(f_r)\right]_{|T} \right|^2, \tag{3.22}$$

where the restrictions of $\partial_i u_{\varepsilon,h}$ and $\boldsymbol{\nabla}\overline{u}_{\varepsilon,h}$ to any element T are constant. This problem is ill-posed if $R < d$, since, in this case, there exist vectors in $\mathbb{R}^d$ orthogonal to all $\left[\boldsymbol{\nabla}\overline{u}_{\varepsilon,h}(f_r)\right]_{|T}$, $1 \leq r \leq R$. We thus always take $R \geq d$. To avoid technicalities related to the $\mathbb{P}^1$ discretization of $\overline{u}_\varepsilon$, only mesh elements *not* contiguous to the boundary of $\mathcal{D}$ are considered in the minimization (3.22).

3.4. The stationary ergodic setting

We have focused in Sections 3.1, 3.2 and 3.3 on the periodic setting. We now briefly turn to the stochastic ergodic setting. We introduce the *modified* cost function $\Phi^{\mathrm{sto}}_\varepsilon$ defined, for any $\overline{A} \in \mathbb{R}^{d\times d}_{\mathrm{sym}}$ and $f \in L^2(\mathcal{D})$, by

$$\Phi^{\mathrm{sto}}_\varepsilon(\overline{A}, f) = \left\| (-\Delta)^{-1} \left[\mathrm{div}\left(\overline{A} \boldsymbol{\nabla} \mathbb{E}(u_\varepsilon(f)) \right) + f \right] \right\|^2_{L^2(\mathcal{D})}. \tag{3.23}$$

Note that $\Phi^{\mathrm{sto}}_\varepsilon$ is a deterministic quantity. The difference with the cost function Φ_ε defined by (3.2) in a deterministic context is that $\Phi^{\mathrm{sto}}_\varepsilon$ involves $\mathbb{E}(u_\varepsilon(f))$ rather than $u_\varepsilon(f)$.

We next amend the inf sup problem (3.1) in the following way. For a given value of ε, we look for a best *deterministic* matrix $\overline{A}_\varepsilon \in \mathcal{S}$ that solves the problem

$$I^{\mathrm{sto}}_\varepsilon = \inf_{\overline{A}\in\mathcal{S}} \sup_{f \in L^2_{\mathrm{n}}(\mathcal{D})} \Phi^{\mathrm{sto}}_\varepsilon(\overline{A}, f). \tag{3.24}$$

All the considerations of Sections 3.1, 3.2 and 3.3 carry over, up to minor adjustments, to the present stochastic setting. Under assumptions (2.2) and (2.3), asymptotic consistency can be proved for any sequence $\{\overline{A}^\flat_\varepsilon \in \mathcal{S}\}_{\varepsilon>0}$ of quasi-minimizers of (3.24). The adaptation of the proof of Proposition 3.2 to the stochastic setting is straightforward. It relies on the fact that, for any $f \in L^2(\mathcal{D})$, $\mathbb{E}(u_\varepsilon(f))$ is bounded in $H^1(\mathcal{D})$. Indeed, using that $\alpha \leq A_\varepsilon(\cdot,\omega) \leq \beta$ almost surely, we have $\|u_\varepsilon(\cdot,\omega)\|_{H^1(\mathcal{D})} \leq \dfrac{C}{\alpha}\|f\|_{L^2(\mathcal{D})}$ almost surely (where C is a deterministic constant only depending on $\mathcal{D}$), hence $\mathbb{E}\left[\|u_\varepsilon\|^2_{H^1(\mathcal{D})}\right]$ is bounded. Using the Cauchy-Schwarz inequality, we infer that $\mathbb{E}(u_\varepsilon(f))$ is indeed bounded in $H^1(\mathcal{D})$. We eventually get that $\boldsymbol{\nabla}\mathbb{E}(u_\varepsilon(f))$ weakly converges, and $\mathbb{E}(u_\varepsilon(f))$ strongly converges, in $L^2(\mathcal{D})$ and when ε goes to zero, to $\boldsymbol{\nabla} u_\star(f)$ and $u_\star(f)$, respectively, where $u_\star(f)$ is the solution to (1.2).

The H^1-reconstruction procedure presented in Section 3.3 is adapted to the stationary ergodic setting as follows. It is known that, almost surely, $u_\varepsilon(\cdot,\omega)$ weakly converges in $H^1(\mathcal{D})$ towards $u_\star$ when ε goes to zero. As in the periodic setting, the correctors allow to obtain a strong convergence in $H^1(\mathcal{D})$, in the sense that (see [18], (Thm. 3))

$$\lim_{\varepsilon\to 0} \mathbb{E}\left[\left\| u_\varepsilon(\cdot,\omega) - u^1_\varepsilon(\cdot,\omega) \right\|^2_{H^1(\mathcal{D})} \right] = 0, \tag{3.25}$$

with

$$u^1_\varepsilon(\boldsymbol{x},\omega) = u_\star(\boldsymbol{x}) + \varepsilon \sum_{i=1}^{d} w_{\boldsymbol{e}_i}(\boldsymbol{x}/\varepsilon,\omega)\, \partial_i u_\star(\boldsymbol{x}), \tag{3.26}$$

where $w_{\boldsymbol{e}_i}$ is the unique solution with vanishing mean value to the stochastic corrector equation (2.7) for $\boldsymbol{p} = \boldsymbol{e}_i$ (in contrast to the periodic case, see Rem. 3.6, we only consider here correctors with *vanishing* mean, for the sake of simplicity).

The equations (3.25)–(3.26) imply that

$$\mathbb{E}\left[\boldsymbol{\nabla} u_\varepsilon(\cdot,\omega)\right] = C_\varepsilon\, \boldsymbol{\nabla} u_\star + \text{h.o.t.},$$

where the $d \times d$ matrix C_ε is given by

$$[C_\varepsilon]_{i,i} = 1 + \mathbb{E}\left[\partial_i w_{\boldsymbol{e}_i}(\cdot/\varepsilon,\omega)\right], \qquad [C_\varepsilon]_{i,j} = \mathbb{E}\left[\partial_i w_{\boldsymbol{e}_j}(\cdot/\varepsilon,\omega)\right] \quad \text{if } j \neq i. \tag{3.27}$$

We have chosen to look for an approximation of $\mathbb{E}(\boldsymbol{\nabla} u_\varepsilon)$ as follows. Once the best matrix $\overline{A}_\varepsilon$ has been computed, we compute a surrogate $\overline{C}_\varepsilon$ of C_ε by solving the least-squares problem

$$\inf_{\overline{C}\in(L^2(\mathcal{D}))^{d\times d}} \sum_{r=1}^{R} \left\| \boldsymbol{\nabla}\mathbb{E}\left[u_\varepsilon(f_r)\right] - \overline{C}\ \boldsymbol{\nabla}\overline{u}_\varepsilon(f_r)\right\|^2_{L^2(\mathcal{D})^d} \tag{3.28}$$

for a given number R of right-hand sides, which are selected as in the periodic setting (see Sect. 3.3). Eventually, $\mathbb{E}\left[\boldsymbol{\nabla} u_\varepsilon(\cdot,\omega)\right]$ is approximated by $\overline{C}_\varepsilon\, \boldsymbol{\nabla}\overline{u}_\varepsilon$.

Remark 3.8. Criteria (3.23) and (3.28) are arbitrary and selected upon practical considerations. Among the possible alternatives, we could have considered

$$\Phi^{\text{sto}}_\varepsilon(\overline{A}, f) = \mathbb{E}\left[\left\|(-\Delta)^{-1}\left[\operatorname{div}\left(\overline{A}\boldsymbol{\nabla} u_\varepsilon(f)\right) + f\right]\right\|^2_{L^2(\mathcal{D})}\right]$$

instead of (3.23), and a similar alternative for the reconstruction (3.28).

We have not proceeded in any of these directions. Note also that, in [16], we defined the minimization problems ω by ω and next took the expectation of the results. Of course, considering expectations in the cost functions results in significant computational savings, besides actually improving accuracy and robustness.

4. Implementation details to solve (3.24)

We detail here how problem (3.24), in the stationary ergodic setting, can be efficiently solved in practice. Problem (3.1), in the periodic setting, is actually simpler to solve, and we skip the easy adaptation to that case.

The minimizer of (3.24) is denoted by $\overline{A}^{P,M}_{\varepsilon,h}$, where $h \ll \varepsilon$ denotes the size of a mesh $\mathcal{T}_h = \{T\}$ of the domain $\mathcal{D}$, P denotes the dimension of the subspace $V^P_{\text{n}}(\mathcal{D})$ of $L^2_{\text{n}}(\mathcal{D})$ used to approximate the sup problem (see (3.5)), and $M \in \mathbb{N}^\star$ denotes the number of Monte Carlo realizations used to approximate $\mathbb{E}(u_\varepsilon)$ in (3.23).
The algorithm consists of three steps:

(1) Compute an approximation of $\{\mathbb{E}[u_\varepsilon(f_p)]\}_{1\le p\le P}$ (see Sect. 4.1). This is the most expensive step, as $M \times P$ oscillatory problems of the type (2.1) are to be solved.
(2) Compute an approximation of $(-\Delta)^{-1} f_p$ and of $\left\{(-\Delta)^{-1}\left(\partial_{ij}\mathbb{E}[u_\varepsilon(f_p)]\right)\right\}_{1\le i,j\le d}$, for any $1 \le p \le P$ (see Sect. 4.2). This amounts to solving $P\,(1 + d(d+1)/2)$ problems with constant coefficients.
(3) Solve problem (3.24) iteratively (see Sect. 4.3). Each iteration involves diagonalizing a $P \times P$ matrix and solving a linear system with $d(d+1)/2$ unknowns. The cost of this third step is negligible.

We now successively detail these three steps.

4.1. Approximation of $\{\mathbb{E}[u_\varepsilon(f_p)]\}_{1\le p\le P}$

For any basis function f_p of $V^P_{\text{n}}(\mathcal{D})$, $1 \le p \le P$, we approximate $\mathbb{E}[u_\varepsilon(f_p)]$ by the empirical mean

$$u^M_{\varepsilon,h}(f_p) = \frac{1}{M}\sum_{m=1}^{M} u_{\varepsilon,h}(f_p;\omega_m), \tag{4.1}$$

where, for $1 \leq m \leq M$, $u_{\varepsilon,h}(f_p;\omega_m)$ is the $\mathbb{P}^1$ approximation on $\mathcal{T}_h$ of $u_\varepsilon(f_p;\omega_m)$, unique solution to (2.1) with the oscillatory matrix-valued coefficient $A_\varepsilon(\cdot,\omega_m)$ and the right-hand side f_p.

To compute (4.1) for all $1 \leq p \leq P$, one has to (i) assemble M random stiffness matrices, (ii) assemble P deterministic right-hand sides, and (iii) solve $M \times P$ linear systems. This step is the only one involving Monte Carlo computations, and is therefore the most expensive part of the whole procedure.

4.2. Precomputation of tensorial quantities

Once the computations of Section 4.1 have been performed, we assemble some tensors that are needed to efficiently solve the sup and inf problems involved in (3.24).

We first compute, for any $1 \leq p \leq P$, the approximations $z_h(f_p)$ and $\{z^{M,ij}_{\varepsilon,h}(f_p)\}_{1\leq i,j\leq d}$ on $\mathcal{T}_h$ of $(-\Delta)^{-1}f_p$ and $\left\{(-\Delta)^{-1}\left(\partial_{ij}\mathbb{E}[u_\varepsilon(f_p)]\right)\right\}_{1\leq i,j\leq d}$. In particular, $z^{M,ij}_{\varepsilon,h}(f_p)$ is such that, for any $\mathbb{P}^1$ function w_h on $\mathcal{T}_h$ that vanishes on $\partial\mathcal{D}$,

$$\int_{\mathcal{D}} \boldsymbol{\nabla} z^{M,ij}_{\varepsilon,h}(f_p)\cdot\boldsymbol{\nabla} w_h = -\int_{\mathcal{D}} \partial_j\left[u^M_{\varepsilon,h}(f_p)\right]\,\partial_i w_h.$$

Note that the following symmetry identity holds: $z^{M,ij}_{\varepsilon,h}(f_p) = z^{M,ji}_{\varepsilon,h}(f_p)$.

We next assemble, for all integers $1 \leq i,j,k,l \leq d$ and $1 \leq p,q \leq P$, the quantities

$$\left[\mathcal{K}^M_{\varepsilon,h}\right]_{i,j,k,l,p,q} = 2\int_{\mathcal{D}} z^{M,ij}_{\varepsilon,h}(f_p)\, z^{M,kl}_{\varepsilon,h}(f_q), \tag{4.2}$$

$$\left[\mathbb{K}^M_{\varepsilon,h}\right]_{i,j,p,q} = -\int_{\mathcal{D}} z^{M,ij}_{\varepsilon,h}(f_p)\, z_h(f_q), \tag{4.3}$$

$$\left[K_h\right]_{p,q} = \int_{\mathcal{D}} z_h(f_p)\, z_h(f_q). \tag{4.4}$$

We emphasize that the cost of this step depends on P but is independent of the number M of Monte Carlo realizations, and thus small in comparison to the cost of the operations described in Section 4.1 for typical values of M and P (in the numerical results reported on in Sect. 5, we have worked with $M = 100$ and $P \leq 9$).

4.3. Solution of the fully discrete problem

4.3.1. Formulation

At this stage, the original problem (3.24) has been approximated by its fully discrete version

$$I^{P,M}_{\varepsilon,h} = \inf_{\overline{A}\in\mathcal{S}} \sup_{\boldsymbol{c}\in\mathbb{R}^P,\,|\boldsymbol{c}|^2=1} \Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}), \tag{4.5}$$

where, for any $\overline{A} \in \mathbb{R}^{d\times d}_{\rm sym}$ and $\boldsymbol{c} = \{c_p\}_{1\leq p\leq P} \in \mathbb{R}^P$,

$$\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}) = \left\| \sum_{p=1}^{P} c_p \left(\sum_{1\leq i,j\leq d} \overline{A}_{i,j}\, z^{M,ij}_{\varepsilon,h}(f_p) + z_h(f_p) \right) \right\|^2_{L^2(\mathcal{D})}. \tag{4.6}$$

Problem (4.5) is solved by iteratively considering the problem

$$\sup_{\boldsymbol{c}\in\mathbb{R}^P,\,|\boldsymbol{c}|^2=1} \Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}) \tag{4.7}$$

with $\overline{A} \in \mathcal{S}$ fixed, and the problem

$$\inf_{\overline{A}\in\mathcal{S}} \Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}) \tag{4.8}$$

with $\boldsymbol{c} \in \mathbb{R}^P$ fixed. We successively explain how we solve the sup problem (4.7) (for $\overline{A} \in \mathcal{S}$ fixed), the inf problem (4.8) (for $\boldsymbol{c} \in \mathbb{R}^P$ fixed), and next describe the iterative algorithm that we have implemented to solve (4.5).

4.3.2. The sup *problem* (4.7)

Let $\overline{A} \in \mathcal{S}$ be fixed. One can easily observe that

$$\Phi_{\varepsilon,h}^{P,M}(\overline{A}, \boldsymbol{c}) = \boldsymbol{c}^T \, G_{\varepsilon,h}^{M}(\overline{A}) \, \boldsymbol{c},$$

where $G_{\varepsilon,h}^{M}(\overline{A})$ is a symmetric, positive semi-definite, $P \times P$ matrix which can be assembled at no additional cost using the precomputed quantities defined in (4.2)–(4.4) (see Appendix B for its exact expression). Solving the sup problem (4.7) (with fixed matrix $\overline{A}$) hence amounts to finding a normalized eigenvector in $\mathbb{R}^P$ associated with the largest eigenvalue of the matrix $G_{\varepsilon,h}^{M}(\overline{A})$. This is reminiscent of the eigenvalue problem discussed in Section 3.1.1. Practically, this eigenvector is computed using the power method. The cost of such a computation is negligible, owing to the small size of the matrix $G_{\varepsilon,h}^{M}(\overline{A})$ (recall that P is typically small in comparison to M). We denote by $\boldsymbol{c}(\overline{A})$ its solution and hence have

$$\sup_{\boldsymbol{c}\in\mathbb{R}^P,\, |\boldsymbol{c}|^2=1} \Phi_{\varepsilon,h}^{P,M}(\overline{A}, \boldsymbol{c}) = \boldsymbol{c}(\overline{A})^T \, G_{\varepsilon,h}^{M}(\overline{A}) \, \boldsymbol{c}(\overline{A}). \tag{4.9}$$

4.3.3. The inf *problem* (4.8)

Let $\boldsymbol{c} \in \mathbb{R}^P$, $|\boldsymbol{c}|^2 = 1$, be fixed. We observe that

$$\Phi_{\varepsilon,h}^{P,M}(\overline{A}, \boldsymbol{c}) = \frac{1}{2} \sum_{1\le i,j,k,l\le d} \left[\mathbb{B}_{\varepsilon,h}^{P,M}(\boldsymbol{c})\right]_{i,j,k,l} \overline{A}_{i,j} \, \overline{A}_{k,l} - \sum_{1\le i,j\le d} \left[B_{\varepsilon,h}^{P,M}(\boldsymbol{c})\right]_{i,j} \overline{A}_{i,j} + b_h^P(\boldsymbol{c}),$$

where $\mathbb{B}_{\varepsilon,h}^{P,M}(\boldsymbol{c})$ is a $d \times d \times d \times d$ fourth-order tensor, $B_{\varepsilon,h}^{P,M}(\boldsymbol{c})$ is a $d \times d$ matrix and $b_h^P(\boldsymbol{c})$ is a scalar that can all be assembled at no additional cost using the precomputed quantities defined in (4.2)–(4.4) (see Appendix B for their exact expressions). We recognize in $\Phi_{\varepsilon,h}^{P,M}$ the discrete equivalent of (3.9). The inf problem (4.8) (with fixed eigenvector $\boldsymbol{c}$) is in practice solved as explained in Section 3.1.2, by considering the linear system (see (3.10))

$$\forall\, 1 \le i,j \le d, \qquad \sum_{1\le k,l\le d} \left[\mathbb{B}_{\varepsilon,h}^{P,M}(\boldsymbol{c})\right]_{i,j,k,l} \overline{A}_{k,l} = \left[B_{\varepsilon,h}^{P,M}(\boldsymbol{c})\right]_{i,j}. \tag{4.10}$$

4.3.4. Iterative algorithm

In the above description, we have considered either the sup problem (on $\boldsymbol{c}$, with fixed $\overline{A}$) or the inf problem (on $\overline{A}$, for fixed $\boldsymbol{c}$) involved in (4.5). We now assemble these two building blocks to build an algorithm to solve (4.5). Introducing

$$\Phi_{\varepsilon,h}^{P,M}(\overline{A}) = \sup_{\boldsymbol{c}\in\mathbb{R}^P,\, |\boldsymbol{c}|^2=1} \Phi_{\varepsilon,h}^{P,M}(\overline{A}, \boldsymbol{c}), \tag{4.11}$$

we recast (4.5) as

$$I_{\varepsilon,h}^{P,M} = \inf_{\overline{A}\in\mathcal{S}} \Phi_{\varepsilon,h}^{P,M}(\overline{A}). \tag{4.12}$$

We have seen (see (4.9)) that $\Phi_{\varepsilon,h}^{P,M}(\overline{A}) = \boldsymbol{c}(\overline{A})^T \, G_{\varepsilon,h}^{M}(\overline{A}) \, \boldsymbol{c}(\overline{A})$, where $\boldsymbol{c}(\overline{A})$ is an eigenvector of the matrix $G_{\varepsilon,h}^{M}(\overline{A})$. One can easily prove that, for any $1 \le i,j \le d$,

$$\left[\boldsymbol{\nabla}_{\overline{A}} \Phi_{\varepsilon,h}^{P,M}(\overline{A})\right]_{i,j} = \boldsymbol{c}(\overline{A})^T \, \partial_{\overline{A}_{i,j}} G_{\varepsilon,h}^{M}(\overline{A}) \, \boldsymbol{c}(\overline{A}),$$

which reads, using the expressions (B.1), (B.2) and (B.3) of $G_{\varepsilon,h}^{M}$, $\mathbb{B}_{\varepsilon,h}^{P,M}$ and $B_{\varepsilon,h}^{P,M}$ given in Appendix B, as

$$\left[\boldsymbol{\nabla}_{\overline{A}} \Phi_{\varepsilon,h}^{P,M}(\overline{A})\right]_{i,j} = \sum_{1\le k,l\le d} \left[\mathbb{B}_{\varepsilon,h}^{P,M}(\boldsymbol{c}(\overline{A}))\right]_{i,j,k,l} \overline{A}_{k,l} - \left[B_{\varepsilon,h}^{P,M}(\boldsymbol{c}(\overline{A}))\right]_{i,j}. \tag{4.13}$$

Let $0 < \mu < 1$. In practice, we iterate as follows to solve problem (4.12). Let $n \in \mathbb{N}$ and $\overline{A}^n \in \mathcal{S}$.

(1) We compute $\boldsymbol{c}^n = \boldsymbol{c}(\overline{A}^n)$ solution to the sup problem (4.11) with fixed matrix $\overline{A}^n$.
(2) We compute $\overline{A}_\flat^{n+1} \in \mathbb{R}^{d\times d}_{\rm sym}$ solution to the linear system (4.10) with fixed eigenvector $\boldsymbol{c}^n$. As pointed out above, we assume that $\overline{A}_\flat^{n+1}$ belongs to the convex subset $\mathcal{S}$ of $\mathbb{R}^{d\times d}_{\rm sym}$. It has always been the case in our numerical experiments.
(3) We define the next iterate as

$$\overline{A}^{n+1} = (1-\mu)\,\overline{A}^n + \mu\,\overline{A}_\flat^{n+1}. \tag{4.14}$$

For the numerical results reported on in Section 5, we have worked with $\mu \leq 0.1$.

Since $\overline{A}^{n+1}$ is a convex combination of $\overline{A}^n \in \mathcal{S}$ and $\overline{A}_\flat^{n+1} \in \mathcal{S}$, we have $\overline{A}^{n+1} \in \mathcal{S}$. The iterations are initialized using, say,

$$\overline{A}^0 = \mathbb{E}\left(\frac{1}{|\mathcal{D}|}\int_{\mathcal{D}} A_\varepsilon(\boldsymbol{x},\cdot)\,\mathrm{d}\boldsymbol{x}\right).$$

Let us briefly explain, at least formally, why the algorithm defined above enables to find a minimizer of (4.12). We assume the linear system (4.10) to be invertible, and we denote by $\left[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\right]^{-1}$ its formal inverse. Since $\overline{A}_\flat^{n+1}$ is defined as the solution to (4.10) with eigenvector $\boldsymbol{c}^n$, we infer from (4.10) and (4.13) that

$$\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c}^n)\overline{A}_\flat^{n+1} = B^{P,M}_{\varepsilon,h}(\boldsymbol{c}^n) = \mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c}^n)\overline{A}^n - \boldsymbol{\nabla}_{\overline{A}}\Phi^{P,M}_{\varepsilon,h}(\overline{A}^n),$$

and thus

$$\overline{A}_\flat^{n+1} = \overline{A}^n - \left[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c}^n)\right]^{-1}\boldsymbol{\nabla}_{\overline{A}}\Phi^{P,M}_{\varepsilon,h}(\overline{A}^n).$$

The iteration (4.14) can be recast under the form

$$\overline{A}^{n+1} = \overline{A}^n - \mu\,\left[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c}^n)\right]^{-1}\boldsymbol{\nabla}_{\overline{A}}\Phi^{P,M}_{\varepsilon,h}(\overline{A}^n).$$

This is a quasi-Newton algorithm for the minimization of the function $\overline{A} \mapsto \Phi^{P,M}_{\varepsilon,h}(\overline{A})$, with a fixed step size μ and where the Hessian of $\Phi^{P,M}_{\varepsilon,h}$ with respect to $\overline{A}$ is approximated by $\mathbb{B}^{P,M}_{\varepsilon,h}$.

Note that each iteration of the algorithm is inexpensive in comparison with the cost of the operations described in Sections 4.1 and 4.2. Consequently, there is no real advantage in improving the optimization algorithm (4.14) (*e.g.* by optimizing the value of μ by a line search).

5. Numerical results

As pointed out in Section 1, our approach targets practical situations where the information on the oscillatory coefficients in the equation may be incomplete, and thus the other available approaches cannot be applied. It is nevertheless a legitimate question to investigate how our approach performs on standard test-cases in the periodic and stationary ergodic settings, and how it compares with the classical homogenization approach for small values of ε. As already pointed out in Section 1.3, and as detailed below (see Sect. 5.2.1), the aim of the numerical tests is different in the periodic setting and in the stochastic setting. It is also different if ε is asymptotically small or if ε takes larger values.

This section is organized as follows. In Section 5.1, we introduce the periodic and the stationary ergodic test cases considered. In Section 5.2, we present the numerical results obtained in the case of small values of ε. In Section 5.3, we address the case of larger values of ε.

5.1. Test-cases

We let $d = 2$ and the domain $\mathcal{D}$ be the unit square $(0,1)^2$. We fix the value of the parameter $\overline{\varepsilon}$ to $\mathrm{size}(\mathcal{D})/10 = 10^{-1}$.

5.1.1. Periodic setting

We consider the test-case introduced in [16], namely

$$A_\varepsilon(x,y) = A^{\mathrm{per}}(x/\varepsilon, y/\varepsilon), \tag{5.1}$$

with A^{per} a $\mathbb{Z}^2$-periodic symmetric matrix field given by

$$\begin{aligned}
[A^{\mathrm{per}}(x,y)]_{1,1} &= 2 + \frac{1}{2\pi}\left(\sin(2\pi x) + \sin(2\pi y)\right), \\
[A^{\mathrm{per}}(x,y)]_{1,2} &= \frac{1}{2\pi}\left(\sin(2\pi x) + \sin(2\pi y)\right), \\
[A^{\mathrm{per}}(x,y)]_{2,2} &= 1 + \frac{1}{2\pi}\left(\sin(2\pi x) + \sin(2\pi y)\right).
\end{aligned} \tag{5.2}$$

The coefficients of the corresponding homogenized matrix (obtained by solving the periodic corrector problem (2.8) on a very fine mesh) are

$$[A_\star]_{1,1} \approx 1.9806, \qquad [A_\star]_{1,2} = [A_\star]_{2,1} \approx -0.019345, \qquad [A_\star]_{2,2} \approx 0.98065. \tag{5.3}$$

5.1.2. Stationary ergodic setting

We consider the random checkerboard test-case (studied *e.g.* in [16]), namely

$$A_\varepsilon(x,y,\omega) = a^{\mathrm{sto}}(x/\varepsilon, y/\varepsilon, \omega)\,\mathrm{Id}_2, \tag{5.4}$$

with a^{sto} a discrete stationary field given by (recall that $Q = (0,1)^2$)

$$a^{\mathrm{sto}}(x,y,\omega) = \sum_{\boldsymbol{k}\in\mathbb{Z}^2} \mathbb{1}_{Q+\boldsymbol{k}}(x,y) X_{\boldsymbol{k}}(\omega), \tag{5.5}$$

where the random variables $X_{\boldsymbol{k}}$ are i.i.d. and such that $\mathbb{P}(X_{\boldsymbol{k}} = 4) = \mathbb{P}(X_{\boldsymbol{k}} = 16) = 1/2$. An explicit expression for the homogenized matrix is known in that case:

$$A_\star = 8\,\mathrm{Id}_2. \tag{5.6}$$

5.2. Results in the case $\varepsilon < \bar{\varepsilon}$

5.2.1. Objectives in the periodic case and in the stochastic case

In the regime $\varepsilon < \bar{\varepsilon}$, we know from Proposition 3.2 that our method can be seen as a practical variational approach for computing the homogenized matrix $A_\star$. The remaining question is whether this approach is efficient or not, and particularly, compared with the classical approach in homogenization.

Our approach (based on (3.1)–(3.2)) requires solving the *highly oscillatory* equations (1.1) set on the domain $\mathcal{D}$, for $P = d(d+1)/2$ right-hand sides. In the periodic setting, the classical homogenization approach requires solving d *non-oscillatory* equations set on the unit cell Q. There is thus no hope to outperform the latter approach in terms of computational time. This setting is nonetheless considered as a validation and we investigate how our approach performs in terms of accuracy, for the approximation of the homogenized matrix, and for the approximation of u_ε in the L^2 and H^1 norms.

The real, discriminating, test-case for our approach is the stationary ergodic setting. Indeed, classical homogenization then requires solving equations that are set on a truncated approximation $Q^N = (-N,N)^d$ of an asymptotically infinitely large domain (see (2.9) in Sect. 2.2). The coefficients of these equations vary at scale 1. In that case, to hope for an accurate approximation of the homogenized matrix, one has to consider a meshsize $H \ll 1$. On the other hand, we consider a meshsize $h \ll \varepsilon$ to solve the highly oscillatory equations (set on the domain $\mathcal{D}$) involved in our approach. We see that, up to an appropriate choice of the parameter H such that

$$\frac{2N}{H} = \frac{\mathrm{size}(\mathcal{D})}{h}, \tag{5.7}$$

where size($\mathcal{D}$) is typically the diameter of $\mathcal{D}$, the classical homogenization approach and ours involve solving linear systems of the same size. The computational workload for the two approaches is thus of the same order of magnitude, although not identical. We have decided to enforce (5.7) and to relate N in (2.9) and ε in (2.1) by

$$N = \text{size}(\mathcal{D})/2\varepsilon. \tag{5.8}$$

Note that imposing (5.8) is equivalent to enforcing $\varepsilon/h = 1/H$. We then compare the two methods in terms of solution time and accuracy. Obviously, for the two methods, the same number M of Monte Carlo realizations is used, and the same M realizations are considered.

Remark 5.1. Another possibility would have been to impose $\varepsilon/h = 1/H$ and to adjust the size N of Q^N in (2.9) so that both approaches exactly share the same workload. We did not pursue in that direction.

The numerical experiments reported in Section 5.2.4 show that, in the stochastic case, and for all the values of $\varepsilon < \overline{\varepsilon}$ that have been considered, the approximation of $A_\star$ obtained by the classical homogenization approach is slightly more accurate than that obtained with our approach. In contrast, our approach provides a better L^2-approximation and a better H^1-approximation of $\mathbb{E}(u_\varepsilon)$. This is somewhat intuitive, as our approach is targeted toward the approximation of u_ε rather than $A_\star$. In terms of computational cost, our approach is slightly less expensive for moderately small values of ε, and slightly more expensive for asymptotically small values of ε (in any cases, the ratio of costs remains close to 1, see Fig. 2 below).

5.2.2. Choice of the numerical parameters

We recall that the integer M denotes the number of i.i.d. realizations used to approximate the expectation in the cost function (3.23) (see (4.1)). We also recall that the integer P denotes the dimension of the set $V_{\rm n}^P(\mathcal{D})$ (defined in (3.5)) that is used to approximate the space $L^2_{\rm n}(\mathcal{D})$ in the sup problem. As explained in Section 3.1.1, we consider as basis functions of the set $V_{\rm n}^P(\mathcal{D})$ the first P normalized eigenvectors of the laplacian operator in the domain $\mathcal{D}$. Because of the simple geometry of $\mathcal{D}$, they are here analytically known. We take here $P = d\,(d+1)/2$, that is $P = 3$, which is the minimum dimension of the search space $V_{\rm n}^P(\mathcal{D})$.

5.2.3. Results in the periodic setting

We consider the parameters $\{\varepsilon_k\}_{0\le k\le 6}$ such that $\varepsilon_0 = 0.4$ and $\varepsilon_k = \varepsilon_{k-1}/2$ for $1 \le k \le 6$. The associated meshsizes are $\{h_k\}_{0\le k\le 6}$ such that $h_k = \varepsilon_k/r$ for $r \approx 43$, unless otherwise mentioned. We focus on the values $\{\varepsilon_k\}_{3\le k\le 6}$, for which we have $\varepsilon_k < \overline{\varepsilon}$.

The error in the approximation of the homogenized matrix is defined by

$$\texttt{err\_per\_mat} = \left(\frac{\sum_{1\le i,j\le d} \left| \left[\overline{A}^P_{\varepsilon,h}\right]_{i,j} - [A_\star]_{i,j} \right|^2}{\sum_{1\le i,j\le d} |[A_\star]_{i,j}|^2} \right)^{1/2}, \tag{5.9}$$

where $A_\star$ is taken equal to its reference value (5.3) and $\overline{A}^P_{\varepsilon,h}$ is the best matrix computed by our approach. The numerical results are collected in Table 1. We observe that our approach provides an accurate approximation of the homogenized matrix. The accuracy of the approximation improves (in the limit of spatial resolution) as ε decreases.

We now examine the approximation of u_ε in the L^2 norm. We denote by:

- $u_{\varepsilon,h}(f)$ the discrete solution to (1.1) with the periodic oscillatory coefficient given by (5.1)–(5.2) and the right-hand side f;
- $u_{\star,h}(f)$ the discrete solution to (1.2) with the homogenized matrix (5.3) and the right-hand side f;

TABLE 1. Approximation of $A_\star$ (`err_per_mat`) in function of ε (each line corresponds to a different value of the ratio ε/h). The test cases with ε too small and ε/h too large are prohibitively expensive to perform. They are marked with an X.

ε	0.05	0.025	0.0125	0.00625
`err_per_mat` ($\varepsilon/h \approx 43$)	1.0145×10^{-3}	7.6477×10^{-4}	6.6613×10^{-4}	6.2881×10^{-4}
`err_per_mat` ($\varepsilon/h \approx 86$)	6.5399×10^{-4}	3.5074×10^{-4}	2.3749×10^{-4}	X

- $u_{\varepsilon,h}^{1,\boldsymbol{\theta}}(f)$ the two-scale expansion (truncated at first-order) built from $u_{\star,h}(f)$ (see (3.17)), where we use the periodic correctors solution to (2.8);
- $\overline{u}_{\varepsilon,h}^{P}(f)$ the discrete solution to (1.3) with the matrix $\overline{A}_{\varepsilon,h}^{P}$ and the right-hand side f (we recall that the matrix $\overline{A}_{\varepsilon,h}^{P}$ has been computed using a small number P of right-hand sides).

To assess the quality of the approximation of $u_{\varepsilon,h}$ by $\widehat{u}_h^{\boldsymbol{\theta}} \in \left\{u_{\star,h},\ u_{\varepsilon,h}^{1,\boldsymbol{\theta}},\ \overline{u}_{\varepsilon,h}^{P}\right\}$ in the L^2 norm, we define the criterion

$$\texttt{err\_per\_L2} = \left(\frac{\inf\limits_{\boldsymbol{\theta}\in\mathbb{R}^2} \left[\sup\limits_{f\in V_{\rm n}^{\mathcal{Q}}(\mathcal{D})} \left\| u_{\varepsilon,h}(f) - \widehat{u}_h^{\boldsymbol{\theta}}(f) \right\|_{L^2(\mathcal{D})}^2 \right]}{\left\| u_{\varepsilon,h}\left(\widehat{f}_\varepsilon\right) \right\|_{L^2(\mathcal{D})}^2} \right)^{1/2}. \tag{5.10}$$

Note that the supremum is taken over $f \in V_{\rm n}^{\mathcal{Q}}(\mathcal{D})$, where $\mathcal{Q} \gg P$. We take $\mathcal{Q} = 16$, and we have checked, in all the cases considered below, that our results do not significantly change for a larger value of $\mathcal{Q}$. The function $\widehat{f}_\varepsilon \in V_{\rm n}^{\mathcal{Q}}(\mathcal{D})$ denotes the argument of the inf sup problem in the numerator of (5.10). We hence compare u_ε with its homogenized limit $u_\star$, its first-order two-scale expansion $u_\varepsilon^{1,\boldsymbol{\theta}}$ (recall in this case that the correctors are defined up to an additive constant $\boldsymbol{\theta}$, over which we minimize the error in (5.10)), and the approximation $\overline{u}_\varepsilon^P$ provided by our approach. The numerical results are collected in Figure 1.

We observe that the solution associated with the best matrix we compute indeed converges towards the exact solution, in the L^2 norm. We however recall that, in the present periodic setting, computing $\overline{u}_{\varepsilon,h}^P$ is much more expensive than computing $u_{\star,h}$ or $u_{\varepsilon,h}^{1,\boldsymbol{\theta}}$.

We next examine the H^1 error. For $f \in L^2(\mathcal{D})$, we denote by $C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h}(f)$ the discrete equivalent of $C_\varepsilon \boldsymbol{\nabla} u_\star(f)$, the homogenization-based approximation of $\boldsymbol{\nabla} u_\varepsilon(f)$, see (3.19)–(3.20) in Section 3.3. We recall that, in our approach, we seek an approximation of $\boldsymbol{\nabla} u_\varepsilon(f)$ under the form $\overline{C}_\varepsilon \boldsymbol{\nabla} \overline{u}_\varepsilon(f)$ (see (3.21)), the discrete equivalent of which is computed as $\overline{C}_{\varepsilon,h}^R \boldsymbol{\nabla} \overline{u}_{\varepsilon,h}^P(f)$. Recall that the integer R is the number of right-hand sides used to define the least-squares minimization problem (3.22) giving $\overline{C}_{\varepsilon,h}^R$. Here, we take $R = P = 3$. To assess the quality of the approximation of $\boldsymbol{\nabla} u_{\varepsilon,h}$, we define, for $\widehat{C}_{\varepsilon,h}\boldsymbol{\nabla}\widehat{u}_h \in \{C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h},\ \overline{C}_{\varepsilon,h}^R \boldsymbol{\nabla}\overline{u}_{\varepsilon,h}^P\}$, the criterion

$$\texttt{err\_per\_H1} = \left(\frac{\sup\limits_{f\in V_{\rm n}^{\mathcal{Q}}(\mathcal{D})} \left\| \boldsymbol{\nabla} u_{\varepsilon,h}(f) - \widehat{C}_{\varepsilon,h}\boldsymbol{\nabla}\widehat{u}_h(f) \right\|_{L^2(\mathcal{D}\setminus\mathcal{B})}^2}{\left\| \boldsymbol{\nabla} u_{\varepsilon,h}\left(\widehat{f}_\varepsilon\right) \right\|_{L^2(\mathcal{D}\setminus\mathcal{B})}^2} \right)^{1/2}, \tag{5.11}$$

where, here again, the supremum is taken over a space $V_{\rm n}^{\mathcal{Q}}(\mathcal{D})$ much larger than $V_{\rm n}^P(\mathcal{D})$ (we take $\mathcal{Q} = 16$), and where $\widehat{f}_\varepsilon \in V_{\rm n}^{\mathcal{Q}}(\mathcal{D})$ denotes the argument of the sup problem. In (5.11), $\mathcal{B}$ represents the subset of $\mathcal{D}$ formed by the boundary elements of the discretization $\mathcal{T}_h$. We remove them in view of the discussion below (3.22). We thus compare $\boldsymbol{\nabla} u_\varepsilon$ with its approximation $C_\varepsilon \boldsymbol{\nabla} u_\star$ provided by the two-scale expansion and with the approximation $\overline{C}_\varepsilon^R\ \boldsymbol{\nabla}\overline{u}_\varepsilon^P$ provided by our approach. The numerical results are collected in Table 2.

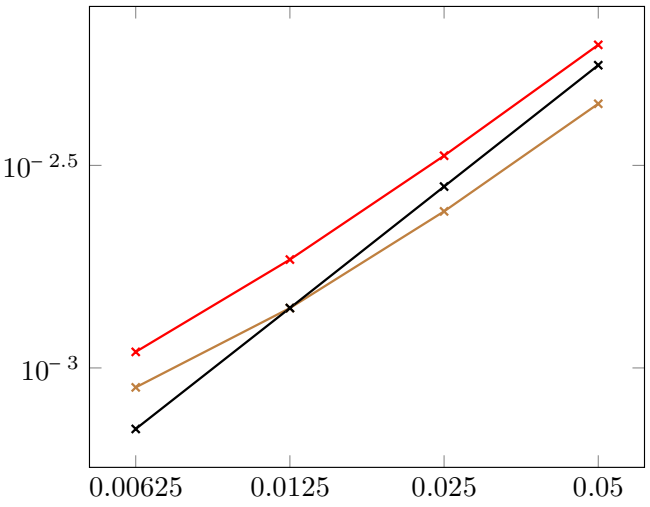

FIGURE 1. Approximation of u_ε in the L^2 norm (`err_per_L2`) by $u_{\star,h}$ (*red*), $u^{1,\boldsymbol{\theta}}_{\varepsilon,h}$ (*brown*) and $\overline{u}^P_{\varepsilon,h}$ (*black*) in function of ε, for h such that $\varepsilon/h \approx 43$.

TABLE 2. Approximation of $\boldsymbol{\nabla} u_\varepsilon$ in the L^2 norm (`err_per_H1`) by $C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h}$ and $\overline{C}^R_{\varepsilon,h}\boldsymbol{\nabla}\overline{u}^P_{\varepsilon,h}$ in function of ε, for h such that $\varepsilon/h \approx 86$. The test cases with ε too small are prohibitively expensive to perform. They are marked with an X.

ε	0.05	0.025	0.0125	0.00625
`err_per_H1` for $C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h}$	2.0906×10^{-2}	1.6461×10^{-2}	1.2513×10^{-2}	X
`err_per_H1` for $\overline{C}^R_{\varepsilon,h}\boldsymbol{\nabla}\overline{u}^P_{\varepsilon,h}$	1.5550×10^{-2}	7.6055×10^{-3}	3.7549×10^{-3}	X

We observe that our approach provides an accurate H^1-approximation of u_ε. As ε goes to zero, the surrogate we compute is (roughly) a first-order convergent approximation of $\boldsymbol{\nabla} u_\varepsilon$ in the L^2 norm. As far as the homogenization-based approximation is concerned, we expect it to converge with order at least one half (see (3.18)). This is what we observe in practice, as long as ε is not too small. Otherwise, the error due to the meshsize dominates, and the error (5.11) does not decrease anymore when ε decreases.

5.2.4. Results in the stationary ergodic setting

We consider the parameters $\{\varepsilon_k\}_{0\le k\le 5}$ such that $\varepsilon_k = 2^{-(k+1)}$ for $0 \le k \le 5$. In agreement with formula (5.8), we couple these parameters to the parameters $\{N_k\}_{0\le k\le 5}$ (defining the domain on which we solve the corrector problems (2.9)) such that $N_k = 2^k$. The associated meshsizes $\{h_k\}_{0\le k\le 5}$ and $\{H_k\}_{0\le k\le 5}$ are computed respectively letting $h_k = \varepsilon_k/r$ for $r \approx 27$ (unless otherwise stated) and using (5.7). We focus on the values $\{\varepsilon_k\}_{3\le k\le 5}$ and $\{N_k\}_{3\le k\le 5}$, for which we have $\varepsilon_k < \overline{\varepsilon}$. We consider $M = 100$ Monte Carlo realizations.

Before discussing the accuracy of our approach, we first compare its cost with that of the classical approach. We show in Figure 2 the ratio of the time needed to compute $\overline{A}^{P,M}_{\varepsilon,h}$ using our approach divided by the time needed to compute $A^{N,M}_{\star,H}$ by the classical homogenization approach. To compare the computational times, we make use of an implementation that does not exploit parallelism, and we solve the linear systems by means of an iterative solver. In view of Figure 2, for the choice of parameters discussed in Section 5.2.1, our method is slightly faster than the standard homogenization approach for values of N up to approximately 14. This observation can be explained as follows. For the number $M = 100$ of Monte Carlo realizations that we consider, we can neglect, in our procedure, the cost of the precomputation and final optimization stages, in comparison to the

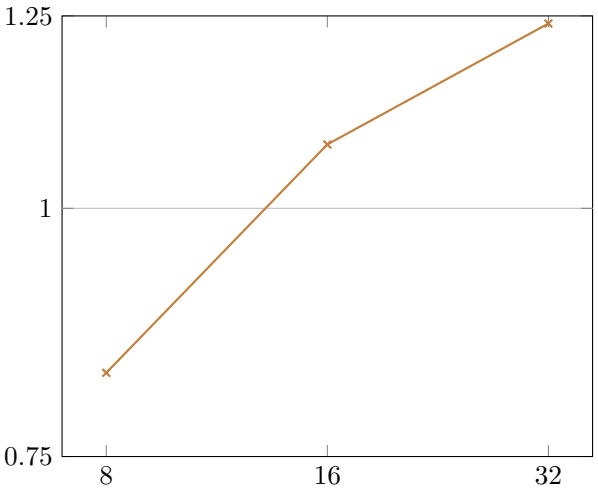

FIGURE 2. Ratio of the computational times between our approach and the classical homogenization approach, in function of N (here $M = 100$ and $\varepsilon/h \approx 27$).

Monte Carlo step (see Sect. 4). Hence, to compute $\overline{A}_{\varepsilon,h}^{P,M}$, we have to (i) assemble $M = 100$ stiffness matrices, (ii) assemble $P = 3$ right-hand sides, and (iii) solve $P \times M = 300$ linear systems. In contrast, to compute $A_{\star,H}^{N,M}$, one has to solve $d \times M = 200$ approximate corrector equations (2.9), that is to say (i) assemble $M = 100$ stiffness matrices, (ii) assemble $d \times M = 200$ right-hand sides, and (iii) solve $d \times M = 200$ linear systems. Consequently, our approach necessitates solving 100 more linear systems, but assembling 200 less right-hand sides, than the classical homogenization approach. This explains what we observe. When the value of N is not too large, the assembly cost is higher than the inversion cost, and our approach is faster.

We adapt to the stationary ergodic setting the accuracy criteria (5.9), (5.10) and (5.11) introduced in the periodic setting. The error in the approximation of the homogenized matrix is defined, for $\widehat{A}^M \in \left\{A_{\star,H}^{N,M}, \overline{A}_{\varepsilon,h}^{P,M}\right\}$, by

$$\texttt{err\_sto\_mat} = \left(\frac{\sum_{1\le i,j\le d} \left| \left[\widehat{A}^M\right]_{i,j} - [A_\star]_{i,j} \right|^2}{\sum_{1\le i,j\le d} |[A_\star]_{i,j}|^2} \right)^{1/2},$$

where $A_\star$ is taken equal to the exact value (5.6). We recall that $A_{\star,H}^{N,M}$ is the practical approximation of $A_\star^{N,M}$ defined in (2.11), and that our approach consists in computing the best matrix $\overline{A}_{\varepsilon,h}^{P,M}$ following the procedure described in Section 3.4.

The numerical results are collected in Figure 3, for several choices of the meshsizes. We observe that the matrix we compute converges to the homogenized matrix as N increases. However, for any value of N in the range we consider, the approximation of $A_\star$ obtained by the classical homogenization approach is slightly more accurate than the one obtained with our approach. As shown in Figure 2, the former approach is as expensive as our approach for $N \approx 14$, and slightly less expensive for larger values of N.

Turning to the approximation of $\mathbb{E}(u_\varepsilon)$ in the L^2 norm, we denote by

- $u_{\varepsilon,h}^M(f)$ the expectation, as defined in (4.1), of the discrete solutions to (2.1) with the oscillatory coefficients given by (5.4)–(5.5) and the right-hand side f;
- $u_{\star,h}(f)$ the discrete solution to (1.2) with the exact homogenized matrix (5.6) and the right-hand side f (note that the exact matrix is usually unknown);

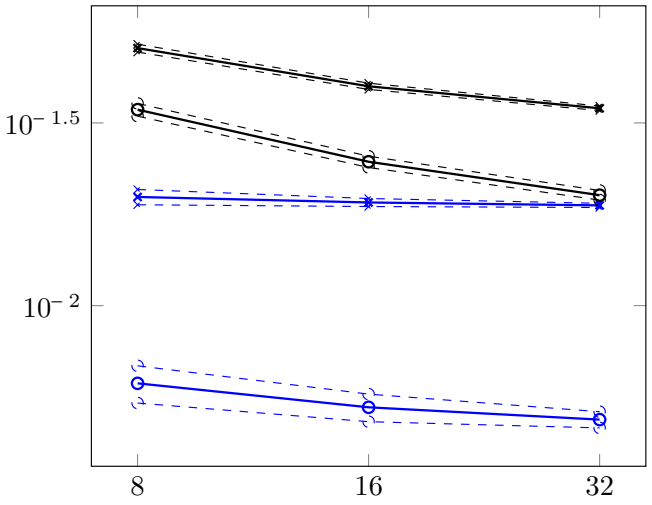

FIGURE 3. Approximation of $A_\star$ by the classical homogenization approach (*blue*) and by our approach (*black*) in function of N, for $M = 100$ realizations. Since M is finite, the error `err_sto_mat` is actually random. We compute it 100 times. The thick line corresponds to the mean value over the 100 computations of the error. The dashed lines show the 95% confidence interval. Results obtained with h such that $\varepsilon/h \approx 27$ (resp. $\varepsilon/h \approx 108$) are denoted with `x` (resp. `o`).

- $u^{N,M}_{\star,h}(f)$ the discrete solution to (1.2) with the matrix $A^{N,M}_{\star,H}$ and the right-hand side f;
- $\overline{u}^{P,M}_{\varepsilon,h}(f)$ the discrete solution to (1.3) with the matrix $\overline{A}^{P,M}_{\varepsilon,h}$ and the right-hand side f.

The M realizations of the field $A(\cdot,\omega)$ we consider to compute $u^M_{\varepsilon,h}(f)$, $u^{N,M}_{\star,h}(f)$ and $\overline{u}^{P,M}_{\varepsilon,h}(f)$ are identical.

To assess the quality of the approximation of $u^M_{\varepsilon,h}$ by $\widehat{u}_h \in \{u_{\star,h},\ u^{N,M}_{\star,h},\ \overline{u}^{P,M}_{\varepsilon,h}\}$ in the L^2 norm, we define the criterion

$$\texttt{err\_sto\_L2} = \left(\frac{\sup\limits_{f \in V^{\mathcal{Q}}_{\rm n}(\mathcal{D})} \left\| u^M_{\varepsilon,h}(f) - \widehat{u}_h(f) \right\|^2_{L^2(\mathcal{D})}}{\left\| u^M_{\varepsilon,h}\left(\widehat{f}_\varepsilon\right) \right\|^2_{L^2(\mathcal{D})}} \right)^{1/2}. \tag{5.12}$$

As in the periodic case, the supremum is taken over $f \in V^{\mathcal{Q}}_{\rm n}(\mathcal{D})$ with $\mathcal{Q} = 16 \gg P$, and $\widehat{f}_\varepsilon \in V^{\mathcal{Q}}_{\rm n}(\mathcal{D})$ denotes the argument of the sup problem. The numerical results are collected in Figure 4, for several choices of the meshsizes and of the total number M of realizations.

We observe that the solution associated with the best matrix we compute is a better L^2-approximation (for the range of parameters considered here) of $\mathbb{E}(u_\varepsilon)$ than the solutions associated with the exact or approximate homogenized matrices. Again, due to the small number P of right-hand sides we consider to compute $\overline{A}^{P,M}_{\varepsilon,h}$, this good accuracy is not an immediate consequence of our practical procedure (it would have been if we had taken P extremely large). We also observe that the accuracy of the three approximations $u_{\star,h}$, $u^{N,M}_{\star,h}$ and $\overline{u}^{P,M}_{\varepsilon,h}$ improves when h decreases or when M increases, in somewhat a complex manner. In terms of cost, our approach is again less expensive than the classical approach for $N \leq 14$.

We next turn to the H^1-error. We denote by $C^{N,M}_{\varepsilon,h}$ the approximation of the deterministic matrix C_ε defined by (3.27) by an empirical mean over M realizations of the corrector functions, solution to (2.9):

$$\left[C^{N,M}_{\varepsilon,h}\right]_{i,j} = \delta_{ij} + \frac{1}{M}\sum_{m=1}^{M} \partial_i w^N_{\boldsymbol{e}_j}(\cdot/\varepsilon, \omega_m).$$

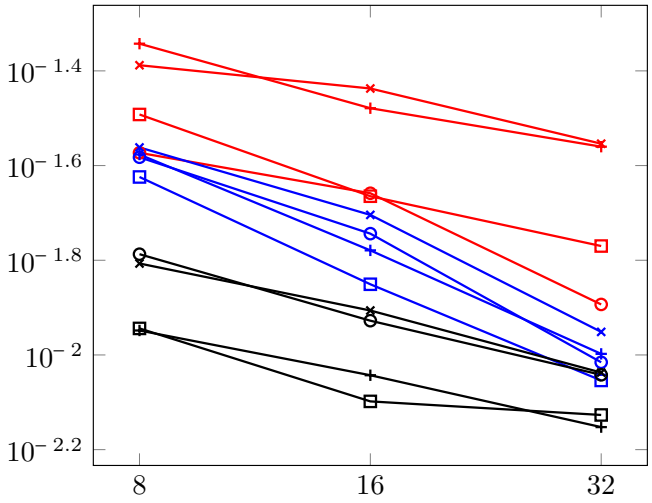

FIGURE 4. Approximation of $\mathbb{E}(u_\varepsilon)$ in the L^2 norm (`err_sto_L2`) by $u_{\star,h}$ (*red*), $u^{N,M}_{\star,h}$ (*blue*) and $\overline{u}^{P,M}_{\varepsilon,h}$ (*black*) in function of N (curves with `x`: $\varepsilon/h \approx 27$ and $M = 100$; curves with `o`: $\varepsilon/h \approx 108$ and $M = 100$; curves with `+`: $\varepsilon/h \approx 27$ and $M = 400$; curves with $\square$: $\varepsilon/h \approx 54$ and $M = 400$).

For $f \in L^2(\mathcal{D})$, we denote by $C^{N,M}_{\varepsilon,h} \boldsymbol{\nabla} u_{\star,h}(f)$ and $C^{N,M}_{\varepsilon,h} \boldsymbol{\nabla} u^{N,M}_{\star,h}(f)$ the two discrete equivalents of $C_\varepsilon \, \boldsymbol{\nabla} u_\star(f)$, the homogenization-based approximation of $\mathbb{E}\left(\boldsymbol{\nabla} u_\varepsilon(f)\right)$, obtained by using the exact homogenized matrix (5.6) and the matrix $A^{N,M}_{\star,H}$, respectively, to compute an approximation of $u_\star(f)$. In our approach, we seek a discrete approximation of $\mathbb{E}\left(\boldsymbol{\nabla} u_\varepsilon\right)$ under the form $\overline{C}^{R,M}_{\varepsilon,h} \boldsymbol{\nabla} \overline{u}^{P,M}_{\varepsilon,h}$, with $R = P = 3$. For

$$\widehat{C}^{M}_{\varepsilon,h} \, \boldsymbol{\nabla} \widehat{u}_h \in \left\{ C^{N,M}_{\varepsilon,h} \, \boldsymbol{\nabla} u_{\star,h}, \; C^{N,M}_{\varepsilon,h} \, \boldsymbol{\nabla} u^{N,M}_{\star,h}, \; \overline{C}^{R,M}_{\varepsilon,h} \, \boldsymbol{\nabla} \overline{u}^{P,M}_{\varepsilon,h} \right\},$$

we define the criterion

$$\texttt{err\_sto\_H1} = \left(\frac{\sup\limits_{f \in V^{\mathcal{Q}}_{\mathrm{n}}(\mathcal{D})} \left\| \boldsymbol{\nabla} u^{M}_{\varepsilon,h}(f) - \widehat{C}^{M}_{\varepsilon,h} \, \boldsymbol{\nabla} \widehat{u}_h(f) \right\|^2_{L^2(\mathcal{D}\setminus\mathcal{B})}}{\left\| \boldsymbol{\nabla} u^{M}_{\varepsilon,h}\left(\widehat{f}_\varepsilon\right) \right\|^2_{L^2(\mathcal{D}\setminus\mathcal{B})}} \right)^{1/2}, \tag{5.13}$$

where, here again, the supremum is taken over the space $V^{\mathcal{Q}}_{\mathrm{n}}(\mathcal{D})$ for $\mathcal{Q} = 16 \gg P$, $\widehat{f}_\varepsilon \in V^{\mathcal{Q}}_{\mathrm{n}}(\mathcal{D})$ denotes the argument of the sup problem, and boundary elements $\mathcal{B}$ are removed from the evaluation criterion, as in the periodic case (5.11). We recall that, in (5.13), $u^{M}_{\varepsilon,h}(f)$ is the empirical mean (4.1) over M realizations of $u_{\varepsilon,h}(f;\omega)$. It is thus an approximation to $\mathbb{E}\left[u_\varepsilon(f)\right]$.

The numerical results are collected in Table 3. We see that our surrogate defines an approximation of $\mathbb{E}(\boldsymbol{\nabla} u_\varepsilon)$ which is systematically better than that provided by the classical homogenization approach, for any choice of h and M.

5.3. Results in the case $\varepsilon \geq \overline{\varepsilon}$

In the regime $\varepsilon \geq \overline{\varepsilon}$, we quantitatively investigate whether the best constant matrix provided by our approach allows for an accurate approximation of the exact solution, in the L^2 norm in the sense of the criteria (5.10) or (5.12), and in the H^1 norm in the sense of the criteria (5.11) or (5.13).

TABLE 3. Approximation of $\mathbb{E}(\nabla u_\varepsilon)$ in the L^2 norm (`err_sto_H1`) by $C^{N,M}_{\varepsilon,h}\nabla u_{\star,h}$, $C^{N,M}_{\varepsilon,h}\nabla u^{N,M}_{\star,h}$ and $\overline{C}^{R,M}_{\varepsilon,h}\nabla \overline{u}^{P,M}_{\varepsilon,h}$ in function of N (the various lines correspond to various values of h and M).

N	8	16	32
`err_sto_H1` for $C^{N,M}_{\varepsilon,h}\nabla u_{\star,h}$ ($\varepsilon/h \approx 27$, $M = 100$)	1.043×10^{-1}	9.635×10^{-2}	9.394×10^{-2}
($\varepsilon/h \approx 108$, $M = 100$)	8.648×10^{-2}	8.120×10^{-2}	8.010×10^{-2}
($\varepsilon/h \approx 27$, $M = 400$)	8.542×10^{-2}	7.828×10^{-2}	7.298×10^{-2}
($\varepsilon/h \approx 54$, $M = 400$)	6.599×10^{-2}	6.222×10^{-2}	6.067×10^{-2}
`err_sto_H1` for $C^{N,M}_{\varepsilon,h}\nabla u^{N,M}_{\star,h}$ ($\varepsilon/h \approx 27$, $M = 100$)	9.799×10^{-2}	9.095×10^{-2}	8.961×10^{-2}
($\varepsilon/h \approx 108$, $M = 100$)	8.620×10^{-2}	8.022×10^{-2}	7.952×10^{-2}
($\varepsilon/h \approx 27$, $M = 400$)	7.605×10^{-2}	7.173×10^{-2}	6.780×10^{-2}
($\varepsilon/h \approx 54$, $M = 400$)	6.142×10^{-2}	5.957×10^{-2}	5.872×10^{-2}
`err_sto_H1` for $\overline{C}^{R,M}_{\varepsilon,h}\nabla \overline{u}^{P,M}_{\varepsilon,h}$ ($\varepsilon/h \approx 27$, $M = 100$)	6.000×10^{-2}	4.542×10^{-2}	3.018×10^{-2}
($\varepsilon/h \approx 108$, $M = 100$)	5.912×10^{-2}	4.657×10^{-2}	3.596×10^{-2}
($\varepsilon/h \approx 27$, $M = 400$)	3.030×10^{-2}	3.814×10^{-2}	2.625×10^{-2}
($\varepsilon/h \approx 54$, $M = 400$)	5.157×10^{-2}	3.613×10^{-2}	2.849×10^{-2}

TABLE 4. Approximation of $A_\star$ (`err_per_mat`) in function of ε (here $\varepsilon/h \approx 43$).

ε	0.4	0.2	0.1
`err_per_mat`	3.8420×10^{-2}	3.7056×10^{-3}	1.8623×10^{-3}

We also consider below the criterion (5.9), only in the periodic setting. It is indeed interesting to quantify the threshold value of ε above which $\overline{A}_\varepsilon$ is significantly different from $A_\star$ (let alone to understand the practical limitation of homogenization theory).

When considering large values of the parameter ε, it is necessary to consider P right-hand sides with P larger than $d(d+1)/2 = 3$, as pointed out in Section 3.1.1. This value depends on ε and is denoted $P(\varepsilon)$.

5.3.1. Results in the periodic setting

We consider the set $\{\varepsilon_k\}_{0\le k\le 2}$ of parameters introduced in Section 5.2.3. For $0 \le k \le 2$, we have $\varepsilon_k \ge \overline{\varepsilon}$. We choose the number of right-hand sides as $P(\varepsilon_0) = 9$ and $P(\varepsilon_1) = P(\varepsilon_2) = 5$ (we recall that $P(\varepsilon_k) = 3$ for $3 \le k \le 6$). Considering less right-hand sides significantly alters the approximation results, while considering more right-hand sides does not significantly improve these results.

We consider the evaluation criteria (5.9)–(5.11). We keep $\mathcal{Q} = 16$ functions in the test-space $V_{\rm n}^{\mathcal{Q}}(\mathcal{D})$. For the H^1-reconstruction, we choose the number of right-hand sides $R(\varepsilon)$ such that $R(\varepsilon_0) = R(\varepsilon_1) = 5$ and $R(\varepsilon_2) = 3$ (which satisfies $R(\varepsilon) \le P(\varepsilon)$). The numerical results for the approximation of the homogenized matrix, the L^2-approximation and the H^1-approximation, are respectively collected in Table 4, Figure 5 and Table 5.

We observe in Table 4 that the approximation of the homogenized matrix provided by our approach highly improves when decreasing ε from $\varepsilon = 0.4$ to $\varepsilon = 0.2$. For $\varepsilon \ge 0.4$, the homogenized matrix does not correctly describe the medium.

Figure 5 confirms this observation when it comes to the solution itself. We have seen that, for $\varepsilon = 0.4$, $A_\star$ and $\overline{A}_\varepsilon$ are significantly different. The solutions $u_\star$ and $\overline{u}_\varepsilon = \overline{u}(\overline{A}_\varepsilon)$ are also significantly different, the latter being a much better L^2-approximation of u_ε than the former or the first-order two-scale expansion. For smaller

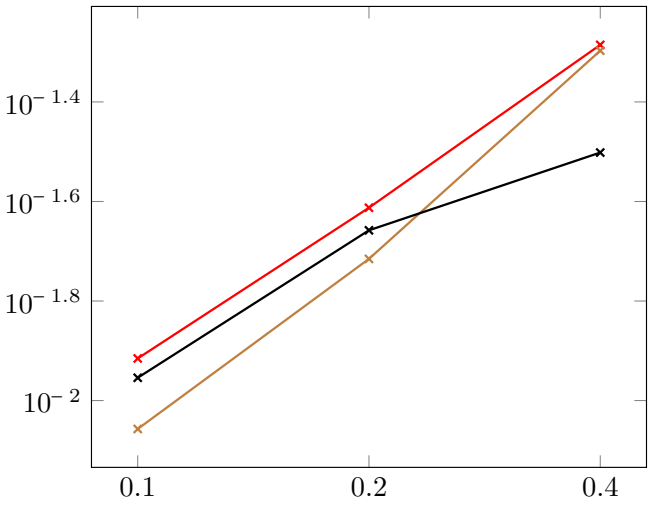

FIGURE 5. Approximation of u_ε in the L^2 norm (`err_per_L2`) by $u_{\star,h}$ (*red*), $u^{1,\boldsymbol{\theta}}_{\varepsilon,h}$ (*brown*) and $\overline{u}^P_{\varepsilon,h}$ (*black*) in function of ε (here $\varepsilon/h \approx 43$). These quantities are defined in Section 5.2.3.

TABLE 5. Approximation of $\boldsymbol{\nabla} u_\varepsilon$ in the L^2 norm (`err_per_H1`) by $C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h}$ and $\overline{C}^R_{\varepsilon,h}\boldsymbol{\nabla}\overline{u}^P_{\varepsilon,h}$ in function of ε (here $\varepsilon/h \approx 43$). See Section 5.2.3 for a definition of these quantities.

ε	0.4	0.2	0.1
`err_per_H1` for $C_{\varepsilon,h}\boldsymbol{\nabla} u_{\star,h}$	9.5890×10^{-2}	4.8421×10^{-2}	3.3923×10^{-2}
`err_per_H1` for $\overline{C}^R_{\varepsilon,h}\boldsymbol{\nabla}\overline{u}^P_{\varepsilon,h}$	8.7591×10^{-2}	5.8225×10^{-2}	3.2373×10^{-2}

values of ε, we already observe the behavior we have described in Section 5.2.3. Similar comments apply to the approximation of $\boldsymbol{\nabla} u_\varepsilon$ (see Tab. 5).

5.3.2. Results in the stationary ergodic setting

We consider the sets $\{\varepsilon_k\}_{0\le k\le 2}$ and $\{N_k\}_{0\le k\le 2}$ of parameters introduced in Section 5.2.4, for which we have $\varepsilon_k > \overline{\varepsilon}$. We choose the number of right-hand sides as $P(\varepsilon_0) = 9$ and $P(\varepsilon_1) = P(\varepsilon_2) = 5$, and fix the number of Monte Carlo realizations to $M = 100$.

We consider the evaluation criteria (5.12) and (5.13), with $\mathcal{Q} = 16$ functions in the test-space $V_{\mathrm{n}}^{\mathcal{Q}}(\mathcal{D})$. For the H^1-reconstruction, the number of right-hand sides is chosen to be $R(\varepsilon_0) = R(\varepsilon_1) = 5$ and $R(\varepsilon_2) = 3$. Note that again $R(\varepsilon) \le P(\varepsilon)$. The numerical results for the L^2- and H^1-approximation are respectively collected in Figure 6 and Table 6.

Remark 5.2. We note that, when working with $\varepsilon = \varepsilon_0 = 1/2$, we have, in view of (5.8), $N = N_0 = 1$. In view of (5.4)–(5.5), it turns out that, in this case, there are only 16 different realizations of the field a^{sto}. For this value of ε, the expectation is computed by a simple enumeration of all the possible realizations. For $\varepsilon = \varepsilon_1 = 1/4$, there are already 65 536 realizations, and expectations are computed by empirical means over M realizations.

In Figure 6, we observe that the solution associated with the best matrix we compute is an approximation of $\mathbb{E}(u_\varepsilon)$ (in the L^2 norm) generally more accurate than the solution associated with the exact homogenized matrix (since here N is small, the approximate matrix $A^{N,M}_\star$ is not expected to be an accurate approximation of $A_\star$). Table 6 shows that our surrogate defines an approximation of $\mathbb{E}(\boldsymbol{\nabla} u_\varepsilon)$, the accuracy of which is comparable,

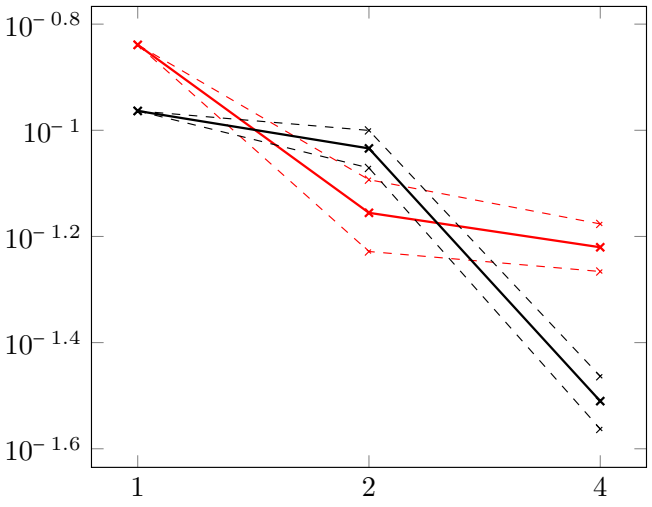

FIGURE 6. Approximation of $\mathbb{E}(u_\varepsilon)$ in the L^2 norm (`err_sto_L2`) by $u_{\star,h}$ (*red*) and $\overline{u}_{\varepsilon,h}^{P,M}$ (*black*) in function of N. For $N \geq 2$, all expectations are approximated by an empirical mean over $M = 100$ realizations. Since M is finite, results are random. We have performed the overall computation 10 times and show the corresponding 95% confidence interval (here $\varepsilon/h \approx 27$).

TABLE 6. Approximation of $\mathbb{E}(\nabla u_\varepsilon)$ in the L^2 norm (`err_sto_H1`) by $C_{\varepsilon,h}^{N,M}\nabla u_{\star,h}$ and $\overline{C}_{\varepsilon,h}^{R,M}\nabla\overline{u}_{\varepsilon,h}^{P,M}$ in function of N, for $M = 100$ and $\varepsilon/h \approx 27$ (see Sect. 5.2.4 for a definition of these quantities).

N	1	2	4
`err_sto_H1` for $C_{\varepsilon,h}^{N,M}\nabla u_{\star,h}$	1.4947×10^{-1}	1.3091×10^{-1}	1.0720×10^{-1}
`err_sto_H1` for $\overline{C}_{\varepsilon,h}^{R,M}\nabla\overline{u}_{\varepsilon,h}^{P,M}$	1.0955×10^{-1}	1.4595×10^{-1}	6.9334×10^{-2}

and often much better, to that provided by the homogenization approach. For the small values of N considered here, our approach is less expensive than the classical homogenization approach.

APPENDIX A. PROOF OF PROPOSITION 3.2

A.1. Preliminary results

Before we are in position to show Proposition 3.2, we first need to prove the following two preliminary lemmas, namely Lemma A.1 and Lemma A.2.

Lemma A.1. *Under the assumptions (2.4) and (2.5), the following convergence holds:*

$$\lim_{\varepsilon\to 0} \Phi_\varepsilon(A_\star) = 0. \tag{A.1}$$

We recall that Φ_ε is defined by (3.11): for any $\overline{A}$,

$$\Phi_\varepsilon(\overline{A}) = \sup_{f\in L^2_{\#}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f) = \sup_{f\in L^2_{\#}(\mathcal{D})} \left\|(-\Delta)^{-1}\left(\operatorname{div}(\overline{A}\nabla u_\varepsilon(f)) + f\right)\right\|^2_{L^2(\mathcal{D})}.$$

Proof of Lemma A.1. We use the notations and results of Section 3.1. Let $f_\star^\varepsilon \in L^2_{\rm n}(\mathcal{D})$ such that

$$\Phi_\varepsilon(A_\star) = \left\|(-\Delta)^{-1}\left(\operatorname{div}(A_\star \nabla u_\varepsilon(f_\star^\varepsilon)) + f_\star^\varepsilon\right)\right\|^2_{L^2(\mathcal{D})}, \tag{A.2}$$

and let $C_{\rm P} > 0$ be a Poincaré constant for $\mathcal{D}$, namely a constant such that, for any $v \in H^1_0(\mathcal{D})$, we have $\|v\|_{L^2(\mathcal{D})} \leq C_{\rm P}\|\nabla v\|_{L^2(\mathcal{D})}$.

Using standard *a priori* estimates, we have, for any $f \in L^2(\mathcal{D})$, that

$$\left\|(-\Delta)^{-1} f\right\|_{L^2(\mathcal{D})} \leq C_{\rm P}^2 \, \|f\|_{L^2(\mathcal{D})}. \tag{A.3}$$

Using that $\alpha \leq A_\varepsilon \leq \beta$ (see (2.5)), we likewise get that, for any $f \in L^2(\mathcal{D})$,

$$\|\nabla u_\varepsilon(f)\|_{L^2(\mathcal{D})} \leq \frac{C_{\rm P}}{\alpha}\|f\|_{L^2(\mathcal{D})}. \tag{A.4}$$

We now estimate $z_\varepsilon = (-\Delta)^{-1}\left(\operatorname{div}(A_\star \nabla u_\varepsilon(f))\right)$. We recall that (2.5) implies that

$$\alpha \leq A_\star \leq \beta. \tag{A.5}$$

From the variational formulation satisfied by z_ε, we obtain $\|\nabla z_\varepsilon\|_{L^2(\mathcal{D})} \leq |A_\star| \, \|\nabla u_\varepsilon(f)\|_{L^2(\mathcal{D})}$, which implies, using (A.4) and (A.5), that $\|\nabla z_\varepsilon\|_{L^2(\mathcal{D})} \leq C_{\rm P}\, \beta/\alpha \, \|f\|_{L^2(\mathcal{D})}$, hence

$$\left\|(-\Delta)^{-1}\left(\operatorname{div}(A_\star \nabla u_\varepsilon(f))\right)\right\|_{L^2(\mathcal{D})} \leq C_{\rm P}^2 \, \frac{\beta}{\alpha} \, \|f\|_{L^2(\mathcal{D})}. \tag{A.6}$$

Using (A.2), (A.6), (A.3) and the fact that $\|f_\star^\varepsilon\|_{L^2(\mathcal{D})} = 1$ for all $\varepsilon > 0$, we deduce that the sequence $\{\Phi_\varepsilon(A_\star)\}_{\varepsilon>0}$ is uniformly bounded. There thus exists a subsequence, that we still denote by $\{\Phi_\varepsilon(A_\star)\}_{\varepsilon>0}$, that converges in $\mathbb{R}$. Let us denote by $\overline{\Phi}$ its limit. We prove in the sequel that $\overline{\Phi} = 0$, which implies (A.1).

Since $\{f_\star^\varepsilon\}_{\varepsilon>0}$ is uniformly bounded in $L^2(\mathcal{D})$, there exists a subsequence, again denoted $\{f_\star^\varepsilon\}_{\varepsilon>0}$, that weakly converges in $L^2(\mathcal{D})$ when $\varepsilon \to 0$ to some function $f_\star^0 \in L^2(\mathcal{D})$ which satisfies $\left\|f_\star^0\right\|_{L^2(\mathcal{D})} \leq 1$. From (A.2), we infer, by the triangle inequality,

$$\left(\Phi_\varepsilon(A_\star)\right)^{1/2} \leq I_1^\varepsilon + I_2^\varepsilon + I_3^\varepsilon, \tag{A.7}$$

with

$$\begin{aligned}
I_1^\varepsilon &= \left\|(-\Delta)^{-1}\left(\operatorname{div}(A_\star \nabla u_\varepsilon(f_\star^\varepsilon - f_\star^0))\right)\right\|_{L^2(\mathcal{D})},\\
I_2^\varepsilon &= \left\|(-\Delta)^{-1}\left(\operatorname{div}(A_\star \nabla u_\varepsilon(f_\star^0)) + f_\star^0\right)\right\|_{L^2(\mathcal{D})},\\
I_3^\varepsilon &= \left\|(-\Delta)^{-1}(f_\star^\varepsilon - f_\star^0)\right\|_{L^2(\mathcal{D})}.
\end{aligned}$$

We successively show that I_1^ε, I_2^ε and I_3^ε vanish with ε.

Step 1. Estimation of I_1^ε. Let $z_\varepsilon = (-\Delta)^{-1}\left(\operatorname{div}\left[A_\star \nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0))\right]\right) \in H^1_0(\mathcal{D})$. We have

$$\|\nabla z_\varepsilon\|^2_{L^2(\mathcal{D})} = -\int_{\mathcal{D}} A_\star \nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0)) \cdot \nabla z_\varepsilon \leq \beta \left\|\nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0))\right\|_{L^2(\mathcal{D})} \|\nabla z_\varepsilon\|_{L^2(\mathcal{D})},$$

where we have used (A.5). Using the Poincaré inequality, we deduce

$$I_1^\varepsilon = \|z_\varepsilon\|_{L^2(\mathcal{D})} \leq C_{\rm P}\, \beta \left\|\nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0))\right\|_{L^2(\mathcal{D})},$$

thus, using (1.1), we get that

$$\left(I_1^\varepsilon\right)^2 \leq C_{\rm P}^2 \, \frac{\beta^2}{\alpha} \int_{\mathcal{D}} A_\varepsilon \nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0)) \cdot \nabla(u_\varepsilon(f_\star^\varepsilon - f_\star^0)) = C_{\rm P}^2 \, \frac{\beta^2}{\alpha} \int_{\mathcal{D}} (f_\star^\varepsilon - f_\star^0)\, u_\varepsilon(f_\star^\varepsilon - f_\star^0). \tag{A.8}$$

From (A.4), we also deduce

$$\left\|\boldsymbol{\nabla}(u_\varepsilon(f_\star^\varepsilon - f_\star^0))\right\|_{L^2(\mathcal{D})} \leq \frac{C_{\mathrm{P}}}{\alpha}\left\|f_\star^\varepsilon - f_\star^0\right\|_{L^2(\mathcal{D})} \leq 2\,\frac{C_{\mathrm{P}}}{\alpha}.$$

Using the Poincaré inequality, we obtain that the sequence $\left\{u_\varepsilon(f_\star^\varepsilon - f_\star^0)\right\}_{\varepsilon>0}$ is uniformly bounded in $H^1(\mathcal{D})$. There thus exists a subsequence, that we again denote $\left\{u_\varepsilon(f_\star^\varepsilon - f_\star^0)\right\}_{\varepsilon>0}$, which is strongly convergent in $L^2(\mathcal{D})$. The right-hand side of (A.8) is therefore the L^2 product of a sequence that weakly converges to 0 times a sequence that strongly converges. We hence deduce from (A.8) that

$$\lim_{\varepsilon\to 0} I_1^\varepsilon = 0. \tag{A.9}$$

Step 2. Estimation of I_2^ε. Let $w_\varepsilon = \mathrm{div}(A_\star \boldsymbol{\nabla} u_\varepsilon(f_\star^0)) + f_\star^0$, $r_\varepsilon = (-\Delta)^{-1} w_\varepsilon \in H_0^1(\mathcal{D})$ and $p_\varepsilon = (-\Delta)^{-1} r_\varepsilon \in H_0^1(\mathcal{D})$. Using the definition of p_ε, we have

$$(I_2^\varepsilon)^2 = \int_{\mathcal{D}} r_\varepsilon^2 = \int_{\mathcal{D}} \boldsymbol{\nabla} r_\varepsilon \cdot \boldsymbol{\nabla} p_\varepsilon. \tag{A.10}$$

Using the definition of r_ε, we have, for any $\phi \in H_0^1(\mathcal{D})$,

$$\int_{\mathcal{D}} \boldsymbol{\nabla} r_\varepsilon \cdot \boldsymbol{\nabla}\phi = -\int_{\mathcal{D}} A_\star \boldsymbol{\nabla} u_\varepsilon(f_\star^0) \cdot \boldsymbol{\nabla}\phi + \int_{\mathcal{D}} f_\star^0\ \phi. \tag{A.11}$$

Using (A.11) for $\phi \equiv p_\varepsilon$, (A.10) reads as

$$(I_2^\varepsilon)^2 = -\int_{\mathcal{D}} A_\star \boldsymbol{\nabla} u_\varepsilon(f_\star^0) \cdot \boldsymbol{\nabla} p_\varepsilon + \int_{\mathcal{D}} f_\star^0\ p_\varepsilon. \tag{A.12}$$

In order to pass to the limit $\varepsilon \to 0$ in (A.12), we establish some bounds. Using (A.11) with $\phi \equiv r_\varepsilon$ and the bounds (A.5), we deduce

$$\|\boldsymbol{\nabla} r_\varepsilon\|_{L^2(\mathcal{D})} \leq \beta \left\|\boldsymbol{\nabla} u_\varepsilon(f_\star^0)\right\|_{L^2(\mathcal{D})} + C_{\mathrm{P}} \left\|f_\star^0\right\|_{L^2(\mathcal{D})},$$

which (together with the Poincaré inequality and (A.4)) implies that r_ε is uniformly bounded in $H^1(\mathcal{D})$. There thus exists $r_0 \in H_0^1(\mathcal{D})$ such that, up to some extraction, r_ε converges to r_0, weakly in $H^1(\mathcal{D})$ and strongly in $L^2(\mathcal{D})$.

Passing to the limit $\varepsilon \to 0$ in (A.11), and using that $\boldsymbol{\nabla} u_\varepsilon(f)$ weakly converges to $\boldsymbol{\nabla} u_\star(f)$, we deduce that, for any $\phi \in H_0^1(\mathcal{D})$,

$$\int_{\mathcal{D}} \boldsymbol{\nabla} r_0 \cdot \boldsymbol{\nabla}\phi = -\int_{\mathcal{D}} A_\star \boldsymbol{\nabla} u_\star(f_\star^0) \cdot \boldsymbol{\nabla}\phi + \int_{\mathcal{D}} f_\star^0\ \phi = 0,$$

in view of the variational formulation of (1.2). We hence get that $r_0 \equiv 0$.

We now turn to p_ε. We have $p_\varepsilon = (-\Delta)^{-1} r_\varepsilon \in H_0^1(\mathcal{D})$ and r_ε converges to $r_0 = 0$, weakly in $H^1(\mathcal{D})$ and strongly in $L^2(\mathcal{D})$. Hence p_ε converges to 0 strongly in $H_0^1(\mathcal{D})$.

We now pass to the limit $\varepsilon \to 0$ in (A.12), and obtain

$$\lim_{\varepsilon\to 0} I_2^\varepsilon = 0. \tag{A.13}$$

Step 3. Estimation of I_3^ε. Let $k_\varepsilon = (-\Delta)^{-1}(f_\star^\varepsilon - f_\star^0)$. We have

$$\|\boldsymbol{\nabla} k_\varepsilon\|_{L^2(\mathcal{D})}^2 = \int_{\mathcal{D}} (f_\star^\varepsilon - f_\star^0) k_\varepsilon, \tag{A.14}$$

hence, using the Poincaré inequality,

$$\|\boldsymbol{\nabla} k_\varepsilon\|_{L^2(\mathcal{D})} \leq C_{\mathrm{P}} \big\| f_\star^\varepsilon - f_\star^0 \big\|_{L^2(\mathcal{D})} \leq 2\, C_{\mathrm{P}}.$$

The sequence $\{k_\varepsilon\}_{\varepsilon>0}$ is thus uniformly bounded in $H^1(\mathcal{D})$ and there exists a subsequence, that we again denote $\{k_\varepsilon\}_{\varepsilon>0}$, which is strongly convergent in $L^2(\mathcal{D})$. Using that $f_\star^\varepsilon - f_\star^0$ weakly converges to 0 in $L^2(\mathcal{D})$, we deduce from (A.14) that $\lim_{\varepsilon\to 0} \|\boldsymbol{\nabla} k_\varepsilon\|^2_{L^2(\mathcal{D})} = 0$, thus, again using the Poincaré inequality,

$$\lim_{\varepsilon\to 0} I_3^\varepsilon = \lim_{\varepsilon\to 0} \|k_\varepsilon\|_{L^2(\mathcal{D})} = 0. \tag{A.15}$$

Conclusion. Collecting (A.7), (A.9), (A.13) and (A.15), we obtain that $\Phi_\varepsilon(A_\star)$ converges to zero as $\varepsilon \to 0$. We thus have shown that $\overline{\Phi} = 0$. The limit being independent of the subsequence that we have considered, we eventually deduce that the whole sequence $\{\Phi_\varepsilon(A_\star)\}_{\varepsilon>0}$ converges to zero. This completes the proof of Lemma A.1. □

In what follows, we identify the set of indices $\{(i,j),\ \ 1 \leq i \leq j \leq d\}$ with the set of indices $\{m,\ \ 1 \leq m \leq \dfrac{d(d+1)}{2}\}$.

Lemma A.2. *There exist* $\dfrac{d\,(d+1)}{2}$ *functions* $f_{\star,k} \in L^2_{\mathrm{n}}(\mathcal{D})$ *and* $\dfrac{d\,(d+1)}{2}$ *functions* $\varphi_{\star,k} \in C_0^\infty(\mathcal{D})$ *such that the matrix* $Z_\star \in \mathbb{R}^{\frac{d(d+1)}{2}\times\frac{d(d+1)}{2}}$ *defined by*

$$\forall\, 1 \leq k \leq \frac{d\,(d+1)}{2}, \quad \forall\, 1 \leq i < j \leq d, \quad \begin{cases} [Z_\star]_{k,(i,i)} = \displaystyle\int_{\mathcal{D}} u_{\star,k}\ \partial_{ii}\varphi_{\star,k}, \\ [Z_\star]_{k,(i,j)} = 2\displaystyle\int_{\mathcal{D}} u_{\star,k}\ \partial_{ij}\varphi_{\star,k}, \end{cases} \tag{A.16}$$

where $u_{\star,k} = u_\star(f_{\star,k})$ *is the solution to* (1.2) *with right-hand side* $f_{\star,k}$*, is invertible.*

Proof of Lemma A.2. In the Steps 1 and 2 below, we construct $f_{\star,k} \in L^2_{\mathrm{n}}(\mathcal{D})$ and $\varphi_{\star,k} \in C_0^\infty(\mathcal{D})$ inductively for $1 \leq k \leq d(d+1)/2$, such that the vector $\boldsymbol{E}_\star^k \in \mathbb{R}^{\frac{d(d+1)}{2}}$ defined by

$$\forall\, 1 \leq i < j \leq d, \qquad \begin{cases} \Big[\boldsymbol{E}_\star^k\Big]_{(i,i)} = \displaystyle\int_{\mathcal{D}} u_{\star,k}\ \partial_{ii}\varphi_{\star,k}, \\ \Big[\boldsymbol{E}_\star^k\Big]_{(i,j)} = 2\displaystyle\int_{\mathcal{D}} u_{\star,k}\ \partial_{ij}\varphi_{\star,k}, \end{cases} \tag{A.17}$$

does not belong to $\mathrm{Span}(\boldsymbol{E}_\star^1, \ldots, \boldsymbol{E}_\star^{k-1})$. The vectors $\boldsymbol{E}_\star^1$, …, $\boldsymbol{E}_\star^{d(d+1)/2}$ being the rows of the matrix $Z_\star$, we deduce that $Z_\star$ is invertible.

Step 1. Construction of $\boldsymbol{E}_\star^1$. Choose $f_{\star,1} \in L^2_{\mathrm{n}}(\mathcal{D})$ and $\varphi_{\star,1} \in C_0^\infty(\mathcal{D})$ such that $\displaystyle\int_{\mathcal{D}} f_{\star,1}\ \varphi_{\star,1} \neq 0$, and consider $\boldsymbol{E}_\star^1 \in \mathbb{R}^{\frac{d(d+1)}{2}}$ defined by (A.17) (where we recall that $u_{\star,1}$ is the solution to (1.2) with right-hand side $f_{\star,1}$). Recalling that $A_\star$ is symmetric and constant, we have

$$\sum_{1\leq i\leq j\leq d} [A_\star]_{i,j}\ \Big[\boldsymbol{E}_\star^1\Big]_{(i,j)} = -\int_{\mathcal{D}} A_\star \boldsymbol{\nabla} u_{\star,1} \cdot \boldsymbol{\nabla}\varphi_{\star,1} = -\int_{\mathcal{D}} f_{\star,1}\ \varphi_{\star,1} \neq 0,$$

hence $\boldsymbol{E}_\star^1 \neq 0$.

Step 2. Induction. We assume that we have constructed $f_{\star,1}, \dots, f_{\star,k-1}$ and $\varphi_{\star,1}, \dots, \varphi_{\star,k-1}$ such that the family $\boldsymbol{E}^1_\star, \dots, \boldsymbol{E}^{k-1}_\star$ is free, for $k \leq d(d+1)/2$. We now construct $f_{\star,k} \in L^2_{\rm n}(\mathcal{D})$ and $\varphi_{\star,k} \in C^\infty_0(\mathcal{D})$ such that the vector $\boldsymbol{E}^k_\star \in \mathbb{R}^{\frac{d(d+1)}{2}}$ defined in (A.17) does not belong to $\mathrm{Span}(\boldsymbol{E}^1_\star, \dots, \boldsymbol{E}^{k-1}_\star)$.

We proceed by contradiction and assume that, for any such $f_{\star,k}$ and $\varphi_{\star,k}$, there exist $\lambda_\ell(f_{\star,k}, \varphi_{\star,k}) \in \mathbb{R}$, $1 \leq \ell \leq k-1$, such that

$$\boldsymbol{E}^k_\star = \sum_{\ell=1}^{k-1} \lambda_\ell(f_{\star,k}, \varphi_{\star,k})\, \boldsymbol{E}^\ell_\star.$$

For any vector $\boldsymbol{S}_\star \in \mathbb{R}^{\frac{d(d+1)}{2}}$, we have

$$\sum_{1\leq i\leq j\leq d} \int_{\mathcal{D}} \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} \partial_{ij} u_{\star,k}\, \varphi_{\star,k} = \sum_{1\leq i\leq j\leq d} [\boldsymbol{S}_\star]_{(i,j)} \left[\boldsymbol{E}^k_\star\right]_{(i,j)} = \sum_{\ell=1}^{k-1} \lambda_\ell(f_{\star,k}, \varphi_{\star,k})\, \boldsymbol{S}_\star \cdot \boldsymbol{E}^\ell_\star,$$

where, for any $\boldsymbol{S} \in \mathbb{R}^{\frac{d(d+1)}{2}}$ and $\boldsymbol{E} \in \mathbb{R}^{\frac{d(d+1)}{2}}$, we denote $\boldsymbol{S} \cdot \boldsymbol{E} = \displaystyle\sum_{m=1}^{d(d+1)/2} [\boldsymbol{S}]_m\, [\boldsymbol{E}]_m$, and where $\widehat{\boldsymbol{S}_\star} \in \mathbb{R}^{\frac{d(d+1)}{2}}$ is defined, for any $1 \leq i < j \leq d$, by

$$\left[\widehat{\boldsymbol{S}}_\star\right]_{(i,i)} = [\boldsymbol{S}_\star]_{(i,i)}\,, \qquad \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} = 2\,[\boldsymbol{S}_\star]_{(i,j)}\,.$$

Since $k-1 < d(d+1)/2$, there exists $\boldsymbol{S}_\star \in \mathbb{R}^{\frac{d(d+1)}{2}}$, $\boldsymbol{S}_\star \neq \boldsymbol{0}$, such that $\boldsymbol{S}_\star \cdot \boldsymbol{E}^\ell_\star = 0$ for all $1 \leq \ell \leq k-1$, and thus

$$\forall \varphi_{\star,k} \in C^\infty_0(\mathcal{D}), \qquad \sum_{1\leq i\leq j\leq d} \int_{\mathcal{D}} \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} \partial_{ij} u_{\star,k}\, \varphi_{\star,k} = 0.$$

Since $\boldsymbol{S}_\star$ (and thus $\widehat{\boldsymbol{S}}_\star$) only depends on $\boldsymbol{E}^1_\star, \dots, \boldsymbol{E}^{k-1}_\star$ and not on $\varphi_{\star,k}$, this implies

$$\sum_{1\leq i\leq j\leq d} \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} \partial_{ij} u_{\star,k} = 0 \quad \text{in the sense of distributions,}$$

thus

$$0 = -\sum_{1\leq i\leq j\leq d} \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} \partial_{ij} \mathrm{div}\,[A_\star \boldsymbol{\nabla} u_{\star,k}] = \sum_{1\leq i\leq j\leq d} \left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)} \partial_{ij} f_{\star,k},$$

for any $f_{\star,k} \in L^2_{\rm n}(\mathcal{D})$. Since $\left[\widehat{\boldsymbol{S}}_\star\right]_{(i,j)}$ does not depend on $f_{\star,k}$, this shows that $\widehat{\boldsymbol{S}}_\star$, and thus $\boldsymbol{S}_\star$, vanishes. We reach a contradiction. We thus obtain the existence of $f_{\star,k} \in L^2_{\rm n}(\mathcal{D})$ and $\varphi_{\star,k} \in C^\infty_0(\mathcal{D})$ such that the vectors $\boldsymbol{E}^1_\star, \dots, \boldsymbol{E}^{k-1}_\star, \boldsymbol{E}^k_\star$ form a free family. □

A.2. Proof of Proposition 3.2

We can now perform the proof of Proposition 3.2. The convergence (A.1) proved in Lemma A.1 readily shows (3.13). We are left with showing (3.14). Using the functions $f_{\star,k} \in L^2_{\rm n}(\mathcal{D})$ and $\varphi_{\star,k} \in C^\infty_0(\mathcal{D})$ defined by Lemma A.2, we introduce the matrix $Z_\varepsilon \in \mathbb{R}^{\frac{d(d+1)}{2} \times \frac{d(d+1)}{2}}$ defined by

$$\forall\, 1 \leq k \leq \frac{d(d+1)}{2}, \quad \forall\, 1 \leq i < j \leq d, \quad \begin{cases} [Z_\varepsilon]_{k,(i,i)} = \displaystyle\int_{\mathcal{D}} u_{\varepsilon,k}\, \partial_{ii} \varphi_{\star,k}, \\ [Z_\varepsilon]_{k,(i,j)} = 2 \displaystyle\int_{\mathcal{D}} u_{\varepsilon,k}\, \partial_{ij} \varphi_{\star,k}, \end{cases}$$

where $u_{\varepsilon,k} = u_\varepsilon(f_{\star,k})$ is the solution to (1.1) with right-hand side $f_{\star,k}$. Note that, for the second index of Z_ε, we have again identified the sets $\{(i,j),\ 1 \le i \le j \le d\}$ and $\{m,\ 1 \le m \le \dfrac{d(d+1)}{2}\}$.

Since $u_{\varepsilon,k}$ converges to $u_{\star,k}$ in $L^2(\mathcal{D})$, the matrix Z_ε converges to the matrix $Z_\star$ defined by (A.16) when ε goes to zero. We have proved in Lemma A.2 that the matrix $Z_\star$ is invertible. This implies that the matrix Z_ε is invertible for ε sufficiently small, and that Z_ε^{-1} is bounded independently of ε.

We now introduce the vectors $\overline{\boldsymbol{V}}_\varepsilon^\flat$ and $\boldsymbol{V}_\star$ in $\mathbb{R}^{\frac{d(d+1)}{2}}$ such that

$$\forall\, 1 \le i \le j \le d, \qquad \left[\overline{\boldsymbol{V}}_\varepsilon^\flat\right]_{(i,j)} = \left[\overline{A}_\varepsilon^\flat\right]_{i,j}, \qquad [\boldsymbol{V}_\star]_{(i,j)} = [A_\star]_{i,j},$$

where we recall that $\overline{A}_\varepsilon^\flat$ is a quasi-minimizing sequence of the functional (3.11) (see (3.12)). It can easily be seen that, for any $\overline{A} \in \mathcal{S}$, denoting $\overline{\boldsymbol{V}} \in \mathbb{R}^{\frac{d(d+1)}{2}}$ the vector such that $[\overline{\boldsymbol{V}}]_{(i,j)} = \overline{A}_{i,j}$ for any $1 \le i \le j \le d$, the following holds: for any $1 \le k \le d(d+1)/2$,

$$\left[Z_\varepsilon\ \overline{\boldsymbol{V}}\right]_k = \int_{\mathcal{D}} u_{\varepsilon,k}\, \mathrm{div}(\overline{A}\boldsymbol{\nabla}\varphi_{\star,k}) = \int_{\mathcal{D}} \mathrm{div}(\overline{A}\boldsymbol{\nabla}u_{\varepsilon,k})\ \varphi_{\star,k} = -\int_{\mathcal{D}} (-\Delta)^{-1}\left[\mathrm{div}(\overline{A}\boldsymbol{\nabla}u_{\varepsilon,k})\right]\ \Delta\varphi_{\star,k}, \tag{A.18}$$

where $Z_\varepsilon\ \overline{\boldsymbol{V}} \in \mathbb{R}^{\frac{d(d+1)}{2}}$ is the product of the matrix $Z_\varepsilon \in \mathbb{R}^{\frac{d(d+1)}{2}\times\frac{d(d+1)}{2}}$ by the vector $\overline{\boldsymbol{V}} \in \mathbb{R}^{\frac{d(d+1)}{2}}$: for any $1 \le k \le d(d+1)/2$, $\left[Z_\varepsilon\ \overline{\boldsymbol{V}}\right]_k = \displaystyle\sum_{1\le i\le j\le d} [Z_\varepsilon]_{k,(i,j)}\, [\overline{\boldsymbol{V}}]_{(i,j)}$.

Now, for any $f \in L^2_{\mathrm{n}}(\mathcal{D})$, we observe that

$$\begin{aligned}\left\|(-\Delta)^{-1}\left[\mathrm{div}\left(\overline{A}_\varepsilon^\flat\boldsymbol{\nabla}u_\varepsilon(f)\right) - \mathrm{div}\Big(A_\star\boldsymbol{\nabla}u_\varepsilon(f)\Big)\right]\right\|^2_{L^2(\mathcal{D})} &\le 2\left(\Phi_\varepsilon(\overline{A}_\varepsilon^\flat) + \Phi_\varepsilon(A_\star)\right)\\ &\le 2\left(I_\varepsilon + \varepsilon + \Phi_\varepsilon(A_\star)\right)\\ &\le 2\left(2\Phi_\varepsilon(A_\star) + \varepsilon\right).\end{aligned} \tag{A.19}$$

Hence, applying this to $f \equiv f_{\star,k}$, and owing to Lemma A.1,

$$\left\|(-\Delta)^{-1}\left[\mathrm{div}\left(\overline{A}_\varepsilon^\flat\boldsymbol{\nabla}u_{\varepsilon,k}\right)\right] - (-\Delta)^{-1}\Big[\mathrm{div}\Big(A_\star\boldsymbol{\nabla}u_{\varepsilon,k}\Big)\Big]\right\|_{L^2(\mathcal{D})}$$

vanishes with ε, for any $1 \le k \le d(d+1)/2$.

We next deduce from (A.18) that $Z_\varepsilon(\overline{\boldsymbol{V}}_\varepsilon^\flat - \boldsymbol{V}_\star)$ vanishes as $\varepsilon \to 0$. Since Z_ε is invertible when ε is sufficiently small (with Z_ε^{-1} bounded independently of ε), we obtain that $\lim_{\varepsilon\to 0} \overline{\boldsymbol{V}}_\varepsilon^\flat = \boldsymbol{V}_\star$, which is exactly the claimed convergence (3.14). This concludes the proof of Proposition 3.2.

Remark A.3. Since the above proof uses (A.19) precisely for the functions $f_{\star,k}$, $1 \le k \le d(d+1)/2$ (and not for all functions $f \in L^2_{\mathrm{n}}(\mathcal{D})$), we observe that, in the inf max formulation introduced in Remark 3.5, we have $\overline{A}_\varepsilon^{\max,\flat} \to A_\star$ when $\varepsilon \to 0$.

Remark A.4. We recall that our approach consists in considering the problem (3.1), that is

$$I_\varepsilon = \inf_{\overline{A}\in\mathcal{S}} \Phi_\varepsilon(\overline{A}),$$

where Φ_ε is defined by (3.11): for any $\overline{A}$,

$$\Phi_\varepsilon(\overline{A}) = \sup_{f\in L^2_{\mathrm{n}}(\mathcal{D})} \Phi_\varepsilon(\overline{A}, f) = \sup_{f\in L^2_{\mathrm{n}}(\mathcal{D})} \left\|(-\Delta)^{-1}\left(\mathrm{div}(\overline{A}\boldsymbol{\nabla}u_\varepsilon(f)) + f\right)\right\|^2_{L^2(\mathcal{D})}.$$

We show here that, when ε is sufficiently small, the minimum I_ε is attained.

Consider indeed a minimizing sequence $\overline{A}_\varepsilon^\eta$, satisfying, for any $\eta > 0$,

$$I_\varepsilon \leq \Phi_\varepsilon(\overline{A}_\varepsilon^\eta) \leq I_\varepsilon + \eta. \tag{A.20}$$

Similarly to (A.19), we observe that, for any $f \in L^2_{\rm n}(\mathcal{D})$,

$$\begin{aligned}\left\|(-\Delta)^{-1}\left[\operatorname{div}\left(\overline{A}_\varepsilon^\eta \boldsymbol{\nabla} u_\varepsilon(f)\right) - \operatorname{div}\left(A_\star \boldsymbol{\nabla} u_\varepsilon(f)\right)\right]\right\|^2_{L^2(\mathcal{D})} &\leq 2\left(\Phi_\varepsilon(\overline{A}_\varepsilon^\eta) + \Phi_\varepsilon(A_\star)\right)\\ &\leq 2\left(I_\varepsilon + \eta + \Phi_\varepsilon(A_\star)\right)\\ &\leq 2\left(2\Phi_\varepsilon(A_\star) + \eta\right).\end{aligned}$$

Using (A.18), we have

$$\left|Z_\varepsilon\left(\overline{\boldsymbol{V}}_\varepsilon^\eta - \boldsymbol{V}_\star\right)\right| \leq C \sup_{f \in L^2_{\rm n}(\mathcal{D})} \left\|(-\Delta)^{-1}\left[\operatorname{div}\left(\overline{A}_\varepsilon^\eta \boldsymbol{\nabla} u_\varepsilon(f)\right) - \operatorname{div}\left(A_\star \boldsymbol{\nabla} u_\varepsilon(f)\right)\right]\right\|_{L^2(\mathcal{D})}$$

where C is a constant independent of ε and η and where the vector $\overline{\boldsymbol{V}}_\varepsilon^\eta \in \mathbb{R}^{\frac{d(d+1)}{2}}$ is defined by $\left[\overline{\boldsymbol{V}}_\varepsilon^\eta\right]_{(i,j)} = \left[\overline{A}_\varepsilon^\eta\right]_{i,j}$ for any $1 \leq i \leq j \leq d$. When ε is sufficiently small, the matrix Z_ε is invertible with Z_ε^{-1} bounded independently of ε. We thus deduce from the two above estimates that

$$\left|\overline{\boldsymbol{V}}_\varepsilon^\eta - \boldsymbol{V}_\star\right|^2 \leq C\left(\Phi_\varepsilon(A_\star) + \eta\right)$$

for some C independent of ε and η. The vector $\overline{\boldsymbol{V}}_\varepsilon^\eta$ (resp. $\boldsymbol{V}_\star$) is the representation (as a vector in $\mathbb{R}^{\frac{d(d+1)}{2}}$) of the symmetric matrix $\overline{A}_\varepsilon^\eta \in \mathbb{R}^{d\times d}$ (resp. $A_\star$). We hence equivalently write that

$$\left|\overline{A}_\varepsilon^\eta - A_\star\right|^2 \leq C\left(\Phi_\varepsilon(A_\star) + \eta\right).$$

This shows that the sequence $\overline{A}_\varepsilon^\eta$ is bounded independently of η. Up to the extraction of a subsequence (that we still denote η for the sake of simplicity), it thus converges to some symmetric matrix $\overline{A}_\varepsilon^0$ when $\eta \to 0$. Since $A_\star$ is positive definite and since $\lim_{\varepsilon\to 0} \Phi_\varepsilon(A_\star) = 0$, we get that $\overline{A}_\varepsilon^0$ is also positive-definite.

Passing to the limit $\eta \to 0$ in (A.20), and temporarily assuming that Φ_ε is continuous, we get that $I_\varepsilon = \Phi_\varepsilon(\overline{A}_\varepsilon^0)$. This concludes the proof that the minimum I_ε is indeed attained when ε is sufficiently small.

We are left with showing the continuity of $\overline{A} \mapsto \Phi_\varepsilon(\overline{A})$. For any two matrices $\overline{A}_1$ and $\overline{A}_2$ and any $f \in L^2(\mathcal{D})$, we compute that

$$\begin{aligned}\Phi_\varepsilon(\overline{A}_1, f) - \Phi_\varepsilon(\overline{A}_2, f) = {}& \left\|(-\Delta)^{-1}\left[\operatorname{div}\left((\overline{A}_1 - \overline{A}_2)\boldsymbol{\nabla} u_\varepsilon(f)\right)\right]\right\|^2_{L^2(\mathcal{D})}\\ &+ 2\left\langle(-\Delta)^{-1}\left[\operatorname{div}\left((\overline{A}_1 - \overline{A}_2)\boldsymbol{\nabla} u_\varepsilon(f)\right)\right], (-\Delta)^{-1}\left[\operatorname{div}\left(\overline{A}_2 \boldsymbol{\nabla} u_\varepsilon(f)\right) + f\right]\right\rangle_{L^2(\mathcal{D})},\end{aligned}$$

hence

$$\left|\Phi_\varepsilon(\overline{A}_1, f) - \Phi_\varepsilon(\overline{A}_2, f)\right| \leq C\left|\overline{A}_1 - \overline{A}_2\right|^2 \|f\|^2_{L^2(\mathcal{D})} + C\left|\overline{A}_1 - \overline{A}_2\right| \|f\|^2_{L^2(\mathcal{D})},$$

where C is independent of f and $\overline{A}_1$. Taking the supremum over $f \in L^2_{\rm n}(\mathcal{D})$, we thus deduce that

$$\left|\Phi_\varepsilon(\overline{A}_1) - \Phi_\varepsilon(\overline{A}_2)\right| \leq C\left|\overline{A}_1 - \overline{A}_2\right|^2 + C\left|\overline{A}_1 - \overline{A}_2\right|,$$

which implies that $\lim_{\overline{A}_1 \to \overline{A}_2} \Phi_\varepsilon(\overline{A}_1) = \Phi_\varepsilon(\overline{A}_2)$, and thus the continuity of Φ_ε.

Appendix B. Details on the algorithm to solve the discrete problem (4.5)

Let $\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c})$ be given by (4.6). Using the fact that $\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c})$ is quadratic with respect to $\boldsymbol{c}\in\mathbb{R}^P$, one can easily observe that

$$\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}) = \boldsymbol{c}^T\, G^M_{\varepsilon,h}(\overline{A})\,\boldsymbol{c},$$

where $G^M_{\varepsilon,h}(\overline{A})$ is the $P\times P$ matrix defined, for any $1\le p,q\le P$, by

$$\big[G^M_{\varepsilon,h}(\overline{A})\big]_{p,q} = \frac{1}{2}\sum_{1\le i,j,k,l\le d}\big[\mathcal{K}^M_{\varepsilon,h}\big]_{i,j,k,l,p,q}\ \overline{A}_{i,j}\ \overline{A}_{k,l} - \sum_{1\le i,j\le d}\Big(\big[\mathbb{K}^M_{\varepsilon,h}\big]_{i,j,p,q}+\big[\mathbb{K}^M_{\varepsilon,h}\big]_{i,j,q,p}\Big)\,\overline{A}_{i,j} + [K_h]_{p,q}\,, \quad \text{(B.1)}$$

where $\mathcal{K}^M_{\varepsilon,h}$, $\mathbb{K}^M_{\varepsilon,h}$ and K_h are defined by (4.2), (4.3) and (4.4), respectively.

Using the fact that $\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c})$ is also quadratic with respect to $\overline{A}$, we have that

$$\Phi^{P,M}_{\varepsilon,h}(\overline{A},\boldsymbol{c}) = \frac{1}{2}\sum_{1\le i,j,k,l\le d}\Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j,k,l}\ \overline{A}_{i,j}\ \overline{A}_{k,l} - \sum_{1\le i,j\le d}\Big[B^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j}\ \overline{A}_{i,j} + b^P_h(\boldsymbol{c}),$$

where $\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})$ is the $d\times d\times d\times d$ fourth-order tensor defined by

$$\Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j,k,l} = \sum_{1\le p,q\le P}\big[\mathcal{K}^M_{\varepsilon,h}\big]_{i,j,k,l,p,q}\ c_p\ c_q, \quad \text{(B.2)}$$

$B^{P,M}_{\varepsilon,h}(\boldsymbol{c})$ is the $d\times d$ matrix defined by

$$\Big[B^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j} = \sum_{1\le p,q\le P}\Big(\big[\mathbb{K}^M_{\varepsilon,h}\big]_{i,j,p,q}+\big[\mathbb{K}^M_{\varepsilon,h}\big]_{i,j,q,p}\Big)\,c_p\ c_q, \quad \text{(B.3)}$$

and

$$b^P_h(\boldsymbol{c}) = \sum_{1\le p,q\le P}[K_h]_{p,q}\ c_p\ c_q,$$

where $\mathcal{K}^M_{\varepsilon,h}$, $\mathbb{K}^M_{\varepsilon,h}$ and K_h are defined by (4.2), (4.3) and (4.4), respectively. We remark, in light of the expressions (4.2) and (4.3), that

$$\Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j,k,l} = \Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{k,l,i,j}, \qquad \Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{i,j,k,l} = \Big[\mathbb{B}^{P,M}_{\varepsilon,h}(\boldsymbol{c})\Big]_{j,i,k,l},$$

and $[B^{P,M}_{\varepsilon,h}(\boldsymbol{c})]_{i,j} = [B^{P,M}_{\varepsilon,h}(\boldsymbol{c})]_{j,i}$.

Acknowledgements. The authors would like to thank Albert Cohen (Université Pierre et Marie Curie) for stimulating and enlightening discussions about the work reported in this article, and in particular for suggesting the cost function in (1.5) in replacement of that in (1.4), for providing the perspective of an optimization upon the class of matrices $\overline{A}$ that are considered, as detailed in Section 1.3, and for carefully reading a preliminary version of this manuscript.

The authors also acknowledge several constructive comments by the two anonymous referees, which have allowed to improve (in particular with Rems. 2.2 and A.4) the original version of this manuscript.

The work of CLB, FL and SL is partially supported by EOARD under Grant FA8655-13-1-3061. The work of CLB and FL is also partially supported by ONR under Grants N00014-12-1-0383 and N00014-15-1-2777.

References

[1] G. Allaire and M. Amar, Boundary layer tails in periodic homogenization. *ESAIM: COCV* **4** (1999) 209–243.

[2] A. Anantharaman, R. Costaouec, C. Le Bris, F. Legoll and F. Thomines, Introduction to numerical stochastic homogenization and the related computational challenges: some recent developments, Multiscale Modeling and Analysis for Materials Simulation. Edited by W. Dao and Q. Du. In Vol. 22 of *Lecture Notes Series*. Institute for Mathematical Sciences, National University of Singapore (2011) 197–272.

[3] A. Bensoussan, J.-L. Lions and G. Papanicolaou. Asymptotic analysis for periodic structures. In Vol. 5 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam-New York (1978).

[4] X. Blanc, C. Le Bris and P.-L. Lions, Une variante de la théorie de l'homogénéisation stochastique des opérateurs elliptiques (A variant of stochastic homogenization theory for elliptic operators). *C.R. Acad. Sci. Paris, Série I* **343** (2006) 717–724.

[5] X. Blanc, C. Le Bris and P.-L. Lions, Stochastic homogenization and random lattices. *J. Math. Pures Appl.* **88** (2007) 34–63.

[6] A. Bourgeat and A. Piatniski, Approximation of effective coefficients in stochastic homogenization. *Ann. Inst. Henri Poincaré Probab. Stat.* **40** (2004) 153–165.

[7] L.J. Durlofsky, Numerical calculation of equivalent grid block permeability tensors for heterogeneous porous media. *Water Resources Res.* **27** (1991) 699–708.

[8] W.E and B. Engquist, The Heterogeneous Multiscale Methods. *Commun. Math. Sci.* **1** (2003) 87–132.

[9] Y. Efendiev and T. Y. Hou. Multiscale Finite Element Methods – Theory and Applications. In Vol. 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer-Verlag, New York (2009).

[10] B. Engquist and P.E. Souganidis, Asymptotic and numerical homogenization. *Acta Numer.* **17** (2008) 147–190.

[11] 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** (1997) 169–189.

[12] T.J.R. Hughes, G.R. Feijó, L.M. Mazzei and J.-B. Quincy, The variational multiscale method – a paradigm for computational mechanics. *Comput. Methods Appl. Mech. Eng.* **166** (1998) 3–24.

[13] V.V. Jikov, S.M. Kozlov and O.A. Oleinik, Homogenization of Differential Operators and Integral Functionals. Springer–Verlag, Berlin Heidelberg (1994).

[14] R. Kornhuber and H. Yserentant, Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.* **14** (2016) 1017–1036.

[15] C. Le Bris, Some numerical approaches for "weakly" random homogenization. Numerical mathematics and advanced applications. In *Proc. of ENUMATH 2009*, of *Lect. Notes Comput. Sci. Eng.* Springer (2010).

[16] C. Le Bris, F. Legoll and K. Li, Approximation grossière d'un problème elliptique à coefficients hautement oscillants (Coarse approximation of an elliptic problem with highly oscillatory coefficients) *C. R. Acad. Sci. Paris, Série I* **351** (2013) 265–270.

[17] A. Målqvist and D. Peterseim, Localization of elliptic multiscale problems. *Math. Comput.* **83** (2014) 2583–2603.

[18] G. C. Papanicolaou and S.R.S. Varadhan, Boundary value problems with rapidly oscillating random coefficients, In *Proc. Colloq. on Random Fields: Rigorous Results in Statistical Mechanics and Quantum Field Theory* **10** (1981) 835–873.

[19] L. Tartar, The general theory of homogenization – A personalized introduction. In Vol. 7 of *Lect. Notes of the Unione Matematica Italiana*. Springer–Verlag, Berlin Heidelberg (2010).