Karlsruhe Institute of Technology

Wave phenomena
analysis and numerics

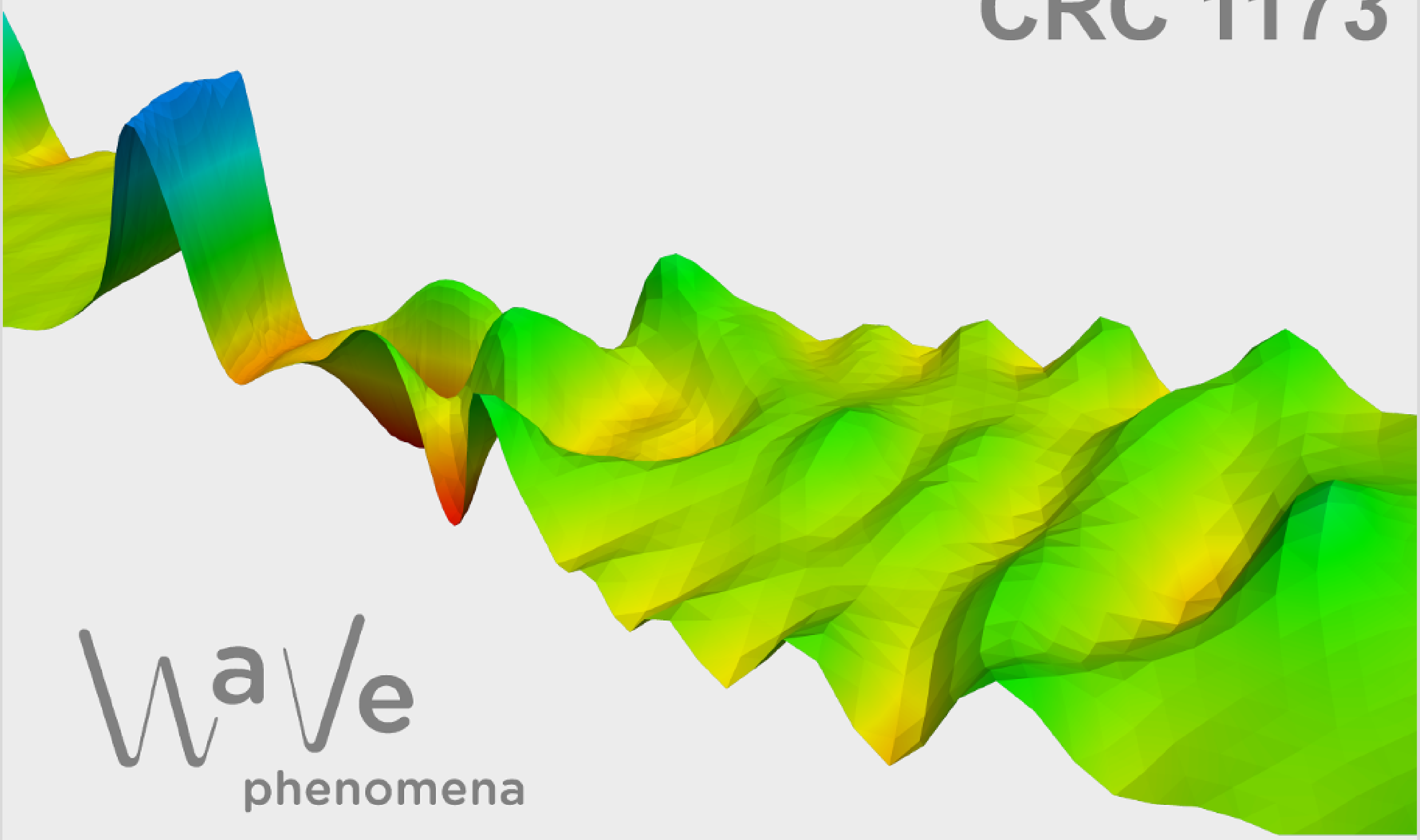
Sparse compression of expected solution operators

Michael Feischl, Daniel Peterseim

CRC Preprint 2018/21, August 2018

KARLSRUHE INSTITUTE OF TECHNOLOGY

CRC 1173



Wave phenomena

Participating universities

Universität Stuttgart

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Funded by

DFG

ISSN 2365-662X

SPARSE COMPRESSION OF EXPECTED SOLUTION OPERATORS

MICHAEL FEISCHL AND DANIEL PETERSEIM

Abstract. We show that the expected solution operator of prototypical linear elliptic partial differential operators with random coefficients is well approximated by a computable sparse matrix. This result is based on a random localized orthogonal multiresolution decomposition of the solution space that allows both the sparse approximate inversion of the random operator represented in this basis as well as its stochastic averaging. The approximate expected solution operator can be interpreted in terms of classical Haar wavelets. When combined with a suitable sampling approach for the expectation, this construction leads to an efficient method for computing a sparse representation of the expected solution operator.

1. Introduction

For a random (or parameterized) family of prototypical linear elliptic partial differential operators $\boldsymbol{\mathcal{A}}(\omega) = -\operatorname{div}(\boldsymbol{A}(\omega)\nabla\bullet)$ and a given deterministic right-hand side f, we consider the family of solutions

$$\boldsymbol{u}(\omega) := \boldsymbol{\mathcal{A}}(\omega)^{-1} f$$

with events $\omega \in \Omega$ in some probability space Ω. We define the harmonically averaged operator

$$\mathcal{A} := \Big(\mathbb{E}[\boldsymbol{\mathcal{A}}(\omega)^{-1}]\Big)^{-1}.$$

The idea behind this definition is that $\mathbb{E}(\boldsymbol{u})$ satisfies

$$\mathbb{E}[\boldsymbol{u}] = \mathcal{A}^{-1} f.$$

In this sense, $\mathcal{A}$ may be understood as a stochastically homogenized operator and $\mathcal{A}^{-1}$ is the effective solution operator. Note that this definition does not rely on probabilistic structures of the random diffusion coefficient $\boldsymbol{A}$ such as stationarity, ergodicity or any characteristic length of correlation. However, we shall emphasize that $\mathcal{A}$ does not coincide with the partial differential operator that would result from the standard theory of stochastic homogenization (under stationarity and ergodicity) [25, 29, 38] (see e.g. [3], [15, 7, 16], [1] for quantitative results). Recent works on discrete random problems on $\mathbb{Z}^d$ with i.i.d. edge conductivies indicate that $\mathcal{A}$ is rather a non-local integral operator [2, 22]. The goal of the present work is to show that, even in the more general PDE setup of this paper without any assumptions on the distribution

---

Date: July 6, 2018.

The authors would like to acknowledge the kind hospitality of the Erwin Schrödinger International Institute for Mathematics and Physics (ESI) where large parts of this research were developed under the frame of the Thematic Programme *Numerical Analysis of Complex PDE Models in the Sciences*. D. Peterseim acknowledges support by Deutsche Forschungsgemeinschaft in the Priority Program 1748 *Reliable simulation techniques in solid mechanics* (PE2143/2-2). M. Feischl acknowledges support by Deutsche Forschungsgemeinschaft (DFG) through CRC 1173.

of the random coefficient, the expected solution operator $\mathcal{A}^{-1}$ can be represented accurately by a sparse matrices R^δ in the sense that

$$\|\mathcal{A}^{-1} - R^\delta\|_{L^2(D)\to L^2(D)} \leq \delta$$

for any $\delta > 0$ while the number of non-zero entries of R^δ scales like δ^{-d} up to logarithmic-in-δ terms (see Theorem 10).

The sparse matrix representation of $\mathcal{A}^{-1}$ is based on multiresolution decompositions of the energy space in the spirit of numerical homogenization by localized orthogonal decomposition (LOD) [26, 19, 30, 12, 23, 13] and, in particular, its multiscale generalization that is popularized under the name gamblets [27]. The latter decomposition of [27] is slightly modified by linking it to classical Haar wavelets via L^2-orthogonal projections and conversely by corrections involving the solution operator (see Section 3). The resulting problem-dependent multiresolution decompositions block-diagonalize the random operator $\boldsymbol{\mathcal{A}}$ for any event in the probability space (see Section 4). The block-diagonal representations (with sparse blocks) are well conditioned and, hence, easily inverted to high accuracy using a few steps of standard linear iterative solvers. The sparsity of the inverted blocks is preserved to the degree that it deteriorates only logarithmically with the accuracy.

While the sparsity pattern of the inverted block-diagonal operator is independent of the stochastic parameter and, hence, not affected when taking the expectation (or any sample mean) the resulting object cannot be interpreted in a known basis. This issue is circumvented by reinterpreting the approximate inverse stiffness matrices in terms of the deterministic Haar basis before stochastic averaging. This leads to an accurate representation of $\mathcal{A}^{-1}$ in terms of piecewise constant functions. Sparsity is not directly preserved by this transformation but can be retained by some appropriate hyperbolic cross truncation which is justified by scaling properties of the multiresolution decomposition (see Section 5).

Apart from the mathematical question of sparse approximability of the expected operator, the above construction leads to a computationally efficient method for approximating $\mathcal{A}^{-1}$ when combined with any sampling approach for the approximation of the expectation (see Section 6). This new sparse compression algorithm for the direct discretization of $\mathcal{A}^{-1}$ may be beneficial if we want to compute $\mathbb{E}[\boldsymbol{u}]$ for multiple right-hand sides f. This, for example, is the case if we have an independent probability space $\xi \in \Xi$ influencing $f = \boldsymbol{f}(\xi)$ as well as the corresponding solution $\boldsymbol{U}(\omega,\xi) := \boldsymbol{\mathcal{A}}(\omega)^{-1}\boldsymbol{f}(\xi)$. Then, we might be interested in the average behavior $\mathbb{E}_{\Omega\times\Xi}[\boldsymbol{U}]$ which is the solution of

$$\mathbb{E}_{\Omega\times\Xi}[\boldsymbol{U}] = \mathbb{E}_{\Xi}[\mathcal{A}^{-1}\boldsymbol{f}] = \mathcal{A}^{-1}\mathbb{E}_{\Xi}[\boldsymbol{f}]. \tag{1.1}$$

As a practical example for this might serve the Darcy flow as a model of ground water flow. Here, $\boldsymbol{\mathcal{A}}$ is a random diffusion process modeling the unknown diffusion coefficient of the ground material. The right-hand side $\boldsymbol{f}$ would be the random (unknown) injection of pollutants into the ground water. Ultimately, the user would be interested in the average distribution of pollutants in the ground. Obviously, computing the right-hand side of (1.1) requires the user to sample Ω and Ξ successively, whereas computing the left-hand side of (1.1) forces the user to sample the much larger product space $\Omega \times \Xi$. Therefore, an accurate discretization of $\mathcal{A}$ can help saving significant computational cost.

2. Model problem

We consider some prototypical linear second order elliptic partial differential equation with random diffusion coefficient. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space with set of events Ω, σ-algebra $\mathcal{F} \subseteq 2^\Omega$ and probability measure $\mathbb{P}$. The expectation operator is denoted by $\mathbb{E}$. Let $D \subseteq \mathbb{R}^d$ for $d \in \{1,2,3\}$ be a bounded Lipschitz polytope with diameter of order 1. The set of admissible coefficients reads

$$\mathcal{M}(D, \gamma_{\min}, \gamma_{\max}) = \left\{ \begin{array}{r} A \in L^\infty(D; \mathbb{R}^{d\times d}_{\mathrm{sym}}) \text{ s.t. } \gamma_{\min}|\xi|^2 \le (A(x)\xi)\cdot\xi \le \gamma_{\max}|\xi|^2 \\ \text{for a.e. } x \in D \text{ and all } \xi \in \mathbb{R}^d \end{array} \right\}$$

for given uniform spectral bounds $0 < \gamma_{\min} \le \gamma_{\max} < \infty$. Here, $\mathbb{R}^{d\times d}_{\mathrm{sym}}$ denotes the set of symmetric $d \times d$ matrices. Let $\boldsymbol{A}$ be an $\mathcal{M}(D, \gamma_{\min}, \gamma_{\max})$-valued random field with $\gamma_{\max} > \gamma_{\min} > 0$. Note that we do not make any structural assumptions regarding the distribution of $\boldsymbol{A}$. Moreover, realizations in $\mathcal{M}(D, \gamma_{\min}, \gamma_{\max})$ are fairly free to vary within the bounds $\gamma_{\min}$ and $\gamma_{\max}$ without any conditions on frequencies of variation or smoothness.

Denote the energy space by $V := H^1_0(D)$ and let $f \in V^* = H^{-1}(D)$ be deterministic. The prototypical second order elliptic variational problem seeks a V-valued random field $\boldsymbol{u}$ such that, for almost all $\omega \in \Omega$,

$$\boldsymbol{a}_\omega(\boldsymbol{u}(\omega), v) := \int_D (\boldsymbol{A}(\omega)(x)\nabla \boldsymbol{u}(\omega)(x)) \cdot \nabla v(x)\, dx = f(v) \quad \text{for all } v \in V. \tag{2.1}$$

The bilinear from $\boldsymbol{a}_\omega$ depends continuously on the coefficient $\boldsymbol{A}(\omega) \in \mathcal{M}(D, \gamma_{\min}, \gamma_{\max})$ and, particularly, is measurable as a function of ω. Hence, the reformulation of this problem in the Hilbert space $L^2(\Omega; V)$ of V-valued random fields with finite second moments shows well-posedness in the sense that there exists a unique solution $\boldsymbol{u} \in L^2(\Omega; V)$ with

$$\|\nabla \boldsymbol{u}\|_{L^2(\Omega;V)} := \left(\int_\Omega \int_D |\nabla(\boldsymbol{u}(\omega))(x)|^2 \, dx \, d\mathbb{P}(\omega) \right)^{1/2} \le \gamma_{\min}^{-1} \|f\|_{V^*}.$$

To connect the model problem to the operator setting of the introduction, we shall introduce the random operator $\boldsymbol{\mathcal{A}} \colon \Omega \to \mathcal{L}(V, V^*)$ by

$$\langle \boldsymbol{\mathcal{A}}(\omega)u, v\rangle_{V^*,V} := \boldsymbol{a}_\omega(u, v)$$

for functions $u, v \in V$ and $\omega \in \Omega$. Then the model problem (2.1) can be rephrased as

$$\boldsymbol{\mathcal{A}}(\omega)\boldsymbol{u}(\omega) = f \quad \text{for almost all } \omega \in \Omega.$$

3. Coefficient-adapted hierarchical bases

Let $\mathcal{T}_\ell$, $\ell = 0, \dots, L$ denote a sequence of uniform refinements with mesh-size h_ℓ of some initial mesh $\mathcal{T}_0$ of D. We allow fairly general meshes in the sense that we only require a reference element T_{ref} together with a family of uniformly bi-Lipschitz maps $\Psi_T \colon T_{\mathrm{ref}} \to T$ for all elements $T \in \mathcal{T}_\ell$, $\ell = 0, \dots, L$. Straightforward examples are simplicial meshes generated from an initial triangulation by red refinement (or newest vertex bisection) or quadrilateral meshes generated by subdividing the elements into 2^d new elements. Particularly, hanging nodes do not pose problems as long as the other properties are observed.

The number of levels (or scales) L will typically be chosen proportional to the modulus of some logarithm of the desired accuracy $1 \gtrsim \delta > 0$. We assume $h_{\ell+1} \le h_\ell/2$. Note that any other fixed factor of mesh width reduction strictly smaller than one would do the job. Define the set of descendants of an element $T \in \mathcal{T}_\ell$ by $\mathrm{ref}(T) := \{T' \in \mathcal{T}_{\ell+1} \,:\, T' \subseteq T\}$. For each $T \in \bigcup_{\ell=0}^{L-1} \mathcal{T}_\ell$, we pick piecewise constant functions $\phi_{T,1}, \phi_{T,2}, \ldots, \phi_{T,\#\mathrm{ref}(T)} \in P^0(\mathrm{ref}(T))$ such that they are pairwise $L^2(T)$-orthogonal and $\int_T \phi_{T,j}\, dx = 0$ for all $j = 1, \ldots, \#\mathrm{ref}(T)$. With the indicator functions $\chi_{(\cdot)}$, we then define $\mathcal{H}_0 := \{\chi_T \,:\, T \in \mathcal{T}_0\}$ and for $\ell \geq 1$

$$\mathcal{H}_\ell := \bigcup_{T \in \mathcal{T}_{\ell-1}} \{\phi_{T,j} \,:\, j = 1, \ldots, \#\mathrm{ref}(T)\}. \tag{3.1}$$

We define a Haar basis via

$$\mathcal{H} := \bigcup_{\ell=0}^{L} \mathcal{H}_\ell.$$

Lemma 1. *The basis $\mathcal{H}$ is L^2-orthogonal and local in the sense that $\phi \in \mathcal{H}_\ell$ satisfies* $\mathrm{supp}(w) = T$ *for some $T \in \mathcal{T}_{\ell-1}$ or $T \in \mathcal{T}_0$ for $\ell = 0$.*

Proof. If $k = \ell$ then the interiors of the supports of any $\phi_1 \neq \phi_2 \in \mathcal{H}_k$ are disjoint which implies $L^2(D)$ orthogonality. If $k < \ell$, we have that ϕ_k is constant on $\mathrm{supp}(\phi_\ell)$. Since $\int_D \phi_\ell\, dx = 0$ by definition, this concludes the proof of L^2-orthogonality. Locality follows readily from the construction. ☐

Remark 2. *For uniform Cartesian meshes, $\mathcal{H}$ is the Haar basis. The choice of the $2^d - 1$ generating functions follows the standard procedure for Haar wavelets (see e.g.* [35]*). The construction is applicable to general meshes that are not based on tensor-product structures.*

Due to the lack of V-conformity, the basis $\mathcal{H}$ is not suited for approximating the solution of model problem (2.1) in a Galerkin approach. It will, however, serve as a companion of certain regularized hierarchical bases $\boldsymbol{\mathcal{B}}(\omega) = \bigcup_{\ell=0}^{L} \boldsymbol{\mathcal{B}}_\ell(\omega) \subset V$ to be defined below. The new bases are connected to $\mathcal{H}$ (and to each other) via L^2-orthogonal projections $\Pi_\ell \colon V \to P^0(\mathcal{T}_\ell)$ onto $\mathcal{T}_\ell$-piecewise constant functions by

$$\Pi_\ell \boldsymbol{\mathcal{B}}_\ell(\omega) = \mathcal{H}_\ell \tag{3.2}$$

for all $\ell = 0, 1, \ldots, L$ and $\omega \in \Omega$. Among the infinitely many possible choices, we define the elements of $\boldsymbol{\mathcal{B}}_\ell(\omega)$ by minimizing the energies $\frac{1}{2}\boldsymbol{a}_\omega(\bullet, \bullet)$ in the closed affine space of preimages of Π_ℓ restricted to V, i.e., given $\phi \in \mathcal{H}_\ell$ and $\omega \in \Omega$, we define $\boldsymbol{b}_\phi(\omega) \in \boldsymbol{\mathcal{B}}_\ell(\omega)$ by

$$\boldsymbol{b}_\phi(\omega) := \operatorname*{argmin}_{v \in V} \tfrac{1}{2}\boldsymbol{a}_\omega(v, v) \quad \text{subject to} \quad \Pi_\ell v = \phi. \tag{3.3}$$

This construction is strongly inspired by numerical homogenization where this sort of orthogonalization of scales in the energy space paved the way to a scheme that works with arbitrary rough coefficients beyond periodicity or scale separation [26, 19, 30, 23]. While most results in the context of this socalled localized orthogonal decomposition (LOD) were based on a conforming companion (the Faber basis), early works also addressed the possibility of using discontinuous companions [8, 9]. The dG version is very useful when taking the step from two levels or scales in numerical homogenization to actual multilevel decomposition. This was first shown in [27] where so-called gamblets are introduced; see also [34, 20, 21, 28]. In particular, piecewise constants induce a natural hierarchical structure with nested kernels of local projection operators (here the Π_ℓ) that is not easily achieved with H^1-conforming functions. The construction of the present paper coincides with the gamblet decomposition of [27] in the sense that the approximation spaces on all levels coincide in some idealized deterministic setting. However, our particular choice of basis is connected to the Haar-wavelets which simplifies the derivation and decouples the computation of the basis across levels so that error amplification through nested iterations is avoided. More importantly, our particular choice of basis is crucial in the context of the random problem at hand because it is exactly the link to the deterministic Haar basis that allows a meaningful interpretation of the averaged object.

We shall express the mapping of bases encoded in (3.2)–(3.3) in terms of two concatenated linear operators. This will be useful for both analysis and actual computations. First, let $\tilde{\Pi}_\ell : L^2(D) \to V$ be such that

$$\Pi_\ell \circ \tilde{\Pi}_\ell = \mathrm{id} \quad \text{on } \mathrm{span}\mathcal{H}_\ell \tag{3.4}$$

In particular, this means that $\tilde{\Pi}_\ell$ maps any $\phi \in \mathcal{H}_\ell$ to some function that is admissible in the sense of the minimization problem (3.3). The operators $\tilde{\Pi}_\ell$ are easily constructed using nonnegative bubble functions $\tilde{\chi}_T$ supported on a element $T \in \mathcal{T}_\ell$ with $\Pi_\ell \tilde{\chi}_T = \chi_T$. Then

$$\tilde{\Pi}_\ell v = \sum_{T \in \mathcal{T}_\ell} (\Pi_\ell v)|_T \tilde{\chi}_T.$$

There is even locality in the sense of

$$\operatorname{supp} \tilde{\Pi}_\ell \phi \subset \operatorname{supp} \phi \tag{3.5}$$

for all $\phi \in \mathrm{span}\mathcal{H}_\ell$. The bubbles can be chosen such that, for some $C > 0$,

$$\|\tilde{\Pi}_\ell \phi\|_{H^m(D)} \leq C h^{-m} \|\phi\|_{L^2(D)} \tag{3.6}$$

holds for $m \in \{0, 1\}$.

The second step involves $\boldsymbol{a}_\omega$-orthogonal projections $\boldsymbol{\mathcal{C}}_\ell(\omega)$ onto the closed subspaces

$$W_\ell := \mathrm{kernel}(\Pi_\ell|_V) = \mathrm{kernel}(\tilde{\Pi}_\ell|_V) \tag{3.7}$$

of V. Given any $u \in V$, define $\boldsymbol{\mathcal{C}}_\ell(\omega) u \in W_\ell$ as the unique solution of the variational problem

$$\boldsymbol{a}_\omega(\boldsymbol{\mathcal{C}}_\ell(\omega) u, v) = \boldsymbol{a}_\omega(u, v) \quad \text{for all } v \in W_\ell. \tag{3.8}$$

With the two operators $\tilde{\Pi}_\ell$ and $\boldsymbol{\mathcal{C}}_\ell$ we rewrite (3.3) as

$$\boldsymbol{b}_\phi = (\mathrm{id} - \boldsymbol{\mathcal{C}}_\ell) \tilde{\Pi}_\ell \phi$$

for all $\phi \in \mathcal{H}_\ell$ and $\ell = 0, 1, \ldots, L$. Actually, for any $\omega \in \Omega$, $(\mathrm{id} - \boldsymbol{\mathcal{C}}_\ell(\omega))\tilde{\Pi}_\ell$ defines a bijection from $\mathcal{H}$ to $\boldsymbol{\mathcal{B}}(\omega)$ with left inverse Π_ℓ.

While the L^2-orthogonality of the Haar basis is not preserved under these mappings, we have achieved $\boldsymbol{a}$-orthogonality between the levels of the hierarchies.

Lemma 3 ($\boldsymbol{a}$-orthogonality and scaling of $\mathcal{B}$). *Any two functions $b_k \in \mathcal{B}_k(\omega)$ and $b_\ell \in \mathcal{B}_\ell(\omega)$ with $k \neq \ell$ satisfy*

$$\boldsymbol{a}_\omega(b_k, b_\ell) = 0.$$

Moreover,

$$C^{-1} \|\phi_k\|_{L^2(D)} \leq C^{-1} \|b_k\|_{L^2(D)} \leq h_k \|\!|\!| b_k |\!|\!\|_\omega \leq C \|\phi_k\|_{L^2(D)}$$

with some generic constant $C > 0$ independent of b_k, the mesh sizes and the event.

Proof. Since

$$\Pi_k(\mathrm{id} - \boldsymbol{\mathcal{C}}_\ell(\omega))\tilde{\Pi}_\ell \mathcal{H}_\ell = \Pi_k \Pi_\ell(\mathrm{id} - \boldsymbol{\mathcal{C}}_\ell(\omega))\tilde{\Pi}_\ell \mathcal{H}_\ell = \Pi_k \mathcal{H}_\ell = \{0\} \tag{3.9}$$

whenever $k < \ell$, we have that

$$\boldsymbol{\mathcal{B}}_\ell(\omega) \subset W_k.$$

This and the orthogonality

$$\boldsymbol{a}_\omega(\boldsymbol{\mathcal{B}}_k(\omega), W_k) = 0$$

from (3.8) proves the orthogonality. The scaling follows from $\Pi_{k-1}\boldsymbol{\mathcal{B}}_k = \{0\}$ (which is a special instance of (3.9)), the Poincaré inequality, (3.6), and the construction. More precisely,

$$\begin{aligned} \|\phi_k\|_{L^2(D)} &= \|\Pi_k b_k\|_{L^2(D)} \leq \|b_k\|_{L^2(D)} = \|(1-\Pi_{k-1})b_k\|_{L^2(D)} \lesssim h_k |||b_k|||_\omega \\ &= h_k |||(1-\boldsymbol{\mathcal{C}}(\omega))\tilde{\Pi}_k \phi_k|||_\omega \leq h_k |||\tilde{\Pi}_k \phi_k|||_\omega \lesssim \|\phi_k\|_{L^2(D)}. \end{aligned} \tag{3.10}$$

This concludes the proof. $\square$

We shall emphasize that, in general, the basis elements $\boldsymbol{b}_\phi(\omega)$ have global support in D. However, their moduli decay exponentially away from $\operatorname{supp}\phi$ in scales of h_ℓ,

$$\|\boldsymbol{b}_\phi(\omega)\|_{H^1(D\setminus B_R(\operatorname{supp}\phi))} \leq C e^{-cR/h_\ell} \|\boldsymbol{b}_\phi(\omega)\|_{H^1(D)} \tag{3.11}$$

with some generic constants $c, C > 0$ that solely depend on the contrast $\gamma_{\max}/\gamma_{\min}$ and the shape regularity of the mesh $\mathcal{T}_\ell$ but not on the mesh size. This is a well established result of numerical homogenization since [26] and valid in many different settings (see [30] and references therein). Here, we will provide some elements of a more recent constructive proof of the decay that provides local approximations by the theory of preconditioned iterative [23] which in turn is based on [24].

We start with introducing an overlapping decomposition of D that we will later use to define the local preconditioner. Let the level $\ell \in \{0, 1, \ldots, L\}$ and the event $\omega \in \Omega$ be arbitrary but fixed. For any vertex of the mesh (including boundary vertices), define the patch

$$D_z := \bigcup \{T \in \mathcal{T}_\ell \mid z \in T\}$$

and a corresponding local subspace

$$V_z := \big\{ v \in V \mid v = 0 \text{ in } D \setminus D_z \big\} \subset V.$$

Note that V_z is equal to $H^1_0(D_Z)$ up to extension by zero outside of D_z. Under the complementary projection $(\mathrm{id} - \tilde{\Pi}_\ell)$ these subspaces are turned into subspaces

$$W_z := (\mathrm{id} - \tilde{\Pi}_\ell) V_z = \big\{ v \in W_\ell \mid v = 0 \text{ in } D \setminus D_z \big\}$$

of W_ℓ. For each $z \in \mathcal{N}(\mathcal{T}_\ell)$ we define the corresponding $\boldsymbol{a}_\omega$-orthogonal projection $\boldsymbol{\mathcal{P}}_z(\omega) \colon V \to W_z \subset W_\ell \subset V$ by the variational problem

$$\boldsymbol{a}_\omega(\boldsymbol{\mathcal{P}}_z(\omega) u, w) = \boldsymbol{a}_\omega(u, w) \quad \text{for all } w \in W_z.$$

The sum of these local Ritz projections

$$\boldsymbol{\mathcal{P}}_\ell(\omega) := \sum\nolimits_{z \in \mathcal{N}(\mathcal{T}_\ell)} \boldsymbol{\mathcal{P}}_z(\omega) \tag{3.12}$$

defines a bounded linear operator from V to W_ℓ that can be seen as a preconditioned version of the correction operator $\boldsymbol{\mathcal{C}}_\ell(\omega)$. The operator $\boldsymbol{\mathcal{P}}_\ell(\omega)$ is quasi-local with respect to the mesh $\mathcal{T}_\ell$ since information can only propagate over distances of order h_ℓ each time $\boldsymbol{\mathcal{P}}_\ell(\omega)$ is applied.

The remaining part of this section aims to show that the preconditioned operators $\boldsymbol{\mathcal{P}}_\ell(\omega)$ serve well within iterative solvers for linear equations. Following the abstract theory for subspace correction or additive Schwarz methods for operator equations [24] (see also [36, 37] for the matrix case) we need to verify that the energy norm of a function $u \in V$ can be bounded in terms of the sum of local contributions from V_T and, for one specific decomposition, we need a reverse estimate.

Lemma 4. *For every decomposition $u = \sum_{z\in\mathcal{N}(\mathcal{T}_\ell)} u_z$ of $u \in W_\ell$ with $u_z \in W_z$ we have*

$$\|\nabla u\|^2_{L^2(D)} \leq K_2 \sum\nolimits_{z\in\mathcal{N}(\mathcal{T}_\ell)} \|\nabla u_z\|^2_{L^2(D)}$$

with constant $K_2 > 0$ depending only on the shape regularity of $\mathcal{T}_\ell$. With the hat-function λ_z associated with the node z in $\mathcal{T}_\ell$, the one decomposition $\sum_{z\in\mathcal{N}(\mathcal{T}_\ell)} u_z = u$ with $u_z := (1-\tilde{\Pi}_\ell)(\lambda_z u) \in W_z$ for $z \in \mathcal{N}(\mathcal{T}_\ell)$ satisfies

$$\sum\nolimits_{z\in\mathcal{N}(\mathcal{T}_\ell)} \|\nabla u_z\|^2_{L^2(D)} \leq K_1 \|\nabla u\|^2_{L^2(D)}$$

with constant $K_1 > 0$ that only depends on the shape regularity of $\mathcal{T}_\ell$ and the contrast $\gamma_{\max}/\gamma_{\min}$.

Proof. With K_2 the maximum number of elements of $\mathcal{T}_\ell$ covered by one patch D_z for $z \in \mathcal{N}(\mathcal{T}_\ell)$, we can estimate on a single element,

$$\|\nabla u\|^2_{L^2(T)} = \|\sum_{z\in\mathcal{N}(\mathcal{T}_\ell)} \nabla u_z\|^2_{L^2(T)} \leq K_2 \sum_{z\in\mathcal{N}(\mathcal{T}_\ell)} \|\nabla u_z\|^2_{L^2(T)}.$$

Due to shape regularity of $\mathcal{T}_\ell$, K_2 is independent of h_ℓ. A summation over all T yields the first inequality. The second one follows from the H^1-stability of $\tilde{\Pi}_\ell$ on W_ℓ, the product rule, (3.6), and the Poincaré inequality. For further details, we refer to [23, Lemma 3.1] where these results are proved in detail in a very similar setting. $\square$

Lemma 4 implies that

$$1/K_1 \boldsymbol{a}_\omega(v,v) \leq \boldsymbol{a}_\omega(\boldsymbol{\mathcal{P}}_\ell(\omega)v, v) \leq K_2 \boldsymbol{a}_\omega(v,v) \tag{3.13}$$

holds for functions v in the kernel W_ℓ of $\Pi_\ell|_V$ and any $\omega \in \Omega$ (cf. [23, Eq. (3.11)]). Following the construction of [24, 23] there exists a localized linear approximation $\boldsymbol{\mathcal{C}}^\delta_\ell(\omega)$ based on $\mathcal{O}(\log(1/\delta))$ steps of some linear iterative solver applied to the preconditioned corrector problems [23, Eqns. (3.8) or (3.18)] such that

$$\|\nabla(\boldsymbol{\mathcal{C}}_\ell(\omega)u - \boldsymbol{\mathcal{C}}^\delta_\ell(\omega)u)\|_{L^2(D)} \leq \delta\|\nabla\boldsymbol{\mathcal{C}}_\ell(\omega)u\|_{L^2(D)}; \tag{3.14}$$

see [23, Lemma 3.2]. With the approximate correctors, we can define modified (localized) bases

$$\mathcal{B}^\delta(\omega) := \bigcup_{\ell=0}^{L} \mathcal{B}^\delta_\ell(\omega) := \bigcup_{\ell=0}^{L} \big\{\boldsymbol{b}^\delta_\phi(\omega) \,:\, \phi \in \mathcal{H}_\ell\big\},$$

where

$$\boldsymbol{b}^\delta_\phi(\omega) := (\mathrm{id} - \boldsymbol{\mathcal{C}}^\delta_\ell(\omega))\tilde{\Pi}_\ell\phi$$

for $\phi \in \mathcal{H}_\ell$. The previous discussion shows that there exist constants $C_1, C_2 > 0$ that only depend on the shape regularity of the meshes $\mathcal{T}_\ell$ and the contrast $\gamma_{\max}/\gamma_{\min}$ of the coefficients such that

$$|||\boldsymbol{b}_\phi(\omega) - \boldsymbol{b}^\delta_\phi(\omega)|||_\omega \leq C_1\delta|||\boldsymbol{b}_\phi(\omega)|||_\omega$$

while

$$\operatorname{supp} \boldsymbol{b}^\delta_\phi(\omega) \subset \big\{x \in D \,:\, \operatorname{dist}(x, \operatorname{supp}\phi) \leq C_2|\log(\delta)|h_\ell\big\}. \tag{3.15}$$

4. Sparse stiffness matrices and basis transformations

With the localized bases of the previous section, we can now study the sparsity of corresponding stiffness matrices and their inverses. We define the level function $\operatorname{lev}(\cdot)$ according to the Haar basis by $\operatorname{lev}(\boldsymbol{b}) = \operatorname{lev}(\boldsymbol{b}^\delta) = \operatorname{lev}(\phi) = \ell$ for $\boldsymbol{b} = \boldsymbol{b}_\phi \in \mathcal{B}_\ell(\omega)$, $b^\delta = b^\delta_\phi \in \mathcal{B}^\delta_\ell(\omega)$ and $\phi \in \mathcal{H}_\ell$. We order the basis functions in $\mathcal{B}$, $\mathcal{B}^\delta$, and $\mathcal{H}$ such that lev is monotonically increasing in the index running from 1 to $N := \#\mathcal{B} = \#\mathcal{B}^\delta = \#\mathcal{H}$. With this convention, we may also write $\operatorname{lev}(i) := \operatorname{lev}(\boldsymbol{b}_i) = \operatorname{lev}(\boldsymbol{b}^\delta_i) = \operatorname{lev}(\phi_i)$. Moreover, we define a (semi-)metric $d(\cdot,\cdot)$ on $\{1,\ldots,N\}$ by

$$d(i,j) := \frac{\operatorname{dist}(\operatorname{mid}(\phi_i),\operatorname{mid}(\phi_j))}{h_{\min\{\operatorname{lev}(i),\operatorname{lev}(j)\}}},$$

where $\operatorname{mid}(w)$ defines the barycenter of $\operatorname{supp}(w)$.

Define the stiffness matrices $\boldsymbol{S}(\omega) \in \mathbb{R}^{N\times N}$ associated with the bases $\boldsymbol{\mathcal{B}}(\omega)$ by

$$\boldsymbol{S}(\omega)_{ij} := a_\omega\Big(\frac{\boldsymbol{b}_j}{|||\boldsymbol{b}_j|||_\omega}, \frac{\boldsymbol{b}_i}{|||\boldsymbol{b}_i|||_\omega}\Big).$$

The orthogonality of the bases $\boldsymbol{\mathcal{B}}$ motivates the approximation of the stiffness matrices by block-diagonal ones even after localization. Given $\delta > 0$, define the block-diagonal stiffness matrices $\boldsymbol{S}^\delta(\omega) \in \mathbb{R}^{N\times N}$ by

$$\boldsymbol{S}^\delta(\omega)_{ij} := \begin{cases} a_\omega\big(\frac{\boldsymbol{b}^\delta_j}{|||\boldsymbol{b}^\delta_j|||_\omega}, \frac{\boldsymbol{b}^\delta_i}{|||\boldsymbol{b}^\delta_i|||_\omega}\big) & \text{for } \operatorname{lev}(i) = \operatorname{lev}(j),\\ 0 & \text{else.}\end{cases}$$

Lemma 5. *There exists a constant $C > 0$ that depends only on D and the shape regularity of $\mathcal{T}_0$ such that, for any $\omega \in \Omega$,*

$$\|\boldsymbol{S}(\omega) - \boldsymbol{S}^\delta(\omega)\|_2 \le C\delta.$$

Moreover, there exists a constant $\zeta > 0$ which depends only on D such that

$$d(i,j) > \zeta(|\log(\delta)| + 1) \text{ or } \operatorname{lev}(i) \ne \operatorname{lev}(j) \quad \Longrightarrow \quad \boldsymbol{S}^\delta_{ij}(\omega) = 0, \tag{4.1}$$

in particular, the number of nonzero entries $\operatorname{nnz}(\boldsymbol{S}^\delta(\omega)) \lesssim N(1 + |\log\delta|)^d$ is bounded uniformly in ω.

Proof. The sparsity of the diagonal blocks follows from (3.15). For the proof of the error bound, define

$$\widetilde{\boldsymbol{S}}^\delta(\omega)_{ij} := \begin{cases} a_\omega\big(\frac{\boldsymbol{b}_j}{|||\boldsymbol{b}_j|||_\omega}, \frac{\boldsymbol{b}^\delta_i}{|||\boldsymbol{b}^\delta_i|||_\omega}\big) & \text{for } \operatorname{lev}(i) = \operatorname{lev}(j),\\ 0 & \text{else.}\end{cases}$$

Since $|\boldsymbol{S}_{ij} - \widetilde{\boldsymbol{S}}^\delta_{ij}| = 0$ whenever $\operatorname{lev}(i) \ne \operatorname{lev}(j)$ it suffices to bound the errors related to the diagonal blocks indexed by $\ell = 1,2,\ldots,L$. We have for any vectors $x, y \in \mathbb{R}^{\#\mathcal{B}^\delta_\ell}$ that

$$x\cdot(\boldsymbol{S}_\ell(\omega) - \widetilde{\boldsymbol{S}}^\delta_\ell(\omega))y = \boldsymbol{a}_\omega\Big(\sum_{\operatorname{lev}(i)=\ell}\frac{\boldsymbol{b}_i}{|||\boldsymbol{b}_i|||_\omega}x_i\;,\;(\boldsymbol{\mathcal{C}}^\delta_\ell - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell\sum_{\operatorname{lev}(j)=\ell}\frac{\phi_j}{|||\boldsymbol{b}_j|||_\omega}y_i\Big).$$

With (3.14) and because $\boldsymbol{\mathcal{C}}_\ell$ is H^1-bounded by construction, this implies

$$|x\cdot(\boldsymbol{S}_\ell(\omega) - \widetilde{\boldsymbol{S}}^\delta_\ell(\omega))y| \lesssim \delta\,\Big\|\tilde{\Pi}_\ell\sum_{\operatorname{lev}(i)=\ell}\frac{\phi_i}{|||\boldsymbol{b}_i|||_\omega}x_i\Big\|_{H^1(D)}\Big\|\tilde{\Pi}_\ell\sum_{\operatorname{lev}(j)=\ell}\frac{\phi_j}{|||\boldsymbol{b}_j|||_\omega}y_j\Big\|_{H^1(D)}.$$

The construction of $\tilde{\Pi}_\ell$ reveals that, for any $v \in \operatorname{span}\mathcal{H}_\ell$, $\|\tilde{\Pi}_\ell v\|_{H^1(D)} \lesssim h_\ell^{-1}\|v\|_{L^2(D)}$. Since $|||\boldsymbol{b}_i|||_\omega \simeq h_\ell^{-1}\|\phi_i\|_{L^2(D)}$ (by Lemma 3), there holds

$$|x \cdot (\boldsymbol{S}_\ell(\omega) - \widetilde{\boldsymbol{S}}_\ell^\delta(\omega))y| \lesssim \delta \Big\| \sum_{\operatorname{lev}(i)=\ell} \frac{\phi_i}{\|\phi_i\|_{L^2(D)}} x_i \Big\|_{L^2(D)} \Big\| \sum_{\operatorname{lev}(j)=\ell} \frac{\phi_j}{\|\phi_j\|_{L^2(D)}} y_j \Big\|_{L^2(D)}$$
$$\simeq \delta \|x\|_{\ell_2} \|y\|_{\ell_2}.$$

The same arguments show $|x \cdot (\boldsymbol{S}_\ell^\delta(\omega) - \widetilde{\boldsymbol{S}}_\ell^\delta(\omega))y| \lesssim \delta\|x\|_{\ell_2}\|y\|_{\ell_2}$ and the triangle inequality readily proves the assertion. □

Lemma 6. *The normalized set $\mathcal{B} = \boldsymbol{\mathcal{B}}(\omega)$ or $\mathcal{B} = \boldsymbol{\mathcal{B}}^\delta(\omega)$ for some $\omega \in \Omega$ is a Riesz bases in the sense that*

$$C^{-1} \sum_{b \in \mathcal{B}} \alpha_b^2 \leq \Big\| \sum_{b \in \mathcal{B}} \alpha_b \frac{b}{|||b|||_\omega} \Big\|_{H^1(D)}^2 \leq C \sum_{b \in \mathcal{B}} \alpha_b^2$$

holds with some constant $C > 0$ which depends only on D and the fact that $\delta^2 L \lesssim 1$.

Proof. Since $\|\cdot\|_{H^1(D)}$ and $|||\cdot|||_\omega$ are equivalent uniformly in ω and the $\boldsymbol{\mathcal{B}}(\omega)$ is $a_\omega(\cdot,\cdot)$-orthogonal across the levels, it suffices to consider one level $k \in \{1, \dots, L\}$. For this on level, the results follows from the $L^2(D)$-orthogonality of the Haar basis, (3.10) (which remains valid when replacing b_k by any linear combination of $\boldsymbol{\mathcal{B}}_k(\omega)$ functions and ϕ_k by the corresponding linear combination of $\mathcal{H}_k$ functions), and proper rescaling by normalization. To see the result for $\mathcal{B} = \boldsymbol{\mathcal{B}}^\delta(\omega)$, note that Lemma 5 imlies

$$\Big\| \sum_{\boldsymbol{b}^\delta \in \boldsymbol{\mathcal{B}}_\ell^\delta(\omega)} \alpha_{\boldsymbol{b}^\delta} \frac{\boldsymbol{b}^\delta}{|||\boldsymbol{b}^\delta|||_\omega} \Big\|_{H^1(D)}^2 \simeq \boldsymbol{S}_\ell^\delta(\omega)\alpha \cdot \alpha \simeq \boldsymbol{S}_\ell(\omega)\alpha \cdot \alpha \pm \delta\|\alpha\|_{\ell_2}^2 \simeq (1 \pm \delta)\|\alpha\|_{\ell_2}^2$$

for $\delta \lesssim 1$. As in the proof of Lemma 5, we see

$$\begin{aligned} &\boldsymbol{a}_\omega\Big(\sum_{\boldsymbol{b}^\delta \in \boldsymbol{\mathcal{B}}_\ell^\delta(\omega)} \alpha_{\boldsymbol{b}^\delta} \frac{\boldsymbol{b}^\delta}{|||\boldsymbol{b}^\delta|||_\omega}, \sum_{\boldsymbol{b}^\delta \in \boldsymbol{\mathcal{B}}_k^\delta(\omega)} \beta_{\boldsymbol{b}^\delta} \frac{\boldsymbol{b}^\delta}{|||\boldsymbol{b}^\delta|||_\omega} \Big) \\ &\quad = \boldsymbol{a}_\omega\Big((\boldsymbol{\mathcal{C}}_\ell^\delta - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell \sum_{\boldsymbol{b}_\phi^\delta \in \boldsymbol{\mathcal{B}}_\ell^\delta(\omega)} \beta_{\boldsymbol{b}_\phi^\delta} \frac{\phi}{|||\boldsymbol{b}_\phi^\delta|||_\omega}, (\boldsymbol{\mathcal{C}}_\ell^\delta - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell \sum_{\boldsymbol{b}_\phi^\delta \in \boldsymbol{\mathcal{B}}_k^\delta(\omega)} \alpha_{\boldsymbol{b}_\phi^\delta} \frac{\phi}{|||\boldsymbol{b}_\phi^\delta|||_\omega} \Big) \\ &\quad \lesssim \delta^2 \|\alpha\|_{\ell_2} \|\beta\|_{\ell_2}. \end{aligned}$$

Altogether, this shows

$$\Big\| \sum_{\boldsymbol{b}^\delta \in \boldsymbol{\mathcal{B}}^\delta(\omega)} \alpha_{\boldsymbol{b}^\delta} \frac{\boldsymbol{b}^\delta}{|||\boldsymbol{b}^\delta|||_\omega} \Big\|_{H^1(D)}^2 \simeq \sum_{\ell=1}^{L} (1 \pm \delta)\|\alpha|_{\boldsymbol{\mathcal{B}}_\ell^\delta}\|_{\ell_2}^2 \pm \delta^2 \sum_{\substack{i,j=1 \\ i \neq j}}^{L} \|\alpha|_{\boldsymbol{\mathcal{B}}_i^\delta}\|_{\ell_2} \|\alpha|_{\boldsymbol{\mathcal{B}}_j^\delta}\|_{\ell_2}$$
$$\simeq (1 \pm \delta \pm \delta^2 L)\|\alpha\|_{\ell_2}^2$$

and concludes the proof. □

In the remaining part of this section we analyze the properties of a certain matrix representation of the $L^2(D)$-orthogonal projections Π_ℓ. Given $\omega \in \Omega$, define the matrix $\boldsymbol{T}(\omega) \in \mathbb{R}^{N \times N}$ by

$$\boldsymbol{T}_{ij}(\omega) := \frac{(\boldsymbol{b}_j, \phi_i)_{L^2(D)}}{|||\boldsymbol{b}_j|||_\omega \|\phi_i\|_{L^2(D)}^2}.$$

Given some $v = \sum_{i=1}^N \alpha_i \frac{\boldsymbol{b}_i}{\|\boldsymbol{b}_i\|_\omega}$ with $\Pi_L v = \sum_{i=1}^N \beta_i \frac{\phi_i}{\|\phi_i\|_{L^2(D)}}$. Then, by definition

$$\beta_i = \frac{(\Pi_L v, \phi_i)}{\|\phi_i\|^2_{L^2(D)}} = \sum_{j=1}^N \alpha_j \boldsymbol{T}_{ij} = (\boldsymbol{T}\alpha)_i,$$

i.e., $\beta = \boldsymbol{T}(\omega)\alpha$. Given $\delta > 0$, a truncated approximation $\boldsymbol{T}^\delta(\omega)$ of $\boldsymbol{T}(\omega)$ is defined by

$$\boldsymbol{T}^\delta_{ij}(\omega) := \begin{cases} \frac{(\boldsymbol{b}_j^\delta, \phi_i)_{L^2(D)}}{\|b_j^\delta\|_\omega \|\phi_i\|^2_{L^2(D)}} & \text{if } \operatorname{lev}(j) \le \operatorname{lev}(i), \\ 0 & \text{otherwise,} \end{cases}$$

for any $i, j \in \{1, \ldots, N\}$. $\boldsymbol{T}^\delta(\omega)$ is a sparse lower block-triangular matrix and the next lemma shows that the error of truncation is at most proportional to δ. To explore the block-structure of matrices we shall introduce the following notation. For any matrix $K \in \mathbb{R}^{N\times N}$, we define sub-blocks $K_{(\ell,k)} \in \mathbb{R}^{\#\mathcal{H}_\ell \times \#\mathcal{H}_k}$ according to the level structure by

$$K_{(\ell,k)} := K|_{\big\{(i,j)\,:\,\operatorname{lev}(i)=\ell,\,\operatorname{lev}(j)=k\big\}}.$$

Thus, we may write

$$K = \begin{pmatrix} K_{(0,0)} & K_{(0,1)} & \cdots & K_{(0,L)} \\ K_{(1,0)} & K_{(1,1)} & \cdots & K_{(1,L)} \\ \vdots & \vdots & \ddots & \vdots \\ K_{(L,0)} & K_{(L,1)} & \cdots & K_{(L,L)} \end{pmatrix}.$$

Lemma 7. *For any $\delta > 0$,*

$$\|\boldsymbol{T}(\omega) - \boldsymbol{T}^\delta(\omega)\|_2 \le CL\delta$$

and, for $0 \le \ell \le k \le L$

$$\|\boldsymbol{T}^\delta(\omega)_{(k,\ell)}\|_2 \le Ch_k. \tag{4.2}$$

Moreover, $\boldsymbol{T}^\delta$ is lower block-triangular with sparse blocks, more precisely,

$$\big(\operatorname{lev}(j) \ge \operatorname{lev}(i) \text{ and } i \ne j\big) \text{ or } d(i,j) > \zeta(1 + |\log(\delta)|) \quad \Longrightarrow \quad \boldsymbol{T}^\delta_{ij} = 0,$$

where $\zeta > 0$ is the bandwidth from Lemma 5. The number of nonzero entries per block is bounded by $\operatorname{nnz}(\boldsymbol{T}^\delta(\omega)_{(k,\ell)}) \lesssim \#\mathcal{H}_k(1 + |\log\delta|)^d$.

Proof. We see immediately $\boldsymbol{T}_{ij}(\omega) = 0$ for all $\operatorname{lev}(j) \ge \operatorname{lev}(i)$ and $i \ne j$ since

$$(\boldsymbol{b}_j, \phi_i)_{L^2(D)} = (\Pi_{\operatorname{lev}(\phi_i)}\boldsymbol{b}_j, \phi_i)_{L^2(D)} = (\phi_j, \phi_i)_{L^2(D)} = 0.$$

Since $\operatorname{supp}(\boldsymbol{b}_i^\delta) \cap \operatorname{supp}(\phi_j) = \emptyset$ as soon as $d(i,j) \gtrsim |\log(\delta)|$, there is some $\zeta > 0$ which depends only on D such that $\boldsymbol{T}(\omega)_{ij} = 0$ for all $d(i,j) > \zeta(1 + |\log(\delta)|)$.

For any vectors $x \in \mathbb{R}^{\#\mathcal{B}_k^\delta}$ and $y \in \mathbb{R}^{\#\mathcal{B}_\ell^\delta}$, we have

$$x \cdot (\boldsymbol{T}_{(k,\ell)} - \boldsymbol{T}^\delta_{(k,\ell)})y = \Bigg(\sum_{\operatorname{lev}(i)=k} \frac{\phi_i}{\|\phi_i\|_{L^2(D)}} x_i, (\boldsymbol{\mathcal{C}}_\ell^\delta - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell \sum_{\operatorname{lev}(j)=\ell} \frac{\phi_j}{\|\boldsymbol{b}_j\|_\omega} y_j\Bigg)_{L^2(D)}.$$

Since $\boldsymbol{T}_{(k,\ell)} \ne 0$ implies $k \ge \ell$, there holds for all $\mathcal{T}_{\ell-1}$-piecewise constants $c_{\ell-1}$

$$\begin{aligned} &|x \cdot (\boldsymbol{T}_{(k,\ell)} - \boldsymbol{T}^\delta_{(k,\ell)})y| \\ &\qquad \lesssim \Big\| \sum_{\operatorname{lev}(i)=k} \frac{\phi_i}{\|\phi_i\|_{L^2(D)}} x_i \Big\|_{L^2(D)} \Big\| (\boldsymbol{\mathcal{C}}_\ell^\delta - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell \sum_{\operatorname{lev}(j)=\ell} \frac{\phi_j}{\|\boldsymbol{b}_j\|_\omega} y_j - c_{\ell-1} \Big\|_{L^2(D)}. \end{aligned}$$

Choosing $c_{\ell-1}$ appropriately, we obtain

$$|x \cdot (\boldsymbol{T}_{(k,\ell)} - \boldsymbol{T}^\delta_{(k,\ell)})y| \lesssim \|x\|_{\ell_2} \left\| (\boldsymbol{\mathcal{C}}^\delta_\ell - \boldsymbol{\mathcal{C}}_\ell)\tilde{\Pi}_\ell \sum_{\mathrm{lev}(j)=\ell} \frac{\phi_j}{|||\boldsymbol{b}_j|||_\omega} y_j \right\|_{H^1(D)}.$$

As in the proof of Lemma 5, we see $\|\boldsymbol{T}(\omega)_{(\ell,k)} - \boldsymbol{T}^\delta(\omega)_{(\ell,k)}\|_2 \lesssim \delta$. Summing up over the levels proves $\|\boldsymbol{T}(\omega) - \boldsymbol{T}^\delta(\omega)\|_2 \lesssim L\delta$.

To see (4.2), note that $w := \sum_{\phi_i \in \mathcal{H}_k} \alpha_i \phi_i$ and $b := \sum_{\boldsymbol{b}_j \in \mathcal{B}_\ell} \beta_j \boldsymbol{b}^\delta_j$ satisfy

$$\begin{aligned} \alpha^T \boldsymbol{T}^\delta(\omega)\beta &= (w, b)_{L^2(D)} = ((1-\Pi_k)w, b)_{L^2(D)} = (w, (1-\Pi_k)b)_{L^2(D)} \\ &\lesssim h_k \|w\|_{L^2(D)} |||b|||_\omega \lesssim h_k \|\alpha\|_{\ell_2} \|\beta\|_{\ell_2} \end{aligned}$$

by Lemma 6. This concludes the proof. □

5. Inverse stiffness matrices and averaging

This section proves that the inverse of the stiffness matrix $\boldsymbol{S}^\delta(\omega)$ (w.r.t. the coefficient adapted bases $\boldsymbol{\mathcal{B}}^\delta(\omega)$) defined in the previous section can be efficiently approximated by a sparse matrix. One possibility to compute an approximate inverse of the matrix $\boldsymbol{S}^\delta(\omega)$ is to apply the conjugate gradient method (CG) to the matrix with unit vectors $e_i \in \mathbb{R}^N$ as right-hand sides. The sparsity structure from Lemma 5 shows that one matrix-vector product with $\boldsymbol{S}^\delta e_i$ increases the number of non-zero entries to $\#\{1 \leq j \leq N \,:\, d(i,j) \lesssim 1 + |\log(\delta)|\}$. Thus, after $k \in \mathbb{N}$ iterations of the CG method, the resulting vector has about $\#\{1 \leq j \leq N \,:\, d(i,j) \lesssim k(1+|\log(\delta)|)\}$ non-zero entries. Since the condition number $\kappa(\boldsymbol{S}^\delta)$ is uniformly bounded, the number of iterations grows only logarithmically in the desired accuracy. Thus, the cost of $k \in \mathbb{N}$ iterations of the CG method can be bounded roughly by $(1 + |\log(\delta)|))^2$.

Lemma 8. *Given $\delta > 0$, there exists a matrix $\boldsymbol{R}^\delta(\omega)$ such that $\|\boldsymbol{S}(\omega)^{-1} - \boldsymbol{R}^\delta(\omega)\|_2 \leq \delta$. Moreover, $\boldsymbol{R}^\delta(\omega)$ satisfies*

$$d(i,j) > C_{\mathrm{inv}}\zeta(|\log(\delta)|^2 + 1) \;\; or \;\; \mathrm{lev}(i) \neq \mathrm{lev}(j) \quad \Longrightarrow \quad \boldsymbol{R}^\delta_{ij}(\omega) = 0, \tag{5.1}$$

for some uniform constant $C_{\mathrm{inv}} > 0$ and ζ from Lemma 5. The number of non zero entries is bounded by $\mathrm{nnz}(\boldsymbol{R}^\delta) \lesssim N(1 + |\log(\delta)|)^d$.

Proof. Due to Lemma 5 and the fact that $\boldsymbol{\mathcal{B}}(\omega)$ is a Riesz basis (Lemma 6), we observe that all eigenvalues of $\boldsymbol{S}^\delta(\omega)$ are of order $\mathcal{O}(1)$ as long as $\delta \lesssim 1$. Therefore, we can obtain $\boldsymbol{R}^\delta(\omega)$ by application of CG steps to $\boldsymbol{S}^{\tilde{\delta}}(\omega)$ (we chose $\tilde{\delta} > 0$ later, see, e.g., [33, Chapter 6]). The convergence properties of CG show

$$\|\boldsymbol{S}^{\tilde{\delta}}(\omega)^{-1} - \boldsymbol{R}^\delta(\omega)\|_2 \leq \delta$$

if we perform $k = \mathcal{O}(|\log(\delta)| + 1)$ CG-steps. This follows since

$$\|\mathrm{res}_k\|_{\ell_2} \simeq \sqrt{\boldsymbol{S}^{\tilde{\delta}}(\omega)\mathrm{res}_k \cdot \mathrm{res}_k}$$

for the residual res_k of the CG method. From Lemma 5, we see that $\boldsymbol{R}^\delta(\omega)$ satisfies

$$d(i,j) > \zeta(|\log(\delta)| + 1)^2 \text{ or } \mathrm{lev}(i) \neq \mathrm{lev}(j) \quad \Longrightarrow \quad \boldsymbol{R}^\delta_{ij}(\omega) = 0,$$

since each CG-step increases the bandwidth by the original bandwidth. With Lemma 5, we conclude the proof by choosing $k \simeq 1 + |\log(\delta)|$ and $\tilde{\delta} \simeq \delta$. □

Lemma 9. *We define a discrete approximation to* $\mathcal{A}^{-1}$ *by*

$$R := \mathbb{E}\big[\big(\boldsymbol{T}^{-T}(\omega)\boldsymbol{S}(\omega)\boldsymbol{T}^{-1}(\omega)\big)^{-1}\big] = \mathbb{E}\big[\boldsymbol{T}(\omega)\boldsymbol{S}(\omega)^{-1}\boldsymbol{T}(\omega)^T\big].$$

We define a perturbed and truncated version of R *by* $R^\delta \in \mathbb{R}^{N\times N}$

$$(R^\delta)_{(\ell,k)} := \begin{cases} \Big(\mathbb{E}\big[\boldsymbol{T}^\delta(\omega)\boldsymbol{R}^\delta(\omega)\boldsymbol{T}^\delta(\omega)^T\big]\Big)_{(\ell,k)} & \ell+k \leq |\log(\delta)|, \\ 0 & \textit{else.} \end{cases} \tag{5.2}$$

which satisfies $\|R - R^\delta\|_2 \leq CL^2\delta$. *The number of non-zero entries in* R^δ *is bounded by* $\mathrm{nnz}(R^\delta) \lesssim L/\delta^d$.

Proof. We define the auxiliary operator

$$\widetilde{R}^\delta := \mathbb{E}\big[\boldsymbol{T}^\delta(\omega)\boldsymbol{R}^\delta(\omega)\boldsymbol{T}^\delta(\omega)^T\big].$$

Following the proofs of Lemma 5, Lemma 7, and Lemma 8, we show

$$\|(R-\widetilde{R}^\delta)x\|_{\ell_2}^2 \lesssim \sum_{\ell=1}^L \Big\| \sum_{k=1}^L (R_{(\ell,k)} - \widetilde{R}^\delta_{(\ell,k)})x_{(k)}\Big\|_{\ell_2}^2 \lesssim \delta^2 L^4 \sum_{\ell=1}^L\sum_{k=1}^L \|x_{(k)}\|_{\ell_2}^2 = \delta^2 L^4 \|x\|_{\ell_2}^2$$

and hence $\|R - \widetilde{R}^\delta\|_2 \lesssim \delta L^2$. The estimate (4.2) implies for $\ell + k > |\log(\delta)|$

$$\begin{aligned} \|(\widetilde{R}^\delta - R^\delta)_{(\ell,k)}\|_2 &\leq \sum_{j=0}^L \|\boldsymbol{T}^\delta(\omega)_{(\ell,j)}\|_2 \|\boldsymbol{R}^\delta(\omega)_{(j,j)}\|_2 \|(\boldsymbol{T}^\delta(\omega)_{(k,j)})\|_2 \\ &\lesssim \sum_{j=0}^L h_\ell(1+\delta)h_k \\ &\lesssim L2^{-\ell-k}. \end{aligned}$$

This implies for $x \in \mathbb{R}^N$

$$\begin{aligned} \|(\widetilde{R}^\delta - R^\delta)x\|_{\ell_2}^2 &\leq \sum_{i,j=0}^L \|(\widetilde{R}^\delta - R^\delta)_{(i,j)} x|_{(j)}\|_{\ell_2}^2 \lesssim \sum_{j=0}^L \|x|_{(j)}\|_{\ell_2}^2 \sum_{i=|\log(\delta)|-j}^L L2^{-i-j} \\ &\lesssim L\delta\|x\|_{\ell_2}^2. \end{aligned}$$

The number of non-zero entries in R^δ can be bounded by

$$\sum_{0\leq i+j\leq |\log(\delta)|} \#(R^\delta)_{(i,j)} \leq \sum_{0\leq i+j\leq|\log(\delta)|} 2^{d(i+j)} \lesssim L\delta^{-d}.$$

This concludes the proof. ☐

To formulate the following main theorem, we identify the matrix R^δ with an operator $\mathcal{R}^\delta\colon L^2(D) \to L^2(D)$ via the natural embedding $\iota\colon \mathbb{R}^N \to \mathrm{span}(\mathcal{H})$, $\iota(\alpha) = \sum_{i=1}^N \alpha_i\phi_i \in L^2(D)$. There holds $\mathcal{R}^\delta := \iota R^\delta \iota^\star$.

Theorem 10. *Given an accuracy* $\delta > 0$, *there exists a finite dimensional operator* $\mathcal{R}^\delta\colon L^2(D) \to L^2(D)$ *which depends only on* δ *such that*

$$\|\mathcal{A}^{-1} - \mathcal{R}^\delta\|_{\mathcal{L}(L^2(D),L^2(D))} \leq \delta.$$

The corresponding operator matrix R^δ *has at most* $\mathcal{O}(|\log(\delta)|^{2d+1}\delta^{-d})$ *non-zero entries.*

Constructive proof. We use the operator matrix $R^\delta \in \mathbb{R}^{N\times N}$ from Theorem 9. Given $f \in L^2(D)$, define $\boldsymbol{F}(\omega) \in \mathbb{R}^N$ by $\boldsymbol{F}_i(\omega) := (f, \boldsymbol{b}_i/|||\boldsymbol{b}_i|||_\omega)$. By definition, there holds $\boldsymbol{S}(\omega)\boldsymbol{\alpha}(\omega) = \boldsymbol{F}(\omega)$ with $\boldsymbol{u}_L(\omega) := \sum_{i=1}^N \boldsymbol{\alpha}_i(\omega)\boldsymbol{b}_i/|||\boldsymbol{b}_i|||_\omega \in \operatorname{span}(\boldsymbol{\mathcal{B}}(\omega))$ being the Galerkin approximation to $\boldsymbol{u}(\omega) \in H_0^1(\Omega)$. Hence, we have

$$\|\boldsymbol{u}(\omega) - \Pi_L \boldsymbol{u}_L(\omega)\|_{L^2(D)} \le \|(1-\Pi_L)\boldsymbol{u}(\omega)\|_{L^2(D)} + \|\boldsymbol{u}(\omega) - \boldsymbol{u}_L(\omega)\|_{L^2(D)}.$$

Since $\Pi_L(\boldsymbol{u}(\omega) - \boldsymbol{u}_L(\omega)) = 0$, we have that

$$|||\boldsymbol{u}(\omega) - \boldsymbol{u}_L(\omega)|||_\omega \lesssim h_L\|(1-\Pi_L)f\|_{L^2(D)} \le h_L\|f\|_{L^2(D)}.$$

Therefore, we obtain

$$\|\boldsymbol{u}(\omega) - \Pi_L\boldsymbol{u}_L(\omega)\|_{L^2(D)} \lesssim h_L\|f\|_{L^2(D)}.$$

With the transfer matrices $\boldsymbol{T}(\omega)$ from Lemma 7, we obtain

$$\widetilde{F} := \iota^\star f = \boldsymbol{T}^{-T}(\omega)\boldsymbol{F}(\omega)$$

and hence $\boldsymbol{\beta} \in \mathbb{R}^N$ with $\boldsymbol{T}^{-T}(\omega)\boldsymbol{S}(\omega)\boldsymbol{T}^{-1}(\omega)\boldsymbol{\beta}(\omega) = \widetilde{F}$ satisfies $\boldsymbol{T}(\omega)\boldsymbol{\alpha}(\omega) = \boldsymbol{\beta}(\omega)$. Together with Lemma 7, this shows that $\Pi_L\boldsymbol{u}_L(\omega) = \sum_{i=1}^N \boldsymbol{\beta}_i(\omega)\phi_i/\|\phi_i\|_{L^2(D)}$. The approximate solution $\mathcal{R}^\delta f = \sum_{i=1}^N \gamma_i\phi_i/\|\phi_i\|_{L^2(D)}$ with $\gamma := R^\delta\widetilde{F}$ satisfies

$$|\gamma - \mathbb{E}[\boldsymbol{\beta}]| \lesssim L^2\delta\|f\|_{L^2(D)},$$

by use of Theorem 9 and since $\mathbb{E}_M[\boldsymbol{\beta}] = R\widetilde{F}$. Since $\mathcal{H}$ is an orthogonal basis, we obtain immediately $\|\mathcal{R}^\delta f - \mathcal{R}f\|_{L^2(D)} \lesssim L^2\delta\|f\|_{L^2(D)}$, where

$$\mathcal{R}^\delta f = \mathbb{E}[\boldsymbol{u}_L].$$

Combining the above error bounds, we conclude

$$\|\mathbb{E}[\boldsymbol{u}] - \mathcal{R}^\delta f\|_{L^2(D)} \lesssim (L^2\delta + h_L)\|f\|_{L^2(D)}.$$

With $L \simeq |\log\delta|$ and $h_L \simeq \delta$ there holds $\|\mathbb{E}[\boldsymbol{u}] - \mathcal{R}^\delta f\|_{L^2(D)} \lesssim (1+|\log(\delta)|^2)\delta\|f\|_{L^2(D)}$. Replacing δ with δ/L^2, we conclude the proof. ☐

6. Sparse operator compression

Theorem 10 shows that the expected operator can indeed be compressed to a sparse matrix. The constructive proof motivates a compression algorithm by simply replacing the expectation by a suitable sample mean. For this purpose, let $\Omega_M \subset \Omega$ be a finite set of sampling points with $|\Omega_M| = M \in \mathbb{N}$ and define the sample mean $\mathbb{E}_M[\mathbf{X}] := M^{-1}\sum_{\omega\in\Omega_M}\mathbf{X}(\omega)$ for a random field $\mathbf{X}$. It is readily seen that Lemma 9 remains valid when $\mathbb{E}$ is replaced by $\mathbb{E}_M$. More precisely, define

$$R_M := \mathbb{E}_M\big[\big(\boldsymbol{T}^{-T}(\omega)\boldsymbol{S}(\omega)\boldsymbol{T}^{-1}(\omega)\big)^{-1}\big] = \mathbb{E}_M\big[\boldsymbol{T}(\omega)\boldsymbol{S}(\omega)^{-1}\boldsymbol{T}(\omega)^T\big]$$

and a perturbed and truncated version of R_M by $R_M^\delta \in \mathbb{R}^{N\times N}$

$$(R_M^\delta)_{(\ell,k)} := \begin{cases} \Big(\mathbb{E}_M\big[\boldsymbol{T}^\delta(\omega)\boldsymbol{R}^\delta(\omega)\boldsymbol{T}^\delta(\omega)^T\big]\Big)_{(\ell,k)} & \ell + k \le |\log(\delta)|, \\ 0 & \text{else.} \end{cases} \tag{6.1}$$

Then

$$\|R_M - R_M^\delta\|_2 \le CL^2\delta \tag{6.2}$$

and the number of non-zero entries in R_M^δ is bounded by $\mathcal{O}(L/\delta^d)$.

Remark 11. *The truncation condition $\ell + k \leq |\log(\delta)|$ in (6.1) can be relaxed to $\ell + k \leq C|\log(\delta)|$ for some $C \simeq 1$ without any harm. In practice, when $L \simeq |\log\delta|$ is chosen, a natural choice would be $\ell + k \leq L$. In the numerical experiment of Section 7 we will see that sometimes it can be advantageous to include also few more blocks of the lower right part of the matrix (see Eq. (7.1)).*

The analog of Theorem (10) in this discrete stochastic setting then reads.

Corollary 12. *Given an accuracy $\delta > 0$ and a set of M samples $\Omega_M \subset \Omega$, $M \in \mathbb{N}$, there exists a finite dimensional operator $\mathcal{R}_M^\delta \colon L^2(D) \to L^2(D)$ which depends only on the sample coefficients $\boldsymbol{A}(\omega)$, $\omega \in \Omega_M$, δ, and D, such that*

$$\|\mathcal{A}^{-1} - \mathcal{R}_M^\delta\|_{\mathcal{L}(L^2(D),L^2(D))} \leq \delta + \|(\mathbb{E} - \mathbb{E}_M)[\boldsymbol{\mathcal{A}}^{-1}]\|_{\mathcal{L}(L^2(D),L^2(D))}.$$

The corresponding operator matrix R_M^δ has $\mathcal{O}(|\log(\delta)|^{2d+1}\delta^{-d})$ non-zero entries.

When using a plain Monte Carlo sampling the mean squared sampling error scales like M^{-1} meaning that $M \simeq \delta^{-2}$ samples suffice to ensure that the sampling error is not dominating the error bound. This is optimal in the present setting with no assumptions on the distribution of the random diffusion coefficient. More advanced sampling techniques such as quasi Monte Carlo methods are certainly possible under additional assumptions such as a rapid decay of eigenvalues of a given Karhunen-Loève expansion of the random parameter (see [6] for a discussion in terms of PDEs with random parameters). Even more promising is in the intertwining of the hierarchical decomposition and the sampling procedure in the spirit of multilevel/multi-index Monte Carlo (see, e.g., [14, 17] for the seminal works as well as [5]). At least in the regime where stochastic homogenization applies, the computation of basis functions is likely to be essentially independent of the parameter ω for levels that are much coarser than the characteristic length scale of random oscillation (or correlation) [13]. The increasing variance for the levels approaching the scale of correlation, stationarity could be exploited to improve the overall complexity.

Another interesting case is the use of log-normal coefficients $\boldsymbol{A}(\omega) = \exp(\boldsymbol{Z}(\omega))$ for a normal random field $\boldsymbol{Z}$. As shown in [10], such random fields can be efficiently generated for general covariance functions and non-uniform grids. The present analysis, however, breaks down since the assumption of bounded contrast in (2) is violated. The authors are confident, however, that the arguments can be modified in the sense that the extreme contrast samples will only appear with very low probability (the tails of the Gaussian density). Thus, a polynomial dependence on the contrast (as is observed for the present construction) will not perturb the final result.

We shall finally mention that so far the construction relies on the exact solution of the (infinite-dimensional) corrector problems (3.8) and their preconditioned variant, respectively. The elegant way to transfer all results to a fully discrete setting is to consider space-discrete problem from the very beginning. It is readily seen that all constructions and results remain valid if we replace the space $V = H_0^1(D)$ by a suitable finite dimensional subspace $V_h \subset V$ throughout the paper. We have in mind some standard V-conforming finite element space V_h that is based on some regular mesh of width h which turns the preconditioned corrector problems into finite element problems on the mesh h restricted to local subdomains of diameter $h_\ell|\log\delta|$. The only restriction that comes with this discretization step is that the mesh size h limits the number of possible levels L in the hierarchical decomposition and, hence, the possible accuracy $\delta \lesssim h$ when the sparse approximation is compared with the reference solution $\mathbb{E}[\boldsymbol{u}_h]$ where $\boldsymbol{u}_h$ solves (2.1) with V replaced with V_h. Clearly, the overall accuracy of the fully discrete method depends on the error $\|\mathbb{E}[\boldsymbol{u} - \boldsymbol{u}_h]\|_{L^2(D)}$ which is a standard finite element error that depends on the spatial regularity of $\boldsymbol{A}$ and also its possible frequencies of oscillations. All this is well understood and implies the usual conditions on the smallness of h so that $\boldsymbol{A}$ is properly resolved (see e.g. [32]).

7. Numerical experiment

This section presents some simple numerical experiments to illustrate the performance of the method. We consider the domain $D = [0,1]^d$ for $d = 1,2$ and the coefficient $\boldsymbol{A}$ is scalar i.i.d. and, on each cell of the uniform Cartesian mesh $\mathcal{T}_\varepsilon$, it is uniformly distributed in the interval $[\gamma_{\min}, \gamma_{\max}] = [0.5, 10]$. The mesh width (scale of oscillation/correlation length) $\varepsilon = 2^{-8}$ $(d = 1)$ and $\varepsilon = 2^{-5}$ $(d = 2)$.

The approximations of the solution operator are based on sequences of uniform Cartesian meshes $\mathcal{T}_\ell$ $(\ell = 0, 1, 2, \ldots, L)$ of mesh width $h_\ell = 2^{-\ell}$ that do not necessarily resolve ε. We compute approximations $\mathcal{R}^L = \mathcal{R}^\delta_{M_L}$ of the expected solution operator depending on the maximal level L which means that we expect errors of order $\delta \approx 2^{-L}$. The truncation of blocks is performed based on the criterion $k + \ell \leq L$ as indicated in Remark 11. For the solution of the corrector problems and the reference solution $\mathbf{u}_h$ we use d-linear finite elements on the mesh $\mathcal{T}_h$ where $h = 2^{-14}$ $(d = 1)$ and $h = 2^{-9}$ $(d = 2)$. To achieve accuracy of order δ (w.r.t. to the reference solution) we perform $\lceil L/2 \rceil$ CG-iterations for both computing the correctors $\boldsymbol{\mathcal{C}}^\delta(\omega)$ and inverting the block-diagonal stiffness matrices $\boldsymbol{S}^\delta(\omega)$. For the approximation of the expected values we use a quasi-Monte Carlo method (particularly a Sobol sequence) with appropriate numbers of sampling points $M_h := h^{-1}$ for the reference solutions and $M_L := 2^L$ for the approximations. While we did not show that the problem is smooth enough to justify the use of quasi-Monte Carlo sampling, we still observe the expected higher convergence rate compared to plain Monte Carlo sampling and thus save significant compute time.

Since the computation of a reference expected operator is hardly feasible we only compute the error for one non-smooth deterministic right-hand side $f = \chi_{[.5,1]\times[0,1]^{d-1}} \in L^2(D) \setminus H^1(D)$. Figures 1-2 (left plots) depict the errors $\|\mathbb{E}_{M_h}[\boldsymbol{u}_h] - \mathcal{R}^L f\|_{L^2(D)}$ versus the number of nonzero entries of $\mathrm{nnz}(R^L)$ for $L = 1, 2, \ldots$. The results are very well in agreement (up to a, possibly pessimistic, logarithmic factor) with the prediction that

$$\|\mathbb{E}_{M_h}[\boldsymbol{u}_h] - \mathcal{R}^L f\|_{L^2(D)} \lesssim M_L^{-1} + \frac{|\log(\mathrm{nnz}(R^L))^{2+1/d}}{\mathrm{nnz}(R^L)^{1/d}}$$

for $d = 1, 2$. This is the optimal rate of convergence (up to a logarithmic factor) given a piecewise constant approximation.

In this setting where the expected solution $\mathbb{E}[\boldsymbol{u}]$ is even $H^2(D)$ regular it would be desirable to recover gradient information from the piecewise constant approximation by suitable postprocessing e.g. in the hierarchical basis associated with a constant coefficient. Figures 1–2 (left plots) indicate that this is not automatically achieved for non-smooth right-hand sides with the present choice of parameters. However, when the truncation in (6.1) is slightly relaxed in the following form

$$(\tilde{R}^L)_{(\ell,k)} := \begin{cases} \left(\mathbb{E}_M\left[\boldsymbol{T}^\delta(\omega)\boldsymbol{R}^\delta(\omega)\boldsymbol{T}^\delta(\omega)^T\right]\right)_{(\ell,k)} & \ell + k \leq L + \max(1, \lceil \log_2 L \rceil), \\ 0 & \text{else,} \end{cases} \tag{7.1}$$

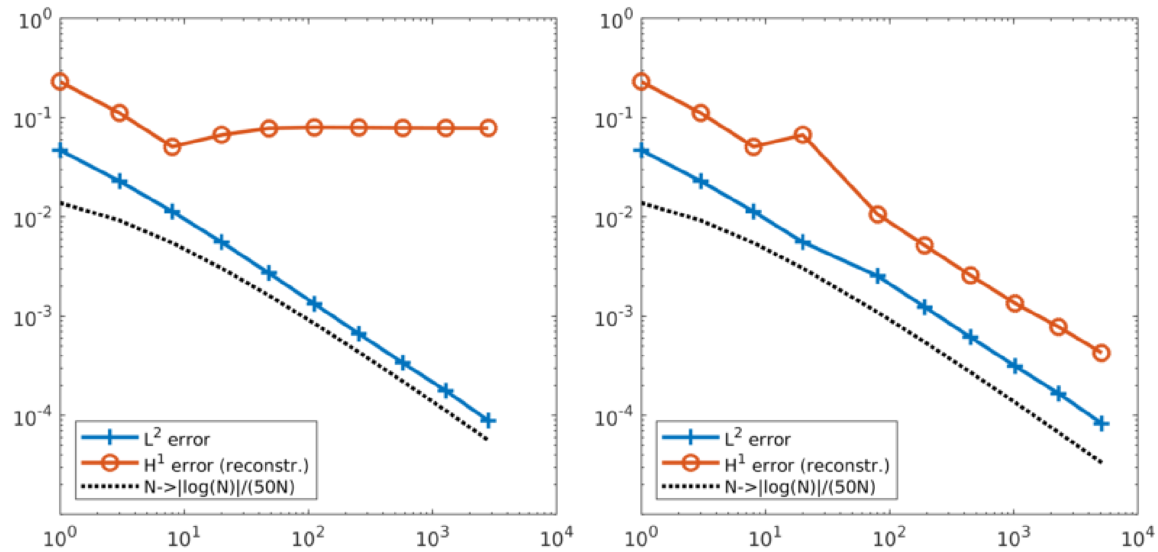


FIGURE 1. Numerical results in $1d$: $L^2(D)$-errors $\|\mathbb{E}_{M_h}[\boldsymbol{u}_h]-\mathcal{R}^L f\|_{L^2(D)}$ and $H^1(D)$-errors $\|\nabla(\mathbb{E}_{M_h}[\boldsymbol{u}_h]-u_L^1)\|_{L^2(D)}$ of post-processed approximation for $L=1,2,\ldots,10$. Left: Errors versus $\mathrm{nnz}(R^L)$ using original approach (6.1). Right: Errors versus $\mathrm{nnz}(\tilde{R}^L)$ using modified approach (7.1).

accurate reconstruction of gradients seems possible. From this slightly more accurate but slightly more dense approximation $\tilde{R}^L$ we can reconstruct the coefficients of a smooth approximation $u_L^1 \in \operatorname{span} \boldsymbol{\mathcal{B}}(\omega_\Delta)$ (with $\omega_\Delta \in \Omega$ such that $\boldsymbol{A}(\omega_\Delta)=1$) in the hierarchical basis that corresponds to the Laplacian by simply applying $T^\delta(\omega_\Delta)^{-1}$ to $\tilde{R}^L f$. The errors of this smooth postprocessing $\|\nabla(\mathbb{E}_{M_h}[\boldsymbol{u}_h]-u_L^1)\|_{L^2(D)}$ are plotted in Figure 1–2 (right plots) against the number of non-zero entries $\mathrm{nnz}(\tilde{R}^L)$. The observed rate of convergence for the H^1-error is $\mathrm{nnz}(\tilde{R}^L)^{-1/d}$ (up to a logarithmic factor) which is nearly optimal. See also the plots on the

These first numerical results support the theoretical findings and indicate the potential of the approach. Since the techniques that were used in the construction of the method and its analysis, in particular the localized orthogonal decomposition, generalize in a straight-forward way to other classes of operators such as linear elasticity [18] or Helmholtz problems [31, 11, 4], we believe that the sparse compression algorithm for the approximation of expected solution operators is applicable beyond the prototypical model problem of this paper.

References

[1] S. Armstrong, T. Kuusi, and J.-C. Mourrat. The additive structure of elliptic homogenization. *Inventiones mathematicae*, 208(3):999–1154, Jun 2017.

[2] J. Bourgain. On a homogenization problem. *Journal of Statistical Physics*, 2018.

[3] A. Bourgeat and A. Piatnitski. Approximations of effective coefficients in stochastic homogenization. *Ann. Inst. H. Poincaré Probab. Statist.*, 40(2):153–165, 2004.

[4] D. L. Brown, D. Gallistl, and D. Peterseim. Multiscale Petrov-Galerkin method for high-frequency heterogeneous Helmholtz equations. In *Meshfree methods for partial differential equations VIII*, volume 115 of *Lect. Notes Comput. Sci. Eng.*, pages 85–115. Springer, Cham, 2017.

[5] J. Dick, M. Feischl, and Ch. Schwab. Improved efficiency of a multi-index FEM. *arXiv preprint 1806.04159*, 2018.

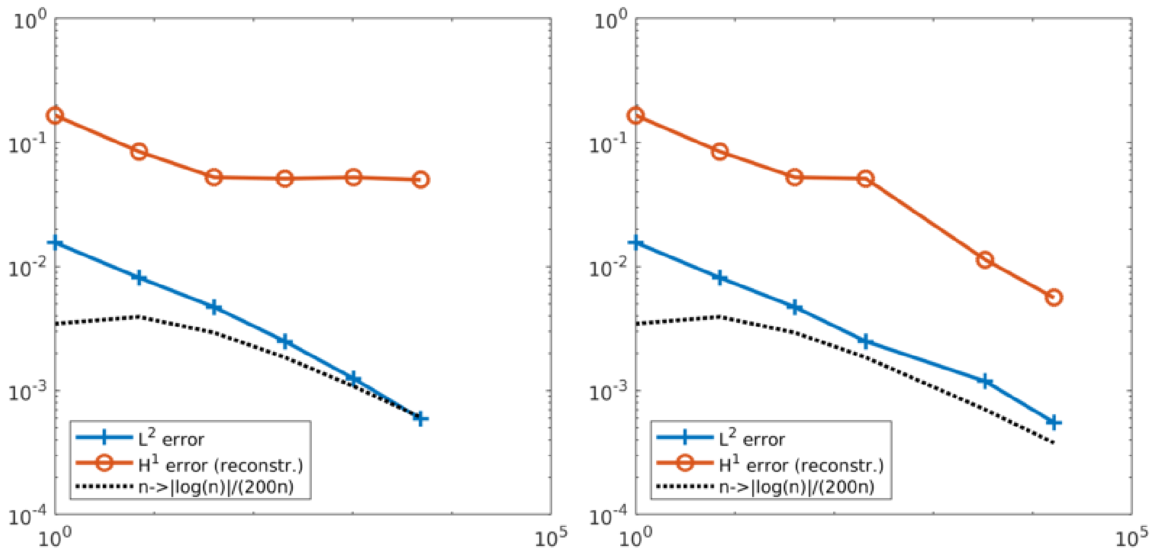


FIGURE 2. Numerical results in $2d$: $L^2(D)$-errors $\|\mathbb{E}_{M_h}[\boldsymbol{u}_h] - \mathcal{R}^L f\|_{L^2(D)}$ and $H^1(D)$-errors $\|\nabla(\mathbb{E}_{M_h}[\boldsymbol{u}_h] - u_L^1)\|_{L^2(D)}$ of post-processed approximation for $L = 1, 2, \ldots, 6$. Left: Errors versus $\mathrm{nnz}(R^L)$ using original approach (6.1). Right: Errors versus $\mathrm{nnz}(\tilde{R}^L)$ using modified approach (7.1).

[6] J. Dick, F. Y. Kuo, Q. T. Le Gia, D. Nuyens, and C. Schwab. Higher order QMC Petrov-Galerkin discretization for affine parametric operator equations with random field inputs. *SIAM J. Numer. Anal.*, 52(6):2676–2702, 2014.
[7] M. Duerinckx, A. Gloria, and F. Otto. The structure of fluctuations in stochastic homogenization. *arXiv e-prints*, 1602.01717 [math.AP], 2016.
[8] D. Elfverson, E. Georgoulis, and A. Målqvist. An adaptive discontinuous galerkin multiscale method for elliptic problems. *Multiscale Modeling & Simulation*, 11(3):747–765, 2013.
[9] D. Elfverson, E. H. Georgoulis, A. Målqvist, and D. Peterseim. Convergence of a discontinuous galerkin multiscale method. *SIAM J. Numer. Anal.*, 51(6):3351–3372, 2013.
[10] M. Feischl, F. Kuo, and I.H. Sloan. Fast random field generation with H-matrices. *Numer. Math. (to appear)*, 2018.
[11] D. Gallistl and D. Peterseim. Stable multiscale Petrov-Galerkin finite element method for high frequency acoustic scattering. *Comput. Methods Appl. Mech. Eng.*, 295:1–17, 2015.
[12] D. Gallistl and D. Peterseim. Computation of quasilocal effective diffusion tensors and connections to the mathematical theory of homogenization. *Multiscale Model. Simul.*, 15(4), 2017.
[13] D. Gallistl and D. Peterseim. Numerical stochastic homogenization by quasilocal effective diffusion tensors. *ArXiv e-prints*, 1702.08858, 2017.
[14] M. B. Giles. Multilevel Monte Carlo path simulation. *Oper. Res.*, 56(3):607–617, 2008.
[15] A. Gloria and F. Otto. The corrector in stochastic homogenization: optimal rates, stochastic integrability, and fluctuations. *arXiv e-prints*, 1510.08290 [math.AP], 2015.
[16] A. Gloria and F. Otto. Quantitative results on the corrector equation in stochastic homogenization. *J. Eur. Math. Soc. (JEMS)*, 19(11):3489–3548, 2017.
[17] Abdul-Lateef Haji-Ali, Fabio Nobile, and Raúl Tempone. Multi-index Monte Carlo: when sparsity meets sampling. *Numer. Math.*, 132(4):767–806, 2016.
[18] P. Henning and A. Persson. A multiscale method for linear elasticity reducing Poisson locking. *Comput. Methods Appl. Mech. Engrg.*, 310:156–171, 2016.
[19] P. Henning and D. Peterseim. Oversampling for the multiscale finite element method. *Multiscale Model. Simul.*, 11(4):1149–1175, 2013.
[20] T. Y. Hou, D. Huang, K. C. Lam, and P. Zhang. An Adaptive Fast Solver for a General Class of Positive Definite Matrices Via Energy Decomposition. *Multiscale Model. Simul.*, 16(2):615–678, 2018.

[21] T. Y. Hou and P. Zhang. Sparse operator compression of higher-order elliptic operators with rough coefficients. *Res. Math. Sci.*, 4:Paper No. 24, 49, 2017.
[22] J. Kim and M. Lemm. On the averaged Green's function of an elliptic equation with random coefficients. *ArXiv e-prints*, April 2018.
[23] R. Kornhuber, D. Peterseim, and H. Yserentant. An analysis of a class of variational multiscale methods based on subspace decomposition. *Math. Comp.*, published electronically, 2018.
[24] R. Kornhuber and H. Yserentant. Numerical homogenization of elliptic multiscale problems by subspace decomposition. *Multiscale Model. Simul.*, 14(3):1017–1036, 2016.
[25] S. M. Kozlov. The averaging of random operators. *Mat. Sb. (N.S.)*, 109(151)(2):188–202, 327, 1979.
[26] A. Målqvist and D. Peterseim. Localization of elliptic multiscale problems. *Math. Comp.*, 83(290):2583–2603, 2014.
[27] H. Owhadi. Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. *SIAM Review*, 59(1):99–149, 2017.
[28] H. Owhadi and L. Zhang. Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic odes/pdes with rough coefficients. *Journal of Computational Physics*, 347:99 – 128, 2017.
[29] G. C. Papanicolaou and S. R. S. Varadhan. Boundary value problems with rapidly oscillating random coefficients. In *Random fields, Vol. I, II (Esztergom, 1979)*, volume 27 of *Colloq. Math. Soc. János Bolyai*, pages 835–873. North-Holland, Amsterdam-New York, 1981.
[30] D. Peterseim. Variational multiscale stabilization and the exponential decay of fine-scale correctors. In Gabriel R. Barrenechea, Franco Brezzi, Andrea Cangiani, and Emmanuil H. Georgoulis, editors, *Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations*, pages 343–369. Springer International Publishing, Cham, 2016.
[31] D. Peterseim. Eliminating the pollution effect in Helmholtz problems by local subscale correction. *Math. Comp.*, 86(305):1005–1036, 2017.
[32] D. Peterseim and S. Sauter. Finite elements for elliptic problems with highly varying, nonperiodic diffusion matrix. *Multiscale Modeling & Simulation*, 10(3):665–695, 2012.
[33] Y. Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, second edition, 2003.
[34] F. Schäfer, T. J. Sullivan, and H. Owhadi. Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *ArXiv e-prints*, June 2017.
[35] P. Wojtaszczyk. *A mathematical introduction to wavelets*, volume 37 of *London Mathematical Society Student Texts*. Cambridge University Press, Cambridge, 1997.
[36] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Rev.*, 34(4):581–613, 1992.
[37] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numer.*, 2:285–326, 1993.
[38] V. V. Yurinskiĭ. Averaging of symmetric diffusion in a random medium. *Sibirsk. Mat. Zh.*, 27(4):167–180, 215, 1986.

Karlsruher Institut für Technologie
E-mail address: michael.feischl@kit.edu

Universität Augsburg
E-mail address: daniel.peterseim@math.uni-augsburg.de