Hou and Zhang *Res Math Sci(2017)4:24*
DOI 10.1186/s40687-017-0113-1

Research in the Mathematical Sciences

RESEARCH Open Access

Sparse operator compression of higher-order elliptic operators with rough coefficients

Thomas Y. Hou and Pengchuan Zhang[*]

[*]Correspondence:
pzzhang@cms.caltech.edu
Applied and Computational Mathematics, Caltech, Pasadena, CA 91125, USA

Abstract

We introduce the sparse operator compression to compress a self-adjoint higher-order elliptic operator with rough coefficients and various boundary conditions. The operator compression is achieved by using localized basis functions, which are energy minimizing functions on local patches. On a regular mesh with mesh size h, the localized basis functions have supports of diameter $O(h \log(1/h))$ and give optimal compression rate of the solution operator. We show that by using localized basis functions with supports of diameter $O(h \log(1/h))$, our method achieves the optimal compression rate of the solution operator. From the perspective of the generalized finite element method to solve elliptic equations, the localized basis functions have the optimal convergence rate $O(h^k)$ for a $(2k)$th-order elliptic problem in the energy norm. From the perspective of the sparse PCA, our results show that a large set of Matérn covariance functions can be approximated by a rank-n operator with a localized basis and with the optimal accuracy.

1 Background

1.1 Main objectives and the problem setting

The main purpose of this paper is to develop a general strategy to compress a class of self-adjoint higher-order elliptic operators by localized basis functions that give optimal approximation property of the solution operator. To be more specific, suppose $\mathcal{L}$ is a self-adjoint elliptic operator in the divergence form

$$\mathcal{L}u = \sum_{0\le|\sigma|,|\gamma|\le k} (-1)^{|\sigma|} D^{\sigma}(a_{\sigma\gamma}(x) D^{\gamma} u), \tag{1.1}$$

where the coefficients $a_{\sigma\gamma} \in L^{\infty}(D)$, D is a bounded domain in $\mathbb{R}^d$, $\sigma = (\sigma_1, \ldots, \sigma_d)$ is a d-dimensional multiindex. We ask the question: given an integer n, what is the best rank-n compression of the operator $\mathcal{L}$ with localized basis functions? This question arises in many different contexts.

Consider the elliptic equation with the homogeneous Dirichlet boundary conditions

$$\mathcal{L}u = f, \quad u \in H_0^k(D), \tag{1.2}$$

Springer Open

where the load $f \in L^2(D)$. For a self-adjoint, positive definite operator $\mathcal{L}$, Eq. (1.2) has a unique weak solution, denoted as $\mathcal{L}^{-1}f$. We define the operator compression error of the basis Ψ as follows:

$$E_{\mathrm{oc}}(\Psi; \mathcal{L}^{-1}) := \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{L}^{-1} - \Psi K_n \Psi^T\|_2, \tag{1.3}$$

which is the optimal approximation error of $\mathcal{L}^{-1}$ among all positive semidefinite operators with range space spanned by Ψ. Using $E_{\mathrm{oc}}(\Psi; (\mathcal{L}+\lambda_G)^{-1})$ for some $\lambda_G > 0$ to quantify the compression error is useful for operators that are not invertible, such as $-\Delta$ with periodic boundary conditions.

Without imposing the sparsity constraints on the basis Ψ, the compression error $E_{\mathrm{oc}}(\Psi; \mathcal{L}^{-1})$ achieves its minimum $\lambda_{n+1}(\mathcal{L}^{-1})$ if we use the first n eigenfunctions of $\mathcal{L}^{-1}$ to form Ψ (λ_n is the nth eigenvalue arranged in a descending order). However, the eigenfunctions are expensive to compute and do not have localized support [20,40,49]. In many cases, localized/sparse basis functions are preferred. For example, in the multiscale finite element method [12], localized basis functions lead to sparse linear systems and thus result in more efficient algorithms, see, e.g., [1,2,5,10,11,22,23,28,36,39,44]. In quantum chemistry, localized basis functions like the Wannier functions have better interpretability of the local interactions between particles (see, e.g., [26,29,30,40,47]), and also lead to more efficient algorithms [15]. In statistics, the sparse principal component analysis (SPCA) looks for sparse vectors to span the eigenspace of the covariance matrix, which leads to better interpretability compared with the PCA, see, e.g., [8,25,45,46,49].

1.2 Summary of our main results

In this paper, we study operator compression for higher-order elliptic operators. We assume that the self-adjoint elliptic operator $\mathcal{L}$ is coercive, bounded and strongly elliptic (to be made precise in Sect. 6.2). Under these assumptions, we construct n basis functions $\Psi^{\mathrm{loc}} = [\psi_1^{\mathrm{loc}}, \ldots, \psi_n^{\mathrm{loc}}]$ that achieve nearly optimal performance on both ends in the accuracy–sparsity trade-off (1.10).

1. They are optimally localized up to a logarithmic factor, i.e.,

$$\left|\mathrm{supp}(\psi_i^{\mathrm{loc}})\right| \le \frac{C_l \log(n)}{n} \quad \forall 1 \le i \le n. \tag{1.4}$$

   Here, $|\mathrm{supp}(\psi_i^{\mathrm{loc}})|$ denotes the area/volume of the support of the localized function ψ_i^{loc} in $\mathbb{R}^d$, and the constant C_l is independent of n.

2. If we use a generalized finite element method [1,10,22,44] to solve the elliptic equations, we achieve the optimal convergence rate in the energy norm, i.e.,

$$\|\mathcal{L}^{-1}f - \Psi^{\mathrm{loc}} L_n^{-1} (\Psi^{\mathrm{loc}})^T f\|_H \le C_e \sqrt{\lambda_n(\mathcal{L}^{-1})} \|f\|_2 \quad \forall f \in L^2(D), \tag{1.5}$$

   where L_n is the stiffness matrix under the basis Ψ^{loc}, $\|\cdot\|_H$ is the associated energy norm, and C_e is independent of n.

3. For the sparse operator compression problem, we achieve the optimal approximation error up to a constant, i.e.,

$$E_{\mathrm{oc}}(\Psi^{\mathrm{loc}};\mathcal{L}^{-1}) \le C_e^2 \lambda_n(\mathcal{L}^{-1}), \tag{1.6}$$

where $E_{\mathrm{oc}}(\Psi^{\mathrm{loc}};\mathcal{L}^{-1})$ is the operator compression error defined in Eq. (1.3).

We will focus on the theoretical analysis of the approximation accuracy (1.5) and the localization of the basis functions (1.4).

1.3 Our construction

To construct such localized basis functions $\Psi^{\mathrm{loc}} = [\psi_1^{\mathrm{loc}}, \ldots, \psi_n^{\mathrm{loc}}]$, we first partition the physical domain D using a regular partition $\{\tau_i\}_{i=1}^m$ with mesh size h. We pick $\{\varphi_{i,q}\}_{q=1}^Q$ to be a set of orthogonal basis functions of $\mathcal{P}_{k-1}(\tau_i)$, which is the space of all d-variate polynomials of degree at most $k-1$ on the patch $\tau_i \subset D$, and $Q = \binom{k+d-1}{d}$ is the dimension of the space $\mathcal{P}_{k-1}(\tau_i)$. For $r > 0$, let S_r be the union of the subdomains τ_j that intersect with $B(x_i, r)$ (for some $x_i \in \tau_i$) and let $\psi_{i,q}^{\mathrm{loc}}$ be the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_{i,q}^{\mathrm{loc}} = \arg\min_{\psi\in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & \int_{S_r} \psi\varphi_{j,q'} = \delta_{iq,jq'} \quad \forall 1 \le j \le m, \quad 1 \le q' \le Q, \\ & \psi(x) \equiv 0, \quad x \in D\backslash S_r. \end{aligned} \tag{1.7}$$

Here, the space $H = \{\mathcal{L}^{-1}f : f \in L^2(D)\}$ is the solution space of the operator $\mathcal{L}$, and $\|\cdot\|_H$ is the energy norm associated with $\mathcal{L}$ and the prescribed boundary condition. It is important to point out that the boundary condition of the elliptic problem is already incorporated in the above optimization problem through the solution space H and the definition of the energy norm $\|\cdot\|_H$. This variational formulation is very general and can take into account lower-order terms very easily.

Collecting all the $\psi_{i,q}^{\mathrm{loc}}$ for $1 \le i \le m$ and $1 \le q \le Q$ together, we get our basis Ψ^{loc}. We will prove that for $r = \mathcal{O}(h\log(1/h))$,

1. they achieve the optimal convergence rate to solve the elliptic equation, i.e.,
$$\|\mathcal{L}^{-1}f - \Psi^{\mathrm{loc}} L_n^{-1}(\Psi^{\mathrm{loc}})^T f\|_H \le C_e h^k \|f\|_2 \quad \forall f \in L^2(D), \tag{1.8}$$
where the constant C_e is independent of n.
2. they achieve the optimal approximation error to approximate the elliptic operator, i.e.,
$$E_{\mathrm{oc}}(\Psi^{\mathrm{loc}};\mathcal{L}^{-1}) \le C_e^2 h^{2k}. \tag{1.9}$$

For $n = mQ$, we can show that the nth largest eigenvalue of $\mathcal{L}^{-1}$ is of the order h^{2k}, i.e., $\lambda_n(\mathcal{L}^{-1}) = \mathcal{O}(h^{2k})$. Therefore, the optimality above is exactly the optimality described in Eqs. (1.5) and (1.6).

1.4 Comparison with other existing methods

Our approach for operator compression originates at the MsFEM and numerical homogenization, where localized multiscale basis functions are constructed to approximate the solution space of some elliptic PDEs with multiscale coefficients; see [1,2,5,10,12,22,28, 35,36,39,44]. Specifically, our work is inspired by the work presented in [28,36], in which multiscale basis functions with support size $O(h\log(1/h))$ are constructed for second-order elliptic equations with rough coefficients and homogeneous Dirichlet boundary conditions. In this paper, we generalize the construction [36] and propose a general framework to compress higher-order elliptic operators with optimal compression accuracy and optimal localization.

We remark that although we use the framework presented in [36] as the direct template for our method, to the best of our knowledge, the local orthogonal decomposition (LOD) [28], in the context of multidimensional numerical homogenization, contains the first rigorous proof of optimal exponential decay rates with a priori estimates (leading to localization to subdomains of size $h\log(1/h)$, with basis functions derived from the Clement interpolation operator). The idea of using the preimage of some continuous or discontinuous finite element space under the partial differential operator to construct localized basis functions in Galerkin-type methods was even used earlier, e.g., in [16], although it did not provide a constructive local basis. In addition to establishing the exponential decay of the basis (for general nonconforming measurements of the solution, we will generalize the proof of this result to higher-order PDEs and measurements formed by local polynomials), a major contribution of [36] was to introduce a multiresolution operator decomposition for second-order elliptic PDEs with rough coefficients.

There are several new ingredients in our analysis that are essential for us to obtain our results for higher-order elliptic operators with rough coefficients. First of all, we prove an inverse energy estimate for functions in Ψ, which is crucial in proving the exponential decay. In particular, Lemma 4.1 is an essential step to obtaining the inverse energy estimate for higher-order PDEs that is not found in [28] nor [36]. We remark that Lemma 3.12 in [36] provides such an estimate for second-order elliptic operators, by utilizing a relation between the Laplacian operator Δ and the d-dimensional Brownian motion. It is not straightforward to extend this probabilistic argument to higher-order cases. In contrast, our inverse energy estimate is valid for any $2k$th-order elliptic operators and is tighter than the estimation in [36] for the second-order case. Secondly, we prove a projection-type polynomial approximation property in $H^k(D)$. This polynomial approximation property plays an essential role in both estimating the compression accuracy and in localizing the basis functions. Thirdly, we propose the notion of the strong ellipticity to analyze the higher-order elliptic operators and show that strong ellipticity is only slightly stronger than the standard uniform ellipticity. Very recently, the authors of [37] introduce the Gaussian cylinder measure and successfully generalize the probabilistic framework in [35, 36] to a much broader class of operators, including higher-order elliptic operators without requiring the strong ellipticity.

As in [28,36], the error bound in our convergence analysis blows up for fixed oversampling ratio r/h. To achieve the desired $O(h^k)$ accuracy in the energy norm, we require $r/h = O(\log(1/h))$. There has been some previous attempt to study the convergence of MsFEM using oversampling techniques with r/h being fixed, see, e.g., [18,41]. In particular, the authors of [18,41] showed that if the oversampling ratio r/h is fixed, the accuracy of the numerical solution will depend on the regularity of the solution and cannot be guaranteed for problems with rough coefficients. By imposing $r/h = O(\log(1/h))$, the authors of [18,41] proved that the MsFEM with constrained oversampling converges with the desired accuracy $O(h)$.

There has been some previous work for second-order elliptic PDEs by using basis functions of support size $O(h)$, see, e.g., [2,21]. However, they need to use $O(\log(1/h))$ basis functions associated with each coarse finite element to recover the $O(h)$ accuracy. The computational complexity of this approach is comparable to the one that we present in this paper. It is worth mentioning that the authors of [21] use a local oversampling operator to construct the optimal local boundary conditions for the nodal multiscale basis and enrich the nodal multiscale basis with optimal edge multiscale basis. Moreover, the method in [21] allows an explicit control of the approximation accuracy in the offline stage by truncating the SVD of the oversampling operator. In [21], the authors demonstrated numerically that this method is robust to high-contrast problems and the number of basis functions per coarse element is typically small. We remark that the recently developed generalized multiscale finite element method (GMsFEM) [5,10] has provided another promising approach in constructing multiscale basis functions with support size $O(h)$.

Another popular way to formulate the operator compression problem is to solve the following l^1 penalized variational problem:

$$\begin{aligned} \min_{\Psi} \quad & \sum_{i=1}^{n} \|\psi_i\|_H^2 + \lambda \sum_{i=1}^{n} \|\psi_i\|_1, \\ \text{s.t.} \quad & (\psi_i, \psi_j) = \delta_{i,j} \quad \forall 1 \leq i, j \leq n, \end{aligned} \tag{1.10}$$

where $\|\psi_i\|_H$ is the energy norm induced by the operator $\mathcal{L}$. In problem (1.10), enforcing $\|\psi_i\|_H$ to be small leads to a small compression error, enforcing $\|\psi_i\|_1$ to be small leads to a sparse basis function, and $\lambda > 0$ is a parameter to control the trade-off between the accuracy and sparsity.

The sparse PCA (SPCA) is closely related to the above l^1-based optimization problem. Given a covariance function $K(x, y)$, the SPCA solves a variational problem similar to Eq. (1.10):

$$\begin{aligned} \min_{\Psi} \quad & -\sum_{i=1}^{n} (\psi_i, \mathcal{K}\psi_i) + \lambda \sum_{i=1}^{n} \|\psi_i\|_1, \\ \text{s.t.} \quad & (\psi_i, \psi_j) = \delta_{i,j} \quad \forall 1 \leq i, j \leq n, \end{aligned} \tag{1.11}$$

where $(\psi_i, \mathcal{K}\psi_i) := \int_D \int_D K(x, y)\psi_i(x)\psi_i(y)\mathrm{d}x\,\mathrm{d}y$. In the SPCA (1.11), we have the minus sign in front the variational term because we are interested in the eigenspace corresponding to the largest n eigenvalues. Although the l^1 approach performs well in practice, neither Problem (1.10) nor the SPCA (1.11) is convex, and one needs to use some sophisticated techniques to solve the non-convex optimization problem or its convex relaxation; see, e.g., [8,26,40,45,49].

In comparison with the l^1-based optimization method or the SPCA, our approach has the advantage that this construction will guarantee that $\psi_{i,q}$ decays exponentially fast away from τ_i. This exponential decay justifies the local construction of the basis functions in Eq. (1.7). Moroever, our construction (1.7) is a quadratic optimization with linear constraints, which can be solved as efficiently as solving an elliptic problem on the local domain S_r. The computational complexity to obtain all n localized basis functions $\{\psi_i^{\mathrm{loc}}\}_{i=1}^n$ is only of order $N \log^{3d}(N)$ if a multilevel construction is employed, where N is the degree of freedom in the discretization of $\mathcal{L}$; see [36]. In contrast, the orthogonality constraint in Eq. (1.10) is not convex, which introduces additional difficulties in solving the problem. Finally, our construction of $\{\psi_i^{\mathrm{loc}}\}_{i=1}^n$ is completely decoupled, while all the basis functions in Eq. (1.10) are coupled together. This decoupling leads to a simple parallel execution and thus makes the computation of $\{\psi_i^{\mathrm{loc}}\}_{i=1}^n$ even more efficient.

The rest of the paper is organized as follows. In Sect. 2, we introduce the abstract framework of the sparse operator compression. In Sect. 3, we prove a projection-type polynomial approximation property for the Sobolev spaces, which can be seen as a generalization of the Poincare inequality for functions with higher regularity. This polynomial approximation property is critical in our analysis of the higher-order case. It plays a role similar to that of the Poincare inequality in the analysis of the second-order elliptic operator. In Sect. 4, we prove the inverse energy estimate by scaling. In Sect. 5, we use the second-order elliptic PDE to illustrate the main idea of our analysis. In Sect. 6, we first introduce the notion of strong ellipticity and then prove the exponential decay of the constructed basis function for strongly elliptic operators. In Sect. 7, we localize the basis functions and provide the convergence rate for the corresponding MsFEM and the compression rate for the corresponding operator compression. Finally, we present several numerical results to support the theoretical findings in Sect. 8. Some concluding remarks are made in Sect. 9 and a few technical proofs are deferred to the "Appendix."

2 Operator compression

In this section, we provide an abstract and general framework to compress a bounded self-adjoint positive semidefinite operator $\mathcal{K} : X \to X$, where X can be any separable Hilbert space with inner product $(\cdot, \cdot)$. In the case of operator compression of an elliptic operator $\mathcal{L}$, $\mathcal{K}$ plays the role of the solution operator $\mathcal{L}^{-1}$ and $X = L^2(D)$. In the case of the SPCA, $\mathcal{K}$ plays the role of the covariance operator. In Sect. 2.1, we introduce the Cameron–Martin space, which plays the role of the solution space of $\mathcal{L}$. In Sect. 2.2, we provide our main theorem to estimate the compression error. We will use this abstract framework to compress elliptic operators in the rest of the paper.

2.1 The Cameron–Martin space

Suppose $\{(\lambda_n, e_n)\}_{n=1}^{\infty}$ are the eigen pairs of the operator $\mathcal{K}$ with the eigenvalues $\{\lambda_n\}_{n=1}^{\infty}$ in a descending order. We have $\lambda_n \geq 0$ for all n since $\mathcal{K}$ is self-adjoint and positive semidefinite. From the spectral theorem of a self-adjoint operator, we know that $\{e_n\}_{n=1}^{\infty}$ forms an orthonormal basis of X.

Lemma 2.1 *Let $\mathcal{K}(X)$ be the range space of $\mathcal{K}$. We have*

1. *$\mathcal{K}(X)$ is an inner product space with inner product defined by*

$$(\mathcal{K}\varphi_1, \mathcal{K}\varphi_2)_H = (\mathcal{K}\varphi_1, \varphi_2) \qquad \forall \varphi_1, \varphi_2 \in X. \tag{2.1}$$

2. *$\mathcal{K}(X)$ is continuously imbedded in X.*
3. *$\mathcal{K}(X)$ is dense in X if the null space of $\mathcal{K}$ only contains the origin, i.e.,* $\mathrm{null}(\mathcal{K}) = \{\mathbf{0}\}$.

Proof 1. Since $\mathcal{K}$ is self-adjoint, we have $(\mathcal{K}\varphi_1, \mathcal{K}\varphi_2)_H = (\mathcal{K}\varphi_2, \mathcal{K}\varphi_1)_H$. The linearity and nonnegativity are obvious. Finally, if $(\mathcal{K}\varphi, \mathcal{K}\varphi)_H = 0$ for some $\varphi \in X$, then $(\mathcal{K}\varphi, \varphi) = 0$. Suppose that $\varphi = \sum_n \alpha_n e_n$ by expanding φ with eigenvectors of $\mathcal{K}$. Then, we have

$(\mathcal{K}\varphi,\varphi) = \sum_n \lambda_n \alpha_n^2 = 0$. Therefore, $\alpha_n = 0$ for all $\lambda_n > 0$. Equivalently, we obtain $\varphi \in \mathrm{null}(\mathcal{K})$, i.e., $\mathcal{K}\varphi = 0$.

2. Since $\lambda_n^2 \le \lambda_1\lambda_n$ for all $n \in \mathbb{N}$, we have $\mathcal{K}^2 \preceq \lambda_1\mathcal{K}$. Then, we obtain

$$\sqrt{(\mathcal{K}\varphi,\mathcal{K}\varphi)} \le \sqrt{\lambda_1(\mathcal{K}\varphi,\varphi)} = \sqrt{\lambda_1}\sqrt{(\mathcal{K}\varphi,\mathcal{K}\varphi)_H}, \tag{2.2}$$

where we have used the definition of $(\cdot,\cdot)_H$ in Eq. (2.1) in the last step.

3. If $\mathrm{null}(\mathcal{K}) = \{\mathbf{0}\}$, we have $\mathrm{span}\{e_n, n \ge 1\} \subset \mathcal{K}(X)$. Then, $\mathcal{K}(X)$ is dense in X. □

We define the Cameron–Martin space H as the completion of $\mathcal{K}(X)$ with respect to the norm $\sqrt{(\cdot,\cdot)_H}$. Then, H is a separable Hilbert space and we have the following lemma.

Lemma 2.2 1. *H can be continuously embedded into X.*

2. *H is dense in X if* $\mathrm{null}(\mathcal{K}) = \{\mathbf{0}\}$*.*
3. *For all $\psi \in X$ and all $f \in H$, we have*

$$(f,\mathcal{K}\psi)_H = (f,\psi). \tag{2.3}$$

Proof 1. By the continuous imbedding from $\mathcal{K}(X)$ to X, we know that a Cauchy sequence in $\mathcal{K}(X)$ is also a Cauchy sequence in X. Therefore, we have $H \subset X$. By Eq. (2.2) and the the continuity of norms, we have $(\psi,\psi) \le \lambda_1(\psi,\psi)_H$ for any $\psi \in H$.

2. It is obvious from item 3 in Lemma 2.1.
3. If $f \in \mathcal{K}(X)$, Eq. (2.3) is exactly the definition of $(\cdot,\cdot)_H$ in Eq. (2.1). By the continuity of the inner product, Eq. (2.3) is true for any $f \in H$. □

2.2 Operator compression

Suppose H is an arbitrary separable Hilbert space and $\Phi \subset H$ is n-dimensional subspace in H with basis $\{\varphi_i\}_{i=1}^n$. In the rest of the paper, $\mathcal{P}_\Phi^{(H)}$ denotes the orthogonal projection from a Hilbert space H to its subspace Φ. With this notation, we present our theorem for error estimates below.

Theorem 2.1 *Suppose there is a n-dimensional subspace $\Phi \subset X$ with basis $\{\varphi_i\}_{i=1}^n$ such that*

$$\|u - \mathcal{P}_\Phi^{(X)}u\|_X \le k_n\|u\|_H \qquad \forall u \in \mathcal{K}(X) \subset H. \tag{2.4}$$

Let Ψ be the n-dimensional subspace in H (also in X) spanned by $\{\mathcal{K}\varphi_i\}_{i=1}^n$. Then

1. *For any $u \in \mathcal{K}(X)$ and $u = \mathcal{K}f$, we have*

$$\|u - \mathcal{P}_\Psi^{(H)}u\|_H \le k_n\|f\|_X. \tag{2.5}$$

2. *For any $u \in \mathcal{K}(X)$ and $u = \mathcal{K}f$, we have*

$$\|u - \mathcal{P}_\Psi^{(H)}u\|_X \le k_n^2\|f\|_X. \tag{2.6}$$

3. *We have*

$$\|\mathcal{K} - \mathcal{P}_\Psi^{(H)}\mathcal{K}\| \le k_n^2, \tag{2.7}$$

where $\|\cdot\|$ *is the induced operator norm on* $\mathcal{B}(X, X)$. *Moreover, the rank-n operator* $\mathcal{P}_\Psi^{(H)}\mathcal{K} : X \to X$ *is self-adjoint.*

In Theorem 2.1, by using a projection-type approximation property of Φ in H, i.e., Eq. (2.4), we obtain the error estimates of the multiscale finite element method with finite element basis $\{\mathcal{K}\varphi_i\}_{i=1}^n$ in the energy norm, i.e., Eq. (2.5). We will take Φ as the discontinuous piecewise polynomial space later, which is a poor finite element space for elliptic equations with rough coefficients. However, after smoothing Φ with the solution operator $\mathcal{K}$, the smoothed basis functions $\{\mathcal{K}\varphi_i\}_{i=1}^n$ have the optimal convergence rate. This data-dependent methodology to construct finite element spaces was pioneered by the generalized finite element (GFEM) [1,44], the multiscale finite element method (MsFEM) [12,22,24], and numerical homogenization [28,36].

Our error analysis is different from the traditional finite element error analysis in two aspects. First of all, the traditional error analysis relies on an interpolation type approximation property where higher regularity is required. For example, the error analysis for the FEM with standard linear nodal basis functions for the Poisson equation requires the following interpolation type approximation:

$$|u - \mathcal{I}_h u|_{1,2,D} \le Ch|u|_{2,2,D} \quad \forall u \in H_0^2(D), \tag{2.8}$$

where $\mathcal{I}_h u$ is the piecewise linear interpolation of the solution u. In Eq. (2.8), one assumes $u \in H^2(D)$, but this is not the case for elliptic operators with rough coefficients. Secondly, in our projection-type approximation property (2.4) the error is measured by the "weaker" $\|\cdot\|_X$ norm, while in the traditional interpolation type approximation property the error is measured by the "stronger" $\|\cdot\|_H$ norm. In this sense, our error estimate relies on weaker assumptions. As far as we know, this kind of error estimate was first introduced in Proposition 3.6 in [36].

Proof of Theorem 2.1 1. For an arbitrary $v \in \Psi$, due to the definition of Ψ, we can write $v = \mathcal{K}(\sum_{i=1}^n c_i\varphi_i)$, and thus we get $u - v = \mathcal{K}(f - \sum_{i=1}^n c_i\varphi_i)$. By Lemma 2.2, we have

$$\begin{aligned}\|u - v\|_H^2 &= \left(u - v, f - \sum_{i=1}^n c_i\varphi_i\right)\\ &= \left(u - v - \mathcal{P}_\Phi^{(X)}(u - v), f - \sum_{i=1}^n c_i\varphi_i\right) + \left(\mathcal{P}_\Phi^{(X)}(u - v), f - \sum_{i=1}^n c_i\varphi_i\right).\end{aligned}$$

By choosing c_i such that $\sum_{i=1}^n c_i\varphi_i = \mathcal{P}_\Phi^{(X)}(f)$, the second term vanishes. Then, we obtain

$$\begin{aligned}\|u - v\|_H^2 &= \left(u - v - \mathcal{P}_\Phi^{(X)}(u - v), f - \sum_{i=1}^n c_i\varphi_i\right)\\ &\le \|u - v - \mathcal{P}_\Phi^{(X)}(u - v)\|_X \|f - \mathcal{P}_\Phi^{(X)}(f)\|_X \le k_n\|u - v\|_H\|f\|_X\end{aligned}$$

Therefore, we conclude $\|u - v\|_H \le k_n\|f\|_X$.

2. We use the Aubin–Nistche duality argument to get the estimation in item 2. Let $v = \mathcal{K}(u - \mathcal{P}_\Psi^{(H)} u)$. On one hand, we get

$$(u-\mathcal{P}_\Psi^{(H)} u, v-\mathcal{P}_\Psi^{(H)} v)_H = (u-\mathcal{P}_\Psi^{(H)} u, v)_H = (u-\mathcal{P}_\Psi^{(H)} u, u-\mathcal{P}_\Psi^{(H)} u)_X = \|u-\mathcal{P}_\Psi^{(H)} u\|_X^2.$$

On the other hand, we obtain

$$(u-\mathcal{P}_\Psi^{(H)} u, v-\mathcal{P}_\Psi^{(H)} v)_H \le \|u-\mathcal{P}_\Psi^{(H)} u\|_H \|v-\mathcal{P}_\Psi^{(H)} v\|_H \le k_n \|f\|_X \, k_n \|u-\mathcal{P}_\Psi^{(H)} u\|_X.$$

We have used the result of item 1 in the last step. Combining these two estimates, the result follows.

3. From the last item, we obtain that $\|\mathcal{K}f - \mathcal{P}_\Psi^{(H)} \mathcal{K}f\|_X \le k_n^2 \|f\|_X$ for any $f \in X$. Therefore, we conclude $\|\mathcal{K} - \mathcal{P}_\Psi^{(H)} \mathcal{K}\| \le k_n^2$. Now, we prove that $\mathcal{P}_\Psi^{(H)} \mathcal{K}$ is self-adjoint. For any $x_1, x_2 \in X$, by definition of H-norm we have

$$(x_1, \mathcal{P}_\Psi^{(H)} \mathcal{K} x_2) = (\mathcal{K} x_1, \mathcal{P}_\Psi^{(H)} \mathcal{K} x_2)_H.$$

Since $\mathcal{P}_\Psi^{(H)}$ is self-adjoint in H, we have

$$(\mathcal{K} x_1, \mathcal{P}_\Psi^{(H)} \mathcal{K} x_2)_H = (\mathcal{P}_\Psi^{(H)} \mathcal{K} x_1, \mathcal{K} x_2)_H = (\mathcal{P}_\Psi^{(H)} \mathcal{K} x_1, x_2),$$

where we have used the definition of H-norm again in the last step. □

Although the basis functions $\{\mathcal{K}\varphi_i\}_{i=1}^n$ have good approximation accuracy, they are typically not localized. Therefore, we construct another set of basis functions $\{\psi_i\}_{i=1}^n$ for Ψ via the following variational approach, which results in basis functions with good localization properties. For any given $i \in \{1, 2, \ldots, n\}$, consider the following quadratic optimization problem

$$\begin{aligned} \psi_i = \arg\min_{\psi \in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_j) = \delta_{i,j}, \quad j = 1, 2, \ldots, n. \end{aligned} \tag{2.9}$$

Define $\Theta \in \mathbb{R}^{n\times n}$ by

$$\Theta_{i,j} := (\mathcal{K}\varphi_i, \varphi_j). \tag{2.10}$$

It is easy to verify that $\{\mathcal{K}\varphi_i\}_{i=1}^n$ are linearly independent if and only if Θ is invertible. We will write Θ^{-1} as its inverse and $\Theta_{i,j}^{-1}$ as the (i, j)th entry of Θ^{-1}. It is not difficult to prove the following properties of ψ_i, which is defined as the unique minimizer of Eq. (2.9).

Theorem 2.2 *If* $\text{null}(\mathcal{K}) \cap \Phi = \{\mathbf{0}\}$ *holds true, then we have*

1. *The optimization problem* (2.9) *admits a unique minimizer* ψ_i*, which can be written as*

$$\psi_i = \sum_{j=1}^n \Theta_{i,j}^{-1} \mathcal{K}\varphi_j. \tag{2.11}$$

2. *For $w \in \mathbb{R}^n$, $\sum_{i=1}^n w_i\psi_i$ is the minimizer of $\|\psi\|_H$ subject to $(\varphi_j, \psi) = w_j$ for $j = 1, 2, \ldots, n$. Moreover, for any ψ which satisfies $(\varphi_j, \psi) = w_j$ for $j = 1, 2, \ldots, n$, we have*
$$\|\psi\|_H^2 = \left\|\sum_{i=1}^n w_i\psi_i\right\|_H^2 + \left\|\psi - \sum_{i=1}^n w_i\psi_i\right\|_H^2. \tag{2.12}$$
3. $(\psi_i, \psi_j)_H = \Theta_{i,j}^{-1}$.

With a good choice of the space Φ and its basis $\{\varphi_i\}_{i=1}^n$, the energy-minimizing basis ψ_i, defined in Eq. (2.9), enjoys good localization properties. We will prove that the energy-minimizing basis function ψ_i decays exponentially fast away from its associated patch. The localization property justifies the following local construction of the basis functions:

$$\begin{aligned} \psi_i^{\text{loc}} = \arg\min_{\psi\in H} \quad & \|\psi\|_H^2 \\ \text{s.t.} \quad & (\psi, \varphi_j) = \delta_{i,j}, \quad j = 1, 2, \ldots, n, \\ & \psi(x) \equiv 0, \quad x \in D\backslash S_i, \end{aligned} \tag{2.13}$$

where $S_i \subset D$ is a neighborhood of the patch that ψ_i is associated with. Compared with Eq. (2.9), the localized basis ψ_i^{loc} is obtained by solving exactly the same quadratic problem but on a local domain S_i.

To compress elliptic operators with order $2k$, we take Φ as the space of (discontinuous) piecewise polynomials, with degree no more than $k-1$. We take its basis as $\{\varphi_{i,q}\}_{i=1,q=1}^{m,Q}$, where $Q := \binom{k+d-1}{d}$ is the dimension of the d-variate polynomial space with degree no more than $k-1$ and $\{\varphi_{i,q}\}_{q=1}^Q$ is an orthonormal basis of the polynomial space on the patch τ_i. Two main theoretical results in this paper are as follows.

1. The basis function ψ_i decays exponentially fast away from its associated patch; see Theorems 6.3 and 6.4.
2. The localized basis function ψ_i^{loc} approximates ψ_i accurately; see Theorem 7.1. Meanwhile, the compression rate $E_{\text{oc}}(\Psi^{\text{loc}}; \mathcal{L}^{-1})$ is the same as $E_{\text{oc}}(\Psi; \mathcal{L}^{-1})$; see Theorem 7.2 and Corollary 7.3.

3 A projection-type polynomial approximation property

The following projection-type polynomial approximation property in the Sobolev space $H^k(D)$ plays an essential role in both obtaining the optimal approximation error and proving the exponential decay of the energy-minimizing basis functions. It can be viewed as a generalized Poincare inequality.

Theorem 3.1 *Suppose $\Omega \subset \mathbb{R}^d$ is affine equivalent to $\widehat{\Omega}$, i.e., there exists an invertible affine mapping*
$$F : \widehat{x} \in \widehat{\Omega} \to F(\widehat{x}) = B\widehat{x} + b \in \Omega \tag{3.1}$$
such that $F(\widehat{\Omega}) = \Omega$. Let h be the diameter of Ω and δh be the maximum diameter of a ball inscribed in Ω. Let the mapping $\Pi : H^{k+1}(\Omega) \to \mathcal{P}_k(\Omega)$ be the projection onto the

polynomial space with degree no greater than k in $L^2(\Omega)$. Then, there exists a constant $C(k, \widehat{\Omega})$ such that for any $u \in H^{k+1}(\Omega)$ and any $0 \le p \le k+1$

$$|u - \Pi u|_{p,2,\Omega} \le C(k, \widehat{\Omega})\delta^{-p}h^{k-p+1}|u|_{k+1,2,\Omega}. \tag{3.2}$$

To prove Theorem 3.1, we use a basic result about the Sobolev spaces, due to J. Deny and J.L. Lions, which pervades the mathematical analysis of the finite element method: over the quotient space $H^{k+1}(D)/\mathcal{P}_k(D)$, the seminorm $|\cdot|_{k+1,D}$ is a norm equivalent to the quotient norm. We will use the following theorem (Theorem 3.1.4 in [6]), to prove Theorem 3.1.

Theorem 3.2 *For some integers $k \ge 0$ and $m \ge 0$, let $H^{k+1}(\widehat{\Omega}) \equiv W^{k+1,2}(\widehat{\Omega})$ and $H^m(\widehat{\Omega}) \equiv W^{m,2}(\widehat{\Omega})$ be Sobolev spaces satisfying the inclusion*

$$H^{k+1}(\widehat{\Omega}) \subset H^m(\widehat{\Omega}),$$

and let $\widehat{\Pi} : H^{k+1}(\widehat{\Omega}) \to H^m(\widehat{\Omega})$ be a continuous linear mapping such that

$$\widehat{\Pi}\widehat{p} = \widehat{p}, \qquad \forall \widehat{p} \in \mathcal{P}_k(\widehat{\Omega}).$$

For any open set Ω which is affine equivalent to the set $\widehat{\Omega}$ (see Eq. (3.1)), let the mapping Π_Ω be defined by

$$\widehat{\Pi_\Omega v} = \widehat{\Pi}\widehat{v},$$

for all functions $\widehat{v} \in H^{k+1}(\widehat{\Omega})$ and $v \in H^{k+1}(\Omega)$ in the correspondence $(\widehat{v} : \widehat{\Omega} \to \mathbb{R}) \to (v = \widehat{v} \circ F^{-1} : \Omega \to \mathbb{R})$. Then, there exists a constant $C(\widehat{\Pi}, \widehat{\Omega})$ such that, for all affine-equivalent sets Ω,

$$|v - \Pi_\Omega v|_{m,2,\Omega} \le C(\widehat{\Pi}, \widehat{\Omega})\delta^{-m}h^{k-m+1}|v|_{k+1,2,\Omega}, \qquad \forall v \in H^{k+1}(\Omega), \tag{3.3}$$

where $h = diam(\Omega)$ and δh is the diameter of the biggest ball contained in Ω.

By specializing the operator $\widehat{\Pi}$ to be the projection of $H^{k+1}(\widehat{\Omega})$ to the polynomial space $\mathcal{P}_k(\widehat{\Omega})$ in $L^2(\widehat{\Omega})$, we can prove Theorem 3.1.

Proof of Theorem 3.1 Let $\widehat{\Pi} : H^{k+1}(\widehat{\Omega}) \to \mathcal{P}_k(\widehat{\Omega})$ be the orthogonal projection in $L^2(\widehat{\Omega})$. Let $F : \widehat{\Omega} \to \Omega$ be the invertible linear map and write $F(\widehat{x}) = B\widehat{x} + b$. Define Π_Ω as

$$\widehat{\Pi_\Omega v} = \widehat{\Pi}\widehat{v},$$

for all functions $\widehat{v} \in H^{k+1}(\widehat{\Omega})$ and $v \in H^{k+1}(\Omega)$ in the correspondence of the linear mapping. In the following, we prove that $\Pi_\Omega : H^{k+1}(\Omega) \to H^{k+1}(\Omega)$ is indeed the orthogonal projection from $H^{k+1}(\Omega)$ to $\mathcal{P}_k(\Omega)$ in $L^2(\Omega)$.

First of all, we have $\Pi_\Omega v = (\widehat{\Pi}\widehat{v}) \circ F^{-1}$ from definition. Since $\widehat{\Pi}\widehat{v} \in \mathcal{P}_k(\widehat{\Omega})$, we have $\Pi_\Omega v \in \mathcal{P}_k(\Omega)$. Secondly, for any $v \in \mathcal{P}_k(\Omega)$, $\widehat{v} = v \circ F \in \mathcal{P}_k(\widehat{\Omega})$, and thus $\widehat{\Pi}\widehat{v} = \widehat{v}$ by the definition of $\widehat{\Pi}$. Therefore, we have $\Pi_\Omega v = \widehat{v} \circ F^{-1} = v$ for any $v \in \mathcal{P}_k(\Omega)$. Thirdly, by changing variable with $x = F(\widehat{x})$, for any $v \in H^{k+1}(\Omega)$ and any $p(x) \in \mathcal{P}_k(\Omega)$, we have

$$\int_\Omega (v(x) - (\Pi_\Omega v)(x))\, p(x)\mathrm{d}x = \int_{\widehat{\Omega}} \left(\widehat{v}(\widehat{x}) - (\widehat{\Pi}\widehat{v})(\widehat{x})\right) \widehat{p}(\widehat{x})\mathrm{d}\widehat{x} \det B = 0.$$

In the last equality, we have used the fact that $\widehat{p} \in \mathcal{P}_k(\widehat{\Omega})$ if $p \in \mathcal{P}_k(\Omega)$ and the fact that $\widehat{\Pi} : H^{k+1}(\widehat{\Omega}) \to \mathcal{P}_k(\widehat{\Omega})$ is the orthogonal projection in $L^2(\widehat{\Omega})$. Therefore, the kernel space of Π_Ω is orthogonal to its range space, i.e., $\mathcal{P}_k(\Omega)$. With the three points above, we have proved that Π_Ω is the orthogonal projection from $H^{k+1}(\Omega)$ to $\mathcal{P}_k(\Omega)$ in $L^2(\Omega)$.

Finally, applying Theorem 3.2 with $\widehat{\Pi}$ and Π_Ω above, we prove Theorem 3.1 with the constant $C(k, \widehat{\Omega}) := C(\widehat{\Pi}, \widehat{\Omega})$ in Eq. (3.3). □

We also give the following theorem, which is a direct result of the Friedrichs' inequality; see, e.g., [34].

Theorem 3.3 *Let Ω_h be a smooth, bounded, open subset of $\mathbb{R}^d$ with diameter at most h. There exists a positive constant C_f such that*

$$|u|_{p,2,\Omega_h} \le C_f h^{k-p} |u|_{k,2,\Omega_h} \quad \forall u \in H_0^k(\Omega_h). \tag{3.4}$$

Here, $C_f = C_f(d,k)$ depends only on the physical dimension d and the order of the derivative k.

4 An inverse energy estimation by scaling

In the sparse operator compression, we will show that for a large set of compact operators, the basis functions $\{\psi_i\}_{i=1}^n$ constructed in (2.9) have exponentially decaying tails, which makes localization of these basis functions possible. The following lemma plays a key role in proving such exponential decay property.

Lemma 4.1 *Let Ω_h be a smooth, bounded, open subset of $\mathbb{R}^d$ with diameter at most h and $B(0, \delta h/2) \subset \Omega_h$ for some $\delta > 0$. For $k \in \mathbb{N}$, consider the operator $\mathcal{L} = (-1)^k \sum_{|\sigma|=k} D^{2\sigma}$ with the homogeneous Dirichlet boundary condition on $\partial\Omega_h$, i.e.,*

$$\begin{aligned} (-1)^k \sum_{|\sigma|=k} D^{2\sigma} u_h(x) &= f(x) \qquad x \in \Omega_h, \\ u_h &\in H_0^k(\Omega_h). \end{aligned} \tag{4.1}$$

Let $\mathcal{P}_s$ be the space of polynomials with order not greater than s. For $\gamma \ge 0$, there exists $C(k,s,d,\delta) > 0$, such that

$$\|\mathcal{L}u_h\|_{L^2(\Omega_h)} \le C(k,s,d,\delta) h^{-k} |u_h|_{k,2,\Omega_h} \quad \forall u_h \in \mathcal{L}^{-1}\mathcal{P}_{s-1}. \tag{4.2}$$

Proof Let G_h be the Green's function of Eq. (4.1). After multiplying u_h on both sides of Eq. (4.1) and integration by parts, we have $|u_h|_{k,2,\Omega_h} = \int_{\Omega_h} u_h(x) f(x) \mathrm{d}x$. Recall that $\mathcal{L}u_h \in \mathcal{P}_{s-1}$, and thus Eq. (4.2) is equivalent to

$$\int_{\Omega_h} p^2(x)\mathrm{d}x \le (C(k,s,d,\delta))^2 h^{-2k} \int_{\Omega_h}\int_{\Omega_h} G_h(x,y)p(x)p(y)\mathrm{d}x\,\mathrm{d}y, \quad \forall p \in \mathcal{P}_{s-1}. \tag{4.3}$$

Let $\{p_1, p_2, \ldots, p_Q\}$ be all the monomials that span $\mathcal{P}_{s-1}$. It is easy to see $Q = \binom{s+d-1}{d}$. For convenience, we assume that $\{p_i\}_{i=1}^Q$ are in non-decreasing order with respect to its degree. Specifically, $p_1 = 1$. Let $u_{h,i}$ be the solution of Eq. (4.1) with right hand side p_i, and $S_h, M_h \in \mathbb{R}^{Q\times Q}$ be defined as follows:

$$S_h(i,j) = \int_{\Omega_h}\int_{\Omega_h} G_h p_i p_j = \int_{\Omega_h} u_{h,i} p_j, \qquad M_h(i,j) = \int_{\Omega_h} p_i p_j. \tag{4.4}$$

Then, Eq. (4.3) is equivalent to

$$M_h \preceq (C(k,s,d,\delta))^2 \, h^{-2k} S_h, \tag{4.5}$$

where $A \preceq B$ means that $B - A$ is positive semidefinite. The change of variable $x = hz$ leads to $u_i(x) = h^{2k+o_i} u_{1,i}(z)$ where $u_{1,i}$ is the solution of the following PDE on $\Omega_1 \equiv \{x/h \,:\, x \in \Omega_h\}$:

$$\begin{aligned} (-1)^k \sum_{|\sigma|=k} D^{2\sigma} u_{1,i}(x) = p_i(x), \qquad & x \in \Omega_1, \\ u_{1,i} \in H_0^k(\Omega_1), \qquad & \end{aligned} \tag{4.6}$$

and o_i is the degree of p_i. Therefore, it is easy to check that

$$S_h(i,j) = h^{2k+o_i+o_j+d} S_1(i,j), \qquad M_h(i,j) = h^{o_i+o_j+d} M_1(i,j), \tag{4.7}$$

where $S_1(i,j) = \int_{\Omega_1}\int_{\Omega_1} G_1 p_i p_j = \int_{\Omega_1} u_{1,i} p_j$ and $M_1(i,j) = \int_{\Omega_1} p_i p_j$, which are independent of h. Notice that both S_1 and M_1 are symmetric positive definite, and let $\lambda_{\max}(M_1, S_1) > 0$ be the largest generalized eigenvalue of M_1 and S_1. By choosing

$$C(k,s,d,\Omega_1) = \sqrt{\lambda_{\max}(M_1,S_1)}, \tag{4.8}$$

we have

$$M_1 \preceq (C(k,s,d,\Omega_1))^2 \, S_1. \tag{4.9}$$

Combining (4.7) and (4.9), Eq. (4.5) naturally follows. In "Appendix A," we prove that $C(k,s,d,\Omega_1)$ can be bounded by $C(k,s,d,\delta)$, and this proves the lemma. □

For the case $s = k = 1$, we can take

$$C(1,1,d,\delta) = 2\sqrt{d(d+2)}\delta^{-1-d/2}.$$

as proved in Proposition (A.1). In this case, we have the estimate

$$|u_h|^2_{1,2,\Omega_h} \geq \frac{\delta^{d+2} h^2 |\Omega_h|}{4d(d+2)},$$

where $|\Omega_h|$ is the volume of Ω_h. The above bound is tight: when Ω_h is a ball with diameter h, the equality holds true. Making use of the mean exit time of a Brownian motion, the author of [36] obtained a different bound

$$|u_h|^2_{1,2,\Omega_h} \geq \frac{\delta^{d+2} h^{2+d} V_d}{2^{5+2d}},$$

where V_d is the volume of a unit d-dimensional ball. The two estimates have the same order of δ and h, but our estimates from Lemma 4.1 is much tighter. Moreover, Lemma 4.1 give estimates for any order k and any degree s, which plays a key role in proving the exponential decay in high-order cases, but the mean exit time of a Brownian motion is difficult to generalize to get these higher-order results.

5 Exponential decay of basis functions: the second-order case

The analysis for a general higher-order elliptic PDE is quite technical. In this section, we will prove that the basis function ψ_i for a second-order elliptic PDE has exponential decay away from τ_i. When $c \equiv 0$, this problem has been studied in [36]. When $c \neq 0$, it has been recently studied in [38] independently of our work. The results presented in this second-order case are not new [36]. We would like to use the simpler second-order elliptic PDE example to illustrate the main ingredients in the proof of exponential decay for a higher-order elliptic PDE, namely the recursive argument, the projection-type approximation property and the inverse energy estimate.

Consider the following second-order elliptic equation:

$$\begin{aligned} \mathcal{L}u := & -\nabla \cdot (a(x)\nabla u(x)) + c(x)u(x) = f(x), \qquad x \in D, \\ & u \in H_0^1(D), \end{aligned} \tag{5.1}$$

where D is an open bounded domain in $\mathbb{R}^d$, the potential $c(x) \geq 0$ and the diffusion coefficient $a(x)$ is a symmetric, uniformly elliptic $d \times d$ matrix with entries in $L^\infty(D)$. For simplicity, we consider the homogeneous Dirichlet boundary condition here. We emphasize that all our analysis can be carried over for other types of homogeneous boundary conditions. We assume that there exist $0 < a_{\min} \leq a_{\max}$ and $c_{\max}$ such that

$$a_{\min} I_d \preceq a(x) \preceq a_{\max} I_d, \quad 0 \leq c(x) \leq c_{\max}, \qquad x \in D. \tag{5.2}$$

To simply our notations, for any $\psi \in H$ and any subdomain $S \subset D$, $\|\psi\|_{H(S)}$ denotes $\left(\int_S \nabla\psi \cdot a\nabla\psi + c\psi^2\right)^{1/2}$. For the second-order case, the projection-type approximation property is simply the Poincare inequality. The following lemma provides us the inverse energy estimate. It is a special case of Lemma 6.2 and can be proved by using Lemma 4.1.

Lemma 5.1 *For any domain partition with $h \leq h_0 \equiv \pi\sqrt{\frac{a_{\max}}{2c_{\max}}}$, we have*

$$\|\mathcal{L}v\|_{L^2(\tau_j)} \leq \sqrt{a_{\max}} C(d,\delta) h^{-1} \|v\|_{H(\tau_j)} \quad \forall v \in \Psi, \quad \forall j = 1, 2, \ldots, m, \tag{5.3}$$

where $C(d,\delta) = \sqrt{8d(d+2)}\delta^{-1-d/2}$. If $c_{\max} = 0$, i.e., $c(x) \equiv 0$, Eq. (5.3) holds true for all $h > 0$ and $C(d,\delta) = \sqrt{4d(d+2)}\delta^{-1-d/2}$.

Now, we are ready to prove the exponential decay of the basis function ψ_i.

Theorem 5.1 *For $h \leq h_0 \equiv \pi\sqrt{\frac{a_{\max}}{2c_{\max}}}$, it holds true that*

$$\|\psi_i\|^2_{H(D\cap(B(x_i,r))^c)} \leq \exp\left(1 - \frac{r}{lh}\right) \|\psi_i\|^2_{H(D)} \tag{5.4}$$

with $l = \frac{e-1}{\pi}(1 + C(d,\delta))\sqrt{\frac{a_{\max}}{a_{\min}}}$ and $C(d,\delta) = \sqrt{8d(d+2)}(1/\delta)^{d/2+1}$. If $c_{\max} = 0$, i.e., $c(x) \equiv 0$, Eq. (5.4) holds true for all $h > 0$ with $l = \frac{e-1}{\pi}(1 + C(d,\delta))\sqrt{\frac{a_{\max}}{a_{\min}}}$ and $C(d,\delta) = \sqrt{4d(d+2)}\delta^{-1-d/2}$.

Proof Let $k \in \mathbb{N}$, $l > 0$ and $i \in \{1, 2, \ldots, m\}$. Let S_0 be the union of all the domains τ_j that are contained in the closure of $B(x_i, klh) \cap D$, let S_1 be the union of all the domains τ_j that

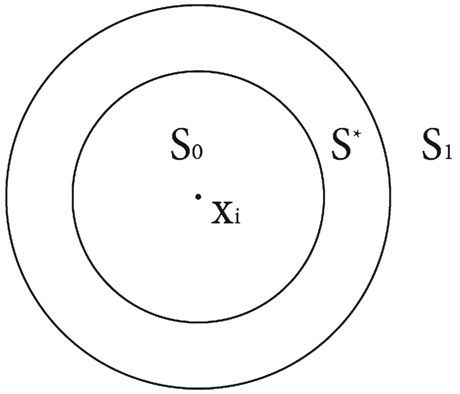

Fig. 1 Illustration of S_0, S_1 and S^*

are not contained in the closure of $B(x_i, (k+1)lh) \cap D$ and let $S^* = S_0^c \cap S_1^c \cap D$ (be the union of all the remaining elements τ_j not contained in S_0 or S_1), as illustrated in Fig. 1.

Let $b_k := \|\psi_i\|^2_{H(S_0^c)}$, and from definition we have $b_0 = \|\psi_i\|^2_{H(D)}$, $b_{k+1} = \|\psi_i\|^2_{H(S_1)}$ and $b_k - b_{k+1} = \|\psi_i\|^2_{H(S^*)}$. The strategy is to prove that for any $k \geq 1$, there exists constant C such that $b_{k+1} \leq C(b_k - b_{k+1})$. Then, we have $b_{k+1} \leq \frac{C}{C+1} b_k$ for any $k \geq 1$ and thus we get the exponential decay $b_k \leq (\frac{C}{C+1})^{k-1} b_1 \leq (\frac{C}{C+1})^{k-1} b_0$. We will choose l such that $C \leq \frac{1}{\mathrm{e}-1}$ and thus get $b_k \leq e^{1-k} b_0$, which gives the result (5.4). We start from $k = 1$ because we want to make sure $\tau_i \in S_0$; otherwise, $S_0 = \emptyset$ and $\tau_i \in S^*$.

Now, we prove that for any $k \geq 1$, there exists constant C such that $b_{k+1} \leq C(b_k - b_{k+1})$, i.e., $\|\psi_i\|^2_{H(S_1)} \leq C\|\psi_i\|^2_{H(S^*)}$. Let η be the function on D defined by $\eta(x) = \mathrm{dist}(x, S_0)/(\mathrm{dist}(x, S_0) + \mathrm{dist}(x, S_1))$. Observe that (1) $0 \leq \eta \leq 1$ (2) η is equal to zero on S_0 (3) η is equal to one on S_1 (4) $\|\nabla\eta\|_{L^\infty(D)} \leq \frac{1}{lh}$.[1]

By integration by parts, we obtain

$$\int_D \eta\nabla\psi_i \cdot a\nabla\psi_i + \int_D \eta c|\psi_i|^2 = \underbrace{\int_D \eta\psi_i(-\nabla\cdot(a\nabla\psi_i) + c\psi_i)}_{I_2} - \underbrace{\int_D \psi_i\nabla\eta\cdot a\nabla\psi_i}_{I_1}. \quad (5.5)$$

Since $a \succeq 0$ and $c \geq 0$, the left-hand side gives an upper bound for $\|\psi_i\|_{H(S_1)}$. Combining $\nabla\eta \equiv 0$ on $S_0 \cup S_1$ and the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} I_1 &\leq \|\nabla\eta\|_{L^\infty(D)}\|\psi_i\|_{L^2(S^*)}\left(\int_{S^*}\nabla\psi_i\cdot a\nabla\psi_i\right)^{1/2}\sqrt{a_{\max}} \\ &\leq \frac{1}{lh}\|\psi_i\|_{L^2(S^*)}\|\psi_i\|_{H(S^*)}\sqrt{a_{\max}}. \end{aligned} \quad (5.6)$$

We have used $c \geq 0$ to get $\left(\int_{S^*}\nabla\psi_i\cdot a\nabla\psi_i\right)^{1/2} \leq \|\psi_i\|_{H(S^*)}$ in the last inequality. By the construction of ψ_i (2.9), we have $\int_D \psi_i\varphi_j = 0$ for $i \neq j$. Thanks to (2.11), we have $-\nabla\cdot(a\nabla\psi_i) + c\psi_i \in \Phi$. Therefore, we have $\int_{S_1}\eta\psi_i(-\nabla\cdot(a\nabla\psi_i) + c\psi_i) = 0$. Denoting η_j as the volume average of η over τ_j, we have

[1] $\|\nabla\eta\|_{L^\infty(D)} := \operatorname*{ess\,sup}_{x\in D}|\nabla\eta(x)|$.

$$
\begin{aligned}
I_2 &= -\int_{S^*} \eta\psi_i(-\nabla\cdot(a\nabla\psi_i)+c\psi_i) = -\sum_{\tau_j\in S^*}\int_{\tau_j}(\eta-\eta_j)\psi_i(-\nabla\cdot(a\nabla\psi_i)+c\psi_i)\\
&\le \frac{1}{l}\sum_{\tau_j\in S^*}\|\psi_i\|_{L^2(\tau_j)}\|\mathcal{L}\psi_i\|_{L^2(\tau_j)}.
\end{aligned}
\tag{5.7}
$$

Up to now, I_1 and I_2 are some quantities of ψ_i purely on S^*, and we only need to prove that both of them can be bounded by $\|\psi_i\|^2_{H(S^*)}$ (up to a constant). By applying the Poincare inequality, we can easily do this for I_1, as we will see soon. However, I_2 involves the high-order term $\|\mathcal{L}\psi_i\|_{L^2(\tau_j)}$ which in general may not be bounded by the lower-order term $\|\psi_i\|_{H(S^*)}$. Fortunately, this can be proved since $\mathcal{L}\psi_i\in\Phi$, the piecewise constant function space. For the current operator $\mathcal{L}u=-\nabla\cdot(a(x)\nabla u)+c(x)u$ with rough coefficient a and nonzero potential c, Lemma 5.1 implies $\|\mathcal{L}\psi_i\|_{L^2(\tau_j)}\le\sqrt{a_{\max}}C(d,\delta)h^{-1}\|\psi_i\|_{H(\tau_j)}$ when $h\le h_0\equiv\pi\sqrt{\frac{a_{\max}}{2c_{\max}}}$. Then, we obtain

$$
I_2\le\frac{\sqrt{a_{\max}}C(d,\delta)}{lh}\|\psi_i\|_{L^2(S^*)}\|\psi_i\|_{H(S^*)}\quad \forall h\le h_0.
\tag{5.8}
$$

By the construction of ψ_i (2.9), we have $\int_{\tau_j}\psi_i=0$ for all $\tau_j\in S^*$. By the Poincare inequality, we have $\|\psi_i\|_{L^2(\tau_j)}\le\|\nabla\psi_i\|_{L^2(\tau_j)}h/\pi$, and then we obtain

$$
\|\psi_i\|^2_{H(S_1)}\le I_1+I_2\le\frac{1+C(d,\delta)}{\pi l}\sqrt{\frac{a_{\max}}{a_{\min}}}\|\psi_i\|^2_{H(S^*)}.
\tag{5.9}
$$

By taking $l\ge\frac{\mathrm{e}-1}{\pi}(1+C(d,\delta))\sqrt{\frac{a_{\max}}{a_{\min}}}$, we have the constant $\frac{1+C(d,\delta)}{\pi l}\sqrt{\frac{a_{\max}}{a_{\min}}}\le\frac{1}{\mathrm{e}-1}$. With the iterative argument given before, we have proved the exponential decay. □

Remark 5.1 We point out that boundary conditions may be important in several applications. For example, the Robin boundary condition is useful in the application of the SPCA. The periodic boundary condition is useful in compressing a Hamiltonian with a periodic boundary condition in quantum physics.

The above proof can be applied to the operator $\mathcal{L}$ in (5.1) with other boundary conditions as long as the corresponding problem $\mathcal{L}u=f$ has a unique solution $u\in H^k(D)$ for every $f\in L^2(D)$. For other homogeneous boundary condition, the Cameron–Martin space is not $H^1_0(D)$. Instead, we should use the solution space associated with the corresponding boundary condition. The proof of Theorem 5.1 can be easily carried over to other homogeneous boundary conditions, and the only difference is that a different boundary condition leads to slightly different integration by parts in (5.5). For the homogeneous Neumann boundary condition or the periodic boundary condition, the proof is exactly the same because the integration by parts (5.5) can be carried out in exactly the same way. For the problems with the Robin boundary condition, i.e.,

$$
\begin{aligned}
\mathcal{L}u &:= -\nabla\cdot(a(x)\nabla u(x))+c(x)u(x)=f(x) && x\in D,\\
&\frac{\partial u}{\partial n}+\alpha(x)u(x)=0 && x\in\partial D,
\end{aligned}
\tag{5.10}
$$

where $\alpha(x) \geq 0$, the Cameron–Martin space is the subspace of $H^1(D)$ in which all elements satisfy the Robin boundary condition and the associated energy norm is defined as

$$\|u\|_H^2 = \int_D \nabla u \cdot a \nabla u + \int_D cu^2 + \int_{\partial D} \alpha u^2. \tag{5.11}$$

In this case, for a subdomain $S \subset D$, the local energy norm on S should be modified as follows:

$$\|u\|_{H(S)}^2 = \int_S \nabla u \cdot a \nabla u + \int_S cu^2 + \int_{\partial D \cap \partial S} \alpha u^2. \tag{5.12}$$

Similarly, we can define the Cameron–Martin space and the associated energy norm for the homogeneous mixed boundary conditions.

6 Exponential decay of basis functions: the higher-order case

In this section, we will study the case when $\mathcal{K} : L^2(D) \to L^2(D)$ is the solution operator of the following higher-order elliptic equation:

$$\begin{aligned} &\mathcal{L}u := \sum_{0 \leq |\sigma|, |\gamma| \leq k} (-1)^{|\sigma|} D^\sigma (a_{\sigma\gamma}(x) D^\gamma u) = f, \\ &f \in L^2(D), \qquad u \in H_0^k(D). \end{aligned} \tag{6.1}$$

Here, we only consider the case when $\mathcal{L}$ (thus $\mathcal{K}$) is self-adjoint, i.e.,

$$\int_D (\mathcal{L}u)v = \int_D u(\mathcal{L}v) \qquad \forall u, v \in H_0^k(D). \tag{6.2}$$

The corresponding symmetric bilinear form on $H_0^k(D)$ is denoted as

$$B(u, v) = \sum_{0 \leq |\sigma|, |\gamma| \leq k} \int_D a_{\sigma\gamma}(x) D^\sigma u D^\gamma v. \tag{6.3}$$

We assume that B is an inner product on $H_0^k(D)$ and the induced norm $(B(u, u))^{1/2}$ is equivalent to the $H_0^k(D)$ norm, i.e., there exists $0 < a_{\min} \leq a_{\max}$ such that

$$a_{\min} |u|_{k,2,D}^2 \leq B(u, u) \leq a_{\max} |u|_{k,2,D}^2 \qquad \forall u \in H_0^k(D). \tag{6.4}$$

Thanks to the Riesz representation lemma, Eq. (6.1) has a unique weak solution in $H_0^k(D)$ for $f \in L^2(D)$.

6.1 Construction of basis functions and the approximation rate

Suppose D is divided into elements $\{\tau_i\}_{1 \leq i \leq m}$, where each element τ_i is a triangle or a quadrilateral in 2D, or a tetrahedron or hexahedron in 3D. Denote the maximum element diameter by h. We also assume that the subdivision is regular [6]. This means that if h_i denotes the diameter of τ_i and ρ_i denotes the maximum diameter of a ball inscribed in τ_i, there is a constant $\delta > 0$ such that

$$\frac{\rho_i}{h_i} \geq \delta \qquad \forall i = 1, 2, \ldots, m.$$

Applying Theorem 3.1 to $\Omega = \tau_j$, for any $u \in H^k(D)$ and any $0 \le p \le k$, we have

$$|u - \Pi_i u|_{p,2,\tau_i} \le C(k-1, \widehat{\tau}_i)\delta^{-p} h^{k-p} |u|_{k,2,\tau_i},$$

where $\Pi_i : H^k(\tau_i) \to \mathcal{P}_{k-1}(\tau_i)$ is the orthogonal projection to the polynomial space $\mathcal{P}_{k-1}(\tau_i)$ in $L^2(\tau_i)$, and $\widehat{\tau}_i$ is some reference domain that is affine equivalent to τ_i. Notice that the constant $C(k-1, \widehat{\tau}_i)\delta^{-p}$ can be bounded from above by a constant C_p for all the elements $\{\tau_i\}_{1 \le i \le m}$, because all elements in $\{\tau_i\}_{1 \le i \le m}$ are affine equivalent to an equilateral triangle or square in 2D, or a equilateral 3-simplex or cubic in 3D. Therefore, for any $u \in H^k(D)$, any $1 \le i \le m$ and any $0 \le p \le k$, we have

$$|u - \Pi_i u|_{p,2,\tau_i} \le C_p h^{k-p} |u|_{k,2,\tau_i}. \tag{6.5}$$

Specifically for $p = 0$, $\widetilde{u} \in L^2(D)$ with $\widetilde{u}|_{\tau_i} = \Pi_i u$, we conclude that

$$\|u - \widetilde{u}\|_{L^2(D)} \le C_p h^k |u|_{k,2,D}. \tag{6.6}$$

Let $X = L^2(D)$ and $H = H_0^k(D)$. We use the standard inner product for $L^2(D)$ and use the inner product $\langle u, v\rangle = B(u, v)$ for H. Further, we denote $\mathcal{K} : L^2(D) \to L^2(D)$ as the operator mapping f to the solution u in Eq. (6.1). Let $\{\varphi_{i,q}\}_{q=1}^{Q}$ be an orthogonal basis of $\mathcal{P}_{k-1}(\tau_i)$ with respect to the inner product in $L^2(\tau_i)$, where $Q = \binom{k+d-1}{d}$ is the number of d-variate monomials with degree at most $k-1$. We take

$$\Phi = \operatorname{span}\{\varphi_{i,q} : 1 \le q \le Q, 1 \le i \le m\}, \quad \Psi = \mathcal{K}\Phi. \tag{6.7}$$

Without loss of generality, we normalize these basis functions such that

$$\int_{\tau_i} \varphi_{i,q}\varphi_{i,q'} = |\tau_i|\delta_{q,q'}. \tag{6.8}$$

A set of basis functions of Ψ is defined by Eq. (2.9) accordingly, i.e.,

$$\begin{aligned} \psi_{i,q} = &\arg\min_{\psi \in H_0^k(D)} \quad \|\psi\|_H^2 \\ &\text{s.t.} \quad \int_D \psi_{i,q}\varphi_{j,q'} = \delta_{iq,jq'} \quad \forall 1 \le q' \le Q, \quad 1 \le j \le m. \end{aligned} \tag{6.9}$$

Combining Eqs. (6.4) and (6.6), we have

$$\|u - \mathcal{P}_{\Phi}^{(X)} u\|_{L^2(D)} \le \frac{C_p h^k}{\sqrt{a_{\min}}} \|u\|_H, \quad \forall u \in H. \tag{6.10}$$

Applying Theorem 2.1 with X and H defined above, we have

1. For any $u \in H$ and $\mathcal{L}u = f$, we have

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_H \le \frac{C_p h^k}{\sqrt{a_{\min}}} \|f\|_{L^2(D)}. \tag{6.11}$$

   Here, C_p plays the role of the Poincare constant $1/\pi$.

2. For any $u \in H$ and $\mathcal{L}u = f$, we have

$$\|u - \mathcal{P}_{\Psi}^{(H)} u\|_{L^2(D)} \le \frac{C_p^2 h^{2k}}{a_{\min}} \|f\|_{L^2(D)}. \tag{6.12}$$

3. We have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \le \frac{C_p^2 h^{2k}}{a_{\min}}. \tag{6.13}$$

Notice that the eigenvalues of the operator $\mathcal{L}$ (with the homogeneous Dirichlet boundary conditions) in (6.1) grow like $\lambda_n(\mathcal{L}) \sim n^{2k/d}$ (see, e.g., [7,32]), and thus, the eigenvalues of $\mathcal{K}$ decay like $\lambda_n(\mathcal{K}) \sim n^{-2k/d}$. Meanwhile, the rank of the operator $\mathcal{P}_{\Psi}^{(H)}\mathcal{K}$, denoted as n, roughly scales like Q/h^d where $1/h^d$ is roughly the number of patches. Plugging $n = Q/h^d$ into Eq. (6.13), we have

$$\|\mathcal{K} - \mathcal{P}_{\Psi}^{(H)} \mathcal{K}\| \le \frac{C_p^2 Q^{2k/d}}{a_{\min}} n^{-2k/d} \lesssim \lambda_n(\mathcal{K}). \tag{6.14}$$

Therefore, our construction of the m-dimensional subspace Ψ approximates $\mathcal{K}$ at the optimal rate. In Sect. 6.2, we introduce the concept of *strong ellipticity* that enables us to prove exponential decay results. In Sect. 6.4, we will prove that the basis functions $\psi_{i,q}$ defined in Eq. (6.9) have exponential decay away from τ_i.

6.2 The strong ellipticity condition

In our proof, we need the following *strong ellipticity condition* of the operator $\mathcal{L}$ to obtain the exponential decay.

Definition 6.1 An operator in the divergence form $\mathcal{L}u := \sum_{0 \le |\sigma|, |\gamma| \le k} (-1)^{|\sigma|} D^{\sigma} (a_{\sigma\gamma}(x) D^{\gamma} u)$ is strongly elliptic if there exists $\theta_{k,\min} > 0$ such that

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x) \boldsymbol{\zeta}_{\sigma} \boldsymbol{\zeta}_{\gamma} \ge \theta_{k,\min} \sum_{|\sigma|=k} \boldsymbol{\zeta}_{\sigma}^2 \quad \forall x \in D, \quad \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}, \tag{6.15}$$

where $\boldsymbol{\zeta}_{\sigma}$ and $\boldsymbol{\zeta}_{\gamma}$ are the σ'th and γ'th entry of $\boldsymbol{\zeta}$, respectively. One can check that $\binom{k+d-1}{k}$ is exactly the number of all possible kth derivatives, i.e., $\#\{D^{\sigma} u : |\sigma| = k\}$.

For a $2k$th-order partial differential operator $\mathcal{L}u = (-1)^k \sum_{|\alpha| \le 2k} a_{\alpha} D^{\alpha} u$, $\mathcal{L}$ is strongly elliptic if there exists a strongly elliptic operator in the divergence form $\widetilde{\mathcal{L}}$ such that $\mathcal{L}u = \widetilde{\mathcal{L}}u$ for all $u \in C^{2k}(D)$.

Remark 6.1 For a $2k$th-order partial differential operator $\mathcal{L}u = (-1)^k \sum_{|\alpha| \le 2k} a_{\alpha} D^{\alpha} u$, its divergence form may not be unique. It is possible that it has two divergence forms, and one does not satisfy the *strong ellipticity condition* (6.1) while the other does. For example, the biharmonic operator $\mathcal{L} = \Delta^2$ in two space dimensions have the following two different divergence forms:

$$\mathcal{L}u = \sum_{|\sigma|=|\gamma|=2} D^{\sigma} (a_{\sigma\gamma} D^{\gamma} u) = \sum_{|\sigma|=|\gamma|=2} D^{\sigma} (\widetilde{a}_{\sigma\gamma}(x) D^{\gamma} u), \tag{6.16}$$

where

$$(a_{\sigma\gamma}) = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad (\widetilde{a}_{\sigma\gamma}) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \tag{6.17}$$

when $\{D^\sigma u : |\sigma| = 2\}$ is ordered as $(\partial_{x_1}^2, \partial_{x_2}^2, \partial_{x_1}\partial_{x_2})$. Obviously, the first one does not satisfy the *strong ellipticity condition* (6.1) while the second one does. These two divergence forms correspond to two bilinear forms on $H_0^2(D)$:

$$B(u, v) = \int_D \Delta u \Delta v, \quad \widetilde{B}(u, v) = \int_D D^2 u : D^2 v, \tag{6.18}$$

where $D^2 u : D^2 v = \sum_{i,j} \frac{\partial^2 u}{\partial x_i \partial x_j} \frac{\partial^2 v}{\partial x_i \partial x_j}$.

The *strong ellipticity condition* guarantees that for any local subdomain $S \subset D$, the seminorm $|\cdot|_{k,2,S}$ can be controlled by the *local energy norm* $\|\cdot\|_{H(S)}$.

Lemma 6.1 *Suppose $\mathcal{L}u = \sum_{0\le|\sigma|,|\gamma|\le k} (-1)^{|\sigma|} D^\sigma (a_{\sigma\gamma}(x) D^\gamma u)$ is self-adjoint. Assume that $a_{\sigma\gamma}(x) \in L^\infty(D)$ for all $0 \le |\sigma|, |\gamma| \le k$ and that for any $x \in D$*

- *$\mathcal{L}$ is nonnegative, i.e.,*

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x) \boldsymbol{\zeta}_\sigma \boldsymbol{\zeta}_\gamma \ge 0 \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}, \tag{6.19}$$

- *$\mathcal{L}$ is bounded, i.e., there exist $\theta_{0,\max} \ge 0$ and $\theta_{k,\max} > 0$ such that*

$$\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x) \boldsymbol{\zeta}_\sigma \boldsymbol{\zeta}_\gamma \le \theta_{k,\max} \sum_{|\sigma|=k} \boldsymbol{\zeta}_\sigma^2 + \theta_{0,\max} \sum_{|\sigma|<k} \boldsymbol{\zeta}_\sigma^2 \qquad \forall x \in D \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}, \tag{6.20}$$

- *and $\mathcal{L}$ is strongly elliptic, i.e., there exists $\theta_{k,\min} > 0$ such that*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x) \boldsymbol{\zeta}_\sigma \boldsymbol{\zeta}_\gamma \ge \theta_{k,\min} \sum_{|\sigma|=k} \boldsymbol{\zeta}_\sigma^2 \qquad \forall x \in D \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}. \tag{6.21}$$

For any subdomain $S \subset D$ and any $\psi \in H^k(D)$, define

$$\|\psi\|_{H(S)}^2 = \sum_{0\le|\sigma|,|\gamma|\le k} \int_S a_{\sigma\gamma}(x) D^\sigma \psi D^\gamma \psi. \tag{6.22}$$

Then, the following two claims hold true.

- *If $\mathcal{L}$ contains only highest order terms, i.e., $\mathcal{L}u = \sum_{|\sigma|=|\gamma|=k} (-1)^{|\sigma|} D^\sigma (a_{\sigma\gamma}(x) D^\gamma u)$, then we have*

$$|\psi|_{k,2,S} \le \theta_{k,\min}^{-1/2} \|\psi\|_{H(S)} \qquad \forall \psi \in H^k(D). \tag{6.23}$$

- *If $\mathcal{L}$ contains low-order terms, for any regular domain partition $D = \cup_{i=1}^m \tau_i$ with diameter $h > 0$ satisfying $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta_{k,\min}^2}{16\theta_{0,\max}\theta_{k,\max}C_p^2}$, and any subdomain $S = \cup_{j\in\Lambda}\tau_j$, we have*

$$|\psi_{i,q}|_{k,2,S} \le \left(2/\theta_{k,\min}\right)^{1/2} \|\psi_{i,q}\|_{H(S)} \qquad \forall \tau_i \notin \mathcal{S}, \quad 1 \le q \le Q. \tag{6.24}$$

Here, Λ is any subset of $\{1, 2, \ldots, m\}$*, and $\psi_{i,q}$ is defined by Eq.* (6.9).

Proof The first point can be obtained directly from the definition of strong ellipticity. In the following, we provide the proof of the second point. For S stated in the second point and any $\psi \in H^k(D)$, we have

$$\|\psi\|_{H(S)}^2 = \underbrace{\sum_{|\sigma|=|\gamma|=k} \int_S a_{\sigma\gamma} D^\sigma \psi D^\gamma \psi}_{J_1} + \underbrace{\sum_{|\sigma|,|\gamma|<k} \int_S a_{\sigma\gamma} D^\sigma \psi D^\gamma \psi}_{J_2}$$
$$+ \underbrace{\sum_{|\sigma|=k,|\gamma|<k} \int_S (a_{\sigma\gamma} + a_{\gamma\sigma}) D^\sigma \psi D^\gamma \psi}_{J_3}. \tag{6.25}$$

From the strong ellipticity (6.21), we have

$$J_1 \ge \theta_{k,\min} |\psi|_{k,2,S}^2. \tag{6.26}$$

From the nonnegativity (6.19), we have

$$J_2 \ge 0. \tag{6.27}$$

Combining the nonnegativity (6.19) and the boundedness (6.20), we can prove that

$$\left| \sum_{|\sigma|=k,|\gamma|<k} (a_{\sigma\gamma} + a_{\gamma\sigma}) D^\sigma \psi D^\gamma \psi \right| \le 2 \left(\theta_{0,\max}\theta_{k,\max} \sum_{|\sigma|=k} |D^\sigma \psi|^2 \sum_{|\sigma|<k} |D^\sigma \psi|^2 \right)^{1/2}.$$

Therefore, using the Cauchy–Schwarz inequality, we obtain

$$|J_3| \le 2\theta_{0,\max}^{1/2}\theta_{k,\max}^{1/2} |\psi|_{k,2,S} \|\psi\|_{k-1,2,S}. \tag{6.28}$$

Thanks to the polynomial approximation property, for any $\tau_i \notin \mathcal{S}$ and $1 \le q \le Q$, we have

$$\|\psi_{i,q}\|_{k-1,2,S}^2 \le C_p^2 \frac{h^2(1-h^{2k})}{1-h^2} |\psi_{i,q}|_{k,2,S}^2. \tag{6.29}$$

Combining Eqs. (6.28) and (6.29), for $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta_{k,\min}^2}{16\theta_{0,\max}\theta_{k,\max}C_p^2}$, we have

$$|J_3| \le \frac{\theta_{k,\min}}{2} |\psi|_{k,2,S}^2. \tag{6.30}$$

Combining Eqs. (6.25), (6.26), (6.27) and (6.30), we prove the second point. □

Remark 6.2 When $\mathcal{L}$ contains low-order terms but there is no crossing term between $D^\sigma u$ ($|\sigma| = k$) and $D^\sigma u$ ($|\sigma| < k$), i.e., $J_3 = 0$, we can directly get the same bound in Eq. (6.23) for all $h > 0$.

The strong ellipticity condition above is different from the standard uniformly elliptic condition (see Definition 9.2 in [42]), i.e., a linear partial differential operator $\mathcal{L}u = (-1)^k \sum_{|\alpha|\le 2k} a_\alpha D^\alpha u$ is uniformly elliptic if there exists a constant $\theta_{k,\min} > 0$ such that

$$\sum_{|\alpha|=2k} a_\alpha(x)\boldsymbol{\xi}^\alpha \ge \theta_{k,\min}|\boldsymbol{\xi}|^{2k}, \quad \forall x \in D, \quad \boldsymbol{\xi} \in \mathbb{R}^d. \tag{6.31}$$

On the one hand, it is obvious that a strongly elliptic operator with smooth coefficients is uniformly elliptic, by taking $\boldsymbol{\zeta}_\sigma := \boldsymbol{\xi}^\sigma$ in Eq. (6.15). On the other hand, the relation between the uniform ellipticity and the strong ellipticity turns out to be closely related to the relation between nonnegative polynomials and sum-of-square (SOS) polynomials. In fact, the strongly ellipticity condition (6.15) is equivalent to that there exists $\theta_{k,\min} > 0$ such that

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\xi}^\sigma\boldsymbol{\xi}^\gamma - \theta_{k,\min}\sum_{|\sigma|=k}|\boldsymbol{\xi}|^{2k} = \text{Sum-Of-Squares (SOS) polynomials.}$$

Using the famous Hilbert's theorem (1888) on nonnegative polynomials and SOS polynomials, we have the following theorem. Readers can find the proof and more discussions in [48].

Theorem 6.2 *Let $a_\alpha \in C^{|\alpha|-k}(\bar{D})$ for $k < |\alpha| \le 2k$, $a_\alpha \in C(\bar{D})$ for $|\alpha| \le k$, and $\mathcal{L}u = (-1)^k \sum_{|\alpha|\le 2k} a_\alpha D^\alpha u$ for all $u \in C^{2k}(D)$. Then in the following two cases, if $\mathcal{L}$ is uniformly elliptic it is also strongly elliptic.*

- *$d = 1$ or 2 : one- or two-dimensional physical domain,*
- *$k = 1$: second-order partial differential operators.*

For the case $(d, k) = (3, 2)$, i.e., fourth-order partial differential operators in 3-dimensional physical domain, all uniformly elliptic operators with constant coefficients are also strongly elliptic.

For the case $(d, k) = (3, 2)$, we are not able to prove that strong ellipticity is equivalent to uniform ellipticity for elliptic operators with smooth and multiscale coefficients, but we suspect that it is true. For all other cases, there are uniformly but not strongly elliptic operators. Fortunately, for small physical dimensions d and differential orders k, strongly elliptic operators approximate uniformly elliptic operators well and counter examples are difficult to construct.

6.3 Exponential decay of basis functions I

In this subsection, we prove the exponential decay of basis functions constructed in Eq. (6.9) for higher-order elliptic operators that contain only the highest order terms. We will leave the proof for the general operators to the next subsection. The proof follows exactly the same structure as that in the second-order elliptic case.

Theorem 6.3 *Let $\mathcal{L}u = (-1)^k \sum_{|\sigma|=|\gamma|=k} D^\sigma(a_{\sigma\gamma}D^\gamma u)$ and $a_{\sigma\gamma}(x) \in L^\infty(D)$ for all $|\sigma| = |\gamma| = k$. Assume that for any $x \in D$*

- *$\mathcal{L}$ is bounded, i.e., there exist nonnegative $\theta_{k,\max}$ such that*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \le \theta_{k,\max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 \qquad \forall\boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}, \tag{6.32}$$

- *and $\mathcal{L}$ is strongly elliptic, i.e., there exists $\theta_{k,\min} > 0$ such that*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \ge \theta_{k,\min}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 \qquad \forall\boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}. \tag{6.33}$$

Then for any $1 \le i \le m$ and $1 \le q \le Q$, it holds true that

$$\|\psi_{i,q}\|^2_{H(D\cap(B(x_i,r))^c)} \le \exp\left(1 - \frac{r}{lh}\right)\|\psi_{i,q}\|^2_{H(D)} \tag{6.34}$$

with $\sqrt{l^2-1} \ge (e-1)C_\eta C_p(C_1 + C(k,d,\delta))\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}$. Here, C_1 and C_η only depends on k and d, C_p is the constant in Eq. (6.5) and $C(k,d,\delta) := C(k,k,d,\delta)$ from Lemma 3.1.

Proof The proof follows the same structure as that of Theorem 5.1 and [36] (Thm. 3.9). Let $k \in \mathbb{N}$, $l > 0$ and $i \in \{1, 2, \ldots, m\}$. Let S_0 be the union of all the domains τ_j that are contained in the closure of $B(x_i, klh) \cap D$, let S_1 be the union of all the domains τ_j that are not contained in the closure of $B(x_i, (k+1)lh) \cap D$ and let $S^* = S_0^c \cap S_1^c \cap D$ (be the union of all the remaining elements τ_j not contained in S_0 or S_1). In the following, we will prove that for any $k \ge 1$, there exists constant C such that $\|\psi_{i,q}\|^2_{H(S_1)} \le C\|\psi_{i,q}\|^2_{H(S^*)}$. Then, the same recursive argument in the proof of Theorem 5.1 can be used to prove the exponential decay.

Let $\eta(x)$ be a smooth function which satisfies (1) $0 \le \eta \le 1$, (2) $\eta|_{B(x_i,klh)} = 0$, (3) $\eta|_{B^c(x_i,(k+1)lh)} = 1$ and (4) $\|D^\sigma\eta\|_{L^\infty(D)} \le \frac{C_\eta}{(lh)^{|\sigma|}}$ for all σ.

By integration by parts, we have

$$\int_D \eta\psi_{i,q}\mathcal{L}\psi_{i,q} = \sum_{|\sigma|=|\gamma|=k}\int_D a_{\sigma\gamma}(x)D^\sigma(\eta\psi_{i,q})D^\gamma\psi_{i,q}.$$

Making use of the binomial theorem $D^\sigma(\eta\psi_{i,q}) = \eta D^\sigma\varphi_{i,q} + \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}$, we obtain

$$\sum_{|\sigma|=|\gamma|=k}\int_D \eta a_{\sigma\gamma}(x)D^\sigma(\psi_{i,q})D^\gamma\psi_{i,q} = \underbrace{\int_D \eta\psi_{i,q}\mathcal{L}\psi_{i,q}}_{I_2} \underbrace{- \sum_{|\sigma|=|\gamma|=k}\sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_D a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma\psi_{i,q}}_{I_1}. \tag{6.35}$$

Since $\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)D^\sigma\psi_{i,q}D^\gamma\psi_{i,q} \ge 0$ for every $x \in D$, the left-hand side gives an upper bound for $\|\psi_{i,q}\|^2_{H(S^1)}$. Since $D^{\sigma_1}\eta = 0$ ($|\sigma_1| \ge 1$) on both S_0 and S_1, we obtain

$$I_1 = -\sum_{|\sigma|=|\gamma|=k}\sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*} a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma\psi_{i,q} \tag{6.36}$$

$$\le \left(\sum_{|\sigma|=k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,\max}} \tag{6.37}$$

$$\le C_1 C_\eta \left(\sum_{s'=1}^{k} (lh)^{-2s'} |\psi_{i,q}|^2_{k-s',2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,\max}}. \tag{6.38}$$

Here, C_1 is a constant only dependent on k and d. We have used the Cauchy–Schwarz inequality and the bound (6.32) in Eq. (6.37). We will defer the proof of the last step in Eq. (6.38) to the "Appendix." Since $\psi_{i,q} \perp \mathcal{P}_{k-1}$ locally in L^2, we obtain from Theorem 3.1 that

$$|\psi_{i,q}|_{k-s',2,S^*} \le C_p h^{s'} |\psi_{i,q}|_{k,2,S^*}.$$

Therefore, we get

$$I_1 \le C_1 C_\eta \sqrt{\theta_{k,\max}} C_p \left(\sum_{s'=1}^{k} l^{-2s'} |\psi_{i,q}|^2_{k,2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \tag{6.39}$$

$$\le \frac{C_1 C_\eta \sqrt{\theta_{k,\max}} C_p}{\sqrt{l^2-1}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{6.40}$$

In the last inequality, we have used $\sum_{s'=1}^{k} l^{-2s'} = \frac{1-l^{-2k}}{l^2-1} \le \frac{1}{l^2-1}$.

By the construction of $\psi_{i,q}$ given in (6.9), we have $\int_D \psi_{i,q}\varphi_{j,q'} = 0$ for $i \ne j$. Thanks to (2.11), we have $\mathcal{L}\psi_{i,q} \in \Phi$. Therefore, we get $\int_{S_1} \eta\psi_{i,q}\mathcal{L}\psi_{i,q} = 0$. Denoting η_j as the volume average of η over τ_j, we obtain

$$I_2 = \int_{S^*} \eta\psi_{i,q}\mathcal{L}\psi_{i,q} = \sum_{\tau_j\in S^*} \int_{\tau_j} (\eta-\eta_j)\psi_{i,q}\mathcal{L}\psi_{i,q} \le \frac{C_\eta}{l} \sum_{\tau_j\in S^*} \|\psi_{i,q}\|_{L^2(\tau_j)} \|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)}. \tag{6.41}$$

By using Lemma 6.2, which is stated in the beginning of Sect. 6.5, we have $\|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)} \le \sqrt{\theta_{k,\max}} C(k,d,\delta) h^{-k} \|\psi_{i,q}\|_{H(\tau_j)}$ for any $h > 0$ because $\mathcal{L}$ contains only the highest order derivatives. Then, we obtain

$$I_2 \le \frac{\sqrt{\theta_{k,\max}} C_\eta C(k,d,\delta)}{lh^k} \|\psi_{i,q}\|_{L^2(S^*)} \|\psi_{i,q}\|_{H(S^*)}$$

$$\le \frac{\sqrt{\theta_{k,\max}} C_\eta C(k,d,\delta) C_p}{l} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}, \tag{6.42}$$

where we have used Eq. (6.5) in the last step.

Combining Eqs. (6.40) and (6.42), we obtain

$$I_1 + I_2 \le \sqrt{\frac{\theta_{k,\max}}{l^2-1}} C_\eta C_p (C_1 + C(k,d,\delta)) |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}.$$

By the strong ellipticity (6.33) and Eq. (6.23), we have $|\psi_{i,q}|_{k,2,S^*} \le \theta_{k,\min}^{-1/2} \|\psi_{i,q}\|_{H(S^*)}$. Therefore, we have

$$\|\psi_{i,q}\|^2_{H(S^1)} \le \sqrt{\frac{\theta_{k,\max}}{(l^2-1)\theta_{k,\min}}} C_\eta C_p (C_1 + C(k,d,\delta)) \|\psi_{i,q}\|^2_{H(S^*)}. \tag{6.43}$$

By taking $\sqrt{l^2-1} \geq (e-1)C_\eta C_p(C_1 + C(k,d,\delta))\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}$, the exponential decay naturally follows. □

6.4 Exponential decay of basis functions II

The following theorem gives the exponential decay property of $\psi_{i,q}$ for an operator $\mathcal{L}$ with lower-order terms. Similar to the proof of Theorem 6.4, we need the polynomial approximation property (6.5) and the Friedrichs' inequality (3.4) to bound the lower-order terms, and we get an extra factor of 2 in our error bound.

Theorem 6.4 *Suppose $\mathcal{L}u = \sum_{0\leq|\sigma|,|\gamma|\leq k}(-1)^{|\sigma|}D^\sigma(a_{\sigma\gamma}(x)D^\gamma u)$ is self-adjoint. Assume that $a_{\sigma\gamma}(x) \in L^\infty(D)$ for all $0 \leq |\sigma|, |\gamma| \leq k$ and that for any $x \in D$*

- *$\mathcal{L}$ is nonnegative, i.e.,*

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \geq 0, \qquad \forall x \in D \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}, \tag{6.44}$$

- *$\mathcal{L}$ is bounded, i.e., there exist $\theta_{0,\max} \geq 0$ and $\theta_{k,\max} > 0$ such that*

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \leq \theta_{k,\max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 + \theta_{0,\max}\sum_{|\sigma|<k}\boldsymbol{\zeta}_\sigma^2 \qquad \forall x \in D, \quad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}, \tag{6.45}$$

- *and $\mathcal{L}$ is strongly elliptic, i.e., there exists $\theta_{k,\min} > 0$ such that*

$$\sum_{|\sigma|=|\gamma|=k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \geq \theta_{k,\min}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2, \qquad \forall \boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d-1}{k}}. \tag{6.46}$$

Then there exists $h_0 > 0$ such that for any $h \leq h_0$, $1 \leq i \leq m$ and $1 \leq q \leq Q$, it holds true that

$$\|\psi_{i,q}\|^2_{H(D\cap(B(x_i,r))^c)} \leq \exp\left(1 - \frac{r}{lh}\right)\|\psi_{i,q}\|^2_{H(D)} \tag{6.47}$$

with $\sqrt{l^2-1} \geq 2(e-1)C_\eta C_p(C_1 + C(k,d,\delta))\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}$. Here, C_1 and C_η depend on k and d only, C_p is the constant given in Eq. (6.5), $C(k,d,\delta) := C(k,k,d,\delta)$ is given in Lemma 4.1 and $\theta_{k,\max} := \max(\theta_{0,\max}, \theta_{k,\max})$. The constant h_0 can be taken as

$$\begin{aligned} h_0 = \sup\Bigg\{h > 0 : \frac{h^2 - h^{2k}}{1-h^2} \leq \frac{1}{C_p^2}, \frac{h^2(1-h^{2k})}{1-h^2} \\ \leq \min\left(\frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}, \frac{\theta_{k,\min}^2}{16\theta_{0,\max}\theta_{k,\max}C_p^2}\right)\Bigg\}, \end{aligned}$$

where C_f is the constant in the Friedrichs' inequality (3.4).

Proof The proof follows the same structure as the proof of Theorem 6.3. All we need to do is to use the polynomial approximation property (6.5) and the Friedrichs' inequality (3.4) to bound the lower-order terms when they appear. First, the I_1 in Eq. (6.35) contains all the lower-order terms and its estimation should be modified as follows:

$$I_1 = -\sum_{0\le|\sigma|,|\gamma|\le k}\ \sum_{\substack{\sigma_1+\sigma_2=\sigma\\ |\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*} a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^{\gamma}\psi_{i,q} \tag{6.48}$$

$$\le \left(\sum_{|\sigma|\le k}\int_{S^*}\left|\sum_{\substack{\sigma_1+\sigma_2=\sigma\\ |\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}\right|^2\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,\max}} \tag{6.49}$$

$$\le C_1C_\eta\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}(lh)^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,\max}}. \tag{6.50}$$

Here, $\theta_{k,\max} := \max(\theta_{0,\max}, \theta_{k,\max})$. We have used the Cauchy–Schwarz inequality and the bound (6.45) in Eq. (6.49). We will defer the proof of the last step in Eq. (6.50) to the "Appendix." Since $\psi_{i,q} \perp \mathcal{P}_{k-1}$ locally in L^2, we obtain from Theorem 3.1 that

$$|\psi_{i,q}|_{s-s',2,S^*} \le C_p h^{s'}|\psi_{i,q}|_{s,2,S^*} \quad \forall\, 0 \le s' \le s \le k.$$

Therefore, we have

$$I_1 \le C_1C_\eta\sqrt{\theta_{k,\max}}C_p\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}l^{-2s'}|\psi_{i,q}|^2_{s,2,S^*}\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)} \tag{6.51}$$

$$\le \frac{C_1C_\eta\sqrt{\theta_{k,\max}}C_p}{\sqrt{l^2-1}}\left(\sum_{s=1}^{k}|\psi_{i,q}|^2_{s,2,S^*}\right)^{1/2}\|\psi_{i,q}\|_{H(S^*)} \tag{6.52}$$

$$\le \frac{C_1C_\eta\sqrt{2\theta_{k,\max}}C_p}{\sqrt{l^2-1}}|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}\|_{H(S^*)}. \tag{6.53}$$

If we compare the above estimate with Eq. (6.40), we conclude that Eq. (6.52) contains all the lower-order terms. We will use the polynomial approximation property (6.5) and take $\frac{h^2-h^{2k}}{1-h^2} \le 1/C_p^2$ to guarantee that Eq. (6.53) is valid. When $\mathcal{L}$ contains lower-order terms, by Lemma 6.2, we have $\|\mathcal{L}\psi_{i,q}\|_{L^2(\tau_j)} \le \sqrt{2\theta_{k,\max}}C(k,d,\delta)h^{-k}\|\psi_{i,q}\|_{H(\tau_j)}$ for any $h > 0$ satisfying $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}$. Therefore, using Eq. (6.42) we get

$$I_2 \le \frac{\sqrt{2\theta_{k,\max}}C_\eta C(k,d,\delta)C_p}{l}|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}\|_{H(S^*)}, \tag{6.54}$$

when h satisfies $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}$. Finally, we need to use Eq. (6.24) instead of Eq. (6.23) to bound $|\psi_{i,q}|_{k,2,S^*}$. We get

$$\|\psi_{i,q}\|^2_{H(S^1)} \le 2\sqrt{\frac{\theta_{k,\max}}{(l^2-1)\theta_{k,\min}}}C_\eta C_p(C_1 + C(k,d,\delta))\|\psi_{i,q}\|^2_{H(S^*)}, \tag{6.55}$$

where we have imposed another condition on h, i.e., $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta^2_{k,\min}}{16\theta_{0,\max}\theta_{k,\max}C_p^2}$. By taking $\sqrt{l^2-1} \ge 2(e-1)C_\eta C_p(C_1 + C(k,d,\delta))\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}$, we prove the exponential decay. □

Remark 6.3 As we have pointed out in Remark 6.2, when $\mathcal{L}$ contains low-order terms but there is no crossing term between $D^\sigma u$ ($|\sigma| = k$) and $D^\sigma u$ ($|\sigma| < k$), Eq. (6.23) can be used to bound $|\psi_{i,q}|_{k,2,S^*}$. In this case, the constraint on l is

$$\sqrt{l^2-1} \geq \sqrt{2}(e-1)C_\eta C_p(C_1 + C(k,d,\delta))\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}$$

and the h_0 can be taken as

$$h_0 = \sup\left\{h > 0 : \frac{h^2 - h^{2k}}{1-h^2} \leq \frac{1}{C_p^2}, \frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}\right\}.$$

6.5 Lemmas

In this subsection, we will prove the following lemma, which is used in the proof of Theorem 6.3 and Theorem 6.4.

Lemma 6.2 *$\mathcal{L}$ is defined in Eq. (6.1) and the space Ψ is defined as above. Assume that for any $x \in D$*

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \leq \theta_{k,\max}\sum_{|\sigma|=k}\boldsymbol{\zeta}_\sigma^2 + \theta_{0,\max}\sum_{|\sigma|<k}\boldsymbol{\zeta}_\sigma^2 \qquad \forall\boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}. \tag{6.56}$$

Let C_f be the constant in the Friedrichs' inequality (3.4). Then for any domain partition with $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}$, we have

$$\|\mathcal{L}v\|_{L^2(\tau_j)} \leq \sqrt{2\theta_{k,\max}}C(k,d,\delta)h^{-k}\|v\|_{H(\tau_j)} \quad \forall v \in \Psi,\ \forall j = 1,2,\ldots,m, \tag{6.57}$$

where $C(k,d,\delta) = C(k,k,d,\delta)$ from Lemma 4.1.

If the operator $\mathcal{L}$ contains only the highest order terms, i.e., $\mathcal{L}u = (-1)^k\sum_{|\sigma|=|\gamma|=k}D^\sigma$ $(a_{\sigma\gamma}D^\gamma u)$, we have $\|\mathcal{L}v\|_{L^2(\tau_j)} \leq \sqrt{\theta_{k,\max}}C(k,d,\delta)h^{-k}\|v\|_{H(\tau_j)}$ for all $h > 0$.

We will use Lemma 4.1 to prove this result, but we need to deal with the variable coefficients $a_{\sigma\gamma}$ and the low-order terms $a_{\sigma\gamma}$ with $|\sigma| + |\gamma| < 2k$ before we can apply Lemma 4.1. Our strategy is to transfer the variable coefficients to constant ones by the variational formulation (see Lemma 6.3) and to use the polynomial approximation property to deal with the low-order terms; see Lemma 6.4. For this purpose, we first introduce the following two lemmas.

Lemma 6.3 *Let Ω be a smooth, bounded, open subset of $\mathbb{R}^d$. $\mathcal{L}u = \sum_{0\leq|\sigma|,|\gamma|\leq k}(-1)^{|\sigma|}D^\sigma$ $(a_{\sigma\gamma}(x)D^\gamma u)$ and $\mathcal{M}u = \sum_{0\leq|\sigma|,|\gamma|\leq k}(-1)^{|\sigma|}D^\sigma(b_{\sigma\gamma}(x)D^\gamma u)$ are two symmetric operators on $H_0^k(\Omega)$. Moreover, we assume that the bilinear forms induced by both $\mathcal{L}$ and $\mathcal{M}$ are equivalent to the standard norm on $H_0^k(\Omega)$. Let $G_\mathcal{L}$ and $G_\mathcal{M}$ be the Green's functions of $\mathcal{L}$ and $\mathcal{M}$, respectively. If for any $x \in D$ we have*

$$\sum_{0\leq|\sigma|,|\gamma|\leq k} a_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma \leq \sum_{0\leq|\sigma|,|\gamma|\leq k} b_{\sigma\gamma}(x)\boldsymbol{\zeta}_\sigma\boldsymbol{\zeta}_\gamma, \qquad \forall\boldsymbol{\zeta} \in \mathbb{R}^{\binom{k+d}{k}}. \tag{6.58}$$

then for all $f \in L^2(\Omega)$,

$$\int_\Omega\int_\Omega G_\mathcal{M}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y \leq \int_\Omega\int_\Omega G_\mathcal{L}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y. \tag{6.59}$$

Proof Let $f \in L^2(\Omega)$. Let $\psi_{\mathcal{L}}$ and $\psi_{\mathcal{M}}$ be the weak solutions of $\mathcal{L}\psi_{\mathcal{L}} = f$ and $\mathcal{M}\psi_{\mathcal{M}} = f$ with the homogeneous Dirichlet boundary conditions on $\partial\Omega$. Observe that $\psi_{\mathcal{L}}$ and $\psi_{\mathcal{M}}$ are the unique minimizers of $I_{\mathcal{L}}(u,f)$ and $I_{\mathcal{M}}(u,f)$ with

$$\begin{aligned} I_{\mathcal{L}}(u,f) &= \frac{1}{2}\sum_{0\leq|\sigma|,|\gamma|\leq k}\int_D a_{\sigma\gamma}(x)D^{\sigma}uD^{\gamma}u - \int_{\Omega} uf, \quad u \in H_0^k(\Omega), \\ I_{\mathcal{M}}(u,f) &= \frac{1}{2}\sum_{0\leq|\sigma|,|\gamma|\leq k}\int_D b_{\sigma\gamma}(x)D^{\sigma}uD^{\gamma}u - \int_{\Omega} uf, \quad u \in H_0^k(\Omega). \end{aligned} \tag{6.60}$$

At the minima $\psi_{\mathcal{L}}$ and $\psi_{\mathcal{M}}$, we have

$$\begin{aligned} I_{\mathcal{L}}(\psi_{\mathcal{L}},f) &= -\frac{1}{2}\int_{\Omega}\psi_{\mathcal{L}}f = -\frac{1}{2}\sum_{0\leq|\sigma|,|\gamma|\leq k}\int_D a_{\sigma\gamma}(x)D^{\sigma}\psi_{\mathcal{L}}D^{\gamma}\psi_{\mathcal{L}}, \\ I_{\mathcal{M}}(\psi_{\mathcal{M}},f) &= -\frac{1}{2}\int_{\Omega}\psi_{\mathcal{M}}f = -\frac{1}{2}\sum_{0\leq|\sigma|,|\gamma|\leq k}\int_D a_{\sigma\gamma}(x)D^{\sigma}\psi_{\mathcal{M}}D^{\gamma}\psi_{\mathcal{M}}. \end{aligned} \tag{6.61}$$

Observe that

$$I_{\mathcal{L}}(\psi_{\mathcal{L}},f) \leq I_{\mathcal{L}}(\psi_{\mathcal{M}},f) \leq I_{\mathcal{M}}(\psi_{\mathcal{M}},f), \tag{6.62}$$

where the first inequality is true because $\psi_{\mathcal{L}}$ is the minimizer of $I_{\mathcal{L}}$, and the second inequality is true because $I_{\mathcal{L}}(u,f) \leq I_{\mathcal{M}}(u,f)$ for any $u \in H_0^k(\Omega)$. Combining Eq. (6.61) and (6.62), we obtain $\int_{\Omega}\psi_{\mathcal{M}}f \leq \int_{\Omega}\psi_{\mathcal{L}}f$. This proves the lemma. □

Lemma 6.4 *Let Ω_h be a smooth, convex, bounded, open subset of $\mathbb{R}^d$ with diameter at most h. Let G_h be the Green's function of $\mathcal{L}u = (-1)^k\sum_{|\sigma|=k}D^{2\sigma}u + c\sum_{|\sigma|<k}(-1)^{\sigma}D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\Omega_h$ and $G_{h,0}$ be the Green's function of $\mathcal{L}_0u = (-1)^k\sum_{|\sigma|=k}D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\Omega_h$. Here, $c > 0$ is a positive constant. Then, for any $f \in L^2(\Omega_h)$*

$$\lim_{h\to 0}\frac{\int_{\Omega_h}\int_{\Omega_h}G_h(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h}G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y} = 1. \tag{6.63}$$

Moreover, $\frac{\int_{\Omega_h}\int_{\Omega_h}G_h(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h}G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y} \geq 1/2$ for all $h > 0$ such that $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{1}{2cC_f^2}$.

Proof Let ψ_h be the solution of $\mathcal{L}\psi_h = f$ with the homogeneous Dirichlet boundary conditions on $\partial\Omega_h$ and $\psi_{h,0}$ be the solution of $\mathcal{L}_0\psi_{h,0} = f$ with the homogeneous Dirichlet boundary conditions on $\partial\Omega_h$. Let

$$\begin{aligned} I_{\mathcal{L}}(u,f) &= \frac{1}{2}|u|_{k,2,\Omega_h}^2 + \frac{c}{2}\|u\|_{k-1,2,\Omega_h}^2 - \int_{\Omega_h} uf, \\ I_{\mathcal{L}_0}(u,f) &= \frac{1}{2}|u|_{k,2,\Omega_h}^2 - \int_{\Omega_h} uf. \end{aligned} \tag{6.64}$$

At the minima ψ_h and $\psi_{h,0}$, we have

$$\begin{aligned} I_{\mathcal{L}}(\psi_h,f) &= -\frac{1}{2}\int_{\Omega_h} \psi_h f = -\frac{1}{2}\left(|\psi_h|^2_{k,2,\Omega_h} + c\|\psi_h\|^2_{k-1,2,\Omega_h}\right), \\ I_{\mathcal{L}_0}(\psi_{\mathcal{L}_0},f) &= -\frac{1}{2}\int_{\Omega_h} \psi_{h,0} f = -\frac{1}{2}|\psi_{h,0}|^2_{k,2,\Omega_h}. \end{aligned} \tag{6.65}$$

Note that Eq. (6.65) implies that $I_{\mathcal{L}_0}(\psi_{h,0},f) < 0$. By the definition of Green's function, we further have

$$\begin{aligned} \int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y &= \int_{\Omega_h} \psi_h f = -2I_{\mathcal{L}}(\psi_h,f) = |\psi_h|^2_{k,2,\Omega_h} + c\|\psi_h\|^2_{k-1,2,\Omega_h}, \\ \int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y &= \int_{\Omega_h} \psi_{h,0} f = -2I_{\mathcal{L}_0}(\psi_{h,0},f) = |\psi_{h,0}|^2_{k,2,\Omega_h}. \end{aligned} \tag{6.66}$$

Since $I_{\mathcal{L}_0}(u,f) \le I_{\mathcal{L}}(u,f)$ for any $u \in H_0^k(\Omega)$, we have $\frac{\int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y} \le 1$ for any $h > 0$. Applying the Friedrich's inequality (3.4) to $\|\psi_{h,0}\|^2_{k-1,2,\Omega_h}$, we get

$$\begin{aligned} -2I_{\mathcal{L}}(\psi_{h,0},f) &\ge -2I_{\mathcal{L}_0}(\psi_{h,0},f) - \frac{cC_f^2h^2(1-h^{2k})}{1-h^2}|\psi_{h,0}|^2_{k,2,\Omega_h} \\ &= -2\left(1 - \frac{cC_f^2h^2(1-h^{2k})}{1-h^2}\right))I_{\mathcal{L}_0}(\psi_{h,0},f). \end{aligned}$$

Here, we have used Eq. (6.66) in the last equality. Therefore, we have

$$\frac{\int_{\Omega_h}\int_{\Omega_h} G_h(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y}{\int_{\Omega_h}\int_{\Omega_h} G_{h,0}(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y} = \frac{-2I_{\mathcal{L}}(\psi_h,f)}{-2I_{\mathcal{L}_0}(\psi_{h,0},f)} \ge \frac{-2I_{\mathcal{L}}(\psi_{h,0},f)}{-2I_{\mathcal{L}_0}(\psi_{h,0},f)} \ge 1 - \frac{cC_f^2h^2(1-h^{2k})}{1-h^2},$$

where we have used $I_{\mathcal{L}}(\psi_h,f) \le I_{\mathcal{L}}(\psi_{h,0},f)$ in the first inequality. By using the above upper bound, we prove the lemma. □

Now, we are ready to prove Lemma 6.2.

Proof of Lemma 6.2 Let $v = \sum_{i=1}^m \sum_{q=1}^Q c_{i,q}\psi_{i,q}$. Thanks to Eq. (2.11), we have

$$\mathcal{L}v = \sum_{i,q}\sum_{j,q'} c_{i,q}\Theta^{-1}_{iq,jq'}\varphi_{j,q'}.$$

Let $g_j = \sum_{q'=1}^Q \sum_{i,q} c_{i,q}\Theta^{-1}_{iq,jq'}\varphi_{j,q'}$. Due to the construction of $\varphi_{j,q'}$, we have

$$\|\mathcal{L}v\|^2_{L^2(\tau_j)} = \|g_j\|^2_{L^2(\tau_j)} \tag{6.67}$$

Furthermore, v can be decomposed over τ_j as $v = v_1 + v_2$, where v_1 solves $\mathcal{L}v_1 = g_j(x)$ in τ_j with $v_1 \in H_0^k(\tau_j)$, and v_2 solves $\mathcal{L}v_2 = 0$ with $v_2 - v \in H_0^k(\tau_j)$. It is easy to check that $\|v\|^2_{H(\tau_j)} = \|v_1\|^2_{H(\tau_j)} + \|v_2\|^2_{H(\tau_j)}$. We denote G_j as the Green's function of the operator $\mathcal{L}$ with the homogeneous Dirichlet boundary condition on τ_j, then

$$\|v_1\|^2_{H(\tau_j)} = \int_{\tau_j} v_1(x)g_j\mathrm{d}x = \int_{\tau_j}\int_{\tau_j} G_j(x,y)g_j(x)g_j(y)\mathrm{d}x\,\mathrm{d}y.$$

Thanks to Lemma 6.3, we have

$$\|v_1\|^2_{H(\tau_j)} \geq \frac{1}{\theta_{k,\max}} \int_{\tau_j}\int_{\tau_j} G_j^*(x,y)g_j(x)g_j(y)\mathrm{d}x\,\mathrm{d}y, \tag{6.68}$$

where G_j^* is the Green's function of the operator $(-1)^k\sum_{|\sigma|=k} D^{2\sigma}u + \frac{\theta_{k,\max}}{\theta_{0,\max}}\sum_{|\sigma|<k}(-1)^\sigma D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\tau_j$. Thanks to Lemma 6.4, for all $h>0$ such that $\frac{h^2(1-h^{2k})}{1-h^2} \leq \frac{\theta_{k,\max}}{2\theta_{0,\max}C_f^2}$ we have

$$\int_{\tau_j}\int_{\tau_j} G_j^*(x,y)g_j(x)g_j(y)\mathrm{d}x\,\mathrm{d}y \geq \frac{1}{2}\int_{\tau_j}\int_{\tau_j} G_{j,0}^*(x,y)g_j(x)g_j(y)\mathrm{d}x\,\mathrm{d}y, \tag{6.69}$$

where $G_{j,0}^*$ is the Green's function of the operator $(-1)^k\sum_{|\sigma|=k} D^{2\sigma}u$ with the homogeneous Dirichlet boundary condition on $\partial\tau_j$. Denote $v_{1,0}$ as the solution of $(-1)^k\sum_{|\sigma|=k} D^{2\sigma}v_{1,0} = g_j$ on τ_j with the homogeneous Dirichlet boundary condition, i.e., $v_{1,0}(x) = \int_{\tau_j} G_{j,0}^*(x,y)g_j(y)\mathrm{d}y$. Since $g_j \in \mathcal{P}_{k-1}$ in τ_j in this case, Lemma 4.1 shows that

$$\|g_j\|^2_{L^2(\tau_j)} \leq (C(k,k,d,\delta))^2\, h^{-2}\int_{\tau_j}\int_{\tau_j} G_{j,0}^*(x,y)g_j(x)g_j(y)\mathrm{d}x\,\mathrm{d}y. \tag{6.70}$$

Combining Eqs. (6.68), (6.69) and (6.70), we have

$$\|g_j\|^2_{L^2(\tau_j)} \leq 2\,(C(k,k,d,\delta))^2\, h^{-2k}\theta_{k,\max}\|v_1\|^2_{H(\tau_j)} \leq 2\,(C(k,k,d,\delta))^2\, h^{-2k}\theta_{k,\max}\|v\|^2_{H(\tau_j)}.$$

Therefore, we have proved Lemma 6.2. We point out that when the operator $\mathcal{L}$ contains only the highest order terms, i.e., $\mathcal{L}u = (-1)^k\sum_{|\sigma|=|\gamma|=k} D^\sigma(a_{\sigma\gamma}D^\gamma u)$, we don't need to pay a factor of 2 in Eq. (6.69), and thus, $\|g_j\|^2_{L^2(\tau_j)} \leq (C(k,k,d,\delta))^2\, h^{-2k}\theta_{k,\max}\|v\|^2_{H(\tau_j)}$ for all $h>0$ in this special case. □

Let $\mathcal{L}_0^{-1}f \in H_0^k(\tau_i)$ be the unique weak solution of the following elliptic equation with the homogeneous Dirichlet boundary condition

$$\mathcal{L}u = f(x) \qquad x\in\tau_i, \quad u \in H_0^k(\tau_i). \tag{6.71}$$

We define $M_0, A_0 \in \mathbb{R}^{Q\times Q}$ as follows:

$$M_0(q,q') = \int_{\tau_i}\varphi_{i,q}\varphi_{i,q'}, \quad A_0(q,q') = \int_{\tau_i}\varphi_{i,q}\mathcal{L}_0^{-1}(a\varphi_{i,q'}). \tag{6.72}$$

Let $\lambda_{\max}(M_0,A_0)$ be the largest generalized eigenvalue of the eigenvalue problem $M_0\alpha = \lambda A_0\alpha$, which can be written as

$$\lambda_{\max}(M_0,A_0) = \sup_{v\in\mathbb{R}^Q}\frac{v^TM_0v}{v^TA_0v} = \sup_{\varphi\in\mathcal{P}_k(\tau_i)}\frac{\|\varphi\|^2_{L^2(\tau_i)}}{\|\mathcal{L}_0^{-1}\varphi\|^2_{H(\tau_i)}}. \tag{6.73}$$

The proof of Lemma 6.2 also implies that

$$\sqrt{\lambda_{\max}(M_0,A_0)} \leq \sqrt{2\theta_{k,\max}}C(k,d,\delta)h^{-k}. \tag{6.74}$$

If the operator $\mathcal{L}$ contains only the highest order terms, we have

$$\sqrt{\lambda_{\max}(M_0,A_0)} \leq \sqrt{\theta_{k,\max}}C(k,d,\delta)h^{-k}. \tag{6.75}$$

7 Localization of the basis functions

Theorem 5.1 or Theorem 6.4 allows us to localize the construction of basis functions $\psi_{i,q}$ as follows. For $r > 0$, let S_r be the union of the subdomains τ_j that intersect with $B(x_i, r)$ (recall that $B(x_i, \delta h_i/2) \subset \tau_i$) and let $\psi_{i,q}^{\mathrm{loc}}$ be the minimizer of the following quadratic problem:

$$\begin{aligned} \psi_{i,q}^{\mathrm{loc}} &= \underset{\psi \in H_0^k(S_r)}{\arg\min} \quad \|\psi\|_H^2 \\ &\text{s.t.} \quad \int \varphi_{j,q'}\psi = \delta_{iq,jq'} \quad \forall 1 \le j \le m, \quad 1 \le q' \le Q. \end{aligned} \tag{7.1}$$

We will naturally identify $\psi_{i,q}^{\mathrm{loc}}$ with its extension to $H_0^k(D)$ by setting $\psi_{i,q}^{\mathrm{loc}} = 0$ outside of S_r.

If the elliptic operator $\mathcal{L}$ is given with some other homogeneous boundary condition, the localized problem (7.1) should be slightly modified as follows such that the basis function $\psi_{i,q}$ honors the given boundary condition on ∂D:

$$\begin{aligned} \psi_{i,q}^{\mathrm{loc}} &= \underset{\psi \in H}{\arg\min} \quad \|\psi\|_H^2 \\ &\text{s.t.} \quad \int \varphi_{j,q'}\psi = \delta_{iq,jq'} \quad \forall 1 \le j \le m, \quad 1 \le q' \le Q, \\ &\qquad \psi(x) \equiv 0 \quad x \in D \backslash S_r. \end{aligned} \tag{7.2}$$

When $\partial S_r \cap \partial D = \emptyset$, Eq. (7.2) is equivalent to Eq. (7.1). However, when $\partial S_r \cap \partial D \neq \emptyset$, Eq. (7.2) only enforces the zero Dirichlet boundary condition on $\partial S_r \backslash \partial D$, but honors the original boundary condition on ∂D.

From now on, to simplify the expression of constants, we will assume without loss of generality that the domain is rescaled so that $\mathrm{diam}(D) \le 1$.

Lemma 7.1 *For any domain partition with* $\frac{h^2(1-h^{2k})}{1-h^2} \le \frac{\theta_{k,\max}}{2\theta_{0,\max} C_f^2}$, *it holds true that*

$$\|\psi_{i,q}^{\mathrm{loc}}\|_H \le C(k,d,\delta)\left(\frac{2^{d+1}\theta_{k,\max}}{V_d \delta^d}\right)^{1/2} h^{-d/2-k}. \tag{7.3}$$

If the operator $\mathcal{L}$ *contains only the highest order terms, it holds true that* $\|\psi_{i,q}^{\mathrm{loc}}\|_H \le C(k,d,\delta)\left(\frac{2^d\theta_{k,\max}}{V_d\delta^d}\right)^{1/2} h^{-d/2-k}$ *for any* $h > 0$.

Proof Consider

$$\zeta_{i,q} = \sum_{q=1}^{Q} A_0^{-1}(q,q')\mathcal{L}_0^{-1}\varphi_{i,q'},$$

where A_0^{-1} is the inverse of A_0 (defined in Eq. (6.74)) and $\mathcal{L}_0^{-1}\varphi_{i,q'}$ is the weak solution of the local problem (6.71) with right-hand side $\varphi_{i,q'}$. From the definition of A_0, we know that $\int_{\tau_i} \varphi_{i,q}\zeta_{i,q'} = \delta_{q,q'}$. Notice that $\zeta_{i,q} \in H_0^k \subset H_0^k(S_r)$. Therefore, $\zeta_{i,q}$ satisfies all constraints of $\psi_{i,q}^{\mathrm{loc}}$ (see Eq. (7.1)), and thus,

$$\|\psi_{i,q}^{\mathrm{loc}}\|_H \le \|\zeta_{i,q}\|_H. \tag{7.4}$$

Making use of $(\mathcal{L}_0^{-1}\varphi_{i,q}, \mathcal{L}_0^{-1}\varphi_{i,q'})_H = \int_{\tau_i} \varphi_{i,q}\mathcal{L}_0^{-1}\varphi_{i,q'} = A_0(q,q')$, we obtain

$$\|\zeta_{i,q}\|_H^2 = A_0^{-1}(q,q) \le \lambda_{\max}(A_0^{-1}) = \frac{\lambda_{\max}(M_0, A_0)}{|\tau_i|}. \tag{7.5}$$

We have used $M_0(q,q') = |\tau_i|\delta_{i,j}$ (due to the normalization (6.8)) in the last inequality. Combining Eq. (6.75) (or (6.74)), (7.4) and (7.5) and $|\tau_i| \ge V_d(\delta h/2)^d$, we complete the proof of Eq. (7.3). □

Theorem 7.1 *Under the same assumptions as those in Theorem 6.4, there exists $h_0 > 0$ such that for any $h \le h_0$, $1 \le i \le m$ and $1 \le q \le Q$, it holds true that*

$$\|\psi_{i,q} - \psi_{i,q}^{\text{loc}}\|_{H(D)} \le C_3 h^{-d/2-k} \exp(-\frac{r-2h}{2lh}), \tag{7.6}$$

where

$$C_3 = C(k,d,\delta)\left(\frac{e2^{d+1}\theta_{k,\max}}{V_d\delta^d}\right)^{1/2} \times \left(\left(2C_1C_\eta C_p\sqrt{\frac{k\theta_{k,\max}}{\theta_{k,\min}}}+1\right)^2 + 2\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}C(k,d,\delta)C_p\right)^{1/2}.$$

Here, all the parameters are the same as those in Theorem 6.4.

When the operator $\mathcal{L}$ contains only the highest order terms, i.e., $\mathcal{L}u = (-1)^k \sum_{|\sigma|=|\gamma|=k} D^\sigma (a_{\sigma\gamma}D^\gamma u)$, Eq. (7.6) holds true for all $h > 0$. In this case, the constant C_3 can be taken as

$$C_3 = C(k,d,\delta)\left(\frac{e2^{d}\theta_{k,\max}}{V_d\delta^d}\right)^{1/2} \times \left(\left(C_1C_\eta C_p\sqrt{\frac{k\theta_{k,\max}}{\theta_{k,\min}}}+1\right)^2 + \sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}C(k,d,\delta)C_p\right)^{1/2}.$$

Proof Let S_0 be the union of the subdomains τ_j that are not contained in S_r and let S_1 be the union of the subdomains τ_j that are at distance at least h from S_0. (We will assume that $S_0 \neq \emptyset$ and $S_1 \neq \emptyset$. If $S_0 \neq \emptyset$, the proof is trivial. We can choose $r \ge 2h$ such that $S_1 \neq \emptyset$.) Let S^* be the union of the subdomains τ_j that are not contained in either S_0 or S_1, as illustrated in Fig. 2. Note that in this case, we have S_1 in the inner region and S_0 in the outer region. This is the opposite of the scenario that we consider in Fig. 1.

Let η be a smooth cut-off function such that $0 \le \eta \le 1$, $\eta|_{S_1} \equiv 1$, $\eta|_{S_0} \equiv 0$ and $\|D^\sigma\eta\|_{L^\infty(D)} \le \frac{C_\eta}{h^{|\sigma|}}$ for all σ. Since $\psi_{i,q}^{\text{loc}}$ satisfies the same constraints as those in the definition of $\psi_{i,q}$, thanks to Eq. (2.12) we have

$$\|\psi_{i,q} - \psi_{i,q}^{\text{loc}}\|_{H(D)}^2 = \|\psi_{i,q}^{\text{loc}}\|_{H(D)}^2 - \|\psi_{i,q}\|_{H(D)}^2. \tag{7.7}$$

Define $\psi_{j,q}^{i,r}$ as the (unique) minimizer of the following quadratic optimization:

$$\begin{aligned} \psi_{j,q}^{i,r} := \arg\min_{\psi \in H_0^k(S_r)} \quad & \|\psi\|_{H(S_r)}^2 \\ \text{s.t.} \quad & \int_{S_r} \psi\varphi_{j',q'} = \delta_{jq,j'q'}, \quad \forall 1 \le j' \le m, \quad 1 \le q' \le Q. \end{aligned} \tag{7.8}$$

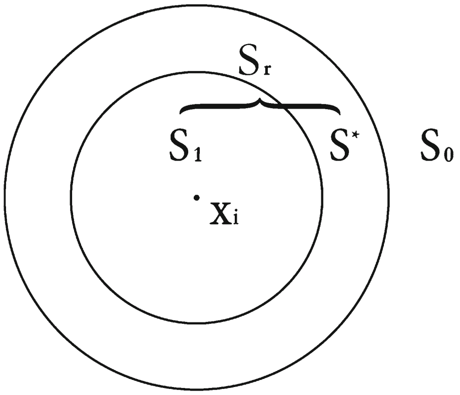

Fig. 2 Illustration of S_r, S_0, S_1 and S^*

Note that $\psi^{\text{loc}}_{i,q} = \psi^{i,r}_{i,q}$. Let $w_{jq'} = \int_D \eta\psi_{i,q}\varphi_{j,q'}$ and $\psi^{iq,r}_w = \sum_{j=1}^m \sum_{q'=1}^Q w_{jq'}\psi^{i,r}_{j,q'}$. Thanks to the orthogonality between $\psi_{i,q}$ and $\varphi_{j,q'}$, i.e., the constraints in Eq. (6.9), we have

$$\psi^{iq,r}_w = \psi^{\text{loc}}_{i,q} + \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q} w_{jq'}\psi^{i,r}_{j,q'}.$$

Using (3) of Theorem 2.2, we have $(\psi^{\text{loc}}_{i,q}, \psi^{i,r}_{j,q'})_H = \Theta^{i,-1}_{iq,jq'}$, where Θ^i is defined by Eq. (2.10) with $\mathcal{K} : L^2(S_r) \to L^2(S_r)$ being the inverse of $\mathcal{L}$ with the homogeneous Dirichlet boundary condition on ∂S_r. Therefore, we have

$$\|\psi^{iq,r}_w\|_H^2 = \|\psi^{\text{loc}}_{i,q}\|_H^2 + \left\| \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q} w_{jq'}\psi^{i,r}_{j,q'} \right\|_H^2 + 2 \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q} w_{jq'}\Theta^{i,-1}_{iq,jq'}. \tag{7.9}$$

By (2) of Theorem 2.2, we know that $\psi^{iq,r}_w$ is the minimizer of the following quadratic problem:

$$\begin{aligned} \psi^{iq,r}_w = \underset{\psi \in H_0^k(S_r)}{\arg\min} \quad & \|\psi\|^2_{H(S_r)} \\ \text{s.t.} \quad & \int_{S_r} \psi\varphi_{j,q'} = \int_D \eta\psi_{i,q}\varphi_{j,q'} \quad \forall 1 \le j \le m, \quad 1 \le q' \le Q. \end{aligned} \tag{7.10}$$

Noting that $\eta\psi_{i,q}$ satisfies the same constraint, we have $\|\psi^{iq,r}_w\|^2_H \le \|\eta\psi_{i,q}\|^2_H$. By using this estimate with (7.7) and (7.9), we obtain

$$\|\psi_{i,q} - \psi^{\text{loc}}_{i,q}\|^2_{H(D)} \le \underbrace{\|\eta\psi_{i,q}\|^2_H - \|\psi_{i,q}\|^2_H}_{I_1} + \underbrace{2\left| \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q} w_{jq'}\Theta^{i,-1}_{iq,jq'} \right|}_{I_2}. \tag{7.11}$$

It turns out that I_1 and I_2 play almost the same role as I_1 and I_2 did in the proof of Theorem 6.4 and can be estimated in a similar way. We will estimate these two terms as follows.

Let's first deal with I_1. Since $\eta|_{S_1} \equiv 1$ and $\eta|_{S_0} \equiv 0$, we have $I_1 = \|\eta\psi_{i,q}\|^2_{H(S^*)} - \|\psi_{i,q}\|^2_{H(S^* \cup S_0)} \le \|\eta\psi_{i,q}\|^2_{H(S^*)}$. In "Appendix B.2," we give a bound for $\|\eta\psi_{i,q}\|_{H(S^*)}$ using a similar technique that we used to obtain Eq. (6.53) from Eq. (6.48) in the proof of Theorem 6.4. With this bound, we obtain

$$I_1 \le \left(\frac{C_3}{2} |\psi_{i,q}|_{k,2,S^*} + \sqrt{\frac{C_3^2}{4} |\psi_{i,q}|_{k,2,S^*}^2 + C_3 |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)} + \|\psi_{i,q}\|_{H(S^*)}^2} \right)^2, \tag{7.12}$$

where $C_3 = C_1 C_\eta C_p \sqrt{2k\theta_{k,\max}}$. With the strong ellipticity (6.46) and the bound (6.24), we conclude

$$I_1 \le \left(2C_1 C_\eta C_p \sqrt{\frac{k\theta_{k,\max}}{\theta_{k,\min}}} + 1 \right)^2 \|\psi_{i,q}\|_{H(S^*)}^2. \tag{7.13}$$

Applying the exponential decay of Theorem 6.4 to $\|\psi_{i,q}\|_{H(S^*)}$, we get

$$I_1 \le \left(2C_1 C_\eta C_p \sqrt{\frac{k\theta_{k,\max}}{\theta_{k,\min}}} + 1 \right)^2 e^{1-\frac{r-2h}{lh}} \|\psi_{i,q}\|_{H(D)}^2. \tag{7.14}$$

We now estimate I_2. Combining (3) of Theorem 2.2 with the definition of H-norm (2.1), we have

$$\Theta_{iq,jq'}^{i,-1} = (\psi_{i,q}^{\text{loc}}, \psi_{j,q'}^{i,r})_{H(S_r)} = (\mathcal{L}\psi_{i,q}^{\text{loc}}, \psi_{j,q'}^{i,r})_{L^2(S_r)}.$$

Thanks to $\mathcal{L}\psi_{i,q}^{\text{loc}} |_{\tau_j} \in \text{span}\{\varphi_{j,q'}\}_{q=1}^{Q}$ and the orthogonality between Φ and $\psi_{j,q'}^{i,r}$, we have

$$\mathcal{L}\psi_{i,q}^{\text{loc}} |_{\tau_j} = \sum_{q'=1}^{Q} \Theta_{iq,jq'}^{i,-1} \varphi_{j,q'}.$$

Since $\{\varphi_{j,q'}\}_{q=1}^{Q}$ is orthogonal and normalized such that $\int \varphi_{j,q}\varphi_{j,q'} = |\tau_j|\delta_{q,q'}$, we get

$$\|\mathcal{L}\psi_{i,q}^{\text{loc}}\|_{L^2(\tau_j)} = |\tau_j|^{1/2} \left(\sum_{q'=1}^{Q} (\Theta_{iq,jq'}^{i,-1})^2 \right)^{1/2}. \tag{7.15}$$

Moreover, we obtain $w_{jq'} = \int_D \eta\psi_{i,q}\varphi_{j,q'}$ by definition, and thus we get

$$|\tau_j|^{-1/2} \left(\sum_{q'=1}^{Q} |w_{jq'}|^2 \right)^{1/2} \le \|\eta\psi_{i,q}\|_{L^2(\tau_j)} \le \|\psi_{i,q}\|_{L^2(\tau_j)}. \tag{7.16}$$

Here, we have made use of $0 \le \eta \le 1$ in the last step. Combining (7.15) and (7.16), we get

$$\begin{aligned} I_2 &= 2 \Big| \sum_{\tau_j \subset S^*} \sum_{q'=1}^{Q} w_{jq'} \Theta_{iq,jq'}^{i,-1} \Big| \\ &\le 2 \sum_{\tau_j \subset S^*} \left(\sum_{q'=1}^{Q} (\Theta_{iq,jq'}^{i,-1})^2 \right)^{1/2} \left(\sum_{q'=1}^{Q} |w_{jq'}|^2 \right)^{1/2} \\ &\le 2 \sum_{\tau_j \subset S^*} \|\mathcal{L}\psi_{i,q}^{\text{loc}}\|_{L^2(\tau_j)} \|\psi_{i,q}\|_{L^2(\tau_j)}. \end{aligned}$$

Now, we arrive at exactly the same situation as I_2 (see (6.41)) in the proof of Theorem 6.3. With the same derivation from Eqs. (6.41) to (6.42), i.e., applying Lemma 6.2 to $\|\mathcal{L}\psi_{i,q}^{\mathrm{loc}}\|_{L^2(\tau_j)}$ and Theorem 3.1 to $\|\psi_{i,q}\|_{L^2(\tau_j)}$, we obtain

$$\begin{aligned} I_2 &\le 2\sqrt{2\theta_{k,\max}}C(k,d,\delta)C_p|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}^{\mathrm{loc}}\|_{H(S^*)} \\ &\le 4\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}C(k,d,\delta)C_p\|\psi_{i,q}\|_{H(S^*)}\|\psi_{i,q}^{\mathrm{loc}}\|_{H(S^*)}, \end{aligned} \tag{7.17}$$

where we have used $\theta_{k,\max} := \max(\theta_{0,\max},\theta_{k,\max})$, the strong ellipticity (6.46) and the bound (6.24) in the last step. Applying the exponential decay of Theorem 6.4 to both $\|\psi_{i,q}\|_{H(S^*)}$ and $\|\psi_{i,q}^{\mathrm{loc}}\|_{H(S^*)}$, we obtain

$$I_2 \le 2\sqrt{\frac{\theta_{k,\max}}{\theta_{k,\min}}}C(k,d,\delta)C_p e^{1-\frac{r-2h}{lh}}\|\psi_{i,q}\|_{H(D)}\|\psi_{i,q}^{\mathrm{loc}}\|_{H(D)}. \tag{7.18}$$

Combining Eqs. (7.11), (7.14) and (7.18), and using Eq. (7.3) to bound $\|\psi_{i,q}^{\mathrm{loc}}\|_{H(D)}$ and $\|\psi_{i,q}\|_{H(D)}$ (recall $\|\psi_{i,q}\|_{H(D)} \le \|\psi_{i,q}^{\mathrm{loc}}\|_{H(D)}$), we complete the proof of Eq. (7.6).

When the operator $\mathcal{L}$ contains only the highest order terms, i.e., $\mathcal{L}u = (-1)^k \sum_{|\sigma|=|\gamma|=k} D^\sigma(a_{\sigma\gamma}D^\gamma u)$, Eqs. (7.14) and (7.18) hold true for all $h > 0$. In this case, we can get rid of the factor "2" in both Eqs. (7.14) and (7.18). Therefore, we obtain the estimate on C_3 stated in the theorem. □

Theorem 7.2 *Let $u \in H_0^k(D)$ be the weak solution of $\mathcal{L}u = f$ and $\psi_{i,q}^{\mathrm{loc}}$ be the localized basis functions defined in Eq. (7.1). Then, for $r \ge (d+4k)lh\log(1/h) + 2(1 + l\log C_4)h$, we have*

$$\inf_{v\in\Psi^{\mathrm{loc}}} \|u-v\|_{H(D)} \le \frac{2C_p}{\sqrt{a_{\min}}}h^k\|f\|_{L^2(D)}, \tag{7.19}$$

where $C_4 = \frac{C_3C_e}{C_p}(Qa_{\min})^{1/2}$, and C_3 is defined in Theorem 7.1, $a_{\min}$ comes from the norm-equivalence (6.4), and C_e is the constant such that $\|u\|_{L^2(D)} \le C_e\|f\|_{L^2(D)}$ holds true.

Proof Let $v_1 := \sum_{i=1}^m\sum_{q=1}^Q c_{iq}\psi_{i,q}$ and $v_2 := \sum_{i=1}^m\sum_{q=1}^Q c_{iq}\psi_{i,q}^{\mathrm{loc}}$ with $c_{iq} = \int_D u\varphi_{i,q}$. Estimation (6.11) gives that

$$\|u-v_1\|_H \le \frac{C_p h^k}{\sqrt{a_{\min}}}\|f\|_{L^2(D)}. \tag{7.20}$$

Using the Cauchy inequality, we have

$$\begin{aligned} \|v_1-v_2\|_H &\le \max_{i,q}\|\psi_{i,q}-\psi_{i,q}^{\mathrm{loc}}\|_H \sum_{i=1}^m\sum_{q=1}^Q |c_{iq}| \\ &\le \max_{i,q}\|\psi_{i,q}-\psi_{i,q}^{\mathrm{loc}}\|_H \sum_{i=1}^m Q^{1/2}\left(\sum_{q=1}^Q |c_{iq}|^2\right)^{1/2}. \end{aligned}$$

Thanks to the orthogonality of $\{\varphi_{i,q}\}_{q=1}^Q$ (6.8), we have $|\tau_i|^{-1/2}(\sum_{q=1}^Q|c_{iq}|^2)^{1/2} \le \|u\|_{L^2(\tau_i)}$. Then we obtain

$$
\begin{aligned}
\|v_1 - v_2\|_H &\leq \max_{i,q} \|\psi_{i,q} - \psi_{i,q}^{\text{loc}}\|_H Q^{1/2} \sum_{i=1}^{m} |\tau_i|^{1/2} \|u\|_{L^2(\tau_i)} \\
&\leq \max_{i,q} \|\psi_{i,q} - \psi_{i,q}^{\text{loc}}\|_H (Q|D|)^{1/2} \|u\|_{L^2(D)}.
\end{aligned}
$$

Using the energy estimation $\|u\|_{L^2(D)} \leq C_e \|f\|_{L^2(D)}$ and Theorem 7.1, we obtain

$$
\|v_1 - v_2\|_H \leq C_3 C_e Q^{1/2} h^{-\frac{d}{2}-k} \exp\left(-\frac{r-2h}{2lh}\right) \|f\|_{L^2(D)}. \tag{7.21}
$$

Combining Eqs. (7.20) and (7.21) together, we conclude the proof. □

By applying the Aubin–Nistche duality argument, we can get the following corollary.

Corollary 7.3 *Let $\psi_{i,q}^{\text{loc}}$ be the localized basis functions defined in Eq. (7.1). Then for $r \geq (d+4k)lh\log(1/h) + 2(1 + l\log C_4)h$, we have*

$$
\|\mathcal{K} - \mathcal{P}_{\Psi^{\text{loc}}}^{(H)}\mathcal{K}\| \leq \frac{4C_p^2}{a_{\min}} h^{2k}, \tag{7.22}
$$

where all the constants are the same as those defined in Theorem 7.2.

Corollary 7.3 shows that we can compress the symmetric positive semidefinite operator $\mathcal{K}$ with the optimal rate h^{2k} and with the nearly optimal localized basis (with support size of order $h\log(1/h)$).

Remark 7.1 All the results and proofs presented above can be carried over to other homogeneous boundary conditions. Given a specific homogeneous boundary condition, one only needs to modify the proof of Lemma 7.1. Specifically, when the patch τ_i intersects with the boundary of D, the constructed function $\zeta_{i,q}$ should honor the same boundary condition on ∂D. The scaling argument in the proof of Lemma 7.1 still works for other homogeneous boundary conditions.

8 Numerical examples

In this section, we present several numerical results to support the theoretical findings and to show how the sparse operator compression is utilized in higher-order elliptic operators. In Sect. 8.1, we apply our method to compress the Matérn covariance function (8.1) with $\nu = 1/2$. We show that our method is able to achieve the optimal compression error with nearly optimally localized basis functions, which means that we are able to get optimality on both ends of the accuracy–sparsity trade-off in the sparse PCA. In Sect. 8.2, we apply our method to a 1D fourth-order elliptic equation with the homogeneous Dirichlet boundary condition and show that our basis functions, when used as multiscale finite element basis, can achieve the optimal h^2 convergence rate in the energy norm. In Sect. 8.3, we apply our method to a 2D fourth-order elliptic equation and show that the energy-minimizing basis functions decays exponentially fast away from its associated patch.

8.1 The compression of a Matérn covariance kernel

In spatial statistics, geostatistics, machine learning and image analysis, the Matérn covariance [31] is used to model random fields with smooth samples; see, e.g., [14,17,43]. The

Matérn covariance between two points $x, y \in D \subset \mathbb{R}^d$ is given by

$$K_\nu(x,y) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{|x-y|}{\rho}\right)^\nu K_\nu\left(\sqrt{2\nu}\frac{|x-y|}{\rho}\right), \tag{8.1}$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind, and ρ and ν are nonnegative parameters of the covariance. Its Fourier transform is given by

$$\widehat{k}(\omega) = c_{\nu,\lambda}\sigma^2 \left(\frac{2\nu}{\lambda^2} + |\omega|^2\right)^{-(\nu+d/2)}, \quad c_{\nu,\lambda} := \frac{2^d \pi^{d/2}\Gamma(\nu + d/2)(2\nu)^\nu}{\Gamma(\nu)\lambda^{2\nu}}, \tag{8.2}$$

where $\widehat{f}(\omega)$ is the Fourier transform of f. For both sampling from the random fields and performing basic computations like marginalization and conditioning, we need to compress the Matérn covariance operator $\mathcal{K} : L^2(D) \to L^2(D)$, which is defined through the Hilbert–Schmidt operator with kernel $K_\nu(x, y)$, by a rank-n covariance operator:

$$E_{\text{oc}}(\Psi; \mathcal{K}) := \min_{K_n \in \mathbb{R}^{n\times n},\ K_n \succeq 0} \|\mathcal{K} - \Psi K_n \Psi^T\|_2, \tag{8.3}$$

where $\Psi = [\psi_1, \ldots, \psi_n]$ spans the range space of the approximate operator $\Psi K_n \Psi^T$. Recent study [4,27] shows that the Matérn covariance and the elliptic operators are closely connected. With proper homogeneous boundary conditions, the Matérn covariance operator with $\nu + d/2$ being an integer is the solution operator of an elliptic operator of order $2\nu + d$. For example, the Matérn covariance operator with $\nu = 1/2$ is the solution operator of a second-order elliptic operator $(2l\sigma^2)^{-1}\left(1 - \rho^2 \frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)$ when the physical dimension $d = 1$ and is the solution operator of a fourth-order elliptic operator $(8\pi\rho^3\sigma^2)^{-1}\left(1 - 2\rho^2\Delta + \rho^4\Delta^2\right)$ when $d = 3$.

Based on Eqs. (2.10) and (2.11), we can also compute the exponentially decaying basis functions from the covariance operator $\mathcal{K}$. In this example, we apply our method to compress the following exponential kernel

$$K(x,y) = \exp(-|x-y|) \qquad x, y \in [0,1], \tag{8.4}$$

which is exactly the Matérn covariance (8.1) with $\nu = 1/2, \sigma = 1$ and $\rho = 1$. This problem has been studied by different groups; see, e.g., [3,9,13,20]. We remark that since the Matérn covariance function corresponds to the solution operator of an elliptic PDE with constant coefficient, one can compress the Matérn covariance kernel by using a piecewise linear polynomial or wavelets with optimal locality and accuracy. It is not necessary to use the exponential decaying basis to perform the operator compression. We use this example to illustrate that our method can be also applied to compress a general kernel function.

We partition the interval $[0, 1]$ uniformly into $m = 2^6$ patches and follow our strategy to construct basis functions. By the Fourier transform, we know that it is associated with the second-order elliptic operator $\frac{1}{2}\left(1 - \frac{\mathrm{d}^2}{\mathrm{d}x^2}\right)$. Therefore, we take Φ as piecewise constant functions and then compute Ψ by Eqs. (2.10) and (2.11). In Fig. 3, we plot φ_{32} and ψ_{32}, which is associated with the patch $[1/2 - h, 1/2]$. We can see that the basis function ψ_{32} clearly has an exponential decay. We take $m = 2^i$ for $0 \le i \le 7$ and compute the compression error $E(\Psi; \mathcal{K})$. The result is shown in Fig. 4. We can see that

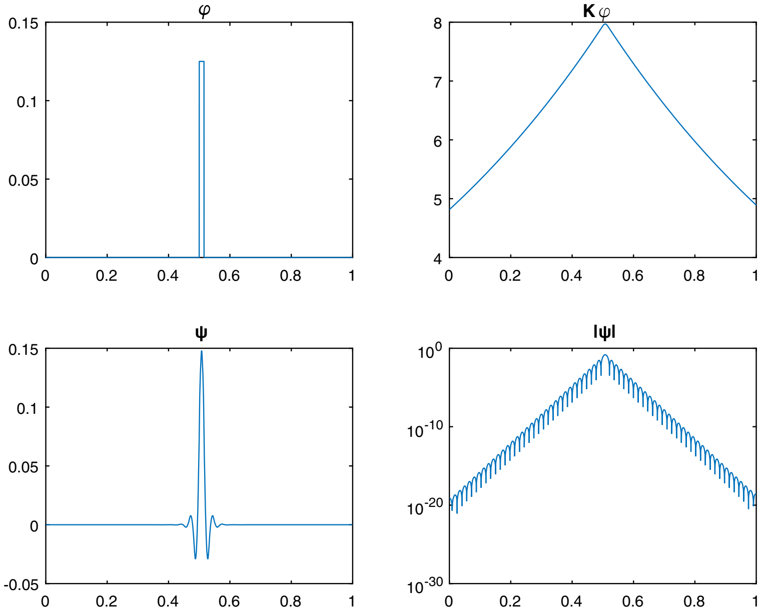

Fig. 3 Basis function associated with patch $[1/2 - h, 1/2]$

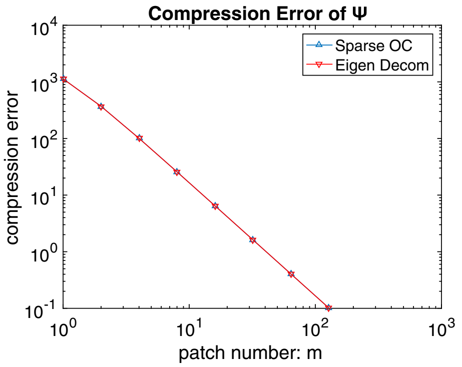

Fig. 4 Operator compression error $E(\Psi; \mathcal{K})$ (8.3) for the exponential kernel (8.4) with exponentially decaying basis functions Ψ. They have nearly same compression error as that given by the global eigenfunctions of $\mathcal{K}$

the exponentially decaying basis functions Ψ have nearly the same compression rate as that of the eigendecomposition.

One can easily verify that the exponential kernel (8.4) is the Green's function of the following second-order elliptic equation

$$-\frac{1}{2}u''(x) + \frac{1}{2}u = f(x), \qquad 0 < x < 1, u(0) - u'(0) = 0, \quad u(1) + u'(1) = 0, \tag{8.5}$$

with boundary condition $u(0) - u'(0) = 0, \quad u(1) + u'(1) = 0$. The associated energy norm is

$$\|u\|_{H(D)}^2 = \frac{1}{2}\left(u(0)^2 + u(1)^2 + \int_0^1 (u')^2 + \int_0^1 u^2\right). \tag{8.6}$$

Solving the localized variational problem (7.2), we can get localized basis functions Ψ^{loc}. With different sizes of the support S_r, we compute the compression error $E(\Psi^{\mathrm{loc}}; \mathcal{K})$ for $m = 2^i$ $(0 \leq i \leq 7)$. The results are summarized in Fig. 5. In the left subfigure of Fig. 5, we take the support with size Ch, for $C = 3, 5, 7, 9$ and 11. In the right subfigure of Fig. 5, we take the support with size $Ch \log_2(1/h)$, for $C = 2, 2.1$ and 2.4. For a support of size $Ch \log_2(1/h)$, it contains $\lceil C \log_2(1/h) \rceil$ patches, where $\lceil C \log_2(1/h) \rceil$ is the smallest integer of $C \log_2(1/h)$. We can see that the oversampling strategy with $r = ch$ does not give the optimal convergence rate , while the oversampling strategy with $r = ch \log_2(1/h)$ gives the optimal second-order convergence rate as guaranteed by Corollary 7.3. For $m = 2^7$ and $r = 2.4h \log_2(1/h)$, the constructed localized basis functions achieves the same operator compression error as that using 128 eignefunctions.

8.2 The 1D fourth-order elliptic operator

Consider the solution operator of the Euler-Bernoulli equation

$$\begin{aligned} &\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(a(x)\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}\right) = f(x), \qquad 0 < x < 1, \\ &\qquad u(0) = u'(0) = 0, \quad u(1) = u'(1) = 0, \end{aligned} \tag{8.7}$$

which describes the deflection u of a clamped beam subject to a transverse force $f \in L^2([0,1])$. The flexural rigidity $a(x)$ of the beam is modeled by

$$a(x) := 1 + \frac{1}{2}\sin\left(\sum_{k=1}^{K} k^{-\alpha}(\zeta_{1k}\sin(kx) + \zeta_{2k}\cos(kx))\right), \tag{8.8}$$

where $\{\zeta_{1k}\}_{k=1}^{K}$ and $\{\zeta_{2k}\}_{k=1}^{K}$ are two independent random vectors with independent entries uniformly distributed in $[-1/2, 1/2]$. This oscillatory coefficient is also used in [22,33,39] and has no scale separation. We choose $\alpha = 0$ and $K = 40$ in the numerical experiment. A sample coefficient is shown in Fig. 6.

We partition the physical space $[0, 1]$ uniformly into $m = 2^6$ patches, where the ith patch $I_i = [(i-1)h, ih]$ with $h = 1/m$. In this fourth-order case, our theory requires the piecewise polynomial space Φ be the space of (discontinuous) piecewise linear functions,

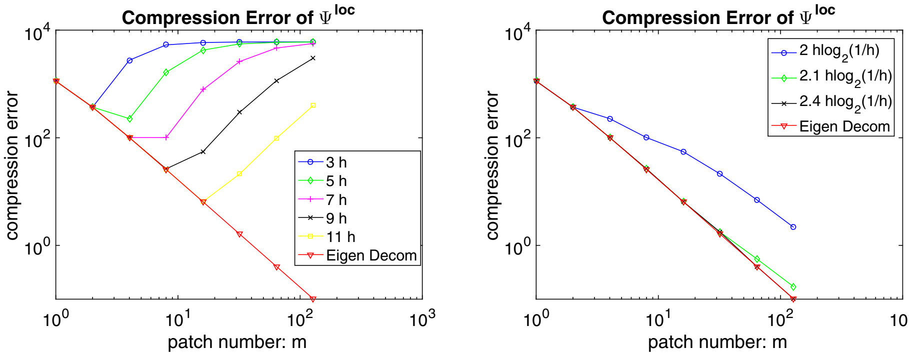

Fig. 5 Operator compression error $E(\Psi^{\mathrm{loc}}; \mathcal{K})$ (8.3) with basis functions Ψ^{loc}. The oversampling strategy with $r = ch$ (left) does not work well, while the oversampling strategy with $r = ch\log_2(1/h)$ (right) gives the optimal second-order convergence rate

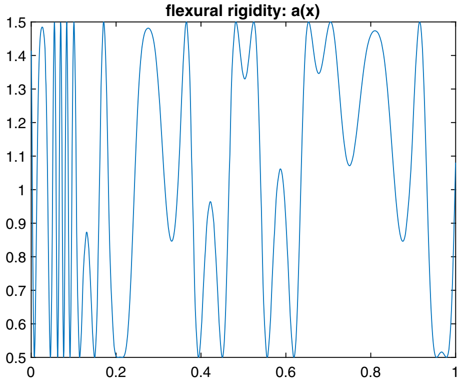

Fig. 6 Highly oscillatory flexural rigidity without scale separation

which has dimension $n = 2m$. We have two φ's, denoted as $\varphi_{i,1}$ and $\varphi_{i,2}$, associated with the patch I_i. Solving the quadratic optimization problem (6.9), we obtain the exponentially decaying basis functions. We also have two ψ's, denoted as $\psi_{i,1}$ and $\psi_{i,2}$, associated with the patch I_i. We plot $\varphi_{i,1}$ and $\varphi_{i,2}$ associated with the patch $I_{32} = [1/2 - h, 1/2]$ in Fig. 7 A. In Fig. 7b, c, we plot the basis functions $\psi_{32,1}$ and $\psi_{32,2}$, which clearly show exponential decay.

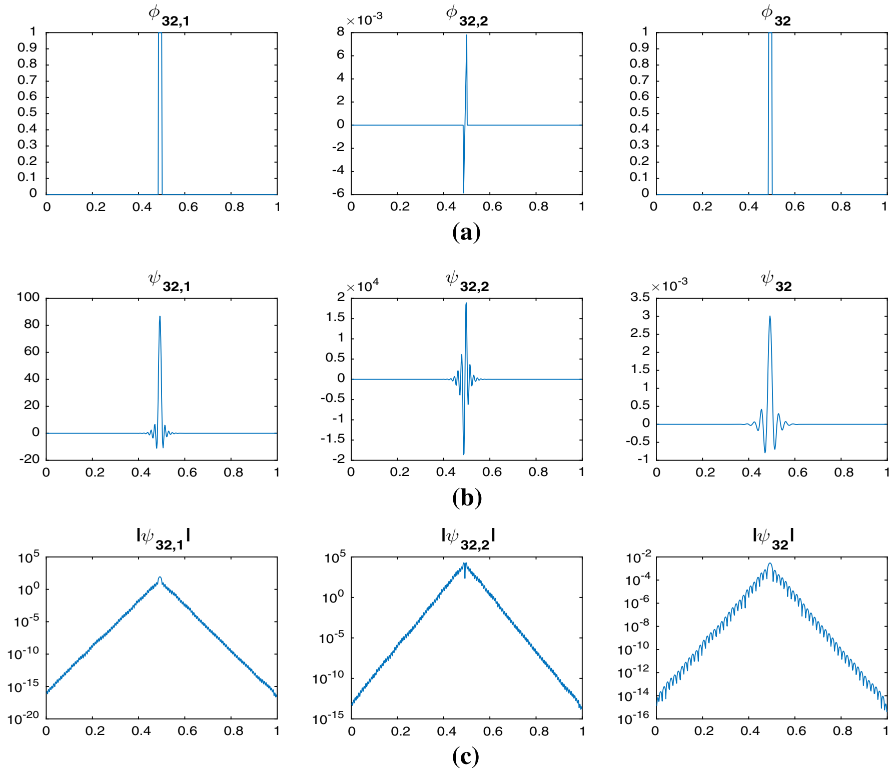

Fig. 7 One-dimensional fourth-order elliptic operator (8.7). **a** $\varphi_{32,1}$, $\varphi_{32,2}$ for piecewise linear Φ and φ_{32} for piecewise constant Φ. **b** $\psi_{32,1}$, $\psi_{32,2}$ for piecewise linear Φ and ψ_{32} for piecewise constant Φ. **c** In log-scale: $\psi_{32,1}$, $\psi_{32,2}$ for piecewise linear Φ and ψ_{32} for piecewise constant Φ

To demonstrate the necessity for Ψ to contain all piecewise linear functions, in the third column of Fig. 7, we also plot the basis functions associated the patch I_{32} when Φ is the space of piecewise constant functions. In this case, we have only one φ, denoted as φ_i, associated with the patch I_i. In the third column of Fig. 7a, b, we plot φ_{32} and ψ_{32}. Solving the quadratic optimization problem (6.9), we obtain only one basis function ψ, denoted as ψ_i, associated with the patch I_i. In Fig. 7c, we plot the basis function ψ_{32} in the third column. Note that ψ_{32} also shows an exponential decay, but its decay rate is much smaller than that of $\psi_{32,1}$ and $\psi_{32,2}$.

We have sampled a force $f \in L^2(D)$ from the same model (8.8) as the flexural rigidity. Using the MsFEM, we use two different sets of basis functions $\{\psi_{i,q}\}_{i=1,q=1}^{m,2}$ and $\{\psi_i\}_{i=1}^{m}$ to solve the corresponding fourth-order elliptic equation (8.7) and get solutions $u_{h,1}$ and $u_{h,0}$ respectively. We show their errors in the energy norm, i.e., $\|u_{h,1} - u\|_H$ and $\|u_{h,0} - u\|_H$ in Fig. 8. We can see that $\|u_{h,1} - u\|_H$ decays quadratically with respect to the patch size h, while $\|u_{h,0} - u\|_H$ decays only linearly. Therefore, to obtain the optimal convergence rate h^2 in the energy norm, it is necessary to include all the piecewise linear functions in the space Φ, as we have proved in Theorem 2.1 and Eq. (6.11).

8.3 The 2D fourth-order elliptic operator

Consider the solution operator of the 2D fourth-order elliptic equation on domain $D = (0, 1)^2$

$$\begin{aligned}&\partial_x^2(a_{20}(x,y)\partial_x^2 u(x,y)) + \partial_y^2(a_{02}(x,y)\partial_y^2 u(x,y))\\&\quad + 2\partial_{xy}(a_{11}(x,y)\partial_{xy}u(x,y)) = f(x,y), \quad u \in H_0^2(D),\end{aligned} \tag{8.9}$$

which describes the vibration u of a clamped plate subject to a transverse force $f \in L^2(D)$. The coefficients in the operator are given by

$$\begin{aligned}a_{20}(x,y) = a_{02}(x,y) &= \frac{1}{6}\left(\frac{1.1+\sin(2\pi x/\epsilon_1)}{1.1+\sin(2\pi y/\epsilon_1)} + \frac{1.1+\sin(2\pi y/\epsilon_2)}{1.1+\cos(2\pi x/\epsilon_2)}\right.\\&\quad\left.+\frac{1.1+\cos(2\pi x/\epsilon_3)}{1.1+\sin(2\pi y/\epsilon_3)} + \frac{1.1+\sin(2\pi y/\epsilon_4)}{1.1+\cos(2\pi x/\epsilon_4)} + \sin(4x^2y^2)+1\right),\\a_{11}(x,y) &= 1 + \frac{1}{2}\sin\left(\sum_{k=1}^{K} k^{-\alpha}(\zeta_{1k}\sin(kx) + \zeta_{2k}\cos(ky))\right),\end{aligned} \tag{8.10}$$

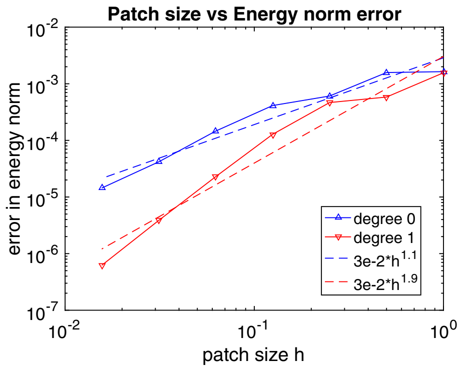

Fig. 8 Error of the finite element solutions: $\|u_{h,0} - u\|_H$ and $\|u_{h,1} - u\|_H$

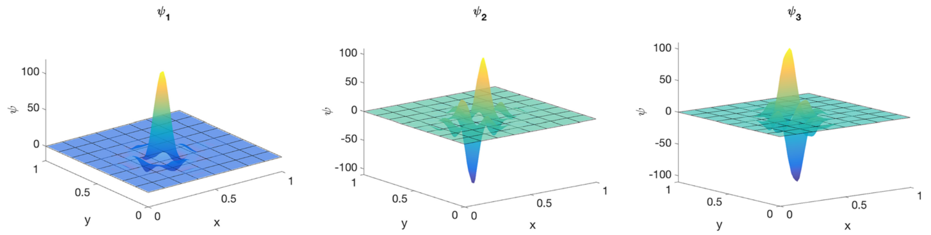

Fig. 9 Three basis functions associated with patch $[1/2 - h_x, 1/2] \times [1/2 - h_y, 1/2]$

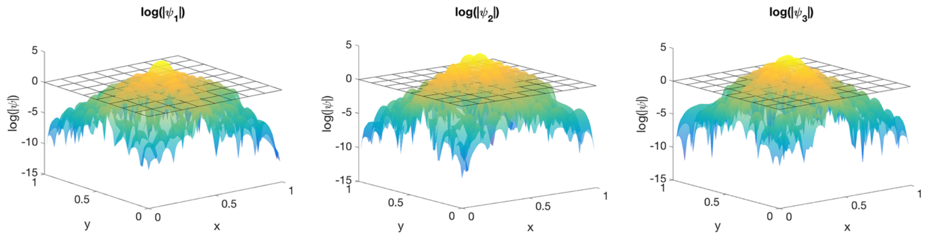

Fig. 10 Three basis functions associated with patch $[1/2 - h_x, 1/2] \times [1/2 - h_y, 1/2]$ in log-scale

where $\epsilon_1 = \frac{1}{5}, \epsilon_2 = \frac{1}{13}, \epsilon_3 = \frac{1}{17}, \epsilon_4 = \frac{1}{31}$, $K = 20$, $\alpha = 0$, and $\{\zeta_{1k}\}_{k=1}^{K}$ and $\{\zeta_{2k}\}_{k=1}^{K}$ are two independent random vectors with independent entries uniformly distributed in $[-1/2, 1/2]$.

Based on the uniform partition with grid size $h_x = h_y = \frac{1}{8}$, we construct the piecewise linear function space Φ, which has dimension $n = 3m = 192$. We solve the quadratic optimization problem (6.9) with the weighted extended B-splines (Web-splines [19]) of degree 3 on the uniform refined grid with grid size $h_{x,f} = h_{y,f} = \frac{1}{32}$. The 2D Gaussian quadrature with 5 points on each axis is utilized to compute the integral on each fine grid cell. The three basis functions associated with the patch $[1/2 - h_x, 1/2] \times [1/2 - h_y, 1/2]$ are shown in Fig. 9. We also show them in the log-scale in Fig. 10. We can clearly see that the basis functions decay exponentially fast away from its associated patch, which validates our Theorem 6.3.

We point out that the stiffness matrix for the fourth-order elliptic operator (8.9) becomes ill-conditioned very quickly when we refine the grid size. A carefully designed numerical strategy is required to validate the optimal convergence rate. We will leave this to our future work.

9 Concluding remarks

In this paper, we have developed a general strategy to compress a class of self-adjoint higher-order elliptic operators by minimizing the energy norm of the localized basis functions. These energy-minimizing localized basis functions are obtained by solving decoupled local quadratic optimization problems with linear constraints, and they give optimal approximation property of the solution operator. For a self-adjoint, bounded and strongly elliptic operator of order $2k$ ($k \geq 1$), we have proved that with support size $O(h \log(1/h))$, our localized basis functions can be used to compress higher-order elliptic operators with the optimal compression rate $O(h^{2k})$. We have applied our new operator compression strategy in different applications. For elliptic equations with rough coefficients, our localized basis functions can be used as multiscale basis functions, which gives the optimal convergence rate $O(h^k)$ in the energy norm. In the application of the sparse PCA, our localized basis functions achieve nearly optimal sparsity and the optimal approximation rate simultaneously when the covariance operator to be compressed is the solution operator of an elliptic operator. We remark that a number of Matérn covariance kernels are related to the Green's functions of some elliptic operators.

There are several directions we can explore in the future work. First of all, the constants in both the compression error and the localization depend on the contrast of the coefficients, which makes the existing methods inefficient for coefficients with high contrast. Other methods (e.g., [16,28,35,36]) also suffer from the same limitation. Our sparse operator compression framework can be used to deal with this high contrast case, and we will report our findings in our upcoming paper. Secondly, in the application of the sparse PCA, our current construction requires the knowledge of the underlying elliptic operator $\mathcal{L}$. We believe that it is possible to construct these localized basis functions using only the covariance function. Moreover, given any covariance operator, which may not be the solution operator of an elliptic operator, we can still define the Cameron–Martin space and the corresponding energy-minimizing basis functions. We are interested in the localization and compression properties of these energy-minimizing basis functions in this general setting. Our preliminary results show that the energy-minimizing basis functions still enjoy fast decay rate away from its associated patch, although the exponential decay may not hold true any more. Thirdly, it is interesting to apply our framework to the graph Laplacians, which can be viewed as discretized elliptic operators. Along this direction, we would like to develop an algorithm with nearly linear complexity to solve linear systems with graph Laplacians. Finally, we are also interested in applying our method to construct localized Wannier functions and to compress the Hamiltonian in quantum chemistry. Unlike the second-order elliptic operators with multiscale diffusion coefficients, all multiscale features of the Hamiltonian $\mathcal{H} = -\Delta + V(x)$ lie in its potential $V(x)$. Some adaptive domain partition strategy may prove to be useful in this application.

Acknowledgements

The research was in part supported by NSF Grants DMS 1318377 and DMS 1613861. We would like to thank Professor Lei Zhang and Venkat Chandrasekaran for several stimulating discussions, and Professor Houman Owhadi for valuable comments.

Dedication

Honor of Bjorn Engquist on the occasion of his 70th birthday.

Appendix A: More on Lemma 4.1

In this section, we prove that $C(k, s, d, \Omega_1)$ can be bounded by $C(k, s, d, \delta)$ and give an explicit formula of $C(k, s, d, \delta)$ for the case $k = s = 1$. Before we do this, we need the following comparison lemma.

Lemma A.1 *Let Ω be a smooth, bounded, open subset of $\mathbb{R}^d$ and S is a smooth subdomain in Ω. Let G_Ω be the Green's function of $\mathcal{L} = (-1)^k \sum_{|\sigma|=k} D^{2\sigma}$ with the homogeneous Dirichlet boundary condition on $\partial\Omega$ and G_S be the Green's function of $\mathcal{L}$ with the homogeneous Dirichlet boundary condition on ∂S. Then for all $f \in L^2(\Omega)$, we have*

$$\int_S \int_S G_S(x, y) f(x) f(y) \mathrm{d}x\, \mathrm{d}y \leq \int_\Omega \int_\Omega G_\Omega(x, y) f(x) f(y) \mathrm{d}x\, \mathrm{d}y. \tag{A.1}$$

Proof Let $f \in L^2(\Omega)$. Let ψ_Ω be the solution of $\mathcal{L}\psi_\Omega = f$ with the homogeneous Dirichlet boundary conditions on $\partial\Omega$ and ψ_S be the solution of $\mathcal{L}\psi_S = f$ with the homogeneous Dirichlet boundary conditions on ∂S. Observe that ψ_Ω and ψ_S are the unique minimizers of $I_\Omega(u,f) = \frac{1}{2}\sum_{|\sigma|=k}\int_\Omega |D^\sigma u|^2 - \int_\Omega uf$ with

$$\begin{aligned} &\psi_\Omega = \underset{u\in H_0^k(\Omega)}{\arg\min}\, I_\Omega(u,f), \qquad \psi_S = \underset{u\in H_0^k(S;\Omega)}{\arg\min}\, I_\Omega(u,f)\\ &H_0^k(S;\Omega) := \{u \in H_0^k(\Omega) : u \equiv 0 \text{ on } \Omega\backslash S\}. \end{aligned} \tag{A.2}$$

Moreover, we have

$$\begin{aligned} I_\Omega(\psi_\Omega,f) &= -\frac{1}{2}\int_\Omega \psi_\Omega f = -\frac{1}{2}\int_\Omega\int_\Omega G_\Omega(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y,\\ I_\Omega(\psi_S,f) &= -\frac{1}{2}\int_S \psi_S f = -\frac{1}{2}\int_S\int_S G_S(x,y)f(x)f(y)\mathrm{d}x\,\mathrm{d}y. \end{aligned} \tag{A.3}$$

Since $H_0^k(S;\Omega)$ is a subset of $H_0^k(\Omega)$, we obtain

$$I_\Omega(\psi_\Omega,f) \le I_\Omega(\psi_S,f), \tag{A.4}$$

which proves the lemma. □

Note that Lemma A.1 in fact holds true for the general operator $\sum_{0\le|\sigma|,|\gamma|\le k}(-1)^{|\sigma|}D^\sigma(a_{\sigma\gamma}(x)D^\gamma u)$ with various boundary conditions. Notice that Ω_1 is a smooth, bounded, open subset of $\mathbb{R}^d$ that satisfies $B(0,\delta/2) \subset \Omega_1 \le B(0,1)$. By Lemma A.1, we are able to bound the energy norm on Ω_1 by that on $B(0,\delta/2)$ and $B(0,1)$. To simplify the notation, we omit the subscript "1" in the rest of this section.

Proposition A.1 *$C(k,s,d,\Omega)$ (defined in Eq. (4.8)) can be bounded by $C(k,s,d,\delta)$ which only depends on k, s, d and δ. Moreover, we can set*

$$C(1,1,d,\delta) = 2\sqrt{d(d+2)}\delta^{-1-d/2}. \tag{A.5}$$

Proof From the definition (4.8), we have

$$(C(k,s,d,\Omega))^2 = \lambda_{\max}(M,S) = \max_{p\in\mathcal{P}_{s-1}} \frac{\int_\Omega p^2(x)\mathrm{d}x}{\int_\Omega\int_\Omega G(x,y)p(x)p(y)\mathrm{d}x\,\mathrm{d}y}, \tag{A.6}$$

where $G(x,y)$ is the Green's function of $\mathcal{L} = (-1)^k\sum_{|\sigma|=k}D^{2\sigma}$ with the homogeneous Dirichlet boundary condition on $\partial\Omega$. Notice that $B(0,\delta/2) \subset \Omega \subset B(0,1)$. Utilizing Lemma A.1, we have

$$\lambda_{\max}(M,S) \le \max_{p\in\mathcal{P}_{s-1}} \frac{\int_{B(0,1)} p^2(x)\mathrm{d}x}{\int_{B(0,\delta/2)}\int_{B(0,\delta/2)} G_{\delta/2}(x,y)p(x)p(y)\mathrm{d}x\,\mathrm{d}y} := \lambda_{\max}(\widehat{M},\widehat{S}),$$

where $G_{\delta/2}$ is the Green's function of $\mathcal{L}$ with the homogeneous Dirichlet boundary condition on $\partial B(0,\delta/2)$, $\lambda_{\max}(\widehat{M},\widehat{S}) > 0$ is the largest generalized eigenvalue of $\widehat{M}$ and $\widehat{S}$ with

$$\widehat{S}(i,j) = \int_{B(0,\delta/2)}\int_{B(0,\delta/2)} G_{\delta/2}p_ip_j = \int_{B(0,\delta/2)} u_{\delta/2,i}p_j, \qquad \widehat{M}(i,j) = \int_{B(0,1)} p_ip_j. \tag{A.7}$$

Here, $\{p_1, p_2, \ldots, p_Q\}$ are all the monomials defined in Lemma 4.1 and $u_{\delta/2,i} = \mathcal{L}^{-1}p_i$ with the homogeneous Dirichlet boundary condition on $\partial B(0, \delta/2)$. It is obvious that $\lambda_{\max}(\widehat{M}, \widehat{S})$ only depends on k, s, d and δ. Thus, we can choose

$$C(k, s, d, \delta) = \sqrt{\lambda_{\max}(\widehat{M}, \widehat{S})}. \tag{A.8}$$

Since Ω has diameter at most 1, there exists $x_0 \in \Omega$ such that $\Omega \subset B(x_0, 1/2)$. Therefore, we have $\int_\Omega p^2(x)\mathrm{d}x \le \int_{B(x_0,1/2)} p^2(x)\mathrm{d}x$, and we have a tighter bound for M in the case $s = 1$: $M \le \widehat{M} := \int_{B(x_0,1/2)} \mathrm{d}x = A_{d-1}/(d2^d)$, where A_{d-1} is the surface area of the $(d-1)$-sphere of radius 1 (set $A_0 = 2$).

For the case $s = k = 1$, $u_{\delta/2,1}$ (defined as $\mathcal{L}^{-1}p_1$ with the homogeneous Dirichlet boundary condition on $\partial B(0, \delta/2)$) can be solved explicitly:

$$u_{\delta/2,1} = \left((\delta/2)^2 - r^2\right)/(2d).$$

Then we have

$$\widehat{S} = \frac{1}{d^2(d+2)}\left(\frac{\delta}{2}\right)^{d+2} A_{d-1}, \quad \widehat{M} = A_{d-1}/(d2^d).$$

Since $\lambda_{\max}(\widehat{M}, \widehat{S}) = \widehat{M}/\widehat{S}$ in the case of $s = 1$, Eq. (A.5) naturally follows. □

Appendix B: Derivations involving I_1

B.1. From Eq. (6.49) to Eq. (6.50) in the proof of Theorem 6.4

We want to prove that there exists a constant $C_1(k, d)$ such that

$$\sum_{|\sigma|\le k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma \\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \le C_1^2 C_\eta^2 \sum_{s=1}^{k}\sum_{s'=1}^{s} (lh)^{-2s'} |\psi_{i,q}|^2_{s-s',2,S^*}. \tag{B.1}$$

Proof We re-arrange terms on the left-hand side with the same $|\sigma|$ and use the Cauchy inequality:

$$\begin{aligned} LHS &= \sum_{s=1}^{k}\sum_{|\sigma|=s} \int_{S^*} \left| \sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \binom{\sigma}{\sigma_1} D^{\sigma_1}\eta D^{\sigma-\sigma_1}\psi_{i,q} \right|^2 \\ &\le \sum_{s=1}^{k}\sum_{|\sigma|=s} \left(\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \binom{\sigma}{\sigma_1}^2 \right) \left(\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \int_{S^*} |D^{\sigma_1}\eta|^2 |D^{\sigma-\sigma_1}\psi_{i,q}|^2 \right) \\ &\le C_{1,1}^2 C_\eta^2 \sum_{s=1}^{k}\sum_{|\sigma|=s}\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \int_{S^*} (lh)^{-2|\sigma_1|} |D^{\sigma-\sigma_1}\psi_{i,q}|^2, \end{aligned} \tag{B.2}$$

where we have used $|D^{\sigma_1}\eta| \le C_\eta (lh)^{-|\sigma_1|}$ and $C_{1,1} := \max_{|\sigma|\le k} \sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \binom{\sigma}{\sigma_1}^2$. We re-arrange the terms in Eq. (B.2) by grouping terms with the same $|\sigma_1|$, and we get

$$\sum_{|\sigma|=s}\sum_{\sigma_1\le\sigma, |\sigma_1|\ge 1} \int_{S^*} (lh)^{-2|\sigma_1|} |D^{\sigma-\sigma_1}\psi_{i,q}|^2 \le \sum_{s'=1}^{s}\sum_{|\sigma_1|=s'} N(s, \sigma_1)(lh)^{-2|\sigma_1|} |D^{\sigma-\sigma_1}\psi_{i,q}|^2,$$

where $N(s,\sigma_1)=\sum_{|\sigma|=s}\sum_{\sigma_1\le\sigma,|\sigma_1|\ge 1}1$. Suppose that $N(s,\sigma_1)\le C_{1,2}$ for all $1\le s\le k$ and $1\le|\sigma_1|\le s$. Then, we have

$$\sum_{|\sigma|=s}\sum_{\sigma_1\le\sigma,|\sigma_1|\ge 1}\int_{S^*}(lh)^{-2|\sigma_1|}|D^{\sigma-\sigma_1}\psi_{i,q}|^2\le C_{1,2}\sum_{s'=1}^{s}(lh)^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}. \tag{B.3}$$

Combining Eqs. (B.2) and (B.3), and denoting $C_1=C_{1,1}C_{1,2}^{1/2}$, we have proved Eq. (B.1). □

Remark B.1 If there are no lower-order terms, we can obtain

$$\sum_{|\sigma|=k}\int_{S^*}\left|\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}\right|^2\le C_1^2C_\eta^2\sum_{s'=1}^{k}(lh)^{-2s'}|\psi_{i,q}|^2_{k-s',2,S^*}. \tag{B.4}$$

Here, we can take $C_1=C_{1,1}C_{1,2}^{1/2}$ with $C_{1,1}:=\max_{|\sigma|=k}\sum_{\sigma_1\le\sigma,|\sigma_1|\ge 1}\binom{\sigma}{\sigma_1}^2$ and $C_{1,2}=\max_{1\le|\sigma_1|\le k}N(k,\sigma_1)$. Of course, we can simply take the same C_1 as in Eq. (B.1).

Equation (B.4) is used from Eqs. (6.37) to (6.38) in the proof of Theorem 6.3.

B.2. Estimation of $\|\eta\psi_{i,q}\|_{H(S^*)}$ in the proof of Theorem 7.1

In this subsection, we will prove the following result that is used in in the proof of Theorem 7.1: for all $h>0$ such that $\frac{1-h^{2k}}{1-h^2}\le 2$, we have

$$\|\eta\psi_{i,q}\|_{H(S^*)}\le\frac{C}{2}|\psi_{i,q}|_{k,2,S^*}+\sqrt{\frac{C^2}{4}|\psi_{i,q}|^2_{k,2,S^*}+C|\psi_{i,q}|_{k,2,S^*}\|\psi_{i,q}\|_{H(S^*)}+\|\psi_{i,q}\|^2_{H(S^*)}}, \tag{B.5}$$

where $C=C_1C_\eta C_p\sqrt{2k\theta_{k,\max}}$.

Proof We begin by expressing the following integral as a sum of two terms:

$$\sum_{0\le|\sigma|,|\gamma|\le k}\int_{S^*}a_{\sigma\gamma}D^\sigma(\eta\psi_{i,q})D^\gamma(\eta\psi_{i,q})=\underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\int_{S^*}\eta a_{\sigma\gamma}(x)D^\sigma\psi_{i,q}D^\gamma(\eta\psi_{i,q})}_{I_3}$$
$$+\underbrace{\sum_{0\le|\sigma|,|\gamma|\le k}\sum_{\substack{\sigma_1+\sigma_2=\sigma\\|\sigma_1|\ge 1}}\binom{\sigma}{\sigma_1}\int_{S^*}a_{\sigma\gamma}(x)D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q}D^\gamma(\eta\psi_{i,q})}_{I_4}. \tag{B.6}$$

Repeating the same argument from Eqs. (6.48) to (6.50), we obtain

$$|I_4|\le C_1C_\eta\left(\sum_{s=1}^{k}\sum_{s'=1}^{s}h^{-2s'}|\psi_{i,q}|^2_{s-s',2,S^*}\right)^{1/2}\|\eta\psi_{i,q}\|_{H(S^*)}\sqrt{\theta_{k,\max}}. \tag{B.7}$$

Since $\psi_{i,q}\perp\mathcal{P}_{k-1}$ locally in L^2, from Eq. (6.5), we have

$$|\psi_{i,q}|_{s-s',2,S^*}\le C_ph^{s'}|\psi_{i,q}|_{s,2,S^*}.$$

Repeating the same argument from Eqs. (6.51) to (6.53), we conclude

$$\begin{aligned} I_4 &\le C_1 C_\eta C_p \sqrt{\theta_{k,\max}} \left(\sum_{s=1}^{k} \sum_{s'=1}^{s} |\psi_{i,q}|^2_{s,2,S^*} \right)^{1/2} \|\eta\psi_{i,q}\|_{H(S^*)} \qquad \text{(B.8)}\\ &\le C_1 C_\eta C_p \sqrt{\theta_{k,\max}} \left(\sum_{s=1}^{k} s|\psi_{i,q}|^2_{s,2,S^*} \right)^{1/2} \|\eta\psi_{i,q}\|_{H(S^*)} \qquad \text{(B.9)}\\ &\le C_1 C_\eta C_p \sqrt{2k\theta_{k,\max}} |\psi_{i,q}|_{k,2,S^*} \|\eta\psi_{i,q}\|_{H(S^*)}. \qquad \text{(B.10)} \end{aligned}$$

In the last inequality (6.53), we have used the polynomial approximation property (6.5) again and take $\frac{h^2-h^{2k}}{1-h^2} \le 1/C_p^2$ to make it true.

Repeating the same process for I_3, we have

$$\begin{aligned} I_3 = &\underbrace{\sum_{0\le|\sigma|,|\gamma|\le k} \int_{S^*} \eta^2 a_{\sigma\gamma}(x) D^\sigma \psi_{i,q} D^\gamma \psi_{i,q}}_{I_5} \\ &+ \underbrace{\sum_{0\le|\sigma|,|\gamma|\le k} \sum_{\substack{\sigma_1+\sigma_2=\sigma\\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} \int_{S^*} \eta a_{\sigma\gamma}(x) D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} D^\gamma \psi_{i,q}}_{I_6}. \end{aligned} \tag{B.11}$$

Here, we have exchanged the index σ and γ so that I_6 has a structure similar to that of I_4. Since $\sum_{0\le|\sigma|,|\gamma|\le k} a_{\sigma\gamma}(x) D^\sigma \psi_{i,q} D^\gamma \psi_{i,q} \ge 0$ and $|\eta(x)| \le 1$ for every $x \in D$, we obtain

$$I_5 \le \|\psi_{i,q}\|^2_{H(S^*)}. \tag{B.12}$$

Repeating the same argument from Eqs. (6.48) to (6.50) again, we obtain

$$\begin{aligned} I_6 &= \sum_{0\le|\sigma|,|\gamma|\le k} \sum_{\substack{\sigma_1+\sigma_2=\sigma\\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} \int_{S^*} a_{\sigma\gamma}(x) \eta D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} D^\gamma \psi_{i,q} \\ &\le \left(\sum_{|\sigma|\le k} \int_{S^*} \left| \sum_{\substack{\sigma_1+\sigma_2=\sigma\\ |\sigma_1|\ge 1}} \binom{\sigma}{\sigma_1} \eta D^{\sigma_1}\eta D^{\sigma_2}\psi_{i,q} \right|^2 \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)} \sqrt{\theta_{k,\max}} \\ &\le C_1 C_\eta \sqrt{\theta_{k,\max}} \left(\sum_{s=1}^{k} \sum_{s'=1}^{s} h^{-2s'} |\psi_{i,q}|^2_{s-s',2,S^*} \right)^{1/2} \|\psi_{i,q}\|_{H(S^*)}. \end{aligned} \tag{B.13}$$

The derivation of Eq. (B.13) is nearly the same as that of Eq. (B.1), and the only difference is that we need to use $|\eta D^{\sigma_1}\eta| \le C_\eta h^{-|\sigma_1|}$ (thanks to $|\eta| \le 1$) in Eq. (B.2). Using exactly the same argument from Eqs. (B.8) to (B.10), we conclude that for all $h > 0$ such that $\frac{1-h^{2k}}{1-h^2} \le 2$,

$$I_6 \le C_1 C_\eta C_p \sqrt{2k\theta_{k,\max}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{B.14}$$

Combining Eqs. (B.11), (B.12) and (B.14), we obtain

$$|I_3| \le \|\psi_{i,q}\|^2_{H(S^*)} + C_1 C_\eta C_p \sqrt{2k\theta_{k,\max}} |\psi_{i,q}|_{k,2,S^*} \|\psi_{i,q}\|_{H(S^*)}. \tag{B.15}$$

Combining Eqs. (B.6), (B.10) and (B.15), we have

$$\|\eta\psi_{i,q}\|_{H(S^*)}^2 \leq \|\psi_{i,q}\|_{H(S^*)}^2 + C_1 C_\eta C_p \sqrt{2k\theta_{k,\max}}|\psi_{i,q}|_{k,2,S^*}(\|\psi_{i,q}\|_{H(S^*)} + \|\eta\psi_{i,q}\|_{H(S^*)}). \tag{B.16}$$

Solving the above quadratic inequality, we have proved the lemma. □

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 9 December 2016 Accepted: 26 June 2017
Published online: 04 December 2017

References

1. Babuška, I., Osborn, J.E.: Generalized finite element methods: their performance and their relation to mixed methods. SIAM J. Numer. Anal. **20**(3), 510–536 (1983)
2. Babuška, I., Lipton, R.: Optimal local approximation spaces for generalized finite element methods with application to multiscale problems. Multiscale Model. Simul. **9**(1), 373–406 (2011)
3. Bachmayr, M., Cohen, A., Migliorati, G.: Representations of gaussian random fields and approximation of elliptic PDEs with lognormal coefficients (2016). arXiv preprint arXiv:1603.05559
4. Bolin, D., Lindgren, F.: Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. Ann. Appl. Stat. **5**(1), 523–550 (2011). www.jstor.org/stable/23024839
5. Chung, E., Efendiev, Y., Hou, T.-Y.: Adaptive multiscale model reduction with generalized multiscale finite element methods. JCP **320**, 69–95 (2016)
6. Ciarlet, P.G.: The Finite Element Method for Elliptic Problems, vol. 40. SIAM, Philadelphia (2002)
7. Dahlke, S., Novak, E., Sickel, W.: Optimal approximation of elliptic problems by linear and nonlinear mappings I. J. Complex. **22**(1), 29–49 (2006)
8. d'Aspremont, A., El Ghaoui, L., Jordan, M., Lanckriet, G.: A direct formulation for sparse PCA using semidefinite programming. SIAM Rev. **49**(3), 434–448 (2007)
9. D'Elia, M., Gunzburger, M.: Coarse-grid sampling interpolatory methods for approximating gaussian random fields. SIAM/ASA J. Uncertain. Quantif. **1**(1), 270–296 (2013)
10. Efendiev, Y., Galvis, J., Hou, T.Y.: Generalized multiscale finite element methods (GMsFEM). J. Comput. Phys. **251**, 116–135 (2013)
11. Efendiev, Y., Galvis, J., Wu, X.-H.: Multiscale finite element methods for high-contrast problems using local spectral basis functions. J. Comput. Phys. **230**(4), 937–955 (2011)
12. Efendiev, Y., Hou, T.Y.: Multiscale Finite Element Methods: Theory and Applications. Springer, New York (2009)
13. Gittelson, C.J.: Representation of Gaussian fields in series with independent coefficients. IMA J. Numer. Anal. **32**(1), 294–319 (2012)
14. Gneiting, T., Kleiber, W., Schlather, M.: Matérn cross-covariance functions for multivariate random fields. J. Am. Stat. Assoc. **105**(491), 1167–1177 (2010). doi:10.1198/jasa.2010.tm09420
15. Goedecker, S.: Linear scaling electronic structure methods. Rev. Mod. Phys. **71**(4), 1085 (1999)
16. Grasedyck, L., Greff, I., Sauter, S.: The AL basis for the solution of elliptic problems in heterogeneous media. Multiscale Model. Simul. **10**(1), 245–258 (2012)
17. Guttorp, P., Gneiting, T.: Studies in the history of probability and statistics xlix on the matern correlation family. Biometrika **93**(4), 989–995 (2006)
18. Henning, P., Peterseim, D.: Oversampling for the multiscale finite element method. Multiscale Model. Simul. **11**(4), 1149–1175 (2013)
19. Höllig, K., Apprich, C., Streit, A.: Introduction to the web-method and its applications. Adv. Comput. Math. **23**(1), 215–237 (2005)
20. Hou, T.Y., Li, Q., Zhang, P.: A sparse decomposition of low rank symmetric positive semidefinite matrices. Multiscale Model. Simul. **15**(1), 410–444 (2017)
21. Hou, T.Y., Liu, P.: Optimal local multi-scale basis functions for linear elliptic equations with rough coefficients. Discrete Contin. Dyn. Syst. A **36**(8), 4451–4476 (2016)
22. Hou, T.Y., Wu, X.-H.: A multiscale finite element method for elliptic problems in composite materials and porous media. J. Comput. Phys. **134**(1), 169–189 (1997)
23. Hou, T.Y., Wu, X.-H., Zhang, Y.: Removing the cell resonance error in the multiscale finite element method via a Petrov–Galerkin formulation. Commun. Math. Sci. **2**(2), 185–205 (2004)
24. Hughes, T.J., Feijóo, G.R., Mazzei, L., Quincy, J.-B.: The variational multiscale methoda paradigm for computational mechanics. Comput. Methods Appl. Mech. Eng. **166**(1), 3–24 (1998)
25. Jolliffe, I.T., Trendafilov, N.T., Uddin, M.: A modified principal component technique based on the LASSO. J. Comput. Graph. Stat. **12**(3), 531–547 (2003)
26. Lai, R., Lu, J., Osher, S.: Density matrix minimization with L_1 regularization. Commun. Math. Sci. **13**(8), 2097–2117 (2015)
27. Lindgren, F., Rue, H., Lindström, J.: An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **73**(4), 423–498 (2011)

28. Målqvist, A., Peterseim, D.: Localization of elliptic multiscale problems. Math. Comput. **83**(290), 2583–2603 (2014)
29. Marzari, N., Mostofi, A.A., Yates, J.R., Souza, I., Vanderbilt, D.: Maximally localized Wannier functions: theory and applications. Rev. Mod. Phys. **84**, 1419–1475 (2012)
30. Marzari, N., Vanderbilt, D.: Maximally localized generalized Wannier functions for composite energy bands. Phys. Rev. B **56**, 12847–12865 (1997)
31. Matérn, B.: Spatial Variation, vol. 36. Springer, New York (2013)
32. Melenk, J.: On n-widths for elliptic problems. J. Math. Anal. Appl. **247**(1), 272–289 (2000)
33. Ming, P., Yue, X.: Numerical methods for multiscale elliptic problems. J. Comput. Phys. **214**(1), 421–445 (2006)
34. Nikol'skii, S.M.: Imbedding Theorems for Different Metrics and Dimensions. Springer, Berlin (1975)
35. Owhadi, H.: Bayesian numerical homogenization. Multiscale Model. Simul. **13**(3), 812–828 (2015)
36. Owhadi, H.: Multigrid with rough coefficients and multiresolution operator decomposition from hierarchical information games. SIAM Rev. **59**(1), 99–149 (2017)
37. Owhadi, H., Scovel, C.: Universal scalable robust solvers from computational information games and fast eigenspace adapted multiresolution analysis (2017). arXiv preprint arXiv:1703.10761
38. Owhadi, H., Zhang, L.: Gamblets for opening the complexity-bottleneck of implicit schemes for hyperbolic and parabolic ODEs/PDEs with rough coefficients (2016). arXiv:1606.07686v1
39. Owhadi, H., Zhang, L., Berlyand, L.: Polyharmonic homogenization, rough polyharmonic splines and sparse super-localization. ESAIM Math. Model. Numer. Anal. **48**(02), 517–552 (2014)
40. Ozoliņš, V., Lai, R., Caflisch, R., Osher, S.: Compressed modes for variational problems in mathematics and physics. Proc. Nat. Acad. Sci. **110**(46), 18368–18373 (2013)
41. Peterseim, D.: Variational multiscale stabilization and the exponential decay of fine-scale correctors. In: Building Bridges: Connections and Challenges in Modern Approaches to Numerical Partial Differential Equations, pp. 343–369. Springer, Cham (2016)
42. Renardy, M., Rogers, R.C.: An Introduction to Partial Differential Equations, vol. 13. Springer, New York (2006)
43. Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (2012)
44. Strouboulis, T., Copps, K., Babuška, I.: The generalized finite element method. Comput. Methods Appl. Mech. Eng. **190**(32), 4081–4193 (2001)
45. Vu, V.Q., Cho, J., Lei, J., Rohe, K.: Fantope projection and selection: A near-optimal convex relaxation of sparse PCA. In: Burges, C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K. (eds.) Advances in Neural Information Processing Systems, vol. 26, pp. 2670–2678. (2013)
46. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**, 515–534 (2009)
47. E, W., Li, T., Lu, J.: Localized bases of eigensubspaces and operator compression. Proc. Natl. Acad. Sci. **107**(4), 1273–1278 (2010)
48. Zhang, P.: Compressing Positive Semidefinite Operators with Sparse/Localized Bases. Ph.D. thesis, California Institute of Technology (2017)
49. Zou, H., Hastie, T., Tibshirani, R.: Sparse principal component analysis. J. Comput. Graph. Stat. **15**(2), 265–286 (2006)