ESAIM: M2AN 55 (2021) S149–S185
https://doi.org/10.1051/m2an/2020024

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

COMPUTING HOMOGENIZED COEFFICIENTS *VIA* MULTISCALE REPRESENTATION AND HIERARCHICAL HYBRID GRIDS

Antti Hannukainen[1], Jean-Christophe Mourrat[2,*] and Harmen T. Stoppels[1]

Abstract. We present an efficient method for the computation of homogenized coefficients of divergence-form operators with random coefficients. The approach is based on a multiscale representation of the homogenized coefficients. We then implement the method numerically using a finite-element method with hierarchical hybrid grids, which is a semi-implicit method allowing for significant gains in memory usage and execution time. Finally, we demonstrate the efficiency of our approach on two- and three-dimensional examples, for piecewise-constant coefficients with corner discontinuities. For moderate ellipticity contrast and for a precision of a few percentage points, our method allows to compute the homogenized coefficients on a laptop computer in a few seconds, in two dimensions, or in a few minutes, in three dimensions.

Mathematics Subject Classification. 65N55, 35B27.

Received June 19, 2019. Accepted April 5, 2020.

1. Introduction

1.1. Statement of the main results

The goal of this paper is to define, study, and implement an efficient approach to the calculation of homogenized coefficients for divergence-form operators with random coefficients. That is, we consider operators of the form $\nabla \cdot \mathbf{a}\nabla$, where $\mathbf{a} = (\mathbf{a}(x))_{x\in\mathbb{R}^d}$ is a random coefficient field on $\mathbb{R}^d$ taking values in the set of symmetric positive definite matrices. We assume that this random coefficient field is uniformly elliptic, $\mathbb{Z}^d$-stationary, and of unit range of dependence; see Section 2.1 for precise statements. Under these assumptions, there exists a *homogenized matrix* $\bar{\mathbf{a}}$ such that the large-scale properties of the heterogeneous operator $\nabla \cdot \mathbf{a}(x)\nabla$ resemble those of the homogeneous operator $\nabla \cdot \bar{\mathbf{a}}\nabla$. We define a multiscale method allowing to compute the homogenized matrix efficiently, and identify rigorously its rate of convergence. We then explain how to implement the algorithm in practice, using the notion of hierarchical hybrid grids, and demonstrate its performance on examples.

For these numerical examples, we consider coefficient fields that are piecewise constant on a square tiling, in two dimensions, or on a cubic tiling, in three dimensions. This class of examples is particularly challenging from a computational perspective. Indeed, solutions develop singularities at the corners of the tiling which are essentially the worst possible in the class of (isotropic) coefficient fields with fixed ellipticity contrast (see Sect. 5.1). Despite

Keywords and phrases. Homogenization, multiscale method, hierarchical hybrid grids.

[1] Department of Mathematics and Systems Analysis, Aalto University, Espoo, Finland.
[2] DMA, Ecole normale supérieure, CNRS, PSL University, Paris, France.
*Corresponding author: mourrat@dma.ens.fr

this, for moderate ellipticity contrast and for a precision of a few percentage points, our algorithm runs on a laptop computer and outputs a satisfactory approximation of the homogenized matrix within a few seconds in two dimensions, and within a few minutes in three dimensions. Our code is written in the Julia language and is freely available online, see the link in (6.1).

The method we explore here was introduced in [52] in the context of discrete finite-difference equations. The main idea is to decompose the homogenized matrix into a series of terms, each of which accounting for a different length scale. The terms associated with short length scales naturally enjoy very small boundary layers and low computational effort. Those associated with larger length scales are *a priori* more demanding, but only appear as small correction terms in the decomposition, and can thus be computed on much smaller sample domains. Overall, this second effect more than compensates for the increase in computational effort, so that the majority of the computational time and memory is spent on the shortest length scales. A prominent feature of the method is that minimal effort is spent on the calculation of boundary layers. An additional benefit is that the method can be refined on the fly: if some calculations have already been performed and one realizes that more precision is necessary, then one does not need to throw these past calculations away and restart from scratch.

We now describe this method more precisely. We fix $\xi \in \mathbb{R}^d$ of unit norm, and introduce the quantities that will allow us to compute $\xi \cdot \bar{\mathbf{a}}\xi$. We let $v_{-1} \in H^{-1}_{\mathrm{loc}}\left(\mathbb{R}^d\right)$ be

$$v_{-1}(x) := \nabla \cdot (\mathbf{a}(x)\xi), \tag{1.1}$$

and for each $k \in \mathbb{N}$, we define inductively $v_k \in H^1_{\mathrm{loc}}\left(\mathbb{R}^d\right)$ to be the unique stationary solution to

$$\left(2^{-k} - \nabla \cdot \mathbf{a}\nabla\right) v_k = 2^{-k}\, v_{k-1} \quad \text{in} \quad \mathbb{R}^d. \tag{1.2}$$

We also give ourselves a bump function $\chi \in C^\infty_c\left(\mathbb{R}^d\right)$ with compact support in the unit ball $B(0,1)$ and such that

$$\int_{\mathbb{R}^d} \chi = 1. \tag{1.3}$$

In (1.3) and throughout the paper, we use the shorthand notation

$$\int_{\mathbb{R}^d} \chi = \int_{\mathbb{R}^d} \chi(x)\,\mathrm{d}x.$$

For every $r \geqslant 1$, we set

$$\chi_r(x) := r^{-d}\chi\left(\tfrac{x}{r}\right). \tag{1.4}$$

The following theorem is our main theoretical result.

Theorem 1.1 (Efficient approximation of $\bar{\mathbf{a}}$)**.** *For every $\varepsilon \in \left(0, \frac{d-1}{2d}\right)$, there exists a constant $c\left(\varepsilon, \|\chi\|_{H^{\lceil \frac{d}{2}+\frac{1}{4} \rceil}(\mathbb{R}^d)}, \Lambda, d\right) > 0$ such that the following holds. Let $n \in \mathbb{N}$, and denote*

$$r_k := 2^{n-\left(\frac{1}{2}-\varepsilon\right)k} \quad (k \in \{0,\dots,n\}), \tag{1.5}$$

$$\widehat{\sigma}^2_n := \int_{\mathbb{R}^d} \left(-\mathbf{a}\xi \cdot \nabla v_0 + v_0^2\right)\chi_{r_0} + \sum_{k=1}^{n} 2^k \int_{\mathbb{R}^d} \left(v_{k-1}v_k + v_k^2\right)\chi_{r_k}. \tag{1.6}$$

For every $t \geqslant 0$, we have

$$\mathbb{P}\left[\left|\xi \cdot \bar{\mathbf{a}}\xi + \widehat{\sigma}^2_n - \int_{\mathbb{R}^d} (\xi \cdot \mathbf{a}\xi)\,\chi_{r_0}\right| \geqslant t2^{-\frac{nd}{2}}\right] \leqslant 2\exp\left(-ct\right). \tag{1.7}$$

Recall that we assume the law of the coefficient field $(\mathbf{a}(x))_{x\in\mathbb{R}^d}$ to be invariant under translations by vectors of $\mathbb{Z}^d$. If we make the stronger hypothesis that the law is invariant under translations by any vector of $\mathbb{R}^d$, then we can replace each average against a smooth mask χ_r in (1.6) by an average over the cube $(-r,r)^d$. However, under our current more restrictive assumption of invariance under translations by vectors of $\mathbb{Z}^d$, this replacement will only work if we make sure that the side length of the box is an integer. In other words, we would need to know the identity of the underlying lattice of periods (which without loss of generality was fixed here to be $\mathbb{Z}^d$) and to make sure that the domain over which we take the average contains an integer number of fundamental cells. In contrast, the formulation in Theorem 1.1 does not require that we identify the lattice of periods.

A result comparable to Theorem 1.1 was proved in [52] in the context of discrete finite-difference equations. Besides the adaptation to the continuous setting, there are two main differences between the present result and the one obtained in [52]. The first one is that the quantities on the right side of (1.6) are averages against a smooth mask, while only box averages could be handled in [52]. The second and most important difference is that Theorem 1.1 gives an exponential tail estimate for the probability in (1.7), while the result in [52] was limited to a variance estimate. We expect the estimate (1.7) to be sharp, in the sense that we do not expect that it is possible to replace t by t^α on the right side of (1.7) for an exponent $\alpha > 1$ that would be independent of $\varepsilon > 0$.

The implementation of the method proposed in Theorem 1.1 requires the accurate calculation of ∇v_0 and of $v_0, \dots, v_n$ in L^2 over the progressively smaller and smaller balls $B(0, r_0), \dots, B(0, r_n)$. As stated in (1.2), the equation satisfied by v_k is posed over the full space $\mathbb{R}^d$. In practice, we can approximate these problems by selecting a sufficiently large constant C_{bl} ("bl" for "boundary layer"), and then solving for $\widetilde{v}_k \in H^1_0\left(B\left(0, r_k + C_{\mathrm{bl}}(1+n)2^{\frac{k}{2}}\right)\right)$ solution to

$$\left(2^{-k} - \nabla\cdot\mathbf{a}\nabla\right)\widetilde{v}_k = 2^{-k}\,\widetilde{v}_{k-1} \quad \text{in} \quad B\left(0, r_k + C_{\mathrm{bl}}(1+n)2^{\frac{k}{2}}\right), \tag{1.8}$$

with null Dirichlet boundary condition on $\partial B\left(0, r_k + C_{\mathrm{bl}}(1+n)2^{\frac{k}{2}}\right)$, and where we have set $\widetilde{v}_{-1} := v_{-1}$. The error in this approximation decays exponentially fast as we increase C_{bl} $\left($this can be proved using that the Green function decays like $\exp\left(-2^{-\frac{k}{2}}|x|\right)\right)$. As a rule of thumb, one should think of choosing C_{bl} of the order of $\sqrt{|\mathbf{a}|}$, where $|\mathbf{a}|$ is a measure of the typical size of the eigenvalues of $\mathbf{a}(x)$ (or, to be more specific, one can take C_{bl} to be of the order of $\sqrt{\Lambda}$). The additional multiplicative factor of $(1+n)$ allows for a progressive increase of the boundary layer as we increase n and aim for finer and finer approximations of $\bar{\mathbf{a}}$. A simple error analysis suggests that the optimal choice for the size of the boundary layer should be an affine function of n, and we chose it to be a multiple of $(1+n)$ for simplicity, but more refined choices can save some computations.

For simplicity, we implemented the version of the method described in Theorem 1.1 with $\varepsilon = 0$. Strictly speaking, this case is not covered by Theorem 1.1, but a modification of the arguments presented below would in this case yield (1.7) with $2^{-\frac{nd}{2}}$ replaced by $2^{-\frac{nd}{2}(1-\delta)}$, for arbitrary $\delta > 0$. (The constant $c > 0$ on the right side would then depend on δ.)

The main power of the method comes from the fact that it splits the problem of calculating $\xi\cdot\bar{\mathbf{a}}\xi$ into multiple scales. Heuristically, the term v_k (or $\widetilde{v}_k$) is meant to capture information related to length scales of the order of $2^{\frac{k}{2}}$. When k is small, the elliptic problem (1.8) is well-conditioned and has a very small boundary layer, of essentially unit size. As k is increased, the elliptic problems in (1.8) become less well-conditioned and involve larger boundary layers. Yet, this is more than compensated by the fact that the domain of interest is rapidly shrinking. In practice, the main part of the computational effort is spent on calculating v_0.

Compared with the discrete setting of finite-difference operators investigated in [52], the case of continuous differential operators considered here poses crucial new challenges from a computational perspective. With applications such as those in materials science in mind, it is natural to consider piecewise constant coefficient fields. We choose to focus more specifically on the case when the coefficient field is constant over each unit square or cube of the form $z + [0,1)^d$, where $z \in \mathbb{Z}^d$. At least in dimension $d = 2$, this class is essentially the most difficult possible, in the sense that solutions then have the worst possible regularity properties, given

the ellipticity contrast – see Section 5.1 for a more precise discussion. As a consequence of the roughness of the solutions, a "coarsest possible" discretization of the coefficient field into finite elements with constant coefficients would yield widely incorrect results. To wit, the algorithm as proposed here would run just fine, but it would compute the homogenized matrix associated with the particular finite-element discretization that is chosen; if the discretization is coarse, then this homogenized matrix will be *far* from the homogenized matrix of the continuous operator.

To remedy this problem, we thus need to rely on much finer discretizations of the coefficient field. Our method for doing so is strongly inspired by the idea of *hierarchical hybrid grids* developed in [14, 15]. In a nutshell, the starting point is the observation that numerical schemes on fully structured grids with constant coefficients are highly efficient, both from the point of view of time and of memory usage. Unfortunately, the problem we wish to solve is not of this type, since the coefficients vary across the domain. The idea then is to deploy a hybrid representation of the problem, using an unstructured coarse grid to represent the variations of the coefficient field on the one hand, and then proceeding to refine each coarse element in a "fully structured" manner. This allows for very significant gains in memory usage, which is otherwise a fundamental bottleneck, and also in execution time.

We did not make any effort to fine-tune the parameters of the method presented in Theorem 1.1. We indicate here some possible directions for doing so. First, the choice to use successive powers of 2 in (1.2) can be replaced by any other real number larger than 1, up to suitable modifications of the expression in (1.6). Second, for the radii r_k appearing in (1.6), we simply followed the prescription of the theoretical result with $\varepsilon = 0$, that is, $r_k = 2^{n-\frac{k}{2}}$. A more fine-tuned method would consist in evaluating the fluctuations of integral averages on the fly and adaptively tune r_k so that the fluctuations of the average get below a certain threshold of the order of $2^{-\frac{nd}{2}}$. Finally, the requirements for accuracy are different for v_0, which needs to be controlled in H^1, and for the subsequent v_k's which only need to be controlled in L^2. We did not try to exploit this feature either, and used approximations of the same quality for all terms.

Although we did not explore this possibility, we point out that the required computations can be performed in parallel in a straightforward way. For instance, instead of computing

$$\int_{\mathbb{R}^d} \left(-\mathbf{a}\xi \cdot \nabla v_0 + v_0^2\right) \chi_{r_0},$$

if one has access to L^d processors, then one can compute

$$\sum_{\ell=1}^{L^d} \int_{\mathbb{R}^d} \left(-\mathbf{a}\xi \cdot \nabla v_0^{(\ell)} + (v_0^{(\ell)})^2\right) \chi_{\frac{r_0}{L}},$$

where $\left(v_0^{(\ell)}\right)_{1 \leqslant \ell \leqslant L^d}$ are versions of v_0 computed on L^d independent realizations of the coefficient field. These computations can obviously be performed without any communication between processors. (If one is given a very large snapshot of a single environment, then effectively independent realizations can be obtained by considering sufficiently distant subregions of the large sample.) See also [39] for more refined techniques allowing for the parallelization of finite-element methods with hierarchical hybrid grids.

For simplicity, we assume here that the coefficient field is uniformly elliptic and with a finite range of dependence. However, we expect the results presented here to hold in much greater generality. In particular, we expect that a result comparable with (1.7), but possibly with a more slowly decaying function of t on the right side, should hold whenever the local statistics of the coefficient field satisfy a central limit theorem. For more strongly correlated coefficient fields, the method is still of interest, but the choice of r_k in (1.5) and the term $2^{-\frac{nd}{2}}$ in (1.7) will have to be suitably modified. (This makes the development of a more adaptive algorithm particularly appealing, since such an algorithm could automatically select the optimal scalings without supervision.) Also, in view of [8, 23], we expect that the results can be generalized to the case of perforated domains of percolation type.

1.2. Related works

Over the last decade, an intensive research effort has been devoted to developing theoretical quantitative results on stochastic homogenization. The multiscale representation of the homogenized coefficients forming the basis of the method is inspired by the "renormalization" approach to quantitative stochastic homogenization, as developed in [9–12]; see also [53] for a gentle introduction to this line of research and [13] for a monograph. A related approach based on the parabolic flow was put forward in [36], see also Chapter 9 of [13], and will give us the most convenient statement for us to build upon here. A different approach based on concentration inequalities was put forward in [33–35, 37, 38, 47], inspired by earlier insights from statistical mechanics [54, 55].

It has been observed long ago that inappropriate boundary conditions for "approximate cell problems" can cause important "resonant errors", and initial attempts at bypassing the problem involved the notion of oversampling [26, 41, 42, 60]. A powerful approach has been studied in [18, 27, 30–34, 50], based on the introduction of a small zero-order term in the equation. The method we propose here, by combining this idea with a multiscale decomposition, enables to take fuller advantage of this idea. We refer to [52] for a detailed comparison between the single-scale and the multiscale approaches. As is shown in [1], the benefits of the multiscale approach can be seen even in the setting of periodic coefficient fields, if we operate under the constraint that the lattice of periods is unknown.

One alternative method for computing homogenized coefficients, based on the idea of an "embedded corrector problem", is proposed in [21, 22]. Well-separated spherical inclusions are considered in the numerical examples. This allows for fairly different approaches to practical calculations than what is pursued in the present paper (and also produces solutions that are more regular than in our examples with corner discontinuities).

For coefficient fields that are very similar to those we investigate numerically here, the standard representative volume method was combined with a tensor-based discretization scheme in [43] to compute homogenized matrices, in two dimensions. The authors of [43] state that their numerical approximation method displays an empirical rate of convergence in L^2 of $O\left(h^{\beta}\right)$ with $\beta \geqslant 3/2$, where h measures the size of a discretized element. We believe that this is an artifact of pre-asymptotic effects and moderate ellipticity contrast. Indeed, for any $\alpha > 0$, solutions can develop singularities that fail to be in $H^{1+\alpha}$, provided that the ellipticity contrast is sufficiently large, and standard finite-element methods provide approximations of these singular solutions that converge in L^2 at a rate that is bounded below by $c\,h^{1+\alpha}$. In fact, for coefficient fields arranged in a checkerboard-type pattern in two dimensions, as considered in [43] and in the present paper, one can identify exactly the optimal exponent of regularity in terms of the ellipticity contrast: solutions are H^{β}-regular if and only if $\beta < 1 + \alpha$, where α is given in (5.5), as proved in [57] and recalled in Section 5.1 below. In particular, an asymptotic rate of convergence in L^2 of $O\left(h^{3/2}\right)$ can only be obtained for values of the ellipticity contrast Λ below $3 + 2\sqrt{2}$. We also refer to the right frame of Figure 6 for an illustration of pre-asymptotic effects and slow rates of convergence, for $\Lambda = 90$.

Several techniques have been explored to reduce the size of the fluctuations of estimators for the homogenized matrix. In particular, control variate techniques and the selection of special realizations of the coefficient field, called "quasi-random structures", have been explored, see [19, 45] for surveys. The latter approach, inspired by [59, 61] and, in the context of the homogenization of elliptic operators, advocated for in [46], has recently received a spectacular theoretical foundation in [28]. We would find it very interesting to investigate how these techniques can be combined with those discussed in the present paper.

In a different direction, several works have considered the question of designing and effectively computing certain expansions of the homogenized matrix, in situations where the random medium can be seen as a small perturbation of a reference medium. The most typical scenario is that of a homogeneous medium with a small density of inclusions [48, 58]. We refer to [3–7, 16, 25, 44, 51, 56] for works in this area.

To conclude this introduction, we mention that the homogenized matrix can also be of use as part of a modified scheme of multigrid type for computing solutions of elliptic equations with rapidly oscillating coefficients. In short, the idea is to use the homogenized operator when operating on the coarser grids [52].

1.3. Organization of the paper

In Section 2, we lay down the notation and make our standing assumptions more precise. We also clarify the meaning of being a stationary solution to (1.2), and recall the definition of the homogenized matrix. We next prove a general multiscale representation of the homogenized matrix in Section 3. By "general", we mean that the finite-range dependence assumption on the coefficient field is not actually used there; assuming ergodicity instead would be sufficient. This is no longer the case in Section 4, where we strongly leverage on the finite-range dependence assumption to obtain sharp quantitative estimates on the different terms appearing in the multiscale decomposition. This allows us to conclude the proof of Theorem 1.1. In Section 5, we explain how to design a finite-element multigrid algorithm using the structure of hierarchical hybrid grids. Finally, we present our numerical results in Section 6. Our code is freely available in the GitHub repository indicated in (6.1).

2. Assumptions, notation, and definition of homogenized matrix

2.1. Precise statement of the assumptions

We fix a constant $\Lambda \in [1, \infty)$ and an integer $d \geqslant 2$ throughout the paper. We denote by Ω the set of measurable mappings from $\mathbb{R}^d$ to the set of d-by-d symmetric matrices which satisfy, for almost every $x \in \mathbb{R}^d$,

$$\forall \xi \in \mathbb{R}^d, \quad \Lambda^{-1}|\xi|^2 \leqslant \xi \cdot \mathbf{a}(x)\xi \leqslant \Lambda|\xi|^2. \tag{2.1}$$

For each Borel set $U \subseteq \mathbb{R}^d$, we denote by $\mathcal{F}_U$ the σ-algebra generated by the mappings

$$\mathbf{a} \mapsto \int_{\mathbb{R}^d} \phi\, \mathbf{a}, \quad \phi \in C_c^\infty(U),$$

where $C_c^\infty(U)$ denotes the set of smooth functions with compact support in U. We also use the shorthand $\mathcal{F} := \mathcal{F}_{\mathbb{R}^d}$. For each $y \in \mathbb{R}^d$, we denote by $T_y : \Omega \to \Omega$ the action of translation by y on Ω, which is such that, for every $x \in \mathbb{R}^d$,

$$T_y\mathbf{a}(x) = \mathbf{a}(x + y).$$

Translations can also operate on events, that is, for every $E \in \mathcal{F}$ and $y \in \mathbb{R}^d$, we set $T_yE := \{T_y\mathbf{a} \ : \ \mathbf{a} \in E\}$.

We assume that we are given a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ that, in addition to (2.1), satisfies the following properties:

- stationarity with respect to $\mathbb{Z}^d$ translations: for every $z \in \mathbb{Z}^d$, we have

$$\mathbb{P} \circ T_z = \mathbb{P}; \tag{2.2}$$

- unit range of dependence: whenever two Borel sets $U, V \subseteq \mathbb{R}^d$ are at least at distance 1 away from one another, we have that $\mathcal{F}_U$ and $\mathcal{F}_V$ are independent.

If the latter condition was satisfied with the constant 1 replaced by another fixed number, then we could reduce to the present setting by scaling. Similarly, if stationarity was known to hold along some lattice of $\mathbb{R}^d$, then we could use a change of coordinates to set it to be $\mathbb{Z}^d$ as in our current assumption.

2.2. General notation and function spaces

We write $\mathbb{N} = \{0, 1, \ldots\}$, and denote the open Euclidean ball centered at $x \in \mathbb{R}^d$ and of radius $r > 0$ by $B(x, r)$. We define the heat kernel at time $t > 0$ and position $x \in \mathbb{R}^d$ by

$$\Phi(t, x) := (4\pi t)^{-\frac{d}{2}} \exp\left(-\frac{|x|^2}{4t}\right). \tag{2.3}$$

For every Borel measurable set $U \subseteq \mathbb{R}^d$, we denote by $|U|$ the Lebesgue measure of U. Whenever $|U| \in (0,\infty)$, we set, for every $f \in L^1(U)$,

$$\fint_U f := \frac{1}{|U|}\int_U f = \frac{1}{|U|}\int_U f(x)\,\mathrm{d}x. \tag{2.4}$$

For each $p \in [1,\infty)$, we define the rescaled L^p norm of a measurable function f as

$$\|f\|_{\underline{L}^p(U)} := \left(\fint_U |f|^p\right)^{\frac{1}{p}}.$$

For each $\ell \in \mathbb{N}\setminus\{0\}$, we denote by $H^\ell(U)$ the classical Sobolev space, with rescaled norm

$$\|f\|_{\underline{H}^\ell(U)} := \sum_{j=0}^{\ell} |U|^{-\frac{\ell-j}{d}} \|\nabla^j f\|_{\underline{L}^2(U)}. \tag{2.5}$$

We denote by $H_0^\ell(U)$ the closure in $H^\ell(U)$ of the space $C_c^\infty(U)$ of smooth functions with compact support in U, and by $H^{-\ell}(U)$ the dual space to $H_0^\ell(U)$, equipped with the rescaled norm

$$\|f\|_{\underline{H}^{-\ell}(U)} := \sup\left\{ \fint_U fg \ : \ g \in H_0^\ell(U) \text{ such that } \|g\|_{\underline{H}^\ell(U)} \leqslant 1 \right\}. \tag{2.6}$$

In the expression above, we used the notation $\fint_U fg$ to denote the duality pairing between $H^{-\ell}(U)$ and $H_0^\ell(U)$ that is normalized in such a way that whenever f and g are smooth, the evaluation of this duality pairing coincides with the value of the integral $\fint_U fg$.

2.3. Notation for random variables

In order to have concise means to express the size of random variables at our disposal, we write, for each real random variable X and $s,\theta > 0$,

$$X \leqslant \mathcal{O}_s(\theta) \quad \Longleftrightarrow \quad \mathbb{E}\left[\exp\left(\theta^{-1}\max(X,0)\right)^s\right] \leqslant 2.$$

We also write

$$X = \mathcal{O}_s(\theta) \quad \Longleftrightarrow \quad X \leqslant \mathcal{O}_s(\theta) \text{ and } -X \leqslant \mathcal{O}_s(\theta).$$

The notation is homogeneous: we have $X \leqslant \mathcal{O}_s(\theta)$ if and only if $\theta^{-1}X \leqslant \mathcal{O}_s(1)$. Informally, the statement that $X \leqslant \mathcal{O}_s(1)$ means that the right tail of the law of X decays like $\exp(-x^s)$. The following lemma makes this precise; see Lemma A.1 of [13] for a proof.

Lemma 2.1. *For every random variable X and $s,\theta \in (0,\infty)$,*

$$X \leqslant \mathcal{O}_s(\theta) \implies \forall x \geqslant 0, \quad \mathbb{P}\left[X \geqslant \theta x\right] \leqslant 2\exp\left(-x^s\right),$$

and

$$\forall x \geqslant 0, \quad \mathbb{P}\left[X \geqslant \theta x\right] \leqslant \exp\left(-x^s\right) \implies X \leqslant \mathcal{O}_s\left(2^{\frac{1}{s}}\,\theta\right).$$

The notion of $\mathcal{O}_s$-bounded random variables is stable under averaging, as the next lemma shows (see [13], Lem. A.4 for a proof).

Lemma 2.2. *Let $s \in [1,\infty)$, μ be a measure over an arbitrary measurable space E, $\theta : E \to (0,\infty)$ be a measurable function and $(X(x))_{x\in E}$ be a jointly measurable family of nonnegative random variables. We have*

$$\forall x \in E, \quad X(x) \leqslant \mathcal{O}_s(\theta(x)) \implies \int X\,\mathrm{d}\mu \leqslant \mathcal{O}_s\left(\int \theta\,\mathrm{d}\mu\right).$$

The key mechanism by which we will witness stochastic cancellations is by appealing to the following lemma.

Lemma 2.3. *For every $s \in (1,2]$, there exists a constant $C(s) < \infty$ such that the following holds. Let $\theta > 0$, $R \geqslant 1$, $\mathcal{Z} \subseteq R\mathbb{Z}^d$, and for each $x \in \mathcal{Z}$, let $X(x)$ be an $\mathcal{F}\left(x + (-R,R)^d\right)$-measurable* centered *random variable such that $X(x) = \mathcal{O}_s(\theta(x))$. We have*

$$\sum_{x\in\mathcal{Z}} X(x) = \mathcal{O}_s\left(C\left(\sum_{x\in\mathcal{Z}} \theta(x)^2\right)^{\frac{1}{2}}\right). \tag{2.7}$$

This lemma is a consequence of Lemmas A.7 and A.11 from [13]. Notice that when specializing Lemma 2.3 to the case when $\theta(x) \equiv \theta$ does not depend on x, and denoting by N the cardinality of $\mathcal{Z}$, we can rewrite the right side of (2.7) as $\mathcal{O}_s\left(CN^{\frac{1}{2}}\theta\right)$. The term $N^{\frac{1}{2}}$ is consistent with the scaling of the central limit theorem.

2.4. Definition of homogenized matrix

We now introduce notions related to stationary random fields and solutions, and recall the definition of the homogenized coefficients in terms of correctors. A *stationary random field* is a measurable mapping $f : \mathbb{R}^d \times \Omega \to \mathbb{R}^n$ (for some $n \in \mathbb{N}$) such that for every $x \in \mathbb{R}^d$, $z \in \mathbb{Z}^d$ and $\mathbf{a} \in \Omega$,

$$f(x+z,\mathbf{a}) = f(x, T_z\mathbf{a}).$$

We may also simply say that the mapping f is *stationary*. For instance, the mapping $x \mapsto \mathbf{a}(x)$ itself is stationary, so the terminology is consistent with the definition in (2.2). As is standard with random objects, most of the time we do not display that a random field f depends explicitly on $\mathbf{a}$, and simply write $f(x)$ in place of $f(x,\mathbf{a})$. Extending the notation introduced in (2.4), whenever $f \in L^1_{\mathrm{loc}}\left(\mathbb{R}^d\right)$ is a stationary random field, we write

$$⨍_{\mathbb{R}^d} f := \lim_{r\to\infty} ⨍_{|x|\leqslant r} f(x)\,\mathrm{d}x = \mathbb{E}\left[\int_{[0,1]^d} f(x)\,\mathrm{d}x\right]. \tag{2.8}$$

That this limit exists and equals the expectation on the right side follows from the ergodic theorem, see [2]. For every $p \in [1,\infty]$, we write

$$\mathcal{L}^p := \left\{f \ : \ f \in L^p_{\mathrm{loc}}\left(\mathbb{R}^d\right) \text{ is a stationary random field}\right\},$$

equipped with the norm

$$\|f\|_{\mathcal{L}^p} := \left(⨍_{\mathbb{R}^d} |f|^p\right)^{\frac{1}{p}}.$$

In the case $p = \infty$, the right side is interpreted as

$$\lim_{r\to\infty} \|f\|_{L^\infty(B(0,r))},$$

which is also the essential supremum of the random variable $\|f\|_{L^\infty([0,1]^d)}$. We denote by $\mathcal{L}^2_{\mathrm{pot}}$ the completion in $\mathcal{L}^2$ of the set

$$\left\{\nabla f \ : \ f \in C^\infty\left(\mathbb{R}^d\right) \text{ is a stationary random field}\right\}.$$

We also define

$$\mathcal{H}^1 := \left\{f \in \mathcal{L}^2 \ : \ \nabla f \in \mathcal{L}^2\right\},$$

equipped with the norm

$$\|f\|_{\mathcal{H}^1} := \left(⨍_{\mathbb{R}^d} \left(|f|^2 + |\nabla f|^2\right)\right)^{\frac{1}{2}}.$$

Any element of $\mathcal{L}^2_{\rm pot}$ can be represented as the gradient of some function $f \in H^1_{\rm loc}\left(\mathbb{R}^d\right)$, and such a function f is defined uniquely up to a constant. However, due to the indeterminacy of this constant, the function f itself may fail to be a stationary field, that is, we do not necessarily have $f \in \mathcal{H}^1$. We will always write elements of $\mathcal{L}^2_{\rm pot}$ in the form ∇f, bearing this caveat in mind. The functions $(v_k, k \in \mathbb{N})$ are defined as elements of $\mathcal{H}^1$. The equation (1.2) is interpreted, for $k = 0$, as

$$\forall w \in \mathcal{H}^1, \quad \fint_{\mathbb{R}^d} \left(w\, v_0 + \nabla w \cdot \mathbf{a} \nabla v_0\right) = -\fint_{\mathbb{R}^d} \nabla w \cdot \mathbf{a}\xi,$$

and, for every $k \geqslant 1$, as

$$\forall w \in \mathcal{H}^1, \quad \fint_{\mathbb{R}^d} \left(2^{-k} w\, v_k + \nabla w \cdot \mathbf{a} \nabla v_k\right) = \fint_{\mathbb{R}^d} 2^{-k} w\, v_{k-1}. \tag{2.9}$$

For each $f \in \mathcal{L}^2$, we define the norm dual to the $\mathcal{H}^1$ norm by setting

$$\|f\|_{\mathcal{H}^{-1}} := \sup\left\{ \fint_{\mathbb{R}^d} fg \ : \ \|g\|_{\mathcal{H}^1} \leqslant 1 \right\},$$

and we denote by $\mathcal{H}^{-1}$ the completion of $\mathcal{L}^2$ with respect to this norm. An example of an element of $\mathcal{H}^{-1}$ is v_{-1}, see (1.1). By definition, for each $g \in \mathcal{H}^1$, the mapping

$$\begin{cases} \mathcal{L}^2 \to \mathbb{R} \\ f \ \mapsto \fint_{\mathbb{R}^d} fg \end{cases} \tag{2.10}$$

extends to a continuous linear functional over $\mathcal{H}^{-1}$. Abusing notation slightly, we keep the same notation for the extension. By an integration by parts, we see that (2.9) also makes sense for $k = 0$, provided that the right side is understood in this extended sense (which is the canonical duality pairing between the spaces $\mathcal{H}^1$ and $\mathcal{H}^{-1}$).

Using the identity (2.8) and stationarity, one can check the following integration by parts formula: for every $f \in \mathcal{H}^1$ and $G \in (\mathcal{H}^1)^d$, we have

$$\fint_{\mathbb{R}^d} \nabla f \cdot G = -\fint_{\mathbb{R}^d} f\, \nabla \cdot G. \tag{2.11}$$

If we only assume $G \in \left(\mathcal{L}^2\right)^d$, then this formula allows to interpret $\nabla \cdot G$ as an element of $\mathcal{H}^{-1}$; similarly, if $f \in \mathcal{L}^2$, then we can interpret ∇f as an element of $\left(\mathcal{H}^{-1}\right)^d$.

The gradient of the *corrector* in the direction of $\xi \in \mathbb{R}^d$ is the unique $\nabla \phi^{(\xi)} \in \mathcal{L}^2_{\rm pot}$ that is a weak solution of

$$-\nabla \cdot \mathbf{a}\left(\xi + \nabla \phi^{(\xi)}\right) = 0. \tag{2.12}$$

This equation is interpreted as

$$\text{for every } \nabla f \in \mathcal{L}^2_{\rm pot}, \quad \fint_{\mathbb{R}^d} \nabla f \cdot \mathbf{a}\left(\xi + \nabla \phi^{(\xi)}\right) = 0.$$

The existence of $\nabla \phi^{(\xi)}$ can be obtained by considering, for every $\lambda \in (0, 1]$, the approximation $\phi^{(\xi)}_\lambda \in \mathcal{H}^1$ which is a weak solution of the equation

$$\lambda \phi^{(\xi)}_\lambda - \nabla \cdot \mathbf{a}\left(\xi + \nabla \phi^{(\xi)}_\lambda\right) = 0. \tag{2.13}$$

It is indeed straightforward to verify that $\nabla \phi^{(\xi)}_\lambda$ is bounded in $\mathcal{L}^2$ uniformly over $\lambda \in (0, 1]$, and that any weak limit must be a solution of (2.12). Moreover, the weak convergence of $\nabla \phi^{(\xi)}_\lambda$ to $\nabla \phi^{(\xi)}$ in $\mathcal{L}^2$ as λ tends to 0 can be improved to strong convergence, see *e.g.* (8.5) of [52]. That is, we have

$$\lim_{\lambda \to 0} \fint_{\mathbb{R}^d} |\nabla \phi^{(\xi)}_\lambda - \nabla \phi^{(\xi)}|^2 = 0. \tag{2.14}$$

By definition, the homogenized matrix $\bar{\mathbf{a}}$ is such that, for every $\xi \in \mathbb{R}^d$,

$$\xi \cdot \bar{\mathbf{a}}\xi = \fint_{\mathbb{R}^d} \left(\xi + \nabla\phi^{(\xi)}\right) \cdot \mathbf{a}\left(\xi + \nabla\phi^{(\xi)}\right). \tag{2.15}$$

For the remainder of the paper, we will keep the unit vector $\xi \in \mathbb{R}^d$ fixed, and drop it from the notation: in particular, we now simply write ϕ in place of $\phi^{(\xi)}$.

3. Multiscale representation

In this section, we give a multiscale representation of the homogenized matrix. That is, we rewrite $\bar{\mathbf{a}}$ as the sum of a first term taking the form of an average of very local objects, and correction terms that involve progressively larger and larger length scales. This increase of the relevant length scale means that producing one relevant sample for the calculation of the correction term becomes progressively more difficult. Yet, the actual size of these correction terms becomes smaller and smaller, and thus fewer samples need to be averaged out in order to approximate the expected value of the quantity up to a given accuracy. Moreover, this beneficial effect more than compensates for the increase in computational effort required to obtain a single sample, and this is the main reason for the efficiency of the approach presented here.

Proposition 3.1 (Multiscale representation of $\bar{\mathbf{a}}$)**.** *Recall that we fixed $\xi \in \mathbb{R}^d$ of unit length, and that $v_{-1}, v_0, v_1, \dots$ are defined in (1.1) and (1.2). For each $n \in \mathbb{N}$, the limit*

$$D_n := \lim_{\lambda \to 0} \fint_{\mathbb{R}^d} v_n \left(\lambda - \nabla \cdot \mathbf{a}\nabla\right)^{-1} v_n \tag{3.1}$$

exists and is finite. Moreover,

$$\xi \cdot \bar{\mathbf{a}}\xi = \fint_{\mathbb{R}^d} \xi \cdot \mathbf{a}\xi - \sum_{k=0}^{n} 2^k \fint_{\mathbb{R}^d} \left(v_{k-1}v_k + v_k^2\right) - D_n. \tag{3.2}$$

Remark 3.2. In the summand indexed by $k = 0$ on the right side of (3.2), we have

$$\fint_{\mathbb{R}^d} v_{-1}v_0 = -\fint_{\mathbb{R}^d} \mathbf{a}\xi \cdot \nabla v_0,$$

and the left side of the identity above is interpreted as the duality pairing between $\mathcal{H}^{-1}$ and $\mathcal{H}^1$, as explained below (2.10). All the other terms on the right side of (3.2) involve functions in $\mathcal{L}^2$ and are thus understood as in (2.8).

Remark 3.3. One can show in great generality (using only the ergodicity of the coefficient field instead of the short-range dependence assumption) that

$$\lim_{n \to \infty} D_n = 0.$$

We do not provide the argument for this fact here. The interested reader can reconstruct it from the quantitative analysis of this term provided in the next section; see also Theorem 5.1 of [52].

Proof of Proposition 3.1. By (2.12), we have

$$\fint_{\mathbb{R}^d} \nabla\phi \cdot \mathbf{a}\left(\xi + \nabla\phi\right) = 0. \tag{3.3}$$

Using also (2.15), we deduce that

$$\xi \cdot \bar{\mathbf{a}}\xi = \fint_{\mathbb{R}^d} \left(\xi + \nabla\phi\right) \cdot \mathbf{a}\left(\xi + \nabla\phi\right) = \fint_{\mathbb{R}^d} \xi \cdot \mathbf{a}\xi - \fint_{\mathbb{R}^d} \nabla\phi \cdot \mathbf{a}\nabla\phi. \tag{3.4}$$

For each $\lambda > 0$, we define the resolvent operator

$$R_\lambda : \begin{cases} \mathcal{H}^{-1} \to \mathcal{H}^1 \\ f \quad \mapsto (\lambda - \nabla \cdot \mathbf{a}\nabla)^{-1} f. \end{cases}$$

The function $u = R_\lambda f$ is interpreted as the unique element of $\mathcal{H}^1$ such that, for every $v \in \mathcal{H}^1$,

$$⨍_{\mathbb{R}^d} (\lambda uv + \nabla u \cdot \mathbf{a}\nabla v) = ⨍_{\mathbb{R}^d} fv,$$

the right side of this identity being understood as explained below (2.10). For every $\lambda, \mu > 0$, we have the resolvent formula

$$R_\lambda = R_\mu + (\mu - \lambda) R_\lambda R_\mu. \tag{3.5}$$

In particular, we have $R_\lambda R_\mu = R_\mu R_\lambda$. Moreover, the operator R_λ is self-adjoint, in the sense that for every $f, g \in \mathcal{H}^{-1}$,

$$⨍_{\mathbb{R}^d} f R_\lambda g = ⨍_{\mathbb{R}^d} g R_\lambda f. \tag{3.6}$$

By (3.3) and (2.14), we have

$$⨍_{\mathbb{R}^d} \nabla\phi \cdot \mathbf{a}\nabla\phi = -⨍_{\mathbb{R}^d} \mathbf{a}\xi \cdot \nabla\phi = -\lim_{\lambda\to 0} ⨍_{\mathbb{R}^d} \mathbf{a}\xi \cdot \nabla\phi_\lambda = \lim_{\lambda\to 0} ⨍_{\mathbb{R}^d} v_{-1} R_\lambda v_{-1}.$$

The completion of the proof will follow from this identity and a repeated application of the resolvent formula. To start with, given any family of numbers $\lambda, \mu_0, \dots, \mu_n \in (0, \infty)$, an inductive argument based on the identity (3.5) yields that

$$R_\lambda = \left(\sum_{k=0}^{n} (\mu_0 - \lambda) \cdots (\mu_{k-1} - \lambda) R_{\mu_0} \cdots R_{\mu_k} \right) + (\mu_0 - \lambda) \cdots (\mu_n - \lambda) R_{\mu_0} \cdots R_{\mu_n} R_\lambda. \tag{3.7}$$

For the summand with $k = 0$, the product $(\mu_0 - \lambda) \cdots (\mu_{k-1} - \lambda)$ appearing above is interpreted as being 1. Note that, for every $k \in \mathbb{N}$, we have

$$v_k = 2^{-k} R_{2^{-k}} v_{k-1}. \tag{3.8}$$

We define recursively

$$v_{-1,\lambda} := v_{-1}, \quad \forall k \in \{0, \dots, n\}, \quad v_{k,\lambda} := (2^{-k} - \lambda) R_{2^{-k}} v_{k-1,\lambda}.$$

For each $\lambda \in (0, 2^{-n})$, we now apply the formula (3.7) with the choice

$$(\mu_0, \mu_1, \dots, \mu_{2n}, \mu_{2n+1}) = \left(1, 1, 2^{-1}, 2^{-1}, \dots, 2^{-n}, 2^{-n}\right).$$

For the summand in (3.7) with k replaced by $2k+1$ (with $k \in \{0, \dots, n\}$), we use the commutation between resolvents and the symmetry (3.6) to obtain

$$⨍_{\mathbb{R}^d} v_{-1} (\mu_0 - \lambda) \cdots (\mu_{2k} - \lambda)\, R_{\mu_0} \cdots R_{\mu_{2k+1}} v_{-1} = \left(2^{-k} - \lambda\right)^{-1} ⨍_{\mathbb{R}^d} v_{k,\lambda}^2.$$

By the same reasoning, we can rewrite the contribution of the summand in (3.7) with k replaced by $2k$ (with $k \in \{0, \dots, n\}$) as

$$⨍_{\mathbb{R}^d} v_{-1} \,(\mu_0 - \lambda) \cdots (\mu_{2k-1} - \lambda)\, R_{\mu_0} \cdots R_{\mu_{2k}} v_{-1} = \left(2^{-k} - \lambda\right)^{-1} ⨍_{\mathbb{R}^d} v_{k-1,\lambda}\, v_{k,\lambda}.$$

Summing over all indices, and using a similar reasoning for the remainder term, we obtain that

$$⨍_{\mathbb{R}^d} v_{-1} R_\lambda v_{-1} = \sum_{k=0}^{n} \left(2^{-k} - \lambda\right)^{-1} ⨍_{\mathbb{R}^d} \left(v_{k-1,\lambda} v_{k,\lambda} + v_{k,\lambda}^2\right) + ⨍_{\mathbb{R}^d} v_{n,\lambda}\, R_\lambda v_{n,\lambda}.$$

Note that $v_{k,\lambda}$ is a scalar multiple of v_k, and that this scalar tends to 1 as λ tends to 0. Hence, the left side and each summand in the sum indexed by k on the right side of the identity above converges as λ tends to 0. It follows that the limit

$$\lim_{\lambda \to 0} ⨍_{\mathbb{R}^d} v_{n,\lambda}\, R_\lambda v_{n,\lambda} = \lim_{\lambda \to 0} ⨍_{\mathbb{R}^d} v_{n,}\, R_\lambda v_n =: D_n$$

is well-defined and finite, and using also (3.4), that the formula (3.2) holds. □

4. Quantitative estimates

Recall that we have fixed a vector $\xi \in \mathbb{R}^d$ of unit norm throughout the paper. The proof of Theorem 1.1 relies on estimates on the solution u of the initial-value problem

$$\begin{cases} \partial_t u - \nabla \cdot \mathbf{a} \nabla u = 0 & \text{in} \quad (0,\infty) \times \mathbb{R}^d, \\ u(0,\cdot) = \nabla \cdot \mathbf{a}\xi & \text{on} \quad \mathbb{R}^d. \end{cases} \tag{4.1}$$

The study of this problem was initiated in [50] where suboptimal estimates were derived. The sharp exponent of decay in time was obtained in [37], with polynomial moments controlled. With a very different proof, the stochastic integrability of this result was improved to almost Gaussian tails in [36]. A variation of this argument is exposed in Chapter 9 of [13].

Theorem 4.1 ([36])**.**

(1) *For every $\sigma \in (0,2)$, there exists $C(\sigma,\Lambda,d) < \infty$ such that for every $t \geqslant 1$ and $x \in \mathbb{R}^d$,*

$$|u(t,x)| \leqslant \mathcal{O}_\sigma\left(C t^{-\frac{1}{2}-\frac{d}{4}}\right).$$

(2) *For every $\delta \in \left(0, \frac{1}{2} + \frac{d}{4}\right)$, there exist $\sigma(\delta,d) > 2$ and $C(\delta,\Lambda,d) < \infty$ such that for every $t \geqslant 1$ and $x \in \mathbb{R}^d$,*

$$|u(t,x)| \leqslant \mathcal{O}_\sigma\left(C t^{-\frac{1}{2}-\frac{d}{4}+\delta}\right). \tag{4.2}$$

We will only use the first part of Theorem 4.1 once, in the course of the proof of Proposition 4.5, in the form of the L^2 estimate

$$⨍_{\mathbb{R}^d} u^2(t,\cdot) \leqslant C t^{-1-\frac{d}{2}}. \tag{4.3}$$

In order to conclude for exponentially decaying tails as in the statement of Theorem 1.1, it is crucial to be able to choose an exponent $\sigma \geqslant 2$ in the estimate on the size of u (and for convenience, we will in fact choose $\sigma > 2$); this is the main motivation for stating the second part of Theorem 4.1. The first part of Theorem 4.1 matches the results found in [36] and Theorem 9.1 of [13]. In order to obtain the second part of the statement as a consequence, we can use the following basic deterministic estimate, a proof of which can be found in Lemma 9.2 of [13].

Lemma 4.2 (Deterministic bounds on u)**.** *There exists $C(\Lambda,d) < \infty$ such that for every $t > 0$,*

$$\|u(t,\cdot)\|_{L^\infty(\mathbb{R}^d)} + t^{\frac{1}{2}} \|\nabla u(t,\cdot)\|_{L^\infty(\mathbb{R}^d)} \leqslant C t^{-\frac{1}{2}}.$$

Proof of part (2) of Theorem 4.1. Let $\sigma > 2$ and $\tau \in (0,2)$ be exponents that will be fixed in terms of $\delta > 0$ and the dimension d in the course of the proof. By the first part of the theorem, there exists a constant $C(\tau, \Lambda, d) < \infty$ such that for every $t \geqslant 1$ and $x \in \mathbb{R}^d$,

$$|u(t,x)| \leqslant \mathcal{O}_\tau\left(Ct^{-\frac{1}{2}-\frac{d}{4}}\right).$$

Explicitly, this means that

$$\mathbb{E}\left[\exp\left(\left(C^{-1}t^{\frac{1}{2}+\frac{d}{4}}|u(t,x)|\right)^\tau\right)\right] \leqslant 2.$$

It follows from Lemma 4.2 that the random variable $|u(t,x)|$ is bounded, uniformly over $t \geqslant 1$ and $x \in \mathbb{R}^d$. Hence, for a constant $C(\tau, \sigma, \Lambda, d) < \infty$,

$$\mathbb{E}\left[\exp\left(C^{-\sigma}t^{\tau\left(\frac{1}{2}+\frac{d}{4}\right)}|u(t,x)|^\sigma\right)\right] \leqslant 2,$$

that is,

$$|u(t,x)| \leqslant \mathcal{O}_\sigma\left(Ct^{-\frac{\tau}{\sigma}\left(\frac{1}{2}+\frac{d}{4}\right)}\right).$$

This is (4.2) with δ given by

$$\delta = \left(1-\frac{\tau}{\sigma}\right)\left(\frac{1}{2}+\frac{d}{4}\right).$$

Since $\sigma > 2$ and $\tau < 2$ can be chosen arbitrarily close to 2, any exponent $\delta \in \left(0, \frac{1}{2}+\frac{d}{4}\right)$ can be represented in this way. □

We also record the following useful lemma allowing to localize the dependency of u on the coefficient field; see Lemma 9.4 of [13] for a proof.

Lemma 4.3 (Localization of u)**.** *There exist a constant $C(\Lambda, d) < \infty$ and, for each $r \in [2,\infty)$, $t \in (0, r^2]$ and $x \in \mathbb{R}^d$, an $\mathcal{F}(B(x,r))$-measurable random variable $u'(r,t,x)$ such that*

$$|u(t,x) - u'(r,t,x)| \leqslant Ct^{-\frac{1}{2}}\exp\left(-\frac{r^2}{Ct}\right), \tag{4.4}$$

and

$$|\nabla u(t,x) - \nabla u'(r,t,x)| \leqslant Ct^{-1}\exp\left(-\frac{r^2}{Ct}\right). \tag{4.5}$$

The function v_k can be represented as a time integral of the function $u(t,\cdot)$, and the main contribution to this integral is for $t \simeq 2^k$. The previous results concerning the function u can thus be translated into information on v_k.

Proposition 4.4 (Quantitative bounds on v_k)**.**

(1) *There exists $C(\Lambda, d) < \infty$ such that for every $k \in \mathbb{N}$,*

$$\|v_k\|_{L^\infty(\mathbb{R}^d)} \leqslant C2^{-\frac{k}{2}}. \tag{4.6}$$

(2) *There exist $C(\Lambda, d) < \infty$ and, for each $r \in [2,\infty)$, $k \in \mathbb{N}$ and $x \in \mathbb{R}^d$, an $\mathcal{F}(B(x,r))$-measurable random variable $v_k(r,x)$ such that*

$$|v_k(x) - v_k(r,x)| \leqslant C2^{-\frac{k}{2}}\exp\left(-C^{-1}2^{-k}r^2\right), \tag{4.7}$$

and

$$|\nabla v_k(x) - \nabla v_k(r,x)| \leqslant C2^{-k}\exp\left(-C^{-1}2^{-k}r^2\right). \tag{4.8}$$

(3) *For every $\delta > 0$, there exist $\sigma(\delta, d) > 2$ and $C(\delta, \Lambda, d) < \infty$ such that for every $k \in \mathbb{N}$ and $x \in \mathbb{R}^d$,*

$$|v_k(x)| \leqslant \mathcal{O}_\sigma\left(C2^{-k\left(\frac{1}{2}+\frac{d}{4}-\delta\right)}\right). \tag{4.9}$$

Proof. We decompose the proof into three steps.

Step 1. Transfering information on $u(t,\cdot)$ onto information on the v_k's relies on the observation that, for every $\lambda > 0$ and $f \in \mathcal{L}^2$,

$$R_\lambda f = (\lambda - \nabla \cdot \mathbf{a}\nabla)^{-1} f = \int_0^{+\infty} e^{-\lambda t} P(t) f \, \mathrm{d}t, \tag{4.10}$$

where $P(t) = \exp(t \nabla \cdot \mathbf{a}\nabla)$ is the semigroup associated with the evolution operator $\partial_t - \nabla \cdot \mathbf{a}\nabla$. This identity can be extended to the case $f = \nabla \cdot \mathbf{a}\xi$, and we thus have in particular that

$$v_0 = \int_0^{\infty} e^{-t} \, u(t,\cdot) \, \mathrm{d}t. \tag{4.11}$$

Since integrals of the form of (4.10) or (4.11) will be iterated multiple times, it is convenient to rewrite them using probabilistic notation. That is, denoting by $T^{(\lambda)}$ an exponential random variable of parameter λ which is independent of any other quantity in the problem, and by E the expectation over this random variable only, we can rewrite (4.10) in the form

$$\lambda R_\lambda f = E\left[P(T^{(\lambda)}) f\right]. \tag{4.12}$$

Denote by $(T_k)_{k\in\mathbb{N}}$ a family of independent exponential random variables, of respective parameters $\left(2^{-k}\right)_{k\in\mathbb{N}}$, and for every $k \in \mathbb{N}$, set

$$S_k := \sum_{j=0}^{k} T_j. \tag{4.13}$$

We keep denoting by E the expectation over these random variables. Recalling (3.8) and using the semigroup property of $(P(t))_{t\geqslant 0}$, we deduce that for every $k \in \mathbb{N}$,

$$v_k = E\left[P\left(S_k\right) v_{-1}\right] = E\left[u\left(S_k, \cdot\right)\right]. \tag{4.14}$$

In view of this relation, we can now transfer information about u onto v_k provided that we have some information on the typical behavior of S_k. As a useful guide for the intuition, we remark that

$$E[S_k] = \sum_{j=0}^{k} 2^j = 2^{k+1} - 1,$$

so heuristically, we hope that any bound on $u(t,\cdot)$ transfers into a bound on v_k after the substitution of t by 2^k.

Step 2. We prove (4.9). In view of Theorem 4.1 and (4.14), we need to know that S_k is rarely much smaller than 2^k. The following result is shown in (5.12) of [52]: for every $\beta > 0$, there exists a constant $C(\beta) < \infty$ such that for every $k \in \mathbb{N}$,

$$E\left[(1 + S_k)^{-\beta}\right] \leqslant C2^{-k\beta}. \tag{4.15}$$

By Theorem 4.1 and Lemma 2.2, we have, for every $x \in \mathbb{R}^d$,

$$E\left[\left|u\left(S_k, x\right)\right| \mathbf{1}_{\{S_k \geqslant 1\}}\right] \leqslant \mathcal{O}_s\left(CE\left[(1+S_k)^{-\frac{1}{2}-\frac{d}{4}+\delta}\right]\right) \leqslant \mathcal{O}_s\left(C2^{-k\left(\frac{1}{2}+\frac{d}{4}-\delta\right)}\right).$$

To control the behavior of this term on the event $S_k \in [0,1]$, we use that the T_j's are nonnegative and independent to write, for every $s \leqslant 1$,

$$P[S_k \leqslant s] \leqslant \prod_{j=0}^{k} P[T_j \leqslant s] \leqslant \prod_{j=0}^{k} \left(1 - \exp\left(-2^{-j} s\right)\right) \leqslant \prod_{j=0}^{k} 2^{-j} s = 2^{-\frac{k(k+1)}{2}} s^{k+1}. \tag{4.16}$$

This and Lemma 4.2 imply that

$$\begin{aligned} E\left[|u(S_k,x)|\, \mathbf{1}_{\{S_k \leqslant 1\}}\right] &\leqslant \sum_{j=0}^{\infty} E\left[|u(S_k,x)|\, \mathbf{1}_{\{2^{-j-1} < S_k \leqslant 2^{-j}\}}\right] \\ &\leqslant C 2^{-\frac{k(k+1)}{2}} \sum_{j=0}^{\infty} 2^{\frac{j}{2}} 2^{-j(k+1)} \leqslant C 2^{-\frac{k(k+1)}{2}}. \end{aligned} \tag{4.17}$$

This is largely sufficient to complete the proof of the estimate in (4.9). The proof of (4.6) is similar, only simpler, using Lemma 4.2 in place of Theorem 4.1.

Step 3. We prove (4.7). Recalling the notation $u'(r,t,x)$ introduced in Lemma 4.3, we define, for each $r \in [2,\infty)$ and $x \in \mathbb{R}^d$, the $\mathcal{F}(B(x,r))$-measurable random variable

$$v_k(r,x) := E\left[u'\left(r, S_k, x\right) \mathbf{1}_{\{S_k \leqslant r^2\}}\right].$$

We need an upper bound for the probability of the event that S_k is large. We obtain this by writing, for every $s \geqslant 0$,

$$\begin{aligned} P\left[S_k \geqslant s\right] \leqslant \mathbb{E}\left[\exp\left(2^{-(k+1)}\left(S_k - s\right)\right)\right] &= \prod_{j=0}^{k} \frac{2^{-j}}{2^{-j} - 2^{-(k+1)}} \exp\left(-2^{-(k+1)} s\right) \\ &= \prod_{j=1}^{k+1} \frac{1}{1 - 2^{-j}} \exp\left(-2^{-(k+1)} s\right) \leqslant C \exp\left(-2^{-(k+1)} s\right). \end{aligned} \tag{4.18}$$

We now decompose the error into

$$\begin{aligned} |v_k(x) - v_k(r,x)| &\leqslant \mathbb{E}\left[|u'\left(r, S_k, x\right) - u\left(S_k, x\right)|\, \mathbf{1}_{\{S_k \leqslant r^2\}}\right] + \mathbb{E}\left[|u\left(S_k, x\right)|\, \mathbf{1}_{\{S_k > r^2\}}\right] \\ &\leqslant \sum_{j=0}^{\lfloor 2^{-k} r^2 \rfloor} \mathbb{E}\left[|u'\left(r, S_k, x\right) - u\left(S_k, x\right)|\, \mathbf{1}_{\{j 2^k \leqslant S_k \leqslant (j+1) 2^k\}}\right] + \mathbb{E}\left[|u\left(S_k, x\right)|\, \mathbf{1}_{\{S_k > r^2\}}\right], \end{aligned}$$

and analyze each of these terms in turn. By Lemma 4.3 and (4.18), we have

$$\begin{aligned} \sum_{j=0}^{\lfloor 2^{-k} r^2 \rfloor} \mathbb{E}\left[|u'\left(r, S_k, x\right) - u\left(S_k, x\right)|\, \mathbf{1}_{\{j 2^k \leqslant S_k \leqslant (j+1) 2^k\}}\right] &\leqslant C 2^{-\frac{k}{2}} \sum_{j=0}^{\lfloor 2^{-k} r^2 \rfloor} \exp\left(-\frac{r^2}{C j 2^k} - \frac{j}{2}\right) \\ &\leqslant C 2^{-\frac{k}{2}} \exp\left(-\frac{r^2}{C 2^k}\right). \end{aligned}$$

Moreover, by Lemma 4.2 and (4.18), we have

$$\mathbb{E}\left[|u\left(S_k, x\right)|\, \mathbf{1}_{\{S_k > r^2\}}\right] \leqslant C r^{-1} \exp\left(-2^{-(k+1)} r^2\right).$$

Combining these estimates yields (4.7). The proof of (4.8) is similar, except that we appeal to (4.5) instead of (4.4). □

We denote

$$w_0 := -\mathbf{a}\xi \cdot \nabla v_0 + v_0^2, \quad \forall k \in \mathbb{N} \setminus \{0\}, \quad w_k := v_{k-1} v_k + v_k^2.$$

For every $k \in \mathbb{N}$, we have that $w_k \in \mathcal{L}^1$, and by Proposition 3.1, for D_n as in (3.1),

$$\xi \cdot \bar{\mathbf{a}}\xi = \fint_{\mathbb{R}^d} \xi \cdot \mathbf{a}\xi - \sum_{k=0}^{n} 2^k \fint_{\mathbb{R}^d} w_k - D_n. \tag{4.19}$$

We next estimate the size of the remainder term D_n.

Proposition 4.5 (Remainder estimate)**.** *There exists $C(\Lambda, d) < \infty$ such that for every $n \in \mathbb{N}$,*

$$0 \leqslant D_n \leqslant C 2^{-\frac{nd}{2}}.$$

Proof. We keep denoting by S_k the random variable defined in (4.13), and we let S_k' be an independent copy of S_k. We also let $T^{(\lambda)}$ be an independent exponential random variable of parameter λ, and denote the expectation with respect to these random variables by E. We recall that the introduction of these random variables allows us for convenient representations such as those in (4.12) and (4.14). Combining these representations with the definition of D_k in (3.1), we obtain that

$$D_k = \lim_{\lambda \to 0} \lambda^{-1} \fint_{\mathbb{R}^d} E\left[P\left(S_k\right) v_{-1} P(T^{(\lambda)} + S_k') v_{-1}\right].$$

Since $P(t)$ is self-adjoint in $\mathcal{L}^2$, we have

$$\begin{aligned} \fint_{\mathbb{R}^d} E\left[P\left(S_k\right) v_{-1} P(T^{(\lambda)} + S_k') v_{-1}\right] &= \fint_{\mathbb{R}^d} E\left[\left(P\left(\frac{S_k + S_k' + T^{(\lambda)}}{2}\right) v_{-1}\right)^2\right] \\ &= \fint_{\mathbb{R}^d} E\left[u^2\left(\frac{S_k + S_k' + T^{(\lambda)}}{2}, \cdot\right)\right], \end{aligned}$$

and thus

$$D_k = \int_0^{+\infty} \fint_{\mathbb{R}^d} E\left[u^2\left(\frac{S_k + S_k' + t}{2}, \cdot\right)\right] \mathrm{d}t \geqslant 0. \tag{4.20}$$

In order to estimate the integral over $t \in [0, 1]$, we use independence to get that

$$P\left[S_k + S_k' \leqslant s\right] \leqslant \left(P\left[S_k \leqslant s\right]\right)^2,$$

and then proceed as in (4.16) and (4.17). For the remaining part, we use (4.3) and the triangle inequality to deduce that

$$D_k \leqslant C \int_1^{\infty} E\left[\left(S_k + S_k' + t\right)^{-1-\frac{d}{2}}\right] \mathrm{d}t \leqslant C \int_1^{\infty} E\left[\left(S_k + \frac{t}{2}\right)^{-1-\frac{d}{2}}\right] \mathrm{d}t.$$

Integrating in t and then appealing to (4.15), we obtain

$$D_k \leqslant C E\left[\left(S_k + 1\right)^{-\frac{d}{2}}\right] \leqslant C 2^{-\frac{kd}{2}},$$

as announced. □

In the next proposition, we replace the global averages $\fint_{\mathbb{R}^d}$ appearing in (4.19) by averages against a heat kernel mask. Recall the definition of Φ in (2.3).

Proposition 4.6 (CLT cancellations)**.** *For every $\delta > 0$, there exist $\sigma(\delta, d) > 1$ and $C(\delta, \Lambda, d) < \infty$ such that for every $k \in \mathbb{N} \setminus \{0\}$, $x \in \mathbb{R}$ and $s > 0$,*

$$\left| \int_{\mathbb{R}^d} w_k(y) \Phi(s, x-y) \, \mathrm{d}y - \fint_{\mathbb{R}^d} w_k \right| \leqslant \mathcal{O}_\sigma \left(C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \left(\frac{s}{2^k \log^2(2+k)} + 1 \right)^{-\frac{d}{4}} \right). \tag{4.21}$$

For $k = 0$, the same estimate holds with the additional restriction $s \geqslant 1$.

Proof. We first record the following elementary observation: for every $\sigma, \theta > 0$ and random variables X and Y, we have

$$|X| \leqslant \mathcal{O}_{2\sigma}(\theta) \text{ and } |Y| \leqslant \mathcal{O}_{2\sigma}(\theta) \implies |XY| \leqslant \mathcal{O}_\sigma\left(\theta^2\right). \tag{4.22}$$

Indeed, this follows from

$$\begin{aligned} \mathbb{E}\left[\exp\left(\left(\theta^{-2}|XY|\right)^\sigma\right)\right] &\leqslant \mathbb{E}\left[\exp\left(\frac{1}{2}\left(\theta^{-1}|X|\right)^{2\sigma} + \frac{1}{2}\left(\theta^{-1}|Y|\right)^{2\sigma}\right)\right] \\ &\leqslant \mathbb{E}\left[\exp\left(\left(\theta^{-1}|X|\right)^{2\sigma}\right)\right]^{\frac{1}{2}} \mathbb{E}\left[\exp\left(\left(\theta^{-1}|Y|\right)^{2\sigma}\right)\right]^{\frac{1}{2}} \leqslant 2. \end{aligned}$$

We decompose the rest of the proof into three steps.

Step 1. We set

$$r_k' := 2^{\frac{k}{2}} \log(2+k). \tag{4.23}$$

In this step, we observe that the statement (4.21) is valid when $\sqrt{s} \leqslant r_k'$. Indeed, for $k \geqslant 1$, the statement (4.21) with $\sqrt{s} \leqslant r_k'$ follows from (4.9), (4.22) and Lemma 2.2. For $k = 0$, we also need a bound on ∇v_0, which is provided by the following deterministic estimate: there exists a constant $C(\Lambda, d) < \infty$ such that for every $x \in \mathbb{R}^d$,

$$\|\nabla v_0\|_{L^2(x+\square_0)} \leqslant C. \tag{4.24}$$

This estimate is a consequence of the Caccioppoli inequality and (4.6).

Step 2. We reformulate (4.21) into an equivalent form that will be more convenient for the analysis. For every $k \in \mathbb{N}$, we denote

$$\widetilde{w}_k := w_k - \mathbb{E}[w_k].$$

We show that it suffices to prove Proposition 4.6 for $\sqrt{s} \geqslant r_k'$ and with (4.21) replaced by

$$\int_{\mathbb{R}^d} \widetilde{w}_k(y) \Phi(s, x-y) \, \mathrm{d}y = \mathcal{O}_\sigma \left(C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \left(\frac{\sqrt{s}}{r_k'} + 1 \right)^{-\frac{d}{2}} \right). \tag{4.25}$$

For every $k \geqslant 1$, the mapping $x \mapsto \mathbb{E}[w_k(x)]$ is $\mathbb{Z}^d$-periodic, and by (4.9), it is uniformly bounded by $C 2^{-k\left(1+\frac{d}{2}-\delta\right)}$. In the case $k = 0$, the mapping $x \mapsto \mathbb{E}[w_0(x)]$ is $\mathbb{Z}^d$-periodic and in $L^2\left([0,1]^d\right)$, as follows from (4.6) and the fact that $v_0 \in \mathcal{H}^1$. Hence, for every $k \in \mathbb{N}$, there exists a constant $C(k, \Lambda, d) < \infty$ such that for every $x \in \mathbb{R}^d$ and $s \geqslant 1$,

$$\left| \int_{\mathbb{R}^d} \mathbb{E}[w_k(y)] \Phi(s, x-y) \, \mathrm{d}y - \fint_{\mathbb{R}^d} w_k \right| \leqslant C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \exp\left(-C^{-1} s\right).$$

See for instance Exercise 3.7 of [13] for a proof. As a consequence, the statements (4.21) and (4.25) are equivalent, up to an adjustment of the constant C.

Step 3. Without loss of generality, it suffices to prove (4.25) for $x = 0$. For each $r \geqslant r_k'$, we define

$$\widetilde{w}_k'(r, x) := w_k'(r, x) - \mathbb{E}\left[w_k'(r, x)\right].$$

By (4.7) and (4.8), there exists $C(\Lambda, d) < \infty$ such that for every $k \in \mathbb{N}$ and $r \geqslant r'_k$,

$$|\widetilde{w}_k(x) - \widetilde{w}'_k(r, x)| \leqslant C2^{-100dk} \exp\left(-C^{-1}2^{-k}r^2\right). \tag{4.26}$$

In this step, we leave aside the case $k = 0$ and show that there exists $C(\delta, \Lambda, d) < \infty$ such that for every $k \geqslant 1$, $r \geqslant r'_k$ and $s > 0$,

$$\int_{\mathbb{R}^d} \widetilde{w}'_k(r, x)\Phi(s, x)\,\mathrm{d}x = \mathcal{O}_\sigma\left(C2^{-k\left(1+\frac{d}{2}-\delta\right)}\left(\frac{s}{r^2}+1\right)^{-\frac{d}{4}}\right). \tag{4.27}$$

We denote the cube of side length r centered at the origin by

$$\square_r := \left(-\frac{r}{2}, \frac{r}{2}\right)^d.$$

By (4.26) and the same argument as in Step 1 of this proof, we may assume that $s \geqslant r^2$. We decompose the left side of (4.27) into

$$\int_{\mathbb{R}^d} \widetilde{w}'_k(r, x)\Phi(s, x)\,\mathrm{d}x = \sum_{z \in r\mathbb{Z}^d} \int_{z+\square_r} \widetilde{w}'_k(r, x)\Phi(s, x)\,\mathrm{d}x.$$

By (4.9), (4.22) and Lemma 2.2, there exists $\sigma > 1$ such that, for each $z \in \mathbb{Z}^d$,

$$\int_{z+\square_r} \widetilde{w}'_k(r, x)\Phi(s, x)\,\mathrm{d}x = \mathcal{O}_\sigma\left(C2^{-k\left(1+\frac{d}{2}-\delta\right)}\|\Phi(s, \cdot)\|_{L^1(z+\square_r)}\right).$$

By Lemma 2.3, we obtain that

$$\sum_{z \in r\mathbb{Z}^d} \int_{z+\square_r} \widetilde{w}'_k(r, x)\Phi(s, x)\,\mathrm{d}x = \mathcal{O}_\sigma\left(C2^{-k\left(1+\frac{d}{2}-\delta\right)}\left(\sum_{z \in r\mathbb{Z}^d} \|\Phi(s, \cdot)\|^2_{L^1(z+\square_r)}\right)^{\frac{1}{2}}\right).$$

We conclude that (4.27) holds by observing that, since $s \geqslant r^2$,

$$\sum_{z \in r\mathbb{Z}^d} \|\Phi(s, \cdot)\|^2_{L^1(z+\square_r)} \leqslant Cr^d s^{-\frac{d}{2}}.$$

Step 4. In this step, we show that there exists $C(\delta, \Lambda, d) < \infty$ such that for every $k \geqslant 1$, $r \geqslant r'_k$ and $s > 0$,

$$\int_{\mathbb{R}^d} \left(\widetilde{w}'_k(2r, x) - \widetilde{w}'_k(r, x)\right)\Phi(s, x)\,\mathrm{d}x = \mathcal{O}_\sigma\left(C2^{-k\left(1+\frac{d}{2}-\delta\right)}\left(\frac{s}{r^2}+1\right)^{-\frac{d}{4}}\exp\left(-\frac{r^2}{C2^k}\right)\right). \tag{4.28}$$

The argument is similar to that of the previous step, only simpler, using only (4.26) and not requiring any appeal to (4.9).

Step 5. We complete the proof. It is clear from (4.26) that

$$\widetilde{w}_k(x) = \lim_{r \to \infty} \widetilde{w}'_k(r, x).$$

We thus decompose $\widetilde{w}_k(x)$ into

$$\widetilde{w}_k(x) = \widetilde{w}'_k(r'_k, x) + \sum_{j=0}^{+\infty} \left(\widetilde{w}'_k(2^{j+1}r'_k, x) - \widetilde{w}'_k(2^j r'_k, x)\right).$$

Applying (4.27) to the first term, (4.28) to each of the summands, and summing the result, we obtain (4.25) for $k \geqslant 1$. In the case $k = 0$, the same reasoning applies, using also the deterministic estimate on ∇v_0 provided in (4.24), and the estimate (4.8) to localize the dependency of this term on the coefficient field. □ We have now obtained almost optimal information on the behavior of w_k when tested against the heat kernel. Since we want to understand the behavior of this field against an arbitrary mask, we now upgrade this information into an $H^{-\ell}$ estimate using the following lemma, which is a rescaled version of Remark D.6 from [13]. In order to keep the presentation of the argument as simple as possible, we only state this lemma for L^2-based Sobolev spaces with integer-valued regularity exponents. Recall the definitions of the rescaled H^ℓ and $H^{-\ell}$ norms in (2.5) and (2.6).

Lemma 4.7 (Sobolev norm from heat-kernel convolutions)**.** *For every $\ell \in \mathbb{N}$, there exists a constant $C(\ell, d) < \infty$ such that for every $f \in H^{-\ell}_{\mathrm{loc}}(\mathbb{R}^d)$ and $r > 0$, we have*

$$\|f\|^2_{\underline{H}^{-\ell}(B(0,r))} \leqslant C r^{-d} \int_{\mathbb{R}^d} \exp\left(-\frac{|x|}{r}\right) \int_0^{r^2} s^{\ell-1} |f * \Phi(s,\cdot)|^2(x)\,\mathrm{d}s\,\mathrm{d}x.$$

Combining this lemma with Proposition 4.6 yields the following estimate.

Lemma 4.8 (Sobolev norm for w_k)**.** *For every $\delta > 0$, there exist $\sigma(\delta, d) > 1$ and, for every $\ell \in \mathbb{N}$ satisfying $\ell > \frac{d}{2}$, a constant $C(\delta, \ell, \Lambda, d) < \infty$ such that for every $k \in \mathbb{N}$ and $r \geqslant 1$, we have*

$$r^{-\ell} \left\| w_k - \fint_{\mathbb{R}^d} w_k \right\|_{\underline{H}^{-\ell}(B(0,r))} \leqslant \mathcal{O}_\sigma\left(C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \left(\frac{r}{2^{\frac{k}{2}} \log(2+k)} + 1 \right)^{-\frac{d}{2}} \right).$$

Proof. For convenience, we set $f := w_k - \fint_{\mathbb{R}^d} w_k$, and use the notation r'_k introduced in (4.23). We first consider the case $k \in \mathbb{N} \setminus \{0\}$. For $r \leqslant r'_k$, by Proposition 4.6, we have for every $x \in \mathbb{R}^d$ that

$$\int_0^{r^2} s^{\ell-1} |f * \Phi(s,\cdot)|^2(x)\,\mathrm{d}s = \mathcal{O}_{\sigma/2}\left(C 2^{-2k\left(1+\frac{d}{2}-\delta\right)} \int_0^{r^2} s^{\ell-1}\,\mathrm{d}s \right) = \mathcal{O}_{\sigma/2}\left(C 2^{-2k\left(1+\frac{d}{2}-\delta\right)} r^{2\ell} \right).$$

For $r \geqslant r'_k$, we have instead that

$$\begin{aligned} \int_0^{r^2} s^{\ell-1} |f * \Phi(s,\cdot)|^2(x)\,\mathrm{d}s &= \mathcal{O}_{\sigma/2}\left(C 2^{-2k\left(1+\frac{d}{2}-\delta\right)} \int_0^{r^2} s^{\ell-1} \left(\frac{\sqrt{s}}{r'_k} + 1 \right)^{-d} \mathrm{d}s \right) \\ &= \mathcal{O}_{\sigma/2}\left(C 2^{-2k\left(1+\frac{d}{2}-\delta\right)} r^{2\ell} \left(\frac{r}{r'_k} \right)^{-d} \right), \end{aligned}$$

where we used that $\ell > \frac{d}{2}$ and $r \geqslant r'_k$ for the second equality. We then obtain the result by an application of Lemma 4.7. The case $k = 0$ is obtained in the same way, except that we treat the integral over $s \in [0, 1]$ separately using the gradient estimate in (4.24). □

We are now ready to complete the proof of Theorem 1.1.

Proof of Theorem 1.1. Recall that we denote by $\chi \in C^\infty_c(\mathbb{R}^d)$ a smooth bump function of unit mass with compact support in the unit ball $B(0,1)$, and that we write $\chi_r := r^{-d}\chi(r^{-1}\cdot)$. By the definition of the rescaled H^ℓ norm in (2.5), for every $\ell \in \mathbb{N}$ and $r \geqslant 1$, we have

$$\|\chi_r\|_{\underline{H}^\ell(B(0,r))} = r^{-\ell} \|\chi\|_{\underline{H}^\ell(B(0,1))}.$$

We select ℓ to be the smallest integer such that $\ell > \frac{d}{2}$ $\left(i.e.\ \ell = \left\lceil \frac{d}{2} + \frac{1}{4} \right\rceil\right)$, and write

$$\left| \int_{\mathbb{R}^d} w_k \chi_r - \fint_{\mathbb{R}^d} w_k \right| \leqslant r^{-\ell} \|\chi\|_{\underline{H}^\ell(B(0,1))} \left\| w_k - \fint_{\mathbb{R}^d} w_k \right\|_{\underline{H}^{-\ell}(B(0,r))}.$$

An application of Lemma 4.8 thus yields, for each $\delta > 0$, that there exists $\sigma(\delta, d) > 1$ and a constant $C(\delta, \|\chi\|_{\underline{H}^\ell(\mathbb{R}^d)}, \Lambda, d) < \infty$ such that for every $k \in \mathbb{N}$ and $r \geqslant 1$,

$$\left| \int_{\mathbb{R}^d} w_k \chi_r - \fint_{\mathbb{R}^d} w_k \right| \leqslant \mathcal{O}_\sigma \left(C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \left(\frac{r}{2^{\frac{k}{2}} \log(2+k)} + 1 \right)^{-\frac{d}{2}} \right).$$

We fix $\varepsilon \in \left(0, \frac{d-1}{2d}\right)$, $n \in \mathbb{N}$, and recall from (1.5) the notation

$$r_k := 2^{n-\left(\frac{1}{2}-\varepsilon\right)k}.$$

Substituting r with r_k in the previous display, we obtain that

$$\begin{aligned} \left| \int_{\mathbb{R}^d} w_k \chi_{r_k} - \fint_{\mathbb{R}^d} w_k \right| &\leqslant \mathcal{O}_\sigma \left(C 2^{-k\left(1+\frac{d}{2}-\delta\right)} \left(\frac{2^{n-k+\varepsilon k}}{\log(2+k)} + 1 \right)^{-\frac{d}{2}} \right) \\ &\leqslant \mathcal{O}_\sigma \left(C 2^{-\frac{nd}{2} - k\left(1-\delta+\frac{\varepsilon d}{2}\right)} \log^{\frac{d}{2}}(2+k) \right). \end{aligned}$$

We select $\delta := \frac{\varepsilon d}{4} > 0$, so that

$$\sum_{k=0}^{n} 2^k \left| \int_{\mathbb{R}^d} w_k \chi_{r_k} - \fint_{\mathbb{R}^d} w_k \right| \leqslant \mathcal{O}_\sigma \left(C 2^{-\frac{nd}{2}} \right).$$

In view of Lemma 2.1 and of the fact that $\sigma > 1$, this implies the existence of a constant $c\left(\varepsilon, \|\chi\|_{\underline{H}^\ell(\mathbb{R}^d)}, \Lambda, d\right) < \infty$ such that for every $n \in \mathbb{N}$ and $t \geqslant 0$,

$$\mathbb{P}\left[\sum_{k=0}^{n} 2^k \left| \int_{\mathbb{R}^d} w_k \chi_{r_k} - \fint_{\mathbb{R}^d} w_k \right| \geqslant t 2^{-\frac{nd}{2}} \right] \leqslant 2 \exp\left(-ct\right).$$

A similar but simpler argument shows that

$$\mathbb{P}\left[\left| \int_{\mathbb{R}^d} (\xi \cdot \mathbf{a}\xi) \chi_{r_0} - \fint_{\mathbb{R}^d} \xi \cdot \mathbf{a}\xi \right| \geqslant t 2^{-\frac{nd}{2}} \right] \leqslant 2 \exp\left(-ct^2\right).$$

Combining these estimates with (4.19) and Proposition 4.5 completes the proof of Theorem 1.1. □

5. Hierarchical hybrid grids

In this section, we explain our strategy for the numerical approximation of solutions of elliptic equations. For definiteness, given a coefficient field $\mathbf{a}(x)$, and a domain $U \subseteq \mathbb{R}^d$, we consider the problem of computing an approximation of the solution $u \in H_0^1(U)$ of the equation

$$\begin{cases} -\nabla \cdot \mathbf{a}\left(\xi + \nabla u\right) = 0 & \text{in } U, \\ u = 0 & \text{on } \partial U. \end{cases} \tag{5.1}$$

Since generalizations such as the addition of lower-order terms or non-zero boundary conditions pose no particular additional difficulty, we will not discuss these further. In the first subsection, we observe the necessity to opt

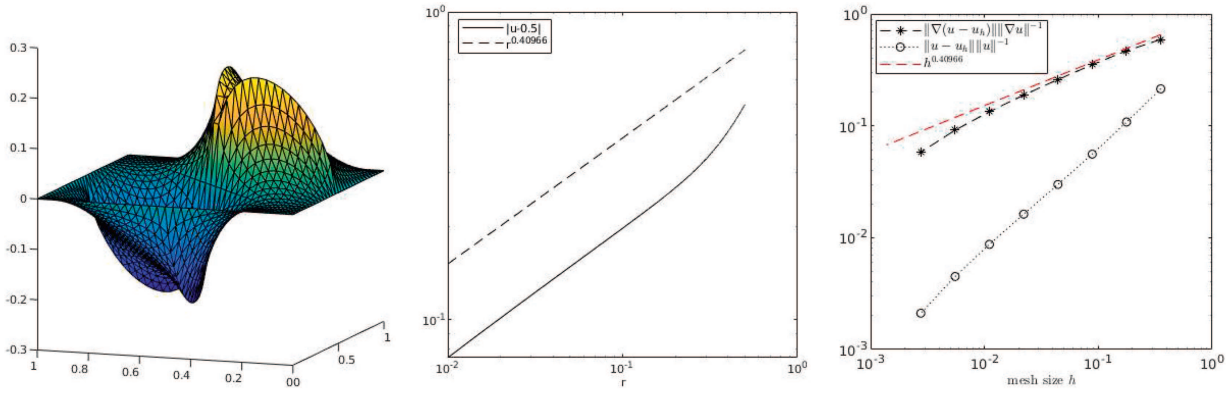

FIGURE 1. *Left*: solution u of (5.2) and (5.3) minus the affine function $x \mapsto x_1 + x_2$, with initial coarse mesh refined five times. *Middle*: solution along the line $x_1 = x_2$ for $x_1 > 0$, compared with the function $r^{0.40966}$, on a logarithmic scale. *Right*: (approximate) relative error, in H^1 and in L^2 respectively, for the problem in (5.2) and (5.3). Successive dots on a given line correspond to successive refinements of the triangular mesh, starting from a coarse mesh of 8 triangles.

for highly refined discretized approximations to the continuous equation. We then explain efficient ways to compute these highly refined approximations. Our approach is in line with the earlier work [14, 15] and based on hierarchical hybrid grids. That is, we start from an unstructured coarse mesh and refine it in a self-similar way a number of times; we then exploit this piecewise-structured hierarchical construction extensively at every step of the algorithm (assembly of the finite-element matrix, matrix-vector products, restriction and interpolation operators).

5.1. Roughness of solutions

In many practical instances, the heterogeneity of the coefficient field is due to the fact that the material of interest is a composite of several different types of substances: see for instance the library of images at [24]. In view of this, we focus on the case of piecewise-constant coefficient fields.

In this case, the discontinuities of the coefficient field compound the difficulties inherent to solving equations with rapidly oscillating coefficients. In order to measure the extent of these difficulties, consider the problem of approximating the solution $u \in H^1\left([-1,1]^2\right)$ to

$$\begin{cases} -\nabla \cdot \mathbf{a}(x) \nabla u = 0 & \text{in} \quad [-1,1]^2, \\ u(x) = x_1 + x_2 & \text{on} \quad \partial\left([-1,1]^2\right), \end{cases} \tag{5.2}$$

where the coefficient field $\mathbf{a}(x)$ is given by

$$\mathbf{a}(x) = \begin{vmatrix} 9\,\mathrm{Id} & \text{if} \quad x \in [-1,0]^2 \cup [0,1]^2, \\ \mathrm{Id} & \text{otherwise.} \end{vmatrix} \tag{5.3}$$

We start from a coarse mesh made of 8 triangles of equal sizes (two triangles in each of the translates of $[0,1]^2$). We then refine a given mesh by subdividing each triangle into 4 smaller triangles, adding a new vertex at the midpoint of each edge. We consider multiple iterations of this refinement procedure, and for each level of refinement, we compute the associated finite-element solution, using piecewise affine elements. The approximation of the solution minus the affine function $x = (x_1, x_2) \mapsto x_1 + x_2$ after five levels of refinement is represented on the left frame of Figure 1. The rough behavior of the solution near the origin is clearly visible. We also display the value of the solution along the line $x_1 = x_2$ on the middle frame of Figure 1 – see below around (5.5) for

the prediction of the exponent $0.4\ldots$ appearing there. The right frame of Figure 1 displays the relative error, measured in H^1 and in L^2 respectively, compared with the true solution. In order for the relative error to be below 10% in the H^1 norm, it is necessary to use at least six levels of refinement. At six levels of refinement, the linear system that needs to be solved already involves 2^{15} unknowns.

We can understand the roughness of the solution theoretically in a precise way. We consider more generally the situation when the coefficient field is given by

$$\mathbf{a}(x) = \begin{vmatrix} \Lambda\,\mathrm{Id} & \text{if} \quad x \in [-1,0]^2 \cup [0,1]^2, \\ \mathrm{Id} & \text{otherwise,} \end{vmatrix} \tag{5.4}$$

for some $\Lambda \in [1,\infty)$. A blow-up analysis near the origin suggests to look for solutions in the unit ball $B(0,1)$ of the form $r^\alpha f(\theta)$, where $r \geqslant 0$ and $\theta \in [0,2\pi)$ are the standard polar coordinates: $x_1 = r\cos\theta$ and $x_2 = r\sin\theta$. Denoting

$$a(\theta) := \begin{vmatrix} \Lambda & \text{if} \quad \theta \in \left[0, \frac{\pi}{2}\right] \cup \left[\pi, \frac{3\pi}{2}\right], \\ 1 & \text{otherwise,} \end{vmatrix}$$

we find that the smallest exponent $\alpha > 0$ such that $r^\alpha f(\theta)$ is a solution in $B(0,1)$ for some function f is given by

$$\alpha^2 = \inf\left\{ \frac{\int_0^{2\pi} (f')^2 a}{\int_0^{2\pi} f^2 a} \;:\; f \in H^1_{\mathrm{per}}([0,2\pi]) \;\; \text{s.t.} \;\; \int_0^{2\pi} fa = 0 \right\}.$$

Moreover, $r^\alpha f(\theta)$ is indeed a solution of the equation in $B(0,1)$ when f is the unique minimizer of the variational problem above. The value of this exponent was computed in [57]: it is

$$\alpha = \frac{4}{\pi} \arctan\left(\frac{1}{\sqrt{\Lambda}}\right). \tag{5.5}$$

Notice that the function $r^\alpha f(\theta)$ belongs to $H^{1+\alpha-\varepsilon}(B(0,1))$ for every $\varepsilon > 0$, but does not belong to $H^{1+\alpha}(B(0,1))$. We therefore expect a finite-element scheme with elements of size h to provide an approximation in H^1 at a precision of the order of h^α (and at precision of the order of $h^{1+\alpha}$ in L^2). The particular case we investigated numerically corresponds to $\Lambda = 9$, which gives

$$\alpha = \frac{4}{\pi} \arctan\left(\frac{1}{3}\right) \simeq 0.4096655294\ldots \tag{5.6}$$

In fact, it was shown in [57] that the exponent in (5.5) is the smallest possible exponent for Hölder regularity one can get if one allows for arbitrary coefficient fields which are everywhere a multiple of the identity and satisfy the ellipticity condition

$$\mathrm{Id} \leqslant \mathbf{a}(x) \leqslant \Lambda\,\mathrm{Id}. \tag{5.7}$$

In this sense, coefficient fields that are piecewise constant on a checkerboard structure are *worst possible* from the point of view of regularity (and therefore of difficulty of numerical approximation). For general coefficient fields satisfying (5.7) but not necessarily being a multiple of the identity matrix at each point, it was shown in [57] that the smallest possible exponent for Hölder regularity is $\alpha = \Lambda^{-\frac{1}{2}}$. An explicit coefficient field satisfying (5.7) and admitting a solution of the form $r^{\Lambda^{-\frac{1}{2}}} f(\theta)$ was first given in [49]. This exponent governs the rate of convergence of the finite-element approximation as the mesh is successively refined: for instance, for an ellipticity contrast of $\Lambda = 100$, we cannot hope for an asymptotic convergence rate better than $h^{0.1}$ in general, and no better than $h^{0.127\cdots}$ in the case of the coefficient field in (5.4). The situation is even worse in dimension $d = 3$, at least from a theoretical point of view. Indeed, to the best of our knowledge, it is an open question to show that when $d = 3$ (or for any $d \geqslant 3$), the regularity exponent can be bounded from below by a negative power of the ellipticity contrast.

5.2. Number of unknowns

We are ultimately interested in solving elliptic equations with *random* coefficients. In order to calculate the homogenized matrix, we will need to average over large domains, so as to tame the fluctuations of the coefficient field. As a toy example, consider the problem of calculating the standard average of the coefficient field, denoted by $\fint_{\mathbb{R}^d} \mathbf{a}$ above, see (2.8). By the scaling of the central limit theorem, in order to measure this quantity within a precision $\delta > 0$, we need to average over at least $C\delta^{-2}$ unit cells. Similarly, as was shown in Proposition 1.1 of [52], it is impossible to compute an approximation of the homogenized matrix at precision δ if one observes only $o(\delta^{-2})$ unit cells (the statement in [52] is written for finite-difference equations, but the proof applies essentially verbatim to the continuous setting).

Roughly speaking, if we want to compute $\bar{\mathbf{a}}$ within a precision of, say, 10%, we are bound to have to examine at least of the order of 10^2 unit cells. In two dimensions, if the mesh we use is refined six times as described in the previous subsection, this means that we must be facing problems involving of the order of $2^{15} \times 10^2 \simeq 3 \times 10^6$ unknowns. Notice that each further refinement of the mesh multiplies this number by 4, and that reducing the size of the fluctuations by a factor of 2 also multiplies this number by 4. Finally, this rough estimation hides multiplicative constants that may be large. (On the other hand, the random coefficient fields we investigate numerically in the next section are not made of a systematic periodic repetition of the worst-case coefficient field in (5.3), and this will mitigate the difficulty somewhat.)

5.3. Motivations for hybrid methods

The upshot of the previous subsections is that we ought to be able to solve for elliptic problems with many degrees of freedom. As is well-known, the numerical approximation of elliptic equations in domains with simple geometry and with constant coefficients can be performed very efficiently using a variety of techniques, including the geometric multigrid method (see [29] for several benchmarks). Indeed, for equations with constant coefficients, stencil-based operations can replace the need to assemble and store the finite-element matrix. Moreover, the data can be organized locally in agreement with the underlying geometry and accessed in a consistent way, resulting in few integer operations and highly optimized usage of the processor cache memory.

For more complex geometries or varying coefficients, completely unstructured approaches can be used instead. In this case, the problem of storing the finite-element matrix in memory becomes a major limitation. Moreover, data access becomes highly unpredictable and requires more integer operations, two factors that cause a dramatic drop in performance [14, 15].

Following [14, 15], we seek to remedy this problem by using a hybrid approach. The idea, called *Hierarchical hybrid grids* in [14, 15], is to proceed as in the completely unstructured case on the coarse mesh, but then rely on structured techniques within each constant-coefficient patch. This approach has multiple advantages. Firstly, we only need to assemble and store the finite-element matrix associated with the coarsest mesh. Similarly, we do not need to store the full computational grid in memory. This results in large gains in memory usage, which is otherwise the main limiting factor on the computing architectures we use. Moreover, we store a vector of the finite-element space in a bi-dimensional array indexed by the identity of the coarse element and then the position within it. This allows to obtain efficiency gains similar to those observed in the completely structured case, in particular regarding fast matrix-vector multiplications, and restriction and interpolation operators in the multigrid method.

5.4. Hierarchical hybrid grids

We now explain how to implement this approach more precisely. We also refer to [40] for a more thorough discussion, as well as [14, 15].

We start with some definitions. We say that $\mathcal{T} = \{K_1, \dots, K_n\}$ is a simplicial partition of the set $U \subseteq \mathbb{R}^d$ if the following three conditions hold: (1) for every $i \in \{1, \dots, n\}$, the set K_i is a simplex in $\mathbb{R}^d$ (*i.e.* the convex envelope of a set of $(d+1)$ points – a triangle in dimension $d = 2$ and a tetrahedron in dimension $d = 3$); (2) for every $i \neq j$, the interiors of K_i and K_j are disjoint; (3) the union $\bigcup_{i=1}^n K_i$ is the closure of the set U. For

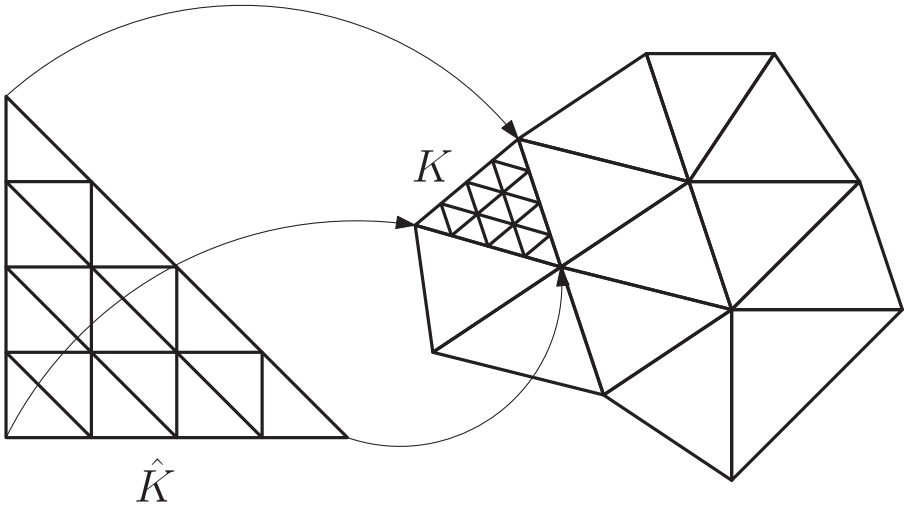

FIGURE 2. *Left*: standard simplex $\widehat{K}$ has been refined twice. *Right*: an unstructured coarse mesh, and the image of the twice-refined standard simplex through the affine mapping F_K for one particular coarse element K.

convenience, we often drop the word "simplicial" and simply say "partition" instead of "simplicial partition". A partition can be represented as a list of nodes $\{\mathbf{n}_i\}_{1\leqslant i\leqslant N} \subseteq \mathbb{R}^d$ and a list of $(d+1)$-tuples of indices that define the identity of the corner points of every simplex in the partition. We say that two partitions $\mathcal{T}_1$ and $\mathcal{T}_2$ are nested, and write $\mathcal{T}_1 \sqsubseteq \mathcal{T}_2$, if for every $K \in \mathcal{T}_1$, there exists $T \in \mathcal{T}_2$ such that $K \subseteq T$.

We denote by $\widehat{K}$ the standard simplex, that is, the convex envelope of the nodes $\mathbf{e}_0, \ldots, \mathbf{e}_d$, where $(\mathbf{e}_1, \ldots, \mathbf{e}_d)$ is the canonical basis of $\mathbb{R}^d$, and $\mathbf{e}_0$ is the null vector. Let $\widehat{\mathcal{T}}$ be a partition of $\widehat{K}$, and $\mathcal{T}_H$ be a partition of an arbitrary domain. We say that a partition $\mathcal{T}_h$ is the *locally uniform partition* associated with $\left(\mathcal{T}_H, \widehat{\mathcal{T}}\right)$, and write $\mathcal{T}_h = \mathsf{lup}\left(\mathcal{T}_H, \widehat{\mathcal{T}}\right)$, if $\mathcal{T}_h \sqsubseteq \mathcal{T}_H$ and, for every $K \in \mathcal{T}_H$, there exists an affine mapping $F_K : \mathbb{R}^d \to \mathbb{R}^d$ such that the image of $\widehat{\mathcal{T}}$ under F_K is $\{T \in \mathcal{T}_h \ : \ T \subseteq K\}$. See Figure 2 for an illustration. Notice that the mapping F_K appearing above must be such that $F_K\left(\widehat{K}\right) = K$. Such an affine mapping is entirely specified by prescribing which nodes of $\widehat{K}$ are sent to which nodes of K.

Note that the locally uniform partition $\mathcal{T}_h$ is completely specified by the knowledge of $\left(\mathcal{T}_H, \widehat{\mathcal{T}}\right)$. This allows for vast memory gains for storing the partition, since only the reference simplex $\widehat{K}$ is meshed finely, while the global partition $\mathcal{T}_H$ remains coarse. In addition, as discussed below, this format will be very convenient for a variety of operations, including for implementing the restriction and interpolation operators in the multigrid method.

5.5. Assembly of the finite-element matrix

We proceed to define the finite-element matrix, and then describe how to store it efficiently using the structure of locally uniform partitions, under the assumption that the coefficient field is constant on each coarse element.

Let $\mathcal{T}$ be a partition of the domain $U \subseteq \mathbb{R}^d$. We think of this partition as being relatively coarse, having a level of detail just sufficient to resolve the variations of the coefficient field. For clarity of exposition, we start by considering the case in which this coarse partition is not refined further. Denote by $\{\mathbf{n}_i\}_{1\leqslant i\leqslant N} \subseteq \mathbb{R}^d$ the nodes of the partition $\mathcal{T}$. We look for an approximation of the solution of (5.1) in the finite-dimensional space

$V(\mathcal{T}) \cap H^1_0(U)$, where

$$V(\mathcal{T}) := \left\{ u \in H^1(U) \ : \ u_{|K} \text{ is affine for every } K \in \mathcal{T} \right\}. \tag{5.8}$$

A standard basis for $V(\mathcal{T})$ is formed by the nodal functions $\{\varphi_i\}_{1 \leqslant i \leqslant N} \subseteq V(\mathcal{T})$, which are specified by the condition

$$\varphi_i(\mathbf{n}_j) = \mathbf{1}_{i=j}, \quad \text{for every } i, j \in \{1, \dots, N\}.$$

Denoting by $\mathbf{x}$ the vector encoding the finite-element approximation of (5.1) in the basis formed by

$$\{\varphi_i \ : \ \mathbf{n}_i \text{ is an interior point of } U\},$$

we identify $\mathbf{x}$ as the solution of the problem

$$A\mathbf{x} = \mathbf{b}, \quad \text{where} \quad A_{ij} = \int_U \nabla\varphi_i \cdot \mathbf{a} \nabla\varphi_j \quad \text{and} \quad \mathbf{b}_i = \int_U \nabla\varphi_i \cdot \mathbf{a}\xi.$$

Notice that the size of the vectors and of the symmetric matrix appearing above is the number of nodes in the interior of U; this is how the null Dirichlet boundary condition is enforced.

For each $K \in \mathcal{T}$, denote by $\left\{\mathbf{n}_i^K\right\}_{0 \leqslant i \leqslant d} \subseteq \mathbb{R}^d$ the extremal points of the simplex K. This defines a mapping $\sigma : \mathcal{T} \times \{0, \dots, d\} \to \{1, \dots N\}$, which to each (K, i) associates the node number of the node $\mathbf{n}_i^K$ in the global ordering $\{\mathbf{n}_j\}_{1 \leqslant j \leqslant N}$. Denote by φ_i^K the restriction to K of the basis function $\varphi_{\sigma(K,i)}$. The functions $\left(\varphi_i^K\right)$ are called *local shape functions*. The contribution of the element $K \in \mathcal{T}$ to the entries of the matrix A can be represented by the matrix $A^{(K)} \in \mathbb{R}^{(d+1)\times(d+1)}$ such that, for every $i, j \in \{0, \dots d\}$,

$$A_{ij}^{(K)} = \int_K \nabla\varphi_i^K \cdot \mathbf{a} \nabla\varphi_j^K. \tag{5.9}$$

The global matrix A can then be reconstructed by the identity

$$A = \sum_{K \in \mathcal{T}} R_K^T A^{(K)} R_K, \tag{5.10}$$

where $R_K \in \mathbb{R}^{N \times (d+1)}$ is the canonical matrix representing the linear mapping

$$\begin{cases} \mathbb{R}^N & \to \mathbb{R}^{d+1} \\ \mathbf{x} & \mapsto \sum_{i=0}^d \mathbf{x}_{\sigma(K,i)} \mathbf{e}_{i+1}. \end{cases}$$

We denote the local shape functions associated with the standard simplex by $\widehat{\varphi}_i := \varphi_i^{\widehat{K}}$, for every $i \in \{0, \dots d\}$, and call them *reference shape functions*. In dimension $d = 2$, these reference shape functions are $x \mapsto 1 - x_1 - x_2$, $x \mapsto x_1$, and $x \mapsto x_2$. We also denote by $F_K : x \mapsto B_K x + v_K$ the unique affine mapping that sends the nodes $\mathbf{e}_0, \dots, \mathbf{e}_d$ of the standard simplex to the nodes $\mathbf{n}_0^K, \dots, \mathbf{n}_d^K$, so that in particular, $F_K\left(\widehat{K}\right) = K$ (see Fig. 2). Notice that, for every $x \in K$,

$$\varphi_i^K\left(F_K(x)\right) = \widehat{\varphi}_i(x), \quad \text{so that} \quad B_K^T\left(\nabla\varphi_i^K\right)\left(F_K(x)\right) = \nabla\widehat{\varphi}_i(x). \tag{5.11}$$

By this change of variables, for every $i, j \in \{0, \dots d\}$, we can rewrite the integral on the right side of (5.9) as

$$A_{ij}^{(K)} = |\det B_K| \int_{\widehat{K}} B_K^{-T} \nabla\widehat{\varphi}_i \cdot \mathbf{a} B_K^{-T} \nabla\widehat{\varphi}_j. \tag{5.12}$$

In the case when the partition $\mathcal{T}$ is sufficiently fine that $\mathbf{a}(x)$ is constant equal to $\mathbf{a}^{(K)}$ when x varies in K, we set

$$\mathbf{c}^{(K)} := |\det B_K| \, B_K^{-1} \mathbf{a}^{(K)} B_K^{-T} \in \mathbb{R}^{d \times d}, \tag{5.13}$$

and the previous display becomes

$$A_{ij}^{(K)} = \int_{\widehat{K}} \nabla \widehat{\varphi}_i \cdot \mathbf{c}^{(K)} \nabla \widehat{\varphi}_j. \tag{5.14}$$

We can expand this expression into

$$A^{(K)} = \sum_{p,q=1}^{d} \mathbf{c}_{p,q}^{(K)} \widehat{A}^{pq}, \tag{5.15}$$

where, for each $p, q \in \{1, \dots d\}$, the matrix $\widehat{A}^{pq} \in \mathbb{R}^{(d+1)\times(d+1)}$ is such that, for every $i, j \in \{0, \dots, d\}$,

$$\widehat{A}_{ij}^{pq} := \int_{\widehat{K}} \partial_{x_p} \widehat{\varphi}_i \, \partial_{x_q} \widehat{\varphi}_j. \tag{5.16}$$

Notice that, using (5.9) and (5.15), we can compute the finite-element matrix A from the knowledge of $\left\{\mathbf{c}^{(K)}\right\}_{K\in\mathcal{T}}$ and $\left\{\widehat{A}^{pq}\right\}_{1\leqslant p,q\leqslant d}$.

We now generalize these observations to the case when the partition $\mathcal{T}$ is locally uniform, say $\mathcal{T} = \mathsf{lup}\left(\mathcal{T}_H, \widehat{\mathcal{T}}\right)$. We keep writing $\{\mathbf{n}_i\}_{1\leqslant i\leqslant N}$ for the nodes of the fine partition $\mathcal{T}$, and we denote by $\{\widehat{\mathbf{n}}_i\}_{1\leqslant i\leqslant \widehat{N}} \subseteq \widehat{K}$ the nodes of the partition $\widehat{\mathcal{T}}$ of the standard simplex. For each $K \in \mathcal{T}_H$, the fine partition $\mathcal{T}$ induces a fine partition of K by restriction; this partition is in fact the image of $\widehat{\mathcal{T}}$ under the mapping F_K appearing in the definition of local uniform partition. Hence, the nodes of this partition are $\mathbf{n}_i^K := F_K\left(\widehat{\mathbf{n}}_i\right)$, where i ranges in $\left\{1, \dots, \widehat{N}\right\}$. This naturally induces a mapping $\sigma : \mathcal{T}_H \times \left\{1, \dots \widehat{N}\right\} \to \{1, \dots N\}$ which, to each (K, i), associates the index of the node $\mathbf{n}_i^K$ in the numbering provided by $\{\mathbf{n}_i\}_{1\leqslant i\leqslant N}$. The mapping σ is clearly surjective, but it is not a bijection: indeed, the nodes that belong to the boundary of multiple coarse elements are represented multiple times. On the other hand, every node that belongs to the interior of a simplex of the coarse partition has a unique representation in the form $\mathbf{n}_i^K$ for some $(K, i) \in \mathcal{T}_H \times \left\{1, \dots, \widehat{N}\right\}$. As the partition $\widehat{\mathcal{T}}$ becomes finer and finer, the approximation $N \simeq |\mathcal{T}_H|\, \widehat{N}$ therefore becomes more and more accurate. (The notation $|\mathcal{T}_H|$ stands for the number of elements in $\mathcal{T}_H$.)

For each $K \in \mathcal{T}_H$ and $i \in \left\{1, \dots, \widehat{N}\right\}$, we denote by φ_i^K the restriction to K of the basis function $\varphi_{\sigma(K,i)}$. The contribution of the coarse element $K \in \mathcal{T}_H$ to the finite-element matrix can represented by the $\widehat{N}$-by-$\widehat{N}$ matrix $A^{(K)}$ such that (5.9) holds for every $i, j \in \left\{0, \dots, \widehat{N}\right\}$. The relation (5.10) still holds, where now $R_K \in \mathbb{R}^{N\times\widehat{N}}$ is the canonical matrix representing the linear mapping

$$\begin{cases} \mathbb{R}^N \to \mathbb{R}^{\widehat{N}} \\ \mathbf{x} \quad \mapsto \sum_{i=1}^{\widehat{N}} \mathbf{x}_{\sigma(K,i)} \mathbf{e}_i. \end{cases} \tag{5.17}$$

For every $i \in \left\{1, \dots, \widehat{N}\right\}$, the reference shape function is defined by setting $\widehat{\varphi}_i := \varphi_i^{\widehat{K}}$. The identities (5.11) to (5.16) still hold, the only difference being that K now ranges in $\mathcal{T}_H$ and the indices i and j now range in $\left\{1, \dots, \widehat{N}\right\}$.

It thus follows that the finite-element matrix associated with the locally uniform partition $\mathcal{T}$ can be represented by storing only the set of d-by-d matrices $\left\{\mathbf{c}^{(K)}\right\}_{K\in\mathcal{T}_H}$ and the set of $\widehat{N}$-by-$\widehat{N}$ matrices $\left\{\widehat{A}^{pq}\right\}_{1\leqslant p,q\leqslant d}$. Moreover, these matrices can be constructed directly in a straightforward manner, without having to construct the fine partition $\mathcal{T}$. Finally, in the practical cases we have in mind, the matrices $\widehat{A}^{pq}$ are highly regular and have only of the order of $C\widehat{N}$ non-zero entries. The amount of memory required to store this data is proportional to

$$d^2\left(|\mathcal{T}_H| + C\widehat{N}\right).$$

If we were to ignore the locally uniform structure of the fine partition $\mathcal{T}$, the cost of storing its finite-element matrix would be proportional to N instead. Recalling that $N \simeq |\mathcal{T}_H|\widehat{N}$, we see that the semi-structured approach results indeed in a significant gain in memory usage.

5.6. Matrix-vector product

Pursuing with the setting of the previous section, we now discuss how to store vectors and perform matrix-vector operations with the finite-element matrix, which we recall is represented in memory by the matrices $\left\{\mathbf{c}^{(K)}\right\}_{K\in\mathcal{T}_H}$ and $\left\{\widehat{A}^{pq}\right\}_{1\leqslant p,q\leqslant d}$. As discussed above,

$$A = \sum_{K\in\mathcal{T}_H}\sum_{p,q=1}^{d} \mathbf{c}_{pq}^{(K)} R_K^T \widehat{A}^{pq} R_K, \tag{5.18}$$

where R_K is the matrix representing the linear mapping in (5.17). In view of (5.18), instead of representing finite-element vectors as N-dimensional vectors, we encode them in an $\widehat{N}$-by-$|\mathcal{T}_H|$ array. That is, we represent each $\mathbf{x}\in\mathbb{R}^N$ by an array X such that the columns of X, denoted by $\{X(:,j)\}_{1\leqslant j\leqslant|\mathcal{T}_H|}\subseteq\mathbb{R}^{\widehat{N}}$, are equal to the vectors $\left\{R_{K_j}\mathbf{x}\right\}_{1\leqslant j\leqslant|\mathcal{T}_H|}$. Here we used the notation $\{K_j\}_{1\leqslant j\leqslant|\mathcal{T}_H|}$ to denote an enumeration of the (unstructured) set $\mathcal{T}_H$. Naturally, the entries that are associated with nodes that belong to multiple coarse elements are repeated in this representation; this parallels the observation that the mapping σ defined in the previous subsection is surjective but not bijective.

The operation of A onto a vector can then be evaluated in two steps: first, we compute the $\widehat{N}$-by-$\widehat{N}$ matrix Y defined, for every $j\in\{1,\dots,|\mathcal{T}_H|\}$, by

$$Y(:,j) = \sum_{p,q=1}^{d} \mathbf{c}_{pq}^{(K_j)} \widehat{A}^{pq} X(:,j).$$

The column $Y(:,j)$ is however not equal to the desired outcome of $R_{K_j}A\mathbf{x}$, due to the presence of nodes that belong to multiple coarse elements. In the second step, we compute

$$(AX)(:,j) = \sum_{K_\ell\in\mathcal{T}} R_{K_j} R_{K_\ell}^T Y(:,\ell). \tag{5.19}$$

In the actual implementation of this second step, we *do not* need to construct the matrices R_{K_j} explicitly. Instead, we implement this formula by identifying the nodes that are found at the interface between two or more elements of the coarse partition. In order to do so, we distinguish between different types of interfaces, according to whether they are to be found on faces, edges, or point vertices. (Naturally, face-type interfaces are only relevant in dimension $d=3$.) For a more precise description of this aspect, we refer to [14, 15, 40].

5.7. Multigrid method

The geometric multigrid method is a technique for the numerical approximation of elliptic problems [20]. It uses a sequence of nested partitions $\mathcal{T}_n \sqsubseteq \cdots \sqsubseteq \mathcal{T}_0$, as well as restriction and interpolation operators which allow to transfer a function defined on a given grid to a function defined on a coarser and finer grids respectively.

The setting of locally uniform partitions is particularly conducive to efficient implementations of the geometric multigrid method. Indeed, we first give ourselves a sequence of nested partitions of the reference element $\widehat{\mathcal{T}}_n \sqsubseteq \cdots \sqsubseteq \widehat{\mathcal{T}}_0$. These nested partitions are constructed as follows: we fix $\widehat{\mathcal{T}}_0 := \left\{\widehat{K}\right\}$ to be the trivial partition, and then inductively construct $\widehat{\mathcal{T}}_{k+1}$ from $\widehat{\mathcal{T}}_k$ by adding new nodes at the middle of the edge of each element of $\widehat{\mathcal{T}}_k$, and, in dimension $d=2$, by replacing each triangle with a partition of this triangle made of 4 triangles, or in

dimension $d = 3$, by replacing each tetrahedron with a partition of this tetrahedron made of 8 tetrahedra [17]. The nested partitions we use to implement the geometric multigrid method are then

$$\mathcal{T}_n = \mathsf{lup}\left(\mathcal{T}_H, \widehat{\mathcal{T}}_n\right) \sqsubseteq \cdots \sqsubseteq \mathcal{T}_1 = \mathsf{lup}\left(\mathcal{T}_H, \widehat{\mathcal{T}}_1\right) \sqsubseteq \mathcal{T}_0 = \mathsf{lup}\left(\mathcal{T}_H, \widehat{\mathcal{T}}_0\right) = \mathcal{T}_H.$$

Recall that we denote by $V(\mathcal{T})$ the finite-element space associated with the partition $\mathcal{T}$, see (5.8). We start by defining interpolation and restriction operators associated with the nested partitions of the standard simplex. For each $k < n$, we define the interpolation operator $\widehat{\mathcal{I}}_k : V\left(\widehat{\mathcal{T}}_k\right) \to V\left(\widehat{\mathcal{T}}_{k+1}\right)$ to be the canonical injection. The restriction operator can then be taken as the transpose of the interpolation operator, up to a normalization constant (see [20], Def. 6.3.1 for more precision). Similarly, we define the interpolation operator $\mathcal{I}_k : V\left(\mathcal{T}_k\right) \to V\left(\mathcal{T}_{k+1}\right)$ to be the canonical injection. Recall that we represent a given vector $\mathbf{x}_k \in V\left(\mathcal{T}_k\right)$ as an $\widehat{N}_k$-by-$|\mathcal{T}_H|$ matrix X_k such that $X_k(:, j) = R_{K_j}^{(k)} \mathbf{x}_k \in \mathbb{R}^{\widehat{N}_k}$ where $\widehat{N}_k$ is the number of vertices of the partition $\widehat{\mathcal{T}}_k$ of the standard simplex, and we wrote $R_{K_j}^{(k)}$ instead of R_{K_j} to emphasize the dependency on k of this operator. In this representation, we can evaluate the interpolation operator very simply by setting, for every $j \in \{1, \ldots, |\mathcal{T}_H|\}$,

$$\left(\mathcal{I}_k X_k\right)(:, j) = \widehat{\mathcal{I}}_k\left(X_k(:, j)\right).$$

Up to a normalization constant, we wish to use the transpose of $\mathcal{I}_k$ as our restriction operator. In view of the format in which we store elements of $V\left(\mathcal{T}_{k+1}\right)$, this is not absolutely straightforward to compute, since it involves some amount of communication between vertices belonging to different elements of the coarse partition. We now explain how to perform this computation efficiently by reducing it to the same calculation as that arising in matrix-vector multiplication, see (5.19). Given a load vector $\mathbf{b}$ and an element $\mathbf{x}_{k+1} \in V\left(\mathcal{T}_{k+1}\right)$, we aim to compute the residual

$$\mathbf{r}_k := \mathcal{I}_k^T\left(\mathbf{b} - A\mathbf{x}_{k+1}\right). \tag{5.20}$$

We use the same data format to store the load vector, that is, we represent it by a family $\left(\mathbf{b}^{(k+1,K)}\right)_{K \in \mathcal{T}_H}$ of vectors of size $\widehat{N}_{k+1}$ such that

$$\mathbf{b} = \sum_{K \in \mathcal{T}_H} \left(R_K^{(k+1)}\right)^T \mathbf{b}^{(k+1,K)}.$$

For the model problem (5.1), this means that we set, for every $i \in \left\{1, \ldots, \widehat{N}_{k+1}\right\}$,

$$\mathbf{b}_i^{(k+1,K)} := \int_K \nabla \varphi_i^{(k+1,K)} \cdot \mathbf{a}\xi,$$

where again we wrote $\varphi_i^{(k+1,K)}$ instead of φ_i^K to make the dependency on k more explicit. Recall that the vector $\mathbf{x}_{k+1}$ in (5.20) is stored in memory as an array whose columns are given by $R_K^{(k+1)}\mathbf{x}_{k+1}$. Using also (5.10), we obtain that

$$\mathbf{r}_k = \sum_{K \in \mathcal{T}_H} \mathcal{I}_k^T \left(R_K^{(k+1)}\right)^T \mathbf{b}^{(k+1,K)} - \sum_{K \in \mathcal{T}_H} \mathcal{I}_k^T \left(R_K^{(k+1)}\right)^T A^{(k+1,K)} R_K^{(k+1)} \mathbf{x}_{k+1}.$$

Moreover, one can verify that

$$R_K^{(k+1)} \mathcal{I}_k = \widehat{\mathcal{I}}_{k+1} R_K^{(k)}.$$

We thus conclude that

$$\mathbf{r}_k = \sum_{K \in \mathcal{T}_H} \left(R_K^{(k)}\right)^T \mathbf{r}_k^{(K)}, \quad \text{with} \quad \mathbf{r}_k^{(K)} := \left(\widehat{\mathcal{I}}_{k+1}\right)^T \left(\mathbf{b}^{(k+1,K)} - A^{(k+1,K)} R_K^{(k+1)} \mathbf{x}_{k+1}\right).$$

Each term $\mathbf{r}_k^{(K)}$ is relatively easy to compute, since $\widehat{\mathcal{I}}_{k+1}$ is an operator of moderate dimension. We have now reached a situation analogous to that in the previous section: the remaining problem is that it is not true in general that $\mathbf{r}_k^{(K)} = R_K^{(k)} \mathbf{r}_k$. This can be arranged by proceeding as in (5.19).

For the smoothing steps in the multigrid method, we use a few steps of conjugate gradient descent. Finally, we use a direct solver for the coarse-grid problem. In our numerical experiments, the above-described implementation of the geometric multigrid method showed robust convergence behavior.

6. Numerical tests

In this section, we report on numerical results for the method presented in this paper. The code was written in the Julia language, and is available at this address:

https://github.com/haampie/Homogenization.jl. (6.1)

In all the examples we consider, the coefficient field is $\mathbb{Z}^d$-stationary, where $d \in \{2, 3\}$ is the dimension. For convenience, we replace averages against the mask χ_{r_k} in (1.6) and (1.7) by averages over the cube $(-r_k, r_k)^d$. While strictly speaking, this situation is not covered by Theorem 1.1, it is not difficult to show that the statement is still correct in this case (in fact, the argument is then somewhat simpler). For simplicity, we also fix $\varepsilon = 0$ in (1.5). As discussed below (1.8), it is not difficult to modify the proof and cover this case as well, at the cost of an arbitrarily small loss of exponent in (1.7). We also slightly modify the definition of the approximations $\widetilde{v}_k$ in (1.8), by using a square or a cube instead of a ball for the domain: that is, for every $k \in \{0, \dots, n\}$, we set

$$L(k, n) := 2^{n - \frac{k}{2}} + C_{\mathrm{bl}}(1 + n) 2^{\frac{k}{2}},$$

and solve for $\widetilde{v}_k \in H_0^1\left((-L(k,n), L(k,n))^d\right)$ solution to

$$(2^{-k} - \nabla \cdot \mathbf{a} \nabla) \widetilde{v}_k = 2^{-k}\, \widetilde{v}_{k-1} \quad \text{in} \quad (-L(k,n), L(k,n))^d, \tag{6.2}$$

with null Dirichlet boundary condition on $\partial\left((-L(k,n), L(k,n))^d\right)$. The estimator we wish to calculate, slightly modified from (1.6), is then defined by

$$\widehat{\sigma}_n^2 := \fint_{(-2^n, 2^n)^d} \left(-\mathbf{a}\xi \cdot \nabla \widetilde{v}_0 + \widetilde{v}_0^2\right) + \sum_{k=1}^{n} 2^k \fint_{\left(-2^{n-\frac{k}{2}}, 2^{n-\frac{k}{2}}\right)^d} \left(\widetilde{v}_{k-1} \widetilde{v}_k + \widetilde{v}_k^2\right). \tag{6.3}$$

In order to obtain numerical approximations of the functions $\widetilde{v}_k$, we use the finite-element method with hierarchical hybrid grids presented in Section 5. In all the examples we consider, the coefficient field is piecewise constant on $z + [0, 1)^d$, for every $z \in \mathbb{Z}^d$. We thus start from a coarse partition of the domain which consists, in dimension $d = 2$, in splitting each unit square into two triangles, or in dimension $d = 3$, in splitting each unit cube into six tetrahedra. This provides us with a coarse partition of the domain, which was denoted by $\mathcal{T}_H$ in Section 5. We then proceed to refine this partition iteratively by decomposing, in dimension $d = 2$, each triangle into four smaller triangles, or in dimension $d = 3$, each tetrahedron into eight smaller tetrahedra (and we do so in practice by constructing a refined partition $\mathcal{T}_h$ of the standard simplex $\widehat{K}$ iteratively, which provides us with an implicit fine partition of the whole domain using the notion of locally uniform partition, see Sect. 5.4). We denote the number of iterative levels of refinement performed in this way by N_{ref}. This defines an approximation of the quantity $\widehat{\sigma}_n^2$ defined in (6.3), which we denote by $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$. Strictly speaking, this quantity also depends on the choice of the boundary layer size C_{bl}, but we keep this implicit in the notation.

Theorem 1.1 bundles together an estimate for the mean error and an estimate for the standard deviation or our approximation $\widehat{\sigma}_n^2$. The approximation has been set up so that both quantities are of the same order, that is, $2^{-\frac{nd}{2}}$. Additionally to this error comes the error due to the finite-element discretization: for each fixed N_{ref}, the quantity $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ computes an approximation of the homogenized matrix of the *discretized system* with N_{ref}

levels of refinement. While we did not prove this, it is clear that all the arguments use to prove Theorem 1.1 would remain valid for the discretized system, and thus $\sigma^2(n, N_{\mathrm{ref}})$ allows to approximate the homogenized matrix $\bar{\mathbf{a}}(N_{\mathrm{ref}})$ of the discretized system with a mean error and a standard deviation that both scale like $2^{-\frac{nd}{2}}$ as n tends to infinity. However, there is also a discrepancy between $\bar{\mathbf{a}}(N_{\mathrm{ref}})$ and the homogenized matrix $\bar{\mathbf{a}}$ of the continuous equation, which is manifested in our algorithm in the fact that we do not have perfect access to the solutions $\widetilde{v}_k$ of (6.2). Moreover, as explained in Section 5.1, the rate of convergence of approximate solutions in terms of N_{ref} can become arbitrarily slow as the ellipticity contrast gets large.

As said above, we consider coefficient fields that are piecewise constant on unit cubes; more precisely, we assume that for every $z \in \mathbb{Z}^d$, we have

$$\forall x \in z + [0,1)^d, \quad \mathbf{a}(x) = \mathbf{b}_z,$$

for some family $(\mathbf{b}_z)_{z\in\mathbb{Z}^d}$. This family is random and constructed in the following way, given two parameters $\alpha \leqslant \beta \in (0,\infty)$: the random variables $(\mathbf{b}_z)_{z\in\mathbb{Z}^d}$ are independent; the matrix $\mathbf{b}_z$ is diagonal; the diagonal entries of $\mathbf{b}_z$, which we denote by $(\mathbf{b}_{z,ii})_{1\leqslant i\leqslant d}$, are independent; and finally, for every $i \in \{1,\dots,d\}$,

$$\mathbb{P}\left[\mathbf{b}_{z,ii} = \alpha\right] = \mathbb{P}\left[\mathbf{b}_{z,ii} = \beta\right] = \frac{1}{2}.$$

As discussed in Section 5.1, this example is particularly interesting since it is in some sense the coefficient field which allows for the most pathological singularities in the solutions for a given ellipticity ratio $\Lambda = \beta/\alpha$. Notice that, in order to demonstrate that our numerical code is not restricted to the case when $\mathbf{a}(x)$ is a multiple of the identity, we have dropped this restriction here (and it would not be difficult to accommodate for matrices that are not diagonal). An additional very interesting feature of this class of examples is that it is one of the very rare cases where the homogenized matrix is known exactly: in dimension $d = 2$, it is given by $\bar{\mathbf{a}} = \sqrt{\alpha\beta}\,\mathrm{Id}$ Exercise 2.10 of [13]. (No such simple formula is expected to exist in dimension $d = 3$, and in fact, we are not aware of any genuinely three-dimensional coefficient field where the homogenized matrix is known exactly.)

6.1. Two-dimensional case, moderate contrast

We fix $d = 2$, $\alpha = 1$, and $\beta = 9$. We thus have in this case that $\bar{\mathbf{a}} = \sqrt{\alpha\beta}\,\mathrm{Id} = 3\,\mathrm{Id}$. Using the notation in (2.8), we also observe that $⨍\,\mathbf{a} = 5\,\mathrm{Id}$, and therefore we expect that $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ converges to 2 as n and N_{ref} tend to infinity. We fix the boundary layer constant $C_{\mathrm{bl}} := 4$, and plot a histogram of $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ for different values of n and N_{ref}, see Figure 3. Each histogram is obtained by sampling 200 realizations of the estimator. For each value of n and N_{ref}, we also report the empirical mean and variance of $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$. Notice that the estimator has a bias to overestimate the value of $\bar{\mathbf{a}}$, which is consistent with the fact that the remainder term D_n in the series expansion (3.2) is nonnegative, see Proposition 4.5 (the sign of the discretization error was not predicted theoretically).

From the results displayed on Figure 3, one can check that the quantity $\widehat{\sigma}^2(n = 4, N_{\mathrm{ref}} = 3)$ falls within the interval $[1.84, 2.02]$ with 95% probability. Taking for granted that we can estimate $⨍_{\mathbb{R}^d}\mathbf{a} = 5\,\mathrm{Id}$ more easily, we obtain an estimation for $\xi\cdot\mathbf{a}\xi$ which falls within the interval $[2.98, 3.16]$ with 95% probability, the true value being 3. This estimator thus produces a result with a relative error of 5% from the true value with 95% probability. It takes about 2 s to compute this quantity on a laptop computer with 16 Go of memory and using a single processor clocking at 2.40 GHz.

We next investigate more precisely the scalings of the standard deviation and mean error of $\sigma^2(n, N_{\mathrm{ref}})$. (By definition, the mean error is $\left|\mathbb{E}\left[\sigma^2(n, N_{\mathrm{ref}})\right] - 2\right|$ for this example). On the left frame of Figure 4, we see that the variance decays like $2^{-dn} = 2^{-2n}$, as predicted by the theoretical results. On the left frame of Figure 4, we display the mean error as a function of n and of the number of refinements. Our theoretical arguments predict that the mean error is the sum of a term of the order of $2^{-\frac{nd}{2}} = 2^{-n}$, of the discretization error which depends on N_{ref}, and of the boundary layer error related to the choice of C_{bl}. We display the dependency of the mean error in these parameters more precisely on Figure 5, for the value $n = 4$. We see on the left frame of Figure 5

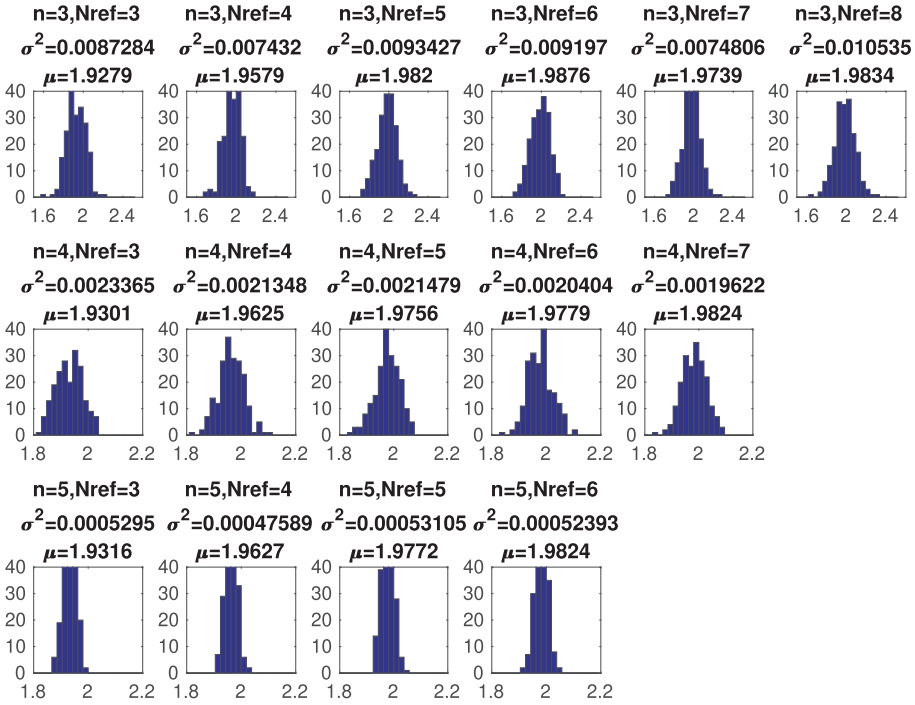

FIGURE 3. Empirical distribution of $\widehat{\sigma}^2\,(n, N_{\mathrm{ref}})$ when $d = 2$, $\alpha = 1$ and $\beta = 9$, for different values of n and N_{ref}. We recall that N_{ref} is the number of times the finite-element mesh has been refined.

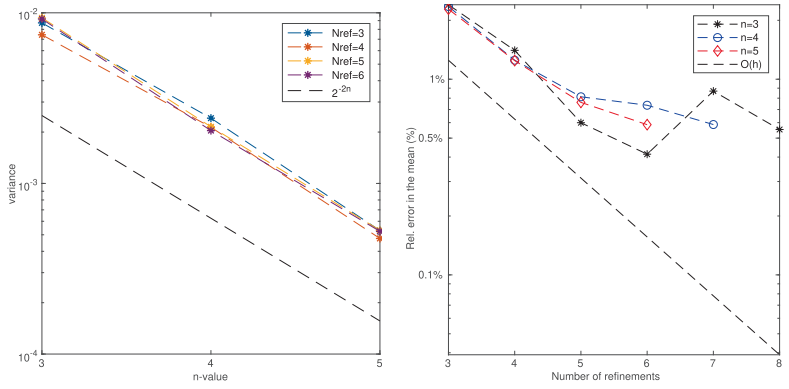

FIGURE 4. Variance (*left*) and mean error (*right*) of $\widehat{\sigma}^2\,(n, N_{\mathrm{ref}})$ for different values of n and number of mesh refinements N_{ref}. The variance decays approximately with the rate 2^{-dn} predicted by Theorem 1.1.

that the choice of $C_{\mathrm{bl}} = 4$ is already sufficient to ensure that the boundary layer error is negligible compared with the discretization error. On the right frame of Figure 5, we observe that the discretization error decays approximately like $h^{0.409}$, where h is the element size, as predicted in the discussion around (5.6).

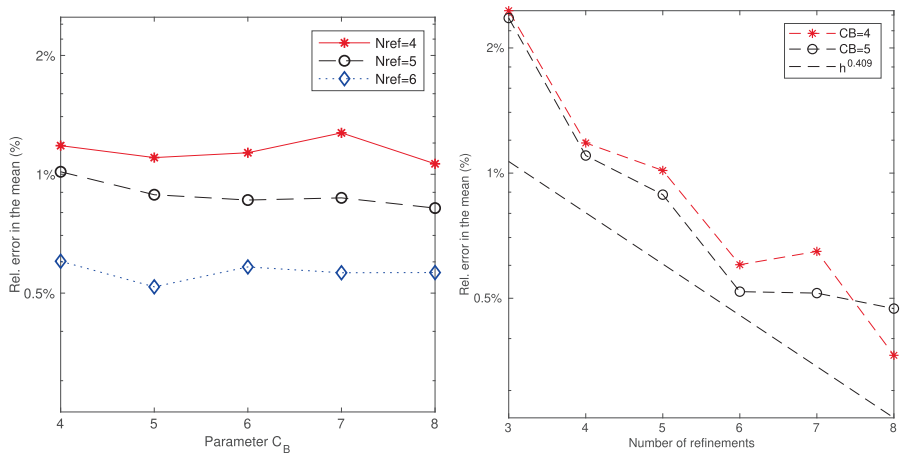

FIGURE 5. Error in the mean, for $n = 4$, as a function of the boundary layer constant C_{bl} (*left*), and as a function of the number of mesh refinements (*right*). The error is essentially independent of the value of $C_{\mathrm{bl}} \geqslant 4$. The dependency in N_{ref} is in good agreement with the predicted convergence rate in h^{α}, for $\alpha \simeq 0.409$.

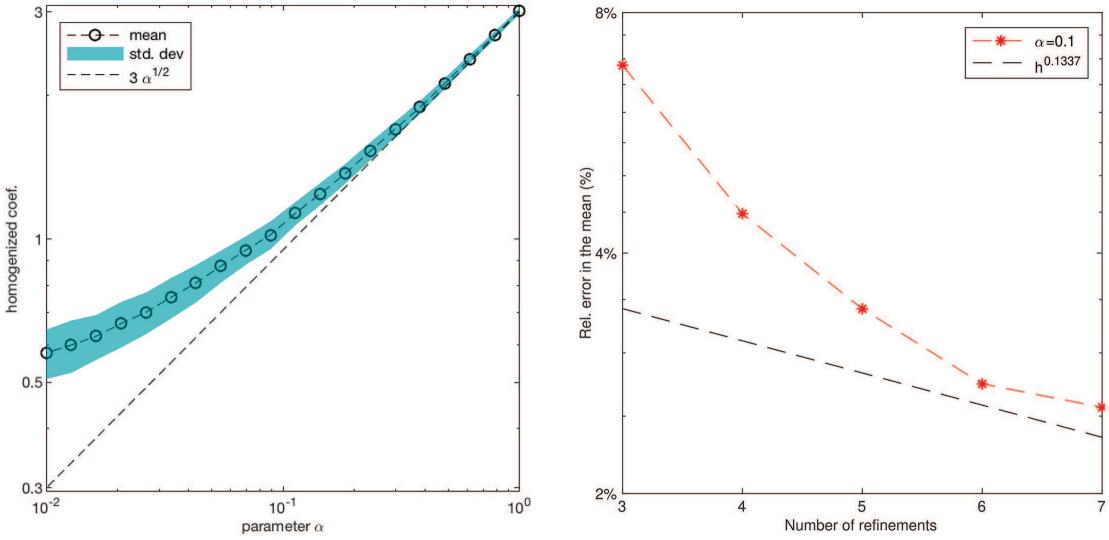

FIGURE 6. *Left*: mean and standard deviation of $\widehat{\sigma}^2$ ($n = 4, N_{\mathrm{ref}} = 5$), for different values of α. Notice that the logarithmic scale inflates the absolute value of the error on the left of the graph. *Right*: the mean error for $\alpha = 0.1$, as a function of N_{ref}. For small values of α, the finite element approximation converges very slowly, due to the singularities at the corners of the cherckerboard tiling. For $\alpha = 0.1$, we expect the asymptotic error rate to scale like $h^{0.1337}$.

6.2. Two-dimensional case, high contrast

We continue with the two-dimensional setting, we also keep $\beta = 9$, but we now progressively decrease α in the interval $[10^{-2}, 1]$. More precisely, we vary α in twenty logarithmically equally spaced steps between the values 1

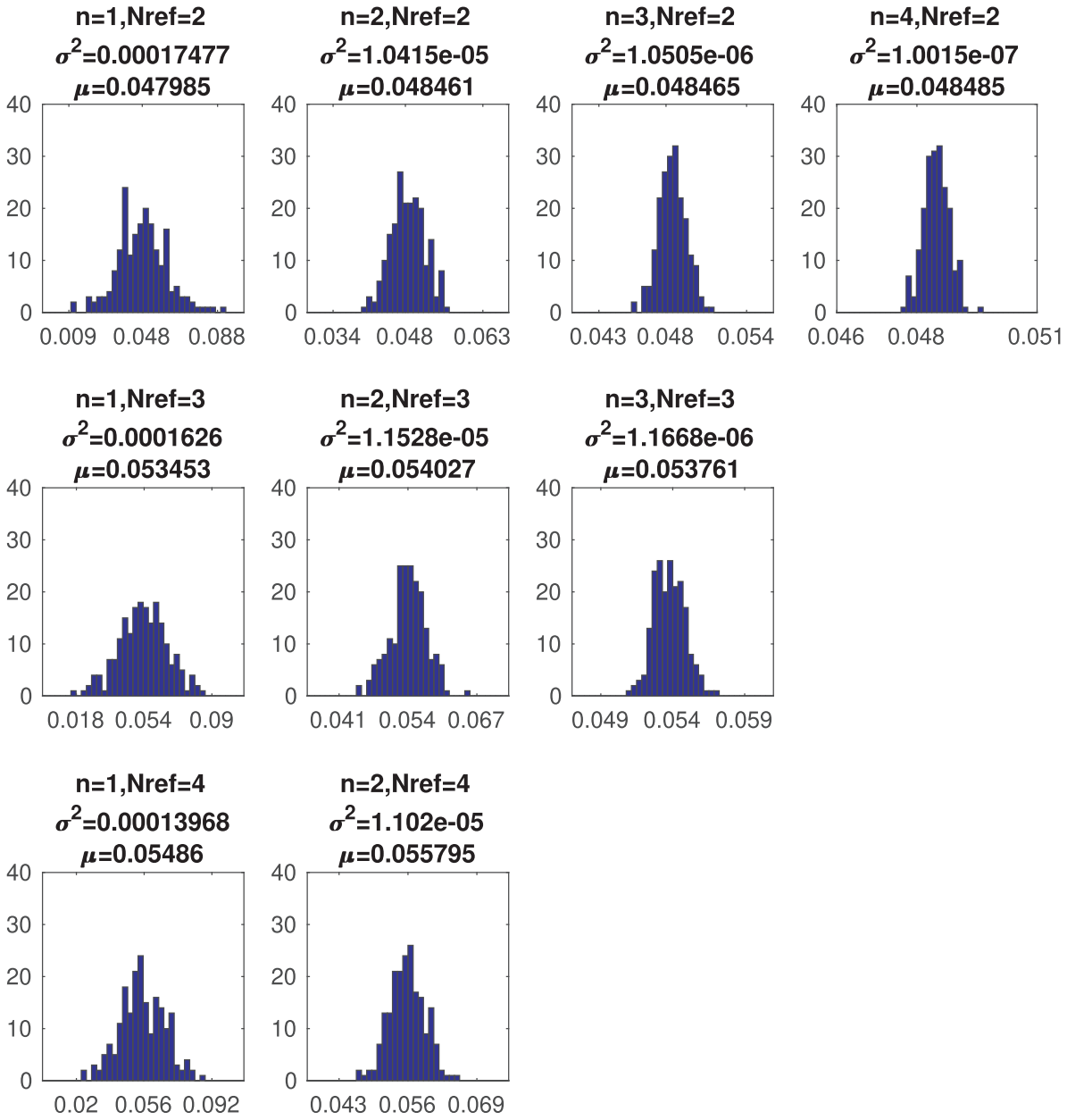

FIGURE 7. Empirical distribution of $\widehat{\sigma}^2(n, N_{\text{ref}})$ when $d = 3$, $\alpha = 1$ and $\beta = 2$, for different values of n and N_{ref}.

and 10^{-2}. We keep the parameters $n = 4$ and $N_{\text{ref}} = 5$ fixed, and use 200 samples of the estimator to compute the empirical average and standard deviation.

In the code provided provided in the GitHub repository (6.1), the choices $\alpha = 1$ and $\beta = 9$ are hard-coded, but these values can be modified by changing line 488 (or, in the three-dimensional case, line 487) of the file src/examples/homogenized_coefficients.jl.

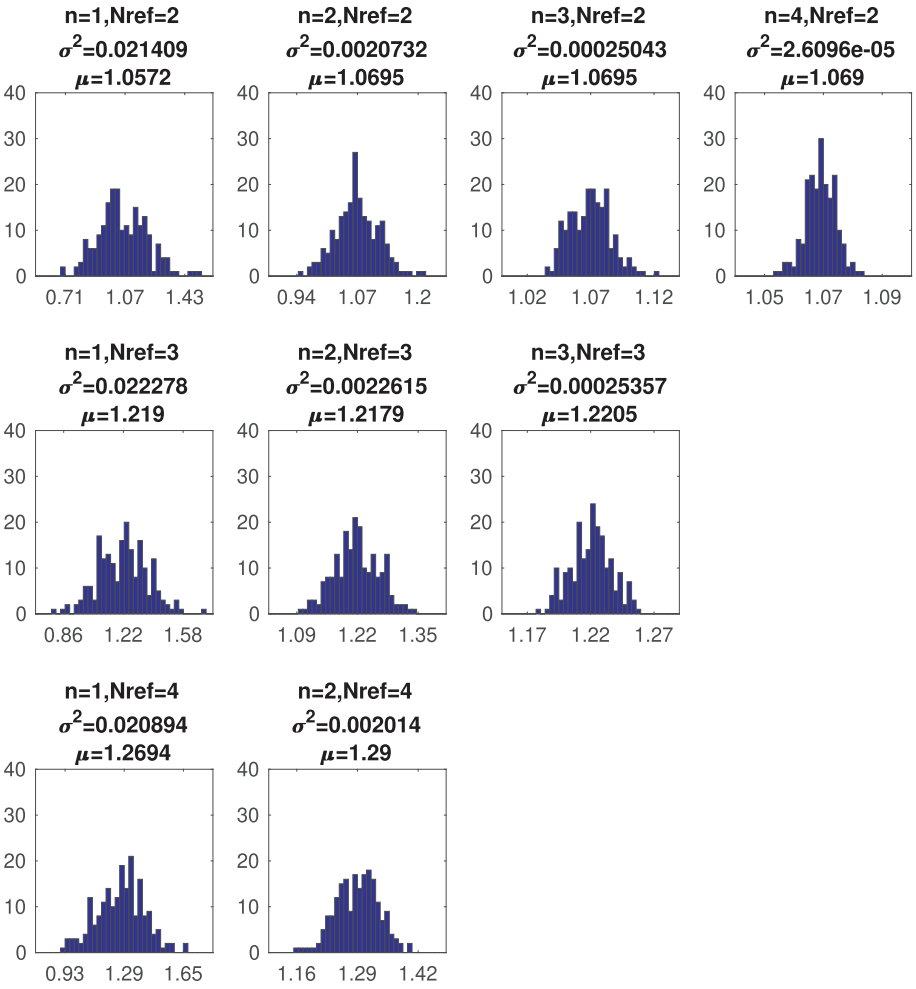

FIGURE 8. Empirical distribution of $\widehat{\sigma}^2(n, N_{\text{ref}})$ when $d = 3$, $\alpha = 1$ and $\beta = 9$, for different values of n and N_{ref}.

For these twenty values of $\alpha \in \left[10^{-2}, 1\right]$, the left frame of Figure 6 displays the mean and the standard deviation of $\widehat{\sigma}^2(n, N_{\text{ref}})$, with the choices of $n = 4$ and $N_{\text{ref}} = 5$. The estimator captures the true value of the homogenized matrix quite well, for a large span of values of α, although relative errors become large when α approaches 10^{-2}. This is in part due to the fact that the true homogenized matrix tends to zero as α is decreased to zero, and thus even a constant error in absolute value would translate into a relative error which blows up. A more fundamental reason for the increase of the error is that solutions become more and more singular, and thus accurate discretizations become more challenging. On the right frame of Figure 6, we plot the relative error in the mean, for $\alpha = 0.1$ and different values of N_{ref}. We expect the asymptotic convergence rate to scale approximately like $h^{0.1337\cdots}$, where h is the size of a finite-element cell. Despite the slow asymptotic rate, a faster pre-asymptotic regime allows to bring the relative error within a few percentage points after five levels of refinement.

6.3. Three-dimensional case, small contrast

We now turn to the investigation of three-dimensional problems. To further make the case that the scaling of the discretization error is strongly affected by the ellipticity contrast, we start by investigating a regime of

relatively small contrast: we fix $\alpha = 1$ and $\beta = 2$. As in the two-dimensional case, we plot a histogram for $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$, for different values of n and N_{ref}, see Figure 7. Each histogram is obtained by combining 200 samples of the estimator.

As a rule of thumb, we expect that the approximation $\bar{\mathbf{a}} \simeq \fint_{\mathbb{R}^d} \mathbf{a}$ improves as we increase the dimension and reduce the contrast. This is confirmed by the numerical results, which suggest that for the example considered, the difference $\fint_{\mathbb{R}^d} \mathbf{a} - \bar{\mathbf{a}}$ is about 4% of the magnitude of the homogenized matrix $\bar{\mathbf{a}}$ itself. We also see that the convergence of $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ is relatively rapid as N_{ref} increases. Finally, the variance decays roughly like $2^{-dn} = 2^{-3n}$, in agreement with the theoretical prediction.

6.4. Three-dimensional case, moderate contrast

We now turn to more sizable values of the ellipticity contrast, in three dimensions, fixing $\alpha = 1$ and $\beta = 9$. Figure 8 displays a histogram of $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ for different values of n and N_{ref}, using 200 samples per histogram.

Notice that the empirical variance of $\widehat{\sigma}^2(n, N_{\mathrm{ref}})$ does not depend much on N_{ref}. A linear regression based on the values for $N_{\mathrm{ref}} = 2$ suggests that this variance decays with n like $C3^{-\gamma n}$ for $\gamma \simeq 3.2$. This is in agreement with the theoretical prediction of $\gamma = d = 3$.

In the three-dimensional case, we are not aware of any analytic expression for the homogenized matrix. The numerical results we obtained and a naive extrapolation suggest that

$$\fint_{\mathbb{R}^d} \mathbf{a} - \bar{\mathbf{a}} \simeq 1.35\,\mathrm{Id}, \quad \text{and thus} \quad \bar{\mathbf{a}} \simeq 3.65\,\mathrm{Id}.$$

Assuming that this is correct, a $\pm 5\%$ error interval for $\bar{\mathbf{a}}$ is $[3.47, 3.83]$. An average of four samples of the quantity $5 - \widehat{\sigma}^2(n = 2, N_{\mathrm{ref}} = 3)$ falls within this interval with probability above 95%, and takes about 20 min to compute on a laptop computer with 16 Go of memory using a single 2.40 GHz processor. A single sample of the quantity $5 - \widehat{\sigma}^2(n = 2, N_{\mathrm{ref}} = 4)$ falls within the smaller interval $[3.62, 3.80]$ with 95% probability, and takes about 38 min to compute with the same piece of hardware. Moreover, the computational time can be significantly reduced by optimizing on the boundary layer size.

Acknowledgements. AH was partially supported by the Stenbäck foundation and Academy of Finland project 312340. JCM was partially supported by the ANR grants LSD (ANR-15-CE40-0020-03) and Malin (ANR-16-CE93-0003) and by a grant from the NYU–PSL Global Alliance. HS was partially supported by Academy of Finland project 305759.

References

[1] A. Abdulle, D. Arjmand and E. Paganoni, Exponential decay of the resonance error in numerical homogenization via parabolic and elliptic cell problems. *C. R. Math. Acad. Sci. Paris* **357** (2019) 545–551.

[2] M.A. Akcoglu and U. Krengel, Ergodic theorems for superadditive processes. *J. Reine Angew. Math.* **323** (1981) 53–67.

[3] Y. Almog, Averaging of dilute random media: a rigorous proof of the Clausius–Mossotti formula. *Arch. Ration. Mech. Anal.* **207** (2013) 785–812.

[4] Y. Almog, The Clausius–Mossotti formula in a dilute random medium with fixed volume fraction. *Multiscale Model. Simul.* **12** (2014) 1777–1799.

[5] Y. Almog, The Clausius–Mossotti formula for dilute random media of perfectly conducting inclusions. *SIAM J. Math. Anal.* **49** (2017) 2885–2919.

[6] A. Anantharaman and C. Le Bris, A numerical approach related to defect-type theories for some weakly random problems in homogenization. *Multiscale Model. Simul.* **9** (2011) 513–544.

[7] A. Anantharaman and C. Le Bris, Elements of mathematical foundations for numerical approaches for weakly random homogenization problems. *Commun. Comput. Phys.* **11** (2012) 1103–1143.

[8] S. Armstrong and P. Dario, Elliptic regularity and quantitative homogenization on percolation clusters. *Commun. Pure Appl. Math.* **71** (2018) 1717–1849.

[9] S.N. Armstrong and J.-C. Mourrat, Lipschitz regularity for elliptic equations with random coefficients. *Arch. Ration. Mech. Anal.* **219** (2016) 255–348.

[10] S.N. Armstrong and C.K. Smart, Quantitative stochastic homogenization of convex integral functionals. *Ann. Sci. Éc. Norm. Supér. (4)* **49** (2016) 423–481.

[11] S. Armstrong, T. Kuusi and J.-C. Mourrat, Mesoscopic higher regularity and subadditivity in elliptic homogenization. *Commun. Math. Phys.* **347** (2016) 315–361.
[12] S. Armstrong, T. Kuusi and J.-C. Mourrat, The additive structure of elliptic homogenization. *Invent. Math.* **208** (2017) 999–1154.
[13] S. Armstrong, T. Kuusi and J.-C. Mourrat, Quantitative Stochastic Homogenization and Large-Scale Regularity. In: Vol. 352 of *Grundlehren der mathematischen Wissenschaften*. Springer Nature (2019).
[14] B.K. Bergen and F. Hülsemann, Hierarchical hybrid grids: data structures and core algorithms for multigrid. *Numer. Linear Algebra App.* **11** (2004) 279–291.
[15] B. Bergen, F. Hülsemann and U. Rüde, Is 1.7×10^{10} unknowns the largest finite element system that can be solved today? In: *SC'05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing. IEEE* (2005).
[16] L. Berlyand and V. Mityushev, Generalized Clausius–Mossotti formula for random composite with circular fibers. *J. Stat. Phys.* **102** (2001) 115–145.
[17] J. Bey, Tetrahedral grid refinement. *Computing* **55** (1995) 355–378.
[18] X. Blanc and C. Le Bris, Improving on computation of homogenized coefficients in the periodic and quasi-periodic settings. *Netw. Heterogen. Media* **5** (2010) 1–29.
[19] X. Blanc, C. Le Bris and F. Legoll, Some variance reduction methods for numerical stochastic homogenization. *Philos. Trans. R. Soc. A* **374** (2016) 15.
[20] S.C. Brenner and L.R. Scott, The mathematical theory of finite element methods, 3rd edition. In: Vol. 15 of *Texts in Applied Mathematics*. Springer, New York (2008).
[21] E. Cancès, V. Ehrlacher, F. Legoll, B. Stamm and S. Xiang, An embedded corrector problem for homogenization. Part II: algorithms and discretization. *J. Comput. Phys.* **407** (2020) 109254, 26.
[22] E. Cancès, V. Ehrlacher, F. Legoll, B. Stamm and S. Xiang, An embedded corrector problem for homogenization. Part I: theory. Preprint arXiv:1807.05131 (2018).
[23] P. Dario, Optimal corrector estimates on percolation clusters. Preprint arXiv:1805.00902 (2020).
[24] DoITPoMS Micrograph Library, University of Cambridge. Available from: https://www.doitpoms.ac.uk/miclib/ (2020).
[25] M. Duerinckx and A. Gloria, Analyticity of homogenized coefficients under Bernoulli perturbations and the Clausius–Mossotti formulas. *Arch. Ration. Mech. Anal.* **220** (2016) 297–361.
[26] Y. Efendiev and T.Y. Hou, Multiscale finite element methods. In: Vol. 4 of *Surveys and Tutorials in the Applied Mathematical Sciences*. Springer, New York (2009).
[27] A.-C. Egloffe, A. Gloria, J.-C. Mourrat and T.N. Nguyen, Random walk in random environment, corrector equation and homogenized coefficients: from theory to numerics, back and forth. *IMA J. Numer. Anal.* **35** (2015) 499–545.
[28] J. Fischer, The choice of representative volumes in the approximation of effective properties of random materials. *Arch. Ration. Mech. Anal.* **234** (2019) 635–726.
[29] A. Gholami, D. Malhotra, H. Sundar and G. Biros, FFT, FMM, or multigrid? A comparative study of state-of-the-art Poisson solvers for uniform and nonuniform grids in the unit cube. *SIAM J. Sci. Comput.* **38** (2016) C280–C306.
[30] A. Gloria, Numerical approximation of effective coefficients in stochastic homogenization of discrete elliptic equations. *ESAIM: M2AN* **46** (2012) 1–38.
[31] A. Gloria and Z. Habibi, Reduction in the resonance error in numerical homogenization II: correctors and extrapolation. *Found. Comput. Math.* **16** (2016) 217–296.
[32] A. Gloria and J.-C. Mourrat, Spectral measure and approximation of homogenized coefficients. *Probab. Theory Relat. Fields* **154** (2012) 287–326.
[33] A. Gloria and F. Otto, An optimal variance estimate in stochastic homogenization of discrete elliptic equations. *Ann. Probab.* **39** (2011) 779–856.
[34] A. Gloria and F. Otto, An optimal error estimate in stochastic homogenization of discrete elliptic equations. *Ann. Appl. Probab.* 22 (2012) 1–28.
[35] A. Gloria and F. Otto, Quantitative results on the corrector equation in stochastic homogenization. *J. Eur. Math. Soc. (JEMS)* **19** (2017) 3489–3548.
[36] A. Gloria and F. Otto, The corrector in stochastic homogenization: optimal rates, stochastic integrability, and fluctuations. Preprint arXiv:1510.08290 (2016).
[37] A. Gloria, S. Neukamm and F. Otto, Quantification of ergodicity in stochastic homogenization: optimal bounds via spectral gap on Glauber dynamics. *Invent. Math.* **199** (2015) 455–515.
[38] A. Gloria, S. Neukamm and F. Otto, A regularity theory for random elliptic operators. Preprint arXiv:1409.2678 (2019).
[39] T. Gradl and U. Rüde, High performance multigrid on current large scale parallel computers. In: *9th Workshop on Parallel Systems and Algorithms* (2008).
[40] A. Hannukainen, J.-C. Mourrat and H. Stoppels, Homogenization.jl tutorial. Available from: https://haampie.github.io/Homogenization.jl/dev/ (2020).
[41] 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.
[42] T.Y. Hou, X.-H. Wu and Z. Cai, Convergence of a multiscale finite element method for elliptic problems with rapidly oscillating coefficients. *Math. Comput.* **68** (1999) 913–943.
[43] V. Khoromskaia, B.N. Khoromskij and F. Otto, Numerical study in stochastic homogenization for elliptic PDEs: convergence rate in the size of representative volume elements. Preprint arXiv:1903.12227 (2019).

[44] S.M. Kozlov, Geometric aspects of averaging. *Uspekhi Mat. Nauk* **44** (1989) 79–120.
[45] C. Le Bris and F. Legoll, Examples of computational approaches for elliptic, possibly multiscale PDEs with random inputs. *J. Comput. Phys.* **328** (2017) 455–473.
[46] C. Le Bris, F. Legoll and W. Minvielle, Special quasirandom structures: a selection approach for stochastic homogenization. *Monte Carlo Methods App.* **22** (2016) 25–54.
[47] D. Marahrens and F. Otto, Annealed estimates on the Green function. *Probab. Theory Relat. Fields* **163** (2015) 527–573.
[48] J.C. Maxwell, Medium in which small spheres are uniformly disseminated, 3rd edition. In: A Treatise on Electricity and Magnetism, part II, chapter IX. Clarendon Press (1891) 314.
[49] N.G. Meyers, An L^p-estimate for the gradient of solutions of second order elliptic divergence equations. *Ann. Scuola Norm. Sup. Pisa* **17** (1963) 189–206.
[50] J.-C. Mourrat, Variance decay for functionals of the environment viewed by the particle. *Ann. Inst. Henri Poincaré Probab. Stat.* **47** (2011) 294–327.
[51] J.-C. Mourrat, First-order expansion of homogenized coefficients under Bernoulli perturbations. *J. Math. Pures Appl.* **103** (2015) 68–101.
[52] J.-C. Mourrat, Efficient methods for the estimation of homogenized coefficients. *Found. Comput. Math.* **19** (2019) 435–483.
[53] J.-C. Mourrat, An informal introduction to quantitative stochastic homogenization. *J. Math. Phys.* **60** (2019) 11.
[54] A. Naddaf and T. Spencer, On homogenization and scaling limit of some gradient perturbations of a massless free field. *Commun. Math. Phys.* **183** (1997) 55–84.
[55] A. Naddaf and T. Spencer, Estimates on the variance of some homogenization problems (1998).
[56] G.C. Papanicolaou, Diffusion in random media. In: Vol. 1 of Surveys in Applied Mathematics. Plenum, New York (1995) 205–253.
[57] L.C. Piccinini and S. Spagnolo, On the Hölder continuity of solutions of second order elliptic equations in two variables. *Ann. Scuola Norm. Sup. Pisa* **26** (1972) 391–402.
[58] J.W. Strutt, 3d Baron Rayleigh, On the influence of obstacles arranged in rectangular order upon the properties of a medium. *Philos. Mag.* **34** (1892) 481–502.
[59] S.-H. Wei, L. Ferreira, J.E. Bernard and A. Zunger, Electronic properties of random alloys: special quasirandom structures. *Phys. Rev. B* **42** (1990) 9622.
[60] X. Yue and E. Weinan, The local microscale problem in the multiscale modeling of strongly heterogeneous media: effects of boundary conditions and cell size. *J. Comput. Phys.* **222** (2007) 556–572.
[61] A. Zunger, S.-H. Wei, L. Ferreira and J.E. Bernard, Special quasirandom structures. *Phys. Rev. Lett.* **65** (1990) 353.